Asymptotic theory

To assess the behavior of an estimator, we would like to know its exact distribution, but this is not always feasible and so we sometimes settle for a convenient approximation. One general way to obtain an approximate distribution for an estimator is the following asymptotic approach.

Let $X_1, \dots, X_n$ be a sample of size n from a distribution parametrized by a parameter vector $\boldsymbol{\theta}$. Suppose $W(X_1, \dots, X_n)$ is an estimator of a parametric function $\tau(\boldsymbol{\theta})$. In order to keep track of the sample size we sometimes write $W_n = W_n(X_1, \dots, X_n)$. Often it is natural to regard W_n as a member of a sequence $W_1, W_2, \dots, W_{n-1}, W_n, W_{n+1}, W_{n+2}, \dots$.

Example. Suppose $X_1, \dots, X_{50}$ are independent observations from a population with mean μ and standard deviation σ. We might use $W = \bar{X} = \sum_{i=1}^{50} X_i/50$ to estimate μ. Even though we actually have a sample size of 50, it is natural to think of $\bar{X}$ as a member of the sequence $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_{49}, \bar{X}_{50}, \bar{X}_{51}, \bar{X}_{52}, \dots$ where $\bar{X}_n = \sum_{i=1}^{n} X_i/n$. The limiting properties of this sequence somehow seem relevant even though we are really concerned only with the properties of $\bar{X}_{50}$. One important limiting property of this sequence is given by the Central Limit Theorem, which says that

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\text{d}} \text{Normal}(0,1) \quad \text{as } n \to \infty,$$

where the d over the arrow denotes convergence in distribution (see below). This is a limiting property (or *asymptotic* property) of the sequence as $n \to \infty$, but experience has shown that for "most" distributions we can say that $(\bar{X}_n - \mu)/(\sigma/\sqrt{n})$ is approximately distributed as Normal$(0,1)$ for n as large as 50, which is a statement that we can apply to our actual sample of data. Δ

The asymptotic properties that we will be looking at are consistency and asymptotic normality, which are defined in terms of, respectively, convergence in probability and convergence in distribution. These concepts were defined in Sections 5.2 and 5.3, but we will review them briefly here.

Review of convergence in probability

Definition 5.2.1. Let $Y_1, \dots, Y_n, \dots$ be a sequence of random variables. The sequence is said to *converge in probability* to a constant c if, for every $\epsilon > 0$,

$$\text{P}\{|Y_n - c| < \epsilon\} \longrightarrow 1 \text{ as } n \to \infty.$$

In symbols, we write $Y_n \xrightarrow{\text{p}} c$. In words, this says that for a large sample size n, there is a high probability that Y_n will be close to c.

Theorem 5.2.2. (a) If $\mathrm{E}[(Y_n - c)^2] \to 0$ as $n \to \infty$, then $Y_n \xrightarrow{\mathrm{P}} c$.
(b) If $\mathrm{E}(Y_n) \to c$ and $\mathrm{Var}(Y_n) \to 0$ as $n \to \infty$, then $Y_n \xrightarrow{\mathrm{P}} c$.
(c) If $\mathrm{E}[|Y_n - c|] \to 0$ as $n \to \infty$, then $Y_n \xrightarrow{\mathrm{P}} c$.
(d) If $\mathrm{E}[|Y_n - c|^r] \to 0$ as $n \to \infty$ for some $r > 0$, then $Y_n \xrightarrow{\mathrm{P}} c$.

Part (d) is proved in the textbook on p. 244 using the Markov Inequality. Parts (a) and (c) are the special cases $r = 2$ and $r = 1$. Part (b) follows from (a) because $\mathrm{E}[(Y_n - c)^2] = \mathrm{Var}(Y_n) + [\mathrm{E}(Y_n) - c]^2$.

Theorem 5.2.1 (Weak Law of Large Numbers). Let $X_1, \ldots, X_n, \ldots$ be i.i.d. with mean μ and standard deviation $\sigma < \infty$. Let $\bar{X}_n = \sum_{i=1}^{n} X_i / n$ be the mean of the first n observations. Then $\bar{X}_n \xrightarrow{\mathrm{P}} \mu$.

This follows from Theorem 5.2.2(b), letting $Y_n = \bar{X}_n$.

Theorem 5.2.4. If $W_n \xrightarrow{\mathrm{P}} b$ and $Y_n \xrightarrow{\mathrm{P}} c$, then (i) $W_n + Y_n \xrightarrow{\mathrm{P}} b + c$, (ii) $W_n - Y_n \xrightarrow{\mathrm{P}} b - c$, (iii) $W_n Y_n \xrightarrow{\mathrm{P}} bc$, and (iv) $W_n / Y_n \xrightarrow{\mathrm{P}} b/c$ provided that $c \neq 0$ and $\mathrm{P}\{Y_n \neq 0\} = 1$.

Theorem 5.2.5. If $Y_n \xrightarrow{\mathrm{P}} c$ and $g(y)$ is a continuous function, then $g(Y_n) \xrightarrow{\mathrm{P}} g(c)$.

Consistency

Let $X_1, \ldots, X_n, \ldots$ be a sequence of i.i.d. observations from a population whose distribution is parameterized by a parameter vector $\boldsymbol{\theta}$. Suppose $W_1, \ldots, W_n, \ldots$ is a sequence of estimators where W_n is calculated from the first n observations, that is, $W_n = W_n(X_1, \ldots, X_n)$.

Definition. W_n is a *consistent* sequence of estimators of $\tau(\boldsymbol{\theta})$ if $W_n \xrightarrow{\mathrm{P}} \tau(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$.

In words, this says that for a large sample size n, there is a high probability that W_n will be close to $\tau(\boldsymbol{\theta})$, no matter what $\boldsymbol{\theta}$ is. We sometimes simply say that W is consistent for $\tau(\boldsymbol{\theta})$, without explicitly mentioning a sequence. "Almost all" reasonable estimators (and even some unreasonable estimators) are consistent.

Theorems 5.2.1 (WLLN), 5.2.2, 5.2.4, and 5.2.5 provide tools for finding and verifying consistent estimators. The WLLN implies that $\bar{X}_n$ is consistent for $\mu(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}(X_1)$. Theorem 5.2.5 implies:

Lemma. If W_n is consistent for $\tau(\boldsymbol{\theta})$ and $g(w)$ is a continuous function, then $g(W_n)$ is consistent for $g(\tau(\boldsymbol{\theta}))$.

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. from a Normal(μ, σ^2) distribution.

(a) $\bar{X}_n$ is consistent for μ (by the WLLN).

(b) $\bar{X}_n^2$ is consistent for μ^2 (by the lemma above).

(c) $e^{\bar{X}_n}$ is consistent for e^{μ} (by the lemma above).

(d) S_n^2 is consistent for σ^2 (by Theorem 5.2.2(b), because $\mathrm{E}(S_n^2) = \sigma^2 \to \sigma^2$ and $\mathrm{Var}(S_n^2) = 2\sigma^4/(n-1) \to 0$).

(e) S_n is consistent for σ (by the lemma above because $g(w) = \sqrt{w}$ is a continuous function).

(f) $\bar{X}_n / S_n$ is consistent for μ/σ (by Theorem 5.2.4). $\triangle$

For a parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, we say that an estimator $\widehat{\boldsymbol{\theta}}_n = (\widehat{\theta}_{1n}, \ldots, \widehat{\theta}_{pn})$ is *consistent* for $\boldsymbol{\theta}$ if every component $\widehat{\theta}_{jn}$ is consistent for θ_j, for $j = 1, \ldots, p$.

Lemma. If $\widehat{\boldsymbol{\theta}}_n$ is consistent for $\boldsymbol{\theta}$ and $\tau(\boldsymbol{\theta})$ is a continuous function, then $\tau(\widehat{\boldsymbol{\theta}}_n)$ is consistent for $\tau(\boldsymbol{\theta})$.

The WLLN can be used to prove:

Theorem. Let $X_1, \ldots, X_n$ be i.i.d. with pmf or pdf $f(x;\boldsymbol{\theta})$. Assume suitable regularity conditions on $f(x;\boldsymbol{\theta})$. There is a solution $\widehat{\boldsymbol{\theta}}_n$ to the likelihood equations (see p. 47 above) that is a consistent estimator of $\boldsymbol{\theta}$.

This is a generalized version of statement (12.2.3) in Chapter 12 on Large-Sample Inference. The regularity conditions required for this theorem are A1, A2, A3 on p. 540 in the textbook. A1 is essentially RC1 and RC2 on p. 21 above. A2 is weaker than RC3 but is similar. The conditions are satisfied by all regular exponential families. For a regular exponential family it is known that there is only one solution to the likelihood equations and it is the MLE.

Corollary. Under the assumptions of the preceding theorem, with the additional assumption that the likelihood equations have a unique solution $\widehat{\boldsymbol{\theta}}_n$, this solution is a consistent estimator of $\boldsymbol{\theta}$.

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. from a Normal(μ, σ^2) distribution. The unique solutions to the likelihood equations are $\widehat{\mu} = \bar{X}$ and $\widehat{\sigma} = \sqrt{\sum (X_i - \bar{X})^2 / n}$. The corollary tells us that $\widehat{\mu}$ is consistent for μ and $\widehat{\sigma}$ is consistent for σ. In the preceding example we saw that S is consistent for σ. This makes sense because $S = \sqrt{n/(n-1)}\,\widehat{\sigma}$ and $\sqrt{n/(n-1)} \to 1$. $\triangle$

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. from a Gamma(α, β) distribution. The solutions to the likelihood equations, which are the MLEs, are not available in the form of an explicit formula. The MLE of α is given implicitly as the solution of the equation

$$\log(\widehat{\alpha}) - \frac{\mathrm{d}}{\mathrm{d}\alpha}\log\Gamma(\widehat{\alpha}) = \log\bar{X} - \frac{1}{n}\sum\log X_i$$

and then the MLE of β can be obtained as $\widehat{\beta} = \bar{X}/\widehat{\alpha}$. The corollary above tells us that $\widehat{\alpha}$ and $\widehat{\beta}$ are consistent estimators of α and β respectively. △

Review of convergence in distribution

Definition 5.3.1. Let $Y_1, \dots, Y_n, \dots$ be a sequence of random variables. The sequence is said to *converge in distribution* to a random variable Y if, for all u,

$$\mathrm{P}\{Y_n \leq u\} \longrightarrow \mathrm{P}\{Y \leq u\} \quad \text{as } n \to \infty.$$

This is the definition when Y has a continuous distribution, but when its distribution is not continuous then convergence is required only for all u at which the cdf $\mathrm{P}\{Y \leq u\}$ is continuous.

In symbols, we write $Y_n \xrightarrow{\mathrm{d}} Y$. In words, we say that for a large sample size n, the distribution of Y_n is approximately the same as the distribution of Y.

Theorem 5.3.4 (Central Limit Theorem). Let $X_1, \dots, X_n, \dots$ be i.i.d. with mean μ and standard deviation $\sigma < \infty$. Let $\bar{X}_n = \sum_{i=1}^n X_i/n$ be the mean of the first n observations. Then $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathrm{d}} \text{Normal}(0,1)$.

Less precisely, we might write, for large sample size n, $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \overset{\text{approx}}{\sim} \text{Normal}(0,1)$ or $\bar{X}_n \overset{\text{approx}}{\sim} \text{Normal}(\mu, \frac{\sigma^2}{n})$.

Theorem 5.3.3 (Slutsky's Theorem). If $W_n \xrightarrow{\mathrm{d}} W$ and $Y_n \xrightarrow{\mathrm{p}} c$, then (i) $W_n + Y_n \xrightarrow{\mathrm{d}} W + c$, (ii) $W_n - Y_n \xrightarrow{\mathrm{d}} W - c$, (iii) $W_nY_n \xrightarrow{\mathrm{d}} Wc$, and (iv) $W_n/Y_n \xrightarrow{\mathrm{d}} W/c$ provided that $c \neq 0$ and $\mathrm{P}\{Y_n \neq 0\} = 1$.

Theorem 5.3.7. If $Y_n \xrightarrow{\mathrm{d}} Y$ and $g(y)$ is a continuous function, then $g(Y_n) \xrightarrow{\mathrm{d}} g(Y)$.

Asymptotic normality

Let $X_1 \dots, X_n, \dots$ be a sequence of i.i.d. observations from a population whose distribution is parameterized by a parameter vector $\boldsymbol{\theta}$. Suppose $W_1, \dots, W_n, \dots$ is a sequence of estimators where W_n is calculated from the first n observations, that is, $W_n = W_n(X_1, \dots, X_n)$.

Definition. W_n is a *consistent asymptotically normal* (*CAN*) sequence of estimators of $\tau(\boldsymbol{\theta})$ if $\frac{W_n - \tau(\boldsymbol{\theta})}{\sigma(\boldsymbol{\theta})/\sqrt{n}} \xrightarrow{\mathrm{d}} \text{Normal}(0,1)$ for all $\boldsymbol{\theta}$ for some $\sigma(\boldsymbol{\theta})$.

We might write, for large sample size n, $W_n \overset{\text{approx}}{\sim} \text{Normal}(\tau(\boldsymbol{\theta}), \frac{\sigma^2(\boldsymbol{\theta})}{n})$. We call $\sigma^2(\boldsymbol{\theta})/n$ the *asymptotic variance* of W_n.

The Central Limit Theorem (CLT) implies that $\bar{X}_n$ is a CAN estimator of $\mu(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}(X_1)$ with asymptotic variance $\sigma^2(\boldsymbol{\theta})/n$ where $\sigma^2(\boldsymbol{\theta}) = \mathrm{Var}_{\boldsymbol{\theta}}(X_1)$.

The CLT can be used to prove the following theorem.

Theorem. Let $X_1, \ldots, X_n$ be i.i.d. with pmf or pdf $f(x\,;\theta)$, θ real-valued. Let $\widehat{\theta}_n$ be a solution of the likelihood equation that is consistent for θ (as in the theorem on p. 66 above). Let $\tau(\theta)$ be a differentiable function of θ. Under certain regularity conditions on $f(x\,;\theta)$, $\tau(\widehat{\theta}_n)$ is a CAN estimator of $\tau(\theta)$ with asymptotic variance $\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\tau(\theta)\right]^2 \Big/ \left[n\mathcal{I}_{X_1}(\theta)\right]$ where $\mathcal{I}_{X_1}(\theta)$ is the Fisher information in X_1.

This theorem generalizes statement (12.2.4) in the textbook. Note that the asymptotic variance is the same as the CRLB (see p. 58 above). A more general version of the theorem is true for a vector-valued parameter.

Theorem. Let $X_1, \ldots, X_n$ be i.i.d. with pmf or pdf $f(x\,;\boldsymbol{\theta})$. Let $\widehat{\boldsymbol{\theta}}_n$ be a solution of the likelihood equations that is consistent for $\boldsymbol{\theta}$ (as in the theorem on p. 66 above). Let $\tau(\boldsymbol{\theta})$ be a differentiable function of $\boldsymbol{\theta}$. Under certain regularity conditions on $f(x\,;\boldsymbol{\theta})$, $\tau(\widehat{\boldsymbol{\theta}}_n)$ is a CAN estimator of $\tau(\boldsymbol{\theta})$ with asymptotic variance

$$\frac{1}{n}\left[D_\tau(\boldsymbol{\theta})\right]'\left[\mathcal{I}_{X_1}(\boldsymbol{\theta})\right]^{-1}\left[D_\tau(\boldsymbol{\theta})\right]$$

where $\mathcal{I}_{X_1}(\boldsymbol{\theta})$ is the Fisher information matrix for a single observation X_1 and $D_\tau(\boldsymbol{\theta}) = \left(\frac{\partial}{\partial\theta_1}\tau(\boldsymbol{\theta}), \ldots, \frac{\partial}{\partial\theta_p}\tau(\boldsymbol{\theta})\right)'$.

Corollary. Under the assumptions of the preceding theorem, for $j = 1, \ldots, p$, the j-th component $\widehat{\theta}_{jn}$ is a CAN estimator of θ_j with asymptotic variance $\mathcal{I}_{X_1}^{jj}(\boldsymbol{\theta})/n$ where $\mathcal{I}_{X_1}^{jj}(\boldsymbol{\theta})$ denotes the (j,j) entry of the inverse of the information matrix for a single observation.

The CRLB is given on p. 58 for models with a real-valued parameter, but the formula can be extended to models with a vector-valued parameter. The formula for the CRLB coincides with the asymptotic variance displayed above. Theorem 7.5.1 (and its generalization to models with vector-valued parameters) are concerned with the exact variances of unbiased estimators and do not apply to the asymptotic variances of CAN estimators. Nevertheless, it turns out that the CRLB is the smallest possible asymptotic variance, ignoring certain artificially constructed exceptions.

Definition. W_n is an *asymptotically efficient* sequence of estimators of $\tau(\boldsymbol{\theta})$ if (i) W_n is a CAN estimator of $\tau(\boldsymbol{\theta})$ and (ii) its asymptotic variance equals the CRLB.

Thus, the preceding theorem can be restated to say that, under the assumptions of the theorem, $\tau(\widehat{\boldsymbol{\theta}}_n)$ is an asymptotically efficient estimator of $\tau(\boldsymbol{\theta})$

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. from a Normal(μ, σ^2) distribution. From p. 27 we know that the information matrix is

$$\mathcal{I}_{X_1}(\mu, \sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}.$$

The inverse of the information matrix is

$$\mathcal{I}_{X_1}^{-1}(\mu, \sigma^2) = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix}.$$

(a) By the first corollary above, the MLE $\widehat{\mu} = \bar{X}$ has the asymptotic distribution Normal$(\mu, \sigma^2/n)$. This is actually the exact distribution of $\widehat{\mu}$.
(b) By the same corollary, the MLE $\widehat{\sigma}^2 = \sum(X_i - \bar{X})^2/n$ has the asymptotic distribution Normal$(\sigma^2, 2\sigma^4/n)$. As in part (a), we do not really need the asymptotic distribution of $\widehat{\sigma}^2$ because we know its exact distribution, but this gives us the opportunity to compare the two distributions. On p. 52 above we found that E$(\widehat{\sigma}^2) = \frac{n-1}{n}\sigma^2 = (1 - \frac{1}{n})\sigma^2$ and Var$(\widehat{\sigma}^2) = \frac{2(n-1)}{n^2}\sigma^4 = \frac{2}{n}(1 - \frac{1}{n})\sigma^4$. For large n, we see that the asymptotic mean and variance are approximately equal to the exact mean and variance. The exact distribution of $\widehat{\sigma}^2$ can be obtained because $\frac{n}{\sigma^2}\widehat{\sigma}^2 = \frac{n-1}{\sigma^2}S^2 \sim \chi^2(n-1)$. By the reproductive property of Chi-squared distributions, the $\chi^2(n-1)$ distribution is the same as that of $\sum_{i=1}^{n-1} Y_i$ where the Y_i's are i.i.d. $\chi^2(1)$. So $\widehat{\sigma}^2$ has the same distribution as $\frac{\sigma^2}{n}\sum_{i=1}^{n-1} Y_i = \sigma^2 \frac{n-1}{n}\frac{1}{n-1}\sum_{i=1}^{n-1} Y_i$ $= \sigma^2(1 - \frac{1}{n})\bar{Y}_{n-1}$. As $n \to \infty$, the distribution of the sample mean $\bar{Y}_{n-1}$ becomes approximately Normal$(1, \frac{2}{n})$, by the CLT. And $1 - \frac{1}{n} \to 1$. So the distribution of $\sigma^2(1 - \frac{1}{n})\bar{Y}_{n-1}$ becomes approximately Normal$(\sigma^2, \frac{2\sigma^4}{n})$, which verifies the theorem in this case.
(c) The MLE of μ/σ is $\widehat{\mu}/\widehat{\sigma}$. Its exact distribution is not easy to deal with, but we can approximate it by its asymptotic distribution when n is large. Here we have $\boldsymbol{\theta} = (\theta_1, \theta_2) = (\mu, \sigma^2)$ and $\tau(\boldsymbol{\theta}) = \tau(\theta_1, \theta_2) = \tau(\mu, \sigma^2) = \mu/\sigma$, that is, $\tau(\theta_1, \theta_2) = \theta_1/\sqrt{\theta_2}$. The partial derivatives of $\tau(\boldsymbol{\theta})$ are $\frac{\partial \tau}{\partial \theta_1} = 1/\sqrt{\theta_2} = 1/\sigma$, $\frac{\partial \tau}{\partial \theta_2} = -\,\theta_1/(2\theta_2\sqrt{\theta_2}) = -\,\mu/(2\sigma^3)$. The asymptotic distribution of $\widehat{\mu}/\widehat{\sigma}$ is Normal with mean μ/σ and variance

$$\frac{1}{n}\begin{bmatrix} \frac{1}{\sigma} & \frac{-\mu}{2\sigma^3} \end{bmatrix} \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma} \\ \frac{-\mu}{2\sigma^3} \end{bmatrix} = \frac{1}{n}(1 + \frac{\mu^2}{2\sigma^2}).\ \triangle$$

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. from a Gamma(α, β) distribution. The MLEs of the parameters, which are the unique solutions to the likelihood equations, are not available in the form of an explicit formula, and so we have no hope of obtaining their exact distribution.

But we can obtain their asymptotic distribution. For this we need the information matrix. From p. 48 above we have

$$\begin{aligned}\frac{\partial}{\partial\alpha}\log f(X_1\,;\alpha,\beta) &= -\,\psi(\alpha)-\log\beta+\log X_1\\ \frac{\partial}{\partial\beta}\log f(X_1\,;\alpha,\beta) &= -\,\frac{\alpha}{\beta}+\frac{1}{\beta^2}X_1\,.\end{aligned}$$

where $\psi(\alpha)=\frac{\partial}{\partial\alpha}\log\Gamma(\alpha)$ is the digamma function. Now

$$\begin{aligned}\frac{\partial^2}{\partial\alpha^2}\log f(X_1\,;\alpha,\beta) &= -\,\psi'(\alpha)\\ \frac{\partial^2}{\partial\beta^2}\log f(X_1\,;\alpha,\beta) &= \frac{\alpha}{\beta^2}-\frac{2}{\beta^3}X_1\\ \frac{\partial^2}{\partial\alpha\partial\beta}\log f(X_1\,;\alpha,\beta) &= -\,\frac{1}{\beta}\,.\end{aligned}$$

The function $\psi'(\alpha)$ is called the trigamma function. The values of the gamma, digamma, and trigamma functions have been tabulated in the Handbook of Mathematical Functions edited by Abramowitz & Stegun. Since $\mathrm{E}(X_1)=\alpha\beta$, we get

$$\mathcal{I}_{X_1}(\alpha,\beta)=\begin{bmatrix}\psi'(\alpha) & \frac{1}{\beta}\\ \frac{1}{\beta} & \frac{\alpha}{\beta^2}\end{bmatrix}$$

and

$$\mathcal{I}_{X_1}^{-1}(\alpha,\beta)=\frac{1}{\alpha\psi'(\alpha)-1}\begin{bmatrix}\alpha & -\,\beta\\ -\,\beta & \beta^2\psi'(\alpha)\end{bmatrix}$$

The asymptotic distribution of $\widehat{\alpha}$ is Normal with mean α and variance $\frac{1}{n}\alpha/(\alpha\psi'(\alpha)-1)$.
The asymptotic distribution of $\widehat{\beta}$ is Normal with mean β and variance $\frac{1}{n}\beta^2\psi'(\alpha)/(\alpha\psi'(\alpha)-1)$. △

Two-sample problems

Two independent samples with unrelated parameters

Suppose X is a data vector with joint density $f(x\,;\theta)$, $\theta \in \Theta$, and Y is a data vector with joint density $g(y\,;\psi)$, $\psi \in \Psi$. Suppose X and Y are independent, and θ and ψ are unrelated. The joint density for the combined data set (X, Y) is

$$f(x, y\,;\theta, \psi) = f(x\,;\theta)g(y\,;\psi)\,, \quad (\theta, \psi) \in \Theta \times \Psi\,.$$

Consider statements of the form:

(*) If the statistic $T = T(X)$ has property P in the model for X,
and the statistic $U = U(Y)$ has property P in the model for Y,
then (T, U) has property P in the model for (X, Y).

Theorem. Statement (*) is true for the following properties: (i) sufficiency, (ii) minimal sufficiency, (iii) completeness, provided that each of the models for X and for Y has common support, and (iv) ancillarity.

Theorem. If each of the models for X and for Y is an exponential (resp., regular exponential) family, then so is the model for (X, Y).

Theorem. $\mathcal{I}_{(X,Y)}(\theta, \psi) = \begin{bmatrix} \mathcal{I}_X(\theta) & 0 \\ 0 & \mathcal{I}_Y(\psi) \end{bmatrix}$.

Theorem. If $\widehat{\theta} = \widehat{\theta}(X)$ is an MLE of θ in the model for X, and $\widehat{\psi} = \widehat{\psi}(Y)$ is an MLE of ψ in the model for Y, then $(\widehat{\theta}, \widehat{\psi})$ is an MLE of (θ, ψ) in the model for (X, Y).

Theorem. Suppose each of the models for X and for Y has a complete sufficient statistic. If $W = W(X)$ is the UMVUE of $\tau_1(\theta)$ in the model for X, and $V = V(Y)$ is the UMVUE of $\tau_2(\psi)$ in the model for Y, then in the model for (X, Y):

(a) $W + V$ is the UMVUE of $\tau_1(\theta) + \tau_2(\psi)$.
(b) $W - V$ is the UMVUE of $\tau_1(\theta) - \tau_2(\psi)$.
(c) WV is the UMVUE of $\tau_1(\theta)\tau_2(\psi)$.

Example. Suppose that $X_1, \ldots, X_n$ are i.i.d. Bernoulli(θ_1) and $Y_1, \ldots, Y_m$ are i.i.d. Bernoulli(θ_2), and that the two samples are independent. In the model for the first sample we know $\sum X_i$ is a complete sufficient statistic and $\bar{X}$ is the UMVUE of θ_1. In the model for the second sample we know $\sum Y_j$ is a complete sufficient statistic and $\bar{Y}$ is the UMVUE of θ_2. In the model for the combined data, the preceding theorem implies that $\bar{X} - \bar{Y}$ is the UMVUE of $\theta_1 - \theta_2$. Since $\bar{X}$ is the MLE of θ_1 in the model for the first sample and $\bar{Y}$ is the MLE of θ_2 in the model for the second sample, then $\bar{X} - \bar{Y}$ is an MLE of $\theta_1 - \theta_2$ in the model for the combined data. △

Example. Suppose that $X_1, \ldots, X_n$ are i.i.d. Normal(μ_1, σ_1^2) and $Y_1, \ldots, Y_m$ are i.i.d. Normal(μ_2, σ_2^2), and that the two samples are independent. We know that the model for each of the two samples has a complete sufficient statistic, namely $(\bar{X}, S_X^2)$ and $(\bar{Y}, S_Y^2)$ respectively.

(a) In the model for the combined data, the theorems above imply that $\bar{X} - \bar{Y}$ is the UMVUE and MLE of $\mu_1 - \mu_2$.

(b) Using the fact that $S_X^2 \sim \frac{\sigma_1^2}{n-1}\chi^2(n-1) = \text{Gamma}\big(\frac{n-1}{2}, \frac{2\sigma_1^2}{n-1}\big)$ and using formula (2.3.26), we find that $\mathrm{E}\big(\frac{b_n}{S_X}\big) = \frac{1}{\sigma_1}$ where $b_n = \Gamma\big(\frac{n-1}{2}\big)/\big[\sqrt{\frac{n-1}{2}}\,\Gamma\big(\frac{n-2}{2}\big)\big]$. Hence $\frac{b_n\bar{X}}{S_X}$ is the UMVUE of $\frac{\mu_1}{\sigma_1}$ in the model for the X_i's. Similarly, $\frac{b_m\bar{Y}}{S_Y}$ is the UMVUE of $\frac{\mu_2}{\sigma_2}$ in the model for the Y_j's. In the combined model, the UMVUE of $\frac{\mu_1}{\sigma_1} - \frac{\mu_2}{\sigma_2}$ is $\frac{b_n\bar{X}}{S_X} - \frac{b_m\bar{Y}}{S_Y}$. The MLE of $\frac{\mu_1}{\sigma_1} - \frac{\mu_2}{\sigma_2}$ is $\frac{\bar{X}}{\widehat{\sigma}_X} - \frac{\bar{Y}}{\widehat{\sigma}_Y}$ where $\widehat{\sigma}_X = \sqrt{\frac{1}{n}\sum(X_i - \bar{X})^2}$ and $\widehat{\sigma}_Y = \sqrt{\frac{1}{m}\sum(Y_j - \bar{Y})^2}$.

(c) In the model for the X_i's, $a_n S_X$ is the UMVUE of σ_1, where $a_n = \sqrt{\frac{n-1}{2}}\,\Gamma\big(\frac{n-1}{2}\big)/\Gamma\big(\frac{n}{2}\big)$. In the model for the Y_j's, $\frac{b_m}{S_Y}$ is the UMVUE of $\frac{1}{\sigma_2}$. Therefore, in the combined model, the UMVUE of $\frac{\sigma_1}{\sigma_2}$ is $a_n b_m \frac{S_X}{S_Y}$. The MLE of $\frac{\sigma_1}{\sigma_2}$ is $\frac{\widehat{\sigma}_X}{\widehat{\sigma}_Y} = \sqrt{\frac{(n-1)m}{n(m-1)}}\frac{S_X}{S_Y}$. △

Two independent normal samples with related parameters

A. Common variance. Suppose that $X_1, \ldots, X_n$ are i.i.d. Normal(μ_1, σ^2) and $Y_1, \ldots, Y_m$ are i.i.d. Normal(μ_2, σ^2), and that the two samples are independent. Note that the two populations have the same variance σ^2.

