Transformations

Feng-Chang Lin

Department of Biostatistics
University of North Carolina at Chapel Hill

flin@bios.unc.edu

(C&B §2.1, §4.3)

Introduction

In this unit we will learn how to answer the following questions:

- $X \sim \mathrm{N}(0, 1)$. Find the distribution of X^2.
- $X \sim \mathrm{U}(0, 1)$. Find the distribution of $-\log X$.
- $X \sim \mathrm{Exp}(\lambda)$. Find the distribution of $X_{(1)} \equiv \min\{X_1, \ldots, X_n\}$.
- $X \sim \mathrm{Exp}(\lambda)$, $Y \sim \mathrm{Exp}(\mu)$, and $X \perp Y$. Find the distribution of $X - Y$. Find the distribution of X/Y.
- $X \sim \mathrm{Exp}(\lambda)$, $Y \sim \mathrm{Exp}(\mu)$, and $X \perp Y$. Find the (joint) distribution of $(X - Y, X/Y)$.
- Table of common distributions in page 627 in C&B may help.

Why Such Questions?

- Summarize data to make statistical inferences.
- Examples include sample mean and sample variance.
- Need to know the distributions under a given *model* in order to use the summary statistics.
- Mathematically, we formulate these statistics as *transformations*.

Transformation of One Variable

- X is a random variable with pdf or pmf f_X and one wants to find the distribution of $Y = g(X)$ where g is a given function.
- We define $\mathcal{X}$ to be the sample space of X

$$\mathcal{X} := \{x : f(x) > 0\},$$

  and $\mathcal{Y}$ to be the sample space of Y, where

$$\mathcal{Y} := \{y : g(x) = y \ \ \textit{for some } x \in \mathcal{X}\}.$$

- A set such as X or Y is called the support set of a distribution, or simply the support of the distribution.

PMF of Discrete Random Variables

- If X is discrete, the pmf of $g(X)$ is no more than simple enumeration.
- **Example**: X is Poisson(λ) and $Y = X^2$, i.e. $g(x) = x^2$. What is $P(Y = 25)$?

$$P(Y = 25) = P(X = 5) = e^{-\lambda}\lambda^5/5!$$

- How about $P(Y = 10)$?
- In this example $X = \{0, 1, 2, 3, 4, \ldots\}$ and $Y = \{0, 1, 4, 9, 16, \ldots\}$.
- For $y \geq 0$, we have $P(Y = y) = P(X = \sqrt{y})$. What if $\sqrt{y}$ is not an integer?

PMF of Discrete Random Variables (cont'd)

- **Example**: X is Poisson(λ) and $Y = X^2 - 7X + 12$. What is $P(Y = 0)$?

$$P(Y = 0) = P(X \in \{3, 4\}) = P(X = 3) + P(X = 4)$$

- For discrete X, to find $P(Y = y)$, find the set

$$A_y = \{x : g(x) = y, x \in \mathcal{X}\}.$$

- $P(Y = y) = P(X \in A_y)$, where A_y may be an empty set.

Transformations of Continuous Random Variables

- Enumeration does not work. Use either *cdf* or *Jacobian*.
- **Example**: Let $X \sim \text{Exp}(\lambda)$ and $Y = g(X) = X^{1/2}$. The cdf of X is $F_X(x) = 1 - e^{-x/\lambda}$, $x \geq 0$. The cdf of Y is

  $$F_Y(y) = P(Y \leq y) = P(X \leq y^2) = F_X(y^2) = 1 - e^{-y^2/\lambda},$$

  and the pdf is

  $$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{2y}{\lambda} e^{-y^2/\lambda}, \quad y \geq 0.$$

- $Y \sim$ Weibull $(2, \lambda)$ (C&B, page 627).

Transformations of Continuous RV (cont'd)

- **Example**: Let $X \sim N(0,1)$ and $Y = g(X) = X^2$. For $y > 0$, the cdf of Y is

$$F_Y(y) = P(Y \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y}),$$

and the pdf is

$$f_Y(y) = \{\phi(\sqrt{y}) + \phi(-\sqrt{y})\}\frac{d}{dy}\sqrt{y} = \frac{1}{\sqrt{2\pi y}}e^{-y/2}, \quad y > 0.$$

- $Y \sim \chi^2(1)$ (C&B, page 626).

Inverse Probability Integral Transform

- **Example**: Suppose that F is a *continuous* and *strictly increasing* cdf, and suppose that U is uniform on (0,1). The distribution of $Y = F^{-1}(U)$ is

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P(F^{-1}(U) \le y) = P(F(F^{-1}(U)) \le F(y)) \\ &= P(U \le F(y)) = F(y). \end{aligned}$$

- Useful in some computer simulation. For example, $X \sim \text{Exp}(1)$ with $F(x) = 1 - e^{-x}$ and $F^{-1}(u) = -log(1-u)$.
- One may generate U from $U(0,1)$ and get $-log(1-U)$ following Exp(1).
- May not work for a normal distribution since F^{-1} is not easy to compute.

Transformation Using Jacobian

- Suppose g is *monotone increasing*. That implies *one-to-one and onto* from $\mathcal{X}$ to $\mathcal{Y}$.
- Then g^{-1} is well-defined *monotone increasing* function. If $X = g^{-1}(Y)$, then

$$P(Y \leq y) = P(g^{-1}(Y) \leq g^{-1}(y)) = P(X \leq g^{-1}(y)).$$

- Hence $F_Y(y) = F_X(g^{-1}(y))$. The pdf of Y is

$$f_Y(y) = \frac{d}{dy} F_Y(y) = f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y).$$

- If g is *monotone decreasing*, then

$$f_Y(y) = -f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y).$$

Transformation Using Jacobian (cont'd)

- For monotone g (either increasing or decreasing),

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y))|\frac{d}{dy}g^{-1}(y)|, & y \in \mathcal{Y}, \\ 0, & otherwise. \end{cases}$$

- The factor $\frac{d}{dy}g^{-1}(y)$ is called *Jacobian* of g^{-1}.
- Only works for monotone g.
- Require $\frac{d}{dy}g^{-1}(y)$ be continuous on $\mathcal{Y}$.
- See Theorems 2.1.3 and 2.1.5 in C&B.

Transformation Using Jacobian (cont'd)

- **Example** Let $X \sim \text{Exp}(\lambda)$ and $Y = g(X) = X^{1/2}$. Note that $f_X(x) = \lambda^{-1} e^{-x/\lambda}$, $\mathcal{X} = [0, \infty)$.
- The function g is monotone on $\mathcal{X}$, $\mathcal{Y} = [0, \infty)$, and $g^{-1}(y) = y^2$ for $y \in \mathcal{Y}$.
- The derivative of $g^{-1}(y)$ is $2y$.
- The density of Y is

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| = \frac{2y}{\lambda} e^{-y^2/\lambda},$$

for $y \geq 0$ and $f_Y(y) = 0$ for $y < 0$.

Transformation Using Jacobian (cont'd)

- What if g^{-1} is not monotone, such as $g(x) = x^2$ on $\mathcal{X} = (-\infty, \infty)$?
- Take advantage of partition: monotone over $(-\infty, 0)$ and monotone over $(0, \infty)$.
- Apply the method of Jacobian to each piece and add up the contributions from all the pieces.

Transformation Using Jacobian (cont'd)

- **Theorem 2.1.8 in C&B**: Suppose that there exists a partition, $A_0, A_1, \ldots, A_k$ of $\mathcal{X}$ such that $P(X \in A_0) = 0$ and $f_X(x)$ is continuous on each A_i, $i = 1, \ldots, k$.
- Further suppose that the function g is monotone over each A_i.
- Let g_i denote the restriction of g to $x \in A_i$, $i > 0$, and suppose that

$$\mathcal{Y} = \{y : g_i(x) = y \textit{ for some } x \in A_i\}, \; 1 \leq i \leq k.$$

- Knowing that $g_i^{-1}(y)$ must be in A_i for $y \in \mathcal{Y}$, one can have

$$f_Y(y) = \begin{cases} \sum_{i=1}^{k} f_X(g_i^{-1}(y)) \left|\frac{d}{dy} g_i^{-1}(y)\right|, & y \in \mathcal{Y}, \\ 0, & \textit{otherwise}. \end{cases}$$

- A_0 is a set for interval endpoints which have zero probability.

Transformation Using Jacobian (cont'd)

- **Example**: Let $X \sim N(0,1)$ and $Y = g(X) = X^2$. Note that $\mathcal{X} = (-\infty, \infty)$ and $\mathcal{Y} = [0, \infty)$.
- We can take $A_1 = (-\infty, 0)$ and $g_1(x) = x^2$ on A_1 and $g_1^{-1}(y) = -\sqrt{y}$. We take $A_2 = (0, \infty)$ and $g_2(x) = x^2$ on A_2 and $g_2^{-1}(y) = \sqrt{y}$.
- The pdf of Y is

$$
\begin{aligned}
f_Y(y) &= \phi(-\sqrt{y})| - \frac{1}{2\sqrt{y}}| + \phi(\sqrt{y})\}|\frac{1}{2\sqrt{y}}| \\
&= \frac{1}{\sqrt{2\pi y}} e^{-y/2}, \quad y > 0.
\end{aligned}
$$

- The method of Jacobian does NOT apply to discrete random variables.

Bivariate and Multivariate Transformations

- X is Poisson(λ_1), Y is Poisson(λ_2), and $X \perp Y$. Let $U = X + Y$.
- Find $P(U = 3)$: The event $\{U = 3\}$ arises when $\{X = 0, Y = 3\}$, $\{X = 1, Y = 2\}$, $\{X = 2, Y = 1\}$, and $\{X = 3, Y = 0\}$.
- Since these four events are mutually exclusive,

$$\begin{aligned} P(U = 3) = {} & P(X = 0, Y = 3) + P(X = 1, Y = 2) \\ & + P(X = 2, Y = 1) + P(X = 3, Y = 0). \end{aligned}$$

- By independence, $P(X = x, Y = y) = P(X = x)P(Y = y)$. Then,

$$P(U = 3) = \sum_{x=0}^{3} P(X = x, Y = 3 - x) = \sum_{x=0}^{3} P(X = x)P(Y = 3 - x).$$

Method of Jacobian

- The method of Jacobian applies with a small adjustment.
- Suppose that the random vector (X, Y) has pdf $f_{X,Y}(x, y)$ and sample space $\mathcal{S}$.
- Consider the transformation of (X, Y) into (U, V) through

$$U = g_1(X, Y), \quad V = g_2(X, Y).$$

- We write $(U, V) = g(X, Y)$. It requires
  - (i) g is one-to-one on $\mathcal{S}$, so its inverse exists and is well-defined.
  - (ii) g has continuous partial derivatives on $\mathcal{S}$.
  - (iii) The Jacobian of g is not zero on $\mathcal{S}$.
- Let h denote the inverse function of g and $x = h_1(u, v)$ and $y = h_2(u, v)$. The density of (U, V) is given by

$$f_{U,V}(u, v) = f_{X,Y}(h_1(u, v), h_2(u, v))|J|.$$

Method of Jacobian (cont'd)

- J is the Jacobian of h; $|\cdot|$ is the *determinant* of the matrix of partial derivatives as

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}.$$

- What is the determinant of a 2×2 matrix?
- **Example** Suppose $X \sim \text{Gamma}(\alpha_1, 1)$, $Y \sim \text{Gamma}(\alpha_2, 1)$, and $X \perp Y$. Let $U = X + Y$ and $V = X/(X+Y)$. The joint pdf of X and Y is

$$f_{X,Y}(x,y) = \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)} e^{-x-y} x^{\alpha_1 - 1} y^{\alpha_2 - 1}, \quad x > 0, \quad y > 0.$$

- Let $(u,v) = g(x,y) = (x+y, x/(x+y))$ and its range is $\{(u,v); u > 0, 0 < v < 1\}$.

Method of Jacobian (cont'd)

- The inverse function is $h(u, v) = (uv, u - uv)$ and the Jacobian is

$$J = \begin{vmatrix} v & u \\ 1 - v & -u \end{vmatrix} = -uv - u(1 - v) = -u.$$

- The joint pdf of (U, V) is

$$f_{U,V}(u, v) = \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)} e^{-u}(uv)^{\alpha_1 - 1}(u - uv)^{\alpha_2 - 1}u, \;\; u > 0, \;\; 0 < v < 1$$

- We can write

$$f_{U,V}(u, v) = \frac{e^{-u}u^{\alpha_1 + \alpha_2 - 1}}{\Gamma(\alpha_1 + \alpha_2)} \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} v^{\alpha_1 - 1}(1 - v)^{\alpha_2 - 1}.$$

- We may claim $U \sim \text{Gamma}(\alpha_1 + \alpha_2, 1)$, $V \sim \text{Beta}(\alpha_1, \alpha_2)$, and $U \perp V$.

Method of CDF

- **Example** A random point in the unit disc has coordinates X and Y where (X, Y) has density

$$f_{X,Y}(x,y) = 1/\pi, \ \textit{for } (x,y) \in \mathcal{S},$$

where $\mathcal{S} = \{(x,y) : x^2 + y^2 < 1\}$. The length of the line from the origin to (X, Y) is

$$U = \sqrt{X^2 + Y^2} = g(X, Y).$$

- The cdf of U is

$$F_U(u) = P(U \leq u) = P(\sqrt{X^2 + Y^2} \leq u) = P(X^2 + Y^2 \leq u^2) = u^2.$$

- The pdf U is $f_U(u) = \frac{d}{du} F_U(u) = 2u$.
- How about the method of Jacobian?

Convolution Formula

- Suppose X and Y are independent continuous random variables with pdf f_X and f_Y. One way to find the density of $Z = X + Y$ is to introduce another variable W so that the transformation from (X, Y) to (Z, W) is one-to-one.
- Choose $W = X$. The inverse transformation, from (Z, W) to (X, Y), is $X = W$ and $Y = Z - W$. The Jacobian is -1.
- Then the density of (Z, W) is $f_{Z,W}(z, w) = f_X(w)f_Y(z - w)$.
- The density of Z is obtained by integrating out w,

$$f_Z(z) = \int_{-\infty}^{\infty} f_{Z,W}(z, w)dw = \int_{-\infty}^{\infty} f_X(w)f_Y(z - w)dw,$$

  which is called *convolution formula*.
- Be careful about the range of W.

Location-Scale Family

- Derivation can be simplified by shifting and scaling.
- Suppose that random variable Z has pdf f, and let $X = \mu + \sigma Z$ where $-\infty < \mu < \infty$ and $0 < \sigma < \infty$.
- Say, $X = g(Z)$ and $Z = g^{-1}(X) = (X - \mu)/\sigma$ with Jacobian $1/\sigma$. The density of X is
$$f_X(x) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right).$$
- Starting with a given density f, the set of distributions generated by all possible (μ, σ) is known as a location-scale family.
- **Example** If $Z \sim N(0,1)$ with density $f(z) = (1/\sqrt{2\pi})e^{-z^2/2}$, then $X = \mu + \sigma Z$ has density
$$f_X(x) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}.$$

Sums of Independent Random Variables

- Use moment generating function (mgf) method.
- Recall that $M_X(t) = Ee^{tX}$.
- **Example** $X \perp Y$ and $X, Y \sim N(0,1)$; $M_X(t) = M_Y(t) = e^{t^2/2}$. Let $U = aX + bY + c$.
- The mgf of U is

$$M_U(t) = Ee^{taX+tbY+tc} = e^{t^2a^2/2}e^{t^2b^2/2}e^{tc} = e^{ct+(a^2+b^2)t^2/2},$$

  which is mgf of $N(c, a^2+b^2)$. Therefore, $U \sim N(c, a^2+b^2)$.
- **Example** $X_1, \ldots X_n$ are mutually independent Bernoulli(θ) random variables. The mgf of each X_i is $M_{X_i}(t) = 1 - \theta + \theta e^t$. The mgf of $U = X_1 + \cdots + X_n$ is $M_U(t) = (1 - \theta + \theta e^t)^n$, which is the mgf of what distribution?

Random Samples

Feng-Chang Lin

Department of Biostatistics
University of North Carolina at Chapel Hill

flin@bios.unc.edu

(C&B §5.1-§5.3)

Introduction

- Statistical inferences are concerned with two entities: *population* and *sample*.
- A sample drawn from a population is used to make *inferences* about the population.
- In this section, we will be concerned with properties of random samples.

Random Sample

- **Example** Suppose a new drug has been developed for the treatment of hypertension. A sample of 50 hypertensive patients from the UNC Hospital is selected and treated by the new treatment.
- The primary outcome is the reduction in DBP after the treatment, which gives 50 numbers $x_1, x_2, \cdots, x_{50}$.
- *A sample of size* $n = 50$.
- Statistician would say: x_i is the *observed value*, or *realized value*, of a random variable X_i.
- If $X_1, X_2, \cdots, X_n$ are mutually independent with the same marginal pdf or pmf $f(x)$ (i.i.d.); $X_1, X_2, \cdots, X_n$ is called a random sample from the population $f(x)$.

Sampling from a Finite Population

- Sampling from a finite populations *with replacement* allows a unit to appear more than once in the sample.
- Sampling from a finite population *without replacement* allows a unit to appear at most once in the sample.
- Assume there are 25 balls in the urn, with 3 blacks and 22 reds.
- $X_1, \cdots, X_5$ ($n = 5$) is a random sample if drawn from a finite population of $N = 25$ *with replacement*.
- $X_1, \cdots, X_5$ is NOT a random sample if drawn *without replacement* because $P(X_2 = 1|X_1 = 1) = \frac{2}{24} \neq P(X_2 = 1) = \frac{3}{25}$, which implies

$$P(X_1 = 1, X_2 = 1) \neq P(X_1 = 1)P(X_2 = 1).$$

- What happen if N is very large?.

