Risk, Scores, Fisher Information, and GLRTs
(Supplementary Material for Math 494)

Stanley Sawyer — Washington University

Vs. April 24, 2010

Table of Contents

1. Statistics and Esimatiors
2. Unbiased Estimators, Risk, and Relative Risk
3. Scores and Fisher Information
4. Proof of the Cramér Rao Inequality
5. Maximum Likelihood Estimators are Asymptotically Efficient
6. The Most Powerful Hypothesis Tests are Likelihood Ratio Tests
7. Generalized Likelihood Ratio Tests
8. Fisher's Meta-Analysis Theorem
9. A Tale of Two Contingency-Table Tests

1. Statistics and Estimators. Let $X_1, X_2, \ldots, X_n$ be an independent sample of observations from a probability density $f(x,\theta)$. Here $f(x,\theta)$ can be either discrete (like the Poisson or Bernoulli distributions) or continuous (like normal and exponential distributions).

In general, a *statistic* is an arbitrary function $T(X_1,\ldots,X_n)$ of the data values $X_1,\ldots,X_n$. Thus $T(X)$ for $X = (X_1, X_2, \ldots, X_n)$ can depend on $X_1,\ldots,X_n$, but cannot depend on θ. Some typical examples of statistics are

$$\begin{aligned} T(X_1,\ldots,X_n) \;&=\; \overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} \\ &=\; X_{\max} = \max\{\, X_k : 1 \le k \le n \,\} \\ &=\; X_{\text{med}} = \text{median}\{\, X_k : 1 \le k \le n \,\} \end{aligned} \tag{1.1}$$

These examples have the property that the statistic $T(X)$ is a symmetric function of $X = (X_1,\ldots,X_n)$. That is, any permutation of the sample $X_1,\ldots,X_n$ preserves the value of the statistic. This is not true in general: For example, for $n = 4$ and $X_4 > 0$,

$$T(X_1, X_2, X_3, X_4) \;=\; X_1 X_2 + (1/2) X_3/X_4$$

is also a statistic.

A statistic $T(X)$ is called an *estimator* of a parameter θ if it is a statistic that we think might give a reasonable guess for the true value of θ. In general, we assume that we know the data $X_1,\ldots,X_n$ but not the value of θ. Thus, among statistics $T(X_1,\ldots,X_n)$, what we call an estimator of θ is entirely up to us.

2. Unbiased Estimators, Risk, and Relative Risk. Assume as before that $X = (X_1, X_2, \dots, X_n)$ is an independent sample where each X_k has density $f(x, \theta)$. An estimator $T(X)$ is *unbiased* if

$$E_\theta\big(T(X)\big) \;=\; \theta \qquad \text{for all values of } \theta \tag{2.1}$$

Here $E_\theta(\cdots)$ means that the sums or integrals involved in calculating the expected value depend on the parameter θ. For example, if θ is the mean of a continuous density $f(x, \theta)$, then

$$E_\theta(X_1) = E_\theta(\overline{X}) = \frac{1}{n}\sum_{k=1}^{n} E_\theta(X_k) = \int x f(x, \theta)\, dx = \theta \tag{2.2}$$

and both of the statistics $T_1 = X_1$ and $T_2 = \overline{X}$ are unbiased estimators of θ. If the density $f(x, \theta)$ is discrete instead of continuous, the integral in (2.2) is replaced by a sum.

The relation (2.1) implies that if we had a large number of different samples $X^{(m)}$, each of size n, then the estimates $T(X^{(m)})$ should cluster around the true value of θ. However, it says nothing about the sizes of the errors $T(X^{(m)}) - \theta$, which are likely to be more important.

The errors of $T(X)$ as an estimator of θ can be measured by a *loss function* $L(x, \theta)$, where $L(x, \theta) \ge 0$ and $L(\theta, \theta) = 0$ (see Larsen and Marx, page 419). The *risk* is the expected value of this loss, or

$$R(T, \theta) \;=\; E_\theta\Big(L\Big(T(X), \theta\Big)\Big)$$

The most common choice of loss function is the *quadratic loss function* $L(x, \theta) = (x - \theta)^2$, for which the risk is

$$R(T, \theta) \;=\; E_\theta\Big(\big(T(X) - \theta\big)^2\Big) \tag{2.3}$$

Another choice is the *absolute value* loss function $L(x, \theta) = |x - \theta|$, for which the risk is $R(T, \theta) = E\Big(|T(X) - \theta|\Big)$.

If $T(X)$ is an unbiased estimator and $L(x, \theta) = (x - \theta)^2$, then the risk (2.3) is the same as the variance

$$R(T, \theta) \;=\; \mathrm{Var}_\theta\big(T(X)\big)$$

but not if $T(X)$ is biased (that is, not unbiased).

Assume $E_\theta\big(T(X)\big) = \psi(\theta)$ for a possibly biased estimator $T(X)$. That is, $\psi(\theta) \ne \theta$ for some or all θ. Let $S = T - \theta$, so that $E_\theta(S) = \psi(\theta) - \theta$.

Then $R(T,\theta) = E_\theta\big((T-\theta)^2\big) = E_\theta(S^2)$ and it follows from the relation $\mathrm{Var}(S) = E(S^2) - E(S)^2$ that

$$\begin{aligned} R(T,\theta) &= E_\theta\big((T(X)-\theta)^2\big) \\ &= \mathrm{Var}_\theta\big(T(X)\big) + \big(\psi(\theta)-\theta\big)^2, \quad \psi(\theta) = E_\theta\big(T(X)\big) \end{aligned} \tag{2.4}$$

In principle, we might be able to find a biased estimator $T(X)$ that outperforms an unbiased estimator $T_0(X)$ if the biased estimator has a smaller variance that more than offsets the term $(\psi(\theta)-\theta)^2$ in (2.4).

Example (1). Suppose that $X_1,\ldots,X_n$ are normally distributed $N(\mu,\sigma^2)$ and we want to estimate μ. Then one might ask whether the biased estimator

$$T(X_1,\ldots,X_n) = \frac{X_1+X_2+\ldots+X_n}{n+1} \tag{2.5}$$

could have $R(T,\mu) < R(\overline{X},\mu)$ for the MLE $\overline{X} = (X_1+\ldots+X_n)/n$. While $T(X)$ is biased, it should also have a smaller variance since we divide by a larger number. As in (2.4)

$$R(\overline{X},\mu) = E\Big((\overline{X}-\mu)^2\Big) = \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n} \tag{2.6}$$

$$\begin{aligned} R(T,\mu) &= E\Big((T-\mu)^2\Big) = \mathrm{Var}(T) + E(T-\mu)^2 \\ &= \mathrm{Var}\left(\frac{X_1+\ldots+X_n}{n+1}\right) + \left(\frac{n}{n+1}\,\mu-\mu\right)^2 \\ &= \frac{n\sigma^2}{(n+1)^2} + \frac{\mu^2}{(n+1)^2} \end{aligned}$$

Comparing $R(T,\mu)$ with $R(\overline{X},\mu)$:

$$\begin{aligned} R(T,\mu) - R(\overline{X},\mu) &= \frac{n}{(n+1)^2}\sigma^2 + \frac{\mu^2}{(n+1)^2} - \frac{1}{n}\sigma^2 \\ &= \frac{\mu^2}{(n+1)^2} - \left(\frac{1}{n} - \frac{n}{(n+1)^2}\right)\sigma^2 \\ &= \frac{1}{(n+1)^2}\left(\mu^2 - \left(\frac{(n+1)^2-n^2}{n}\right)\sigma^2\right) \\ &= \frac{1}{(n+1)^2}\left(\mu^2 - \left(\frac{2n+1}{n}\right)\sigma^2\right) \end{aligned} \tag{2.7}$$

Thus $R(T,\mu) < R(\overline{X},\mu)$ if $\mu^2 < ((2n+1)/n)\sigma^2$, which is guaranteed by $\mu^2 < 2\sigma^2$. In that case, $R(T,\mu)$ is less risky than $R(\overline{X},\mu)$ (in the sense of having smaller expected squared error) even though it is biased.

2.1. Shrinkage Estimators. The estimator $T(X)$ in (2.5) can be written

$$T(X_1,\ldots,X_n) \;=\; \frac{n}{n+1}\overline{X} \;+\; \frac{1}{n+1}0$$

which is a convex combination of $\overline{X}$ and 0. A more general estimator is

$$T(X_1,\ldots,X_n) \;=\; c\overline{X} \;+\; (1-c)a \tag{2.8}$$

where a is an arbitrary number and $0 < c < 1$. Estimators of the form (2.5) and (2.8) are called *shrinkage estimators*. While shrinkage estimators are biased unless $E(X_i) = \mu = a$, the calculation above shows that they have smaller risk if $\mu^2 < 2\sigma^2$ for (2.5) or $(\mu-a)^2 < ((1+c)/(1-c))(\sigma^2/n)$ for (2.8).

On the other hand, $R(T,\mu)$ and $R(\overline{X},\mu)$ are of order $1/n$ and, by arguing as in (2.6) and (2.7), the difference between the two is of order $1/n^2$ for fixed μ, a, and $0 < c < 1$. (*Exercise*: Prove this.) Thus one cannot go too far wrong by using $\overline{X}$ instead of a shrinkage estimator.

2.2. Ridge Regression. In *ridge regression* (which is discussed in other courses), the natural estimator $T_1(X)$ of certain parameters is unbiased, but $\mathrm{Var}(T_1)$ is very large because $T_1(X)$ depends on the inverse of a matrix that is very close to being singular.

The method of ridge regression finds *biased* estimators $T_2(X)$ that are similar to $T_1(X)$ such that $E\big(T_2(X)\big)$ is close to $E\big(T_1(X)\big)$ but $\mathrm{Var}\big(T_2(X)\big)$ is of moderate size. If this happens, then (2.4) with $T(X) = T_2(X)$ implies that the biased ridge regression estimator $T_2(X)$ can be a better choice than the unbiased estimator $T_1(X)$ since it can have much lower risk and give much more reasonable estimates.

2.3. Relative Efficiency. Let $T(X)$ and $T_0(X)$ be estimators of θ, where $T_0(X)$ is viewed as a standard estimator such as $\overline{X}$ or the MLE (maximum likelihood estimator) of θ (see below). Then, the *relative risk* or *relative efficiency* of $T(X)$ with respect to $T_0(X)$ is

$$RR(T,\theta) \;=\; \frac{R(T_0,\theta)}{R(T,\theta)} \;=\; \frac{E\big((T_0(X)-\theta)^2\big)}{E\big((T(X)-\theta)^2\big)} \tag{2.9}$$

Note that $T_0(X)$ appears in the numerator, not the denominator, and $T(X)$ appears in the denominator, not the numerator. If $RR(T,\theta) < 1$, then $R(T_0,\theta) < R(T,\theta)$ and $T(X)$ can be said to be less efficient, or more risky, than $T_0(X)$. Conversely, if $RR(T,\theta) > 1$, then $T(X)$ is more efficient (and less risky) than the standard estimator $T_0(X)$.

2.4. MLEs are Not Always Sample Means even if $E_\theta(X) = \theta$.

The most common example with $\widehat{\theta_{\text{MLE}}} = \overline{X}$ is the normal family $N(\theta, 1)$. In that case, $\text{Var}_\theta(\overline{X}) = 1/n$ attains the Cramér-Rao lower bound (see Section 4) and thus is the unbiased estimator of θ with the smaller possible variance.

A second example with $E_\theta(X) = \theta$ (so that $\overline{X}$ is an unbiased estimator of θ) is the *Laplace distribution* $L(\theta, c)$. This has density

$$f(x, \theta, c) \;=\; \frac{1}{2c} e^{-|x-\theta|/c}, \qquad -\infty < x < \infty \tag{2.10}$$

Since the density $f(x, \theta, c)$ is symmetric about $x = \theta$, $E_\theta(X) = E_\theta(\overline{X}) = \theta$.

If Y has the density (2.10), then Y has the same distribution as $\theta + cY_0$ where $Y_0 \approx L(0, 1)$. (*Exercise:* Prove this.) Thus the Laplace family (2.10) is a *shift-and-scale* family like the normal family $N(\theta, \sigma^2)$, and is similar to $N(\theta, \sigma^2)$ except that the probability density of $X \approx L(\theta, c)$ decays exponentially for large x instead of faster than exponentially as is the case for the normal family. (It also has a non-differentiable cusp at $x = \theta$.)

In any event, one might expect that the MLE of θ might be less willing to put as much weight on large sample values than does the sample mean $\overline{X}$, since these values may be less reliable due to the relatively heavy tails of the Laplace distribution. In fact

Lemma 2.1. Let $X = (X_1, \ldots, X_n)$ be an independent sample of size n from the Laplace distribution (2.10) for unknown θ and c. Then

$$\widehat{\theta_{\text{MLE}}}(X) \;=\; \text{median}\{\, X_1, \ldots, X_n \,\} \tag{2.11}$$

Remark. That is, if

$$X_{(1)} < X_{(2)} < \ldots X_{(n)} \tag{2.12}$$

are the *order statistics* of the sample $X_1, \ldots, X_n$, then

$$\widehat{\theta_{\text{MLE}}}(X) \;=\; \begin{cases} X_{(k+1)} & \text{if } m = 2k+1 \text{ is odd} \\ \big(X_{(k)} + X_{(k+1)}\big)/2 & \text{if } m = 2k \text{ is even} \end{cases} \tag{2.13}$$

Thus $\widehat{\theta_{\text{MLE}}} = X_{(2)}$ if $n = 3$ and $X_{(1)} < X_{(2)} < X_{(3)}$, and $\widehat{\theta_{\text{MLE}}} = (X_{(2)} + X_{(3)})/2$ if $n = 4$ and $X_{(1)} < X_{(2)} < X_{(3)} < X_{(4)}$.

