Abstract

BOYER, JOSEPH G. Topics Involving the Gamma Distribution: The Normal Coefficient of Variation and Conditional Monte Carlo. (Under the direction of William H. Swallow).

A transformation of the sample coefficient of variation \( CV \) for normal data is shown to be nearly proportional to a \( \chi^2 \) random variable. The associated density is applied to inference on the common \( CV \) of \( k \) populations, testing \( CV \) homogeneity across populations, and confidence intervals for the ratio of two \( CVs \). The resulting tests and confidence intervals are shown via theory and simulation to be valid and powerful.

In other work on the coefficient of variation, a sample of scientific abstracts is used to characterize the values of the \( CV \) encountered in practice, point estimation for a common \( CV \) in normal populations is studied, and the literature on testing \( CV \) homogeneity in normal populations is reviewed.

There is very little literature on the problem of conducting inference in models for continuous data conditional on sufficient statistics for nuisance parameters. This thesis explores Monte Carlo approaches to conditional \( p \)-value calculation in such models, including Dirichlet data generation, importance sampling, Markov chain Monte Carlo, and a method related to fiducial inference. Importance sampling is used to create a conditional test of \( CV \) homogeneity in normal populations using the \( \chi^2 \) approximation mentioned above. A Markov chain Monte Carlo solution is given to the long-standing problem of testing the homogeneity of exponential populations subject to Type I censoring. Conditional Monte Carlo algorithms are also applied to testing for an effect of a factor in an experiment with exponential data, testing for a dispersion effect in a replicated experiment with normal data, and testing a null value of a coefficient in exponential regression with an inverse link; brief consideration is also given to the problem of testing the homogeneity of \( k \) gamma distributions.
Topics Involving the Gamma Distribution: the Normal Coefficient of Variation and Conditional Monte Carlo

by

Joseph G. Boyer

A dissertation submitted to the graduate faculty of North Carolina State University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Statistics

Raleigh, NC
January 18, 2007

Approved by:

Dr. William H. Swallow (Chair)  Dr. Thomas Gerig

Dr. Dennis D. Boos  Dr. Cavell Brownie
Dedication

To Lori, for making me finish. To Ellen, my kindred spirit. We continue to overcome. And to Monahan’s prisoners, all current and future, for the fight that cannot be won, but not being fought cannot be lost, for the lock that cannot be broken but is not fastened, and for the journey that no perseverance can ever end, but that can be simply ended, by the power within you that courage shall unleash to turn the hand upon the doorknob, open to an anticlimactic day.
Biography

Joe was born in New Orleans, LA in 1969. He has lived a scandalously interesting but unfortunately secretive life.
I would like to thank first of all my advisor, Bill Swallow. Without his experience, guidance, encouragement, and direction of my efforts, this project would have meandered into a much less fruitful conclusion. Second, I would like to thank my committee members Dennis Boos, Cavell Brownie, and Tom Gerig for giving their attention to my research, resulting in beneficial feedback.

I don’t know what I would have done without the conscientious computational support of Terry Byron. On this topic, I owe a dinner to Jimmy Doi for going far out of his way to help me learn software, even supplying me with manuals. I say this in the full knowledge that Jimmy is safely a continent away from claiming that dinner. For assistance in helping me through the NC State bureaucracy, I thank Adrian Blue, whose job I did not make easier.

I must acknowledge the skilled listening of Mary Huelsbeck, who helped me to figure out how to catch a whale by digging in the desert.

I should give thanks here for the animated cakes supplied to the department by Justin Shows and to Mike “Sox in 7” Rubitski and Jeff “Yankees in 6” Desimone for keeping me abreast of all sports information while my attention was temporarily diverted to this thesis. I would also like to thank Alvin “Van” Orden for helping to create the office Olympics, “The Sandpiper”, and the psychic energy that brought about national championships for my Texas Longhorns and LSU Tigers.

Finally, I am very thankful for the comradery of all of my colleagues from whom I have learned and drawn support, especially Clay “Off the Sliver” Barker, Karen “All Fortune Cookies Are True” Chiswell, Jimmy “Chic” Cho, Hugh “Level of Difficulty” Crews, Michael “Straight Shot” Crotty, Venita “Back Rub Angel” Depuy, Kirsten “St. Jude’s Protege” Doehler, Kristen “Madsen” Foley, Ross “Am I the Only Mature One Around Here” Gosky, Amanda “4 Years” Hepler, Shenek “Secretly Republican” Heyward, Emily Hohohohmeister, Marti “Legs” Jones, Hye “Young” Lee, Shiufang “Claudia” Liu, Amy “Rather Be a Hammer” Nail, and Lavanya “Ramanan” Noodle.
2.1.3 The Distribution of \( \hat{CV} \) ........................................ 50
2.1.4 The bias of \( \hat{CV} \) ........................................ 51

2.2 Approximations to the Distribution of \( \hat{CV} \) ................. 53
2.2.1 \( \chi^2 \) approximations ........................................ 53
2.2.2 Delta-method approximations ..................................... 55
2.2.3 Numerical evaluation of the approximations ................. 56
2.2.4 An exponential family model for inference on \( CV \) ......... 57
2.2.5 Asymptotic comparison of approximations .................... 60

2.3 Inference on a common \( CV \) .................................. 61
2.3.1 Notation ....................................................... 63
2.3.2 Previous literature ............................................. 63
2.3.3 Inference based on the normal approximation ............... 65
2.3.4 Inference based on the \( \chi^2 \) approximation ............... 68
2.3.5 Simulation comparison of two-sided confidence intervals for \( CV \) ...................................................... 73

2.4 Testing the Equality of the Coefficients of Variation of \( k \) Normal Populations .................................................... 77
2.4.1 Notation ....................................................... 78
2.4.2 Previous literature ............................................. 78
2.4.3 Applying exponential family model to testing \( CV \) equality ... 81
2.4.4 Simulation results for equal sample sizes ..................... 86
2.4.5 One-sided tests for \( k = 2 \) ..................................... 89

2.5 Confidence Intervals for Differences Between Two Coefficients of Variation ........................................... 90
2.5.1 Previous literature ............................................. 91
2.5.2 Confidence Interval for the ratio of two \( CVs \) ............... 92
2.5.3 Simulation results ............................................. 92

2.6 The Normality Assumption ........................................ 94
2.6.1 Robustness of inference that assumes normality ............. 94
2.6.2 The relevance of normal-theory inference ..................... 96

3 Additional Work on the Coefficient of Variation in Normal Populations .......................................................... 97
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Sample of coefficients of variation from the scientific literature</td>
<td>97</td>
</tr>
<tr>
<td>3.2</td>
<td>Point estimation of a common coefficient of variation in normal samples</td>
<td>101</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Estimators</td>
<td>101</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Theoretical comparison of estimators</td>
<td>105</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Simulation comparison of point estimators</td>
<td>109</td>
</tr>
<tr>
<td>3.3</td>
<td>Literature Review on Test of CV Homogeneity in Normal Populations</td>
<td>112</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Tests based on the likelihood of the full data</td>
<td>114</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Tests based on the likelihood of the sample CVs.</td>
<td>117</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Tests based on the delta method approximation</td>
<td>118</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Tests based on ( \frac{1}{CV} )</td>
<td>120</td>
</tr>
<tr>
<td>3.3.5</td>
<td>Evaluating existing tests</td>
<td>121</td>
</tr>
<tr>
<td>3.4</td>
<td>Using a stochastic representation to obtain confidence intervals</td>
<td>124</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Exact confidence interval for common CV</td>
<td>125</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Monte Carlo approximate intervals for CV (_2 - CV_1 )</td>
<td>126</td>
</tr>
<tr>
<td>3.5</td>
<td>Convenient Inference on a Common CV Using the ( \chi^2 ) Approximation</td>
<td>127</td>
</tr>
<tr>
<td>4</td>
<td>Monte Carlo Conditional ( p )-value Calculation for Continuous Data</td>
<td>130</td>
</tr>
<tr>
<td>4.1</td>
<td>Reduction to Data Generation on ( \mathbb{R}^{N-d_{np}} )</td>
<td>132</td>
</tr>
<tr>
<td>4.2</td>
<td>Assessment of the literature</td>
<td>135</td>
</tr>
<tr>
<td>4.3</td>
<td>Special case: gamma distribution, known shape parameter – Dirichlet data generation</td>
<td>137</td>
</tr>
<tr>
<td>4.4</td>
<td>Importance Sampling</td>
<td>143</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Importance sampling with linear sufficient statistics: generating data on a hyperplane</td>
<td>146</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Implementing importance sampling for testing CV equality</td>
<td>150</td>
</tr>
<tr>
<td>4.4.3</td>
<td>Simulation results for testing CV equality, unequal sample sizes</td>
<td>151</td>
</tr>
<tr>
<td>4.4.4</td>
<td>Implementing importance sampling for testing the scale parameter of a gamma distribution</td>
<td>156</td>
</tr>
<tr>
<td>4.5</td>
<td>Gibbs sampling</td>
<td>160</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Implementation of Gibbs sampling with linear sufficient statistics</td>
<td>163</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Application: exponential regression with inverse link</td>
<td>165</td>
</tr>
</tbody>
</table>
4.5.3 Simulation evaluation of Gibbs sampling for inverse link exponential regression ........................................ 170
4.5.4 Application: comparing exponential populations with Type I censoring ............................................. 173

4.6 Future research: approaches for nonlinear sufficient statistics .......................................................... 180
4.6.1 Fiducial Monte Carlo ................................................. 180
4.6.2 Importance Sampling ................................................. 181
4.6.3 Gibbs sampling ..................................................... 185

Literature Cited ......................................................... 193
List of Tables

2.1 Bias of \( \hat{CV} \) as percentage of true \( CV \) ........................................ 52
2.2 Exact cdf minus approximate cdf, \( CV = 0.33, N = 10 \) ...................... 57
2.3 Coverage probability (\( CP \)) and width (\( W \)) of confidence intervals for
\( CV = 0.05 \). \( SE = 0.002 \) for \( CP \), < 0.001 for width. ....................... 74
2.4 Coverage probability (\( CP \)) and width (\( W \)) of confidence intervals for
\( CV = 0.33 \). \( SE = 0.002 \) for \( CP \), < 0.01 for width. ....................... 75
2.5 Size of two-sided tests of \( CV \) homogeneity. (\( SE = 0.002 \)) ............... 87
2.6 Power of two-sided tests of \( CV \) homogeneity. (\( SE = 0.005 \)) ............ 88
2.7 Size of one-sided tests of \( CV \) homogeneity, \( k = 2 \). (\( SE = 0.002 \)) ....... 89
2.8 Power of one-sided tests of \( CV \) homogeneity, \( k = 2 \). (\( SE = 0.005 \)) ....... 89
2.9 Coverage probability and average width of two-sided confidence intervals.
\( SE \) for coverage = 0.001, for width < 0.001. ......................... 93
2.10 Ratio of confidence interval length assuming normality to valid length,
\( CV = 0.05 \) .................................................. 95
2.11 Ratio of confidence interval length assuming normality to valid length,
\( CV = 0.33 \) .................................................. 95

3.1 Articles in random sample – coefficients of variation keyword ............ 98
3.2 Articles in random sample – coefficient of variation keyword ............ 99
3.3 Articles in random sample – relative standard deviation keyword ....... 100
3.4 Bias in point estimate as percentage of common \( CV \), \( CV = 0.05 \) ........ 111
3.5 Bias in point estimate as percentage of common \( CV \), \( CV = 0.40 \) ........ 112
3.6 Bias in point estimate as percentage of common \( CV \), sample sizes not
identical ................................................. 113
3.7 Root MSE as percentage of common \( CV \), \( CV = 0.05 \) ...................... 114
3.8 Root MSE as percentage of common \( CV \), \( CV = 0.40 \) ...................... 115
3.9 Root MSE as percentage of common \( CV \) ...................................... 116
4.1 Size of Bartlett-Kendall test for marginal dispersion effect. \((SE = 0.002)\) 142
4.2 Size of two-sided tests of \(CV\) homogeneity ................................. 152
4.3 Power of two-sided tests of \(CV\) homogeneity ............................... 153
4.4 Properties of importance sampling test ....................................... 154
4.5 Size of test of \(CV_1 = CV_2\) versus \(CV_2 > CV_1\). \((SE = 0.002)\) ....... 155
4.6 Power of one-sided tests of \(CV\) homogeneity, \(k = 2\). \((SE = 0.005)\) ... 156
4.7 Performance of ELS to test \(\beta = \beta_o\). \((SE\ of\ size: 0.002)\) ............. 158
4.8 Type I error of one-sided test of no effect of continuous predictor ......... 170
4.9 Power of Gibbs sampling test with and without burn-in. \((SE = 0.025)\) 171
4.10 Autocorrelations for \(Ind(z(i))\) .................................................. 172
4.11 Size of LR test for equality of scale parameters under Type I censoring.
\((SE = 0.002)\) ................................................................. 176
4.12 Power of \(MCMC\) test for equality of scale parameters under Type I
censoring. \((SE = 0.02)\) ................................................................. 177
List of Figures

2.1 Histogram of CVs from the Scientific Literature .......................... 49
2.2 Comparison of exact cdf, $\Phi_V$, and $\Phi_Z$ ............................. 58
2.3 Comparison of exact cdf, $\Phi_V$, and $\Phi_Z$ ............................. 59
4.1 A one-dimensional surface in $\mathbb{R}^2$ ................................. 190
Exponential families (Definition 1.3.4.1) are convenient probability models because the theory of exponential families allows us to easily find similar, unbiased, and uniformly most powerful inferential procedures (Definitions 1.3.3.1 - 1.3.3.3). This thesis is concerned with two inquiries, one which applies the theory of exponential families to an inferential problem, another which seeks to make the theory of exponential families more practically useful. The gamma distribution (Definition 1.3.5.10), an example of an exponential family, is another thread that unifies the inquiries. The gamma distribution has a number of applications in statistics [1]. It’s simplicity and flexibility make it a convenient model for right-skewed nonnegative data, and it has become an important model for life-testing experiments. It arises theoretically in several scenarios: the sample variance from a normal sample is gamma; the exponential distribution, a lifetime distribution with a constant hazard rate, is gamma; and the gamma distribution has been suggested as an appropriate model for survival time in a system with continuous maintenance [1].

1.1 Chapters

1.1.1 Chapter 1

Sections 2 and 3 serve as a reference to the reader. Section 2 explains the notation used, while Section 3 contains a list of definitions and theorems which I shall cite throughout the dissertation.

1.1.2 Chapter 2

The coefficient of variation, \( \frac{\sigma}{\mu} \), measures variability in a way that is invariant to changes of scale. Chapter 2 studies inference on the coefficient of variation (CV) in normal
Chapter 1. Introduction

populations, which is based on the sample $CV$, $S/\bar{X} \equiv \hat{CV}$.

Chapter 2 discovers that a transformation of the sample $CV$ for a normal population has an approximate gamma density and that the approximation is nearly exact. This allows the development of convenient, valid, and theoretically powerful approaches to three problems: inference on the common $CV$ of $k$ populations, constructing confidence intervals for the ratio of two $CV$s, and testing $CV$ homogeneity – the assumption that the $CV$s of $k$ populations are the same.

As regards inference on the common $CV$, tests are developed that are uniformly most powerful (Definition 1.3.3.3) taking the approximating density as exact, and the associated confidence intervals are shown via simulation to be shorter than those of the existing fiducial approach and to provide confidence levels that are closer to the nominal level than those of the existing approach. If the sizes of the different samples are equal, the confidence limits based on the $UMP$ test are analytical functions of $\chi^2$ percentiles; a Monte Carlo algorithm is provided to calculate the limits with unequal sample sizes.

The confidence interval developed for ratios of $CV$s allows the assessment of bioequivalence of two drug formulations to include consideration of the $CV$ of quantities such as the $AUC$. The interval is shown via simulation to have coverage close to nominal, although they are slightly conservative if the sample sizes are unequal. The confidence limits are analytical functions of $F$ distribution percentiles.

Taking the approximating density to be exact, Chapter 2 obtains four useful results for testing $CV$ homogeneity:

- An existing test known as the modified Bennett ($MB$) test is unbiased if the sample sizes are equal.
- If the sample sizes are equal, the $MB$ test for $k = 2$ is $UMP$ unbiased among the class of tests that are invariant to the sample means and the common $CV$.
- The chapter provides a Monte Carlo algorithm to obtain accurate $p$-values for the $MB$ test with equal sample sizes.
- Chapter 2 derives the $UMP$ one-sided similar test for $k = 2$ with potentially different sample sizes.
Chapter 1. Introduction

These results on testing CV homogeneity under normality bring theoretical closure to that literature, which contains over a dozen papers starting in 1976.

Chapter 2 also explores inference based on the delta-method approximation to the distribution of the sample CV. It develops delta-based confidence intervals for a common CV and proves the asymptotic legitimacy of existing delta-based approaches to assess differences among CVs. Simulations show that the delta-based confidence interval for the common CV is a reasonable alternative to the UMP interval and can be used to simplify calculations for unequal sample sizes. Delta-based inference concerning differences among CVs is shown via simulations to possess competitive power and to be accurate even for small sample sizes, though confidence intervals for differences among CVs can have low coverage if sample sizes are unequal.

Asymptotic calculations confirm that like inference on the variance, inference on the coefficient of variation is not robust to violations of the normality assumption. For skewed populations, inference based on normal distribution theory will be slightly conservative, while excess kurtosis can make normal-theory inference quite liberal.

1.1.3 Chapter 3

Chapter 3 presents additional work related to inference on the coefficient of variation. It reports on a sample of the values of coefficients of variation drawn from scientific abstracts. The sample provides justification for viewing $0 < CV < 0.33$ as the “practical range” for values of CV; a very high percentage of the CVs in the sample are in this range.

The chapter conducts a study of point estimation for a common coefficient of variation in normal populations. It derives a bias-corrected weighted average of sample CVs with variance-minimizing weights, the maximum-likelihood estimator (MLE) of the common CV from the approximate marginal density of the sample CVs, and a nearly-exact analytical approximation for the MLE from the full likelihood. It compares the theoretical bias, variance, and consistency of these estimators with an existing one, and also compares the estimators via simulation. All of the estimators considered are reasonably effective except for the full-likelihood MLE with small sample sizes. The MLE from the marginal likelihood has the lowest mean squared error in simulations.
Chapter 1. Introduction

Another contribution in Chapter 3 is a review of the literature on testing CV homogeneity across normal samples. The chapter argues that, of the dozen tests that have appeared in the literature, two stand out as superior to the others. These are the MB test and the delta-based test discussed above.

Finally, Chapter 3 explains how to use a stochastic representation for the sample CV from a normal population in Monte Carlo calculations to obtain exact confidence intervals for a common CV and approximate confidence intervals for the difference between CVs. These methods are not competitive with the methods in Chapter 2; they are reported simply to provide an example of how one might use Monte Carlo methods to create confidence intervals.

1.1.4 Chapter 4

Chapter 2 solves the problem of getting accurate p-values for the MB test of CV homogeneity only for the equal-sample size case. If sample sizes are not equal, a problem arises: the common CV becomes a nuisance parameter; it does not vanish from the distribution of the MB statistic.

In principle, one can always deal with nuisance parameters in exponential families by conditioning them away. That is, one can find a statistic that is sufficient for the nuisance parameter, and then calculate the p-value for a test statistic conditional on the value of the sufficient statistic. By Definition 1.3.1.1, this p-value will be invariant to the nuisance parameter, and the resulting test will be similar (Definition 1.3.3.1).

This is an effective way to deal with nuisance parameters as long as one can calculate the conditional p-value. But this can be a difficult problem. If the nuisance parameter vector is of dimension $d_{np}$ and there are $N$ observations in the dataset, the conditional support of the data will typically be an $N - d_{np}$-dimensional surface, perhaps oddly shaped, in $\mathbb{R}^N$. Calculating the p-value is equivalent to integrating a function over this surface, which is typically a difficult problem.

Chapter 4 explores Monte Carlo approaches to this problem for continuous data (see Definition 1.3.3.7). Monte Carlo is nontrivial here for the same reason that numerical integration is difficult; the unconditional density assigns the conditional support measure 0, so rejection sampling (Definition 1.3.5.15) will not work. Chapter 4 does
explain how one can turn the problem into one of generating data on a subset of $\mathbb{R}^{N-d_{np}}$ so that one does not have to deal with a support that has lower dimension than the data vector. But this creates additional problems, so that neither Monte Carlo nor numerical integration is straightforward in the transformed problem.

While a lot of attention has been paid to Monte Carlo conditional $p$-value estimation in certain models for discrete data – logistic regression, log-linear models, and contingency tables – very little has been written on this problem concerning continuous data. There are no review articles, and while tools developed for other purposes can be applied, their application is not straightforward, and one must dig them out from scattered sources. Chapter 4 fulfills the useful purposes of listing available options, explaining how to implement them, discussing some of their strengths and weaknesses, deriving some necessary formulas (including the standard errors of the $p$-value estimates), suggesting solutions to some problems that arise in implementation, and evaluating the options with simulations.

The Monte Carlo approaches discussed in Chapter 4 include Dirichlet data generation for the special case of gamma distributions with known shape parameter, importance sampling (Definition 1.3.9.1), Markov Chain Monte Carlo (MCMC), and a method that I shall call “fiducial Monte Carlo” due to its relationship to fiducial inference (Definition 1.3.3.5). Except for fiducial Monte Carlo, these methods were developed for problems other than the one considered in Chapter 4 and must be adapted to that particular problem.

The conditional distribution of gamma data given a value for $T_\alpha$ (see Definition 1.3.5.10) is Dirichlet (Definition 1.3.5.4), allowing convenient Monte Carlo $p$-value calculation via Theorem 1.3.5.4. For a couple of problems concerning gamma data with known shape, the Monte Carlo option has been overlooked.

The strategy behind importance sampling is to generate from a convenient generating density with the same support as the target density and then weight the data by the ratio of the target density to the generating density; the weights serve a purpose much like Census weights correcting for undercounting of certain populations. (In our case the target density is that of the data conditional on the observed values of the sufficient statistics.) To implement importance sampling one needs a generating
density (Definition 1.3.9.1). Chapter 4 suggests four ways to come up with one. If
the sufficient statistic is linear and one-dimensional, one can use the Dirichlet distribution. As has been pointed out in the discrete-data literature, the multivariate normal distribution can serve as a generating density for the case where the sufficient statistics are linear and multidimensional, because it is well-known how to generate normal data subject to a linear constraint. Chapter 4 develops a third way, called estimated likelihood sampling (ELS), that can be used for both linear and nonlinear sufficient statistics; its generating density is essentially the unconditional likelihood with the nuisance parameters set equal to their maximum likelihood values. ELS turns out to be an implementation of a method called “conditional Monte Carlo” (CMC) that has been described in general terms in the literature. To evaluate the generating density for CMC, one needs to evaluate the Jacobian of a transformation between $\mathbb{R}^{N-d_{op}}$ and a surface in $\mathbb{R}^N$ (see Theorem 1.3.5.1); Chapter 4 provides a formula for calculating Jacobians of such functions.

The Gibbs sampler (Definition 1.3.7.2) is a Markov chain (Definition 1.3.7.1) which in principle can produce data that resembles a random sample from the target density. To generate the $j$th element of the data vector in the $i$th step of the Gibbs sampler, one generates a data point from the jumping density. Chapter 4 derives the jumping density when the target is the density of the data conditional on a linear sufficient statistic and discusses the evaluation of the jumping density for the case of nonlinear sufficient statistics, where one may not be able to obtain it analytically. If such is the case, Gibbs sampling cannot be implemented, but one can still construct a Markov chain to provide an approximate random sample from the target density using the Metropolis-within-Gibbs algorithm (Definition 1.3.7.9). However, such an algorithm may be highly computationally intensive.

Fiducial Monte Carlo is related to fiducial inference, itself a Monte Carlo method for unconditional inference (Definition 1.3.3.5). Fiducial Monte Carlo is a promising strategy, but it does not generate from the exact conditional distribution in general, and it has potentially prohibitive computational challenges.

One important benefit of these Monte Carlo methods is that typically the resulting $p$-value estimate theoretically converges almost surely to the true $p$-value; the estimated
Chapter 1. Introduction

$p$-value can be made as close as desired to exact by taking a large enough number of draws. However, the number of draws necessary for acceptable precision may be impractically large, as demonstrated in some of the examples below. With importance sampling, there are two key problems: the support of the target density may be only a small subset of the support of the generating density, and the generating density may be a poor match for the target. With Gibbs sampling, the correlation across steps may be so high that it takes too many steps to approximate a random sample.

Chapter 4 illustrates methods for linear sufficient statistics by applying them to several problems involving the gamma distribution. The chapter explains how Dirichlet data generation can be applied to the problem of testing for a factor effect with exponential experimental data and to testing for dispersion effects in normal data from experiments with replication. Simulations suggest that the standard test for dispersion effects with replicated normal experimental data, which relies on the approximate normality of the log sample variance, is liberal.

Importance sampling with a normal generating density is used to solve the problem that motivated Chapter 4, that of executing a similar test of CV homogeneity with normal samples of differing size. The approximate gamma density from Chapter 2 is used to identify a sufficient statistic for the common CV.

ELS is used to calculate precise similar $p$-values for testing a null value for the scale parameter of a gamma distribution with unknown shape. A method for doing this exists in the literature, but it requires a complicated algorithm to evaluate messy integrals.

Gibbs sampling is applied to a long-standing problem in life testing: testing the equality of the scale parameter in exponential samples subject to Type I censoring. For testing against a general alternative, the current test relies on the asymptotic distribution of the likelihood ratio statistic (Theorem 1.3.6.13). Simulations indicate that the $p$-value calculated via Gibbs sampling is slightly more accurate for the general test; Gibbs sampling also allows straightforward calculation of a $p$-value for any specific test. Chapter 4 also applies Gibbs sampling to exponential regression with an inverse link function, where it can provide accurate similar $p$-values for testing a null value for any coefficient vector of interest. Currently, exact testing of a coefficient in this model
Chapter 1. Introduction

requires there to be no other predictors in the model.

In all these applications, the Monte Carlo methods are able to give precise estimates of the true conditional \( p \)-values in a reasonable number of steps, although for some applications a bound on sample size is identified beyond which the Monte Carlo method chosen would not be practical. The tests conducted by utilizing Monte Carlo conditional \( p \)-values have actual size very close to nominal size in simulations. Another benefit of the Monte Carlo methods is that in all of the examples above, they can be used to calculate a similar \( p \)-value for any goodness-of-fit statistic, allowing one to test the validity of the \textit{gamma} model used. Currently, only asymptotic goodness-of-fit tests of the \textit{gamma} distribution exist in the literature.

The chapter stops short of applying Monte Carlo methods to examples with nonlinear statistics, but does derive the kernel of the jumping density for Gibbs sampling from a \textit{gamma} distribution conditional on \( T_\alpha \) and \( T_\beta \) (Definition 1.3.5.10). The fact that the target density may not be available analytically if the sufficient statistics are nonlinear creates computational challenges for both importance sampling and \textit{MCMC} and weakens the key selling point of the latter – that it can be done without the need to match a generating density to a target.

Applications and theoretical musings in Chapter 4 lead to the following tentative conclusions:

• Gibbs sampling appears to be effective in that the \( p \)-value can apparently converge in a practical number of steps even for large sample sizes.

• Two aspects of this application of Gibbs sampling increase its effectiveness. First, under the null hypothesis the Gibbs sampler starts out in its stationary distribution. Second, if the data are from a random sample under the null, one can randomly reorder the elements of the vectors drawn in the steps of the Gibbs sampler, reducing correlation across steps.

• Importance sampling is subject to the curse of dimensionality – lack of a match between the marginals of the generating and target densities becomes a greater and greater problem as the sample size increases. For Gibbs sampling, whether correlation across steps increases or decreases with sample size is case dependent.
Chapter 1. Introduction

- For linear sufficient statistics, the estimated likelihood generating density produces a better match for importance sampling than does the normal generating density, one that may not deteriorate as the sample size grows. The match between the marginals of the ELS generating density and the target density improves with sample size, potentially offsetting the curse of dimensionality.

- ELS appears to be effective for large samples, but no comparison to Gibbs sampling can be made because the sample size in the application of ELS was smaller than in the applications of Gibbs.

- ELS will often be easier to implement than Gibbs sampling because it will usually be easier to draw from the estimated likelihood generating density for ELS than from the jumping density for Gibbs.

### 1.2 Notation and Conventions

Generic random variables will be denoted by upper case letters. Realizations of a random variable from its support will be denoted by lowercase letters. If the random variable is multivariate, $a$ will represent its dimension.

Vectors (including vector-valued functions) and matrices will be denoted in bold-face, while scalars will be denoted in plain text.

$X$ will represent a data vector $\{X_1, \ldots, X_N\}$ of independent random variables, and $X_i$ will refer to an element from a data vector. If the data are identically distributed, $X$ will be referred to as a “random sample from the density $f_X$”. The population mean will be denoted $\mu$, and the variance and third and fourth moments of a univariate random sample will be denoted $\sigma^2$, $\mu_3$, and $\mu_4$ respectively. $W$ will represent a matrix of predictor variables, either design points or covariates or both. $W_{ij}$ will be the value of the $j$th predictor variable for observation $i$.

If the data are a random sample, $\bar{X}$ will refer to the sample mean, $S^2$ will refer to the sample variance $\frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N-1}$ of a univariate random sample, $CV$ will refer to the population coefficient of variation $\frac{\sigma}{\mu}$ for a univariate random sample, and $\hat{CV}$ to the sample coefficient of variation $S/\bar{X}$ for a univariate random sample.
Chapter 1. Introduction

\( \theta \) will denote the \( d \times 1 \) parameter vector of a density and consists of two subvectors \( \theta_{pi} \) \( (d - d_{np} \times 1) \) and \( \theta_{np} \) \( (d_{np} \times 1) \). (The latter can be thought of as nuisance parameters, and the former can be thought of as parameters of interest.) \( \hat{\theta}_{MLE} \) represents the maximum likelihood estimate of \( \theta \).

Taking the random variable \( Z \) as an example, here is how I shall denote various concepts. \( z \) represents a realized value from the support, \( f_Z(\cdot; \theta) \) is the parameterized density function, \( F_Z(\cdot; \theta) \) is the cumulative distribution function, \( f_{Z|Y}(\cdot|y; \theta) \) is the density of \( Z \) conditional on \( Y = y \), \( E_Z(\cdot) \) is the expectation of the function \( h \) taken with respect to \( Z \), \( E_{Z|Y}(g(Z)|y) \) is the expectation of \( g(Z) \) conditional on \( Y = y \), \( \text{Var}(Z) \) indicates the variance of \( Z \), \( \text{Var}_{Z|Y}(Z|y) \) is the variance of \( Z \) conditional on \( Y = y \), and the expression \( \int_A h(z)dz \) indicates the integration of the function \( h \) over the range \( A \) in the support of \( Z \). \( \text{Prob}_{\theta^*}(B) \) can be interpreted as the probability, if the true parameter vector is \( \theta^* \), that the event \( B \) occurs. \( \text{Prob}_{\theta^*}(B|G) \) indicates the conditional probability of the event \( B \) given the event \( G \). \( I_{B}(z) \) is equal to 1 if \( z \) is in the event \( B \) and 0 otherwise.

For any vector, a subscript simply indicates an element of the vector. For instance, \( \theta_i \) represents the \( i \)th element of \( \theta \), \( \theta_{np,i} \) represents the \( i \)th element of \( \theta_{np} \), and \( X_{ij} \) represents the \( j \)th element of \( X_i \). For a matrix, a subscript indicates a column of the matrix. For instance, \( W_i \) is the \( i \)th column of \( W \). The notation \( Y[r:p] \) will indicate the vector formed by the \( r \)th through \( p \)th elements of the vector \( Y \), and \( W[r:p] \) will indicate the matrix formed by the \( r \)th through \( p \)th columns of the matrix \( W \). If I want to describe a \( p \times 1 \) vector \( Y \) by reference to its individual elements, I shall denote the vector \( \{Y_1, \ldots, Y_p\} \). Similarly, \( \{Y, Z\} \) is the vector formed by concatenating the vectors \( Y \) and \( Z \).

Functions of a data vector denoted by \( T, T_1, \) etc will be understood to be statistics. A lowercase \( t \) will denote the observed value of the statistic.

In discussing hypothesis tests, rejection regions will have forms similar to \( T > b \). In such expressions, \( b \), adorned by various stars or subscripts, will be understood to be a constant that is determined by the significance level of the test. \( b \) does not retain value across rejection regions; that is, the \( b \) in the discussion of one rejection region is not the same \( b \) as in another rejection region.
Chapter 1. Introduction

In all simulations to determine the actual size and power of hypothesis tests or the coverage probability of confidence intervals, the tests will be conducted with nominal size of 0.05 and the confidence intervals will be constructed with nominal confidence level 0.95. Reported size is the proportion of datasets simulated under the null for which the test rejects, reported power is the proportion of datasets simulated under the alternative for which the test rejects, and reported coverage is the proportion of simulated datasets for which the interval contains the true value. The default formula for standard errors (SE) of all estimated sizes and coverage probabilities will be \( \sqrt{0.05(0.95)/s} \), where \( s \) is the number of simulated datasets on which the estimate is based. This is justified by substituting \( \hat{p} = 0.05 \) in the standard error formula in Theorem 1.3.6.9. The default formula for the standard error of all estimated powers will be \( 0.5/\sqrt{s} \), which is justified by substituting \( \hat{p} = 0.5 \) in the standard error formula in Theorem 1.3.6.9; this gives an upper bound on the true standard deviation of the estimated power.

When I wish to emphasize the fact that a random variable \( Z \) is the realization of a random sequence indexed by \( m \), I shall denote it \( Z(m) \). For instance, \( \bar{X}(N) \) is a random sequence indexed by \( N \). The \( m \)th Monte Carlo draw of \( Z \) will be \( Z(m) \).

If \( g(y) \) is a differentiable function from \( \mathbb{R}^a \) to \( \mathbb{R}^b \), \( \frac{\partial g(y)}{\partial y} \) will be the matrix whose \( ij \)th element is the partial derivative of the \( i \)th element of \( g \) with respect to the \( j \)th element of \( y \).

1.3 Background Theory

Some of the definitions and theorems in this section are slightly modified from their widespread use to make them more convenient for this thesis. For all theorems, I either give sources or a short proof or justification.

The results in this section are either well-known or fairly obvious from well-known results, except for Theorem 1.3.5.3, for which I absolutely take credit.
Chapter 1. Introduction

1.3.1 Properties of statistics

Definition 1.3.1.1: Sufficient statistic

If \( f_{X|T}(X|t; \theta) \) is invariant to \( \theta_{np} \), then \( T(X) \) is **sufficient** for \( \theta_{np} \).

Definition 1.3.1.2: The Sufficiency Principle ([2], page 272)

Using the notation of Definition 1.3.1.1, the **Sufficiency Principle** states that since the data contains no information about \( \theta_{np} \) apart from that contained in \( T \), all inference on \( \theta_{np} \) should depend on \( T \) alone.

Theorem 1.3.1.1: Sufficient statistics for normal data ([2], page 279)

If the data are a normal random sample, \( \{\bar{X}, S\} \) is sufficient for \( \{\mu, \sigma\} \).

Definition 1.3.1.3: Minimal sufficient statistic

\( T \) is **minimal sufficient** for \( \theta_{np} \) if it is a sufficient statistic for \( \theta_{np} \) (Definition 1.3.1.1) and it is a function of every other sufficient statistic for \( \theta_{np} \).

Theorem 1.3.1.2: Finding minimal sufficient statistics ([2], page 281)

\( T \) is minimal sufficient for \( \theta_{np} \) if for every \( z \) and \( y \) in the support of \( X \), \( \frac{f_X(z; \theta_{np}, \theta_{pl})}{f_X(y; \theta_{np}, \theta_{pl})} \) is constant as a function of \( \theta_{np} \) if and only if \( T(z) = T(y) \).

Definition 1.3.1.4: Complete statistic

\( T(X) \) is **complete** if the only functions \( g \) which satisfy \( E_T(g(T)) = 0 \) for every value of \( \theta \) in the parameter space are equal to \( 0 \) except on a set of measure 0.

Definition 1.3.1.5: Ancillary statistic

\( T(X) \) is **ancillary** for \( \theta_{np} \) if the distribution of \( T \) is invariant to \( \theta_{np} \).
Chapter 1. Introduction

Theorem 1.3.1.3: Basu’s Theorem ([2], page 287)

If $T_1$ is complete (Definition 1.3.1.4) and minimal sufficient (Definition 1.3.1.3) for $\theta_{np}$ and $T_2$ is ancillary (Definition 1.3.1.5) for $\theta_{np}$, then $T_1$ and $T_2$ are independent.

Theorem 1.3.1.4: Factorization Theorem ([2], page 276)

If $f_X(x; \theta) = g(T(x), \theta_{np})h(x, \theta_{pi})$, then $T$ is sufficient for $\theta_{np}$.

1.3.2 Invariance

Let $\mathcal{F}$ be a group of transformations with the properties that

1. For each transformation in $\mathcal{F}$, its inverse is in $\mathcal{F}$
2. If $g$ is in $\mathcal{F}$ and $h$ is in $\mathcal{F}$ then $h \circ g$ is in $\mathcal{F}$.

For example, $\mathcal{F}$ could be the group of positive scale transformations, which multiply all the elements of $X$ by the same positive scaling factor.

Definition 1.3.2.1: Orbit

An orbit $\mathcal{O}$ in the support $\mathcal{X}$ of $X$ with respect to $\mathcal{F}$ is formed by applying all the transformations in $\mathcal{F}$ to a point $x$ in $\mathcal{X}$ and keeping the resulting points that are in $\mathcal{X}$. By the properties of $\mathcal{F}$, if $y$ is in $\mathcal{O}$: it can be obtained from any other element in $\mathcal{O}$ by applying a transformation in $\mathcal{F}$, and if $y = g(z)$ where $z$ is in $\mathcal{X}$ and $g$ is in $\mathcal{F}$, then $z$ is in $\mathcal{O}$. For example, if $\mathcal{F}$ is the group of positive scale transformations above, the orbit that contains $x$ would be the set of all points $y$ in $\mathcal{X}$ satisfying $y = \alpha x$ for some constant $\alpha > 0$.

Definition 1.3.2.2: Invariance Principle ([3], page 41)

If applying a transformation from the group $\mathcal{F}$ to $X$ does not change the value of a parameter, then inference on that parameter should depend on $X$ only through which orbit in the support $\mathcal{X}$ with respect to $\mathcal{F}$ that $X$ falls into. In other words, all the points in the same orbit (Definition 1.3.2.1) should provide the same inference. For
example, if the data is a random sample, since the coefficient of variation of \( \alpha X \) is
\[
\frac{\sigma}{\mu} = \frac{\sigma}{\alpha \mu} = CV,
\]
inference on \( CV \) should be invariant to scale transformations.

**Definition 1.3.2.3: Maximal invariant**

A **maximal invariant** for a group \( F \) on \( X \) is a statistic \( T \) for which \( T(x) = T(y) \) if and only if \( x \) and \( y \) are in the same orbit (Definition 1.3.2.1) in the support of \( X \) with respect to \( F \). For the example of positive scale transformations, \( T = \{ 1, \frac{x_2}{\mu}, \ldots, \frac{x_N}{\mu} \} \) is a maximal invariant ([3], page 161). The Invariance Principle (Definition 1.3.2.2) implies that if a parameter is invariant to \( F \), inference on that parameter should be based solely on a maximal invariant for \( F \).

Applying a member of the group \( F \) to the data will not change the value of a maximal invariant.

Instead of considering the entire sample space, by the Sufficiency Principle (Definition 1.3.1.2) we can conduct all inference via a statistic \( T \) that is sufficient for \( \theta \). Let \( B(s) \) be the set of all points \( x \) in the support \( X \) of \( X \) for which \( T(x) = s \). An orbit in the support of the sufficient statistic with respect to \( F \) is the set of all points \( t \) for which \( t = T(y) \) for some \( y \) in the same orbit as an element of \( B(s) \) for some \( s \). A maximal invariant on \( T \) is a function \( T_1 \) defined on the support of the sufficient statistic for which \( T_1(t) = T_1(s) \) if and only if \( t \) and \( s \) are in the same orbit. Now if \( \theta_{np} \) is invariant to \( F \), the Invariance Principle implies that inference on \( \theta_{np} \) should be based solely on a “maximal invariant” for \( F \) on \( T \).

**Definition 1.3.2.4: Invariant inference**

Inference based on a maximal invariant (Definition 1.3.2.3) for a group \( F \) is said to be **invariant** to \( F \), because it is not changed by transforming \( X \) by a member of \( F \).

### 1.3.3 Inference

**Definition 1.3.3.1: Similar test or confidence region**

If the \( p \)-value of a test does not depend on the value of \( \theta_{np} \), the test is called a **similar test** and the \( p \)-value is called a **similar \( p \)**-value. A **similar region** is a confidence
region whose confidence level does not depend on the value of $\theta_{np}$.

**Definition 1.3.3.2: Unbiased test and confidence interval**

A hypothesis test is **unbiased** if for every value of $\theta$ that satisfies the null, the probability of rejecting the null is smaller than for every value of $\theta$ that does not satisfy the null. That is, the power function is always lower for a point in the null set than in the alternative set.

The probability that an **unbiased confidence interval** contain any false value of the parameter is less than the probability that it contains the true value.

**Theorem 1.3.3.1: Similar unbiased test in gamma populations [4]**

Let $X_i \sim \text{Gamma}(\alpha_i, \beta_i)$, with the shape parameters known. Let $H_o$ be that $\beta_1 = \ldots = \beta_N = \beta$. Here $\theta_{np} = \beta$. An unbiased similar test rejects the null if

$$\frac{\prod_{i=1}^{N} \left( \frac{X_i}{\alpha_i} \right)^{\alpha_i}}{\left( \frac{\sum_{i=1}^{N} X_i}{\sum_{i=1}^{N} \alpha_i} \right)^{\sum_{i=1}^{N} \alpha_i}} < b,$$

where $b$ is chosen to give the test the desired size.

**Definition 1.3.3.3: Uniformly most powerful (UMP) test**

If $\theta$ is one-dimensional, a **most powerful size $\alpha$ test** of $H_o : \theta = \theta_o$ against $H_a : \theta = \theta_a$ is a test that has greater power than any other size $\alpha$ test against that alternative. A **UMP size $\alpha$ test** against a class of alternatives is a most powerful size $\alpha$ test for all the alternatives in that class.

For two-sided alternatives typically no UMP test will exist. In this case, we would look for tests that are UMP within the class of unbiased tests.

If $\theta$ is multidimensional we can refer to tests concerning a one-dimensional $\theta_{pi}$ that are UMP within a certain class of tests, such as the class of similar tests or invariant tests.
Chapter 1. Introduction

**Theorem 1.3.3.2: Reparameterization in UMP tests**

Let $\theta_{pi}$ be one-dimensional. Let $g$ be a monotonically increasing function. Let $A$ be the rejection region for the uniformly most powerful (UMP) size $\alpha$ test in some class of $\theta_{pi} = \theta_o$ against the alternatives $\theta_{pi} > \theta_o$. $A$ is also the rejection region for the UMP size $\alpha$ test in that class of $g(\theta_{pi}) = g(\theta_o)$ against the alternative $g(\theta_{pi}) > g(\theta_o)$.

**Proof.** The null hypotheses are the same, and the set of alternatives for the latter test is also the set of alternatives for the former test.

The analogous lemmas for decreasing functions and for alternatives in the other direction are also true. It is also true, by the very same proof, that the UMP tests within a class of $H_o : \theta_{pi}^1 = \theta_{pi}^2$ against two-sided alternatives are identical to tests of $H_o : g(\theta_{pi}) = g(\theta_o)$.

**Definition 1.3.3.4: Uniformly most accurate (UMA) confidence intervals.**

Again let $\theta$ be one-dimensional.

Suppose the statistics $T_l(X), T_u(X)$ create a confidence interval for $\theta$ with confidence level $1 - \alpha$. They create a **UMA confidence interval of level** $1 - \alpha$ if for every value $\theta^*$ in the parameter space and $\theta^* \neq \theta^*$, $Prob_{\theta^*}(T_l < \theta^* < T_u)$ is smaller than $Prob_{\theta^*}(T_l^* < \theta^* < T_u^*)$ for any other confidence interval $T_l^*, T_u^*$ of level $1 - \alpha$.

In less precise English, the UMA interval has the smallest probability of admitting any wrong value.

Usually, we would only speak of intervals that are UMA within a certain class, most commonly the class of all unbiased intervals.

A **UMA lower confidence bound** $T_l$ of level $1 - \alpha$ satisfies the property that for all $\theta^*$ in the parameter space and $\theta^* < \theta^*$, $Prob_{\theta^*}(T_l < \theta^*)$ is smaller than $Prob_{\theta^*}(T_l^* < \theta^*)$ for any other $1 - \alpha$ lower confidence bound $T_l^*$. Such a bound would minimize any risk function for the underestimation of $\theta^*$ ([5], page 90). The definition of the UMA **upper confidence bound** is analogous to the lower bound.

If $\theta$ is multidimensional, we would look for intervals and bounds for a one-dimensional $\theta_{pi}$ that are UMA within a certain class, such as the class of similar regions or the class of invariant regions.
Chapter 1. Introduction

Theorem 1.3.3.3: UMP test inversion ([5], page 91)

Again, let \( \theta \) be one-dimensional.

Inverting the UMP test of size \( \alpha \) of \( H_o : \theta = \theta_o \) against \( H_a : \theta > \theta_o \) will produce the UMA lower confidence bound for \( \theta \) of confidence level \( 1 - \alpha \).

Inverting the UMP test of \( H_o : \theta = \theta_o \) against \( H_a : \theta < \theta_o \) will produce the UMA upper confidence bound for \( \theta \) of confidence level \( 1 - \alpha \).

Inverting the UMP unbiased size \( \alpha \) test of \( H_o : \theta = \theta_o \) against \( H_a : \theta \neq \theta_o \) will produce the UMA unbiased confidence interval.

If there are nuisance parameters, and \( \theta_{pi} \) is one-dimensional, inverting a test that is UMP for its class among similar tests of the hypothesis \( H_o : \theta_{pi} = \theta_o \) will produce confidence bounds that are UMA as just described among similar confidence bounds.

This theorem assumes that inverting the tests actually produces intervals. Examples where inverting UMP tests do not produce intervals are difficult to construct.

Definition 1.3.3.5: Fiducial inference

Suppose \( X = g(U, \theta) \), where the generating vector \( U \) is a random vector whose distribution is known. We can always find such a stochastic representation by letting \( U \) be a vector of iid uniform(0,1) random variables and letting \( g_i(U) = F_{X_i}^{-1}(U_i; \theta) \) ([2], page 54). I shall call \( U \) the vector of latent variables. Suppose that \( T \) is a statistic for inference on \( \theta \).

Traditional inference treats \( T \) as random and \( \theta \) as fixed, and evaluates a null hypothesis by calculating a \( p \)-value for \( t \). Fiducial inference treats \( t \) as fixed and \( \theta \) as random, and evaluates a null hypothesis \( H_o : \theta = \theta_o \) by comparing \( \theta_o \) to a reference distribution of \( \theta \). The difference between Bayesian inference and fiducial inference is that Bayesian inference uses a posterior distribution as the reference distribution, while fiducial inference compares \( \theta_o \) to draws from the fiducial distribution.

To draw a value \( \theta(j) \) from the fiducial distribution, one draws a value \( u(j) \) of the latent vector from distribution of \( U \), then solves \( T(g(u(j), \theta(j))) = t \).
Chapter 1. Introduction

Definition 1.3.3.6: More extreme symbol

The expression \( y <> x \) will mean “\( y \) is more extreme than \( x \).” The definition of extreme will depend on the context. The pronunciation of the expression will be “\( y \) is extremer than \( x \).”

Theorem 1.3.3.4: \( p \)-value as an expectation

The \( p \)-value associated with \( t \) for a test of the hypothesis \( \theta = \theta_0 \) based on \( T \) is \( E_X(I_{T<>t}) \). (see Definition 1.3.3.6).

Proof. The \( p \)-value is \( \text{Prob}_{\theta_0}(T <> t) \). The proof follows from the fact that probabilities of events are just the expected value of the indicator function that is 1 where the event occurs.

Definition 1.3.3.7: Monte Carlo \( p \)-value

If one takes \( s \) independent draws of \( X \) from its known distribution under the null hypothesis, the Monte Carlo \( p \)-value \( \hat{p}_{MC} \) associated with a value \( t \) for a statistic \( T \) is \( \frac{1}{s} \sum_{i=1}^{s} I_{T(x(i))<>t} \) (see Definition 1.3.3.6).

Since the Monte Carlo \( p \)-value is a sample proportion from a binomial experiment (see Definition 1.3.5.1), it converges almost surely to the actual \( p \)-value \( p \), and its standard deviation is \( \sqrt{\frac{p(1-p)}{s}} \), by Theorems 1.3.6.9 and 1.3.6.10.

The Monte Carlo test rejects if the Monte Carlo \( p \)-value is less than the nominal size.

1.3.4 Exponential families

In this section, I occasionally assume the existence of certain derivatives. These derivatives will exist under broad regularity conditions.
Chapter 1. Introduction

Definition 1.3.4.1: Exponential family

\( f_X(x; \theta) \) belongs to an exponential family if it can be written as

\[
h(x)C(\theta) \exp \left( \sum_{i=1}^{d} \theta_i T_i(x) \right).
\]

\( T \) will denote the vector \( \{T_1, \ldots, T_d\} \)

Theorem 1.3.4.1: Mean and variance of statistics in exponential families ([2], page 112)

In the exponential family of Definition 1.3.4.1,

\[
E(T_i) = - \frac{\partial}{\partial \theta_i} \ln(c(\theta)).
\]

\[
Var(T_i) = - \frac{\partial^2}{\partial \theta_i^2} \ln(c(\theta)).
\]

Theorem 1.3.4.2: Complete statistics in the exponential family ([2], page 288)

In the exponential family of Definition 1.3.4.1, as long as the parameter space contains an open set in \( \mathbb{R}^d \), \( T_i \) is complete (Definition 1.3.1.4).

Theorem 1.3.4.3: Minimal sufficient statistics in exponential family

In the exponential family of Definition 1.3.4.1, \( T_i \) is minimal sufficient for \( \theta_i \).

Proof.

\[
\frac{f_X(z; \theta_1, \theta_2, \ldots, \theta_d)}{f_X(y; \theta_1, \theta_2, \ldots, \theta_d)} = \frac{h(z) \exp \left( \sum_{i=2}^{d} \theta_i T_i(z) \right)}{h(y) \exp \left( \sum_{i=2}^{d} \theta_i T_i(z) \right)} \exp \left( \theta_1 (T_1(z) - T_1(y)) \right).
\]

This will be constant as a function of \( \theta_1 \) if and only if \( T_1(z) = T_1(y) \). Then by Theorem 1.3.1.2, \( T_1 \) is minimal sufficient for \( \theta_1 \).
Chapter 1. Introduction

Theorem 1.3.4.4: Uniformly most powerful tests in a one-parameter exponential family ([5], page 80)

Assume in the exponential family of Definition 1.3.4.1 that \( \theta \) is one-dimensional. If \( T_1 \) is a continuous random variable, the UMP size \( \alpha \) test of \( \theta = \theta_o \) against the alternatives \( \theta > \theta_o \) has the rejection region \( T_1 > b \), where \( b \) is chosen to give the test size \( \alpha \). The UMP test of \( \theta = \theta_o \) against the alternatives \( \theta < \theta_o \) has a rejection region of the form \( T_1 < b \).

This is slightly different from the form of the theorem stated in [5], but his result implies this theorem here via Theorem 1.3.3.2.

Theorem 1.3.4.5: UMP unbiased tests in a one-parameter exponential family ([5], page 136)

If in the exponential family of Definition 1.3.4.1 \( \theta \) is one-dimensional, there exists a UMP size \( \alpha \) unbiased test of \( H_o : \theta = \theta_o \) against \( H_a : \theta \neq \theta_o \), and the rejection region is \( T_1 > b_u \cup T_1 < b_l \). If \( \phi \) is the power function of the test, \( b_u \) and \( b_l \) solve \( \phi(\theta_o = \alpha) \) and \( \frac{d}{d\theta_o} \phi(\theta_o) = 0 \).

Theorem 1.3.4.6: UMP similar test in exponential families ([5], page 147)

In the exponential family of Definition 1.3.4.1, let \( \theta_{np} = \{\theta_2, \ldots, \theta_d\} \). Then UMP similar tests and UMP unbiased similar tests for \( H_o : \theta_{pi} = \theta_o \) have the forms in Theorem 1.3.4.4 and Theorem 1.3.4.5, with the constants determined by the distribution of \( T_1 \) conditional on \( T_2 = t_2, \ldots, T_d = t_d \).

Theorem 1.3.4.7: Finding UMP unbiased tests

Suppose we want to test \( H_o : \theta = \theta_o \) against \( H_a : \theta \neq \theta_o \) in a one-parameter exponential family. If a test of size \( \alpha \) has rejection region of the form \( T_1 > b_u \cup T_1 < b_l \), and the test is unbiased, then it is the UMP unbiased size \( \alpha \) test.

Proof. By Theorem 1.3.4.5, this lemma will be true if the derivative of the power function is 0 at \( \theta_o \). But unbiasedness implies that the power function has a minimum at \( \theta_o \); thus, the derivative must be 0. □
This lemma also applies to finding UMP unbiased similar tests; in this case one need only check that the test has size \( \alpha \) conditional on \( T_2 = t_2, \ldots, T_d = t_d \) and the test is unbiased conditional on \( T_2 = t_2, \ldots, T_d = t_d \).

**Theorem 1.3.4.8: Distribution conditional on sufficient statistics**

In the exponential family of Theorem 1.3.4.1, \( f_{X|\{T_{d-dnp+1},\ldots,T_d\}}(x|\{t_{d-dnp+1},\ldots,t_d\}; \theta) \) has the form

\[
I_{T_{d-dnp+1}=t_{d-dnp+1},\ldots,T_d=t_d} B(\theta_1, \ldots, \theta_{d-dnp}, t_{d-dnp+1}, \ldots, t_d) H(x) \exp\left( \sum_{i=1}^{d-dnp} \theta_i T_i \right).
\]

Also, \( f_{X|T}(x|t; \theta) = I_{T=t}(x) B(t) H(x) \).

