Lecture 15

SURVEY OF ORTHODOX PRINCIPLES

Now I want to turn to a few other topics which come under the heading of clearing up various questions that were left dangling in previous lectures. We need to have an understanding of the terminology and the various concepts and principles of orthodox statistics in order to make comparisons and refer easily to the existing literature. We have already examined the principle of maximum likelihood in Lecture 9, and in the last two lectures we saw something of the orthodox principles for point estimation of parameters, and the orthodox approach to decision theory. This seems like as good a time as any to extend the list.

The methods to be described are now obsolete, in the sense that Bayesian methods either include them as special cases, or improve on them. Nevertheless, they exist, the literature is full of them, and they will continue to appear in the literature throughout our lifetimes, because many Statistics Departments are still teaching them to their students as if Bayesian methods didn't exist. So, we have no choice but to learn the terminology of orthodox statistics.

However, don't get the impression that there exists any definite monolithic "orthodox theory." In fact, orthodox statistics is a mish-mash of mutually contradictory ad hoc principles, and there are just as many--and just as bitter-controversies between different workers within the orthodox school as between orthodox and Bayesian advocates.

15.1. Sufficient Statistics.

Given a sampling distribution function $(x_1 \ldots x_n|\alpha)$ and a proposed estimator $\beta(x_1 \ldots x_n)$ of α, let us carry out a change of variables $(x_1 \ldots x_n) \to (y_1 \ldots y_n)$ such that $y_1 = \beta(x_1 \ldots x_n)$ and the jacobian $J = \partial(y_1 \ldots y_n)/\partial(x_1 \ldots x_n)$ is finite and not identically zero. Then the sampling distribution function of the y_i is

$$(y_1 \ldots y_n|\alpha) = (x_1 \ldots x_n|\alpha)\,|J|^{-1} \tag{15-1}$$

By our Rule 1, this can be factored:

$$(y_1 \ldots y_n|\alpha) = (\beta y_2 \ldots y_n|\alpha) = (\beta|\alpha)(y_2 \ldots y_n|\beta\alpha) \tag{15-2}$$

Suppose now that $(y_2 \ldots y_n|\beta\alpha)$ turns out to be independent of α. This is equivalent to saying that the original sampling distribution can be factored in the form

$$(x_1 \ldots x_n|\alpha) = g(x_1 \ldots x_n)\,(\beta|\alpha) \tag{15-3}$$

where $g(x_1 \ldots x_n) = (y_2 \ldots y_n|\beta)|J|$ can be expressed as a function of the x_i, not involving α. Therefore, if β is known, knowing the value of α would give us no more information about the sample. Conversely, it seems intuitively that if β is known, then knowledge of $(y_2 \ldots y_n)$ could give us no further information about α; i.e. <u>all</u> the information in the sample, that is relevant for inference about α, is contained summarized in the single function $\beta(x_1 \ldots x_n)$.

Let us check whether this is true. The ultimate criterion is, of course, whether the conjectured property can be derived from Bayes' theorem; i.e. whether the posterior distribution $(\alpha|x_1 \ldots x_n)$ depends on the sample only through the function $\beta(x_1 \ldots x_n)$.

This distribution has the form

$$(\alpha|x_1 \ldots x_n) = \frac{(x_1 \ldots x_n|\alpha)\ f(\alpha)}{\int (x_1 \ldots x_n|\alpha)\ f(\alpha)\ d\alpha} \tag{15-4}$$

where $f(\alpha)$ is a prior probability density.

Substituting (15-3) into this, we obtain

$$(\alpha|x_1 \ldots x_n) = \frac{(\beta|\alpha)\ g(x_1 \ldots x_n)\ f(\alpha)}{\int (\beta|\alpha)\ g(x_1 \ldots x_n)\ f(\alpha)\ d\alpha} \tag{15-5}$$

Since $g(x_1 \ldots x_n)$ does not depend on α, it cancels out, leaving us with

$$(\alpha|x_1 \ldots x_n) = \frac{(\beta|\alpha)\ f(\alpha)}{\int (\beta|\alpha)\ f(\alpha)\ d\alpha} \tag{15-6}$$

which says, as conjectured, that the posterior probability distribution of α depends only on the particular function $\beta(x_1 \ldots x_n)$ of the sample values. All other properties of the sample are irrelevant for inference about α. In this case, β is said to be a sufficient statistic for α, a terminology introduced by Fisher. More generally, any function $f(x_1 \ldots x_n)$ of the sample values is called a "statistic."

For example, let α be the unknown mean value of a gaussian distribution of known variance σ^2. Then

$$(x_1 \ldots x_n|\alpha) = A \exp\left[- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \alpha)^2 \right] \tag{15-7}$$

where A is a normalizing constant. Rearranging, we have

$$\begin{aligned}(x_1 \ldots x_n|\alpha) &= A \exp\left[- \frac{n}{2\sigma^2} (\overline{x^2} - 2\alpha\overline{x} + \alpha^2) \right] \\ &= A \exp\left[- \frac{ns^2}{2\sigma^2} \right] \exp\left[- \frac{n}{2\sigma^2} (\overline{x} - \alpha)^2 \right]\end{aligned} \tag{15-8}$$

where

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{15-9}$$

$$\overline{x^2} = \frac{1}{n} \sum_{i=1}^{n} x_i^2 \tag{15-10}$$

$$s^2 = \overline{x^2} - \overline{x}^2 \tag{15-11}$$

are the sample mean, mean square, and variance respectively.

Suppose we propose the sample mean as our estimator; i.e. we take

$$\beta(x_1 \ldots x_n) \equiv \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{15-12}$$

The sampling distribution of β is

$$(\beta|\alpha) = \int \cdots \int dx_1 \ldots dx_n \; \delta\left(\beta - \frac{1}{n}\sum x_i\right) (x_1 \ldots x_n|\alpha) \qquad (15\text{-}13)$$

To evaluate this, it is easier to take first its Fourier transform, or <u>characteristic function</u>:

$$\phi(k) \equiv \left\langle e^{ik\beta} \right\rangle = \int_{-\infty}^{\infty} (\beta|\alpha)\; e^{ik\beta}\, d\beta$$

$$= \int_{-\infty}^{\infty} dx_1 \cdots \int_{-\infty}^{\infty} dx_n \; e^{i\frac{k}{n}(x_1+\ldots+x_n)} \; (x_1 \ldots x_n|\alpha)$$

The integration is elementary, and we find

$$\phi(k) = \exp\left[\, ik\alpha - \frac{k^2\sigma^2}{n}\right] \qquad (15\text{-}14)$$

Then, inverting the Fourier integral, we have

$$(\beta|\alpha) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi(k)\; e^{-ik\beta}\, dk$$

$$= \sqrt{\frac{n}{2\pi\sigma^2}}\, \exp\left[\, -\frac{n}{2\sigma^2}\, (\beta - \alpha)^2\right] \qquad (15\text{-}15)$$

But, comparing with (14-8), we see that the factorization property (15-3) does hold for this estimator, and consequently β is a sufficient statistic for estimation of α.

Conversely, applying Bayes' theorem (14-4), we find

$$(\alpha|x_1 \ldots x_n) = \frac{\exp\left[\, -\frac{n}{2\sigma^2}\, (\alpha - \overline{x})^2\right] f(\alpha)}{\int \exp\left[\, -\frac{n}{2\sigma^2}\, (\alpha - \overline{x})^2\right] f(\alpha)\; d\alpha} \qquad (15\text{-}16)$$

which says again that the sample mean $\overline{x}$ is a sufficient statistic for estimation of α. The parameter α would be termed the "population mean" by the statistician. However, this underlying "population" is entirely fictitious in most real problems.