Proof of Lemma 2.1. By (2.10), the likelihood of θ is

$$L(\theta, X_1, \ldots, X_n) \;=\; \prod_{i=1}^{n} \left(\frac{1}{2c} e^{-|X_i-\theta|/c} \right) \;=\; \frac{1}{(2c)^n} \exp\left(-\sum_{i=1}^{n} \frac{|X_i - \theta|}{c} \right)$$

It follows that the likelihood $L(\theta, X)$ is *maximized* whenever the sum

$$M(\theta) \;=\; \sum_{i=1}^{n} |X_i - \theta| \;=\; \sum_{i=1}^{n} |X_{(i)} - \theta| \tag{2.14}$$

is *minimized*, where $X_{(i)}$ are the order statistics in (2.12).

The function $M(\theta)$ in (2.14) is continuous and piecewise linear. If $X_{(m)} \le \theta \le X_{(m+1)}$ (that is, if θ lies between the m^{th} and the $(m+1)^{\text{st}}$ order statistics of $\{\, X_i \,\}$), then $X_{(i)} \le X_{(m)} \le \theta$ if $i \le m$ and $\theta \le X_{(m+1)} \le X_{(i)}$ if $m+1 \le i \le n$. Thus

$$M(\theta) \;=\; \sum_{i=1}^{n} |X_{(i)} - \theta| \;=\; \sum_{i=1}^{m} (\theta - X_{(i)}) \;+\; \sum_{i=m+1}^{n} (X_{(i)} - \theta)$$

and if $X_{(m)} < \theta < X_{(m+1)}$

$$\frac{d}{d\theta} M(\theta) \;=\; M'(\theta) \;=\; m - (n-m) = 2m - n$$

It follows that $M'(\theta) < 0$ (and $M(\theta)$ is decreasing) if $m < n/2$ and $M'(\theta) > 0$ (and $M(\theta)$ is increasing) if $m > n/2$. If $n = 2k+1$ is odd, then $n/2 = k+(1/2)$ and $M(\theta)$ is strictly decreasing if $\theta < X_{(k+1)}$ and is strictly increasing if $\theta > X_{(k+1)}$. It follows that the minimum value of $M(\theta)$ is attained at $\theta = X_{(k+1)}$.

If $n = 2k$ is even, then, by the same argument, $M(\theta)$ is minimized at any point in the interval $(X_{(k)}, X_{(k+1)})$, so that any value in that interval maximizes the likelihood. When that happens, the usual convention is to set the MLE equal to the center of the interval, which is the average of the endpoints. Thus $\widehat{\theta_{\text{MLE}}} = X_{(k+1)}$ if $n = 2k+1$ is odd and $\widehat{\theta_{\text{MLE}}} = (X_{(k)} + X_{(k+1)})/2$ if $n = 2k$ is even, which implies (2.13).

A third example of a density with $E_\theta(X) = E_\theta(\overline{X}) = \theta$ is

$$f(x, \theta) = (1/2) I_{(\theta-1, \theta+1)}(x) \tag{2.15}$$

which we can call the *centered uniform* distribution *of length 2*. If X has density (2.15), then X is uniformly distributed between $\theta - 1$ and $\theta + 1$ and $E_\theta(X) = \theta$. The likelihood of an independent sample $X_1, \ldots, X_n$ is

$$\begin{aligned} L(\theta, X_1, \ldots, X_n) \;&=\; \prod_{i=1}^{n} \Big(\frac{1}{2} I_{(\theta-1,\theta+1)}(X_i) \Big) \;=\; \frac{1}{2^n} \prod_{i=1}^{n} I_{(X_i-1, X_i+1)}(\theta) \qquad (2.16) \\ &=\; \frac{1}{2^n} I_{(X_{\max}-1, X_{\min}+1)}(\theta) \end{aligned}$$

since (i) $\theta - 1 < X_i < \theta + 1$ if and only if $X_i - 1 < \theta < X_i + 1$, so that $I_{(\theta-1,\theta+1)}(X_i) = I_{(X_i-1,X_i+1)}(\theta)$, and (ii) the product of the indicator functions is non-zero if and only $X_i < \theta + 1$ and $\theta - 1 < X_i$ for all i, which is equivalent to $\theta - 1 < X_{\min} \le X_{\max} < \theta + 1$ or $X_{\max} - 1 < \theta < X_{\min} + 1$.

Thus the likelihood is zero except for $\theta \in (X_{\max} - 1, X_{\min} + 1)$, where the likelihood has the constant value $1/2^n$. Following the same convention as in (2.13), we set

$$\widehat{\theta_{\mathrm{MLE}}}(X) \;=\; \frac{X_{\max} + X_{\min}}{2} \tag{2.17}$$

(*Exercise*: Note that normally $X_{\min} < X_{\max}$. Prove that the interval $(X_{\max} - 1, X_{\min} + 1)$ is generally nonempty for the density (2.15).)

2.5. Relative Efficiencies of Three Sample Estimators. We can use computer simulation to compare the relative efficiencies of the sample mean, the sample median, and the average of the sample minima and maxima for the three distributions in the previous subsection. Recall that, while all three distributions are symmetric about a shift parameter θ, the MLEs of θ are the sample mean, the sample median, and the average of the sample minimum and maximum, respectively, and are not the same.

It is relatively easy to use a computer to do random simulations of n random samples $X^{(j)}$ $(1 \le j \le n)$ for each of these distributions, where each random sample $X^{(j)} = \big(X_1^{(j)}, \dots, X_m^{(j)}\big)$ is of size m. Thus the randomly simulated data for each distribution will involve generating $n \times m$ random numbers.

For each set of simulated data and each sample estimator $T(X)$, we estimate the risk by $(1/n)\sum_{j=1}^n \big(T(X^{(j)}) - \theta\big)^2$. Analogously with (2.9), we estimate the relative risk with respect to the sample mean $\overline{X}$ by

$$RR(T,\theta) \;=\; \frac{(1/n)\sum_{j=1}^n \Big(\overline{X}^{(j)} - \theta\Big)^2}{(1/n)\sum_{j=1}^n \Big(T(X^{(j)}) - \theta\Big)^2}$$

Then $RR(T,\theta) < 1$ means that the sample mean has less risk, while $RR(T,\theta) > 1$ implies that it is riskier. Since all three distributions are shift invariant in θ, it is sufficient to assume $\theta = 0$ in the simulations.

The simulations show that, in each of the three cases, the MLE is the most efficient of the three estimators of θ. Recall that the MLE is the sample mean only for the normal family. Specifically, we find

Table 2.1: Estimated relative efficiencies with respect to the sample mean for $n = 1{,}000{,}000$ simulated samples, each of size $m = 10$:

Distrib	Mean	Median	AvMinMax	Most Efficient
CentUnif	1.0	0.440	2.196	AvMinMax
Normal	1.0	0.723	0.540	Mean
Laplace	1.0	1.379	0.243	Median

The results are even more striking for samples of size 30:

Table 2.2: Estimated relative efficiencies with respect to the sample mean for $n = 1{,}000{,}000$ simulated samples, each of size $m = 30$:

Distrib	Mean	Median	AvMinMax	Most Efficient
CentUnif	1.0	0.368	5.492	AvMinMax
Normal	1.0	0.666	0.265	Mean
Laplace	1.0	1.571	0.081	Median

Table 2.2 shows that the sample mean has a 3:2 advantage over the sample median for normal samples, but a 3:2 deficit for the Laplace distribution. Averaging the sample minimum and maximum is 5-fold better than the sample mean for the centered uniforms, but is 12-fold worse for the Laplace distribution. Of the three distributions, the Laplace has the largest probability of large values.

3. Scores and Fisher Information. Let $X_1, X_2, \ldots, X_n$ be an independent sample of observations from a density $f(x, \theta)$ where θ is an unknown parameter. Then the *likelihood function* of the parameter θ given the data $X_1, \ldots, X_n$ is

$$L(\theta, X_1, \ldots, X_n) \;=\; f(X_1, \theta) f(X_2, \theta) \ldots f(X_n, \theta) \tag{3.1}$$

where the observations $X_1, \ldots, X_n$ are used in (3.1) instead of dummy variables x_k. Since the data $X_1, \ldots, X_n$ is assumed known, $L(\theta, X_1, \ldots, X_n)$ depends only on the parameter θ.

The *maximum likelihood estimator* of θ is the value $\theta = \widehat{\theta}(X)$ that maximizes the likelihood (3.1). This can often be found by forming the partial derivative of the logarithm of the likelihood

$$\frac{\partial}{\partial\theta} \log L(\theta, X_1, \ldots, X_n) \;=\; \sum_{k=1}^{n} \frac{\partial}{\partial\theta} \log f(X_k, \theta) \tag{3.2}$$

and setting this expression equal to zero. The sum in (3.2) is sufficiently important in statistics that not only the individual terms in the sum, but also their variances, have names.

Specifically, the *scores* of the observations $X_1, \ldots, X_n$ for the density $f(x, \theta)$ are the terms

$$Y_k(\theta) \;=\; \frac{\partial}{\partial\theta} \log f(X_k, \theta) \tag{3.3}$$

Under appropriate assumptions on $f(x, \theta)$ (see Lemma 3.1 below), the scores $Y_k(\theta)$ have mean zero. (More exactly, $E_\theta\big(Y_k(\theta)\big) = 0$, where the same value of θ is used in both parts of the expression.)

The *Fisher information* of an observation X_1 from $f(x, \theta)$ is the variance of the scores

$$I(f, \theta) \;=\; \mathrm{Var}_\theta\big(Y_k(\theta)\big) \;=\; \int \left(\frac{\partial}{\partial\theta} \log f(x, \theta)\right)^2 f(x, \theta)\, dx \tag{3.4}$$

Under an additional hypothesis (see Lemma 3.2 below), we also have

$$I(f, \theta) \;=\; -\int \left(\frac{\partial^2}{\partial\theta^2} \log f(x, \theta)\right) f(x, \theta)\, dx \tag{3.5}$$

which is often easier to compute since it involves a mean rather than a second moment.

For example, assume $X_1, \ldots, X_n$ are normally distributed with unknown mean θ and known variance σ_0^2. Then

$$f(x, \theta) \;=\; \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-(x-\theta)^2/2\sigma_0^2}, \qquad -\infty < x < \infty \tag{3.6}$$

Thus

$$\log f(x, \theta) \;=\; -\frac{1}{2}\log(2\pi\sigma_0^2) \;-\; \frac{(x-\theta)^2}{2\sigma_0^2}$$

It follows that $(\partial/\partial\theta)\log f(x, \theta) = (x - \theta)/\sigma_0^2$, and hence the k^{th} score is

$$Y_k(\theta) \;=\; \frac{\partial}{\partial\theta} \log f(X_k, \theta) \;=\; \frac{X_k - \theta}{\sigma_0^2} \tag{3.7}$$

In particular $E_\theta(Y_k(\theta)) = 0$ as expected since $E_\theta(X_k) = \theta$, and, since $E\big((X_k - \theta)^2\big) = \sigma_0^2$, the scores have variance

$$I(f, \theta) \;=\; E_\theta\big(Y_k(\theta)^2\big) \;=\; \frac{E\big((X_k - \theta)^2\big)}{(\sigma_0^2)^2} \;=\; \frac{1}{\sigma_0^2} \tag{3.8}$$

In this case, the relation

$$\frac{\partial^2}{\partial\theta^2}\log f(X,\theta) \;=\; -\frac{1}{\sigma_0^2}$$

from (3.7) combined with (3.5) gives an easier derivation of (3.8).

The Fisher information $I(f,\theta) = 1/\sigma_0^2$ for (3.6) is large (that is, each X_k has "lots of information") if σ_0^2 is small (so that the error in each X_k is small), and similarly the Fisher information is small if σ_0^2 is large. This may have been one of the original motivations for the term "information".

We give examples below of the importance of scores and Fisher information. First, we give a proof that $E_\theta\big(Y_k(\theta)\big) = 0$ under certain conditions.

Lemma 3.1. Suppose that $K = \{\, x : f(x,\theta) > 0 \,\}$ is the same bounded or unbounded interval for all θ, that $f(x,\theta)$ is smooth enough that we can interchange the derivative and integral in the first line of the proof, and that $(\partial/\partial\theta)\log f(x,\theta)$ is integrable on K. Then

$$E_\theta\big(Y_k(\theta)\big) \;=\; \int\left(\frac{\partial}{\partial\theta}\log f(x,\theta)\right) f(x,\theta)\,dx \;=\; 0 \qquad (3.9)$$

Proof. Since $\int f(x,\theta)\,dx = 1$ for all θ, we can differentiate

$$\begin{aligned}\frac{d}{d\theta}\int f(x,\theta)dx \;&=\; 0 \;=\; \int\frac{\partial}{\partial\theta}f(x,\theta)\,dx \;=\; \int\frac{(\partial/\partial\theta)f(x,\theta)}{f(x,\theta)}\,f(x,\theta)\,dx\\ &=\; \int\left(\frac{\partial}{\partial\theta}\log f(x,\theta)\right) f(x,\theta)\,dx \;=\; 0\end{aligned}$$

Lemma 3.2. Suppose that $f(x,\theta)$ satisfies the same conditions as in Lemma 3.1 and that $\log f(x,\theta)$ has two continuous partial derivatives that are continuous and bounded on K. Then

$$I(f,\theta) \;=\; E_\theta\big(Y_k(\theta)^2\big) \;=\; -\int\left(\frac{\partial^2}{\partial\theta^2}\log f(x,\theta)\right) f(x,\theta)\,dx \qquad (3.10)$$

Proof. Extending the proof of Lemma 3.1,

$$\begin{aligned}\frac{d^2}{d\theta^2}\int f(x,\theta) \;&=\; 0 \;=\; \frac{d}{d\theta}\int\left(\frac{\partial}{\partial\theta}\log f(x,\theta)\right) f(x,\theta)\,dx \qquad (3.11)\\ &=\; \int\left(\frac{\partial^2}{\partial\theta^2}\log f(x,\theta)\right) f(x,\theta)\,dx \;+\; \int\left(\frac{\partial}{\partial\theta}\log f(x,\theta)\right)\frac{\partial}{\partial\theta}f(x,\theta)\,dx\end{aligned}$$

The last term equals

$$\begin{aligned} &\int \left(\frac{\partial}{\partial\theta}\log f(x,\theta)\right) \frac{(\partial/\partial\theta)f(x,\theta)}{f(x,\theta)}\, f(x,\theta)\,dx \\ &\qquad = \int \left(\frac{\partial}{\partial\theta}\log f(x,\theta)\right)^2 f(x,\theta)\,dx \;=\; I(f,\theta) \end{aligned}$$

by (3.4). Since the left-hand side of (3.11) is equal to zero, Lemma 3.2 follows.