**Proof.** I’ll present the proof for the latter statement. In the notation of Theorem 1.3.5.14, we can derive that \( g(t) = C(\theta) \exp(\sum_{i=1}^d \theta_i t_i) \int_{T(y)=t} H(y)dy \). Then from that Theorem 1.3.5.14 and Definition 1.3.4.1 we can derive the result. We get \( B(t) = \frac{1}{\int_{T(y)=t} H(y)dy} \).

### 1.3.5 Probability Theory

**Definition 1.3.5.1: Bernoulli trial and binomial experiment**

A Bernoulli trial is a random variable with two outcomes – 1 (success) or 0 (failure). Let the probability of success be \( p \). The mean of a Bernoulli trial is \( p \) and the variance is \( p(1-p) \) ([2], page 89). A binomial experiment is a random sample of \( N \) Bernoulli trials. The sample proportion \( \hat{p} \) in a binomial experiment is the number of successes over \( N \). The mean of the sample proportion is \( p \) and the variance is \( \frac{p(1-p)}{N} \).

**Definition 1.3.5.2: Jacobian**

In abstract terms, the Jacobian of a one-to-one function \( f \) from a domain in \( \mathbb{R}^n \) to a range in \( \mathbb{R}^m \) evaluated at a point \( z \), denoted \( J_f(z) \), is the inverse of the ratio of the
Chapter 1. Introduction

volume of an epsilon cube in the domain with a corner at \( z \) to the volume of the image of the cube. By definition, \( J_{f^{-1}}(y) = \frac{1}{J_f(f^{-1}(y))} \)

For a one-to-one function \( g \) from an \( n \)-dimensional subset of \( \mathbb{R}^n \) to an \( n \)-dimensional subset of \( \mathbb{R}^n \), \( J_g(z) \) is the absolute value of the determinant of \( \frac{\partial g}{\partial z}(z) \).

**Theorem 1.3.5.1: Multivariate transformations ([2], page 185)**

Suppose \( g \) is a one-to-one function from the support \( Z \) of \( Z \) onto some range. For \( Y = g(Z) \), \( f_Y(y) = f_z(g^{-1}(y))J_{g^{-1}}(y) \).

This theorem assumes the existence of the Jacobian \( J_{g^{-1}} \) at the point \( y \). (see Definition 1.3.5.2)

**Theorem 1.3.5.2: Generating uniform data on a sphere ([6], page 227)**

If \( X \) is a random sample from a \( N(0, 1) \) population, the vector \( \{ \frac{X_1}{\sqrt{\sum_{i=1}^{N} X_i^2}}, \ldots, \frac{X_N}{\sqrt{\sum_{i=1}^{N} X_i^2}} \} \) is uniformly distributed on the unit sphere.

**Theorem 1.3.5.3 Time management theorem**

You are wasting your time right now.

*Proof.* You are reading this theorem and its proof, which are irrelevant.

**Definition 1.3.5.3: Marksmanship**

Execution of a practical joke that would be worthy of Crosby Marks.

**Definition 1.3.5.4: Dirichlet distribution**

\( Z \) has a Dirichlet(\( \theta \)) distribution if \( f_z(z; \theta_1, \ldots, \theta_a) = I_{\sum_{i=1}^{a} z_i = 1} c(\theta) \prod_{i=1}^{a} z_i^{\theta_i-1} \).

**Theorem 1.3.5.4: Generating a Dirichlet random variable ([7], page 582)**

Let \( X_i \sim Gamma(\theta_i, 1) \) (Definition 1.3.5.10). Then the random vector whose \( i \)th element is \( \frac{X_i}{\sum_{i=1}^{N} X_i} \) is Dirichlet(\( \theta \)).
Theorem 1.3.5.5: Moment existence for bounded random variables

If \(Z\) is bounded, all of its moments exist.

Proof. For simplicity, I will assume \(Z\) is univariate. By definition, moments “exist” if their absolute values are less than \(\infty\). We can write the absolute value for the \(i\)th moment as

\[
\left| \int_{b_1}^{b_u} z^i f_Z(z)dz \right| \leq \int_{b_1}^{b_u} \left| z^i \right| f_Z(z)dz \leq M \int_{b_1}^{b_u} f_Z(z)dz = M
\]

for some bounds \(b_l\) and \(b_u\), where \(M\) is the upper bound on \(|Z^i|\).

\[\square\]

Theorem 1.3.5.6: Chebyshev’s Inequality ([2], page 122)

\[
Prob\left( \frac{|X-\mu|}{\sigma} > \epsilon \right) \leq \frac{1}{\epsilon^2}.
\]

Definition 1.3.5.5: Information matrix

The information matrix \(I(\theta)\) for a data vector is

\[
-EX\left( \frac{\partial^2}{\partial \theta \partial \theta^T} \ln (f_X(x; \theta)) \right).
\]

Theorem 1.3.5.7: Information matrix for normal random sample ([8])

Letting \(\theta = \{\mu, \sigma\}\), the information matrix (Definition 1.3.5.5) for a normal random sample is

\[
\begin{pmatrix}
\frac{N}{\sigma^2} & 0 \\
0 & \frac{2N}{\sigma^2}
\end{pmatrix}.
\]

Definition 1.3.5.6: Noncentral t distribution

Let \(Z \sim N(0, 1)\) and \(U \sim \chi_\nu^2\). Then \(\frac{Z+\delta}{\sqrt{U}}\) has a noncentral t distribution with \(\nu\) degrees of freedom and noncentrality parameter \(\delta\).

Theorem 1.3.5.8: \(\chi^2\) distribution ([2], page 623)

If \(Y \sim \chi_\nu^2\), then

\[
f_Y(y) = I_{y \geq 0} \frac{1}{(\nu)^{2\pi}} y^{\nu-1} \exp \left( -\frac{y}{2} \right).
\]

Theorem 1.3.5.9: Distribution of a linear combination of normal variables ([9], page 62)

If \(X_i \sim N(\mu_i, \sigma_i^2)\) then \(\sum_{i=1}^{N} \alpha_i X_i \sim N \left( \sum_{i=1}^{N} \alpha_i \mu_i; \sum_{i=1}^{N} \alpha_i^2 \sigma_i^2 \right)\).
Chapter 1. Introduction

In general, if \( Y \sim N(\mu, \Sigma) \), then \( AY + c \sim N(A\mu + c, A\Sigma A^T) \).

**Theorem 1.3.5.10: Distribution of a quadratic form ([10], page 135)**

If \( Y \) is normal with mean vector 0 and covariance matrix \( \Sigma \), and \( A \Sigma \) has rank \( r \) and is such that \( A\Sigma A\Sigma = A\Sigma \), then \( Y^T AY \sim \chi_r^2 \).

**Theorem 1.3.5.11: Consequences of Theorem 1.3.5.10**

1. If \( X_i \sim N(0, 1) \), then \( \sum_{i=1}^{N} X_i^2 \sim \chi_r^2 \).

2. Let \( X_i \sim \chi_{r_i}^2 \), \( \sum_{i=1}^{N} X_i \sim \chi_{\sum_{i=1}^{N} r_i}^2 \).

3. If \( Y \), an \( r \times 1 \) vector, is \( N(\{\mu, \ldots, \mu\}, V) \), where \( V \) is diagonal, and \( \tilde{Y} \) be a weighted average of the elements of \( Y \), where the \( i \)th weight is \( \frac{1}{V_{ii}} \). Then \( \sum_{i=1}^{k} \frac{(Y_i - \tilde{Y})^2}{V_{ii}} \sim \chi_{r-1}^2 \).

These results are immediate except for 3. A proof for 3 can be found in [11].

**Theorem 1.3.5.12: Multivariate normal conditional distribution ([9], page 63)**

Let \( \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim N \left( \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21}^T & \Sigma_{22} \end{pmatrix} \right) \).

Then assuming \( \Sigma_{22} \) is nonsingular, conditional on \( Z_2 = z_2 \),

\[ Z_1 \sim N \left( \mu_1 + \Sigma_{12} \Sigma_{22}^{-1}(z_2 - \mu_2), \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{12}^T \right). \]

**Definition 1.3.5.7: Generalized linear model**

The data are generated by a **generalized linear model** if the density is

\[ h(x) \exp \left( \sum_{i=1}^{N} Q(W_i^T \beta)x_i + K(W_i^T \beta) \right), \]

where \( Q \) and \( K \) are functions, \( \beta \) is a vector of parameters.
Definition 1.3.5.8: Canonical link

In Definition 1.3.5.7, the generalized linear model is said to have the canonical link if \( Q(W_i^T \beta) = W_i^T \beta \).

Definition 1.3.5.9: Skewness and Kurtosis

The population skewness \( \gamma_1 \equiv \frac{E(Y^3)}{\sigma^3} \), and the population kurtosis \( \gamma_2 \equiv \frac{E(Y^4)}{\sigma^4} - 3 \). Here we are defining kurtosis as the excess over that of the normal distribution.

Definition 1.3.5.10: Gamma distribution

If \( Y \sim \text{gamma}(\alpha, \beta) \), then \( f_Y(y) = \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} \exp\left(-\frac{y}{\beta}\right) \). \( \alpha \) is known as the shape parameter, and \( \beta \) is the scale parameter because changing \( \beta \) is tantamount to re-scaling \( Y \) via multiplication by a positive number.

From Theorem 1.3.4.3, we can find sufficient statistics for \( \alpha \) and \( \beta \) in a random sample. These are \( T_\alpha \equiv \sum_{i=1}^N \ln(X_i) \) and \( T_\beta \equiv \sum_{i=1}^N X_i \).

Definition 1.3.5.11: Kernel of distribution

If \( f_Y(y) = Ck(y) \), where \( C \) is a constant with respect to \( y \), then \( k(y) \) is a kernel of \( f_Y(y) \).

\( C \) is called the constant of proportionality.

Theorem 1.3.5.13: Expected standard deviation of Monte Carlo \( p \)-value under the null

Under the null, the expected value of the standard deviation of the Monte Carlo \( p \)-value (Definition 1.3.3.7) is \( \frac{\pi}{8\sqrt{\pi}} \).

Proof. Under the null, the distribution of the \( p \)-value, which is a random variable that I shall call \( U \), is \( U(0, 1) \) (Theorem 1.3.5.19). The expected value of the standard deviation of the Monte Carlo \( p \)-value is \( E_{\hat{p}_{MC}}(\sqrt{\text{Var}(\hat{p}_{MC})}) = E_U(E_{\hat{p}_{MC}|U}(\sqrt{\text{Var}(\hat{p}_{MC})|U})) = E_U\left(\frac{\sqrt{U(1-U)}}{\sqrt{\pi}}\right)\), by Definition 1.3.5.1. This is \( \frac{1}{\sqrt{\pi}} \int_0^1 \sqrt{z(1-z)}dz = \frac{1}{\sqrt{\pi}} \frac{\pi}{8} \).
Chapter 1. Introduction

Theorem 1.3.5.14: Distribution of data conditional on a statistic

The distribution of \( X \) conditional on \( T(X) = t \) is

\[
\frac{I_{T(X)=t}f_X(x)}{g(t)},
\]

where \( g(t) = \int_{T(y)=t} f_X(y)dy \).

Of note, \( I_{T(X)=t}f_X(x) \) is a kernel (Definition 1.3.5.11) of the conditional distribution.

Proof. The joint distribution of \( X \) and \( T \) is \( I_{T(X)=t}f_X(x) \). The conditional is this over the marginal distribution of \( T \) evaluated at \( t \), which is \( g(t) \).

\( \square \)

Theorem 1.3.5.15: Uniformity of data conditional on statistic

If the density is a function of the data only through a statistic \( T \), then the distribution of the data conditional on \( T = t \) is uniform on the subset of the support of \( X \) on which \( T = t \).

Proof. We can write \( f_X(x) = g(t) \). From Theorem 1.3.5.14, on the conditional support the density is proportional to \( g(t) \), which is constant on that support.

\( \square \)

Theorem 1.3.5.16: Marginal distribution of a coordinate of a one-dimensional surface

Suppose that a surface \( S \) in \( \mathbb{R}^p \) is the set of all points that can be written as

\[
y = \{x_1, g_1(x_1), \ldots, g_{p-1}(x_1)\}, x_1 \in \mathcal{X}
\]

where \( g(x_1) \equiv \{g_1(x_1), \ldots, g_{p-1}(x_1)\} \) is differentiable and one-to-one. The probability density \( I_Y \in sf(y) \) implies a marginal density for \( X_1 \). That density is

\[
I_{x_1 \in \mathcal{X}} f_x(\{x_1, g(x_1)\}) \sqrt{1 + \sum_{i=1}^{p-1} \left( \frac{dg_i}{dx_1}(x_1) \right)^2}.
\]
Proof. The probability that $x_1$ is in an interval $[a, b]$ is equal to the probability that a point on $S$ is in the interval defined by $\{a, g(a)\}$ and $\{b, g(b)\}$. From Theorem 1.3.8.14, we can deduce that this probability is

$$
\int_a^b I_{t \in X} f(\{t, g(t)\}) \sqrt{1 + \sum_{i=1}^{p-1} \left( \frac{dg_i}{dx_1}(t) \right)^2} \, dt.
$$

Then the integrand must be the marginal density of $X_1$. \hfill \Box

Definition 1.3.5.12: Homogeneous function

A function $g(x)$ is **homogeneous of degree** $k$ if $g(\alpha x) = \alpha^k g(x)$ for any positive constant $\alpha$.

Definition 1.3.5.13: Accept/reject method of generating a random variable

Suppose we want to generate a random sample from a target density $f_X$, and we have a generating density $f_Z$ that has the same support. Let $M = \max \left( \frac{f_X}{f_Z} \right)$ over the support. One step of the accept/reject algorithm would generate $z$ from $f_Z$ and accept this as an observation of $X$ with probability $\frac{f_X(z)}{M f_Z(z)}$.

Theorem 1.3.5.17: Independence and factorization ([2], page 153)

$X_1, \ldots, X_N$ are independent if and only if $f_X(x_1, \ldots, x_N) = \prod_{i=1}^N f_i(x_i)$ for some functions $f_1, \ldots, f_N$.

Definition 1.3.5.14: Memoryless property

If the distribution of $Y$ conditional on $Y > c$ is the same for all $c$, then the distribution of $Y$ is said to possess the **memoryless property**.

Theorem 1.3.5.18: Conditional variance identity ([2], page 167)

If $Y$ and $Z$ are any univariate two random variables, $Var(Z) = E_Y(Var_{Z|Y}(Z|Y)) + Var(E_{Z|Y}(Z|Y))$. 

27
Chapter 1. Introduction

Definition 1.3.5.15: Rejection sampling

One method of generating a random variable $Z$ conditional on some event is rejection sampling. One step of this algorithm generates a $z$ from the unconditional distribution of $Z$, but this value is included in the final sample only if it is in the event.

Theorem 1.3.5.19: Distribution of $p$-value under the null

The true $p$-value for any test is itself a random variable. If the test statistic is continuous, the true $p$-value associated with it will be $U(0, 1)$ under the null.

Proof. This is consequence of the probability integral transformation (see Casella and Berger [2], page 54). \qed

1.3.6 Asymptotic theory

Definition 1.3.6.1: Convergence in probability

$L(n)$ converges in probability with $n$ to $\text{plim}(L)$, a random variable or constant, if for any $\epsilon > 0$ $\lim_{n \to \infty} (\text{Prob}(\sqrt{\sum_{i=1}^{n} (L(n)_i - \text{plim}(L)_i)^2} > \epsilon)) = 0$.

I shall write $L(n) \to_p \text{plim}(L)$.

Definition 1.3.6.2: Almost sure convergence

$L(n)$ converges almost surely with $n$ to $\text{alim}(L)$, a random variable or constant, if $\text{Prob}(\lim_{n \to \infty} (L(n)) = \text{alim}(L)) = 1$.

One trivial example is that $\frac{Z}{g(N)}$ converges almost surely to 0 if $g$ is a monotonically increasing, unbounded function of $N$.

I shall write $L(n) \to_{as} \text{alim}(L)$.

Definition 1.3.6.3: Multivariate convergence in distribution

$L(n)$ converges in distribution with $n$ to a random variable $cd(L)$ if for any region $\mathcal{A}$ in the support of $cd(L)$, $\lim_{n \to \infty} (\text{Prob}(L(n) \in \mathcal{A})) = \text{Prob}(cd(L) \in \mathcal{A})$.

I shall write $L(n) \to_d cd(L)$.
Chapter 1. Introduction

Definition 1.3.6.4: Asymptotic normality

$L(n)$ is asymptotically normal with $n$ if there is a vector of constants $L^*$ such that $\sqrt{n}(L(n) - L^*)$ converges in distribution with $n$ to a multivariate normal random variable with mean 0 and some constant covariance matrix $\Sigma_L$. We will write $L(n) \sim AN(L^*, \frac{1}{n}\Sigma_L)$.

Theorem 1.3.6.1: Relationship among types of convergence

1. If $L(n) \rightarrow_{as} \text{plim}(L)$, then $L_n \rightarrow_p \text{plim}(L)$. ([12], page 70)

2. If $L(n) \sim AN(L^*, \frac{1}{n}\Sigma_L)$, then $L(n) \rightarrow_p L^*$.

Proof. I will give a handwaving argument here, using the univariate case for simplicity. For large $n$, $\text{Prob}(\|L(n) - L^*\| > \epsilon) = \text{Prob}\left(\frac{|L(n) - L^*|}{\sqrt{n}\Sigma_L} > \sqrt{n}\frac{\epsilon}{\Sigma_L}\right) \approx \text{Prob}(\|N(0,1)\| > \sqrt{nc})$ for some constant $c$, which converges to 0.

Theorem 1.3.6.2: Asymptotic normality of the maximum likelihood estimator $MLE$ [8]

Under regularity conditions, $\hat{\theta}_{MLE}(N) \sim AN(\theta, I^{-1}(\theta))$.

($I(\theta)$ is defined in Definition 1.3.5.5.)

Theorem 1.3.6.3: Multivariate delta method ([9], page 52)

If $L(n) \sim AN(\text{plim}(L), \frac{1}{n}\Sigma_L)$, and $g$ is differentiable at $\text{plim}(L)$,

$$g(L(n)) \sim AN_n\left(g(\text{plim}(L)), \frac{1}{n} \frac{\partial}{\partial L} g(\text{plim}(L)) \Sigma_L \frac{\partial}{\partial L} g(\text{plim}(L))^T\right).$$

Theorem 1.3.6.4: Slutsky’s Theorem ([2], page 239)

If $L(n) \rightarrow_d cd(L)$, and $Q(n) \rightarrow_p \alpha$, a constant, then $L(n)Q(n) \rightarrow_d \alpha cd(L)$ and $L(n) + Q(n) \rightarrow_d cd(L) + \alpha$. 

29
Chapter 1. Introduction

Theorem 1.3.6.5: Continuity and convergence

1. If \( L(n) \to_p \text{plim}(L) \), and \( g \) is a continuous function, then \( g(L(n)) \to_p g(\text{plim}(L)) \). ([2], page 233)

2. If \( L(n) \to_d \text{cd}(L) \), and \( g \) is continuous on the support of \( \text{cd}(L) \), then \( g(L(n)) \to_d g(\text{cd}(L)) \). ([8])

3. If \( L(n) \to_{as} \text{alim}(L) \) and \( Q(n) \to_{as} \text{alim}(Q) \), and \( g \) is a continuous function, then \( g(L(n), Q(n)) \to_{as} g(\text{plim}(L), \text{plim}(Q)) \).

Proof. (Of part 3): This follows from Theorem 1.3.8.10.

By simple iteration, one can extend this theorem to any number of sequences.

Theorem 1.3.6.6: Slutsky-delta method of obtaining standard normal statistics

If \( L(n) \sim AN(\mu, \frac{1}{n}g(\theta)) \), where \( g \) is continuous at \( \theta \), and \( \hat{\theta}(n) \to_p \theta \), then

\[
\sqrt{n} \frac{L(n) - \mu}{\sqrt{g(\hat{\theta}(n))}} \to_d N(0, 1).
\]

Proof. From Definition 1.3.6.4, \( \sqrt{n} \frac{L(n) - \mu}{\sqrt{g(\theta)}} \to_d N(0, 1) \). Also note that by Theorem 1.3.6.5 part 1, \( \frac{\sqrt{g(\theta)}}{\sqrt{g(\hat{\theta})}} \to_p 1 \). Then by Slutsky’s Theorem (1.3.6.4),

\[
\sqrt{n} \frac{L(n) - \mu}{\sqrt{g(\hat{\theta})}} = \sqrt{n} \frac{L(n) - \mu \sqrt{g(\theta)}}{\sqrt{g(\theta)}} \to_d N(0, 1).
\]

\[\square\]

Theorem 1.3.6.7: Central Limit Theorem ([9], page 51)

If \( \Sigma \) is the covariance matrix for a random sample, \( \mathbf{X}(N) \sim AN(\mu, \frac{1}{N}\Sigma) \).
Chapter 1. Introduction

Many extensions to the Central Limit Theorem have been developed (see [5], Chapter 11). Extensions have been given for nonindependent data. Extensions have also been given for cases where the $X_i$s are not identically distributed. If the process generating the data is such that the average of the population means converges to a constant and the average of the population covariance matrices converges to a constant, the sample mean will be asymptotically normal. And if the process generating the population means and covariance matrices can be thought of as a random sample, this condition will be fulfilled by the Strong Law of Large Numbers (Theorem 1.3.6.8.)

Theorem 1.3.6.8: Strong Law of Large Numbers ([2], page 235)

If $X$ is a random sample, and its covariance matrix exists, then $\bar{X}(N) \to_{as} \mu$.

Theorem 1.3.6.9: Asymptotic normality of sample proportion

In a binomial experiment, $\hat{p}(N) \sim AN \left( p, \frac{p(1-p)}{N} \right)$.

Proof. Realizing that the sample proportion is the sample mean in a binomial experiment, in which the individual observations have mean $p$ and variance $p(1 - p)$, this follows from Definition 1.3.5.1 and the Central Limit Theorem (Theorem 1.3.6.7).

This gives an expression for the standard error for the sample proportion: $\sqrt{\hat{p}(1-\hat{p}) \frac{1}{N}}$.

Theorem 1.3.6.10: Convergence of sample proportion

In a binomial experiment (Definition 1.3.5.1), $\hat{p}(N) \to_{as} p$.

Proof. The sample proportion is also the sample mean in a binomial experiment, in which each observation has mean $p$ and variance $p(1 - p)$. Then the result follows from the Strong Law of Large Numbers (Theorem 1.3.6.7).

Theorem 1.3.6.11: Eventuality of success in a binomial experiment

If a binomial experiment (Definition 1.3.5.1) is repeated until the first success, the probability that it eventually stops is 1 as long as the probability of success is greater than 0.
Chapter 1. Introduction

Proof. If this were not true, there would be a positive probability that the sample proportion would not converge to the true proportion, which would violate Theorem 1.3.6.10.

Theorem 1.3.6.12: Asymptotics of $\chi^2$ distribution

If $Y \sim \chi^2_m$, then $\frac{Y}{m}(m) \sim AN(1, \frac{2}{m})$, and $\frac{Y}{m}$ converges almost surely with $m$ to 1.

Proof. We can write $Y = \sum_{i=1}^{m} Z_i^2$, where the $Z_i$s are iid $N(0, 1)$. Now, $E(Z_i^2) = 1$ and $Var(Z_i^2) = E(Z_i^4) - E(Z_i^2)^2 = 2$. Then the results follow from the Central Limit Theorem (Theorem 1.3.6.7) and the Strong Law of Large Numbers (Theorem 1.3.6.8).

Theorem 1.3.6.13: Asymptotic distribution of the likelihood ratio statistic [8]

If $L(N)$ is the likelihood ratio statistic for testing $H_o : \theta_{\mu} = \theta_o$, and $\theta_{\mu}$ is $r \times 1$, under the null hypothesis (and certain regularity conditions) $-2 \ln(L(N))$ converges in distribution (Definition 1.3.6.3) to a $\chi^2_r$.

Theorem 1.3.6.14: Joint asymptotic normality of sample mean and variance [5]

Under regularity conditions, for a univariate random sample

$$
\left(\begin{array}{c}
\bar{X}(N) \\
S^2(N)
\end{array}\right) \sim AN\left(\begin{array}{c}
\mu \\
\sigma^2
\end{array}\right), 1 \frac{1}{N} \left(\begin{array}{cc}
\sigma^2 & \mu_3 \\
\mu_3 & \mu_4 - \sigma^4
\end{array}\right).
$$

Proof. Without loss of generality, I shall replace the $N - 1$ in the denominator of $S^2$ with an $N$. We can rewrite the vector on the left hand side of the expression as

$$
\left(\begin{array}{c}
\frac{1}{N} \sum_{i=1}^{N} X_i \\
\frac{1}{N} \sum_{i=1}^{N} (X_i - \mu)^2
\end{array}\right) + \left(\begin{array}{c}
0 \\
(\mu - \bar{X})^2
\end{array}\right).
$$
Chapter 1. Introduction

Now by the Strong Law of Large Numbers (Theorem 1.3.6.8), Theorem 1.3.6.5, part 3, and Theorem 1.3.6.1, part 1, the second term converges in probability to the 0 vector. So by Slutsky’s Theorem (1.3.6.4), the result can be shown by proving the asymptotic normality of the first term. Now the vector

\[
\begin{pmatrix}
X_i \\
(X_i - \mu)^2
\end{pmatrix}
\]

has mean

\[
\begin{pmatrix}
\mu \\
\sigma^2
\end{pmatrix}
\]

and covariance

\[
\begin{pmatrix}
\sigma^2 & \mu_3 \\
\mu_3 & \mu_4 - \sigma^4
\end{pmatrix}.
\]

Then the result follows via the Multivariate Central Limit Theorem (1.3.6.7).

\[\square\]

Theorem 1.3.6.15: Asymptotic unbiasedness of normal sample standard deviation [13]

\[
\frac{\sqrt{2\pi}}{\sqrt{N - 1} \Gamma\left(\frac{N - 1}{2}\right)}
\]

converges to 1 as \(N \to \infty\).

1.3.7 Markov Chains

Definition 1.3.7.1: Markov chain

A Markov chain \(\mathcal{C}\) is a stochastic process that produces, at step \(n\), a random variable \(Y(n)\) whose density function \(f_{\mathcal{C}}(y(n)\mid y(n - 1))\) depends on the previous steps only through the value of \(Y(n - 1)\).

The sequence of densities that define a chain, the rules for generating \(Y(n)\) once \(Y(n - 1)\) has been observed, imply a density \(f_{\mathcal{C},n}\) for \(Y(n)\) which specifies probabilities of events concerning \(Y(n)\) before \(Y(2)\) has been observed. I shall call this the predictive density of the chain.
Chapter 1. Introduction

**Definition 1.3.7.2: Gibbs sampling Markov chain**

The **Gibbs sampler** is a Markov chain (Definition 1.3.7.1) constructed in the following way:

1. $Y_1(n)$ is drawn from the density $f_{Z_1|\{Z_2,\ldots,Z_a\}}(y_1(n)|\{y_2(n-1),\ldots,y_a(n-1)\})$, for some random variable $Z$.

2. $Y_2(n)$ is drawn from the density $f_{Z_2|\{Z_1,Z_3,\ldots,Z_a\}}(y_2(n)|\{y_1(n),y_3(n-1),\ldots,y_a(n-1)\})$.

3. And so on until $Y_a(n)$ is drawn from $f_{Z_a|\{Z_1,\ldots,Z_{a-1}\}}(y_a(n)|\{y_1(n),\ldots,y_{a-1}(n)\})$.

I shall say that $Y_i(n)$ is drawn in the $i$th substep of the $n$th step. The density from which $Y_i(n)$ is drawn in the $i$th substep of the $n$th step will be called the $i$th jumping density. $f_Z$ will be called the **target** density.

**Definition 1.3.7.3: Su sampling Markov chain**

I know of no terminology for this useful construct, so I will propose one here.

The **Su sampler** is a Markov chain (Definition 1.3.7.1) constructed in the following way:

\[ \{Y_s(n), Y_{s+1}(n), \ldots Y_{s+u}(n)\} \text{ is drawn from} \]
\[ f_{Z_{su}|Z_{-su}}(\{y_s(n),\ldots,y_{s+u}(n)\}|\{y_1(n-1),\ldots,y_{s-1}(n-1),y_{s+u+1}(n-1),\ldots y_a(1)\}), \]

where $Z_{su} = \{Z_s,\ldots,Z_u\}$ and $Z_{-su}$ consists of the elements of $Z$ not in $Z_{su}$. The other elements of $Y(n)$ are constant as a function of $n$.

**Definition 1.3.7.4 Chaining of Markov chains**

Two Markov chains $\mathcal{C}_1$ and $\mathcal{C}_2$ are **chained** into a stochastic process if the first substep of the process produces $Y(1)$ via the density $f_{\mathcal{C}_1}(y(1)|y(0))$, the second substep produces $Y(2)$ via the density $f_{\mathcal{C}_2}(y(2)|y(1))$, the third substep produces $Y(3)$ via the density $f_{\mathcal{C}_1}(y(3)|y(2))$, etc.
Chapter 1. Introduction

The even substeps of a stochastic process produced in this way are a Markov chain (Definition 1.3.7.1), and the density function defining the chain is

$$\int_{\mathcal{E}(y(n-1))} f_{C_2}(y(n)|y)f_{C_1}(y|y(n-1))dy,$$

where $\mathcal{E}(y(n-1))$ is the support of the density $f_{C_1}(\cdot|y(n-1))$.

More than two Markov chains can be combined in this way to create a new Markov chain; if $b$ chains are combined, the $b$th substeps compose a chained Markov chain. I shall refer to the outcome of the $b$th substep as a step.

**Definition 1.3.7.5: Stationary distribution of Markov Chain**

The density $f_Z$ is the stationary distribution of a Markov Chain $C$ if

$$\int_Z f_C(y(n)|z)f_Z(z)dz = f_Z(y(n)),$$

where $Z$ is the support of $Z$.

**Theorem 1.3.7.1: Stationary distribution of the Su sampler**

The stationary distribution (Definition 1.3.7.5) of the Su sampler is $f_Z$, where $Z$ is as defined in Definition 1.3.7.3.

**Proof.** We have $f_C(y(n)|z) = I_{y(1)=z_1, \ldots, y_{s-1}(n)=z_{s-1}, y_{s+u+1}(n)=z_{s+u+1}, \ldots, y_a(n)=z_a} \times f_{Z_{su}|Z_{-su}}\left(\{y_s(n), \ldots, y_{s+u}(n)\}\{z_1, \ldots, z_{s-1}, z_{s+u+1}, \ldots, z_a\}\right)$.

Now this is $f_{Z|Z_{-su}}(y(n)|\{z_1, \ldots, z_{s-1}, z_{s+u+1}, \ldots, z_a\})$.

To prove the result, we need to evaluate the integral in Definition 1.3.7.5. It is

$$\int_Z f_{Z|Z_{-su}}(y(n)|\{z_1, \ldots, z_{s-1}, z_{s+u+1}, \ldots, z_a\})f_Z(z)dz =$$

$$\int_Z f_{Z|Z_{-su}}(y(n)|\{y_s(n), \ldots, y_{s+u}(n)\}|Z_{-su})f_{Z_{su}}(z_{-su})dz_1 \ldots dz_{s-1}dz_{s+u+1} \ldots dz_a.$$
Chapter 1. Introduction

The integrand is the joint density of $Z$ and $Z_{-su}$, so integrating out the latter random variable will leave the marginal density of $Z$, which is the result we wanted. □

**Theorem 1.3.7.2: Support of stationary distribution**

Let $f_Z$ be the stationary density of a Markov chain $C$ and let $S$ be the set of all points in the support of $f_C(y(n)|z)$ for at least one value $z$ in the support $Z$ of $Z$. Then $S$ equals $Z$ up to a set which has measure 0 under the measure defined by $f_Z$.

*Proof.* To see this, realize that one way to interpret Definition 1.3.7.5 is that the marginal distribution of $Y(n)$ from the random variable $\{Y(n), Z\}$ is the same as the marginal distribution of $Z$. □

**Theorem 1.3.7.3: Stationary distribution of chained Markov chains**

If Markov chains $C_1, \ldots, C_b$ have the same stationary distribution, then the Markov chain that results from chaining them will have the same stationary distribution.

*Proof.* I will do the proof for the case $b = 2$. For general $b$, simply apply this proof iteratively.

Let $f_Z$ be the stationary distribution of both chains. Call the new Markov chain $C_{12}$.

From Definition 1.3.7.4, $f_{C_{12}}(y(n)|z) = \int_{E(z)} f_{C_2}(y(n)|y)f_{C_1}(y|z) dy$.

To complete the proof, we need to evaluate the integral in Definition 1.3.7.5. It is

$$\int_Z \int_{E(z)} f_{C_2}(y(n)|y)f_{C_1}(y|z)f_Z(z) dydz,$$

where $Z$ is the support of $Z$.

As a technical point, we can replace $E(z)$ with the set $S_1$ of all points in the support of $f_{C_1}(y|z)$ for at least one value of $z$, since $f_{C_1}(y|z)$ is 0 outside $E(z)$. That means by Theorem 1.3.7.2, we can replace $E(z)$ with $Z$.

Now write the expression as

$$\int_Z \int_Z f_{C_2}(y(n)|y)f_{C_1}(y|z)f_Z(z) dydz$$
Chapter 1. Introduction

Changing the order of integration, and we get the expression

$$\int_Z f_{c_2}(y(n)|y) \left( \int_Z f_{c_1}(y|z)f_Z(z)dz \right) dy,$$

which by Definition 1.3.7.5 equals

$$\int_Z f_{c_2}(y(n)|y)f_Z(y)dy,$$

which by Definition 1.3.7.5 is $f_Z(y(n))$, which completes the proof. \(\square\)

**Theorem 1.3.7.4: Stationary distribution of Gibbs sampler**

Using the notation of Definition 1.3.7.2, the stationary distribution of the Gibbs sampler is $f_Z$.

**Proof.** This follows from Theorem 1.3.7.1 and 1.3.7.3, since the Gibbs sampler is a chain of the $a$ Su samplers formed by setting $s = 1$, $s = 2$, etc. and $u = 0$. \(\square\)

**Definition 1.3.7.6: Aperiodic**

If there are no disjoint subsets $D_1, \ldots, D_c$ such that $\text{Prob}(Y(n) \in D_1|Y(n-1) \in D_c) = 1$, $\text{Prob}(Y(n) \in D_b|Y(n-1) \in D_{b-1}) = 1$, and the probability that the chain ever enters $D_1$ is positive for some starting point, the chain is called periodic. A chain that is not periodic is called aperiodic.

Basically, an aperiodic chain is one that does not get caught up into a recognizable cycle.

**Definition 1.3.7.7: Irreducibility**

A Markov chain $C$ with a stationary distribution $f_Z$ is irreducible if for any set $E$ for which $\int_E f_Z(y)dy > 0$ there exists an $n$ such that $\int_E f_{c,n}(y)dy > 0.$ (See Definition 1.3.7.1.)
Chapter 1. Introduction

Theorem 1.3.7.5: Convergence to stationary distribution [14]

If a Markov chain whose stationary distribution is $f_Z$ is aperiodic and irreducible, $Y_n \rightarrow_d Z$.

Actually, a result that is stronger than the simple pointwise convergence of the distribution function of $Y(n)$ can be given, but this is sufficient for our purposes.

This result is often used to justify taking the realized values of a Markov chain as the values of a random sample for $Z$. Even if $n$ is large enough that the distribution of $Y_n$ is indistinguishable from that of $Z$, this is not precisely valid because the steps of the chain are correlated. However, usually if we take a large enough sample it will behave like a random sample. In any case, usually we are interested in the expected value of a function of $Z$, and we can use the following theorems to justify the presumption of independence for the purpose of approximating that expected value.

Theorem 1.3.7.6: Ergodic Theorem [14]

Let $N_{mc}$ be the total number of steps taken in a Markov chain. If a Markov chain with stationary distribution $f_Z$ is aperiodic and irreducible, $\frac{1}{N_{mc}} \sum_{n=1}^{N_{mc}} g(Y(n))$ converges almost surely with $n$ to $E_Z(g(Z))$.

Theorem 1.3.7.7 Central Limit Theorem for Markov chains [14]

Let $N_{mc}$ be the total number of steps taken by a Markov chain.

Under regularity conditions, if a Markov chain with stationary distribution $f_Z$ is aperiodic and irreducible, then $\frac{1}{N_{mc}} \sum_{n=1}^{N_{mc}} g(Y(n)) \sim \mathcal{N} \left( E_Z(g(Z)), \frac{1}{N_{mc}} \tau Var(g(Z)) \right)$, where $\tau$ is a correction factor that accounts for the fact that the draws from the Markov chain are correlated. If the first draw of the chain is from its stationary distribution, it is possible to derive $\tau = \lim_{N_{mc} \to \infty} \sum_{n=1}^{N_{mc}} Corr(g(Y(n)), g(Y(1)))$.

This theorem requires a regularity condition called geometric ergodicity [14]. It’s difficult to check this condition, so often the theorem is invoked as a justification for the assumption of normality rather than as a proof.

Central limit theorems for correlated data can also be used to justify normality, since for large $N_{mc}$ Theorem 1.3.7.5 indicates that the marginals of the $Y_n$ behave like...
Chapter 1. Introduction

the stationary density.

Definition 1.3.7.8: Metropolis-Hastings Markov chain

Let $f_Z(z)$ be a target density, and let $Q(n)$ be drawn from a generating density $f_Q(n)$ with the same support as $f_Z$. The Metropolis-Hastings Markov chain sets $Y(n) = Q(n)$ if $B(n) = 1$ and $Y(n) = y(n-1)$ if $B(n) = 0$, where $B(n)$ is a Bernoulli random variable with probability of success $\min\left(\frac{f_Z(q(n))f_Q(n)}{f_Z(y(n-1))f_Q(n-1)}, 1\right)$.

Definition 1.3.7.9: Metropolis-within-Gibbs Markov chain

The Metropolis-within-Gibbs Markov chain chains together Metropolis-Hastings chains, where the equivalent of $f_Z$ (Definition 1.3.7.8) in the $i$th substep is the density of a random variable $V$ conditional on $V_1 = v_1(n), \ldots, V_{i-1} = v_{i-1}(n), V_{i+1} = v_{i+1}(n-1), \ldots, V_a = v_a(n-1)$. This is the jumping density for a Gibbs sampling Markov Chain. (see Definition 1.3.7.2).

Theorem 1.3.7.10: Stationary distribution of Metropolis-Hastings chain ([2], page 255)

The stationary distribution of the Metropolis-Hastings chain is $f_Z$ from Definition 1.3.7.8. By Theorems 1.3.7.1 and Theorem 1.3.7.3, the stationary distribution of the Metropolis-within-Gibbs Markov chain is $f_V$.

1.3.8 Basic Math

Definition 1.3.8.1: Compact set

For our purposes, a compact set is a bounded set in $\mathbb{R}^n$ that contains its boundary.

Theorem 1.3.8.1: The minimum-maximum theorem ([15], page 189)

A continuous function on a compact set is bounded.
Chapter 1. Introduction

Definition 1.3.8.2: Convex set

A set is **convex** if the line segment connecting two points in the set is also contained in the set.

Definition 1.3.8.3: Generalized inverse

A matrix $A^g$ is a **generalized inverse** of a matrix $A$ if $AA^gA = A$.

Theorem 1.3.8.2: Generalized inverses of symmetric matrices ([16], page 115)

Every symmetric matrix has a symmetric generalized inverse.

Theorem 1.3.8.3: Particular generalized inverse ([16], page 167)

The matrix $(A^TA)^gA^T$, where $(A^TA)^g$ is a generalized inverse of $A^TA$, is a generalized inverse of $A$.

Combined with Theorem 1.3.8.2, this ensures that a generalized inverse always exists.

Theorem 1.3.8.4: Solution to set of linear equations ([16], page 141)

Let $A$ be an $r \times n$ matrix of rank $r$, let $v$ be an $r \times 1$ vector, and let $A^g$ be a generalized inverse of $A$. A vector $y$ satisfies

$$Ay = v$$

if and only if can be written

$$y = A^gv + (I - A^gA)z,$$

for some $z$ in $\mathbb{R}^n$.

The set of solutions described by the equation above is a hyperplane in $\mathbb{R}^n$. 

40
Chapter 1. Introduction

Definition 1.3.8.4: Orthogonal matrix

An \( r \times n \) matrix \( A \) is an **orthogonal matrix** if the columns are orthogonal to each other and each column has norm 1. \( A^T A = I_{n \times n} \).

Theorem 1.3.8.5: Eigenvectors and eigenvalues of symmetric matrix ([17], page 294)

Every \( r \times r \) symmetric matrix \( A \) can be written in the form \( PDP^T \), where \( P \) is an \( r \times n \) orthogonal matrix and \( D \) is an \( n \times n \) diagonal matrix with real nonzero elements on the diagonal.

This theorem implies \( AP = PD \), because \( P^TP = I_{n \times n} \).

Definition 1.3.8.5: Eigenvectors and eigenvalues

The columns of \( P \) in Theorem 1.3.8.5 are the **eigenvectors** of \( A \), and the elements of \( D \) are the **eigenvalues** of \( A \).

Definition 1.3.8.6: Column space

The column space of a matrix \( A \) is the set of all points that can be written \( Ay \) for some vector \( y \). The column space of a collection of column vectors as the column space of the matrix that they form.

Theorem 1.3.8.6: Column space as a subset

If all of the columns of \( A \) are contained within the column space of \( B \), then the column space of \( A \) is contained in the column space of \( B \).

*Proof.* \( A = BZ \) for some matrix \( Z \). Then \( Ay = BZy = By^* \) for some vector \( y^* \). \( \square \)

Theorem 1.3.8.7: Unique representation via an orthogonal matrix

For a point \( b \) in the column space of an orthogonal matrix \( P \), there is a unique \( y \) that solves \( Py = b \).
Chapter 1. Introduction

Proof. First realize that such a \( y \) exists by the definition of a column space (Definition 1.3.8.7). Next realize that \( P^T \) is a generalized inverse of \( P \) (Definition 1.3.8.3). By Theorem 1.3.8.4, all solutions are of the form \( y = P^Tb + (I - P^TP)z = P^Tb. \)

Definition 1.3.8.7: Taylor expansion

The Taylor expansion of a function \( g(y) \) around a point \( a \) is

\[
g(a) + \sum_{i=1}^{\infty} \frac{d^i g(a)}{i!} (y-a)^i.
\]

For example, the Taylor expansion of \( \frac{1}{1+y} \) around 0 is

\[
\sum_{i=0}^{\infty} (-1)^i y^i.
\]

The Taylor expansion, usually taken out to one or two terms, is used as an approximation of \( g(y) \).

Definition 1.3.8.8: Method of bisection

Let \( b \) be a constant and \( g \) be a continuous monotonic function from \( \mathbb{R} \) to \( \mathbb{R} \). To solve the equation \( g(y) = b \), one can employ the method of bisection, which I will illustrate for an increasing \( g \):

1. Choose \( U \) and \( L \) such that \( g(U) > b \) and \( g(L) < b \), and choose a convergence criterion \( cc > 0 \).

2. Let \( m = \frac{U + L}{2} \).

3. If \( g(m) > b \), set \( U = m \). Otherwise, set \( L = m \).

4. Repeat steps 2 and 3 until \( |g(m) - b| < cc \).

Theorem 1.3.8.8: Convergence of method of bisection

If there exist a starting \( U \) and \( L \) in Definition 1.3.8.8, the method of bisection always stops.

Proof. By the fact that the function is monotonic, the solution is always between \( U \) and \( L \) at every iteration. The distance between \( U \) and \( L \) as a function of the number \( s \) of iterations is \( \left( \frac{1}{2} \right)^s \) times a constant. This converges to zero with \( s \). Thus, \( m \) converges to the solution. Then by Theorem 1.3.8.10, in some finite step \( |g(m) - b| < cc \).

42
Chapter 1. Introduction

Theorem 1.3.8.9: Shape of $y + \frac{1}{y}$

The function $g(y) = y + \frac{1}{y}$ is $U$-shaped on the positive numbers. That is, it decreases from infinity at $y = 0$ to a minimum at $y = 1$, then increases to infinity.

Proof. The function asymptotes to infinity as it approaches zero. The first derivative is $1 - \frac{1}{y^2}$, which is negative for $y < 1$, 0 at $y = 1$, and positive for $y > 1$. Then the function decreases to a minimum at 1 and increases thereafter. \qed

Theorem 1.3.8.10: Convergence of a function of a sequence [15]

If $\lim_{n \to \infty} (Y(n)) = y$, and $g$ is continuous at $y$, then $\lim_{n \to \infty} (g(Y(n))) = g(y)$.

This theorem is so basic that it is sometimes given as the definition of continuity.

Theorem 1.3.8.11: Eigenvectors as a basis

The column space (Definition 1.3.8.6) of $P$ from Definition 1.3.8.5 is the same as the column space of $A$ from Theorem 1.3.8.5.

Proof. If $y = Az$, then $y = PDP^Tz = Pz^*$ for some $z^*$, so the column space of $A$ is contained within that of $P$. Also, $APD^{-1} = P$, so by Theorem 1.3.8.6, the column space of $P$ is also contained within that of $A$. \qed

Theorem 1.3.8.12: Nonnegativity and constrained sum imply boundedness

The set of points $\{z_1, \ldots, z_p\}$ that satisfy $\sum_{i=1}^{p} \alpha_i z_i = c$ and $z_1 \geq 0, \ldots, z_p \geq 0$, where the $\alpha_i$s are nonnegative constants, is bounded.

Proof. Suppose not and let $z_1$ shoot off to $\infty$. Because of the constraint on the sum, clearly at least one of the other terms must shoot off to negative $\infty$, so the proof follows by contradiction. \qed

Theorem 1.3.8.13: Sum of a geometric series ([2], page 31)

If $0 < t < 1$, $\sum_{i=1}^{\infty} t^{i-1} = \frac{1}{1-t}$.
Chapter 1. Introduction

Theorem 1.3.8.14: Line integral ([18], page 1081)

Consider the one-dimensional surface $S$ in $\mathbb{R}^p$ that consists of all points that can be written as $g(t)$, where $g$ is a differentiable, one-to-one function from an interval in $\mathbb{R}$ to $\mathbb{R}^p$. Then the area underneath a scalar-valued function $f$ between the points $x$ and $y$ on $S$, called the line integral of $f$ along $S$ between $x$ and $y$, is

$$\int_{g^{-1}(x)}^{g^{-1}(y)} f(g(t)) \sqrt{\sum_{i=1}^{p} \left( \frac{dg_i}{dt}(t) \right)^2} dt.$$ 

Theorem 1.3.8.15: Implicit function theorem ([19], page 202)

Let $h(x, y)$ be a function from a domain in $\mathbb{R}^{n+m}$ to a domain in $\mathbb{R}^m$, where $x$ is $n \times 1$ and $y$ is $m \times 1$. The equation $h(x, y) = 0$ often defines $y$ as an implicit function $l$ of $x$, at least locally. The implicit function theorem allows us to calculate the partial derivatives of $l$ with respect to $x$:

If $h(x^*, y^*) = 0$ and $\det \left( \frac{\partial h(x^*, y^*)}{\partial y} \right) \neq 0$, there is a unique local differentiable solution $y = l(x)$, and $\frac{\partial l(x^*)}{\partial x} = -\left( \frac{\partial h(x^*, y^*)}{\partial y} \right)^{-1} \frac{\partial h(x^*, y^*)}{\partial x}$.

In the statement of the theorem, all derivatives are assumed to exist.

Definition 1.3.8.9: The orthogonal projection matrix

Let $Z$ be an $l \times n$ matrix. The orthogonal projection matrix onto the column space (Definition 1.3.8.6) of $Z$ is $Z(Z^TZ)^gZ^T$, where $(Z^TZ)^g$ is any generalized inverse of $(Z^TZ)$. For any vector $Y$ in $\mathbb{R}^l$, $Z(Z^TZ)^gZ^TY$ is the orthogonal projection of $Y$ onto the column space of $Z$.

1.3.9 Importance Sampling

Definition 1.3.9.1: Importance sampling estimator

Suppose $X$ is a random sample. The importance sampling estimator $\hat{g}_Z$ of the expected value of a scalar-valued function $g$ of $Z$ is

$$\sum_{i=1}^{N} \frac{g(X_i)}{f_Z(X_i)} \frac{f_Z(X_i)}{f_X(X_i)}.$$ 

$f_X$ is called the generating density and $f_Z$ is the target random variable.
Chapter 1. Introduction

Theorem 1.3.9.1: Convergence of importance sampling estimator

If \( g(Z) \) is a bounded random variable, \( \frac{f_Z(x)}{f_X(x)} \) is bounded on \( \mathbb{R}^1 \), and \( E_X \left( \frac{f_Z(x)}{f_X(x)} \right) \neq 0 \), then \( \hat{g}_Z(N) \) (Definition 1.3.9.1) converges almost surely with \( N \) to \( E_Z(g(Z)) \).

Proof. For ease of notation, let all random variables be one-dimensional.

Rewrite \( \hat{g}(Z) \) as

\[
\frac{1}{N} \sum_{i=1}^{N} g(X_i) \frac{f_Z(X_i)}{f_X(X_i)} = \frac{1}{N} \sum_{i=1}^{N} \frac{f_Z(X_i)}{f_X(X_i)}.
\]

The boundedness assumptions imply that all moments exist for both numerator and denominator, by Theorem 1.3.5.5.

The numerator is the sample mean of the variable \( g(X_1) \frac{f_Z(X_1)}{f_X(X_1)} \). Then by the Strong Law of Large Number (Theorem 1.3.6.8), the numerator converges almost surely to

\[
E_X \left( g(X) \frac{f_Z(X)}{f_X(X)} \right) = \int_{\mathcal{X}} g(x) \frac{f_Z(x)}{f_X(x)} f_X(x) dx,
\]

where \( \mathcal{X} \) is the support of \( X_1 \). This integral is simply

\[
\int_{\mathcal{X}} g(x) f_Z(x) dx.
\]

Now by the fact that \( \frac{f_Z(\cdot)}{f_X(\cdot)} \) is bounded, the support of \( Z \) is contained within \( \mathcal{X} \), so the integral is simply \( E_Z(g(Z)) \).

The denominator is the sample mean of the variable \( \frac{f_Z(X)}{f_X(X)} \). Also by the Strong Law of Large Numbers, the denominator converges almost surely to

\[
E_X \left( \frac{f_Z(X)}{f_X(X)} \right) = \int_{\mathcal{X}} \frac{f_Z(x)}{f_X(x)} f_X(x) dx = \int_{\mathcal{X}} f_Z(x) dx = 1.
\]

(In the last step we need to use the fact that the support of \( Z \) is in \( \mathcal{X} \).)

Then Theorem 1.3.6.5 part 3, \( \hat{g}_Z \) converges almost surely to \( \frac{E_Z(g(Z))}{1} \).

\( \square \)
Chapter 1. Introduction

Theorem 1.3.9.2: Asymptotic normality of importance sampling estimator

If \( g(Z) \) is a bounded random variable, \( \frac{f_Z(X)}{f_X(X)} \) is bounded on \( \mathbb{R}^1 \), and \( E_X \left( \frac{f_Z(X)}{f_X(X)} \right) \neq 0 \), then \( \hat{g}_Z(N) \) (Definition 1.3.9.1) \( \sim AN \left( E_Z(g(Z)), \frac{1}{N} V \right) \), where

\[
V = E_X \left( \frac{g^2(X) f_Z^2(X)}{f_X^2(X)} \right) - 2 E_Z(g(Z)) E_X \left( \frac{g(X) f_Z^2(X)}{f_X^2(X)} \right) + E_Z(g(Z))^2 E_X \left( \frac{f_Z^2(X)}{f_X^2(X)} \right).
\]

Proof. Rewrite \( \hat{g}(Z) \) as

\[
\frac{1}{N} \sum_{i=1}^N g(X_i) \frac{f_Z(X_i)}{f_X(X_i)}.
\]

The numerator \( Num \) is the sample mean of the variable \( g(X) \frac{f_Z(X)}{f_X(X)} \). We saw in the proof of Theorem 1.3.9.1 that this has expected value \( E_Z(g(Z)) \). It’s variance is \( V_{num} \equiv E_X \left( g^2(X) \frac{f_Z^2(X)}{f_X^2(X)} \right) - E_Z(g(Z))^2 \).

The denominator \( Den \) is the sample mean of \( \frac{f_Z(X)}{f_X(X)} \). We saw in the proof of Theorem 1.3.9.1 that this has expected value \( 1 \). It’s variance is \( V_{den} \equiv E_X \left( \frac{f_Z^2(X)}{f_X^2(X)} - 1 \right) \).

The covariance of \( g(X) \frac{f_Z(X)}{f_X(X)} \) and \( \frac{f_Z(X)}{f_X(X)} \) is \( Cov \equiv E_X \left( g(X) \frac{f_Z^2(X)}{f_X^2(X)} - E_Z(g(Z)) \right) \).

The boundedness assumptions imply that all moments exist for both numerator and denominator, by Theorem 1.3.5.5. Then by the Central Limit Theorem (Theorem 1.3.6.7),

\[
\left( \begin{array}{c} Num(N) \\ Den(N) \end{array} \right) \sim AN \left( \left( \begin{array}{c} E_Z(g(Z)) \\ 1 \end{array} \right), \frac{1}{N} \left( \begin{array}{cc} V_{num} & Cov \\ Cov & V_{den} \end{array} \right) \right).
\]

Now \( \hat{g}_Z \) is a differentiable function of \( \left( \begin{array}{c} Num \\ Den \end{array} \right) \) at \( Den = 1 \). Then by the delta theorem (Theorem 1.3.6.3), \( \hat{g}_Z \) is asymptotically normal, with mean and variance given by the delta theorem. Working out those means and variances will produce the result above.

\( \square \)
A population’s CV provides a unitless measure of the variability in a population. It is useful when relative variability is of more interest than absolute variability. This can happen when the scales of measurement are arbitrary, or the populations being compared are measured in different units entirely. For example, in a diet study the variability in the ratio of total to HDL cholesterol might be compared to the variability of blood vessel diameter [20].

