Preface

These are lecture notes from the first Harold Hotelling Lectures, presented by Professor D.R. Cox on April 3, 4, 8 and 9, 1980. The notes were taken by John Castellana, Patrick Crockett, Jed Frees, Robert Frimmel, Paul Gallo, David Kikuchi, and Robert Smith. I then edited the notes, which received a final editing by Professor Cox, who also supplied the references.

David Ruppert

ASYMPTOTIC AND CONDITIONAL INFERENCE
SOME GENERAL CONCEPTS AND RECENT DEVELOPMENTS

by

D.R. Cox
Imperial College, London

I. Introduction

Before beginning discussion it is helpful to say some words concerning the topics which the theory of statistics should cover in order to give a basic background. Consider analysis of *data* (we do not include here the study of experimental design). We must describe its structure (e.g. the individuals and the variables). Another aspect is the *quantity* of data.

There are two broad aspects of statistical analysis: (1) the aim of the statistical procedure, including editing the data, preliminary and definitive analysis, the presentation of conclusions, and the interpretation of results; and (2) the basis of the analysis distinguishing (i) systematic variation and (ii) haphazard variation. This description of variation may be model-free or model-specific.

There are detailed considerations of technique: (1) the conceptual complexity of the technique; (2) the mathematical or numerical analytic complexity of the technique; (3) some kind of sensitivity analysis (i.e. power); (4) the style of presentation, which may range from a graphical (and probably nonmathematical) form to a numerical form; and (5) the type of answer, descriptive or probabilistic.

There are computational aspects which have been made vastly easier in recent years by the advent of the computer.

A. Preliminaries
   1. Data/Structure (individuals, variables)
   2. Quantity
B. Broad aspects of the statistical analysis
   1. Aim
      a. Editing
      b. Preliminary and definitive analysis
      c. Presentation of conclusions
      d. Interpretation
   2. Basis for
      a. Systematic variation
      b. Haphazard variation
      in an environment that may range between
      i. Model-free
      ii. Model-specific
   3. Detailed consideration of technique
      a. Conceptual complexity
      b. Mathematical or numerical analytic complexity
      c. Sensitivity analysis
      d. Style
         i. Graphical
         ii. Numerical
      e. Type of answer
         i. Descriptive
         ii. Probabilistic
   4. Computational aspects

The theory of statistics should cover all these topics. Theory is not synonymous with mathematics.

This series of lectures will focus on parametric inference so that we will be model-specific (instead of model-free). In definitive analysis, we may be interested in confidence intervals and the estimation of parameters, or in preliminary analysis we will want to use significance tests as guides for model choice. We will require a probabilistic, rather than descriptive, type of answer, and our style will be numerical rather than graphical.

There is a base problem that underlies the discussion: We have an observed random vector Y assuming values y in some sample space S_y of dimension n. There is a *model function*, i.e. a density function $f_Y(y,\theta)$ with respect to a measure μ. The parameter space Ω_θ is a subset of $\mathbb{R}^q$.

We shall assume that parameters θ may be written as an ordered pair $\theta = (\psi,\lambda)$ where λ is a nuisance parameter. We shall assume that the variation is *independent*, i.e. $\Omega_\theta = \Omega_\psi \times \Omega_\lambda$, which means that any value ψ may be combined with any value λ to form a parameter θ. We shall consider the null hypothesis $\psi = \psi_0$. We shall be interested in finding size critical regions denoted by $w_\alpha(\psi_0)$. For all (or many) $0 < \alpha < 1$, we shall want to require that:

(i) $w_{\alpha_1}(\psi_0) \subset w_{\alpha_2}(\psi_0)$ if $\alpha_1 < \alpha_2$;

(ii) $P(Y \in w_\alpha(\psi_0) ; \psi_0,\lambda) = \alpha$ for all $\lambda \in \Omega_\lambda$.

We shall be not be concerned with sensitivity (i.e. power) of such critical regions.

There are three applications which are special cases of this base problem.

(1) A genuine hypothesis test of a null hypothesis. Then we want to calculate the significance level
$P = p(Y) = \min\{\alpha: Y \in w_\alpha(\psi_0)\}$.

(2) A $1-\alpha$ confidence region $\{\psi: Y \notin w(\psi)\}$.

(3) Prediction (and parametric empirical Bayesian statistics). Let $Y = (Y_0,Y^\dagger)$ where Y_0 is observed, and $Y^\dagger$ is what we want to predict. Assume that there is a similar decomposition in the parameters $(\theta,\theta^\dagger)$. Let $\psi = \theta^\dagger - \theta$. Assume $\psi_0 = 0$. Then the prediction region is
$\{y^\dagger: (y_0,y^\dagger) \notin w_\alpha(0)\}$.

All three cases are covered by the base problem. (We should probably include a fourth case, namely a straight Neyman rule of behavior: e.g. we want to come up with a procedure that will reject H_0 falsely 5% of the time, say.)

II. Exponential Families

We consider a function $S = s(Y)$ and $\phi = \phi(\theta)$. The family of densities of Y has the form $\exp(-s'\phi - K\phi + a(y))$. The family of densities of S is

$$f_S(s,\theta) = \frac{e^{-s'\phi} f_0(s)}{M(\phi)}$$

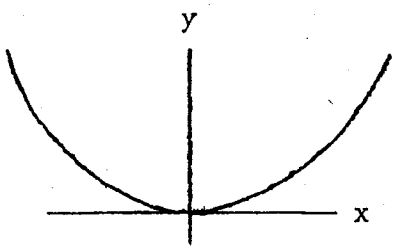
$$= \exp(-s'\phi - K(\phi))f_0(s) \ .$$

Usually Ω_θ is determined by all ϕ such that the moment-generating function (i.e. Laplace transform) satisfies $M(\phi) < \infty$. Here ϕ is the canonical parameter and $\mu = ES$ is the expectation parameter. Let p be the number of components of the minimal sufficient statistic. Recall that q is the dimension of the sample space (Barndorff-Nielsen, 1978; Johansen, 1979).