Statistics

- Let $X_1, \cdots, X_n$ be a random sample with $E(X_1) = \mu$ and $Var(X_1) = \sigma^2$.
- A statistic is denoted by $T(x_1, \cdots, x_n)$, which can be real-valued or vector-valued.
- **Example:** *Sample mean* $\bar{X}$ and *sample variance* S^2, which are defined by, respectively,

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$S^2 = \frac{(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

- The probability distribution of T is called the sampling distribution of T.

Computational Formula

$$
\begin{aligned}
(n-1)S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 &= \sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\
&= \sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2 \\
&= \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 = \sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2/n.
\end{aligned}
$$

- The last expression is sometimes described as a “computational formula" for S^2.

Sums of $X_1, \cdots, X_n$

- Sums are attractive mathematically because their means and variances can be calculated using simple rules, like

$$E\bar{X} = n^{-1}E(X_1 + \cdots + X_n) = n^{-1}nEX_1 = \mu,$$
$$Var\bar{X} = n^{-2}Var(X_1 + \cdots + X_n) = n^{-2}nVarX_1 = n^{-1}\sigma^2.$$

- For S^2, $E[(n-1)S^2] = E(\sum_{i=1}^n X_i^2 - n\bar{X}^2) = \sum_{i=1}^n EX_i^2 - nE\bar{X}^2$.
- We have

$$EX_i^2 = VarX_i + (EX_i)^2 = \sigma^2 + \mu^2,$$

and

$$E\bar{X}^2 = Var\bar{X} + (E\bar{X})^2 = n^{-1}\sigma^2 + \mu^2$$

- We get $E[(n-1)S^2] = (n-1)\sigma^2$ and $ES^2 = \sigma^2$.

Unbiased Estimator

- If $ET(X_1, \cdots, X_n) = \theta$, we say that T is an unbiased estimator of θ.
- **Example** If $EX_1 = \mu$ and $VarX_1 = \sigma^2$, then $\bar{X}$ is an unbiased estimator of μ, and S^2 is an unbiased estimator of σ^2.
- If one defines

$$T = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

  is T an unbiased estimator of σ^2?
- What happen if $n \to \infty$?

Samples from Normal Distribution

- If X has mgf $M_X(t)$, then $M_{\bar{X}}(t) = \{M_X(t/n)\}^n$.
- Suppose that $X_1, X_2, \cdots, X_n$ is a random sample from $N(\mu, \sigma^2)$.
- Then,

$$M_{\bar{X}}(t) = [\exp\{\mu t/n + \sigma^2 (t/n)^2/2\}]^n = \exp\{\mu t + (\sigma^2/n)t^2/2\}.$$

- Thus, $\bar{X} \sim N(\mu, \sigma^2/n)$.
- In some cases, the mgf of $\bar{X}$ may not correspond to any distribution we know, or the mgf of X may not exist (e.g. Cauchy).

Samples from Normal Distribution (cont'd)

- Suppose that $X_1, X_2, \cdots, X_n$ is a random sample from $N(\mu, \sigma^2)$.
- We know that $\bar{X} \sim N(\mu, \sigma^2/n)$, and that $ES^2 = \sigma^2$.
- If $\mu = 0$ and $\sigma = 1$, the density of $X_1, X_2, \cdots, X_n$ is

$$\prod_{i=1}^{n} \phi(x_i) = \prod_{i=1}^{n} \frac{e^{-x_i^2/2}}{\sqrt{2\pi}} = (2\pi)^{-n/2} e^{-\sum_{i=1}^{n} x_i^2/2}.$$

- Consider a transformation from $X_1, \cdots, X_n$ to $Y_1, \cdots, Y_n$, where $Y_1 = \bar{X}$ and $Y_i = X_i - \bar{X}$, $2 \le i \le n$, with inverse transformations $X_1 = Y_1 - \sum_{i=2}^{n} Y_i$ and $X_i = Y_i + Y_1$, $2 \le i \le n$.
- The joint density of $Y_1, \cdots, Y_n$ is

$$f_{Y_1,\cdots,Y_n}(y_1, \cdots, y_n) = n\phi(y_1 - \sum_{i=2}^{n} y_i) \prod_{i=2}^{n} \phi(y_i + y_1),$$

Samples from Normal Distribution (cont'd)

which can be expressed as

$$\left\{ \frac{1}{\sqrt{2\pi(1/n)}} e^{-y_1^2/(2/n)} \right\} \left\{ \frac{n^{1/2}}{(2\pi)^{(n-1)/2}} e^{-c/2} \right\},$$

where $c = \sum_{i=2}^n y_i^2 + (\sum_{i=2}^n y_i)^2$.

- This implies: $Y_1 = \bar{X}$ is independent of $Y_2, \cdots, Y_n$.
- Since

$$\begin{aligned} S^2 &= \frac{1}{n-1} \left\{ (X_1 - \bar{X})^2 + \sum_{i=2}^n (X_i - \bar{X})^2 \right\} \\ &= \frac{1}{n-1} \left[\left\{ -\sum_{i=2}^n (X_i - \bar{X}) \right\}^2 + \sum_{i=2}^n (X_i - \bar{X})^2 \right], \end{aligned}$$

one can claim S^2 is a function of $Y_2, \cdots, Y_n$.

Distribution of S^2

- This tells you $\bar{X} \perp S^2$.
- What is the distribution of S^2? Consider $n = 2$. In this case,

$$S^2 = \left(X_1 - \frac{X1 + X2}{2}\right)^2 + \left(X_2 - \frac{X1 + X2}{2}\right)^2 = \left(\frac{X_1}{\sqrt{2}} - \frac{X_2}{\sqrt{2}}\right)^2$$

- Since $X_1 \perp X_2$, $\frac{X_1}{\sqrt{2}} - \frac{X_2}{\sqrt{2}} \sim N(0, 1)$ and $S^2 \sim \chi_1^2$.
- How about $n = k$?
- Let $\bar{X}_k$ and S_k^2 denote the sample mean and sample variance, respectively.
- A method of **induction** will be shown to prove that S_{k+1}^2 follows χ_k^2, assuming S_k^2 follows χ_{k-1}^2.

Distribution of S^2 (cont'd)

- One can show

$$\bar{X}_{k+1} = \frac{k\bar{X}_k + X_{k+1}}{k+1}$$

$$kS^2_{k+1} = (k-1)S^2_k + \frac{k}{k+1}(X_{k+1} - \bar{X}_k)^2,$$

- The second equation is proved in the following page.

Distribution of S^2 (cont'd)

$$
\begin{aligned}
kS_{k+1}^2 &= \sum_{i=1}^{k+1} X_i^2 - (k+1)\bar{X}_{k+1}^2 = \sum_{i=1}^{k+1} X_i^2 - (k+1)\left(\frac{X_{k+1}+k\bar{X}_k}{k+1}\right)^2 \\
&= \sum_{i=1}^{k+1} X_i^2 - \frac{1}{k+1}(X_{k+1}^2 + 2kX_{k+1}\bar{X}_k + k^2\bar{X}_k^2) \\
&= \sum_{i=1}^{k} X_i^2 - k\bar{X}_k^2 + X_{k+1}^2 + k\bar{X}_k^2 - \frac{1}{k+1}(X_{k+1}^2 + 2kX_{k+1}\bar{X}_k + k^2\bar{X}_k^2) \\
&= (k-1)S_k^2 + \frac{k}{k+1}(X_{k+1}^2 + 2X_{k+1}\bar{X}_k + \bar{X}_k^2) \\
&= (k-1)S_k^2 + \frac{k}{k+1}(X_{k+1} + \bar{X}_k)^2
\end{aligned}
$$

Distribution of S^2 (cont'd)

- The distribution of $kS_{k+1}^2 = (k-1)S_k^2 + \frac{k}{k+1}(X_{k+1} - \bar{X}_k)^2$ is derived as follows:

(1) First, we have known that S_k^2, X_{k+1}, $\bar{X}_k$ are independent.

(2) Since $X_{k+1} - \bar{X}_k \sim N(0, 1 + 1/k)$, $\frac{k}{k+1}(X_{k+1} - \bar{X}_k)^2 \sim \chi_1^2$.

(3) Assuming the statement $(k-1)S_k^2 \sim \chi_{k-1}^2$ is true, one shall show that kS_{k+1}^2 follows χ_k^2.

(4) By induction, we need to show when $k = 2$, S_2^2 follows χ_1^2, which was proved in the previous page.

Extension to $X_i \sim N(\mu, \sigma^2)$

- What if $X_1, \cdots, X_n$ from $N(\mu, \sigma^2)$?
- Define $Z_i = (X_i - \mu)/\sigma$ and let S_X^2 and S_Z^2 denote sample variance of X and Z, respectively.
- We know that $Z_i \sim N(0, 1)$, $i = 1, \cdots, n$.
- Also, $\bar{Z} = (\bar{X} - \mu)/\sigma$, and $S_Z^2 = S_X^2/\sigma^2$.
- Therefore, $((\bar{X} - \mu)/\sigma, S_X^2/\sigma^2)$ has the same distribution as $(\bar{Z}, S_Z^2)$.
- One has

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

$$\frac{S_X^2}{\sigma^2} \sim \frac{\chi^2_{n-1}}{n - 1},$$

and $\bar{X}$ and S_X^2 are independent.

More Transformations

- χ^2_{n-1} distribution has mean $n-1$ and variance $2(n-1)$.
- We have $E(S_X^2) = \sigma^2$ and $Var(S_X^2) = 2\sigma^4/(n-1)$ since

$$Var\{(n-1)S_X^2/\sigma^2\} = 2(n-1).$$

- A test statistic $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ can't be computed if σ unknown.
- If σ unknown, look for the distribution of $\frac{\bar{X}-\mu}{S/\sqrt{n}}$.
- What is the distribution of

$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}}?$$

Student's t Distribution

- If $U \sim N(0,1)$, $V \sim \chi^2_p$ and U and V are independent, then the distribution of $T = U/\sqrt{V/p}$ known as *Student's t distribution with p degrees of freedom*, abbreviated as t_p, with density

$$f_T(t) = \frac{\Gamma(\frac{p+1}{2})}{\Gamma(\frac{p}{2})}(p\pi)^{-1/2}\left(1+\frac{t^2}{p}\right)^{-(p+1)/2}, \quad t \in (-\infty, \infty)$$

- Since U and V are independent, the joint density of (U, V) is

$$f_{U,V}(u,v) = f_U(u)f_V(v) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2}\frac{1}{\Gamma(p/2)2^{p/2}}v^{p/2-1}e^{-v/2}.$$

- The transformation from (u, v) to $t = \frac{u}{\sqrt{v/p}}$ and $w = v$.

Student's t Distribution (cont'd)

- The inverse is $u = t\sqrt{w/p}$ and $v = w$ with Jacobian $\sqrt{w/p}$.
- The joint density of (T, W) is then

$$f_{T,W} = f_{U,V}(t\sqrt{w/p}, w)\sqrt{w/p}.$$

- The marginal density of T is obtained by integrating out w.
- The t_p density is symmetric about 0.
- It does not have an mgf. In fact, only the first $p - 1$ moments exist.
- The mean is 0 if $p > 1$, and the variance is $p/(p - 2)$ if $p > 2$.
- The case $p = 1$ is Cauchy distribution ($\Gamma(\frac{1}{2}) = \sqrt{\pi}$).

Student's t Distribution (cont'd)

- If $X_1, \cdots, X_n$ is a random sample from the $N(\mu, \sigma^2)$ density, and we define $U = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ and $V = (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.
- $U \sim N(0,1)$, $V \sim \chi^2_{n-1}$, and $U \perp V$. That shows

$$T = \frac{U}{\sqrt{V/(n-1)}} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

- 95% CI: $\bar{x} \pm t_{n-1,1-\alpha/2} s/\sqrt{n}$.
- How about 95% CI for σ^2? We will talk about pivotal quantity in the future.

F Distribution

- If $U \sim \chi^2_p$, $V \sim \chi^2_q$, and $U \perp V$. The distribution of $X = (U/p)/(V/q)$, which is known as *Snedecor's F distribution with p and q degrees of freedom*, abbreviated as $F_{p,q}$.
- This distribution arises in the study of ratios of sample variances. Such ratios arise in the analysis of variance (ANOVA) and in regression analysis.
- What is the distribution of $1/X$?

Other Properties of Normal Variates

- If X has a normal distribution and Y has a normal distribution, then X and Y are independent if and only if $Cov(X, Y) = 0$.
- $\bar{X}$ is normal, $X_i - \bar{X}$ is normal, and

$$Cov(\bar{X}, X_i - \bar{X}) = Cov(\bar{X}, X_i) - Cov(\bar{X}, \bar{X}) = \sigma^2/n - \sigma^2/n = 0.$$

- We can conclude $\bar{X}$ and $X_i - \bar{X}$ are independent, for $1 \leq i \leq n$.
- That can help show $\bar{X}$ is independent of $X_i - \bar{X}$ (check the notes).
- The “zero covariance implies independence" property generally does not apply to other distributions.
- For example, if $X \sim N(0, 1)$ and $Y = X^2 \sim \chi_1^2$, then clearly X and Y are not independent. However,

$$Cov(X, Y) = Cov(X, X^2) = EX^3 - (EX)(EX^2) = 0 - (0)(1) = 0.$$

Order Statistics

Feng-Chang Lin

Department of Biostatistics
University of North Carolina at Chapel Hill

flin@bios.unc.edu

(C&B §5.4)

Introduction

- A useful statistic of a random sample is to order the sample values in ascending order.
- This is called order statistics, denoted by $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$, distinguishing from the original values $x_1, x_2, \cdots, x_n$.
- The *sample minimum*, $x_{(1)}$, and the *sample maximum*, $x_{(n)}$, are also order statistics.
- The *sample median* is the middle order statistic, $x_{(m+1)}$, if $n = 2m + 1$ (n is odd).
- If n is even, the sample median is usually taken to be the average of the two middle order statistics, $(x_{(n/2)} + x_{(n/2+1)})/2$.

Introduction (cont'd)

- The *sample range*, $R = x_{(n)} - x_{(1)}$, is the distance between the smallest and largest observations.
- $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ are not independent since

$$X_{(1)} < X_{(2)} < \cdots < X_{(n)}.$$

- They are not identically distributed as well since

$$EX_{(1)} < EX_{(2)} < \cdots < EX_{(n)}.$$

Sample Maximum

- The distribution of the sample maximum can be easily derived since

$$\{X_{(n)} \le x\} = \{X_1 \le x, \cdots, X_n \le x\}$$

- This implies

$$F_{X_{(n)}}(x) = P(X_{(n)} \le x) = P(X_1 \le x, \cdots, X_n \le x) = \{F(x)\}^n$$

- If X is continuous,

$$f_{X_{(n)}}(x) = \frac{d}{dx} F_{X_{(n)}}(x) = nf(x)\{F(x)\}^{n-1}.$$

- **Example** If f is the uniform(0,1) pdf, then

$$f_{X_{(n)}}(x) = nx^{n-1}, \quad x \in (0, 1).$$

Sample Minimum

- Similarly,
$$\{X_{(1)} > x\} = \{X_1 > x, \cdots, X_n > x\}$$
- This implies
$$F_{X_{(1)}}(x) = P(X_{(1)} \leq x) = 1 - P(X_{(1)} > x) = 1 - \prod_{i=1}^{n} P(X_i > x) = 1 - \{1 -$$
- If X is continuous,
$$f_{X_{(1)}}(x) = \frac{d}{dx} F_{X_{(1)}}(x) = nf(x)\{1 - F(x)\}^{n-1}.$$
- **Example** If f is the $\exp(\beta)$ pdf, then
$$f_{X_{(1)}}(x) = n\beta^{-1} e^{-x/\beta}\{1 - 1 + e^{-x/\beta}\}^{n-1} = (\beta/n)^{-1} e^{-x/(\beta/n)}.$$

Joint Distribution of Order Statistics

- The vector of order statistics is a function of the sample values, $(x_{(1)}, \ldots, x_{(n)}) = g(x_1, \cdots, x_n)$.
- The inverse transformation, from order statistics to sample values, does not exist (not 1-to-1).
- What did we learn from "not 1-to-1" previously? Partition!!!
- Restrict the sample to, for example, the set

$$\{(x_1, x_2, x_3) : x_2 < x_3 < x_1\}.$$

  We would be able to compute the inverse of $(x_{(1)} = 2, x_{(2)} = 5, x_{(3)} = 9)$ as $(x_1 = 9, x_2 = 2, x_3 = 5)$.
- How may such sets? It's $3! = 6$.

Joint Distribution of Order Statistics (cont'd)

- Keep in mind that the order statistics are a permutation of the sample values.
- Partition: $A_1, \cdots, A_{n!}$. Let g_i be the transformation on A_i and g^{-1} be its inverse.
- Each row and column of Jacobian matrix (or called *permutation matrix* here) consists of 1 one and $n-1$ zeros, so $|J| = 1$.
- The joint pdf of the order statistics is

$$f_{X_{(1)},\cdots,X_{(n)}}(y_1, \cdots, y_n) = \sum_{j=1}^{n!} f_{X_1,\cdots,X_n}(g_j^{-1}(y_1, \cdots, y_n)) = n! \prod_{i=1}^{n} f_X(y_i),$$

for $y_1 < \cdots < y_n$.