(ii) $f(\boldsymbol{x}, \boldsymbol{y}; \mu_1, \mu_2, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(x_i-\mu_1)^2} \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(y_j-\mu_2)^2}$

$$= \Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^{n+m} \exp\Big\{-\frac{1}{2\sigma^2}\Big[\sum_{i=1}^{n}(x_i-\mu_1)^2 + \sum_{j=1}^{m}(y_j-\mu_2)^2\Big]\Big\}.$$

$$= \Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^{n+m} \exp\Big\{-\frac{1}{2\sigma^2}\big[n\mu_1^2 + m\mu_2^2\big]\Big\} \times$$

$$\exp\Big\{-\frac{1}{2\sigma^2}\big[\sum x_i^2 + \sum y_j^2\big] + \frac{\mu_1}{\sigma^2}\sum x_i + \frac{\mu_2}{\sigma^2}\sum y_j\Big\}.$$

Letting $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma^2)$, the density has the form

$$f(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta}) = a(\boldsymbol{\theta})\exp\{b_1(\boldsymbol{\theta})R_1(\boldsymbol{x},\boldsymbol{y}) + b_2(\boldsymbol{\theta})R_2(\boldsymbol{x},\boldsymbol{y}) + b_3(\boldsymbol{\theta})R_3(\boldsymbol{x},\boldsymbol{y})\}$$

where

$$\begin{aligned} b_1(\boldsymbol{\theta}) &= \tfrac{-1}{2\sigma^2}, & R_1(\boldsymbol{x},\boldsymbol{y}) &= \textstyle\sum x_i^2 + \sum y_j^2 \\ b_2(\boldsymbol{\theta}) &= \tfrac{\mu_1}{\sigma^2}, & R_2(\boldsymbol{x},\boldsymbol{y}) &= \textstyle\sum x_i \\ b_3(\boldsymbol{\theta}) &= \tfrac{\mu_2}{\sigma^2}, & R_3(\boldsymbol{x},\boldsymbol{y}) &= \textstyle\sum y_j . \end{aligned}$$

In the notation used on p. 20 above, $k = p = 3$. The other two conditions for a regular exponential family are also satisfied. From Theorem 6.6.2 we conclude that $\boldsymbol{T} = (\sum X_i^2 + \sum Y_j^2, \sum X_i, \sum Y_j)$ is a complete sufficient statistic. Since $(\bar{X},\bar{Y},S_{\mathrm{p}}^2)$ where $S_{\mathrm{p}}^2 = \frac{1}{n+m-2}\left[\sum(X_i-\bar{X})^2 + \sum(Y_j-\bar{Y})^2\right]$ is a one-to-one function of $\boldsymbol{T}$, it also complete and sufficient..

Thus we find that

$\bar{X}$ is the UMVUE of μ_1.

$\bar{Y}$ is the UMVUE of μ_2.

$\bar{X}-\bar{Y}$ is the UMVUE of $\mu_1-\mu_2$.

S_{p}^2 is the UMVUE of σ^2.

$c\frac{\bar{X}-\bar{Y}}{S_{\mathrm{p}}}$ is the UMVUE of $\frac{\mu_1-\mu_2}{\sigma}$ where

$$c = \Gamma\left(\tfrac{n+m-2}{2}\right)/\left[\sqrt{\tfrac{n+m-2}{2}}\,\Gamma\left(\tfrac{n+m-3}{2}\right)\right].$$

(ii) To find MLEs, we calculate

$$\begin{aligned} \frac{\partial}{\partial\mu_1}\log f(\boldsymbol{x},\boldsymbol{y};\mu_1,\mu_2,\sigma^2) &= \tfrac{n}{\sigma^2}(\overline{x}-\mu_1) \\ \frac{\partial}{\partial\mu_2}\log f(\boldsymbol{x},\boldsymbol{y};\mu_1,\mu_2,\sigma^2) &= \tfrac{m}{\sigma^2}(\overline{y}-\mu_2) \\ \frac{\partial}{\partial\sigma^2}\log f(\boldsymbol{x},\boldsymbol{y};\mu_1,\mu_2,\sigma^2) &= -\frac{n+m}{2\sigma^2} + \frac{1}{2\sigma^4}\left[\textstyle\sum(x_i-\mu_1)^2 + \sum(y_j-\mu_2)^2\right] \end{aligned}$$

and obtain

$$\widehat{\mu}_1 = \bar{X},\quad \widehat{\mu}_2 = \bar{Y},\quad \widehat{\sigma}^2 = \left(\tfrac{n+m-2}{n+m}\right)S_{\mathrm{p}}^2 .\ \triangle$$

B. <u>Common mean</u>. Suppose that $X_1,\ldots,X_n$ are i.i.d. Normal(μ,σ_1^2) and $Y_1,\ldots,Y_m$ are i.i.d. Normal(μ,σ_2^2), and that the two samples are independent. Note that the two populations have the same mean μ.

(i) $f(\boldsymbol{x},\boldsymbol{y};\mu,\sigma_1^2,\sigma_2^2) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma_1^2}}e^{-\frac{1}{2\sigma_1^2}(x_i-\mu)^2}\prod_{j=1}^{m}\frac{1}{\sqrt{2\pi\sigma_2^2}}e^{-\frac{1}{2\sigma_2^2}(y_j-\mu)^2}$

$$= \left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right)^n \left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\right)^m \exp\Big\{-\frac{1}{2\sigma_1^2}\sum_{i=1}^n (x_i-\mu)^2 - \frac{1}{2\sigma_2^2}\sum_{j=1}^m (y_j-\mu)^2\Big\}.$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right)^n \left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\right)^m \exp\Big\{-\frac{n\mu^2}{2\sigma_1^2} - \frac{m\mu^2}{2\sigma_2^2}\Big\} \times$$

$$\exp\Big\{-\frac{1}{2\sigma_1^2}\sum x_i^2 + \frac{\mu}{\sigma_1^2}\sum x_i - \frac{1}{2\sigma_2^2}\sum y_j^2 + \frac{\mu}{\sigma_2^2}\sum y_j\Big\}.$$

Letting $\boldsymbol{\theta} = (\mu, \sigma_1^2, \sigma_2^2)$, the density has the form

$$f(\boldsymbol{x}, \boldsymbol{y}\,; \boldsymbol{\theta}) = a(\boldsymbol{\theta})\exp\{b_1(\boldsymbol{\theta})R_1(\boldsymbol{x},\boldsymbol{y}) + b_2(\boldsymbol{\theta})R_2(\boldsymbol{x},\boldsymbol{y}) + b_3(\boldsymbol{\theta})R_3(\boldsymbol{x},\boldsymbol{y}) + b_4(\boldsymbol{\theta})R_4(\boldsymbol{x},\boldsymbol{y})\}$$

where

$$b_1(\boldsymbol{\theta}) = \frac{-1}{2\sigma_1^2}, \qquad R_1(\boldsymbol{x},\boldsymbol{y}) = \sum x_i^2$$

$$b_2(\boldsymbol{\theta}) = \frac{\mu}{\sigma_1^2}, \qquad R_2(\boldsymbol{x},\boldsymbol{y}) = \sum x_i$$

$$b_3(\boldsymbol{\theta}) = \frac{-1}{2\sigma_2^2}, \qquad R_3(\boldsymbol{x},\boldsymbol{y}) = \sum y_j^2$$

$$b_4(\boldsymbol{\theta}) = \frac{\mu}{\sigma_2^2}, \qquad R_4(\boldsymbol{x},\boldsymbol{y}) = \sum y_j.$$

In the notation used on p. 20 above, $p = 3 \neq 4 = k$. This is an exponential family but it is not regular. It can be shown, using Theorem 6.3.1, that $R = (\sum X_i^2, \sum X_i, \sum Y_j^2, \sum Y_j)$ is a minimal sufficient statistic. It is not complete because $\mathrm{E}(\frac{1}{n}\sum X_i - \frac{1}{m}\sum Y_j) = \mu - \mu = 0$.

(ii) To find MLEs, we calculate

$$\frac{\partial}{\partial\mu}\log f(\boldsymbol{x},\boldsymbol{y}\,;\mu,\sigma_1^2,\sigma_2^2) = \frac{n}{\sigma_1^2}(\overline{x}-\mu) + \frac{m}{\sigma_2^2}(\overline{y}-\mu)$$

$$\frac{\partial}{\partial(\sigma_1^2)}\log f(\boldsymbol{x},\boldsymbol{y}\,;\mu,\sigma_1^2,\sigma_2^2) = -\frac{n}{2\sigma_1^2} + \frac{1}{2\sigma_1^4}\sum(x_i-\mu)^2$$

$$\frac{\partial}{\partial(\sigma_2^2)}\log f(\boldsymbol{x},\boldsymbol{y}\,;\mu,\sigma_1^2,\sigma_2^2) = -\frac{m}{2\sigma_2^2} + \frac{1}{2\sigma_2^4}\sum(y_j-\mu)^2.$$

The likelihood equations can be manipulated to obtain

$$(1) \qquad \widehat{\mu} = \frac{\frac{n}{\widehat{\sigma}_1^2}\overline{x} + \frac{m}{\widehat{\sigma}_2^2}\overline{y}}{\frac{n}{\widehat{\sigma}_1^2} + \frac{m}{\widehat{\sigma}_2^2}}$$

$$(2) \qquad \widehat{\sigma}_1^2 = \frac{1}{n}\sum(x_i - \widehat{\mu})^2$$

$$(3) \qquad \widehat{\sigma}_2^2 = \frac{1}{m}\sum(y_j - \widehat{\mu})^2.$$

These can be solved iteratively. Start with, say, an initial estimate $\widehat{\mu} = (\sum x_i + \sum y_j)/(n+m)$, the average of both samples combined. Plug this into (2) and (3) to obtain initial estimates $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$. Now plug these into (1) to obtain an improved estimate $\widehat{\mu}$. Next plug this into (2) and (3) to obtain improved estimates $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$. Continue this procedure until the estimates converge.

To estimate the SEs of these estimators, we can use the square roots of their asymptotic variances. For this we need the information matrix.

$$
\begin{aligned}
\frac{\partial^2}{\partial\mu^2}\log f(\boldsymbol{x},\boldsymbol{y}\,;\mu,\sigma_1^2,\sigma_2^2) &= -\frac{n}{\sigma_1^2} - \frac{m}{\sigma_2^2} \\
\frac{\partial^2}{\partial(\sigma_1^2)^2}\log f(\boldsymbol{x},\boldsymbol{y}\,;\mu,\sigma_1^2,\sigma_2^2) &= \frac{n}{2\sigma_1^4} - \frac{1}{\sigma_1^6}\sum(x_i-\mu)^2 \\
\frac{\partial^2}{\partial(\sigma_2^2)^2}\log f(\boldsymbol{x},\boldsymbol{y}\,;\mu,\sigma_1^2,\sigma_2^2) &= \frac{m}{2\sigma_2^4} - \frac{1}{\sigma_2^6}\sum(y_j-\mu)^2 . \\
\frac{\partial^2}{\partial\mu\partial\sigma_1^2)}\log f(\boldsymbol{x},\boldsymbol{y}\,;\mu,\sigma_1^2,\sigma_2^2) &= -\frac{n}{\sigma_1^4}(\overline{x}-\mu) \\
\frac{\partial^2}{\partial\mu\partial\sigma_2^2)}\log f(\boldsymbol{x},\boldsymbol{y}\,;\mu,\sigma_1^2,\sigma_2^2) &= -\frac{m}{\sigma_2^4}(\overline{y}-\mu) \\
\frac{\partial^2}{\partial(\sigma_1^2)\partial(\sigma_2^2)}\log f(\boldsymbol{x},\boldsymbol{y}\,;\mu,\sigma_1^2,\sigma_2^2) &= 0 .
\end{aligned}
$$

Thus we obtain

$$
\mathcal{I}_{X,Y}(\mu,\sigma_1^2,\sigma_2^2) = \begin{bmatrix} \frac{n}{\sigma_1^2} + \frac{m}{\sigma_2^2} & 0 & 0 \\ 0 & \frac{n}{2\sigma_1^4} & 0 \\ 0 & 0 & \frac{m}{2\sigma_2^4} \end{bmatrix} .
$$

The second theorem on p. 68, which is for i.i.d. samples, generalizes to the two-sample situation. It is easy to invert this diagonal matrix:

$$
\mathcal{I}_{X,Y}^{-1}(\mu,\sigma_1^2,\sigma_2^2) = \begin{bmatrix} \frac{1}{\frac{n}{\sigma_1^2} + \frac{m}{\sigma_2^2}} & 0 & 0 \\ 0 & \frac{2\sigma_1^4}{n} & 0 \\ 0 & 0 & \frac{2\sigma_2^4}{m} \end{bmatrix} .
$$

For large n and m, the solution $\widehat{\mu}$ to the likelihood equations has approximately a normal distribution with mean μ and variance $1/(\frac{n}{\sigma_1^2} + \frac{m}{\sigma_2^2})$. Let us compare this with what we could achieve if we knew the values of σ_1^2 and σ_2^2. It can be shown that we would then have a 1-parameter regular exponential family and that an MLE (and UMVUE) for μ would be

$$\widehat{\mu}_* = \frac{\frac{n}{\sigma_1^2}\overline{x} + \frac{m}{\sigma_2^2}\overline{y}}{\frac{n}{\sigma_1^2} + \frac{m}{\sigma_2^2}} .$$

The variance of $\widehat{\mu}_*$ is $1/(\frac{n}{\sigma_1^2} + \frac{m}{\sigma_2^2})$. So the asymptotic variance of the likelihood-equation solution $\widehat{\mu}$ when σ_1^2 and σ_2^2 are unknown is approximately the same as the UMVUE of μ when σ_1^2 and σ_2^2 are known. This provides support for the claim that that likelihood-equation estimators are often efficient, at least for large sample sizes.

```
% Script air.m

% Data from Cox & Snell (1981) Example T.
% Intervals in service-hours between failures
%   of the air-conditioning equipment in a Boeing 720 jet aircraft.
% Aircraft # 2.
% From Proschan (1963) Technometrics, 375-383.

y1 = [ 90   10   60 186   61   49   14   24   56   20 ] ;
y2 = [ 79   84   44   59   29 118   25 156 310   76 ] ;
y3 = [ 26   44   23   62 130 208   70 101 208 ] ;
y = [y1 y2 y3] ;

n = length(y)
hist(y)
mle = gamfit(y) ;
ahat = mle(1)
d = ahat*trigamma(ahat)-1 ;
varahat = (1/n)*ahat/d ;
seahat = sqrt(varahat)
bhat = mle(2)
varbhat = (1/n)*bhat^2*trigamma(ahat)/d ;
sebhat = sqrt(varbhat)
```

```
function p = trigamma(x)

% Approximation to the trigamma function

c1 = 0.0757575757575757576 ;
c2 = -0.0333333333333333 ;
c3 = 0.0238095238095238 ;
c4 = 0.166666666666667 ;

x=x+6;
p=1/(x.*x);
p=(((((c1*p+c2).*p+c3).*p+c2).*p+c4).*p+1)./x+0.5*p;
for i = 0:5 ;
  x=x-1;
  p=1/(x.*x)+p;
end
```

Birkes, Spring 2001
Textbook by Mukhopadhyay

Notes on Test of Hypotheses

(see Ch. 8 in Mukhopadhyay)

Formulation of a probability model

Let $\boldsymbol{x}$ be a vector of observed data. Sometimes it is appropriate to formulate a probability model for the data. To be mathematically precise, a probability model is regarded as a family of possible probability distributions that could have generated the data. That is, we suppose $\boldsymbol{x}$ is the observed result of a random experiment that produces a random data vector $\boldsymbol{X}$ according to a probability distribution with pmf or pdf $f(\boldsymbol{x}\,;\boldsymbol{\theta})$ for some parameter (or parameter vector) $\boldsymbol{\theta}$ in a specified parameter space Θ.

Notation. Upper case $\boldsymbol{X}$ represents the outcome before it is observed and lower case $\boldsymbol{x}$ represents the outcome after it is observed. Lower case $\boldsymbol{x}$ is also used as a mathematical dummy variable, as in $f(\boldsymbol{x}\,;\boldsymbol{\theta})$. To distinguish the two uses of lower case $\boldsymbol{x}$, we sometimes write $\boldsymbol{x}_{\text{obs}}$ for a vector of observed data.

We assume that one of the parameter vectors in Θ is the true parameter vector indexing the true probability distribution that governed the generation of $\boldsymbol{x}$. But we do not know which is the true $\boldsymbol{\theta}$; we know only that it is in Θ. Using the observed data $\boldsymbol{x}$, our goal is to make an inference about the true $\boldsymbol{\theta}$.

Example 1. A polling organization selects a simple random sample of 100 voters in Corvallis and asks them whether they are in favor of a school tax bond. The data can be expressed as $\boldsymbol{x} = (x_1, x_2, \ldots, x_{100})$ where $x_i = 1$ if the i-th voter is in favor and $x_i = 0$ if not. A sensible model for these data is that they are the observed outcome of $\boldsymbol{X} = (X_1, X_2, \ldots, X_{100})$ where the X_i's are i.i.d. Bernoulli(θ) for some $0 < \theta < 1$. Then $\Theta = (0,1)$ and

$$f(\boldsymbol{x}\,;\theta) = \prod_{i=1}^{100} f(x_i\,;\theta) = \prod_{i=1}^{100} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\Sigma x_i}(1-\theta)^{100-\Sigma x_i}. \qquad \|$$

Example 2. A rat is weighed 5 times using a scale whose standard deviation is known to be 3 grams. The data are $\boldsymbol{x} = (x_1, x_2, x_3, x_4, x_5)$ where x_i is the weight obtained from the i-th weighing. A possible model for these data is that they are the observed outcome of $\boldsymbol{X} = (X_1, X_2, X_3, X_4, X_5)$ where the X_i's are i.i.d. Normal($\mu, 9$) for some $\mu > 0$. Then $\boldsymbol{\theta} = \mu$, $\Theta = (0, \infty)$ and

$$f(\boldsymbol{x}\,;\mu) = \prod_{i=1}^{5} f(x_i\,;\mu) = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi}\,3}\exp\left[-\tfrac{1}{18}(x_i-\mu)^2\right]$$

$$= (3\sqrt{2\pi})^{-5}\exp\left[-\tfrac{1}{18}\sum_{i=1}^{5}(x_i - \mu)^2\right] . \qquad \|$$

Example 3. A polling organization selects a simple random sample of 100 voters in Corvallis and another simple random sample of 100 voters in Bend and asks them whether they are in favor of removing dams to aid salmon. The data can be expressed as $\boldsymbol{x} = (x_{11}, x_{12}, \ldots, x_{1,100}, x_{21}, x_{22}, \ldots, x_{2,100})$ where $x_{1i} = 1$ if the i-th voter in Corvallis is in favor and $x_{1i} = 0$ if not and where $x_{2i} = 1$ if the i-th voter in Bend is in favor and $x_{2i} = 0$ if not. A sensible model for these data is that they are the observed outcome of $\boldsymbol{X} = (X_{11}, X_{12}, \ldots, X_{1,100}, X_{21}, X_{22}, \ldots, X_{2,100})$ where the X_{1i}'s are i.i.d. Bernoulli(θ_1) for some $0 < \theta_1 < 1$, the X_{2i}'s are i.i.d. Bernoulli(θ_2) for some $0 < \theta_2 < 1$, and the X_{1i}'s are independent of the X_{2i}'s. Then $\boldsymbol{\theta} = (\theta_1, \theta_2)$, $\Theta = (0,1) \times (0,1)$, and

$$f(\boldsymbol{x}\,;\boldsymbol{\theta}) = \theta_1^{\sum x_{1i}}(1-\theta_1)^{100-\sum x_{1i}}\,\theta_2^{\sum x_{2i}}(1-\theta_2)^{100-\sum x_{2i}} . \qquad \|$$

Example 4. A polling organization selects a simple random sample of 100 voters in Corvallis and asks them whether they are in favor of removing dams to aid salmon and whether they are in favor of a school tax bond. The data can be expressed as $\boldsymbol{x} = (x_{11}, x_{12}, x_{21}, x_{22}, \ldots, x_{100,1}, x_{100,2})$ where $x_{i1} = 1$ if the i-th voter is in favor of removing dams and $x_{i1} = 0$ if not and where $x_{i2} = 1$ if the i-th voter is in favor of the bond and $x_{i2} = 0$ if not. A sensible model for these data is that they are the observed outcome of $\boldsymbol{X} = (X_{11}, X_{12}, X_{21}, X_{22}, \ldots, X_{100,1}, X_{100,2})$ where the pairs (X_{i1}, X_{i2}) are i.i.d. with pmf defined by $f(1,1) = \theta_{11}$, $f(1,0) = \theta_{10}$, $f(0,1) = \theta_{01}$, and $f(0,0) = \theta_{00}$ for parameters satisfying $0 < \theta_{rs} < 1$ and $\theta_{11} + \theta_{10} + \theta_{01} + \theta_{00} = 1$. Then $\boldsymbol{\theta} = (\theta_{11}, \theta_{10}, \theta_{01}, \theta_{00})$, $\Theta = \{\, \boldsymbol{\theta} \in (0,1) \times (0,1) \times (0,1) \times (0,1) : \theta_{11} + \theta_{10} + \theta_{01} + \theta_{00} = 1 \,\}$, and

$$\begin{aligned} f(\boldsymbol{x}\,;\boldsymbol{\theta}) &= \prod_{i=1}^{100} f(x_{i1}, x_{i2}\,;\boldsymbol{\theta}) = \prod_{i=1}^{100} \theta_{11}^{x_{i1}x_{i2}}\,\theta_{10}^{x_{i1}(1-x_{i2})}\,\theta_{01}^{(1-x_{i1})x_{i2}}\,\theta_{00}^{(1-x_{i1})(1-x_{i2})} \\ &= \theta_{11}^{y_{11}}\,\theta_{10}^{y_{10}}\,\theta_{01}^{y_{01}}\,\theta_{00}^{y_{00}} \end{aligned}$$

where $y_{rs} = \#\{\, i : (x_{i1}, x_{i2}) = (r, s) \,\}$. $\|$

Example 5. 100 rats are taken from a colony of rats and a simple random sample of 50 of them are selected to be fed a high-sodium diet. The remaining 50 rats are fed a low-sodium diet. After six months their blood pressures are measured. The data can be expressed as $\boldsymbol{x} = (x_{11}, x_{12}, \ldots, x_{1,50}, x_{21}, x_{22}, \ldots, x_{2,50})$ where x_{1i} is the blood pressure of the i-th rat on the high-sodium diet and x_{2i} is the blood pressure of the i-th rat on the low-sodium diet. A possible model for these data is that they are the observed outcome of $\boldsymbol{X} = (X_{11}, X_{12}, \ldots, X_{1,50}, X_{21}, X_{22}, \ldots, X_{2,50})$ where the X_{1i}'s are i.i.d. Normal(μ_1, σ_1^2), the

X_{2i}'s are i.i.d. Normal(μ_2, σ_2^2), and the X_{1i}'s are independent of the X_{2i}'s. Then $\boldsymbol{\theta} = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$, $\Theta = (0,\infty) \times (0,\infty) \times (0,\infty) \times (0,\infty)$, and

$$\begin{aligned} f(\boldsymbol{x}\,;\boldsymbol{\theta}) &= \prod_{i=1}^{50} f(x_{1i}\,;\mu_1,\sigma_1^2) \prod_{i=1}^{50} f(x_{2i}\,;\mu_2,\sigma_2^2) \\ &= \prod_{i=1}^{50} \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\big[-\frac{1}{2\sigma_1^2}(x_{1i}-\mu_1)^2\big] \prod_{i=1}^{50} \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\big[-\frac{1}{2\sigma_2^2}(x_{2i}-\mu_2)^2\big] \\ &= (2\pi\sigma_1\sigma_2)^{-50} \exp\big[-\frac{1}{2\sigma_1^2}\sum_{i=1}^{50}(x_{1i}-\mu_1)^2 - \frac{1}{2\sigma_2^2}\sum_{i=1}^{50}(x_{2i}-\mu_2)^2\big]\,. \qquad \| \end{aligned}$$

Formulation of ~~a test of~~ a hypothesis

In order to perform a statistical test of a hypothesis, the hypothesis should be expressed in terms of the parameters of the probability model. The general form of a hypothesis is $\mathrm{H}_0 : \boldsymbol{\theta} \in \Theta_0$ where Θ_0 is a subset of the parameter space Θ. Using the data vector $\boldsymbol{x}$, we want to decide whether or not the true parameter vector $\boldsymbol{\theta}$ is in Θ_0. That is, we want to test the null hypothesis $\mathrm{H}_0 : \boldsymbol{\theta} \in \Theta_0$ versus the alternative hypothesis $\mathrm{H}_1 : \boldsymbol{\theta} \notin \Theta_0$. The null hypothesis is given preferred status in the sense that it is accepted as a plausible statement unless the data show strong evidence against it. Sometimes the alternative hypothesis is written as $\mathrm{H}_1 : \boldsymbol{\theta} \in \Theta_1$ where $\Theta_1 = \Theta \setminus \Theta_0 = \{\boldsymbol{\theta} \in \Theta : \boldsymbol{\theta} \notin \Theta_0\}$.

Sometimes we want to focus on a certain subset of the null parameter vectors or of the alternative parameter vectors. To do this we test $\mathrm{H}_0 : \boldsymbol{\theta} \in \Theta_0$ versus $\mathrm{H}_1 : \boldsymbol{\theta} \in \Theta_1$ where Θ_0 and Θ_1 are disjoint (that is, $\Theta_0 \cap \Theta_1 = \emptyset$) but their union does not contain all the parameter vectors in Θ. Such a focus reduces the parameter space from Θ to $\Theta^\dagger = \Theta_0 \cup \Theta_1$.

Example 1 (cont'd). Suppose we want to test whether or not the percentage of voters in favor of the school tax bond is greater than 50%. Recall that the parameter space is $\Theta = (0,1)$. We might set $\Theta_0 = (0.5, 1)$, which corresponds to $\mathrm{H}_0 : \theta > 0.5$. Then $\mathrm{H}_1 : \theta \le 0.5$ and $\Theta_1 = (0, 0.5]$. However, as will be seen later, we want the null hypothesis to be as specific as possible, and hence we want H_0 to contain the equality sign rather than H_1. So we should let the hypotheses be $\mathrm{H}_0 : \theta \le 0.5$ versus $\mathrm{H}_1 : \theta > 0.5$. Then $\Theta_0 = (0, 0.5]$ and $\Theta_1 = (0.5, 1)$.

It is also sensible to formulate the hypotheses as $\mathrm{H}_0 : \theta = 0.5$ versus $\mathrm{H}_1 : \theta > 0.5$. Then we would be reducing the parameter space to $\Theta^\dagger = [0.5, 1)$.

In developing the theory of testing, it is convenient to also consider simplified hypotheses such as $\mathrm{H}_0 : \theta = 0.5$ versus $\mathrm{H}_1 : \theta = 0.6$. This reduces the parameter space to $\Theta^\dagger = \{0.5, 0.6\}$.

One approach to testing the value of a parameter

Suppose that we have been given an observed data vector $\boldsymbol{x}$ and that we have formulated a probability model $f(\boldsymbol{x};\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, for the data. Suppose we want to test a simple null hypothesis $\mathrm{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. A null hypothesis is called *simple* if Θ_0 contains only one member. A sensible approach to testing a simple null hypothesis is to first estimate the true parameter by some reasonable estimator $\widehat{\boldsymbol{\theta}}$ and then ~~accept~~ H_0 if $\widehat{\boldsymbol{\theta}}$ is "close" to $\boldsymbol{\theta}_0$ and ~~accept~~ $\mathrm{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ if $\widehat{\boldsymbol{\theta}}$ is "far" from $\boldsymbol{\theta}_0$.

f.t.r. rej. Ho

Example 2 (cont'd). We assume that X_1, X_2, X_3, X_4, X_5 are i.i.d. Normal$(\mu, 9)$ for some μ. Let us test $\mathrm{H}_0 : \mu = 200$. First let us choose an estimator for μ. A good choice is the sample mean $\overline{X}$. This is a very reasonable choice because it is the method-of-moments estimator (Example 7.2.2), the maximum likelihood estimator (Example 7.2.6), and the uniformly minimum variance unbiased estimator (Example 7.5.3).

We will accept $\mathrm{H}_0 : \mu = 200$ if $\overline{X}$ is "close" to 200. We could specify "closeness" in terms of $|\overline{X} - 200|$. Thus, we accept $\mathrm{H}_0 : \mu = 200$ if $|\overline{X} - 200| \leq b$ (for some $b > 0$, the choice of which is discussed below), and we accept $\mathrm{H}_1 : \mu \neq 200$ if $|\overline{X} - 200| > b$. Ho rej.

The value of b determines the probabilities of two types of errors. *Type I error* occurs if ~~H_1 is accepted~~ when H_0 is true. *Type II error* occurs if ~~H_0 is accepted~~ when H_1 is true.

rej. Ho. f.t.r. Ho

$$\begin{aligned} \mathrm{P}\{\text{Type I error}\} &= \mathrm{P}\{\overset{\text{rej. Ho.}}{\cancel{\text{accept } \mathrm{H}_1}} \mid \mathrm{H}_0 \text{ is true}\} \\ &= \mathrm{P}\{|\overline{X} - 200| > b \mid \mu = 200\}\ . \end{aligned}$$

Since X_1, X_2, X_3, X_4, X_5 are i.i.d. Normal$(\mu, 9)$, then $\overline{X} \sim$ Normal$(\mu, 1.8)$, by Theorem 4.4.2. If $\mu = 200$, then $\overline{X} \sim$ Normal$(200, 1.8)$, so $\overline{X} - 200 \sim$ Normal$(0, 1.8)$ and $(\overline{X} - 200)/\sqrt{1.8} \sim$ Normal$(0,1)$. Therefore

$$\begin{aligned} \mathrm{P}\{\text{Type I error}\} &= \mathrm{P}\{|\overline{X} - 200|/\sqrt{1.8} > b\,/\sqrt{1.8} \mid \mu = 200\} \\ &= \mathrm{P}\{|Z| > b\,/\sqrt{1.8}\} \text{ where } Z \sim \text{Normal}(0,1) \\ &= \mathrm{P}\{Z < -b\,/\sqrt{1.8} \text{ or } Z > b\,/\sqrt{1.8}\} \\ &= \Phi(-b\,/\sqrt{1.8}) + 1 - \Phi(b\,/\sqrt{1.8}) \\ &= 2\Phi(-b\,/\sqrt{1.8})\ . \end{aligned}$$

(because $\Phi(-\infty) = 0$)

If we choose b very large, then P{Type I error} will be close to $\underset{\wedge}{0}$. However, we are also concerned about Type II error.

$$\mathrm{P}\{\text{Type II error}\} = \mathrm{P}\{\overset{\text{f.t.r.}}{\cancel{\text{accept}}}\ \mathrm{H}_0 \mid \mathrm{H}_1 \text{ is true}\}$$

$$= \mathrm{P}\{\, |\overline{X} - 200| \le b \,|\, \mu \ne 200 \,\} \ .$$

This should be interpreted as a collection of probabilities, because there is a different probability for each different value of $\mu \ne 200$.

$$\mathrm{P}_\mu\{\, |\overline{X} - 200| \le b \,\} \;=\; \mathrm{P}_\mu\{\, 200 - b \le \overline{X} \le 200 + b \,\} \ .$$

As above, $\overline{X} \sim \mathrm{Normal}(\mu, 1.8)$ and so $(\overline{X} - \mu)/\sqrt{1.8} \sim \mathrm{Normal}(0, 1)$. Now

$$\begin{aligned}
&\mathrm{P}_\mu\{\, 200 - b \le \overline{X} \le 200 + b \,\} \\
&\quad = \mathrm{P}_\mu\{\, (200 - b - \mu)/\sqrt{1.8} \le (\overline{X} - \mu)/\sqrt{1.8} \le (200 + b - \mu)/\sqrt{1.8} \,\} \\
&\quad = \mathrm{P}_\mu\{\, (200 - b - \mu)/\sqrt{1.8} \le Z \le (200 + b - \mu)/\sqrt{1.8} \,\} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{where } Z \sim \mathrm{Normal}(0, 1) \\
&\quad = \Phi((200 + b - \mu)/\sqrt{1.8}\,) - \Phi((200 - b - \mu)/\sqrt{1.8}\,) \ .
\end{aligned}$$

If b is very large, then $\mathrm{P}\{\text{Type II error}\}$ will be close to 1 (because $\Phi(\infty) - \Phi(-\infty) = 1 - 0 = 1$). The probability of Type II error can be made small by choosing b close to 0, (because $\Phi((200 - \mu)/\sqrt{1.8}\,) - \Phi((200 - \mu)/\sqrt{1.8}\,) = 0$), but unfortunately, for b close to 0, $\mathrm{P}\{\text{Type I error}\}$ will be close to 1. We must choose b to balance our concerns about Type I and Type II error.