Remarks. The hypotheses of Lemma 3.1 are satisfied for the normal density (3.6), for which $E_\theta\big(Y_k(\theta)\big) = 0$ by (3.7). However, the hypotheses are not satisfied for the uniform density $f(x,\theta) = (1/\theta)I_{(0,\theta)}(x)$ since the supports $K(\theta) = (0,\theta)$ depend on θ.

For $f(x,\theta) = (1/\theta)I_{(0,\theta)}(x)$, the scores $Y_k(\theta) = -(1/\theta)I_{(0,\theta)}(X_k)$ have means $E_\theta\big(Y_k(\theta)\big) = -1/\theta \neq 0$, so that the proof of Lemma 3.1 breaks down at some point. (*Exercise*: Show that this is the correct formula for the score for the uniform density and that this is the mean value.)

4. The Cramér-Rao Inequality. Let $X_1, X_2, \ldots, X_n$ be an independent random sample from the density $f(x,\theta)$, where $f(x,\theta)$ satisfies the conditions of Lemma 3.1. In particular,

(i) The set $K = \{\, x : f(x,\theta) > 0 \,\}$ is the same for all values of θ and

(ii) The function $\log f(x,\theta)$ has two continuous partial derivatives in θ that are integrable on K.

We then have

Theorem 4.1. (Cramér-Rao Inequality) Let $T(X_1, X_2, \ldots, X_n)$ be an arbitrary unbiased estimator of θ. Then, under the assumptions above,

$$E_\theta\big((T-\theta)^2\big) \;\geq\; \frac{1}{n\,I(f,\theta)} \tag{4.1}$$

for all values of θ, where $I(f,\theta)$ is the Fisher information defined in (3.4).

Remark. Note that (4.1) need not hold if $T(X_1, \ldots, X_n)$ is a *biased* estimator of θ, nor if the assumptions (i) or (ii) fail.

Proof of Theorem 4.1. Let $T = T(X_1, \ldots, X_n)$ be any unbiased estimator of θ. Then

$$\begin{aligned} \theta \;&=\; E_\theta\big(T(X_1, \ldots, X_n)\big) && (4.2) \\ &=\; \int \!\ldots\! \int T(y_1, \ldots, y_n)\, f(y_1,\theta) \ldots f(y_n,\theta)\, dy_1 \ldots dy_n \end{aligned}$$

Differentiating (4.2) with respect to θ

$$1 \;=\; \int\!\dots\!\int T(y_1,\dots,y_n)\,\frac{\partial}{\partial\theta}\Bigl(\prod_{k=1}^{n} f_k\Bigr)\,dy_1\dots dy_n$$

where $f_k = f(y_k,\theta)$. By the chain rule

$$1 \;=\; \int\!\dots\!\int T\;\sum_{k=1}^{n}\left(\left(\prod_{j=1}^{k-1} f_j\right)\Bigl(\frac{\partial}{\partial\theta}f_k\Bigr)\left(\prod_{j=k+1}^{n} f_j\right)\right)dy_1\dots dy_n$$

for $T = T(y_1,y_2,\dots,y_n)$ and

$$\begin{aligned}
1 \;&=\; \int\!\dots\!\int T\;\sum_{k=1}^{n}\left(\left(\prod_{j=1}^{k-1} f_j\right)\Bigl(\frac{(\partial/\partial\theta)f_k}{f_k}\Bigr)f_k\left(\prod_{j=k+1}^{n} f_j\right)\right)dy_1\dots dy_n\\
&=\; \int\!\dots\!\int T\;\left(\sum_{k=1}^{n}\frac{\partial}{\partial\theta}\log\bigl(f(y_k,\theta)\bigr)\right)\left(\prod_{j=1}^{n} f_j\right)\;dy_1\dots dy_n\\
&=\; E_\theta\left(T(X_1,\dots,X_n)\left(\sum_{k=1}^{n}Y_k\right)\right) \qquad (4.3)
\end{aligned}$$

where $Y_k = (\partial/\partial\theta)\log f(X_k,\theta)$ are the scores defined in (3.3). Since $E_\theta(Y_k(\theta)) = 0$ by Lemma 3.1, it follows by subtraction from (4.3) that

$$1 \;=\; E_\theta\left(\Bigl(T(X_1,\dots,X_n)-\theta\Bigr)\left(\sum_{k=1}^{n}Y_k\right)\right) \qquad (4.4)$$

By Cauchy's inequality (see Lemma 4.1 below),

$$E(XY) \le \sqrt{E(X^2)}\sqrt{E(Y^2)}$$

for any two random variables X, Y with $E(|XY|) < \infty$. Equivalently $E(XY)^2 \le E(X^2)E(Y^2)$. Applying this in (4.4) implies

$$1 \;\le\; E_\theta\Bigl(\bigl(T(X_1,\dots,X_n)-\theta\bigr)^2\Bigr)\;E_\theta\left(\left(\sum_{k=1}^{n}Y_k\right)^2\right) \qquad (4.5)$$

The scores $Y_k = (\partial/\partial\theta)\log f(X_k,\theta)$ are independent with the same distribution, and have mean zero and variance $I(f,\theta)$ by Lemma 3.1 and (3.4). Thus

$$E_\theta\left(\left(\sum_{k=1}^{n} Y_k\right)^2\right) = \mathrm{Var}_\theta\left(\sum_{k=1}^{n} Y_k\right) = n\,\mathrm{Var}_\theta(Y_1) = nI(f,\theta)$$

Hence $1 \le E_\theta\big((T-\theta)^2\big)\,nI(f,\theta)$, which implies the lower bound (4.1).

Definition. The *efficiency* of an estimator $T(X_1,\ldots,X_n)$ is

$$RE\,(T,\theta) \;=\; \frac{1/\big(nI(f,\theta)\big)}{E_\theta((T-\theta)^2)} \;=\; \frac{1}{nI(f,\theta)\,E_\theta((T-\theta)^2)} \tag{4.6}$$

Note that this is the same as the *relative risk* or *relative efficiency* (2.9) with $R(T_0,\theta)$ replaced by the Cramér-Rao lower bound (4.1). Under the assumptions of Theorem 4.1, $RE\,(T,\theta) \le 1$.

An unbiased estimator $T(X_1,\ldots,X_n)$ is called *efficient* if $RE\,(T,\theta) = 1$; that is, if its variance attains the lower bound in (4.1). This means that any other unbiased estimator of θ, no matter how nonlinear, must have an equal or larger variance.

An estimator $T(X)$ of a parameter θ is *super-efficient* if its expected squared error $E\big((T(X)-\theta)^2\big)$ is *strictly less* than the Cramér-Rao lower bound. Under the assumptions of Theorem 4.1, this can happen only if $T(X)$ is biased, and typically holds for some parameter values θ but not for others. For example, the shrinkage estimator of Section 2.1 is super-efficient for parameter values θ that are reasonably close to the value a but not for other θ.

Examples (1). Assume $X_1, X_2, \ldots, X_n$ are $N(\theta,\sigma_0^2)$ (that is, normally distribution with unknown mean θ and known variance σ_0^2). Then $E_\theta(X_k) = E_\theta(\overline{X}) = \theta$, and $\overline{X}$ is an unbiased estimator of θ. Its variance is

$$\mathrm{Var}_\theta(\overline{X}) \;=\; (1/n)\,\mathrm{Var}_\theta(X_1) \;=\; \frac{\sigma_0^2}{n}$$

By (3.8), the Fisher information is $I(f,\theta) = 1/\sigma_0^2$, so that $1/(nI(f,\theta)) = \sigma_0^2/n$. Thus $\mathrm{Var}_\theta(\overline{X})$ attains the Cramér-Rao lower bound for unbiased estimators of θ, so that $\overline{X}$ is an *efficient* unbiased estimator of θ.

(2). Assume that $X_1,\ldots,X_n$ are uniformly distributed in $(0,\theta)$ for some unknown value of θ, so that they have density $f(x,\theta) = (1/\theta)I_{(0,\theta)}(x)$.

Then $X_1, \ldots, X_n$ do not satisfy condition (i) at the beginning of Section 4, but we can see if Theorem 4.1 holds anyway.

As in a remark at the end of Section 3, the scores are $Y_k(\theta) = (-1/\theta)I_{(0,\theta)}(X_k)$. Thus $Y_k(\theta) = -1/\theta$ whenever $0 < X_k < \theta$ (that is, with probability one), so that $E_\theta\big(Y_k(\theta)\big) = -1/\theta$ and $E_\theta\big(Y_k(\theta)^2\big) = 1/\theta^2$. Hence $I(f,\theta) = \mathrm{Var}\big(Y_k(\theta)\big) = 0$, so that the Cramér-Rao lower bound for unbiased estimators in Theorem 4.1 is ∞. (If we can use Lemma 3.2, then $I(f,\theta) = -1/\theta^2$ and the lower bound is negative. These are not contradictions, since the density $f(x,\theta) = (1/\theta)I_{(0,\theta)}(x)$ does not satisfy the hypotheses of either Lemma 3.1 or 3.2.)

Ignoring these awkwardnesses for the moment, let $X_{\max} = \max_{1\le i\le n} X_i$. Then $E_\theta(X_{\max}) = (n/(n+1))\theta$, so that if $T_{\max} = ((n+1)/n)X_{\max}$

$$E_\theta(2\overline{X}) \;=\; E_\theta(T_{\max}) \;=\; \theta$$

Thus both $T_1 = 2\overline{X}$ and $T_2 = T_{\max}$ are unbiased estimators of θ. However, one can show

$$\mathrm{Var}_\theta(2\overline{X}) = \frac{2\theta^2}{3n} \qquad \text{and} \qquad \mathrm{Var}_\theta(T_{\max}) = \frac{\theta^2}{n(n+2)}$$

Assuming $I(f,\theta) = 1/\theta^2$ for definiteness, this implies that

$$RE\,(2\overline{X},\theta) = 3/2 \quad \text{and} \quad RE\,(T_{\max},\theta) = n+2 \;\to\; \infty \quad \text{(if n is large)}$$

Thus the conclusions of Theorem 4.1 are either incorrect or else make no sense for either unbiased estimator in this case.

We end this section with a proof of Cauchy's inequality.

Lemma 4.1 (Cauchy-Schwartz-Bunyakowski) Let X, Y be any two random variables such that $E(|XY|) < \infty$. Then

$$E(XY) \le \sqrt{E(X^2)}\sqrt{E(Y^2)} \tag{4.7}$$

Proof. Note $\big((\sqrt{a})\,x - (1/\sqrt{a})\,y\big)^2 \ge 0$ for arbitrary real numbers x, y, a with $a > 0$. Expanding the binomial implies $ax^2 - 2xy + (1/a)y^2 \ge 0$, or

$$xy \;\le\; (1/2)\Big(ax^2 \;+\; \frac{1}{a}\,y^2\Big)$$

for all real x, y and any $a > 0$. It then follows that for any values of the random variables X, Y

$$XY \;\le\; \frac{1}{2}\left(aX^2 \;+\; \frac{1}{a}\,Y^2\right)$$

In general, if $Y_1 \le Y_2$ for two random variables Y_1, Y_2, then $E(Y_1) \le E(Y_2)$. This implies

$$E(XY) \;\le\; \frac{1}{2}\Big(aE(X^2) \;+\; \frac{1}{a}\,E(Y^2)\Big), \qquad \text{any } a > 0 \tag{4.8}$$

If we minimize the right-hand side of (4.8) as a function of a, for example by setting the derivative with respect to a equal to zero, we obtain $a^2 = E(Y^2)/E(X^2)$ or $a = \sqrt{E(Y^2)/E(X^2)}$. Evaluating the right-hand side of (4.8) with this value of a implies (4.7).

5. Maximum Likelihood Estimators are Asymptotically Efficient. Let $X_1, X_2, \ldots, X_n, \ldots$ be independent random variables with the same distribution. Assume $E(X_j^2) < \infty$ and $E(X_j) = \mu$. Then the central limit theorem implies

$$\lim_{n\to\infty} P\left(\frac{X_1+X_2+\cdots+X_n - n\mu}{\sqrt{n\sigma^2}} \le y\right) \;=\; \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y} e^{-(1/2)x^2}\,dx \tag{5.1}$$

for all real values of y. Is there something similar for MLEs (maximum likelihood estimators)? First, note that (5.1) is equivalent to

$$\lim_{n\to\infty} P\left(\sqrt{\frac{n}{\sigma^2}}\left(\frac{X_1+X_2+\cdots+X_n}{n} - \mu\right) \le y\right) \;=\; \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y} e^{-(1/2)x^2}\,dx \tag{5.2}$$

If $X_1, \ldots, X_n$ were normally distributed with mean θ and variance σ^2, then $\widehat{\theta}_n(X) = \overline{X} = (X_1 + \ldots + X_n)/n$. This suggests that we might have a central limit theorem for MLEs $\widehat{\theta}_n(X)$ of the form

$$\lim_{n\to\infty} P\Big(\sqrt{nc(\theta)}\,\big(\widehat{\theta}_n(X) - \theta\big) \le y\Big) \;=\; \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y} e^{-(1/2)x^2}\,dx$$

where θ is the true value of θ and $c(\theta)$ is a constant depending on θ. In fact

Theorem 5.1. Assume

(i) The set $K = \{\, x : f(x,\theta) > 0 \,\}$ is the same for all values of θ,

(ii) The function $\log f(x,\theta)$ has two continuous partial derivatives in θ that are integrable on K, and

(iii) $E(Z) < \infty$ for $Z = \sup_\theta |(\partial^2/\partial\theta^2)\log f(X,\theta)|$, and

(iv) the MLE $\widehat{\theta}(X)$ is attained in the *interior* of K.