If $p = q$, then we say that the exponential family is *full* of order p, and we usually know what to do in terms of analysis. We will concentrate on the case when $p > q$ and call this exponential family a *curved* family of order (p,q).

Example 1 (Efron, 1975). Let $(Y_{11},Y_{21}),\ldots,(Y_{1n},Y_{2n}) \sim N_2\left(\begin{bmatrix}\theta \\ a\theta^2\end{bmatrix}, \begin{bmatrix}1 & 0 \\ 0 & 1\end{bmatrix}\right)$

where $a \neq 0$ is known. That means that the mean of these random vectors must fall on the parabola pictured below:



$y = a\theta^2$

$x = \theta$

and thus $q = 1$. But each vector itself may assume any value in $\mathbb{R}^2$, and the minimal sufficient statistic is the mean $S = n(\bar{Y}_{1\cdot}, \bar{Y}_{2\cdot})$, so $p = 2$. The full family generated by S is $N_2\left(\begin{bmatrix}\theta_1\\ \theta_2\end{bmatrix}, \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\right)$.

Example 2. Let $(Y_{11},Y_{21}),\ldots,(Y_{1n},Y_{2n}) \sim N_2\left(\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right)$. Then the density of a vector (Y_{1i},Y_{2i}) is

$$f_Y(y,\rho) \alpha \exp\left(-\frac{1}{2(1-\rho^2)}\left(\Sigma(Y_{1i}^2 + Y_{2i}^2) - 2\rho\Sigma Y_{1i}Y_{2i}\right)\right).$$

The minimal sufficient statistic is $S = \begin{bmatrix}\Sigma(Y_{1i}^2 + Y_{2i}^2)\\ \Sigma Y_{1i}Y_{2i}\end{bmatrix}$. Let $\phi = \begin{bmatrix}\frac{1}{2(1-\rho^2)}\\ \frac{\rho}{1-\rho^2}\end{bmatrix}$. This is another (2,1) family.

Example 3. Consider local inference near $\theta = \theta_0$. Consider the scalar case first. A first-order approximation of the log-likelihood function may look like

$$\log f_Y(y,\theta) = \log f_Y(y,\theta_0) + Ku(y,\theta_0) + o(K) ,$$

and a second-order approximation may look like

$$\log f_Y(y,\theta) = \log f_Y(y,\theta_0) + Ku(y,\theta_0) + \tfrac{1}{2}K^2 v(y,\theta_0) + o(K^2) .$$

Here $K = (\theta-\theta_0)$, $u(y,\theta_0) = \left\{\frac{\partial \log f_Y(y,\theta)}{\partial\theta}\right\}_{\theta=\theta_0}$, and $v(y,\theta_0) = \left\{\frac{\partial^2 \log f_Y(y,\theta)}{\partial\theta^2}\right\}_{\theta=\theta_0}$. By taking exponentials, the first-order approximation gives rise to a (1,1) family, the second-order approximation gives rise to a (2,1) family, etc.

One has a similar situation for the vector case ($q > 1$), except that now

$$u(y,\theta_0) = \nabla \log f_Y(y,\theta_0) ,$$

$$v(y,\theta_0) = \nabla\nabla' \log f_Y(y,\theta_0) .$$

The maximum likelihood estimator in a full exponential family is $\hat{\mu} = S$.

III. The Role of Asymptotic Theory

1. *Objectives.* One looks at the case when $n = \dim(S_y) \to \infty$. We do not actually consider taking repeated experiments of increasing sample sizes that get larger and larger; this is simply a technical device for producing approximations. That means that the question of the adequacy of the approximation has to be always considered though not necessarily answered! This occurs in the context of a two-fold role, each of which will be illustrated by an example.

(1) Asymptotic expansion may be a mathematical (i.e. numerical) simplification, even though an *exact* solution may be available in principle. This term *exact* makes sense only in a certain mathematical formulation, but then one must ask the question whether the mathematical formulation is appropriate for the underlying real problem.

(2) Asymptotic theory may establish the form of the solution, e.g. the shape of w_α.

Example 4. This illustrates (1) above. Consider the truncated Poisson density, the Poisson density conditioned on the fact that the random variable never assumes the value 0. The density function looks like

$$P(Y=r) = \frac{e^{-\theta}\theta^r}{r!(1-e^{-\theta})}, \qquad r = 1,2,3,\ldots .$$

This is a full exponential family. The minimal sufficient statistic for a random sample of size n from this distribution is simply the sample mean. But numerical work with regard to the study of this sample mean is difficult. So we might consider asymptotic theory.

If we work in a Bayesian framework, we may develop the role of asymptotic theory along the same lines as in the classical case, but we are entirely in case (1) above.

Example 5. This illustrates (2) above. Let $Y_1,\ldots,Y_n \sim N(\mu,\sigma^2)$. Consider the null hypothesis H: $\mu=0$, $\sigma^2=1$. The form of the critical

region is not obvious. We may use asymptotic theory to establish the form of the critical region, and having done so may calculate the size of the test exactly or approximately.

Example 6. Let $MS_1,\ldots,MS_m$ be independent normal theory estimates of $\sigma_1^2,\ldots,\sigma_m^2$ corresponding to $d_1,\ldots,d_m$ degrees of freedom, respectively. Consider the null hypothesis

$$H_0\colon \sigma_1^2=\ldots=\sigma_m^2 .$$

Again we may wish to use asymptotic theory to establish the shape of the critical region and then calculate the size of the test.

2. *Likelihood-based methods.* Let $l(\theta,Y) = \log f_Y(Y,\theta)$. Let

$$U = U(\theta_0) = u(Y,\theta_0) = \nabla l(\theta_0,Y) ,$$

$$V = V(\theta_0) = \nabla\nabla' l(\theta_0,Y) \stackrel{\text{def}}{=\!=} -n\bar{i}(\theta_0,Y) .$$

We will define $\bar{i}(\theta_0,Y)$ by this last equation and call it the observed information matrix. Let $EV(\theta_0) = -\bar{i}(\theta_0)$. Then $\bar{i}(\theta_0)$, thus defined, is called the expected information matrix.