Also, it is frequently the case that the variability of a population is expected to increase with its mean, so that one might want to measure variability relative to the mean rather than absolute variability. For instance, the larger the average price of a stock, the larger the variance of the price day to day; the greater a person’s average weight, the greater the variance of that weight month to month; and the higher the average blood concentration of HDL cholesterol, the greater the variance of that concentration hour to hour. These common phenomena are consistent with the laws of probability, which imply that the variances of sums of independent components increase with the number of components, and that the variance of a scaled up version of a random variable is larger than that of the original random variable.

The coefficient of variation is used in chemistry and medicine as a measure of the reliability of an assay [21], in finance to quantify the riskiness of stocks [22], in clinical trials to account for baseline variability of measurements [23], in ecology to assess year-to-year variability in populations [24], in psychology in the study of choice under uncertainty [25], in speech pathology to diagnose apraxia of speech [26], in meteorology to compare rainfall variability over time [27], in physical therapy to determine sincerity of effort [28], in genetics as a measure of evolvability [29], in quality control to seek production processes with minimal dispersion [30], and in many other fields. Typing “coefficient of variation” OR ”coefficients of variation” OR ”relative standard
deviation” into the topic box in ISI Web of Science’s General Search page pulls up over 24,600 hits, which is \(\frac{2}{3}\) the number of hits (37,200) pulled up by “standard deviation.”

In this chapter, we shall be concerned about inference on the CV in normal populations.

### 2.1 Useful Facts about the CV of a Normal Population

#### 2.1.1 The CV in practice

What I shall call the “practical range” for the CV for normal populations is \(0 < CV < 0.33\). Cases where an investigator would be interested in CVs outside this range for normal data are rare.

Figure 1 is a histogram of the values of 92 CVs reported in the abstracts of a sample of 60 papers (see Chapter 4). The lion’s share of the CVs fall between 0 and 33.

In almost every case that occurs in practice where the CV is of interest, the data are necessarily positive. In the sample of 60 studies, in 57 of the studies the variable in question was necessarily positive (in the other 3, I could not understand the definition of the variable). Any normal model for positive data would have to have a negligible fraction of its density below zero. If the CV of a normal random variable is \(\frac{1}{3}\), then 0 is 3 standard deviations from the mean, and 0.13% of the density is below 0. Increasing CV increases the negative portion of the density. \(\frac{1}{3}\) is generally taken as the upper bound on CV for a normal model to be acceptable for nonnegative data. ([31], [32], [33]).

#### 2.1.2 The sample CV as a maximal invariant

\(\bar{X}, S\) is sufficient (Definition 1.3.1.1) for the parameters of a normal random sample (Theorem 1.3.1.1). Also, CV is invariant to positive scale transformations of the data (ie, multiplying \(X\) by a scalar \(\alpha > 0\) will not change CV). This is what makes it a good measure of relative variability. Lehman ([5], page 294) showed that \(\tilde{CV}\) is a maximal invariant (Definition 1.3.2.3) for scale transformations on the space of the sufficient
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

![Histogram of CVs from the Scientific Literature](image)

Figure 2.1: Histogram of CVs from the Scientific Literature

statistics. Then by the Invariance Principle (Definition 1.3.2.2), inference on CV from a normal random sample should be based solely on $\hat{CV}$.

Also, note that if $X$ is a univariate normal random sample with mean $\mu$ and coefficient of variation $CV$, then $X_i \sim N(\mu, \mu^2 CV^2)$, which means that $X = \mu Z$, where $Z$ is a random sample with $Z_i \sim N(1, CV^2)$. In other words, when the normal population is parameterized by $\{\mu, CV\}$ rather than $\{\mu, \sigma^2\}$, $\mu$ becomes a scaling factor. The distribution of the maximal invariant $\hat{CV}$ does not depend on $\mu$.

All of the inferential procedures we shall consider in this chapter satisfy the requirement of depending on $\hat{CV}$ alone. In addition to guaranteeing invariant inference, this also allows us to ignore $\mu$ in simulations, since it does not affect the size or power of
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

procedures depending only on $\hat{CV}$.

Since $CV$ is the standard deviation of a population that has been rescaled to have mean 1, a naive way to do inference on the $CV$ in a normal sample would be to scale the sample to have mean 1 by dividing each observation by $\bar{X}$, and then assume that the sample variance of the rescaled sample is $\frac{CV^2}{N-1}$ times a $\chi^2_{N-1}$ variable, as it would be if the population were known to have mean 1. Unfortunately, this does not take into account the variation in the scaling factor $\bar{X}$, and leads to incorrect inference in small samples.

2.1.3 The Distribution of $\hat{CV}$

We can write

$$\frac{1}{\hat{CV}} = \frac{\bar{X}}{S} = \frac{\mu + (\bar{X} - \mu)}{(\sigma/\sqrt{N-1})\sqrt{(N-1)S^2}} = \frac{\mu + \frac{\sigma}{\sqrt{N}}Z}{\frac{\sigma}{\sqrt{N-1}}\sqrt{U}},$$

where $Z$ is an $N(0,1)$ random variable independent of $U$, which is $\chi^2_{N-1}$.

Dividing top and bottom by $\sigma$ and multiplying by $\sqrt{N}$, we see that

$$\sqrt{N}\frac{1}{\hat{CV}} = \sqrt{N}\frac{1}{CV} \frac{1}{\sqrt{\frac{U}{N-1}}} + \frac{Z}{\sqrt{\frac{U}{N-1}}}.$$ (2.1)

Equation 2.1 was derived by Johnson and Welch [31]. From Definition 1.3.5.6, we can see that $\sqrt{N}\frac{1}{\hat{CV}}$ is a noncentral $t$ with $N - 1$ degrees of freedom and noncentrality parameter $\frac{\sqrt{N}}{CV}$.

We can also get a stochastic representation for $\hat{CV}$:

$$\hat{CV} = \frac{CV}{1 + \frac{CV}{\sqrt{N}}Z}.$$ (2.2)

Most statistical software packages have a noncentral $t$ probability function. In SAS-IML, to find $P \equiv \text{Prob}(T < a)$, where $T$ has a noncentral $t$ distribution with $\nu$ degrees of freedom and noncentrality parameter $\delta$, the function is “$P = \text{cdf}(T', a, \nu, \delta)$”. One
can use such functions or noncentral $t$ tables in this context via the following equations:

\[
\text{Prob}\left(\frac{1}{\hat{CV}} < a\right) = \Phi_{N\text{CT}}^{NCT}(\sqrt{N}a)
\]

\[
\text{Prob}(\hat{CV} < a) = 1 - \Phi_{N\text{CT}}^{NCT}(\sqrt{N}a)
\]

(2.3)

where $\Phi_{\nu,\delta}^{NCT}$ is the cdf of the noncentral $t$ with $\nu$ degrees of freedom and noncentrality parameter $\delta$.

The latter equation is valid only if there is a negligible probability that $\bar{X}$ and thus $\hat{CV}$ is negative. If $\mu > 0$ and either $CV$ is small or $N$ large, there is virtually no chance for a negative sample $CV$. Through simulations, I have verified that this is not a problem for $CV$ in the practical range; even for $CV$ as high as 0.33 and $N = 2$, $\text{Prob}(\hat{CV} < 0) = 0.00001$. If there is a non-negligible probability that $\hat{CV}$ is negative, to obtain the distribution function for $\hat{CV}$ one should take many draws of $Z$ and $U$ and calculate an empirical distribution function using Equation 2.2.

### 2.1.4 The bias of $\hat{CV}$

If we look at the stochastic representation of $\hat{CV}$ from Equation 2.2 as a function of $\frac{CV}{\sqrt{N}}Z$ and take a Taylor expansion (Definition 1.3.8.7) around the point $\frac{CV}{\sqrt{N}}Z = 0$ we get

\[
\hat{CV} = CV \sqrt{\frac{U}{N - 1}} \sum_{i=0}^{\infty} (-1)^i \frac{CV^i}{N^{\frac{i}{2}}} Z^i.
\]

Using the independence of $U$ and $Z$, we get

**Proposition 1. Expected value of $\hat{CV}$**

\[
E(\hat{CV}) \approx \gamma(CV, N)CV,
\]

where

\[
\gamma(CV, N) = \frac{\sqrt{2\Gamma\left(\frac{N}{2}\right)}}{\sqrt{N - 1}\Gamma\left(\frac{N-1}{2}\right)} \left(1 + \frac{CV^2}{N}\right).
\]
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

Table 2.1: Bias of $\hat{CV}$ as percentage of true CV

<table>
<thead>
<tr>
<th>$CV$</th>
<th>$N$</th>
<th>Proposition 1</th>
<th>simulated pct bias</th>
<th>standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>2</td>
<td>-20.1</td>
<td>-20.0</td>
<td>0.2</td>
</tr>
<tr>
<td>0.05</td>
<td>5</td>
<td>-5.9</td>
<td>-5.9</td>
<td>0.1</td>
</tr>
<tr>
<td>0.05</td>
<td>20</td>
<td>-1.3</td>
<td>-1.2</td>
<td>0.1</td>
</tr>
<tr>
<td>0.33</td>
<td>2</td>
<td>-15.9</td>
<td>-14.9</td>
<td>0.2</td>
</tr>
<tr>
<td>0.33</td>
<td>5</td>
<td>-4.0</td>
<td>-4.0</td>
<td>0.1</td>
</tr>
<tr>
<td>0.33</td>
<td>20</td>
<td>-0.8</td>
<td>-0.7</td>
<td>0.1</td>
</tr>
</tbody>
</table>

The first factor in $\gamma(CV, N)$ is $\frac{E(S)}{\sigma}$, and the second factor is approximately $E\left(\frac{\mu}{\bar{X}}\right)$. By Jentzen’s inequality, we know that $S$ and $1/\bar{X}$ are both biased for their population counterparts; these biases offset each other so that the overall bias in $\hat{CV}$ depends on the value of $CV$. For values of $CV$ in the practical range, Proposition 1 indicates the contribution of the bias in $1/\bar{X}$ will be very small.

Table 2.1.4 was calculated by averaging the percentage $\frac{\hat{CV}-CV}{CV}$ from a large number of normal samples to obtain the simulated percentage bias of $\hat{CV}$ to compare to $\gamma(CV, N) - 1$, the percentage bias from Proposition 1. The table indicates that the overall bias is downward in the practical range, is small for $N > 5$, but can be substantial for small $N$. The table also demonstrates that the approximate formula for the bias agrees well with the simulated bias.

In any normal population, there will be some small probability that the sample mean will be in a neighborhood of zero. The result of such an outcome would be a huge sample $CV$. The possibility of such huge values is such that the moments of $\hat{CV}$ do not exist, as shown by Iglewicz [34].

However, the probability of enormous values of $\hat{CV}$ is small enough for $CV$ in the practical range that the distribution function can be well-approximated by one that has all its moments, as we shall see below; the discrepancy would come only in the extreme upper tail, even with small samples. Table 2.1.4 confirms that Proposition 1
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

accurately predicts the average value for \( \hat{CV} \) over a large number of samples, and the small standard errors for the simulated bias indicate that \( \hat{CV} \) is reasonably stable.

2.2 Approximations to the Distribution of \( \hat{CV} \).

For the definitions of the different convergence concepts in Section 2.2, refer to Section 1.3.6.

2.2.1 \( \chi^2 \) approximations

By working directly with the density function of \( \hat{CV} \), McKay [33] derived

\[
\frac{1 + CV^2}{CV^2} (N - 1) M \approx \chi^2_{N-1}, \tag{2.4}
\]

where

\[
M = \frac{\hat{CV}^2}{1 + \frac{N-1}{N} CV^2}.
\]

Computations by a number of authors ([35], [36], [37], [38]) have shown that McKay’s approximation is quite good, even for \( N \) as small as 5, for \( CV \) in the practical range. Warren [39] reported results that seemed to show that McKay’s approximation was not as close as had been previously thought, but these are due to a misunderstanding of the definition of McKay’s statistic [37].

Vangel [38] created an even more accurate approximation by numerically demonstrating that

\[
\text{Prob} \left( \frac{1 + CV^2}{CV^2} (N - 1) \left( 1 + \left( \frac{N-1}{N} \left( \frac{\hat{CV}^2}{\chi^2_{N-1, \alpha}} + 1 \right) \right) \right) \hat{CV}^2 < \chi^2_{N-1, \alpha} \right) \tag{2.5}
\]

is for all intents and purposes exactly \( \alpha \) for \( N \geq 5 \) and \( 0 < CV < 0.33 \). (Here \( \chi^2_{N-1, \alpha} \) is the 100\( \alpha \)th percentile of the \( \chi^2_{N-1} \) distribution.)

Both McKay’s and Vangel’s approximations imply distribution functions for \( \hat{CV} \):
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

Proposition 2. Distribution functions from $\chi^2$ approximations

McKay’s approximation implies:

$$\text{Prob}(\hat{CV} < a) = \Phi_{\chi^2}^{N-1} \left( \frac{(N-1)^{1+CV^2}}{\frac{N-1}{N} + \frac{1}{a^2}} \right) \equiv \Phi_M(a),$$

Vangel’s approximation implies:

$$\text{Prob}(\hat{CV} < a) = \Phi_{\chi^2}^{N-1} \left( \frac{(N-1)^{1+CV^2} - 2^{N-1}}{\frac{N-1}{N} + \frac{1}{a^2}} \right) \equiv \Phi_V(a),$$

where $\Phi_{\chi^2}^{N-1}$ is the cdf for the $\chi^2$ random variable with $N-1$ degrees of freedom.

Proof. For McKay’s approximation, write

$$\text{Prob}(\hat{CV} < a) = \text{Prob}(\hat{CV}^2 < a^2) = \text{Prob} \left( \frac{1}{a^2} < \frac{1}{\hat{CV}^2} \right)$$

$$= \text{Prob} \left( \frac{1}{a^2} + \frac{N-1}{N} < \frac{1}{\hat{CV}^2} + \frac{N-1}{N} \right) = \text{Prob} \left( \frac{1}{a^2} + \frac{N-1}{N} > \frac{1}{\hat{CV}^2} + \frac{N-1}{N} \right)$$

$$= \text{Prob} \left( \frac{(N-1)^{1+CV^2}}{\frac{1}{a^2} + \frac{N-1}{N}} > (N-1) \frac{1 + CV^2}{CV^2} \frac{CV^2}{1 + \frac{N-1}{N} CV^2} \right),$$

which by McKay’s approximation is $\Phi_{\chi^2}^{N-1} \left( \frac{(N-1)^{1+CV^2}}{\frac{N-1}{N} + \frac{1}{a^2}} \right)$.

Vangel’s approximation can be written in a form slightly different Equation 2.5:

$$\text{Prob} \left( (N-1) \frac{1 + CV^2}{CV^2} \frac{\hat{CV}^2}{1 + \left( \frac{N-1}{N} \left( \frac{1}{z} + 1 \right) \right) \hat{CV}^2} \leq z \right) = \Phi_{\chi^2}^{N-1}(z).$$

Algebraic manipulations yield

$$\text{Prob} \left( \hat{CV} \leq \sqrt{\frac{1}{ \left( \frac{1+CV^2}{CV^2} (N-1) - 2^{N-1} \frac{N-1}{N} \right) \frac{1}{z} - \frac{N-1}{N}}} \right) = \Phi_{\chi^2}^{N-1}(z).$$
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

Now \( \text{Prob}(\hat{CV} \leq a) = \Phi_{\chi^2_{N-1}}(z^*) \), where \( z^* \) is the value of \( z \) that solves

\[
a = \sqrt{\frac{1}{\left( \frac{1+CV^2}{CV^2} (N - 1) - 2 \frac{N-1}{N} \right) \frac{1}{z} - \frac{N-1}{N}}}.
\]

Algebraic manipulations yield

\[
z^* = \frac{\frac{1+CV^2}{CV^2} (N - 1) - 2 \frac{N-1}{N}}{\frac{N-1}{N} + \frac{1}{\sigma^2}}.
\]

\[\square\]

2.2.2 Delta-method approximations

If the normal distribution is parameterized by \( \{\mu, \sigma\} \), \( \hat{\theta}_{MLE} = \{\bar{X}, S\} \). Then by Theorem 1.3.6.2 and Theorem 1.3.5.7,

\[
\begin{pmatrix}
\bar{X}(N) \\
S(N)
\end{pmatrix}
\sim AN
\begin{pmatrix}
\mu \\
\sigma
\end{pmatrix},
\begin{pmatrix}
\frac{\sigma^2}{N} & 0 \\
0 & \frac{\sigma^2}{2N}
\end{pmatrix}.
\]

Then by the delta method (Theorem 1.3.6.3) we can deduce

\[
\hat{CV}(N) \sim AN\left( CV, \frac{CV^2}{2N} + \frac{CV^4}{N} \right). \tag{2.6}
\]

In simulations, I found that using \( N - 1 \) rather than \( N \) in the denominator of the approximate variance gave a better approximation, so henceforth I shall use that.

The delta-method approximation was derived by Cramér [40], but may have appeared in a textbook before that. It implies an approximate cdf:

\[
\text{Prob}(\hat{CV} < a) = \Phi\left( \frac{a - CV}{\sqrt{\frac{CV^2}{2(N-1)} + \frac{CV^4}{(N-1)}}} \right) \equiv \Phi_Z(a), \tag{2.7}
\]

where \( \Phi \) is the standard normal cdf. Iglewicz and Myers [36] suggested on the basis
of numerical calculations that the delta-method approximation is reasonable even for small $N$.

One might question whether Equation 2.6 is inconsistent with the fact that the moments of $CV$ do not exist for any $N$, as explained in Section 2.1.3. There is no inconsistency, because asymptotic normality means pointwise convergence of the distribution. (see Definitions 1.3.6.3 and 1.3.6.4). For any $N$, there is enough of a discrepancy between the actual and asymptotic cdf that the moments of the actual cdf do not exist, but the discrepancy gets pushed farther and farther out into the tail as $N$ increases.

It is convenient here to note that

**Proposition 3. Consistency of sample coefficient of variation**

$$\hat{CV}(N) \to_p CV.$$  

*(see Definition 1.3.6.1.)*

**Proof.** This follows from by Equation 2.6 and Theorem 1.3.6.1 part 2.

2.2.3 Numerical evaluation of the approximations

Using Equation 2.3, I calculated the exact cdf over a range of possible values for $\hat{CV}$ for various combinations of $CV$ and $N$ and compared them to $\Phi_M$, $\Phi_V$, and $\Phi_Z$.

Table 2.2 illustrates two findings, that $\Phi_M$ and $\Phi_V$ are highly accurate and that $\Phi_V$ has accuracy of a higher order of magnitude. ($CV = 0.33$ actually minimizes the accuracy of the approximations over the practical range.)

Figure 2.2 and Figure 2.3 compare $\Phi_V$ and $\Phi_Z$ to the exact cdf. We see that for small values of $N$, there is substantial skewness in the distribution of $\hat{CV}$, so that $\Phi_Z$ is inadequate. Nonetheless, for $N = 10$ the delta method approximation appears to come close in the tails. Most importantly, $\Phi_V$ is indistinguishable from the exact cdf even for values of $N$ as small as 2.

The findings in Table and the figures hold for all values of $CV$ in the practical range.
Table 2.2: Exact cdf minus approximate cdf, \( CV = 0.33, \) \( N = 10 \)

<table>
<thead>
<tr>
<th>Exact cdf</th>
<th>( \Phi_M )</th>
<th>( \Phi_V )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>(-8 \times 10^{-4})</td>
<td>(-8 \times 10^{-5})</td>
</tr>
<tr>
<td>0.05</td>
<td>(-4 \times 10^{-3})</td>
<td>(-2 \times 10^{-4})</td>
</tr>
<tr>
<td>0.10</td>
<td>(-6 \times 10^{-3})</td>
<td>(-3 \times 10^{-4})</td>
</tr>
<tr>
<td>0.25</td>
<td>(-1 \times 10^{-2})</td>
<td>(-3 \times 10^{-4})</td>
</tr>
<tr>
<td>0.50</td>
<td>(-2 \times 10^{-2})</td>
<td>(4 \times 10^{-4})</td>
</tr>
<tr>
<td>0.75</td>
<td>(-1 \times 10^{-2})</td>
<td>(1 \times 10^{-3})</td>
</tr>
<tr>
<td>0.90</td>
<td>(-7 \times 10^{-3})</td>
<td>(1 \times 10^{-3})</td>
</tr>
<tr>
<td>0.95</td>
<td>(-4 \times 10^{-3})</td>
<td>(1 \times 10^{-3})</td>
</tr>
<tr>
<td>0.99</td>
<td>(-1 \times 10^{-3})</td>
<td>(5 \times 10^{-4})</td>
</tr>
</tbody>
</table>

### 2.2.4 An exponential family model for inference on \( CV \)

With some algebra we can derive

\[
Prob(M < m) = Prob \left( \sqrt{\frac{1}{m}} \leq \frac{1}{N-1} \frac{1}{N} \right).
\]

Using \( \Phi_V \) as an approximate cdf for \( \hat{CV} \), we get

\[
Prob(M < m) \approx \Phi_{\chi^2} \left( \left( 1 + \frac{CV^2}{CV^2} (N-1) - 2 \frac{N-1}{N} \right) m \right).
\]

Recalling that \( U \sim \chi^2_{N-1} \), we can write this as

\[
Prob(M < m) \approx Prob \left( U < \left( \frac{1 + CV^2}{CV^2} (N-1) - 2 \frac{N-1}{N} \right) m \right)
\]

\[
= Prob \left( \frac{CV^2}{1 + \frac{N-1}{N} CV^2} U < m \right),
\]

57
so that $\Phi_V$ yields a stochastic representation for $M$:

$$M \approx \frac{CV^2}{1 + \frac{N-2}{N} CV^2} \frac{U}{N-1}. \tag{2.8}$$

From this we can derive the density of $M$ using Theorem 1.3.5.1 and Theorem 1.3.5.8.

**Proposition 4.** An exponential family for $\hat{CV}$

$$f_M(m) \approx \phi_M(m) \equiv C_M(\theta)H_M(m) \exp(\theta T(m)),$$
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

Figure 2.3: Comparison of exact cdf, \( \Phi_V \), and \( \Phi_Z \)

where

\[
\theta = \frac{1 + CV^2}{CV^2},
\]

\[
C_M(\theta) = \frac{1}{\Gamma\left(\frac{N-1}{2}\right)} \left( (N-1) \frac{\theta - \frac{2}{N}}{2} \right)^{\frac{N-1}{2}},
\]

\[
H_M(m) = I_{m \geq 0} m^{\frac{N-1}{2}-1} \exp \left( \frac{N-1}{N} m \right),
\]

\[
T(m) = -\frac{N-1}{2} m.
\]
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

Since $\Phi_V$ is a very accurate approximation for the cdf of $\widehat{CV}$, the resulting approximation for the cdf of $M$ should be very accurate and so should the probability models in Equation 2.8 and Proposition 4. Since $M$ is a monotonically increasing function of $\widehat{CV}$ and thus a one-to-one function, it is a maximal invariant (Definition 1.3.2.3) for scale transformations, and by the Invariance Principle (Definition 1.3.2.2), all inference on $CV$ can be based on it. Furthermore, $\theta$ is a monotonically decreasing function of $CV$, so all inferential questions concerning $CV$ can be based on inference concerning $\theta$.

$\phi_M$ belongs to an exponential family (Definition 1.3.4.1), which will allow us to apply the associated theory to inference on $CV$.

2.2.5 Asymptotic comparison of approximations

Vangel [38] claims that his approximation and McKay’s are asymptotically exact, but a formal analysis has not been done.

Rewrite Equation 2.4 and Equation 2.8 as

$$\frac{1 + CV^2}{CV^2}(N - 1)M(N) \approx U(N),$$

$$\left(\frac{1 + CV^2}{CV^2} - \frac{2}{N}\right)(N - 1)M(N) \approx U(N)$$

where the $N$ subscript simply emphasizes the fact that $M$ is a random sequence indexed by $N$.

From Equation 2.2,

$$\left(\frac{1 + CV^2}{CV^2} - \frac{c}{N}\right)(N - 1)M(N) =$$

$$\frac{1 + CV^2}{CV^2} \left(1 + \frac{CV}{\sqrt{N}}Z\right)^2 + \frac{CV^2U(N)}{N}$$

$$- \frac{c}{N} \left(1 + \frac{CV}{\sqrt{N}}Z\right)^2 + \frac{CV^2U(N)}{N}.$$
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

where \( c \) is any constant. Using algebra, Theorem 1.3.6.5, and Theorem 1.3.6.12, we can show that

**Proposition 5.** Asymptotic behavior of McKay’s and Vangel’s approximations

\[
\left( \frac{1 + CV^2}{CV^2} - \frac{c}{N} \right) \frac{(N - 1)M(N)}{U(N)} \to_{as} 1
\]

Thus, in some sense, both Equation 2.4 and Equation 2.8 are asymptotically justified.

We might be drawn to the conclusion, from Equation 2.8 and Theorem 1.3.6.12, that

\[ M(N) \sim AN \left( \frac{CV^2}{1 + CV^2}, \frac{CV^4}{(1 + CV^2)^2} \right). \quad (2.9) \]

However, Proposition 5 is about the ratio of 2 random variables that are both functions of \( N \); it does not give us a function of \( M(N) \) that converges to a stationary distribution. We can get a convergence in distribution result for \( M(N) \) from the delta theorem (Theorem 1.3.6.3):

\[ M(N) \sim AN \left( \frac{CV^2}{1 + CV^2}, \frac{CV^4 + 2CV^6}{(1 + CV^2)^4} \right). \quad (2.10) \]

So Equation 2.9 and by extension Equation 2.8 actually does not lead to the right inference asymptotically, but it is very close. The ratio of the true asymptotic variance to that in Equation 2.9 is \( \frac{1+2CV^2}{(1+CV^2)^2} \), which is bounded below by 0.99 for \( CV \) in the practical range.

### 2.3 Inference on a common \( CV \)

The first problem we shall apply the exponential family model to is inference on the common \( CV \) of populations with possibly different means.

Hypothesis tests of \( CV = CV_o \) against either one-sided or two-sided alternatives are occasionally useful.
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

These tests are relevant when previous investigations have provided expectations about what the $CV$ should be. For instance, statisticians have come up with ways to use prior knowledge of the $CV$ to improve the efficiency of estimation ([41], [42], [43], [44], [45]). Naturally, one would want to test the validity of any assumptions made about $CV$ for a certain population.

A variable that is logically non-negative should have a small $CV$ in order to be modelable as a normal variable, preferably with $CV < \frac{1}{3}$, as we have seen. This means that a hypothesis test of the form $H_0 : CV = \frac{1}{3}$ against $H_a : CV > \frac{1}{3}$ can be a quick check of normality for positive variables.

Another example of the usefulness of hypothesis tests on $CV$ is in verifying that variability meets a certain standard. Quan and Shih [46] report that for biochemical assays, $CV$'s in the 0.10 to 0.20 range are considered “good”. Bohidar and Bohidar [47] suggest a test for the purpose of certifying the content uniformity of pharmaceuticals. To pass the requirement, a population of dosage units needs to have a $CV$ for the active ingredients that is “demonstrably less than 6%.” This can be done by rejecting the null $CV = 0.06$ against the alternative $CV < 0.06$.

Confidence intervals provide more information than hypothesis tests, and give the investigator an idea of the possible range of the relative variability in the populations he is working with. Confidence intervals for $CV$ can be used in sensitivity analyses for statistical methods that assume a value for $CV$ or for methods, as in [48], [49], [50], and [51], whose performance depends on the $CV$. Also, $CV$'s are used in sample-size planning for future studies [52], so confidence intervals can help provide upper and lower bounds for sample sizes.

It is frequently the case that the $CV$ we wish to do inference on is the common $CV$ of populations with different means. For instance, researchers might want to conduct inference on the $CV$ of a certain type of variable using available data from different studies, or there might be more than one observation on each subject in the same study. The latter situation arises frequently in medical studies, where the $CV$ for the distribution of a subject’s repeated measure is called the “within-subject coefficient of variation” ($WSCV$). In such studies, often the sample sizes (number of observations per subject) are small, while the number of subjects is still large enough to allow for
powerful inference; inferential procedures for the common \( CV \) need to work effectively in this case.

### 2.3.1 Notation

In the remainder of Section 2.3, \( \mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_k\} \), where \( \mathbf{X}_i \) is an \( N_i \times 1 \) vector of \( iid \) normal random variables with mean \( \mu_i \) and common coefficient of variation \( CV \); The data can be thought of as arising from a one-way ANOVA setup where the cell means are considered fixed effects. \( \hat{CV}_i \) will refer to the \( i \)th sample coefficient of variation, and \( M_i = \frac{CV_i^2}{1 + \frac{2}{N_i - 1} CV_i^2} \). Since these maximal invariants are invariant to the cell means, the analysis is applicable to random effects models as well.

The constant \( \alpha \) will refer to the size of a hypothesis test or to 1 minus the confidence level of a confidence interval. If the \( N_i \)s are equal, the common value will be called \( N^* \). \( Z_i \) and \( U_i \) will refer to the \( i \)th sample’s value for the unobserved variables \( Z \) and \( U \) from Equation 2.2.

### 2.3.2 Previous literature

There has been surprisingly little work on inference for a common \( CV \). Zeigler \[53\] explores the bias, variance, and mean squared error of various point estimates of the common \( CV \), and Chow and Tse \[54\] consider these issues in random effects models. Neither treats the issue of how to construct confidence intervals or hypothesis tests.

Both Quan and Shih \[46\] and Tian \[55\] assume the random effects model \( \mu_i = \mu + \alpha_i \), where \( \alpha_i \sim N(0, \sigma_\alpha^2) \), and that \( \sigma_i \) is equal to the constant \( \sigma \). They call \( \frac{\sigma}{\mu} \) the “within-subject coefficient of variation” and study inference on this parameter. This is not how \( WSCV \) is defined above, or how it is defined by Chow and Tse \[54\]. The \( WSCV \) thus defined is not a common \( CV \); in fact the \( CV \)s differ across the populations in this setup, while the variances are the same. It’s questionable whether the \( WSCV \) of Quan and Shih would be considered relevant to researchers, because if researchers are interested in the \( CV \) at all, they would usually not be willing to assume constant variance.

The only paper that studies hypothesis testing and confidence intervals for a com-
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

mon $CV$ is Tian [20], which uses fiducial inference. (see Definition 1.3.3.5.) Using Equation 2.2, Weerahandi [56] derived

$$\frac{1}{CV} = \frac{1}{CV_i} \sqrt{\frac{U_i}{N_i - 1}} - \frac{Z_i}{\sqrt{N_i}}.$$  

From this, Tian [20] observed that

$$CV = \frac{CV_i}{\sqrt{\frac{U_i}{N_i - 1} - CV_i} \frac{Z_i}{\sqrt{N_i}}}.$$  

and thus

$$CV = \frac{1}{\sum_{i=1}^{k} (N_i - 1)} \sum_{i=1}^{k} (N_i - 1) \frac{CV_i}{\sqrt{\frac{U_i}{N_i - 1} - CV_i} \frac{Z_i}{\sqrt{N_i}}}.$$  \hspace{1cm} (2.11)

The fiducial inference approach treats the $CV_i$s as fixed quantities, and $CV$ as a random variable whose distribution can be calculated by taking repeated draws of the $Z_i$s and $U_i$s.

As explained by Tian, one can conduct a hypothesis test of $H_0: CV = CV_o$ against $H_a: CV > CV_o$ using this approach:

1. Calculate the $CV_i$s.

2. Generate a large number of the vectors $\{Z_1, \ldots, Z_k\}$ and $\{U_1, \ldots U_k\}$.

3. For each draw of $Z_i$s and $U_i$s, calculate a value of the right-hand side of Equation 2.11, treating the $CV_i$s as fixed. This will give you a set of “draws” of $CV$.

4. The $p$-value of the test is the proportion of the draws that are less than $CV_o$.

Tian erroneously reports in her paper that the $p$-value is the proportion of draws that are greater than $CV_o$. For a test with power against $H_a: CV < CV_o$, the $p$-value would be the proportion of draws that are greater than $CV_o$. For a two-sided test, the $p$-value is 2 times the $min$ of the proportion of draws less than $CV_o$ and the proportion of draws greater than $CV_o$. 

64
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

To construct a $1 - \alpha$ confidence interval, take repeated draws of the $U_i$s and $Z_i$s. The $\frac{\alpha}{2}$th and $1 - \frac{\alpha}{2}$th percentiles of the right hand side of Equation 2.11 are the lower and upper bounds for a fiducial interval.

Fiducial inference is not guaranteed to give exact tests and confidence intervals. The true significance level of the test needs to be evaluated by simulation.

Fiducial inference was introduced by Fisher [57]. Tian does not refer to her approach as “fiducial” inference but rather as “generalized” inference based on the “generalized” pivot $CV - \frac{1}{\sum_{i=1}^{k} (N_i - 1)} \sum_{i=1}^{k} (N_i - 1) \frac{\hat{CV}_i}{\sqrt{\frac{1}{N_i - 1} - CV_i}} \cdot \frac{Z_i}{\sqrt{\frac{1}{N_i}}}$. Generalized inference was introduced by Tsui and Weerahandi [58]. Hannig et al. [59] later pointed out that generalized inference is essentially identical to fiducial inference.

Naturally, in order to do inference on a common $CV$, one would have to assume that the $CV$s are equal. If the assumption is in doubt, one way to proceed is by first conducting an hypothesis test of the assumption, and then either accepting the assumption as fact if the test fails to reject, or treating each $CV$ as unique if the test rejects. Alternative ways of “shrinking” individual $\hat{CV}_i$s toward a pooled estimate based on the strength of evidence from the test are suggested by Ahmed and co-authors ([60], [61], [62]), who explore the bias and variance of the resulting estimators under various conditions.

2.3.3 Inference based on the normal approximation

Equation 2.6 suggests that a weighted average

$$\bar{CV} = \frac{\sum_{i=1}^{k} (N_i - 1) \hat{CV}_i}{\sum_{i=1}^{k} (N_i - 1)}$$  \hspace{1cm} (2.12)

should be approximately normal. Here I weight by degrees of freedom $N_i - 1$ rather than sample size $N_i$, since the information we have about variation in the sample comes from differences from the sample mean rather than from individual observations. In order to avoid creating tests that are biased in small $N$, large $k$ situations, we could bias-correct by multiplying $\hat{CV}_i$ by $B_i = \frac{\sqrt{N_i - 1} \Gamma(\frac{N_i - 1}{2})}{\sqrt{2} \Gamma(\frac{k}{2})}$. This term corrects for the bias induced by the fact that $S$ tends to underestimate $\sigma$, which accounts for most of the
bias in the sample \( CV \). This bias correction factor converges to 1 with \( N_i \) (Theorem 1.3.6.15).

The weighted average with the adjusted weights is

\[
\hat{CV} \equiv \frac{\sum_{i=1}^{k} (N_i - 1) B_i \hat{C}V_i}{\sum_{i=1}^{k} (N_i - 1)}.
\]  

(2.13)

We could further bias adjust by multiplying by the additional factor \( C_i = \frac{1}{1 + \frac{CV^2}{N_i}} \), to correct for the second factor in the bias coefficient in Proposition 1.

To test the hypothesis that \( CV = CV_o \), one would use a standardized, bias-corrected weighted average

\[
Z_{\text{common}} \equiv \frac{\sum_{i=1}^{k} (N_i - 1) (B_i C_i \hat{C}V_i - CV_o)}{\sqrt{\sum_{i=1}^{k} (N_i - 1)^2 B_i^2 C_i^2 \left( \frac{\hat{C}V^2}{2(N_i-1)} + \frac{\hat{C}V^4}{(N_i-1)} \right)}}.
\]  

(2.14)

Since \( CV \) is known under the null, in this expression we would set \( C_i = \frac{1}{1 + CV^2/N_i} \). Since we have subtracted off the mean and standardized by an approximate variance, this statistic will be approximately \( N(0, 1) \). Then the \( p \)-values for the test statistic in Equation 2.14 are approximately \( \Phi(Z_{\text{common}}) \) for \( H_a : CV < CV_o \), \( 1 - \Phi(Z_{\text{common}}) \) for \( H_a : CV > CV_o \), and \( 2\min(\Phi(Z_{\text{common}}), 1 - \Phi(Z_{\text{common}})) \) for \( H_a : CV \neq CV_o \).

For a two-sided confidence interval for \( CV \), we can make use of the approximate pivotal quantity

\[
\frac{\sum_{i=1}^{k} (N_i - 1) (B_i C_i \hat{C}V_i - CV)}{\sqrt{\sum_{i=1}^{k} (N_i - 1)^2 B_i^2 C_i^2 \left( \frac{\hat{C}V^2}{2(N_i-1)} + \frac{\hat{C}V^4}{(N_i-1)} \right)}} = Z_{\gamma} \pm \frac{Z_{\gamma}}{\sqrt{\sum_{i=1}^{k} (N_i - 1)^2 B_i^2 C_i^2 \left( \frac{\hat{C}V^2}{2(N_i-1)} + \frac{\hat{C}V^4}{(N_i-1)} \right)}}Z_{1-\frac{\gamma}{2}},
\]  

(2.15)

which uses an estimate of the asymptotic variance of \( \hat{C}V_i \) in place of the true variance. This will yield the approximate confidence interval

\[
\frac{\sum_{i=1}^{k} (N_i - 1) B_i C_i \hat{C}V_i}{\sum_{i=1}^{k} (N_i - 1)} \pm \frac{\sqrt{\sum_{i=1}^{k} (N_i - 1)^2 B_i^2 C_i^2 \left( \frac{\hat{C}V^2}{2(N_i-1)} + \frac{\hat{C}V^4}{(N_i-1)} \right)}}{\sum_{i=1}^{k} (N_i - 1)}Z_{1-\frac{\gamma}{2}},
\]  

(2.16)

where \( Z_{\gamma} \) is the \( \gamma \)th percentile of the standard normal distribution. The confidence
level for the confidence interval in Equation 2.16 is approximately $1 - \alpha$.

To analyze the asymptotics of this approach, let the sample sizes be equal for simplicity. With sample sizes equal, $B_1 = \ldots = B_k = B$, and $C_1 = \ldots = C_k = C$.

We can write the approximate pivotal quantity in Equation 2.15 as

$$P = \sqrt{k(N^* - 1)} \frac{1}{BC \sqrt{\left(\frac{\hat{CV}^2}{2} + \hat{CV}^4\right)}} \frac{1}{k} \sum_{i=1}^{k} (BC\hat{CV}_i - CV).$$

By Theorem 1.3.6.15, $B(N) \rightarrow 1$. Also, by Proposition 3 and Theorem 1.3.6.5 part 1, $C(N) \rightarrow_p 1$. So for simplicity we will simply set them equal to 1 for the large $N^*$ case.

Equation 2.6, Proposition 3, and the Slutsky-delta theorem (Theorem 1.3.6.6) imply that $i$th term $D_i \equiv \sqrt{N^* - 1} \frac{\hat{CV}_i - CV}{\sqrt{\left(\frac{\hat{CV}^2}{2} + \hat{CV}^4\right)}}$ converges in distribution with $N^*$ to a $N(0, 1)$. By Theorem 1.3.6.5 part 2 and Theorem 1.3.5.9, $\frac{1}{k} \sum_{i=1}^{k} D_i$ will converge in distribution with $N^*$ to a $N(0, 1)$ variable.

Now thinking of $P$ as a sequence indexed by $k$ rather than $N^*$, the last factor $\frac{1}{k} \sum_{i=1}^{k} (BC\hat{CV}_i - CV)$ is a sample mean and thus should approach normality as $k$ grows. We cannot technically use the Central Limit Theorem (1.3.6.7) to prove it, however, since the moments of $\hat{CV}$ do not exist. Similarly, with sample sizes equal, from Equation 2.13, $\hat{CV}$ is a sample mean of $k$ sample coefficients of variation, so it’s density should collapse around a constant as $k$ grows, although we cannot use the Strong Law of Large Numbers (Theorem 1.3.6.8) to give a technical proof. Then Theorem 1.3.6.5, part 1, indicates that $\sqrt{\left(\frac{\hat{CV}^2}{2} + \hat{CV}^4\right)}$ should become highly concentrated around a constant. Then the definition of asymptotic normality (Definition 1.3.6.4), Slutsky’s Theorem (1.3.6.4), and Theorem 1.3.5.9, indicate that $P(k)$ should be approximately normal with large $k$. The mean will be close to zero, because we have corrected for bias, but the variance will not in general be 1, because in small samples the delta-method expression for the variance is not correct.

The implication of this analysis is that the confidence interval in Equation 2.16 will be valid for large $N^*$, but for small $N^*$, even for large $k$, may either under or over
2.3.4 Inference based on the $\chi^2$ approximation

With $k$ samples, since $CV_i$ is invariant to transformations that multiply each $X_i$ by a nonnegative scalar $\beta_i$, inference should be based on a maximal invariant (Definition 1.3.2.3) to that group of transformations, which we have seen is $\mathbf{M} = \{M_1, \ldots, M_k\}$. The distribution of this statistic depends on only one parameter, $CV$. For simplicity of exposition, in the rest of Section 2.3.4 I shall speak of Proposition 4 as if it were exact. It implies a density $\phi_{\mathbf{M}}$ for $\mathbf{M}$:

$$\phi_{\mathbf{M}}(\mathbf{m}) = C_{\mathbf{M}}(\theta)H_{\mathbf{M}}(\mathbf{m}) \exp(\theta T(\mathbf{m})), \tag{2.17}$$

where

$$\theta = \frac{1 + CV^2}{CV^2},$$

$$C_{\mathbf{M}}(\theta) = \prod_{i=1}^{k} \frac{1}{\Gamma\left(\frac{N_i-1}{2}\right)} \left(\frac{\theta - \frac{2}{N_i}}{2}\right)^{\frac{N_i-1}{2}},$$

$$H_{\mathbf{M}}(\mathbf{m}) = I_{m_1 \geq 0, \ldots, m_k \geq 0} \prod_{i=1}^{k} m_i^{\frac{N_i-1}{2}-1} \exp\left(\sum_{i=1}^{k} \frac{N_i - 1}{N_i} m_i\right),$$

$$T(\mathbf{m}) = -\sum_{i=1}^{k} \frac{N_i - 1}{2} m_i.$$

**Proposition 6.** Uniformly most powerful invariant test

The UMP invariant test (Definition 1.3.3.3 and Definition 1.3.2.4) of $H_0 : CV = CV_o$ against $H_a : CV > CV_o$ has rejection region

$$\sum_{i=1}^{k} (N_i - 1)M_i > b_u,$$

where $\text{Prob}_{CV_o}\left(\sum_{i=1}^{k} (N_i - 1)M_i > b_u\right) = \alpha$, and the UMP invariant test of $H_0$ :
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

CV = CV₀ against Hₐ : CV < CV₀ has rejection region

\[ \sum_{i=1}^{k} (N_i - 1)M_i < b_t, \]

where \( \text{Prob}_{CV₀} \left( \sum_{i=1}^{k} (N_i - 1)M_i < b_t \right) = \alpha. \)

Proof. I will prove the latter statement.

From Definition 1.3.4.1 and Theorem 1.3.4.4, the UMP invariant test of \( H_0 : \theta = \theta_0 \) against \( H_a : \theta > \theta_0 \) is of the form \( T(M) > b^*_u \), where \( T \) is taken from Equation 2.17. Multiply both sides of this inequality by \(-2\) to get a rejection region of the form \( \sum_{i=1}^{k} (N_i - 1)M_i < b^{**}_u \).

Since \( \theta \) is a monotonically decreasing function of \( CV \), the result we are trying to prove follows from Theorem 1.3.3.2.

For a two-sided test, we shall employ the rejection region

\[ \sum_{i=1}^{k} (N_i - 1)M_i > b_{2u} \cup \sum_{i=1}^{k} (N_i - 1)M_i < b_{2l}, \]

where

\[ \text{Prob}_{CV₀} \left( \sum_{i=1}^{k} (N_i - 1)M_i > b_{2u} \right) = \frac{\alpha}{2}, \text{Prob}_{CV₀} \left( \sum_{i=1}^{k} (N_i - 1)M_i < b_{2l} \right) = \frac{\alpha}{2}. \]

Now the power of a two-sided test against an alternative of interest is approximately the probability under this alternative that the statistic is in the part of the rejection region on the same side as the alternative. So Proposition 6 indicates that the two-sided test above should have good power among two-sided tests with an equal probability of being in either tail. However, the two-sided test above is not unbiased (Definition 1.3.2.2), so for detecting alternatives very near the null this test would not be optimal.

The flip side of UMP tests is uniformly most accurate confidence intervals (Definition 1.3.3.4). Theorem 1.3.3.3 and Proposition 6 imply that inverting the one-sided tests above would give UMA upper and lower confidence bounds for \( CV \). Inverting
the two-sided test should also lead to accurate intervals, though they would not be unbiased.

Theorem 1.3.4.5 gives us a way to find a UMP unbiased test based on \( \sum_{i=1}^{k} (N_i - 1)M_i \), by allowing the probability of rejection to differ across the tails. While unbiasedness is a good property, the property of equal probabilities of rejection in both tails is also a good property; it creates confidence intervals for which the probability of underestimating the parameter is the same as the probability of overestimating the parameter.

To obtain p-values, Equation 2.8 gives us
\[
\sum_{i=1}^{k} (N_i - 1)M_i = \sum_{i=1}^{k} \frac{CV_o^2}{1 + \frac{N_i - 2}{N_i} CV_o^2} U_i. \tag{2.18}
\]

If the sample sizes are equal,
\[
\sum_{i=1}^{k} (N^* - 1)M_i = \frac{CV_o^2}{1 + \frac{N^* - 2}{N^*} CV_o^2} \sum_{i=1}^{k} U_i = \frac{CV_o^2}{1 + \frac{N^* - 2}{N^*} CV_o^2} U^*, \tag{2.19}
\]
where \( U^* \sim \chi^2_{k(N^*-1)} \). This is convenient for calculating p-values.

**Proposition 7.** P-values for UMP test with equal sample sizes.

For \( H_a : CV < CV_o \), the p-value for the UMP test is
\[
\Phi^{\chi^2_{k(N^*-1)}} \left( \left( 1 + \frac{CV_o^2}{CV_a^2} \right) - \frac{2}{N^*} \right) \sum_{i=1}^{k} (N^* - 1)m_i. \tag{2.18a}
\]

For \( H_a : CV > CV_o \), the p-value for the UMP test is
\[
1 - \Phi^{\chi^2_{k(N^*-1)}} \left( \left( 1 + \frac{CV_o^2}{CV_a^2} \right) - \frac{2}{N^*} \right) \sum_{i=1}^{k} (N^* - 1)m_i. \tag{2.18b}
\]

For \( H_a : CV \neq CV_o \), the p-value for the UMP test is 2 times the min of
\[
\Phi^{\chi^2_{k(N^*-1)}} \left( \left( 1 + \frac{CV_o^2}{CV_a^2} \right) - \frac{2}{N^*} \right) \sum_{i=1}^{k} (N^* - 1)m_i. \tag{2.18c}
\]
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

\[ 1 - \Phi_{k(N^*)-1} \left( \left( \frac{1 + CV_o^2}{CV_o^2} - \frac{2}{N^*} \right) \sum_{i=1}^{k} (N^*-1) m_i \right). \]

If the sample sizes are not equal, one either sacrifices convenience or accuracy to calculate the \( p \)-values. Section 3.5 presents some convenient approaches that may sacrifice accuracy. One can obtain accurate Monte Carlo \( p \)-values (Definition 1.3.3.7) in the following way:

**Algorithm 1.** \( p \)-values for UMP test with unequal sample sizes.

1. Take \( s \) independent draws of the vectors \( \{Z_1, \ldots, Z_k\} \) and \( \{U_1, \ldots, U_k\} \).
2. Substitute \( CV_o \) in for \( CV \) in Equation 2.2, and use that equation along with the randomly drawn \( Z_i \)s and \( U_i \)s to calculate \( s \) independent draws of the vector \( \{\hat{CV}_1, \ldots, \hat{CV}_k\} \).
3. From the randomly drawn \( \hat{CV}_i \)s, calculate \( s \) independent draws of the vector \( \{M_1, \ldots, M_k\} \).
4. Calculate \( s \) independent draws of \( \sum_{i=1}^{k} (N_i - 1) M_i \).
5. Record the number of times \( NME \) an independent draw of \( \sum_{i=1}^{k} (N_i - 1) M_i \) is more extreme than the observed value.
6. The \( p \)-value is approximately \( \frac{NME}{s} \).

If the sample sizes are equal, we have a convenient way to create confidence intervals. From Equation 2.19,

\[
\text{Prob} \left( \chi^2_{(N^*-1)k, 1-\frac{\alpha}{2}} < \left( \frac{1 + CV_o^2}{CV_o^2} - \frac{2}{N^*} \right) (N^* - 1) \sum_{i=1}^{k} M_i < \chi^2_{(N^*-1)k, 1-\frac{\alpha}{2}} \right) = 1 - \alpha.
\]

From this, a \((1 - \alpha)\) interval for \( CV \) can be calculated:

**Proposition 8.** Confidence intervals from UMP test inversion, equal sample sizes
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

The interval
\[
\sqrt{\frac{\sum_{i=1}^{k} (N_i - 1) M_i}{\chi^2(N^* - 1, k, 1 - \alpha / 2) \sum_{i=1}^{k} (N_i - 1) M_i}},
\]
\[
\sqrt{\frac{\sum_{i=1}^{k} (N_i - 1) M_i}{\chi^2(N^* - 1, k, \alpha / 2) \sum_{i=1}^{k} (N_i - 1) M_i}},
\]

is a 1 − α confidence interval for CV.

To create confidence intervals with differing sample sizes, we again have a tradeoff between convenience and accuracy. Convenient methods are described in Section 3.5. Here I shall describe a simulation-based approach. I shall describe how to compute an upper confidence bound, a value UB that satisfies \( \text{Prob}_{UB} \left( \sum_{i=1}^{k} (N_i - 1) M_i < O \right) = \alpha \), where \( O \) is the observed value of the test statistic. A lower bound would be computed analogously, and for a two-sided interval, upper and lower bounds associated with confidence level \( \alpha / 2 \) would be computed.

Algorithm 2. Confidence intervals from UMP test inversion, unequal sample sizes

1. Take s independent draws of the vectors \( \{Z_1, \ldots, Z_k\} \) and \( \{U_1, \ldots, U_k\} \).
2. Set \( UB_u \) equal to a number that is higher than thought possible for the upper confidence bound to be.
3. Set \( UB_l \) equal to a number that is lower than thought possible for the upper confidence bound to be.
4. Determine a convergence criterion \( cc \).
5. Set midpoint = \( \frac{UB_u + UB_l}{2} \).
6. Calculate the p-value \( p_{mid} \) for a test of \( H_o : CV = \text{midpoint} \) against \( H_a : CV < \text{midpoint} \) via steps 2 through 6 of Algorithm 1.
7. If |\( p_{mid} - \alpha \)| < \( cc \), stop and set \( UB = \text{midpoint} \).
8. If \( p_{mid} < \alpha \), then set \( UB_u = \text{midpoint} \). Otherwise set \( UB_l = \text{midpoint} \).
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

9. Repeat steps 5 through 8 until a stop occurs.

With $s$ large enough the $p$-value calculated in step 6 is essentially a continuous decreasing function of midpoint via Equation 2.2; then by Theorem 1.3.8.8 the method of bisection in steps 5 through 8 will converge. (The $p$-value is actually a discrete random variable, but as long as $cc$ is not set ridiculously small we can ignore this technicality.) The confidence level of the resulting interval can be made as close as desired to $\alpha$ by setting $s$ high enough and $cc$ low enough. Notice that one does not need to repeat step 1 of Algorithm 1 in order to carry out step 6 of Algorithm 2; one can simply store the values of the $Z_i$s and $U_i$s in a vector to allow rapid computation of the $p$-value as a function of $CV$.

One should keep in mind that the results of Section 2.3.4 are approximate because they are based on Proposition 4, which depends on the accuracy of $\Phi_V$. But since we have seen that $\Phi_V$ is a nearly exact cdf, the procedures in this section should yield accurate and powerful inference.

One could also use asymptotics to calculate approximate $p$-values. For example, let $T^* \equiv \frac{\sum_{i=1}^{k}(N^*-1)M_i}{k(N^*-1)} = \frac{1}{k} \sum_{i=1}^{k} M_i$, where we have assumed constant sample sizes for simplicity. From Equation 2.10 above we can deduce

$$T^*(N^*) \sim AN\left(\frac{CV^2}{1 + CV^2}, \frac{CV^4 + 2CV^6}{(1 + CV^2)^4 \cdot kN^*}\right). \quad (2.20)$$

2.3.5 Simulation comparison of two-sided confidence intervals for $CV$

Table 2.3 and Table 2.4 display simulation results for two-sided 95% confidence intervals, for the fiducial interval of Tian, the approximate normal interval, and the interval based on the inversion of the $UMP$ test. In each simulation, 10,000 sets of $k$ populations with $N^*$ observations each were generated from normal distributions with a common $CV$. Recall that since all of the procedures depend on the data only through the $\hat{CV}_i$s, their performance is invariant to the population means.

The $UMP$ interval is the best overall interval, maintaining a coverage probability very close to nominal even for small $N^*$ and $k$ and for high $CV$, but having favorable
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

Table 2.3: Coverage probability (CP) and width (W) of confidence intervals for \( CV = 0.05 \). \( SE = 0.002 \) for CP, < 0.001 for width.