If the mean α and standard deviation σ are both unknown, we can apply Bayes' theorem to find their joint posterior probability density $(\alpha\sigma|x_1 \ldots x_n)$.

In this case we need the correct normalization constant for the sample distribution function:

$$(x_1 \ldots x_n | \alpha,\sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i - \alpha)^2}{2\sigma^2}\right\} \qquad (15\text{-}17)$$

Bayes' theorem then yields, with prior probability density $f(\alpha,\sigma)$:

$$(\alpha\sigma | x_1 \ldots x_n) = \frac{A\, f(\alpha,\sigma)}{\sigma^n} \exp\left\{-\frac{n}{2\sigma^2}\left[s^2 + (\alpha - \bar{x})^2\right]\right\} \qquad (15\text{-}18)$$

where A is a normalizing constant independent of α and σ. Since only the sample mean and variance $\bar{x}$, s^2 appear here, $\bar{x}$ and s^2 are <u>jointly sufficient</u> for α and σ, a fact that I mentioned briefly at the end of Lecture 6. In general (15-18) will show some correlation between α and σ; but if we just want the best estimates of each independently of the other, we get them from the <u>marginal distributions</u> obtained by integrating out the unwanted parameter:

$$(\alpha | x_1 \ldots x_n) = \int (\alpha\sigma | x_1 \ldots x_n)\, d\sigma \qquad (15\text{-}19)$$

$$(\sigma | x_1 \ldots x_n) = \int (\alpha\sigma | x_1 \ldots x_n)\, d\alpha \qquad (15\text{-}20)$$

Similarly, let $0 < \alpha < 1$, $0 \le x_i < \infty$, and consider the distribution

$$(x_1 \ldots x_n | \alpha) = A \prod_{i=1}^{n} x_i^p\, \alpha^{x_i} \qquad (15\text{-}21)$$

Since this factors:

$$(x_1 \ldots x_n | \alpha) = A\, (x_1 \ldots x_n)^p\, \alpha^{n\bar{x}} \qquad (15\text{-}22)$$

we find as before that the sample mean $\bar{x}$ is a sufficient statistic for α, and the best estimate of α, by the criterion of any loss function, will be some function $\beta(\bar{x})$ of the sample mean only.

Likewise, consider the rectangular distribution

$$(x_1 \ldots x_n | \alpha) = \prod_{i=1}^{n} f_\alpha(x_i) \qquad (15\text{-}23)$$

where

$$f_\alpha(x) \equiv \begin{cases} 0, & x < 0 \\ \alpha^{-1}, & 0 \le x \le \alpha \\ 0, & \alpha < x \end{cases} \tag{15-24}$$

Thus,

$$(x_1 \ldots x_n | \alpha) = \begin{cases} 0, & x_{min} < 0 \\ \alpha^{-n}, & 0 \le x_{min} \le x_{max} \le \alpha \\ 0, & \alpha < x_{max} \end{cases} \tag{15-25}$$

where x_{min}, x_{max} are the minimum and maximum observed sample values. The posterior distribution $(\alpha | x_1 \ldots x_n)$ depends on the x_i only through x_{max}, (and, of course, on the number n of observations). Consequently x_{max} is a sufficient statistic for estimation of α; or, in a little different terminology often found in the literature, x_{max} and n are <u>jointly sufficient</u>.

Evidently, the condition for existence of a sufficient statistic is that a single function $\gamma(x_1 \ldots x_n)$ of the sample values must exist such that $(x_1 \ldots x_n | \alpha)$ factors into the form

$$(x_1 \ldots x_n | \alpha) = g(x_1 \ldots x_n)\ h(\gamma, \alpha). \tag{15-26}$$

For the rectangular distribution, this is the case with $\gamma(x_1 \ldots x_n) = x_{max}$, $g(x_1 \ldots x_n) = 1$, and

$$h(\gamma, \alpha) = \begin{cases} \alpha^{-n}, & \alpha \ge \gamma \\ 0, & \alpha < \gamma \end{cases} \tag{15-27}$$

A sufficient statistic does not always exist. For example, the Cauchy distribution $(x_1 \ldots x_n | \alpha) = A \prod_{i=1}^{n} [1 + (x_i - \alpha)^2]^{-1}$ does not admit any factorization of the form (15-26), nor does the truncated exponential distribution $(x_1 \ldots x_n | \alpha) = A \exp[-\alpha(x_1 + \ldots + x_n)]$, $0 \le x_1 \ldots x_n \le \alpha$. But in the latter case $\bar{x}$ and x_{max} are jointly sufficient for α.

15.2. Efficient Estimates.

I have already pointed out [Eq. (13-38)] that the criterion of minimum α-expected loss does not in general lead to any specific "best" estimator $\beta(x_1 \ldots x_n)$, but it may do so in some special cases. We can now exhibit one such special case. Consider a quadratic loss function $L(\alpha,\beta) = (\beta - \alpha)^2$, and independent sampling so that

$$(x_1 \ldots x_n | \alpha) = f(x_1,\alpha)\ f(x_2,\alpha)\ \ldots\ f(x_n,\alpha) \tag{15-28}$$

An estimator β which minimizes the α-expected loss was called "efficient" by R. A. Fisher. In some of the later literature, however, the term "efficient" is taken to mean only that this condition is approached asymptotically, in the limit of large samples. This is the condition called "asymptotic efficiency" by Cramér (1946). A famous inequality associated with the names of Fréchet, Darmois, Rao, Cramér, and others, places a lower limit on the α-expected loss with any estimator $\beta(x_1 \ldots x_n)$:

$$\langle(\beta - \alpha)^2\rangle \geq \frac{\left(\frac{d\langle\beta\rangle}{d\alpha}\right)^2}{n\int\left(\frac{\partial \log f}{\partial\alpha}\right)^2 f(x,\alpha)\ dx} \tag{15-29}$$

with equality when and only when the following two conditions are met:

(1) β is a sufficient statistic for estimation of α, i.e.

$$(x_1 \ldots x_n | \alpha) = g(x_1 \ldots x_n)\ h(\beta,\alpha) \tag{15-30}$$

(2) the function $h(\beta,\alpha)$ satisfies

$$\frac{\partial \log h}{\partial\alpha} = k(\alpha)(\beta - \alpha) \tag{15-31}$$

for some function $k(\alpha)$. A simple proof of this theorem is given by Cramér (1946, Sec. 32.3). From (15-30) and (15-31) it is seen that the sampling distribution function must also satisfy

$$\frac{\partial \log (x_1 \ldots x_n | \alpha)}{\partial \alpha} = k(\alpha)(\beta - \alpha) \tag{15-32}$$

or, on integration, it must have the form

$$(x_1 \ldots x_n | \alpha) = \frac{m(x_1 \ldots x_n) \exp[-\lambda(\alpha)\ \beta(x_1 \ldots x_n)]}{Z(\lambda)} \tag{15-33}$$

where λ depends only on α, and

$$Z(\lambda) \equiv \int m(x_1 \ldots x_n) \exp[-\lambda \beta(x_1 \ldots x_n)]\ dx_1 \ldots dx_n \tag{15-34}$$

Since this is just the canonical distribution of statistical mechanics, we may restate the theorem as follows: The best estimator $\beta(x_1 \ldots x_n)$ by the criterion of minimum α-expected loss, which achieves equality in (15-29), exists when and only when the sampling distribution function has the canonical form with maximum entropy, relative to some weighting function $m(x_1 \ldots x_n)$, for a given expectation value $<\beta>$.