Distribution of $X_{(j)}$

- $\{X_{(j)} \leq x\} = \{$at least j of the sample vales are $\leq x\}$.
- If $Z_i = I(X_i \leq x)$ and $Y_i = \sum_{i=1}^{n} Z_i$, then $\{X_{(j)} \leq x\} = \{Y \geq j\}$.
- Let $A = F(x)$ and $a = f(x)$. We have

$$\begin{aligned} F_{X_{(j)}}(x) &= P(X_{(j)} \leq x) = P(Y \geq j) \\ &= \sum_{k=j}^{n} P(Y = k) = \sum_{k=j}^{n} \binom{n}{k} A^k (1-A)^{n-k}. \end{aligned}$$

- The pdf of $X_{(j)}$ is

$$\begin{aligned} f_{X_{(j)}}(x) &= \frac{d}{dx} F_{X_{(j)}}(x) \\ &= \sum_{k=j}^{n} \binom{n}{k} kaA^{k-1}(1-A)^{n-k} - \sum_{k=j}^{n} \binom{n}{k} A^k (n-k)a(1-A \\ &= C - D \end{aligned}$$

Distribution of $X_{(j)}$ (cont'd)

- C can be expressed by $C = C_1 + C_2$, where

$$
\begin{aligned}
C_1 &= \binom{n}{j} jaA^{j-1}(1-A)^{n-j} \\
C_2 &= \sum_{k=j+1}^{n} \binom{n}{k} kaA^{k-1}(1-A)^{n-k} \\
&= \sum_{t=j}^{n-1} \binom{n}{t+1} (t+1)aA^{t}(1-A)^{n-t-1} \\
&= \sum_{t=j}^{n-1} \binom{n}{t} (n-t)aA^{t}(1-A)^{n-t-1}.
\end{aligned}
$$

- One can show $C_2 = D$ since the last term in D ($j = n$) is 0.
- $f_{X_{(j)}}(x) = C_1 = \binom{n}{j} jaA^{j-1}(1-A)^{n-j}$

Distribution of $X_{(j)}$ (cont'd)

$$
\begin{aligned}
f_{X_{(j)}}(x) = C_1 &= \begin{pmatrix} n \\ j \end{pmatrix} jaA^{j-1}(1-A)^{n-j} \\
&= \frac{n!}{(j-1)!(n-j)!} f(x)\{F(x)\}^{j-1}\{1-F(x)\}^{n-j}
\end{aligned}
$$

- Intuitive interpretation: $(j-1)$ observations are on the left of $X_{(j)}$, contributing $\{F(x)\}^{j-1}$, $X(j)$ itself, contributing $f(x)$, and $(n-j)$ observations are on the right of $X_{(j)}$, contributing $\{1-F(x)\}^{n-j}$.
- The combinatorial factor is the number of ways in which n observations can be grouped into three sets containing $j-1$, 1, and $n-j$ observations.

Distribution of $X_{(j)}$ (cont'd)

- **Example** Suppose that $X_1, \cdots, X_n$ are iid from the uniform density on $(0, 1)$. Then for $1 \le j \le n$,

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} x^{j-1}(1-x)^{n-j}$$
$$= \frac{\Gamma(n+1)}{\Gamma(j)\Gamma(n-j+1)} x^{j-1}(1-x)^{n-j}, \quad x \in (0, 1)$$

- This is the pdf of $Beta(j, n-j+1)$ with $EX_{(j)} = \frac{j}{n+1}$ and $VarX_{(j)} = \frac{j(n-j+1)}{(n+1)^2(n+2)}$.
- If $n = 2m+1$ (n is odd), it follows that the sample median, $X_{(m+1)}$, has a $Beta(m+1, m+1)$ density with mean 1/2 and variance $1/\{4(n+2)\}$.
- The expected value of sample mean is 1/2 and variance $1/(12n)$.

Distribution of $(X_{(i)}, X_{(j)})$

- This follows the same lines as the derivation of $f_{X_{(j)}}$.
- The joint distribution of $(X_{(i)}, X_{(j)})$ is

$$f_{X_{(i)},X_{(j)}}(u, v) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} f(u)f(v) \\ \times F(u)^{i-1}\{F(v) - F(u)\}^{j-i-1}\{1 - F(v)\}^{n-j}$$

- **Example** Suppose that $X_1, \cdots, X_n$ are iid from the uniform density on $(0, a)$, $a > 0$. For $0 < x < y < a$,

$$f_{X_{(1)},X_{(n)}}(x, y) = \frac{n(n-1)(y-x)^{n-2}}{a^n}.$$

- One may be interested in the distribution of the range variable $R = X_{(n)} - X_{(1)}$ and midrange variable $V = (X_{(n)} + X_{(1)})/2$,

Distribution of $(X_{(i)}, X_{(j)})$ (cont'd)

- One has $X_{(n)} = V + R/2$, $X_{(1)} = V - R/2$, and $|J| = 1$. The joint pdf of (R, V) is

$$f_{R,V}(r, v) = f_{X_{(1)},X_{(n)}}(v + r/2, v - r/2) = \frac{n(n-1)r^{n-2}}{a^n},$$

  for $0 < r < a$ and $r/2 < v < a - r/2$ since $0 < x_{(1)} < x_{(n)} < a$.
- The support region of (R, V) is a triangle.
- The marginal pdf of R can be obtained as

$$f_R(r) = \int_{r/2}^{a-r/2} f_{R,V}(r, v)dv = \frac{n(n-1)r^{n-2}(a-r)}{a^n}, \quad 0 < r < a.$$

Distribution of $(X_{(i)}, X_{(j)})$ (cont'd)

- If $Z = R/a$, then $Z \sim Beta(n-1, 2)$ since

$$f_Z(z) = n(n-1)z^{n-2}(1-z)$$
$$= \frac{1}{B(n-1,2)} z^{n-2}(1-z), \quad z \in (0,1).$$

Convergence Concepts

Feng-Chang Lin

Department of Biostatistics
University of North Carolina at Chapel Hill

flin@bios.unc.edu

(C&B §5.5)

Introduction

- For random samples from the normal distribution, we derived the *exact* joint distribution of $(\bar{X}, S^2)$.
- For other distributions, the *exact* distribution may be too complicated to be of practical use.
- Instead, *approximate* distribution may be easier to derive or be computed.
- In this section, we will study the behavior of sample statistics in large samples, or say, $n \to \infty$.
- The term *large sample theory* or *asymptotic theory* refers to this approach.

Two Basic Tools

- Law of large numbers (LLN) and central limit theorem (CLT)
- That says, loosely, when sample size is large, the sample mean is close to the population mean (LLN) and the sample mean is approximately normally distributed (CLT).
- We will need Taylor's expansion from calculus, Slutsky's theorem, and delta method.
- All these tools rely on the mathematical notion of convergence.

Convergent Non-Random Sequences

- Sequences will be denoted by either $a_1, a_2, \cdots$ or by $\{a_n\}$.
- A sequence $\{a_n\}$ of real numbers is said to *converge* if there is a point a with the following property:
- For every $\epsilon > 0$, there is an integer N such that $n \geq N$ implies that $|a_n - a| < \epsilon$.
- In this case we say that $\{a_n\}$ converges to a, or that a is the limit of $\{a_n\}$, and we write $\lim_{n\to\infty} a_n = a$, or $a_n \to a$ as $n \to \infty$.
- If $\{a_n\}$ does not converge, it is said to *diverge*.
- The above definitions apply as well to sequences in R^k (finite k), with $|\cdot|$ replaced by Euclidean distance $\|\cdot\|$.

Convergent Random Sequences

- Does a sequence $\{X_n\}$ of random variables converge to a limit random variable X?
- Is there a meaningful way to say that "$X_n \to X$ as $n \to \infty$"?
- Remember $\{X_n\}$ is a "random" sequence, so whether $\{X_n\}$ converges to X or not is a "random" event.
- That means, some sequences converge, others do not.
- Since $\{X_n \to X \;\; as \;\; n \to \infty\}$ is a random event, we can put

$$P(X_n \to X \;\; as \;\; n \to \infty) = 1.$$

- We claim "the sequence of random variables $X_1, X_2, \cdots,$ *converges almost surely* to a random variable X.
- Written as $P(\lim_{n\to\infty} |X_n - X| = 0) = 1$.

Converge Almost Surely

- Recall, a random variable is a real-value function defined on the sample space S.
- One may also write almost sure convergence as

$$P(\{s : \lim_{n\to\infty} |X_n(s) - X(s)| = 0\}) = 1$$

- Notation: $X_n \to_{a.s.} X$ as $n \to \infty$.
- Almost sure convergence means that $X_n(s) - X(s)$ for all $s \in S$, except possibly for a subset of S that has zero probability.
- **Example** S is uniform on $[0, 1]$, and define $X_n(s) = s + s^n$. For every $s \in [0, 1)$, $X_n(s) \to s$. But for $s = 1$, $s^n \to 1$, and $X_n(1) \to 2 \neq 1$.
- One can still claim $X_n \to_{a.s.} s = X(s)$ as $n \to \infty$ since $P(S = 1) = 0$.

Strong Law of Large Numbers (SLLN)

- Let $X_1, \cdots, X_n$ be iid random variables with $EX_i = \mu$ and $VarX_i = \sigma^2 < \infty$. Let $\bar{X}_n = \sum_{i=1}^n X_i/n$. Then, for every $\epsilon > 0$,

$$P(\lim_{n\to\infty} |\bar{X}_n - \mu| < \epsilon) = 1.$$

- That is, $\bar{X}_n$ converges almost surely to μ.
- The property $\bar{X}_n \rightarrow_{a.s.} \mu$ is called *strong consistency* of $\bar{X}_n$ as an estimator of μ.
- One may also say that $\bar{X}_n$ is a *strongly consistent estimator* of μ.

Converge in Probability

- A weaker form of convergence.
- A sequence of random variables X_n converges in probability to a random variable X if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1.$$

- One may say, for $\epsilon > 0$, define $a_n(\epsilon) = P(|X_n - X| < \epsilon)$.
- Convergence in probability means that $a_n(\epsilon) \to 1$ as $n \to \infty$, for every $\epsilon > 0$.
- Notation: $X_n \to_p X$ as $n \to \infty$.
- **Convergence in probability, not almost surely** see example 5.5.8 in C&B.

Weak Law of Large Numbers (WLLN)

- Let $X_1, \cdots, X_n$ be iid random variables with $EX_i = \mu$ and $VarX_i = \sigma^2 < \infty$. Let $\bar{X}_n = \sum_{i=1}^{n} X_i/n$. Then, for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \epsilon) = 1.$$

- That is, $\bar{X}_n$ converges in probability to μ.
- The property $\bar{X}_n \to_p \mu$ is called *consistency* of $\bar{X}_n$.
- Comment: The condition that EX_i exists and is finite is *sufficient* in both WLLN and SLLN.

Converge in Distribution

- Let F_{X_n} be the cdf of X_n.
- A sequence of random variables X_n converges in distribution to a random variable X if,

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

  at all points x where $F_X(x)$ is continuous.
- Notation: $X_n \to_d X$ as $n \to \infty$.
- Convergence in distribution does not imply that X_n and X approximate each other.
- It only says that, for large n, the cdf of X_n becomes close to the cdf of X.

Central Limit Theorem (CLT)

- Let $X_1, \cdots, X_n$ be iid random variables with $EX_i = \mu$ and $VarX_i = \sigma^2 < \infty$. Define $\bar{X}_n = \sum_{i=1}^n X_i/n$, $Z_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma$, and let G_n denote the cdf of Z_n. For any $-\infty < z < \infty$,

$$lim_{n\to\infty} G_n(z) = \Phi(z).$$

- That is, Z_n has a limiting standard normal distribution, $Z_n \to_d N(0,1)$ as $n \to \infty$.

Central Limit Theorem (cont'd)

- **Example** Suppose that $X_1, \cdots, X_n$ are iid *Bernoulli*(p), and define $Y = \sum_{i=1}^{n} X_i$. The CLT states that $Z_n = \sqrt{n}(\bar{X}_n - p)/\sqrt{p(1-p)}$ is approximately $N(0, 1)$ for large n.
- Since $Z_n = (Y - np)/\sqrt{np(1-p)}$, that shows one can use a normal approximation to the binomial distribution of Y.
- Suppose $n = 100$ and $p = 0.5$. One can calculate $P(Y \leq 57) = 0.933$, which is close to $\Phi(z) = \Phi(1.4) = 0.919$.

Relationships between Modes of Convergence

- $X_n \to_{a.s.} X \Rightarrow X_n \to_p X \Rightarrow X_n \to_d X$.
- The converse statements are "generally" not true.
- **Example** A special case for $X_n \to_d X \Rightarrow X_n \to_p X$: If c is a non-random constant, $P(X = c) = 1$ then $X_n \to_d X$ implies that $X_n \to_p c$ (proofs in C&B).
- That is, convergence in distribution to a degenerate one-point distribution implies convergence in probability.

Slutsky's Theorem

- If $X_n \to_d X$ and $Y_n \to_p a$, where a is a finite constant, then
  1. $Y_n X_n \to_d aX$;
  2. $Y_n + X_n \to_d a + X$;
  3. $X_n / Y_n \to_d X/a$ if $a \neq 0$;
- Slutsky's theorem allows substituting consistent estimators when proving convergence in distribution.
- X_n and Y_n need not be independent.
- **Example** Suppose that $X \sim N(0, \sigma^2)$ and $T_n \to_d X$ as $n \to \infty$. By Slutsky's theorem, $T_n/\sigma \to_d X/\sigma$. Since $X/\sigma \sim N(0, 1)$, we conclude that $T_n/\sigma \to_d N(0, 1)$.

Convergence of Transformed Sequences

- Suppose that h is a continuous function.
- One has
  1. If $X_n \to_{a.s.} X$ then $h(X_n) \to_{a.s.} h(X)$.
  2. If $X_n \to_p X$ then $h(X_n) \to_p h(X)$.
  3. If $X_n \to_d X$ then $h(X_n) \to_d h(X)$.
- h needs be continuous only on the range of X. For example, if X is non-negative, the behavior of $h(x)$ for $x < 0$ does not matter.
- **Example** Let $X_1, \cdots, X_n$ be iid random variables with mean μ and variance $\sigma^2 < \infty$. Does the sample variance S_n^2 converge to σ^2 in some sense? Write

$$S_n^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2\right\} = \frac{n}{n-1}\frac{\sum_{i=1}^{n} X_i^2}{n}. - \frac{n}{n-1}\bar{X}_n^2$$

Convergence of Transformed Sequences (cont'd)

- As $n \to \infty$, $n/(n-1) \to 1$,

$$\frac{\sum_{i=1}^{n} X_i^2}{n} \to_{a.s.} EX_1^2 = \mu^2 + \sigma^2,$$

and

$$\bar{X}_n^2 \to_{a.s.} \mu^2.$$

- Slutsky's theorem and convergence of transformed random sequences lead to the result that $S_n^2 \to_{a.s.} \sigma^2$ as $n \to \infty$.
- **Example** Suppose that $\{T_n\}$ is a random sequence with $\sqrt{n}(T_n - \theta) \to_d N(0, \sigma^2)$. The asymptotic distribution of T_n is centered about θ. But, does T_n converge to θ in some sense? That is, is $T_n \to_p \theta$?

Convergence of Transformed Sequences (cont'd)

- Let $Z_n = \sqrt{n}(T_n - \theta)/\sigma \to_d N(0,1)$
- Given $\epsilon > 0$, one has

$$
\begin{aligned}
P(|T_n - \theta| < \epsilon) &= P(-\sqrt{n}\epsilon/\sigma < Z_n < \sqrt{n}\epsilon/\sigma) \\
&< P(-\sqrt{n}\epsilon/\sigma < Z_n \leq \sqrt{n}\epsilon/\sigma) \\
&= P(Z_n \leq \sqrt{n}\epsilon/\sigma) - P(Z_n \leq -\sqrt{n}\epsilon/\sigma).
\end{aligned}
$$

- Since Z_n converges in distribution, $P(Z_n \leq \sqrt{n}\epsilon/\sigma) \to 1$ and $P(Z_n \leq -\sqrt{n}\epsilon/\sigma) \to 0$.
- Hence $P(|T_n - \theta| < \epsilon) \to 1$ as $n \to \infty$. That means, $T_n \to_p \theta$.

Delta Method - Univariate

- Suppose that $\{T_n\}$ is a random sequence with $\sqrt{n}(T_n - \theta) \to_d N(0, \sigma^2)$, and g is a function with $g'(\theta)$ exists and is not 0. Then

$$\sqrt{n}\{g(T_n) - g(\theta)\} \to_d N(0, \{g'(\theta)\}^2\sigma^2).$$

- We say that θ is the *asymptotic mean* of T_n. However θ may or may not be the mean of T_n. In fact, the mean of T_n may not even exist (example below).
- **Example** Suppose that $X_1, \cdots, X_n$ are iid Bernoulli(θ), $0 < \theta < 1$, and we want to make statistical inferences about the log-odds, which is defined by

$$\psi = \log\left(\frac{\theta}{1-\theta}\right).$$

Delta Method - Univariate (cont'd)

- Define $g(u) = \log\{u/(1-u)\}$ for $u \in (0,1)$, so $\psi = g(\theta)$.
- By SLLN, $\bar{X}_n \to_{a.s.} \theta$. Since g is continuous at $\theta \in (0,1)$, one has that $g(X_n) \to_{a.s.} g(\theta)$.
- Since $g'(\theta) = 1/\{\theta(1-\theta)\} \neq 0$ for $\theta \in (0,1)$, the delta method gives
$$\sqrt{n}(g(\bar{X}_n) - g(\theta)) \to_d N(0, \{g'(\theta)\}^2\theta(1-\theta)),$$
or, equivalently,
$$\sqrt{n}(g(\bar{X}_n) - \psi) \to_d N\left(0, \frac{1}{\theta(1-\theta)}\right).$$
- The *asymptotic mean* of $g(\bar{X}_n)$ is ψ.
- The *exact mean* $Eg(\bar{X}_n)$ does not exist because $g(0) = -\infty$, $P(\bar{X}_n = 0) > 0$, $g(1) = \infty$, $P(\bar{X}_n = 1) > 0$, $Eg(\bar{X}_n) = \infty - \infty$.