<table>
<thead>
<tr>
<th>( N^* )</th>
<th>( k )</th>
<th>CP fid</th>
<th>CP norm</th>
<th>CP UMP</th>
<th>W fid</th>
<th>W norm</th>
<th>W UMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>0.999</td>
<td>0.848</td>
<td>0.952</td>
<td>1.042</td>
<td>0.123</td>
<td>0.255</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>1.00</td>
<td>0.902</td>
<td>0.951</td>
<td>1.207</td>
<td>0.087</td>
<td>0.107</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>1.00</td>
<td>0.935</td>
<td>0.951</td>
<td>1.296</td>
<td>0.061</td>
<td>0.060</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>1.00</td>
<td>0.953</td>
<td>0.950</td>
<td>1.315</td>
<td>0.043</td>
<td>0.038</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.965</td>
<td>0.890</td>
<td>0.948</td>
<td>0.219</td>
<td>0.078</td>
<td>0.107</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.947</td>
<td>0.921</td>
<td>0.951</td>
<td>0.178</td>
<td>0.055</td>
<td>0.060</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>0.889</td>
<td>0.943</td>
<td>0.951</td>
<td>0.145</td>
<td>0.039</td>
<td>0.038</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>0.739</td>
<td>0.951</td>
<td>0.947</td>
<td>0.119</td>
<td>0.028</td>
<td>0.026</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>0.958</td>
<td>0.915</td>
<td>0.946</td>
<td>0.082</td>
<td>0.052</td>
<td>0.060</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>0.950</td>
<td>0.938</td>
<td>0.950</td>
<td>0.060</td>
<td>0.037</td>
<td>0.038</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>0.910</td>
<td>0.944</td>
<td>0.945</td>
<td>0.044</td>
<td>0.026</td>
<td>0.026</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>0.805</td>
<td>0.948</td>
<td>0.948</td>
<td>0.032</td>
<td>0.018</td>
<td>0.018</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>0.955</td>
<td>0.932</td>
<td>0.947</td>
<td>0.041</td>
<td>0.034</td>
<td>0.036</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>0.950</td>
<td>0.944</td>
<td>0.949</td>
<td>0.029</td>
<td>0.024</td>
<td>0.024</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>0.933</td>
<td>0.946</td>
<td>0.950</td>
<td>0.021</td>
<td>0.017</td>
<td>0.017</td>
</tr>
<tr>
<td>10</td>
<td>16</td>
<td>0.889</td>
<td>0.95</td>
<td>0.950</td>
<td>0.015</td>
<td>0.012</td>
<td>0.012</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>0.950</td>
<td>0.938</td>
<td>0.946</td>
<td>0.030</td>
<td>0.027</td>
<td>0.028</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>0.953</td>
<td>0.944</td>
<td>0.95</td>
<td>0.021</td>
<td>0.019</td>
<td>0.019</td>
</tr>
<tr>
<td>15</td>
<td>8</td>
<td>0.944</td>
<td>0.949</td>
<td>0.952</td>
<td>0.015</td>
<td>0.013</td>
<td>0.013</td>
</tr>
<tr>
<td>15</td>
<td>16</td>
<td>0.910</td>
<td>0.954</td>
<td>0.951</td>
<td>0.011</td>
<td>0.009</td>
<td>0.009</td>
</tr>
</tbody>
</table>
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

Table 2.4: Coverage probability (\(CP\)) and width (\(W\)) of confidence intervals for \(CV = 0.33\). \(SE = 0.002\) for \(CP\), < 0.01 for width.

<table>
<thead>
<tr>
<th>(N^*)</th>
<th>(k)</th>
<th>(CP) fid</th>
<th>(CP) norm</th>
<th>(CP) UMP</th>
<th>(W) fid</th>
<th>(W) norm</th>
<th>(W) UMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>0.999</td>
<td>0.850</td>
<td>0.951</td>
<td>8.69</td>
<td>0.93</td>
<td>1.69</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>1.00</td>
<td>0.907</td>
<td>0.951</td>
<td>8.71</td>
<td>0.65</td>
<td>0.70</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>1.00</td>
<td>0.942</td>
<td>0.951</td>
<td>8.75</td>
<td>0.46</td>
<td>0.40</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>1.00</td>
<td>0.960</td>
<td>0.952</td>
<td>8.72</td>
<td>0.32</td>
<td>0.25</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.996</td>
<td>0.890</td>
<td>0.947</td>
<td>3.80</td>
<td>0.59</td>
<td>1.01</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>1.00</td>
<td>0.919</td>
<td>0.951</td>
<td>3.89</td>
<td>0.41</td>
<td>0.43</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>1.00</td>
<td>0.943</td>
<td>0.949</td>
<td>3.99</td>
<td>0.29</td>
<td>0.27</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>1.00</td>
<td>0.964</td>
<td>0.952</td>
<td>4.01</td>
<td>0.20</td>
<td>0.18</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>0.987</td>
<td>0.914</td>
<td>0.950</td>
<td>0.91</td>
<td>0.39</td>
<td>0.48</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>0.990</td>
<td>0.940</td>
<td>0.949</td>
<td>0.79</td>
<td>0.27</td>
<td>0.28</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>0.996</td>
<td>0.955</td>
<td>0.952</td>
<td>0.70</td>
<td>0.19</td>
<td>0.18</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>0.999</td>
<td>0.960</td>
<td>0.949</td>
<td>0.70</td>
<td>0.14</td>
<td>0.13</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>0.958</td>
<td>0.934</td>
<td>0.954</td>
<td>0.33</td>
<td>0.25</td>
<td>0.27</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>0.951</td>
<td>0.949</td>
<td>0.954</td>
<td>0.24</td>
<td>0.17</td>
<td>0.18</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>0.920</td>
<td>0.952</td>
<td>0.949</td>
<td>0.18</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>10</td>
<td>16</td>
<td>0.844</td>
<td>0.954</td>
<td>0.950</td>
<td>0.13</td>
<td>0.09</td>
<td>0.08</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>0.957</td>
<td>0.943</td>
<td>0.954</td>
<td>0.23</td>
<td>0.20</td>
<td>0.21</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>0.948</td>
<td>0.951</td>
<td>0.953</td>
<td>0.17</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>15</td>
<td>8</td>
<td>0.925</td>
<td>0.950</td>
<td>0.948</td>
<td>0.12</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>15</td>
<td>16</td>
<td>0.878</td>
<td>0.953</td>
<td>0.949</td>
<td>0.09</td>
<td>0.07</td>
<td>0.07</td>
</tr>
</tbody>
</table>
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

width, as we would expect.

The fiducial interval breaks down for \( N = 2 \), and if \( CV = 0.33 \), the interval is impractically wide for \( N < 10 \). This is due to the fact that, in Equation 2.11, the denominators of the terms on the right-hand side can be quite small if \( CV \) is this large. Even for the cases for which the interval has close to nominal coverage, it is substantially wider on average than the \( UMP \) interval.

Hannig et al [59] proved that under regularity conditions, the nominal \( p \)-values and confidence levels for fiducial inference converge to the true \( p \)-values and confidence levels as \( N^* \) grows with \( k \) fixed. And we do see in Table that with \( k \) fixed, the confidence level of the fiducial interval gets closer to the nominal level as \( N^* \) increases. However, the coverage and width of the fiducial interval decline with \( k \), with the coverage eventually dropping far below the nominal level. The central point of the fiducial distribution of \( CV \) is not equal to the true \( CV \) for small samples, so as the distribution collapses around it by the addition of samples with fixed \( N^* \), the probability increases that the true \( CV \) is in the tails.

The findings for the fiducial interval are inconsistent with those in Tian’s original paper, which examines the cases \( k = 3 \) and \( k = 5 \) for \( N \geq 10 \). For comparable cases, the coverage probabilities are slightly larger and the widths are substantially smaller in the tables above than in Tian’s simulations; that is, I am finding that the fiducial interval’s properties are better than those found by the original paper. There is an objective reason not to trust Tian’s results, since the widths do not decrease with \( k \) in her results, when clearly increasing the amount of data should cause the percentiles of the fiducial distribution to move closer to its mode.

For small \( N^* \) or \( k \) the approximate normal interval has coverage probability below nominal. As \( N^* \) increases for fixed \( k \), the coverage appears to converge to the nominal level, as predicted in Section 2.3.3. As \( k \) increases, the coverage probability initially improves – an indication the estimator is converging toward the normality that the nominal coverage assumes. For small \( N^* \), the interval appears to become conservative as \( k \) grows, an indication that the delta method variance expression is an underestimate in small samples, but the deviation from the nominal coverage is slight. The width of the normal interval is comparable to that of the \( UMP \) interval in cases where they
have similar coverage.

In conclusion, if the sample sizes are equal and \( CV \) is in the practical range, for accurate size and optimal power, the \( UMP \) tests and corresponding interval should be used. If the sample sizes are unequal, confidence intervals and hypothesis tests based on the \( UMP \) test can still be created, but the computation is sufficiently complicated that it makes sense to use the approximate normal inference if the sample sizes and \( k \) are large enough. The fiducial interval is not competitive.

### 2.4 Testing the Equality of the Coefficients of Variation of \( k \) Normal Populations

The need to compare relative variability across populations arises occasionally in statistics.

In agriculture, medicine, and other fields, within-study \( CV \) serves as an indication of whether two studies are comparable [11]; if one has much higher variability than the other, it can be taken as an indication that different unobserved variables were at play in the different studies.

Usually, the question of interest to experimenters is the effect of a treatment on central tendency, but occasionally the response is relative variability as measured by the \( CV \). Researchers compare the precision of different assays ([63], [64]), financial analysts compare the riskiness of different stocks, and quality control managers might want to compare the consistency of different production processes. Paleontologists compare the \( CVs \) of dental data from fossil finds to those of extant species in order to determine if the fossils are from more than one species [65]. Epidemiologists have considered the effects of smoking on the \( CVs \) of cardiac variables [66]. Climatologists keep track of changes in rainfall \( CV \) over time [27]. Clinicians are sometimes concerned about the effect of a disease or drug on the \( CV \) of a biological variable ([54], [67], [68], [69]). Other examples where researchers consider the effect of some treatment on \( CV \) can be found in [28], [29], [70], and [71].

In comparing relative variability across \( k \) populations, the first question to ask is usually whether the \( CVs \) are different. This would mean conducting a hypothesis test
of $H_0: CV_1 = \ldots = CV_k$ against the alternative that there is at least one difference, which is the test we shall consider in Section 2.4.

Very often, this null hypothesis is reasonable. With regard to the WSCV, the assumption of constant $CV$ across subjects appears so natural that often researchers don’t realize they are making an assumption. The equality null is consistent with the assumption that one population is a rescaled version of another population. For instance, one might model a body with larger mass as just a scaled-up version of a body with smaller mass, although Allen’s Rule in biology is an example where this is not the case.

In addition to being a reasonable assumption, it is one that researchers can make use of to make more accurate or powerful inference ([72], [73], [74], [75], [76], [77]). For instance, it allows a convenient answer to the question of how to model variances that apparently differ across populations, which comes up time and again (see [78]). Tests of the assumption are needed to ensure that convenience does not come at the expense of validity.

### 2.4.1 Notation

Section 2.4 will use the notation of Section 2.3, with the exception that $CV_i$ will refer to the $i$th sample’s coefficient of variation, while $CV$ will refer to the common value if $CV_1 = \ldots = CV_k$ and a central value if there are differences.

### 2.4.2 Previous literature

There is a substantial literature on testing $CV$ homogeneity against general alternatives in normal populations, totalling 18 different papers. A dozen different tests have appeared, but two clearly stand out as superior to the others. A thorough review is given in Section 3.3.
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

Feltz-Miller test

Feltz and Miller [11] proposed the statistic

\[ FM = \sum_{i=1}^{k} \left( \frac{\hat{CV}_i - CV}{\sqrt{\frac{CV^2}{2(N_i - 1)} + \frac{CV^4}{N_i - 1}}} \right)^2, \]  

(2.21)

where \( CV \) is taken from Equation 2.12.

Feltz and Miller argued for using \( \chi^2_{k-1} \) tables to evaluate the \( p \)-value for this statistic. Here I shall give a formal asymptotic justification.

**Proposition 9. Asymptotic \( \chi^2 \) distribution for \( FM \) statistic**

Under the null hypothesis, \( FM \) converges in distribution to a \( \chi^2_{k-1} \) as the sample sizes converge to \( \infty \).

**Proof.** For simplicity let \( N_i - 1 = \beta_i(N^* - 1) \).

By Equation 2.6 and Definition 1.3.6.4,

\[ Y(N^*) \equiv \frac{\sqrt{N^* - 1}}{\sqrt{\frac{CV^2}{2} + CV^4}} \begin{pmatrix} \hat{CV}_1 - CV \\ \vdots \\ \hat{CV}_k - CV \end{pmatrix} \rightarrow_d N(0, V(\beta)), \]

where \( V(\beta) \) is the diagonal matrix with \( V(\beta)_{ii} = \frac{1}{\beta_i} \). Now by Proposition 3, clearly \( CV_{N^*} \rightarrow_p CV \), so by the Slutsky-delta method (Theorem 1.3.6.6), we can write

\[ Y^*(N^*) \equiv \frac{\sqrt{N^* - 1}}{\sqrt{\frac{CV^2}{2} + CV^4}} \begin{pmatrix} \hat{CV}_1 - CV \\ \vdots \\ \hat{CV}_k - CV \end{pmatrix} \rightarrow_d N(0, V(\beta)). \]

Let \( \tilde{Y} \) be the weighted average of the elements of \( Y^* \), where the weights are the \( \beta_i \)'s. Consider the function \( g(Y^*) = \sum_{i=1}^{k} \frac{(Y_i^* - \hat{Y})^2}{V(\beta)_{ii}} \). Algebra shows that \( g(Y^*) = FM \). Also, since \( g \) is continuous, by Theorem 1.3.6.5 part 2, \( g(Y^*_{N^*}) \) converges distribution to
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

$g(\mathbf{Z}_\beta)$, where $\mathbf{Z}_\beta \sim N(0, \mathbf{V}(\beta))$. But Theorem 1.3.5.11, part 3 tells us that $g(\mathbf{Z}_\beta) \sim \chi^2_{k-1}$.

The asymptotic results that justify Proposition 9 apply to increasing sample sizes, not increasing populations with sample sizes fixed. Increasing $k$ with sample sizes fixed may or may not improve the accuracy of the $FM$ test; we would need simulations to tell us.

**Modified Bennett test**

McKay’s approximation yields $(N_i - 1)M_i \sim \Gamma \left( \frac{N_i - 1}{2}, \frac{2}{\omega_i} \right)$, where $\omega_i$ is the monotonic function $\frac{1 + CV^2_i}{CV^2_i}$. Bennett [79] pointed out that with this approximation, the null hypothesis of equal coefficients of variation is equal to the null that the scale parameters of the $k$ gamma variables $(N_i - 1)M_i$ are the same, and suggested using Pitman’s [4] test. The version of Pitman’s test in Theorem 1.3.3.1 also turns out to be the likelihood ratio test.

Bennett’s expression for the test statistic had a slight mistake; the correction, known as the modified Bennett test statistic, was presented in Shafer and Sullivan [80]:

$$M_B = \frac{\prod_{i=1}^{k} M_i^{N_i-1}}{\left( \sum_{i=1}^{k} (N_i-1)M_i \right) \sum_{i=1}^{k} N_i-1}. \quad (2.22)$$

Letting $\omega_i = \omega^* + c_i$, where we restrict $c_k = - \sum_{i=1}^{k-1} c_i$, the null hypothesis becomes $c_1 = \ldots = c_{k-1} = 0$, and we can see by Theorem 1.3.6.14 that

$$-2 \ln(M_B) \approx \chi^2_{k-1}, \quad (2.23)$$

which is used to obtain the $p$-value for the $M_B$ test.

The $\chi^2$ approximation for the likelihood ratio test improves with increasing sample size, and McKay’s $\chi^2$ approximation improves with increasing sample size, so we should expect that the accuracy of the $p$-value of the $M_B$ test should improve with increasing
sample size. As with the FM test, however, asymptotic theory does not guide us as to the behavior of the test as $k$ increases with the sample sizes fixed.

Simulations done by Feltz and Miller indicated that the MB test and the FM test had about the same size and power for the scenarios considered.

### 2.4.3 Applying exponential family model to testing $CV$ equality

We can apply the results in Section 2.2.4 to the problem of testing equality. For simplicity of exposition, in Section 2.4.3 I shall speak of Section 2.2.4 as if it were exact. Extending Equation 2.17 to the case of unequal coefficients of variation, the density $\phi^*_M$ of $M$ is

$$\phi^*_M(m) = C_M(\theta)H_M(m) \exp\left(\sum_{i=1}^{k} \theta_i T_i(m)\right), \quad (2.24)$$

where

$$\theta = \{c_1, \ldots, c_{k-1}, \omega^*\}$$

$$c_k = -\sum_{i=1}^{k-1} c_i,$$

$$C_M(\theta) = \prod_{i=1}^{2} \frac{1}{\Gamma\left(\frac{N_i-1}{2}\right)} \left[(N_i - 1) \frac{\omega^* + c_i - \frac{2}{N_i}}{2}\right]^{\frac{N_i-1}{2}},$$

$$H_M(m) = I_{m_1 \geq 0, \ldots, m_k \geq 0} \prod_{i=1}^{k} \frac{m_i^{N_i - 1}}{m_i^{2}} \exp\left(\sum_{i=1}^{k} \frac{N_i - 1}{N_i} m_i\right),$$

$$T_k(m) = -\sum_{i=1}^{k} \frac{N_i - 1}{2} m_i,$$

$$T_i(m) = \frac{N_k - 1}{2} m_k - \frac{N_i - 1}{2} m_i, \quad i \neq k.$$}

Here I have used a parameterization in which $\omega^* = \frac{1 + CV^2}{CV^2}$ and $\omega_i = \omega^* + c_i$. 

81
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

Properties of $MB$ test with equal sample sizes

For the likelihood in Equation 2.24, if the sample sizes are equal, the likelihood ratio statistic for testing the null turns out to be $MB$. Also, by Equation 2.8, if the sample sizes are equal, $(N^*-1)M_i \sim \Gamma\left(\frac{N^*-1}{2}, \frac{2}{\omega^*+c_i-2/N^*}\right)$ so testing the null hypothesis is equivalent to testing the equality of the scale parameters of $k$ gamma variables, so that the likelihood ratio test is identical to Pitman’s test, which has the advantage that it is unbiased (Theorem 1.3.3.1). Unbiasedness is more important in this context than in testing the value of a common $CV$ because the possible alternatives are more complex here; in the Section 2.3, we would be worried about unbiasedness only for the insignificant case of values of $CV$ near the null.

Obtaining accurate $p$-values for $MB$ statistic

The $MB$ test relies on the large-sample theory for its $p$-values. The impediment to calculating $p$-values that are valid in small samples is the existence of the nuisance parameter $CV$, or $\omega^*$ to be more exact.

If the sample sizes are equal, from Equation 2.8:

$$MB = \frac{\prod_{i=1}^k \left(\frac{U_i}{N^*-1}\right)^{N^*-1}}{\left(\frac{1}{k} \sum_{i=1}^k \frac{U_i}{N^*-1}\right)^{k(N^*-1)}}.$$  \hspace{1cm} (2.25)

Since $CV$ does not appear in this expression, a similar test (Definition 1.3.3.1) can be conducted by Monte Carlo:

**Algorithm 3.** $p$-values for $MB$ test, equal sample sizes

1. Find the observed value $mb$ of $MB$.
2. Take $s$ draws of $\{U_1, \ldots, U_k\}$.
3. Calculate $s$ draws of $MB$ by using the draws of the $U_i$s in Equation 2.25.
4. Record the number of times $NMBE$ that a drawn value of $MB$ is less than $mb$.
5. The $p$-value for the test is approximately $\frac{NMBE}{s}$.
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

\(NMBE_s\) is a Monte Carlo \(p\)-value (Definition 1.3.3.7), and for \(s\) large enough will be highly accurate.

If the sample sizes are unequal, we can condition on the sufficient statistic \(T_k\) to “get rid” of the nuisance parameter. This is discussed in Chapter 4.

**Most powerful similar inference for \(k = 2\).**

With \(k = 2\), there are two parameters in \(\phi^*_M\), the nuisance parameter \(\omega^*\) and \(c_1\). Here we shall derive hypothesis tests that are similar (as regards the nuisance parameter), invariant to positive scale transformations (because they are based on the maximal invariant \(M\)), and \(UMP\) among tests that have those properties.

The null hypothesis is identical to \(H_o : c_1 = 0\). \(H_a : c_1 > 0\) is identical to \(H_a : \frac{1+CV_1}{CV_1} > \frac{1+CV_2}{CV_2}\), or \(H_a : CV_2 > CV_1\).

**Proposition 10.** **UMP similar invariant tests for equality of two normal coefficients of variation**

If \(k = 2\), the \(UMP\) similar invariant test of \(H_o : CV_1 = CV_2\) against \(H_a : CV_1 < CV_2\) rejects \(H_o\) if \(M_1 < b_l\), where \(Prob_{c_1=0}(M_1 < b_l|T_2 = t_2) = \alpha\).

If \(k = 2\), the \(UMP\) similar invariant test of \(H_o : CV_1 = CV_2\) against \(H_a : CV_1 > CV_2\) rejects \(H_o\) if \(M_1 > b_u\), where \(Prob_{c_1=0}(M_1 > b_u|T_2 = t_2) = \alpha\).

**Proof.** For brevity I shall prove just the first result.

By Theorem 1.3.4.6, the \(UMP\) similar invariant test of \(H_o : c_1 = 0\) against \(H_a : c_1 > 0\) has rejection region of the form \(T_1 > b^*_l\), where \(Prob_{c_1=0}(T_1 > b^*_l|T_2 = t_2) = \alpha\). (Recall that since we are conditioning on a sufficient statistic for \(\omega^*\), that probability will not depend on \(\omega^*\).) We can rewrite the rejection region as \(\frac{N_2-1}{2}M_2 - \frac{N_1-1}{2}M_1 > b^*_l\), which is \(-t_2 - (N_1 - 1)M_1 > b^*_l\) since we are conditioning on \(T_2 = t_2\), or \(M_1 < \frac{b^*_l - t_2}{N_1 - 1}\).

We can view the right hand side as simply the constant \(b_l\).

Now \(H_o : c_1 = 0\) is identical to \(H_o : CV_1 = CV_2\), and the set of alternatives of the form \(c_1 > 0\) is identical to the set of alternatives of the form \(CV_1 < CV_2\). Then the rejection region we have just described is also the \(UMP\) similar invariant rejection region for testing \(H_o : CV_1 = CV_2\) against \(H_a : CV_1 < CV_2\). \(\square\)
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

To determine \( p \)-values with unequal sample sizes will require the use of a Monte Carlo method described in Chapter 4. However, with equal sample sizes we get a test that is like the \( F \) test for equal variances.

**Proposition 11.** Rejection region for UMP test of equality with \( k = 2, N_1 = N_2 \)

If \( k = 2 \) and \( N_1 = N_2 \):

The UMP similar invariant test implied by Equation 2.24 of \( H_0 : CV_1 = CV_2 \) against \( H_a : CV_1 < CV_2 \) rejects \( H_0 \) if \( \frac{M_2}{M_1} > F_{1-\alpha}^{N^*-1,N^*-1}, \)

The UMP similar invariant test implied by Equation 2.24 of \( H_0 : CV_1 = CV_2 \) against \( H_a : CV_1 > CV_2 \) rejects \( H_0 \) if \( \frac{M_1}{M_2} > F_{1-\alpha}^{N^*-1,N^*-1}, \)

where \( F_{\gamma_1,\gamma_2} \) is the \( \gamma \)th percentile of the \( F \) distribution with \( \nu_1, \nu_2 \) degrees of freedom.

**Proof.** For brevity, I shall prove the former result.

From Proposition 10, the rejection region is of the form \( M_1 < b_l \). Recalling that we are conditioning on \( t_2 \), we can write this as \( M_1 - t_2 < b_l - t_2 \), where we can treat the right hand side as a constant \( b_l^* \). We can rewrite this as \( M_1 \left( \frac{N^*}{N^*-1} \right) M_2 = \frac{1}{\frac{N^*}{N^*-1} + \frac{N^*-1}{N^*-1} \frac{M_2}{M_1}} < b_l^* \),

or \( \frac{M_2}{M_1} > \frac{N^*-1}{N^*-1} \equiv b_l^{**} \), where \( \text{Prob}_{c_1=0} \left( \frac{M_2}{M_1} > b_l^{**} | T_2 = t_2 \right) = \alpha. \)

Now by Theorem 1.3.4.3 \( T_2 \) is complete. Also, by Theorem 1.3.4.2, \( T_2 \) is minimal sufficient for \( \omega^* \), and thus for \( CV \). From Equation 2.8, under the null, \( \frac{M_2}{M_1} = \frac{U_2^{\frac{1}{N^*}-1}}{U_1^{\frac{1}{N^*}-1}} \), an \( F_{N^*-1,N^*-1} \) variable, so its distribution does not depend on \( CV \) and it is ancillary for \( CV \). Then by Basu’s Theorem (Theorem 1.3.1.3), \( \frac{M_2}{M_1} \) is independent of \( T_2 \), so that \( \text{Prob}_{c=0} \left( \frac{M_2}{M_1} > b_l^{**} | T_2 = t_2 \right) \) is just the unconditional probability.

Then the result follows by the fact that \( \frac{U_2}{U_1^{\frac{1}{N^*}-1}} \) is \( F_{N^*-1,N^*-1}. \)

From this Theorem, \( p \)-values for the UMP test with equal sample sizes would be

\[
1 - \phi_{N^*-1,N^*-1} \left( \frac{m_2}{m_1} \right), \quad H_a : CV_1 < CV_2
\]  

\[
1 - \phi_{N^*-1,N^*-1} \left( \frac{m_1}{m_2} \right), \quad H_a : CV_1 < CV_2,
\]

\[\text{Equation 2.26}\]
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

where \( \phi_{N^*-1,N^*-1}^F \) is the cdf of an \( F_{N^*-1,N^*-1} \) random variable.

Usually, we would like to test a two-sided alternative rather than a one-sided alternative. The two-sided test with equal probabilities of underestimating and overestimating \( c \) has the rejection region

\[
M_1 < b_l \cup M_1 > b_u,
\]

where \( \text{Prob}_{c=0} (M_1 < b_l | T_2 = t_2) = \frac{\alpha}{2} \) and \( \text{Prob}_{c=0} (M_1 > b_u | T_2 = t_2) = \frac{\alpha}{2} \). For unequal sample sizes, this test will not be unbiased, though theory would lead us to believe that it should be quite powerful against alternatives of interest, as explained in Section 2.3.4.

If the sample sizes are equal, we can calculate the two-sided bounds in Equation 2.27 explicitly using Proposition 11, and the \( p \)-value of the test is

\[
2\min \left( \left( 1 - \phi_{N^*-1,N^*-1}^F \left( \frac{m_2}{m_1} \right) \right), \left( 1 - \phi_{N^*-1,N^*-1}^F \left( \frac{m_1}{m_2} \right) \right) \right). \tag{2.28}
\]

It turns out that with equal sample sizes that this is the \( \text{UMP} \) unbiased similar test based on the maximal invariant.

**Proposition 12.** \( \text{UMP} \) unbiased similar invariant test of \( CV \) equality for \( k = 2 \), \( N_1 = N_2 \)

The \( \text{UMP} \) unbiased similar invariant test implied by Equation 2.24 of \( H_o: CV_1 = CV_2 \) against \( H_a: CV_1 \neq CV_2 \) rejects \( H_o \) if \( \frac{M_2}{M_1} > F_{1-\frac{\alpha}{2}}^{N^*-1,N^*-1} \) or if \( \frac{M_1}{M_2} > F_{1-\frac{\alpha}{2}}^{N^*-1,N^*-1} \).

**Proof.** For \( k = 2 \) and \( N_1 = N_2 \), the rejection region for a test based on the \( MB \) statistic can be written as \( \frac{(M_1 M_2)}{(M_1 + M_2)^2} < b \), or \( \frac{M_1 M_2}{M_1^2 + 2M_1 M_2 + M_2^2} < b \). We can manipulate this algebraically to get \( \frac{M_1}{M_2} + \frac{M_2}{M_1} > b^* \). By Theorem 1.3.8.9, this is equal to a region of the form \( \frac{M_2}{M_1} < b_l^* \cup \frac{M_1}{M_2} > b_u^* \) or \( \frac{M_1}{M_2} > b_{21}^* \cup \frac{M_2}{M_1} > b_{21}^* \).

By the fact that \( M_1 \) and \( M_2 \) enter the \( MB \) statistic symmetrically, the rejection region must be of the form \( \frac{M_2}{M_1} > b^* \cup \frac{M_1}{M_2} > b^* \). So a test based on the \( MB \) statistic has the same form as the test we are contemplating if \( k = 2 \), \( N_1 = N_2 \).

Recall that the test based on the \( MB \) statistic (with accurate, not asymptotic, \( p \)-values) is unbiased with equal sample sizes, so the test we are contemplating is unbiased. Then by Theorem 1.3.4.7 and Proposition 11 the test is the \( \text{UMP} \) unbiased similar invariant test.
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

The proof also shows that the $F$ test in Proposition 12 is based on the same test statistic as the $MB$ test; the difference is only in how the $p$-values are calculated.

One should keep in mind that the results in Section 2.4.3 are ultimately based on the approximate $cdf$ $\Phi_V$. But since we have seen that this approximation is highly accurate, we can have a high degree of confidence in them.

2.4.4 Simulation results for equal sample sizes

For equal sample sizes, the discussion above suggests a third test to compare to the $MB$ and the $FM$, and that is the $MB$ test with $p$-values corrected by either Algorithm 3 (for $k > 2$) or Equation 2.28.

Table 2.5 compares the sizes of the three tests. For each scenario, 10,000 data sets of $k$ populations of sample size $N^*$ were generated from normal distributions with coefficient of variation equal to the $CV$ column.

The $MB$ test and $FM$ tests are reasonably accurate, but are slightly liberal for small $N$, and this problem is worse with large $k$. The alternative approach for calculating $p$-values for the $MB$ statistic corrects the liberality, but may slightly overcorrect at the extreme upper end of the practical range.

The discussion above would lead us to believe that tests based on the $MB$ statistic should have very good properties. In the model of Equation 2.24 with equal sample sizes, it is the likelihood ratio test, it is unbiased in that model, and it is the $UMP$ unbiased similar invariant test for $k = 2$ once the $p$-values are corrected. Table 2.6 compares the powers of the $MB$ test and the $FM$ test. The appropriate comparison here is between the uncorrected version of the $MB$ test and the $FM$, since these two tests have comparable sizes. Table 2.6 was computed in the same way as Table 2.5, except that half of the populations had coefficient of variation equal to $CV_1$ and half had $CV_2$.

The main result from Table 2.6 is that the $MB$ test and the $FM$ test have similar power; the theoretical power advantages that come from using the $MB$ statistic appear to be minor.
Table 2.5: Size of two-sided tests of CV homogeneity. \((SE = 0.002)\)

<table>
<thead>
<tr>
<th>CV</th>
<th>k</th>
<th>N</th>
<th>MB</th>
<th>FM</th>
<th>Corrected MB</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>2</td>
<td>10</td>
<td>0.057</td>
<td>0.057</td>
<td>0.051</td>
</tr>
<tr>
<td>0.05</td>
<td>2</td>
<td>20</td>
<td>0.051</td>
<td>0.050</td>
<td>0.048</td>
</tr>
<tr>
<td>0.05</td>
<td>2</td>
<td>30</td>
<td>0.052</td>
<td>0.052</td>
<td>0.051</td>
</tr>
<tr>
<td>0.05</td>
<td>4</td>
<td>8</td>
<td>0.064</td>
<td>0.061</td>
<td>0.053</td>
</tr>
<tr>
<td>0.05</td>
<td>4</td>
<td>16</td>
<td>0.055</td>
<td>0.054</td>
<td>0.049</td>
</tr>
<tr>
<td>0.05</td>
<td>4</td>
<td>24</td>
<td>0.049</td>
<td>0.048</td>
<td>0.046</td>
</tr>
<tr>
<td>0.05</td>
<td>8</td>
<td>6</td>
<td>0.071</td>
<td>0.062</td>
<td>0.050</td>
</tr>
<tr>
<td>0.05</td>
<td>8</td>
<td>12</td>
<td>0.065</td>
<td>0.062</td>
<td>0.055</td>
</tr>
<tr>
<td>0.05</td>
<td>8</td>
<td>18</td>
<td>0.056</td>
<td>0.055</td>
<td>0.051</td>
</tr>
<tr>
<td>0.33</td>
<td>2</td>
<td>10</td>
<td>0.056</td>
<td>0.051</td>
<td>0.049</td>
</tr>
<tr>
<td>0.33</td>
<td>2</td>
<td>20</td>
<td>0.050</td>
<td>0.049</td>
<td>0.047</td>
</tr>
<tr>
<td>0.33</td>
<td>2</td>
<td>30</td>
<td>0.046</td>
<td>0.045</td>
<td>0.044</td>
</tr>
<tr>
<td>0.33</td>
<td>4</td>
<td>8</td>
<td>0.061</td>
<td>0.055</td>
<td>0.051</td>
</tr>
<tr>
<td>0.33</td>
<td>4</td>
<td>16</td>
<td>0.051</td>
<td>0.049</td>
<td>0.046</td>
</tr>
<tr>
<td>0.33</td>
<td>4</td>
<td>24</td>
<td>0.049</td>
<td>0.049</td>
<td>0.046</td>
</tr>
<tr>
<td>0.33</td>
<td>8</td>
<td>6</td>
<td>0.066</td>
<td>0.062</td>
<td>0.046</td>
</tr>
<tr>
<td>0.33</td>
<td>8</td>
<td>12</td>
<td>0.063</td>
<td>0.063</td>
<td>0.053</td>
</tr>
<tr>
<td>0.33</td>
<td>8</td>
<td>18</td>
<td>0.056</td>
<td>0.057</td>
<td>0.050</td>
</tr>
</tbody>
</table>
Table 2.6: Power of two-sided tests of CV homogeneity. ($SE = 0.005$)

<table>
<thead>
<tr>
<th>$CV_1$</th>
<th>$CV_2$</th>
<th>k</th>
<th>N</th>
<th>MB</th>
<th>FM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>2</td>
<td>10</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>2</td>
<td>20</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>2</td>
<td>30</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>4</td>
<td>8</td>
<td>0.55</td>
<td>0.55</td>
</tr>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>4</td>
<td>16</td>
<td>0.89</td>
<td>0.89</td>
</tr>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>4</td>
<td>24</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>8</td>
<td>6</td>
<td>0.58</td>
<td>0.59</td>
</tr>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>8</td>
<td>12</td>
<td>0.94</td>
<td>0.94</td>
</tr>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>8</td>
<td>18</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>0.165</td>
<td>0.33</td>
<td>2</td>
<td>10</td>
<td>0.483</td>
<td>0.472</td>
</tr>
<tr>
<td>0.165</td>
<td>0.33</td>
<td>2</td>
<td>20</td>
<td>0.814</td>
<td>0.811</td>
</tr>
<tr>
<td>0.165</td>
<td>0.33</td>
<td>2</td>
<td>30</td>
<td>0.941</td>
<td>0.940</td>
</tr>
<tr>
<td>0.165</td>
<td>0.33</td>
<td>4</td>
<td>8</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>0.165</td>
<td>0.33</td>
<td>4</td>
<td>16</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>0.165</td>
<td>0.33</td>
<td>4</td>
<td>24</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>0.165</td>
<td>0.33</td>
<td>8</td>
<td>6</td>
<td>0.54</td>
<td>0.55</td>
</tr>
<tr>
<td>0.165</td>
<td>0.33</td>
<td>8</td>
<td>12</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>0.165</td>
<td>0.33</td>
<td>8</td>
<td>18</td>
<td>0.99</td>
<td>0.99</td>
</tr>
</tbody>
</table>
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

Table 2.7: Size of one-sided tests of CV homogeneity, $k = 2$. ($SE = 0.002$)

<table>
<thead>
<tr>
<th>CV</th>
<th>$N_1$</th>
<th>$N_2$</th>
<th>Miller</th>
<th>UMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>10</td>
<td>10</td>
<td>0.054</td>
<td>0.049</td>
</tr>
<tr>
<td>0.33</td>
<td>10</td>
<td>10</td>
<td>0.051</td>
<td>0.049</td>
</tr>
</tbody>
</table>

Table 2.8: Power of one-sided tests of CV homogeneity, $k = 2$. ($SE = 0.005$)

<table>
<thead>
<tr>
<th>CV_1</th>
<th>CV_2</th>
<th>$N_1$</th>
<th>$N_2$</th>
<th>Miller</th>
<th>UMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.05</td>
<td>10</td>
<td>10</td>
<td>0.64</td>
<td>0.62</td>
</tr>
<tr>
<td>0.33</td>
<td>0.165</td>
<td>10</td>
<td>10</td>
<td>0.61</td>
<td>0.59</td>
</tr>
</tbody>
</table>

2.4.5 One-sided tests for $k = 2$

The literature focuses almost exclusively on two-sided tests. For testing the null against $H_a: CV_1 > CV_2$, Miller [81] introduced the statistic

$$Z_2 = \frac{\bar{CV}_1 - \bar{CV}_2}{\sqrt{CV^2 \left( \frac{1}{2(N_1-1)} + \frac{1}{2(N_2-1)} \right) + CV^4 \left( \frac{1}{N_1-1} + \frac{1}{N_2-1} \right)}}.$$  

By Equation 2.6, Theorem 1.3.6.5 part 2, and the Slutsky-delta method (Theorem 1.3.6.6), $Z_2$ is asymptotically standard normal, so that the $p$-value of Miller’s one-sided test is $1 - \Phi(Z_2)$.

Table 2.7 shows simulation results for the $N_1 = N_2$ case concerning the size of two different tests of CV equality against $H_a: CV_1 > CV_2$: Miller’s test and the test of Proposition 11. To generate the table, 10,000 datasets with $k = 2$ and $CV$, $N_1$, and $N_2$ as shown were randomly created. Table 2.7 indicates that both tests have close to nominal size, even for fairly small sample sizes.

Table 2.8 was created in the same way as Table 2.7, except that $CV_1$ was allowed
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

to differ from $CV_2$. The tests have essentially identical power.

2.5 Confidence Intervals for Differences Between Two Coefficients of Variation

While the test whether population $CV$s are equal has been thoroughly studied, confidence intervals for differences between $CV$s have received almost no attention. Miller and Feltz [82] is the only paper that addresses the subject; it uses the delta-method approximation to create a confidence interval for the difference between two coefficients of variation.

There are several reasons why one might be interested in the ratio of two population $CV$s rather than their difference. First, as shown in Section 2.1.2, the $CV$ is the standard deviation adjusted for mean differences, and usually confidence intervals for ratios of standard deviations, not differences between them, are constructed. Second, it is evidently the ratio of two $CV$s, not their difference, which determines whether they will be distinguishable in practice; in Table 2.6, the difference between 0.165 and 0.33 is greater than that between 0.025 and 0.05, but we have about the same power for both sets of $CV$s. Third, ratios are more important than differences in some applications. For example, in deciding whether two formulations of the same drug are bioequivalent, the customary procedure is to verify that confidence intervals for the ratios of certain parameters for drug A to those of drug B fall between 0.8 and 1.25. Chow and Tse [54] suggested that the $CV$ of quantities such as $AUC$ should be included among the parameters examined.

The notation in Section 2.5 will be the same as in Section 2.4.
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

2.5.1 Previous literature

Miller and Feltz propose a confidence interval based on the normal approximation of Section 2.2.2. Consider the approximate pivot

$$\frac{\hat{CV}_2 - \hat{CV}_1 - (CV_2 - CV_1)}{\sqrt{\frac{CV_1^2}{2(N_1-1)} + \frac{CV_1^4}{N_1-1} + \frac{CV_2^2}{2(N_2-1)} + \frac{CV_2^4}{N_2-1}}}$$

(2.29)

Miller and Feltz argued that this should be approximately $N(0,1)$. I shall prove this formally here.

**Proposition 13.** The approximate pivot in Equation 2.29 is asymptotically $N(0,1)$ with increasing sample size.

**Proof.** For simplicity, let $N_1 = N_2$. From Equation 2.6 and Definition 1.3.6.4,

$$\sqrt{N^*-1(\hat{CV}_1 - CV_1)} \to_d N \left(0, \frac{CV_1^2}{2} + CV_1^4\right)$$

and

$$\sqrt{N^*-1(\hat{CV}_2 - CV_2)} \to_d N \left(0, \frac{CV_2^2}{2} + CV_2^4\right).$$

Then by Theorem 1.3.6.5 part 2, $\sqrt{N^*-1(\frac{\hat{CV}_2 - \hat{CV}_1 - (CV_2 - CV_1)}{\sqrt{\frac{CV_1^2}{2} + CV_1^4 + \frac{CV_2^2}{2} + CV_2^4}}}) \to_d N(0,1)$. Then the result follows from Proposition 3 and the Slutsky-delta method (Theorem 1.3.6.6). \qed

The upper and lower limits of the $1 - \alpha$ interval for $CV_2 - CV_1$ by Miller and Feltz are:

$$\hat{CV}_2 - \hat{CV}_1 \pm Z_{1-\alpha/2} \sqrt{\frac{\hat{CV}_1^2}{2(N_1-1)} + \frac{\hat{CV}_1^4}{N_1-1} + \frac{\hat{CV}_2^2}{2(N_2-1)} + \frac{\hat{CV}_2^4}{N_2-1}}.$$  

(2.30)

No simulation results have been published regarding the performance of this interval.

Differencing alleviates the asymmetry and bias in $\hat{CV}_i$, improving the accuracy of the normal approximation. To see this, consider that in the case where $CV_2 = CV_1$ and the sample sizes are equal, $\hat{CV}_2 - \hat{CV}_1$ will be symmetric around $CV_2 - CV_1 = 0$. If the
sample sizes are vastly different, we wind up differencing two variables with different
degrees of asymmetry and bias; we would expect that the coverage of this interval with
different sample sizes would not be as close to nominal as with similar sample sizes.

2.5.2 Confidence Interval for the ratio of two CVs

As with testing equality, the problem in constructing confidence intervals for ratios of
CVs is the presence of nuisance parameters, in this case \( CV_1 \) and \( CV_2 \).

From Equation 2.8 we get

\[
\frac{CV_2^2}{CV_1^2} \left( 1 + \frac{N_1-2}{N_1} CV_1^2 \right) \left( 1 + \frac{N_2-2}{N_2} CV_2^2 \right) \approx W, \tag{2.31}
\]

where \( W \sim F_{N_1-1, N_2-1} \).

To get a pivot for \( \frac{CV_2}{CV_1} \), we need to estimate the factor \( \frac{1 + \frac{N_1-2}{N_1} CV_1^2}{1 + \frac{N_2-2}{N_2} CV_2^2} \). We could estimate it by replacing \( CV_2 \) and \( CV_1 \) with \( \hat{CV}_2 \) and \( \hat{CV}_1 \). This approach has the potential to produce intervals with inadequate coverage, because the resulting random variable will have extra variability induced by the variability in \( \hat{CV}_1 \) and \( \hat{CV}_2 \), above that modeled by \( W \). Another option, since this factor is close to 1 for CV in the practical range, is simply to set it equal to 1. Simulation results not shown here indicate that the second approach is slightly preferable. The resulting 1 – \( \alpha \)th interval for \( \frac{CV_2}{CV_1} \) is

\[
upper \ bound = \sqrt{\frac{M_2}{M_1} F_{N_1-1,N_2-1}^{1-\frac{\alpha}{2}}}, \tag{2.32}
\]

\[
lower \ bound = \sqrt{\frac{M_2}{M_1}} \frac{1}{\sqrt{F_{N_1-1,N_2-1}^{1-\frac{\alpha}{2}}}}. \tag{2.33}
\]

2.5.3 Simulation results

Simulation results for the two proposals above are shown in Table 2.9. For the case
where the sample sizes are not equal, it matters for the results which sample is labeled
1 and which is labeled 2.
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

Table 2.9: Coverage probability and average width of two-sided confidence intervals. SE for coverage = 0.001, for width < 0.001.

<table>
<thead>
<tr>
<th>CV₁</th>
<th>CV₂</th>
<th>N₁</th>
<th>N₂</th>
<th>MF coverage</th>
<th>width</th>
<th>Corrected-F coverage</th>
<th>width</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>10</td>
<td>10</td>
<td>0.934</td>
<td>0.10</td>
<td>0.952</td>
<td>0.11</td>
</tr>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>21</td>
<td>21</td>
<td>0.943</td>
<td>0.07</td>
<td>0.950</td>
<td>0.07</td>
</tr>
<tr>
<td>0.165</td>
<td>0.33</td>
<td>10</td>
<td>10</td>
<td>0.931</td>
<td>0.37</td>
<td>0.949</td>
<td>0.36</td>
</tr>
<tr>
<td>0.165</td>
<td>0.33</td>
<td>21</td>
<td>21</td>
<td>0.946</td>
<td>0.250</td>
<td>0.950</td>
<td>0.225</td>
</tr>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>10</td>
<td>21</td>
<td>0.954</td>
<td>0.08</td>
<td>0.968</td>
<td>0.10</td>
</tr>
<tr>
<td>0.10</td>
<td>0.05</td>
<td>10</td>
<td>21</td>
<td>0.915</td>
<td>0.10</td>
<td>0.967</td>
<td>0.10</td>
</tr>
<tr>
<td>0.165</td>
<td>0.33</td>
<td>10</td>
<td>21</td>
<td>0.958</td>
<td>0.28</td>
<td>0.969</td>
<td>0.34</td>
</tr>
<tr>
<td>0.33</td>
<td>0.165</td>
<td>10</td>
<td>21</td>
<td>0.912</td>
<td>0.35</td>
<td>0.965</td>
<td>0.33</td>
</tr>
</tbody>
</table>

All simulations are based on 50,000 randomly-generated datasets of two populations with coefficients of variation CV₁ and CV₂ and sample sizes N₁ and N₂, for which confidence intervals of level 0.95 were calculated by each method.

The width column is the average confidence interval width over the 50,000 datasets. The “width” associated with the F interval is calculated so as to be comparable to that of the Miller-Feltz interval. Let Ru be the upper bound for the confidence interval for the ratio CV₂/CV₁ calculated from one set of simulated data, and let Rd be the lower. The width of the confidence interval for that dataset was calculated as $rac{1}{2}(Ru\widehat{CV}_1 - \widehat{CV}_1 - (R_d\widehat{CV}_1 - \widehat{CV}_1) + \frac{1}{2}(\widehat{CV}_2 - \frac{1}{Ru}\widehat{CV}_2 - (\widehat{CV}_2 - \frac{1}{R_d}\widehat{CV}_2)) = \frac{1}{2}(Ru - R_d)\widehat{CV}_1 + \frac{1}{2}(\frac{1}{Ru} - \frac{1}{R_d})\widehat{CV}_2$.

The Miller-Feltz interval has close to nominal coverage but can under cover with small samples or if there is a difference in sample sizes. The F interval for the ratios has coverage close to nominal, but is slightly conservative if the sample sizes differ.
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

2.6 The Normality Assumption

Boos and Brownie [83] do an asymptotic analysis that indicates that inference on the variance that assumes normality is not robust to the underlying distribution. Assuming that robustness in large samples is a reasonable guide to robustness in small samples, I follow a similar approach here.

2.6.1 Robustness of inference that assumes normality

Applying the delta method (Theorem 1.3.6.3) to the result in Theorem 1.3.6.14, we can derive the asymptotic distributions for \( \hat{CV} \) and \( M \) from a general population.

**Proposition 14.** For a general population, under regularity conditions

\[
\hat{CV}(N) \sim AN\left(CV, \frac{1}{N} \left(CV^4 - \frac{\mu^4}{\mu^3} + \frac{\mu^4 - \sigma^4}{4\mu^2\sigma^2}\right)\right),
\]

Equation 2.6 gives the large-\( N \) distribution of \( \hat{CV} \) under normality. As shown in Section 2.2.3, the asymptotic distribution of the transformation \( M(N) \) derived from Equation 2.6 is practically the same as that derived from either of the \( \chi^2 \) approximations, which are asymptotically equivalent. So the inferential procedures that depend on the \( \chi^2 \) approximations for \( M(N) \) will produce the same inference asymptotically as those that depend on the delta-method approximation for \( \hat{CV} \). Since the fiducial approach is known to be asymptotically valid with sample size, it is safe to assume that it will also produce the same inference with large \( N \).

Now the delta method allows us to infer that the large-\( N \) standard deviation of a function of \( \hat{CV} \) will be proportional to the large-\( N \) standard deviation of \( \hat{CV} \), with the constant being a function of the true \( CV \) (Theorem 1.3.6.3). Thus, to conclude how well any of the inferential procedures in this chapter would do in large samples with a non-normal population, we can use the ratio of the asymptotic standard deviation from Equation 2.6 to the standard deviation from Proposition 14 as a guide. We can interpret this ratio as the ratio of the length a \( 1 - \alpha \) confidence interval assuming normality to that of a valid interval.
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

Table 2.10: Ratio of confidence interval length assuming normality to valid length, $CV = 0.05$

<table>
<thead>
<tr>
<th>$\gamma_1$</th>
<th>0</th>
<th>1.5</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.00</td>
<td>0.76</td>
<td>0.63</td>
</tr>
<tr>
<td>1</td>
<td>1.05</td>
<td>0.78</td>
<td>0.65</td>
</tr>
<tr>
<td>2</td>
<td>1.12</td>
<td>0.80</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Table 2.11: Ratio of confidence interval length assuming normality to valid length, $CV = 0.33$

<table>
<thead>
<tr>
<th>$\gamma_1$</th>
<th>0</th>
<th>1.5</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.00</td>
<td>0.79</td>
<td>0.67</td>
</tr>
<tr>
<td>1</td>
<td>1.48</td>
<td>0.96</td>
<td>0.77</td>
</tr>
<tr>
<td>2</td>
<td>NA</td>
<td>1.37</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Writing this ratio in terms of the skewness ($\gamma_1$) and kurtosis ($\gamma_2$) (Definition 1.3.5.9) of the underlying distribution, we get

$$\sqrt{\frac{1+2CV^2}{1 + 2CV^2 + \frac{\gamma_2}{2} - 2\gamma_1 CV}}$$  \hspace{1cm} (2.34)

With a skewed population of normal kurtosis the procedures in this chapter will produce conservative inference – confidence intervals that are too large. With symmetric leptokurtotic distributions, inference will be liberal.

We can get an idea of the quantitative effect of altering the underlying distribution by plugging in values of $\gamma_1$ and $\gamma_2$ to Equation 2.34. The Weibull distribution is viewed as having moderate skewness (1) and kurtosis (1.5), while the double exponential is viewed as having high kurtosis (3) and the exponential distribution has extreme skew-
Chapter 2. An Exponential Family for a Normal Coefficient of Variation

ness (2). Table 2.10 and Table 2.11 give us an idea of how well the normal-theory procedures do with such values of skewness and kurtosis. The numbers reported are the values of Equation 2.34.

With low $CV$, the effect of introducing skewness is small, but normal-theory inference appears unacceptably liberal with even moderate kurtosis.

With a high value of $CV$, skewness has a large effect, and normal-theory inference is unacceptably conservative with moderate skewness if the underlying population has normal kurtosis. Kurtosis appears to “cancel out” skewness, so that normal-theory inference is reasonable for distributions with the same degree of skewness and kurtosis. ($\gamma_1 = 2$ and $\gamma_2 = 0$ is theoretically impossible with $CV = 0.33$.)

2.6.2 The relevance of normal-theory inference

Although Table 2.10 and Table 2.11 indicate that normal-theory inference will not be acceptable in a number of cases, inferential procedures that assume normality are undoubtedly useful. It is frequently the case that the underlying distribution has only mild skewness or kurtosis; and we have seen that in some cases, normal-theory inference is reasonable even with substantial skewness or kurtosis. Often, we have good a priori reasons to expect normality. A random variable (eg, adult height) which can be thought of as a linear function of a number of random variables (eg, a large number of genes from both the mother and father, diet, environment) can be expected, under regularity conditions, to be approximately normal due to various central limit theorems (see Theorem 1.3.6.7).

Where the normal model clearly does not apply, it is often still useful for modeling either the transformed data or the bulk of the data once outliers are removed. And robust and nonparametric methods frequently sacrifice power. The disadvantages of using an incorrect probability model may be outweighed by the unsatisfactory power of the alternative procedures. In any event, researchers typically conduct two analyses – one that assumes normality, one robust approach – and so it is important that we develop accurate and powerful inferential methods for the former case.
CHAPTER 3

Additional Work on the Coefficient of Variation in Normal Populations

3.1 Sample of coefficients of variation from the scientific literature

In Chapter 2, it was useful to know something about the values of the coefficient of variation in situations that are encountered in practice. Presumably, the set of CVs reported in the abstracts of papers in the ISI Web of Knowledge database are representative of the CVs of interest in practice. Adopting that assumption, I created a sample of scientific papers in the following way:

- I typed “coefficient of variation” into the topic box on ISI Web of Knowledge’s General Search page and found that the current number of articles was 9,785. For “relative standard deviation,” the number was 8,496, and for “coefficients of variation NOT coefficient of variation,” the number was 5,207.
- I decided to sample 60 papers referenced by ISI Web of Knowledge. So that the proportions in the sample equaled the proportions in the population, I sampled 25 papers from the “coefficient of variation” keyword search, 22 from the “relative standard deviation” search and 13 from “coefficients of variation NOT coefficient of variation”.
- I randomly ordered the 9,785 “coefficient of variation” papers, then went to the abstracts in order. If there was a number reported for a coefficient of variation in the abstract, I recorded it; if there was more than one number, I recorded the max and the min. I continued until I had recorded at least 1 CV from 25 different papers. This required reading 36 abstracts.
Chapter 3.  Additional Work on the Coefficient of Variation in Normal Populations

Table 3.1: Articles in random sample – coefficients of variation keyword

<table>
<thead>
<tr>
<th>Journal</th>
<th>Vol, page</th>
<th>CVmin</th>
<th>CVmax</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thorax</td>
<td>49, 500-503</td>
<td>0.075</td>
<td>0.173</td>
</tr>
<tr>
<td>Reproduction</td>
<td>121, 905-913</td>
<td>0.013</td>
<td>0.11</td>
</tr>
<tr>
<td>Biology of Reproduction</td>
<td>54, 1252-1260</td>
<td>0.03</td>
<td>0.125</td>
</tr>
<tr>
<td>Journal of Applied Physiology</td>
<td>94, 2448-2455</td>
<td>0.15</td>
<td>0.19</td>
</tr>
<tr>
<td>Environmental Toxicology and Water Quality</td>
<td>6, 63-75</td>
<td></td>
<td>0.11</td>
</tr>
<tr>
<td>Journal of Chromatography B</td>
<td>745, 373-388</td>
<td>0.03</td>
<td>0.06</td>
</tr>
<tr>
<td>Journal of AOAC International</td>
<td>85, 333-340</td>
<td>0.0037</td>
<td>0.0168</td>
</tr>
<tr>
<td>J of Irrig. and Drain. Eng.</td>
<td>117, 361-376</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>Remote Sensing of Environment</td>
<td>37, 181-191</td>
<td>0.25</td>
<td>0.5</td>
</tr>
<tr>
<td>Journal of Physiology</td>
<td>471, 637-657</td>
<td>0.06</td>
<td>0.16</td>
</tr>
<tr>
<td>Journal of Chromatography B</td>
<td>761, 237-246</td>
<td>0.07</td>
<td>0.21</td>
</tr>
<tr>
<td>Analytica Chimica Acta</td>
<td>276, 3-13</td>
<td>0.072</td>
<td>0.299</td>
</tr>
<tr>
<td>Eur. J of Clin. Chem and Clin Biochem.</td>
<td>30, 837-845</td>
<td>0.043</td>
<td>0.058</td>
</tr>
</tbody>
</table>

- I repeated the above procedure for “relative standard deviation”, randomly ordering 8,496 papers. It turned out that the first 22 abstracts all reported at least one CV. I repeated the procedure to obtain CVs from 13 “coefficients of variation NOT coefficient of variation” papers. Again, I needed to read the minimum number of abstracts.