Consider a simple null hypothesis $H_0\colon \theta=\theta_0$ (with no nuisance parameters). Let $\hat{\theta}$ be the maximum likelihood estimator of θ (i.e. $\hat{\theta}$ maximizes $l(\theta,Y)$). There are many procedures that may be used to test this null hypothesis, including

(1) $W = 2(l(\hat{\theta},Y) - l(\theta_0,Y))$ (Wilks)

(2) $W_u = U'\,\Omega_u^{-1}\,U$ (Rao-Bartlett)

(3) $W_e = (\hat{\theta}-\theta_0)'\,\Omega_e^{-1}(\hat{\theta}-\theta_0)$ (Wald)

Usually we let $\Omega_u = n\bar{i}(\theta_0)$. For Ω_e^{-1} we may take any of four expressions:

$$n\bar{i}(\theta_0)\ ,\ n\bar{i}(\hat{\theta}) \qquad\qquad n\bar{i}(\theta_0,Y)\ ,\ n\bar{i}(\hat{\theta},Y)$$

The first two expressions are related to the expected second derivatives of the log-likelihood function, and the last two expressions are related to the observed second derivatives of the log-likelihood function.

There is a generalization to the more interesting case $\theta = (\psi,\lambda)$ involving nuisance parameters. For W_u, use U_{ψ_0} where the relevant components of $\nabla\ell$ with $\psi=\psi_0$, $\lambda=\hat{\lambda}_0$ are the maximum likelihood estimator of λ at ψ_0. This might involve the Lagrange multiplier test or a $C(\alpha)$ test.

We can use any function $\tilde{\ell}$ whose first two derivatives are sufficiently close to ℓ. The situation above may be visualized by the graph

$\frac{1}{2}W$ is the height of the line, $W_u^{\frac{1}{2}}$ is proportional to the slope of the line, and $W_e^{\frac{1}{2}}$ is proportional to the horizontal distance.

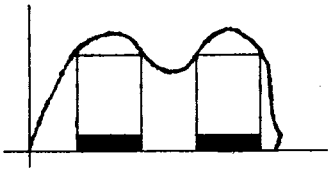
There are various considerations in the choice of a procedure.

(a) Computational. W_u has an advantage in the case of nuisance parameters, since one may want to vary the nuisance parameters, keeping the same form of the test statistic.

(b) General qualitative arguments for the choice.

(i) Invariance under changes in the parameter. As we change from θ to $f(\theta)$, W and W_u do not change since maximum likelihood estimators are transformed to maximum likelihood estimators. But W_e will change.

(ii) Nonmonotonicity. Consider a log-likelihood function with two peaks.



This situation is likely to lead to a confidence interval that consists of two disjoint intervals. Here we see that W is preferred to W_e.

(iii) Ancillarity. This is conditional inference, which will be mentioned later.

(c) Distributional approximations under H_0. W, W_u, and W_e are approximately chi-squared. How good is the approximation?

(d) Higher-order approximations.

(e) Sensitivity considerations.

Dr. Wassily Hoeffding (1965) was one of the first people to look at this situation, and he showed in a different context that W is advantageous.

There may be many nuisance parameters. The standard theory covers n: $\dim(\mathcal{L}_Y) \to \infty$ with q: $\dim(\Omega_\theta)$ fixed. There are major difficulties if number of parameters is comparable with number of observations.

Example 7.

$$\begin{array}{ll} (Y_{11}, Y_{12}) & N(\mu_m, \sigma^2) \\ \vdots \quad \vdots & \vdots \quad \vdots \\ (Y_{m1}, Y_{m2}) & N(\mu_m, \sigma^2) \end{array}$$

all r.v.'s independent; $\theta = (\sigma^2, \mu_1, \ldots, \mu_m)$, $n = 2m$, and $q = m + 1$. Then

$$\hat{\sigma}^2 = \frac{1}{2m}\Sigma(Y_{ij} - \bar{Y}_{i\cdot})^2 \xrightarrow{p} \tfrac{1}{2}\sigma^2 .$$

There are similar phenomena in other exponential family problems, e.g. logistic models for binary data. For the normal theory linear model, the MLE $\hat{\sigma}^2$ will be inconsistent unless $\dim(\theta) = o(n)$.

What can one do? Among possible devices are:

(a) Other limiting operations (e.g. as $n \to \infty$, $q \to \infty$ as well). Thus for predictive problems, we usually want

$$\dim(\mathcal{L}_{Y^\dagger}) \text{small (1, perhaps)}, \quad \dim(\mathcal{L}_{Y_0}) \to \infty .$$

(This is largely an open field for study.)

(b) Modified likelihood functions (Bartlett, 1937).

In Example 7, one can apply an orthogonal transformation to get

$$\begin{matrix} \frac{1}{\sqrt{2}}(Y_{11}+Y_{12}) , & \frac{1}{\sqrt{2}}(Y_{11}-Y_{12}) \\ \vdots & \vdots \\ \frac{1}{\sqrt{2}}(Y_{m1}+Y_{m2}) , & \frac{1}{\sqrt{2}}(Y_{m1}-Y_{m2}) \end{matrix}$$

Maximum likelihood applied to the second column removes the difficulty with large numbers of nuisance parameters. One could act similarly for the normal theory linear model. The idea has been extended to *partial likelihood*. (Here one factors the likelihood, "keeps" factors depending heavily on parameters of interest, and "ignores" those factors depending heavily on nuisance parameters.)

IV. Role of Conditional Inference

1. *General idea*. Take a statistic C and consider the base problem conditional on C = c, i.e.

$$\Pr\{Y \epsilon w_\alpha(\psi_0) \,|\, C=c \,;\, \psi_0, \lambda\} = \alpha \ .$$

P-values are calculated in the conditional distribution. Why?

(i) Neyman structure;

(ii) Ancillarity as a matter of principle;

(iii) "Removal" of large numbers of nuisance parameters.