Thus, for example, the energy of a system at thermal equilibrium is always a sufficient and efficient statistic for estimation of the temperature of the heat-bath surrounding it, all other details of its state being irrelevant for that purpose.

We examined the notion of sufficiency in Lecture 14, from the standpoint of "information" in the sense of entropy, and saw in Eq. (14-14) the exact sense in which the colloquial term "information" is related to entropy. Although "sufficiency" was introduced by R. A. Fisher within the context of orthodox statistics, we saw in Eq. (15-6) that it is exactly derivable from Bayes' theorem. Therefore, it remains a valid and useful notion in Bayesian statistics; any problem of inference in which a single sufficient statistic exists, will be vastly simpler mathematically, and will lead to much shorter calculations in applications. Generally, in nontrivial real problems where a sufficient statistic does not exist, we will be driven to approximations

in reducing data.

The notion of efficiency, however, is not of any particular value in Bayesian statistics, because Bayes' theorem automatically gives us the best estimator by the criterion of any loss function. Thus the need to compare different estimators doesn't arise unless the equations are so complicated that we have to resort to approximations. But then it is the x-expected loss [as defined in Eq. (13-15)] rather than the α-expected loss that provides our criterion of good approximation.

Furthermore, the notion of efficiency doesn't really have any "objective" meaning, because it depends on the particular way you or I choose to define our parameters. For example, instead of the parameter α, there is no reason why we couldn't use just as well, the parameter $\gamma \equiv \alpha^2$, or $\delta \equiv \log \alpha$, etc., and of course any satisfactory statistical methods ought to lead us to the same final conclusions however we have defined our parameters. But the Fisher definition of efficiency is so parameter-dependent that if an efficient estimator of α exists, then an efficient estimator of α^2 does <u>not</u> exist! For these reasons, we will have no further use for the concept of efficiency.

15.3. Tests of Goodness of Fit.

Back in Lecture 7, when we discussed the application of Bayes' theorem to such problems as the validity of Newtonian celestial mechanics, we noted this: Bayes' theorem tells us that we cannot say how the observed facts affect the probability of some hypothesis H, until we state some specific alternatives against which H is to be tested. For example. suppose there are only two possible hypotheses, H and H', to be considered. Then, on any data D, we must always have $(H|D) + (H'|D) = 1$, and in terms of our logarithmic measure of plausibility in decibels, Bayes' theorem becomes

$$e(H|D) = e(H|X) + 10 \log_{10} \frac{(D|H)}{(D|H')} \qquad (15\text{-}35)$$

which we might describe in words by saying that, "data D supports hypothesis H relative to H', by 10 $\log_{10}$ $(D|H)/(D|H')$ decibels." The phrase relative to H' is essential here, since with some other alternative H", the change in evidence for H, $[e(H|D) - e(H|X)]$ might be entirely different; it does not make sense to ask how much the observed facts tend "in themselves" to support or refute H (except, of course, in the case where D is absolutely impossible on hypothesis H, so deductive reasoning can take over).

Now as long as we talk only in these generalities, our common sense readily assents to this. But if we consider specific problems, we may have some doubts. For example, in the particle counter problem of Lecture 8 we had a case (known source strength s and known counter efficiency a) where the probability of getting c counts in any one second is a Poisson distribution (8-5) with mean value $\bar{c} = sa$:

$$(c|s,a) = e^{-sa}\frac{(sa)^c}{c!} \tag{15-36}$$

Although it wasn't necessary for the problem we were considering then, we can still ask: what can we infer from this about the relative frequencies with which we would see c counts if we repeat the measurement in many different seconds, with the result $\{c_1 c_2 \ldots c_n\}$? If the probability of any particular event (say the event c = 12) is independently equal to

$$p = e^{-sa}\frac{(sa)^{12}}{12!} \tag{15-37}$$

at each trial, then the probability that the event will occur exactly r times in n trials is the binomial distribution

$$(r|n) = \binom{n}{r} p^r (1-p)^{n-r} \tag{15-38}$$

or, the probability that it will occur with frequency f = r/n, is

$$(f|n) = \frac{n!}{(nf)!\,(n-nf)!}\, p^{nf}(1-p)^{n-nf} \tag{15-39}$$

When n is very large, we can use the Stirling approximation (10-16) to get

$$L \equiv \frac{1}{n} \log (f/n)$$

$$\simeq - f \log f - (1-f) \log (1-f) + f \log p + (1-f) \log (1-p) \qquad (15\text{-}40)$$

Treating f as a continuous variable,

$$\frac{\partial L}{\partial f} = \log \frac{1-f}{f} - \log \frac{1-p}{p}$$

$$\frac{\partial^2 L}{\partial f^2} = - \frac{1}{f(1-f)}$$

So L reaches a maximum at $f = p$, and we have the Taylor series expansion about that point:

$$L(f) = L(p) - \frac{(f-p)^2}{2p(1-p)} + \ldots$$

Therefore, an approximation (which is actually much better than you might guess from this simple derivation) to (15-39) is

$$(f|n) \simeq (\text{const.}) \cdot \exp\left\{ - \frac{n(f-p)^2}{2p(1-p)} \right\} \qquad (15\text{-}41)$$

Thus the most likely frequency to be observed is numerically equal to the probability; and the (mean ± standard deviation) estimate of the frequency is

$$(f)_{\text{est}} = p \pm \sqrt{\frac{p(1-p)}{n}} \qquad (15\text{-}42)$$

Here is another connection between probability and frequency which common sense could have anticipated, except that it would hardly give us a quantitative interval of reasonable "error." The result (15-42) will be generalized to a wider class of probability models in the next two lectures.

In the long run, therefore, we expect that the actual frequencies of various counts will be distributed in a manner approximating the Poisson distribution (15-36). Now we can perform the experiment, and the experimental frequencies either will or will not be a reasonable approximation to the predicted values. If, by the time we have observed a few thousand counts, the observed frequencies are wildly different from a Poisson distribution,

our common sense will tell us that the theory which led to Poisson prediction must be wrong. Yet we have not said anything about any alternatives! Is our common sense wrong here, or is there some way we can reconcile the theory with common sense?

Let's look again at equation (15-35). No matter what H' is, we must have $(D|H') \geq 1$, and therefore a statement which is independent of any alternative hypotheses is

$$e(H|D) \geq e(H|X) + 10 \log_{10}(D|H) = e(H|X) - \psi_\infty \qquad (15\text{-}43)$$

where

$$\psi_\infty = -10 \log_{10}(D|H) \geq 0. \qquad (15\text{-}44)$$

Thus, <u>there is no possible alternative which data D could support, relative to H, by more than ψ_∞ decibels</u>.

This suggests a solution to our paradox: in judging the amount of agreement between theory and observations, the proper question to ask is not, "How well does data D support hypothesis H?" A much better question is, "Are there any alternatives H' which data D would support relative to H, and how much support is possible?" Probability theory can give no meaningful answer to the first question, but it can give a very definite answer to the second.