A common convention is to choose b so that $\mathrm{P}\{\text{Type I error}\} = .05$. In this example this requires $2\Phi(-b/\sqrt{1.8}\,) = .05$, hence $\Phi(-b/\sqrt{1.8}\,) = .025$, hence $-b/\sqrt{1.8} = \Phi^{-1}(.025) = -1.96$ or $b = 1.96\sqrt{1.8} = 2.63$.

Thus we have derived the following test procedure.

Calculate $\overline{X}$.

f.t.r

If $|\overline{X} - 200| \le 2.63$, accept $\mathrm{H}_0 : \mu = 200$.

rej. Ho

If $|\overline{X} - 200| > 2.63$, ~~accept H_1~~ : $\mu \ne 200$.

This test procedure has the following properties.

$$\mathrm{P}\{\text{Type I error}\} \;=\; .05 \ .$$

$$\mathrm{P}\{\text{Type II error}\} \;=\; \Phi((202.63 - \mu)/\sqrt{1.8}\,) - \Phi((197.37 - \mu)/\sqrt{1.8}\,) \ \text{ for } \mu \ne 200 \,.$$

Let $\beta(\mu) = \mathrm{P}_\mu\{\text{Type II error}\}$. The probabilities of Type II error for some selected values of μ are shown below.

μ	$\beta(\mu)$		μ	$\beta(\mu)$
200.1	.9494		199.9	.9494
200.2	.9475		199.8	.9475
201	.8844		199	.8844
202	.6804		198	.6804
205	.0387		195	.0387
210	.00000002		190	.00000002

Instead of looking at the probability of making a Type II error, we can take a more positive viewpoint and consider the probability of not making a Type II error, which is called the *power*.

$$\begin{aligned} \text{power} &= \mathrm{P}\{\text{no Type II error}\} \\ &= \mathrm{P}\{\overset{\text{rej } H0}{\cancel{\text{accept}}}\ \mathrm{H}_1 \mid \mathrm{H}_1 \text{ is true}\}\ . \end{aligned}$$

In our example,

$$\text{power} = \mathrm{P}\{\,|\overline{X} - 200| > b \mid \mu \neq 200\,\}\ .$$

This is actually a collection of probabilities, one for each $\mu \neq 200$.

Let $Q(\mu) = 1 - \beta(\mu) =$ power. (Caution: Some books let $\beta(\mu)$ denote the power.) The power function $Q(\mu)$ can be defined for all values of μ,

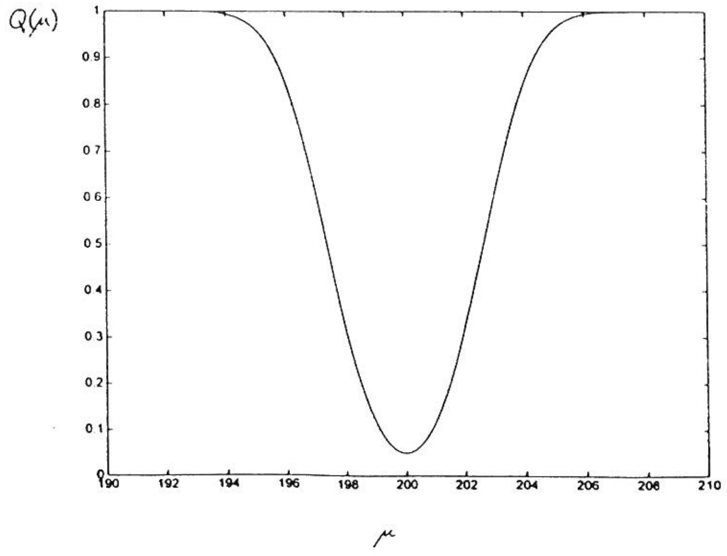
$$Q(\mu) = \mathrm{P}\{\overset{\text{rej. Ho,}}{\cancel{\text{accept}}}\ \mathrm{H}_1 \mid \mu \text{ is the true parameter}\}\ .$$

Then for $\mu = 200$,

$$Q(200) = \mathrm{P}\{\overset{\text{rej. Ho.}}{\cancel{\text{accept}}}\ \mathrm{H}_1 \mid \mu = 200\,\} = \mathrm{P}\{\text{Type I error}\}\ ,$$

and for $\mu \neq 200$,

$$Q(\mu) = \mathrm{P}\{\overset{\text{rej. H0}}{\cancel{\text{accept}}}\ \mathrm{H}_1 \mid \mu\ (\neq 200)\} = 1 - \mathrm{P}\{\text{Type II error}\}\ .$$



General concept of a test

A test of a hypothesis is a procedure for deciding, based on the data, whether to accept the hypothesis or reject it. Let $\mathcal{X}$ be the sample space of the probability model; that is, $\mathcal{X}$ is the set of all possible values that the random vector $\boldsymbol{X}$ might take. In Example 1 the sample space is $\mathcal{X} = \{0,1\} \times \cdots \times \{0,1\}$ $(100$ times$) = \{0,1\}^{100}$. In Example 2 the sample space is $\mathcal{X} = (-\infty,\infty) \times \cdots \times (-\infty,\infty)$ $(5$ times$) = (-\infty,\infty)^5 = \mathbb{R}^5$. (Note that the weight of a rat is always positive and so it seems that a more suitable sample space might be $\mathcal{X} = (0,\infty)^5$, but then instead of the Normal distribution we would need to use a distribution on the positive real numbers such as a truncated Normal distribution or a Gamma distribution. A truncated Normal distribution is difficult to work with, and it is often the case that the amount of truncation is very small and not worth bothering with.)

A testing procedure divides the sample space into two parts, $\mathcal{R}$ and its complement $\mathcal{R}^c$, where $\mathcal{R} = \{\boldsymbol{x} \in \mathcal{X} : \mathrm{H}_1$ is accepted if $\boldsymbol{x}$ is observed$\}$. Since H_1 is accepted if and only if H_0 is rejected, we call $\mathcal{R}$ the rejection region of the test. Any subset $\mathcal{R}$ in $\mathcal{X}$ can be regarded as the rejection region of a test. (Technically, $\mathcal{R}$ should be a "measurable" subset of $\mathcal{X}$, but we can safely assume that all sets we will encounter will be measurable.) Of course some of these tests are ridiculous ones that no one would ever seriously consider using.

The effectiveness of a test in coming to a correct decision can be evaluated by its *power function*

$$Q(\boldsymbol{\theta}) = \mathrm{P}\{\overset{\text{rej } H_0.}{\cancel{\text{accept}}}\ \mathrm{H}_1 \mid \boldsymbol{\theta} \text{ is the true parameter}\} = \mathrm{P}_{\boldsymbol{\theta}}\{\boldsymbol{X} \in \mathcal{R}\}\ ,$$

defined for all $\boldsymbol{\theta} \in \Theta$. For $\boldsymbol{\theta} \in \Theta_0$, we have $Q(\boldsymbol{\theta}) = \mathrm{P}\{\text{Type I error}\}$. And for $\boldsymbol{\theta} \in \Theta_1$, we have $Q(\boldsymbol{\theta}) = 1 - \mathrm{P}\{\text{Type II error}\} =$ power. Therefore, for $\boldsymbol{\theta} \in \Theta_0$, we want $Q(\boldsymbol{\theta})$ to be small, and for $\boldsymbol{\theta} \in \Theta_1$, we want $Q(\boldsymbol{\theta})$ to be large.

The *size* of a test is the maximum probability of Type I error, that is, $\sup\{Q(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_0\}$. If the size of a test is $\leq \alpha$, we say it has *level* α.

The usual approach to testing is to first be concerned with Type I error by choosing an acceptably low value of α and requiring that a test have level α. That is, if the true parameter vector is in Θ_0, the probability of a wrong decision can be no more than α. Next, subject to having level α, we look for a test with high power.

In some special testing situations, we can find a level α test that has the highest power $Q(\boldsymbol{\theta})$ among all level α tests and for all $\boldsymbol{\theta} \in \Theta_1$. Such a test is called a *best* level α test or a *uniformly most powerful* (UMP) level α test. The word "uniformly" refers to the property of having the highest power “for all $\theta \in \Theta_1$”. If the alternative hypothesis is simple, i.e., $\mathrm{H}_1 : \theta = \theta_1$, then a test having the highest power is called simply a *most powerful* (MP) test.

Most powerful tests

In order to develop some theory that can help us in finding tests of high power, we first consider the simple, artificial situation in which the null and alternative hypotheses are both simple, that is, $\mathrm{H}_0 : \theta = \theta_0$ versus $\mathrm{H}_1 : \theta = \theta_1$. Thus we are momentarily supposing that there are only two possible distributions that the data could have been generated from. Let $f(x\,; \theta_0)$ and $f(x\,; \theta_1)$ be the pmf's or pdf's of the two distributions.

A pmf or pdf $f(x\,; \theta)$ is a function of x for each fixed value of θ. If we view it as a function of θ for a fixed value of x, then it is a *likelihood function*, denoted $L(\theta\,; x)$. (This is slightly different from Mukhopadhyay's notation.) Many statisticians like to regard $L(\theta\,; x)$ as the likelihood of θ being the true parameter if the data vector x is observed. From this viewpoint, the likelihood ratio $f(x\,; \theta_1)/f(x\,; \theta_0) = L(\theta_1\,; x)/L(\theta_0\,; x)$ measures how much more likely θ_1 is to be the true parameter than θ_0 is. It turns out that the MP level α test of $\mathrm{H}_0 : \theta = \theta_0$ versus $\mathrm{H}_1 : \theta = \theta_1$ is one that accepts H_1 when the likelihood ratio is larger than some critical value k. That is, the MP test accepts H_1 when θ_1 is more than k times more likely to be the true parameter than θ_0 is. The value of k is chosen to obtain the desired level. Traditionally, scientists take the position that the null hypothesis will be rejected only if there is strong evidence that it is wrong. For a larger value of k, stronger evidence against H_0 is required to reject it, which implies a smaller probability of Type I error. Thus a larger value of k gives a smaller level.

Theorem. (Neyman-Pearson Lemma — version 1) Suppose $f(x\,; \theta_0) > 0$ for all x. Let $k > 0$ and $\mathcal{R}^* = \{x : f(x\,; \theta_1)/f(x\,; \theta_0) > k\}$. Then $\mathcal{R}^*$ is the rejection region of an MP level α test of $\mathrm{H}_0 : \theta = \theta_0$ versus $\mathrm{H}_1 : \theta = \theta_1$ where $\alpha = \mathrm{P}_{\theta_0}\{X \in \mathcal{R}^*\}$.

The requirement that $f(x\,; \theta_0) > 0$ for all x can be dropped.

Theorem. (Neyman-Pearson Lemma — version 2) Let $k > 0$ and suppose $\mathcal{R}^*$ is a subset of the sample space such that

$$f(x\,; \theta_1) > kf(x\,; \theta_0) \Rightarrow x \in \mathcal{R}^*$$

$$f(x\,; \theta_1) < kf(x\,; \theta_0) \Rightarrow x \notin \mathcal{R}^* .$$

(no requirement on x when $=$)

Then $\mathcal{R}^*$ is the rejection region of an MP level α test of $\mathrm{H}_0 : \theta = \theta_0$ versus $\mathrm{H}_1 : \theta = \theta_1$ where $\alpha = \mathrm{P}_{\theta_0}\{X \in \mathcal{R}^*\}$.

To obtain a desired level α, one tries to find the appropriate value of k. When the distribution of X is discrete, however, it may not be possible to find such a k.

Example 1 (cont'd). Suppose $X_1, X_2, \ldots, X_{100}$ are i.i.d. Bernoulli(θ) for some $0 < \theta < 1$. Their joint pmf is $f(\boldsymbol{x}\,;\theta) = \theta^{\sum x_i}(1-\theta)^{100-\sum x_i}$. Let us test $\mathrm{H}_0 : \theta = \theta_0$ versus $\mathrm{H}_1 : \theta = \theta_1$. The likelihood ratio is

$$f(\boldsymbol{x}\,;\theta_1)/f(\boldsymbol{x}\,;\theta_0) \;=\; \theta_1^{\sum x_i}(1-\theta_1)^{100-\sum x_i}\Big/\theta_0^{\sum x_i}(1-\theta_0)^{100-\sum x_i}$$

According to the Neyman-Pearson lemma, for every $k > 0$, the region $\mathcal{R}^* = \{\boldsymbol{x} : \theta_1^{\sum x_i}(1-\theta_1)^{100-\sum x_i}\Big/\theta_0^{\sum x_i}(1-\theta_0)^{100-\sum x_i} > k\}$ is the rejection region of an MP level α test where $\alpha = \mathrm{P}_{\theta_0}\{X \in \mathcal{R}^*\}$.

It would be nice if we could re-express $\mathcal{R}^*$ in a simpler form. What does the inequality defining $\mathcal{R}^*$ say about $\boldsymbol{x}$? Note that $\boldsymbol{x}$ is involved in the inequality only through the value of $\sum x_i$. By collecting terms involving $\sum x_i$, we can write the inequality as

$$\left[\frac{\theta_1/(1-\theta_1)}{\theta_0/(1-\theta_0)}\right]^{\sum x_i}\left(\frac{1-\theta_1}{1-\theta_0}\right)^{100} > k$$

or

$$(\textstyle\sum x_i)\log\left[\dfrac{\theta_1/(1-\theta_1)}{\theta_0/(1-\theta_0)}\right] + 100\log\left(\dfrac{1-\theta_1}{1-\theta_0}\right) > \log k\ .$$

or

$$\textstyle\sum x_i > \left[\log k - 100\log\left(\dfrac{1-\theta_1}{1-\theta_0}\right)\right]\Big/\log\left[\dfrac{\theta_1/(1-\theta_1)}{\theta_0/(1-\theta_0)}\right]$$

provided that $\log\left[\dfrac{\theta_1/(1-\theta_1)}{\theta_0/(1-\theta_0)}\right] > 0$, which is true if and only if $\dfrac{\theta_1/(1-\theta_1)}{\theta_0/(1-\theta_0)} > 1$, that is, if and only if $\theta_1 > \theta_0$.

So consider the case $\theta_1 > \theta_0$. Then the MP rejection region can be expressed as $\mathcal{R}^* = \{\boldsymbol{x} : \sum x_i > c\}$ where $c = \left[\log k - 100\log\left(\frac{1-\theta_1}{1-\theta_0}\right)\right]\Big/\log\left[\frac{\theta_1/(1-\theta_1)}{\theta_0/(1-\theta_0)}\right]$. Considering all values of $k > 0$ is equivalent to considering all values of c. Restating what was said above: for every c, the region $\mathcal{R}^* = \{\boldsymbol{x} : \sum x_i > c\}$ is the rejection region of an MP level α test where $\alpha = \mathrm{P}_{\theta_0}\{\sum X_i > c\}$.

Note that the description of the MP rejection region $\mathcal{R}^*$ does not involve the specific value of θ_1. It only involves the inequality $\theta_1 > \theta_0$ (so that $\log\left[\frac{\theta_1/(1-\theta_1)}{\theta_0/(1-\theta_0)}\right] > 0$). Therefore $\mathcal{R}^*$ is the MP rejection region for testing $\mathrm{H}_0 : \theta = \theta_0$ versus $\mathrm{H}_1 : \theta = \theta_1$ for all $\theta_1 > \theta_0$. In other words, $\mathcal{R}^*$ is the rejection region of a UMP level α test of $\mathrm{H}_0 : \theta = \theta_0$ versus $\mathrm{H}_1 : \theta > \theta_0$, where $\alpha = \mathrm{P}_{\theta_0}\{\sum X_i > c\}$.

To be more specific, let $\theta_0 = 0.5$ and suppose we want an MP test of level $.05$ for testing $\mathrm{H}_0 : \theta = 0.5$ versus $\mathrm{H}_1 : \theta > 0.5$. Such a test should reject the null hypothesis if and only if $\sum x_i > c$, and c should be chosen so that $\mathrm{P}_{\theta=0.5}\{\sum X_i > c\} = .05$. When $\theta = 0.5$, then $\sum X_i \sim \mathrm{Binomial}(100, 0.5)$. So c should be chosen so that $\mathrm{P}\{Y > c\} = .05$ where $Y \sim \mathrm{Binomial}(100, 0.5)$. We may as well suppose c is an integer. Note that $\mathrm{P}\{Y > c\} = 1 - \mathrm{P}\{Y \le c\}$. Using the Matlab function `binocdf(x,100,0.5)` for integers `x` between 0 and 100, we find that $\mathrm{P}\{Y > 57\} = .0666$ and $\mathrm{P}\{Y > 58\} = .0443$. Due to the discreteness of the binomial distribution, we cannot achieve an MP level $.05$ test. (We can achieve it if we allow the use of randomized tests, as in Example 8.3.6, but randomized tests are unappealing to most statisticians and are not used in practice.) But we can say that the test that rejects H_0 when $\sum x_i > 58$ is an MP level $.0443$ test for testing $\mathrm{H}_0 : \theta = 0.5$ versus $\mathrm{H}_1 : \theta > 0.5$. ||

It is often true, as in the example above, that the likelihood ratio that is used to define an MP rejection region is a somewhat complicated expression. Since the rejection region is a set of vectors $\boldsymbol{x}$, we want to re-express the condition $f(\boldsymbol{x}\,;\theta_1)/f(\boldsymbol{x}\,;\theta_0) > k$ in a form that is as simple as possible with regard to $\boldsymbol{x}$. Thus we want to collect together, to the extent possible, all terms involving $\boldsymbol{x}$.

Example 2 (cont'd). Suppose X_1, X_2, X_3, X_4, X_5 are i.i.d. $\mathrm{Normal}(\mu, 9)$ for some $\mu > 0$. Their joint pdf is given at the top of p. 2. Let us test $\mathrm{H}_0 : \mu = \mu_0$ versus $\mathrm{H}_1 : \mu = \mu_1$. The likelihood ratio is

$$\frac{f(\boldsymbol{x}\,;\mu_1)}{f(\boldsymbol{x}\,;\mu_0)} = \frac{(3\sqrt{2\pi})^{-5}\exp\left[-\frac{1}{18}\sum_{i=1}^{5}(x_i-\mu_1)^2\right]}{(3\sqrt{2\pi})^{-5}\exp\left[-\frac{1}{18}\sum_{i=1}^{5}(x_i-\mu_0)^2\right]} = \frac{\exp\left[-\frac{1}{18}\sum_{i=1}^{5}(x_i-\mu_1)^2\right]}{\exp\left[-\frac{1}{18}\sum_{i=1}^{5}(x_i-\mu_0)^2\right]}\,.$$

It is not clear what this says about $\boldsymbol{x}$, and so we try algebraic manipulations in an attempt to simplify it with regard to $\boldsymbol{x}$. Manipulations that help in this case are $\exp(a)/\exp(b) = \exp(a-b)$ and $\sum_{i=1}^{k} a_i + \sum_{i=1}^{k} b_i = \sum_{i=1}^{k}(a_i + b_i)$ and $(x-a)^2 - (x-b)^2 = 2(b-a)x + a^2 - b^2$. This leads to

$$\frac{f(\boldsymbol{x}\,;\mu_1)}{f(\boldsymbol{x}\,;\mu_0)} = \exp\left[\frac{\mu_1-\mu_0}{9}\sum_{i=1}^{5}x_i - \frac{5}{18}(\mu_1^2 - \mu_0^2)\right]\,.$$

Now we see that

$$\frac{f(\boldsymbol{x}\,;\mu_1)}{f(\boldsymbol{x}\,;\mu_0)} > k \quad \text{iff} \quad \exp\Big[\tfrac{\mu_1-\mu_0}{9}\sum_{i=1}^{5} x_i - \tfrac{5}{18}(\mu_1^2-\mu_0^2)\Big] > k$$

$$\text{iff} \quad \tfrac{\mu_1-\mu_0}{9}\sum_{i=1}^{5} x_i - \tfrac{5}{18}(\mu_1^2-\mu_0^2) > \log k$$

$$\text{iff} \quad \tfrac{\mu_1-\mu_0}{9}\sum_{i=1}^{5} x_i > \log k + \tfrac{5}{18}(\mu_1^2-\mu_0^2)\ .$$

If $\mu_1 > \mu_0$, then $\mu_1 - \mu_0 > 0$, and so the inequality is equivalent to

$$\sum_{i=1}^{5} x_i > \frac{9}{\mu_1-\mu_0}\left[\log k + \tfrac{5}{18}(\mu_1^2-\mu_0^2)\right] = \text{(say) } c\ .$$

Therefore, the MP rejection region can be expressed as $\mathcal{R}^* = \{\boldsymbol{x} : \sum x_i > c\}$. This defines a MP level α test where $\alpha = \mathrm{P}_{\mu_0}\{\sum X_i > c\}$.

Suppose we want a particular level α. We need to find a suitable value of c. Note that $\sum_{i=1}^{5} X_i \sim \text{Normal}(5\mu, 45)$, so $\mathrm{P}_{\mu_0}\{\sum X_i > c\} = \mathrm{P}_{\mu_0}\{(\sum X_i - 5\mu_0)/\sqrt{45} > (c-5\mu_0)/\sqrt{45}\} = \mathrm{P}\{Z > (c-5\mu_0)/\sqrt{45}\} = 1 - \Phi((c-5\mu_0)/\sqrt{45}\,)$. Setting $1 - \Phi((c-5\mu_0)/\sqrt{45}\,) = \alpha$, we can solve for $c = 5\mu_0 + \sqrt{45}\,\Phi^{-1}(1-\alpha)$. Thus, if $\mu_1 > \mu_0$, the following test is a MP level α test of $\mathrm{H}_0 : \mu = \mu_0$ versus $\mathrm{H}_1 : \mu = \mu_1$: reject H_0 if and only if $\sum X_i > 5\mu_0 + \sqrt{45}\,\Phi^{-1}(1-\alpha)$. Note that this does not involve the specific value of μ_1. So the test is a UMP level α test of $\mathrm{H}_0 : \mu = \mu_0$ versus $\mathrm{H}_1 : \mu > \mu_0$. ||

Note that an MP test always exists for testing a simple null hypothesis versus a simple alternative hypothesis (although one may be limited to certain values of α when dealing with discrete distributions as in Example 1 above, unless one is willing to do a randomized test). But only in special cases do UMP tests exist. In most of these special cases, the ~~set of~~ OK OK ~~alternative~~ parameter~~s~~ has dimension 1 and ~~the hypothesis~~ is one-sided. Even then, however, a UMP test may not exist. For example, given a single observation x from a Cauchy distribution with unknown median θ, i.e., with pdf $f(x\,;\theta) = 1/\{\pi[1+(x-\theta)^2]\}$, there is no UMP test of $\mathrm{H}_0 : \theta = \theta_0$ versus $\mathrm{H}_1 : \theta > \theta_0$.

The textbook presents several examples of MP tests. These are summarized below.

Example 8.3.1. Suppose $X_1, \ldots, X_n$ are i.i.d. Normal(μ, σ_0^2) with unknown mean μ and known variance σ_0^2. We want to test $\mathrm{H}_0 : \mu = \mu_0$ versus $\mathrm{H}_1 : \mu = \mu_1$. Suppose $\mu_1 > \mu_0$. Then the likelihood ratio can be simplified so that the MP level α test can be expressed as: reject H_0 if $\overline{X} > \mu_0 + (\sigma_0/\sqrt{n})\Phi^{-1}(1-\alpha)$. Note that this is a UMP level α test for testing $\mathrm{H}_0 : \mu = \mu_0$ versus $\mathrm{H}_1 : \mu > \mu_0$. ||

Example 8.3.2. Suppose $X_1, \dots, X_n$ are i.i.d. Normal(μ, σ_0^2) with unknown mean μ and known variance σ_0^2. We want to test $\mathrm{H}_0 : \mu = \mu_0$ versus $\mathrm{H}_1 : \mu = \mu_1$. Suppose $\mu_1 < \mu_0$. Then the MP level α test can be expressed as: reject H_0 if $\overline{X} < \mu_0 - (\sigma_0/\sqrt{n})\Phi^{-1}(1-\alpha)$. Note that this is a UMP level α test for testing $\mathrm{H}_0 : \mu = \mu_0$ versus $\mathrm{H}_1 : \mu < \mu_0$. ||

Example 8.3.3. Suppose $X_1, \dots, X_n$ are i.i.d. Exponential(β) with unknown mean β. We want to test $\mathrm{H}_0 : \beta = \beta_0$ versus $\mathrm{H}_1 : \beta = \beta_1$. Suppose $\beta_1 > \beta_0$. Then the likelihood ratio can be simplified so that the MP level α test can be expressed as: reject H_0 if $\sum X_i > c$. Note that this is a UMP level α test for testing $\mathrm{H}_0 : \beta = \beta_0$ versus $\mathrm{H}_1 : \beta > \beta_0$.

In order to find c, we must know the distribution of $\sum X_i$ when $\beta = \beta_0$. Since $X_i \sim$ Exponential(β) = Gamma$(1, \beta)$, it follows from Theorem 4.3.2(ii) that $\sum X_i \sim$ Gamma(n, β). For any desired level α, we could use the `gaminv` function in Matlab to calculate a suitable value of c. Or if a computer function was not available, we could express the test as: reject H_0 if $T = (2/\beta_0)\sum X_i > c'$. Then $T \sim$ Gamma$(n, (2/\beta_0)\beta)$, and so under H_0, $T \sim$ Gamma$(n, 2) = \chi^2(2n)$. Hence the critical value c' can be obtained from a chi-squared table. ||

(✱)

Example 8.3.4. Suppose $X_1, \dots, X_n$ are i.i.d. Uniform$(0, \theta)$ with unknown parameter θ. We want to test $\mathrm{H}_0 : \theta = \theta_0$ versus $\mathrm{H}_1 : \theta = \theta_1$. Suppose $\theta_1 > \theta_0$. In this example the two pdf's do not have the same support, and so we should use version 2 of the Neyman-Pearson Lemma. Using the N-P Lemma in this example turns out to be tricky. First let us guess what a good test might be. A sufficient statistic for the sample is the maximum $X_{(n)}$ (see Example 6.2.13). Since $\theta_1 > \theta_0$, a sensible test is to reject H_0 when $X_{(n)} > c$. The value of c must be chosen so that $\mathrm{P}_{\theta_0}\{\, X_{(n)} > c \,\} = \alpha$. We have $\mathrm{P}_{\theta_0}\{\, X_{(n)} > c \,\} = 1 - \mathrm{P}_{\theta_0}\{\, X_{(n)} \le c \,\}$ $= 1 - \prod_{i=1}^n \mathrm{P}_{\theta_0}\{\, X_i \le c \,\} = 1 - \prod_{i=1}^n (c/\theta_0) = 1 - (c/\theta_0)^n$. Setting $1 - (c/\theta_0)^n = \alpha$, we solve to get $c = \theta_0(1-\alpha)^{1/n}$.

$X_{(n)} = X_{n:r}$

Now we want to find ~~a~~ some $k > 0$ such that the conditions of the N-P Lemma, version 2, are satisfied by $\mathcal{R}^* = \{\, \boldsymbol{x} : x_{(n)} > c \,\}$. The joint pdf is $f(\boldsymbol{x}\,; \theta) = \prod_{i=1}^n f(x_i\,; \theta) =$ $\prod_{i=1}^n (1/\theta) I_{(0,\theta)}(x_i) = (1/\theta^n) I_{(0,\theta)}(x_{(n)})$. The conditions are

$$(1/\theta_1^n) I_{(0,\theta_1)}(x_{(n)}) > k(1/\theta_0^n) I_{(0,\theta_0)}(x_{(n)}) \Rightarrow x_{(n)} > c$$

$$(1/\theta_1^n) I_{(0,\theta_1)}(x_{(n)}) < k(1/\theta_0^n) I_{(0,\theta_0)}(x_{(n)}) \Rightarrow x_{(n)} \le c$$

The trick is to set $k = \theta_0^n/\theta_1^n$. Then the conditions become

$$I_{(0,\theta_1)}(x_{(n)}) > I_{(0,\theta_0)}(x_{(n)}) \Rightarrow x_{(n)} > c$$

$$I_{(0,\theta_1)}(x_{(n)}) < I_{(0,\theta_0)}(x_{(n)}) \Rightarrow x_{(n)} \le c$$

which reduce to

$$\theta_0 \leq x_{(n)} < \theta_1 \;\Rightarrow\; x_{(n)} > c\,.$$

This is true because $c = \theta_0(1-\alpha)^{1/n} < \theta_0$.