2. *Neyman-Pearson similarity*. If when $\psi = \psi_0$, $S(\psi_0)$ is minimal sufficient for λ and complete, i.e. $\dim(S(\psi_0)) = \dim(\Omega_\lambda)$, then the *only* way to achieve *exact* similarity is to take $C = S(\psi_0)$.

Example 8. Y_1 bin(θ_1, n_1) Y_2 bin(θ_2, n_2)

$$\psi = \log \frac{\theta_2}{1-\theta_2} - \log \frac{\theta_1}{1-\theta_1}$$

Condition on $Y_1 + Y_2 = y_1 + y_2$ (Fisher, 1935). Note: There has always been some controversy over whether to use continuity correction.

3. *Ancillarity.*

3.1. Definition. Consider a curved exponential family (p,q), $p > q$. Let $A = a(S)$ have same distribution as $\forall \theta \in \Omega_\theta$:

$$f_A(a;\theta) = f_A(a) \ .$$

Then A is called (simply) ancillary.

The principle of ancillarity is to base all calculations on $f_{S|A}(s|a\,;\theta)$. If $\dim(A) = p - q$, we have in effect reduced the problem to the full exponential family.

3.2. Examples.

Example 9. Let Y be $N(\theta,\sigma_{10}^2)$ with probability ½ and $N(\theta,\sigma_{00}^2)$ with probability ½, and let A = 1,0 respectively. Then $S = \begin{pmatrix} Y \\ A \end{pmatrix}$ and A is ancillary.

Example 10. N is a r.v. with known distribution on $\mathbb{Z}^+$. Given $N = n$, $Y_1,\ldots,Y_n$ are i.i.d. $N(\theta,1)$. Then $S = (\Sigma Y_i, N)$ and $A = N$.

Example 11. $Y_1,\ldots,Y_n$ i.i.d. rect$(\theta\,,\theta+1)$

$$S = (Y_{(\min)}, Y_{(\max)}) \text{ and } A = Y_{(\max)} - Y_{(\min)}$$

Example 12. $Y_1,\ldots,Y_n$ i.i.d. $N(\theta, K^2\theta^2)$, K is known.

$$S = (\Sigma Y_j\ ,\ \Sigma(Y_j-\bar{Y})^2) \quad \text{and} \quad A = \Sigma(Y_j-\bar{Y})^2/(\Sigma Y_j)^2$$

3.3. Applications. There may be a loss of power when conditioning is used (as in Examples 9 and 11). The argument for conditioning is to ensure applicability to the "unique case" under study.

3.4. Further points. The general idea of ancillarity raises various problems:

(i) Nonuniqueness. One can restrict attention to functions of S and choose the most selective (i.e. the one which sorts out the data "most sharply").

(ii) Nonexistence. There is a discontinuity on introduction of slight dependence (e.g. note that in Example 10 the distribution of the ancillary statistic depends slightly on θ). There is here a need for approximate theory.

(iii) Generalizations. One situation where we would like to generalize the previous concepts is where the sample size is a random variable whose distribution is unknown, but the parameters determining this distribution are independent of the parameters determining the distribution of the observed random variables. One way to generalize to account for this case is to let $S = (S', A)$, $\theta = (\psi,\lambda,y)$, and $\Omega_\theta = \Omega_{\psi,\lambda} \times \Omega_y$. Now we generalize simple ancillarity as follows: let

$$f_A(a;\theta) = f_A(a;y) \ ,$$

$$f_{S'|A}(s'|a\,;\theta) = f_{S'|A}(s'|a\,;\psi,\lambda) \ .$$

Then we may argue conditionally on A = a. This copes with the situation where the sample size has an unknown distribution. The conditional distribution (given A = a) depends only on ψ,λ. (Note: Ordinary linear regression is another example of generalized ancillarity.)

We have other generalizations of ancillarity, most of which are directed at a target problem, namely the comparison of two binomial parameters, an example we considered last time.

(iv) Implications for alternative asymptotics. Procedures using W and those using observed rather than expected second derivatives have some qualitative advantages. (They are in some sense more *conditional*.)

Example 9a. (This is an extension of Example 9.) Let $(Y_1,A_1),\ldots,(Y_n,A_n)$ be independent as in Example 9. Let

$$\hat{\theta} = (\Sigma Y_i/\sigma^2_{A_{i_0}})/(\Sigma 1/\sigma^2_{A_{i_0}}) \ .$$

Then we have

$$n\bar{i}(\theta,Y) = \Sigma(1/\sigma^2_{A_{i_0}}) \qquad \text{(observed information)}$$

whereas

$$n\bar{i}(\theta) = \tfrac{1}{2} n\left(\frac{1}{\sigma_{10}^2} + \frac{1}{\sigma_{00}^2}\right) \qquad \text{(expected information)} .$$

This is one situation where the observed information is preferable to the expected information. In the simple exponential family, the observed and expected informations are the same.

3.5. Location problems.

Example 13 (Fisher, 1934). Let $Y_1,\ldots,Y_n$ be i.i.d. random variables with density $g(y-\theta)$ where $g(\cdot)$ is a known density. Then it turns out that $S = (Y_{(1)},\ldots,Y_{(n)})$ (i.e. the order statistics). So we have an (n,1)-system. (We usually have no systematic procedure to construct ancillary statistics.) In this case, take $A_2 = Y_{(2)} - Y_{(1)},\ldots,A_n = Y_{(n)} - Y_{(n-1)}$. Let $Z = Y_{(1)}$. Now $f_Y(y;\theta) = g(y_1-\theta)\ldots g(y_n-\theta)$, which we may write as $lik(\theta;y)$ or $L(\theta\,;z,a)$, a function of $z-\theta$. Then $f_{Z,A}(z,a\,;\theta) = n!\,L(\theta\,;z,a)$. $f_A(a;\theta) = n!\int_{-\infty}^{\infty} L(\theta\,;z,a)dz = n!\int_{-\infty}^{\infty} L(\theta';z,a)d\theta'$ because $L(\theta\,;z,a)$ is a function of $z-\theta$, so that $f_{Z|A}(z|a\,;\theta) = L^+(\theta\,;z,a)$ where

$$L^+(\theta\,;z,a) = \frac{L(\theta\,;z,a)}{\int_{-\infty}^{\infty} d\theta'\, L(\theta\,;z,a)} .$$

Likelihood functions are usually determined to within a constant, but if we divide by an appropriate normalizing constant we obtain a probability density function. This is the context by which we mean that $L^+(\theta\,;z,a)$ is a normalized likelihood function. Given $\theta = \theta_0$ the upper α point of z is $\alpha = \int_z^{\infty} L^+(\theta_0\,;z',a)dz' = \int_{-\infty}^{\theta_0} L^+(\theta\,;z,a)d\theta$.