We might be tempted to conclude that the proper criterion of "goodness of fit" is simply ψ_∞, or what is the same thing, the probability $(D|H)$. This is not so, however, as the following argument shows. <u>After</u> we have obtained data D, it is always possible to invent a strange, "sure thing" hypothesis H_s, according to which D was inevitable: $(D|H_s) = 1$, and H_s will always be supported relative to H by exactly ψ_∞ decibels. Let us see what this implies. Suppose I toss a die N = 10,000 times, and record the result of each toss. Then, on the hypothesis H ≡ "the die is honest," each of the

6^N possible results has probability 6^{-N}, or $\psi_\infty = 10 \log_{10}(6^N) = 77{,}815$ decibels! No matter what I observe in the 10,000 tosses, there is always an hypothesis H_s that will be supported relative to H by this enormous amount. If, after performing this experiment, I continue to believe that the die is honest, it can be only because I considered the prior probability of H_s to be very much lower than minus 77,815 decibels. Otherwise, I am reasoning inconsistently.

This is, I think, all perfectly correct and we have to accept the conclusion. The prior probability of H_s was indeed much lower than 6^{-N}, simply because there were 6^N different "sure thing" hypotheses which were all on the same footing before I observed D. But it is obvious that in practice we don't want to bother with this kind of hypothesis; even though it is supported by the data more than any other, its prior probability is so low that we are not going to accept it anyway.

In practice we are not interested in comparing H to all conceivable alternatives, but only to all those in some restricted class Ω, consisting of hypotheses which we consider to be in some sense "reasonable" a priori. Let me give one example (by far the most common and useful one) of a test relative to such a restricted class of hypotheses.

We consider some experiment, which has r possible outcomes, $A_1, A_2, \ldots, A_r$. Define the quantities

$$x_n \equiv m, \text{ if } A_m \text{ is true on the n'th trial} \tag{15-45}$$

Thus each x_n can take on the values $x_n = 1, 2, \ldots, r$. If the experiment consists of tossing a die, then $r = 6$, and x_n is the number of spots up on the n'th toss. Suppose now we wish to take into account only the hypotheses belonging to the "Bernoulli class" B_r, in which the probabilities of the A_m on successive repetitions of the experiment are considered independent and stationary; thus, when H is in B_r, the probability, conditional on H, of any specific sequence $\{x_1 \ldots x_N\}$ of observations has the form

$$(x_1 \ldots x_N | H) = p_1^{n_1} \ldots p_r^{n_r} \tag{15-46}$$

where p_m is the probability of result A_m in any trial, and n_m is the number of times A_m was true in the sequence. Of course, $\sum_m n_m = N$. To every hypothesis in B_r there corresponds a set of numbers $\{p_1 \ldots p_r\}$ such that $p_m \geq 0$, $\sum p_m = 1$, and for our present purposes these numbers completely characterize the hypothesis. Conversely, every such set of numbers defines an hypothesis belonging to the Bernoulli class B_r.

Now let's note an important lemma, which we have used before to establish some properties of entropy. Using the fact that $\log x \geq (1 - x^{-1})$, with equality if and only if $x = 1$, we find at once that

$$\sum_{i=1}^{r} n_i \log \left(\frac{n_i}{Np_i} \right) \geq 0, \tag{15-47}$$

with equality if and only if $p_i = n_i/N$ for all i. This inequality is the same as

$$\log (x_1 \ldots x_N | H) \leq N \sum f_i \log f_i \tag{15-48}$$

where $f_i = n_i/N$ is the observed frequency of result A_i. The righthand side of (15-48) depends only on the observed sample, so if we consider various hypotheses H_1, H_2, ... in B_r in the light of this particular sample, the quantity (15-47) gives us a measure of how well the different hypotheses fit the data; the nearer to equality, the better the fit.

For convenience in numerical work, let's express the quantity (15-47) in decibel units:

$$\psi_B \equiv 10 \sum_{i=1}^{r} n_i \log_{10} \left(\frac{n_i}{Np_i} \right) \tag{15-49}$$

To see the exact significance of ψ_B, suppose we apply Bayes' theorem in the form of Equation (15-35). There are only two hypotheses, $H = \{p_1 \ldots p_r\}$, and

$H' = \{p_1', \ldots, p_r'\}$ to be considered. Let the values of (15-49) computed according to H and H' be ψ_B, ψ_B' respectively. Then Bayes' theorem reads

$$e(H|x_1 \ldots x_N) = e(H|X) + 10 \log_{10} \frac{(x_1 \ldots x_N|H)}{(x_1 \ldots x_N|H')}$$

$$= e(H|X) + \psi_B' - \psi_B \qquad (15\text{-}50)$$

Now we can always find an hypothesis H' in B_r, for which $p_i' = n_i/N$, and $\psi_B' = 0$; therefore ψ_B has the following meaning:

> Given an hypothesis H and the observed data $\{x_1 \ldots x_N\}$, compute ψ_B from (15-49). Then given any $D \leq \psi_B$, it is possible to find an alternative hypothesis H' in B_r such that the data will support H' relative to H by D decibels. There is no H' in B_r which is supported relative to H by more than ψ_B decibels.

Thus, ψ_B is exactly the appropriate measure of "goodness of fit" relative to the class of Bernoulli alternatives.

We can also interpret ψ_B in this manner: we may regard the observed results $\{x_1 \ldots x_N\}$ as a "message" consisting of N symbols chosen from an alphabet of r letters. On each repetition of the experiment, Nature transmits to us one more letter of the message. How much information is transmitted by this message, under the Bernoulli probability assignment with independence of successive symbols? Note that

$$\psi_B/N = 10 \sum_{i=1}^{r} f_i \log_{10}(f_i/p_i) \qquad (15\text{-}51)$$

with $f_i = n_i/N$. Thus, $(-\psi_B/N)$ is the entropy per symbol of the observed message distribution $\{f_1 \ldots f_r\}$ relative to the "expected distribution" $\{p_1 \ldots p_r\}$. This shows that the notion of entropy is, in a sense, "inherent" in probability theory. Independently of Shannon's theorem, entropy or some monotonic function of entropy will appear automatically in the equations of

anyone who is willing to use Bayes' theorem for hypothesis testing.

Historically, a slightly different test was introduced by Karl Pearson. We expect that, if hypothesis H is true, then n_i will be close to Np_i, in the sense that the difference $|n_i - Np_i|$ will grow with N only as $\sqrt{N}$. Call this "condition A." Using the expansion $\log x = (x-1) - (x-1)^2/2 + \ldots$, we easily find that

$$\sum_{i=1}^{r} n_i \log \frac{n_i}{Np_i} = \frac{1}{2} \sum_{i=1}^{r} \frac{(n_i - Np_i)^2}{Np_i} + O\left(\frac{1}{\sqrt{N}}\right) \qquad (15\text{-}52)$$

the quantity designated as $O(1/\sqrt{N})$ tending to zero as indicated <u>provided</u> that the observed sample does in fact satisfy condition A. The quantity

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - Np_i)^2}{Np_i} = N \sum_{i=1}^{r} \frac{(f_i - p_i)^2}{p_i} \qquad (15\text{-}53)$$

is thus very nearly proportional to ψ_B, if the sample frequencies are close to the expected values:

$$\psi_B = (10 \log_{10} e) \frac{1}{2} \chi^2 + O(1/\sqrt{N}) = 2.1715 \chi^2 + O(1/\sqrt{N}) \qquad (15\text{-}54)$$

Pearson suggested that the quantity χ^2 be used as a criterion of goodness of fit, and this has led to the "Chi-squared" test, one of the most used techniques of orthodox statistics. Before describing the test, let's examine first its theoretical basis and suitability as a criterion. Evidently, $\chi^2 \geq 0$, and $\chi^2 = 0$ only if the observed frequencies agree exactly with those expected if the hypothesis is true. So, larger values of χ^2 correspond in some way to greater deviations between prediction and observation, and too large a value of χ^2 should lead us to doubt the truth of the hypothesis. But these qualitative properties are possessed also by ψ--and by any number of other quantities we could define. We have seen the theoretical basis, and precise significance, of ψ; so we ask (noting the comments of Pratt and Bross, as quoted in the preface) whether there exists any "connected argument" leading to χ^2.

