Chapter 10
Conditional Inference

10.1 Mixtures of Experiments

The present chapter has a somewhat different character from the preceding ones. It is concerned with problems regarding the proper choice and interpretation of tests and confidence procedures, problems which—despite a large literature—have not found a definitive solution. The discussion will thus be more tentative than in earlier chapters, and will focus on conceptual aspects more than on technical ones.

Consider the situation in which either the experiment $\mathcal{E}$ of observing a random quantity X with density p_θ (with respect to μ) or the experiment $\mathcal{F}$ of observing an X with density q_θ (with respect to ν) is performed with probability p and $q = 1 - p$, respectively. On the basis of X, and knowledge of which of the two experiments was performed, it is desired to test $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. For the sake of convenience it will be assumed that the two experiments have the same sample space and the same σ-field of measurable sets. The sample space of the overall experiment consists of the union of the sets

$$\mathcal{X}_0 = \{(I, x) : I = 0, \ x \in \mathcal{X}\} \quad \text{and} \quad \mathcal{X}_1 = \{(I, x) : I = 1, \ x \in \mathcal{X}\},$$

where I is 0 or 1 as $\mathcal{E}$ or $\mathcal{F}$ is performed.

A level-α test of H_0 is defined by its critical function

$$\phi_i(x) = \phi(i, x)$$

and must satisfy

$$p\,E_0\big[\phi_0(X) \mid \mathcal{E}\big] + q\,E_0\big[\phi_1(X) \mid \mathcal{F}\big] = \\ p \int \phi_0 p_{\theta_0} \, d\mu + q \int \phi_1 q_{\theta_0} \, d\nu \leq \alpha. \tag{10.1}$$

Suppose that p is unknown, so that H_0 is composite. Then a level-α test of H_0 satisfies (10.1) for all $0 < p < 1$, and must therefore satisfy

$$\alpha_0 = \int \phi_0 p_{\theta_0} \, d\mu \le \alpha \quad \text{and} \quad \alpha_1 = \int \phi_1 q_{\theta_0} \, d\nu \le \alpha. \tag{10.2}$$

As a result, a UMP test against H_1 exists and is given by

$$\phi_0(x) = \begin{cases} 1 & \text{when } p_{\theta_1}(x) > c_0 p_{\theta_0}(x), \\ \gamma_0 & \text{when } p_{\theta_1}(x) = c_0 p_{\theta_0}(x), \\ 0 & \text{when } p_{\theta_1}(x) < c_0 p_{\theta_0}(x), \end{cases} \quad \text{and} \quad \phi_1(x) = \begin{cases} 1 & \text{when } q_{\theta_1}(x) > c_1 q_{\theta_0}(x), \\ \gamma_1 & \text{when } q_{\theta_1}(x) = c_1 q_{\theta_0}(x), \\ 0 & \text{when } q_{\theta_1}(x) < c_1 q_{\theta_0}(x), \end{cases} \tag{10.3}$$

where the c_i and γ_i are determined by

$$E_{\theta_0}\big[\phi_0(X) \mid \mathcal{E}\big] = E_{\theta_0}\big[\phi_1(X) \mid \mathcal{F}\big] = \alpha. \tag{10.4}$$

The power of this test against H_1 is

$$\beta(p) = p\beta_0 + q\beta_1 \tag{10.5}$$

with

$$\beta_0 = E_{\theta_1}\big[\phi_0(X) \mid \mathcal{E}\big], \qquad \beta_1 = E_{\theta_1}\big[\phi_1(X) \mid \mathcal{F}\big]. \tag{10.6}$$

The situation is analogous to that of Section 4.4 and, as was discussed there, it may be more appropriate to consider the conditional power β_i when $I = i$, since this is the power pertaining to the experiment that has been performed. As in the earlier case, the conditional power β_I can also be interpreted as an estimate of the unknown $\beta(p)$, which is unbiased, since

$$E(\beta_I) = p\beta_0 + q\beta_1 = \beta(p).$$

So far, the probability p of performing experiment $\mathcal{E}$ has been assumed to be unknown. Suppose instead that the value of p is known, say $p = \frac{1}{2}$. The hypothesis H can be tested at level α by means of (10.3) as before, but the power of the test is now known to be $\frac{1}{2}(\beta_0 + \beta_1)$. Suppose that $\beta_0 = 0.3$, $\beta_1 = 0.9$, so that at the start of the experiment the power is $\frac{1}{2}(0.3 + 0.9) = 0.6$. Now a fair coin is tossed to decide whether to perform $\mathcal{E}$ (in case of heads) or $\mathcal{F}$ (in case of tails). If the coin shows heads, should the power be reassessed and scaled down to 0.3?

Let us postpone the answer and first consider another change resulting from the knowledge of p. A level-α test of H now no longer needs to satisfy (10.2) but only the weaker condition

$$\frac{1}{2}\left[\int \phi_0 p_{\theta_0} \, d\mu + \int \phi_1 q_{\theta_0} \, d\nu\right] \le \alpha. \tag{10.7}$$

The most powerful test against K is then again given by (10.3), but now with $c_0 = c_1 = c$ and $\gamma_0 = \gamma_1 = \gamma$ determined by (Problem 10.3)

$$\tfrac{1}{2}(\alpha_0 + \alpha_1) = \alpha, \tag{10.8}$$

where

$$\alpha_0 = E_{\theta_0}\big[\phi_0(X) \mid \mathcal{E}\big], \qquad \alpha_1 = E_{\theta_0}\big[\phi_1(X) \mid \mathcal{F}\big]. \tag{10.9}$$

As an illustration of the change, suppose that experiment $\mathcal{F}$ is reasonably informative, say that the power β_1 given by (10.6) is 0.8, but that $\mathcal{E}$ has little ability to distinguish between p_{θ_0} and p_{θ_1}. Then it will typically not pay to put much of the rejection probability into α_0; if β_0 [given by (10.6)] is sufficiently small, the best choice of α_0 and α_1 satisfying (10.8) is approximately $\alpha_0 \approx 0$, $\alpha_1 \approx 2\alpha$. The situation will be reversed if $\mathcal{F}$ is so informative that $\mathcal{F}$ can attain power close to 1 with an α_1 much smaller than $\alpha/2$.

When p is known, there are therefore two issues. Should the procedure be chosen which is best on the average over both experiments, or should the best conditional procedure be preferred; and, for a given test or confidence procedure, should probabilities such as level, power, and confidence coefficient be calculated conditionally, given the experiment that has been selected, or unconditionally? The underlying question is of course the same: Is a conditional or unconditional point of view more appropriate?

The answer cannot be found within the model but depends on the context. If the overall experiment will be performed many times, for example in an industrial or agricultural setting, the average performance may be the principal feature of interest, and an unconditional approach suitable. However, if repetitions refer to different clients, or are potential rather than actual, interest will focus on the particular event at hand, and conditioning seems more appropriate. Unfortunately, as will be seen in later sections, it is then often not clear how the conditioning events should be chosen.

The difference between the conditional and the unconditional approaches tends to be most striking, and a choice between them therefore most pressing, when the two experiments $\mathcal{E}$ and $\mathcal{F}$ differ sharply in the amount of information they contain, if for example the difference $|\beta_1 - \beta_0|$ in (10.6) is large. To illustrate an extreme situation in which this is not the case, suppose that $\mathcal{E}$ and $\mathcal{F}$ consist in observing X with distribution $N(\theta, 1)$ and $N(-\theta, 1)$ respectively, that one of them is selected with known probabilities p and q, respectively, and that it is desired to test $H : \theta = 0$ against $K : \theta > 0$. Here $\mathcal{E}$ and $\mathcal{F}$ contain exactly the same amount of information about θ. The unconditional most powerful level-α test of H against $\theta_1 > 0$ is seen to reject (Problem 10.5) when $X > c$ if $\mathcal{E}$ is performed, and when $X < -c$ if $\mathcal{F}$ is performed, where $P_0(X > c) = \alpha$. The test is UMP against $\theta > 0$, and happens to coincide with the UMP conditional test.

The issues raised here extend in an obvious way to mixtures of more than two experiments. As an illustration of a mixture over a continuum, consider a regression situation. Suppose that $X_1, \ldots, X_n$ are independent, and that the conditional density of X_i given t_i is

$$\frac{1}{\sigma} f\left(\frac{x_i - \alpha - \beta t_i}{\sigma}\right).$$

The t_i themselves are obtained with error. They may for example be independently normally distributed with mean c_i and known variance τ^2, where the c_i are the intended values of the t_i. Then it will again often be the case that the most appropriate inference concerning α, β, and σ is conditional on the observed values of the t's (which represent the experiment actually being performed). Whether this is the case will, as before, depend on the context.