V. Asymptotic Expansions

1. *Edgeworth series, univariate.* Let $U_1,\ldots, U_n$ be i.i.d. random variables with density $f(\cdot)$, moment-generating function

$$M(\xi) = E \exp(-\xi U) ,$$

and cumulant-generating function

$$K(\xi) = \log M(\xi) .$$

Denote the cumulants by K_ℓ. We may standardize the cumulants to obtain

$$\rho_\ell = K_\ell / K_2^{\frac{1}{2}\ell} .$$

Let $R_n = U_1 +\ldots+ U_n$. We may normalize these last expressions by defining

$$X_n = (R_n - nK_1)/(K_2 n)^{\frac{1}{2}} .$$

The moment-generating function of X_n is

$$M_{X_n}(\xi) = \exp\left(\frac{nK_1 \xi}{(K_2 n)^{\frac{1}{2}}}\right)\left(M\left(\frac{\xi}{(K_2 n)^{\frac{1}{2}}}\right)\right)^n .$$

Denote the cumulant-generating function of X_n by K_{X_n}. Expand $K_{X_n}(\xi)$ in powers of $\frac{1}{\sqrt{n}}$. Obtain

$$M_{X_n}(\xi) = e^{-\frac{1}{2}\xi^2}\left(1 - \frac{\rho_3}{6\sqrt{n}}\xi^3 + \frac{\rho_4}{24n}\xi^4 + \frac{\rho_3^2}{72n}\xi^6 + O\left(\frac{1}{n^{3/2}}\right)\right) .$$

We invert this last expression to obtain

$$f_{X_n}(x) = (\exp(-\tfrac{1}{2}x^2))\frac{1}{\sqrt{2\pi}}\left(1 + \frac{\rho_3}{6\sqrt{n}}H_3(x) + \frac{\rho_4}{24n}H_4(x) + \frac{\rho_3^2}{72n}H_6(x) + O\left(\frac{1}{n^{3/2}}\right)\right)$$

where $H_n(\cdot)$ are the Hermite polynomials, which may be defined by

$$\frac{d^r}{dx^r}e^{-\frac{1}{2}x^2} = (-1)^r H_r(x)e^{-\frac{1}{2}x^2} .$$

There are several considerations. One is rigor, as to whether the expression is valid as an asymptotic expansion. (Also see Feller, 1971.) We may want to obtain an expansion for $F_{X_n}(x)$, the cumulative distribution function of X_n. Also we may need the Fisher-Cornish expansion.

We may want to obtain a similar expansion for functions other than sums of i.i.d. random variables. For example, often maximum likelihood estimators are almost (but not quite) sums of i.i.d. random variables.

2. *Bivariate Edgeworth.* Let $(U_1,V_1),\ldots,(U_n,V_n)$ be i.i.d. random vectors corresponding to moment-generating function $M(\xi,\eta)$ and cumulative generating function $K(\xi,\eta)$. And we have similar expressions $K_{\ell n},\rho_{\ell n}$. Let us employ some mathematical symbolism to simplify some of the expressions:

$$(H'\rho)^{[2]}(x,y) = \rho_{20}H_2(x) + 2\rho_{11}H_1(x)H_1(y) + \rho_{02}H_2(y)$$

$$(H'\rho)^{[3]}(x,y) = \rho_{30}H_3(x) + 3\rho_{21}H_2(x)H_1(y) + 3\rho_{12}H_1(x)H_2(y) + \rho_{03}H_3(y)$$

$$\vdots$$

etc.

$$((H'\rho)^{[2]})^2\ (x,y) = \rho_{20}^2 H_4(x) + \ldots + \rho_{02}^2 H_4(y)$$

$$((H'\rho)^{[3]})^2\ (x,y) = \rho_{30}^2 H_6(x) + \ldots + \rho_{03}^2 H_6(y)$$

$$\vdots$$

etc.

Let

$$R_n = U_1 + \ldots + U_n\ ,$$

$$S_n = V_1 + \ldots + V_n\ ,$$

$$X_n = \frac{R_n - nK_{10}}{\sqrt{nK_{20}}}\ ,$$

$$Y_n = \frac{S_n - nK_{01}}{\sqrt{nK_{02}}}\ .$$

If $\rho = 0$ (con(U,V) = 0), then

$$f_{X_n,Y_n}(x,y) = \frac{1}{2\pi}\exp(-\tfrac{1}{2}x^2 - \tfrac{1}{2}y^2)\left(1 + \frac{1}{6\sqrt{n}}(H'\rho)^{[3]}\ (x,y)\right.$$

$$+ \frac{1}{24n}(H'\rho)^{[4]}\ (x,y) + \frac{1}{72n}((H'\rho)^{[3]})^2\ (x,y)$$

$$\left. + O\left(\frac{1}{n^{3/2}}\right)\right)\ .$$

For general ρ (i.e. $\neq 0$), we may have to introduce new kinds of polynomials or consider the transformation $X_n, Y_n' = \frac{Y_n - \rho X_n}{\sqrt{1-\rho^2}}$. In this last case,

$$f_{Y_n|X_n}(y|x) = \frac{1}{\sqrt{2\pi(1-\rho^2)}}\exp(-\tfrac{1}{2}y'\,y)\left(1 + \frac{1}{6\sqrt{n}}((H'\rho')^{[3]}\ (x,y')\right.$$

$$\left. - \rho_{30}H_3(x)\right) + \ldots)\ .$$

We have studied the bivariate generalization. There are multivariate generalizations as well, but the algebra becomes unwieldy. Perhaps computers may be employed to handle some of the algebra.