The results of a search for this connected argument are disappointing. Scanning a number of orthodox textbooks, we find that χ^2 is often introduced as a straight deux ex machina; but Cramér (1946) does attempt to prepare the way for the idea, in these words: "It will then be in conformity with the general principle of least squares to adopt as measure of deviation an expression of the form $\sum c_i(n_i/N - p_i)^2$ where the coefficients c_i may be chosen more or less arbitrarily. It was shown by K. Pearson that if we take $c_i = N/p_i$, we shall obtain a deviation measure with particularly simple properties." In other words, χ^2 is adopted, not because of any connected argument but because it has, in Pratt's words, "some pleasant properties."

We have seen that in some cases χ^2 is nearly a multiple of ψ and in such cases they will of course lead to essentially the same conclusions. But let's try to understand the quantitative difference in these criteria by a technique that I want to use a lot from now on, in comparing orthodox and Bayesian methods. As discussed in the preface, we often find a small quantitative difference between Bayesian and orthodox results, which would be of no consequence in most practical problems, and is so small that our common sense is unable to pass judgment on which result is preferable. But when this happens, we can understand the difference by "magnifying" it--by finding some extreme problem where the difference is so great that our common sense can tell us which theory is giving sensible results, and which is not.

As our first example of this magnification technique, let's compare ψ and χ^2 to see which is the more reasonable criterion of goodness of fit.

15.4. Comparison of ψ and Chi-squared.

A coin toss can give three different outcomes: (1) heads, (2) tails, (3) it may stand on edge. Suppose that Mr. A's knowledge of coins is such that he assigns probabilities $p_1 = p_2 = 0.499$, $p_3 = 0.002$ to these cases.

We are in communication with Mr. B on the planet Mars, who has never seen a coin and doesn't have the slightest idea what a coin is. So, when told that there are three possible outcomes at each trial, and nothing more, he can only assign equal probabilities, $p_1' = p_2' = p_3' = 1/3$.

Now we want to test Mr. A's hypothesis against Mr. B's by doing a "random" experiment. We toss the coin 29 times and observe the outcomes: $n_1 = n_2 = 14$; $n_3 = 1$. So, we have for the two hypotheses:

$$\psi_A = 10\left[28 \log_{10}\left(\frac{14}{29 \times .499}\right) + \log_{10}\left(\frac{1}{29 \times .002}\right)\right] = 8.34 \text{ db}$$

$$\psi_B = 10\left[28 \log_{10}\left(\frac{14 \times 3}{29}\right) + \log_{10}\left(\frac{3}{29}\right)\right] = 35.19 \text{ db.}$$

From this experiment the man on Mars thus learns that (a) there is another hypothesis about the coin that is 35.2 db better than his (35.2 db corresponds to odds of over 3,300:1) and so unless he can justify an extremely low prior probability for that alternative, he cannot reasonably adhere to his first theory. (b) Mr. A's hypothesis is better than his by some 26.8 db, and in fact is within about 8 db of the best hypothesis that could be made, under our assumption of independent Bernoulli trials B_3. Here the ψ-test tells us pretty much what our common sense does.

But suppose that the man on Mars knew only about "orthodox" statistical principles as usually taught; and therefore believed that χ^2 was the proper criterion of goodness of fit. He would find that

$$\chi_A^2 = 2\,\frac{(14 - 29 \times .499)^2}{29 \times .499} + \frac{(1 - 29 \times .002)^2}{29 \times .002} = 15.33$$

$$\chi_B^2 = 2\,\frac{(14 - 29 \times .333)^2}{29 \times .333} + \frac{(1 - 29 \times .333)^2}{29 \times .333} = 11.65$$

and he would report back delightedly: "My hypothesis, by the accepted statistical test, is shown to be slightly preferable to yours!"

I think that many persons trained to use χ^2 will find this comparison startling, and will immediately try to find the error in my numerical work

above. We have here still another fulfillment of our robot's prediction back in Lecture 4. The ψ criterion is exactly derivable from Bayes' theorem; therefore any criterion which is only an approximation to it must contain either an inconsistency or a qualitative violation of common sense, which can be exhibited by producing special cases.

We can learn an important lesson about the practical use of χ^2 by looking more closely at what is happening here. On hypothesis A, the "expected" number of heads or tails in 29 tosses was $Np_1 = 14.471$. The actual observed number must be an integer; and we supposed above that in each case it was the closest possible integer, namely 14. This certainly seems a mild assumption, not harmful to hypothesis A. Yet this small discrepancy between expected and observed sample numbers, in a sense the smallest it could possibly be, nevertheless had an enormous effect on χ^2. The spook lies entirely in the fact that ${\chi_A}^2$ turned out so much larger than seems reasonable; there is nothing surprising about the other numerical values. Evidently, it is the last term in ${\chi_A}^2$, which refers to the fact that the coin stood on edge once in 29 tosses, that is causing the trouble. On hypothesis A, the probability that this would happen exactly n times in 29 tosses is our binomial distribution

$$(n|N,p) = \binom{N}{n} p^n (1-p)^{N-n}$$

with N = 29, p = 0.002. From this, we find that the probability of seeing the coin on edge one or more times in 29 trials is about (1/18); i.e. the fact that we saw it even once is a bit unexpected, and constitutes some evidence against A, that contributes 8 db to the value of ψ_A. But this amount of evidence is certainly not overwhelming; if our travel guide tells us that London has fog, on the average, one day in 18, we are hardly astonished to see fog on the day we arrive.

It is the $(1/p_i)$ weighting factor in the summand of χ^2 that causes this anomaly. Because of it, the χ^2 criterion essentially concentrates its attention on the extremely unlikely possibilities, if the hypothesis contains them; and the slightest discrepancy between expected and observed sample numbers for the unlikely events severely penalizes the hypothesis. The ψ-test also contains this effect, but in a much milder form, the $(1/p_i)$ factor appearing only in the logarithm.

To see this effect more clearly, suppose now that the experiment had yielded the results $n_1 = 14$, $n_2 = 15$, $n_3 = 0$. Evidently, by either the χ^2 or ψ criterion, this ought to make hypothesis A look better, B worse, than in the first example. Repeating the calculations, we now find

$$\psi_A = 0.30 \text{ db} \qquad \chi_A^2 = 0.0925$$

$$\psi_B = 51.2 \text{ db} \qquad \chi_B^2 = 14.55$$

You see that by far the greatest relative change was in χ_A^2; both criteria now agree that hypothesis A is far superior to B.