Delta Method - Univariate (cont'd)

- Can the distribution above be used in practice? Why?
- We know that if a random variable Z follows a $N(0, 1/\{\theta(1-\theta)\})$, then $\sqrt{\theta(1-\theta)}Z$ follows a $N(0, 1)$.
- Is the following statement true?

$$\sqrt{\theta(1-\theta)}\sqrt{n}(g(\bar{X}_n) - \psi) \to_d N(0, 1),$$

and

$$\sqrt{\bar{X}(1-\bar{X})}\sqrt{n}(g(\bar{X}_n) - \psi) \to_d N(0, 1).$$

- To construct a 95% CI for log-odds ψ, which one to use?

Second-order Delta Method

- Suppose that T_n is a random sequence with $\sqrt{n}(T_n - \theta) \rightarrow_d N(0, \sigma^2)$, and g is a function with $g'(\theta) = 0$ and $g''(\theta)$ exists and is not 0. Then

$$n\{g(T_n) - g(\theta)\} \rightarrow_d \sigma^2 \frac{g''(\theta)}{2} \chi_1^2$$

- **Example** $g(T_n) = \bar{X}_n(1 - \bar{X}_n)$, $g(\theta) = \theta(1 - \theta)$, $g'(\theta) = 1 - 2\theta$, $g''(\theta) = -2$. If $\theta = 1/2$, one can have

$$n\left\{\bar{X}_n(1 - \bar{X}_n) - \frac{1}{4}\right\} \rightarrow_d -\frac{1}{4}\chi_1^2.$$

Delta Method - multivariate

- Let the p-dimensional random vectors $X_1, \cdots, X_n$ be a random sample with $EX_{ij} = \mu_j$ ($j = 1, \cdots, p$) and $Cov(X_{ij}, X_{ik}) = \sigma_{jk}^2$.
- The population mean vector will be denoted by $\mu = (\mu_1, \cdots, \mu_p)$.
- If a function g maps R^p into R and has continuous first partial derivatives, $\partial g(t)/\partial t_j$, then

$$\sqrt{n}\{g(\bar{X}_1, \cdots, \bar{X}_p) - g(\mu_1, \cdots, \mu_p)\} \rightarrow_d N(0, \tau^2),$$

where

$$\tau^2 = \sum_{j=1}^{p} \sum_{k=1}^{p} \sigma_{jk}^2 \frac{\partial g(\mu)}{\partial \mu_j} \frac{\partial g(\mu)}{\partial \mu_k},$$

provided that $\tau^2 > 0$.

Pair-Matched Case-Control Study

- A case (i.e., a diseased person, denoted D) is ”matched" (on covariates such as age, race, and sex) to a control (i.e., non-diseased person, denoted $\bar{D}$).
- Each member of the pairs is then interviewed as to the presence (E) or absence ($\bar{E}$) of a history of exposure to some harmful substance (e.g., cigarette smoke, asbestos, benzene, etc.)
- The data from such study involving n case-control pairs can be presented in tabular form as follows:

$$
\begin{array}{cc|cc|c}
 & & \multicolumn{2}{c}{\bar{D}} & \\
 & & E & \bar{E} & \\
\hline
\multirow{2}{*}{D} & E & Y_{11} & Y_{10} & \\
 & \bar{E} & Y_{01} & Y_{00} & \\
\hline
 & & & & n
\end{array}
$$

Pair-Matched Case-Control Study (cont'd)

- Y_{11} is the number of pairs where *both* case *and* control are exposed (i.e., both have a history of exposure).
- Y_{10} is the number of pairs where case is exposed but the control is not, and so on.
- Clearly $\sum_{j=0}^{1}\sum_{k=0}^{1} Y_{jk} = n$.
- Assume that $\{Y_{ij}\}$ have a multinomial distribution with sample size n and associated cell probabilities $\{\pi_{ij}\}$, where

$$\sum_{j=0}^{1}\sum_{k=0}^{1} \pi_{jk} = 1.$$

- The interpretation is that π_{10} is the probability of obtaining a pair in which the case is exposed and its matched control is not.

Pair-Matched Case-Control Study (cont'd)

- In such study, the parameter measuring the association between exposure and disease is the odds ratio $\psi = \pi_{10}/\pi_{01}$. Intuitively, the estimator for ψ is $\hat{\psi} = Y_{10}/Y_{01}$.
- To derive the large sample distribution of $\hat{\psi}$, it will be easier to work on $\log(\hat{\psi})$, instead of $\hat{\psi}$.
- By the delta method in the multivariate case, think about $g(\pi_{10}, \pi_{01}) = \log(\pi_{01}/\pi_{01})$.
- Then, one has

$$\sqrt{n}\{\log(Y_{10}/Y_{01}) - \log(\pi_{10}/\pi_{01})\} \rightarrow_d N(0, \tau^2),$$

where

$$\tau^2 = A + B + C,$$

Pair-Matched Case-Control Study (cont'd)

- With

$$A = \left\{ \frac{\partial g(\pi_{10}, \pi_{01})}{\partial \pi_{10}} \right\}^2 \sigma_{10}^2 = \frac{1}{\pi_{10}^2} \pi_{10}(1 - \pi_{10}),$$

$$B = \left\{ \frac{\partial g(\pi_{10}, \pi_{01})}{\partial \pi_{01}} \right\}^2 \sigma_{01}^2 = \frac{1}{\pi_{01}^2} \pi_{01}(1 - \pi_{01})$$

and

$$C = 2 \frac{\partial g(\pi_{10}, \pi_{01})}{\partial \pi_{01}} \frac{\partial g(\pi_{10}, \pi_{01})}{\partial \pi_{10}} \sigma_{10,01}^2 = \frac{-2}{\pi_{10}\pi_{01}} (-\pi_{10}\pi_{01}).$$

- That concludes,

$$\tau^2 = \frac{1}{\pi_{10}}(1 - \pi_{10}) + \frac{1}{\pi_{01}}(1 - \pi_{01}) + 2 = \frac{1}{\pi_{10}} + \frac{1}{\pi_{01}}.$$

Pair-Matched Case-Control Study (cont'd)

- A common way to express $Var\{\log(\hat{\psi})\}$ is

$$Var\{\log(\hat{\psi})\} \approx \frac{1}{n\pi_{10}} + \frac{1}{n\pi_{01}}.$$

- That gives a common estimator for the variance of $\log(\hat{\psi})$ as

$$\widehat{Var}\{\log(\hat{\psi})\} \approx \frac{1}{Y_{10}} + \frac{1}{Y_{01}}.$$

- And, the large sample distribution of $\log(\hat{\psi})$ is

$$\frac{\log(\hat{\psi}) - \log(\psi)}{\sqrt{\widehat{Var}\{\log(\hat{\psi})\}}} \sim N(0, 1).$$

Data Reduction

Feng-Chang Lin

Department of Biostatistics
University of North Carolina at Chapel Hill

flin@bios.unc.edu

(C&B §6)

Introduction

- Suppose that we are interested in estimating a parameter θ.
- If there is a random sample, X, whose pdf or pmf does not depend on θ, one would say "X does not contain any information about θ".
- On the other hand, it is possible to have a brief summary statistic that contains all the information about θ.
- We call this "data reduction", which summarizes a large number of observations into a small number of summary statistics.
- Our ultimate goal is to find the "smallest", most concise, summary statistics.

Sufficient Statistics

- Principle: If $T(X)$ is a sufficient statistic for θ, then it is sufficient to do any inference about θ through $T(X)$.
- That is, if x and y are two sample values such that $T(x) = T(y)$, then inference about θ should be the same whether $X = x$ or $X = y$ is observed.
- **Sufficient statistics**: A statistic $T(X)$ is a *sufficient statistic* for θ if the conditional distribution of the sample X given the value of $T(X)$ does not depend on θ.

Sufficient Statistics (cont'd)

- **Example** Let $X_1, \cdots, X_n$ be iid random variables distributed as bernoulli(θ), $0 < \theta < 1$. Show that $T(X) = \sum_{i=1}^n X_i$ a sufficient statistic for θ.
- **Proof** Since

$$P(X = x | T(X) = t) = \frac{P(X = x, T(X) = t)}{P(T(X) = t)},$$

where

$$P(T(x) = t) = \left(\begin{array}{c} n \\ t \end{array} \right) \theta^t (1-\theta)^{n-t},$$

and

$$P(X = x, T(X) = t) = P(X = x) = \prod_{i=1}^n P(X_i = x_i) = \theta^t (1-\theta)^{n-t}.$$

Sufficient Statistics (cont'd)

- Hence, $P(X = x|T(X) = t) = t!(n-t)!/n!$, for those $x_i's$ with $\sum_{i=1}^{n} x_i = t$, and $P(X = x|T(X) = t) = 0$, otherwise.

Sufficient Statistics (cont'd)

- For θ, the sufficiency statistics may not be unique.
- In this case, $\bar{X}$, $(X_1, \bar{X})$, $(X_1, \cdots, X_n)$ are all sufficient statistics.
- **Theorem 6.2.2** If $p(x|\theta)$ is the joint pdf or pmf of X and $q(t|\theta)$ is the pdf or pmf of $T(X)$. $T(X)$ is a sufficient statistic for θ if, for every x in the sample space, the ratio $p(x|\theta)/q(T(x)|\theta)$ does not depend on θ.

Finding Sufficient Statistics

- So far, we only show whether $T(X)$ is a sufficient statistic.
- The question here is "how to find one"?

Theorem (Factorization Theorem)

Let $f(x|\theta)$ be the joint pdf or pmf of X. A statistic $T(X)$ is a sufficient statistic for θ if and only if there exist functions $g(t|\theta)$ and $h(x)$ such that, for all sample points x and all parameter points θ,

$$f(x|\theta) = g(T(x)|\theta)h(x).$$

Finding Sufficient Statistics (cont'd)

- **Example** Let $X_1, \cdots, X_n$ be iid random variables distributed as Bernoulli(θ), $0 < \theta < 1$. Show that $T(x) = \sum_{i=1}^n x_i$ is a sufficient statistic using Factorization Theorem.
- **Proof** We first write the joint pmf

$$\begin{aligned} P(X = x) &= \prod_{i=1}^n P(X_i = x_i) \\ &= \prod_{i=1}^n \theta^{x_i}(1-\theta)^{(1-x_i)} I(x_i \in \{0,1\}) \\ &= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \prod_{i=1}^n I(x_i \in \{0,1\}). \end{aligned}$$

- We can have $g(T(x)|\theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$ as a function of $T(x) = \sum_{i=1}^n x_i$ and $h(x) = \prod_{i=1}^n I(x_i \in \{0,1\})$.

Finding Sufficient Statistics (cont'd)

- **Example** Let $X_1, \cdots, X_n$ be iid random variables distributed as Uniform$(0, \theta)$. Find a sufficient statistic for θ.
- **Solution** To apply the factorization theorem, we first write the joint pdf

$$f_X(x) = \theta^{-n} \prod_{i=1}^{n} I(0 < x_i < \theta) = \theta^{-n} I(0 < x_{(n)} < \theta) I(0 < x_{(1)})$$

- Take $T(x) = x_{(n)}$, $g(T(x)|\theta) = \theta^{-n} I(0 < T(x) < \theta)$, and $h(x) = I(0 < x_{(1)})$.
- We can conclude $T(X) = X_{(n)}$ is a sufficient statistic for θ.

Sufficiency in Exponential Family

- **Theorem 6.2.10** Let $X_1, \cdots, X_n$ be iid random variables from a pdf or pmf $f(x|\theta)$ that belongs to the exponential family given by

$$f(x|\theta) = h(x)c(\theta)\exp\left(\sum_{j=1}^{k} w_j(\theta)t_j(x)\right),$$

where $\theta = (\theta_1, \cdots, \theta_d)$, $d \leq k$. Then,

$$T(X) = \left(\sum_{i=1}^{n} t_1(X_i), \cdots, \sum_{i=1}^{n} t_k(X_i)\right)$$

is a sufficient statistic for θ.

Sufficiency in Exponential Family (cont'd)

- **Example** Let $X_1, \cdots, X_n$ be iid random variables distributed as Bernoulli(θ), $0 < \theta < 1$. Show that $T(X) = \sum_{i=1}^n X_i$ is a sufficient statistic for θ.
- **Solution** The pmf for one observation is

$$
\begin{aligned}
P(X_1 = x) &= \theta^x (1-\theta)^{1-x} I(x \in \{0,1\}) \\
&= I(x \in \{0,1\})(1-\theta) \exp\left(x \log \frac{\theta}{1-\theta} \right).
\end{aligned}
$$

- Take $h(x) = I(x \in \{0,1\})$, $c(\theta) = (1-\theta)$, $w_1(\theta) = \log \frac{\theta}{1-\theta}$, $t_1(x) = x$.
- By the sufficiency theorem in exponential family, one can conclude $T(X) = \sum_{i=1}^n t_1(X_i) = \sum_{i=1}^n X_i$ is a sufficient statistic for θ.

Minimal Sufficient Statistics

- In the Bernoulli example, there is a large number of sufficient statistics: $\sum_{i=1}^{n} X_i$, $\bar{X}$, $(X_1, \bar{X})$,$\cdots$,$(X_1, \cdots, X_n)$.
- Apparently, some of these can be reduced to a simpler form that is still sufficient for θ.
- **Minimal Sufficient Statistics**: A sufficient statistic is a minimal sufficient statistic if it is a function of every other sufficient statistic.
- Any one-to-one transformation of a minimal sufficient statistic is also a minimal sufficient statistic (still not unique).

Minimal Sufficient Statistics (cont'd)

- **Theorem 6.2.13** Let $f(x|\theta)$ be the joint pdf or pmf of X. Suppose that there exists a function $T(X)$ such that, for every two sample points x and y, the ratio $f(x|\theta)/f(y|\theta)$ does not depend on θ if and only if $T(x) = T(y)$. Then $T(X)$ is a minimal sufficient statistic for θ.

Minimal Sufficient Statistics (cont'd)

- **Example** Let $X_1, \cdots, X_n$ be iid random variables distributed as Bernoulli(θ), $0 < \theta < 1$. Show that $T(x) = \sum_{i=1}^n x_i$ is a minimal sufficient statistic.
- **Proof** To apply the above theorem, we first write the joint pmf

$$P(X = x) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \prod_{i=1}^n I(x_i \in \{0, 1\}).$$

- If $T(x) = \sum_{i=1}^n x_i$, one can have

$$P(X = x) = \left(\frac{\theta}{1-\theta}\right)^{T(x)} (1-\theta)^n \prod_{i=1}^n I(x_i \in \{0, 1\}).$$

Minimal Sufficient Statistics (cont'd)

- Taking two points, x and y, in the sample space for X. One has

$$\frac{P(X=x)}{P(X=y)} = \left(\frac{\theta}{1-\theta}\right)^{T(x)-T(y)}.$$

- The ratio does not depend on θ if and only of $T(x) = T(y)$.

Ancillary Statistics

- Sample values may contain some additional information that is redundant of θ.
- For example, suppose that X_1, X_2 are iid as $N(\theta, 1)$. The random variable $X_1 - X_2$ is distributed as $N(0, 2)$.
- Is $X_1 - X_2$ expected to provide any information about θ?
- How about $(X_1 - X_2, X_2)$?
- **Ancillary Statistics**: A statistic whose distribution does not depend on the parameter θ is called an *ancillary statistic* (for θ).

Ancillary Statistics (cont'd)

- Let $X_1, \cdots, X_n$ be iid from a *scale* parameter family with cdf $F(x/\sigma)$, $\sigma > 0$.
- Any statistic that depends on $X_1/X_n, \cdots, X_{n-1}/X_n$ is an ancillary statistic.
- For example, $(X_1 + \cdots + X_n)/X_n = X_1/X_n + \cdots + X_{n-1}/X_n + 1$ is an ancillary statistic.
- Let $Z_i = X_i/\sigma$. We know that Z_i does not depend on σ.
- Since the joint cdf of $X_1/X_n, \cdots, X_{n-1}/X_n$ is

$$\begin{aligned} F(y_1, \ldots, y_{n-1}|\sigma) &= P(X_1/X_n \leq y_1, \ldots, X_{n-1}/X_n \leq y_{n-1}) \\ &= P(\sigma Z_1/(\sigma Z_n) \leq y_1, \ldots, \sigma Z_{n-1}/(\sigma Z_n) \leq y_{n-1}) \\ &= P(Z_1/Z_n \leq y_1, \ldots, Z_{n-1}/Z_n \leq y_{n-1}) \end{aligned}$$

- The last line shows the cdf does not depend on σ and $(X_1 + \cdots + X_n)/X_n$ is an ancillary statistic of σ.