3. *Indirect Edgeworth expansions* (also called the saddle-point method). Edgeworth expansions are the most direct and simple technique where large sample theory leads to a normal distribution and you want to ask the question: Can you do better? But there exist more powerful mathematical methods introduced by Daniels (1954), using complex-variable techniques, called the saddle-point method. Khinchin (1949) also uses a similar technique in the context of statistical mechanics. We can by this method obtain precision out of one term in the expansion that Edgeworth gets out of several terms.

Note that the Edgeworth expansion has an error term $O\left(\frac{1}{n^2}\right)$ at $y = 0$ even if only one correction term is used. Note that the Hermite polynomials vanish at the origin for odd order. This means that at $x = 0$ the expression behaves like $1/n$, but at the tail it behaves like $1/n^{\frac{1}{2}}$. The Edgeworth expansion works well in the center but bad at the tails. In this new technique we consider

$$f_{\chi_n}(x;\lambda_0) = e^{-u\lambda} f(u)/M(\lambda) = \exp(-u\lambda - K(\lambda))f(u) .$$

Then we have the identity

$$f_{\chi_n}(x;\lambda_0) = \exp(nK(\lambda) - nK(\lambda_0) + r(\lambda-\lambda_0)) \cdot f_{\chi_n}(x;\lambda)$$

where

$$x = (r-nK_{1(0)})/(K_{2(0)}n)^{\frac{1}{2}} .$$

The idea is to apply the Edgeworth expansion to the term $f_{\chi_n}(x;\lambda_0)$. We choose λ_0 so that the expansion occurs at the middle of the distribution. To approximate $f_{\chi_n}(x;\lambda_0)$ for given x and λ_0, choose λ so that x is at "origin" under λ. This leads to $\hat{\lambda}$, the maximum likelihood estimate of λ, based on considering x as an observation with density

$$f_{R_n}(r;\lambda_0) = \frac{\exp(nK(\hat{\lambda})-nK(\lambda_0)+r(\hat{\lambda}-\lambda_0))}{\sqrt{2\pi n\hat{K}_2}}\left(1+0\left(\frac{1}{n}\right)\right) .$$

We may also perform multivariate and conditional versions of this approximation.

4. *Higher-order likelihood theory.* In this section we examine some applications of the Edgeworth and indirect Edgeworth theory developed in the previous section. We have concentrated on the problem of choosing among several test statistics that are asymptotically equivalent, e.g. W, W_u, and W_e. Now we want to look at a different problem: Can we do better than just an asymptotic chi-square distribution for these statistics? That is, W, W_u, and W_e are asymptotically chi-square, but can we get a more refined approximation to their distributions, especially under the null hypothesis?

Bartlett (1937) touched upon this problem in his test of the homogeneity of variances. There is the Bartlett correction factor: If $d = \dim(\psi)$, suppose that

$$E(W;\psi=\psi_0) = \left(1+\frac{c}{n}\right)d + O\left(\frac{1}{n^2}\right) .$$

Put

$$W' = \frac{W}{1+\frac{c}{n}} .$$

In general if we apply this kind of technique, c may have to be estimated. This might seem bad from a mathematical point of view, but we usually want to approximate the tails of distributions, so if we are very careful (from a mathematical point of view) in our choice of the mean (i.e. in our choice of c) we may get bad tails. Lawley (1956) appeared to prove a remarkable result: If this correction term is used, then the approximation to the cumulative distribution is improved to the same order. That is, if the mean is fixed up, then all higher-order cumulants are fixed up to the same order. But Hayakawa (1977) showed that this is true only in certain cases. Hayakawa showed that if $g_d(\cdot)$ is the $\chi^2(d)$ function, then under H_0

$$f_W(W) = \left(1-\frac{a}{n}-\frac{b}{n}\right)g_d(W) + \frac{a}{n}\, g_{d+2}(W) + \frac{b}{n}\, g_{d+4}(W) + O\left(\frac{1}{n^{3/2}}\right) .$$

Then $b = 0$ if and only if W' has the density $g_d(\cdot)$ to order $O\left(\frac{1}{n}\right)$. If, in fact, the expansion had not had the term $g_{d+4}(\cdot)$, then Lawley would have been unconditionally correct.

In fact, $b = 0$ for

(i) Simple hypotheses (perhaps this was the case Lawley had in mind);

(ii) Hypotheses about canonical parameters in the exponential family.

The simplest proof of (i) is via the saddle-point method.

If $b = 0$, we may take the asymptotic chi-square approximation, divide by a suitable factor, and then we may obtain an asymptotic chi-square approximation to a higher order.

Totally unsolved problem: All these calculations are unconditional. If these calculations were appropriately conditioned, then would Hayakawa's result be true?

The central problem is that we have asymptotically equivalent statistics and we want to choose between them. The choice between W, W_u, W_e, etc. on the basis of adequacy of the chi-square approximation to null distributions is one of convenience rather than "principle." Power calculations tend to be inconclusive.

VI. Approximate Ancillarity

1. *General*. We shall assume

(i) Simple hypotheses;

(ii) Scalar parameter θ;

(iii) We do not have a full exponential family.

The problem is tractable if exact ancillarity can be used to effect reduction, as in the location problem. Otherwise one must look for approximate ancillarity (in some sense).

One approach (Hinkley, 1980; Barndorff-Nielsen, 1980) assumes that

$$\{\bar{i}(\hat{\theta};Y) - \bar{i}(\hat{\theta})\}/\text{st. dev.}_Y\ \bar{i}(\hat{\theta};Y)$$

is approximate ancillary. Alternatively, we can try to construct an approximate ancillary by arguing locally near $\theta = \theta_0$ where θ_0 is

arbitrary but fixed and known (cf. Example 3, page 5).