This shows what can happen through uncritical use of χ^2. Professor Q believes in extrasensory perception, and undertakes to prove it to us poor benighted, intransigent doubters. So he plays card games. On the "null hypothesis" that only chance is operating, it is extremely unlikely that the subject will guess many cards correctly.

The first few hundred times he plays, the results are disappointing; but these are readily explained away on the ground that the subject is not in a "receptive" mood. [The literature of parapsychology abounds with wistful complaints about the difficulty of reproducing the phenomenon; indeed, just the kind of difficulty one would expect if the phenomenon did not exist!]

But one day providence smiles on Mr. Q; the subject comes through handsomely. Immediately he calls in the statisticians, the mathematicians,

the notary publics, and the newspaper reporters. An extremely improbable event has at last occurred; and χ^2 is enormous. Now he can publish the results and assert: "The validity of the data is certified by reputable, disinterested persons, the statistical analysis has been under the supervision of recognized statisticians, the calculations have been checked by competent mathematicians. By the accepted statistical test, the null hypothesis has been decisively rejected." And everything he has said is absolutely true!

Moral: For testing hypotheses involving moderately large probabilities, which agree moderately well with observation, it won't make much difference whether we use ψ or χ^2. But for testing hypotheses involving extremely unlikely events, we had better use ψ; or life might become too exciting for us.

Now let's describe briefly the Chi-squared test as done in practice. We have the so-called "null hypothesis" H to be tested, and no alternative is stated. The null hypothesis predicts certain relative frequencies $\{p_1 \ldots p_r\}$ and corresponding sample numbers $\{Np_1, \ldots, Np_r\}$ where N is the number of trials. We observe the actual sample numbers $\{n_1, \ldots, n_r\}$. If some of the n_i are very small, we group categories together so that each n_i is at least, say, five. For example, in a case with r = 6, if the observed sample numbers were {6, 11, 14, 7, 3, 2} we would group the last two categories together, making it equivalent to a problem with r' = 5 distinguishable outcomes per trial, with sample numbers {6, 11, 14, 7, 5}, and null hypothesis H' which predicts frequencies $\{p_1, p_2, p_3, p_4, p_5+p_6\}$.

We then calculate the observed value of χ^2:

$$\chi^2_{obs} = \sum_{i=1}^{r'} \frac{(n_i - Np_i)^2}{Np_i} \tag{15-55}$$

as our measure of deviation of observation from prediction. Evidently, it is very unlikely that we would find $\chi^2_{obs} = 0$ even if the hypothesis is true.

So, goes the orthodox reasoning, we should calculate the probability that χ^2 would have various values, given H', and reject H if the probability of a deviation as great or greater than χ^2_{obs} is sufficiently small; usually one takes 5 per cent as the threshold of rejection.

Now the n_i are integers, so χ^2 is capable of taking on only a discrete set of numerical values, at most $(N+r-1)!/N!(r-1)!$ different values, if the p_i are all different and incommensurable. Therefore, the exact χ^2 distribution is necessarily discrete and defined at only a finite number of points. However, for sufficiently large N, the number and density of points becomes so large that we may approximate the χ^2 distribution by a continuous one. The "pleasant property" referred to by Cramér and Pratt, is then the fact, at first glance surprising, that in the limit of large N, we obtain a universal distribution law: the probability that χ^2 lies in the interval $d(\chi^2)$ is

$$g(\chi^2)\, d(\chi^2) = \frac{\chi^{f-2}}{2^{f/2}\left(\frac{f-2}{2}\right)!} \exp\left\{-\frac{1}{2}\chi^2\right\} d(\chi^2) \qquad (15\text{-}56)$$

where f is called the "number of degrees of freedom" of the χ^2-distribution. If the null hypothesis H was completely specified (i.e. if it contained no variable parameters), then $f = r' - 1$, where r' is the number of categories used in the sum of (15-55). But if H contains unspecified parameters which must be estimated from the data, we take $f = r' - 1 - m$, where m is the number of parameters estimated.

We readily calculate the expectation and variance of χ^2 over this distribution: $\langle\chi^2\rangle = f$, $\mathrm{var}(\chi^2) = 2f$; so if we were given H but didn't have the data of the experiment, the (mean ± standard deviation) estimate of the χ^2 we expect to see, would be just

$$(\chi^2)_{est} = f \pm \sqrt{2f} \qquad (15\text{-}57)$$

The reason usually given for grouping categories for which the sample numbers

are small, is that the approximation (15-56) would otherwise be bad. But grouping inevitably throws away some of the relevant evidence of the sample, and there is never any reason to do this when using ψ.

The probability that we would see a deviation as great or greater than χ^2_{obs} is then

$$P(\chi^2_{obs}) = \int_{\chi^2_{obs}}^{\infty} g(\chi^2)\, d(\chi^2)$$

$$= \int_{q_{obs}}^{\infty} \frac{q^k}{k!} e^{-q}\, dq \qquad (15\text{-}58)$$

where $q \equiv \frac{1}{2}\chi^2$, $k \equiv (f-2)/2$. If $P(\chi^2_{obs}) < 0.05$, we reject the null hypothesis at the 5% "significance level" (sometimes called the 95% level). Tables of χ^2_{obs} for which P = 0.01, 0.05, 0.10, 0.50 for various numbers of degrees of freedom, are given in most orthodox textbooks and collections of statistical tables.

Note the traditional procedure here: we choose some basically arbitrary significance level first, then report only whether the null hypothesis was or was not rejected at this level. Evidently, this doesn't tell us very much about the real import of the data; if you tell me that the hypothesis was rejected at the 5% level, then I don't know from this whether it would have been rejected at the 2%, or 1%, level. If you tell me it was not rejected at the 5% level, then I don't know whether is would have been rejected at the 10%, or 20%, level. The orthodox statistician would tell us far more about what the data really indicates if he would report instead the significance level $P(\chi^2_{obs})$ at which the null hypothesis is just barely rejected; for then we know what the verdict would be at all levels. But, for reasons totally incomprehensible to me, orthodox practice never does this, on the Chi-squared or any other significance test. In fact, the orthodox χ^2 and other tables are so constructed that you can't report the conclusions in

this more informative way, because they give numerical values only at such widely separated values of the significance level that interpolation isn't possible.

So, let me show you how to find numerical values of $P(\chi^2_{obs})$ from (15-58) without using the Chi-squared tables. Writing $q = q_o + t$, we have

$$P = \int_{q_o}^{\infty} \frac{q^k}{k!} e^{-q}\, dq = \int_0^{\infty} \frac{(q_o+t)^k}{k!} e^{-(q_o+t)}\, dt$$

$$= \frac{1}{k!} \sum_{m=0}^{k} \binom{k}{m} \int_0^{\infty} q_o^{\,m}\, t^{k-m}\, e^{-(q_o+t)}\, dt$$

$$= \sum_{m=0}^{k} e^{-q_o} \frac{q_o^{\,m}}{m!} \qquad (15\text{-}59)$$

But this is just the cumulative Poisson distribution; i.e. the probability

$$(m \le k | q_o) = \sum_{m=0}^{k} (m | q_o)$$

that $m \le k$, if m has a Poisson distribution with mean value $\langle m \rangle = q_o$:

$$(m | q_o) = e^{-q_o} \frac{q_o^{\,m}}{m!} \qquad (15\text{-}60)$$

Numerical values of (15-59) for all values of q_o, k of usual interest are given in the graph of the cumulative Poisson distribution in Appendix C. Use of this will somewhat improve the value of the Chi-squared test.