The argument for conditioning also applies when the probabilities of performing the various experiments are unknown, say depend on a parameter ϑ, provided ϑ is unrelated to θ, so that which experiment is chosen provides no information concerning θ. A more precise statement of this generalization is given at the end of the next section.

10.2   Ancillary Statistics

Mixture models can be described in the following general terms. Let $\{\mathcal{E}_z, z \in \mathcal{Z}\}$ denote a collection of experiments of which one is selected according to a known probability distribution over $\mathcal{Z}$. For any given z, the experiment $\mathcal{E}_z$ consists in observing a random quantity X, which has a distribution $P_\theta(\cdot \mid z)$. Although this structure seems rather special, it is common to many statistical models.

Consider a general statistical model in which the observations X are distributed according to P_θ, $\theta \in \Omega$, and suppose there exists an *ancillary statistic*, that is, a statistic Z whose distribution F does not depend on θ. Then one can think of X as being obtained by a two-stage experiment: observe first a random quantity Z with distribution F; given $Z = z$, observe a quantity X with distribution $P_\theta(\cdot \mid z)$. The resulting X is distributed according to the original distribution P_θ. Under these circumstances, the argument of the preceding section suggests that it will frequently be appropriate to take the conditional point of view.[1] (Unless Z is discrete, these definitions involve technical difficulties concerning sets of measure zero and the existence of conditional distributions, which we shall disregard.)

An important class of models in which ancillary statistics exist is obtained by invariance considerations. Suppose the model $\mathcal{P} = \{P_\theta, \theta \in \Omega\}$ remains invariant under the transformations

$$X \to gX, \quad \theta \to \bar{g}\theta; \qquad g \in G, \quad \bar{g} \in \bar{G},$$

and that $\bar{G}$ is transitive over Ω.[2]

Theorem 10.2.1 *If $\mathcal{P}$ remains invariant under G and if $\bar{G}$ is transitive over Ω, then a maximal invariant T (and hence any invariant) is ancillary.*

---

[1] A distinction between experimental mixtures and the present situation, relying on aspects outside the model, is discussed by Basu (1964) and Kalbfleisch (1975).

[2] The family $\mathcal{P}$ is then a group family; see Lehmann and Casella (1998), Section 1.3.

PROOF. It follows from Theorem 6.3.2 that the distribution of a maximal invariant under G is invariant under $\bar{G}$. Since $\bar{G}$ is transitive, only constants are invariant under $\bar{G}$. The probability $P_\theta(T \in B)$ is therefore constant, independent of θ, for all B, as was to be proved. ∎

As an example, suppose that $X = (X_1, \ldots, X_n)$ is distributed according to a location family with joint density $f(x_1 - \theta, \ldots, x_n - \theta)$. The most powerful test of $H : \theta = \theta_0$ against $K : \theta = \theta_1 > \theta_0$ rejects when

$$\frac{f(x_1 - \theta_1, \ldots, x_n - \theta_1)}{f(x_1 - \theta_0, \ldots, x_n - \theta_0)} \geq c. \tag{10.10}$$

Here the set of differences $Y_i = X_i - X_n$ ($i = 1, \ldots, n-1$) is ancillary. This is obvious by inspection and follows from Theorem 10.2.1 in conjunction with Example 6.2.1(i). It may therefore be more appropriate to consider the testing problem conditionally given $Y_1 = y_1, \ldots, Y_{n-1} = y_{n-1}$. To determine the most powerful conditional test, transform to $Y_1, \ldots, Y_n$, where $Y_n = X_n$. The conditional density of Y_n given $y_1, \ldots, y_{n-1}$ is

$$p_\theta(y_n \mid y_1, \ldots, y_{n-1}) = \frac{f(y_1 + y_n - \theta, \ldots, y_{n-1} + y_n - \theta, y_n - \theta)}{\int f(y_1 + u, \ldots, y_{n-1} + u, u)\, du} \tag{10.11}$$

and the most powerful conditional test rejects when

$$\frac{p_{\theta_1}(y_n \mid y_1, \ldots, y_{n-1})}{p_{\theta_0}(y_n \mid y_1, \ldots, y_{n-1})} > c(y_1, \ldots, y_{n-1}). \tag{10.12}$$

In terms of the original variables this becomes

$$\frac{f(x_1 - \theta_1, \ldots, x_n - \theta_1)}{f(x_1 - \theta_0, \ldots, x_n - \theta_0)} > c(x_1 - x_n, \ldots, x_{n-1} - x_n). \tag{10.13}$$

The constant $c(x_1 - x_n, \ldots, x_{n-1} - x_n)$ is determined by the fact that the conditional probability of (10.13), given the differences of the x's, is equal to α when $\theta = \theta_0$.

For describing the conditional test (10.12) and calculating the critical value $c(y_1, \ldots, y_{n-1})$, it is useful to note that the statistic $Y_n = X_n$ could be replaced by any other Y_n satisfying the equivariance condition[3]

$$Y_n(x_1 + a, \ldots, x_n + a) = Y_n(x_1, \ldots, x_n) + a \quad \text{for all } a. \tag{10.14}$$

This condition is satisfied for example by the mean of the X's, the median, or any of the order statistics. As will be shown in Lemma 10.2.1, any two statistics Y_n and Y_n' satisfying (10.14) differ only by a function of the differences $Y_i = X_i - X_n$

---

[3] For a more detailed discussion of equivariance, see Lehmann and Casella (1998), Chapter 3.

($i = 1, \ldots, n - 1$). Thus conditionally, given the values $y_1, \ldots, y_{n-1}$, Y_n and Y_n' differ only by a constant, and their conditional distributions (and the critical values $c(y_1, \ldots, y_{n-1})$) differ by the same constant. One can therefore choose Y_n, subject to (10.14), to make the conditional calculations as convenient as possible.

Lemma 10.2.1 *If Y_n and Y_n' both satisfy (10.14), then their difference $\Delta = Y_n' - Y_n$ depends on $(x_1, \ldots, x_n)$ only through the differences $(x_1 - x_n, \ldots, x_{n-1} - x_n)$.*

PROOF. Since Y_n and Y_n' satisfy (10.14),

$$\Delta(x_1 + a, \ldots, x_n + a) = \Delta(x_1, \ldots, x_n) \qquad \text{for all } a.$$

Putting $a = -x_n$, one finds

$$\Delta(x_1, \ldots, x_n) = \Delta(x_1 - x_n, \ldots, x_{n-1} - x_n, 0),$$

which is a function of the differences. ∎

The existence of ancillary statistics is not confined to models that remain invariant under a transitive group $\bar{G}$. The mixture and regression examples of Section 10.1 provide illustrations of ancillaries without the benefit of invariance. Further examples are given in Problems 10.8–10.13.

If conditioning on an ancillary statistic is considered appropriate because it makes the inference more relevant to the situation at hand, it is desirable to carry the process as far as possible and hence to condition on a *maximal* ancillary. An ancillary Z is said to be maximal if there does not exist an ancillary U such that $Z = f(U)$ without Z and U being equivalent. [For a more detailed treatment, which takes account of the possibility of modifying statistics on sets of measure zero without changing their probabilistic properties, see Basu (1959).]

Conditioning, like sufficiency and invariance, leads to a reduction of the data. In the conditional model, the ancillary is no longer part of the random data but has become a constant. As a result, conditioning often leads to a great simplification of the inference. Choosing a maximal ancillary for conditioning thus has the additional advantage of providing the greatest reduction of the data.

Unfortunately, maximal ancillaries are not always unique, and one must then decide which maximal ancillary to choose for conditioning. [This problem is discussed by Cox (1971) and Becker and Gordon (1983).] If attention is restricted to ancillary statistics that are invariant under a given group G, the maximal ancillary of course coincides with the maximal invariant.

Another issue concerns the order in which to apply reduction by sufficiency and ancillarity.

Example 10.2.1 Let (X_i, Y_i), $i = 1, \ldots, n$, be independently distributed according to a bivariate normal distribution with $E(X_i) = E(Y_i) = 0$, $\text{Var}(X_i) = \text{Var}(Y_i) = 1$,

and unknown correlation coefficient ρ. Then $X_1, \ldots, X_n$ are independently distributed as $N(0, 1)$ and are therefore ancillary. The conditional density of the Y's given $X_1 = x_i, \ldots, X_n = x_n$ is

$$C \exp\left(-\frac{1}{2(1-\rho^2)} \sum (y_i - \rho x_i)^2\right),$$

with the sufficient statistics $(\sum Y_i^2, \sum x_i Y_i)$.