Let $I(f,\theta)$ be the Fisher information (3.8) in Section 3. Then

$$\lim_{n\to\infty} P_\theta\Big(\sqrt{n\,I(f,\theta)}\,\Big(\widehat{\theta}_n(X) - \theta\Big) \le y\Big) \;=\; \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y} e^{-(1/2)x^2}\,dx \tag{5.3}$$

for all real values of y. (Condition (iii) is more than is actually required.)

Remarks (1). The relation (5.3) says that the MLE $\widehat{\theta}_n(X)$ is approximately normally distributed with mean θ and variance $1/\big(nI(f,\theta)\big)$, or symbolically

$$\widehat{\theta}_n(X) \;\approx\; N\bigg(\theta, \frac{1}{n\,I(f,\theta)}\bigg) \tag{5.4}$$

If (5.4) held exactly, then $E_\theta\big(\widehat{\theta}_n(X)\big) = \theta$ and $\mathrm{Var}_\theta\big(\widehat{\theta}_n(X)\big) = 1/\big(nI(f,\theta)\big)$, and $\widehat{\theta}_n(X)$ would be an unbiased estimator whose variance was equal to the Cramér-Rao lower. We interpret (5.3)–(5.4) as saying that $\widehat{\theta}_n(X)$ is *asymptotically* normal, is *asymptotically* unbiased, and is *asymptotically* efficient in the sense of Section 4, since its *asymptotic* variance is the the Cramér-Rao lower bound. However, (5.3) does not exclude the possibility that $E_\theta\big(|\widehat{\theta}_n(X)|\big) = \infty$ for all finite n, so that $\widehat{\theta}_n(X)$ need not be unbiased nor efficient nor even have finite variance in the usual senses for any value of n.

(2). If $f(x,\theta) = (1/\theta)I_{(0,\theta)}(x)$, so that $X_i \approx U(0,\theta)$, then the order of the rate of convergence in the analog of (5.3) is n instead of $\sqrt{n}$ and the limit is a one-sided exponential, not a normal distribution. (*Exercise*: Prove this.) Thus the conditions of Theorem 5.1 are essential.

Asymptotic Confidence Intervals. We can use (5.3) to find asymptotic confidence intervals for the true value of θ based on the MLE $\widehat{\theta}_n(X)$. It follows from (5.3) and properties of the standard normal distribution that

$$\begin{aligned}&\lim_{n\to\infty} P_\theta\left(\frac{-1.96}{\sqrt{nI(f,\theta)}} \;<\; \widehat{\theta}_n - \theta \;<\; \frac{1.96}{\sqrt{nI(f,\theta)}}\right) \qquad (5.5)\\ &= \lim_{n\to\infty} P_\theta\left(\widehat{\theta}_n(X) - \frac{1.96}{\sqrt{nI(f,\theta)}} \;<\; \theta \;<\; \widehat{\theta}_n(X) + \frac{1.96}{\sqrt{nI(f,\theta)}}\right) \;=\; 0.95\end{aligned}$$

Under the assumptions of Theorem 5.1, we can approximate the Fisher information $I(f,\theta)$ in (3.8) by $I\big(f,\widehat{\theta}_n(X)\big)$, which does not depend explicitly on θ. The expression $I\big(f,\widehat{\theta}_n(X)\big)$ is called the *empirical Fisher information* of θ depending on $X_1,\ldots,X_n$. This and (5.5) imply that

$$\left(\widehat{\theta}_n(X) - \frac{1.96}{\sqrt{nI(f,\widehat{\theta}_n(X))}}\;,\;\; \widehat{\theta}_n(X) + \frac{1.96}{\sqrt{nI(f,\widehat{\theta}_n(X))}}\right) \tag{5.6}$$

is an *asymptotic* 95% confidence interval for the true value of θ.

Examples (1). Let $f(x,p) = p^x(1-p)^{1-x}$ for $x = 0,1$ for the Bernoulli distribution. (That is, tossing a biased coin.) Then

$$\begin{aligned} \log f(x,p) &= x\log(p) + (1-x)\log(1-p) \\ \frac{\partial}{\partial\theta}\log f(x,p) &= \frac{x}{p} - \frac{1-x}{1-p} \qquad \text{and} \qquad \frac{\partial^2}{\partial\theta^2}\log f(x,p) \ = \ -\frac{x}{p^2} - \frac{1-x}{(1-p)^2} \end{aligned}$$

Thus by Lemma 3.2 the Fisher information is

$$\begin{aligned} I(f,p) &= -E\left(\frac{\partial^2}{\partial\theta^2}\log f(X,p)\right) = \frac{E(X)}{p^2} + \frac{E(1-X)}{(1-p)^2} \\ &= \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)} \end{aligned}$$

This implies

$$\frac{1}{\sqrt{nI(f,p}} = \sqrt{\frac{p(1-p)}{n}}$$

Hence in this case (5.6) is exactly the same as the usual (approximate) 95% confidence interval for the binomial distribution.

(2). Let $f(x,\theta) = \theta x^{\theta-1}$ for $0 \le x \le 1$. Then

$$\begin{aligned} Y_k(\theta) &= (\partial/\partial\theta)\log f(X_k,\theta) = (1/\theta) + \log(X_k) \\ W_k(\theta) &= (\partial^2/\partial\theta^2)\log f(X_k,\theta) = -1/\theta^2 \end{aligned}$$

Since $(\partial/\partial\theta)\log L(\theta,X) = \sum_{k=1}^n Y_k(\theta) = (n/\theta) + \sum_{k=1}^n \log(X_k)$, it follows that

$$\widehat{\theta}_n(X) = -\frac{n}{\sum_{k=1}^n \log(X_k)} \tag{5.7}$$

Similarly, $I(f,\theta) = -E_\theta\big(W_k(\theta)\big) = 1/\theta^2$ by Lemma 3.2. Hence by (5.6)

$$\left(\widehat{\theta}_n(X) \ - \ \frac{1.96\,\widehat{\theta}_n(X)}{\sqrt{n}}\ , \ \ \widehat{\theta}_n(X) \ + \ \frac{1.96\,\widehat{\theta}_n(X)}{\sqrt{n}}\right) \tag{5.8}$$

is an asymptotic 95% confidence interval for θ.

Proof of Theorem 5.1. Let

$$M(\theta) \ = \ \frac{\partial}{\partial\theta}\log L(\theta, X_1, \ldots, X_n) \tag{5.9}$$

where $L(\theta, X_1, \ldots, X_n)$ is the likelihood function defined in (3.1). Let $\widehat{\theta}_n(X)$ be the maximum likelihood estimator of θ. Since $\widehat{\theta}_n(X)$ is attained in the interior of K by condition (iv),

$$M(\widehat{\theta}_n) \;=\; \frac{\partial}{\partial\theta}\log L(\widehat{\theta}_n, X) \;=\; 0$$

and by Lemma 3.1

$$M(\theta) \;=\; \frac{\partial}{\partial\theta}\log L(\theta, X) \;=\; \sum_{k=1}^{n}\frac{\partial}{\partial\theta}\log f(X_k,\theta) \;=\; \sum_{k=1}^{n} Y_k(\theta)$$

where $Y_k(\theta)$ are the scores defined in Section 3. By the mean value theorem

$$\begin{aligned} M(\widehat{\theta}_n) - M(\theta) \;&=\; (\widehat{\theta}_n - \theta)\,\frac{d}{d\theta}M(\widetilde{\theta}_n) \;=\; (\widehat{\theta}_n - \theta)\,\frac{\partial^2}{\partial\theta^2}\log L(\widetilde{\theta}_n, X) \\ &=\; (\widehat{\theta}_n - \theta)\sum_{k=1}^{n}(\partial^2/\partial\theta^2)\log f\big(X_k, \widetilde{\theta}_n(X)\big) \end{aligned}$$

where $\widetilde{\theta}_n(X)$ is a value between θ and $\widehat{\theta}_n(X)$. Since $M(\widehat{\theta}_n) = 0$

$$\widehat{\theta}_n - \theta \;=\; \frac{-M(\theta)}{(d/d\theta)M(\widetilde{\theta}_n)} \;=\; \frac{\sum_{k=1}^{n} Y_k(\theta)}{-\sum_{k=1}^{n}(\partial^2/\partial\theta^2)\log f(X_k, \widetilde{\theta}_n)} \tag{5.10}$$

Thus

$$\sqrt{nI(f,\theta)}\,\big(\widehat{\theta}_n - \theta\big) \;=\; \frac{\dfrac{1}{\sqrt{nI(f,\theta)}}\displaystyle\sum_{k=1}^{n} Y_k(\theta)}{-\frac{1}{nI(f,\theta)}\sum_{k=1}^{n}(\partial^2/\partial\theta^2)\log f(X_k, \widetilde{\theta}_n(X))} \tag{5.11}$$

By Lemma 3.1, the $Y_k(\theta)$ are independent with the same distribution with $E_\theta\big(Y_k(\theta)\big) = 0$ and $\mathrm{Var}_\theta\big(Y_k(\theta)\big) = I(f,\theta)$. Thus by the central limit theorem

$$\lim_{n\to\infty} P_\theta\left(\frac{1}{\sqrt{nI(f,\theta)}}\sum_{k=1}^{n} Y_k(\theta) \;\le\; y\right) \;=\; \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y} e^{-(1/2)x^2}\,dx \tag{5.12}$$

Similarly, by Lemma 3.2, $W_k(\theta) = (\partial^2/\partial\theta^2)\log f(X_k,\theta)$ are independent with $E_\theta\big(W_k(\theta)\big) = -I(f,\theta)$. Thus by the law of large numbers

$$\lim_{n\to\infty}\frac{1}{nI(f,\theta)}\sum_{k=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(X_k,\theta) \;=\; -1 \tag{5.13}$$

in the sense of convergence in the law of large numbers. One can show that, under the assumptions of Theorem 5.1, we can replace $\widetilde{\theta}_n(X)$ on the right-hand side of (5.11) by θ as $n \to \infty$. It can then be shown from (5.11)–(5.13) that

$$\lim_{n\to\infty} P_\theta\Big(\sqrt{n\,I(f,\theta)}\,\big(\widehat{\theta}_n(X)-\theta\big) \le y\Big) \;=\; \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y} e^{-(1/2)x^2}\,dx$$

for all real values of y. This completes the proof of Theorem 5.1.

6. The Most Powerful Hypothesis Tests are Likelihood Ratio Tests. The preceding sections have been concerned with estimation and interval estimation. These are concerned with finding the most likely value or range of values of a parameter θ, given an independent sample $X_1,\ldots,X_n$ from a probability density $f(x,\theta)$ for an unknown value of θ.

In contrast, *hypothesis testing* has a slightly different emphasis. Suppose that we want to use data $X_1,\ldots,X_n$ to decide between two different hypotheses, which by convention are called hypotheses H_0 and H_1. The hypotheses are not treated in a symmetrical manner. Specifically,

H_0: What one would believe if one had no additional data

H_1: What one would believe if the data $X_1,\ldots,X_n$ makes the alternative hypothesis H_1 significantly more likely.

Rather than estimate a parameter, we decide between two competing hypotheses, or more exactly decide (yes or no) whether the data $X_1,\ldots,X_n$ provide sufficient evidence to reject the conservative hypothesis H_0 in favor of a new hypothesis H_1.

This is somewhat like an an estimation procedure with $D(X) = D(X_1,\ldots,X_n) = 1$ for hypothesis H_1 and $D(X_1,\ldots,X_n) = 0$ for H_0. However, this doesn't take into account the question of whether we have sufficient evidence to reject H_0.

A side effect of the bias towards H_0 is that choosing H_1 can be viewed as "proving" H_1 in some sense, while choosing H_0 may just mean that we do not have enough evidence one way or the other and so stay with the more conservative hypothesis.

Example. (Modified from Larsen and Marx, pages 428–431.) Suppose that it is generally believed that a certain type of car averages 25.0 miles per gallon (mpg). Assume that measurements $X_1,\ldots,X_n$ of the miles per gallon are normally distributed with distribution $N(\theta,\sigma_0^2)$ with $\sigma_0 = 2.4$. The conventional wisdom is then $\theta = \theta_0 = 25.0$.

A consumers' group suspects that the current production run of cars actually has a higher mileage rate. In order to test this, the group runs $n = 30$ cars through a typical course intended to measure miles per gallon. The results are observations of mpg $X_1, \ldots, X_{30}$ with sample mean $\overline{X} = (1/n)\sum_{i=1}^{30} X_i = 26.50$. Is this sufficient evidence to conclude that mileage per gallon has improved?

In this case, the "conservative" hypothesis is

$$H_0 : X_i \approx N(\theta_0, \sigma_0^2) \tag{6.1}$$

for $\theta_0 = 25.0$ and $\sigma_0 = 2.40$. The alternative hypothesis is

$$H_1 : X_i \approx N(\theta, \sigma_0^2) \quad \text{for some } \theta > \theta_0 \tag{6.2}$$

A standard statistical testing procedure is, in this case, first to choose a "level of significance" α that represents the degree of confidence that we need to reject H_0 in favor of H_1. The second step is to choose a "critical value" $\lambda = \lambda(\alpha)$ with the property that

$$P(\overline{X} \ge \lambda) \;=\; P\big(\overline{X} \ge \lambda(\alpha)\big) \;=\; \alpha \tag{6.3}$$

Given α, the value $\lambda = \lambda(\alpha)$ in (6.3) can be determined from the properties of normal distributions and the parameters in (6.1), and is in fact $\lambda = 25.721$ for $\alpha = 0.05$ and $n = 30$. (See below.)