Note that this is a UMP level α test for testing $\mathrm{H}_0 : \theta = \theta_0$ versus $\mathrm{H}_1 : \theta > \theta_0$. ||

Example 8.3.5. Suppose $X_1, \dots, X_n$ are i.i.d. Gamma(δ_0, β) where $\delta_0 > 0$ is known and $\beta > 0$ is an unknown parameter. Let us test $\mathrm{H}_0 : \beta = \beta_0$ versus $\mathrm{H}_1 : \beta = \beta_1$. Suppose $\beta_1 > \beta_0$. This generalizes Example 8.3.3 which is the special case in which $\delta_0 = 1$. The likelihood ratio can be simplified so that the MP level α test can be expressed as: reject H_0 if $\sum X_i > c$. Note that this is a UMP level α test for testing $\mathrm{H}_0 : \beta = \beta_0$ versus $\mathrm{H}_1 : \beta > \beta_0$. For finding the value of c to achieve a given level α, we can use the fact that $\sum X_i \sim \text{Gamma}(n\delta_0, \beta)$. ||

Example. Suppose $X_1, \dots, X_n$ are i.i.d. Gamma(δ, β_0) where $\delta > 0$ is an unknown parameter and $\beta_0 > 0$ is known. Let us test $\mathrm{H}_0 : \delta = \delta_0$ versus $\mathrm{H}_1 : \delta = \delta_1$. Suppose $\delta_1 > \delta_0$. Then the likelihood ratio can be simplified so that the MP level α test can be expressed as: reject H_0 if $\sum \log X_i > c$. Note that this is a UMP level α test for testing $\mathrm{H}_0 : \delta = \delta_0$ versus $\mathrm{H}_1 : \delta > \delta_0$. For finding the value of c to achieve a given level α, we would like to know the cdf of $\sum \log X_i$, which is a sum of log-gamma random variables, but as far as I know, there are no computer functions or tables for it. One could use a normal approximation if n was large. If n was small, one could compute a saddlepoint approximation, or one could simulate the distribution of $\sum \log X_i$ under H_0 by simulating Gamma(δ_0, β_0) random variables (e.g., use `gamrnd` in Matlab). ||

Example 8.3.6. Suppose $X_1, \dots, X_n$ are i.i.d. Bernoulli(θ) with unknown proportion θ, $0 < \theta < 1$. We want to test $\mathrm{H}_0 : \theta = \theta_0$ versus $\mathrm{H}_1 : \theta = \theta_1$. Suppose $\theta_1 > \theta_0$. Then the likelihood ratio can be simplified so that the MP level α test can be expressed as: reject H_0 if $\sum X_i > c$. Note that this is a UMP level α test for testing $\mathrm{H}_0 : \theta = \theta_0$ versus $\mathrm{H}_1 : \theta > \theta_0$. For finding a suitable c, we can use the fact that $\sum X_i \sim \text{Binomial}(n, \theta)$. Due to the discreteness of the Binomial distribution, only a limited number of levels α can be achieved (unless a randomized test is used). ||

Example 8.3.7. Suppose $X_1, \dots, X_n$ are i.i.d. Poisson(λ) with unknown mean $\lambda > 0$. We want to test $\mathrm{H}_0 : \lambda = \lambda_0$ versus $\mathrm{H}_1 : \lambda = \lambda_1$. Suppose $\lambda_1 > \lambda_0$. Then the likelihood ratio can be simplified so that the MP level α test can be expressed as: reject H_0 if $\sum X_i > c$. Note that this is a UMP level α test for testing $\mathrm{H}_0 : \lambda = \lambda_0$ versus $\mathrm{H}_1 : \lambda > \lambda_0$. For finding a suitable c, we can use the fact that $\sum X_i \sim \text{Poisson}(n\lambda)$. Due to the discreteness of the

Poisson distribution, only a limited number of levels α can be achieved (unless a randomized test is used). $\|$

Example. Suppose $X_1, \dots, X_n$ are i.i.d. Normal(μ, σ^2) with unknown mean μ and unknown variance σ^2. Suppose we want to test $\mathrm{H}_0 : \mu = 0 \,,\, \sigma^2 = 1$ versus $\mathrm{H}_1 : \mu = 2 \,,$ $\sigma^2 = 1$. This is the same as testing $\mathrm{H}_0 : \mu = 0$ versus $\mathrm{H}_1 : \mu = 2$ when σ^2 is known to be equal to 1. As seen in Example 8.3.1 above, an MP level α has the form: reject H_0 iff $\sum x_i > c$.

Now suppose we want to test $\mathrm{H}_0 : \mu = 0 \,,\, \sigma^2 = 1$ versus $\mathrm{H}_1 : \mu = 2 \,,\, \sigma^2 = 4$. The likelihood ratio is

$$\frac{f(\boldsymbol{x}\,;2{,}4)}{f(\boldsymbol{x}\,;0{,}1)} = \frac{(\sqrt{8\pi})^{-n}\exp[-\frac{1}{8}\sum(x_i-2)^2]}{(\sqrt{2\pi})^{-n}\exp[-\frac{1}{2}\sum x_i^2]}$$

$$= 2^{-n}\exp[\,\tfrac{3}{8}\textstyle\sum x_i^2 + \tfrac{1}{2}\sum x_i - \tfrac{n}{2}] \,.$$

Thus we see that the MP test rejects H_0 iff $\frac{3}{8}\sum x_i^2 + \frac{1}{2}\sum x_i > c$. By "completing the square", we can re-express the test as rejecting iff $\sum(x_i + \frac{2}{3})^2 > c'$.

Suppose we want to test $\mathrm{H}_0 : \mu = 0 \,,\, \sigma^2 = 1$ versus $\mathrm{H}_1 : \mu = 2 \,,\, \sigma^2 = 9$. By manipulating the likelihood ratio, the MP test can be expressed as: reject H_0 iff $\sum(x_i + \frac{1}{4})^2 > c$. Therefore, there is no UMP test of $\mathrm{H}_0 : \mu = 0 \,,\, \sigma^2 = 1$ versus $\mathrm{H}_1 : \mu > 0 \,,\, \sigma^2 > 1$. As noted earlier, we would not expect a UMP test here, because it is very rare for a UMP test to exist if the alternative hypothesis is 2-dimensional. $\|$

Example. Suppose $X_1, \dots, X_n$ are i.i.d. from a Cauchy distribution with unknown median θ, i.e., with pdf $f(x\,;\theta) = 1/\{\pi[1 + (x - \theta)^2]\}$. Let us test $\mathrm{H}_0 : \theta = 0$ versus $\mathrm{H}_1 : \theta = \theta_1$ for a value $\theta_1 > 0$. The likelihood ratio is

$$\frac{f(\boldsymbol{x}\,;0)}{f(\boldsymbol{x}\,;\theta_1)} = \frac{(\pi)^{-n}\prod_{i=1}^{n}[1+(x_i-\theta_1)^2]^{-1}}{(\pi)^{-n}\prod_{i=1}^{n}[1+x_i^2]^{-1}} = \prod_{i=1}^{n}\frac{1+x_i^2}{1+(x_i-\theta_1)^2} \,.$$

One tries to simplify the likelihood ratio with regard to $\boldsymbol{x}$, but it seems that there is no "nice" expression for the MP rejection region. It turns out that the MP rejection region is different for different values of θ_1. Therefore there is no UMP test of $\mathrm{H}_0 : \theta = 0$ versus $\mathrm{H}_1 : \theta > 0$. Finding the value of c to achieve a given level α would be difficult. One could simulate the distribution of the likelihood ratio under H_0 by simulating Cauchy random variables. $\|$

The N-P Lemma does not require that the distribution in the alternative hypothesis be in the same family as the distribution in the null hypothesis.

Example 8.3.10. Suppose $X_1, \ldots, X_n$ are i.i.d. from some population with an unknown distribution D. Let us test $\mathrm{H}_0 : D = \text{Normal}(0, \frac{1}{2})$ versus $\mathrm{H}_1 : D = \text{Cauchy with median } 0$. Under H_0, the joint pdf is $f_0(\boldsymbol{x}) = \pi^{-n/2}\exp(-\sum x_i^2)$, and under H_1, the joint pdf is $f_1(\boldsymbol{x}) = \pi^{-n}\prod(1+x_i^2)^{-1}$. An MP rejection region is defined by $f_1(\boldsymbol{x})/f_0(\boldsymbol{x}) > k$, which is equivalent to $\prod[\exp(x_i^2)/(1+x_i^2)] > k'$, or $\sum[x_i^2 - \log(1+x_i^2)] > c$. This rejection region does not have a "nice" form unless $n = 1$. For $n = 1$, the MP region is $\mathcal{R}^* = \{x : x^2 - \log(1+x^2) > c\}$. Let $h(t) = t - \log(1+t)$, so that $\mathcal{R}^* = \{x : h(x^2) > c\}$. The derivative of $h(t)$ is $\frac{\mathrm{d}}{\mathrm{d}t}h(t) = 1 - 1/(1+t) = t/(1+t) > 0$ for all $t > 0$. Therefore, $h(t)$ is a strictly increasing function, so $h(x^2) > c$ iff $x^2 > c'$ iff $|x| > c''$. ||

The N-P Lemma is not restricted to i.i.d. samples.

Example. Suppose $X_1, \ldots, X_n$ are independent with distributions $X_i \sim \text{Normal}(\beta w_i, \sigma^2)$. This is the model for "regression through the origin". That is, the data are assumed to follow a linear regression model with intercept 0. Then

$$f(\boldsymbol{x}\,;\beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}(x_i - \beta w_i)^2\right]$$

and the likelihood ratio for testing $\mathrm{H}_0 : \beta = \beta_0$ versus $\mathrm{H}_1 : \beta = \beta_1$ can be simplified to

$$\frac{f(\boldsymbol{x}\,;\beta_1)}{f(\boldsymbol{x}\,;\beta_0)} = \exp\left[\frac{1}{\sigma^2}\left\{(\beta_1 - \beta_0)\sum_{i=1}^{n} w_i x_i - \tfrac{1}{2}(\beta_1^2 - \beta_0^2)\sum_{i=1}^{n} w_i^2\right\}\right].$$

For $\beta_1 > \beta_0$, the MP rejection region can be expressed as $\sum_{i=1}^{n} w_i x_i > c$. Note that this is UMP for testing $\mathrm{H}_0 : \beta = \beta_0$ versus $\mathrm{H}_1 : \beta > \beta_0$. ||

The N-P Lemma is not restricted to samples of independent observations, as seen in Example 8.3.12.

Lemma. Let $\boldsymbol{x}$ be a data vector with joint pmf or pdf $f(\boldsymbol{x}\,;\boldsymbol{\theta})$ for some $\boldsymbol{\theta} \in \Theta$. Suppose $T(\boldsymbol{x})$ is a sufficient statistic for $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ be in Θ. An MP test of $\mathrm{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $\mathrm{H}_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$ can be expressed so that it involves $\boldsymbol{x}$ only through $T(\boldsymbol{x})$.

Proof. By the Factorization Theorem (Theorem 6.2.1), the joint pmf or pdf can be factored as $f(\boldsymbol{x}\,;\boldsymbol{\theta}) = g(T(\boldsymbol{x});\boldsymbol{\theta})h(\boldsymbol{x})$. Now the likelihood ratio is

$$\frac{f(\boldsymbol{x}\,;\boldsymbol{\theta}_1)}{f(\boldsymbol{x}\,;\boldsymbol{\theta}_0)} = \frac{g(T(\boldsymbol{x});\boldsymbol{\theta}_1)h(\boldsymbol{x})}{g(T(\boldsymbol{x});\boldsymbol{\theta}_0)h(\boldsymbol{x})} = \frac{g(T(\boldsymbol{x});\boldsymbol{\theta}_1)}{g(T(\boldsymbol{x});\boldsymbol{\theta}_0)},$$

which is function of $\boldsymbol{x}$ only through $T(\boldsymbol{x})$. □

Uniformly most powerful tests

In several of the examples above we noticed that an MP test derived from the N-P Lemma was actually a UMP test. In all of these examples the model has the following property.

Definition. Let $f(\boldsymbol{x}\,;\theta)$, $\theta \in \Theta \subset \mathbb{R}^1$, be a 1-parameter family of pdf's or pmf's and let $T(\boldsymbol{x})$ be a real-valued statistic. We say that the family has the *monotone likelihood ratio* (MLR) property in $T(\boldsymbol{x})$ if, for all $\theta_0 < \theta_1$, the likelihood ratio $f(\boldsymbol{x}\,;\theta_1)/f(\boldsymbol{x}\,;\theta_0)$ is an increasing (or more generally, nondecreasing) function of $T(\boldsymbol{x})$.

Note that the MLR property can be verified without finding the pdf or pmf of T . This property allows one to construct UMP tests.

Karlin-Rubin Theorem. Suppose $f(\boldsymbol{x}\,;\theta)$, $\theta \in \Theta \subset \mathbb{R}^1$, is a 1-parameter family of pdf's or pmf's that has the MLR property in a statistic $T(\boldsymbol{x})$. A UMP level α test of $\mathrm{H}_0 : \theta \le \theta_0$ versus $\mathrm{H}_1 : \theta > \theta_0$ is obtained by rejecting H_0 iff $T(\boldsymbol{x}) > c$ where $\mathrm{P}_{\theta_0}\{T(\boldsymbol{x}) > c\} = \alpha$.

Sketch of the proof. To get an MP test of $\mathrm{H}_0 : \theta = \theta_0$ versus $\mathrm{H}_1 : \theta = \theta_1$, we apply the N-P Lemma, which tells us to reject H_0 iff $f(\boldsymbol{x}\,;\theta_1)/f(\boldsymbol{x}\,;\theta_0) > k$. The MLR property implies that if $\theta_1 > \theta_0$, then the likelihood ratio is greater than k if and only if $T(\boldsymbol{x})$ is greater than some c . By expressing the test in terms of $T(\boldsymbol{x})$, we see that it does not depend on the particular value of θ_1 but only requires that $\theta_1 > \theta_0$. Therefore, the test described in the theorem is a UMP test of $\mathrm{H}_0 : \theta = \theta_0$ versus $\mathrm{H}_1 : \theta > \theta_0$.

It can also be shown that this test is a UMP test of $\mathrm{H}_0 : \theta \le \theta_0$ versus $\mathrm{H}_1 : \theta > \theta_0$. The main issue here is to show that the test has the correct level, that is, $\mathrm{P}_{\theta}\{T(\boldsymbol{x}) > c\} \le \mathrm{P}_{\theta_0}\{T(\boldsymbol{x}) > c\}$ for all $\theta \le \theta_0$. □

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Normal(μ, σ_0^2) with σ_0^2 known. Take any $\mu_0 < \mu_1$.

$$\frac{f(\boldsymbol{x}\,;\mu_1)}{f(\boldsymbol{x}\,;\mu_0)} = \frac{\left(1/\sqrt{2\pi\sigma_0^2}\right)^n \exp[-\frac{1}{2\sigma_0^2}\sum(x_i-\mu_1)^2]}{\left(1/\sqrt{2\pi\sigma_0^2}\right)^n \exp[-\frac{1}{2\sigma_0^2}\sum(x_i-\mu_0)^2]} = \cdots (\textstyle\sum x_i^2 \text{ cancels out}) \cdots$$

$$= \exp\Big[\Big(\frac{\mu_1-\mu_0}{\sigma_0^2}\Big)\sum x_i - \frac{n(\mu_1^2-\mu_0^2)}{2\sigma_0^2}\Big] \ .$$

Note that $(\mu_1 - \mu_0)/\sigma_0^2 > 0$ because $\mu_0 < \mu_1$. Therefore, as $\sum x_i$ increases, the exponent increases, and so the likelihood ratio increases. Thus this family of pdf's has the MLR property. Now we can apply the Karlin-Rubin Theorem and conclude that a UMP level α test of $\mathrm{H}_0 : \mu \le \mu_0$ versus $\mathrm{H}_1 : \mu > \mu_0$ is obtained by rejecting H_0 iff $\sum x_i > c$ where $\mathrm{P}_{\theta_0}\{\sum X_i > c\} = \alpha$. ||

The Normal(μ, σ_0^2) family of pdf's is a special case of a *1-parameter full-rank exponential family*, which (see Section 3.8) is a family of pdf's or pmf's of the form

$$(*) \qquad f(\boldsymbol{x}\,;\theta) = a(\theta)g(\boldsymbol{x})\exp[b(\theta)R(\boldsymbol{x})]$$

for $\theta \in \Theta \subset \mathbb{R}^1$. For an i.i.d. sample from a Normal(μ, σ_0^2) population,

$$f(\boldsymbol{x}\,;\mu) = \left(1/\sqrt{2\pi\sigma_0^2}\right)^n \exp\Big[-\tfrac{1}{2\sigma_0^2}\textstyle\sum(x_i-\mu)^2\Big]$$

$$= \left(1/\sqrt{2\pi\sigma_0^2}\right)^n \exp\Big[-\tfrac{n\mu^2}{2\sigma_0^2}\Big]\exp\Big[-\tfrac{1}{2\sigma_0^2}\textstyle\sum x_i^2\Big]\exp\Big[\tfrac{\mu}{\sigma_0^2}\textstyle\sum x_i\Big]$$

which has the form in (*) with $a(\mu) = \left(1/\sqrt{2\pi\sigma_0^2}\right)^n \exp[-n\mu^2/2\sigma_0^2]$, $g(\boldsymbol{x}) = \exp[-(1/2\sigma_0^2)\sum x_i^2]$, $b(\mu) = \mu/\sigma_0^2$, and $R(\boldsymbol{x}) = \sum x_i$. Note that the expression in (*) is not entirely unique. For example, we could also let $b(\mu) = n\mu/\sigma_0^2$, and $R(\boldsymbol{x}) = \overline{x}$.

Lemma. A 1-parameter full-rank exponential family, for which $b(\theta)$ is a strictly increasing function, has the MLR property in $R(\boldsymbol{x})$.

Proof. For $\theta_0 < \theta_1$,

$$\frac{f(\boldsymbol{x}\,;\theta_1)}{f(\boldsymbol{x}\,;\theta_0)} = \frac{a(\theta_1)g(\boldsymbol{x})\exp[b(\theta_1)R(\boldsymbol{x})]}{a(\theta_0)g(\boldsymbol{x})\exp[b(\theta_0)R(\boldsymbol{x})]} = \frac{a(\theta_1)}{a(\theta_0)}\exp[[b(\theta_1)-b(\theta_0)]R(\boldsymbol{x})]\ .$$

Since $\theta_0 < \theta_1$, then $b(\theta_1) - b(\theta_0) > 0$. Therefore, as $R(x)$ increases, the exponent increases, and so the likelihood ratio increases. □

Examples of 1-parameter full-rank exponential families for which $b(\theta)$ is a strictly increasing function occur for i.i.d. samples from:

family	$b(\theta)$	$R(\boldsymbol{x})$
Normal(μ, σ_0^2), $-\infty < \mu < \infty$	μ/σ_0^2	$\sum x_i$
Normal(μ_0, σ^2), $\sigma^2 > 0$	$-1/2\sigma^2$	$\sum(x_i - \mu_0)^2$
Exponential(β), $\beta > 0$	$-1/\beta$	$\sum x_i$
Gamma(δ_0, β), $\beta > 0$	$-1/\beta$	$\sum x_i$
Gamma(δ, β_0), $\delta > 0$	$-1/\beta$	$\sum \log x_i$
Bernoulli(θ), $0 < \theta < 1$	$\log(\theta/(1-\theta))$	$\sum x_i$
Poisson(λ), $\lambda > 0$	$\log(\lambda)$	$\sum x_i$

Some families that are not 1-parameter full-rank exponential families nevertheless have the MLR property.

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Uniform$(0,\theta)$, $\theta > 0$. The joint pdf is $f(\boldsymbol{x};\theta) = (1/\theta^n) I_{(0,\theta)}(x_{(n)})$. This is not an exponential family because the pdf's do not all have the same support. Take any $\theta_0 < \theta_1$. The likelihood ratio is

$$\frac{f(\boldsymbol{x};\theta_1)}{f(\boldsymbol{x};\theta_0)} = \frac{(1/\theta_1^n) I_{(0,\theta_1)}(x_{(n)})}{(1/\theta_0^n) I_{(0,\theta_0)}(x_{(n)})} \,.$$

If $0 < x_{(n)} < \theta_0$, then both indicators are 1, so the value of the likelihood ratio is θ_0^n/θ_1^n. If $\theta_0 \le x_{(n)} < \theta_1$, then the numerator is positive and the denominator is 0 and so the ratio is ∞. If $\theta_1 \le x_{(n)}$, then both indicators are 0, so the ratio is $0/0$, which is indeterminate. But we can ignore the case when $\theta_1 \le x_{(n)}$ because this event has probability 0 under both distributions. Thus we see that the likelihood is nondecreasing as a function of $x_{(n)}$ and hence the family of joint pdf's has the MLR property.

By the Karlin-Rubin Theorem, a UMP test of $\mathrm{H}_0 : \theta \le \theta_0$ versus $\mathrm{H}_1 : \theta > \theta_0$ is obtained by rejecting H iff $x_{(n)} > c$. As noted on p. 12 above, to obtain a UMP level α test, let $c = \theta_0(1-\alpha)^{1/n}$. ||

Numerical example. Suppose $X_1, \ldots, X_5$ are i.i.d. Normal$(\mu, 9)$. From the example on p. 16 above, we know that a UMP level $.05$ test of $\mathrm{H}_0 : \mu \le 200$ versus $\mathrm{H}_1 : \mu > 200$ is obtained by rejecting H_0 iff $\sum x_i > c$ or, equivalently, iff $\overline{x} > b$ where $\mathrm{P}_{\mu=200}\{\overline{X} > b\} = .05$. From the fact that $\overline{X} \sim$ Normal$(200, 1.8)$ when $\mu = 200$, we find that $c = 202.2068$. Let us check the power of this UMP test and compare it to some other reasonable tests. Its power at $\mu = 205$ is $\mathrm{P}_{\mu=205}\{\overline{X} > 202.2068\} =$ P{Normal$(205, 1.8) > 202.2068\} = .9813$.

Since the sample mean $\overline{X}$ is a reasonable estimator of the population mean μ, it makes sense to reject the null hypothesis in favor of $\mu > 200$ if $\overline{X} > b$ for a suitable value of b. Since μ is also the population median, the sample median $X_{(3)}$ is also a reasonable estimator of μ. A sensible test would be to reject the null hypothesis if $X_{(3)} > d$ for a suitable value of d. The distribution of $X_{(3)}$ is not a well-known distribution, and so to obtain d we might resort to simulation. That is, we can simulate a sample of 5 independent observations from the Normal$(200, 9)$ distribution and calculate the median and repeat this until we have, say, $10,000$ medians. The 95-th percentile of these $10,000$ numbers is an estimate of d. In one such simulation, this produced $d \approx 202.6$. To calculate an estimate of the power of this test at $\mu = 205$, we can simulate a sample of 5 independent observations from the Normal$(205, 9)$ distribution and calculate the median and repeat this until we have, say, $10,000$ medians. The proportion of these medians that are greater than 202.6 is an estimate of the power. In one such simulation, this produced a estimated power of $.93$. Of course this si less than the power of the UMP test.

Another test to try could be based on the sample maximum. Reject the null hypothesis if $X_{(5)} > k$ for a suitable value of k. Here again we could use simulation. One such simulation produced $k \approx 207.0$ with the power at $\mu = 205$ estimated to be $.76$. $\|$

Two-sided alternative hypotheses

Suppose we have a data vector $\boldsymbol{x}$ and have formulated a probability model with joint pdf or pmf $f(\boldsymbol{x}\,;\theta)$ for some 1-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^1$. The Karlin-Rubin Theorem concerns testing against a one-sided alternative. Sometimes we are interested in testing against a two-sided alternative, $\mathrm{H}: \theta = \theta_0$ versus $\mathrm{H}_1: \theta \neq \theta_0$. Is there a UMP level α test? To be a UMP level α test of $\mathrm{H}: \theta = \theta_0$ versus $\mathrm{H}_1: \theta \neq \theta_0$, the test must be a MP level α test of $\mathrm{H}_0: \theta = \theta_0$ versus $\mathrm{H}_1: \theta = \theta_1$ for all $\theta_1 \neq \theta_0$. Said another way, this same test must simultaneously be (a) a UMP level α test of $\mathrm{H}: \theta = \theta_0$ versus $\mathrm{H}_1: \theta > \theta_0$ and (b) a UMP level α test of $\mathrm{H}: \theta = \theta_0$ versus $\mathrm{H}_1: \theta < \theta_0$.

Only in special cases (for example, for 1-parameter families with the MLR property) does there exist a UMP test as in (a) or as in (b). In these special cases, typically the two tests are different and hence (if the two UMP tests are unique, which they typically are), there is no UMP test against the two-sided alternative.

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Normal(μ, σ_0^2) with σ_0^2 known. The UMP level α test of $\mathrm{H}_0: \mu = \mu_0$ versus $\mathrm{H}_1: \mu > \mu_0$ is unique, and it rejects H_0 iff $\overline{x} > c$ for a suitable c. The UMP level α test of $\mathrm{H}_0: \mu = \mu_0$ versus $\mathrm{H}_1: \mu < \mu_0$ is unique, and it rejects H_0 iff $\overline{x} < c'$ for a suitable c'. We see that these two tests are different, and so there is no UMP test of $\mathrm{H}_0: \mu = \mu_0$ versus $\mathrm{H}_1: \mu \neq \mu_0$. $\|$

The following example is an unusual instance in which there is a UMP test against a two-sided alternative.

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Uniform$(0, \theta)$. On p. 12 above we found a UMP level α test of $\mathrm{H}_0: \theta = \theta_0$ versus $\mathrm{H}_1: \theta > \theta_0$. Is this test also UMP versus $\mathrm{H}_1: \theta < \theta_0$? No, it is not, but in this example a UMP level α test of $\mathrm{H}_0: \theta = \theta_0$ versus $\mathrm{H}_1: \theta > \theta_0$ is not unique. Consider the following two tests:

Test 1. Reject $\theta = \theta_0$ in favor of $\theta > \theta_0$ if $x_{(n)} > \theta_0(1-\alpha)^{1/n}$.

Test 2. Reject $\theta = \theta_0$ in favor of $\theta > \theta_0$ if $x_{(n)} > \theta_0$ or $x_{(n)} < \theta_0\alpha^{1/n}$.

Test 1 is the UMP test from p. 12. To calculate its power, we can use the fact that $P_\theta\{X_{(n)} \le t\} = [P_\theta\{X_1 \le t\}]^n = (t/\theta)^n$. Its power at an alternative parameter $\theta_1 > \theta_0$ is $P_{\theta_1}\{X_{(n)} > \theta_0(1-\alpha)^{1/n}\} = 1 - P_{\theta_1}\{X_{(n)} \le \theta_0(1-\alpha)^{1/n}\} =$ $1 - [\theta_0(1-\alpha)^{1/n}/\theta_1]^n = 1 - (\theta_0/\theta_1)^n(1-\alpha)$. Next, the power of Test 2 at θ_1 is $P_{\theta_1}\{X_{(n)} > \theta_0 \text{ or } X_{(n)} < \theta_0\alpha^{1/n}\} = P_{\theta_1}\{X_{(n)} < \theta_0\alpha^{1/n}\} + 1 - P_{\theta_1}\{X_{(n)} \le \theta_0\} =$ $(\theta_0\alpha^{1/n}/\theta_1)^n + 1 - (\theta_0/\theta_1)^n = 1 - (\theta_0/\theta_1)^n(1-\alpha)$. We see that Test 2 has the same power as Test 1, so it is also a UMP level α test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$. It can be shown, by using version 2 of the N-P Lemma, that Test 2 is an MP level α test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ for all $0 < \theta_1 < \theta_0$. Therefore, Test 2 is a UMP level α test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \ne \theta_0$. ||

In situations in which there is no UMP test, we may decide to restrict our attention to a particular subclass of tests. We may be able to find a test that is UMP in the smaller class of tests. A sensible property for a test to have is to be unbiased. This means that there is a greater chance of rejecting the null hypothesis when it is false than when it is true. More precisely, a test is said to be *level α unbiased* if its power function satisfies $Q(\boldsymbol{\theta}) \le \alpha$ for all $\boldsymbol{\theta} \in \Theta_0$ and $Q(\boldsymbol{\theta}) \ge \alpha$ for all $\boldsymbol{\theta} \in \Theta_1$.

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Normal(μ, σ_0^2) with σ_0^2 known. As seen on p. 19 above, there is no UMP level α test of $H_0 : \mu = \mu_0$ versus $H_1 : \mu \ne \mu_0$. There is, however, a UMP level α unbiased test, i.e., a test that is UMP among all level α unbiased tests. This test is the one that rejects H_0 iff $|\overline{X} - \mu_0| > c$ for a suitable c. Equivalently, this is the two-sided z-test that rejects H_0 iff $z > k$ where $z = |\overline{X} - \mu_0|/(\sigma_0/\sqrt{n})$. ||

Notes on Confidence Interval Estimation

(see Ch. 9 in Mukhopadhyay)

Basic definitions

Let $\boldsymbol{x}$ be a vector of observed data. As a probability model for the data, suppose we have assumed a family of pdf's $f(\boldsymbol{x}\,;\theta)$ parameterized by a real-valued parameter $\theta \in \Theta \subset \mathbb{R}^1$. An *upper confidence limit* for θ is a real-valued statistic $T_{\mathrm{U}}(\boldsymbol{x})$. Actually it can be any statistic, but we call it an upper confidence limit if it is chosen with the intention that there is a high probability that $\theta < T_{\mathrm{U}}(\boldsymbol{X})$. Similarly, a *lower confidence limit* for θ is a real-valued statistic $T_{\mathrm{L}}(\boldsymbol{x})$ which is chosen with the intention that there is a high probability that $T_{\mathrm{L}}(\boldsymbol{X}) < \theta$. A *confidence interval* for θ is a pair $(T_{\mathrm{L}}(\boldsymbol{x}), T_{\mathrm{U}}(\boldsymbol{x}))$ consisting of a lower confidence limit and an upper confidence limit.

The *coverage probability* of a confidence interval is $\mathrm{P}_\theta\{T_{\mathrm{L}}(\boldsymbol{x}) < \theta < T_{\mathrm{U}}(\boldsymbol{x})\}$. This probability may depend on θ, but in many common examples, the coverage probability is the same for all θ. The *confidence coefficient* of a confidence interval is the infimum of the coverage probabilities for all θ.

Example. Suppose $X_1, \ldots, X_5$ are i.i.d. Normal$(\mu, 9)$.

(a) Consider $(\overline{X}-1, \overline{X}+1)$ as a confidence interval for μ. Its coverage probability is $\mathrm{P}_\mu\{\overline{X}-1 < \mu < \overline{X}+1\} = \mathrm{P}_\mu\{-1 < \overline{X}-\mu < 1\} = \mathrm{P}\{-1 < \mathrm{Normal}(0, 1.8) < 1\} =$ $.5439$. This does not depend on μ and so it is also the confidence coefficient. In this experiment we can be $54\,\%$ confident that the true μ lies between $\overline{X}-1$ and $\overline{X}+1$.