Alternatively, one could begin by noticing that $(Y_1, \ldots, Y_n)$ is ancillary. The conditional distribution of the X's given $Y_1 = y_1, \ldots, Y_n = y_n$ then admits the sufficient statistics $(\sum X_i^2, \sum X_i y_i)$. A unique maximal ancillary V does not exist in this case, since both the X's and Y's would have to be functions of V. Thus V would have to be equivalent to the full sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, which is not ancillary.

Suppose instead that the data are first reduced to the sufficient statistics $T = (\sum X_i^2 + \sum Y_i^2, \sum X_i Y_i)$. Based on T, no nonconstant ancillaries appear to exist.[4] This example and others like it suggest that it is desirable to reduce the data as far as possible through sufficiency, before attempting further reduction by means of ancillary statistics. ∎

Note that contrary to this suggestion, in the location example at the beginning of the section, the problem was not first reduced to the sufficient statistics $X_{(1)} < \cdots < X_{(n)}$. The omission can be justified in hindsight by the fact that the optimal conditional tests are the same whether or not the observations are first reduced to the order statistics.

In the structure described at the beginning of the section, the variable Z that labels the experiment was assumed to have a known distribution. The argument for conditioning on the observed value of Z does not depend on this assumption. It applies also when the distribution of Z depends on an unknown parameter ϑ, which is independent of θ and hence by itself contains no information about θ, that is, when the distribution of Z depends only on ϑ, the conditional distribution of X given $Z = z$ depends only on θ, and the parameter space Ω for (θ, ϑ) is a Cartesian product $\Omega = \Omega_1 \times \Omega_2$, with

$$(\theta, \vartheta) \in \Omega \quad \Leftrightarrow \quad \theta \in \Omega_1 \quad \text{and} \quad \vartheta \in \Omega_2 . \tag{10.15}$$

(the parameters θ and ϑ are then said to be *variation independent*, or unrelated.)

Statistics Z satisfying this more general definition are called *partial ancillary* or *S-ancillary*. (The term ancillary without modification will be reserved here for a statistic that has a known distribution.) Note that if $X = (T, Z)$ and Z is a partial ancillary, then T is a partial sufficient statistic in the sense of Problem 3.65. For a more detailed discussion of this and related concepts of partial ancillarity, see for example Basu (1978) and Barndorff-Nielsen (1978).

---

[4] So far, nonexistence has not been proved. It seems likely that a proof can be obtained by the methods of Unni (1978).

Example 10.2.2 Let X and Y be independent with Poisson distributions $P(\lambda)$ and $P(\mu)$, and let the parameter of interest be $\theta = \mu/\lambda$. It was seen in Section 10.4 that the conditional distribution of Y given $Z = X + Y = z$ is binomial $b(p, z)$ with $p = \mu/(\lambda + \mu) = \theta/(\theta + 1)$ and therefore depends only on θ, while the distribution of Z is Poisson with mean $\vartheta = \lambda + \mu$. Since the parameter space $0 < \lambda, \mu < \infty$ is equivalent to the Cartesian product of $0 < \theta < \infty$, $0 < \vartheta < \infty$, it follows that Z is S-ancillary for θ.

The UMP unbiased level-α test of $H : \mu \leq \lambda$ against $\mu > \lambda$ is UMP also among all tests whose conditional level given z is α for all z. (The class of conditional tests coincides exactly with the class of all tests that are similar on the boundary $\mu = \lambda$.) ∎

When Z is S-ancillary for θ in the presence of a nuisance parameter ϑ, the unconditional power $\beta(\theta, \vartheta)$ of a test φ of $H : \theta = \theta_0$ may depend on ϑ as well as on θ. The conditional power $\beta(\vartheta \mid z) = E_\theta[\varphi(X) \mid z]$ can then be viewed as an unbiased estimator of the (unknown) $\beta(\theta, \vartheta)$, as was discussed at the end of Section 4.4. On the other hand, if no nuisance parameters ϑ are present and Z is ancillary for θ, the unconditional power $\beta(\theta) = E_\theta \varphi(X)$ and the conditional power $\beta(\theta \mid z)$ provide two alternative evaluations of the power of φ against θ, which refer to different sampling frameworks, and of which the latter of course becomes available only after the data have been obtained.

Surprisingly, the S-ancillarity of $X + Y$ in Example 10.2.2 does not extend to the corresponding binomial problem.

Example 10.2.3 Let X and Y have independent binomial distributions $b(p_1, m)$ and $b(p_2, n)$ respectively. Then it was seen in Section 4.5 that the conditional distribution of Y given $Z = X + Y = z$ depends only on the cross-product ratio $\Delta = p_2 q_1 / p_1 q_2$ ($q_i = 1 - p_i$). However, Z is not S-ancillary for Δ. To see this, note that S-ancillarity of Z implies the existence of a parameter ϑ unrelated to Δ and such that the distribution of Z depends only on ϑ. As Δ changes, the family of distributions $\{P_\vartheta, \vartheta \in \Omega_2\}$ of Z would remain unchanged. This is not the case, since Z is binomial when $\Delta = 1$ and not otherwise (Problem 10.15). Thus Z is not S-ancillary.

In this example, all unbiased tests of $H : \Delta = \Delta_0$ have a conditional level given z that is independent of z, but conditioning on z cannot be justified by S-ancillarity. ∎

Closely related to this example is the situation of the multinomial 2×2 table discussed from the point of view of unbiasedness in Section 4.6.

Example 10.2.4 In the notation of Section 4.6, let the four cell entries of a 2×2 table be X, X', Y, Y' with row totals $X + X' = M$, $Y + Y' = N$, and column totals $X + Y = T$, $X' + Y' = T'$, and with total sample size $M + N = T + T' = s$. Here it is easy to check that (M, N) is S-ancillary for $\theta = (\theta_1, \theta_2) = (p_{AB}/p_B, p_{A\tilde{B}}/p_{\tilde{B}})$ with $\vartheta = p_B$. Since the cross-product ratio Δ can be expressed as a function of (θ_1, θ_2), it may be appropriate to condition a test of $H : \Delta = \Delta_0$ on (M, N). Exactly analogously one finds that (T, T') is S-ancillary for $\theta' = (\theta_1', \theta_2') = (p_{AB}/p_A, p_{\tilde{A}B}/p_{\tilde{A}})$,

and since Δ is also a function of (θ_1', θ_2'), it may be equally appropriate to condition a test of H on (T, T'). One might hope that the set of all four marginals $(M, N, T, T') = Z$ would be S-ancillary for Δ. However, it is seen from the preceding example that this is not the case.

Here, all unbiased tests have a constant conditional level given z. However, S-ancillarity permits conditioning on only one set of margins (without giving any guidance as to which of the two to choose), not on both. ■

Despite such difficulties, the principle of carrying out tests and confidence estimation conditionally on ancillaries or S-ancillaries frequently provides an attractive alternative to the corresponding unconditional procedures, primarily because it is more appropriate for the situation at hand. However, insistence on such conditioning leads to another difficulty, which is illustrated by the following example.

Example 10.2.5 Consider N populations $\prod_i$, and suppose that an observation X_i from $\prod_i$ has a normal distribution $N(\xi_i, 1)$. The hypothesis to be tested is $H : \xi_1 = \cdots = \xi_N$. Unfortunately, N is so large that it is not practicable to take an observation from each of the populations; the total sample size is restricted to be $n < N$. A sample $\prod_{J_1}, \ldots, \prod_{J_n}$ of n of the N populations is therefore selected at random, with probability $1/\binom{N}{n}$ for each set of n, and an observation X_{j_i} is obtained from each of the populations $\prod_{j_i}$, in the sample.

Here the variables $J_1, \ldots, J_n$ are ancillary, and the requirement of conditioning on ancillaries would restrict any inference to the n populations from which observations are taken. Systematic adherence to this requirement would therefore make it impossible to test the original hypothesis H.[5] Of course, rejection of the partial hypothesis $H_{j_1, \ldots, j_n} : \xi_{j_1} = \cdots = \xi_{j_n}$ would imply rejection of the original H. However, acceptance of $H_{j_1, \ldots, j_n}$ would permit no inference concerning H.

The requirement to condition in this case runs counter to the belief that a sample may permit inferences concerning the whole set of populations, which underlies much of statistical practice.

With an unconditional approach such an inference is provided by the test with rejection region

$$\sum \left[X_{j_i} - \left(\frac{1}{n} \sum_{k=1}^n X_{j_k} \right) \right]^2 \geq c,$$

where c is the upper α-percentage point of χ^2 with $n-1$ degrees of freedom. Not only does this test actually have unconditional level α, but its conditional level given $J_1 = j_1, \ldots, J_n = j_n$ also equals α for all $(j_1, \ldots, j_n)$. There is in fact no difference in the present case between the conditional and the unconditional tests: they will accept or reject for the same sample points. However, as has been pointed out, there is a crucial difference between the conditional and unconditional interpretations of the results.