But if you use the ψ-test instead, you don't need any tables or graphs at all. The evidential meaning of the sample is then described simply by the numerical value of ψ; and not by a further arbitrary constraint such as tail areas. Of course, the numerical value of ψ doesn't in itself tell you whether to reject the hypothesis (although we could, with just as much justification as in the Chi-squared test, prescribe some definite "level" at which to reject). From the Bayesian point of view, there is simply no use in "rejecting" any hypothesis unless we can replace it with a definite alternative

known to be better; and whether this is justified must obviously depend not only on ψ, but also on the prior probability of the alternative (recall our quotation from E. L. Lehmann on p. 90), and on the consequences of making wrong decisions.

In spite of the difference in viewpoints, there is often not much difference in the actual conclusions reached. For example, as the number of degrees of freedom f increases, the orthodox statistician will accept a higher value of χ^2 [roughly proportional to f, as (15-57) indicates] before rejecting the hypothesis, on the grounds that such a high value is quite likely to occur if the hypothesis is true; but the Bayesian who will reject it only in favor of a definite alternative, must also accept a proportionally higher value of ψ, because the number of reasonable alternatives is increasing exponentially with f, and the prior probability of any one of them is correspondingly decreasing. So, in either case we reject the hypothesis if ψ or χ^2 exceeds some limit, with an enormous difference in the philosophy of how we choose that limit, but not necessarily a big difference in its actual location.

Although the point isn't made in the orthodox literature which just doesn't mention alternatives at all, we see from the above that χ^2 is <u>not</u> a measure of goodness of fit relative to all conceivable alternatives; but only relative to those in the same Bernoulli class. More generally, given any well-defined class C of alternatives, if we can write Bayes' theorem (describing the effect of new data D on the plausibility of two hypotheses H_1, H_2) in the form

$$e(H_1|DX) - e(H_1|X) = \psi_1 - \psi_2$$

where ψ_i depends only on the sample and H_i, is non-negative over C, and vanishes for some H_i in C, then we have constructed the appropriate ψ which

measures goodness of fit relative to the class of alternatives C.

In a recent article, Anscombe (1963) holds it to be a weakness of the Bayesian method that we had to introduce a specific class of alternatives. It seems to me, however, that it is entirely meaningless to speak of "goodness of fit" without reference to definite alternatives. For example, if you ask a scientist, "How well did the Zilch experiment fit the Bong theory?" you may get this reply: "Well, if you had asked me last week, I would have said it fits the Bong theory very handsomely; the experimental points lie much closer to Bong's curve than to the old Smith theory curve. But just yesterday I learned that this fellow Jones has worked out a new theory based on entirely different assumptions; and his curve goes right through the experimental points. So, now I'm afraid I have to say that the Zilch experiment pretty well demolishes the Bong theory."

Whether given data support or refute an hypothesis depends entirely on which alternatives we have in mind; if we fail to specify any alternatives we cannot hope to get a meaningful significance test, because we have not asked a well-posed question. The question when we should seek new alternatives must involve our knowledge about the "mechanism" being studied, and the line of reasoning which led to formulation of the null hypothesis in the first place; it cannot be answered merely from examining the null hypothesis and the sample. I would hold it to be a great merit of the Bayesian approach that it forces us to recognize these things, which have apparently not been obvious to statisticians (although qualitatively they are part of the elementary common sense which any scientist uses constantly in judging his theories).

This is a good example of what, I suggest, is the general situation; the Bayesian approach to statistics supplies the missing theoretical basis for, and often improvements on, orthodox methods which had long been, just as Pratt says, "ad hoc procedures with some pleasant properties."

15.5. An Acceptance Test.

Here is another very interesting example of a useful significance test. The probability that a certain machine will operate without failure for a time t is, by hypothesis, $\exp(-\lambda t)$. We test n units for a time t, and observe r failures; what assurance do we then have that the mean life $\theta = \lambda^{-1}$ exceeds a preassigned value θ_o? Let us examine the orthodox solution based on the same kind of philosophy that we just saw in the Chi-squared test (i.e. it is taboo to speak of the probability that θ has various values, because θ isn't a "random variable"; so we can use only the probability of getting various sample values, or the probability distribution of some "statistic"); and also give the Bayesian solution.

Sobel and Tischendorf (1959) (hereafter denoted ST) give an orthodox solution with tables that are reproduced in Roberts (1963). The test is to have a critical number C (i.e. we accept only if $r \leq C$). On the hypothesis that we have the maximum tolerable failure rate, $\lambda_o = \theta_o^{-1}$, the probability that we shall see r or fewer failures is the binomial sum

$$W(n,r) = \sum_{k=0}^{r} \binom{n}{k} e^{-(n-k)\lambda_o t} \left(1 - e^{-\lambda_o t}\right)^k \qquad (15\text{-}61)$$

and so, setting $W(n,C) \leq 1 - P$ gives us the sample size n required in order that this test will assure $\theta \geq \theta_o$ at the 100 P per cent significance level. From the ST tables we find, for example, that if we wish to test only for a time $t = 0.01\ \theta_o$ with C = 3, then at the 90 per cent significance level we shall require a test sample of n = 668 units; while if we are willing to test for a time $t = \theta_o$ with C = 1, we need test only 5 units.

The amount of testing called for is appalling if $t \ll \theta_o$; and out of the question if the units are complete systems. For example, if we want to have 95 per cent confidence (synonymous with significance) that a space vehicle has $\theta_o \geq 10$ years, but the test must be made in six months, then

with C = 1, the ST tables say that we must build and test 97 vehicles! Suppose that, nevertheless, it had been decreed on the highest policy level that this degree of confidence must be attained, and you were in charge of the testing program. If a more careful analysis of the statistical problem, requiring a few man-years of statisticians' time, could reduce the test sample by only one or two units, it would be well justified economically. Scrutinizing the test more closely, we note four points:

(1) We know from the experiment not only the total number r of failures, but also the particular times $\{t_1 \ldots t_r\}$ at which failure occurred. This information is clearly relevant to the question being asked; but the ST test makes no use of it.

(2) The test has a "quasi-sequential" feature; if we adopt an acceptance number C = 3, then as soon as the fourth failure occurs, we know that the units are going to be rejected. If no failures occur, the required degree of confidence will be built up long before the time t specified in the ST tables. In fact, t is the maximum possible testing time, which is actually required only in the marginal case where we observe exactly C failures. A test which is "quasi-sequential" in the sense that it terminates when a clear rejection or the required confidence is attained, will have an expected length less than t; conversely, such a test with the expected length set at t will require fewer units tested.

(3) We have relevant prior information; after all, the engineers who designed the space vehicle knew in advance what degree of reliability was needed. They have chosen the quality of materials and components, and the construction methods, with this in mind. Each sub-unit has had its own tests. The vehicles would never have reached the final testing stage unless the engineers knew that they were operating satisfactorily. In other words, we are not testing a completely unknown entity. These facts constitute prior

information about the reliability, just as cogent as anything we can learn from a random experiment.