(b) Next consider $(\frac{1}{2}\overline{X}, 2\overline{X})$ as a confidence interval for μ. Its coverage probability is

$$\begin{aligned}\mathrm{P}_\mu\{\tfrac{1}{2}\overline{X} < \mu < 2\overline{X}\} &= \mathrm{P}_\mu\{\tfrac{1}{2}\mu < \overline{X} < 2\mu\} \\ &= \mathrm{P}_\mu\{-\tfrac{1}{2}\mu < \overline{X}-\mu < \mu\} \\ &= \mathrm{P}_\mu\{-\tfrac{1}{2}\mu/\sqrt{1.8} < (\overline{X}-\mu)/\sqrt{1.8} < \mu/\sqrt{1.8}\} \\ &= \mathrm{P}_\mu\{-\tfrac{1}{2}\mu/\sqrt{1.8} < \mathrm{Normal}(0,1) < \mu/\sqrt{1.8}\} \\ &= \Phi(\mu/\sqrt{1.8}) - \Phi(-\tfrac{1}{2}\mu/\sqrt{1.8})\,.\end{aligned}$$

The coverage probability depends on μ. At $\mu = 0$, the coverage probability is $\Phi(0) - \Phi(0)$ $= 0$. So the confidence coefficient is 0. However, if we are able to assume that $\mu \geq 5$, then the confidence coefficient is $.9687$. ||

Inversion of a test

There is a close connection between the concept of hypothesis testing and the concept of confidence interval estimation. Suppose $(T_{\mathrm{L}}(\boldsymbol{X}), T_{\mathrm{U}}(\boldsymbol{X}))$ is a confidence interval for θ with confidence coefficient $.95$. Then we are 95% confident that the true θ will lie in the interval. A sensible test is obtained by accepting $\mathrm{H}_0 : \theta = \theta_0$ iff θ_0 is in the interval. Equivalently, we reject $\mathrm{H}_0 : \theta = \theta_0$ iff θ_0 is outside the interval. Such a test has level $.05$ because $\mathrm{P}_{\theta_0}\{\text{reject } \mathrm{H}_0\} = 1 - \mathrm{P}_{\theta_0}\{\text{accept } \mathrm{H}_0\} = 1 - \mathrm{P}_{\theta_0}\{T_{\mathrm{L}}(\boldsymbol{X}) < \theta_0 < T_{\mathrm{U}}(\boldsymbol{X})\}$. Since the confidence coefficient is $.95$, the coverage probability is $\geq .95$ for all θ. In particular, $\mathrm{P}_{\theta_0}\{T_{\mathrm{L}}(\boldsymbol{X}) < \theta_0 < T_{\mathrm{U}}(\boldsymbol{X})\} \geq .95$, and so $\mathrm{P}_{\theta_0}\{\text{reject } \mathrm{H}_0\} \leq 1 - .95 = .05$.

It is convenient to generalize the concept of a confidence interval. A *confidence region* for a real-valued parameter θ is a random set $\mathcal{C}(\boldsymbol{X})$ in the real line $\mathbb{R}^1$ defined in terms of the data vector. In practice we prefer regions that are intervals, $\mathcal{C}(\boldsymbol{X}) = (T_{\mathrm{L}}(\boldsymbol{X}), T_{\mathrm{U}}(\boldsymbol{X}))$ or $\mathcal{C}(\boldsymbol{X}) = (-\infty, T_{\mathrm{U}}(\boldsymbol{X}))$ or $\mathcal{C}(\boldsymbol{X}) = (T_{\mathrm{L}}(\boldsymbol{X}), \infty)$. But when discussing the theory of confidence intervals, it is convenient to allow more general regions. The *coverage probability* of the confidence region is $\mathrm{P}_\theta\{\theta \in \mathcal{C}(\boldsymbol{X})\}$. The *confidence coefficient* of the confidence region is the infimum of the coverage probabilities for all θ.

As with a confidence interval, a confidence region can be used to test hypotheses. Given a confidence region $\mathcal{C}(\boldsymbol{X})$ for θ, we can test $\mathrm{H}_0 : \theta = \theta_0$ by accepting it iff $\theta_0 \in \mathcal{C}(\boldsymbol{X})$. Let γ be the confidence coefficient of the confidence region, so that $\mathrm{P}_{\theta_0}\{\theta_0 \in \mathcal{C}(\boldsymbol{X})\} \geq \gamma$. Then the size of the corresponding test is $\mathrm{P}_{\theta_0}\{\text{reject } \mathrm{H}_0\} = 1 - \mathrm{P}_{\theta_0}\{\text{accept } \mathrm{H}_0\} = 1 - \mathrm{P}_{\theta_0}\{\theta_0 \in \mathcal{C}(\boldsymbol{X})\} \leq 1 - \gamma$, so the test has level $1 - \gamma$.

Note that the rejection region of the test can be described as $\mathcal{R}(\theta_0) = \{\boldsymbol{x} : \theta_0 \notin \mathcal{C}(\boldsymbol{x})\}$. This says that $\boldsymbol{x} \in \mathcal{R}(\theta_0)$ iff $\theta_0 \notin \mathcal{C}(\boldsymbol{x})$, which expresses the "inverse" relationship between rejection regions and confidence regions.

In the preceding paragraph we started with a confidence region with confidence coefficient γ and "inverted" it to obtain a test of level $1 - \gamma$. Conversely, given a testing procedure of level α for testing any simple null hypothesis, we can "invert" it to obtain a confidence region with confidence coefficient $\geq 1 - \alpha$. (Sometimes, but not always, the region is an interval.) The idea is that $\theta_0 \in \mathcal{C}(\boldsymbol{x})$ iff $\boldsymbol{x} \notin \mathcal{R}(\theta_0)$, or $\mathcal{C}(\boldsymbol{x}) = \{\theta_0 : \boldsymbol{x} \notin \mathcal{R}(\theta_0)\}$. That is, the confidence region consists of those values of θ_0 that would not be rejected by the given testing procedure for the observed data vector $\boldsymbol{x}$. Note that the coverage probability is $\mathrm{P}_{\theta_0}\{\theta_0 \in \mathcal{C}(\boldsymbol{X})\} = \mathrm{P}_{\theta_0}\{\boldsymbol{X} \notin \mathcal{R}(\theta_0)\} = 1 - \mathrm{P}_{\theta_0}\{\boldsymbol{X} \in \mathcal{R}(\theta_0)\} = 1 - \mathrm{P}_{\theta_0}\{\text{Type I error}\} \geq 1 - \alpha$.

The condition for θ_0 to be in the confidence region is $\boldsymbol{x} \notin \mathcal{R}(\theta_0)$. On the one hand, this condition completely defines the confidence region, but on the other hand, it can usually be re-expressed in a simpler form by "isolating" θ_0.

Example. Consider the level .05 test in the example on pp. 4-5 above. The test rejects $\mathrm{H}_0 : \mu = 200$ iff $|\overline{X} - 200| > 2.63$. Can we invert this test to obtain a 95% confidence region? Not exactly — we need a testing procedure for testing $\mathrm{H}_0 : \mu = \mu_0$ for an arbitrary μ_0. The testing procedure is to reject $\mathrm{H}_0 : \mu = \mu_0$ iff $|\overline{X} - \mu_0| > 2.63$. For any μ_0, this test has level .05. The rejection region is $\mathcal{R}(\mu_0) = \{\boldsymbol{x} : |\overline{x} - \mu_0| > 2.63\}$. The corresponding confidence region is $\mathcal{C}(\boldsymbol{x}) = \{\mu_0 : \boldsymbol{x} \notin \mathcal{R}(\mu_0)\}$. Now note that $\boldsymbol{x} \notin \mathcal{R}(\mu_0)$ iff $|\overline{x} - \mu_0| \le 2.63$ iff $\overline{x} - 2.63 \le \mu_0 \le \overline{x} + 2.63$. Thus we see that $\mathcal{C}(\boldsymbol{x}) = [\overline{x} - 2.63\,, \overline{x} + 2.63]$. This is a 95% confidence interval for the true μ. ||

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Gamma(δ_0, β) where $\delta_0 > 0$ is known and $\beta > 0$ is an unknown parameter. On p. 13 above we saw that a UMP level α test for testing $\mathrm{H}_0 : \beta = \beta_0$ versus $\mathrm{H}_1 : \beta > \beta_0$ rejects H_0 iff $\sum x_i > c$, where $\mathrm{P}_{\beta_0}\{\sum X_i > c\} = \alpha$. We know that $\sum X_i \sim$ Gamma$(n\delta_0, \beta)$, and so $c = \texttt{gaminv}(1 - \alpha, n\delta_0, \beta_0)$. This is Matlab notation; in S-Plus $c = \beta_0{*}\texttt{qgamma}(1 - \alpha, n\delta_0)$. We can invert this test to obtain a confidence region (hopefully an interval) with confidence coefficient $1 - \alpha$. The confidence region is $\mathcal{C}(\boldsymbol{x}) = \{\beta_0 : \boldsymbol{x} \notin \mathcal{R}(\beta_0)\}$. Note that $\boldsymbol{x} \notin \mathcal{R}(\beta_0)$ iff $\sum x_i \le c = \texttt{gaminv}(1 - \alpha, n\delta_0, \beta_0)$.

Given a numerical value of $\sum x_i$, a straightforward but computationally intensive way to determine the confidence region is to calculate $\texttt{gaminv}(1 - \alpha, n\delta_0, \beta_0)$ for many values of β_0 and notice when it is $\ge \sum x_i$. This gives only an approximation to the confidence region, the goodness of the approximation depending on how many values of β_0 are tried.

A more precise way to determine the confidence region is to use the fact that $(2/\beta)\sum X_i \sim \text{Gamma}(n\delta_0, 2) = \chi^2(2n\delta_0)$. Then $\sum x_i > c$ iff $(2/\beta_0)\sum x_i > k$, where $\mathrm{P}_{\beta_0}\{(2/\beta_0)\sum X_i > k\} = \mathrm{P}\{\chi^2(2n\delta_0) > k\} = \alpha$. So $k = \texttt{chi2inv}(1 - \alpha, 2n\delta_0)$, which does not depend on β_0. Now $\boldsymbol{x} \notin \mathcal{R}(\beta_0)$ iff $(2/\beta_0)\sum x_i \le k$ iff $\beta_0 \ge (2/k)\sum x_i$. Therefore, $(2/k)\sum x_i$ is a lower $(1 - \alpha)$-confidence limit for β. ||

Pivotal approach

Let $\boldsymbol{x}$ be a data vector whose distribution is assumed to have pdf $f(\boldsymbol{x}\,; \theta)$ for some real-valued parameter θ. A pivot can be used to construct a confidence region for θ. A random variable $U = U(\boldsymbol{X}, \theta)$, which is a function of both the random data vector $\boldsymbol{X}$ and the parameter θ, is a *pivot* if its distribution under the distribution with pdf $f(x\,;\theta)$ does not depend on θ. Of course the distribution of X does depend on θ and so the function $U(X,\theta)$ must involve θ in such a way that it cancels out the distributional effect of θ in the distribution of X. This will become clearer in the examples.

The most useful pivots are functions of a minimal sufficient statistic. So when looking for a pivot, a good thing to do first is to find a minimal sufficient statistic, say $T(X)$. Then consider the distribution of $T(X)$ and try to figure out how the dependence of its distribution on θ might be cancelled out by some function $U(T(X),\theta)$.

The most common pivots are of the following three kinds.

(i) U is a *location pivot* if it has the form $U = T(X) - a(\theta)$ for some statistic $T(X)$.

(ii) U is a *scale pivot* if it has the form $U = T(X)/b(\theta)$.

(iii) U is a *location-scale pivot* if it has the form $U = [T(X) - a(\theta)]/b(\theta)$.

Examples. (a) Suppose $X_1,\ldots,X_n$ are i.i.d. Normal(μ,σ_0^2). Then $\overline{X}$ is a minimal sufficient statistic (by Theorem 6.3.3). Its distribution is $\overline{X} \sim$ Normal$(\mu,\sigma_0^2/n)$. Note that $\overline{X} - \mu \sim$ Normal$(0,\sigma_0^2/n)$, which is a distribution not depending on μ. Therefore, $\overline{X} - \mu$ is a location pivot. Also, $\sqrt{n}(\overline{X} - \mu)/\sigma_0$ is a pivot because its distribution is Normal$(0,1)$.

(b) Suppose $X_1,\ldots,X_n$ are i.i.d. Normal(μ_0,σ^2). Then $T = \sum(X_i - \mu_0)^2$ is a minimal sufficient statistic (by Theorem 6.3.3). To describe the distribution of T, we might note that $X_i \sim$ Normal$(\mu_0,\sigma^2) \Rightarrow (X_i - \mu_0)/\sigma \sim$ Normal$(0,1) \Rightarrow (X_i - \mu_0)^2/\sigma^2 \sim \chi^2(1)$ (see Example 4.4.3) $\Rightarrow \sum_{i=1}^{n}(X_i - \mu_0)^2/\sigma^2 = T/\sigma^2 \sim \chi^2(n)$ (see Theorem 4.3.2(iii)). At this point we see that T/σ^2 is a scale pivot.

(c) Suppose $X_1,\ldots,X_n$ are i.i.d. Gamma(δ_0,β). Then $T = \sum X_i$ is a minimal sufficient statistic (by Theorem 6.3.3). We know $T \sim$ Gamma$(n\delta_0,\beta)$ (by Theorem 4.3.2(ii)). Also, $T/\beta \sim$ Gamma$(n\delta_0,1)$ (by using Theorem 4.4.1 or Theorem 4.3.1). Therefore T/β is a scale pivot. ||

Next we show how a pivot can be used to obtain a confidence region for the parameter. Since the distribution of a pivot $U = U(X,\theta)$ does not depend on θ, its quantiles do not depend on θ. Let a be the $\frac{1}{2}\alpha$-quantile of the distribution of U, that is, $\mathrm{P}\{U \le a\} = \frac{1}{2}\alpha$, and let b be the $(1 - \frac{1}{2}\alpha)$-quantile of the distribution of U, that is, $\mathrm{P}\{U \le b\} = 1 - \frac{1}{2}\alpha$. Then $\mathrm{P}\{a < U(X,\theta) \le b\} = 1 - \alpha$ for all θ. Define a confidence region to be $C(x) = \{\theta : a < U(x,\theta) \le b\}$. Its coverage probability is $\mathrm{P}_\theta\{\theta \in C(X)\} = \mathrm{P}_\theta\{a < U(X,\theta) \le b\} = 1 - \alpha$ for all θ.

Although the condition $a < U(\boldsymbol{x},\theta) \le b$ completely defines the confidence region, one should try to manipulate the inequalities to achieve a more direct description of the region. Often this leads to an interval $T_{\mathrm{L}}(\boldsymbol{x}) < \theta \le T_{\mathrm{U}}(\boldsymbol{x})$ (or $T_{\mathrm{L}}(\boldsymbol{x}) \le \theta < T_{\mathrm{U}}(\boldsymbol{x})$). However, it is not always possible to find a pivot that allows easy manipulation of the condition $a < U(\boldsymbol{x},\theta) \le b$.

Examples. (a) Suppose $X_1,\ldots,X_n$ are i.i.d. Normal(μ,σ_0^2). Above we saw that $\overline{X}-\mu$ is a pivot with distribution Normal$(0,\sigma_0^2/n)$. Its $\frac{1}{2}\alpha$-quantile and $(1-\frac{1}{2}\alpha)$-quantile are $a = \texttt{norminv}(\frac{1}{2}\alpha, 0, \sigma_0/\sqrt{n})$ and $b = \texttt{norminv}(1-\frac{1}{2}\alpha, 0, \sigma_0/\sqrt{n})$ respectively. This yields the $(1-\alpha)$-confidence region $\{\mu : a < \overline{X}-\mu < b\}$. Since $a < \overline{X}-\mu < b$ iff $\overline{X}-b < \mu < \overline{X}-a$, the confidence region for μ is the interval $(\overline{X}-b\,,\overline{X}-a)$. Since the Normal$(0,\sigma_0^2/n)$ distribution is symmetric about 0, we see that $a = -b$, so the interval is $(\overline{X}-b\,,\overline{X}+b)$. The quantile b could be obtained from a standard normal table by re-expressing it as $b = (\sigma_0/\sqrt{n})\Phi^{-1}(1-\frac{1}{2}\alpha)$.

(b) Suppose $X_1,\ldots,X_n$ are i.i.d. Normal(μ_0,σ^2). Above we saw that $\sum(X_i-\mu_0)^2/\sigma^2$ is a pivot with distribution $\chi^2(n)$. Its $\frac{1}{2}\alpha$-quantile and $(1-\frac{1}{2}\alpha)$-quantile are $a = \texttt{chi2inv}(\frac{1}{2}\alpha, n)$ and $b = \texttt{chi2inv}(1-\frac{1}{2}\alpha, n)$ respectively. This yields the $(1-\alpha)$-confidence region $\{\sigma^2 : a < \sum(X_i-\mu_0)^2/\sigma^2 < b\}$. Since $a < \sum(X_i-\mu_0)^2/\sigma^2 < b$ iff $\sum(X_i-\mu_0)^2/b < \sigma^2 < \sum(X_i-\mu_0)^2/a$, the confidence region for σ^2 is the interval $(\sum(X_i-\mu_0)^2/b\,,\sum(X_i-\mu_0)^2/a)$.

(c) Suppose $X_1,\ldots,X_n$ are i.i.d. Gamma(δ_0,β). Above we saw that $\sum X_i/\beta$ is a pivot with distribution Gamma$(n\delta_0,1)$. Its $\frac{1}{2}\alpha$-quantile and $(1-\frac{1}{2}\alpha)$-quantile are $a = \texttt{gaminv}(\frac{1}{2}\alpha, n\delta_0, 1)$ and $b = \texttt{gaminv}(1-\frac{1}{2}\alpha, n\delta_0, 1)$ respectively. This yields the $(1-\alpha)$-confidence region $\{\beta : a < \sum X_i/\beta < b\}$. Since $a < \sum X_i/\beta < b$ iff $\sum X_i/b < \beta < \sum X_i/a$, the confidence region for β is the interval $(\sum X_i/b\,,\sum X_i/a)$. The quantiles could be obtained from a chi-squared table by re-expressing them as $a = \frac{1}{2}\texttt{chi2inv}(\frac{1}{2}\alpha, 2n\delta_0)$ and $b = \frac{1}{2}\texttt{chi2inv}(1-\frac{1}{2}\alpha, 2n\delta_0)$, provided that $2n\delta_0$ is an integer. ||

Confidence regions in models with a vector-valued parameter

Suppose the model is parameterized by a parameter vector $\boldsymbol{\theta}$. Suppose we are interested in a particular real-valued parametric function $\tau(\boldsymbol{\theta})$. A confidence region for $\tau(\boldsymbol{\theta})$ can be obtained either by inverting a test or by pivoting.

If we have a test for testing $\mathrm{H}_0 : \tau(\boldsymbol{\theta}) = \tau_0$ for an arbitrary τ_0, then the test can be inverted to obtain a confidence region for $\tau(\boldsymbol{\theta})$. Let $\mathcal{R}(\tau_0)$ be the rejection region of the test. As before, we define the confidence region to be $\mathcal{C}(\boldsymbol{x}) = \{\tau_0 : \boldsymbol{x} \notin \mathcal{R}(\tau_0)\}$.

A *pivot* in this situation is a real-valued random variable $U = U(\boldsymbol{X}, \tau(\boldsymbol{\theta}))$ whose distribution does not depend on $\boldsymbol{\theta}$. Given a pivot, a confidence region can be constructed from the quantiles of U, as before.

Examples. (a) Suppose $X_1, \ldots, X_n$ are i.i.d. Normal(μ, σ^2) with both parameters unknown, and suppose we want to focus our attention on μ. We know that $(\overline{X}, S^2)$ is a minimal sufficient statistic (Example 6.3.4). We would like to find a function $U(\overline{X}, S^2, \mu)$ whose distribution does not depend on μ or σ^2. Such a function was discovered by W.S. Gosset (Student) in 1908. The pivot is $U = (\overline{X} - \mu)/(S/\sqrt{n})$, which has a $t(n-1)$ distribution. Its $\frac{1}{2}\alpha$-quantile and $(1 - \frac{1}{2}\alpha)$-quantile are $a = \mathtt{tinv}(\frac{1}{2}\alpha, n-1)$ and $b = \mathtt{tinv}(1 - \frac{1}{2}\alpha, n-1)$ respectively. Since a t distribution is symmetric about 0, then $a = -b$. This yields the $(1-\alpha)$-confidence region $\{\mu : -b < (\overline{X} - \mu)/(S/\sqrt{n}) < b\}$. Since $-b < (\overline{X} - \mu)/(S/\sqrt{n}) < b$ iff $\overline{X} - bS/\sqrt{n} < \mu < \overline{X} + bS/\sqrt{n}$, the confidence region for μ is the interval $(\overline{X} - bS/\sqrt{n}, \overline{X} + bS/\sqrt{n})$.

(b) Again suppose $X_1, \ldots, X_n$ are i.i.d. Normal(μ, σ^2) with both parameters unknown, but now suppose we want to focus on σ^2. We would like to find a function $U(\overline{X}, S^2, \sigma^2)$ whose distribution does not depend on μ or σ^2. We can use the pivot $U = (n-1)S^2/\sigma^2$, which has a $\chi^2(n-1)$ distribution. Its $\frac{1}{2}\alpha$-quantile and $(1 - \frac{1}{2}\alpha)$-quantile are $a = \mathtt{chi2inv}(\frac{1}{2}\alpha, n-1)$ and $b = \mathtt{chi2inv}(1 - \frac{1}{2}\alpha, n-1)$ respectively. This yields the $(1-\alpha)$-confidence region $\{\sigma^2 : a < (n-1)S^2/\sigma^2 < b\}$. Since $a < (n-1)S^2/\sigma^2 < b$ iff $(n-1)S^2/b < \sigma^2 < (n-1)S^2/a$, the confidence region for σ^2 is the interval $((n-1)S^2/b, (n-1)S^2/a)$. ||

Continue to suppose the parameter is a vector $\boldsymbol{\theta}$, and now suppose we are interested in several real-valued parametric functions $\tau_1(\boldsymbol{\theta}), \ldots, \tau_m(\boldsymbol{\theta})$. Suppose we know how to construct a confidence interval for $\tau_i(\boldsymbol{\theta})$ for $i = 1, \ldots, m$. Let $J_i(\boldsymbol{x})$ be a confidence interval for $\tau_i(\boldsymbol{\theta})$ with confidence coefficient γ_i. Then $\mathrm{P}_{\boldsymbol{\theta}}\{\tau_i(\boldsymbol{\theta}) \in J_i(\boldsymbol{X})\} = \gamma_i$. Each individual statement $\tau_i(\boldsymbol{\theta}) \in J_i(\boldsymbol{X})$ is true with probability γ_i, but what is the probability that all m statements are true? By the Bonferroni Inequality (Theorem 3.9.10),

$$\mathrm{P}_{\boldsymbol{\theta}}\{\tau_i(\boldsymbol{\theta}) \in J_i(\boldsymbol{X}) \text{ for all } i = 1, \ldots, m\} \geq \gamma_1 + \cdots + \gamma_m - (m-1)$$

and so the joint confidence coefficient of these intervals is at least $\gamma_1 + \cdots + \gamma_m - (m-1)$. If we write $\gamma_i = 1 - \alpha_i$, then the joint coefficient becomes $1 - (\alpha_1 + \cdots + \alpha_m)$. To achieve a joint confidence coefficient of at least $1-\alpha$, we could construct intervals J_i with confidence coefficients $1-\alpha/m$.

Example. Suppose $X_1,\ldots,X_n$ are i.i.d. Normal(μ,σ^2) with both parameters unknown. From part (a) of the preceding example, with $\alpha=.025$, we know that a 97.5%-confidence interval for μ is $(\overline{X}-bS/\sqrt{n}\,,\overline{X}+bS/\sqrt{n})$ where $b=\texttt{tinv}(.9875,n-1)$. From part (b) of the preceding example, with $\alpha=.025$, we also know that a 97.5%-confidence interval for σ^2 is $((n-1)S^2/d\,,(n-1)S^2/c)$ where $c=\texttt{chi2inv}(.0125,n-1)$ and $d=\texttt{chi2inv}(.9875,n-1)$. So we have joint confidence of at least 95% that $\overline{X}-bS/\sqrt{n}<\mu<\overline{X}+bS/\sqrt{n}$ and $(n-1)S^2/d<\sigma^2<(n-1)S^2/c$. $\|$

The accuracy of a confidence interval

One way to measure the accuracy of a confidence interval is by its width. Given two 95%-confidence intervals for a parameter θ, we prefer the one with the shorter width. If the intervals have random widths, we might prefer the one with the shortest expected width.

Example. Suppose $X_1,\ldots,X_n$ are i.i.d. Normal(μ,σ_0^2). Above we obtained a $(1-\alpha)$-confidence interval for μ to be $(\overline{X}-b\,,\overline{X}+b)$ where $b=\texttt{norminv}(1-\frac{1}{2}\alpha,0,\sigma_0/\sqrt{n})$. This interval was derived from the fact that $\mathrm{P}_\mu\{-b<\overline{X}-\mu<b\}=1-\alpha$. More generally, if $b_1=\texttt{norminv}(1-\alpha_1,0,\sigma_0/\sqrt{n})$ and $b_2=\texttt{norminv}(1-\alpha_2,0,\sigma_0/\sqrt{n})$ where $\alpha_1+\alpha_2=\alpha$, then $\mathrm{P}_\mu\{-b_1<\overline{X}-\mu<b_2\}=1-\alpha$, which leads to $(\overline{X}-b_2,\overline{X}+b_1)$ as a $(1-\alpha)$-confidence interval for μ. The width of the interval is b_1+b_2. Which choice of α_1,α_2 gives us the shortest width? The answer is $\alpha_1=\alpha_2=\frac{1}{2}\alpha$. This is because the distribution of the pivot $\overline{X}-\mu$, namely Normal$(0,\sigma_0^2/n)$, is symmetric around 0 and is unimodal. See Figure 9.2.7. $\|$

Example. Suppose $X_1,\ldots,X_n$ are i.i.d. Normal(μ,σ^2) with μ and σ^2 unknown. Previously we obtained a $(1-\alpha)$-confidence interval for μ to be $(\overline{X}-bS/\sqrt{n}\,,\overline{X}+bS/\sqrt{n})$ where $b=\texttt{tinv}(1-\frac{1}{2}\alpha,n-1)$. If $b_1=\texttt{tinv}(1-\alpha_1,n-1)$ and $b_2=\texttt{tinv}(1-\alpha_1,n-1)$ where $\alpha_1+\alpha_2=\alpha$, then $(\overline{X}-b_2S/\sqrt{n}\,,\overline{X}+b_1S/\sqrt{n})$ is also a $(1-\alpha)$-confidence interval for μ. Its width is $(b_1+b_2)S/\sqrt{n}$ and its expected width is $(b_1+b_2)\mathrm{E}(S)/\sqrt{n}$. Among all the possible choices of α_1,α_2, the width is smallest when b_1+b_2 is smallest. Since the distribution of the pivot is $t(n-1)$, which is symmetric around 0 and is unimodal, the best choice is $\alpha_1=\alpha_2=\frac{1}{2}\alpha$. $\|$

Confidence intervals in some two-sample problems

Let us review the pivotal approach to constructing a confidence region for a real-valued parametric function $\tau(\boldsymbol{\theta})$ of a parameter vector $\boldsymbol{\theta}$. There are four steps.

(a) Find a minimal sufficient statistic $\boldsymbol{T}$ for $\boldsymbol{\theta}$.

(b) Determine the distribution of $\boldsymbol{T}$, perhaps after taking a one-to-one transformation to obtain a new minimal sufficient statistic whose distribution is easier to describe.

(c) Using the information in (b), find a pivot $U(\boldsymbol{T},\tau)$. The function U should involve θ only through $\tau(\boldsymbol{\theta})$. Preferably, the distribution of U should be one whose quantiles are available (in a formula or a table or a computer package).

(d) Letting a and b be the $\alpha/2$ and $1-\alpha/2$ quantiles of the distribution of U, manipulate the inequalities $a < U(\boldsymbol{T},\tau) < b$ to isolate τ. This produces a $(1-\alpha)$-confidence region (~~[illegible]~~ an interval) for τ.

Example 9.3.1. Suppose $X_{11},\ldots,X_{1n_1}$ are an i.i.d. sample from a Normal(μ_1,σ^2) population, $X_{21},\ldots,X_{2n_2}$ are an i.i.d. sample from a Normal(μ_2,σ^2) population, and the two samples are independent. This model has three unknown parameters μ_1, μ_2 and σ^2. Note that the two populations are assumed to have a common variance. Our goal is to construct a confidence interval for $\mu_1-\mu_2$.

(a) Since the two samples are independent, the joint pdf of the data is the product of the joint pdf's of the two samples:

$$\begin{aligned}
f(\boldsymbol{x};\mu_1,\mu_2,\sigma^2) &= \Big(\tfrac{1}{\sqrt{2\pi\sigma^2}}\Big)^{n_1}\exp\Big[-\tfrac{1}{2\sigma^2}\textstyle\sum(x_{1i}-\mu_1)^2\Big]\\
&\qquad\times\Big(\tfrac{1}{\sqrt{2\pi\sigma^2}}\Big)^{n_2}\exp\Big[-\tfrac{1}{2\sigma^2}\textstyle\sum(x_{2i}-\mu_2)^2\Big]\\
&= \Big(\tfrac{1}{\sqrt{2\pi\sigma^2}}\Big)^{n_1+n_2}\exp\Big[-\tfrac{1}{2\sigma^2}\textstyle\sum(x_{1i}-\mu_1)^2+\sum(x_{2i}-\mu_2)^2\Big]\\
&= \Big(\tfrac{1}{\sqrt{2\pi\sigma^2}}\Big)^{n_1+n_2}\exp\Big[-\tfrac{n_1\mu_1^2}{2\sigma^2}-\tfrac{n_2\mu_2^2}{2\sigma^2}\Big]\\
&\qquad\times\exp\Big[-\tfrac{1}{2\sigma^2}\big(\textstyle\sum x_{1i}^2+\sum x_{2i}^2\big)+\tfrac{\mu_1}{\sigma^2}\sum x_{1i}+\tfrac{\mu_2}{\sigma^2}\sum x_{2i}\Big]\,.
\end{aligned}$$

We see that this is a full-rank exponential family. Theorem 6.3.3 implies that $\boldsymbol{T} = (\sum X_{1i}, \sum X_{2i}, \sum X_{1i}^2 + \sum X_{2i}^2)$ is a minimal sufficient statistic.

(b) The distribution of $\sum X_{1i}^2+\sum X_{2i}^2$ is difficult to describe, so we make a one-to-one transformation from $\boldsymbol{T}$ to $\boldsymbol{W} = (\overline{X}_1, \overline{X}_2, \sum(X_{1i}-\overline{X}_1)^2+\sum(X_{2i}-\overline{X}_2)^2)$. To see how $\boldsymbol{W}$ is a transformation of $\boldsymbol{T}$, recall that $\sum(X_i-\overline{X})^2 = \sum X_i^2 - n\overline{X}^2$ (see formula (4.4.9)).