---

[5] For other implications of this requirement, called the weak conditionality principle, see Birnbaum (1962) and Berger and Wolpert (1988).

If $\beta_{j_1,\ldots,j_n}(\xi_{j_1},\ldots,\xi_{j_n})$ denotes the conditional power of this test given $J_1 = j_1,\ldots,J_n = j_n$, its unconditional power is

$$\frac{\sum \beta_{j_1,\ldots,j_n}(\xi_{j_1},\ldots,\xi_{j_n})}{\binom{N}{n}}$$

summed over all $\binom{N}{n}$ n-tuples $j_1 < \ldots < j_n$. As in the case with any test, the conditional power given an ancillary (in the present case $J_1,\ldots,J_n$) can be viewed as an unbiased estimate of the unconditional power. ∎

10.3 Optimal Conditional Tests

Although conditional tests are often sensible and are beginning to be employed in practice [see for example Lawless (1972, 1973, 1978) and Kappenman (1975)], not much theory has been developed for the resulting conditional models. Since the conditional model tends to be simpler than the original unconditional one, the conditional point of view will frequently bring about a simplification of the theory. This possibility will be illustrated in the present section on some simple examples.

Example 10.3.1 Specializing the example discussed at the beginning of Section 10.1, suppose that a random variable is distributed according to $N(\theta, \sigma_1^2)$ or $N(\theta, \sigma_0^2)$ as $I = 1$ or 0, and that $P(I = 1) = P(I = 0) = \frac{1}{2}$. Then the most powerful test of $H : \theta = \theta_0$ against $\theta = \theta_1 (> \theta_0)$ based on (I, X) rejects when

$$\frac{x - \frac{1}{2}(\theta_0 + \theta_1)}{2\sigma_i^2} \geq k.$$

A UMP test against the alternatives $\theta > \theta_0$ therefore does not exist. On the other hand, if H is tested conditionally given $I = i$, a UMP conditional test exists and rejects when $X > c_i$ where $P(X > c_i \mid I = i) = \alpha$ for $i = 0, 1$. ∎

The nonexistence of UMP unconditional tests found in this example is typical for mixtures with known probabilities of two or more families with monotone likelihood ratio, despite the existence of UMP conditional tests in these cases.

Example 10.3.2 Let $X_1,\ldots,X_n$ be a sample from a normal distribution $N(\xi, a^2\xi^2)$, $\xi > 0$, with known coefficient of variation $a > 0$, and consider the problem of testing $H : \xi = \xi_0$ against $K : \xi > \xi_0$. Here $T = (T_1, T_2)$ with $T_1 = \bar{X}, T_2 = \sqrt{(1/n)\sum X_i^2}$ is sufficient, and $Z = T_1/T_2$ is ancillary. If we let $V = \sqrt{n}T_2/a$, the conditional density of V given $Z = z$ is equal to (Problem 10.18)

$$p_\xi(v \mid z) = \frac{k}{\xi^n}v^{n-1}\exp\left\{-\frac{1}{2}\left[\frac{v}{\xi} - \frac{z\sqrt{n}}{a}\right]^2\right\}. \tag{10.16}$$

The density has monotone likelihood ratio, so that the rejection region $V > C(z)$ constitutes a UMP conditional test.

Unconditionally, $Y = \bar{X}$ and $S^2 = \sum(X_i - \bar{X})^2$ are independent with joint density

$$cs^{(n-3)/2} \exp\left(-\frac{n}{2a^2\xi^2}(y-\xi)^2 - \frac{1}{2a^2\xi^2}s^2\right), \tag{10.17}$$

and a UMP test does not exist. [For further discussion of this example, see Hinkley (1977).] ∎

An important class of examples is obtained from situations in which the model remains invariant under a group of transformations that is transitive over the parameter space, that is, when the given class of distributions constitutes a group family. The maximal invariant V then provides a natural ancillary on which to condition, and an optimal conditional test may exist even when such a test does not exist unconditionally. Perhaps the simplest class of examples of this kind are provided by location families under the conditions of the following lemma.

Lemma 10.3.1 *Let $X_1, \ldots, X_n$ be independently distributed according to $f(x_i - \theta)$, with f strongly unimodal. Then the family of conditional densities of $Y_n = X_n$ given $Y_i = X_i - X_n$ $(i = 1, \ldots, n-1)$ has monotone likelihood ratio.*

PROOF. The conditional density (10.11) is proportional to

$$f(y_n + y_1 - \theta) \cdots f(y_n + y_{n-1} - \theta)f(y_n - \theta). \tag{10.18}$$

By taking logarithms and using the fact that each factor is strongly unimodal, it is seen that the product is also strongly unimodal, and the result follows from Example 8.2.1. ∎

Lemma 10.3.1 shows that for strongly unimodal f there exists a UMP conditional test of $H : \theta \le \theta_0$ against $K : \theta > \theta_0$ which rejects when

$$X_n > c(X_1 - X_n, \ldots, X_{n-1} - X_n). \tag{10.19}$$

Conditioning has reduced the model to a location family with sample size one. The double exponential and logistic distributions are both strongly unimodal (Section 9.2), and thus provide examples of UMP conditional tests. In neither case does there exist a UMP unconditional test unless $n = 1$.

As a last class of examples, we shall consider a situation with a nuisance parameter. Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent samples from location families with densities $f(x_1 - \xi, \ldots, x_m - \xi)$ and $g(y_1 - \eta, \ldots, y_n - \eta)$ respectively, and consider the problem of testing $H : \eta \le \xi$ against $K : \eta > \xi$. Here the differences $U_i = X_i - X_m$ and $V_j = Y_j - Y_n$ are ancillary. The conditional density of $X = X_m$ and $Y = Y_n$ given the u's and v's is seen from (10.18) to be of the form

$$f_u^*(x - \xi) g_v^*(y - \eta), \tag{10.20}$$

where the subscripts u and v indicate that f^* and g^* depend on the u's and v's respectively. The problem of testing H in the conditional model remains invariant under the transformations: $x' = x + c$, $y' = y + c$, for which $Y - X$ is maximal invariant. A UMP invariant conditional test will then exist provided the distribution of $Z = Y - X$, which depends only on $\Delta = \eta - \xi$, has monotone likelihood ratio. The following lemma shows that a sufficient condition for this to be the case is that f_u^* and g_v^* have monotone likelihood ratio in x and y respectively.

Lemma 10.3.2 *Let X, Y be independently distributed with densities $f^*(x - \xi)$, $g^*(y - \eta)$ respectively. If f^* and g^* have monotone likelihood with respect to ξ and η, then the family of densities of $Z = Y - X$ has monotone likelihood ratio with respect to $\Delta = \eta - \xi$.*

PROOF. The density of Z is

$$h_\Delta(z) = \int g^*(y - \Delta) f^*(y - z) \, dy. \tag{10.21}$$

To see that $h_\Delta(z)$ has monotone likelihood ratio, one must show that for any $\Delta < \Delta'$, $h_{\Delta'}(z)/h_\Delta(z)$ is an increasing function of z. For this purpose, write

$$\frac{h_{\Delta'}(z)}{h_\Delta(z)} = \int \frac{g^*(y - \Delta')}{g^*(y - \Delta)} \cdot \frac{g^*(y - \Delta) f^*(y - z)}{\int g^*(u - \Delta) f(u - z) \, du} \, dy.$$

The second factor is a probability density for Y,

$$p_z(y) = C_z g^*(y - \Delta) f^*(y - z), \tag{10.22}$$

which has monotone likelihood ratio in the parameter z by the assumption made about f^*. The ratio

$$\frac{h_{\Delta'}(z)}{h_\Delta(z)} = \int \frac{g^*(y - \Delta')}{g^*(y - \Delta)} p_z(y) \, dy \tag{10.23}$$

is the expectation of $g^*(Y - \Delta')/g^*(Y - \Delta)$ under the distribution $p_z(y)$. By the assumption about g^*, $g^*(y - \Delta')/g^*(y - \Delta)$ is an increasing function of y, and it follows from Lemma 3.4.2 that its expectation is an increasing function of z. ∎

It follows from (10.18) that $f_u^*(x - \xi)$ and $g_v^*(y - \eta)$ have monotone likelihood ratio provided this condition holds for $f(x - \xi)$ and $g(y - \eta)$, i.e., provided f and g are strongly unimodal. Under this assumption, the conditional distribution $h_\Delta(z)$ then has monotone likelihood ratio by Lemma 10.3.2, and a UMP conditional test exists and rejects for large values of Z. (This result also follows from Problem 8.12.)