(4) In practice, we are usually concerned with a different question than the one the ST test answers. An astronaut starting a five-year flight to Mars would not be particularly comforted to be told, "We are 95 per cent confident that the average life of an imaginary population of space vehicles like yours, is at least ten years." He would much rather hear, "There is 95 per cent probability that this vehicle will operate without breakdown for ten years." Such a statement might appear meaningless to an orthodox statistician who holds that (probability) $\equiv$ (frequency). But such a statement would be very meaningful indeed to the astronaut. This is hardly a trivial point; for if it were known that $\lambda^{-1} = 10$ years, the probability that a particular vehicle will actually run for 10 years would be only $1/e = 0.368$; and the period for which we are 95 per cent sure of success would be only $-10 \ln(0.95)$ years, or 6.2 months. Reports which concern only the "mean life" can be rather misleading!

Let us first compare the ST test with a Bayesian test which makes use of exactly the same information; i.e. we are allowed to use only the total number of failures, not the actual failure times. On the hypothesis that the failure rate is λ, the probability that exactly r units fail in time t is

$$p(r|n,\lambda,t) = \binom{n}{r} e^{-(n-r)\lambda t} (1 - e^{-\lambda t})^r \qquad (15\text{-}62)$$

I want to defer discussion of nonuniform priors to a later section; for the time being suppose we assign a uniform prior to λ. This amounts to saying that, before the test, we consider it extremely unlikely that our space vehicles have a mean life as long as a microsecond; nevertheless it will be of interest to see the result of using this prior. The posterior distribution of λ is then

$$p(d\lambda|n,r,t) = \frac{n!}{(n-r-1)!\ r!}\, e^{-(n-r)\lambda t}\, (1 - e^{-\lambda t})^r\, d(\lambda t) \qquad (15\text{-}63)$$

The Bayesian acceptance criterion, which ensures $\theta \geq \lambda_o^{-1}$ with 100 P per cent probability, is then

$$\int_{\lambda_o}^{\infty} p(d\lambda|n,r,t) \leq 1 - P \qquad (15\text{-}64)$$

But the left-hand side of (15-64) is identical with W(n,r) given by (15-61); this is just the well-known identity of the incomplete Beta function and the incomplete binomial sum, given already in the original memoir of Bayes (1762). In this first comparison we therefore find that the ST test is mathematically identical with a Bayesian test in which (1) we are denied use of the actual failure times; (2) because of this it is not possible to take advantage of the quasi-sequential feature; (3) we assign a ridiculously pessimistic prior to λ; (4) we still are not answering the question of real interest for most applications.

Of these shortcomings, (2) is readily corrected, and (1) undoubtedly could be corrected, without departing from orthodox principles. On the hypothesis that the failure rate is λ, the probability that r specified units fail in the time intervals $\{dt_1 \ldots dt_r\}$ respectively, and the remaining (n-r) units do not fail in time t, is

$$p(dt_1 \ldots dt_r|n,\lambda,t) = [\lambda^r\, e^{-\lambda r\bar{t}}\, dt_1 \ldots dt_r][e^{-(n-r)\lambda t}] \qquad (15\text{-}65)$$

where $\bar{t} \equiv r^{-1} \sum t_i$ is the mean life of the units which failed. There is no single "statistic" which conveys all the relevant information; but r and $\bar{t}$ are jointly sufficient, and so an optimal orthodox test must somehow make use of both. When we seek their joint sampling distribution $p(r,d\bar{t}|n,\lambda,t)$ we find, to our dismay, that for given r the interval $0 < \bar{t} < t$ is broken up into r equal intervals, with a different analytical expression for each. Evidently a decrease in r, or an increase in $\bar{t}$, should incline us in the

direction of acceptance; but at what rate should we trade off one against the other? To specify a definite critical region in both variables would seem to imply some postulate as to their relative importance. The problem does not appear simple, either mathematically or conceptually; and I would not presume to guess how an orthodox statistician would solve it.

The relative simplicity of the Bayesian analysis is particularly striking in this problem; for all four of the above shortcomings are corrected effortlessly. For the time being, we again assign the pessimistic uniform prior to λ; from (15-65), the posterior distribution of λ is then

$$p(d\lambda \,|\, n,t,t_1 \ldots t_r) = \frac{(\lambda T)^r}{r!} e^{-\lambda T} \, d(\lambda T) \qquad (15\text{-}66)$$

where

$$T \equiv r\bar{t} + (n-r)t \qquad (15\text{-}67)$$

is the total unit-hours of failure-free operation observed. The posterior probability that $\lambda \geq \lambda_o$ is now

$$B(n,r) = \frac{1}{r!} \int_{\lambda_o T}^{\infty} x^r e^{-x} \, dx = e^{-\lambda_o T} \sum_{k=0}^{r} \frac{(\lambda_o T)^k}{k!} \qquad (15\text{-}68)$$

and so, $B(n,r) \leq 1 - P$ is the new Bayesian acceptance criterion at the 100 P per cent level; the test can terminate with acceptance as soon as this inequality is satisfied.

Numerical analysis shows little difference between this test and the ST teat in the usual range of practical interest, where we test for a time short compared to θ_o and observe only a very few failures. For, if $\lambda_o t << 1$, and $r << n$, then the Poisson approximation to (15-61) will be valid (as in Lecture 8); but this is just the expression (15-68) except for the replacement of T by nt, which is itself a good approximation. In this region the Bayesian test (15-68) with maximum possible duration t generally calls for a test sample one or two units smaller than the ST test. Our common sense readily

assents to this; for if we see only a few failures, then information about the actual failure time adds little to our state of knowledge.

Now let us magnify. The big differences between (15-61) and (15-68) will occur when we find many failures; if all n units fail, the ST test tells us to reject at all confidence levels, even though the observed mean life may have been thousands of times our preassigned θ_o. The Bayesian test (15-68) does not break down in this way; thus if we test 9 units and all fail, it tells us to accept at the 90 per cent level if the observed mean life $\bar{t} \geq 1.58\ \theta_o$. If we test 10 units and 9 fail, the ST test says we can assert with 90 per cent confidence that $\theta \geq 0.22t$; the Bayesian test (15-68) says there is 90 per cent probability that $\theta \geq 0.63\ \bar{t} + 0.07\ t$. Our common sense has no difficulty in deciding which result we should prefer; thus taking the actual failure times into account leads to a clear, although usually not spectacular, improvement in the test. The person who rejects the use of Bayes' theorem in the manner of Eq. (15-66) will be able to obtain a comparable improvement only with far greater difficulty.

But the Bayesian test (15-68) can be further improved in two respects. To correct shortcoming (4), and give a test which refers to the reliability of the individual unit instead of the mean life of an imaginary "population" of them, we note that if λ were known, then by our original hypothesis the probability that the lifetime θ of a given unit is at least θ_o, is

$$p(\theta \geq \theta_o | \lambda) = e^{-\lambda \theta_o} \tag{15-69}$$

The probability that $\theta \geq \theta_o$, conditional on the evidence of the test, is therefore

$$p(\theta \geq \theta_o | n, t_1 \ldots t_r) = \int_0^\infty e^{-\lambda \theta_o}\, p(d\lambda | n, t_1 \ldots t_r)$$

$$= \left(\frac{T}{T + \theta_o} \right)^{r+1} \tag{15-70}$$

Thus, the Bayesian test which ensures, with 100 P per cent probability, that the life of an individual unit is at least θ_o, has an acceptance criterion that the expression (15-70) is $\geq$ P; a result which is simple, sensible, and as far as I can see, utterly beyond the reach of orthodox statistics.

The Bayesian tests (15-68) and (15-70) are, however, still based on a ridiculous prior for λ; another improvement, even further beyond the reach of orthodox statistics, will be found presently, as a result of using a reasonable prior.