We can write the last component of W as $W_3 = (n_1-1)S_1^2 + (n_2-1)S_2^2$. We know $(n_1-1)S_1^2/\sigma^2 \sim \chi^2(n_1-1)$ and $(n_2-1)S_2^2/\sigma^2 \sim \chi^2(n_2-1)$ (Theorem 4.4.2(ii)). Since the two samples are independent, Theorem 4.3.2(iii) implies that $W_3/\sigma^2 \sim \chi^2(n_1+n_2-2)$. We also know $\overline{X}_1 \sim \text{Normal}(\mu_1, \sigma^2/n_1)$ and $\overline{X}_2 \sim \text{Normal}(\mu_2, \sigma^2/n_2)$.

(c) We want a pivot for $\mu_1-\mu_2$. We see $\overline{X}_1 - \overline{X}_2 \sim \text{Normal}(\mu_1-\mu_2, \sigma^2/n_1+\sigma^2/n_2)$ and so

$$\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim \text{Normal}(0,1) .$$

This is a pivot, but it is not a pivot for $\mu_1-\mu_2$, because it also involves σ. A pivot for $\mu_1-\mu_2$ should involve the parameters only through $\mu_1-\mu_2$. What we can do is use W_3 to cancel out σ. Recall Definition 4.5.1, which says that if $Z \sim \text{Normal}(0,1)$, $V \sim \chi^2(m)$, and Z and V are independent, then $Z/\sqrt{V/m} \sim t(m)$. By Theorem 4.4.2(i) and the independence of the two samples, W_3 is independent of $\overline{X}_1$ and $\overline{X}_2$. Therefore

$$\left[\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right] \Bigg/ \sqrt{(W_3/\sigma^2)/(n_1+n_2-2)} \sim t(n_1+n_2-2) .$$

Note that σ cancels out. We have

$$\frac{W_3}{n_1+n_2-2} = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} = S_\text{P}^2 ,$$

the pooled sample variance (formula (4.5.7)). The pivot can be written as

$$U = \frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{S_\text{P}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} .$$

(d) Let $b = \texttt{tinv}(1-\frac{1}{2}\alpha, n_1+n_2-2)$. Then $\text{P}\{-b < U < b\} = 1-\alpha$ for all μ_1, μ_2 and σ^2. Manipulate $-b < U < b$, that is,

$$-b < \frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{S_\text{P}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} < b ,$$

to obtain the $(1-\alpha)$-confidence interval

$$\overline{X}_1-\overline{X}_2 - b\,S_\text{P}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} < \mu_1-\mu_2 < \overline{X}_1-\overline{X}_2 + b\,S_\text{P}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} . \quad \|$$

Example. Suppose $X_{11}, \ldots, X_{1n_1}$ are an i.i.d. sample from a Normal(μ_1, σ_1^2) population, $X_{21}, \ldots, X_{2n_2}$ are an i.i.d. sample from a Normal(μ_2, σ_2^2) population, and the two samples are independent. The model has four unknown parameters μ_1, μ_2, σ_1^2 and σ_2^2. In this example we are not assuming that the two populations have a common variance. Our goal again is to construct a confidence interval for $\mu_1 - \mu_2$. We try to follow the same steps as in the preceding example. The steps are summarized below.

(a) A minimal sufficient statistic is $T = (\overline{X}_1, \overline{X}_2, S_1^2, S_2^2)$.

(b) $\overline{X}_1 \sim \text{Normal}(\mu_1, \sigma_1^2/n_1)$, $\overline{X}_2 \sim \text{Normal}(\mu_2, \sigma_2^2/n_2)$,

$(n_1 - 1)S_1^2/\sigma_1^2 \sim \chi^2(n_1 - 1)$, $(n_2 - 1)S_2^2/\sigma_2^2 \sim \chi^2(n_2 - 1)$

(c) $\overline{X}_1 - \overline{X}_2 \sim \text{Normal}(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$ and so

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \text{Normal}(0, 1) .$$

This is a pivot, but it is not a pivot for $\mu_1 - \mu_2$, because it also involves σ_1^2 and σ_2^2. At this point we would like to find a function $g(S_1^2, S_2^2)$ such that

$$g(S_1^2, S_2^2) \Big/ \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \sim \chi^2 .$$

Unfortunately, it seems that no such function exists. This is one way in which the analysis of two normal samples is more difficult without the assumption of common variance. ||

Example 9.3.4. Suppose $X_{11}, \ldots, X_{1n_1}$ are an i.i.d. sample from a Normal(μ_1, σ_1^2) population, $X_{21}, \ldots, X_{2n_2}$ are an i.i.d. sample from a Normal(μ_2, σ_2^2) population, and the two samples are independent. This model has four unknown parameters μ_1, μ_2, σ_1^2 and σ_2^2. In this example we are not assuming that the two populations have a common variance. Our goal is to construct a confidence interval for the variance ratio σ_1^2/σ_2^2. We follow the same four steps as in Example 9.3.1. The steps are summarized below.

(a) A minimal sufficient statistic is $T = (\overline{X}_1, \overline{X}_2, S_1^2, S_2^2)$.

(b) $\overline{X}_1 \sim \text{Normal}(\mu_1, \sigma_1^2/n_1)$, $\overline{X}_2 \sim \text{Normal}(\mu_2, \sigma_2^2/n_2)$,

$(n_1 - 1)S_1^2/\sigma_1^2 \sim \chi^2(n_1 - 1)$, $(n_2 - 1)S_2^2/\sigma_2^2 \sim \chi^2(n_2 - 1)$

The four components of T are independent of one another.

(c) Recall Definition 4.5.2, which says that if $V \sim \chi^2(m)$, $W \sim \chi^2(p)$, and V and W are independent, then $(V/m)/(W/p) \sim F(m, p)$. Therefore $U = (S_1^2/\sigma_1^2)/(S_2^2/\sigma_2^2) = (S_1^2/S_2^2)/(\sigma_1^2/\sigma_2^2) \sim F(n_1 - 1, n_2 - 1)$, so U is a pivot for σ_1^2/σ_2^2.

(d) Let $a = \texttt{finv}(\frac{1}{2}\alpha , n_1 - 1 , n_2 - 1)$ and $b = \texttt{finv}(1 - \frac{1}{2}\alpha , n_1 - 1 , n_2 - 1)$. Then $\mathrm{P}\{a < U < b\} = 1 - \alpha$ for all μ_1 , μ_2 , σ_1^2 and σ_2^2 . manipulate $a < U < b$ to obtain the $(1 - \alpha)$-confidence interval

$$\frac{S_1^2}{bS_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{aS_2^2} \; . \qquad \|$$

Example 9.3.6. Suppose $X_{11}, \ldots , X_{1n_1}$ are an i.i.d. sample from a Uniform$(0, \theta_1)$ population, $X_{21}, \ldots , X_{2n_2}$ are an i.i.d. sample from a Uniform$(0, \theta_2)$ population, and the two samples are independent. This model has two unknown parameters θ_1 and θ_2 . Our goal is to construct a confidence interval for the ratio θ_1/θ_2 .

(a) A minimal sufficient statistic is $T = (T_1 , T_2)$ where $T_1 = \max\{X_{11}, \ldots , X_{1n_1}\}$ and $T_2 = \max\{X_{21}, \ldots , X_{2n_2}\}$.

(b) From section 4.2.3 we know that the cdf of T_1 is $F_1(t) = (t/\theta_1)^{n_1}$ and the cdf of T_2 is $F_2(t) = (t/\theta_2)^{n_2}$. Another way to describe the distributions of T_1 and T_2 is to say that T_1/θ_1 has cdf $\overline{F_1}(u) = u^{n_1}$ and T_2/θ_2 has cdf $\overline{F_2}(u) = u^{n_2}$.

(c) We see that $U = (T_1/\theta_1)/(T_2/\theta_2) = (T_1/T_2)/(\theta_1/\theta_2)$ is a pivot for θ_1/θ_2 . However, it is not obvious how to determine the quantiles of the distribution of U . From the boxed statement on p. 182 of the textbook, we know that $(T_1/\theta_1)^{n_1} \sim$ Uniform$(0, 1)$ and $-\log[(T_1/\theta_1)^{n_1}] = -n_1\log(T_1/\theta_1) \sim$ Exponential(1) . Similarly, $-n_2\log(T_2/\theta_2) \sim$ Exponential(1) . In order to be able to obtain a nice distribution for the pivot, let us now suppose the sample sizes are equal, $n_1 = n_2 = n$. Let $W = -n \log U = -n\log(T_1/\theta_1) + n\log(T_2/\theta_2)$. The distribution of W is that of the difference of two independent Exponential(1) random variables. This distribution can be derived by the method in section 4.4.1. Its pdf is $f(w) = \frac{1}{2}e^{-|w|}$. The $(1 - \frac{1}{2}\alpha)$-quantile of W is q satisfying $\int_q^\infty f(w)dw = \frac{1}{2}\alpha$, which can be solved to obtain $q = -\log \alpha$.

(d) $\mathrm{P}\{\log \alpha < W < -\log \alpha\} = 1 - \alpha$ for all θ_1 and θ_2 . (Note that $\log \alpha < 0$ because $0 < \alpha < 1$.) Manipulate $\log \alpha < W < -\log \alpha$, that is,

$$\log \alpha < -n \log\left(\frac{T_1/T_2}{\theta_1/\theta_2}\right) < -\log \alpha \; .$$

We obtain

$$\log \alpha < \log\left(\frac{T_1/T_2}{\theta_1/\theta_2}\right)^{-n} < \log \alpha^{-1} \; ,$$

or $$\alpha < \left(\frac{T_1/T_2}{\theta_1/\theta_2}\right)^{-n} < \alpha^{-1} \; ,$$

or $$\alpha^{\frac{1}{n}}\left(\frac{T_1}{T_2}\right) < \frac{\theta_1}{\theta_2} < \alpha^{-\frac{1}{n}}\left(\frac{T_1}{T_2}\right) .$$

This is a $(1-\alpha)$-confidence interval. ||

Multiple comparisons using the Bonferroni Inequality

Consider an experiment to compare four formulas for making cement. Using each formula, n specimens were made and after a month their compressive strengths were measured. Denote the strength measurements by $X_{11}, \ldots, X_{1n}, X_{21}, \ldots, X_{2n}, X_{31}, \ldots, X_{3n}, X_{41}, \ldots, X_{4n}$. We will assume that these are four independent samples from normal populations having means that are possibly different but with a common variance. That is, the X_{ij}'s are all independent with $X_{ij} \sim \mathrm{N}(\mu_i , \sigma^2)$.

For comparing cement formulas 1 and 2, we can construct a confidence interval for $\mu_1 - \mu_2$ as in Example 9.3.1. Using only the samples for these two cement formulas, we can construct a $(1-\alpha)$-confidence interval for $\mu_1 - \mu_2$ to be

$$\text{from} \quad (\overline{X}_1 - \overline{X}_2) - t_{(2n-2),\alpha/2} S_{\mathrm{P}}\sqrt{\frac{2}{n}} \quad \text{to} \quad (\overline{X}_1 - \overline{X}_2) + t_{(2n-2),\alpha/2} S_{\mathrm{P}}\sqrt{\frac{2}{n}}$$

where $S_{\mathrm{P}}^2 = (S_1^2 + S_2^2)/2$. Another interval, which is likely to be narrower, can be obtained by using all four samples to estimate σ^2. That is, let $S^2 = (S_1^2 + S_2^2 + S_3^2 + S_4^2)/4$ and construct the interval

$$\text{from} \quad (\overline{X}_1 - \overline{X}_2) - t_{(4n-4),\alpha/2} S\sqrt{\frac{2}{n}} \quad \text{to} \quad (\overline{X}_1 - \overline{X}_2) + t_{(4n-4),\alpha/2} S\sqrt{\frac{2}{n}} .$$

We have confidence $1-\alpha$ in this interval if we are focusing on the difference $\mu_1 - \mu_2$ alone. But often we would be interested in the differences between all 6 pairs of means: $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, $\mu_1 - \mu_4$, $\mu_2 - \mu_3$, $\mu_2 - \mu_4$, $\mu_3 - \mu_4$. Suppose we want to construct confidence intervals for the differences between all pairs of means and that we want our joint confidence to be at least $1-\alpha$. One way to do this is to use the Bonferroni Inequality as on pp. 26-27 above. For each of the 6 pairs i, k of distinct indices $1, 2, 3, 4$, construct the interval

$$\text{from} \quad (\overline{X}_i - \overline{X}_k) - t_{(4n-4),\alpha/12} S\sqrt{\frac{2}{n}} \quad \text{to} \quad (\overline{X}_i - \overline{X}_k) + t_{(4n-4),\alpha/12} S\sqrt{\frac{2}{n}} .$$

Note that the t quantile is for $\alpha/12$ (not $\alpha/2$). Therefore each individual interval has confidence coefficient $1-\alpha/6$, and hence our joint confidence in the 6 intervals is, according to Bonferroni's Inequality, $\geq 1-\alpha$.

In some situations, we are not interested in all the pairwise comparisons between treatments but only the comparisons involving a standard treatment. Suppose cement formula 1 is the one currently being used by a cement manufacturer and that formulas 2, 3, and 4 are new formulas that this experiment is designed to test. Then the only differences of interest are the 3 differences $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, and $\mu_1 - \mu_4$. For $i = 2, 3, 4$, construct the interval

$$\text{from} \quad (\overline{X}_1 - \overline{X}_i) - t_{(4n-4),\alpha/6} S\sqrt{\frac{2}{n}} \quad \text{to} \quad (\overline{X}_1 - \overline{X}_i) + t_{(4n-4),\alpha/6} S\sqrt{\frac{2}{n}} \,.$$

Note that each individual interval has confidence coefficient $1 - \alpha/3$, and hence our joint confidence in the 3 intervals is, according to Bonferroni's Inequality, $\geq 1 - \alpha$. By focusing on these 3 differences, we obtain slightly narrower intervals than in the preceding paragraph where all 6 differences were of interest.

Multiple comparisons using the multivariate t distribution

Section 9.4.2 presents another way to obtain joint confidence intervals for $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, and $\mu_1 - \mu_4$. We will again use the pivotal approach. A good starting point is a minimal sufficient statistic. A minimal sufficient statistic for this model is $(\overline{X}_1, \overline{X}_2, \overline{X}_3, \overline{X}_4, S^2)$ where $S^2 = (S_1^2 + S_2^2 + S_3^2 + S_4^2)/4$.

We want to find a pivot for $\boldsymbol{\delta} = (\delta_1, \delta_2, \delta_3)' = (\mu_1 - \mu_2, \mu_1 - \mu_3, \mu_1 - \mu_4)'$. Of course the differences of the corresponding sample means must be relevant, so let $\boldsymbol{Y} = (Y_1, Y_2, Y_3)'$ $= (\overline{X}_1 - \overline{X}_2, \overline{X}_1 - \overline{X}_3, \overline{X}_1 - \overline{X}_4)'$. The random vector $\boldsymbol{Y}$ has a multivariate normal distribution $\mathrm{N}_3(\boldsymbol{\delta}, \sigma^2 H)$ where the entries of the matrix $\sigma^2 H$ are the covariances $\mathrm{Cov}(Y_i, Y_k) = \sigma^2 h_{ik}$. For $i = k$, this says that $\mathrm{Var}(Y_i) = \sigma^2 h_{ii}$. Check that $h_{ii} = 2/n$ and $h_{ik} = 1/n$ for $i \neq k$. We see that $(\boldsymbol{Y} - \boldsymbol{\delta})/\sigma \sim \mathrm{N}_3(\boldsymbol{0}, H)$, which is a distribution not depending on the parameters. So $(\boldsymbol{Y} - \boldsymbol{\delta})/\sigma$ is a pivot (a vector-valued pivot rather than a real-valued pivot). But it is not a pivot for $\boldsymbol{\delta}$ alone because it involves σ.

To get rid of σ, we can try substituting the estimate S. Let $\boldsymbol{U} = (\boldsymbol{Y} - \boldsymbol{\delta})/S$. This is a pivot because it can be written as $\boldsymbol{U} = [(\boldsymbol{Y} - \boldsymbol{\delta})/\sigma]/[S/\sigma]$ in which (i) the numerator $(\boldsymbol{Y} - \boldsymbol{\delta})/\sigma$ has a distribution, $\mathrm{N}_3(\boldsymbol{0}, H)$, that does not depend on the parameters, (ii) the denominator S/σ has a distribution, $\sqrt{\chi^2_{(4n-4)}/(4n-4)}$, that does not depend on the parameters, and (iii) the numerator and denominator are independent. Fact (iii) follows from Theorem 4.4.2(i) and the independence of the four samples.

Note that facts (i) and (ii) are not sufficient to establish that $\boldsymbol{U}$ is a pivot. In general, if X and Y are two random variables whose distributions do not depend on the parameters, it is still possible for the distribution of a function $g(X, Y)$ to depend on the parameters. This is because the distribution of $g(X, Y)$ depends on the joint distribution of (X, Y) and not just on their marginal distributions. For example, suppose (X, Y) has a bivariate normal distribution $\mathrm{N}_2(0, 0, 1, 1, \rho)$ as in section 3.6. Then $X \sim \mathrm{N}(0, 1)$ and $Y \sim \mathrm{N}(0, 1)$, so their distributions do not depend on the parameter ρ. Consider $g(X,Y) = XY$. Its distribution depends on the parameter because $\mathrm{E}(XY) = \mathrm{Cov}(X,Y) = \rho$.

Now we want to use U to construct a confidence region for δ. We can express its distribution in terms of a multivariate t distribution, introduced in section 4.6.2. We have $\sqrt{n/2}(Y - \delta)/\sigma \sim \mathrm{N}_3(0\,, G)$ where $G = (n/2)H$ is a correlation matrix, that is, having all its diagonal entries 1. All the off-diagonal entries of G are $1/2$. According to section 4.6.2, facts (i), (ii) and (iii) imply that $\sqrt{n/2}\,U$ has a multivariate t distribution $Mt_3(4n-4, G)$. For some values of p, m, G and α, tables are available that give critical values c such that, if $T = (T_1, \ldots, T_p) \sim Mt_p(m, G)$, then $\mathrm{P}\{|T_i| < c \text{ for all } i = 1, \ldots, p\} = 1 - \alpha$. Let c be the critical value in our case, that is, for $p = 3$, $m = 4n - 4$, and G with all off-diagonal entries $1/2$. Then $\mathrm{P}\{|\sqrt{n/2}\,[(\overline{X}_1 - \overline{X}_i) - (\mu_1 - \mu_i)]/S\,| < c \text{ for } i = 1\,,2\,,3\} = 1 - \alpha$. This defines 3 intervals, the one for $\mu_1 - \mu_i$ having limits $(\overline{X}_1 - \overline{X}_i) \pm cS\sqrt{2/n}$.

Numerical example. Suppose the four samples are of size $n = 3$ with sample means and standard deviations as follows:

$\overline{X}_1 = 5635$	$\overline{X}_2 = 5753$	$\overline{X}_3 = 4527$	$\overline{X}_4 = 3442$
$S_1 = 966$	$S_2 = 432$	$S_3 = 510$	$S_4 = 356$

Pooling the standard deviations from all four samples, we obtain $S = \sqrt{(966^2 + 432^2 + 510^2 + 356^2)/4} = 614$.

(A) Suppose we focus on the difference $\mu_1 - \mu_2$ and use only the 6 observations from samples 1 and 2. A 95%-confidence interval for $\mu_1 - \mu_2$ has limits $(5635 - 5753) \pm (2.776)(748)\sqrt{2/3}$, that is, $(\,-1813\,, 1577)$. See p. 32 above. Its width is 3390.

(B) Suppose we continue to focus on the difference $\mu_1 - \mu_2$ but now we use all 12 observations to estimate σ. A 95%-confidence interval for $\mu_1 - \mu_2$ has limits $(5635 - 5753) \pm (2.306)(614)\sqrt{2/3}$, that is, $(\,-1274\,, 1038)$. See p. 32 above. Its width is 2312. This interval is narrower for two reasons. First, the t critical value is smaller, which is due to the increase in degrees of freedom for estimating the error variance. Second, the estimate of σ is smaller, which is just by chance. The estimate of σ based on 8 observations is better than the one based on 4 observations, but it could just as easily have been larger rather than smaller.

(C) Now suppose we want intervals for all 6 pairwise differences and we want our joint confidence in them to be at least 95%. Using the Bonferroni Inequality as on p. 32 above, the interval for $\mu_1 - \mu_2$ has limits $(5635 - 5753) \pm (3.479)(614)\sqrt{2/3}$, that is, $(- 1862 , 1626)$. Its width is 3488. It makes sense that this interval is wider than the one in (B) where our confidence is focused on one interval rather encompassing 6 intervals.

(D) Suppose we want intervals only for the 3 pairwise differences $\mu_1 - \mu_i$ for $i = 2 , 3 , 4$, and suppose we want our joint confidence in them to be at least 95%. Using the Bonferroni Inequality as on p. 33 above, the interval for $\mu_1 - \mu_2$ has limits $(5635 - 5753) \pm (3.016)(614)\sqrt{2/3}$, that is, $(- 1630 , 1394)$. Its width is 3024. It makes sense that this interval is narrower than the one in (C) where our confidence encompasses 6 intervals rather than 3 intervals.

(E) Now let us use the method of section 9.4.2. The multivariate t distribution that is relevant here is $Mt_3(8, G)$. The critical value for $\alpha = .05$ can be found in a table referenced in section 9.4.2; it is $c = 2.88$. The interval for $\mu_1 - \mu_2$ has limits $(5635 - 5753) \pm (2.88)(614)\sqrt{2/3}$, that is, $(- 1562 , 1326)$. Its width is 2888, which is narrower than in (D). So this method is an improvement over the more general Bonferroni method, but it requires special tables. $\|$

Notes on Likelihood Ratio Tests

(see Ch. 11 in Mukhopadhyay)

Introduction

Let $\boldsymbol{x}$ be a vector of observed data. As a probability model for the data, suppose we have assumed a family of pdf's or pmf's $f(\boldsymbol{x}\,;\boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$. In section 8.3 we looked at the problem of testing a simple null hypothesis versus a simple alternative hypothesis, that is, $\mathrm{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $\mathrm{H}_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$. Intuitively, the likelihood ratio

$$\mathrm{LR} = \frac{f(\boldsymbol{x}\,;\boldsymbol{\theta}_1)}{f(\boldsymbol{x}\,;\boldsymbol{\theta}_0)}$$

is a sensible test statistic. From the idea that $f(\boldsymbol{x}\,;\boldsymbol{\theta})$ represents the likelihood of the parameter $\boldsymbol{\theta}$ if the data vector $\boldsymbol{x}$ is observed, it is natural to construct a test that rejects H_0 iff $\mathrm{LR} > k$, where k is chosen to achieve the desired probability of Type I error. Our intuition is confirmed by the Neyman-Pearson Lemma, which says that this test is the most powerful one among all tests having the same level.

More generally, suppose we want to test $\mathrm{H}_0 : \boldsymbol{\theta} \in \Theta_0$ versus $\mathrm{H}_1 : \boldsymbol{\theta} \in \Theta_1$. The idea of the likelihood ratio can be generalized by considering a "generalized likelihood ratio"

$$\text{“GLR”} = \frac{\sup\{f(\boldsymbol{x}\,;\boldsymbol{\theta}) : \boldsymbol{\theta}\in\Theta_1\}}{\sup\{f(\boldsymbol{x}\,;\boldsymbol{\theta}) : \boldsymbol{\theta}\in\Theta_0\}} \;.$$

(We have put "GLR" in quotation marks because shortly we will be modifying its definition.) A natural test is constructed by rejecting H_0 iff "GLR" $> k$ for a suitable k. It is a little more convenient to change the definition of the generalized likelihood ratio (GLR), also called the *likelihood ratio test* (LRT) statistic or simply the *likelihood ratio* (LR), to be

$$\mathrm{GLR} = \Lambda = \frac{\sup\{f(\boldsymbol{x}\,;\boldsymbol{\theta}) : \boldsymbol{\theta}\in\Theta_0\}}{\sup\{f(\boldsymbol{x}\,;\boldsymbol{\theta}) : \boldsymbol{\theta}\in\Theta\}} \;.$$

The LR test rejects H_0 iff $\Lambda < c$ for a suitable c. This is essentially equivalent to "GLR" $> k$. To see this, note that the level α is almost always chosen to be small, which means that we reject H_0 only when there is strong evidence against it. This means that we will have $k > 1$, in which case $\sup\{f(\boldsymbol{x}\,;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_1\} = \sup\{f(\boldsymbol{x}\,;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, and hence GLR $= 1/$"GLR".

One-sample problems

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Normal(μ, σ_0^2) where σ_0^2 is known. Consider the problem of testing $\mathrm{H}_0 : \mu = \mu_0$ versus $\mathrm{H}_1 : \mu \neq \mu_0$. The LRT statistic is

$$\Lambda = \frac{\sup\{f(\boldsymbol{x}\,;\mu) : \mu{=}\mu_0\}}{\sup\{f(\boldsymbol{x}\,;\mu) : -\infty{<}\mu{<}\infty\}} = \frac{f(\boldsymbol{x}\,;\mu_0)}{f(\boldsymbol{x}\,;\widehat{\mu})}$$

where $\widehat{\mu}$ is the MLE of μ. We know that $\widehat{\mu} = \overline{x}$ (Example 7.2.6). As in the example on p. 16 above,

$$\Lambda = \frac{\left(1/\sqrt{2\pi\sigma_0^2}\right)^n \exp\left[-\frac{1}{2\sigma_0^2}\sum(x_i-\mu_0)^2\right]}{\left(1/\sqrt{2\pi\sigma_0^2}\right)^n \exp\left[-\frac{1}{2\sigma_0^2}\sum(x_i-\widehat{\mu})^2\right]}$$
$$= \exp\left[-\frac{1}{2\sigma_0^2}\left\{\sum(x_i-\mu_0)^2 - \sum(x_i-\overline{x})^2\right\}\right] .$$

We reject H$_0$ iff $\Lambda < c$ iff $\sum(x_i-\mu_0)^2 - \sum(x_i-\overline{x})^2 > c'$. This latter statistic can be re-expressed in a nicer form by noting that

$$(x_i-\mu_0)^2 = [(x_i-\overline{x}) + (\overline{x}-\mu_0)]^2 = (x_i-\overline{x})^2 + 2(x_i-\overline{x})(\overline{x}-\mu_0) + (\overline{x}-\mu_0)^2$$

and hence

$$\sum(x_i-\mu_0)^2 = \sum(x_i-\overline{x})^2 + 0 + n(\overline{x}-\mu_0)^2 .$$

The 0 term comes from the fact that $\sum(x_i-\overline{x}) = 0$. So the test rejects H$_0$ iff $n(\overline{x}-\mu_0)^2 > c'$ iff $|\overline{x}-\mu_0| > \sqrt{c'/n} = c''$. We prefer to standardize the distribution of the test statistic. Since $\overline{X} \sim \text{Normal}(\mu, \sigma_0^2/n)$, we have $(\overline{X}-\mu)/(\sigma_0/\sqrt{n}) \sim \text{Normal}(0,1)$. Under H$_0$, $(\overline{X}-\mu_0)/(\sigma_0/\sqrt{n}) \sim \text{Normal}(0,1)$. Thus we obtain what seems to be the most implementable and interpretable form of the LR test for this testing problem:

$$\text{reject H}_0 \quad \text{iff} \quad \left|\frac{\overline{x}-\mu_0}{\sigma_0/\sqrt{n}}\right| > z_{\alpha/2} . \quad \|$$

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Normal(μ, σ^2) where μ and σ^2 are both unknown parameters. Let us test H$_0$: $\mu = \mu_0$ versus H$_1$: $\mu \neq \mu_0$.

- The LRT statistic is

$$\Lambda = \frac{\sup\{f(\boldsymbol{x}\,;\mu,\sigma^2) : \mu{=}\mu_0\,, \sigma^2{>}0\}}{\sup\{f(\boldsymbol{x}\,;\mu,\sigma^2) : -\infty{<}\mu{<}\infty\,, \sigma^2{>}0\}} = \frac{f(\boldsymbol{x}\,;\mu_0,\widehat{\sigma}_0^2)}{f(\boldsymbol{x}\,;\widehat{\mu},\widehat{\sigma}^2)}$$

where $(\widehat{\mu}, \widehat{\sigma}^2) = (\overline{x}, \frac{1}{n}\sum(x_i-\overline{x})^2)$ is the MLE of (μ, σ^2) (Example 7.2.7) in the full model, and $\widehat{\sigma}_0^2$ is the MLE of σ^2 in the hypothesized submodel in which $\mu = \mu_0$.

- To obtain $\widehat{\sigma}_0^2$, we want to maximize

$$f(\boldsymbol{x}\,;\, \mu_0, \sigma^2) = \left(1/\sqrt{2\pi\sigma^2}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum(x_i-\mu_0)^2\right] .$$

or, equivalently,

$$\log f(\boldsymbol{x}\,;\, \mu_0, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum(x_i-\mu_0)^2 .$$

Its derivative is

$$\begin{aligned}\frac{\partial}{\partial(\sigma^2)}\log f(\boldsymbol{x}\,;\,\mu_0,\sigma^2) &= -\frac{n}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}\sum(x_i-\mu_0)^2\\ &= \frac{n}{2(\sigma^2)^2}\Big[-\sigma^2+\frac{1}{n}\sum(x_i-\mu_0)^2\Big]\,.\\ &= \frac{n}{2(\sigma^2)^2}\left(\widehat{\sigma}_0^2-\sigma^2\right)\end{aligned}$$

where $\widehat{\sigma}_0^2=\frac{1}{n}\sum(x_i-\mu_0)^2$. The derivative is positive for $\sigma^2<\widehat{\sigma}_0^2$, is 0 for $\sigma^2=\widehat{\sigma}_0^2$, and is negative for $\sigma^2>\widehat{\sigma}_0^2$. This implies that a global maximum of $\log f(\boldsymbol{x}\,;\,\mu_0,\sigma^2)$ is attained at $\sigma^2=\widehat{\sigma}_0^2$.

- Now

$$\Lambda = \frac{f(\boldsymbol{x}\,;\mu_0,\widehat{\sigma}_0^2)}{f(\boldsymbol{x}\,;\widehat{\mu},\widehat{\sigma}^2)} = \frac{\left(1/\sqrt{2\pi\widehat{\sigma}_0^2}\right)^n\exp\left[-\frac{1}{2\widehat{\sigma}_0^2}\sum(x_i-\mu_0)^2\right]}{\left(1/\sqrt{2\pi\widehat{\sigma}^2}\right)^n\exp\left[-\frac{1}{2\widehat{\sigma}^2}\sum(x_i-\overline{x})^2\right]}\,.$$

Note that $-\frac{1}{2\widehat{\sigma}_0^2}\sum(x_i-\mu_0)^2=-\frac{n}{2}$ and $-\frac{1}{2\widehat{\sigma}^2}\sum(x_i-\overline{x})^2=-\frac{n}{2}$. Therefore

$$\Lambda = \left(\frac{\widehat{\sigma}^2}{\widehat{\sigma}_0^2}\right)^{\frac{n}{2}}\,.$$

- The LR test rejects H_0 iff $\Lambda<c$, that is, iff $\frac{\widehat{\sigma}^2}{\widehat{\sigma}_0^2}=\frac{\sum(x_i-\overline{x})^2}{\sum(x_i-\mu_0)^2}<c'$. This test statistic can be interpreted as follows. The x_i's are centered around $\overline{x}$, which tends to be near μ. If $\mu\neq\mu_0$, then the x_i's will tend not to be centered around μ_0, and so the squared deviations $(x_i-\mu_0)^2$ will tend to be larger than the squared deviations $(x_i-\overline{x})^2$, which makes the test statistic small. If the test statistic is significantly small, this supports the conclusion that $\mu\neq\mu_0$. But the distribution of the test statistic $\sum(X_i-\overline{X})^2\Big/\sum(X_i-\mu_0)^2$ is not well known.