Complete Statistics

- **Complete Statistics**: Let $\{f(t|\theta) : \theta \in \Theta\}$ be a family of pdfs or pmfs for $T(X)$. The family is called complete if $E_\theta g(T) = 0$ for all $\theta \in \Theta$ implies that $P_\theta(g(T) = 0) = 1$ for all $\theta \in \Theta$.
- Completeness means that the only function of T with mean 0 is the 0 function.
- **Example** Let $X_1, \cdots, X_n$ be iid random variables distributed as $N(\theta, \theta^2)$, $-\infty < \theta < \infty$. Is $T = (\bar{X}, S^2)$ complete? Since $E_\theta \bar{X}^2 = \theta^2 + \theta^2/n = (1 + 1/n)\theta^2$ and $E_\theta S^2 = \theta^2$, one can have $g(T) = \bar{X}^2 - (1 + 1/n)S^2$ and $E_\theta g(T) = 0$ for all $\theta \in \Theta$.
- Here $g(T)$ is not a zero function (with probability 1) and does not involve θ. Hence T is NOT complete.

Complete Statistics (cont'd)

- **Example** Let $X \sim \text{Bernoulli}(\theta)$, $\theta \in (0, 1)$. Take $T(X) = X$. Is T complete? This is equivalent to find out if $g = 0$ is the only function that has $E_\theta g(T) = 0$ for all $\theta \in (0, 1)$.
- **Solution** Since X follows Bernoulli, one only has $g(0)$ and $g(1)$ for $g(T)$. Then, if

$$E_\theta g(T) = g(0)(1 - \theta) + g(1)\theta = g(0) + \{g(1) - g(0)\}\theta = 0,$$

the only solution for g function is $g(0) = g(1) = 0$ for $\theta \in (0, 1)$.

Complete Statistics (cont'd)

- **Example** Similarly, let $X \sim \text{Binomial}(2, \theta)$, $\theta \in \Theta$, where $\Theta = \{1/3, 2/3\}$. Take $T(X) = X$. Is T complete? One can see $X = 0, 1, 2$. Follow the same approach,

$$E_\theta g(T) = (4/9)g(0) + (4/9)g(1) + (1/9)g(2), \text{ if } \theta = 1/3,$$
$$E_\theta g(T) = (1/9)g(0) + (4/9)g(1) + (4/9)g(2), \text{ if } \theta = 2/3.$$

  If $E_\theta g(T) = 0$, one can find $g(0) = g(2) = 4$, $g(1) = -5$ as a solution, which shows g function can be non-zero

- **Example** Let $X \sim \text{Binomial}(2, \theta)$, $\theta \in \Theta$, where $\Theta = \{1/3, 1/2, 2/3\}$. Take $T(X) = X$. Is T complete? Yes.
- That tells you the completeness highly depends on the parameter space.

Completeness in Exponential Families

- Let $X_1, \cdots, X_n$ be iid random variables from a pdf or pmf $f(x|\theta)$ that belongs to the exponential family given by

$$f(x|\theta) = h(x)c(\theta)\exp\left(\sum_{j=1}^{k} w_j(\theta)t_j(x)\right),$$

  where $\theta = (\theta_1, \cdots, \theta_k)$. Then

$$T(X) = \left(\sum_{i=1}^{n} t_1(X_i), \cdots, \sum_{i=1}^{n} t_k(X_i)\right)$$

  is complete if $\{(w_1(\theta), \cdots, w_k(\theta)) : \theta \in \Theta\}$ contains an open set in R^k.
- **Example**: The family $\{N(\mu, \sigma^2) : -\infty < \mu < \infty\}$ with a fixed $\sigma^2 < \infty$ is complete.

Exponential Families

- **Example**: Let $f(x|\mu, \sigma^2)$ be the $N(\mu, \sigma^2)$ family of pdfs where $\theta = (\mu, \sigma^2)$, $-\infty < \mu < \infty$, $\sigma > 0$. Then

$$
\begin{aligned}
f(x|\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \exp\left(-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2}\right).
\end{aligned}
$$

- Take $h(x) = 1$ for all x,

$$
\begin{aligned}
& c(\theta) = c(\mu, \sigma) = (\sqrt{2\pi}\sigma)^{-1} \exp(-\mu^2/(2\sigma^2)), -\infty < \mu < \infty, \sigma > 0, \\
& w_1(\mu, \sigma) = \sigma^{-2}, \sigma > 0, w_2(\mu, \sigma) = \mu/\sigma^{-2}, \sigma > 0, \\
& t_1(x) = -x^2/2, \text{ and } t_2(x) = x.
\end{aligned}
$$

Exponential Families (cont'd)

- **Example** If $f(x|\theta) = \theta^{-1} \exp(1 - (x/\theta))$, $0 < \theta < x < \infty$, it is not an exponential family since

$$f(x|\theta) = \theta^{-1} \exp\left(1 - \left(\frac{x}{\theta}\right)\right) I_{[\theta,\infty)}(x).$$

- The indicator function is not a function of x alone, and cannot be expressed as an exponential.

Basu's theorem

Theorem (Basu's Theorem)

If $T(X)$ is a complete and minimal sufficient statistic, then $T(X)$ is independent of every ancillary statistic.

Proof: (only for discrete distributions) Let $S(X)$ be any ancillary statistic, so $P(S(X) = s)$ does not depend on θ. Since $T(X)$ is a sufficient statistic,

$$P(S(X) = s|T(X) = t) = P(X \in \{x : S(x) = s\}|T(X) = t),$$

does not depend on θ. For independence, we owe to show

$$P(S(X) = s|T(X) = t) = P(S(X) = s)$$

for all possible values of $t \in \mathcal{T}$.

Basu's theorem (cont'd)

- Marginalizing the joint probability of $S(X)$ and $T(X)$, one can have

$$
\begin{aligned}
P(S(X) = s) &= \sum_{t \in \mathcal{T}} P(S(X) = s, T(X) = t) \\
&= \sum_{t \in \mathcal{T}} P(S(X) = s | T(X) = t) P_\theta(T(X) = t). \qquad (1)
\end{aligned}
$$

- Since $\sum_{t \in \mathcal{T}} P_\theta(T(X) = t) = 1$, one can also write

$$
\begin{aligned}
P(S(X) = s) &= P(S(X) = s) \sum_{t \in \mathcal{T}} P_\theta(T(X) = t) \\
&= \sum_{t \in \mathcal{T}} P(S(X) = s) P_\theta(T(X) = t). \qquad (2)
\end{aligned}
$$

Basu's theorem (cont'd)

- By (1) and (2), we can have

$$
\begin{aligned}
0 &= P(S(X)=s) - P(S(X)=s) \\
&= \sum_{t\in\mathcal{T}} \{P(S(X)=s|T(X)=t) - P(S(X)=s)\} P_\theta(T(X)=t)
\end{aligned}
$$

- If we let $g(t) = P(S(X)=s|T(X)=t) - P(S(X)=s)$, then

$$
0 = \sum_{t\in\mathcal{T}} g(t) P_\theta(T(X)=t) = E_\theta g(T), \quad \text{for all } \theta.
$$

- Since $T(X)$ is a complete statistic, the equation above implies that $g(t) = 0$ for all possible values of $t \in \mathcal{T}$.
- Hence, we can claim $P(S(X)=s|T(X)=t) = P(S(X)=s)$.

Basu's theorem (cont'd)

- Did we use "minimality" of the sufficient statistics in the proof?
- For the problems we will consider, a sufficient statistic will be complete only if it is minimal.
- **Theorem 6.2.28** If a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistics.

Practical Use of Basu's theorem

- **Example** Let $X_1, \cdots, X_n$ be iid Exponential(θ). Compute the expected value of

$$S(X) = \frac{X_n}{X_1 + \cdots + X_n}.$$

- We can show that $S(X)$ is an ancillary statistic (How?)
- Since Exponential(θ) belongs to the exponential family (homework) with $t(x) = x$, so $T(X) = \sum_{i=1}^n X_i$ is a (minimal) sufficient statistic.
- Hence by Basu's theorem, $T(X)$ and $S(X)$ are independent and

$$\theta = E_\theta X_n = E_\theta T(X) S(X) = E_\theta T(X) E_\theta S(X) = n\theta E_\theta S(X).$$

One has $E_\theta S(X) = 1/n$.

Point Estimation

Feng-Chang Lin

Department of Biostatistics
University of North Carolina at Chapel Hill

flin@bios.unc.edu

(C&B §7)

Introduction

- Random sample $X_1, \cdots, X_n$ from $f(x|\theta)$, where θ is either a scalar or vector.
- We want to estimate θ or $\tau(\theta)$.
- **Example** If $X \sim N(\mu, \sigma^2)$, how do we estimate $\theta = (\mu, \sigma^2)$?
- **Example** If $X \sim N(\mu, \sigma^2)$, how do we estimate $\tau(\theta) = \mu/\sigma^2$?
- **Example** If $X \sim N(\mu, \sigma^2)$, how do we estimate $\tau(\theta) = P(X_1 > 100) = \Phi((100 - \mu)/\sigma)$?

Introduction (cont'd)

- *Point estimator*: Any function of the sample, a statistic, $W(X_1, \cdots, X_n)$, also simply called *estimator*. Specifically, an estimator can not be a function of θ. It must be a statistic.
- *Estimator*: The random variable $W(X_1, \cdots, X_n)$.
- *Estimate*: The realized value $W(x_1, \cdots, x_n)$.
- We want a good point estimator.
- How to find good estimators?
- What is a "good" estimator?

Method of Moments

- Match sample moments with population moments.
- Use as many sample moments as needed. Start with lower order moments first.
- The kth population moment: $\mu_k = EX_1^k$.
- The kth sample moment: $M_k = \frac{1}{n}\sum_{i=1}^n X_i^k$. What is M_1?
- Finding the moment estimator: Set $M_1 = \mu_1$, $M_2 = \mu_2$, $\cdots$, and solve for θ.
- The moment estimator will be denoted by $\hat{\theta}_{MM}$.
- **Example** $X_1, \cdots, X_n$ iid Bernoulli(θ), $\theta \in [0, 1]$. $M_1 = \mu_1$ gives $\hat{\theta}_{MM} = \bar{X}$.
- **Example** $X_1, \cdots, X_n$ iid $N(0, \theta)$, $M_1 = \mu_1 = 0$ is not usable. $M_2 = \mu_2 = \theta$ gives $\hat{\theta}_{MM} = \frac{1}{n}\sum_{i=1}^n X_i^2$.

Method of Moments (cont'd)

- **Example** $X_1, \cdots, X_n$ iid $N(\mu, \sigma^2)$, both μ and σ^2 unknown.

$$M_1 = \mu, \quad \text{and} \quad M_2 = \mu^2 + \sigma^2.$$

$$\hat{\mu}_{MM} = \bar{X}, \hat{\sigma}^2_{MM} = M_2 - M_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}S^2.$$

- **Example** $X_1, \cdots, X_n$ iid binomial(m, p), both m and p unknown, $p \in [0, 1]$, $m \in \{0, 1, \cdots\}$.

$$M_1 = mp, M_2 = (mp)^2 + mp(1-p).$$

$$\frac{M_2}{M_1} - M_1 = 1 - p, \hat{p}_{MM} = 1 - \frac{M_2 - M_1^2}{M_1}, \hat{m}_{MM} = \frac{M_1}{\hat{p}_{MM}}.$$

- Negative $\hat{p}_{MM}$ and $\hat{m}_{MM}$ is possible. Out of range moment estimators are not rare in applications.

Maximum Likelihood

- The *likelihood function* is the joint pdf or pmf, but viewed as a function of θ with the sample x being fixed.
- If X is a random vector representing the observable data, then

$$L(\theta|x) = f(x|\theta).$$

- If $X_1, \cdots, X_n$ is a random sample from a pdf or pmf $f(x|\theta)$, then

$$L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta),$$

with the log-likelihood function

$$\ell(\theta|x) = \log L(\theta|x) = \sum_{i=1}^{n} \log f(x_i|\theta).$$

Maximum Likelihood (cont'd)

- For a given sample x, the maximum likelihood estimator (MLE), denoted $\hat{\theta}(x)$ is a value of θ at which $L(\theta|x)$ attains its maximum over the parameter space.
- The abbreviation MLE is used for both maximum likelihood *estimator* and maximum likelihood *estimate*.
- If the range of x depends on θ, that dependence should be built into $L(\theta|x)$.
- **Example** $X_1, \cdots, X_n$ iid uniform on $[0, \theta]$.

$$L(\theta|x) = \theta^{-n} \prod_{i=1}^{n} I(0 \le x_i \le \theta) = \theta^{-n} I(x_{(n)} \le \theta).$$

One has $\hat{\theta} = X_{(n)}$.

Maximum Likelihood (cont'd)

- Multiplied by a positive constant that does not involve the unknown parameters does not change the final answers.
- **Example** $X_1, \cdots, X_n$ iid Binomial(m, θ), with m known and $\theta \in [0, 1]$ unknown. The likelihood is

$$L(\theta|x) = \prod_{i=1}^{n} \binom{m}{x_i} \theta^{x_i}(1-\theta)^{m-x_i} = C(x) \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{m-x_i},$$

  where $C(x)$ depends on x but not θ.
- Dropping $C(x)$ does not affect the maximization over θ.
- A value of θ that maximizes the log-likelihood $\ell(\theta|x)$ will also maximize the likelihood $L(\theta|x)$.

Maximum Likelihood (cont'd)

- There is no single simple procedure that is applicable to all types of problems for finding MLE.
- **Example** X is a single observation from the Binomial(m, θ), with unknown $\theta \in [0, 1]$, and known $m \geq 1$.

$$L(\theta|x) = \left(\begin{array}{c} m \\ x \end{array} \right) \theta^x (1-\theta)^{m-x}.$$

- If $x = 0$, the likelihood $L(\theta|0) = (1-\theta)^m$, which is monotone decreasing in θ. One would say $\hat{\theta} = 0$.
- If $x = m$, the likelihood $L(\theta|m) = \theta^m$, which is monotone increasing in θ. One would say $\hat{\theta} = 1$.
- If $0 < x < m$, the likelihood $L(\theta|x)$ is maximized at $\hat{\theta} = x/m$.
- In all cases, $\hat{\theta} = x/m$.

Maximum Likelihood (cont'd)

- **Example** Let $X_1, \cdots, X_n$ be iid random variables distributed as $N(\theta, 1)$, $\theta \in (-\infty, \infty)$.

$$L(\theta|x) = (2\pi)^{-n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (x_i - \theta)^2 \right\},$$

$$\ell(\theta|x) = (-n/2)\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 - \frac{n}{2}(\bar{x} - \theta)^2,$$

- The log-likelihood is a quadratic function in θ that has a unique global maximum at $\theta = \bar{x}$, so $\hat{\theta} = \bar{x}$.

Maximum Likelihood (cont'd)

- **Example (restricted range)** Let $X_1, \cdots, X_n$ be iid random variables distributed as $N(\theta, 1)$, $\theta \in [0, \infty)$. If $\bar{x} \geq 0$, then $\bar{x}$ is the MLE. If $\bar{x} < 0$, the log-likelihood

$$\ell(\theta|x) = (-n/2)\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 - \frac{n}{2}(\bar{x} - \theta)^2,$$

  will be monotone decreasing over $[0, \infty)$, hence its maximum will be at $\hat{\theta} = 0$.

- **Example (flat likelihood function)** $X_1, \cdots, X_n$ iid Uniform$(\theta - 1/2, \theta + 1/2)$.

$$L(\theta|x) = I(x_{(1)} > \theta - \frac{1}{2}) I(x_{(n)} < \theta + \frac{1}{2})$$

  The likelihood $L(\theta|x) = 1$ over $\theta \in (x_{(n)} - 1/2, x_{(1)} + 1/2)$ and $L(\theta|x) = 0$ otherwise.

Maximum Likelihood (cont'd)

- **Discrete parameter, MLE of the binomial m with known p**
  Consider a single observation X from the Binomial(m, p), with p known and m unknown. We want to find the MLE of m.
- The parameter space is the set of integers $\{1, 2, \cdots\}$.
- Suppose that $p = 0.71$ and the observed value is $x = 7$. What is the MLE of m?
- Because $P(X = x|m) = 0$ if $x > m$, we get $L(m|x) = 0$ for $m < 7$ and $L(m|x) = \left(\begin{array}{c} m \\ x \end{array} \right) p^x(1-p)^{(m-x)}$ for integer $m \geq 7$.
- $L(m|x)$ is increasing for $7 \leq m \leq 9$ and decreasing for for $m \geq 9$. We can conclude that the MLE is $\hat{m} = 9$.
- Since $EX = mp$, the moment estimate is $\hat{m}_{MM} = x/p = 7/0.71 \approx 9.86$, which is not far from the MLE.

MLE for a 2-dimensional Parameter

- A two-dimensional parameter, and the likelihood is twice-differentiable.
- Use rules of calculus to find "local" maximum.
- The rules for a local maximum:
  a) Two first-order partial derivatives are zero.
  b) At least one second-order partial derivatives is negative.
  c) The Jacobian of the second-order partial derivatives is positive.
- **Example**: The $N(\mu, \sigma^2)$ model with both parameters unknown.

$$\ell(\mu, \sigma^2|x) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2/\sigma^2.$$

Example (Normal Distribution)

$$\frac{\partial}{\partial \mu}\ell(\mu, \sigma^2|x) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)$$

$$\frac{\partial}{\partial \sigma^2}\ell(\mu, \sigma^2|x) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\frac{\partial^2}{\partial \mu^2}\ell(\mu, \sigma^2|x) = -\frac{n}{\sigma^2}$$

$$\frac{\partial^2}{\partial (\sigma^2)^2}\ell(\mu, \sigma^2|x) = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\frac{\partial^2}{\partial \mu \partial \sigma^2}\ell(\mu, \sigma^2|x) = -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu)$$

$$J(\hat{\mu}, \hat{\sigma}^2) = \frac{1}{\hat{\sigma}^6}\frac{n^2}{2} > 0.$$

Successive 1-dimensional Maximization

- To maximize $L(\alpha, \beta|x)$ over both α and β, we can proceed as follows.
- First, for a fixed α, we maximize $L(\alpha, \beta|x)$ over β.
- Let $\hat{\beta}(\alpha)$ be the value of β that maximizes $L(\alpha, \beta|x)$ for a fixed α.
- The function
$$H(\alpha|x) = L(\alpha, \hat{\beta}(\alpha)|x)$$
depends on α.
- We call this kind of function $H(\alpha|x)$ as the *profiled likelihood* for α.
- Then, the MLE of β is simply $\hat{\beta}(\hat{\alpha}_H)$, where $\hat{\alpha}_H$ is the maximizer of $H(\alpha|x)$.