Assume $Y_1,\ldots,Y_n$ are independent. Use of the local approximation leads to the statistics

$$\overline{T}_1 = n^{-1} \Sigma \frac{\partial \log f_{Y_j}(Y_j;\theta_0)}{\partial\theta_0} = n^{-1} \Sigma U_j ,$$

$$\overline{T}_2 = n^{-1} \Sigma \frac{\partial^2 \log f_{Y_j}(Y_j;\theta_0)}{\partial\theta_0^2} = n^{-1} \Sigma V_j .$$

2. *Second-order local ancillarity.* Put $i_{\ell m} = E(U_j^\ell V_j^m ; \theta_0)$ (taking the i.i.d. case for simplicity) where the $i_{\ell m}, i_{\ell mp}, \ldots$ are generalizations of Fisher information with mixed moments. Then, subject to standard regularity conditions (for example those used in the derivation of the Cramér-Rao lower bound),

$$E\left(U_j^\ell V_j^m ; \theta_0 + \frac{\delta}{\sqrt{n}}\right) = i_{\ell m} + i_{\ell+1,m} \frac{\delta}{\sqrt{n}} + \tfrac{1}{2}\left(i_{\ell+2,m} + i_{\ell,m+1}\right)\frac{\delta^2}{n} + O\left(\frac{1}{n^{3/2}}\right) .$$

We first apply a linear transformation $(\overline{T}_1, \overline{T}_2) \to (\overline{S}_1, \overline{S}_2)$ so that

$$\operatorname{var}(\overline{S}_1;\theta_0) = \operatorname{var}(\overline{S}_2;\theta_0) = \frac{1}{n} ,$$

$$\operatorname{cov}(\overline{S}_1, \overline{S}_2 ; \theta_0) = 0 ,$$

$$E\left(\overline{S}_2 ; \theta_0 + \frac{\delta}{\sqrt{n}}\right) = \beta_{01} \frac{\delta^2}{n} + O\left(\frac{1}{n^{3/2}}\right) ,$$

$$E\left(\overline{S}_1 ; \theta_0 + \frac{\delta}{\sqrt{n}}\right) = \alpha_{10} \frac{\delta}{\sqrt{n}} + \beta_{10} \frac{\delta^2}{n} + O\left(\frac{1}{n^{3/2}}\right) .$$

We obtain

$$\overline{S}_1 = \frac{\overline{U}}{\sqrt{i_{20}}} \qquad \text{(normalized first derivative)}$$

$$\overline{S}_2 = \frac{\overline{V} - i_{01} - \overline{U}\,\frac{i_{11}}{i_{20}}}{\sigma_{2\cdot 1}} \qquad \text{(deviation of the second derivative from its linear regression on the first derivative suitably normalized)}$$

where $\sigma^2_{2\cdot 1} = i_{02} - i^2_{01} - \dfrac{i^2_{11}}{i_{20}}$.

3. *Distributional aspects.* Put $S_1 = \overline{S}_1\sqrt{n}$, $S_2 = \overline{S}_2\sqrt{n}$, and $A = \overline{A}\sqrt{n}$. Using the Edgeworth expansion for $f_{S_1,S_2}(s_1,s_2)$, and transforming from (S_1,S_2) to (S_1,A), we obtain, at $\theta = \theta_0$,

$$f_{S_1|A}(s_1|a) = \frac{e^{-\frac{1}{2}s_1^2}}{\sqrt{2\pi}}\left\{1 + \frac{K_{30}}{6\sqrt{n}}H_3(s_1) - \tfrac{1}{2}\frac{\sigma_{2\cdot 1}}{i_{20}\sqrt{n}}H_2(s_1)H_1(a) + O\left(\frac{1}{n}\right)\right\},$$

and thus a two-sided 2ε critical region for $\theta = \theta_0$ is

$$(*) \qquad \frac{\overline{U}(\theta_0)\sqrt{n}}{\sqrt{i_{20}(\theta_0)}} \notin I\left[\pm k^*_\varepsilon\left\{1 - \frac{\overline{V}(\theta_0) - i_{01}(\theta_0) - \overline{U}(\theta_0)\frac{i_{11}(\theta_0)}{i_{20}(\theta_0)}}{2i_{20}(\theta_0)}\right\} + \frac{1}{6}(k^{*2}_\varepsilon - 1)\frac{i_{30}(\theta_0)}{i^{3/2}_{20}(\theta_0)\sqrt{n}}\right]$$

where $\Phi(-k^*_\varepsilon) = \varepsilon$ and $I[\ ]$ denotes the appropriate interval. Hence we can obtain confidence intervals for θ as those values not rejected by formula (*). This confidence interval is equivalent to

$$\theta \pm k_\varepsilon^* \frac{1}{\sqrt{n}\sqrt{-\bar{V}(\hat{\theta})}} + \frac{i'_{20}(\hat{\theta})}{2ni_{20}^2(\hat{\theta})} + \frac{i_{001}(\hat{\theta})}{3ni_{20}^2(\hat{\theta})}$$

$$+ \frac{k_\varepsilon^{*2}}{6ni_{20}^2(\hat{\theta})} \{i_{001}(\hat{\theta}) - 3i'_{20}(\hat{\theta})\} .$$

Now consider

$$\bar{A} = \bar{S}_2 + c_{20}\bar{S}_1^2 + c_{11}\bar{S}_1\bar{S}_2$$

(where without loss of generality we omit the term $c_{02}\bar{S}_2^2$) and choose c_{20}, c_{11} to obtain a function even less sensitive than $\bar{S}_2$ to movement away from θ_0. In fact

$$E\left(\bar{A}\,;\,\theta_0 + \frac{\delta}{\sqrt{n}}\right) = \beta_{01}\frac{\delta^2}{n} + \frac{c_{20}(\alpha_{10}^2\delta^2 + 1)}{n} + O\left(\frac{1}{n^3}\right) ,$$

$$E\left(\bar{A}^2\,;\,\theta_0 + \frac{\delta}{\sqrt{n}}\right) = \frac{1}{n} + (\alpha_{02} + 2c_{11}\alpha_{10})\frac{1}{n^{3/2}} + O\left(\frac{1}{n^2}\right) ;$$

hence taking $c_{20} = -\frac{\beta_{01}}{\alpha_{10}^2}$, $c_{11} = -\frac{\alpha_{02}}{2\alpha_{10}}$ we obtain, at $\theta_0 + \frac{\delta}{\sqrt{n}}$,

$$E(\bar{A}) = -\frac{\beta_{10}}{n\alpha_{10}^2} + O\left(\frac{1}{n^{3/2}}\right) ,$$

$$\mathrm{var}(\bar{A}) = \frac{1}{n} + O\left(\frac{1}{n^2}\right) ,$$

$$\rho_3(\bar{A}) = \frac{K_{03}}{\sqrt{n}} + O\left(\frac{1}{n}\right)$$

independently of δ. We call such a statistic "second order locally ancillary."