- We prefer a test statistic whose the distribution is "standard". Recall

$$\sum(x_i-\mu_0)^2 = \sum(x_i-\overline{x})^2+n(\overline{x}-\mu_0)^2\,.$$

So

$$\frac{\sum(x_i-\overline{x})^2}{\sum(x_i-\mu_0)^2} = \frac{\sum(x_i-\overline{x})^2}{\sum(x_i-\overline{x})^2+n(\overline{x}-\mu_0)^2} = \frac{1}{1+\frac{n(\overline{x}-\mu_0)^2}{\sum(x_i-\overline{x})^2}}\,.$$

Therefore an equivalent expression for the rejection region is

$$\frac{n(\overline{x}-\mu_0)^2}{\sum(x_i-\overline{x})^2} > c''$$

or

$$\frac{n(\overline{x}-\mu_0)^2}{\frac{1}{n-1}\sum(x_i-\overline{x})^2} \;=\; \frac{(\overline{x}-\mu_0)^2}{s^2/n} \;>\; c'''$$

or

$$\left|\frac{\overline{x}-\mu_0}{s/\sqrt{n}}\right| > b$$

This is the usual t test. For a level α test we take $b = \texttt{tinv}(1-\frac{\alpha}{2}, n-1)$. ||

In general, suppose we have a data vector $\boldsymbol{x}$ which is assumed to have been randomly generated from a distribution with pdf or pmf $f(\boldsymbol{x}\,;\boldsymbol{\theta})$ parameterized by an unknown parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$. Suppose we want to test $\mathrm{H}_0 : \boldsymbol{\theta} \in \Theta_0$ versus $\mathrm{H}_1 : \boldsymbol{\theta} \in \Theta_1$. The steps to perform an LR test are the following.

(a) Find the MLE $\widehat{\boldsymbol{\theta}}$ in the full model parameterized by $\boldsymbol{\theta} \in \Theta$.

(b) Find the MLE $\widehat{\boldsymbol{\theta}}_0$ in the hypothesized submodel parameterized by $\boldsymbol{\theta} \in \Theta_0$.

(c) Form the LRT statistic $\Lambda = f(\boldsymbol{x}\,;\widehat{\boldsymbol{\theta}}_0)/f(\boldsymbol{x}\,;\widehat{\boldsymbol{\theta}})$.

(d) Manipulate the inequality $\Lambda < c$ defining the rejection region to obtain an equivalent expression in terms of a statistic W (perhaps (i) $W > k$ or (ii) $W < k$ or (iii) $W < k_1$ or $W > k_2$ or (iv) $k_1 < W < k_2$) where, as far as possible, W is interpretable and either its distribution under H_0 is "standard" or its quantiles are relatively easy to calculate.

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Normal(μ, σ^2) where μ and σ^2 are both unknown parameters. Let us test $\mathrm{H}_0 : \sigma^2 = \sigma_0^2$ versus $\mathrm{H}_1 : \sigma^2 \neq \sigma_0^2$.

(a) We know that the MLEs in the full model are $\widehat{\mu} = \overline{x}$ and $\widehat{\sigma}^2 = \frac{1}{n}\sum(x_i - \overline{x})^2$ (Example 7.2.7).

(b) In the hypothesized submodel, $\sigma^2 = \sigma_0^2$ and the MLE of μ is $\widehat{\mu}_0$ that maximizes

$$f(\boldsymbol{x}\,;\,\mu, \sigma_0^2) \;=\; \left(1/\sqrt{2\pi\sigma_0^2}\right)^n \exp\left[-\frac{1}{2\sigma_0^2}\sum(x_i-\mu)^2\right] . \quad \text{(Example 7.2.6)}$$

or, equivalently,

$$\log f(\boldsymbol{x}\,;\,\mu, \sigma_0^2) \;=\; -\frac{n}{2}\log 2\pi\sigma_0^2 - \frac{1}{2\sigma_0^2}\sum(x_i-\mu)^2 .$$

Its derivative is

$$\begin{aligned}\frac{\partial}{\partial\mu}\log f(\boldsymbol{x}\,;\,\mu, \sigma_0^2) \;&=\; -\frac{1}{2\sigma_0^2}\sum 2(x_i-\mu)(-1) \\ &=\; \frac{n}{\sigma_0^2}(\overline{x}-\mu)\end{aligned}$$

The derivative is positive for $\mu < \overline{x}$, is 0 for $\mu = \overline{x}$, and is negative for $\mu > \overline{x}$. This implies that a global maximum of $\log f(\boldsymbol{x}\,;\ \mu, \sigma_0^2)$ is attained at $\mu = \overline{x}$. Therefore, $\widehat{\mu}_0 = \overline{x}$.

(c) $$\Lambda = \frac{f(\boldsymbol{x}\,;\widehat{\mu}_0, \sigma_0^2)}{f(\boldsymbol{x}\,;\widehat{\mu}, \widehat{\sigma}^2)} = \frac{\left(1/\sqrt{2\pi\sigma_0^2}\right)^n \exp\left[-\frac{1}{2\sigma_0^2}\sum(x_i - \widehat{\mu}_0)^2\right]}{\left(1/\sqrt{2\pi\widehat{\sigma}^2}\right)^n \exp\left[-\frac{1}{2\widehat{\sigma}^2}\sum(x_i - \widehat{\mu})^2\right]}\,.$$

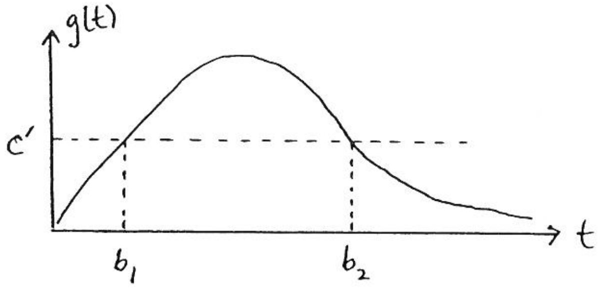
Since $\widehat{\mu} = \widehat{\mu}_0 = \overline{x}$ and $\sum(x_i - \overline{x})^2 = n\,\widehat{\sigma}^2$,

$$\Lambda = \left(\frac{\widehat{\sigma}^2}{\sigma_0^2}\right)^{\frac{n}{2}} \exp\left[\left(-\frac{1}{2\sigma_0^2} + \frac{1}{2\widehat{\sigma}^2}\right) n\,\widehat{\sigma}^2\right]$$

$$= \left(\frac{\widehat{\sigma}^2}{\sigma_0^2}\right)^{\frac{n}{2}} \exp\left[\ \# \frac{n}{2}\left(1 - \frac{\widehat{\sigma}^2}{\sigma_0^2}\right)\right].$$

(d) Note that if $0 < a < b$ and $p > 0$, then $a^p < b^p$. Therefore, $\Lambda < c$ iff $\Lambda^{\frac{2}{n}} < c'$ iff $g(T) < c'$ where $T = \widehat{\sigma}^2/\sigma_0^2$ and $g(t) = te^{1-t}$.

- Let us investigate the behavior of $g(t)$ for $t > 0$. Its derivative is $g'(t) = e^{1-t} - te^{1-t}$ $= e^{1-t}(1-t)$, which is positive for $t < 1$, is 0 for $t = 1$, and is negatrive for $t > 1$. Thus, as t increases from 0 to 1, $g(t)$ increases from $g(0) = 0$ to $g(1) = 1$, and as t increases from 1 to ∞, $g(t)$ decreases from $g(1) = 1$ to $g(\infty) = 0$. Drawing a picture of the graph of this function, we see that $g(T) < c'$ iff $T < b_1$ or $T > b_2$ for b_1 and b_2 satisfying $g(b_1) = g(b_2)$. So we can use $T = \widehat{\sigma}^2/\sigma_0^2 = \frac{1}{n}\sum(x_i - \overline{x})^2/\sigma_0^2$ as our test statistic.



- To get a "standard" distribution, recall that $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$. We have $(n-1)S^2 = \sum(X_i - \overline{X})^2$, so $W = nT = \sum(X_i - \overline{X})^2/\sigma_0^2$ can be used for the test statistic. Its distribution under H_0 is $\chi^2(n-1)$. The LR test rejects H_0 iff $W < nb_1$ or $W > nb_2$.

- Rather than worry about the condition $g(b_1) = g(b_2)$, Mukhopadhyay drops this condition and instead takes the rejection region to be defined by $W < k_1$ or $W > k_2$ where $k_1 =$ `chi2inv`$(\frac{1}{2}\alpha, n-1)$ and $k_2 =$ `chi2inv`$(1 - \frac{1}{2}\alpha, n-1)$. We can call this the *equal-tailed modification of the LR test.* ||

Two-sample problems

Example. Suppose $X_{11}, \ldots, X_{1n_1}$ are an i.i.d. sample from a Normal(μ_1, σ^2) population and $X_{21}, \ldots, X_{2n_2}$ are an i.i.d. sample from a Normal(μ_2, σ^2) population, where μ_1, μ_2 and σ^2 are unknown parameters. Note that the two populations are assumed to have a common variance. The two samples are assumed to be independently selected. Let us do an LR test of $\mathrm{H}_0 : \mu_1 = \mu_2$ versus $\mathrm{H}_1 : \mu_1 \neq \mu_2$.

(a) First we find the MLEs in the full model. The joint pdf is $f(\boldsymbol{x}_1\,;\, \mu_1, \sigma^2)\, f(\boldsymbol{x}_2\,;\, \mu_2, \sigma^2)$, so the log-likelihood is

$$\log L = \log f(\boldsymbol{x}_1\,;\, \mu_1, \sigma^2) + \log f(\boldsymbol{x}_2\,;\, \mu_2, \sigma^2) \ .$$

Now $\frac{\partial}{\partial\mu_1} \log L = \frac{\partial}{\partial\mu_1} \log f(\boldsymbol{x}_1\,;\, \mu_1, \sigma^2)$, which is the same as if only the first sample had been observed. Therefore, as in Example 7.2.7, $\widehat{\mu}_1 = \overline{x}_1$. Similarly, $\widehat{\mu}_2 = \overline{x}_2$. Next we can plug these into the log-likelihood and maximize the resulting function of σ^2. Its derivative is

$$\begin{aligned} &\frac{\partial}{\partial(\sigma^2)} \log f(\boldsymbol{x}_1\,;\, \overline{x}_1, \sigma^2) + \frac{\partial}{\partial(\sigma^2)} \log f(\boldsymbol{x}_2\,;\, \overline{x}_2, \sigma^2) \\ &= -\frac{n_1}{2\sigma^2} + \frac{1}{2\sigma^4} \sum (x_{1i} - \overline{x}_1)^2 - \frac{n_2}{2\sigma^2} + \frac{1}{2\sigma^4} \sum (x_{2i} - \overline{x}_2)^2 \\ &= \frac{n_1+n_2}{2\sigma^4} \left(\widehat{\sigma}^2 - \sigma^2 \right) \end{aligned}$$

where $\widehat{\sigma}^2 = \frac{1}{n_1+n_2} \left[\sum (x_{1i} - \overline{x}_1)^2 + \sum (x_{2i} - \overline{x}_2)^2 \right]$. Note that $\widehat{\sigma}^2 = \frac{n_1+n_2-2}{n_1+n_2} s_{\mathrm{P}}^2$ where s_{P}^2 is the pooled estimator of variance (4.5.7). From the derivative we see that $\widehat{\sigma}^2$ is the MLE of σ^2.

(b) Next we find the MLEs $\widehat{\mu}_0$ and $\widehat{\sigma}_0^2$ under the null hypothesis. Under the null hypothesis that $\mu_1 = \mu_2$, the two samples together constitute a single sample of size $n_1 + n_2$ from a Normal$(\mu\,, \sigma^2)$ population, where μ is the common value of μ_1 and μ_2. So $\widehat{\mu}_0 = \overline{x}$, which can be expressed as $\overline{x} = (\sum x_{1i} + \sum x_{2i})/(n_1 + n_2) = (n_1\overline{x}_1 + n_2\overline{x}_2)/(n_1 + n_2)$. And $\widehat{\sigma}_0^2 = \frac{1}{n_1+n_2} \left[\sum (x_{1i} - \overline{x})^2 + \sum (x_{2i} - \overline{x})^2 \right]$.

$$\begin{aligned} \text{(c) } \Lambda &= \frac{f(\boldsymbol{x}_1\,;\overline{x},\widehat{\sigma}_0^2)\, f(\boldsymbol{x}_2\,;\overline{x},\widehat{\sigma}_0^2)}{f(\boldsymbol{x}_1\,;\overline{x}_1,\widehat{\sigma}^2)\, f(\boldsymbol{x}_2\,;\overline{x}_2,\widehat{\sigma}^2)} \\ &= \frac{\left(1/\sqrt{2\pi\widehat{\sigma}_0^2}\right)^{n_1+n_2} \exp\left[-\frac{1}{2\widehat{\sigma}_0^2} \left\{ \sum (x_{1i} - \overline{x})^2 + \sum (x_{2i} - \overline{x})^2 \right\}\right]}{\left(1/\sqrt{2\pi\widehat{\sigma}^2}\right)^{n_1+n_2} \exp\left[-\frac{1}{2\widehat{\sigma}^2} \left\{ \sum (x_{1i} - \overline{x}_1)^2 + \sum (x_{2i} - \overline{x}_2)^2 \right\}\right]} \ . \\ &= \left(\frac{\widehat{\sigma}^2}{\widehat{\sigma}_0^2} \right)^{\frac{n_1+n_2}{2}} \frac{\exp\left[-\frac{1}{2\widehat{\sigma}_0^2} \left\{ (n_1+n_2)\widehat{\sigma}_0^2 \right\}\right]}{\exp\left[-\frac{1}{2\widehat{\sigma}^2} \left\{ (n_1+n_2)\widehat{\sigma}^2 \right\}\right]} \end{aligned}$$

$$= \left(\frac{\widehat{\sigma}^2}{\widehat{\sigma}_0^2}\right)^{\frac{n_1+n_2}{2}} .$$

(d) The LR test rejects H_0 iff $\Lambda < c$, that is, iff $\dfrac{\widehat{\sigma}^2}{\widehat{\sigma}_0^2} < c'$.

Let us re-express $\widehat{\sigma}_0^2$ in terms of $\widehat{\sigma}^2$.
Recall that $\sum(x_i - a)^2 = \sum(x_i - \overline{x})^2 + n(\overline{x} - a)^2$. Hence

$$\sum(x_{1i} - \overline{x})^2 = \sum(x_{1i} - \overline{x}_1)^2 + n_1(\overline{x}_1 - \overline{x})^2$$

and

$$\sum(x_{2i} - \overline{x})^2 = \sum(x_{2i} - \overline{x}_2)^2 + n_2(\overline{x}_2 - \overline{x})^2 ,$$

so

$$\begin{aligned}(n_1 + n_2)\,\widehat{\sigma}_0^2 &= \sum(x_{1i} - \overline{x})^2 + \sum(x_{2i} - \overline{x})^2 \\ &= \sum(x_{1i} - \overline{x}_1)^2 + \sum(x_{2i} - \overline{x}_2)^2 + n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2 \\ &= \sum(x_{1i} - \overline{x}_1)^2 + \sum(x_{2i} - \overline{x}_2)^2 + \frac{n_1 n_2}{n_1+n_2}(\overline{x}_1 - \overline{x}_2)^2 \\ &= (n_1 + n_2)\,\widehat{\sigma}^2 + \frac{n_1 n_2}{n_1+n_2}(\overline{x}_1 - \overline{x}_2)^2 .\end{aligned}$$

The last equality uses the equations $\overline{x}_1 - \overline{x} = \frac{n_2}{n_1+n_2}(\overline{x}_1 - \overline{x}_2)$ and $\overline{x}_2 - \overline{x} = \frac{n_1}{n_1+n_2}(\overline{x}_2 - \overline{x}_1)$, which follow from the expression $\overline{x} = (n_1\overline{x}_1 + n_2\overline{x}_2)/(n_1 + n_2)$.

The LR test rejects H_0 iff $\dfrac{\widehat{\sigma}^2}{\widehat{\sigma}_0^2} < c'$ iff $\dfrac{\widehat{\sigma}_0^2}{\widehat{\sigma}^2} > c''$, and we now see that

$$\begin{aligned}\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}^2} = \frac{(n_1+n_2)\widehat{\sigma}_0^2}{(n_1+n_2)\widehat{\sigma}^2} &= \frac{(n_1+n_2)\,\widehat{\sigma}^2 + \frac{n_1 n_2}{n_1+n_2}(\overline{x}_1 - \overline{x}_2)^2}{(n_1+n_2)\widehat{\sigma}^2} \\ &= 1 + \frac{\frac{n_1 n_2}{n_1+n_2}(\overline{x}_1 - \overline{x}_2)^2}{(n_1+n_2-2)s_{\mathrm{P}}^2} .\end{aligned}$$

So the LR test rejects H_0 iff $\dfrac{(\overline{x}_1 - \overline{x}_2)^2}{s_{\mathrm{P}}^2} > c'''$ iff $\dfrac{|\overline{x}_1 - \overline{x}_2|}{s_{\mathrm{P}}} > c''''$ iff $\dfrac{|\overline{x}_1 - \overline{x}_2|}{s_{\mathrm{P}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > b$.

This is the usual t statistic for comparing two samples.
Let $b = \mathtt{tinv}(1 - \frac{\alpha}{2}, n_1 + n_2 - 2)$. ||

Example. Suppose $X_{11}, \ldots, X_{1n_1}$ are an i.i.d. sample from a Normal(μ_1, σ_1^2) population and $X_{21}, \ldots, X_{2n_2}$ are an i.i.d. sample from a Normal(μ_2, σ_2^2) population, where μ_1, μ_2, σ_1^2 and σ_2^2 are unknown parameters. This differs from the preceding example in that the variances of the two populations may be different. The two samples are assumed to be independently selected. Let us do an LR test of $\mathrm{H}_0 : \mu_1 = \mu_2$ versus $\mathrm{H}_1 : \mu_1 \neq \mu_2$.

(a) In the full model we have two independent samples with no parameters in common, so the MLEs can be obtained separately from the two samples. Thus

$$\widehat{\mu}_1 = \overline{x}_1\,, \quad \widehat{\sigma}_1^2 = \frac{1}{n_1}\sum(x_{1i}-\overline{x}_1)^2, \quad \widehat{\mu}_2 = \overline{x}_2\,, \quad \widehat{\sigma}_2^2 = \frac{1}{n_2}\sum(x_{2i}-\overline{x}_2)^2\,.$$

(b) When we try to obtain the MLEs under H$_0$, we find that there are no explicit formulas for them. Given the observed values of the data, $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$, one can form the log-likelihood function $\log f(\boldsymbol{x}\,;\,\mu,\sigma_1^2,\sigma_2^2)$ under H$_0$ and use numerical methods to maximize this function over the three variables μ, σ_1^2 and σ_2^2. (See the optimization toolbox in Matlab.) One reason that there are no explicit formulas for the MLEs in this model is that it is not a full-rank exponential family. It is an exponential family, but when the pdf is written as in (3.8.4), there are 4 "canonical" statistics $R_i(\boldsymbol{x})$ but there are 3 parameters. The well-behaved exponential families are those having full rank, i.e., those for which the number of canonical statistics is the same as the number of parameters.

(c) The generalized likelihood ratio is

$$\Lambda = \frac{f(\boldsymbol{x}_1\,;\widehat{\mu}_0\,,\widehat{\sigma}_{10}^2)\,f(\boldsymbol{x}_2\,;\widehat{\mu}_0\,,\widehat{\sigma}_{20}^2)}{f(\boldsymbol{x}_1\,;\overline{x}_1,\widehat{\sigma}_1^2)\,f(\boldsymbol{x}_2\,;\overline{x}_2,\widehat{\sigma}_2^2)}$$

where, as mentioned in (b), the values of $\widehat{\mu}_0$, $\widehat{\sigma}_{10}^2$ and $\widehat{\sigma}_{20}^2$ must be obtained by numerical methods. We should reject H$_0$ if Λ is significantly small, but to judge significance we need to know, at least approximately, the distribution of Λ under H$_0$. For large sample sizes n_1 and n_2, this distribution can be approximated. We will return to this problem later. $\|$

Example. As in the preceding example, suppose we have two independent samples, one i.i.d. sample of size n_1 from a Normal(μ_1,σ_1^2) population and another i.i.d. sample of size n_2 from a Normal(μ_2,σ_2^2) population. Let us do an LR test of H$_0 : \sigma_1^2 = \sigma_2^2$ versus H$_1 : \sigma_1^2 \neq \sigma_2^2$.

(a) The MLEs in the full model are given in the preceding example.

(b) Under H$_0$ we have two independent samples from normal populations having a common variance. This is the full model in the example on p. 41 above. The MLEs are

$$\widehat{\mu}_{10} = \overline{x}_1\,, \quad \widehat{\mu}_{20} = \overline{x}_2\,, \quad \widehat{\sigma}_0^2 = \frac{n_1+n_2-2}{n_1+n_2}\,s_{\text{p}}^2$$

where s_{p}^2 is the pooled estimate of variance.

(c) $\Lambda = \cdots = \dfrac{\widehat{\sigma}_1^{n_1}\widehat{\sigma}_2^{n_2}}{\widehat{\sigma}_0^{n_1+n_2}}\,.$

(d) We can express the likelihood ratio as $\Lambda = g(T)$ where $T = s_1^2/s_2^2$ and $g(t) = kt^{n_1}/((n_1-1)t+(n_2-1))^{n_1+n_2}$ where k is a positive constant not involving t (it is some function of n_1 and n_2). To determine the shape of the function $g(t)$, we take its derivative. It is easier to take the derivative of $\log g(t)$, and it has the same sign as the derivative of $g(t)$. We find that the shape of the function is the same as in the picture on p. 40; that is, the function increases to a unique maximum and then decreases. Therefore, $\Lambda < c$ iff $T < a$ or $T > b$ for suitable a and b satisfying $g(a) = g(b)$.

Recall from Example 4.5.3 that $(S_1^2/\sigma_1^2)\big/(S_2^2/\sigma_2^2) \sim F(n_1-1, n_2-1)$. Under H$_0$, $\sigma_1^2 = \sigma_2^2$ and so $T = S_1^2/S_2^2 \sim F(n_1-1, n_2-1)$. As before, we will ignore the condition $g(a) = g(b)$ and use the equal-tailed modification of the LRT. Let $a = \mathtt{finv}(\frac{1}{2}\alpha, n_1-1, n_2-1)$ and $b = \mathtt{finv}(1-\frac{1}{2}\alpha, n_1-1, n_2-1)$. ||

Bivariate data

Example. Suppose $(X_{11}, X_{21}), \ldots, (X_{1n}, X_{2n})$ are an i.i.d. sample from a bivariate normal distribution $\mathrm{N}_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Here ρ denotes the correlation coefficient $\rho = \mathrm{Cov}(X_{i1}, X_{i2})/\sigma_1\sigma_2$. Let us test $\mathrm{H}_0 : \mu_1 = \mu_2$ versus $\mathrm{H}_1 : \mu_1 \neq \mu_2$. It turns out that the MLEs under H$_0$ are not explicit and would need to be calculated by numerical methods. So the LR test is difficult to perform in this example.

An approach that leads to a nicer test is to reduce the data to the differences $Y_i = X_{1i} - X_{2i}$. The vector $Y = (Y_1, \ldots, Y_n)$ is not a sufficient statistic; we are restricting our attention to Y only for convenience. The differences $Y_1, \ldots, Y_n$ are an i.i.d. sample from a $\mathrm{N}(\mu_1 - \mu_2, \sigma_*^2)$ population. (We could express σ_*^2 as a function of σ_1^2, σ_2^2 and ρ but this is not necessary for deriving our test.) Based on $\boldsymbol{Y}$, the LR test of $\mathrm{H}_0 : \mu_1 - \mu_2 = 0$ versus $\mathrm{H}_1 : \mu_1 - \mu_2 \neq 0$ is the one-sample t test that rejects H$_0$ iff $|\overline{Y}|/(S_Y/\sqrt{n}) > \mathtt{tinv}(\frac{1}{2}\alpha, n-1)$. ||

Example. Consider an i.i.d. sample from a bivariate normal distribution as in the preceding example. Let us test $\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2$ versus $\mathrm{H}_1 : \sigma_1^2 \neq \sigma_2^2$. It turns out that the MLEs under H$_0$ are difficult to derive directly, but we can use the following trick. Transform the pairs (X_{i1}, X_{i2}) to pairs (Y_{i1}, Y_{i2}) where $Y_{i1} = X_{1i} + X_{2i}$ and $Y_{i2} = X_{1i} - X_{2i}$. This is a one-to-one transformation and so the n transformed pairs constitute a sufficient statistic. They are an i.i.d. sample from a bivariate normal distribution $\mathrm{N}_2(\mu_1^*, \mu_2^*, \sigma_1^{*2}, \sigma_2^{*2}, \rho^*)$ where $\rho^* = \mathrm{Cov}(Y_{i1}, Y_{i2})/\sigma_1^*\sigma_2^*$. (It is not necessary to obtain formulas for μ_1^*, μ_2^*, σ_1^{*2} and σ_2^{*2}.) Note that $\mathrm{Cov}(Y_{i1}, Y_{i2}) = \mathrm{Cov}(X_{1i} + X_{2i}, X_{1i} - X_{2i}) = \mathrm{Cov}(X_{1i}, X_{1i}) - \mathrm{Cov}(X_{1i}, X_{2i}) + \mathrm{Cov}(X_{2i}, X_{1i}) - \mathrm{Cov}(X_{2i}, X_{2i}) = \sigma_1^2 - \sigma_2^2$. Therefore, $\sigma_1^2 = \sigma_2^2$ iff $\rho^* = 0$. What we want now is the LR test for $\mathrm{H}_0 : \rho^* = 0$ versus $\mathrm{H}_1 : \rho^* \neq 0$. We address this in the next example. ||

Example. Consider an i.i.d. sample from a bivariate normal distribution $N_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ as in the preceding two examples. Let us do an LR test of $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$.

(a) It can be shown that the MLEs in the full model are

$$\widehat{\mu}_1 = \overline{x}_1 , \quad \widehat{\sigma}_1^2 = \frac{1}{n}\sum(x_{1i} - \overline{x}_1)^2, \quad \widehat{\mu}_2 = \overline{x}_2 , \quad \widehat{\sigma}_2^2 = \frac{1}{n}\sum(x_{2i} - \overline{x}_2)^2 ,$$
$$\text{and} \quad \widehat{\rho} = \frac{1}{n}\sum(x_{1i} - \overline{x}_1)(x_{2i} - \overline{x}_2)/\widehat{\sigma}_1\widehat{\sigma}_2 .$$

Note that the MLEs of μ_1, μ_2, σ_1^2 and σ_2^2 are the same as they are in the case of independent samples (when $\rho = 0$).

(b) Under H_0 we have two independent samples and the MLEs are the same as in (a).

(c) $\Lambda = \cdots = \left(\sqrt{1 - \widehat{\rho}^2}\right)^n$.

(d) $\Lambda < c$ iff $|\widehat{\rho}| > k$. This expression for the test is interpretable, but for the test to be implementable we need to know the distribution of $\widehat{\rho}$ under H_0, or the distribution of some one-to-one function of $\widehat{\rho}$. Fisher discovered that $T = \sqrt{n-2}\,\widehat{\rho}\Big/\sqrt{1 - \widehat{\rho}^2}$ has a $t(n-2)$ distribution under H_0. So the LR test rejects H_0 iff $|T| > \mathtt{tinv}(\frac{1}{2}\alpha, n-2)$. ||

LR tests of one-sided alternatives

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Normal(μ, σ_0^2) where σ_0^2 is known. Consider the problem of testing $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$. Let us do an LR test. The LRT statistic is $\Lambda = f(\boldsymbol{x}\,;\, \widehat{\mu}_0)/f(\boldsymbol{x}\,;\, \widehat{\mu})$ where $\widehat{\mu}$ is the MLE of μ in the full model and $\widehat{\mu}_0$ is the MLE of μ under H_0. We know that $\widehat{\mu} = \overline{x}$, but we need to figure out what $\widehat{\mu}_0$ is.

In deriving the MLE in the full model, one uses the derivative

$$\frac{\partial}{\partial\mu}\log f(\boldsymbol{x}\,;\, \mu) = \frac{n}{\sigma_0^2}(\overline{x} - \mu) .$$

From the sign of the derivative we see that the log-likelihood is increasing for $\mu < \overline{x}$, attains a global maximum at $\mu = \overline{x}$, and decreases for $\mu > \overline{x}$. Under H_0, $\mu \leq \mu_0$. If $\overline{x} \leq \mu_0$, then the maximum of $\log f(\boldsymbol{x}\,;\, \mu)$ under H_0 occurs at $\mu = \overline{x}$. However, if $\overline{x} > \mu_0$, then $\log f(\boldsymbol{x}\,;\, \mu)$ is increasing for all $\mu \leq \mu_0$, which implies that the maximum occurs at $\mu = \mu_0$. Thus we have

$$\widehat{\mu}_0 = \begin{cases} \overline{x} & \text{if } \overline{x} \leq \mu_0 \\ \mu_0 & \text{if } \overline{x} > \mu_0 \end{cases} .$$

Now

$$\Lambda = \begin{cases} 1 & \text{if } \overline{x} \leq \mu_0 \\ f(\boldsymbol{x}\,;\,\mu_0)/f(\boldsymbol{x}\,;\overline{x}) & \text{if } \overline{x} > \mu_0 \end{cases} .$$

We reject H_0 iff $\Lambda < c$. Since the values of Λ are always between 0 and 1, we would have $c < 1$ and we would never reject if $\overline{x} \leq \mu_0$. (This is sensible because $\overline{x} \leq \mu_0$ is entirely consistent with the null hypothesis that $\mu \leq \mu_0$.) So, $\Lambda < c$ iff $\overline{x} > \mu_0$ and $f(\boldsymbol{x}\,;\,\mu_0)/f(\boldsymbol{x}\,;\overline{x}) < c$. On p. 37 above, when deriving the LRT for a two-sided alternative, we saw that the likelihood ratio is less than c iff $|\,\overline{x} - \mu_0| > b$ for a suitable b. With the added condition $\overline{x} > \mu_0$, we see that $\Lambda < c$ iff $\overline{x} - \mu_0 > b$. This is equivalent to rejecting H_0 iff the z statistic $(\overline{x} - \mu_0)/(\sigma_0/\sqrt{n})$ is significantly large. Noting that the z statistic is a strictly increasing function of $\sum x_i$, we see that the LR test is the same as the UMP test we derived on p. 16 above. ||

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Normal(μ, σ^2) where μ and σ^2 are both unknown parameters. Consider the problem of testing $\mathrm{H}_0 : \mu \leq \mu_0$ versus $\mathrm{H}_1 : \mu > \mu_0$. In a way similar to the preceding example, it can be shown that the LR test rejects H_0 iff $T >$ $\mathtt{tinv}(1 - \alpha, n-1)$ where $T = (\overline{X} - \mu_0)/(S/\sqrt{n})$. This is the usual one-sided t test. ||

Notes on Large-Sample Inference

(see Ch. 12 in Mukhopadhyay)

Large-sample distribution of an MLE

Suppose $X_1, \dots, X_n$ are i.i.d. with pmf or pdf $f(x\,;\theta)$ where θ is a real-valued parameter. We will assume some "regularity" conditions:

RC1. $f(x\,;\theta)$ has the same support for all θ.