The final step is to compare the measure $\overline{X} = 26.50$ with λ. If $\overline{X} \ge \lambda$, we reject H_0 and conclude that the mpgs of the cars have improved. If $\overline{X} < \lambda$, we assume that, even though $\overline{X} > \theta_0$, we do not have sufficient evidence to conclude that mileage has improved. Since $\overline{X} = 26.50 > 25.721$, we reject H_0 in favor of H_1 for this value of α, and conclude that the true $\theta > 25.0$.

Before determining whether or not this is the best possible test, we first need to discuss what is a test, as well as a notion of "best".

6.1. What is a Test? What Do We Mean by the Best Test?

The standard test procedure leading up to (6.3) leaves open a number of questions. Why should the best testing procedure involve $\overline{X}$ and not a more complicated function of $X_1, \ldots, X_n$? Could we do better if we used more of the data? Even if the best test involves only $\overline{X}$, why necessarily the simple form $\overline{X} > \lambda$?

More importantly, what should we do if the data $X_1, \ldots, X_n$ are not normal under H_0 and H_1, and perhaps involve a family of densities $f(x, \theta)$ for which the MLE is not the sample mean? Or if H_0 is expressed in terms of one family of densities (such as $N(\theta, \sigma_0^2)$) and H_1 in terms of a different family, such as gamma distributions?

Before proceeding, we need a general definition of a test, and later a definition of "best".

Assume for definiteness that $\theta, X_1, \ldots, X_n$ are all real numbers. We then define (an abstract) *test* to be an arbitrary subset $\mathcal{C} \subseteq R^n$, with the convention that we choose H_1 if the data $X = (X_1, \ldots, X_n) \in \mathcal{C}$ and otherwise choose H_0. (The set $\mathcal{C} \subseteq R^n$ is sometimes called the *critical region* of the test.) Note that the decision rule $D(X)$ discussed above is now the indicator function $D(X) = I_{\mathcal{C}}(X)$.

In the example (6.1)–(6.3), $\mathcal{C} = \left\{ \widetilde{x} : \overline{x} = \frac{1}{n}\sum_{i=1}^n x_i \geq \lambda \right\}$ for $\widetilde{x} = (x_1, \ldots, x_n)$, so that $X \in \mathcal{C}$ if and only if $\overline{X} \geq \lambda$.

Later we will derive a formula that gives the best possible test in many circumstances. Before continuing, however, we need some more definitions.

6.2. Simple vs. Composite Tests. In general, we say that a hypothesis (H_0 or H_1) is a *simple hypothesis* or is *simple* if it uniquely determines the density of the random variables X_i. The hypothesis is *composite* otherwise.

For example, suppose that the X_i are known to have density $f(x, \theta)$ for unknown θ for a family of densities $f(x, \theta)$, as in (6.1)–(6.2) for a normal family with known variance. Then

$$H_0 : \theta = \theta_0 \qquad \text{and} \qquad H_1 : \theta = \theta_1 \tag{6.4}$$

are both simple hypotheses. If as in (6.1)–(6.2)

$$H_0 : \theta = \theta_0 \qquad \text{and} \qquad H_1 : \theta > \theta_0 \tag{6.5}$$

then H_0 is simple but H_1 is composite.

Fortunately, if often turns out that the best test for $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ is the same test for all $\theta_1 > \theta_0$, so that it is also the best test against $H_1 : \theta > \theta_0$. Thus, in this case, it is sufficient to consider simple hypotheses as in (6.4).

6.3. The Size and Power of a Test. If we make a decision between two hypotheses H_0 and H_1 on the basis of data $X_1, \ldots, X_n$, then there are two types of error that we can make.

The first type (called a Type I error) is to reject H_0 and decide on H_1 when, in fact, the conservative hypothesis H_0 is true. The probability of a Type I error (which can only happen if H_0 is true) is called the *false positive* rate. The reason for this is that deciding on the *a priori* less likely hypothesis H_1 is called a *positive* result. (Think of proving H_1 as the first step towards a big raise, or perhaps towards getting a Nobel prize. On the other hand, deciding on H_1 could mean that you have a dread disease, which you might not consider a positive result at all. Still, it is a positive result for the *test*, if not necessarily for you.)

Suppose that H_0 and H_1 are both simple as in (6.4). Then the probability of a Type I error for the test $\mathcal{C}$, or equivalently the false positive rate, is

$$\begin{aligned} \alpha \;&=\; P(\text{reject } H_0 \mid H_0 \text{ is true}) \;=\; P(\text{choose } H_1 \mid H_0) \\ &=\; P(X \in \mathcal{C} \mid H_0) \;=\; \int_{\mathcal{C}} f(\widetilde{x}, \theta_0)\, d\widetilde{x} \end{aligned} \tag{6.6}$$

where

$$f(\widetilde{x}, \theta) = f(x_1, \theta) f(x_2, \theta) \ldots f(x_n, \theta) \tag{6.7}$$

is the joint probability density of the sample $X = (X_1, \ldots, X_n)$ and $\int_C f(\widetilde{x}, \theta_0)\, d\widetilde{x}$ is an n-fold integral.

As the form of (6.6) indicates, α depends only on the hypothesis H_0 and not on H_1, since it is given by the integral of $f(\widetilde{x}, \theta_0)$ over $\mathcal{C}$ and does not involve θ_1. Similarly, the "critical value" $\lambda = \lambda(\alpha)$ in (6.3) in the automobile example depends only on α and n and the parameters involved in H_0.

The value α in (6.6) is also called the *level of significance* of the test $\mathcal{C}$ (or, more colloquially, of the test with critical region $\mathcal{C}$). As mentioned above, α depends only on the hypothesis H_0 and is given by the integral of a probability density over $\mathcal{C}$. For this reason, α is also called the *size* of the test $\mathcal{C}$. That is,

$$\text{Size}(\mathcal{C}) \;=\; \int_{\mathcal{C}} f(\widetilde{x}, \theta_0)\, d\widetilde{x} \tag{6.8}$$

Note that we have just given four different verbal definitions for the value α in (6.6) or the value of the integral in (6.8). This illustrates the importance of α for hypothesis testing.

Similarly, a Type II error is to reject H_1 and choose H_0 when the alternative H_1 is correct. The probability of a Type II error is called the *false negative* rate, since it amounts to failing to detect H_1 when H_1 is correct. This is

$$\beta \;=\; P(\text{reject } H_1 \mid H_1) \;=\; \int_{R^n - \mathcal{C}} f(\widetilde{x}, \theta_1)\, d\widetilde{x} \tag{6.9}$$

for θ_1 in (6.4) and $f(\widetilde{x}, \theta_1)$ in (6.7). Note that β depends only on H_1 and not on H_0.

The *power* of a test is the probability of deciding correctly on H_1 if H_1 is true, and is called the *true positive* rate. It can be written

$$\text{Power}(\theta_1) \;=\; 1 - \beta \;=\; P(\text{choose } H_1 \mid H_1) \;=\; \int_{\mathcal{C}} f(\widetilde{x}, \theta_1)\, d\widetilde{x} \tag{6.10}$$

The power Power(θ) is usually written as a function of θ since the hypothesis H_1 is more likely to be composite. Note that both the level of significance α and the power $P(\theta_1)$ involve integrals over the same critical region $\mathcal{C}$, but with different densities.

To put these definitions in a table:

Table 6.1. Error Type and Probabilities

	What We Decide	
Which is True	H_0	H_1
H_0	OK	Type I α
H_1	Type II β	OK Power

If H_0 and/or H_1 are composite, then α, β, and the power are replaced by their worst possible values. That is, if for example

$$H_0 : X_i \approx f_0(x) \text{ for some density } f_0 \in T_0$$
$$H_1 : X_i \approx f_1(x) \text{ for some density } f_1 \in T_1$$

for two classes of densities T_0, T_1 on R, then

$$\alpha \;=\; \sup_{f_0 \in T_0} \int_{\mathcal{C}} f_0(\widetilde{x})\, d\widetilde{x}, \qquad \beta \;=\; \sup_{f_1 \in T_1} \int_{R^n - \mathcal{C}} f_1(\widetilde{x})\, d\widetilde{x}$$

and

$$\text{Power} \;=\; \inf_{f_1 \in T_1} \int_{\mathcal{C}} f_1(\widetilde{x})\, d\widetilde{x}$$

6.4. The Neyman-Pearson Lemma. As suggested earlier, a standard approach is to choose a highest acceptable false positive rate α (for rejecting H_0) and restrict ourselves to tests $\mathcal{C}$ with that false positive rate or smaller.

Among this class of tests, we would like to find the test that has the highest probability of detecting H_1 when H_1 is true. This is called (reasonably enough) the most powerful test of H_0 against H_1 among tests $\mathcal{C}$ of a given size or smaller.

Assume for simplicity that H_0 and H_1 are both simple hypotheses, so that

$$H_0 : X_i \approx f(x, \theta_0) \qquad \text{and} \qquad H_1 : X_i \approx f(x, \theta_1) \tag{6.11}$$

where $X_i \approx f(x)$ means that the observations X_i are independently chosen from the density $f(x)$ and $f(x,\theta)$ is a family of probability densities. As mentioned above, both the size and power of a test $\mathcal{C} \subseteq R^n$ can be expressed as n-dimensional integrals over $\mathcal{C}$:

$$\text{Size}(\mathcal{C}) \;=\; \int_C f(\widetilde{x},\theta_0)\,d\widetilde{x} \quad\text{and}\quad \text{Power}(\mathcal{C}) \;=\; \int_C f(\widetilde{x},\theta_1)\,d\widetilde{x} \tag{6.12}$$

The next result uses (6.12) to find the most powerful tests of one simple hypothesis against another at a fixed level of significance α.

Theorem 6.1. (Neyman-Pearson Lemma) Assume that the set

$$\mathcal{C}_0 \;=\; \mathcal{C}_0(\lambda) \;=\; \left\{ \widetilde{x} \in R^n : \frac{f(\widetilde{x},\theta_1)}{f(\widetilde{x},\theta_0)} \;\ge\; \lambda \right\} \tag{6.13}$$

has $\text{Size}(\mathcal{C}_0) = \alpha$ for some constant $\lambda > 0$. Then

$$\text{Power}(\mathcal{C}) \;\le\; \text{Power}\big(\mathcal{C}_0(\lambda)\big) \tag{6.14}$$

for any other subset $\mathcal{C} \subseteq R^n$ with $\text{Size}(\mathcal{C}) \le \alpha$.

Remarks (1). This means that $\mathcal{C}_0(\lambda)$ is the most powerful test of H_0 against H_1 with size $\text{Size}(\mathcal{C}) \le \alpha$.

(2). If $\widetilde{x} = X$ for data $X = (X_1,\ldots,X_n)$, then the ratio in (6.13)

$$L(\widetilde{x},\theta_1,\theta_0) \;=\; \frac{f(\widetilde{x},\theta_1)}{f(\widetilde{x},\theta_0)} \;=\; \frac{L(\theta_1,X)}{L(\theta_0,X)} \tag{6.15}$$

is a ratio of likelihoods. In this sense, the tests $\mathcal{C}_0(\lambda)$ in Theorem 6.1 are *likelihood-ratio* tests.

(3). Suppose that the likelihood $L(\theta,X) = f(X_1,\theta)\ldots f(X_n,\theta)$ has a sufficient statistic $S(X) = S(X_1,\ldots,X_n)$. That is,

$$L(\theta,X) = f(X_1,\theta)\ldots f(X_n,\theta) = g\big(S(X),\theta\big)A(X) \tag{6.16}$$

Then, since the factors $A(\widetilde{x})$ cancel out in the likelihood ratio, the most-powerful tests

$$\mathcal{C}_0(\lambda) \;=\; \left\{ \widetilde{x} \in R^n : \frac{f(\widetilde{x},\theta_1)}{f(\widetilde{x},\theta_0)} \;\ge\; \lambda \right\} \;=\; \left\{ \widetilde{x} \in R^n : \frac{g\big(S(\widetilde{x}),\theta_1\big)}{g\big(S(\widetilde{x}),\theta_0\big)} \;\ge\; \lambda \right\}$$

depend only on the sufficient statistic $S(X)$.

(4). By (6.12) and (6.15)

$$\text{Size}(\mathcal{C}) = \int_C f(\widetilde{x}, \theta_0)\, d\widetilde{x} \quad \text{and} \quad \text{Power}(\mathcal{C}) = \int_C L(\widetilde{x}, \theta_1, \theta_0) f(\widetilde{x}, \theta_0)\, d\widetilde{x}$$

for $L(\widetilde{x}, \theta_1, \theta_0)$ in (6.15). Intuitively, the set $\mathcal{C}$ that maximizes Power($\mathcal{C}$) subject to Size($\mathcal{C}$) $\le \alpha$ should be the set of size α with the largest values of $L(\widetilde{x}, \theta_1, \theta_0)$. This is essentially the proof of Theorem 6.1 given below.

Before giving a proof of Theorem 6.1, let's give some examples.