Successive 1-dimensional Maximization (cont'd)

- **Example (MLE of the Weibull parameters)** Let $X_1, \cdots, X_n$ are iid Weibull(α, β) with density

$$f(x|\alpha, \beta) = \frac{\alpha}{\beta} x^{\alpha-1} \exp\left(-\frac{x^\alpha}{\beta}\right), x \geq 0, \alpha > 0, \beta > 0.$$

- The log-likelihood

$$\ell(\alpha, \beta|x) = n \log \alpha - n \log \beta + (\alpha - 1) \sum_{i=1}^{n} \log x_i - \frac{1}{\beta} \sum_{i=1}^{n} x_i^\alpha.$$

- Maximized over β by setting the derivative

$$\frac{d}{d\beta} \ell(\alpha, \beta|x) = -\frac{n}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^{n} x_i^\alpha = 0.$$

- The solution is $\hat{\beta}(\alpha) = n^{-1} \sum_{i=1}^{n} x_i^\alpha$.

Successive 1-dimensional Maximization (cont'd)

- The solution is verified to be a maximum since

$$\frac{d^2}{d\beta^2}\ell(\alpha,\beta|x)\Big|_{\beta=\hat{\beta}(\alpha)} = -\frac{n}{\hat{\beta}(\alpha)^2} < 0.$$

- The profile log-likelihood for α is

$$\begin{aligned} h(\alpha|x) &= \ell(\alpha,\hat{\beta}(\alpha)|x) \\ &= n\left\{\log\alpha - \log\frac{\sum_{i=1}^n x_i^\alpha}{n} + (\alpha-1)\frac{\sum_{i=1}^n \log x_i}{n} - 1\right\}. \end{aligned}$$

- There is no "closed" form for α. Maximization over α can be done either graphically or by numerical methods.

Invariance Property of MLE

Theorem (Theorem 7.2.10)

If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

- If the mapping $\theta \to \tau(\theta)$ is one-to-one, then it is easy to see that the MLE of $\eta = \tau(\theta)$ is the same since

$$L^*(\eta|x) = \prod_{i=1}^{n} f(x_i|\tau^{-1}(\eta)) = L(\tau^{-1}(\eta)|x),$$

  and

$$\sup_{\eta} L^*(\eta|x) = \sup_{\eta} L(\tau^{-1}(\eta)|x) = \sup_{\theta} L(\theta|x).$$

- The proof is more complicated if the τ function is not one-to-one. Check p. 320 in C&B.

Invariance Property of MLE (cont'd)

- **Example** What is the MLE of θ^2 if $X_1, \cdots, X_n$ follows $N(\theta, \sigma^2)$?
- **Example** What is the MLE of $\sqrt{p(1-p)}$ if $X_1, \cdots, X_n$ follows Binomial(n, p)?

$$
\begin{aligned}
L(p|x) &= \prod_{i=1}^{n} \left(\begin{array}{c} n \\ x_i \end{array} \right) p^{x_i}(1-p)^{n-x_i} \\
&= C(x) p^{\sum_{i=1}^{n} x_i}(1-p)^{n^2 - \sum_{i=1}^{n} x_i}.
\end{aligned}
$$

$$
\ell(p|x) \propto \sum_{i=1}^{n} x_i \log p + (n^2 - \sum_{i=1}^{n} x_i) \log(1-p).
$$

- How do we find the MLE of p?

Instability of MLE

- The MLE can be highly unstable if the likelihood function is very flat in the neighborhood of its maximum.
- **Example** $X_1, \cdots, X_5$ follows Binomial(n, p) with both n and p unknown.

$$\text{Sample 1: } (16, 18, 22, 25, 27) \Rightarrow \hat{n} = 99;$$
$$\text{Sample 2: } (16, 18, 22, 25, 28) \Rightarrow \hat{n} = 190.$$

- Even worse, there is no finite maximum. The MLE doesn't exist.

Method of Evaluating Estimators

- Bias: $\text{Bias}_\theta W(X) = E_\theta W(X) - \theta$.
- Variance: $Var_\theta W(X)$.
- Mean Squared Error (MSE): $E_\theta(W(X) - \theta)^2$=Bias^2 + Variance.
- Other: $E_\theta g(|W - \theta|)$.
- **Example**: Let $X_1, \cdots, X_n$ be iid N(μ, σ^2). Since,

$$E\bar{X} = \mu, \quad ES^2 = \sigma^2,$$

for all μ and σ^2. The MSE of these estimators are given by

$$E(\bar{X} - \mu)^2 = Var\bar{X} = \frac{\sigma^2}{n},$$
$$E(S^2 - \sigma^2)^2 = VarS^2 = \frac{2\sigma^4}{n-1}.$$

Method of Evaluating Estimators (cont'd)

- What is the MLE of μ and σ^2?

$$E(\hat{\sigma}^2) = E\left(\frac{n-1}{n}S^2\right) = \frac{n-1}{n}\sigma^2,$$
$$Var(\hat{\sigma}^2) = Var\left(\frac{n-1}{n}S^2\right) = \frac{2(n-1)}{n^2}\sigma^4.$$

- The MSE of $\hat{\sigma}^2$ is given by

$$E(\hat{\sigma}^2 - \sigma^2)^2 = \frac{2(n-1)}{n^2}\sigma^4 + \left(\frac{n-1}{n}\sigma^2 - \sigma^2\right)^2 = \left(\frac{2n-1}{n^2}\right)\sigma^4.$$

- We have

$$E(\hat{\sigma}^2 - \sigma^2)^2 = \left(\frac{2n-1}{n^2}\right)\sigma^4 < \left(\frac{2}{n-1}\right)\sigma^4 = E(S^2 - \sigma^2)^2.$$

Best Unbiased Estimators

- Having a “biased" estimator may not be acceptable.
- Finding an estimator that minimizes MSE may not be reasonable for “scale" parameters.
- One can restrict their searching for the “best" estimator only form those “unbiased" estimators.
- **Uniformly Minimum Variance Unbiased Estimators (UMVUE)**: An estimator W^* is a best unbiased estimator of $\tau(\theta)$ if it satisfies $E_\theta W^* = \tau(\theta)$ for all θ and, for any other estimator W with $E_\theta W = \tau(\theta)$, we have $Var_\theta W^* \leq Var_\theta W$ for all θ.
- W^* is called *uniformly minimum variance unbiased estimators* (UMVUE) of $\tau(\theta)$.
- “Uniformly" means the statement holds for all $\theta \in \Theta$.

Best Unbiased Estimators (cont'd)

- **Example** Let $X_1, \cdots, X_n$ be iid Poisson(λ), and let $\bar{X}$ and S^2 be the sample mean and variance, respectively. One has

$$E_\lambda \bar{X} = \lambda, \quad \text{and} \quad E_\lambda S^2 = \lambda,$$

  so both $\bar{X}$ and S^2 are unbiased estimators of λ. By linear combinations of $\bar{X}$ and S^2, we can create infinitely many unbiased estimators. Do we have the best one?

Cauchy-Schwarz Inequality

- For random variables X and Y,

$$Cov(X,Y) \leq \sqrt{Var(X)Var(Y)}.$$

or, equivalently,

$$VarX \geq \frac{\{Cov(X,Y)\}^2}{VarY}.$$

Cramér-Rao Lower Bound (CRLB)

- **Cramér-Rao Inequality** Let $X_1, \ldots, X_n$ be a sample with pdf $f(x|\theta)$, and let $W(\boldsymbol{X}) = W(X_1, \ldots, X_n)$ be any unbiased estimator of $\tau(\theta)$ satisfying

$$\frac{d}{d\theta} E_\theta W(\boldsymbol{X}) = \int_{\mathcal{X}} \frac{\partial}{\partial\theta}[W(\boldsymbol{x}) f(\boldsymbol{x}|\theta)] d\boldsymbol{x},$$

and

$$Var_\theta W(\boldsymbol{X}) < \infty.$$

Then

$$Var_\theta W(\boldsymbol{X}) \geq \frac{\{d\tau(\theta)/d\theta\}^2}{E_\theta\{U(\theta|\boldsymbol{x})\}^2},$$

where $U(\theta|\boldsymbol{x}) = \frac{\partial}{\partial\theta} \log f(\boldsymbol{x}|\theta)$ is called score function.

Cramér-Rao Lower Bound (cont'd)

- **Proof**: Note that,

$$\begin{aligned}\frac{d}{d\theta}E_\theta W(\boldsymbol{X}) &= \int_{\mathcal{X}} W(\boldsymbol{x})\left\{\frac{\partial}{\partial\theta}f(\boldsymbol{x}|\theta)\right\}d\boldsymbol{x} \\ &= E_\theta\left\{W(\boldsymbol{X})\frac{\frac{\partial}{\partial\theta}f(\boldsymbol{X}|\theta)}{f(\boldsymbol{X}|\theta)}\right\} \\ &= E_\theta\left\{W(\boldsymbol{X})\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right\}. \qquad (1)\end{aligned}$$

- If $W(\boldsymbol{X}) = 1$ in (1), one can have

$$E_\theta\left\{\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right\} = \frac{d}{d\theta}E_\theta(1) = 0.$$

Cramér-Rao Lower Bound (cont'd)

- According to (1), we have

$$Cov_\theta \left\{ W(\boldsymbol{X}), \frac{\partial}{\partial\theta} \log f(\boldsymbol{X}|\theta) \right\} = E_\theta \left\{ W(\boldsymbol{X}) \frac{\partial}{\partial\theta} \log f(\boldsymbol{X}|\theta) \right\}$$
$$= \frac{d}{d\theta} E_\theta W(\boldsymbol{X}).$$

- Also, we have

$$Var_\theta \left\{ \frac{\partial}{\partial\theta} \log f(\boldsymbol{X}|\theta) \right\} = E_\theta \left[\left\{ \frac{\partial}{\partial\theta} \log f(\boldsymbol{X}|\theta) \right\}^2 \right].$$

- By the Cauchy-Schwarz Inequality, we have

$$Var_\theta W(\boldsymbol{X}) \geq \frac{\left\{ \frac{d}{d\theta} E_\theta W(\boldsymbol{X}) \right\}^2}{E_\theta \left[\left\{ \frac{\partial}{\partial\theta} \log f(\boldsymbol{X}|\theta) \right\}^2 \right]}.$$

Cramér-Rao Lower Bound (cont'd)

- If $X_1, \cdots, X_n$ are iid with pdf $f(x|\theta)$, then

$$Var_\theta W(\boldsymbol{X}) \geq \frac{\left\{\frac{d}{d\theta} E_\theta W(\boldsymbol{X})\right\}^2}{nE_\theta\left[\left\{\frac{\partial}{\partial\theta}\log f(X_1|\theta)\right\}^2\right]}.$$

- If $f(x|\theta)$ satisfies

$$\frac{d}{d\theta} E_\theta\left\{\frac{\partial}{\partial\theta}\log f(X_1|\theta)\right\} = \int \frac{\partial}{\partial\theta}\left[\left\{\frac{\partial}{\partial\theta}\log f(x_1|\theta)\right\} f(x_1|\theta)\right] dx_1,$$

  then

$$E_\theta\left[\left\{\frac{\partial}{\partial\theta}\log f(X_1|\theta)\right\}^2\right] = -E_\theta\left\{\frac{\partial^2}{\partial\theta^2}\log f(X_1|\theta)\right\}.$$

- Proof can be found in Exercise 7.39 of C&B.

Cramér-Rao Lower Bound (cont'd)

- In the Poisson example, $\tau(\lambda) = \lambda$ so $\tau'(\lambda) = 1$.
- One can show

$$E_\lambda\{U(\lambda|\boldsymbol{X})\}^2 = -nE_\lambda\left\{\frac{\partial^2}{\partial\lambda^2}\log f(X_1|\lambda)\right\} = \frac{n}{\lambda}.$$

- Hence for any unbiased estimator, W, of λ, we must have

$$Var_\lambda W \geq \frac{\lambda}{n}.$$

- Since $Var_\lambda \bar{X} = \lambda/n$, $\bar{X}$ is the best unbiased estimator of λ.

Violation of the Assumption in CRLB

- Let $X_1, \cdots, X_n$ be iid pdf $f(x|\theta) = 1/\theta$, $0 < x < \theta$. Since $\frac{\partial}{\partial\theta} \log f(x|\theta) = -1/\theta$. We have

$$E_\theta \left[\left\{ \frac{\partial}{\partial\theta} \log f(X_1|\theta) \right\}^2 \right] = \frac{1}{\theta^2}.$$

- The CRLB indicates $Var_\theta W \geq \theta^2/n$.
- However, $EX_{(n)} = \frac{n}{n+1}\theta$, and

$$Var_\theta \left(\frac{n+1}{n} X_{(n)} \right) = \frac{1}{n(n+2)}\theta^2 < \frac{1}{n}\theta^2.$$

- The problem is $\frac{d}{d\theta} \int_0^\theta h(x) f(x|\theta) dx \neq \int_0^\theta h(x) \frac{d}{d\theta} f(x|\theta) dx$.

Uniqueness of UMVUE

Theorem (7.3.19 in C&B)

If W is the best unbiased estimator of $\tau(\theta)$, then W is unique.

- Suppose that W' is another best unbiased estimator of $\tau(\theta)$, i.e., $Var(W') = Var(W)$.
- Take $W^* = (W + W')/2$; one can easily see $EW^* = \tau(\theta)$.
- Using covariance inequality, one can show $Var(W^*) \leq Var(W)$.
- However, since W is the best, $Var(W^*)$ can only equal $Var(W)$.
- When the equality stands, it implies that $W' = a + bW$.
- One can show that $a = 0$, $b = 1$, and $W' = W$.

Sufficiency and Unbiasedness

Theorem (Rao-Blackwell Theorem)

Let W be any unbiased estimator of $\tau(\theta)$, and let T be a sufficient statistic for θ. Define $\phi(T) = E(W|T)$. Then $E_\theta \phi(T) = \tau(\theta)$ and $Var_\theta \phi(T) \leq Var_\theta W$ for all θ.

- **Proof** We have $\phi(T)$ as an unbiased estimator of $\tau(\theta)$ since

$$\tau(\theta) = E_\theta W = E_\theta\{E(W|T)\} = E_\theta \phi(T).$$

- Also,

$$\begin{aligned} Var_\theta W &= Var_\theta\{E(W|T)\} + E_\theta\{Var(W|T)\} \\ &= Var_\theta\{\phi(T)\} + E_\theta\{Var(W|T)\} \\ &\geq Var_\theta \phi(T). \end{aligned}$$

- We must show that $\phi(T) = E(W|T)$ is a function of only the sample and independent of θ (sufficiency!!).

Sufficiency/Completeness and Unbiasedness

Theorem (Lehmann-Sheffe Theorem)

Let W be any unbiased estimator of $\tau(\theta)$, and let T be a sufficient and complete statistic for θ. Then $\phi(T) = E(W|T)$ is the UMVUE for $\tau(\theta)$ and is unique.

- **Proof** Assume both W_1 and W_2 are unbiased estimator of $\tau(\theta)$.
- If we let $\phi_1(T) = E(W_1|T)$ and $\phi_2(T) = E(W_2|T)$, then

$$E\{\phi_1(T) - \phi_2(T)\} = E(W_1) - E(W_2) = 0.$$

- By the definition of completeness, $\phi_1 - \phi_2$ is a zero function.
- Hence $\phi_1(T) = \phi_2(T)$ (uniqueness).

Find UMVUE

- Method 1:
  - Find an unbiased estimator W for $\tau(\theta)$.
  - Look for a complete sufficient statistic for θ.
  - Derive $\phi(t) = E(W|T = t)$.
  - Then $\phi(T)$ is the UMVUE of $\tau(\theta)$.
- Method 2:
  - **Theorem 7.3.23** Let T be a complete sufficient statistic for a parameter θ, and let $\phi(T)$ be any estimator based only on T. Then $\phi(T)$ is the unique best unbiased estimator of its expected value.
  - Adjusting a complete sufficient statistic to be unbiased gives the UMVUE.

Find UMVUE (cont'd)

Example Assume $X_1, \cdots, X_n$ are iid and follow Poisson(θ).

(1) Show that $I(X_1 = 0)$ is an unbiased estimator for $e^{-\theta}$.

(2) Find UMVUE for $e^{-\theta}$.

Solution

- Since $E\{I(X_1 = 0)\} = P(X_1 = 0) = e^{-\theta}$, $I(X_1 = 0)$ is an unbiased estimator for $e^{-\theta}$.
- Since the Poisson distribution belongs to an exponential family, $\sum_{i=1}^n X_i$ is a complete sufficient statistic.
- By the Lehmann-Scheffe Theorem, we know

$$\phi\left(\sum_{i=1}^n X_i\right) = E\left\{I(X_1 = 0) | \sum_{i=1}^n X_i\right\}$$

is the UMVUE for $e^{-\theta}$.