The difference between conditional tests of the kind considered in this section and the corresponding (e.g., locally most powerful) unconditional tests typically

disappears as the sample size(s) tend(s) to infinity. Some results in this direction are given by Liang (1984); see also Barndorff-Nielsen (1983).

The following multivariate example provides one more illustration of a UMP conditional test when unconditionally no UMP test exists. The results will only be sketched. The details of this and related problems can be found in the original literature reviewed by Marden and Perlman (1980) and Marden (1983).

Example 10.3.3 Suppose you observe $m + 1$ independent normal vectors of dimension $p = p_1 + p_2$,

$$Y = (Y_1 \ \ Y_2) \quad \text{and} \quad Z_1, \dots, Z_m,$$

with common covariance matrix Σ and expectations

$$E(Y_1) = \eta_1, \qquad E(Y_2) = E(Z_1) = \cdots = E(Z_m) = 0.$$

(The normal multivariate two-sample problem with covariates can be reduced to this canonical form.) The hypothesis being tested is $H : \eta_1 = 0$. Without the restriction $E(Y_2) = 0$, the model would remain invariant under the group G of transformations: $Y^* = YB, Z^* = ZB$, where B is any nonsingular $p \times p$ matrix. However, the stated problem remains invariant only under the subgroup G' in which B is of the form (Problem 10.22(i))

$$B = \begin{pmatrix} B_{11} & 0 \\ B_{21} & B_{22} \end{pmatrix} \begin{matrix} p_1 \\ p_2 \end{matrix}.$$
$$\hphantom{B = (} {\scriptstyle p_1 \quad p_2}$$

If

$$Z'Z = S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

the maximal invariants under G' are the two statistics $D = Y_2 S_{22}^{-1} Y_2'$ and

$$N = \frac{(Y_1 - S_{12} S_{22}^{-1} Y_2)(S_{11} - S_{12} S_{22}^{-1} S_{21})^{-1}(Y_1 - S_{12} S_{22}^{-1} Y_2)'}{1 + D},$$

and the joint distribution of (N, D) depends only on the maximal invariant under G',

$$\Delta = \eta_1 (\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21})^{-1} \eta_1'.$$

The statistic D is ancillary (Problem 10.22(ii)), and the conditional distribution of N given $D = d$ is that of the ratio of two independent χ^2-variables: the numerator noncentral χ^2 with p degrees of freedom and noncentrality parameter $\Delta/(1 + d)$, and the denominator central χ^2 with $m + 1 - p$ degrees of freedom. It follows from Section 7.1 that the conditional density has monotone likelihood ratio. A conditionally UMP invariant test therefore exists, and rejects H when $(m + 1 - p)N/p > C$,

where C is the critical value of the F-distribution with p and $m + 1 - p$ degrees of freedom. On the other hand, a UMP invariant (unconditional) test does not exist; comparisons of the optimal conditional test with various competitors are provided by Marden and Perlman (1980). ∎

10.4   Relevant Subsets

The conditioning variables considered so far have been ancillary statistics, i.e., random variables whose distribution is fixed, independent of the parameters governing the distribution of X, or at least of the parameter of interest. We shall now examine briefly some implications of conditioning without this constraint. Throughout most of the section we shall be concerned with the simple case in which the conditioning variable is the indicator of some subset C of the sample space, so that there are only two conditioning events $I = 1$ (i.e., $X \in C$) and $I = 0$ (i.e., $X \in C^c$, the complement of C). The mixture problem at the beginning of Section 10.1, with $\mathcal{X}_1 = C$ and $\mathcal{X}_0 = C^c$, is of this type.

Suppose X is distributed with density p_θ, and R is a level-α rejection region for testing the simple hypothesis $H : \theta = \theta_0$ against some class of alternatives. For any subset C of the sample space, consider the conditional rejection probabilities

$$\alpha_C = P_{\theta_0}(X \in R \mid C) \quad \text{and} \quad \alpha_{C^c} = P_{\theta_0}(X \in R \mid C^c), \qquad (10.24)$$

and suppose that $\alpha_C > \alpha$ and $\alpha_{C^c} < \alpha$. Then we are in the difficulty described in Section 10.1. Before X was observed, the probability of falsely rejecting H was stated to be α. Now that X is known to have fallen into C (or C^c), should the original statement be adjusted and the higher value α_C (or lower value α_{C^c}) be quoted? An extreme case of this possibility occurs when C is a subset of R or R^c, since then $P(X \in R \mid X \in C) = 1$ or 0.

It is clearly always possible to choose C so that the conditional level α_C exceeds the stated α. It is not so clear whether the corresponding possibility always exists for the levels of a family of confidence sets for θ, since the inequality must now hold for all θ.

Definition 10.4.1   A subset C of the sample space is said to be a *negatively biased relevant subset* for a family of confidence sets $S(X)$ with unconditional confidence level $\gamma = 1 - \alpha$ if for some $\epsilon > 0$

$$\gamma_C(\theta) = P_\theta[\theta \in S(X) \mid X \in C] \leq \gamma - \epsilon \qquad \text{for all } \theta, \qquad (10.25)$$

and a *positively biased relevant subset* if

$$P_\theta[\theta \in S(X) \mid X \in C] \geq \gamma + \epsilon \qquad \text{for all } \theta. \qquad (10.26)$$

The set C is *semirelevant, negatively or positively biased*, if respectively

$$P_\theta[\theta \in S(X) \mid X \in C] \le \gamma \qquad \text{for all } \theta \tag{10.27}$$

or

$$P_\theta[\theta \in S(X) \mid X \in C] \ge \gamma \qquad \text{for all } \theta, \tag{10.28}$$

with strict inequality holding for at least some θ.

Obvious examples of relevant subsets are provided by the subsets $\mathcal{X}_0$ and $\mathcal{X}_1$ of the two-experiment example of Section 10.1.

Relevant subsets do not always exist. The following four examples illustrate the various possibilities.

Example 10.4.1 Let X be distributed as $N(\theta, 1)$, and consider the standard confidence intervals for θ:

$$S(X) = \{\theta : X - c < \theta < X + c\},$$

where $\Phi(c) - \Phi(-c) = \gamma$. In this case, there exists not even a semirelevant subset.

To see this, suppose first that a positively biased semirelevant subset C exists, so that

$$A(\theta) = P_\theta[X - c < \theta < X + c \text{ and } X \in C] - \gamma P_\theta[X \in C] \ge 0$$

for all θ, with strict inequality for some θ_0. Consider a prior normal density $\lambda(\theta)$ for θ with mean 0 and variance τ^2, and let

$$\beta(x) = P[x - c < \Theta < x + c \mid x],$$

where Θ has density $\lambda(\theta)$. The posterior distribution of Θ given x is then normal with mean $\tau^2 x/(1 + \tau^2)$ and variance $\tau^2/(1 + \tau^2)$ (Problem 10.24(i)), and it follows that

$$\beta(x) = \Phi\left[\frac{x}{\tau\sqrt{1 + \tau^2}} + \frac{c\sqrt{1 + \tau^2}}{\tau}\right] - \Phi\left[\frac{x}{\tau\sqrt{1 + \tau^2}} - \frac{c\sqrt{1 + \tau^2}}{\tau}\right]$$

$$\le \Phi\left[\frac{c\sqrt{1 + \tau^2}}{\tau}\right] - \Phi\left[\frac{-c\sqrt{1 + \tau^2}}{\tau}\right] \le \gamma + \frac{c}{\sqrt{2\pi\tau^2}}.$$

Next let $h(\theta) = \sqrt{2\pi}\tau\lambda(\theta) = e^{-\theta^2/2\tau^2}$ and

$$D = \int h(\theta)A(\theta)\, d\theta \le \sqrt{2\pi}\tau \int \lambda(\theta)\{P_\theta[X - c < \theta < X + c \text{ and } X \in C]$$

$$-E_\theta[\beta(X)I_C(X)]\}\, d\theta + \frac{c}{\tau}.$$

The integral on the right side is the difference of two integrals each of which equals $P[X - c < \Theta < X + c \text{ and } X \in C]$, and is therefore 0, so that $D \leq c/\tau$.

Consider now a sequence of normal priors $\lambda_m(\theta)$ with variances $\tau_m^2 \to \infty$, and the corresponding sequences $h_m(\theta)$ and D_m. Then $0 \leq D_m \leq c/\tau_m$ and hence $D_m \to 0$. On the other hand, D_m is of the form $D_m = \int_{-\infty}^{\infty} A(\theta) h_m(\theta) \, d\theta$, where $A(\theta)$ is continuous, nonnegative, and > 0 for some θ_0. There exists $\delta > 0$ such that $A(\theta) \leq \frac{1}{2} A(\theta_0)$ for $|\theta - \theta_0| < \delta$ and hence

$$D_m \geq \int_{\theta_0 - \delta}^{\theta_0 + \delta} \frac{1}{2} A(\theta_0) h_m(\theta) \, d\theta \to \delta A(\theta_0) > 0 \qquad \text{as} \quad m \to \infty.$$

This provides the desired contradiction. ∎

That also no negatively semirelevant subsets exist is a consequence of the following result.