For each paper, I also recorded (if I could determine it) whether the variable in question was one that necessarily had to be positive.

A histogram of values of the coefficient of variation is reported in Figure 2.1. The lowest CV in the survey was 0.0017. There were four CVs less than 0.01. Over \( \frac{1}{3} \) were between 0.01 and 0.05, and over \( \frac{4}{5} \) of the reported CVs were between 0.01 and 0.15. Only three CVs were greater than 0.34, with the largest being 1.0.

Tables 3.1, 3.2, and 3.3 contain a full listing of the papers in the survey along with their reported CVs. They provide further demonstration that CV is important.
### Table 3.2: Articles in random sample – coefficient of variation keyword

<table>
<thead>
<tr>
<th>Journal</th>
<th>Vol, page</th>
<th>CVmin</th>
<th>CVmax</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zeitschrift fur Pflanzenernahrung und Bode.</td>
<td>161, 51-58</td>
<td>0.03</td>
<td>0.34</td>
</tr>
<tr>
<td>Journal of Pediatrics</td>
<td>144, 169-176</td>
<td>0.027</td>
<td></td>
</tr>
<tr>
<td>Cement Concrete and Aggregates</td>
<td>21, 23-30</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>Monatschrift Kinderheilkunde</td>
<td>150, 1095+</td>
<td>0.26</td>
<td>0.3</td>
</tr>
<tr>
<td>Progress in Biochemistry and Biophysics</td>
<td>28, 118-120</td>
<td>0.096</td>
<td></td>
</tr>
<tr>
<td>Journal of Agricultural and Food Chemistry</td>
<td>54, 2154-2161</td>
<td>0.248</td>
<td>0.26</td>
</tr>
<tr>
<td>Therapeutic Drug Monitoring</td>
<td>16, 293-297</td>
<td>0.0413</td>
<td></td>
</tr>
<tr>
<td>Journal of Pediatric Hematology Oncology</td>
<td>25, 33-37</td>
<td>0.049</td>
<td>0.223</td>
</tr>
<tr>
<td>European J of Clin Chem and Clin Biochem</td>
<td>29, 549-554</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Blood</td>
<td>85, 1897-1902</td>
<td>0.135</td>
<td></td>
</tr>
<tr>
<td>European Journal of Oral Sciences</td>
<td>105, 67-73</td>
<td>0.019</td>
<td>0.199</td>
</tr>
<tr>
<td>Tumor Biology</td>
<td>3, 169-175</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>Bulletin Du Cancer</td>
<td>80, 431-438</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>Drugs of Today</td>
<td>34, 141-152</td>
<td>0.065</td>
<td>0.073</td>
</tr>
<tr>
<td>Analytical Chemistry</td>
<td>69, 1038-1044</td>
<td>0.0042</td>
<td></td>
</tr>
<tr>
<td>Journal of Chromatography – Biomedical Apps</td>
<td>573, 43-48</td>
<td>0.065</td>
<td></td>
</tr>
<tr>
<td>Journal of the Pharma. Soc. of Japan</td>
<td>124, 135-139</td>
<td>0.0608</td>
<td></td>
</tr>
<tr>
<td>Japanese Journal of Crop Science</td>
<td>68, 63-70</td>
<td>0.07</td>
<td>0.11</td>
</tr>
<tr>
<td>Arztliche Laboratorium</td>
<td>37, 39-44</td>
<td>0.1</td>
<td>0.23</td>
</tr>
<tr>
<td>International Journal of Cancer</td>
<td>114, 791-796</td>
<td>0.0017</td>
<td>0.0029</td>
</tr>
<tr>
<td>Journal of Clinical Pathology</td>
<td>52, 430-434</td>
<td>0.11</td>
<td>0.195</td>
</tr>
<tr>
<td>Acta Pharmalogica Sinica</td>
<td>15, 197-201</td>
<td>0.0333</td>
<td>0.0697</td>
</tr>
<tr>
<td>Physics in Medicine and Biology</td>
<td>43, 2325-2336</td>
<td>0.044</td>
<td>0.128</td>
</tr>
<tr>
<td>Journal of Liquid Chromatography</td>
<td>17, 855-865</td>
<td>0.02</td>
<td>0.05</td>
</tr>
<tr>
<td>International Journal of Parasitology</td>
<td>31, 87-91</td>
<td>0.10</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 3. Additional Work on the Coefficient of Variation in Normal Populations

<table>
<thead>
<tr>
<th>Journal</th>
<th>Vol, page</th>
<th>CVmin</th>
<th>CVmax</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journal of Chromatography</td>
<td>1035, 277-279</td>
<td>0.075</td>
<td></td>
</tr>
<tr>
<td>Journal of Analytical Atomic Spectrometry</td>
<td>9, 285-290</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Talanta</td>
<td>52, 181-188</td>
<td>0.013</td>
<td></td>
</tr>
<tr>
<td>Chinese Chemical Letters</td>
<td>12, 799-782</td>
<td>0.024</td>
<td>0.041</td>
</tr>
<tr>
<td>Spectroscopy and Spectral Analysis</td>
<td>25, 113-115</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>Journal of the American Oil Chemists Society</td>
<td>69, 174-177</td>
<td>0.035</td>
<td></td>
</tr>
<tr>
<td>Spectroscopy Letters</td>
<td>29, 69-85</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Bunseki Kagaku</td>
<td>40, T5-T8</td>
<td>0.018</td>
<td></td>
</tr>
<tr>
<td>International Journal of Mass Spectrometry</td>
<td>178, 73-79</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Analyst</td>
<td>129, 15-19</td>
<td>0.04</td>
<td>0.15</td>
</tr>
<tr>
<td>Thermochemistry Acta</td>
<td>224, 271-279</td>
<td>0.011</td>
<td></td>
</tr>
<tr>
<td>Analytical and Bioanalytical Chemistry</td>
<td>379, 764-769</td>
<td>0.031</td>
<td></td>
</tr>
<tr>
<td>Journal of Chromatography</td>
<td>828, 95-103</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>Thermochemistry Acta</td>
<td>326, 53-67</td>
<td>0.017</td>
<td>0.02</td>
</tr>
<tr>
<td>Fresenius Journal of Analytical Chemistry</td>
<td>366, 504-507</td>
<td>0.021</td>
<td>0.027</td>
</tr>
<tr>
<td>Analytical Letters</td>
<td>25, 1687-1692</td>
<td>0.015</td>
<td></td>
</tr>
<tr>
<td>Applied Microbiology and Biotechnology</td>
<td>46, 10-14</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>Rapid Communications in Mass Spectrometry</td>
<td>10, 1017-1023</td>
<td>0.071</td>
<td>0.63</td>
</tr>
<tr>
<td>Analytica Chimica Acta</td>
<td>404, 151-157</td>
<td>0.01</td>
<td>0.086</td>
</tr>
<tr>
<td>Talanta</td>
<td>50, 819-826</td>
<td>0.024</td>
<td>0.026</td>
</tr>
<tr>
<td>Mikrochimica Acta</td>
<td>111, 207-213</td>
<td>0.011</td>
<td></td>
</tr>
<tr>
<td>Analytica Chimica Acta</td>
<td>369, 157-161</td>
<td>0.03</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 3. Additional Work on the Coefficient of Variation in Normal Populations

in a wide variety of fields. The papers for which I could not determine whether the variable in question was necessarily positive were *Cement, Concrete, and Aggregates, Monatschrift Kinderheilkunde*, and *Journal of Chromatography* (V. 828, p.95-103).

### 3.2 Point estimation of a common coefficient of variation in normal samples

In this section, the data come from $k$ normal populations with different means but the same $CV$, and the object is to come up with an effective point estimator of the common $CV$. The notation used in this section will be the same as that in Section 2.3.

#### 3.2.1 Estimators

The different point estimators can be grouped into the three categories: weighted averages, maximum likelihood estimators from the full likelihood, and maximum likelihood estimators from the marginal likelihood of the sample $CV$s.

**Weighted averages**

$CV$ from Equation 2.12 is a simple weighted average. $\hat{CV}$ from Equation 2.13 is a bias-corrected alternative, originally suggested by Zeigler [53]. A third alternative is $\frac{\sum_{i=1}^{k}(N_i-1)B_iC_i\hat{CV}_i}{\sum_{i=1}^{k}N_i-1}$, where $B_i$ and $C_i$ are described in Section 2.3.3.

Rather than simply use the sample sizes as weights, we can find the variance-minimizing weights $\omega_1, \ldots, \omega_k$ to use with the bias-corrected sample $CV$s. The weights minimize

$$\sum_{i=1}^{k} \omega_i^2 B_i^2C_i^2Var(\hat{CV}_i) \text{ s.t. } \sum_{i=1}^{k} \omega_i = 1.$$

As discussed in Chapter 2, $Var(\hat{CV}_i) \approx \frac{1}{N_i-1} \left( \frac{CV^2}{2} + CV^4 \right)$. Plugging this in and using the rules for minimizing a quadratic function with respect to a linear constraint
Chapter 3. Additional Work on the Coefficient of Variation in Normal Populations

[16], we can derive that the optimal weights are

$$\omega_j^* = \frac{N_j - 1}{\sum_{i=1}^k N_i - 1} \frac{C_j^2 B_j^2}{\sum_{i=1}^k C_j^2 B_j^2}.$$  

I shall call the weighted, bias-corrected average with the variance minimizing weights $\bar{CV}_{MV}$.

Chow and Tse [54] suggested taking the square root of the weighted average of the squared sample CVs:

$$\bar{CV}_S \equiv \sqrt{\frac{\sum_{i=1}^k N_i \hat{CV}_i^2}{\sum_{i=1}^k N_i}}.$$  

(3.1)

Maximum likelihood estimators from likelihood of $\hat{CV}_i$s

One has a choice of maximizing the full likelihood of the data or maximizing the marginal likelihood of the sample CVs. This is the product of the densities of the individual sample CVs. The exact density of the sample CV can be obtained from Equation 2.3, but it is inconvenient to work with, so we shall work with Equation 2.17.

The value of the common CV, $\bar{CV}_{MLE}$, that maximizes Equation 2.17 solves

$$\sum_{i=1}^k (N_i - 1) \frac{C_i^2 B_i^2}{1 + \frac{N_i - 2}{N_i} CV^2} = \sum_{i=1}^k (N_i - 1) \frac{\bar{CV}_i^2}{1 + \frac{N_i - 1}{N_i} \bar{CV}_i^2} .$$  

(3.2)

Note that as CV tends to infinity, the left-hand side tends to $\sum_{i=1}^k \frac{N_i (N_i - 1)}{N_i - 2}$. For any practical situation, this will be larger than the right-hand side. As CV tends to 0, the left-hand side tends to 0. Furthermore, the left-hand side is monotonic in CV. Thus, for any practical situation, there will be a unique solution which can be found by the efficient method of bisection (Definition 1.3.8.8).
Chapter 3. Additional Work on the Coefficient of Variation in Normal Populations

If \( N_1 = \ldots = N_k = N^* \), the analytical solution is

\[
\hat{CV}_{MLE}^M = \frac{\sum_{i=1}^k (N^* - 1) \frac{\hat{CV}_i^2}{k(N^* - 1) - \frac{N^* - 2}{N^*} \sum_{i=1}^k (N^* - 1) \frac{\hat{CV}_i^2}{1 + \frac{N^* - 1}{N^*} \hat{CV}_i}}}{k(N^* - 1) - \frac{N^* - 2}{N^*} \sum_{i=1}^k (N^* - 1) \frac{\hat{CV}_i^2}{1 + \frac{N^* - 1}{N^*} \hat{CV}_i}}. 
\]

(3.3)

Zeigler [53] suggested a similar estimator based on Mckay’s approximation, but he gave a formula only for the constant sample size case. We will work with \( \hat{CV}_{MLE}^M \) because it is based on a more accurate approximate density.

**Maximum likelihood estimators for full likelihood**

The full log likelihood function for the data is a constant plus:

\[
-\ln(CV) \sum_{i=1}^k N_i - \sum_{i=1}^k N_i \ln(\mu_i) - \frac{1}{CV^2} \sum_{i=1}^k \frac{1}{2\mu_i^2} \left( (N_i - 1)S_i^2 + N_i(\bar{X}_i - \mu_i)^2 \right). 
\]

(3.4)

Setting the partial derivatives of Equation 3.4 equal to zero, after some algebra we get the following equations for \( \hat{CV}_{MLE}^2 \) and \( \hat{\mu}_{i, MLE} \):

\[
\hat{CV}_{MLE}^2 = \frac{\sum_{i=1}^k \frac{1}{\hat{\mu}_{i, MLE}^2} \left( (N_i - 1)S_i^2 + N_i(\bar{X}_i - \hat{\mu}_{i, MLE})^2 \right)}{\sum_{i=1}^k N_i}, 
\]

(3.5)

\[
\hat{CV}_{MLE}^2\hat{\mu}_{i, MLE}^2 + \bar{X}_i\hat{\mu}_{i, MLE} - \frac{(N_i - 1)}{N_i} S_i^2 - \bar{X}_i^2 = 0, \ i \in \{1, \ldots, k\}. 
\]

(3.6)

One root of Equation 3.6 will be negative. Since we are concentrating on the practical range, where \( \mu > 0 \), we can ignore this root. So we can rewrite the second likelihood equation as

\[
\hat{\mu}_{i, MLE} = \bar{X}_i \left( \frac{\sqrt{1 + 4\hat{CV}_{MLE}^2 \left( \frac{N_i - 1}{N_i} \hat{CV}_i^2 + 1 \right)} - 1}{2\hat{CV}_{MLE}^2} \right), \ i \in \{1, \ldots, k\}. 
\]

(3.7)

Lohrding [72] gave an analytical formula for \( CV_{MLE} \) for the case \( k = 2, N_1 = N_2 \equiv \)
Chapter 3. Additional Work on the Coefficient of Variation in Normal Populations

\( N^* \). His solution is

\[
\hat{CV}^{L}_{MLE} = \sqrt{\frac{2\sqrt{1 + a(\hat{CV}_1)}\sqrt{1 + a(\hat{CV}_2)} \left( \sqrt{1 + a(\hat{CV}_1)}\sqrt{1 + a(\hat{CV}_2)} - 1 \right)}{\sqrt{1 + a(\hat{CV}_1)} + \sqrt{1 + a(\hat{CV}_2)}}}
\]

where \( a(\hat{CV}) \equiv 2\frac{N^* - 1}{N^*} \hat{CV}^2 \). Zeigler \([53]\) reports a slightly different result, but that is a typo.

Gerig and Sen \([73]\) and Sinha et al. \([84]\) discuss the solution for the case \( k = 2 \), \( N_1 \neq N_2 \). A complicated analytical solution can be obtained using the equations they provide.

Doornbos and Dijkstra \([85]\) presented the first algorithm for obtaining an exact MLE estimator for \( CV \) for the general case. It is inconvenient in that it requires iteratively solving equation that are nonlinear in the parameters.

Gupta and Ma \([86]\) developed a simpler alternative algorithm. They derived that \( \hat{CV}_{MLE} \) solves the following equation:

\[
\sum_{i=1}^{k} N_i \left(1 + \sqrt{1 + 4 \left(1 + \frac{N_i - 1}{N_i} \hat{CV}_i^2 \right) (CV)^2}\right) = \sum_{i=1}^{k} N_i. \tag{3.8}
\]

Now if \( k = 1 \), one can show \( \hat{CV}_{MLE} = \sqrt{\frac{N_1 - 1}{N} \hat{CV}} \). So the following result by Gupta and Ma stands to reason:

\[
\min \left( \sqrt{\frac{N_1 - 1}{N_1} \hat{CV}_1} \ldots \sqrt{\frac{N_k - 1}{N_k} \hat{CV}_k} \right) \leq \hat{CV}_{MLE} \leq \max \left( \sqrt{\frac{N_1 - 1}{N_1} \hat{CV}_1} \ldots \sqrt{\frac{N_k - 1}{N_k} \hat{CV}_k} \right). \tag{3.9}
\]

Furthermore, the left-hand side of Equation 3.8 is monotonic in \( CV \). Thus, Equation 3.8 can be solved by the efficient method of bisection, with the intial upper and lower bounds given by Equation 3.9.
Chapter 3. Additional Work on the Coefficient of Variation in Normal Populations

Several authors have suggested approximating the MLE. Bennett’s [87] approximation requires solving a nonlinear equation, just like the exact solution of Gupta and Ma. Nairy and Rao [88] suggested using the second step in a Newton-Raphson algorithm as an approximate solution for Equations 3.5 and Equation 3.6. Sinha et al. [84] claim that for general \( k \),

\[
\hat{CV}_{MLE} \approx \sqrt{1 - \frac{1}{\sum_{i=1}^{k} \frac{N_i}{1 + \frac{N_i - 1}{N_i} \hat{CV}_i^2}}}
\]

This approximation underestimates the MLE slightly. The difference is noticeable for \( CV \) near the upper edge of the practical range.

An improved analytical approximation is

\[
\hat{CV}_{MLE} \approx \sqrt{\frac{\sum_{i=1}^{k} \frac{N_i}{1 + \frac{N_i - 1}{N_i} \hat{CV}_i^2}}{\sum_{i=1}^{k} \frac{N_i}{1 + \frac{N_i - 1}{N_i} \hat{CV}_i^2}}}
\]

(3.10)

Extensive computations indicate that for \( CV \leq 0.4 \), this approximation is accurate to the fourth nonzero decimal place, even for small \( N_i \)s and \( k \). Because of this accuracy, I shall treat Equation 3.10 as the exact MLE.

3.2.2 Theoretical comparison of estimators

One good property that all the estimators share is that they are all functions of the data only through the maximal invariant.

It is possible to derive approximate expected values and variances for each estimator. These expressions are complicated and not all that informative, so I shall not rely heavily on them. In any case, they are only approximations, and so the small-sample biases and variances would have to be determined by simulation. However, we can still make a few theoretical points.
Chapter 3. Additional Work on the Coefficient of Variation in Normal Populations

Bias of estimators

By design, the bias in $\hat{CV}$ should be small, in all samples.

Now

$$\hat{CV}^2_i = \frac{CV^2 \frac{U_i}{N_i}}{1 + 2 \frac{CV}{\sqrt{N_i}} z_i + CV^2 \frac{z_i^2}{N_i}}. $$

Then $E(\hat{CV}^2_i) = E \left( CV^2 \frac{1}{1 + 2 \frac{CV}{\sqrt{N_i}} z_i + CV^2 \frac{z_i^2}{N_i}} \right)$, and taking a second-order Taylor expansion (Definition 1.3.8.7) of the second factor around $\frac{1}{1+CV^2}$, we get

$$E(\hat{CV}^2_i) \approx CV^2 \frac{1}{1 + CV^2 \frac{1}{N_i}} \left( 1 + \frac{2CV^2}{N_i} + \frac{CV^4}{(1 + CV^2 \frac{1}{N_i})^2} \right) \equiv CV^2 a(CV, N_i).$$

Now $a(CV, N_i)$ will be slightly greater than 1, and will converge to 1 as $N_i$ increases. It’s largest value in the range of CV’s we are considering is 1.07 for $CV = 0.4$, $N_i = 2$. So $\sum_{i=1}^{k} \hat{CV}^2_i$ is essentially an unbiased estimator of $CV^2$; this makes sense since we saw in the last chapter that the bias in $\hat{CV}$ is due largely to the bias in $S$, and $S^2$ is actually unbiased. However, by Jentzen’s inequality, $E\sqrt{Y} < \sqrt{E(Y)}$, so $\hat{CV}_S$ should have some downward bias, except perhaps for small $N_i$ situations with a high $CV$.

From Equation 2.17 we can get

$$\frac{\partial \ln(C_M(\theta))}{\partial \theta} = \sum_{i=1}^{k} \frac{N_i - 1}{2} \frac{CV^2}{1 + \frac{N_i - 2}{N_i} CV^2}. $$

Then by Equation 2.17, Equation 3.2, and Theorem 1.3.4.1, we can derive

$$E \left( \sum_{i=1}^{k} \frac{(N_i - 1) (\hat{CV}_{MLE}^M)^2}{1 + \frac{N_i - 2}{N_i} (\hat{CV}_{MLE}^M)^2} \right) \equiv E(b(\hat{CV}_{MLE}^M))$$

$$\approx b(CV) \equiv \sum_{i=1}^{k} (N_i - 1) \frac{CV^2}{1 + \frac{N_i - 2}{N_i} CV^2}. $$

This approximation will be extremely close, since it is based ultimately on Proposition 4.

Now we essentially have $E(b(\hat{CV}_{MLE}^M)) = b(CV)$. If the sample sizes are equal,
Chapter 3. Additional Work on the Coefficient of Variation in Normal Populations

Equation 3.3 gives us $b^{-1}$. Taking the second derivative of $b^{-1}$, we find that it will be negative for all but extreme values of the $\hat{CV}_i$s. Then $b^{-1}$ is concave over the relevant part of its domain, and then by Jentzen’s inequality

$$b^{-1}(E(b(\hat{CV}_{MLE}^M))) = CV > E(b^{-1}(b(\hat{CV}_{MLE}^M))) = E(\hat{CV}_{MLE}^M), \tag{3.12}$$

and we would expect $\hat{CV}_{MLE}^M$ to be downward biased.

To compare $\hat{CV}_S$ and $\hat{CV}_{MLE}^M$, set $\hat{CV}_1 = \ldots = \hat{CV}_k = \hat{CV}$ and $N_1 = \ldots = N_k = N^*$ and take the ratio of Equation 3.1 over Equation 3.3. With some algebra, that ratio is $1 + \frac{1}{N^*} \hat{CV}^2$. The implications are that these two estimators should be very close and that $\hat{CV}_S > \hat{CV}_{MLE}^M$, and therefore presumably less biased.

From Equation 3.2 and Equation 3.10, we can derive

$$\sum_{i=1}^{k} \frac{N_i \hat{CV}_{MLE}^2}{1 + \hat{CV}_{MLE}^2} = \sum_{i=1}^{k} \frac{(N_i - 1)(\hat{CV}_{MLE}^M)^2}{1 + \frac{N_i - 2}{N_i}(\hat{CV}_{MLE}^M)^2}. \tag{3.13}$$

Consider the ratio

$$\frac{\frac{N_i \hat{CV}_{MLE}^2}{1 + \hat{CV}_{MLE}^2}}{\frac{(N_i - 1)(\hat{CV}_{MLE}^M)^2}{1 + \frac{N_i - 2}{N_i}(\hat{CV}_{MLE}^M)^2}}. \tag{3.14}$$

If $\hat{CV}_{MLE} = \hat{CV}_{MLE}^M$, then this ratio would be

$$\frac{\frac{N_i}{N_i - 1} \left(1 + \frac{N_i - 2}{N_i}(\hat{CV}_{MLE}^M)^2\right)}{1 + (\hat{CV}_{MLE}^M)^2} = \frac{\frac{N_i}{N_i - 1} + \frac{N_i - 2}{N_i - 1}(\hat{CV}_{MLE}^M)^2}{1 + (\hat{CV}_{MLE}^M)^2} \tag{3.15}$$

This will be at least 1 as long as $2(\hat{CV}_{MLE}^M)^2 \leq 1 + (\hat{CV}_{MLE}^M)^2$ or $\hat{CV}_{MLE}^M \leq 1$. But this ratio increases if $\hat{CV}_{MLE}$ increases, meaning that for values of $CV$ near the practical range, if $\hat{CV}_{MLE} \geq \hat{CV}_{MLE}^M$, then each term of the right hand side of Equation 3.13 will be greater than the corresponding term on the left hand side. Then in order for
the equality to hold, $\hat{CV}_{MLE} < \hat{CV}_M$. So $\hat{CV}_{MLE}$ will have a greater downward bias than $\hat{CV}_M$.

**Consistency of estimators**

There are two different ways for the number of observations in our data to converge to infinity: via increasing the sample sizes for individual populations, or by increasing the number of populations. Here I shall simplify the reasoning by assuming that $N_1 = \ldots = N_k = N^*$. Estimators that converge in probability with $N^*$ will be called $N^*$-consistent, and those that converge in probability with $k$ will be called $k$-consistent.

Each individual $\hat{CV}_i$ has a bias that converges upward to 0 with $N^*$. Thus, increasing $k$ without increasing $N^*$ will cause the uncorrected weighted average to converge in probability to a quantity that is less than the true $CV$, while increasing $N^*$ will bring about convergence to the desired quantity.

Since $\hat{CV}_{MV}$ is bias-corrected, the weighted average that we are working with should be consistent whether we increase $N^*$ or $k$.

Now $\hat{CV}_i^2$ converges in probability with $N^*$ to $CV^2$, so $\hat{CV}_S$ has $N^*$-consistency. As we have seen, there is essentially no bias in $\hat{CV}_i^2$, so the weighted average should converge in probability to a quantity very close to $CV^2$ as $k$ increases, imparting $k$-consistency to $\hat{CV}_S$ by Theorem 1.3.6.5 part 1. (There might be a slight upward bias for high $CV$, low $N^*$ cases.)

By the basic properties of the $MLE$ (Theorem 1.3.6.2 and Theorem 1.3.6.1 part 2), $\hat{CV}_{MLE}$ is $N^*$-consistent and $\hat{CV}_M$ is $k$-consistent. The $N^*$-consistency of $\hat{CV}_{MLE}$ can be shown arithmetically from Equation 3.3 after applying Proposition 3.

We can show, however, that $\hat{CV}_{MLE}$ is not $k$ consistent. (The reason that $k$-consistency of $\hat{CV}_{MLE}$ does not follow from Theorem 1.3.6.2 is that the Theorem assumes that the number of parameters does not grow with the sample size, but there will be $k$ sample means.) From Equation 3.13, setting $N_i = N^*$, conducting some algebra, and using the fact that $\hat{CV}_M$ converges with $k$ to $CV$, we get that $\hat{CV}_{MLE}$
converges with \( k \) to

\[
\sqrt{\frac{N^* - 1}{N^* - \frac{CV^2}{1 + \frac{N^* - 2CV^2}{1 + \frac{N^* - 1}{N^* - CV^2}}}}} = CV \sqrt{\frac{N^* - 1}{N^* - CV^2}}. \tag{3.16}
\]

\( \hat{CV}_{MLE} \) maintains a downward bias as \( k \) grows. This result is not surprising. Full-data maximum likelihood underestimates standard deviations for a normal population by a factor of \( \sqrt{\frac{N^* - 1}{N^*}} \); it ignores the loss in degrees of freedom from estimating the mean. Adding a number of independent samples from different populations to the data will not correct this bias, due to the need to estimate each mean. Using the marginal likelihood of the \( \hat{CV}_i \)s avoids this problem.

In summary, we would expect that for cases where \( N^* \) is large, the various estimators would be very similar, while for small \( N^* \), large \( k \) situations, the downward bias in \( \hat{CV}_{MLE} \) would make it less desirable.

**Variance of Estimators**

Under regularity conditions, the MLE is asymptotically efficient (see [2], Chapter 10). This would lead us to expect that the MLEs and \( \hat{CV}_S \), which is an approximation of \( \hat{CV}_{MLE} \), will have a smaller variance than \( \hat{CV}_{MV} \). This expectation is strengthened by the fact that, if Equation 2.17 is valid, \( b(\hat{CV}_{MLE}^M) \) is the minimum-variance unbiased estimator of \( b(CV) \), which can be shown using Corollary 7.3.15 of Casella and Berger [2].

**3.2.3 Simulation comparison of point estimators**

The only simulation evidence in the literature is in Zeigler [53]. He compared the bias and variance of \( \frac{1}{\sum_{i=1}^{N} N_i \hat{CV}_i} \), \( \hat{CV} \) from Equation 2.13, and the MLE based on McKay’s approximation. His simulations did not include small \( N^* \), large \( k \) cases (defined here as \( N^* < 5 \)) or cases where the sample sizes differed.

The simulation study here will compare the bias, standard deviation, and mean squared error of \( \hat{CV}_{MV}, \hat{CV}_S, \hat{CV}_{MLE}^M, \) and \( \hat{CV}_{MLE} \), with combinations of \( k = \)
Chapter 3. Additional Work on the Coefficient of Variation in Normal Populations

2, 4, 8, 16, 32 and \(N^* = 2, 3, 5, 10, 15, 20\). These combinations were chosen to encompass the range of situations where there might be noticeable difference among the estimators, including the small \(N^*\), large \(k\) case. I will show results for a small value of \(CV (0.05)\), and a value that is above the upper end of the practical range (0.4). Recall that since the estimators are functions of the data through the \(\hat{CV}_i\)s only, the simulation results will not depend on the \(\mu_i\)s.

The data in the simulations have been generated from a normal distribution that has been truncated at 0. For the \(CV = 0.4\), small \(N^*\) cases, this was necessary to avoid the problem of negative sample \(CV\)s. Also, because of the fact that in most cases the data will have nonnegative support, a normal model for the data is effectively a truncated normal model.

Tables 3.4 and 3.5 show the percentage bias in the various estimators for cases where \(N_1 = \ldots = N_k = N^*\). The largest standard error for any of the quantities reported in the two tables is 0.4.

As expected, \(\hat{CV}_{MV}\) has almost no bias for small \(CV\). But there is some bias in \(\hat{CV}_{MV}\) due to the need to estimate the correction factor \(1 + \frac{CV^2}{N_i}\), and as it turns out, this source of bias is noticeable for high values of \(CV\). \(\hat{CV}_S\) generally has the smallest bias for high \(CV\). The convergence of \(\hat{CV}_{MLE}^M\) with \(k\) is slow with high \(CV\); we notice some differences between \(\hat{CV}_S\) and \(\hat{CV}_{MLE}^M\) with high \(CV\), but not with low \(CV\). \(\hat{CV}_{MLE}^M\) turns out to have unacceptable bias; as predicted, increases in \(k\) do not eliminate the bias.

Table 3.6 looks at the bias in cases where half the \(N_i\)s equal \(N_i\), and half equal \(N_h\). This table does not alter the conclusions from the equal-sample size case. The highest standard error in Table 3.6 is 0.25.

Tables 3.7 through 3.9 have the square root of the mean squared errors of the point estimators, expressed as a percentage of the underlying \(CV\). The standard error of the mean squared error for a scenario is \(\sqrt{\frac{1}{s} \sum_{i=1}^{s} (T_i - CV)^2} \times MSE^2\), where \(T_i\) is the value of the point estimate of \(CV\) from the \(i\)th simulated dataset and \(s\) is the number of simulated datasets. The standard error of the root mean squared error is \(\frac{1}{2RMSE}\) times the standard error of the mean squared error. The largest standard error for a quantity in the \(RMSE\) tables in 0.3. For low values of \(CV\), \(\hat{CV}_S\) and \(\hat{CV}_{MLE}^M\) have slightly
lower RMSE than $\overline{CV}_{MV}$. For high values, the two low-RMSE estimators are $\overline{CV}_{MV}$ and $\overline{CV}^M_{MLE}$. In light of the bias in $\overline{CV}_{MLE}$, it is not surprising that it has the highest RMSE.

The estimators in the first three columns in the tables above are all very close in terms of their bias and RMSE. $\overline{CV}^M_{MLE}$ can be a bit more tricky to compute if the sample sizes are not equal. $\overline{CV}_{S}$ is the most easily computed. On the whole, the properties of the estimators in the first three columns offer only slight improvements
Table 3.5: Bias in point estimate as percentage of common \( CV \), \( CV = 0.40 \)

<table>
<thead>
<tr>
<th>( N^* )</th>
<th>( k )</th>
<th>( \hat{CV}_{MV} )</th>
<th>( \hat{CV}_S )</th>
<th>( \hat{CV}_{MLE} )</th>
<th>( \hat{CV}_{MLE} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4</td>
<td>-6.9</td>
<td>-0.2</td>
<td>-9.8</td>
<td>-33.1</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>-5.0</td>
<td>4.0</td>
<td>-6.8</td>
<td>-31.3</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>-4.4</td>
<td>5.5</td>
<td>-5.8</td>
<td>-30.7</td>
</tr>
<tr>
<td>2</td>
<td>32</td>
<td>-4.2</td>
<td>6.2</td>
<td>-5.5</td>
<td>-30.5</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>-5.7</td>
<td>-3.7</td>
<td>-8.9</td>
<td>-23.2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>-3.9</td>
<td>0.6</td>
<td>-5.9</td>
<td>-20.9</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>-3.6</td>
<td>2.1</td>
<td>-5.0</td>
<td>-20.3</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>-3.5</td>
<td>2.8</td>
<td>-4.7</td>
<td>-20.1</td>
</tr>
<tr>
<td>3</td>
<td>32</td>
<td>-3.3</td>
<td>3.4</td>
<td>-4.3</td>
<td>-19.8</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>-3.9</td>
<td>-2.9</td>
<td>-5.9</td>
<td>-14.4</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>-3.6</td>
<td>-1.1</td>
<td>-4.9</td>
<td>-13.5</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>-3.4</td>
<td>-0.2</td>
<td>-4.3</td>
<td>-13.1</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>-3.1</td>
<td>0.5</td>
<td>-3.9</td>
<td>-12.7</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>-3.2</td>
<td>-2.8</td>
<td>-4.3</td>
<td>-8.5</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>-2.9</td>
<td>-1.7</td>
<td>-3.5</td>
<td>-7.8</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>-3.0</td>
<td>-1.4</td>
<td>-3.5</td>
<td>-7.7</td>
</tr>
<tr>
<td>10</td>
<td>16</td>
<td>-3.0</td>
<td>-1.3</td>
<td>-3.4</td>
<td>-7.7</td>
</tr>
</tbody>
</table>

over those explored by Zeigler [53].

3.3 Literature Review on Test of \( CV \) Homogeneity in Normal Populations

Section 3.3 will use the notation of Section 2.4.

There is a rather large literature on testing \( CV \) homogeneity against a general
## Additional Work on the Coefficient of Variation in Normal Populations

### Table 3.6: Bias in point estimate as percentage of common \( CV \), sample sizes not identical

<table>
<thead>
<tr>
<th>( CV )</th>
<th>( N_l )</th>
<th>( N_h )</th>
<th>( k )</th>
<th>( CV_{MV} )</th>
<th>( CV_S )</th>
<th>( CV_{MLE}^M )</th>
<th>( CV_{MLE} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>2</td>
<td>10</td>
<td>2</td>
<td>-0.2</td>
<td>-2.7</td>
<td>-2.6</td>
<td>-11.1</td>
</tr>
<tr>
<td>0.05</td>
<td>2</td>
<td>10</td>
<td>4</td>
<td>0.0</td>
<td>-1.3</td>
<td>-1.3</td>
<td>-9.9</td>
</tr>
<tr>
<td>0.05</td>
<td>2</td>
<td>10</td>
<td>8</td>
<td>0.0</td>
<td>-0.6</td>
<td>-0.6</td>
<td>-9.3</td>
</tr>
<tr>
<td>0.05</td>
<td>2</td>
<td>10</td>
<td>16</td>
<td>0.0</td>
<td>-0.2</td>
<td>-0.3</td>
<td>-9.0</td>
</tr>
<tr>
<td>0.05</td>
<td>2</td>
<td>10</td>
<td>32</td>
<td>-0.1</td>
<td>-0.2</td>
<td>-0.2</td>
<td>-8.9</td>
</tr>
<tr>
<td>0.05</td>
<td>10</td>
<td>20</td>
<td>2</td>
<td>-0.1</td>
<td>-0.9</td>
<td>-0.9</td>
<td>-4.3</td>
</tr>
<tr>
<td>0.05</td>
<td>5</td>
<td>10</td>
<td>4</td>
<td>0.0</td>
<td>-0.9</td>
<td>-0.9</td>
<td>-7.7</td>
</tr>
<tr>
<td>0.05</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>0.1</td>
<td>-1.1</td>
<td>-1.1</td>
<td>-16.4</td>
</tr>
<tr>
<td>0.05</td>
<td>2</td>
<td>5</td>
<td>16</td>
<td>0.1</td>
<td>-0.4</td>
<td>-0.5</td>
<td>-15.9</td>
</tr>
<tr>
<td>0.05</td>
<td>2</td>
<td>5</td>
<td>32</td>
<td>0.0</td>
<td>-0.1</td>
<td>-0.2</td>
<td>-15.6</td>
</tr>
<tr>
<td>0.4</td>
<td>2</td>
<td>10</td>
<td>2</td>
<td>-3.4</td>
<td>-2.8</td>
<td>-5.2</td>
<td>-12.6</td>
</tr>
<tr>
<td>0.4</td>
<td>2</td>
<td>10</td>
<td>4</td>
<td>-3.3</td>
<td>-1.2</td>
<td>-4.3</td>
<td>-11.8</td>
</tr>
<tr>
<td>0.4</td>
<td>2</td>
<td>10</td>
<td>8</td>
<td>-3.2</td>
<td>-0.5</td>
<td>-4.0</td>
<td>-11.5</td>
</tr>
<tr>
<td>0.4</td>
<td>2</td>
<td>10</td>
<td>16</td>
<td>-2.9</td>
<td>0.1</td>
<td>-3.6</td>
<td>-11.2</td>
</tr>
<tr>
<td>0.4</td>
<td>2</td>
<td>10</td>
<td>32</td>
<td>-3.0</td>
<td>0.2</td>
<td>-3.6</td>
<td>-11.1</td>
</tr>
<tr>
<td>0.4</td>
<td>10</td>
<td>20</td>
<td>2</td>
<td>-3.0</td>
<td>-2.7</td>
<td>-3.6</td>
<td>-6.4</td>
</tr>
<tr>
<td>0.4</td>
<td>5</td>
<td>10</td>
<td>4</td>
<td>-3.2</td>
<td>-1.6</td>
<td>-4.1</td>
<td>-9.8</td>
</tr>
<tr>
<td>0.4</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>-3.5</td>
<td>1.3</td>
<td>-4.8</td>
<td>-17.9</td>
</tr>
<tr>
<td>0.4</td>
<td>2</td>
<td>5</td>
<td>16</td>
<td>-3.4</td>
<td>1.9</td>
<td>-4.4</td>
<td>-17.7</td>
</tr>
<tr>
<td>0.4</td>
<td>2</td>
<td>5</td>
<td>32</td>
<td>-3.3</td>
<td>2.4</td>
<td>-4.2</td>
<td>-17.5</td>
</tr>
</tbody>
</table>
alternative that there is at least some difference. Proposed tests can be divided into 4 categories.

### 3.3.1 Tests based on the likelihood of the full data

The likelihood ratio (LR) test was first proposed by Miller and Karson [22] and further explored by Bennett [87], Doornbos and Dijkstra [85], and Bhoj and Ahsanullah [89]. In this chapter, when I refer to the “likelihood ratio statistic,” I actually mean the
transformation of the statistic that is asymptotically $\chi^2$ (Theorem 1.3.6.13).

Gupta and Ma [86] also proposed a score test.

All proposed likelihood-based tests in the literature rely on the well-known asymptotic $\chi^2$ distributions of likelihood-based statistics for their $p$-values (see [8]). Also, all of these tests utilize the full likelihood rather than the likelihood of the sample $CV$s.

Using the solutions to Equation 3.8 and Equation 3.6, Gupta and Ma’s formula for
Table 3.9: Root MSE as percentage of common CV

<table>
<thead>
<tr>
<th>CV</th>
<th>N_t</th>
<th>N_h</th>
<th>k</th>
<th>( \bar{CV}_{MV} )</th>
<th>( \bar{CV}_S )</th>
<th>( \bar{CV}_{MLE} )</th>
<th>( \bar{CV}_{MLE} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>2</td>
<td>10</td>
<td>2</td>
<td>22.8</td>
<td>22.6</td>
<td>22.2</td>
<td>23.0</td>
</tr>
<tr>
<td>0.05</td>
<td>2</td>
<td>10</td>
<td>4</td>
<td>16.0</td>
<td>16.1</td>
<td>15.7</td>
<td>17.4</td>
</tr>
<tr>
<td>0.05</td>
<td>2</td>
<td>10</td>
<td>8</td>
<td>11.6</td>
<td>11.6</td>
<td>11.3</td>
<td>13.9</td>
</tr>
<tr>
<td>0.05</td>
<td>2</td>
<td>10</td>
<td>16</td>
<td>8.1</td>
<td>8.1</td>
<td>8.0</td>
<td>11.5</td>
</tr>
<tr>
<td>0.05</td>
<td>2</td>
<td>10</td>
<td>32</td>
<td>5.7</td>
<td>5.7</td>
<td>5.6</td>
<td>10.3</td>
</tr>
<tr>
<td>0.05</td>
<td>10</td>
<td>20</td>
<td>2</td>
<td>13.4</td>
<td>13.3</td>
<td>13.3</td>
<td>13.5</td>
</tr>
<tr>
<td>0.05</td>
<td>5</td>
<td>10</td>
<td>4</td>
<td>14.1</td>
<td>13.9</td>
<td>13.8</td>
<td>15.0</td>
</tr>
<tr>
<td>0.05</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>16.4</td>
<td>16.2</td>
<td>15.8</td>
<td>21.2</td>
</tr>
<tr>
<td>0.05</td>
<td>2</td>
<td>5</td>
<td>16</td>
<td>11.7</td>
<td>11.5</td>
<td>11.3</td>
<td>18.5</td>
</tr>
<tr>
<td>0.05</td>
<td>2</td>
<td>5</td>
<td>32</td>
<td>8.3</td>
<td>8.1</td>
<td>8.0</td>
<td>17.0</td>
</tr>
<tr>
<td>0.4</td>
<td>2</td>
<td>10</td>
<td>2</td>
<td>22.9</td>
<td>24.9</td>
<td>22.9</td>
<td>24.5</td>
</tr>
<tr>
<td>0.4</td>
<td>2</td>
<td>10</td>
<td>4</td>
<td>16.3</td>
<td>17.6</td>
<td>16.2</td>
<td>18.8</td>
</tr>
<tr>
<td>0.4</td>
<td>2</td>
<td>10</td>
<td>8</td>
<td>11.6</td>
<td>12.4</td>
<td>11.6</td>
<td>15.4</td>
</tr>
<tr>
<td>0.4</td>
<td>2</td>
<td>10</td>
<td>16</td>
<td>8.6</td>
<td>8.9</td>
<td>8.6</td>
<td>13.3</td>
</tr>
<tr>
<td>0.4</td>
<td>2</td>
<td>10</td>
<td>32</td>
<td>6.4</td>
<td>6.4</td>
<td>6.6</td>
<td>12.3</td>
</tr>
<tr>
<td>0.4</td>
<td>10</td>
<td>20</td>
<td>2</td>
<td>14.0</td>
<td>14.2</td>
<td>14.0</td>
<td>14.8</td>
</tr>
<tr>
<td>0.4</td>
<td>5</td>
<td>10</td>
<td>4</td>
<td>14.3</td>
<td>14.9</td>
<td>14.3</td>
<td>16.4</td>
</tr>
<tr>
<td>0.4</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>15.9</td>
<td>18.1</td>
<td>15.6</td>
<td>22.3</td>
</tr>
<tr>
<td>0.4</td>
<td>2</td>
<td>5</td>
<td>16</td>
<td>11.6</td>
<td>13.0</td>
<td>11.5</td>
<td>20.1</td>
</tr>
<tr>
<td>0.4</td>
<td>2</td>
<td>5</td>
<td>32</td>
<td>8.5</td>
<td>9.4</td>
<td>8.5</td>
<td>18.7</td>
</tr>
</tbody>
</table>
Chapter 3. Additional Work on the Coefficient of Variation in Normal Populations

the actual likelihood ratio statistic to carry out the test is

$$-2 \ln \lambda \equiv \sum_{i=1}^{k} N_i \ln \left( \frac{\hat{\mu}_{i, MLE}^2 \hat{CV}_{MLE}^2}{\frac{N_i - 1}{N_i} S_i^2} \right).$$

(3.17)

This statistic is asymptotically $\chi^2_{k-1}$.

Gupta and Ma also derived the score statistic:

$$\hat{CV}_{MLE}^2 \left( 2 \hat{CV}_{MLE}^2 + 1 \right) \sum_{i=1}^{k} \left( \frac{1}{N_i} \left( \frac{\Sigma_{j=1}^{N_i} (X_{ij} - \hat{\mu}_{i, MLE})^2}{\hat{\mu}_{i, MLE}^2 \hat{CV}_{MLE}^3} - \frac{N_i}{\hat{CV}_{MLE}} \right) \right)^2,$$

(3.18)

which again is approximately $\chi^2_{k-1}$.

Pardo and Pardo [90] developed tests based on Renyi’s divergence, which is a measure of the distance between two densities. For two multivariate densities on $\mathbb{R}^p$ from the same family $f$ parameterized by $\theta_1$ and $\theta_2$ respectively, Renyi’s divergence between them is

$$\frac{1}{b(b-1)} \ln \left( \int_{\mathbb{R}^p} f_{Y}(y; \theta_1)^b f_{Y}(y; \theta_2)^{1-b} dy \right),$$

where $b$ is arbitrarily chosen. If $\hat{\theta}_{MLE}$ and $\hat{\theta}_{R}$ are substituted for $\theta_1$ and $\theta_2$, where $\hat{\theta}_{R}$ is the maximum likelihood estimate of the parameters under $r$ restrictions, Renyi’s divergence becomes a statistic for testing the restrictions. Morales et al [91] proved that if certain regularity conditions hold, the Renyi’s divergence statistic, multiplied by $2N$, is asymptotically $\chi^2_r$.

If $f$ represents an exponential family, then as $b \rightarrow 0$, the scaled Renyi’s statistic converges to $-2 \ln \lambda$, where $\lambda$ is the likelihood ratio. Thus, the test based on Renyi’s divergence is a generalization of the likelihood ratio test. Different choices of $b$ generate different tests.

### 3.3.2 Tests based on the likelihood of the sample $CV$s.

Bennett’s test and the modified Bennett test were mentioned in Section 2.4.2. The “slight mistake” in Bennett’s formula comes from the fact that he substitutes $\sqrt{\frac{N}{N-1} \hat{CV}}$ for $\hat{CV}$ in McKay’s approximation (Equation 2.4) before deriving the test statistic using Pitman’s formula. The test statistics are likelihood ratios from an approximate marginal likelihood for the sample $CV$s.
Chapter 3. Additional Work on the Coefficient of Variation in Normal Populations

3.3.3 Tests based on the delta method approximation

The Wald test of a hypothesis is based on a quadratic form of $h(\hat{\theta}_{MLE}, \hat{\theta}_R)$, a function that indicates how far away $\hat{\theta}_{MLE}$ is from the set of parameter vectors that would fulfill the null hypothesis. One of the keys to the Wald test is that any nuisance parameters in the expression for the statistic are replaced by a consistent estimate. Theorem 1.3.6.2, Theorem 1.3.5.11, part 1, Theorem 1.3.6.5, and Theorem 1.3.6.6 together imply that the Wald statistic for testing the null of CV homogeneity is asymptotically $\chi^2_{k-1}$, which is how the $p$-value is calculated for the Wald test. The Wald statistic draws independent justification in this case from the delta-method approximation for the sample CV in Equation 2.6.

If $k = 2$, the Wald test statistic is

$$\sqrt{\frac{N_1-1}{N_1} CV_1 - \frac{N_2-2}{N_2} CV_2} \times \frac{\sqrt{\frac{N_1-1}{N_1} CV_1^2} + \left(\frac{N_1-1}{N_1}\right)^2 \frac{CV_1^4}{N_1} + \frac{N_2-1}{N_2} CV_2^2 + \left(\frac{N_2-1}{N_2}\right)^2 \frac{CV_2^4}{N_2}}{\sqrt{N_1 \cdot CV_1^2}}.$$

Closely-related test statistics were proposed independently by Rao and Vidya [92] and Bhoj and Ahsanullah [89]. Gupta and Ma [86] extended the Wald test to general $k$ and sample sizes. Their statistic is

$$h^T \left( H^T G^{-1} H \right)^{-1} h,$$

where I have derived that

$$h \equiv \left( \begin{array}{c} \sqrt{\frac{N_1-1}{N_1} CV_1} - \sqrt{\frac{N_2-2}{N_2} CV_2} \\ \vdots \\ \sqrt{\frac{N_{k-1}-1}{N_{k-1}} CV_{k-1}} - \sqrt{\frac{N_k-1}{N_k} CV_k} \end{array} \right),$$

118
where

\[
H = \begin{pmatrix}
-\sqrt{\frac{N_1-1}{N_1}}\hat{CV}_1 \left(\frac{1}{\bar{X}_1}\right) & \ldots & 0 \\
\sqrt{\frac{N_2-1}{N_2}}\hat{CV}_2 \left(\frac{1}{\bar{X}_2}\right) & \ldots & 0 \\
0 & \ldots & 0 \\
\vdots & \ldots & \vdots \\
0 & \ldots & -\sqrt{\frac{N_k-1}{N_k-1}}\hat{CV}_{k-1} \left(\frac{1}{\bar{X}_{k-1}}\right) \\
0 & \ldots & \sqrt{\frac{N_k-1}{N_k}}\hat{CV}_k \left(\frac{1}{\bar{X}_k}\right) \\
1/\bar{X}_1 & \ldots & 0 \\
-1/\bar{X}_2 & \ldots & 0 \\
0 & \ldots & 0 \\
\vdots & \ldots & \vdots \\
0 & \ldots & 1/\bar{X}_{k-1} \\
0 & \ldots & 1/\bar{X}_k
\end{pmatrix}
\]

and

\[
G^{-1} \equiv \begin{pmatrix}
\frac{N_1-1}{N_1}S^2 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
0 & \ddots & \ldots & 0 & 0 & 0 & \ldots & 0 \\
0 & \ldots & \ddots & \ldots & 0 & 0 & \ldots & 0 \\
0 & \ldots & 0 & \frac{N_k-1}{N_k}S^2 & 0 & 0 & \ldots & 0 \\
0 & 0 & \ldots & 0 & \frac{N_1-1}{N_1}S^2 & 0 & \ldots & 0 \\
0 & 0 & \ldots & 0 & 0 & \ddots & \ldots & 0 \\
0 & 0 & \ldots & 0 & 0 & \ldots & \ddots & 0 \\
0 & 0 & \ldots & 0 & 0 & 0 & \ldots & \frac{N_k-1}{N_k}S^2
\end{pmatrix}
\]

Since the CVs are equal under \(H_0\), the efficiency of the Wald statistic can be improved by substituting a weighted average of the \(\hat{CV}_i\)s for each \(\hat{CV}_i\) in the expres-
Chapter 3. Additional Work on the Coefficient of Variation in Normal Populations

sions for $H$ and $G^{-1}$, thereby reducing the variation of the estimated variances of the $\tilde{CV}_i$s. Miller [81] proposes a variant of the Wald statistic that does just this. He uses a slightly different $h$, in which the $i$th element is $\tilde{CV}_i - \tilde{CV}_k$, with the $H$ matrix modified appropriately.

Feltz and Miller [11] pointed out that there is no “correct” $h$ vector and that the Wald test is not invariant to the choice of $h$. A researcher using the Wald test must not change the $h$ vector looking for the result he wants. The $h$ vector whose $i$th element is the difference of $\tilde{CV}_i$ from a sample size-weighted average would seem to be reasonable.

In keeping with this, Feltz and Miller proposed the statistic for their test mentioned in Chapter 2. $FM$ (Equation 2.21) is not technically a Wald statistic, because it does not take covariance between terms of the form $\tilde{CV}_i - \tilde{CV}$ into account. Ahmed [60] independently proposed a statistic that is identical to Feltz and Miller’s except that $\frac{\sum_{i=1}^{k} N_i \tilde{CV}_i}{\sum_{i=1}^{k} N_i}$ is substituted for $\bar{CV}$, and in keeping with this, $N_i - 1$ is replaced by $N_i$ in the denominator.

Both Ahmed’s and Feltz-Miller’s statistics are asymptotically $\chi^2_{k-1}$ under the null as the sample sizes grow, as discussed in Chapter 2.

### 3.3.4 Tests based on $\frac{1}{CV}$

Nairy and Rao [88] claim to create new tests by deriving the likelihood ratio and score tests for the null hypothesis $\frac{1}{CV_1} = \ldots = \frac{1}{CV_k}$. However, the likelihood ratio and score tests are invariant to reparameterization, so their tests should not perform any differently than Gupta and Ma’s.

They also present a Wald test based on the row vector

$$NR \equiv \left( \begin{array}{c} \frac{1}{CV_1} - \frac{1}{CV_2} \\ \cdots \\ \frac{1}{CV_1} - \frac{1}{CV_k} \end{array} \right).$$

The delta-method approximation of $Var\left(\frac{1}{CV_i}\right)$ is $\frac{1}{N_i} \left(1 + \frac{1}{CV_i^2}\right)$. Thus, an estimate of
the covariance matrix of $NR$ is

$$
\begin{pmatrix}
\widehat{\text{Var}}\left(\frac{1}{CV_1}\right) + \widehat{\text{Var}}\left(\frac{1}{CV_2}\right) & \widehat{\text{Var}}\left(\frac{1}{CV_1}\right) & \cdots & \widehat{\text{Var}}\left(\frac{1}{CV_1}\right) \\
\vdots & \ddots & \vdots \\
\widehat{\text{Var}}\left(\frac{1}{CV_1}\right) & \cdots & \widehat{\text{Var}}\left(\frac{1}{CV_1}\right) + \widehat{\text{Var}}\left(\frac{1}{CV_k}\right)
\end{pmatrix},
$$

where $\widehat{\text{Var}}\left(\frac{1}{CV_i}\right) = \frac{1}{N_i} \left(1 + \frac{1}{CV_i^2}\right)$.