RC2. $f(x\,;\theta)$ is differentiable.

RC3. Other technical conditions concerning derivatives and integrals.

All the common pdf's that satisfy RC1 also satisfy RC2 and RC3. Note that the pdf of Uniform$(0, \theta)$ does not satisfy RC1.

The MLE $\widehat{\theta}$, by definition, maximizes the joint pdf $f(\boldsymbol{x}\,;\theta) = \prod_{i=1}^{n} f(x_i\,;\theta)$. Equivalently, $\widehat{\theta}$ maximizes the log-likelihood $\log f(\boldsymbol{x}\,;\theta) = \sum_{i=1}^{n} \log f(x_i\,;\theta)$. (In order for the log-likelihood to be well-defined, we need RC1.) The derivative of a differentable function is 0 at a maximum, so $\widehat{\theta}$ satisfies the *likelihood equation*:

$$\frac{\partial}{\partial\theta} \log f(\boldsymbol{x}\,;\theta) = 0 \ .$$

A solution of the likelihood equation is usually an MLE, but to be sure, we must check that it is a global maximum.

Notation. We will be studying samples as the sample size increases, and so we should keep track of n. Let us write $\boldsymbol{X} = \boldsymbol{X}_n = (X_1, \dots, X_n)$ and $\widehat{\theta} = \widehat{\theta}_n = \widehat{\theta}_n(X_1, \dots, X_n)$.

Theorem. (12.2.3) Assume conditions RC1, RC2, RC3. Then the MLE $\widehat{\theta}_n$ is a consistent estimator of θ.

This means that $\widehat{\theta}_n \xrightarrow{\text{P}} \theta$ as $n \to \infty$; that is, for all $\epsilon > 0$, $\text{P}_\theta\{|\widehat{\theta}_n - \theta| < \epsilon\}$ as $n \to \infty$. In words, for large sample sizes, the MLE $\widehat{\theta}_n$ is close to θ with high probability. The proof of this theorem is based on the WLLN (Theorem 5.2.1).

Theorem. (12.2.4) Assume conditions RC1, RC2, RC3. Then the MLE $\widehat{\theta}_n$ has an asymptotic normal distribution. Specifically, as $n \to \infty$,

$$\frac{\widehat{\theta}_n - \theta}{\sqrt{\mathcal{I}_1(\theta)^{-1}/n}} \xrightarrow{\mathcal{L}} \text{N}(0, 1)$$

where $\mathcal{I}_1(\theta)$ is the Fisher information in a single observation.

This theorem tells us that, for large sample sizes,

$$\widehat{\theta}_n \underset{\text{approx}}{\sim} \mathrm{N}(\theta, \frac{1}{n\mathcal{I}_1(\theta)}) .$$

The proof of this theorem is based on the CLT (Theorem 5.3.4).

Recall the Cramer-Rao Inequality in Theorem 7.5.1. Consider the case $\tau(\theta) = \theta$, so that $\tau'(\theta) = 1$. Using Definition 6.4.1, the theorem says that, under the conditions RC1, RC2, RC3, for a fixed sample size n, the Cramer-Rao Lower Bound $1/[n\mathcal{I}_1(\theta)]$ is the smallest possible variance among all unbiased estimators of θ. This is the same variance as the asymptotic variance of the MLE. It can shown that $1/[n\mathcal{I}_1(\theta)]$ is the smallest possible variance among all consistent asymptotically normal estimators of θ. An estimator that has the smallest possible variance is said to be *efficient*. So, under the regularity conditions, we can say that the MLE is asymptotically efficient.

We should recall the definition of Fisher information from section 6.4. Consider a data vector $\boldsymbol{x}$ with joint pmf or pdf $f(\boldsymbol{x}\,;\theta)$, where θ is a real-valued parameter. Assume conditions RC1, RC2, RC3. We define

$$\mathcal{I}(\theta) \;=\; \mathrm{E}_\theta\big[\{\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}\,;\theta)\}^2\big] \;.$$

Two other ways to calculate it are:

$$\mathcal{I}(\theta) \;=\; \mathrm{Var}_\theta\big[\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}\,;\theta)\big] \;,$$

$$\mathcal{I}(\theta) \;=\; -\,\mathrm{E}_\theta\big[\frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{X}\,;\theta)\big] \;.$$

Suppose $\boldsymbol{X} = (X_1, \ldots, X_n)$ where the X_i's are i.i.d. Then $\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}\,;\theta) = \sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(X_i\,;\theta)$, and so, keeping track of the sample size by a subscript, we have $\mathcal{I}_n(\theta) = \mathrm{Var}_\theta\big[\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}\,;\theta)\big] = n\mathrm{Var}_\theta\big[\frac{\partial}{\partial\theta}\log f(X_1\,;\theta)\big] = n\mathcal{I}_1(\theta)$.

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Normal(μ, σ_0^2) where σ_0^2 is known. One finds that for a single x, $\frac{\partial}{\partial\mu}\log f(x\,;\mu) = \sigma_0^{-2}(x-\mu)$. Therefore, $\mathcal{I}_1(\mu) = \mathrm{E}_\mu\big[\{\sigma_0^{-2}(X-\mu)\}^2\big] = \sigma_0^{-4}\mathrm{E}_\mu\big[(X-\mu)^2\big] = \sigma_0^{-4}\sigma_0^2 = \sigma_0^{-2}$. Alternatively, $\mathcal{I}_1(\mu) = \mathrm{Var}_\mu\big[\sigma_0^{-2}(X-\mu)\big] = \sigma_0^{-4}\mathrm{Var}_\mu(X) = \sigma_0^{-4}\sigma_0^2 = \sigma_0^{-2}$, or $\mathcal{I}_1(\mu) = -\,\mathrm{E}_\mu\big[\frac{\partial^2}{\partial\mu^2}\log f(X\,;\mu)\big] = -\,\mathrm{E}_\mu\big[-\sigma_0^{-2}\big] = \sigma_0^{-2}$.

The two theorems above imply that, as $n \to \infty$,

$$\widehat{\mu}_n \xrightarrow{\mathrm{p}} \mu$$

and

$$\frac{\widehat{\mu}_n-\mu}{\sqrt{\sigma_0^2/n}} \xrightarrow{\mathcal{L}} \mathrm{N}(0,1) \ .$$

We know that $\widehat{\mu} = \overline{x}$, the sample mean, and so these two results are instances of the WLLN and the CLT. For a sample from a normal population, the random variable $(\overline{X}-\mu)/\sqrt{\sigma_0^2/n}$ has a standard normal distribution exactly for all n, not only in the limit. $\|$

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Bernoulli(θ), $0 < \theta < 1$. The pmf is $f(x\,;\theta) = \theta^x(1-\theta)^{1-x}$. We find that

$$\frac{\partial}{\partial\theta}\log f(x\,;\theta) \ = \ \cdots \ = \ \frac{x}{\theta} - \frac{1-x}{1-\theta} \ = \ \frac{x}{\theta(1-\theta)} - \frac{1}{1-\theta} \ ,$$

$$\mathcal{I}_1(\theta) \ = \ \mathrm{Var}\big[\frac{\partial}{\partial\theta}\log f(X\,;\theta)\big] \ = \ \frac{1}{\theta(1-\theta)} \ .$$

By (12.2.4), for large n,

(*) $\quad \widehat{\theta}_n \underset{\text{approx}}{\sim} \mathrm{N}(\theta, \frac{\theta(1-\theta)}{n})$.

We know that $\widehat{\theta}_n = \overline{X} =$ the sample mean $=$ the sample proportion, $\theta =$ the population mean $=$ the population proportion, and $\mathrm{Var}(\widehat{\theta}_n) = \mathrm{Var}(\overline{X}) = \theta(1-\theta)/n$. So (*) could also be concluded directly from the CLT.

Two uses of (*) are given next.

(A) To construct an approximate 95%-confidence interval for θ, re-express (*) as

$$\frac{\widehat{\theta}_n-\theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \underset{\text{approx}}{\sim} \mathrm{N}(0,1) \ .$$

This is an approximate pivot for θ. Now, for large n,

$$\mathrm{P}_\theta\{-1.96 < \frac{\widehat{\theta}_n-\theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} < 1.96\} \ = \ 0.95 \ .$$

Unfortunately, we cannot manipulate these inequalities to get nice confidence limits for θ. This is because θ appears not only in the numerator of the approximate pivot but also in the denominator. To get around this problem, we substitute $\widehat{\theta}_n$ for θ in the denominator. Using the re-expression of (*) and the fact that $\widehat{\theta}_n$ is a consistent estimator of θ and applying Theorem 5.2.5 and Slutsky's Theorem 5.3.3, we can show that

$$\frac{\widehat{\theta}_n-\theta}{\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}}} \underset{\text{approx}}{\sim} \mathrm{N}(0,1) \ .$$

This is another approximate pivot for θ. For large n,

$$\mathrm{P}_\theta\Big\{ -1.96 < \frac{\widehat{\theta}-\theta}{\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}}} < 1.96\Big\} \approx 0.95 \ .$$

By manipulating these inequalities we obtain an approximate 95%-confidence interval for θ:

$$\widehat{\theta} - 1.96\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}} < \theta < \widehat{\theta} + 1.96\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}} \ .$$

(B) To do a test of $\mathrm{H}_0 : \theta = \theta_0$ versus $\mathrm{H}_1 : \theta \neq \theta_0$ with level approximately 0.05, reject H_0 iff

$$\left| \frac{\widehat{\theta}-\theta_0}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}} \right| > 1.96 \ .$$

The critical value 1.96 is appropriate in so far as the test statistic has approximately a $\mathrm{N}(0, 1)$ distribution under H_0. Note that in forming the test statistic, we have taken full advantage of the null hypothesis by putting θ_0 in the denominator rather than $\widehat{\theta}$. How close the actual level is to 0.05 depends on the sample size n and also on the population proportion θ. ||

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. with pdf $f(x\,;\theta) = \theta e^{-\theta x}$ for $x > 0$ for some $\theta > 0$. (This is the Exponential distribution parameterized by a rate parameter.) We find that

$$\frac{\partial}{\partial\theta}\log f(x\,;\theta) \;=\; \cdots \;=\; \frac{1}{\theta} - x \ ,$$

$$\mathcal{I}_1(\theta) \;=\; \mathrm{Var}(\tfrac{1}{\theta} - X) \;=\; \mathrm{Var}(X) \;=\; \frac{1}{\theta^2} \ .$$

The MLE is the solution of $\sum\limits_{i=1}^{n} \frac{\partial}{\partial\theta}\log f(x_i\,;\theta) \;=\; \frac{n}{\theta} - \sum\limits_{i=1}^{n} x_i \;=\; 0$, that is, $\widehat{\theta} = \frac{1}{\overline{x}}$.

By (12.2.4), for large n,

(*) $\quad \widehat{\theta} \underset{\text{approx}}{\sim} \mathrm{N}(\theta, \frac{\theta^2}{n})$.

The approximation (*) could also be concluded from the CLT and the Mann-Wald Theorem 5.3.5 for $g(\overline{x}) = 1/\overline{x}$. As in the preceding example, one can use (*) to obtain approximate confidence intervals and tests for θ. ||

In this example the CLT implies that $\overline{X}$ has approximately a normal distribution for large n, and the Mann-Wald Theorem implies that $1/\overline{X}$ also has approximately a normal distribution for large n. How can this be? Although a linear function of a normal random variable is normal, a nonlinear function of a normal random variable is nonnormal. However, in a small interval, a differentiable function is approximately linear. That is, a differentiable function is well approximated in a small neighborhood of a point by the tangent line at that point. For large n, the variance of $\overline{X}$ is small and so with high probability its values occur in a small interval around its mean.

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Gamma(α,1), $\alpha > 0$. The pdf is

$$f(x\,;\alpha) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x} \text{ for } x > 0 \ ,$$

$$\log f(x\,;\alpha) = -\log\Gamma(\alpha) + (\alpha-1)\log x - x \ ,$$

$$\frac{\partial}{\partial\alpha}\log f(x\,;\alpha) = -\psi(\alpha) + \log x$$

where $\psi(\alpha) = \frac{\partial}{\partial\alpha}\log\Gamma(\alpha)$, the digamma function. A table of this function is available, for example, in Abramowitz & Stegun (1974).

The MLE is the solution of $\sum_{i=1}^{n}\frac{\partial}{\partial\alpha}\log f(x_i\,;\alpha) = -n\psi(\alpha) + \sum_{i=1}^{n}\log x_i = 0$. That is, $\psi(\widehat{\alpha}) = \frac{1}{n}\sum_{i=1}^{n}\log x_i$. There is no explicit solution for $\widehat{\alpha}$. For a given data set, the value of $\widehat{\alpha}$ must be calculated by numerical methods. Even though there is no explicit formula for it, we can approximate its distribution by (12.2.4) if n is large.

The Fisher information in a single observation is

$$\mathcal{I}_1(\alpha) = \text{Var}[-\psi(\alpha) + \log X] = \text{Var}[\log X] = ?\,.$$

Alternatively,

$$\mathcal{I}_1(\alpha) = -\text{E}[\frac{\partial^2}{\partial\alpha^2}\log f(X\,;\alpha)] = -\text{E}[-\psi'(\alpha)] = \psi'(\alpha)\,,$$

the trigamma function. A table of this function is available in Abramowitz & Stegun.

By (12.2.4), for large n,

$$(*) \qquad \widehat{\alpha} \underset{\text{approx}}{\sim} \text{N}(\alpha, \frac{1}{n\psi'(\alpha)}) \ .$$

(A) To calculate an approximate 95%-confidence interval for α, first use a digamma table to calculate the MLE $\widehat{\alpha}$. Then use a trigamma table to obtain $\psi'(\widehat{\alpha})$. The interval has limits

$$\widehat{\alpha} \pm 1.96\sqrt{\frac{1}{n\psi'(\widehat{\alpha})}} \ .$$

(B) To test $\text{H}_0 : \alpha = \alpha_0$ versus $\text{H}_1 : \alpha \neq \alpha_0$, reject H_0 iff

$$\left| \frac{\widehat{\alpha}-\alpha_0}{\sqrt{\frac{1}{n\psi'(\alpha_0)}}} \right| > 1.96 \ . \quad \|$$

The same sort of approach used in the preceding three examples to obtain approximate confidence intervals and tests can be followied whenever we have an estimator that has approximately a normal distribution with the parameter of interest as the mean. Three ways to establish that an estimator has approximately a normal distribution are: (1) the CLT, (2) the Mann-Wald Theorem, and (3) the asymptotic MLE result (12.2.4).

Note that not all estimators have approximately a normal distribution, even for large n. For example, given an i.i.d. sample from a Uniform$(0, \theta)$ distribution, the MLE of θ is the sample maximum $X_{(n)}$, which does not have approximately a normal distribution.

Example. Suppose $X_{11}, \dots, X_{1n_1}$ are an i.i.d. sample from a Bernoulli(θ_1) population and $X_{21}, \dots, X_{2n_2}$ are an i.i.d. sample from a Bernoulli(θ_2) population. The two samples are assumed to have been selected independently of one another.

(A) Find an approximate 95%-confidence interval for $\theta_1 - \theta_2$. The MLEs of θ_1 and θ_2 are the samples proportions $\widehat{\theta}_1 = \overline{X}_1$ and $\widehat{\theta}_2 = \overline{X}_2$. By the CLT,

$$\widehat{\theta}_1 \underset{\text{approx}}{\sim} \mathrm{N}\Big(\theta_1, \tfrac{\theta_1(1-\theta_1)}{n_1}\Big) \quad \text{and} \quad \widehat{\theta}_2 \underset{\text{approx}}{\sim} \mathrm{N}\Big(\theta_2, \tfrac{\theta_2(1-\theta_2)}{n_2}\Big) \ .$$

The two samples are independent, so

$$\widehat{\theta}_1 - \widehat{\theta}_2 \underset{\text{approx}}{\sim} \mathrm{N}\Big(\theta_1 - \theta_2, \tfrac{\theta_1(1-\theta_1)}{n_1} + \tfrac{\theta_2(1-\theta_2)}{n_2}\Big) \ .$$

We can substitute consistent estimators of θ_1 and θ_2 into the variance. (This is justified by Slutsky's Thorem and Theorem 5.2.5.) Thus

$$\widehat{\theta}_1 - \widehat{\theta}_2 \underset{\text{approx}}{\sim} \mathrm{N}\Big(\theta_1 - \theta_2, \tfrac{\widehat{\theta}_1(1-\widehat{\theta}_1)}{n_1} + \tfrac{\widehat{\theta}_2(1-\widehat{\theta}_2)}{n_2}\Big) \ .$$

An approximate 95%-confidence interval for $\theta_1 - \theta_2$ has limits

$$\widehat{\theta}_1 - \widehat{\theta}_2 \pm 1.96\sqrt{\tfrac{\widehat{\theta}_1(1-\widehat{\theta}_1)}{n_1} + \tfrac{\widehat{\theta}_2(1-\widehat{\theta}_2)}{n_2}} \ .$$

The confidence coefficient is approximately 0.95 if n_1 and n_2 are large.

(B) Test $\mathrm{H}_0 : \theta_1 = \theta_2$ versus $\mathrm{H}_1 : \theta_1 \neq \theta_2$ at level approximately $.05$.

In calculating the test statistic it is a good idea to take full advantage of the null hypothesis. Thus, when estimating the SE of $\widehat{\theta}_1 - \widehat{\theta}_2$, let us suppose $\theta_1 = \theta_2 =$ (say) θ, so that the approximate SE of $\widehat{\theta}_1 - \widehat{\theta}_2$ is $\sqrt{\frac{\theta(1-\theta)}{n_1} + \frac{\theta(1-\theta)}{n_2}} = \sqrt{\theta(1-\theta)\big(\frac{1}{n_1} + \frac{1}{n_2}\big)}$. Under H_0 the MLE of θ is $\widehat{\theta}_0 = \overline{x} = \frac{n_1\widehat{\theta}_1 + n_2\widehat{\theta}_2}{n_1+n_2}$. We reject H_0 iff

$$\left|\frac{\widehat{\theta}_1-\widehat{\theta}_2}{\sqrt{\widehat{\theta}_0(1-\widehat{\theta}_0)(\frac{1}{n_1}+\frac{1}{n_2})}}\right| > 1.96 \ .$$

The level of the test is approximately 0.05 if n_1 and n_2 are large. ||

Mukhopadhyay considers $n \geq 30$ to be "large".

Transformations of the MLE

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Bernoulli(θ) for some $0 < \theta < 1$. By result (12.2.4) (or by the CLT), for large n, the MLE $\widehat{\theta} = \overline{X}$ is approximately distributed as

$$\widehat{\theta} \underset{\text{approx}}{\sim} \mathrm{N}(\theta, \frac{\theta(1-\theta)}{n}) .$$

By the Mann-Wald Theorem 5.3.5, for any differentiable function $g(\theta)$ that is strictly increasing on $(0, 1)$,

$$g(\widehat{\theta}) \underset{\text{approx}}{\sim} \mathrm{N}(g(\theta), \{g'(\theta)\}^2 \frac{\theta(1-\theta)}{n}) .$$

Therefore, for large n, an approximate 95%-confidence interval for $g(\theta)$ is given by

$$g(\widehat{\theta}) - 1.96\, g'(\widehat{\theta})\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}} \ < \ g(\theta) \ < \ g(\widehat{\theta}) + 1.96\, g'(\widehat{\theta})\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}}$$

and so an approximate 95%-confidence interval for θ is given by

$$g^{-1}\left(g(\widehat{\theta}) - 1.96\, g'(\widehat{\theta})\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}}\right) \ < \ \theta \ < \ g^{-1}\left(g(\widehat{\theta}) + 1.96\, g'(\widehat{\theta})\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}}\right) .$$

For example, for $g(\theta) = \theta^2$, we get the interval

$$\sqrt{\widehat{\theta}^2 - 1.96(2\widehat{\theta})\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}}} \ < \ \theta \ < \ \sqrt{\widehat{\theta}^2 + 1.96(2\widehat{\theta})\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}}} \ .$$

For very large n, all such intervals, for various choices of $g(\theta)$, should be approximately valid in the sense of having a true confidence coefficient that is approximately equal to the nominal value 95%. But for sample sizes n that are only moderately large, some intervals will be more valid than others. And among those that are approximately valid, some intervals will be more precise (narrower) than others. Which choice of $g(\theta)$ is best? From the standpoint of simplicity, $g(\theta) = \theta$ is preferred, but in some cases (depending on the value of the true θ and the value of n), other choices may give higher validity or higher precision.

A transformation that Mukhopadhyay introduces is $g(\theta) = \arcsin(\sqrt{\theta})$. Note, however, that the motivation for this transformation is not necessarily to obtain confidence intervals that are more valid or more precise, but rather to stabilize the variance of $g(\widehat{\theta})$ so that it is approximately the same for all θ. This property is desirable when doing inference in an analysis-of-variance setting. By the Mann-Wald Theorem,

$$\arcsin(\sqrt{\widehat{\theta}}) \underset{\text{approx}}{\sim} \mathrm{N}(\arcsin(\sqrt{\theta}), \tfrac{1}{4n}) .$$

This leads to the following approximate 95%-confidence interval for θ:

$$\sin^2\left(\arcsin(\sqrt{\widehat{\theta}}) - 1.96\sqrt{\tfrac{1}{4n}}\right) < \theta < \sin^2\left(\arcsin(\sqrt{\widehat{\theta}}) + 1.96\sqrt{\tfrac{1}{4n}}\right) . \quad \|$$

Suppose

$$T_n \underset{\text{approx}}{\sim} \mathrm{N}(\theta, \tfrac{\sigma^2}{n})$$

and suppose $g(t)$ is a strictly increasing differentiable function. By the Mann-Wald Theorem,

$$g(T_n) \underset{\text{approx}}{\sim} \mathrm{N}(g(\theta), \{g'(\theta)\}^2 \tfrac{\sigma^2}{n}) .$$

If the variance of T_n depends on θ (i.e., $\sigma^2 = \sigma^2(\theta)$), one can try to stabilize the variance by choosing $g(\theta)$ so that $\{g'(\theta)\}^2\sigma^2(\theta) = c^2$, a constant. For such a $g(\theta)$,

$$g(T_n) \underset{\text{approx}}{\sim} \mathrm{N}(g(\theta), \tfrac{c^2}{n}) .$$

We want $g'(\theta) = c/\sigma(\theta)$, or

$$g(\theta) = \int \tfrac{c}{\sigma(\theta)} \mathrm{d}\theta .$$

Example. Consider the preceding example of a random sample from a Bernoulli population. We know that

$$\widehat{\theta} \underset{\text{approx}}{\sim} \mathrm{N}(\theta, \tfrac{\theta(1-\theta)}{n}) .$$

Here we have $\sigma^2(\theta) = \theta(1-\theta)$, so a variance-stabilizing transformation is

$$g(\theta) = \int \tfrac{c}{\sqrt{\theta(1-\theta)}} \mathrm{d}\theta = 2c \arcsin(\sqrt{\theta}) .$$

The integral may be found in a table of integrals or at the web site integrals.com. We may as well let $c = \frac{1}{2}$, so that $g(\theta) = \arcsin(\sqrt{\theta})$. As noted above, the large-sample variance of $\arcsin(\sqrt{\widehat{\theta}})$ is $c^2/n = 1/4n$. ||

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Poisson(λ) for some $\lambda > 0$. The MLE of λ is $\widehat{\lambda} = \overline{X}$ (Example 7.2.10). By the CLT (or result (12.2.4)), for large n, the MLE is approximately distributed as

$$\widehat{\lambda} \underset{\text{approx}}{\sim} \mathrm{N}(\lambda, \tfrac{\lambda}{n}).$$

Recall that $\mathrm{Var}(X_1) = \lambda$. The variance-stabilizing transformation is

$$g(\lambda) = \int \frac{c}{\sqrt{\lambda}} \mathrm{d}\lambda = 2c\sqrt{\lambda}.$$

We may as well let $c = \frac{1}{2}$ so that $g(\lambda) = \sqrt{\lambda}$. For large n we have

$$\sqrt{\widehat{\lambda}} \underset{\text{approx}}{\sim} \mathrm{N}(\sqrt{\lambda}, \tfrac{1}{4n}).$$

This is a useful transformation when doing analysis of variance on Poisson counts.

This transformation also improves the normal approximation, but it has been found that the transformation $\widehat{\lambda}^{\frac{2}{3}}$ gives an even better normal approximation. By using the Mann-Wald Theorem, we find that

$$\widehat{\lambda}^{\frac{2}{3}} \underset{\text{approx}}{\sim} \mathrm{N}(\lambda^{\frac{2}{3}}, \tfrac{4\lambda^{\frac{1}{3}}}{9n}).$$

Thus we have the following approximate 95%-confidence intervals for λ:

(1) $\widehat{\lambda} \pm 1.96\sqrt{\frac{\widehat{\lambda}}{n}}$

(2) $\left[\sqrt{\widehat{\lambda}} \pm 1.96\sqrt{\frac{1}{4n}}\,\right]^2$

(3) $\left[\widehat{\lambda}^{\frac{2}{3}} \pm 1.96\sqrt{\frac{4\widehat{\lambda}^{\frac{1}{3}}}{9n}}\,\right]^{\frac{3}{2}}$

In a simulation of these intervals for $n = 30$ and $\lambda = 1$, the coverage probability of interval (3) was closest to the nominal 95%.

Example. Suppose $(X_{11}, X_{21}), \ldots, (X_{1n}, X_{2n})$ are i.i.d. from a bivariate normal distribution, $\mathrm{N}_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ where ρ is the population correlation coefficient. The MLE of ρ is $\widehat{\rho} = \frac{1}{n}\sum(X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2)/\widehat{\sigma}_1\widehat{\sigma}_2$. When $\rho = 0$, Fisher found that $\sqrt{(n-2)}\,\widehat{\rho}/\sqrt{1-\widehat{\rho}^2} \sim t(n-1)$. When $\rho \neq 0$, the distribution of $\widehat{\rho}$ is more complicated. Fisher found that, for large n,

$$\widehat{\rho} \underset{\text{approx}}{\sim} \mathrm{N}(\rho, \tfrac{(1-\rho^2)^2}{n}).$$

An approximate 95%-confidence interval for ρ is

(1) $\widehat{\rho} \pm 1.96\frac{1-\widehat{\rho}^2}{\sqrt{n}}$.

Now try a variance-stabilizing transformation.

$$g(\lambda) = \int \frac{c}{1-\rho^2} d\lambda = \frac{c}{2} \log\left(\frac{1+\rho}{1-\rho}\right) .$$

Let $c = 2$ and let τ denote $\log((1+\rho)/(1-\rho))$, so that $\rho = (e^\tau - 1)/(e^\tau + 1)$.

$$\log\left(\frac{1+\widehat{\rho}}{1-\widehat{\rho}}\right) \underset{\text{approx}}{\sim} \mathrm{N}(\log\left(\frac{1+\rho}{1-\rho}\right), \frac{4}{n}) .$$

This yields an approximate 95%-confidence interval for ρ,

(2) $\rho_L < \rho < \rho_U$

where $\rho_L = (e^{\tau_L} - 1)/(e^{\tau_L} + 1)$, $\rho_U = (e^{\tau_U} - 1)/(e^{\tau_U} + 1)$, ←

$\tau_L = \widehat{\tau} - 1.96\sqrt{\frac{4}{n}}$, $\tau_U = \widehat{\tau} + 1.96\sqrt{\frac{4}{n}}$.

In a simulation of these intervals for $n = 30$ and $\rho = 0.5$, the coverage probability of interval (2) was closer to the nominal 95%.

Numerical example of binomial regression

The following data is from The Statistical Sleuth, p. 602.

At location i $(i = 1, \ldots, 7)$, m_i moths of a type common in Liverpool were placed on a tree. After 24 hours the experimenters counted the number y_i of moths that had been taken by predators. The distance in kilometers of each location from Liverpool is also listed.

Location	x	m	y
1	0.0	56	14
2	7.2	80	20
3	24.1	52	22
4	30.2	60	16
5	36.4	60	23
6	41.5	84	40
7	51.2	92	39

MLEs: $\widehat{\alpha} = -1.1290$, $\widehat{\beta} = 0.01850$

$$\bar{y} = 0.3595\,, \quad \widehat{\alpha}_0 = \log\left(\frac{\bar{y}}{1-\bar{y}}\right) = -0.5775$$

(A) LR test

$$-2\log\Lambda = 11.14 > 3.84$$

Reject $\beta = 0$ at level $.05$.

(B) Wald test

The information matrix is $\mathcal{I} = \begin{bmatrix} 108.927 & 3341.23 \\ 3341.23 & 133867 \end{bmatrix}$

$$\mathcal{I}^{-1} = \begin{bmatrix} 0.03917 & -0.009776 \\ -0.009776 & 0.00003187 \end{bmatrix}$$

$$\text{Var}(\widehat{\beta}) \approx 0.00003187\,, \quad \text{est.SE}(\widehat{\beta}) = \sqrt{0.00003187} = 0.005645$$

$$W = \frac{\widehat{\beta}}{\text{est.SE}(\widehat{\beta})} = \frac{0.01850}{0.005645} = 3.277 > 1.96$$

Reject $\beta = 0$ at level $.05$.