Theorem 10.4.1 *Let $S(x)$ be a family of confidence sets for θ such that $P_\theta[\theta \in S(X)] = \gamma$ for all θ, and suppose that $0 < P_\theta(C) < 1$ for all θ.*

(i) If C is semirelevant, then its complement C^c is semirelevant with opposite bias.

(ii) If there exists a constant a such that

$$1 > P_\theta(C) > a > 0 \qquad \text{for all } \theta$$

and C is relevant, then C^c is relevant with opposite bias.
PROOF. The result is an immediate consequence of the identity

$$P_\theta(C)[\gamma_C(\theta) - \gamma] = [1 - P_\theta(C)][\gamma - \gamma_{C^c}(\theta)]. \quad \blacksquare$$

The next example illustrates the situation in which a semirelevant subset exists but no relevant one.

Example 10.4.2 Let X be $N(\theta, 1)$, and consider the uniformly most accurate lower confidence bounds $\underline{\theta} = X - c$ for θ, where $\Phi(c) = \gamma$. Here $S(X)$ is the interval $[X - c, \infty)$ and it seems plausible that the conditional probability of $\theta \in S(X)$ will be lowered for a set C of the form $X \geq k$. In fact

$$P_\theta(X - c \leq \theta \mid X \geq k) = \begin{cases} \frac{\Phi(c) - \Phi(k - \theta)}{1 - \Phi(k - \theta)} & \text{when} \quad \theta > k - c, \\ 0 & \text{when} \quad \theta < k - c. \end{cases} \qquad (10.29)$$

The probability (10.29) is always $< \gamma$, and tends to γ as $\theta \to \infty$. The set $X \geq k$ is therefore semirelevant negatively biased for the confidence sets $S(X)$.

We shall now show that no relevant subset C with $P_\theta(C) > 0$ exists in this case. It is enough to prove the result for negatively biased sets; the proof for positive bias

is exactly analogous. Let A be the set of x-values $-\infty < x < c + \theta$, and suppose that C is negatively biased and relevant, so that

$$P_\theta[X \in A \mid C] \leq \gamma - \epsilon \quad \text{for all } \theta.$$

If

$$a(\theta) = P_\theta(X \in C), \qquad b(\theta) = P_\theta(X \in A \cap C),$$

then

$$b(\theta) \leq (y - \epsilon)\, a(\theta) \quad \text{for all } \theta. \tag{10.30}$$

The result is proved by comparing the integrated coverage probabilities

$$A(R) = \int_{-R}^{R} a(\theta)\, d\theta, \qquad B(R) = \int_{-R}^{R} b(\theta)\, d\theta$$

with the Lebesgue measure of the intersection $C \cap (-R, R)$,

$$\mu(R) = \int_{-R}^{R} I_C(x)\, dx,$$

where $I_C(x)$ is the indicator of C, and showing that

$$\frac{A(R)}{\mu(R)} \to 1, \quad \frac{B(R)}{\mu(R)} \to \gamma \quad \text{as} \quad R \to \infty. \tag{10.31}$$

This contradicts the fact that by (10.30),

$$B(R) \leq (\gamma - \epsilon) A(R) \quad \text{for all } R,$$

and so proves the desired result.

To prove (10.31), suppose first that $\mu(\infty) < \infty$. Then if ϕ is the standard normal density

$$A(\infty) = \int_{-\infty}^{\infty} d\theta \int_C \phi(x - \theta)\, dx = \int_C dx = \mu(\infty),$$

and analogously $B(\infty) = \gamma\mu(\infty)$, which establishes (10.31).

When $\mu(\infty) = \infty$, (10.31) will be proved by showing that

$$A(R) = \mu(R) + K_1(R), \qquad B(R) = \gamma\mu(R) + K_2(R), \tag{10.32}$$

where $K_1(R)$ and $K_2(R)$ are bounded. To see (10.32), note that

$$\mu(R) = \int_{-R}^{R} I_C(x)\,dx = \int_{-R}^{R} I_C(x) \left[\int_{-\infty}^{\infty} \phi(x-\theta)\,d\theta \right] dx$$

$$= \int_{-\infty}^{\infty} \left[\int_{-R}^{R} I_C(x)\phi(x-\theta)\,dx \right] d\theta,$$

while

$$A(R) = \int_{-R}^{R} \left[\int_{-\infty}^{\infty} I_C(x)\phi(x-\theta)\,dx \right] d\theta. \tag{10.33}$$

A comparison of each of these double integrals with that over the region $-R < x < R, -R < \theta < R$, shows that the difference $A(R) - \mu(R)$ is made up of four integrals, each of which can be seen to be bounded by using the fact that $\int |t|\phi(t)\,dt < \infty$ (Problem 10.24(ii)). This completes the proof. ∎

Example 10.4.3 Let $X_1, \ldots, X_n$ be independently normally distributed as $N(\xi, \sigma^2)$, and consider the uniformly most accurate equivariant (and unbiased) confidence intervals for ξ given by (5.36).

It was shown by Buehler and Feddersen (1963) and Brown (1967) that in this case there exist positively biased relevant subsets of the form

$$C : \frac{|\bar{X}|}{S} \leq k. \tag{10.34}$$

In particular, for confidence level $\gamma = .5$ and $n = 2$, Brown shows that with $C : |\bar{X}|/|X_2 - X_1| \leq \frac{1}{2}(1 + \sqrt{2})$, the conditional level is $> \frac{2}{3}$ for all values of ξ and σ. Goutis and Casella (1992) provide detailed values for general n.

It follows from Theorem 10.4.1 that C^c is negatively biased semirelevant, and Buehler (1959) shows that any set $C^* : S \leq k$ has the same property. These results are intuitively plausible, since the length of the confidence intervals is proportional to S, and one would expect short intervals to cover the true value less often than long ones.

Theorem 10.4.1 does not show that C^c is negatively biased relevant, since the probability of the set (10.34) tends to zero as $\xi/\sigma \to \infty$. It was in fact proved by Robinson (1976) that no negatively biased relevant subset exists in this case.

The calculations for C^c throw some light on the common practice of stating confidence intervals for ξ only when a preliminary test of $H : \xi = 0$ rejects the hypothesis. For a discussion of this practice see Olshen (1973), and Meeks and D'Agostino (1983). ∎

The only type of example still missing is that of a negatively biased relevant subset. It was pointed out by Fisher (1956a, 1956b,1959,1973) that the Welch–Aspin solution of the Behrens–Fisher problem (discussed in Sections 6.6 and 13.2) provides an illustration of this possibility. The following are much simpler examples of both negatively and positively biased relevant subsets.

Example 10.4.4 An extreme form of both positively and negatively biased subsets was encountered in Section 7.7, where lower and upper confidence bounds $\underline{\Delta} < \Delta$ and $\Delta < \bar{\Delta}$ were obtained in (7.42) and (7.43) for the ratio $\Delta = \sigma_A^2/\sigma^2$ in a model II one-way classification. Since

$$P(\underline{\Delta} \leq \Delta \mid \underline{\Delta} < 0) = 1 \quad \text{and} \quad P(\Delta \leq \bar{\Delta} \mid \bar{\Delta} < 0) = 0,$$

the sets $C_1 : \underline{\Delta} < 0$ and $C_2 : \bar{\Delta} < 0$ are relevant subsets with positive and negative bias respectively. ∎

The existence of conditioning sets C for which the conditional coverage probability of level-γ confidence sets is 0 or 1, such as in Example 10.4.4 or Problems 10.27, 10.28 are an embarrassment to confidence theory, but fortunately they are rare. The significance of more general relevant subsets is less clear,[6] particularly when a number of such subsets are available. Especially awkward in this connection is the possibility [discussed by Buehler (1959)] of the existence of two relevant subsets C and C' with nonempty intersection and opposite bias.

If a conditional confidence level is to be cited for some relevant subset C, it seems appropriate to take account also of the possibility that X may fall into C^c and to state in advance the three confidence coefficients γ, γ_C, and γ_{C^c}. The (unknown) probabilities $P_\theta(C)$ and $P_\theta(C^c)$ should also be considered. These points have been stressed by Kiefer, who has also suggested the extension to a partition of the sample space into more than two sets. For an account of these ideas, see Kiefer (1977a, 1977b), Brownie and Kiefer (1977), and Brown (1978).