This statistic is asymptotically $\chi^2_{k-1}$.

Doornbos and Dijkstra [85] use a related statistic for their test: $\sum_{i=1}^{k} N_i \left(\frac{1}{CV_i} - \frac{1}{CV}\right)^2$, where $\frac{1}{CV}$ is the sample-size weighted average of the $\frac{1}{CV_i}$s. They derive the mean and variance of this statistic using the noncentral $t$ distribution of the $\frac{\sqrt{N_i}}{CV_i}$s, and create a standardized version of their statistic which has an asymptotic $\chi^2_{k-1}$ distribution.

Hedges and Olkin [93] follow a similar strategy. The difference is that their statistic is a sum of squared standardized deviations, while Doornbos and Dijkstra’s is a standardized sum of squared deviations.

### 3.3.5 Evaluating existing tests

Summarizing, we have a dozen tests.

- **Likelihood-based tests:** Likelihood ratio, Score, Renyi’s Divergence.
- **McKay’s approximation:** Bennett, Modified Bennett.
- **Delta method tests:** Wald, Miller, Feltz and Miller, Ahmed.
- **Inverse $CV$ tests:** Wald, Doornbos and Dijkstra, Hedges and Olkin.

For practical reasons, a practitioner would report only a few of these in any analysis. Here, I shall argue that there is no reason to keep more than two in the toolbox.

Three review studies have been done on the topic of testing $CV$ homogeneity against a general alternative in normal populations: Gupta and Ma [86], Fung and Tsang [94], and Nairy and Rao [88]. These studies are helpful but do not present a complete picture. In none of Gupta and Ma’s reported simulations are all of the population $CV$s in the practical range. Also, they present their data in charts rather than in tables,
so it is impossible to ascertain whether differences between tests in Type I or Type II error are statistically significant. Fung and Tsang consider only two of the above tests – the likelihood ratio test and Feltz/Miller. Nairy and Rao’s simulations consider more tests, but only one of the two tests we shall identify as most promising.

We can rule out Renyi’s divergence statistic because it is not scale invariant. Pardo and Pardo [90] show that Renyi’s divergence statistic is a monotonic function of

\[
\sum_{i=1}^{k} \frac{N_i(\bar{X}_i - \hat{\mu}_{i,MLE})^2}{(1 - b)\frac{N_i-1}{N_i}S_i^2 + b\hat{\mu}_{i,MLE}^2CV_{MLE}^2} + \frac{N_i}{b(1 - b)} \ln \left( \frac{(1 - b)\frac{N_i-1}{N_i}S_i^2 + b\hat{\mu}_{i,MLE}^2CV_{MLE}^2}{(N_i-1)^{1-b}S_i^{2-2b} + \hat{\mu}_{i,MLE}^{2b}CV_{MLE}^{2b}} \right).
\]

Via simulation, I have verified that for general \(b\), the distribution of this statistic depends on the \(\mu_i\)s. Recall from Section 2.1.2 that from the invariance principle, inference on the \(CV\) should not be dependent on the population means.

However, the remaining 11 test statistics are scale invariant. For the latter three categories of tests, this can be seen from the fact that all are functions of the data only through the \(\hat{CV}_i\)s. (For the Wald test, this takes some straightforward matrix multiplication to show.)

Consider the likelihood ratio statistic in Equation 3.17. Now \(\hat{CV}_{MLE}\) is a function of the \(\hat{CV}_i\)s alone, so we can write it as \(\hat{CV}_{MLE}(\hat{CV}_1, \ldots, \hat{CV}_k)\). Plugging this into Equation 3.7, we can then write \(\hat{\mu}_{i,MLE} = a(N_i, \hat{CV}_i, CV)\bar{X}_i\). Then from 3.17 the likelihood ratio statistic is

\[
-2\ln\lambda = \sum_{i=1}^{k} N_i \ln \left( \frac{a(N_i, \hat{CV}_i, CV)\bar{X}_i^2\hat{CV}_{MLE}(\hat{CV}_1, \ldots, \hat{CV}_k)}{\frac{N_i-1}{N_i}S_i^2} \right),
\]

which is a function of the data only through the \(\hat{CV}_i\)s.
Chapter 3. Additional Work on the Coefficient of Variation in Normal Populations

To show that the distribution of the score statistic does not depend on the \( \mu_i \)'s, from Equation 3.18 it is sufficient to show that \( \frac{(X_{ij} - \hat{\mu}_i, MLE)^2}{\hat{\mu}_i^2, MLE} \) has a distribution that does not depend on the \( \mu_i \)'s. We can write this as

\[
\frac{(\mu_i \frac{X_{ij}}{\mu_i} - \hat{\mu}_i, MLE)^2}{\hat{\mu}_i^2, MLE} = \left( \frac{\mu_i}{\hat{\mu}_i, MLE} \frac{X_{ij}}{\mu_i} - 1 \right)^2 = \frac{\mu_i X_{ij}}{\bar{X}_i \mu_i} \frac{1}{a(N_i, \hat{CV}_i, CV)} - 1.
\]

And from Section 2.1.2, we can deduce that the distribution of \( \frac{X_{ij}}{\mu_i} \) and thus \( \frac{\bar{X}_i}{\mu_i} \) is free of \( \mu_i \).

However, both the score and LR tests have other shortcomings. Simulation results from Doornbos and Dijkstra [85], Fung and Tsang [94], and Nairy and Rao [88] indicate that the likelihood ratio test is unacceptably liberal – it over-rejects in small samples \( (N_i \leq 15) \). Nairy and Rao’s simulations indicate that the score test is the opposite – unacceptably conservative.

We can also rule out Bennett’s test, since the modified Bennett test is more consistent with the original motivation of the test. Simulations in Shafer and Sullivan [80] and elsewhere indicate that in any case, the difference between the two versions is miniscule.

Since the Feltz-Miller test bases its delta-method variance estimates on an estimate of the common \( CV \) while the Wald test bases its variance estimates on the individual \( \hat{CV}_i \)'s, the Feltz-Miller test will be less variable and presumably more powerful. For the exact same reason, we would expect the Feltz-Miller test to be preferable to the inverse Wald test.

Feltz-Miller is preferable to Miller’s original test because of the implicit choice of \( h \). Their \( h \) corrects the problem in Miller’s original test that the value of the statistic depends on how one numbers the populations.

The differences between Ahmed’s test and Feltz and Miller’s will be small. Which is preferable depends on whether weighting by degrees of freedom or by sample size produces a better \( \chi^2 \) approximation. The fact that using degrees of freedom in the expression for the approximate variance produces a more accurate delta-method ap-
proximation would favor Feltz and Miller’s test.

Doornbos and Dijkstra as well as Feltz and Miller have done simulation studies of the Doornbos-Dijkstra test based on the non-central $t$. One set of authors has a mistake in their code, because Doornbos and Dijkstra find that their test is too conservative while Feltz and Miller find that the test is too liberal. For instance, if $CV = 0.1$, $k = 4$, and $N_1 = \ldots = N_4 = 10$, for a test of nominal size 0.05, Doornbos and Dijkstra find that the actual size is 0.034 (standard error 0.007) while Feltz and Miller’s results show an actual size of 0.156 (standard error 0.002). (The conservatism of the Doornbos and Dijkstra test is apparent in all the scenarios in their paper, so it is safe to conclude that their code yields a conservative test despite the high standard error for that one scenario.) Whichever set of authors is right, the Doornbos and Dijkstra test does not have the right size.

Feltz and Miller provide simulation results for the Hedges and Olkin test. For $k = 4$, it appears to have about the same size as the variants of Bennett’s test, but for $k = 2$ it has much smaller power.

Thus, we have seen strikes against the Likelihood ratio test, Renyi’s divergence test, the Bennett test, the Wald test, Miller’s test, Ahmed’s test, the inverse $CV$ Wald test, Doornbos and Dijkstra’s test, and the Hedges and Olkin test. This leaves us with the modified Bennett test and Feltz-Miller. Feltz and Miller found that their test and the modified Bennett test had about the same Type I error and power.

### 3.4 Using a stochastic representation to obtain confidence intervals

Section 3.4 explains an idea to create an exact (up to Monte Carlo error) confidence interval for the common $CV$ of $k$ normal samples and a Monte Carlo idea to create an approximate confidence interval for the difference between two $CV$s.
3.4.1 Exact confidence interval for common CV

While Tian [20] based an approximate test on the weighted average of the sample CVs, a fiducial test that turns out to be exact can be based on the weighted average of the inverse sample CVs.

From Equation 2.1, we get

\[
\overline{CV} = \sum_{i=1}^{k} \frac{(N_i - 1)}{CV_i} = \left( \frac{1}{CV} \sum_{i=1}^{k} \frac{(N_i - 1)^{3/2}}{\sqrt{U_i}} + \sum_{i=1}^{k} \frac{(N_i - 1)^{3/2}}{\sqrt{N_i}} \frac{Z_i}{\sqrt{U_i}} \right)
\]

Let the observed value of \( \sum_{i=1}^{k} \frac{N_i - 1}{CV_i} \) be denoted \( \overline{CV}_O \). Then under \( H_0 : CV = CV_o \),

\[
\text{Prob}_{CV_o} \left( \sum_{i=1}^{k} \frac{N_i - 1}{CV_i} > \overline{CV}_O \right) = \text{Prob} \left( \frac{1}{CV_o} \sum_{i=1}^{k} \frac{(N_i - 1)^{3/2}}{\sqrt{U_i}} + \sum_{i=1}^{k} \frac{(N_i - 1)^{3/2} Z_i}{\sqrt{N_i}} \sqrt{U_i} > \overline{CV}_O \right) = \text{Prob} \left( \frac{1}{CV_o} > \frac{\overline{CV}_O - \sum_{i=1}^{k} \frac{(N_i - 1)^{3/2}}{\sqrt{U_i}}}{\sum_{i=1}^{k} \frac{(N_i - 1)^{3/2}}{\sqrt{N_i}} \frac{Z_i}{\sqrt{U_i}}} \right) = \text{Prob} \left( CV_o < \frac{\overline{CV}_O - \sum_{i=1}^{k} \frac{(N_i - 1)^{3/2}}{\sqrt{U_i}}}{\sum_{i=1}^{k} \frac{(N_i - 1)^{3/2}}{\sqrt{N_i}} \frac{Z_i}{\sqrt{U_i}}} \equiv E \right). \tag{3.21}
\]

This is an exact p-value for testing the hypothesis \( H_o : CV = CV_o \) against \( H_a : CV < CV_o \) using the rejection region of the form \( \sum_{i=1}^{k} \frac{N_i - 1}{CV_i} > c \). For the p-value for the alternative \( H_a : CV > CV_o \), reverse the sign of the final inequality. For two-sided p-values, multiply the smaller one-sided p-value by 2. Calculating these p-values can be done via simulation, using a large number of draws of \( \{Z_1, \ldots, Z_k\} \) and \( \{U_1, \ldots, U_k\} \).

To create a confidence interval of level \( 1 - \alpha \), take a large number of draws of \( \{Z_1, \ldots, Z_k\} \) and \( \{U_1, \ldots, U_k\} \) and simply treat the \( \frac{\alpha}{2} \)th and \( 1 - \frac{\alpha}{2} \)th percentiles of the right-hand side of Equation 3.21 as the upper and lower limits of the interval.

Simulation results indicate that this interval is noncompetitively wide as compared
to the intervals in the last two columns of Table 2.3.

3.4.2 Monte Carlo approximate intervals for $CV_2 - CV_1$

Suppose we seek to create a confidence interval by pivoting around $\hat{CV}_1 - \hat{CV}_2$. The Miller-Feltz interval utilizes the normal approximation, inserting $\hat{CV}_1$ and $\hat{CV}_2$ for the nuisance parameters $CV_1$ and $CV_2$. Here, I shall also utilize the sample $CV$s as estimates of the true $CV$s, but I shall use the exact distribution rather than the normal approximation as the basis for the calculations.

We would like to find $\Delta_l$ and $\Delta_u$ such that

$$Pr(\Delta_l < \hat{CV}_1 - \hat{CV}_2 - (CV_1 - CV_2) < \Delta_u) = 1 - \alpha.$$ 

Plugging in Equation 2.2, this equation solves

$$Pr(\Delta_l < CV_1 \left(\frac{1 + CV_1 \frac{Z_1}{\sqrt{N_1}}}{1 + CV_1^2 \frac{Z_1^2}{N_1}} - 1\right) - CV_2 \left(\frac{1 + CV_2 \frac{Z_2}{\sqrt{N_2}}}{1 + CV_2^2 \frac{Z_2^2}{N_2}} - 1\right) < \Delta_u) = 1 - \alpha.$$ 

Now substituting in $\hat{CV}_1$ and $\hat{CV}_2$, $\Delta_l$ and $\Delta_u$ are approximately equal to the $\frac{\alpha}{2}$th and $1 - \frac{\alpha}{2}$th percentiles of

$$\hat{CV}_1 \left(\frac{1 + CV_1 \frac{Z_1}{\sqrt{N_1}}}{1 + CV_1^2 \frac{Z_1^2}{N_1}} - 1\right) - \hat{CV}_2 \left(\frac{1 + CV_2 \frac{Z_2}{\sqrt{N_2}}}{1 + CV_2^2 \frac{Z_2^2}{N_2}} - 1\right),$$

which we can find via simulation, taking a large number of draws of $U_1$, $Z_1$, $U_2$, and $Z_2$.

Then the confidence interval for $CV_2 - CV_1$ is

$$lower bound = \hat{CV}_2 - \hat{CV}_1 + \Delta_l,$$

$$upper bound = \hat{CV}_2 - \hat{CV}_1 + \Delta_u.$$ 

Simulations indicate that this interval performs about as well in terms of width
and coverage as the Miller-Feltz interval. But since this interval requires simulation to compute, the Miller-Feltz interval is more convenient.

One might also consider fiducial intervals as a strategy for assessing differences between CVs. The strategy here is to obtain fiducial distributions (Definition 1.3.3.5) for $CV_1 - CV_2$ or $\frac{CV_2}{CV_1}$ by taking fiducial draws of $CV_1$ and $CV_2$ using Equation 2.11. The confidence interval bounds for either quantity would be the $\frac{\alpha}{2}$th and $1 - \frac{\alpha}{2}$th percentiles of these distributions.

### 3.5 Convenient Inference on a Common CV Using the $\chi^2$ Approximation

Section 2.3.4 presented Monte Carlo approaches to inference on a common CV with unequal sample sizes based on the $\chi^2$ approximations in Section 2.2.4. The advantage of the Monte Carlo approach is that the size of a test and the coverage probability for an interval can be made as close as desired to nominal by choosing the number of Monte Carlo draws large enough. Here I present approaches that are more convenient but less true to Section 2.2.4.

One option with differing $N_i$s would be to ignore the $\frac{N_i - 2}{N_i}$ in Equation 2.18, which would imply that $\frac{1 + CV_2^2}{CV_2} \sum_{i=1}^{k} (N_i - 1)M_i \approx \chi^2_{\sum_{i=1}^{k}(N_i - 1)}$. This gives us a convenient pivot for conducting tests and creating confidence intervals using $\chi^2$ tables. Here I have essentially used McKay’s approximation to get an approximate distribution of $\sum_{i=1}^{k}(N_i - 1)M_i$.

Alternatively, adopting a Satterthwaite-type approach, we could assume $\sum_{i=1}^{k}(N_i - 1)M_i$ is distributed as a constant $A$ times a $\chi^2_B$ random variable, where $A$ and $B$ are
calibrated to the expected value and variance of \( \sum_{i=1}^{k} (N_i - 1)M_i \). This would give us

\[
A = \frac{\sum_{i=1}^{k} \left( \frac{CV_i^2}{1 + \frac{N_i - 2}{N_i} CV_i^2} \right)^2 (N_i - 1)}{\sum_{i=1}^{k} \frac{CV_i^2}{1 + \frac{N_i - 2}{N_i} CV_i^2} (N_i - 1)} \tag{3.22}
\]

\[
B = \frac{\left( \sum_{i=1}^{k} \frac{CV_i^2}{1 + \frac{N_i - 2}{N_i} CV_i^2} (N_i - 1) \right)^2}{\sum_{i=1}^{k} \left( \frac{CV_i^2}{1 + \frac{N_i - 2}{N_i} CV_i^2} \right)^2 (N_i - 1)} \tag{3.23}
\]

We would round \( B \) to the nearest integer. We could obtain \( p \)-values for tests using the \( \chi^2 \) tables, and we could obtain confidence limits by solving nonlinear equations for \( CV_o \).

Limited simulations indicate that these two methods give outputs that are very close to the Monte Carlo \( p \)-values and confidence limits.

Finally, from Equation 2.8 we get the following approximate pivot

\[
\sum_{i=1}^{k} \left( \frac{1 + CV_o^2}{CV_i^2} - \frac{2}{N_i} \right) (N_i - 1)M_i \approx \chi^2_{\sum_{i=1}^{k}(N_i - 1)}.
\]

The \( p \)-value for testing \( H_o : CV > CV_o \) would be

\[
\Phi_{\sum_{i=1}^{k}(N_i - 1)} \left( \sum_{i=1}^{k} \left( \frac{1 + CV_o^2}{CV_i^2} - \frac{2}{N_i} \right) (N_i - 1)m_i \right).
\]

The lower and upper confidence limits for \( CV \) would respectively solve the nonlinear equations

\[
\sum_{i=1}^{k} \left( \frac{1 + CV^2}{CV_i^2} - \frac{2}{N_i} \right) (N_i - 1)m_i = \chi^2_{\sum_{i=1}^{k}(N_i - 1), 1 - \alpha}.
\]

\[
\sum_{i=1}^{k} \left( \frac{1 + CV^2}{CV_i^2} - \frac{2}{N_i} \right) (N_i - 1)m_i = \chi^2_{\sum_{i=1}^{k}(N_i - 1), \alpha}.
\]

Although such inference is quite convenient, it is not based on the \( UMP \) test in Pro-
postion 6. Simulations would be needed to determine if these procedures sacrificed power as compared to Section 2.3.4.
Consider models of the form

\[ f_X(x) = C(\theta)h(x)K(T_{np}(x); \theta_{np})G(x; \theta_{pi}). \quad (4.1) \]

Suppose one wants to test \( H_0 : \theta_{pi} = \theta_o \). By the Factorization Theorem (1.3.1.4), \( T_{np} \) is sufficient for \( \theta_{np} \). Then by the definition of a sufficient statistic (1.3.1.1), the distribution of \( X \) conditional on \( T_{np} = t_{np} \) is invariant to \( \theta_{np} \). So we can conduct a similar test using a test statistic \( T_{pi} \) (Definition 1.3.3.1) by calculating the \( p \)-value associated with \( t_{pi} \) conditional on \( T_{np} = t_{np} \). I shall refer to this \( p \)-value as \( pv_{T_{pi}|T_{np}}(t_{pi}|t_{np}) \). If \( f \) belongs to an exponential family, conditioning on \( T_{np} \) will lead to uniformly most powerful similar inference, by Theorem 1.3.4.6.

If we have a goodness-of-fit statistic for testing the model in Equation 4.1, and \( G = 1 \) (i.e., we have a nontrivial sufficient statistic for the entire vector of unknown parameters), then conditioning on \( T_{np} \) will also allow us to conduct a pure (similar) goodness-of-fit test.

But calculating \( p \)-values from the conditional distribution is problematic. Using Theorem 1.3.5.14, we can derive the conditional distribution of \( X \) implied by Equation 4.1:

\[ f_{X|T_{np}}(x|t_{np}; \theta) = \frac{I_{T_{np}=t_{np}}h(x)G(x; \theta_{pi})}{\int_{T_{np}(y)=t_{np}} h(y)G(y; \theta_{pi})dy}. \quad (4.2) \]

By Theorem 1.3.3.4, calculation of a conditional \( p \)-value requires evaluating an expectation using this density. If Equation 4.1 is well-parameterized, the indicator function in Equation 4.2 will define an \((N - d_{np})\)-dimensional surface in \( \mathbb{R}^N \), which I shall call the “support surface.” So calculating the conditional \( p \)-value would require integration over this surface, which may be nonlinear. Such integration is not straightforward and
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

is considered an “extremely difficult” [95] problem for conditional inference in general. The support of the data may be further restricted by $h$ and $G$, making the problem even more complex.

This chapter will explore Monte Carlo approaches to calculating conditional $p$-values. (see Definition 1.3.3.7). Just like numerical integration, Monte Carlo calculations are nontrivial because of the complicated support. Since the conditional support has measure 0 under the unconditional distribution of the data, we cannot simply employ rejection sampling (Definition 1.3.5.15). While data generation on $\mathbb{R}^p$, a box in $\mathbb{R}^p$, or a $p$-dimensional sphere (see Theorem 1.3.5.2) is manageable, and a few algorithms have been published for generating data on more complicated geometric constructs [96], there is no omnibus approach for generating data on arbitrarily-constrained sets. WinBugs, a popular software for MCMC, limits the ability to constrain the support to very special cases.

A special case of Equation 4.1 is

$$f_X(x; \theta) = C(\theta)h(x)K\left(\theta_1 \sum_{i=1}^N W_{i1}x_i, \ldots, \theta_{dnp} \sum_{i=1}^N W_{idnp}x_i\right)G(x; \theta_{pi}), \quad (4.3)$$

where $W$ would be full rank in a well-parameterized model. Such models encompass generalized linear models (Definition 1.3.5.7) with canonical link functions (Definition 1.3.5.8). Here $T_{np}$ is the linear function $W^TX$ (Apply Theorem 1.3.1.4). The conditional distribution of $X$ given $T_{np} = t_{np}$ has the support $W^TX = t_{np}$, which is an $(N - d_{np})$-dimensional hyperplane in $\mathbb{R}^N$ if the density is well-parameterized. I shall call this the “support hyperplane.” A kernel (Definition 1.3.5.11) for Equation 4.3 is

$$I_{W^TX = t_{np}}h(x)G(x; \theta_{pi}). \quad (4.4)$$

(Apply the proof of Theorem 1.3.4.8). $h$ and $G$ may further restrict the support to only part of the hyperplane.

This chapter will focus on models of the form of Equation 4.3. Monte Carlo calculation of conditional $p$-values will turn out to be tractable in these cases. Section 4.6 will discuss Monte Carlo conditional $p$-value calculation in the more general model of
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

Equation 4.1.

4.1 Reduction to Data Generation on $\mathbb{R}^{N-d_{np}}$

Equation 4.2 and Equation 4.4 describe the conditional density of the entire data vector. But if the density is well-parameterized, it will be sufficient to generate from the conditional density of the first $N-d_{np}$ elements of the data vector (the meaningful variables), since the value of the sufficient statistic determines the rest (the residual variables). I shall call this the “marginal” conditional distribution.

Let $Y = g(X)$, where $Y_1 = X_1, \ldots, Y_{N-d_{np}} = X_{N-d_{np}}$ and $Y[N-d_{np}+1 : N] = T_{np}$. The density of $Y[1 : N-d_{np}]$ conditional on $Y[N-d_{np}+1 : N] = t_{np}$ will be the marginal conditional density we seek to derive.

If $g$ is one-to-one, from Equation 4.1 and Theorem 1.3.5.1, the density of $Y$ is

$$I_{g^{-1}(y) \in \mathbb{R}^N} C(\theta) h(g^{-1}(y)) K(y[N-d_{np}+1 : N], \theta_{np}) G(g^{-1}(y), \theta_{pi}) J_{g^{-1}(y)}.$$

The indicator function merely emphasizes that $g^{-1}$ must exist at the given point. Then using Theorem 1.3.5.14 we can deduce the kernel of the conditional distribution of $X[1 : N-d_{np}]$ given $T_{np} = t_{np}$ is

$$I_{g^{-1}(\{x[1:N-d_{np}]/t_{np}\}) \in \mathbb{R}^N} h(g^{-1}(\{x[1:N-d_{np}]/t_{np}\})) \times G(g^{-1}(\{x[1:N-d_{np}]/t_{np}\}); \theta_{pi}) J_{g^{-1}(\{x[1:N-d_{np}]/t_{np}\})}.$$  (4.5)

We can simplify computations by noting that the first $N-d_{np}$ elements of

$$g^{-1}(\{x[1:N-d_{np}]/t_{np}\})$$

will simply be $x[1:N-d_{np}]$.

If $g$ is not one-to-one,

$$g(x) = y$$  (4.6)

may have more than one solution for $x$. In this case we shall usually be able to divide the domain of $g$ up into $m$ pieces $B_1, \ldots, B_m$ on which it is one-to-one (see Casella 132
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

and Berger [2] page 185). Then we can write \( g(x) \equiv \sum_{i=1}^{m} I_{x \in B_{i}} g_{i}(x) \), where \( g_{i} \) is one-to-one on \( B_{i} \). The kernel of the marginal conditional is then

\[
\sum_{i=1}^{m} I_{g_{i}^{-1}(\{x[1:N-d_{np}], t_{np}\}) \in B_{i}} h(g_{i}^{-1}(\{x[1:N-d_{np}], t_{np}\})) \times G(g_{i}^{-1}(\{x[1:N-d_{np}], t_{np}\}); \theta_{pi}) J_{g_{i}^{-1}(\{x[1:N-d_{np}], t_{np}\})}.
\]

The support of the marginal conditional is a subset of \( \mathbb{R}^{N-d_{np}} \), not an \((N-d_{np})\)-dimensional surface in \( \mathbb{R}^{N} \). This can potentially simplify the problem, but difficulties remain. First, the \( N-d_{np} \) elements of the data vector are not independent, as one can see from Equation 4.5 via Theorem 1.3.5.17. Second, especially if the \( h \) and \( G \) functions embody constraints, the support may be an inconvenient subset of \( \mathbb{R}^{N-d_{np}} \); but since it will not have measure 0, rejection sampling from the unconditional support of \( X[1:N-d_{np}] \) might be a viable option. Third, Equation 4.5 cannot be written analytically if \( g^{-1} \) cannot be solved for analytically, and even then the analytical expression might be impractical to write down because it requires evaluating a Jacobian.

With linear sufficient statistics, at least the third issue will not be a problem. In that case, \( g \) will be the linear function \( AX \), where the first \( N-d_{np} \) rows of \( A \) will be the first \( N-d_{np} \) rows of the identity matrix, and the last \( d_{np} \) rows will be \( W^{T} \). If the density is well-parameterized, \( A \) will be invertible and \( g \) will be one-to-one, and \( g^{-1}(y) = A^{-1}y, J_{g^{-1}}(y) = |det(A^{-1})| \) by Theorem 1.3.5.1. Then from Equation 4.5 the kernel of the marginal conditional will be

\[
h(\{x[1:N-d_{np}], A^{-1}[N-d_{np} + 1:N]\{x[1:N-d_{np}], t_{np}\}\}) \times G(\{x[1:N-d_{np}], A^{-1}[N-d_{np} + 1:N]\{x[1:N-d_{np}], t_{np}\}; \theta_{pi}). \]

For an alternative way to reduce the problem if the sufficient statistic is linear, let \((WW^{T})^{g} \) be a generalized inverse of \( WW^{T} \) (Definition 1.3.8.3). By Theorem 1.3.8.2, this will always exist. Computation of generalized inverses in matrix programming languages such as SAS-IML is standard. By Theorem 1.3.8.3, \((WW^{T})^{g}W \) is a generalized inverse of \( W^{T} \). Then by Theorem 1.3.8.4, the support \( W^{T}X = t_{np} \) can be
written as the set of all points that satisfy

$$\mathbf{x} = (\mathbf{WW}^T)^g\mathbf{W}_t + (\mathbf{I} - (\mathbf{W}W^T)^g\mathbf{WW}^T)\mathbf{y}$$

(4.9)

for some $\mathbf{y}$ in $\mathbb{R}^N$.

Now let $\mathbf{P}$ be the matrix of eigenvectors (Definition 1.3.8.5) of the matrix $\left(\mathbf{I} - (\mathbf{WW}^T)^g\mathbf{WW}^T\right)$. By the way I have defined it, $\mathbf{P}$ is $N \times N - d_{np}$. Again, calculation of $\mathbf{P}$ in matrix programming languages is standard. By Theorem 1.3.8.11, we can rewrite Equation 4.9 as

$$\mathbf{x} = (\mathbf{W}W^T)^g\mathbf{W}_{t_{np}} + \mathbf{Pz}$$

(4.10)

for some $\mathbf{z}$ in $\mathbb{R}^{N-d_{np}}$. Furthermore, by Theorem 1.3.8.7, there is a unique $\mathbf{z}$ in $\mathbb{R}^{N-d_{np}}$ that satisfies Equation 4.10 for a given value of $\mathbf{x}$. Thus, Equation 4.10 defines a one-to-one and onto transformation from the support hyperplane of $\mathbf{X}$ to $\mathbb{R}^{N-d_{np}}$. We can generate $\mathbf{X}$ by first generating $\mathbf{Z}$ and then using Equation 4.10 to get $\mathbf{X}$.

We can use Theorem 1.3.5.1 to get the kernel of the density of $\mathbf{Z}$ implied by Equation 4.10 and Equation 4.4. The kernel of the density must have the form

$$h((\mathbf{WW}^T)^g\mathbf{W}_{t_{np}} + \mathbf{Pz})G((\mathbf{WW}^T)^g\mathbf{W}_{t_{np}} + \mathbf{Pz}, \theta_{pi})J(\mathbf{z}).$$

The indicator function from Equation 4.4 is dropped because it is guaranteed by the transformation. Because the transformation is between spaces of different dimension, we cannot use the simple formula in Definition 1.3.5.2 to get the Jacobian. But since the transformation is linear, $J(\mathbf{z})$ will be a constant, so that the kernel of the density of $\mathbf{Z}$ will be

$$h((\mathbf{WW}^T)^g\mathbf{W}_{t_{np}} + \mathbf{Pz})G((\mathbf{WW}^T)^g\mathbf{W}_{t_{np}} + \mathbf{Pz}, \theta_{pi}).$$

(4.11)

In fact, $J(\mathbf{z}) = 1$. This is due to the fact that $\mathbf{P}$ is an orthonormal set. By the fact that they are orthogonal to each other, the columns of $\mathbf{P}$ point in the direction of coordinate axes for the conditional support of $\mathbf{X}$; by the fact that the columns have length 1, unit movements along the axes of $\mathbb{R}^{N-d_{np}}$ induce unit movements along the axes of the range of Equation 4.10. In other words, consider the unit hyperbox defined by starting at $\mathbf{z}$ and moving one unit along each axis in $\mathbb{R}^{N-d_{np}}$. The image of this hyperbox in the
conditional support of $X$ resulting from the transformation in Equation 4.10 will also be a unit hyperbox.

To generate $X$ from its conditional density, we can generate $Z$ in $\mathbb{R}^{N-d_{np}}$ from a density whose kernel is Equation 4.11 and then obtain $X$ using Equation 4.10. As with the marginal conditional density, to generate directly from Equation 4.11 we would have to solve the problem of generating correlated multivariate data with possibly complicated support restrictions.

### 4.2 Assessment of the literature

There is a large literature on Monte Carlo methods for conditional inference, but it is focused largely on problems involving discrete data (see [97] [98], [99]). I shall draw on this literature and cite the relevant papers, but for the most part the methods are either nongeneralizable to the continuous case or involve complications that can be avoided with continuous data. There are two conditional Monte Carlo approaches of which I am aware that have their origins outside the discrete-data literature, one which has claimed the name “conditional Monte Carlo” and one which I shall call fiducial Monte Carlo. Both of these have limitations; they will be discussed further below.

Statisticians have been largely content with approximate methods of calculating conditional $p$-values for problems involving continuous data. Likelihood methods such as the likelihood ratio test are asymptotically similar. (See Theorem 1.3.6.13). Higher-order asymptotics (HOA), discussed in [100], were developed to improve the precision of likelihood-based methods. A popular HOA approach to inference involving a univariate $T_{pi}$ conditional on $T_{np}$ in an exponential family uses an approximation to the conditional $cdf$ of $T_{pi}$ (see [101], [102], [103], [104]). Other applications of HOA to conditional inference are discussed in Bellio and Brazzale [100] and Reid [105].

There are several reasons why one might prefer Monte Carlo methods to approximate methods:

1. Approximate methods do not provide exact $p$-values.

   $p$-values calculated via Monte Carlo methods typically converge almost surely (see Definition 1.3.3.7), which means that they can be made as close as desired
Chapter 4. Monte Carlo Conditional $p$-value Calculation for Continuous Data

to the true $p$-values.

Certain $HOA$ methods have been shown to be highly accurate in examples, and have high orders of convergence theoretically. Claims of the near-exactness of $HOA$ are often made in the literature (see [100]). However, examples have also been found where the accuracy of at least some popular $HOA$ methods is inadequate [106].

2. Diagnostics are not available for how close the approximate $p$-values are, while standard error formulas are available to let the investigator decide where to stop Monte Carlo algorithms.

3. Goodness-of-fit testing with $HOA$ methods is not straightforward.

   But the Monte Carlo calculation of a $p$-value of a goodness-of-fit statistic conditional on a sufficient statistic for $\theta$ is no different in principle from the Monte Carlo calculation $pv_{T_p|T,n}(t_p|t_n)$.

4. Monte Carlo methods are more broadly applicable

   Equation 4.1 is more general than the model assumed by some higher-order asymptotic methods.

5. $HOA$ are complicated to use.

   Quoting from [100],

   ...the derivation and evaluation of asymptotic expressions is typically a direct but laborious task as it presents the combination of higher-order expansions and ... multivariate situations which make it difficult to perform the calculations by hand. Due to the forbidding mathematics involved ... higher-order asymptotics are still under-used in practical work.

   Lengthy calculations require time, skill, and the risk of making a mistake. And if expressions necessary for $HOA$ such as the profiled log-likelihood or $MLE$s

136
cannot be obtained analytically, the investigator is required to calculate them numerically, which can be a nontrivial programming exercise.

The Monte Carlo methods in this chapter will typically require only minor hand calculations. The programming can be nontrivial, but in the applications below can be done in a surprisingly small number of lines. One might think that Monte Carlo methods would require more computer time, but this may not be true if the HOA expressions are not available analytically. Unfortunately, the more accurate HOA methods are usually more difficult or computationally intensive.

Comparing HOA to Monte Carlo methods for conditional logistic regression, which is a generalized linear model for discrete data with a canonical link, Corcoran et. al. [107] preferred Monte Carlo methods; they found that a highly accurate approximate method was more computationally intensive than Monte Carlo while a less accurate approximate method was unreliable. The lack of accessibility of HOA methods for dealing with hypotheses about a multidimensional $\theta_{pi}$ has led to attempts to build Monte Carlo algorithms ([108], [109]).

### 4.3 Special case: gamma distribution, known shape parameter – Dirichlet data generation

Suppose $X_1, \ldots, X_N$ are independent $\text{gamma}(\alpha_i, \beta)$ with the $\alpha_i$s known (Definition 1.3.5.10). To generate $X$ from its distribution conditional on $T_\beta = t_\beta$, the following algorithm will suffice:

**Algorithm 4. Generating gamma variates conditional on their sum**

1. Generate $Y_1, \ldots, Y_N$, independent $\text{gamma}(\alpha_i, 1)$ variables.
2. Obtain $X_i = \frac{Y_i}{\sum_{i=1}^N Y_i} \cdot t_\beta$.

**Proof.** To see that this is the same as drawing from the distribution of $X$ conditional

137
on $T_{\beta} = t_{\beta}$, from Equation 4.2 and Definition 1.3.5.10, the kernel of this distribution is

$$I_{\sum_{i=1}^{N} x_i = t_{\beta}} \prod_{i=1}^{N} x_i^{\alpha_i - 1}.$$ 

Then from Theorem 1.3.5.1 we get that the kernel of the conditional distribution of

$$\{Z_1, \ldots, Z_N\} \equiv \left\{ \frac{X_1}{t_{\beta}}, \ldots, \frac{X_N}{t_{\beta}} \right\}$$

is

$$I_{\sum_{i=1}^{N} z_i = 1} \prod_{i=1}^{N} z_i^{\alpha_i - 1},$$

which is Dirichlet (Definition 1.3.5.4). So we can generate $X$ by generating a Dirichlet $Z$ and then multiplying $Z$ by $t_{\beta}$. This is exactly what Algorithm 4 is doing, by Theorem 1.3.5.4.

By using algorithm 4, we can estimate a $p$-value for any statistic conditional on $T_{\beta}$ via Monte Carlo. This is useful when $\beta$ is a nuisance parameter and we want to conduct a similar (Definition 1.3.5.1) test.

Although Algorithm 4 is straightforward, for certain problems that involve testing the equality of gamma distributions with $\alpha$ known, conditional Monte Carlo has been ignored as an option for calculating $p$-values. One such problem is testing for the marginal effect of an experimental variable when there are nuisance factors.

Let $i$ index a cell defined by the levels of factors whose effects are not of immediate interest. This could represent the $i$th level of a single factor, or it could represent a factor-level combination. Let there be $I$ such cells. Let $j$ index the levels of the factor of interest, with $k$ such levels, and let $r$ index the replicates for the $ij$th combination, with $R_{ij}$ in total. Suppose that $\alpha$ is known and we wish to test the null hypothesis that the factor of interest has no effect on the scale parameter. The density of the data under the null is a constant times

$$\prod_{i=1}^{I} \prod_{j=1}^{k} \prod_{r=1}^{R_{ij}} x_{ijr}^{\alpha_i - 1} \exp \left( - \sum_{i=1}^{I} \sum_{j=1}^{k} \sum_{r=1}^{R_{ij}} \frac{1}{\beta_i} x_{ijr} \right).$$

(4.12)
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

In other words, under the null the data from the $ij$th cell are drawn from a $\text{gamma}(\alpha, \beta_i)$ distribution. Under a general alternative, $\beta_i$ would be replaced by $\beta_{ij}$.

Suppose we want to compute a similar $p$-value for a test statistic such as the likelihood ratio statistic. We can do this by Monte Carlo, using Algorithm 4 to generate data within each cell conditional on the value of the sufficient statistics

$$T_{\beta_i} = \sum_{j=1}^{k} \sum_{r=1}^{R_{ij}} X_{ijr}.\$$

Note that nothing requires the factor of interest to be discrete; ie, Monte Carlo $p$-values will be valid if $R_{ij} = 1$ (see Definition 1.3.3.7).

Since the exponential distribution is $\text{gamma}$ with $\alpha = 1$, one can apply this approach to testing for effects in survival experiments where the underlying data is assumed exponential. Exponential models are frequently adopted for life-testing experiments in industry ([110], [111], [112]) and occasionally for survival experiments in medicine as well. With only one parameter, a constant hazard rate, and the memoryless property (Definition 1.3.5.14), the exponential model is the simplest and most convenient survival model. The memoryless assumption is often a good approximation for actual data.

Very often, survival data are censored because it is expensive to carry out the experiment until the last failure. But even though the censored data do not have an exact exponential distribution, Algorithm 4 can still be used to conduct similar tests for factor effects if the censoring is of Type II – observation of the $ij$th factor level combination is terminated after the first $f_{ij}$ failures are observed.

For exponential data subject to Type II censoring, Equation 4.12 becomes

$$\exp \left( - \sum_{i=1}^{l} \sum_{j=1}^{k} \frac{1}{\beta_i} T_{ij} \right),$$

where $T_{ij}$ is the sum of the first $f_{ij}$ failure times plus $R_{ij} - f_{ij}$ times the $f_{ij}$th failure time ([111], page 101). Under the general alternative, $\beta_i$ would be replaced by $\beta_{ij}$.

The likelihood is a function of the data only through the $T_{ij}$s. Furthermore, the $T_{ij}$s are independent $\text{gamma}(f_{ij}, \beta_i)$ random variables under the null ([111], page 103). Thus, we can calculate a similar $p$-value for any likelihood-based statistic for testing a factor effect by using Algorithm 4 to generate the $T_{ij}$s conditional on the sufficient
statistics $\sum_{j=1}^{k} T_{ij}$ for the $\beta$s.

Exact tests of a factor effect for a one-way exponential experiment are well known [113], but the ability to obtain an exact (up to Monte Carlo error) test of a marginal effect with nuisance factors appears to have been overlooked. The standard practice is to use Theorem 1.3.6.2 or Theorem 1.3.6.13 to conduct likelihood-based inference ([111], page 285). Lawless [114] creates a way to do an exact test of the marginal effect of a single continuous factor when there are no covariates or nuisance factors; he opines that the extension to more general cases is “not feasible” ([111], page 291).

If one wants to determine the effects of factors in an experiment with replication on the variability of the response — “dispersion effects” — one can use the sample variance calculated from the $N_{ij}$ observations for the $ij$th factor combination as the response variable for that combination. This would create a new dataset where there is one observation per factor combination. If the underlying data is normal, then

\[(N_{ij} - 1)S_{ij}^2 \sim \text{gamma} \left( \frac{N_{ij} - 1}{2}, \frac{2\sigma_{ij}^2}{\sigma_{ij}^2} \right), \tag{4.13} \]

where $\sigma_{ij}^2$ is the population variance for the $ij$th factor combination. The likelihood of the variability response would have the form of Equation 4.12 with $R_{ij} = 1$. So we can use Algorithm 4 to ascertain the significance of the dispersion effect of a factor of interest in the presence of nuisance factors.

Bartlett’s test [115], is commonly used for testing the equality of variances of normal populations. This problem is identical in our setup to testing for dispersion effects due to a categorical variable within a single cell — $I = 1$. Bartlett’s test compares

$$\sum_{j=1}^{k} (N_j - 1) \ln \left( \frac{\sum_{l=1}^{k}(N_l - 1)S_{lj}^2}{(N_j - 1)S_{j}^2} \right)$$

to a scaled $\chi^2_{k-1}$ random variable. This approximation is often very good.

However, we can easily obtain a similar $p$-value for Bartlett’s statistic by using Algorithm 4 to generate the $(N_j - 1)S_{j}^2$ variables conditional on their sum. Now since Bartlett’s statistic is a function of a maximal invariant to positive scale transformations, its distribution is invariant to $\sigma^2$ (Definitions 1.3.2.3 and 1.3.2.4), so it is ancillary
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

for \( \sigma^2 \) (Definition 1.3.1.5). Then by Basu’s Theorem (1.3.1.3), Bartlett’s statistic is independent of \( \sum_{j=1}^{k} (N_j - 1)S_j^2 \), so drawing it from its conditional distribution is the same as drawing it from its unconditional distribution. Thus, the Monte Carlo p-value calculated with Algorithm 4 is actually an unconditional p-value.

Three authors ([116], [117], [118]) have published papers deriving the exact distribution of Bartlett’s statistic and computing critical values; the Monte Carlo approach here is simple and effective.

Testing dispersion effects in normal replicated experiments using sample variances has long been of interest ([119], [120], [121], [122], [123]). Even in experiments without replication, such as fractional factorial quality control studies popular in industrial statistics, the issue often becomes relevant when insignificant factors are dropped from the model. Davidian and Carroll [124] point out that using sample variances rather than individual observations to test for dispersion effects entails a loss of efficiency. The advantage, however, is that it makes the test invariant to location effects in the experiment; to use individual observations requires using a model that is potentially misspecified or inaccurately estimated [125].

Exact methods are available for a few special cases of testing dispersion effects. For a one-way setup, I have mentioned Bartlett’s test. Interestingly, 4 other tests for the one-way setup are commonly cited (see [123]). For all 4, researchers have derived exact distributions and tabulated critical values; but all 4 could easily be handled by the Monte Carlo strategy here. Wludyka and Nelson [123] suggest another one-way test; they recognize the connection to the Dirichlet distribution, but use a more complicated method to tabulate critical values. In the same paper, they suggest a way to handle the test of a main dispersion effect while eliminating the effect of a second factor, but it requires the assumption that the dispersion effects are additive. For testing a dispersion effect in the presence of a single nuisance factor without the additivity assumption, two tests have been proposed that are exact: an F test [126] and a Monte Carlo method that generates the residuals of the experimental model from their null distribution [127].

The Monte Carlo conditional approach of Algorithm 4 allows exact testing (up to Monte Carlo error) for cases when there is more than one nuisance factor. Also,
since it generates from the null conditional distribution, it is very simple to use the approach here to calculate the $p$-value for ANY statistic that tests for differences, including specific differences. Most of the exact tests in the literature are tests for general differences. An example where a test is needed for specific differences can be found in [128]. There the factor of interest is quantitative; if the levels of a factor can be given quantitative values, we might want to test ordered hypotheses or to assume a functional form for its relationship to dispersion.

Bartlett and Kendall’s [119] approach to testing for dispersion effects in replicated factorial experiments with more than one nuisance factor is the longest standing and possibly the most frequently-used in practice for normal data. They assume the logarithm of the sample variance for the $ij$th variable combination is normal with variance $\frac{\sigma^2_i}{N_{ij} - 1}$ and regress this on the original design matrix. They use the standard $F$ statistic for testing the significance of a factor, modified for the fact that the variance of each observation is known.

Table 4.1 displays the estimated size of this dispersion effect test. In the simulations the degrees of freedom $N_{ij} - 1$ were kept the same for all sample variances. The design matrix contained $I$ columns, each of which contained a 1 for observations in the corresponding cell and a 0 otherwise, plus an additional column for the factor of interest. It was assumed that $\sigma^2_i$ was constant. The factor of interest was assumed to have two levels. In the cases shown in the table, the estimated power to detect a doubling of the standard deviation across levels of the factor of interest ranged from around 50% to 1. The estimated sizes were the about the same if the factor of interest

<table>
<thead>
<tr>
<th>$N_{ij}$</th>
<th>1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.210</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.127</td>
<td>0.125</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>0.087</td>
<td>0.081</td>
<td>0.084</td>
<td>0.084</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.066</td>
<td>0.065</td>
<td>0.064</td>
<td>0.066</td>
<td>0.061</td>
<td></td>
</tr>
</tbody>
</table>
was modeled as quantitative with a linear dispersion effect.

The table indicates that the size has considerable room for correction via Monte Carlo calculation of the $p$-value. The validity of the Bartlett-Kendall test improves as the individual sample sizes increase and the normal approximation of the log sample variance becomes more accurate.

Another application of Algorithm 4 is testing the distributional assumption. As long as the data are gamma with known shape parameters drawn from $I$ different populations, Algorithm 4 allows us to calculate a Monte Carlo $p$-value for any goodness-of-fit test statistic conditional on the value of the $T_{\beta_i}$s. The resulting goodness-of-fit test will be pure in the sense that it will be invariant to the unknown parameters.

A variety of exact one-sample goodness-of-fit tests for the exponential density have been explored ([111], page 444), but it is not well appreciated that pure goodness-of-fit testing for exponential regression models in factorial experiments is straightforward. The existence of an exact pure goodness of fit test makes it almost obligatory try out the exponential model in situations where it might apply, since frequentist statistics proceeds by the progressive testing of simplifying assumptions.

For a goodness-of-fit test of Equation 4.12, we need $R_{ij} > 1$. In the case of a test of the $\chi^2$ assumption for sample variances, one must accept the null of no dispersion effect of at least one factor in order to obtain the necessary replication.

While time is occasionally viewed as a drawback in Monte Carlo-based inference, Algorithm 4 is very fast; with $s = 1,000,000$ and $N = 100$, it finishes in two minutes in SAS-IML.

### 4.4 Importance Sampling

The idea behind importance sampling is that as long as the data has the correct support, if we simply reweight it to reflect the correct density, it will tell us what we want to know. More specifically, consider the following algorithm:

**Algorithm 5. Importance Sampling**

1. Take $s$ independent draws $y_1, \ldots, y_s$ of a random vector $Y$ of the same dimension as $X$ and whose support is the support surface of $X$. 

143
2. Let $g(y_i) = 1$ if $T_{pi}(y_i) \leftrightarrow t_{pi}$, $g(y_i) = 0$ otherwise (see Definition 1.3.3.6).

3. Let $\hat{pv}_{IS} = \frac{\sum_{i=1}^{s} g(y_i) f_{X|T_{np}}(y_i|t_{np})}{\sum_{i=1}^{s} f_{X|T_{np}}(y_i|t_{np})}$, where $k_{X|T_{np}}$ is a kernel (Definition 1.3.5.11) of $f_{X|T_{np}}(x|t_{np})$ and $k_Y$ is a kernel of $f_Y(y)$.

From Definition 1.3.5.11, $f_{X|T_{np}}(y|t_{np}) = C_n k_{X|T_{np}}(y|t_{np})$, and $f_Y(y) = C_d k_Y(y)$ for some constants $C_n$ and $C_d$. Multiplying the numerator and denominator of $\hat{pv}_{IS}$ by $rac{C_n}{C_d}$, it is clear that $\hat{pv}_{IS} = \frac{\sum_{i=1}^{s} g(y_i) f_{X|T_{np}}(y_i|t_{np})}{\sum_{i=1}^{s} f_{X|T_{np}}(y_i|t_{np})}$, so that by Definition 1.3.9.1, $\hat{pv}_{IS}$ is an importance sampling estimator of $E_{X|T_{np}}(g(X)|t_{np})$, with $f_Y$ as the generating distribution and $f_{X|T_{np}}$ as the target distribution. Furthermore, $g(y)$ is simply $I_{T_{pi}(y) \leftrightarrow t_{pi}}$, so by Theorem 1.3.9.2, $\hat{pv}_{IS}$ is estimating the conditional $p$-value for $t_{pi}$. For the moment, assume $\frac{f_{X|T_{np}}(y|t_{np})}{f_Y(y)}$ is bounded; then by Theorem 1.3.9.1 $\hat{pv}_{IS}$ converges almost surely with $s$ to $\hat{pv}_{T_{pi}|T_{np}}(t_{pi}|t_{np})$, and Theorem 1.3.9.2 gives us a variance for $\hat{pv}_{IS}$.

In order to use Theorem 1.3.9.2, we need to estimate the quantities in the expression for $V$. First, we need to estimate $\frac{C_n}{C_d}$, because this does not drop out of the expression for the variance as it does in the expression for the point estimate itself. To do so, realize that

$$\int_{X|t_{np}} \frac{C_n k_{X|T_{np}}(y|t_{np})}{C_d k_Y(y)} f_Y(y) dy = \int_{X|t_{np}} f_{X|T_{np}}(y|t_{np}) dy = 1,$$

where $X|t_{np}$ is the support of $f_{X|T_{np}}(x|t_{np})$. So $\frac{C_n}{C_d} = \frac{1}{E_Y \left( \frac{k_{X|T_{np}}(Y|t_{np})}{k_Y(Y)} \right)}$. Now assuming that $\frac{k_{X|T_{np}}(Y|t_{np})}{k_Y(Y)}$ is bounded, by Theorem 1.3.5.5, all of its moments exist. Also, the reciprocal function is continuous. So by the Strong Law of Large Numbers (Theorem 1.3.6.8) and Theorem 1.3.6.5, part 3,

$$\hat{c} = \frac{s}{\sum_{i=1}^{s} \frac{k_{X|T_{np}}(y_i|t_{np})}{k_Y(Y_i)}} \rightarrow_{as} \frac{C_n}{C_d}.$$

Now we can estimate $V$ with its sample counterpart, giving us an expression for
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

the standard error of \( \hat{p}_v \):

\[
\frac{1}{\sqrt{s}} \left[ \frac{1}{s} \sum_{i=1}^{s} (1 - 2\hat{p}_v) \left( I_{T_{\mu} < t_{\mu}(y_i)} - \frac{\hat{c}_2 k_{X|T_{np}}^2(y_i|t_{np})}{k_Y(y_i)} \right) \right] + \hat{p}_v \frac{\hat{c}_2 k_{X|T_{np}}^2(y_i|t_{np})}{k_Y(y_i)}.
\]  \[ (4.14) \]

By the Strong Law of Large Numbers and Theorem 1.3.6.5, part 3, the term under the radical sign converges almost surely to \( V \). (The SLLN applies because all the random variables are bounded, and thus have all of their moments.)

To justify the use of importance sampling and the standard error estimate, we have relied on the assumption that \( \frac{k_{X|T_{np}}(Y_i|t_{np})}{k_Y(Y_i)} \) is bounded. This is stronger than we need; the existence of \( E_Y \left( \left( \frac{k_{X|T_{np}}(y|t_{np})}{k_Y(y)} \right)^2 \right) \) is sufficient, and even this is not necessary. Proving Theorems 1.3.9.1, 1.3.9.2, and the asymptotic validity of Equation 4.14 without boundedness can be quite tricky, while cases where they do not hold are the exception, so I shall not be overly concerned about justifying them.

The validity of Theorem 1.3.9.1 says nothing about the finite-\( s \) properties of Algorithm 5. A practitioner can use Equation 4.14 to let him know when he has achieved enough precision to stop Algorithm 5. But nothing guarantees that the \( s \) that provides acceptable precision will be practically small.

In this chapter, we care not just about implementing importance sampling once, but also about assessing its performance at a prespecified \( s \) using simulations. To this end we shall use two diagnostics.

1. Type I error of importance sampling test. The “importance sampling test” is analogous to the Monte Carlo test (Definition 1.3.3.7). The properties of this test can be evaluated in simulations just like any other test. For infinite \( s \), the size will be equal to the nominal size, but for finite \( s \) may be larger or smaller.

The Monte Carlo test (Definition 1.3.3.7) has size approximately equal to

\[
\int_0^1 \Phi \left( \frac{\sqrt{s}(\alpha - p)}{\sqrt{p(1-p)}} \right) dp,
\]

where \( \alpha \) is the nominal size and \( \Phi \) is the standard normal cdf. This converges
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

with $s$ to the nominal size, but it turns out that the Monte Carlo test is liberal with small $s$. However, for any reasonable $s$ the deviation from nominal size is tiny. For $s = 500$, $\alpha = 0.05$, the size of the Monte Carlo test is approximately 0.0509, and for $s = 1000$, the size is approximately 0.0504.

A deviation of the size of the importance sampling test from 0.05 will be partly due to the liberality inherent in Monte Carlo tests but mainly due to bias in $\hat{pv}_{IS}$. If the size of the importance sampling test is close to 0.05, we can take it as a sign that $s$ is large enough that the bias has been eliminated.

2. Average standard error (ASE) of $\hat{pv}_{IS}$ under the null. We can evaluate the average standard error from Equation 4.14 across simulated datasets. For comparison purposes, under the null hypothesis, the expected value of the standard deviation of the Monte Carlo $p$-value is $\frac{1}{s} \frac{\pi}{8}$ (Theorem 1.3.5.13).