Example (1). Continuing the example (6.1)–(6.2) where $f(x, \theta)$ is the normal density $N(\theta, \sigma_0^2)$, the joint density (or likelihood) is

$$\begin{aligned} f(X_1, \dots, X_n, \theta) &= \frac{1}{\sqrt{2\pi\sigma_0^2}^{\,n}} \exp\left(-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (X_i - \theta)^2 \right) \\ &= C_1(\theta, \sigma_0, n) \exp\left(-\frac{1}{2\sigma_0^2} \left(\sum_{i=1}^n X_i^2 - 2\theta \sum_{i=1}^n X_k \right) \right) \end{aligned}$$

Since the factor containing $\sum_{i=1}^n X_i^2$ is the same in both likelihoods, the likelihood ratio is

$$\frac{f(X_1, \dots, X_n, \theta_1)}{f(X_1, \dots, X_n, \theta_0)} = C_2 \exp\left(\frac{(\theta_1 - \theta_0)}{\sigma_0^2} \sum_{j=1}^n X_j \right) \tag{6.17}$$

where $C_2 = C_2(\theta_1, \theta_0, \sigma_0, n)$. If $\theta_0 < \theta_1$ are fixed, the likelihood-ratio sets $\mathcal{C}_0(\lambda)$ in (6.13) are

$$\mathcal{C}_0(\lambda) = \left\{ \widetilde{x} : \ C_2 \exp\left(\frac{(\theta_1 - \theta_0)}{\sigma_0^2} \sum_{j=1}^n x_j \right) \ge \lambda \right\} \tag{6.18a}$$

$$= \left\{ \widetilde{x} : \ \frac{1}{n} \sum_{i=1}^n x_i \ge \lambda_m \right\} \tag{6.18b}$$

where λ_m is a monotonic function of λ. Thus the most powerful tests of H_0 against H_1 for any $\theta_1 > \theta_0$ are tests of the form $\overline{X} \ge \lambda_m$. As in (6.3), the constants $\lambda_m = \lambda_m(\alpha)$ are determined by

$$\text{Size}\big(\mathcal{C}(\lambda)\big) = \alpha = P_{\theta_0}\big(\overline{X} \ge \lambda_m(\alpha)\big)$$

Since $X_i \approx N(\theta_0, \sigma_0^2)$ and $\overline{X} \approx N(\theta_0, \sigma_0^2/n)$, this implies

$$\lambda_m(\alpha) \;=\; \theta_0 + \frac{\sigma_0}{\sqrt{n}} z_\alpha \tag{6.19}$$

where $P(Z \geq z_\alpha) = \alpha$. In particular, the most powerful test $\mathcal{C} \subseteq R^n$ of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ at level of significance α or smaller is $\mathcal{C} = \{\, x : \overline{x} \geq \lambda_m(\alpha) \,\}$ for $\lambda_m(\alpha)$ in (6.19).

Since exactly the same tests are most powerful for any $\theta_1 > \theta_0$, the test (6.18b) are called *uniformly most powerful* (UMP) for all $\theta_1 > \theta_0$. Note that $\lambda_m(\alpha)$ in (6.19) depends on θ_0 but not on θ_1. In this sense, these tests are also most powerful for $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$.

Example (2). Let $f(x, \theta) = \theta x^{\theta - 1}$ for $0 \leq x \leq 1$ and $\theta > 0$. Suppose that we want to test $H_0 : \theta = 1$ against $H_1 : \theta = \theta_1$ for some $\theta_1 > 1$. If $\theta = 1$, random variables X with distribution $f(x, \theta)$ are uniformly distributed in $(0, 1)$, while if $\theta_1 > 1$ a sample $X_1, X_2, \ldots, X_n$ will tend to be more bunched towards $x = 1$. We would like to find the most powerful test for detecting this, at least against the alternative $\theta x^{\theta-1}$ for $\theta > 1$.

The joint density here is

$$f(\widetilde{x}, \theta) \;=\; \prod_{j=1}^{n} f(x_j, \theta) \;=\; \prod_{j=1}^{n} \theta x_j^{\theta-1} \;=\; \theta^n \Big(\prod_{j=1}^{n} x_j \Big)^{\theta - 1}$$

In general if $\theta_0 < \theta_1$, the likelihood ratio is

$$\frac{f(\widetilde{x}, \theta_1)}{f(\widetilde{x}, \theta_0)} \;=\; \frac{\theta_1^n \Big(\prod_{j=1}^{n} x_j \Big)^{\theta_1 - 1}}{\theta_0^n \Big(\prod_{j=1}^{n} x_j \Big)^{\theta_0 - 1}} \;=\; C \Big(\prod_{j=1}^{n} x_j \Big)^{\theta_1 - \theta_0} \tag{6.20}$$

for $C = C(\theta_0, \theta_1, n)$. Thus the most powerful tests of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ for $\theta_0 < \theta_1$ are

$$\mathcal{C}_0(\lambda) \;=\; \left\{ \widetilde{x} : \; C \Big(\prod_{j=1}^{n} x_j \Big)^{\theta_1 - \theta_0} \geq \lambda \right\} \tag{6.21a}$$

$$= \Big\{ \widetilde{x} : \; \prod_{j=1}^{n} x_j \;\geq\; \lambda_m \Big\} \tag{6.21b}$$

where λ_m is a monotonic function of λ.

Note that the function $\lambda_m = \lambda_m(\alpha)$ in (6.21b) depends on H_0 but not on H_1. Thus the tests (6.21b) are UMP for $\theta_1 > \theta_0$ as in Example 1.

Exercise. For $H_0 : \theta = \theta_0$, prove that the tests

$$\mathcal{C}_0(\lambda) \ = \ \Big\{ \ \widetilde{x} : \ \prod_{j=1}^{n} x_j \ \le \ \lambda_m \ \Big\} \tag{6.22}$$

are UMP against $H_1 : \theta = \theta_1$ for all $\theta_1 < \theta_0$.

6.5. P-values. The nested structure of the likelihood-ratio sets in (6.13) and (6.14)

$$\mathcal{C}(\lambda_\alpha) = \left\{ \widetilde{x} \in R^n : \frac{f(\widetilde{x},\theta_1)}{f(\widetilde{x},\theta_0)} \ \ge \ \lambda_\alpha \right\} \quad \text{where} \tag{6.23}$$

$$\text{Size}\big(\mathcal{C}(\lambda_\alpha)\big) \ = \ P\left(\left\{ X : \frac{f(X,\theta_1)}{f(X,\theta_0} \ \ge \ \lambda_\alpha \right\}\right) = \alpha$$

means that we can give a single number that describes the outcome of the tests (6.23) for all α. Specifically, let

$$P \ = \ P\left(\frac{f(X,\theta_1)}{f(X,\theta_0)} \ \ge \ T_0 \ \middle| \ H_0 \right) \tag{6.24}$$

where $T_0 = T_0(X) = f(X,\theta_1)/f(X,\theta_0)$ for the observed sample. Note that the X in (6.24) is random with distribution H_0, but the X in $T_0(X)$ is the observed sample and assumed constant. Then

Lemma 6.1. Suppose that $X = (X_1,\ldots,X_n)$ is an independent sample with density $f(x,\theta)$. Suppose that we can find constants λ_α such that (6.23) holds for all α, $0 < \alpha < 1$. Define P by (6.24).

Then, if $P < \alpha$, the observed $X \in \mathcal{C}(\lambda_\alpha)$ and we reject H_0. If $P > \alpha$, then the observed $X \notin \mathcal{C}(\lambda_\alpha)$ and we accept H_0.

Proof. If $P < \alpha$, then the observed $T_0(X) > \lambda_\alpha$ by (6.23) and (6.24), and thus $X \in \mathcal{C}(\lambda_\alpha)$. Hence we reject H_0. If $P > \alpha$, then the observed $T_0(X) < \lambda_\alpha$ by the same argument and $X \notin \mathcal{C}(\lambda,\alpha)$. Hence in this case we accept H_0.

We still need to prove Theorem 6.1:

Proof of Theorem 6.1. For $\mathcal{C}_0$ in (6.13) and an arbitrary test $\mathcal{C} \subseteq R^n$ with Size$(\mathcal{C}) \le \alpha$, let $A = \mathcal{C}_0 - \mathcal{C}$ and $B = \mathcal{C} - \mathcal{C}_0$. Then by (6.12)

$$\begin{aligned} \text{Power}(\mathcal{C}_0) - \text{Power}(\mathcal{C}) \ &= \ \int_{\mathcal{C}_0} f(\widetilde{x},\theta_1)d\widetilde{x} \ - \ \int_{\mathcal{C}} f(\widetilde{x},\theta_1)d\widetilde{x} \\ &= \ \int_A f(\widetilde{x},\theta_1)d\widetilde{x} \ - \ \int_B f(\widetilde{x},\theta_1)d\widetilde{x} \end{aligned} \tag{6.25}$$

by subtracting the integral over $\mathcal{C}_0 \cap \mathcal{C}$ from both terms.

By the definition in (6.13), $f(\widetilde{x},\theta_1) \ge \lambda f(\widetilde{x},\theta_0)$ on $A \subseteq \mathcal{C}_0$ and $f(\widetilde{x},\theta_1) < \lambda f(\widetilde{x},\theta_0)$ on $B \subseteq R^n - \mathcal{C}_0$. Thus by (6.25)

$$\begin{aligned}\mathrm{Power}(\mathcal{C}_0) - \mathrm{Power}(\mathcal{C}) &\ \ge\ \int_A \lambda f(\widetilde{x},\theta_0)\,d\widetilde{x}\ -\ \int_B \lambda f(\widetilde{x},\theta_0)\,d\widetilde{x} \\ &\ =\ \int_{\mathcal{C}_0} \lambda f(\widetilde{x},\theta_0)\,d\widetilde{x}\ -\ \int_{\mathcal{C}} \lambda f(\widetilde{x},\theta_0)\,d\widetilde{x}\end{aligned}$$

by adding the integral of $\lambda f(\widetilde{x},\theta_0)$ over $\mathcal{C}_0 \cap \mathcal{C}$ to both integrals. Hence again by (6.12)

$$\mathrm{Power}(\mathcal{C}_0) - \mathrm{Power}(\mathcal{C})\ \ge\ \lambda\big(\mathrm{Size}(\mathcal{C}_0)\ -\ \mathrm{Size}(\mathcal{C})\big)\ \ge\ 0$$

since $\lambda > 0$ and $\mathrm{Size}(\mathcal{C}) \le \mathrm{Size}(\mathcal{C}_0) = \alpha$ by assumption. Thus $\mathrm{Power}(\mathcal{C}) \le \mathrm{Power}(\mathcal{C}_0)$, which completes the proof of (6.14) and hence of Theorem 6.1.

7. Generalized Likelihood Ratio Tests. Suppose that an independent sample of observations $X_1, \dots, X_n$ are known to have density $f(x,\theta)$ for some unknown (vector) parameter $\theta \in R^m$, and that we want to test the hypothesis

$$H_0 : \theta \in \Theta_0 \qquad \text{against} \qquad H_1 : \theta \in \Theta_1 \tag{7.1}$$

for subsets $\Theta_0, \Theta_i \subseteq R^m$. Some examples are

Example (1). Assume X_j is uniformly distributed $U(0,\theta)$ for some unknown θ. That is, X_j have density $f(x,\theta) = (1/\theta)I_{(0,\theta)}(x)$ and test

$$H_0 : \theta = 1 \qquad \text{against} \qquad H_1 : \theta < 1 \tag{7.2a}$$

This is of the form (7.1) with $\Theta_0 = \{\,1\,\}$ and $\Theta_1 = (0,1)$.

(2). Assume X_j are normally distributed $N(\mu,\sigma^2)$ and we want to test

$$H_0 : \mu = \mu_0 \qquad \text{against} \qquad H_1 : \mu \ne \mu_0 \tag{7.2b}$$

without any restrictions on σ. This is of the form (7.1) with $\Theta_0 = \{\,(\mu_0,\sigma^2)\,\}$, which is a half-line in R^2, and $\Theta_1 = \{\,(\mu,\sigma^2) : \mu \ne \mu_0\,\}$, which is a half-plane minus a half-line.

(3). Assume $X_1, \dots, X_{n_1}$ are independent normal $N(\mu_1,\sigma_1^2)$ and $Y_1, \dots, Y_{n_2}$ are $N(\mu_2,\sigma_2^2)$ and that we want to test

$$\begin{aligned}&H_0 : \mu_1 = \mu_2 \text{ and } \sigma_1^2 = \sigma_2^2 \qquad \text{against} \\ &H_1 : \text{All other } (\mu_1,\mu_2,\sigma_1^2,\sigma_2^2)\end{aligned} \tag{7.2c}$$

In this case, $\Theta_0 = \{\,(\mu,\mu,\sigma^2,\sigma^2)\,\}$ is a two-dimensional subset of R^4 and Θ_1 is four-dimensional.

In any of these examples, if we choose particular $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$ ($\theta_1 \neq \theta_0$) and wanted to test

$$H_0 : \theta = \theta_0 \qquad \text{against} \qquad H_1 : \theta = \theta_1 \tag{7.3}$$

then, by the Neyman-Pearson Lemma (Theorem 6.1 in Section 6.5 above), the most powerful test of H_0 against H_1 at any level of significance α is the likelihood-ratio set

$$\mathcal{C}_{\lambda_\alpha} \;=\; \left\{ \widetilde{x} \in R^n : \frac{f(\widetilde{x},\theta_1)}{f(\widetilde{x},\theta_0)} \;\geq\; \lambda_\alpha \right\} \tag{7.4}$$

where $\widetilde{x} = (x_1,\dots,x_n)$. That is, we reject H_0 in favor of H_1 if $X = (X_1,\dots,X_n) \in \mathcal{C}_{\lambda_\alpha}$.

The idea behind *generalized* likelihood-ratio tests (abbreviated GLRTs) is that we use the likelihood-ratio test (7.4) with our best guesses for $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$. That is, we define

$$\widehat{LR}_n(X) \;=\; \frac{\max\limits_{\theta\in\Theta_1} L(\theta, X_1,\dots,X_n)}{\max\limits_{\theta\in\Theta_0} L(\theta, X_1,\dots,X_n)} \;=\; \frac{L(\widehat{\theta}_{H_1}(X), X_1,\dots,X_n)}{L(\widehat{\theta}_{H_0}(X), X_1,\dots,X_n)} \tag{7.5}$$

where $\widehat{\theta}_{H_1}(X)$ and $\widehat{\theta}_{H_1}(X)$ are the maximum-likelihood estimates for $\theta \in \Theta_0$ and $\theta \in \Theta_1$, respectively. Note that $\widehat{LR}_n(X)$ depends on X but not on θ (except indirectly from the sets Θ_0 and Θ_1). We then use the tests with critical regions

$$\mathcal{C}_{\lambda_\alpha} \;=\; \left\{ \widetilde{x} \in R^n : \widehat{LR}_n(\widetilde{x}) \;\geq\; \lambda_\alpha \right\} \quad \text{with} \quad \text{Size}(\mathcal{C}_{\lambda_\alpha}) = \alpha \tag{7.6}$$

Since the maximum likelihood estimates $\widehat{\theta}_{H_1}(X), \widehat{\theta}_{H_1}(X)$ in (7.5) depend on $X = (X_1,\dots,X_n)$, the Neyman-Pearson lemma does not guarantee that (7.6) provides the most powerful tests. However, the asymptotic consistency of the MLEs (see Theorem 5.1 above) suggests that $\widehat{\theta}_{H_0}, \widehat{\theta}_{H_1}$ may be close to the "correct" values.