Kiefer's theory does not consider the choice of conditioning set or statistic. The same question arose in Section 10.2 with respect to conditioning on ancillaries. The problem is similar to that of the choice of model. The answer depends on the context and purpose of the analysis, and must be determined from case to case.

10.5 Problems

Section 10.1

Problem 10.1 Let the experiments of $\mathcal{E}$ and $\mathcal{F}$ consist in observing $X : N(\xi, \sigma_0^2)$ and $X : N(\xi, \sigma_1^2)$ respectively ($\sigma_0 < \sigma_1$), and let one of the two experiments be performed, with $P(\mathcal{E}) = P(\mathcal{F}) = \frac{1}{2}$. For testing $H : \xi = 0$ against $\xi = \xi_1$, determine values $\sigma_0, \sigma_1, \xi_1$, and α such that

$$\text{(i)} \quad \alpha_0 < \alpha_1; \quad \text{(ii)} \quad \alpha_0 > \alpha_1,$$

where the α_i are defined by (10.9).

---

[6] For a discussion of this issue, see Buehler (1959), Robinson (1976, 1979a), and Bondar (1977).

Problem 10.2 Under the assumptions of Problem 10.1, determine the most accurate invariant (under the transformation $X' = -X$) confidence sets $S(X)$ with

$$P(\xi \in S(X) \mid \mathcal{E}) + P(\xi \in S(X) \mid \mathcal{F}) = 2\gamma.$$

Find examples in which the conditional confidence coefficients γ_0 given $\mathcal{E}$ and γ_1 given $\mathcal{F}$ satisfy

(i) $\gamma_0 < \gamma_1$;   (ii) $\gamma_0 > \gamma_1$.

Problem 10.3 The test given by (10.3), (10.8), and (10.9) is most powerful under the stated assumptions.

Problem 10.4 Let $X_1, \ldots, X_n$ be independently distributed, each with probability p or q as $N(\xi, \sigma_0^2)$ or $N(\xi, \sigma_1^2)$.

(i) If p is unknown, determine the UMP unbiased test of $H : \xi = 0$ against $K : \xi > 0$.
(ii) Determine the most powerful test of H against the alternative ξ_1 when it is known that $p = \frac{1}{2}$, and show that a UMP unbiased test does not exist in this case.
(iii) Let α_k $(k = 0, \ldots, n)$ be the conditional level of the unconditional most powerful test of part (ii) given that k of the X's came from $N(\xi, \sigma_0^2)$ and $n - k$ from $N(\xi, \sigma_1^2)$. Investigate the possible values $\alpha_0, \alpha_1, \ldots, \alpha_n$.

Problem 10.5 With known probabilities p and q perform either $\mathcal{E}$ or $\mathcal{F}$, with X distributed as $N(\theta, 1)$ under $\mathcal{E}$ or $N(-\theta, 1)$ under $\mathcal{F}$. For testing $H : \theta = 0$ against $\theta > 0$ there exist a UMP unconditional and a UMP conditional level-α test. These coincide and do not depend on the value of p.

Problem 10.6 In the preceding problem, suppose that the densities of X under $\mathcal{E}$ and $\mathcal{F}$ are $\theta e^{-\theta x}$ and $(1/\theta)e^{-x/\theta}$ respectively. Compare the UMP conditional and unconditional tests of $H : \theta = 1$ against $K : \theta > 1$.

Section 10.2

Problem 10.7 Let X, Y be independently normally distributed as $N(\theta, 1)$, and let $V = Y - X$ and

$$W = \begin{cases} Y - X \text{ if } X + Y > 0, \\ X - Y \text{ if } X + Y \leq 0. \end{cases}$$

(i) Both V and W are ancillary, but neither is a function of the other.
(ii) (V, W) is not ancillary. [Basu (1959).]

Problem 10.8 An experiment with n observations $X_1, \ldots, X_n$ is planned, with each X_i distributed as $N(\theta, 1)$. However, some of the observations do not materialize (for

example, some of the subjects die, move away, or turn out to be unsuitable). Let $I_j = 1$ or 0 as X_j is observed or not, and suppose the I_j are independent of the X's and of each other and that $P(I_j = 1) = p$ for all j.

(i) If p is known, the effective sample size $M = \sum I_j$ is ancillary.
(ii) If p is unknown, there exists a UMP unbiased level-α test of $H : \theta \leq 0$ versus $K : \theta > 0$. Its conditional level (given $M = m$) is $\alpha_m = \alpha$ for all $m = 0, \ldots, n$.

Problem 10.9 Consider n tosses with a biased die, for which the probabilities of $1, \ldots, 6$ points are given by

1	2	3	4	5	6
$\frac{1-\theta}{12}$	$\frac{2-\theta}{12}$	$\frac{3-\theta}{12}$	$\frac{1+\theta}{12}$	$\frac{2+\theta}{12}$	$\frac{3+\theta}{12}$

and let X_i be the number of tosses showing i points.

(i) Show that the triple $Z_1 = X_1 + X_5, Z_2 = X_2 + X_4, Z_3 = X_3 + X_6$ is a maximal ancillary; determine its distribution and the distribution of $X_1, \ldots, X_6$ given $Z_1 = z_1, Z_2 = z_2, Z_3 = z_3$.
(ii) Exhibit five other maximal ancillaries. [Basu (1964).]

Problem 10.10 In the preceding problem, suppose the probabilities are given by

1	2	3	4	5	6
$\frac{1-\theta}{6}$	$\frac{1-2\theta}{6}$	$\frac{1-3\theta}{6}$	$\frac{1+\theta}{6}$	$\frac{1+2\theta}{6}$	$\frac{1+3\theta}{6}$

Exhibit two different maximal ancillaries.

Problem 10.11 Let X be uniformly distributed on $(\theta, \theta + 1), 0 < \theta < \infty$, let $[X]$ denote the largest integer $\leq X$, and let $V = X - [X]$.

(i) The statistic $V(X)$ is uniformly distributed on $(0, 1)$ and is therefore ancillary.
(ii) The marginal distribution of $[X]$ is given by

$$[X] = \begin{cases} [\theta] & \text{with probability } 1 - V(\theta), \\ [\theta] + 1 & \text{with probability } V(\theta). \end{cases}$$

(iii) Conditionally, given that $V = v$, $[X]$ assigns probability 1 to the value $[\theta]$ if $V(\theta) \leq v$ and to the value $[\theta] + 1$ if $V(\theta) > v$. [Basu (1964).]

Problem 10.12 Let X, Y have joint density

$$p(x, y) = 2f(x)f(y)F(\theta xy),$$

where f is a known probability density symmetric about 0, and F its cumulative distribution function. Then

(i) $p(x, y)$ is a probability density.
(ii) X and Y each have marginal density f and are therefore ancillary, but (X, Y) is not.
(iii) $X \cdot Y$ is a sufficient statistic for θ. [Dawid (1977).]

Problem 10.13 A sample of size n is drawn with replacement from a population consisting of N distinct unknown values $\{a_1, \ldots, a_N\}$. The number of distinct values in the sample is ancillary.

Problem 10.14 Assuming the distribution (4.22) of Section 4.9, show that Z is S-ancillary for $p = p_+/(p_+ + p_-)$.

Problem 10.15 In the situation of Example 10.2.3, $X + Y$ is binomial if and only if $\Delta = 1$.

Problem 10.16 In the situation of Example 10.2.2, the statistic Z remains S-ancillary when the parameter space is $\Omega = \{(\lambda, \mu) : \mu \leq \lambda\}$.

Problem 10.17 Suppose $X = (U, Z)$, the density of X factors into

$$p_{\theta, \vartheta}(x) = c(\theta, \vartheta)g_\theta(u; z)h_\vartheta(z)k(u, z),$$

and the parameters θ, ϑ are unrelated. To see that these assumptions are not enough to insure that Z is S-ancillary for θ, consider the joint density

$$C(\theta, \vartheta)e^{-\frac{1}{2}(u-\theta)^2 - \frac{1}{2}(z-\vartheta)^2} I(u, z),$$

where $I(u, z)$ is the indicator of the set $\{(u, z) : u \leq z\}$. [Basu (1978).]

Section 10.3

Problem 10.18 Verify the density (10.16) of Example 10.3.2.

Problem 10.19 Let the real-valued function f be defined on an open interval.

(i) If f is logconvex, it is convex.
(ii) If f is strongly unimodal, it is unimodal.