As an aside, an alternative way to compare the precision of any $p$-value estimator $\hat{p}$ with that of a Monte Carlo $p$-value would be to simulate $s$ datasets under the null, calculate $\hat{p}$ for each, and then compare the sample variance of these values with the theoretical value of the variance of the Monte Carlo $p$-value across simulated datasets. From Definition 1.3.5.1, Theorem 1.3.5.18 and Theorem 1.3.5.19, we can derive that this variance is $\frac{s+2}{12s}$.

To implement importance sampling in practice requires one to obtain the kernel of the conditional density and to find a suitable $Y$. For the former problem we can use Equation 4.2, Equation 4.4, Equation 4.5, Equation 4.7, Equation 4.8, or Equation 4.11. For the latter problem, with linear sufficient statistics one of several approaches can be tried. As we shall see, the performance of $\hat{pv}_{IS}$ depends on the match between the generating density and the target, so one would choose the approach that provides the best match.

4.4.1 Importance sampling with linear sufficient statistics: generating data on a hyperplane

Equation 4.4 gives us the kernel of the full conditional distribution for the case of linear sufficient statistics. For a generating density to be suitable, it’s support needs to be
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

the hyperplane $\mathbf{W}^T\mathbf{X} = t_{np}$. Here I suggest three practical-to-generate distributions whose support is a hyperplane with dimension lower than the size of the data vector.

**Dirichlet distribution**

If the dimension of $T_{np}$ is 1, we have the option of using the Dirichlet distribution as our generating distribution. If an $N$-dimensional $\mathbf{Z} \sim \text{Dirichlet}(\gamma)$, then a $\mathbf{Y}$ with the correct support can be obtained via the transformation $Y_1 = \frac{t_{np}Z_1}{W_{11}}, \ldots, Y_N = \frac{t_{np}Z_N}{W_{N1}}$.

Using Theorem 1.3.5.1 and Definition 1.3.5.4, $k_Y(y) = I_{W^T y = t_{np}} \prod_{i=1}^N \left( \frac{W_{ii}y_i}{t_{np}} \right)^{\gamma_i - 1}$.

**Conditional normal distribution**

The second candidate for a generating distribution (Definition 1.3.9.1) is that of an $N$-dimensional normal variable conditional on its belonging to the support hyperplane. This is the idea of Booth and Butler ([129]), who used it for conditional inference in a log-linear model of discrete data.

Consider the random variable $\mathbf{Y}^*$, a normal $N \times 1$ vector. Let $\mathbf{A} = \begin{pmatrix} \mathbf{I}_N \\ \mathbf{W}^T \end{pmatrix}$ where $\mathbf{I}_N$ is the $N \times N$ identity matrix. By Theorem 1.3.5.9,

$$\mathbf{A} \mathbf{Y}^* \sim N \left( \begin{pmatrix} E_{\mathbf{Y}^*}(\mathbf{Y}^*) \\ \mathbf{W}^T E_{\mathbf{Y}^*}(\mathbf{Y}^*) \end{pmatrix}, \begin{pmatrix} \text{Var}(\mathbf{Y}^*) & \text{Var}(\mathbf{Y}^*) \mathbf{W} \\ \mathbf{W}^T \text{Var}(\mathbf{Y}^*) & \mathbf{W}^T \text{Var}(\mathbf{Y}^*) \mathbf{W} \end{pmatrix} \right)$$

And by Theorem 1.3.5.12, assuming that the relevant inverse exists, the distribution of $\mathbf{Y}^*$ conditional on $\mathbf{W}^T \mathbf{Y}^* = t_{np}$ is normal with mean

$$E_{\mathbf{Y}^*}(\mathbf{Y}^*) + \text{Var}(\mathbf{Y}^*) \mathbf{W} (\mathbf{W}^T \text{Var}(\mathbf{Y}^*) \mathbf{W})^{-1} (t_{np} - \mathbf{W}^T E_{\mathbf{Y}^*}(\mathbf{Y}^*)) \quad (4.15)$$

and variance

$$\text{Var}(\mathbf{Y}^*) - \text{Var}(\mathbf{Y}^*) \mathbf{W} (\mathbf{W}^T \text{Var}(\mathbf{Y}^*) \mathbf{W})^{-1} \mathbf{W}^T \text{Var}(\mathbf{Y}^*). \quad (4.16)$$

Let $\mathbf{Y}$ be the random variable whose distribution is the same as the conditional
distribution of $Y^*$. To actually draw from the distribution of $Y$, we can employ the following algorithm:

**Algorithm 6. Drawing from conditional distribution of a normal random variable**

1. Find $B$, a lower triangular matrix such that
   \[ BB^T = Var(Y^*) - Var(Y^*)W(W^TVar(Y^*)W)^{-1}W^TVar(Y^*). \]
   This can be done by Cholesky decomposition, a standard matrix-programming language algorithm.

2. Draw $Z$ from a $N(0, I_N)$ distribution using a standard built-in algorithm.

3. Let $y_i = E_{Y^*}(Y^*) + Var(Y^*)W(W^TVar(Y^*)W)^{-1}(t_{np} - W^TE_{Y^*}(Y^*)) + Bz_i$

$Y$ will have the correct distribution by Theorem 1.3.5.9.

We can choose $E(Y^*)$ and $Var(Y^*)$ to improve the agreement between the generating density and the target density; a reasonable choice is to set $E(Y^*)$ equal to an estimate of $E(X)$ and $Var(Y^*)$ equal to an estimate of $Var(X)$.

We can choose $Var(Y^*)$ so that the support of $Y$ will be the entire support hyperplane; this will be true if the $Y^*_i$s are independent. Now the function $f_X|_{t_{np}}(y|t_{np})/f_{Y}(y)$ will be continuous on that hyperplane. If $h$ and $G$ restrict the conditional support of $X$ to a compact (Definition 1.3.8.1) subset of the hyperplane, by Theorem 1.3.8.1 the function $f_X|_{t_{np}}(y|t_{np})/f_{Y}(y)$ will be bounded, and Theorems 1.3.9.1 and 1.3.9.2, which guarantee convergence of $\hat{pv}_{IS}$ and justify Equation 4.14, will hold. Furthermore, Equation 4.14 will be asymptotically valid. One important case where the conditional support of $X$ is compact is if the unconditional support of $X$ is nonnegative and $W$ contains a column of 1s (see Theorem 1.3.8.12).

If the conditional support of $X$ is a subset of the supporting hyperplane, we will need to throw out the draws of $Y$ that are not in the correct subset of the hyperplane, increasing the standard deviation of $\hat{pv}_{IS}$ by effectively reducing the number of draws. There’s no guarantee that the proportion of throwaway draws of $Y$ will be small enough to allow the algorithm to converge in a reasonable amount of time.
In Algorithm 5, throwing out will be done automatically by indicator functions in $k_{X|T, np}$ that will have the value 0 outside the compact subset, and the effect on the standard error will be captured by an increase $\hat{c}$. We see that the support requirement for Algorithm 5 is not that generating and target density have the same support, but that the support of the target density be contained in the support of the generating density (and have nonzero measure under the generating density).

To implement Algorithm 5 using Algorithm 6, we need the density of $Y$. This is ([9], page 41)

$$f_Y(y) = \frac{(2\pi)^{-\frac{N-d_{np}-1}{2}}}{\sqrt{\lambda_1 \times \ldots \times \lambda_{N-d_{np}}}} \exp \left( -\frac{1}{2}(y - E_Y(y))^T(Var(Y)^g)(y - E_Y(y)) \right),$$

(4.17)

where $(Var(Y))^g$ is any generalized inverse (Definition 1.3.8.3) of $Var(Y)$, and the vector of $\lambda$s are the eigenvalues (Definition 1.3.8.5) of $Var(Y)$. (Since all of the values drawn will be on the correct hyperplane, we don’t need an indicator function.)

**Estimated likelihood sampling**

The idea here is simple. Generate the first $N - d_{np}$ elements of $Y$ by the unconditional likelihood of the first $N - d_{np}$ elements of $X$, with $\theta = \{\hat{\theta}_{np, \text{MLE}}, \theta_o\}$. Let the last $d_{np}$ elements of $Y$ solve $W^T\{y[1 : N - d_{np}], y[N - d_{np} + 1 : N]\} = t_{np}$. If $A$ is as defined in Section 4.1, the last $d_{np}$ elements will solve $y[N - d_{np} + 1 : d_{np}] = A^{-1}[N - d_{np} + 1 : N]\{y[1 : N - d_{np}], t_{np}\}$. The density of $Y$ will simply be the estimated unconditional likelihood, ie $f_{X[1 : N-d_{np}]}(y[1 : N - d_{np}]; \{\hat{\theta}_{np, \text{MLE}}, \theta_o\})$.

If $h$ and $G$ restrict the conditional support of $X$ to a subset of the hyperplane, then we may have the same problem as with the use of a normal generating density, in that the conditional support of $X$ may be a subset of the support of $Y$. While this might allow us to prove Theorems 1.3.9.1 and 1.3.9.2 easily, it also forces us to throw out some draws of $Y$.

ELS turns out to be a specific implementation of the method known as “conditional Monte Carlo” referred to above.

Incidentally, we could use the marginal conditional distribution in Equation 4.8 as the target rather than the full conditional. In this case, we would not need to calculate
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

\[ y[N - d_{np} + 1] \] after drawing \( y[1 : N - d_{np}] \).

4.4.2 Implementing importance sampling for testing CV equality

Section 4.4.2 will speak as if Equation 2.24 were exact.

Section 2.4.3 derived the UMP similar invariant one-sided test of CV homogeneity for normal data with \( k = 2 \) and discussed p-value calculation for this test and the MB test with equal sample sizes. Here we shall apply importance sampling using a normal generating density to the problem of calculating conditional p-values for the MB and UMP test with unequal sample sizes.

A test of CV equality is a test of \( H_0 : c_1 = \ldots = c_{k-1} = 0 \) in Equation 2.24. By Definition 1.3.1.3 and Theorem 1.3.4.3, \( T_k \) is a sufficient statistic for the nuisance parameter \( \omega^* \). Then by Definition 1.3.1.1, the distribution of the \( M_i \)'s conditional on \( \sum_{i=1}^{k} (N_i - 1)M_i \) will not depend on the nuisance parameter. So we can turn the MB test into a similar test by calculating the p-value from the conditional distribution.

From Theorem 1.3.4.8, under the null, the kernel of this conditional distribution is

\[
I_{T_k=t_k}(m)H_M(m),
\]  

(4.18)

where \( H_M \) is as defined in Equation 2.24.

Because \( T_k \) is a linear function of the \( M_i \)'s, the density in Equation 4.18 is of the form in Equation 4.4; \( W \) will be the column vector whose \( i \)th element is \( N_i - 1 \). We can use a normal generating density as described in Section 4.4.1. Although we could use the Dirichlet generating density here, we shall use the normal generating density because Section 2.2.2 and the delta Theorem (1.3.6.3) suggest that the sample \( M_i \)'s should be approximately normal.

Now the unconditional support of the data is \( M_i \geq 0 \); the conditional support will therefore be not only closed but bounded, by Theorem 1.3.8.12. With a compact conditional support, Theorems 1.3.9.1 and 1.3.9.2 will hold and Equation 4.14 will be asymptotically valid, as explained in the Section 4.4.1.

Under the null we are trying to test, \( CV_i = CV \), and we can estimate this with \( \hat{CV} \)
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

from Equation 2.13. \( Y^\ast \) will be \( k \times 1 \). From Theorem 1.3.4.1 and Proposition 4 we can derive that

\[
E(M_i) \approx \frac{CV^2}{1 + \frac{N_i - 2}{N_i} CV^2} \quad (4.19)
\]

\[
Var(M_i) \approx \frac{CV^4}{(1 + \frac{N_i - 2}{N_i} CV^2)^2 N_i - 1} \quad (4.20)
\]

So for \( E(Y_i^\ast) \) it makes sense to use \( \frac{CV^2}{1 + \frac{N_i - 2}{N_i} CV^2} \) and for \( Var(Y_i^\ast)_{ii} \) it makes sense to use \( \frac{CV^4}{(1 + \frac{N_i - 2}{N_i} CV^2)^2 N_i - 1} \), and it makes sense to set \( Var(Y_i^\ast)_{ij} \) = 0 for \( i \neq j \) since the \( M_i \)’s are independent. We can then get the mean and variance of \( Y \) using Equation 4.15 and Equation 4.16.

To calculate a conditional p-value for a test statistic, implement Algorithm 5 by drawing from \( Y \) using Algorithm 6, obtaining the kernels for the weights from Equation 4.17 and Equation 4.18. For the MB test, \( T_{pi} = MB \) (Equation 2.22). For the one-sided UMP test with \( H_a : CV_2 > CV_1 \), \( T_{pi} = M_2 \).

Essentially what we are doing here is testing for an effect on \( CV \) in a one-way experiment. By conducting importance sampling within cells in a way similar to that explored in Section 4.3, we can calculate similar p-values for main effects on \( CV \) in experiments with nuisance factors. Zacks [130] calculates such p-values by assuming a normal approximation for the log sample \( CV \); his methodology requires the experiment to be balanced. Wilson and Payton [131] derive the version of Equation 2.24 implied by McKay’s approximation, then use Theorem 1.3.6.13 to get the p-value for a likelihood ratio statistic. No exact method exists in the literature.

4.4.3 Simulation results for testing \( CV \) equality, unequal sample sizes

First I examine the properties of the MB test (using the asymptotic p-value) and the FM test with unequal sample sizes. Table 4.2 was formed by simulating independent datasets of \( k \) populations with coefficients of variation equal to \( CV \), half with sample size equal to \( N_1 \) and half with sample size equal to \( N_2 \) and . The sizes in this table
Table 4.2: Size of two-sided tests of CV homogeneity

<table>
<thead>
<tr>
<th>CV</th>
<th>k</th>
<th>N₁</th>
<th>N₂</th>
<th>MB</th>
<th>FM</th>
<th>std err</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>2</td>
<td>14</td>
<td>7</td>
<td>0.057</td>
<td>0.053</td>
<td>0.002</td>
</tr>
<tr>
<td>0.05</td>
<td>4</td>
<td>10</td>
<td>6</td>
<td>0.060</td>
<td>0.055</td>
<td>0.001</td>
</tr>
<tr>
<td>0.05</td>
<td>8</td>
<td>8</td>
<td>4</td>
<td>0.073</td>
<td>0.059</td>
<td>0.001</td>
</tr>
<tr>
<td>0.33</td>
<td>2</td>
<td>14</td>
<td>7</td>
<td>0.051</td>
<td>0.046</td>
<td>0.002</td>
</tr>
<tr>
<td>0.33</td>
<td>4</td>
<td>10</td>
<td>6</td>
<td>0.056</td>
<td>0.047</td>
<td>0.001</td>
</tr>
<tr>
<td>0.33</td>
<td>8</td>
<td>8</td>
<td>4</td>
<td>0.071</td>
<td>0.059</td>
<td>0.003</td>
</tr>
</tbody>
</table>

are similar to those in Table 2.5. Like Table 2.5, we see that the sizes diverge from nominal as sample sizes decrease. The divergence is more rapid for the MB test. For the large k, small N situations, MB appears to be unacceptably liberal.

Table 4.3 displays estimated powers of the two tests in various scenarios with unequal samples. For each scenario, a number of independent datasets were simulated, some populations with CV₁ and the rest with CV₂; the entries under N₁ and N₂ indicate the number of populations and the size of the sample drawn from each. If the sample sizes in the populations with high CVs are similar to the sample sizes for the populations with low CVs, the two tests have similar power, as in Table 2.6. For both tests, introducing correlation between sample size and power changes the power of the test, with the power of the FM test changing more dramatically. It appears to be easier to discern differences in CV if there is negative correlation between CV and sample size. The power in all of these scenarios hovers around 0.5.

Table 4.4 looks at the properties of the importance sampling test using the MB statistic. The datasets for the simulation reported in Table 4.4 were created as in Table 4.2 and 4.3. The size/power column indicates the proportion of datasets for each scenario for which the test rejected the null. The standard error column is the standard error of the estimate in the size/power column. With k = 2 or k = 4, the test appears to have acceptable Type I error for relatively small s. (The s values in Table 7 are much smaller than the s an investigator would choose in practice in order
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

Table 4.3: Power of two-sided tests of CV homogeneity

<table>
<thead>
<tr>
<th>CV_1</th>
<th>N_1</th>
<th>CV_2</th>
<th>N_2</th>
<th>MB</th>
<th>FM</th>
<th>std err</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>14</td>
<td>0.10</td>
<td>7</td>
<td>0.50</td>
<td>0.53</td>
<td>0.005</td>
</tr>
<tr>
<td>0.05</td>
<td>7</td>
<td>0.10</td>
<td>14</td>
<td>0.44</td>
<td>0.39</td>
<td>0.005</td>
</tr>
<tr>
<td>0.05</td>
<td>10, 10</td>
<td>0.10</td>
<td>6, 6</td>
<td>0.54</td>
<td>0.59</td>
<td>0.005</td>
</tr>
<tr>
<td>0.05</td>
<td>6, 6</td>
<td>0.10</td>
<td>10, 10</td>
<td>0.47</td>
<td>0.41</td>
<td>0.005</td>
</tr>
<tr>
<td>0.05</td>
<td>10, 6</td>
<td>0.10</td>
<td>10, 6</td>
<td>0.54</td>
<td>0.54</td>
<td>0.005</td>
</tr>
<tr>
<td>0.05</td>
<td>8, 8, 8, 8</td>
<td>0.10</td>
<td>4, 4, 4, 4</td>
<td>0.56</td>
<td>0.64</td>
<td>0.01</td>
</tr>
<tr>
<td>0.05</td>
<td>4, 4, 4, 4</td>
<td>0.10</td>
<td>8, 8, 8, 8</td>
<td>0.43</td>
<td>0.30</td>
<td>0.01</td>
</tr>
<tr>
<td>0.05</td>
<td>8, 4, 8, 4</td>
<td>0.10</td>
<td>8, 4, 8, 4</td>
<td>0.56</td>
<td>0.56</td>
<td>0.01</td>
</tr>
<tr>
<td>0.165</td>
<td>14</td>
<td>0.33</td>
<td>7</td>
<td>0.47</td>
<td>0.51</td>
<td>0.005</td>
</tr>
<tr>
<td>0.165</td>
<td>7</td>
<td>0.33</td>
<td>14</td>
<td>0.41</td>
<td>0.31</td>
<td>0.005</td>
</tr>
<tr>
<td>0.165</td>
<td>10, 10</td>
<td>0.33</td>
<td>6, 6</td>
<td>0.49</td>
<td>0.54</td>
<td>0.005</td>
</tr>
<tr>
<td>0.165</td>
<td>6, 6</td>
<td>0.33</td>
<td>10, 10</td>
<td>0.42</td>
<td>0.33</td>
<td>0.005</td>
</tr>
<tr>
<td>0.165</td>
<td>10, 6</td>
<td>0.33</td>
<td>10, 6</td>
<td>0.50</td>
<td>0.50</td>
<td>0.005</td>
</tr>
<tr>
<td>0.165</td>
<td>8, 8, 8, 8</td>
<td>0.33</td>
<td>4, 4, 4, 4</td>
<td>0.54</td>
<td>0.65</td>
<td>0.01</td>
</tr>
<tr>
<td>0.165</td>
<td>4, 4, 4, 4</td>
<td>0.33</td>
<td>8, 8, 8, 8</td>
<td>0.39</td>
<td>0.24</td>
<td>0.01</td>
</tr>
<tr>
<td>0.165</td>
<td>8, 4, 8, 4</td>
<td>0.33</td>
<td>8, 4, 8, 4</td>
<td>0.53</td>
<td>0.54</td>
<td>0.01</td>
</tr>
</tbody>
</table>

to attain a p-value estimate with an acceptably small standard error. The sizes less than 0.05 for the $CV = 0.33$ case are consistent with the results reported in Section 2.4.4 and probably reflect a slight inaccuracy in the approximate density in Equation 2.24 rather than an inadequacy of importance sampling. For $k = 8$, the importance sampling estimate of the p-value is biased downward for small $s$. $s$ needs to be quite large $- 100,000$ – to eliminate the bias.

The ASE column shows the average (over the simulated datasets) of the standard errors of $\hat{pv}_{IS}$ calculated using Equation 4.14. The expected standard deviation of the
### Table 4.4: Properties of importance sampling test

<table>
<thead>
<tr>
<th>CV₁</th>
<th>CV₂</th>
<th>N₁</th>
<th>N₂</th>
<th>s</th>
<th>size/power</th>
<th>stderr</th>
<th>ASE</th>
<th>rej</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.05</td>
<td>14</td>
<td>7</td>
<td>1,000</td>
<td>0.050</td>
<td>0.002</td>
<td>0.013</td>
<td>0.02</td>
</tr>
<tr>
<td>0.05</td>
<td>0.05</td>
<td>10,6</td>
<td>10, 6</td>
<td>3,000</td>
<td>0.052</td>
<td>0.001</td>
<td>0.010</td>
<td>0.07</td>
</tr>
<tr>
<td>0.05</td>
<td>0.05</td>
<td>8, 4, 8, 4</td>
<td>8, 4, 8, 4</td>
<td>3,000</td>
<td>0.066</td>
<td>0.001</td>
<td>0.022</td>
<td>0.34</td>
</tr>
<tr>
<td>0.05</td>
<td>0.05</td>
<td>8, 4, 8, 4</td>
<td>8, 4, 8, 4</td>
<td>100,000</td>
<td>0.055</td>
<td>0.003</td>
<td>0.007</td>
<td>0.34</td>
</tr>
<tr>
<td>0.33</td>
<td>0.05</td>
<td>14</td>
<td>7</td>
<td>1,000</td>
<td>0.045</td>
<td>0.002</td>
<td>0.013</td>
<td>0.02</td>
</tr>
<tr>
<td>0.33</td>
<td>0.05</td>
<td>10, 6</td>
<td>10, 6</td>
<td>3,000</td>
<td>0.046</td>
<td>0.001</td>
<td>0.01</td>
<td>0.08</td>
</tr>
<tr>
<td>0.33</td>
<td>0.05</td>
<td>8, 4, 8, 4</td>
<td>8, 4, 8, 4</td>
<td>100,000</td>
<td>0.049</td>
<td>0.003</td>
<td>0.006</td>
<td>0.38</td>
</tr>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>14</td>
<td>7</td>
<td>1,000</td>
<td>0.48</td>
<td>0.005</td>
<td>0.009</td>
<td>0.01</td>
</tr>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>7</td>
<td>14</td>
<td>1,000</td>
<td>0.41</td>
<td>0.005</td>
<td>0.009</td>
<td>0.01</td>
</tr>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>10, 10</td>
<td>6, 6</td>
<td>3,000</td>
<td>0.51</td>
<td>0.005</td>
<td>0.008</td>
<td>0.05</td>
</tr>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>6, 6</td>
<td>10, 10</td>
<td>3,000</td>
<td>0.43</td>
<td>0.005</td>
<td>0.009</td>
<td>0.05</td>
</tr>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>10, 6</td>
<td>10, 6</td>
<td>3,000</td>
<td>0.51</td>
<td>0.005</td>
<td>0.008</td>
<td>0.05</td>
</tr>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>8, 8, 8, 8</td>
<td>4, 4, 4, 4</td>
<td>3,000</td>
<td>0.54</td>
<td>0.005</td>
<td>0.017</td>
<td>0.25</td>
</tr>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>4, 4, 4, 4</td>
<td>8, 8, 8, 8</td>
<td>3,000</td>
<td>0.40</td>
<td>0.005</td>
<td>0.02</td>
<td>0.30</td>
</tr>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>8, 4, 8, 4</td>
<td>8, 4, 8, 4</td>
<td>3,000</td>
<td>0.56</td>
<td>0.005</td>
<td>0.017</td>
<td>0.26</td>
</tr>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>8, 8, 8, 8</td>
<td>4, 4, 4, 4</td>
<td>100,000</td>
<td>0.51</td>
<td>0.01</td>
<td>0.007</td>
<td>0.25</td>
</tr>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>4, 4, 4, 4</td>
<td>8, 8, 8, 8</td>
<td>100,000</td>
<td>0.37</td>
<td>0.01</td>
<td>0.008</td>
<td>0.30</td>
</tr>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>8, 4, 8, 4</td>
<td>8, 4, 8, 4</td>
<td>100,000</td>
<td>0.50</td>
<td>0.01</td>
<td>0.007</td>
<td>0.27</td>
</tr>
<tr>
<td>0.165</td>
<td>0.05</td>
<td>14</td>
<td>7</td>
<td>1,000</td>
<td>0.45</td>
<td>0.005</td>
<td>0.009</td>
<td>0.01</td>
</tr>
<tr>
<td>0.165</td>
<td>0.05</td>
<td>7</td>
<td>14</td>
<td>1,000</td>
<td>0.37</td>
<td>0.005</td>
<td>0.013</td>
<td>0.02</td>
</tr>
<tr>
<td>0.165</td>
<td>0.05</td>
<td>10, 6</td>
<td>6, 6</td>
<td>3,000</td>
<td>0.46</td>
<td>0.005</td>
<td>0.008</td>
<td>0.05</td>
</tr>
<tr>
<td>0.165</td>
<td>0.05</td>
<td>6, 6</td>
<td>10, 10</td>
<td>3,000</td>
<td>0.38</td>
<td>0.005</td>
<td>0.009</td>
<td>0.06</td>
</tr>
<tr>
<td>0.165</td>
<td>0.05</td>
<td>10, 6</td>
<td>10, 6</td>
<td>3,000</td>
<td>0.46</td>
<td>0.005</td>
<td>0.008</td>
<td>0.06</td>
</tr>
<tr>
<td>0.165</td>
<td>0.05</td>
<td>8, 8, 8, 8</td>
<td>4, 4, 4, 4</td>
<td>100,000</td>
<td>0.48</td>
<td>0.01</td>
<td>0.007</td>
<td>0.28</td>
</tr>
<tr>
<td>0.165</td>
<td>0.05</td>
<td>4, 4, 4, 4</td>
<td>8, 8, 8, 8</td>
<td>100,000</td>
<td>0.33</td>
<td>0.01</td>
<td>0.007</td>
<td>0.23</td>
</tr>
<tr>
<td>0.165</td>
<td>0.05</td>
<td>8, 4, 8, 4</td>
<td>8, 4, 8, 4</td>
<td>100,000</td>
<td>0.48</td>
<td>0.01</td>
<td>0.007</td>
<td>0.30</td>
</tr>
</tbody>
</table>
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

Table 4.5: Size of test of $CV_1 = CV_2$ versus $CV_2 > CV_1$. ($SE = 0.002$)

<table>
<thead>
<tr>
<th>CV</th>
<th>$N_1$</th>
<th>$N_2$</th>
<th>Miller</th>
<th>UMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>14</td>
<td>7</td>
<td>0.054</td>
<td>0.050</td>
</tr>
<tr>
<td>0.33</td>
<td>14</td>
<td>7</td>
<td>0.047</td>
<td>0.050</td>
</tr>
</tbody>
</table>

Monte Carlo p-value (Theorem 1.3.5.13) under the null is 0.0124 for $s = 1000$, 0.007 for $s = 3000$, and 0.001 for $s = 100000$. For $k = 2$, the importance sampling estimator is about as efficient as the Monte Carlo estimator would be. For $k = 4$, it is slightly less efficient; having the Monte Carlo estimator would reduce the standard error by about 30%. For $k = 8$, the importance sampling estimator is highly inefficient; it has a standard error orders of magnitude above the Monte Carlo standard deviation.

The last column tells us the proportion of times we have to throw out a draw of $Y$ because one of the drawn elements is negative. This is rare for $k \leq 4$, but not uncommon for $k = 8$.

The results for the importance sampling test when the null does not hold show that the power for the importance sampling tests tends to be a few percentage points below the power of the $MB$ test, which is indicative of the fact that the importance sampling test is correcting for the liberality of the $MB$ test.

Table 4.5 looks at the actual size under the null of two one-sided tests of $CV$ equality: Miller’s test (Section 2.4.5) and the importance sampling test with $T_{\psi} = M_2$ with $s = 1000$. Both tests have close to nominal size.

Table 4.6 was created in the same way as Table 4.5, except that $CV_1$ was allowed to differ from $CV_2$. The tests have essentially identical power.

As with the two-sided test, importance sampling seems to work well as a way to calculate the $p$-values. The importance sampling test with $s = 1000$ has close to nominal size. For the simulations under the null, $ASE \approx 0.013$, essentially identical to the expected value of the standard deviation for the Monte Carlo $p$-value of 0.012. As in the applications of importance sampling for $k = 2$ above, fewer than 2% of the samples drawn from the generating normal distribution need to be thrown out because
Table 4.6: Power of one-sided tests of CV homogeneity, \(k = 2\). \((SE = 0.005)\)

<table>
<thead>
<tr>
<th>(CV_1)</th>
<th>(CV_2)</th>
<th>(N_1)</th>
<th>(N_2)</th>
<th>Miller</th>
<th>UMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.05</td>
<td>14</td>
<td>7</td>
<td>0.56</td>
<td>0.54</td>
</tr>
<tr>
<td>0.10</td>
<td>0.05</td>
<td>7</td>
<td>14</td>
<td>0.64</td>
<td>0.63</td>
</tr>
<tr>
<td>0.33</td>
<td>0.165</td>
<td>14</td>
<td>7</td>
<td>0.51</td>
<td>0.51</td>
</tr>
<tr>
<td>0.33</td>
<td>0.165</td>
<td>7</td>
<td>14</td>
<td>0.62</td>
<td>0.61</td>
</tr>
</tbody>
</table>

of negative values.

4.4.4 Implementing importance sampling for testing the scale parameter of a gamma distribution

This example illustrates ELS and also shows that by using transformations, we can turn some nonlinear sufficient statistics into linear ones.

Suppose \(U_i \sim \text{gamma}(\alpha, \beta)\), and we want to conduct a test of \(H_0 : \beta = \beta_0\) against \(H_a : \beta > \beta_0\). From Definition 1.3.5.10, we can write the density of the data under the null as a constant times

\[
I_{u_1 \geq 0, \ldots, u_N \geq 0} \exp \left( (\alpha - 1) \sum_{i=1}^{N} \ln(u_i) - \frac{1}{\beta_0} \sum_{i=1}^{N} u_i \right).
\]

From Theorem 1.3.4.3, \(T_\alpha = \sum_{i=1}^{N} \ln(U_i)\) is sufficient for \(\alpha\). This is not a linear sufficient statistic. But consider the data vector formed by log-transforming \(U\): \(X_i = \ln(U_i)\). By Theorem 1.3.5.1, the density of the new dataset is a constant times

\[
\exp \left( \alpha \sum_{i=1}^{N} x_i - \frac{1}{\beta_0} \sum_{i=1}^{N} \exp(x_i) \right). \quad (4.21)
\]

For this density, the sufficient statistic for the nuisance parameter is linear: \(T_{n\beta} = \)
Chapter 4. Monte Carlo Conditional \( p \)-value Calculation for Continuous Data

\[ \sum_{i=1}^{N} X_i; \]  Equation 4.21 is an example of Equation 4.3 where

\[ h(x) = 1, \quad G(x, \theta_{pi}) = \exp \left( -\frac{1}{\beta_o} \sum_{i=1}^{N} \exp(x_i) \right), \]

and \( W \) is simply a row vector of 1s. Notice that we have gotten rid of the nonnegativity constraints via this transformation.

By Theorem 1.3.4.4, it would make sense to choose \( T_{pi} \equiv \sum_{i=1}^{N} \exp(X_i) \) as the test statistic.

Equation 4.4 yields the kernel for the conditional density:

\[ \exp \left( -\frac{1}{\beta_o} \sum_{i=1}^{N-1} \exp(x_i) - \frac{1}{\beta_o} \exp(t_{np} - \sum_{i=1}^{N-1} x_i) \right). \]

The indicator function has been incorporated here by using it to eliminate \( x_N \).

It can be shown that the \( MLE \) for \( \alpha \) solves

\[ \sum_{i=1}^{N} X_i - N \ln \beta_o = N \psi(\alpha), \]

where \( \psi \) is the \textit{digamma} function. This equation is readily solved by the method of bisection (Definition 1.3.8.8).

To implement \( ELS \), we would generate each draw of \( Y[1 : N-1] \) from the likelihood

\[ \exp \left( \hat{\alpha}_{MLE} \sum_{i=1}^{N-1} y_i - \frac{1}{\beta_o} \sum_{i=1}^{N-1} \exp(y_i) \right). \] (4.22)

But recalling that this is the likelihood of \( iid \) log-transformed \textit{gamma} variables, we can do this by generating a random sample of \( N-1 \) \textit{gamma} variables with scale parameter \( \beta_o \) and shape parameter \( \hat{\alpha}_{MLE} \) and then log transforming them. We would obtain \( Y_N \) by \( t_{np} - \sum_{i=1}^{N-1} y_i = y_N \).

We now have everything we need to use Algorithm 5 to calculate a \( p \)-value for \( T_{np} \) conditional on the observed value of \( T_{pi} \): the form of \( T_{pi} \) and \( T_{np} \), a way to generate a suitable \( Y \), and formulas for the kernels of the generating and target density. The
importance sampling test that results will converge to the *UMP* similar test.

Table 4.7 displays the performance of the *ELS* test. For each scenario in the table, 10,000 datasets were generated with the specified $N$ and $\alpha$, with $\beta = 1$, and for each dataset the estimated likelihood sampling test of the null $\beta = 1$ was conducted with $s$ at a prespecified level. The results of the simulations do not depend on the value of $\beta$; since $\beta$ is a scale parameter, one can test $H_0$ is $\beta = \beta_0$, by multiplying the data by $\frac{1}{\beta_0}$ and then testing $\beta = 1$. The values of $\alpha$ were chosen to create a wide range for the coefficient of variation $\frac{1}{\sqrt{\alpha}}$ for a *gamma* variable, from $\frac{1}{4}$ to 4.

The size of the *ELS* test is close to nominal even for small values of $s$, indicating that there is little bias in the *ELS* $p$-value.

The last column reports the *ASE*, which is discussed in the introduction to Section 4.4. For comparison, the expected standard deviation of a hypothetical Monte Carlo estimator is 0.018 for $s = 500$ and 0.012 for $s = 1000$. For $N = 5$, the *ELS* $p$-value
appears to have about 1.3 times the standard deviation of the Monte Carlo estimator, and for $N = 30$ it has about 2.1 times the Monte Carlo standard deviation. These suggest that $s$ should be set at 2 to 5 times the number of draws one would desire for a traditional Monte Carlo estimator if $N \leq 30$. Estimated likelihood sampling is more computationally-intensive than traditional Monte Carlo would be, but is feasible in this example.

As in the previous example, the properties of importance sampling appear to improve as the sample size decreases, a characteristic which could potentially carve a niche for it in statistics. The decrease in performance with increased dimension can be interpreted as reflecting a deteriorating match between the target density and the generating density. Table 4.4 highlights the fact that, with a large enough sample, the match can be so poor that the $s$ required for convergence can be impractical.

Notice that the upper bound on sample size for importance sampling to be feasible in this example is clearly much higher than the upper bound in the previous example. The question is raised whether ELS will be more efficient in general than importance sampling with a normal generating density. We have some theoretical reasons to think so. First, if $h$ and $G$ in Equation 4.3 put constraints on the support of $X$, those same constraints will be built into the estimated likelihood generating density, but not into the normal generating density. This means the discrepancy between the support of the target density in Algorithm 5 and the support of the estimated likelihood generating density should be less than the discrepancy between the target and a normal generating density, leading to fewer throwaway draws. Second, with the normal generating density, the difference between the generating density and the target grows as $N - d_{np}$ increases due to the curse of dimensionality. But with ELS, opposing forces are at work so that it’s not at all clear that the match should continue to get bad as $N$ grows. By Theorems 1.3.6.2 and 1.3.6.1 part 1, as $N - d_{np}$ increases, $\hat{\theta}_{np,MLE} \rightarrow_p \theta_{np}$, meaning that for large $N$ we are at least drawing from the “correct” marginal unconditional distribution. Because of this, by Theorem 1.3.6.1 part 2 and Central Limit Theorems (see Theorem 1.3.6.7), for ELS $\frac{1}{N-d_{np}}W^T Y$ should converge in probability to a constant which of necessity must be consistent with the observed values of the sufficient statistics; ie, for large samples, drawing from the unconditional marginal parameterized by
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

\{\theta_0, \hat{\theta}_{np, \text{MLE}}\} should reproduce the observed values of the sufficient statistics, so the "effect" of the condition \(T_{np} = t_{np}\) on the marginal distribution of a given \(X_i\) should be minor, giving us reason to think that the estimated likelihood generating density should be a good match for the marginal conditional.

Table 4.7 indicates that the estimate likelihood sampling test provides an effective way to get a valid test of \(\beta = \beta_0\). Engelhardt and Bain ([132]) took a different approach to an exact test, effectively deriving the joint density of the sufficient statistics \(T_\alpha\) and \(T_\beta\) for \(\alpha\) and \(\beta\) in the model of Definition 1.3.5.10; with this, using Theorem 1.3.5.14, one can get the conditional distribution of \(T_\beta\) given \(T_\alpha = t_\alpha\), and then by numerical integration obtain a conditional p-value for \(T_\beta\). To do the integration requires advanced methods; one needs to track down several different sources to understand Engelhardt and Bain’s calculations. The method here is simple by comparison. If nothing else, it can be used to provide a check on Engelhardt and Bain’s tables.

4.5 Gibbs sampling

Gibbs sampling (Definition 1.3.7.2) is an attractive algorithm for generating from a conditional distribution because if the first step of the sampler is in the support of the target density, each step in the resulting Markov chain (Definition 1.3.7.1) will produce a point in this support. Because it is designed to mimic sampling from the target density, it ameliorates the concern about choosing a correct reference or generating distribution, which we have seen causes problems for importance sampling. Typically, the stationary distribution of the Gibbs sampler will be the target density (Definition 1.3.7.5 and Theorem 1.3.7.4), and it is rarely the case that it violates the assumptions of the theorems (1.3.7.5, 1.3.7.6, and 1.3.7.7) that allow us to treat the steps of the chain as a random sample from the stationary distribution. Theorem 1.3.7.7 allows us to assess the standard error of the p-value estimate obtained with a finite number of steps of the Gibbs sampler and thus to figure out when to stop.

Furthermore, in the problem of this chapter, under the null hypothesis the Gibbs sampler starts out in the stationary distribution; theoretically, the predictive distribution (Definition 1.3.7.1) for any step in the chain would be the conditional distribution
we are trying to draw from, although the resulting sample will not be random because of correlation across steps. In other words, if the null is true, we should not need to throw out many steps of the Gibbs sampler. When the Gibbs sampler is started from an arbitrary point, as is the case in Bayesian applications, typically an investigator will throw out the first \( b_i \) steps and treat the remaining steps as a random sample from the target density; steps 1 to \( bi \) are called a “burn-in” period.

In the Gibbs sampling algorithm of Kolassa and Tanner [108], the target density is the approximate joint density of \( T_{pi} \) conditional on \( T_{np} = t_{np} \) in exponential families implied by the approximate univariate conditional cdfs of the elements of \( T_{pi} \) mentioned in Section 4.2. The problem the paper addresses is that there is no approximation for the joint density of \( T_{pi} \), requiring MCMC to supplement higher-order asymptotics. Here I am setting up Gibbs sampling algorithms whose target is in effect the actual conditional density of \( T_{pi} \).

As with importance sampling, we shall be interested in using simulations to assess the performance of Gibbs sampling with a finite number \( s \) of steps. We shall be especially interested in three outputs of these simulations: the Type I error of the associated test, as explained in Section 4.4, the ratio of the standard error of the Gibbs sampling \( p \)-value to the standard error of the Monte Carlo \( p \)-value (see Definition 1.3.3.7), and the effect of a burn-in period. As mentioned, a burn-in period might not be needed under the null, but under the alternative using a burn-in period may increase the power of the associated test by increasing the difference between the observed sample and the sample taken as representative of the stationary distribution.

In our context, \( g(Y(i)) \) in Theorem 1.3.7.7 is the indicator function \( I_{T_{pi}(Y(i))<t_{pi}} \). The approximate ratio of the standard error of the Gibbs sampling \( p \)-value to the Monte Carlo standard error is given by \( \sqrt{\tau} \) from Theorem 1.3.7.7. To calculate \( \tau \), I shall assume that the correlation of \( g(Y(i)) \) across steps satisfies

\[
\text{Corr}(g(Y(i)), g(Y(j))) = \rho^{|i-j|}, \tag{4.23}
\]
so that a reasonable estimate for $\tau$ is

$$\hat{\tau} = \frac{1}{1 - \hat{\rho}},$$

(4.24)

where $\hat{\rho}$ is the average of the first-order autocorrelations of $g(Y(i))$ from the simulated datasets.

Theorem 1.3.7.7 is taken from Roberts and Rosenthal [14]. (I have corrected a typo that appears in the original.) Here I shall derive a slightly different approximation for $\tau$, but I shall use Theorem 1.3.7.7 in subsequent calculations.

Letting $g$ and $Z$ be as defined in Theorem 1.3.7.7,

$$\text{Var}\left(\frac{1}{s} \sum_{i=1}^{s} g(Y(i))\right) \approx E\left(\frac{1}{s} \sum_{i=1}^{s} (g(Y(i)) - E(g(Z)))^2\right)$$

$$= \frac{1}{s^2} \sum_{i=1}^{s} E(g(Y(i)) - E(g(Z)))^2 + \frac{2}{s^2} \sum_{i=1}^{s} \sum_{j>i} E\{(g(Y(i)) - E(g(Z)))(g(Y(j)) - E(g(Z)))\}. \quad (4.25)$$

Now letting $\text{Var}(g(Z)) = \sigma^2$ and accepting the approximation

$$\text{Cov}(g(Y(i)), g(Y(j))) = \rho_{|i-j|}\sigma^2,$$

Equation 4.25 becomes

$$\sigma^2 \left(\frac{1}{s} + \frac{2}{s^2} \sum_{i=1}^{s} \sum_{j>i} \rho_{|i-j|}\right)$$

$$= \sigma^2 \left(\frac{1}{s} + \frac{2}{s^2}(s-1)\rho_1 + \frac{2}{s^2}(s-2)\rho_2 + \ldots + \frac{2}{s^2}\rho_{s-1}\right)$$

$$= \sigma^2 \left(\frac{1}{s} + \frac{2}{s^2} \sum_{i=1}^{s-1} (s-i)\rho_i\right)$$

$$= \sigma^2 \left(\frac{1}{s} + 2 \sum_{i=1}^{s-1} \rho_i - \frac{2}{s} \sum_{i=1}^{s-1} i\rho_i\right). \quad (4.26)$$

Now if the variance exists, the last term in parentheses converges to 0. Equation 4.26

162
suggests the following approximation for $\tau$:

$$\tau \approx 1 + 2 \sum_{i=2}^{\infty} \text{Corr}(g(Y(i)), g(Y(1))).$$

This is larger than the correction factor in Theorem 1.3.7.7.

### 4.5.1 Implementation of Gibbs sampling with linear sufficient statistics

If we use the full conditional distribution as the target density (Definition 1.3.7.1), we shall run into problems, because the jumping density (Definition 1.3.7.2) will be degenerate due to the fact that the indicator function in Equation 4.4 determines $x_i$ once the other elements of $x$ are known. Furthermore, using the full conditional distribution would lead to $d_{np}$ unnecessary substeps (Definition 1.3.7.2) at each step, since we can get the last $d_{np}$ elements of $x$ from the first $N - d_{np}$. We can get around this problem either by using the marginal conditional distribution or the distribution of $Z$ from Equation 4.10 as the target.

If the marginal conditional density is the target, by Theorem 1.3.5.14 the kernel of the $ij$th jumping density is simply Equation 4.8 with $x_1 = x_1(j), \ldots, x_{i-1} = x_{i-1}(j), x_{i+1} = x_{i+1}(j-1), \ldots, x_{N-d_{np}} = x_{N-d_{np}}(j-1)$. If the distribution of $Z$ from Equation 4.10 is the target, again by Theorem 1.3.5.14 the kernel of the jumping density will be Equation 4.11 with $z_1 = z_1(j), \ldots, z_{i-1} = z_{i-1}(j), z_{i+1} = z_{i+1}(j-1), \ldots, z_{N-d_{np}} = z_{N-d_{np}}(j-1)$. Generating from the jumping densities may not be straightforward, but they are univariate densities, and the problem of generating from univariate densities is much easier than that of generating correlated multivariate data with potentially complicated support restrictions, which we would be faced with if we wanted to generate directly from either of these target densities. The fact that we know only the kernel can be dealt with either by using a generating method that does not require knowledge of the constant of proportionality, or by using (unidimensional) integration to calculate that constant. Note that the upper and lower bounds of the support of the jumping density will not in general be infinite due to the constraints.
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

We can derive a simpler expression for the jumping density for a marginal conditional target by what I shall call “minimal degrees of freedom” derivation. To simplify the notation I’ll describe how it would work for the first substep of a step. The jumping density we seek for the $j$th step would be the distribution of $X_1$ conditional on $W^T X = t_{np}, X[2 : N - d_{np}] = x[2 : N - d_{np}](j - 1)$. This is the marginal distribution of $X_1$ implied by what I shall call the “joint conditional”: the joint distribution of $\{X_1, X_{N-d_{np}+1}, \ldots, X_N\}$ conditional on $W^T X = t_{np}, X[2 : N - d_{np}] = x[2 : N - d_{np}](j - 1)$. Looked at this way, we are solving the problem of defining an appropriate jumping density in the presence of constraints by allowing the minimum degrees of freedom in the first substep necessary to prevent degeneracy, hence the name of the approach. The minimal degrees of freedom approach to Gibbs sampling is very similar to the approach for testing for independence in square contingency tables in Smith et al. [133].

By Theorem 1.3.5.14, the kernel of the joint conditional will be proportional to Equation 4.4 with $X[2 : N - d_{np}]$ set equal to $x[2 : N - d_{np}](j - 1)$. If the Equation 4.3 has been well-parameterized, the support of the joint conditional will be a one-dimensional hyperplane in $\mathbb{R}^{d_{np}+1}$. That hyperplane is described by the equation $W_1^T X_1 + \sum_{i=N-d_{np}+1}^{N} W_i^T X_i = t_{np} - \sum_{i=2}^{N-d_{np}} W_i^T x_i(j - 1)$. If we pick a value $x_1$ for $X_1$, we fix a point on the hyperplane, and the corresponding value of $X[N - d_{np} + 1 : N]$ will be

$$h(V(x_1))G(V(x_1), \theta_p),$$

(4.28)
where $V(x_1)$ is

$$\{x_1, x[2 : N - d_{np}] (j - 1), (W^T_{N - d_{np} + 1 : N})^{-1} (t_{np} - \sum_{i=2}^{N-d_{np}} W_i^T x_i (j - 1) - W_1^T x_1)\}.$$ 

Because of the linearity of Equation 4.27, the term under the radical in Theorem 1.3.5.16 will be a constant, so we don’t need to include it in the kernel. The jumping densities for the other substeps will have analogous forms.

To obtain $\tau$ in order to use Theorem 1.3.7.7 to judge when to stop the Gibbs sampler, a practitioner could use Equation 4.24, where $\hat{\rho}$ would simply be the first-order autocorrelation estimated from the steps of the Gibbs sampler. To decide on the length of a burn-in period, a practitioner could simply increase $b_i$ from 1 until the estimated $p$-value stabilized.

Throughout Section 4.5 I shall refer to the “Gibbs sampling $p$-value” and the “Gibbs sampling test”. These terms are analogous to “Monte Carlo $p$-value” and “Monte Carlo test” (Definition 1.3.3.7).

### 4.5.2 Application: exponential regression with inverse link

Suppose that $X_i$ is exponential with mean $\frac{1}{\sum_{j=1}^{p} W_{ij} \beta_j}$. Such a model was proposed by Davidov and Zelen [134]. The resulting density of the data is

$$f_X(x; \beta) = I_{x_1 \geq 0, \ldots, x_N \geq 0} \prod_{i=1}^{N} \left( \sum_{j=1}^{p} W_{ij} \beta_j \right) \exp \left( -\beta_1 \sum_{i=1}^{N} W_{i1} x_i - \ldots - \beta_p \sum_{i=1}^{N} W_{ip} x_i \right)$$  

(4.29)

This model satisfies Equation 4.3. Using the inverse link creates a convenient model in which the sufficient statistics are linear functions of the data and uniformly most powerful tests of hypotheses can be readily identified using Theorems 1.3.4.4 through 1.3.4.7.

Section 4.3 obtained similar tests for a factor in an exponential-data experiment without assuming any structure for the factor effects. Adopting the inverse link creates a more versatile model. It allows us to condition away a nuisance factor even when there is no replication within levels of the factor. It can reduce the number of sufficient
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

statistics we need to condition on. For instance, if there are two nuisance factors with 3 levels each, the analysis of Section 4.3 requires conditioning on 9 sufficient statistics; but if we adopt the inverse link model with only main effects, we need to condition on only 2 sufficient statistics. Finally, with no assumed structure for the factor effects, no UMP test exists except in very special cases.

For the case where \( E(X_i) = \frac{1}{\beta_0 + \beta_1 W_i} \), Davidov and Zelen derived the exact distribution of the sufficient statistic \( \sum_{i=1}^{N} W_i X_i \) for \( \beta_1 \) conditional on the sufficient statistic \( \sum_{i=1}^{N} X_i \) for \( \beta_0 \). Their derivation involves knowledge of Laplace transforms and clever derivation. It is interesting to note that, for testing the null that \( \beta_1 = 0 \) in their problem, one can simply use Algorithm 4 to condition away \( \beta_0 \). But the larger point here is that by implementing Gibbs sampling conditional Monte Carlo, without any tedious calculations, we shall be able to conduct similar tests for hypotheses that place null values on any number of \( \beta \)s of interest in the presence of any number of nuisance predictors.

For a generalized linear model for any type of gamma data with known \( \alpha \), the inverse link is the canonical link (Definition 1.3.5.8), resulting in exponential-family models with trivial UMP tests and sufficient statistics. In particular, one might choose inverse links to relate sample variances or sample coefficients of variation for normal data to predictor variables; ie, if one wanted to model \( \sigma_{ij}^2 \) from Equation 4.13 or \( \omega^* + c_i \) from Equation 2.24 as the inverse of \( \sum_{j=1}^{p} W_{ij} \beta_j \) one would get an exponential family density with linear sufficient statistics for \( \beta \). The Gibbs sampling approach developed below for conducting similar tests can in principle be applied to any canonical-link glm for gamma data, making such models worth exploring.

At least two other authors have pointed out the convenience of the inverse link in modeling gamma variables ([122], [135]). An obvious question arises as to how often it will be appropriate; after all, the linear link seems simpler, and the log link \((\ln(E(X_i)) = \sum_{j=1}^{p} W_{ij}^T \beta_j)\) ensures nonnegativity. But very often, there is no compelling reason to choose one monotonic link over another. Nair and Pregibon [122] argued that in modeling normal dispersion effects with sample variances,

If the sample variances are all of the same order of magnitude, any link function can be reasonably approximated by a linear function. So although
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

the interpretation of the dispersion effects may change as the link changes, the broad qualitative conclusions concerning model identification will be the same.

There is something to be said for using a model for which exact inference is possible; of the two sources of error in inference – model misspecification and method inexactness – one is eliminated as a concern. In any event, conditional Monte Carlo makes possible a pure goodness of fit test of the inverse link model for gamma data with known shape since there is a linear sufficient statistic for every unknown parameter, so one could readily test the inverse link if one were to try it.

In this application of Gibbs sampling we seek to test $H_0: \beta_{p+1} = 0$ against $H_a: \beta_{p+1} < 0$ (The alternative is that the $p$th variable has a positive effect on the mean.) Equation 4.29 is then the model for the data under $H_0$. The UMP similar test has the form $T_{pi} \equiv \mathbf{W}_{p+1}^T \mathbf{X} > b$ (Theorem 1.3.4.6), where $\mathbf{W}_{p+1}$ is a column we would add to $\mathbf{W}$ for the alternative model, with the $p$-value calculated from the conditional distribution of $\mathbf{X}$ given the value of $\mathbf{W}^T \mathbf{X}$.

We shall use the distribution of $\mathbf{Z}$ in Equation 4.10 as the target density. In Equation 4.29, under the null the density is a function of the data only through the sufficient statistics for the nuisance parameters. Then applying Theorem 1.3.5.15, we find that the conditional density of the data under the null is

$$I_{(\mathbf{W}^T \mathbf{X} = \mathbf{t}_{np}, x_1 \geq 0, \ldots, x_N \geq 0)}.$$ (4.30)

In other words, $\mathbf{Z}$ is uniform on the set $SZ$ of points in $\mathbb{R}^{N-p}$ that satisfies

$$(\mathbf{W}^T \mathbf{W})^\theta \mathbf{W}t_{np} + \mathbf{P} \mathbf{z} \geq 0.$$

From Section 4.5.1, the $ij$th jumping density will be constant on the set

$$I_{(\mathbf{W}^T \mathbf{X} = \mathbf{t}_{np}, x_1 \geq 0, \ldots, x_N \geq 0)}(z_1(j), \ldots, z_{i-1}(j), z_i, z_{i+1}(j-1), \ldots, z_{N-dnp(j-1)}(j-1) \geq 0).$$ (4.31)

We can expect $SZ$ to be compact (Definition 1.3.8.1) and convex (Definition 1.3.8.2).
For compactness, closedness follows from the inclusiveness of the constraint; to see that the set is bounded, consider the set $SX$ of points in $\mathbb{R}^N$ that satisfy

$$I_{W^TX=t_{np}, \ x_1 \geq 0, \ldots, x_N \geq 0} = 1$$

(4.32)

In Section 4.4.1 I argued that this set is compact if $W_{i1} = 1$ for all $i$ — if there is a constant in the regression, which merely implies that the mean cannot be infinite. Now from Theorem 1.3.8.7, $SZ$ is the set of points satisfying

$$z = P^T(x - (WW^T)^{g}W_{np})$$

for some $x$ in $SX$. So the boundedness of $SX$ implies the boundedness of $SZ$.