Warning: Some statisticians, such as the authors of the textbook Larsen and Marx, use an alternative version of the likelihood ratio

$$\widehat{LR}_n^{\text{alt}}(X) \;=\; \frac{\max\limits_{\theta\in\Theta_0} L(\theta, X_1,\dots,X_n)}{\max\limits_{\theta\in\Theta_1} L(\theta, X_1,\dots,X_n)} \;=\; \frac{L(\widehat{\theta}_{H_0}(X), X_1,\dots,X_n)}{L(\widehat{\theta}_{H_1}(X), X_1,\dots,X_n)} \tag{7.7}$$

with the maximum for H_1 in the denominator instead of the numerator and the maximum for H_0 in the numerator instead of the denominator. One then tests for small values of the GLRT statistic instead of large values. Since $\widehat{LR}_n^{\rm alt}(X) = 1/\widehat{LR}_n(X)$, the critical tests for (7.7) are

$$\begin{aligned} \mathcal{C}_{\lambda_\alpha}^{\rm alt} &= \Big\{ \widetilde{x} \in R^n : \widehat{LR}_n^{\rm alt}(\widetilde{x}) \ \le \ \lambda_\alpha \Big\} \qquad (7.8) \\ &= \Big\{ \widetilde{x} \in R^n : \widehat{LR}_n(\widetilde{x}) \ \ge \ 1/\lambda_\alpha \Big\} \quad \text{with} \quad \text{Size}(\mathcal{C}_{\lambda_\alpha}^{\rm alt}) = \alpha \end{aligned}$$

Thus the critical regions (7.8) are exactly the same as those for (7.6) except for a transformation $\lambda_\alpha \to 1/\lambda_\alpha$ and the direction of the inequality. However, one has to be aware of small differences between how the tests as described.

7.1. Examples. In Example 1 (see (7.2a) above), $\Theta_0 = \{1\}$ and $\Theta_1 = (0,1)$. This corresponds to the hypotheses $H_0 : \theta = 1$ and $H_1 : \theta < 1$ where $X_1, \ldots, X_n$ are $U(0,\theta)$. Then one can show

$$\widehat{LR}_n(X) = (1/X_{\rm max})^n, \qquad X_{\rm max} = \max_{1 \le j \le n} X_j$$

(Argue as in Section 6.5 in the text, Larsen and Marx.) The GLRTs (7.6) in this case are equivalent to

$$\begin{aligned} \mathcal{C}_{\lambda_\alpha}(X) &= \{ X : \widehat{LR}_n(X) \ge \lambda_\alpha \} = \{ X : (1/X_{\rm max})^n \ge \lambda_\alpha \} \qquad (7.9) \\ &= \{ X : X_{\rm max} \le \mu_\alpha \} \quad \text{where} \quad P(X_{\rm max} \le \mu_\alpha \mid H_0) = \alpha \end{aligned}$$

In Example 2 (see (7.2b) above), $\Theta_0 = \{ (\mu_0, \sigma^2) : \mu = \mu_0 \}$ and $\Theta_1 = \{ (\mu, \sigma^2) : \mu \ne \mu_0 \}$. This corresponds to the hypotheses $H_0 : \mu = \mu_0$ and $H_1 : \mu \ne \mu_0$ where $X_1, \ldots, X_n$ are $N(\mu, \sigma^2)$ with σ^2 unspecified. One can show in this case that

$$\widehat{LR}_n(X) \ = \ \left(1 \ + \ \frac{T(X)^2}{n-1} \right)^{n/2} \qquad \text{where} \qquad (7.10)$$

$$T(X) = \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sqrt{S(X)^2}}, \qquad S(X) = \frac{1}{n-1} \sum_{j=1}^n (X_j - \overline{X})^2$$

(See Appendix 7.A.4, pages 519–521, in the textbook, Larsen and Marx. They obtain (7.10) with $-n/2$ in the exponent instead of $n/2$ because they use (7.7) instead of (7.5) to define the GLRT statistic, which is $\widehat{LR}_n(X)$ here but λ in their notation.)

Since $\widehat{LR}_n(X)$ is a monotonic function of $|T(X)|$, the GLRT test (7.6) is equivalent to

$$\mathcal{C}_{\mu_\alpha}(X) = \{ X : |T(X)| \ge \mu_\alpha \} \quad \text{where} \quad P\big(|T(X)| \ge \mu_\alpha \ \big| \ H_0 \big) = \alpha \quad (7.11)$$

This is the same as the classical two-sided one-sample Student-t test.

There is a useful large sample asymptotic version of the GLRT, for which it is easy to find the critical values λ_α. First, we need a reformulation of the problem $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$.

7.2. Nested Hypothesis Tests. Each of Examples 1–3 can be reformulated as

$$H_0 : \theta \in \Theta_0 \qquad \text{against} \qquad H_1 : \theta \in \Theta_1 \tag{7.12}$$

where $\Theta_0 \subseteq \Theta_1$ and Θ_1 has more "free parameters" than H_0.

Equivalently, "more free parameters" means $m_0 < m_1$, where $m_0 = \dim(\Theta_0)$ and $m_1 = \dim(\Theta_1)$ are the "local dimensions" of Θ_0 and Θ_1. In Example 1, $m_0 = 0$ (a point has "no free parameters" and is "of dimension 0") and $m_1 = 1$, where we now take $\Theta_1 = (0, 1]$. In Example 2, $m_1 = 1$ (there is one free parameter, $\sigma^2 > 0$), $m_2 = 2$, and $\Theta_2 = \{\,(\mu, \sigma^2)\,\}$ is now a full half-plane. In Example 3, $m_1 = 2$ and $m_2 = 4$, since the sets $\Theta_0, \Theta_1 \subseteq R^4$.

Since $\Theta_0 \subseteq \Theta_1$, a test of H_0 against H_1 cannot be of the form "either-or" as in Section 6, since $\theta \in \Theta_0$ implies $\theta \in \Theta_1$. Instead, we view (7.12) with $\Theta_0 \subseteq \Theta_1$ as a test of whether we really need the additional $d = m_1 - m_0$ parameter or parameters. That is, if the data $X = (X_1, \ldots, X_n)$ does not fit the hypothesis H_1 sufficiently better than H_0 (as measured by the relative size of the fitted likelihoods in (7.5)) to provide evidence for rejecting H_0. Then, to be conservative, we accept H_0 and conclude that there is not enough evidence for the more complicated hypothesis H_1.

A test of the form (7.12) with $\Theta_0 \subseteq \Theta_1$ is called a *nested hypothesis test*. Note that, if $\Theta_0 \subseteq \Theta_1$, then (7.5) implies that $\widehat{LR}_n(X) \geq 1$.

Under the following assumptions for a nested hypothesis test, we have the following general theorem. Assume as before that $X_1, \ldots, X_n$ is an independent sample with density $f(x, \theta)$ where $f(x, \theta)$ satisfies the conditions of the Cramér-Rao lower bound (Theorem 4.1 in Section 4 above) and of the asymptotic normality of the MLE (Theorem 5.1 in Section 5 above). Then we have

Theorem 7.1. ("Twice the Log-Likelihood Theorem") Under the above assumptions, assume that $\Theta_0 \subseteq \Theta_1$ in (7.12), that $d = m_1 - m_0 > 0$, and that the two maximum-likelihood estimates $\widehat{\theta}_{H_0}(X)$ and $\widehat{\theta}_{H_1}(X)$ in (7.5) are attained in the interior of the sets Θ_0 and Θ_1, respectively. Then, for $\widehat{LR}_n(X)$ in (7.5),

$$\lim_{n \to \infty} P\left(2 \log\bigl(\widehat{LR}_n(X)\bigr) \leq y \;\middle|\; H_0\right) \;=\; P\left(\chi^2_d \;\leq\; y\right) \tag{7.13}$$

for $y \geq 0$, where χ^2_d represents a random variable with a χ^2 distriution with $d = m_1 - m_0$ degrees of freedom.

Proof. The proof is similar to the proof of Theorem 5.1 in Section 5, but uses an m-dimensional central limit theorem for vector-valued random variables in R^m and Taylor's Theorem in R^m instead of in R^1.

Remarks (1). The analog of Theorem 7.1 for the alternative ("upside-down") definition of $\widehat{LR}_n(X)$ in (7.7) has $-2\log\bigl(\widehat{LR}_n(X)^{\mathrm{alt}}\bigr)$ instead of $2\log\bigl(\widehat{LR}_n(X)\bigr)$.

(2). There is no analog of Theorem 7.1 if the hypotheses H_0 and H_1 are not nested. Finding a good asymptotic test for general non-nested composite hypotheses is an open question in Statistics that would have many important applications.

8. Fisher's Meta-Analysis Theorem. Suppose that we are interested in whether we can reject a hypothesis H_0 in favor of a hypothesis H_1. Assume that six different groups have carried out statistical analyses based on different datasets with mixed results. Specifically, assume that they have reported the six P-values (as in Section 6.5 above)

$$0.06 \qquad 0.02 \qquad 0.13 \qquad 0.21 \qquad 0.22 \qquad 0.73 \tag{8.1}$$

While only one of the six groups rejected H_0 at level $\alpha = 0.05$, and that was with borderline significance ($0.01 < P < 0.05$), five of the six P-values are rather small. Is there a way to assign an aggregated P-value to the six P-values in (8.1)? After reading these six studies (and finding nothing wrong with them), should we accept H_0 or reject H_1 at level $\alpha = 0.05$?

The first step is to find the random distribution of P-values that independent experiments or analyses of the same true hypothesis H_0 should attain. Suppose that each experimenter used a likelihood-ratio test of the Neyman-Pearson form (6.23) or GLRT form (7.6) where it is possible to find a value λ_α for each α, $0 < \alpha < 1$.

Lemma 8.1. Under the above assumptions, assuming that the hypothesis H_0 is true, the P-values obtained by random experiments are uniformly distributed in $(0, 1)$.

Proof. Choose α with $0 < \alpha < 1$. Since α is the false positive rate, the fraction of experimenters who reject H_0, and consequently have $P < \alpha$ for their computed P-value, is α. In other words, treating their P-values as observations of a random variable $\widetilde{P}$, then $P\bigl(\widetilde{P} < \alpha\bigr) = \alpha$ for $0 < \alpha < 1$. This means that $\widetilde{P}$ is uniformly distributed in $(0, 1)$.

Given that the numbers in (8.1) should be uniformly distributed if H_0 is true, do these numbers seem significantly shifted towards smaller values, as they might be if H_1 were true? The first step towards answering this is to find a reasonable alternative distribution of the P-values given H_1.

Fisher most likely considered the family of distributions $f(p,\theta) = \theta p^{\theta-1}$ for $0 < \theta \le 1$, so that H_0 corresponds to $\theta = 1$. For $\theta < 1$, not only is $E(P) = \theta/(\theta+1) < 1$, but the density $f(p,\theta)$ has an infinite cusp at $\theta = 0$. For this family, the likelihood of random P-values $P_1, \ldots, P_n$ given θ is

$$L(\theta, P_1, \ldots, P_n) = \prod_{j=1}^{n} f(\theta, P_j) = \theta^n \left(\prod_{j=1}^{n} P_j \right)^{\theta-1}$$

Thus $Q = \prod_{j=1}^{n} P_j$ is a sufficient statistic for θ, and we have at least a single number to summarize the six values in (8.1).

Morever, it follows as in Example 2 in Section 6.4 above that tests of the form $\{\, P : \prod_{j=1}^{n} P_j \le \lambda_\alpha \,\}$ are UMP for $H_0 : \theta = 1$ against the alternatives $\theta < 1$. (See Exercise (6.22).) The distribution of $\prod_{j=1}^{n} P_j$ given $\theta = 1$ can be obtained from

Lemma 8.2. Assume that U is uniformly distributed in $(0,1)$. Then

(a) The random variable $Y = -A \log U = A \log(1/U)$ has an exponential distribution with rate $1/A$.

(b) $Y = 2\log(1/U) \approx \chi^2_2$ (that is, Y has a chi-square distribution with 2 degrees of freedom).

(c) If $U_1, U_2, \ldots, U_n$ are independent and uniformly distributed in $(0,1)$, then $Q = \sum_{j=1}^{n} 2\log(U_j) \approx \chi^2_{2n}$ has a chi-square distribution with $2n$ degrees of freedom.

Proof. (a) For $A, t > 0$

$$\begin{aligned} P(Y > t) \;&=\; P\big(-A\log(U) > t\big) \;=\; P\big(\log(U) < -t/A\big) \\ &=\; P\big(U < \exp(-t/A)\big) \;=\; \exp(-t/A) \end{aligned}$$

This implies that Y has a probability density $f_Y(t) = -(d/dt)\exp(-t/A) = (1/A)\exp(-t/A)$, which is exponential with rate A.

(b) A χ^2_d distribution is gamma$(d/2, 1/2)$, so that $\chi^2_2 \approx$ gamma$(1, 1/2)$. By the form of the gamma density, gamma$(1,\beta)$ is exponential with rate β. Thus, by part (a), $2\log(1/U) \approx$ gamma$(1,1/2) \approx \chi^2_2$.