Problem 10.20 Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be positive, independent random variables distributed with densities $f(x/\sigma)$ and $g(y/\tau)$, respectively. If f and g have monotone likelihood ratios in (x, σ) and (y, τ), respectively, there exists a UMP conditional test of $H : \tau/\sigma \leq \Delta_0$ against $\tau/\sigma > \Delta_0$ given the ancillary statistics $U_i = X_i/X_m$ and $V_j = Y_j/Y_n$ $(i = 1, \ldots, m - 1; j = 1, \ldots, n - 1)$.

Problem 10.21 Let $V_1, \ldots, V_n$ be independently distributed as $N(0, 1)$, and given $V_1 = v_1, \ldots,$
$V_n = v_n$, let X_i $(i = 1, \ldots, n)$ be independently distributed as $N(\theta v_i, 1)$.

(i) There does not exist a UMP test of $H : \theta = 0$ against $K : \theta > 0$.
(ii) There does exist a UMP conditional test of H against K given the ancillary $(V_1, \ldots, V_n)$. [Buehler (1982).]

Problem 10.22 In Example 10.3.3,

 (i) the problem remains invariant under G' but not under G;
(ii) the statistic D is ancillary.

Section 10.4

Problem 10.23 In Example 10.4.1, check directly that the set $C = \{x : x \le -k$ or $x \ge k\}$ is not a negatively biased semirelevant subset for the confidence intervals $(X - c, X + c)$.

Problem 10.24 (i) Verify the posterior distribution of Θ given x claimed in Example 10.4.1.
(ii) Complete the proof of (10.32).

Problem 10.25 Let X be a random variable with cumulative distribution function F. If $E|X| < \infty$, then $\int_{-\infty}^{0} F(x)\, dx$ and $\int_{0}^{\infty} [1 - F(x)]\, dx$ are both finite. [Apply integration by parts to the two integrals.]

Problem 10.26 Let X have probability density $f(x - \theta)$, and suppose that $E|X| < \infty$. For the confidence intervals $X - c < \theta$ there exist semirelevant but no relevant subsets. [Buehler (1959).]

Problem 10.27 Let $X_1, \ldots, X_n$ be independently distributed according to the uniform distribution $U(\theta, \theta + 1)$.

 (i) Uniformly most accurate lower confidence bounds $\underline{\theta}$ for θ at confidence level $1 - \alpha$ exist and are given by

$$\underline{\theta} = \max(X_{(1)} - k, X_{(n)} - 1),$$

where $X_{(1)} = \min(X_1, \ldots, X_n)$, $X_{(n)} = \max(X_1, \ldots, X_n)$, and $(1 - k)^n = \alpha$.
 (ii) The set $C : x_{(n)} - x_{(1)} \ge 1 - k$ is a relevant subset with $P_\theta(\underline{\theta} \le \theta \mid C) = 1$ for all θ.
(iii) Determine the uniformly most accurate conditional lower confidence bounds $\underline{\theta}(v)$ given the ancillary statistic $V = X_{(n)} - X_{(1)} = v$, and compare them with $\underline{\theta}$. [The conditional distribution of $Y = X_{(1)}$ given $V = v$ is $U(\theta, \theta + 1 - v)$.]

[Pratt (1961a), Barnard (1976).]

Problem 10.28   (i) Under the assumptions of the preceding problem, the uniformly most accurate unbiased (or invariant) confidence intervals for θ at confidence level $1 - \alpha$ are

$$\underline{\theta} = \max(X_{(1)} + d, X_{(n)}) - 1 < \theta < \min(X_{(1)}, X_{(n)} - d) = \bar{\theta},$$

where d is the solution of the equation

$$2d^n = \alpha \quad \text{if} \quad \alpha < 1/2^{n-1},$$
$$2d^n - (2d - 1)^n = \alpha \quad \text{if} \quad \alpha > 1/2^{n-1}.$$

(ii) The sets $C_1 : X_{(n)} - X_{(1)} > d$ and $C_2 : X_{(n)} - X_{(1)} < 2d - 1$ are relevant subsets with coverage probability

$$P_\theta[\underline{\theta} < \theta < \bar{\theta} \mid C_1] = 1 \quad \text{and} \quad P_\theta[\underline{\theta} < \theta < \bar{\theta} \mid C_2] = 0.$$

(iii) Determine the uniformly most accurate unbiased (or invariant) conditional confidence intervals $\underline{\theta}(v) < \theta < \bar{\theta}(v)$ given $V = v$ at confidence level $1 - \alpha$, and compare $\underline{\theta}(v)$, $\bar{\theta}(v)$, and $\bar{\theta}(v) - \underline{\theta}(v)$ with the corresponding unconditional quantities.
   [Welch (1939), Pratt (1961a), Kiefer (1977a).]

Problem 10.29   Suppose X_1 and X_2 are i.i.d. with

$$P\{X_i = \theta - 1\} = P\{X_i = \theta + 1\} = \frac{1}{2}.$$

Let C be the confidence set consisting of the single point $(X_1 + X_2)/2$ if $X_1 \neq X_2$ and $X_1 - 1$ if $X_1 = X_2$. Show that, for all θ,

$$P_\theta\{\theta \in C\} = 0.75,$$

but

$$P_\theta\{\theta \in C | X_1 = X_2\} = 0.5$$

and

$$P_\theta\{\theta \in C | X_1 \neq X_2\} = 1.$$

[Berger and Wolpert (1988)]

Problem 10.30   Instead of conditioning the confidence sets $\theta \in S(X)$ on a set C, consider a randomized procedure which assigns to each point x a probability $\psi(x)$ and makes the confidence statement $\theta \in S(x)$ with probability $\psi(x)$ when x is observed.[7]

---

[7] Randomized and nonrandomized conditioning is interpreted in terms of betting strategies by Buehler (1959) and Pierce (1973).

(i) The randomized procedure can be represented by a nonrandomized conditioning set for the observations (X, U), where U is uniformly distributed on $(0, 1)$ and independent of X, by letting $C = \{(x, u) : u < \psi(x)\}$.

(ii) Extend the definition of relevant and semirelevant subsets to randomized conditioning (without the use of U).

(iii) Let $\theta \in S(X)$ be equivalent to the statement $X \in A(\theta)$. Show that ψ is positively biased semirelevant if and only if the random variables $\psi(X)$ and $I_{A(\theta)}(X)$ are positively correlated, where I_A denotes the indicator of the set A.

Problem 10.31 The nonexistence of (i) semirelevant subsets in Example 10.4.1 and (ii) relevant subsets in Example 10.4.2 extends to randomized conditioning procedures.

10.6 Notes

Conditioning on ancillary statistics was introduced by Fisher (1934a, 1935b, 1936).[8] The idea was emphasized in Fisher (1956b,1959,1973) and by Cox (1958), who motivated it in terms of mixtures of experiments providing different amounts of information. The consequences of adopting a general principle of conditioning in mixture situations were explored by Birnbaum (1962) and Durbin (1970). Following Fisher's suggestion (1934), Pitman (1938b) developed a theory of conditional tests and confidence intervals for location and scale parameters. For recent paradox concerning conditioning on an ancillary statistic, see Brown (1990) and Wang (1999).

The possibility of relevant subsets was pointed out by Fisher (1956a, 1956b,1959,1973) (who called them *recognizable*). Its implications (in terms of betting procedures) were developed by Buehler (1959), who in particular introduced the distinction between relevant and semirelevant, positively and negatively biased subsets, and proved the nonexistence of relevant subsets in location models. The role of relevant subsets in statistical inference, and their relationship to Bayes and admissibility properties, was discussed by Pierce (1973), Robinson (1976, 1979a, 1979b), Bondar (1977), and Casella (1988), among others.

Fisher (1956a, 1956b,1959,1973) introduced the idea of relevant subsets in the context of the Behrens–Fisher problem. As a criticism of the Welch–Aspin solution, he established the existence of negatively biased relevant subsets for that procedure. It was later shown by Robinson (1976) that no such subsets exist for Fisher's preferred solution, the so-called Behrens–Fisher intervals. This fact may be related to the conjecture [supported by substantial numerical evidence in Robinson (1976) but so far unproved] that the unconditional coverage probability of the Behrens–Fisher intervals always exceeds the nominal level. For a review of these issues, see Wallace (1980) and Robinson (1982).

---

[8] Fisher's contributions to this topic are discussed in Savage (1976, pp. 467–469).

Maatta and Casella (1987) examine the conditional properties of some confidence intervals for the variance in the one-sample normal problem. The conditional properties of some confidence sets for the multivariate normal mean, including confidence sets centered at James-Stein or shrinkage estimators, see Casella (1987) and George and Casella (1994). The conditional properties of the standard confidence sets in a normal linear model are studied in Hwang and Brown (1991).

In testing a simple hypothesis against a simple alternative, Berger et al. (1994) present a conditional frequentist methodology that agrees with a Bayesian approach.