Find UMVUE (cont'd)

$$\phi(t) = E\left\{I(X_1 = 0) | \sum_{i=1}^{n} X_i = t\right\} = P\left(X_1 = 0 | \sum_{i=1}^{n} X_i = t\right)$$
$$= \frac{P\left(X_1 = 0, \sum_{i=1}^{n} X_i = t\right)}{P\left(\sum_{i=1}^{n} X_i = t\right)} = \frac{P\left(X_1 = 0\right) P\left(\sum_{i=2}^{n} X_i = t\right)}{P\left(\sum_{i=1}^{n} X_i = t\right)}$$
$$= \left(1 - \frac{1}{n}\right)^{t}.$$

- One can conclude $\phi(\sum_{i=1}^{n} X_i) = (1 - 1/n)^{\sum_{i=1}^{n} X_i}$ is the UMVUE for $e^{-\theta}$.
- What is the MLE for $e^{-\theta}$?
- What does the $\phi(\sum_{i=1}^{n} X_i)$ converge to when $n \to \infty$?

Hypothesis Testing

Feng-Chang Lin

Department of Biostatistics
University of North Carolina at Chapel Hill

flin@bios.unc.edu

(C&B §8)

Introduction

- **Example** Current standard treatment for a given disease has success probability 0.7. A new drug has success probability θ (unknown). Is the new drug better than the current treatment?
- Hypothesis: $H_0 : \theta \leq 0.7$ and $H_1 : \theta > 0.7$.
- A *hypothesis* is a statement about a population parameter.
- The parameter space is divided into two disjoint sets:

$$\Theta = \Theta_0 \cup \Theta_0^c$$

- The *null hypothesis* is $H_0 : \theta \in \Theta_0$.
- The *alternative hypothesis* is $H_1 : \theta \in \Theta_0^c$.

Introduction (cont'd)

- Assume $X_1, \ldots, X_n \sim \text{Bernoulli}(\theta)$:
  - **Simple versus simple** $\Theta = \{1/4, 3/4\}$, $\Theta_0 = \{1/4\}$, $\Theta_0^c = \{3/4\}$.
  - **Simple versus composite, one-sided** $\Theta = [1/4, 1]$, $\Theta_0 = \{1/4\}$, $\Theta_0^c = (1/4, 1]$.
  - **Simple versus composite, two-sided** $\Theta = [0, 1]$, $\Theta_0 = \{1/4\}$, $\Theta_0^c = [0, 1/4) \cup (1/4, 1]$.
  - **Composite versus composite** $\Theta = [0, 1]$, $\Theta_0 = [0, 1/4]$, $\Theta_0^c = (1/4, 1]$.
  - **Composite versus composite** $\Theta = [0, 1/4] \cup [3/4, 1]$, $\Theta_0 = [0, 1/4]$, $\Theta_0^c = [3/4, 1]$.

Hypothesis Testing

- *Hypothesis testing*: Use data to decide whether to reject H_0 as false or accept H_0 as true (do not reject H_0).
- A *hypothesis testing procedure* is a rule that specifies for which values of $\boldsymbol{X}$ are to reject H_0 or not.
- *Test function*: $\delta(\boldsymbol{X})$ is either 0 or 1.
- Decision rule: If $\delta(\boldsymbol{X}) = 1$, H_0 is rejected; If $\delta(\boldsymbol{X}) = 0$, H_0 is not rejected.

Rejection Region

- $\delta(\boldsymbol{X})$ divides the sample space into two regions.
- The *rejection region* or *critical region* R is the region over which $\delta(\boldsymbol{x}) = 1$, and H_0 is rejected.
- The acceptance region is the region R^c (the complement of R) over which $\delta(\boldsymbol{x}) = 0$, and H_0 is accepted.

$$R = \{\boldsymbol{x} : \delta(\boldsymbol{x}) = 1\}, \text{ and, } R^c = \{\boldsymbol{x} : \delta(\boldsymbol{x}) = 0\}.$$

- *Type I error*: Reject H_0 when it is true.
- Size of the test (the largest type-I error one can make):

$$\alpha = \sup_{\theta \in \Theta_0} P(\delta(\boldsymbol{x}) = 1).$$

- *Type II error*: Do not reject H_0 when it is false.

Likelihood Ratio Test

- The *likelihood ratio statistic* for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_0^c$ is

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\theta \in \Theta_0} L(\theta|\boldsymbol{x})}{\sup_{\theta \in \Theta} L(\theta|\boldsymbol{x})}$$

- let $\hat{\theta}_0$ denote the restricted MLE over Θ_0, and Let $\hat{\theta}$ denote the unrestricted MLE over $\Theta = \Theta_0 \cup \Theta_0^c$.
- The likelihood ratio statistic:

$$\lambda(\boldsymbol{x}) = \frac{L(\hat{\theta}_0|\boldsymbol{x})}{L(\hat{\theta}|\boldsymbol{x})}.$$

- A *likelihood ratio test* (LRT) is any test with

$$R = \{\boldsymbol{x} : \lambda(\boldsymbol{x}) \leq c\} \text{ for some } c \in [0, 1].$$

Likelihood Ratio Test (cont'd)

- A test with test function $\delta(\boldsymbol{x}) = I(\lambda(\boldsymbol{x}) \leq c)$ for some $c \in [0, 1]$.
- Use the test size, say α, to find c, where

$$\alpha = \sup_{\theta \in \Theta_0} P(\lambda(\boldsymbol{X}) \leq c).$$

- However, we usually do not know about the distribution of $\lambda(\boldsymbol{X})$.
- We intend to find an equivalent region using unrestricted MLE $\hat{\theta}$ with

$$R = \{\boldsymbol{x} : \lambda(\boldsymbol{X}) \leq c\} \Longleftrightarrow R^* = \{\boldsymbol{x} : \hat{\theta} \geq c^* \ \text{ or } \ \hat{\theta} \leq c^*\}.$$

- $\hat{\theta} \geq c^*$ or $\hat{\theta} \leq c^*$ follows the direction of H_1.

LRT under Normal Distribution

- Let $X_1, \cdots, X_n$ be iid $\mathrm{N}(\theta, 1)$. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.
- Under H_0, θ_0 is a fixed number determined by the researcher, so the numerator of $\lambda(\boldsymbol{x})$ is

$$\sup_{\theta \in \Theta_0} L(\theta|\boldsymbol{x}) = L(\theta_0|\boldsymbol{x}).$$

- Under the unrestricted parameter space $\Theta = \Theta \cup \Theta^c$, the MLE is $\bar{X}$, so the denominator of $\lambda(\boldsymbol{x})$ is

$$\sup_{\theta \in \Theta} L(\theta|\boldsymbol{x}) = L(\bar{x}|\boldsymbol{x}).$$

LRT under Normal Distribution (cont'd)

- The LRT statistic is

$$\lambda(\boldsymbol{x}) = \frac{(2\pi)^{-n/2}\exp\{-\sum_{i=1}^n (x_i-\theta_0)^2/2\}}{(2\pi)^{-n/2}\exp\{-\sum_{i=1}^n (x_i-\bar{x})^2/2\}}$$
$$= \exp\left\{\left[-\sum_{i=1}^n (x_i-\theta_0)^2 + \sum_{i=1}^n (x_i-\bar{x})^2\right]/2\right\}.$$

- Since $\sum_{i=1}^n (x_i-\theta_0)^2 = \sum_{i=1}^n (x_i-\bar{x})^2 + n(\bar{x}-\theta_0)^2$, the LRT statistic is simplified to

$$\lambda(\boldsymbol{x}) = \exp\{-n(\bar{x}-\theta_0)^2/2\}.$$

- The rejection region is $\{\boldsymbol{x} : \lambda(\boldsymbol{x}) \le c\}$, which can be written by

$$\{\boldsymbol{x} : |\bar{x}-\theta_0| \ge \sqrt{-2(\log c)/n}\}.$$

LRT under Exponential Distribution

- Let $X_1, \cdots, X_n$ be a random sample from an exponential distribution with pdf

$$f(x|\theta) = \begin{cases} e^{-(x-\theta)} & x \geq \theta \\ 0 & x < \theta, \end{cases}$$

where $-\infty < \theta < \infty$. The likelihood function is

$$L(\theta|\boldsymbol{x}) = \begin{cases} e^{-\sum_{i=1}^{n} x_i + n\theta} & \theta \leq x_{(1)} \\ 0 & \theta > x_{(1)}. \end{cases}$$

- Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, where θ_0 is a value specified by the researcher.
- Unrestricted MLE (denominator) is more straightforward. The unrestricted maximum of $L(\theta|x)$ is $L(x_{(1)}|\boldsymbol{x}) = e^{-\sum x_i + n x_{(1)}}$.

LRT under Exponential Distribution (cont'd)

- Under H_0, finding maximum of $L(\theta|\boldsymbol{x})$ is more complicated. Drawing $L(\theta|\boldsymbol{x})$ helps.
- If $x_{(1)} \leq \theta_0$, the numerator of $\lambda(\boldsymbol{x})$ is also $L(x_{(1)}|\boldsymbol{x})$.
- If $x_{(1)} > \theta_0$, the numerator of $\lambda(\boldsymbol{x})$ is $L(\theta_0|\boldsymbol{x})$.
- Therefore, the likelihood ratio test statistic is

$$\lambda(\boldsymbol{x}) = \begin{cases} 1 & x_{(1)} \leq \theta_0 \\ e^{-n(x_{(1)}-\theta_0)} & x_{(1)} > \theta_0. \end{cases}$$

- One rejects H_0 if $\lambda(\boldsymbol{x}) \leq c$.
- The rejection region $\{\boldsymbol{x} : x_{(1)} \geq \theta_0 - \log(c)/n\}$.

Evaluating Tests

- The *power function* of a hypothesis test is

$$\beta(\theta) = P_\theta(\boldsymbol{X} \in R) = E_\theta \delta(\boldsymbol{X})$$

- *Type I error*: $\beta(\theta)$, $\theta \in \Theta_0$.
- *Type II error*: $1 - \beta(\theta)$, $\theta \in \Theta_0^c$.
- A **size** α test if

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha.$$

- A **level** α test if

$$\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha.$$

Power Function under Normal Distribution

- Let $X_1, \ldots, X_n$ be a random sample from N(θ, σ^2) with known σ^2.
- To test $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, one would reject H_0 if $(\bar{X} - \theta_0)/(\sigma/\sqrt{n}) > c$ by LRT.
- The power function of this test is

$$\begin{aligned}
\beta(\theta) &= P_\theta\left(\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} > c\right) = P_\theta\left(\frac{\bar{X} - \theta}{\sigma/\sqrt{n}} > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \\
&= P_\theta\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right).
\end{aligned}$$

- $\lim_{\theta \to -\infty} \beta(\theta) = 0$ and $\lim_{\theta \to \infty} \beta(\theta) = 1$
- $\beta(\theta_0) = \alpha$ if $\Phi(c) = 1 - \alpha$.

Power Function under Binomial Distribution

- $X \sim \text{Binomial}(3, \theta)$, $\Theta = (0, 1)$,
- $H_0 : \theta \leq 1/4$ versus $H_1 : \theta > 1/4$.
- The test defined by $\delta(x) = I(x = 3)$ has a power function

$$\beta(\theta) = P_\theta(X = 3) = \theta^3.$$

- The size of $\delta(x)$ is $\sup_{\theta \in \Theta_0} \beta(\theta) = \beta(1/4) = 1/64$.
- Another test defined by $\delta^*(x) = I(x \geq 2)$ has a power function

$$\beta^*(\theta) = P_\theta(X \in \{2, 3\}) = 3\theta^2(1 - \theta) + \theta^3.$$

- The size of $\delta^*(x)$ is $\beta^*(1/4) = 10/64$.
- Clearly, $\beta^*(\theta) > \beta(\theta)$ for all $\theta \in (0, 1)$.

Size of a Binomial Test

- $X \sim \text{Binomial}(3, \theta)$, $\Theta = \{1/4, 3/4\}$.
- $H_0 : \theta = \theta_0 = 1/4$ versus $H_1 : \theta = \theta_1 = 3/4$.
- Under H_0, $P_{\theta_0}(X = 0) = 27/64$, $P_{\theta_0}(X = 1) = 27/64$, $P_{\theta_0}(X = 2) = 9/64$, and $P_{\theta_0}(X = 3) = 1/64$.
- Any test function $\delta(x)$ will simply partition the set $\{0, 1, 2, 3\}$ into two subsets.
- Hence, no matter what $\delta(X)$ is, the test size

$$\sup_{\theta \in \Theta_0} \beta(\theta) = P_{\theta_0}(\delta(X) = 1)$$

  will be the sum of one or more of the numbers in $\{0, 27/64, 27/64, 9/64, 1/64\}$.
- Can we have the test size exactly equals $\alpha = 0.05$?

Nonexistence of a Size α Test

- A size α test may not always exist (for example, discreteness).
- Solutions:
  (1) Practical: Settle for a size α^* test with α^* being the largest possible size that is less than or equal to α.
  (2) Mathematical: Randomized tests. Find c such that $\alpha^* + c(1 - \alpha^*) = \alpha$. If the test with size α^* does not reject H_0, draw $U \sim \text{Uniform}(0, 1)$ and reject H_0 if $U < c$.

Desirable Properties of Tests

- Error probabilities as small as possible.
- Error probabilities that are uniformly 0 are impossible except in trivial cases.
- **Example of a trivial case**: $X \sim \text{Bernoulli}(\theta)$, $\Theta = \{0, 1\}$. If $H_0 : \theta = 0$ against $H_1 : \theta = 1$.
- The test $\delta(X) = X$ has error probabilities uniformly 0. Why?

Uniformly Most Powerful (UMP) Level α Test

- Fix type I error at α, then minimize type II error uniformly in θ.
- Restrict to the class of level α tests, then find the uniformly most powerful test.
- **Neyman-Pearson Lemma**: X (scalar or vector) has pdf or pmf $f(x|\theta)$, $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. Suppose a test has a rejection region

$$R = \left\{ x \,\middle|\, \frac{f(x|\theta_1)}{f(x|\theta_0)} > c \right\},$$

and acceptance region

$$R^c = \left\{ x \,\middle|\, \frac{f(x|\theta_1)}{f(x|\theta_0)} < c \right\},$$

for some $c \geq 0$ and has size $\alpha = P_{\theta_0}(X \in R)$.

UMP Level α Test (cont'd)

- Then,
  (a) Any such test is a UMP level α test.
  (b) If such a test exists with $c > 0$ then every UMP size α test has the same test function (except on a set that has probability 0).
- Proof of (a): Given a level α test $\delta^*(x)$, we want to show that $\beta(\theta_1) - \beta^*(\theta_1) \geq 0$. The inequality

$$\{\delta(x) - \delta^*(x)\}\{f(x|\theta_1) - cf(x|\theta_0)\} \geq 0.$$

  holds for each of the four cases: $\delta(x)$, $\delta^*(x) = 0, 1$.
- Integrating out x gives

$$\beta(\theta_1) - c\beta(\theta_0) - \beta^*(\theta_1) + c\beta^*(\theta_0) \geq 0,$$

  which can be written as

$$\beta(\theta_1) - \beta^*(\theta_1) \geq c\{\beta(\theta_0) - \beta^*(\theta_0)\}.$$

UMP Level α Test (cont'd)

- Since $\beta(\theta_0) = \alpha$, $\beta^*(\theta_0) \leq \alpha$ and $c \geq 0$, it follows that $c\{\beta(\theta_0) - \beta^*(\theta_0)\} \geq 0$ and

$$\beta(\theta_1) \geq \beta^*(\theta_1).$$

- **Example** $X_1, \ldots, X_n \sim \mathrm{N}(\theta, 1)$. Find the UMP level α test for $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1 > \theta_0$ (simple versus simple).
- This test is also UMP test for $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ (simple versus composite) since the rejection region does not depend on $\theta_1 > \theta_0$.

UMP Level α Test (cont'd)

- The statement of the Neyman-Pearson Lemma does not say what to do (reject or accept H_0) if

$$X \in \left\{ x \,\middle|\, \frac{f(x|\theta_1)}{f(x|\theta_0)} = c \right\}.$$

- If X is continuous, the probability of this event is zero, and we do not need to worry about it.
- If X is discrete, the event may or may not have positive probability.
- If it does have positive probability, the lemma does not say anything about such tests.
- The implication is that, when deriving UMP tests based on discrete X, we simply avoid using such values of c.

Monotone Likelihood Ratio (MLR)

- The MLR property is said to hold if the likelihood ratio

$$\frac{L(\theta_2|x)}{L(\theta_1|x)} = \frac{f_X(x|\theta_2)}{f_X(x|\theta_1)},$$

depends on x only through a statistic $T(x)$, and is monotone increasing function of $T(x)$ for every $\theta_2 > \theta_1$.

- We will say that the likelihood has a MLR property in $T(X)$.
- **Example**: $X_1, \cdots, X_n$ are iid Poisson(θ), $\theta > 0$. The likelihood ratio

$$\frac{f_X(x|\theta_2)}{f_X(x|\theta_1)} = \exp\left\{ \left(\log\frac{\theta_2}{\theta_1}\right)\left(\sum_{i=1}^n x_i\right) - n(\theta_2 - \theta_1)\right\},$$

is clearly a monotone increasing function in $T(X) = \sum_{i=1}^n X_i$ since $\log(\theta_2/\theta_1) > 0$ for all $\theta_2 > \theta_1 > 0$.