(c) Each $2\log(1/P_j) \approx \chi^2_2$, which implies that

$$Q \;=\; \sum_{j=1}^{n} 2\log(1/P_j) \;\approx\; \chi^2_{2n} \tag{8.2}$$

Putting these results together,

Theorem 8.1 (Fisher). Assume independent observations $U_1, \ldots, U_n$ have density $f(x, \theta) = \theta x^{\theta-1}$ for $0 \le x \le 1$, and in particular are independent and uniformly distributed in $(0, 1)$ if $\theta = 1$. Then, the P-value of the UMP test for $H_0 : \theta = 1$ against $H_1 : \theta < 1$ is

$$P \;=\; P(\chi^2_{2n} \;\ge\; Q_0)$$

where Q_0 is the observed value of $2\sum_{j=1}^n \log(1/U_j)$ and χ^2_{2n} represents a chi-square distribution with $2n$ degrees of freedom.

Proof. By Lemma 8.1 and (8.2).

Example. The numbers $P_1, \ldots, P_6$ in (8.1) satisfy

$$\sum_{j=1}^{6} 2\log(1/P_j) \;=\; 5.63 + 7.82 + 4.08 + 3.12 + 3.03 + 0.63 \;=\; 24.31$$

Thus the P-value in Theorem 8.1 is

$$P \;=\; P(\chi^2_{12} \ge 24.31) \;=\; 0.0185$$

Thus the net effect of the six tests with P-values in (8.1) is $P = 0.0185$, which is significant at $\alpha = 0.05$ but not at $\alpha = 0.01$.

9. Two Contingency-Table Tests. Consider the following contingency table for $n = 1033$ individuals with two classifications A and B:

Table 9.1. A Contingency Table for *A* and *B*

		B: 1	2	3	4	5	6	Sums:
	1	29	11	95	78	50	47	310
A:	2	38	17	106	105	74	49	389
	3	31	9	60	49	29	28	206
	4	17	13	35	27	21	15	128
Sums:		115	50	296	259	174	139	1033

It is assumed that the data in Table 9.1 comes from independent observations $Y_i = (A_i, B_i)$ for $n = 1033$ individuals, where A_i is one of $1, 2, 3, 4$ and B_i is one of $1, 2, 3, 4, 5, 6$. Rather than write out the $n = 1033$ values, it is more convenient to represent the data as 24 counts for the 4×6 possible A, B values, as we have done in Table 9.1.

Suppose we want to test the hypothesis that the Y_i are sampled from a population for which A and B are independent. (Sometimes this hypothesis is stated that "rows and columns" are independent, but this doesn't make very much sense if you analyze it closely.)

If the sample is homogeneous, each observation $Y_i = (A_i, B_i)$ has a multivariate Bernoulli distribution with probability function $P\big(Y = (a,b)\big) = p_{ab}$ for $1 \le a \le s$ and $1 \le b \le t$, where $s = 4$ is the number of rows in Table 9.1 and $t = 6$ is the number of columns, and $\sum_{a=1}^{s}\sum_{b=1}^{t} p_{ab} = 1$. If the random variables A and B are independent, then $P\big(Y = (a,b)\big) = P(A = a)P(B = b)$. If $P(A = a) = p_a^A$ and $P(B = b) = p_b^B$, then $p_{ab} = p_a^A p_b^B$. This suggests the two nested hypotheses

$$H_1 : p_{ab} > 0 \text{ are arbitrary subject with } \quad \sum_{a=1}^{s}\sum_{b=1}^{t} p_{ab} = 1 \tag{9.1}$$

$$H_0 : p_{ab} = p_a^A\, p_b^B \quad \text{where} \quad \sum_{a=1}^{s} p_a^A = \sum_{b=1}^{t} p_b^B = 1$$

9.1. Pearson's Chi-Square Test.

We first consider the GLRT test for (9.1). Writing p for the matrix $p = p_{ab}$ ($1 \le a \le s$, $1 \le b \le t$), the likelihood of $Y = (Y_1, Y_2, \ldots, Y_n)$ is

$$L(p, Y) \;=\; \prod_{i=1}^{n} \{q_i = p_{ab} : Y_i = (a,b)\} \;=\; \prod_{a=1}^{s}\prod_{b=1}^{t} p_{ab}^{X_{ab}} \tag{9.2}$$

where X_{ab} are the counts in Table 9.1. The MLE $\widehat{p}_{H_1}$ for hypothesis H_1 can be found by the method of Lagrange multipliers by solving

$$\frac{\partial}{\partial p_{ab}} \log L(p, Y) = 0 \quad \text{subject to} \quad \sum_{a=1}^{s}\sum_{b=1}^{t} p_{ab} = 1$$

This leads to $(\widehat{p}_{H_1})_{ab} = X_{ab}/n$. The MLE p_{H_0} can be found similarly as the solution of

$$\frac{\partial}{\partial p_a^A} \log L(p, Y) = \frac{\partial}{\partial p_b^B} \log L(p, Y) = 0$$

$$\text{subject to} \quad \sum_{a=1}^{s} p_a^A \;=\; \sum_{b=1}^{t} p_b^B \;=\; 1$$

This implies $\widehat{p_a^A} = X_{a+}/n$ and $\widehat{p_b^B} = X_{+a}/n$ where $X_{a+} = \sum_{c=1}^t X_{ac}$ and $X_{+b} = \sum_{c=1}^s X_{cb}$. This in turn implies $(\widehat{p}_{H_0})_{ab} = (X_{a+}/n)(X_{+b}/n)$. Thus the GLRT statistic for (9.1) is

$$\widehat{LR}_n(Y) \;=\; \frac{L(\widehat{p}_{H_1}, Y)}{L(\widehat{p}_{H_0}, Y)} \;=\; \frac{\prod_{a=1}^s \prod_{b=1}^t \left(\frac{X_{ab}}{n}\right)^{X_{ab}}}{\prod_{a=1}^s \left(\frac{X_{a+}}{n}\right)^{X_{a+}} \prod_{b=1}^t \left(\frac{X_{+b}}{n}\right)^{X_{+b}}} \tag{9.3}$$

Note that hypothesis H_1 in (9.1) has $m_1 = st - 1$ free parameters, while hypothesis H_0 has $m_0 = (s-1) + (t-1)$ free parameters. The difference is $d = m_1 - m_0 = st - 1 - (s-1) - (t-1) = st - s - t + 1 = (s-1)(t-1)$. Thus by Theorem 7.1 in Section 7

$$\lim_{n\to\infty} P\Big(2\log\big(\widehat{LR}_n(X)\big) \le y \;\Big|\; H_0\Big) \;=\; P\left(\chi_d^2 \;\le\; y\right) \tag{9.4}$$

where $d = (s-1)(t-1)$. The test of H_0 against H_1 based on (9.4) is often called the G-test.

Pearson's "Sum of (Observed − Expected)^2/Expected" statistic is

$$D_n(y) \;=\; \sum_{a=1}^s \sum_{b=1}^t \frac{\left(X_{ab} - n\widehat{p}_a^A \widehat{p}_b^B\right)^2}{n\widehat{p}_a^A \widehat{p}_b^B} \;=\; \sum_{a=1}^s \sum_{b=1}^t \frac{\left(X_{ab} - (X_{a+}X_{+b}/n)\right)^2}{(X_{a+}X_{+b}/n)}$$

It was proven in class in a more general context that

$$E\Big(\big|2\log\widehat{LR}_n(Y) - D_n(Y)\big|\Big) \;\le\; \frac{C}{\sqrt{n}}$$

for $n \ge 1$. It can be show that this in combination with (9.4) implies

$$\lim_{n\to\infty} P\big(D_n(Y) \le y \;\big|\; H_0\big) \;=\; P\left(\chi_d^2 \;\le\; y\right) \tag{9.5}$$

Thus the GLRT test for H_0 within H_1 in (9.1) is asymptotically equivalent to a test on $D_n(Y)$, for which the P-value can be written asymptotically

$$P \;=\; P(\chi_d^2 \;\ge\; D_n(Y)_{\text{Obs}})$$

where "Obs" stands for "Observed value of".

For the data is Table 9.1, $D_n(Y) = 19.33$ and $P = 0.199$ for $d = (4-1)(6-1) = 15$ degrees of freedom. Thus, the data in Table 9.1 is not significant using the Pearson's chi-square test.

9.2. The Pearson Test is an Omnibus Test.

The GLRT test of (9.1) is sometimes called a test of H_0 against an "omnibus" alternative, since it is designed to have power against any alternative p_{ab} for which A and B fail to be independent.

A test that is sensitive to a particular way in which H_0 may fail can have much greater power against that alternative than an omnibus test, which much guard against any possible failure of H_0. Conversely, a test that is "tuned" towards a particular alternative can fail miserably when H_0 is false for other reasons.

The shrinkage estimator in Section 2.1 provides a somewhat similar example. If we make even a crude guess what the true mean of a normal sample might be, then a shrinkage estimator towards that value can have smaller expected squared error than the sample mean estimator, which is the minimum-variance unbiased estimator for all possible true means. Conversely, if we guess wrongly about the true mean, the shrinkage estimator may have a much larger expected squared error.

9.3. The Mantel-Haenszel Trend Test.

Suppose that one suspects that the random variables A and B in Table 9.1 are correlated as opposed to being independent. In particular, we would like a test of H_0 in (9.1) whose implicit alternative is that A, B are correlated, which may have greater power if A and B are in fact correlated. We understand that this test may have much less power against an alternative to independence in which A and B are close to being uncorrelated.

The Mantel-Haenszel trend test does exactly this. (Note: This test is also called the Mantel trend test. The "trend" is necessary here because there is a contingency table test for stratified tables that is also called the Mantel-Haenszel test.)

Specifically, let r be the sample Pearson correlation coefficient of A_i and B_i for the sample $Y_i = (A_i, B_i)$. That is,

$$r = \frac{\sum_{i=1}^n (A_i - \overline{A})(B_i - \overline{B})}{\sqrt{\sum_{i=1}^n (A_i - \overline{A})^2}\sqrt{\sum_{i=1}^n (B_i - \overline{B})^2}} \tag{9.6}$$

Recall that A_i takes on integer values with $1 \le A_i \le s$ and B_i takes on integer values with $1 \le B_i \le t$. Then

Theorem 9.1 (Mantel-Haenszel). Under the assumptions of this section, using a permutation test based on permuted the values B_i in $Y_i = (A_i, B_i)$ for $1 \le i \le n$ among themselves while holding A_i fixed,

$$\lim_{n\to\infty} P\big((n-1)r^2 \ge y \bigm| H_0\big) = P(\chi_1^2 \ge y) \tag{9.7}$$

Remarks. The limits in Theorem 7.1 and (9.4) are based on a probability space that supports independent random variables with a given probability density $f(x, \theta)$.

In contrast, the underlying probability space in Theorem 9.1, in common with permutation tests in general, is defined by a set of permutations of the data under which the distribution of a sample statistic is the same as if H_0 is true. For example, in this case, if we choose A_i at random from $A_1, \ldots, A_n$ and match it with a randomly permuted B_i at that i, then

$$P(A_i = a, B_i = b) \; = \; P(A_i = a)P(B_i = a)$$

and A_i, B_i are independent. (In contrast, B_1 and B_2 are *not* independent. If B_1 happened to be a large value, then the value B_2 at a different offset in the permuted values, conditional on B_1 already have been chosen, would be drawn from values with a smaller mean. Thus B_1 and B_2 are negatively correlated.)

Since the pairs A_i, B_i are independent in this permutation probability space, if the observed value of r in (9.6) is far out on the tail of the statistics r calculated by randomly permuted the B_i in this manner, then it is likely that the observed A_i and B_i were not chosen from a distribution in which A and B were independent. We needn't worry that the set of possible P-values is overly discrete if n is large, since in that case the number of permutations ($n!$) is truly huge. Since the test statistic (9.7) is the sample correlation itself, if we reject H_0 then it is likely that A and B are correlated.

Example. For the data in Table 9.1, the sample correlation coefficient $r = -0.071$ and $X_{\mathrm{obs}} = (n-1)r^2 = (1032)(-0.071)^2 = 5.2367$. The P-value is $P = P(\chi_1^2 \geq 5.2367) = 0.0221$. Thus Table 9.1 shows significant departure from independence by the Mantel test, but not for the standard Pearson test.

In general, one can get P-values for a χ_1^2 distribution from a standard normal table, since it is the square of a standard normal. Thus

$$\begin{aligned} P \; &= \; P(\chi_1^2 \geq 5.2367) \; = \; P(Z^2 \geq 5.2367) \\ &= \; P(|Z| \geq \sqrt{5.2367}) \; = \; 2P(Z \geq 2.2884) \; = \; 0.0221 \end{aligned}$$

Proof of Theorem 9.1 (Remarks). The usual central limit theorem is for independent random variables $X_1, \ldots, X_n, \ldots$ drawn from the same probability space. There are analogous limit theorems (also giving normal distributions) for sums and linear combinations of values in segments $X_{k+1}, \ldots, X_{k+m}$ (that is, for offsets between $k+1$ and $k+m$) of a large, randomly permuted sequence of numbers $X_1, \ldots, X_n$, assuming that $\epsilon n \leq m \leq (1-\epsilon)n$ as $n \to \infty$ for some $\epsilon > 0$.

These theorems are sometimes referred to collectively as "Hajek's Central Limit Theorem" and can be used to prove not only Theorem 9.1, but also a large variety of large-sample limit theorems in nonparametric statistics.

Remark. Even though (I believe) the Friedman of the Freidman rank-sum test (a nonparametric analog of the two-way ANOVA) is the same as the economist Friedman, the Hajek of Hajek's central limit theorem (Jaroslav Hájek, 1926–1974) is not the same person as the famous conservative economist (Friedrich von) Hajek (1899–1992). In particular, it is not true that all famous results in nonparametric statistics are named after Nobel-prize-winning economists.