If we reparameterize $\theta \to \phi$ to keep i constant we get

$$\hat{\phi} \pm \frac{k_\varepsilon^*}{\sqrt{-\ell''(\hat{\phi})}} + \frac{\ell'''(\hat{\phi})}{6\{\ell''(\hat{\phi})\}^2}(k_\varepsilon^{*2} + 2) \ .$$

Up to $O\left(\frac{1}{n}\right)$, this confidence interval is equivalent to that obtained by treating the normalized likelihood function for ϕ as a formal probability function and calculating the appropriate endpoints.

Summary (and open topics)

Ancillary statistics:
- conditions for existence
- philosophical and mathematical roles
- approximate ancillary statistics
- nuisance parameters

Asymptotic theory:
- higher-order problems (need more refined techniques)
- power
- nonstandard limiting operations
  - number of parameters $\to \infty$
  - number of observations $\to \infty$
  - large number of nuisance parameters

Nonindependent and nonregular cases:
- little done other than by Durbin (1980) and Akahira and Takeuchi (1980)

References

Aitchison, J. and Silvey, S.D. (1970). Maximum likelihood estimation procedures and associated tests of significance. *J. R. Statist. Soc. B 22*, 154-171.

Akahira, M. and Takeuchi, K. (1980). The concept of asymptotic efficiency and higher order asymptotic efficiency in statistical estimation theory. Preprint.

Anderson, E.B. (1973). *Conditional Inference and Models for Measuring*. Copenhagen: Mentalhygiejnsk Forlag.

Barndorff-Nielsen, O. (1978). *Information and Statistical Families*. Wiley.

Barndorff-Nielsen, O. (1980). On conditionality resolutions. *Biometrika*, to appear.

Barndorff-Nielsen, O. and Cox, D.R. (1979). Edgeworth and saddle-point approximations with statistical applications (with discussion). *J. R. Statist. Soc. B 41*, 279-312.

Bartlett, M.S. (1937). Properties of sufficiency and statistical tests. *Proc. Roy. Soc. A 160*, 268-282.

Cox, D.R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist. 29*, 357-372.

Cox, D.R. (1971). The choice between alternative ancillary statistics. *J. R. Statist. Soc. B 33*, 251-255.

Cox, D.R. (1975). Partial likelihood. *Biometrika 62*, 269-276.

Cox, D.R. (1975). Prediction intervals and empirical Bayes confidence intervals. In *Perspectives in Probability and Statistics*, ed. J. Gani, pp. 47-55. Academic Press.

Cox, D.R. (1980). Local ancillarity. *Biometrika*, to appear.

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman and Hall.

Daniels, H.E. (1954). Saddle point approximations in statistics. *Ann. Math. Statist. 25*, 631-650.

Durbin, J. (1969). Inferential aspects of the randomness of sample size in survey sampling. In *New Developments in Survey Sampling*, ed. N.L. Johnson and H. Smith, pp. 629-651. Wiley.

Durbin, J. (1980). Approximations for densities of sufficient estimators. *Biometrika*, to appear.

Efron, B. (1975). Defining the curvature of a statistical problem (with discussion). *Ann. Statist. 3*, 1189-1242.

Efron, B. and Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator (with discussion). *Biometrika 65*, 457-487.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications, Vol. 2.* 2nd ed. Wiley.

Fisher, R.A. (1934). Two new properties of mathematical likelihood. *Proc. Roy. Soc. A 144*, 285-307.

Fisher, R.A. (1935). The logic of inductive inference. *J. R. Statist. Soc. 98*, 39-54.

Fraser, D.A.S. (1968). *The Structure of Inference.* Wiley.

Ghosh, J.K. and Bhattacharya, R.N. (1978). On the validity of the formal Edgeworth expansion. *Ann. Statist. 28*, 861-881.

Hayakawa, T. (1977). The likelihood ratio criterion and the asymptotic expansion of its distribution. *Ann. Inst. Statist. Math. 29*, 359-378.

Hinkley, D.V. (1980). Likelihood as approximate pivotal distribution. *Biometrika*, to appear.

Hoeffding, W. (1965). Asymptotically optimal tests for multinomial distributions (with discussion). *Ann. Math. Statist. 36*, 369-408.

Huber, P.J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist. 1*, 799-821.

Jeffreys, H. (1968). *Theory of Probability.* 3rd ed. Oxford.

Johansen, S. (1979). Introduction to the theory of regular exponential families. Lecture notes, University of Copenhagen.

Khinchin, A.J. (1949). *Mathematical Foundations of Statistical Mechanics.* New York: Dover.

Lawley, D.N. (1956). A general method for approximating to the distribution of likelihood ratio criteria. *Biometrika 43*, 295-303.

Lehmann, E.L. (1959). *Testing Statistical Hypotheses.* Wiley.

Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. In *Probability and Statistics*, ed. U. Grenander, pp. 213-234. Almquist and Wiksall.

Neyman, J. and Scott, E.L. (1948). Consistent estimates based on partially consistent observations. *Econometrica 16*, 1-32.

Welch, B.L. (1939). On confidence limits and sufficiency, with particular reference to parameters of location. *Ann. Math. Statist. 10*, 58-69.

Welch, B.L. and Peers, H.W. (1963). On formulae for confidence points based on integrals of weighted likelihood. *J. R. Statist. Soc. B 25*, 318-329.