For convexity, let $z_1 \in SZ$ and let $z_2 \in SZ$. Convexity will be proven if we can show $H(\lambda) = (WW^T)^{g}W_{np} + P(\lambda z_1 + (1 - \lambda)z_2) \geq 0$ for $0 \leq \lambda \leq 1$. It is trivially true for $\lambda = 0$ and $\lambda = 1$. Now $\frac{\partial H}{\partial \lambda} = P(z_1 - z_2) - a$ constant. So for each element, $H(\lambda)$ has its smallest value at $H(0)$ or $H(1)$, and both of these are greater than 0.

Convexity allows us to infer that the support of the jumping density for $z_i(j)$, the set described in Equation 4.31, is an interval containing $z_i(j - 1)$. Compactness guarantees that the uniform density on this support is well-defined. The following algorithm finds the upper and lower limits on that interval and fills in the details of how the Gibbs sampling $p$-value will actually be calculated.

**Algorithm 7. Gibbs Sampling for conditional inference in exponential regression with an inverse link**

1. Solve $x = (WW^T)^{g}W_{np} + Pz$ for $z$. Call the solution $z(1)$.

2. Draw $z_1(j)$ from a uniform distribution with upper bound $U_1(j)$ and lower bound $L_1(j)$.

   To find the upper and lower bounds for $z_1(j)$, find all values of $z_1$ that set at least one element of $(WW^T)^{g}W_{np} + \sum_{i=2}^{N-p} P_{1}z_i(j - 1) + P_{1}z_1$ equal to 0. $U_1(j)$ will be the smallest such value of $z_1$ that is greater than $z_1(j - 1)$, and $L_1(j)$ will be the largest such value of $z_1$ that is less than $z_1(j - 1)$.
3. Draw $z_2(j)$ from a uniform distribution with upper bound $U_2(j)$ and lower bound $L_2(j)$. To find the upper and lower bounds of the uniform distribution of $z_2(j)$, you would proceed as in step 2, finding all values of $z_2$ that set at least one element of $(WW^T)^{p}W_{tp} + \sum_{i=3}^{N-p} P_i z_i (j-1) + P_1 z_1(j) + P_2 z_2$ equal to 0.

4. And so on until $z_{N-p}(j)$ is drawn from a uniform distribution with upper bound $U_{N-p}(j)$ and lower bound $L_{N-p}(j)$ calculated similarly.

5. Repeat steps 2-4 $s$ - 1 times.

6. For each $z(i)$ drawn in this way calculate $\text{Ind}(z(i)) \equiv I_{t_{pi}}((WW^T)^{p}W_{tp} + P z(i)) \geq t_{pi}$, where $t_{pi}$ is the observed value of $T_{pi}$.

7. Take $\hat{p}_{GS} \equiv \frac{\sum_{i=1}^{s} \text{Ind}(z(i))}{s}$ as the estimate of the $p$-value.

Steps 1 – 5 of Algorithm 7 form a Gibbs sampling algorithm (Definition 1.3.7.2) whose target density is defined by Equation 4.30 and whose stationary distribution (Definition 1.3.7.5) is therefore that of the variable $Z$, by Theorem 1.3.7.4. Algorithm 7 is clearly aperiodic (Definition 1.3.7.6). For a quick proof, for any epsilon ball around the point $z(i)$, there is a positive probability that $z(i+1)$ will be inside that ball. The algorithm is also irreducible (Definition 1.3.7.7). Formally proving this is tedious, but it is not hard to see. Recall that the support $SZ$ is convex. For any convex set in $\mathbb{R}^{N-p}$, one can move from any point $A$ to any point $B$ by a series of movements along the axes. Now Algorithm 7 does precisely that – makes probabilistic movements along the axes; clearly, if one takes an epsilon ball around any point $A$ in the support of $Z$ and an epsilon ball around another point $B$ in the support and starts Algorithm 7 at $A$, there would be some positive probability of making a finite number of movements that brought the algorithm into the epsilon ball around $B$. From Theorem 1.3.3.4 and Equation 4.10, the $p$-value associated with $t_{pi}$ is $E_{Z}(\text{Ind}(Z))$. Then $\hat{p}_{GS}(s)$ converges almost surely to the true conditional $p$-value by the Ergodic Theorem (1.3.7.6), we can use Theorem 1.3.7.7 to obtain its standard error, and by transforming each $(i)$ by Equation 4.10 we obtain what we can treat as a random sample of $X$ from its conditional distribution.
Chapter 4. Monte Carlo Conditional \( p \)-value Calculation for Continuous Data

Table 4.8: Type I error of one-sided test of no effect of continuous predictor

<table>
<thead>
<tr>
<th>( p )</th>
<th>( r )</th>
<th>( s )</th>
<th>size</th>
<th>std error</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5</td>
<td>500</td>
<td>0.052</td>
<td>0.001</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>500</td>
<td>0.047</td>
<td>0.002</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>500</td>
<td>0.053</td>
<td>0.002</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>500</td>
<td>0.062</td>
<td>0.002</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>1000</td>
<td>0.054</td>
<td>0.002</td>
</tr>
</tbody>
</table>

4.5.3 Simulation evaluation of Gibbs sampling for inverse link exponential regression

In the simulations in Section 4.5.3 under the null hypothesis, \( W \) will be a design matrix for a factorial experiment where each of the \( p - 1 \) factors has 2 levels, only main effects are included in the model, and there are \( r \) replicates in each of the \( 2^{p-1} \) cells. Such a matrix will have \( p \) linearly independent columns; the dimension of the sufficient statistic will be \( p \times 1 \), and \( X \), the response, will be an \( 2^{p-1}r \times 1 \) vector. Each step of Algorithm 7 will draw an \( 2^{p-1}r - p \)-dimensional \( z \) vector.

Under the alternative hypothesis, \( W \) will contain an additional column \( W_{p+1} \), a continuous variable that takes the values 1, 2, \ldots, \( r \) within each cell, and the coefficient associated with that variable will negative.

Table 4.8 displays the Type I error of the Gibbs sampling test implemented with the reported value of \( s \). To calculate Table 4.8, 10,000 or more datasets were generated for each scenario under the null hypothesis, with the \( \beta \)s set so that the mean response is 1 and no factors have any effects.

We can see that even for a fairly small \( s \), the test has approximately the correct size. This is not surprising. Recall that under the null, the Gibbs sampling Markov chain starts off in its stationary distribution, so we would not expect a systematic bias in the \( p \)-value, although it would have a different variance than that given by Theorem 1.3.6.9. As the number of degrees of freedom in the data grows, the value of \( s \) required
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

Table 4.9: Power of Gibbs sampling test with and without burn-in. \((SE = 0.025)\)

<table>
<thead>
<tr>
<th>(p)</th>
<th>(r)</th>
<th>(s)</th>
<th>(bi)</th>
<th>power</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5</td>
<td>500</td>
<td>0</td>
<td>0.46</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>750</td>
<td>250</td>
<td>0.44</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>500</td>
<td>0</td>
<td>0.53</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>750</td>
<td>250</td>
<td>0.53</td>
</tr>
</tbody>
</table>

to obtain a test with the correct size grows.

Table 4.9 reports some preliminary findings as to the need for a burn-in period. This table was generated in the same way as Table 4.8, except that the data were generated under the alternative hypothesis, with the coefficient on the continuous predictor chosen to make the power about 0.5. A burn-in period of size \(bi\) was thrown out before conducting the Gibbs sampling test. The number of datasets for each scenario in this table – 400 – was much smaller than the number of datasets for Table 4.8, because the goal here was just to get a general feel for the power as a function of \(bi\).

Table 4.9 indicates that the burn-in period need not be very large. To be safe, one could choose an \(s\) that gives a desired \(p\)-value standard error, then double it and throw out the first half. Alternatively, there is no need to use the observed data as the starting point of Algorithm 7. If one wanted a starting point that would be typical of the null distribution, one could pick any nonnegative point \(x\) that solved \(W_1^T x = t_{np}, \ldots, W_p^T x = t_{np}, W_{p+1}^T x = W_{p+1}^T \bar{X}\). With a starting point that is typical of the null, there is no reason to expect that any burn-in period would be needed.

Table 4.10 displays the average across 20 simulated datasets of the estimated autocorrelations of \(\text{Ind}(\mathbf{z}(i))\); for each scenario in the table the data were generated under the null and \(s = 500\) so that the estimated autocorrelations would be precise. We can see that there is substantial autocorrelation which increases with \(p\) and \(r\). Applying Equation 4.24, our estimate of \(\tau\) for the last scenario in Table 4.10 would be \(\frac{1}{1-0.8}\), so that the ratio of the standard error of the Gibbs sampling \(p\)-value to that of a hypothet-
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

Table 4.10: Autocorrelations for $\text{Ind}(z(i))$

<table>
<thead>
<tr>
<th>$p$</th>
<th>$r$</th>
<th>lag</th>
<th>1</th>
<th>2</th>
<th>10</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5</td>
<td>0.3</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>0.3</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.4</td>
<td>0.3</td>
<td>0.1</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>0.7</td>
<td>0.6</td>
<td>0.2</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>0.7</td>
<td>0.6</td>
<td>0.3</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>0.8</td>
<td>0.7</td>
<td>0.4</td>
<td>0.2</td>
<td></td>
</tr>
</tbody>
</table>

Trivial traditional Monte Carlo $p$-value would be $\sqrt{5}$; to obtain a desired standard error would require 5 times as many steps in the Gibbs sampler (post burn-in) as traditional Monte Carlo draws.

One caveat here is that the estimated autocorrelations in Table 4.10 are not consistent with Equation 4.23. $0.8^{10} = 0.1$, and $0.8^{20} = 0.0$, lower than the estimated autocorrelations in the Table. The autocorrelations assumed by Equation 4.23 evidently decay too fast, so that our estimate of $\tau$ will be an underestimate. Practitioners should be advised that they may need an approximation more complicated than Equation 4.24 to estimate $\tau$.

An alternative assumption that would allow more flexibility for matching the observed autocorrelations would be that $\text{Corr}(\text{Ind}(z(i)), \text{Ind}(z(i))) = \rho_1$ for $|i - j| = 1$ and $\text{Corr}(\text{Ind}(z(i)), \text{Ind}(z(i))) = \rho_1 \rho^{|i-j|-1}$ for $|i - j| > 1$ for some constants $\rho_1$ and $\rho$. For the last scenario in Table 4.10, $\rho_1 = 0.8$ and $\rho = 0.95$ would be a conservative choice – producing autocorrelations that are slightly larger than observed. Using Theorem 1.3.8.13, this would yield an estimate $\hat{\tau} = 17$. This suggests that in order for the investigator to obtain the same degree of precision as obtained by traditional Monte Carlo with $s$ draws, the investigator would require $17s$ steps for Algorithm 7.

The effect of the correlation across draws in the Gibbs sampler on the standard error
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

is quite dramatic in the last scenario in Table 4.10. One step of Algorithm 7 takes 0.23 seconds in SAS-IML. If one desired to estimate the p-value with a standard error of 0.002, one would need 10,000 draws with traditional Monte Carlo and 170,000 draws with Algorithm 7, requiring almost 11 hours of computer time. This is not infeasible for an investigator that needs just one run of the algorithm, but would require an inconvenience such as running the algorithm overnight. This scenario can be thought of as a minimally-practical scenario; problems with more than 6 nuisance parameters and 150 degrees of freedom may be too much for Gibbs sampling to handle.

4.5.4 Application: comparing exponential populations with Type I censoring

With survival data subject to Type I censoring, there is an upper bound $L$ on the lifetime. This would occur when there is a date at which observation must cease, so that if an individual survives from the start of the experiment until the end date, all we know about him is that his survival time is at least $L$.

If data are drawn from an exponential distribution with scale parameter $\beta$, under Type I censoring the likelihood of the data is ([111], page 105)

$$I_{0 \leq x_1 \leq \ldots \leq x_N \leq L} \beta^{-\sum_{i=1}^{N} I_{x_i < L}} \exp \left( -\frac{1}{\beta} \sum_{i=1}^{N} I_{x_i = L} L + I_{x_i < L} x_i \right).$$

For ease of notation, I shall call the number of fully-observed lifetimes $R$, and $O$ will be the set of individuals whose lifetime is fully observed. Then the likelihood is

$$\frac{1}{\beta^r} \exp \left( -\frac{1}{\beta} \left( (N - r)L + \sum_{O} x_i \right) \right).$$

There is a small literature on such data (see [111], [113], [136], [137]). To conduct similar inference, that does not depend on the value of the nuisance parameter $\beta$, is surprisingly tricky in this model. The issue arises in testing whether the value of $\beta$ is the same across $k$ populations and in goodness-of-fit testing of the exponential assumption. No exact homogeneity or goodness of fit tests have appeared.
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

As can be verified using Theorem 1.3.1.2, \(R\) and \(\sum O X_i\) form a minimal sufficient statistic for \(\beta\) ([111], page 106). Thus, the distribution of the data conditional on \(R = r, \sum O X_i = t_2\) does not depend on \(\beta\). In fact, by Theorem 1.3.5.15, it is uniform on the set of datasets yielding \(R = r, \sum O X_i = t_2\). Elements of this set satisfy three properties:

1. Exactly \(r\) of the observations are \(L\).
2. The sum of the fully-observed lifetimes is \(t_2\).
3. Each fully-observed lifetime lies between 0 and \(L\).

To generate uniform data satisfying these constraints is a two-part problem: selecting which observations are fully-observed and generating the fully-observed data. First I shall consider the latter problem.

Let \(Z\) be a random variable with the same distribution as that of the fully-observed data conditional on \(R = r, \sum O X_i = t_2\). Consider an \(r\)-dimensional \(U\) that is uniform on the support \(0 \leq U_1 \leq L, \ldots, 0 \leq U_r \leq L\). The density has the form of Equation 4.3 with \(h(u) = I_{0 \leq u_1 \leq L, \ldots, 0 \leq u_r \leq L,} K = 1, G = 1\). Let \(W = \{1, \ldots, 1\}\). By Theorem 1.3.5.14, we can deduce that the distribution of \(U\) conditional on \(W^T U = t_2\) is

\[
I_{\sum_{i=1}^r u_i = t_2, \ 0 \leq u_1 \leq L, \ldots, 0 \leq u_r \leq L}.
\]

This is a uniform density on the set satisfying items 2 and 3 above, so we shall be generating \(Z[1 : r - 1]\) if we draw from the marginal conditional distribution of \(U\).

We shall use Gibbs sampling to generate from this distribution. Applying Equation 4.28, the \(ij\)th jumping density is the indicator function for the event

\[
0 \leq u_1(j) \leq L, \ldots, 0 \leq u_{i-1}(j) \leq L, 0 \leq u_i \leq L, 0 \leq u_{i+1}(j) \leq L, \ldots, 0 \leq u_{r-1}(j) \leq L,
\]

\[
0 \leq t_2 - \sum_{i=1}^{i-1} u_i(j) - \sum_{i=i+1}^{r-1} u_i(j-1) - u_i \leq L.
\]

In other words, the jumping density is uniform on the intersection of the intervals

\[
0 \leq u_i \leq L \text{ and } t_2 - \sum_{i=1}^{i-1} u_i(j) - \sum_{i=i+1}^{r-1} u_i(j-1) - L \leq u_i \leq t_2 - \sum_{i=1}^{i-1} u_i(j) - L.
\]
\[\sum_{i=1}^{r-1} u_t(j - 1). \text{ The lower limit is } \max(0, t_2 - \sum_{i=1}^{r-1} u_t(j) - \sum_{i=i+1}^{r-1} u_t(j - 1) - L) \text{ and} \]
\[\text{the upper limit is } \min(L, t_2 - \sum_{i=1}^{r-1} u_t(j) - \sum_{i=i+1}^{r-1} u_t(j - 1)). \]

By the fact that the resulting Markov chain is a Gibbs sampler, Theorem 1.3.7.4 holds. Aperiodicity follows from the same argument as given for Algorithm 7. It’s also not hard to see that the support of the chain is convex (use the reasoning in Section 4.5.2), so that irreducibility follows from the Algorithm 7 argument as well. Then by Theorem 1.3.7.5, if we choose the number of steps \( s \) large enough we shall be on solid ground for treating the draws as a random sample of \( Z[1 : r - 1] \).

For the first part of the problem of generating censored exponential data conditional on a minimal sufficient statistic, the uniformity of the conditional distribution implies that each set of \( r \) observations is equally likely to be the fully-observed observations. To complete the job of generating the data, choose one of the \( \frac{1}{r} \) sets at random. The next few paragraphs formally describe an MCMC algorithm for calculating the \( p \)-value associated with the value \( t \) of a test statistic \( T(X) \) conditional on \( R = r, \sum_O X_i = t_2 \).

Let \( A \) be the vector created by placing the digits 1 to \( N \) in random order in a vector and then slicing off the last \( r \) elements. Let \( U_{mc} \) be a random variable whose distribution is that of \( Z[1 : r - 1] \). Then, conditional on \( R = r, \sum_O X_i = t_2 \) the data follow the stochastic representation \( X = g(A, U_{mc}) \), where \( g \) can be described in the following way: the \( N - r \) elements of \( X \) indexed by the values in \( A \) are equal to \( L \), and the values of the vector \( \{U_{mc}, t_2 - \sum_{i=1}^{r-1} U_{mc, i}\} \) are assigned in order to the remaining values of \( X \).

Let \( A(i) \) be the \( i \)th of \( s \) independent draws of the random vector \( A \), and let \( U_{mc(i)} \) represent the \( i \)th step in the Gibbs sampler for the fully-observed data. \( \{A(i), U_{mc(i)}\} \) form an irreducible, aperiodic Markov chain whose stationary distribution is that of \( \{A, Z\} \), so Theorem 1.3.7.5 will apply. Then for \( s \) large enough, we have justification for treating \( x(i) \equiv g(a(i), u_{mc(i)}) \) as a random sample from the conditional distribution of \( X \). Furthermore, by Theorem 1.3.7.6, the average of \( Ind(a(i), u_{mc(i)}) \equiv I_{T(x(i))<t} \) across the \( s \) steps in the chain will converge almost surely to the conditional \( p \)-value associated with the observed value \( t \), and we can use Theorem 1.3.7.7 to obtain the standard error of this estimated \( p \)-value.

Suppose the data are from \( k \) populations, and \( X_i \) is a vector of data from the
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

Table 4.11: Size of LR test for equality of scale parameters under Type I censoring. ($SE = 0.002$)

<table>
<thead>
<tr>
<th>$k$</th>
<th>$N^*$</th>
<th>Asy</th>
<th>MC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>40</td>
<td>0.052</td>
<td>0.050</td>
</tr>
<tr>
<td>4</td>
<td>26</td>
<td>0.049</td>
<td>0.048</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>0.058</td>
<td>0.053</td>
</tr>
</tbody>
</table>

nth population, where all samples are subject to the Type I censoring with the same maximum observed lifetime $L$. This kind of situation would arise in a time-constrained one-way experiment. With the exponential model being the simplest survival model and with Type I censoring a common reality, clearly it is of interest to be able to test whether the scale parameter is the same across populations.

Let $\beta_i$ be the scale parameter for population $i$, $r_i$ be the number of fully-observed lifetimes from the $i$th population, $O_i$ be the set of observations from the $i$th population for which lifetimes are fully-observed, and $N_i$ be the sample size of the $i$th population. The (log-transformed) likelihood ratio statistic for testing $H_0: \beta_1 = \ldots = \beta_k$ against the alternative that there is at least one inequality is ([111], page 116)

$$LR - ET \equiv 2 \left( \sum_{i=1}^{k} r_i \right) \ln \left( \frac{\sum_{i=1}^{k} ((N_i - r_i)L + \sum_{O_i} X_{ij})}{\sum_{i=1}^{k} r_i} \right)$$

$$- 2 \sum_{i=1}^{k} r_i \ln \left( \frac{(N_i - r_i)L + \sum_{O_i} X_{ij}}{r_i} \right).$$

We would reject the null for large values. Lawless [111] reports that it is “satisfactory” to use the approximate $\chi^2_{k-1}$ distribution (Theorem 1.3.6.13) for this test unless the number of fully-observed lifetimes is “quite small”. Under the null hypothesis, the data are from one homogeneous exponential population subject to Type I censoring. Thus, we can estimate similar $p$-values for $LR - ET$ via $MCMC$ as just described.

Table 4.11 reports the estimated Type I error of the tests resulting from two different
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

Table 4.12: Power of MCMC test for equality of scale parameters under Type I censoring. ($SE = 0.02$)

<table>
<thead>
<tr>
<th>$k$</th>
<th>$N^*$</th>
<th>burn-in</th>
<th>$s$</th>
<th>power</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>40</td>
<td>50</td>
<td>500</td>
<td>0.49</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>250</td>
<td>750</td>
<td>0.51</td>
</tr>
<tr>
<td>4</td>
<td>26</td>
<td>50</td>
<td>500</td>
<td>0.43</td>
</tr>
<tr>
<td>4</td>
<td>26</td>
<td>250</td>
<td>750</td>
<td>0.43</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>50</td>
<td>500</td>
<td>0.47</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>250</td>
<td>750</td>
<td>0.49</td>
</tr>
</tbody>
</table>

methods of calculating the p-values for $LR - ET$. For each scenario, 10,000 datasets were generated, each with $k$ samples of size $N^*$, from exponential distributions with $\beta = 1$ and $L = 1.4$, chosen so that $E(R) = 0.75N^*$. The “Asy column” reports the estimated Type I error associated with the $\chi^2_{k-1}$ approximation, and the “MC” column reports the Type I error associated with the MCMC conditional p-value with $s = 500$. The scenarios in the table have been chosen so that the power to detect a difference when half the populations have a scale parameter equal to 0.75 and half have one equal to 1.25 is approximately 0.5.

The table shows that even for small values of $s$, conditional Monte Carlo produces a test with accurate size. This is not surprising; it reflects the fact that under the null, the Markov chain starts out in the stationary distribution and so should produce p-values with low bias. For the sample sizes in the table, the asymptotic approximation provides a valid test if $k$ is small, but can be slightly liberal for large $k$.

Table 4.12 was generated in the same way as Table 4.11 except that half the populations had a scale parameter equal to 0.75 and half had one equal to 1.25, and the MCMC test was run with a burn-in period. Also, only 1000 datasets were generated for each scenario. The fact that the proportion of rejections is not sensitive to the size of the burn-in period indicates no apparent need for a large burn-in period.

Finally, in order to get an idea of how large $s$ should be for the standard error of...
the estimated \( p \)-value to be reasonably small, we need the look at the autocorrelations of the indicator function \( \text{Ind}(a(i), u_{mc}(i)) \). Averaging the estimated autocorrelations obtained for each simulated dataset in the simulations in Table 4.11, we find that there is no evidence that there is any autocorrelation at all. The average first-order autocorrelation is less than 0.004 in absolute value. This indicates that the Monte Carlo algorithm we have described should be about as efficient as drawing random samples from the actual conditional distribution of the data, so that Theorem 1.3.6.9 should provide a valid standard error for the \textit{MCMC} \( p \)-value.

The \textit{MCMC} algorithm runs in a reasonable time. With total sample size equal to 500, one step, including the generation of both \( A \) and \( U_{mc} \), runs in under 0.003 seconds in SAS-IML. This means that 100,000 steps can be done in about 5 minutes.

Comparing the two applications we have seen, we have some indication that using a marginal conditional target for Gibbs sampling may be preferable to using the distribution of \( Z \) from Equation 4.10, judging from the relationship between the computational intensity of the various methods and the amount of data. If \( Z \) is the target, typically each substep of the Gibbs sampler must satisfy \( N \) constraints – one for each element of the original data vector. If we were to use the marginal conditional as the target, typically we would need to satisfy only \( d_{np} + 1 \) constraints in each substep, one for the generated variable and one for each sufficient statistic. Not only would this reduce the computing time, but it might also “free up” the algorithm and allow for less correlation across steps.

In the censored data example the correlation across steps is certainly reduced by the random reallocation of fully-observed lifetimes. This creates a caveat to the conclusion that the marginal conditional target eliminates correlation. However, it also suggests a way to reduce such correlation if \( X \) is a random sample under the null: random reallocation of the elements drawn in each step to indices of the data vector.

For the problem of testing scale equality against general alternatives, our simulations indicate that for sample sizes allowing reasonable power to detect differences the benefit of \textit{MCMC} over the existing approximate approach is minor. However, the \textit{MCMC} algorithm can be quite useful. First, for very small samples the \( \chi^2 \) approximation is inadequate ([111], page 116). Second, we can use our algorithm to estimate...
Chapter 4. Monte Carlo Conditional \( p \)-value Calculation for Continuous Data

a conditional \( p \)-value for \textit{any} statistic designed to detect differences across populations, which means that if we want to test the null against a specific hypothesis, as long as we can come up with a reasonable statistic, we can get an essentially exact \( p \)-value. For example, the problem of detecting an ordered alternative with exponential data has arisen [138]. Even more generally, in any situation in which the investigator wants to model the scale parameter as a function of predictors, we can use the \textit{MCMC} algorithm to get an exact \( p \)-value for testing whether the model has any explanatory power. Third, we can use conditional Monte Carlo to conduct a pure goodness-of-fit test. As emphasized in Section 4.3, such a test is extremely useful for a model as simple as the exponential. It allows the investigator to try what is in effect the simplest possible life-testing model before moving to more complicated models.

An extension of Type I censoring allows the upper bound to differ across observations. This type of censoring would apply to time-constrained experiments where individuals enter at different dates. Introducing different time constraints for different observations vastly complicates the problem of drawing from the distribution of exponential data conditional on the minimal sufficient statistic for the scale parameter, so it will be left to future research.

Commentary on Fang et. al. [139]

By setting \( \gamma_i = 1 \) in Definition 1.3.5.4, we can see that uniform data on the set \( \sum_{i=1}^{r} Z_i = 1 \) is a special case of the Dirichlet distribution; placing the limit \( Z_1 \leq L/t_2, \ldots, Z_r \leq L/t_2 \) on such data creates a special case of what is known as the truncated Dirichlet distribution. The Gibbs sampling algorithm above for generating the fully-observed data conditional on their sum can be extended to generate truncated Dirichlet data: simply divide the data vector by \( t_2 \). This algorithm can also be obtained as a special case of the Gibbs sampling algorithm for generating truncated Dirichlet data suggested by Fang et. al. [139] for Bayesian purposes.

As an aside, that paper proposes a method for generating truncated Dirichlet data directly rather than through a Markov chain. The method hinges on the theorem provided in the paper that the marginal distribution of an element of a truncated Dirichlet vector is a truncated \textit{beta} distribution. Here I note that the theorem is not
true. The mistake the authors make is that they integrate out the other variables over the support of the non-truncated Dirichlet. Finding an analytical expression for the integral over the support of the truncated Dirichlet would in fact be quite an achievement. Thus, one cannot use the method in [139] for generating data directly from the truncated Dirichlet; Gibbs sampling is the only existing way.

To provide a quick counterexample to the theorem in the paper, consider the Dirichlet distribution with $\gamma_i = 1$ and $a = 3$, truncated so that $Z_1 < 0.5$, $Z_2 < 0.5$, $Z_3 < 0.5$. This can be generated via rejection sampling (Definition 1.3.5.15). The sample mean of $Z_1$ from 25,119 draws was 0.332; the sample standard deviation was 0.118. The theorem in [139] implies that $Z_1$ is beta(1, 2) (see [2], page 623) truncated at 0.5. Generating 75,096 values from this distribution via rejection sampling, the sample mean was 0.222 and the sample standard deviation was 0.142. From the theorems in Section 1.3.6, it is acceptable to treat the difference between the two sample means as normal with standard deviation $\sqrt{\frac{0.118^2}{25119} + \frac{0.142^2}{75096}} = 0.0009$. The difference in the sample means is many, many times the standard deviation of that difference. So clearly, the two distributions have different means, and the theorem in the paper is incorrect.

4.6 Future research: approaches for nonlinear sufficient statistics

If the sufficient statistics are nonlinear, the problem becomes more difficult. There are three different approaches we can turn to, none of which is a panacea.

4.6.1 Fiducial Monte Carlo

The fiducial Monte Carlo algorithm for drawing from $f_{X|T_{np}}(x|t_{np})$ works in the following way.

**Algorithm 8. Fiducial Monte Carlo**

1. Draw $\theta_{np}(i)$ from the fiducial distribution (Definition 1.3.3.5) implied by $t_{np}$. Record the latent variables $U(i)$. (Definition 1.3.3.5)
2. Calculate $X(i)$ using $\theta_{np}(i)$ and $U(i)$. ($\theta pi$ must be known.)

3. Repeat steps one and two $s$ times for a sample of size $s$.

Fiducial Monte Carlo was pioneered by Engen and Lillegard [95]. Like fiducial inference, this method is exact (draws from the actual conditional distribution) if one can create a pivotal quantity out of $T_{np}$ and $\theta_{np}$ [140], but it is not exact in general. Engen and Lillegard provided a “proof” that the method is exact if $\theta_{np}(i)$ is a unique function of $U(i)$ and $t_{np}$, but Lindqvist et. al. [141] subsequently disproved the result by presenting a counterexample; they also stated that the gamma distribution provided another example of where fiducial Monte Carlo is not exact.

However, since fiducial inference often gives close to exact inference, one might hope that the approximate inference that stems from this method will be fairly accurate. This was the finding in an application of this method to the Behrens-Fisher problem [142].

Another drawback in addition to inexactness is that drawing from the fiducial distribution is potentially nontrivial. While shortcuts may be available in special cases, in general each draw would require the solution of a system of nonlinear equations where one side may contain integrals that cannot be obtained analytically and must be reevaluated each time a new solution is proposed. Solving these equations may require programming skill and may be quite computationally intensive.

While I have listed this method in the nonlinear section, nothing prevents us from using it with linear sufficient statistics. I have not used it in the applications above because the methods chosen were more convenient.

### 4.6.2 Importance Sampling

**Marginal conditional as the target**

One option is importance sampling using the marginal conditional as a target. As mentioned in Section 4.4, finding an acceptable generating density is not guaranteed; among the problems are constraints on the support of the marginal conditional. The estimated likelihood (defined in Section 4.4.1) is a promising candidate, but the theoretical musings in Section 4.4.4 are far from rock-solid.
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

With nonlinear sufficient statistics, we often cannot obtain an analytical expression for the kernel of the marginal conditional. Importance sampling, can still be done because we just need to evaluate that kernel at the draws taken from the generating density rather than to draw from that kernel. We can evaluate Equation 4.5 or Equation 4.7 by finding \( g^{-1} \) numerically and by employing the formula \( J_{g^{-1}}(y) = \frac{1}{J_g(g^{-1}(y))} \), which we should be able to calculate via an analytical expression for \( J_g \).

But not having an analytical target still creates complications. The algorithm to evaluate Equation 4.5 numerically would result in a computationally-intensive algorithm. And \( g \) from Section 4.1 need not be one to one; the need to find multiple inverse functions can complicate numerical computation. Finally, if we cannot derive the support of the marginal conditional, finding a suitable target may be difficult; we would have to choose one with a conservatively large support, leading to a high proportion of throwaway draws.

Conditional Monte Carlo

The method that bears the name “conditional Monte Carlo” (CMC) due to its early origins [143], is essentially a form of importance sampling [144] with the full conditional as the target. CMC was not invented for the problem of interest in this chapter; in fact, it seems to have been forgotten by the statistics literature, while having been picked up on by physicists [145].

Suppose we want to estimate the p-value associated with \( t_{pi} \) conditional on \( T_{np} = t_{np} \). Suppose we can find a one-to-one differentiable function \( g \) such that \( g(X) = \{T(X), T_{np}\} \). With \( T_{np} \) fixed at \( t_{np} \), \( g \) defines a one-to-one transformation between the conditional support of \( X \) and the support of \( T \). I shall write this function as \( T = h_{t_{np}}(X) \); the function is undefined outside the conditional support of \( X \).

In order to do importance sampling, at a minimum we need to be able to first, generate data on the conditional support of \( X \), and second, calculate the density of the data so generated. If we can find a random variable \( Y \) with known density \( f_Y \) that has the same support as the range of \( h_{t_{np}} \), then \( Z = h_{t_{np}}^{-1}(Y) \) will have the same support as that of \( X \) conditional on \( T_{np} = t_{np} \), and by Theorem 1.3.5.1 will have the density \( f_Y(h_{t_{np}}(z))J_{h_{t_{np}}}(z) \).
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

Whether we can find a suitable \( g \) and a suitable \( Y \) is not at all guaranteed in general. No guidance is provided in the statistics literature. However, one does have the freedom to choose them to suit the purposes of importance sampling. One might reasonably try \( g(X) = \{X_1, \ldots, X_N, T_{np}\} \) and \( f_Y = f_{X[1:N-d_{np}]}(y; \{\theta_o, \hat{\theta}_{np, MLE}\}) \). This would produce the same algorithm as estimated likelihood sampling. (Since the range of \( h \) is most likely a subset of the unconditional support of \( X[1: N - d_{np}] \), one would need to throw out a number of draws of \( Y \).

Having a \( g \) that is not one-to-one is not prohibitive, if it is one-to-one in pieces (see Section 4.1). In this case, we could create a generating density out of \( f_Y \) by probabilistically determining which value in the support of \( X \) to map a drawn value of \( Y \) to.

Note that we must find a \( g \) and a \( Y \) that will not only create a random variable whose support is the conditional support of \( X \) but that will also create one whose density reasonably matches the full conditional. The difficulty in achieving all of this probably accounts for the fact that there is only one application of this approach of which I am aware in the statistics literature, which is to generate normal data subject to a constraint that is homogeneous of degree 1 [143] (see Definition 1.3.5.12).

Calculating the Jacobian when the range is a surface.

An additional difficulty with CMC is that calculating \( J_{h_{tnp}} \) can be tricky. \( h_{tnp} \) goes from \( \mathbb{R}^N \) to \( \mathbb{R}^{N-d_{np}} \), so the formula in Definition 1.3.5.2 cannot be used. Since formulas for Jacobians of functions between spaces of different dimension are hard to find – I couldn’t even find them in texts entitled “Advanced Calculus” – I shall explain their calculation here. I have no reference to provide, having not come across these results in any publication, though would be surprised if they are original.

Consider a function \( f \) from \( \mathbb{R}^p \) to a surface in \( \mathbb{R}^l \). Following Stewart ([18], page 1127), who handles the case of transformations from \( \mathbb{R}^2 \) to surfaces in \( \mathbb{R}^3 \), we would like to know, for a small hypercube \( A(y) \) in the domain of \( f \) with a vertex at \( y \), the ratio of the volume of the corresponding hyperparallogram \( f(A(y)) \) anchored at \( f(y) \) to the volume of \( A(y) \). We can calculate this with the following algorithm.

**Algorithm 9.** Calculating the Jacobian of a one-to-one function \( f \) from \( \mathbb{R}^p \) to a surface
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

in $\mathbb{R}^l$ at a point $y$.

1. Calculate the $l \times 1$ vector $\frac{\partial f(y)}{\partial y_1}$. Call this $D^1$.

2. Calculate the $l \times 1$ vector $\frac{\partial f(y)}{\partial y_2}$.

3. Find the orthogonal projection of $\frac{\partial f(y)}{\partial y_2}$ onto $\frac{\partial f(y)}{\partial y_1}$. This can be done by means of the orthogonal projection matrix (Definition 1.3.8.9).

4. Find the vector $D^2_1$ that is the difference between $\frac{\partial f(y)}{\partial y_2}$ and its orthogonal projection onto $\frac{\partial f(y)}{\partial y_1}$.

5. Calculate $\frac{\partial f(y)}{\partial y_3}$, its projection onto the column space of $\frac{\partial f(y)}{\partial y_1}$ and $\frac{\partial f(y)}{\partial y_2}$ (see Definition 1.3.8.6), and the difference $D^3_{12}$ between $\frac{\partial f(y)}{\partial y_3}$ and its projection.

6. Likewise, obtain $D^4_{123}, \ldots, D^p_{12 \ldots (p-1)}$.

Speaking very loosely, if we made unit perturbations along each axis in the domain starting at the point $y$, and we “straightened out” the corresponding hyperparallelogram in the range, the lengths of the sides of the hyperbox in the range would be the lengths of the $D$ vectors. Therefore:

7. The value of the Jacobian is the product of the lengths of the $D$ vectors.

We can actually conduct this algorithm by evaluating a single mathematical expression for $J_f(y)$ for the Jacobian of a one-to-one function:

$$J_f(y) = \sqrt{\prod_{i=1}^{p} \left( \frac{\partial f(y)}{\partial y_i} \right)^T (I - P_{1 \ldots i-1}) \frac{\partial f(y)}{\partial y_i}},$$

(4.33)

where $P_0$ is the zero matrix and $P_{1 \ldots i-1}$ is the orthogonal projection matrix (Definition 1.3.8.6) onto the column space of $\frac{\partial f(y)}{\partial y_1}, \ldots, \frac{\partial f(y)}{\partial y_{i-1}}$.

Notice that if $f$ is linear, then $y$ does not appear in the expression for the Jacobian, and the Jacobian will be a constant, a fact we have used above.

To calculate the Jacobian for a transformation $f$ from a surface $S_1$ in $\mathbb{R}^p$ to a surface $S_2$ in $\mathbb{R}^l$ is more complicated. However, if we can write $f = goh$, where $g$
is a transformation from \( \mathbb{R}^k \) to \( S_2 \) and \( h \) is a transformation from \( S_1 \) to \( \mathbb{R}^k \), then \( J_f(z) = \frac{1}{J_{h^{-1}(h(z))}} J_g(h(z)) \), and both Jacobians on the right hand side can be evaluated by Equation 4.33 (see Definition 1.3.5.1). Since \( h_{\text{top}} \) will be a function from a surface to \( \mathbb{R}^{N-d_{np}} \), we would calculate \( J_{h_{\text{top}}}(z) \) via the expression \( \frac{1}{J_{h_{\text{top}}^{-1}(h_{\text{top}}(z))}} \), which we can obtain using Equation 4.33.

To execute the calculation, we must evaluate \( \frac{\partial h_{\text{top}}^{-1}(h_{\text{top}}(z))}{\partial y} \). Using the implicit function theorem (1.3.8.15), this will be the first \( N-d_{np} \) columns of \( -\left( \frac{\partial g(z)}{\partial x} \right)^{-1} \). We can calculate \( \frac{\partial g(z)}{\partial x} \) analytically, and the inverse numerically.

### 4.6.3 Gibbs sampling

**Gibbs sampling when the marginal conditional is analytical**

If the marginal conditional is the target, by Theorem 1.3.5.14, the kernel of the jumping density will be either Equation 4.5 or Equation 4.7 with the values of all but one of the elements of \( X[1:N-d_{np}] \) fixed. For a concrete example, I shall consider the problem of testing whether two or more samples are from the same gamma distribution.

Under the null hypothesis of homogeneity, the combined sample comes from one gamma distribution, with the sufficient statistics \( T_\beta \) and \( T_\alpha \) for the nuisance parameters \( \alpha \) and \( \beta \) (see Definition 1.3.5.10). Here the goal of Gibbs sampling would be to calculate a \( p \)-value for a test statistic (with power to detect differences) conditional on \( T_\alpha = t_\alpha \) and \( T_\beta = t_\beta \).

We could also use data generated by the Gibbs sampler to conduct a pure goodness-of-fit test for the gamma distribution.

Bhattacharya [1] reviews the literature on comparing gamma distributions with unknown parameters. No exact tests exist. Also, no pure goodness-of-fit test has been proposed for the gamma distribution. Pettit [146] and Dahiya and Gurland [147] studied applications of common gof tests to the gamma distribution; these tests require the estimation of the gamma parameters and rely on asymptotic theory for their \( p \)-values.

Let \( g, X, \) and \( Y \) be as defined in Section 4.1. We have \( Y_1 = X_1, \ldots, Y_{N-2} = X_{N-2}, Y_{N-1} = T_\beta(X) \equiv \sum_{i=1}^{N-2} X_i + X_{N-1} + X_N, Y_N = T_\alpha \equiv \sum_{i=1}^{N-2} \ln(X_i) + \ln(X_{N-1}) + \sum_{i=1}^{N-2} \ln(X_i) + \sum_{i=1}^{N-2} \ln(X_{N-1}) + \)
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

\[ \ln(X_N). \] Then \( g^{-1}(Y)[1] = Y_1, \ldots, g^{-1}(Y)[N-2] = Y_{N-2}. \] To find \( g^{-1}(Y)[N-1] \) and \( g^{-1}(Y)[N], \) we find the \( X_{N-1} \) and the \( X_N \) that solve the equations

\[
\begin{align*}
X_{N-1} + X_N &= Y_{N-1} - \sum_{i=1}^{N-2} Y_i \\
X_{N-1}X_N &= \frac{\exp(Y_N)}{\prod_{i=1}^{N-2} Y_i}.
\end{align*}
\] (4.34)

Solving the second equation for \( X_{N-1} \) and substituting into the first, we get a quadratic equation whose solution is

\[
X_N = \frac{Y_{N-1} - \sum_{i=1}^{N-2} Y_i \pm \sqrt{(Y_{N-1} - \sum_{i=1}^{N-2} Y_i)^2 - 4 \exp(Y_N) / \prod_{i=1}^{N-2} Y_i}}{2}.
\] (4.35)

By the symmetry of the problem, we can see that there are two solutions to Equation 4.34, of the form

\[
\begin{align*}
g_1^{-1}(Y)[N-1] &= \frac{Y_{N-1} - \sum_{i=1}^{N-2} Y_i + A(Y_{N-1}, Y_N, \sum_{i=1}^{N-2} Y_i, \prod_{i=1}^{N-2} Y_i)}{2}, \\
g_1^{-1}(Y)[N] &= \frac{Y_{N-1} - \sum_{i=1}^{N-2} Y_i - A(Y_{N-1}, Y_N, \sum_{i=1}^{N-2} Y_i, \prod_{i=1}^{N-2} Y_i)}{2},
\end{align*}
\] (4.36)

and

\[
\begin{align*}
g_2^{-1}(Y)[N-1] &= \frac{Y_{N-1} - \sum_{i=1}^{N-2} Y_i - A(Y_{N-1}, Y_N, \sum_{i=1}^{N-2} Y_i, \prod_{i=1}^{N-2} Y_i)}{2}, \\
g_2^{-1}(Y)[N] &= \frac{Y_{N-1} - \sum_{i=1}^{N-2} Y_i + A(Y_{N-1}, Y_N, \sum_{i=1}^{N-2} Y_i, \prod_{i=1}^{N-2} Y_i)}{2}.
\end{align*}
\]

In other words, \( g \) is not one to one, but we can define \( g_1 \) on \( X_N \leq X_N - 1 \) and \( g_2 \) on \( X_N > X_N - 1 \) that will be one to one, so we would use Equation 4.7 rather than Equation 4.5 to get the marginal conditional.

Now for the \textit{gamma} distribution, in the notation of Equation 4.1, \( h = I_{x_1 \geq 0, \ldots, x_N \geq 0} \) and \( G = 1. \) We can see that the data will enter Equation 4.7 only through the indicator function \( h, \) the Jacobians of \( g_1^{-1} \) and \( g_2^{-1}, \) and the indicator function in Equation 4.7. In order for either \( g_1^{-1}(Y) \) or \( g_2^{-1}(Y) \) to exist, we require the term under the radical
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

in Equation 4.35 to be positive. So the expression that determines the indicator from Equation 4.7 is

\[(t_\beta - \sum_{i=1}^{N-2} x_i)^2 \geq 4 \frac{\exp(t_\alpha)}{\prod_{i=1}^{N-2} x_i}. \quad (4.37)\]

Using the fact that Equation 4.35 is necessarily nonnegative and taking advantage of symmetry, we get the following expression for the kernel of the marginal conditional:

\[2I_{(t_\beta - \sum_{i=1}^{N-2} x_i)^2 > \frac{\exp(t_\alpha)}{\prod_{i=1}^{N-2} x_i}} \times I_{x_1 \geq 0, \ldots, x_{N-2} \geq 0} \times J_{g^{-1}}(\{x[1 : N - 2], t_\beta, t_\alpha\}) \quad (4.38)\]

To evaluate the Jacobian in Equation 4.38, we can employ the formula

\[J_{g^{-1}}(\{x[1 : N - 2], t_\beta, t_\alpha\}) = \frac{1}{J_{g^{-1}}(\{x[1 : N - 2], t_\beta, t_\alpha\})}. \]

After some derivation, we find that the Jacobian in Equation 4.38 is proportional to

\[\frac{1}{\prod_{i=1}^{N-2} x_i \sqrt{(t_\beta - \sum_{i=1}^{N-2} x_i)^2 - 4 \frac{\exp(t_\alpha)}{\prod_{i=1}^{N-2} x_i}}}. \quad (4.39)\]

By Theorem 1.3.5.14, the \(ij\)th jumping density will be proportional to Equation 4.38, with all of the variables except the \(i\)th fixed. This will yield a kernel of the form

\[I_{L(i,j) < x_i < U(i,j)} J_{g^{-1}}(\{x_1(j), \ldots, x_{i-1}(j), x_i, x_i(j + 1), \ldots, x_{N-2}(j - 1), t_\beta, t_\alpha\}), \quad (4.40)\]

where \(L(i,j)\) and \(U(i,j)\) are the smaller and larger positive solutions for

\[(t_\beta - \sum_{k=1}^{i-1} x_k(j) - x_i - \sum_{k=i+1}^{N-2} x_k(j - 1))^2 = 4 \frac{\exp(t_\alpha)}{x_i \prod_{k=1}^{i-1} x_k(j) \prod_{k=i+1}^{N-2} x_k(j - 1)}. \quad (4.41)\]

The constraint in Equation 4.40 is derived from the need to satisfy the first indicator function in Equation 4.38. It can be shown, using graphical arguments, that except on a set which the Gibbs sampler has probability 0 of reaching, Equation 4.41 will always have two solutions and that \(x_i\) must be between those two solutions to satisfy
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

the constraint in Equation 4.37.

This example demonstrates that implementing Gibbs sampling can be challenging even when the marginal conditional is analytically tractable. It may be difficult to derive $g^{-1}$, the possibility of multiple solutions to Equation 4.6 creates additional work, it may be tedious to evaluate $h$ or $G$ at $g^{-1}(y)$ if either embodies constraints, it may be difficult to express the indicator function in Equation 4.7, the expression for the Jacobian may be quite long and complicated, especially if $d_{np}$ is large, and one must work to translate constraints on the support of Equation 4.7 into bounds on the univariate jumping density. Finally, one needs to come up with a strategy for drawing from that density.

When the marginal conditional cannot be written analytically

With nonlinear sufficient statistics, we might not be able to write the marginal conditional analytically. If the marginal conditional kernel cannot be written analytically, Gibbs sampling cannot be done, because we don’t have a jumping density to draw from. But for the Metropolis-Hastings-within-Gibbs algorithm (Definition 1.3.7.9) rather than pure Gibbs sampling, we don’t need to draw directly from the jumping density; we just need to be able to evaluate it. By Theorem 1.3.7.10, the Metropolis-within-Gibbs Markov Chain has the right stationary distribution, and typically Theorems 1.3.7.5, 1.3.7.6, and 1.3.7.7 will hold so that the properties of the algorithm will be similar to those of Gibbs sampling if we do enough steps.

Using Theorem 1.3.5.14, evaluating the marginal conditional kernel from Equations 4.5 or 4.7 will often be feasible, since we shall often be able to solve for $g^{-1}$ numerically, and since we can evaluate the Jacobian as $J_g(g^{-1}(\{x|N-d_{np},t_{np}\}))$, for which we can derive the necessary derivatives analytically (see Definition 1.3.5.1). If $g$ is not one-to-one, finding multiple solutions without some analytical tractability can become quite troublesome, however, especially if $d_{np}$ is large.

Moreover, the Metropolis-within-Gibbs algorithm is not guaranteed to converge in a practical number of steps. Like importance sampling, its efficiency depends crucially how well the generating density matches the jumping density target (see Definition 1.3.7.8). We saw with importance sampling that picking a generating density can be
hit or miss even if we have an analytical expression for the target, and it is made more difficult here by the fact that we may have not even know the support, though we might be able to identify bounds numerically. One aspect of this problem that might make it easier than the problem of finding a generating density for importance sampling is that the target in each substep is univariate. Even with a good generating density, the resulting algorithm, which will have higher correlation across steps than Gibbs sampling and which will require numerical calculations at each step, may be quite computationally intensive if we desire a precise \( p \)-value estimate.

Minimal degrees of freedom approach

Recall from Section 4.4.1 that we can look at the jumping density as the marginal distribution of \( x_i(j) \) from the joint conditional, giving us an alternative way of deriving it. Doing so will provide us with some beneficial insights.

For notational simplicity, here I shall assume that we are in the first substep of the \( j \)th step. The jumping distribution for \( X_1(j) \) is the marginal distribution of \( X_1 \) from the joint distribution of \( \{X_1, X_{N-d_{np}+1}, \ldots, X_N\} \) conditional on \( T_{np} = t_{np} \) and \( X_2 = x_2(j-1), \ldots, X_{N-d_{np}} = x_{N-d_{np}}(j-1) \). From Theorem 1.3.5.14 and Equation 4.2 the kernel of the joint conditional is

\[
I_{T_{np}=t_{np}}(\{x_1, x[2 : N - d_{np}](j - 1), x_{N-d_{np}+1}, \ldots, x_N\})
\]

\[
h(\{x_1, x[2 : N - d_{np}](j - 1), x_{N-d_{np}+1}, \ldots, x_N\})
\]

\[
G(\{x_1, x[2 : N - d_{np}](j - 1), x_{N-d_{np}+1}, \ldots, x_N; \theta_{pi}\}).
\]

The support of this distribution will be a one-dimensional surface in \( \mathbb{R}^{d_{np}+1} \). This surface is described by the equation

\[
T_{np}(\{x_1, x[2 : N - d_{np}](j - 1), x_{N-d_{np}+1}, \ldots, x_N\}) = t_{np}.
\]

Figure 4.1 provides an example of a one-dimensional surface in a higher-dimensional space.

Fixing a value of \( x_1 \) in Equation 4.43 will produce an equation that has a number of solutions in general. In a well-behaved problem, by Theorem 1.3.8.15 the \( x[N -
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

Figure 4.1: A one-dimensional surface in $\mathbb{R}^2$

$d_{np} + 1 : N]$ that solves the equation will be a local differentiable function of $x_1$ in the neighborhood of any particular point on the surface except on a set of measure 0. In other words, the surface will consist of $m$ identifiable pieces for the $i$th of which we can write $x[N - d_{np} + 1 : N] = f_i(x_1)$, and

$$\frac{\partial f_i(x_1)}{\partial x_1} = -\frac{\partial T_{np}(\{x_1, x[2 : N - d_{np}](j - 1), f_i(x_1)\})^{-1}}{\partial x[N - d_{np} + 1 : N]} \times \frac{\partial T_{np}(\{x_1, x[2 : N - d_{np}](j - 1), f_i(x_1)\})}{\partial x_1}. \quad (4.44)$$

For the surface in Figure 2, for each value of $x$ there are two possible values of $y$.
Chapter 4. Monte Carlo Conditional p-value Calculation for Continuous Data

(except for \(x = 0\), which would have measure 0). The first piece of the surface runs from the point \((0, 0)\) to the point \((1, 0.25)\), the second piece runs from \((1, 0.25)\) to the point \((-1, 0.75)\), and the third piece from \((-1, 0.75)\) to \((0, 1)\). Along each piece, \(y\) is a differentiable function of \(x\). Clearly, along the \(i\)th piece the surface can be described by a function \(y = f_i(x)\). At the points \(x = 1\) and \(x = -1\), the derivative of \(y\) with respect to \(x\) blows up, but this would be a set of measure 0.

Using Theorem 1.3.5.16 and simplifying, the jumping density for \(X_1(j)\) will be proportional to

\[
\sum_{i=1}^{m} I_{x_1 \in X_i} \times \\
\quad h(\{x_1, \mathbf{x}[2 : N - d_{np}](j - 1), f_i(x_1)\}) \times \\
\quad G(\{x_1, \mathbf{x}[2 : N - d_{np}](j - 1), f_i(x_1)\}; \theta_{pi}) \times \\
\quad \sqrt{1 + \frac{\partial f_i(x_1)^T}{\partial x_1} \frac{\partial f_i(x_1)}{\partial x_1}},
\]

where \(X_i\) is the set of all values of \(x_1\) that map to the \(i\)th piece of the surface (which will depend on the values of \(\mathbf{x}[2 : N - d_{np}](j - 1)\)). The \(m\) points \(f_i(x_1)\) are the possible solutions to Equation 4.43 for the fixed value of \(x_1\).

Of necessity, this will be proportional to the expression that we would get by applying Theorem 1.3.5.14 to Equation 4.7, but one might allow for easier derivation than the other. For instance, we saw in Section 4.4.1 that the minimal-degrees-of-freedom approach allowed us to derive a more convenient expression for the jumping density in the case of linear sufficient statistics. The last line of Equation 4.45 provides an explicit expression for the Jacobian denoted \(J_{g_{i-1}}\) in Equation 4.7. This may be more convenient than direct derivation of the Jacobian, which requires the evaluation of a determinant.

Recall from the previous discussion that if we need to use Equation 4.7 to get the marginal conditional and it is not available analytically, we are faced with the computational problem of finding multiple solutions to Equation 4.43. In the minimal degrees of freedom approach, this becomes the problem of finding multiple solutions.
to Equation 4.43 for a given value of $x_1$. But notice from Figure 4.1 that the “folds” in the surface – the endpoints of the pieces discussed above – will all occur where at least one element of $\frac{\partial f_i(x_1)}{\partial x_1}$ is infinite, which by Equation 4.44 would be all the points on the surface for which $\det \left( \frac{\partial T_{np}(\{x_1,x[2:N-d_{np}](j-1),f_i(x_1(j))\})}{\partial x[N-d_{np}+1:N]} \right) = 0$ or for which $\frac{\partial T_{np}(\{x_1,x[2:N-d_{np}](j-1),f_i(x_1)\})}{\partial x_1} = \infty$. Presumably, we could find these points numerically. Knowing where the folds in the surface are would give us some idea of what the surface looks like and of how many pieces it has. The minimal degrees of freedom approach has provided us with a potentially useful strategy for one of the computational challenges of MCMC.
Literature Cited


Literature Cited


Literature Cited


Literature Cited


Literature Cited


198
Literature Cited


Literature Cited


Literature Cited


Literature Cited


Literature Cited


Literature Cited