Karlin-Rubin Theorem

- Consider testing $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$.
- Suppose that T is *sufficient*, and the *MLR property holds*, then $\delta(X) = I(T > c)$ defines a UMP level α test.
- The theorem can be restated for the reversed testing problem.
- For testing $H_0 : \theta \geq \theta_0$ against $H_1 : \theta < \theta_0$, a UMP level α test has test function $\delta(X) = I(T < t_0)$.
- The value of t_0 needs to be chosen so that the test has the desired size α in the continuous case.
- Or, the largest possible size $\alpha^* \leq \alpha$ in the discrete case.

Unbiased Tests

- Uniformly most powerful (UMP) level α tests do not always exist.
- **Example 8.3.19** Let $X_1, \cdots, X_n$ be iid N(θ, σ^2), σ^2 known. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.
- A size α test that rejects for large values of $\bar{X}$ is most powerful for $\theta > \theta_0$ but not for $\theta < \theta_0$.
- One way out of the nonexistence of UMP is to restrict to smaller classes of tests.

Unbiased Tests (cont'd)

- We define *unbiased tests* as:

$$\sup_{\theta \in \Theta_0} \beta(\theta) \leq \inf_{\theta \in \Theta_0^c} \beta(\theta).$$

- **Example** X is a random sample of size n from the N$(\theta, 1)$ distribution. $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.
- A uniformly most powerful (UMP) unbiased level α test is defined by $\delta(X) = I(\sqrt{n}|\bar{X} - \theta_0| > c)$ for some $c \geq 0$.
- Since $\Theta_0 = \{\theta_0\}$, the size of the test is $E_{\theta_0}\delta(X) = 2\Phi(-c)$.
- If we want the size to be 0.05, we choose $c = 1.96$.

P-value

- **Definition**: Given a sample $\boldsymbol{X}$, a *p-value* is a test statistic $p(\boldsymbol{X}) \in [0, 1]$ such that small values support H_1 over H_0.
- A p-value is *valid* if, for every $\theta \in \Theta_0$, and every $0 \leq \alpha \leq 1$,

$$P_\theta(p(\boldsymbol{X}) \leq \alpha) \leq \alpha.$$

- That means, if $p(\boldsymbol{X})$ is a valid p-value, a test that rejects H_0 if $p(\boldsymbol{X}) \leq \alpha$ is a level α test.

P-value (cont'd)

- **Example** $\boldsymbol{X}$ is a random sample of size n from the $\mathrm{N}(\theta, 1)$ distribution. $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$.
- Let $\boldsymbol{x}$ be an observed sample and $\bar{x}$ be the observed sample mean.
- Consider a function:

$$p(\boldsymbol{x}) = 1 - \Phi\left(\sqrt{n}(\bar{x} - \theta_0)\right).$$

- $p(\boldsymbol{X})$ is a statistic since θ_0 is a specified known number (not an unknown parameter), and n is also known.
- $p(\boldsymbol{x})$ can be interpreted as the probability that the random variable $\bar{X}$ exceeds the observed value $\bar{x}$ if $\theta = \theta_0$.

P-value (cont'd)

- $p(\boldsymbol{x})$ is decreasing in $\bar{x}$, so large values of $\bar{x}$, which would support H_1 over H_0, go with small values of $p(\boldsymbol{x})$.
- Also,

$$\begin{aligned} P_\theta(p(\boldsymbol{X}) \leq \alpha) &= P_\theta(\bar{X} \geq \theta_0 + \Phi^{-1}(1-\alpha)/\sqrt{n}) \\ &= P_\theta(\sqrt{n}(\bar{X}-\theta) \geq \sqrt{n}(\theta_0 - \theta) + \Phi^{-1}(1-\alpha)) \\ &= 1 - \Phi(\sqrt{n}(\theta_0 - \theta) + \Phi^{-1}(1-\alpha)). \end{aligned}$$

- That means $P_{\theta_0}(p(\boldsymbol{X}) \leq \alpha) = \alpha$, and $P_\theta(p(\boldsymbol{X}) \leq \alpha) < \alpha$ for $\theta < \theta_0$.
- Hence, this is a valid p-value, and the test with test function $\delta(X) = I(p(\boldsymbol{X}) \leq \alpha)$ has size α.

P-value (cont'd)

- In general, if a hypothesis test rejects $H_0 : \theta = \theta_0$ for large values of a statistic $T(X)$, the p-value can be defined to be

$$p(x) = P_{\theta_0}(T(X) \geq T(x)),$$

where $T(x)$ is observed value of $T(X)$.

Union-Intersection Test

- Union-intersection and intersection-union tests are ways of combining many simpler hypothesis tests into a single more complicated test.
- In some problems, the null hypothesis is the intersection of two or more simpler null hypotheses,

$$H_0 : \theta \in \bigcap_{j \in J} \Theta_j \text{ against } H_1 : \theta \in \bigcup_{j \in J} \Theta_j^c.$$

- J may be finite or infinite.

Union-Intersection Test (cont'd)

- Suppose that for each individual problem of testing

$$H_{0j} : \theta \in \Theta_j \ \text{ against } \ H_{1j} : \theta \in \Theta_j^c,$$

  $j \in J$, with rejection region R_j. Then, the union-intersection test has rejection region

$$R = \bigcup_{j \in J} R_j$$

- That is, the union-intersection test rejects H_0 if any of the individual hypotheses H_{0j} is rejected.
- The null hypothesis is an intersection while the rejection region is a union.

Example for Union-Intersection Test

- $X \sim \mathrm{N}(\theta, 1)$. Test $H_0 : \theta = 1$ against $H_1 : \theta \neq 1$.
- Suppose that the simpler hypothesis tests are

$$H_{01} : \theta \geq 1 \text{ against } H_{11} : \theta < 1,$$

with rejection region $R_1 = \{x : x < a\}$, and

$$H_{02} : \theta \leq 1 \text{ against } H_{12} : \theta > 1,$$

with rejection region $R_2 = \{x : x > b\}$, where a and b are specified constants with $a < b$.

- Then the union-intersection test has critical region

$$R = R1 \bigcup R2 = \{x : x \notin [a, b]\}.$$

Intersection-Union Test

- In intersection-union tests, the null hypothesis is a union while the rejection region is an intersection,

$$H_0 : \theta \in \bigcup_{j \in J} \Theta_j \text{ against } H_1 : \theta \in \bigcap_{j \in J} \Theta_j^c,$$

and

$$R = \bigcap_{j \in J} R_j.$$

Example for Intersection-Union Test

- Suppose we observe a pair of random variables for each patient.
- X is an indicator of response to treatment, while Y is an indicator of severe side effects.
- Let $\theta_1 = P(X = 1)$ and $\theta_2 = P(Y = 1)$.
- One may test

$$H_0 : \theta_1 < 0.8 \ \text{ or } \ \theta_2 > 0.15 \ \text{ against } \ H_1 : \theta_1 \geq 0.8 \ \text{ and } \ \theta_2 \leq 0.15.$$

- Suppose that the simpler hypothesis tests are

$$H_{01} : \theta_1 < 0.8 \ \text{ against } \ H_{11} : \theta_1 \geq 0.8,$$

with rejection region $R_1 = \{x : \sum_{i=1}^n x_i > a\}$,

Example for Intersection-Union Test (cont'd)

- and
$$H_{02} : \theta_2 > 0.15 \text{ against } H_{12} : \theta_2 \le 0.15,$$
with rejection region $R_2 = \{y : \sum_{i=1}^n y_i < b\}$.
- Then the intersection-union test has critical region
$$R = R_1 \bigcap R_2 = \{(x, y) : \sum_{i=1}^n x_i > a \text{ and } \sum_{i=1}^n y_i < b\},$$
and it rejects H_0 if the observed (x, y) falls within R, i.e. if both simpler null hypotheses are rejected.

Interval Estimation

Feng-Chang Lin

Department of Biostatistics
University of North Carolina at Chapel Hill

flin@bios.unc.edu

(C&B §9)

Introduction

- **Example 1** Suppose $X_1, \ldots, X_n$ are iid from N$(\theta, 1)$.
- We know that $P_\theta(\bar{X} = \theta) = 0$ since $\bar{X}$ is a continuous random variable.
- Therefore, even though $\bar{X}$ is a good estimator of θ, it is never equal to θ.

Introduction (cont'd)

- **Example 2** $X \sim \text{Binomial}(n, \theta)$, $\theta \in (0, 1)$.
- X/n is the MLE of θ.
- $P_\theta(X/n = \theta)$ will be 0 unless θ is one of $\{1/n, 2/n, \ldots, (n-1)/n\}$.
- If $\theta = i/n$ for some $i \in \{1, 2, \ldots, n-1\}$, then

$$P_\theta(X/n = \theta) = P(X = i) = \binom{n}{i} \left(\frac{i}{n}\right)^i \left(1 - \frac{i}{n}\right)^{n-i}.$$

- This probability can be very small, especially for large n. For example, if $n = 20$, $\theta = 1/2$, then $P_\theta(X/n = \theta)$ is about 0.18, and if $n = 100$, it is about 0.08.

Introduction (cont'd)

- In many situations point estimators have low (or zero) probability of being equal to the parameter they estimate.
- If one considers estimators that are intervals instead of single points, that shortcoming can be overcome.
- In the normal mean problem, the interval $(\bar{X} - 1.96/\sqrt{n}, \bar{X} + 1.96/\sqrt{n})$ has probability 0.95 of containing the true parameter value θ.

Interval Estimator

- **Interval Estimator** $(L(X), U(X))$, where $L(X)$ and $U(X)$ are statistics, $L(X) < U(X)$.
- Denoted by either $(L(X), U(X))$ or $[L(X), U(X)]$.
- One-sided intervals: e.g. $L(X) = -\infty$ or $U(X) = \infty$ (depending on Θ).
- **Coverage probability** for $(L(X), U(X))$:

$$CP(\theta) = P_\theta(\theta \in (L(X), U(X))),$$

  as a function of θ.
- **Confidence Coefficient (Confidence Level)**: $\inf_{\theta \in \Theta} CP(\theta)$.

Interval Estimator (cont'd)

- **Example** $X \sim$ Bernoulli(θ), $\theta \in [0, 1]$. If one has a confidence interval $[0.4, 0.5 + 0.2X]$

$$CP(\theta) = \begin{cases} 0, & 0 \leq \theta < 0.4, \\ 1, & 0.4 \leq \theta \leq 0.5, \\ \theta, & 0.5 < \theta \leq 0.7, \\ 0, & 0.7 < \theta \leq 1. \end{cases}$$

- Confidence coefficient = 0.

How to Find a Confidence Interval

- **Inverting a test**: Consider $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. By inverting the acceptance region of a level α test

  $$A(\theta_0) = \{x : \delta(x, \theta_0, \alpha) = 0\}$$

  with a test function $\delta(x)$, written as $\delta(x, \theta, \alpha)$, one can define

  $$C(x) = \{\theta \in \Theta : \delta(x, \theta, \alpha) = 0\},$$

  as a subset of Θ.
- Then,

  $$\begin{aligned} P_\theta(\theta \in C(x)) &= P_{\theta_0}(\delta(X, \theta_0, \alpha) = 0) \\ &= 1 - P_{\theta_0}(\delta(X, \theta_0, \alpha) = 1) \geq 1 - \alpha. \end{aligned}$$

- Thus $C(x)$ is a $1 - \alpha$ confidence interval of θ.

How to Find a Confidence Interval (cont'd)

- **Example** X is a random sample of size n from $\mathrm{N}(\theta, 1)$. $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ with $\alpha = 0.05$.

$$\delta(X, \theta_0, \alpha) = I(|\bar{X} - \theta_0| > 1.96/\sqrt{n}).$$

- That means,

$$P_{\theta_0}(\bar{X} - 1.96\frac{1}{\sqrt{n}} \leq \theta_0 \leq \bar{X} + 1.96\frac{1}{\sqrt{n}}) = 0.95.$$

- The statement is true for every θ_0. Hence, we can write

$$P_{\theta}(\bar{X} - 1.96\frac{1}{\sqrt{n}} \leq \theta \leq \bar{X} + 1.96\frac{1}{\sqrt{n}}) = 0.95.$$

- Hence, the 0.95 confidence interval is

$$C(x) = (\bar{x} - 1.96/\sqrt{n}, \bar{x} + 1.96/\sqrt{n}).$$

How to Find a Confidence Interval (cont'd)

- **Example** X is a random sample of size n from Exponential(θ).
- To test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, the acceptance region of the likelihood ratio test (LRT) statistic is

$$A(\theta_0) = \left\{ \boldsymbol{x} : \left(\frac{\sum_{i=1}^n x_i}{\theta_0} \right)^n e^{-\sum_{i=1}^n x_i/\theta_0} \geq c \right\}$$

- Inverting this acceptance region gives the $1 - \alpha$ confidence interval

$$C(x) = \left\{ \theta : \left(\frac{\sum_{i=1}^n x_i}{\theta} \right)^n e^{-\sum_{i=1}^n x_i/\theta} \geq c \right\}.$$

- Check C&B on how to find the upper and lower bound for θ.

How to Find a Confidence Interval (cont'd)

- **Lower confidence bounds**: $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$. Inverting a test gives the interval $[L(X), \infty)$.
- **Upper confidence bounds**: $H_0 : \theta = \theta_0$ versus $H_1 : \theta < \theta_0$. Inverting a test gives the interval $(-\infty, U(X)]$.
- **Example** Let $X_1, \cdots, X_n$ be a random sample from $\mathrm{N}(\theta, \sigma^2)$.
- Consider constructing a $1-\alpha$ upper confidence bound for μ.
- The size α test acceptance region is

$$A(\theta_0) = \left\{ \boldsymbol{x} : \frac{\bar{x} - \theta_0}{s/\sqrt{n}} \geq t_{n-1,\alpha} \right\}.$$

- The $1-\alpha$ confidence region (or set) is

$$C(x) = \left\{ \theta : \bar{x} - t_{n-1,\alpha} \frac{s}{\sqrt{n}} \geq \theta \right\}$$

How to Find a Confidence Interval (cont'd)

- **Pivot (Pivotal Quantity)** $Q(X, \theta)$ is a pivot if the distribution of $Q(X, \theta)$ does not depend on θ.
- **Examples**

Family	Density	Pivot
Location	$f(x-\mu)$	$\bar{X}-\mu$
Scale	$\frac{1}{\sigma}f(\frac{1}{\sigma})$	$\frac{\bar{X}}{\sigma}$
Location-Scale	$\frac{1}{\sigma}f(\frac{x-\mu}{\sigma})$	$\frac{\bar{X}-\mu}{\sigma}$

Pivotal Quantity

- **Example** If $X_1, \cdots, X_n$ are iid N(μ, σ^2), then $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ is a pivot.
- If σ^2 is known, we can use this pivot to calculate a confidence interval for μ.
- Let $z_{1-\alpha/2}$ be the $(1 - \alpha/2)$th percentile of a standard norm distribution. One has

$$\begin{aligned} 1 - \alpha &= P\left(-z_{1-\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) \\ &= P\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

- The $1 - \alpha$ confidence interval is $(\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}})$.

Pivotal Quantity (cont'd)

- What if σ^2 is unknown, what pivot can we use to calculate a confidence interval for μ?

Pivotal Quantity (cont'd)

- **Example** X is a random sample of size n from exponential(θ).
- Construct a 95% ($1-\alpha=0.95$) confidence interval for θ.
- This is a scale family. Why?
- Let $Q(X,\theta)=2n\bar{X}/\theta\sim\chi^2_{2n}$. Then,

$$\begin{aligned}1-\alpha=P\left(a<Q(X,\theta)<b\right)&=P\left(a<2n\bar{X}/\theta<b\right)\\&=P\left(2n\bar{X}/b<\theta<2n\bar{X}/a\right).\end{aligned}$$

- Hence, the $1-\alpha$ confidence interval for θ is $(2n\bar{X}/b, 2n\bar{X}/a)$.
- How to choose a and b? One may let $a=F^{-1}(\alpha_1)$ and $b=F^{-1}(1-\alpha_2)$, where $\alpha_1+\alpha_2=\alpha$.

Minimization of Expected Length

- How to choose α_1 and α_2? A convenient choice is $\alpha_1 = \alpha_2 = \alpha/2$.
- One possible criterion is “the shortest interval".
- Since the length can be considered as a function of $\bar{X}$, we may calculate the “expected length"

$$E\left(2n\bar{X}/a - 2n\bar{X}/b\right) = 2n\theta\left(\frac{1}{a} - \frac{1}{b}\right).$$

- We choose a and b (or equivalently, α_1 and α_2) such that the expected length is minimized.
- For a fixed θ, the solution depends on n.
- Examples: for $n = 1$, $\alpha_1 = 0.05$, $\alpha_2 = 0$; for $n = 10$, $\alpha_1 = 0.044, \alpha_2 = 0.006$; for $n = 20$, $\alpha_1 = 0.04$, $\alpha_2 = 0.01$.

Another Example from Scale Family

- **Example** X is a random sample of size n from N(μ, σ^2).
- How do we construct a $1 - \alpha$ confidence interval for σ^2?
- If μ is unknown, the pivot is

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

- What if μ is known? What is the pivot?

Pivoting the CDF

- Suppose T is a statistic with cdf F_T. Using $F_T(t|\theta)$ as a pivot is feasible if $F_T(t|\theta)$ is a decreasing or increasing function in θ for each fixed t.
- If $F_T(t|\theta)$ is a decreasing function of θ, to construct a $1-\alpha$ confidence interval, we find $U(t)$ and $L(t)$ such that

$$P(T \leq t|\theta = U(t)) = \alpha_1, \text{ and } P(T \geq t|\theta = L(t)) = \alpha_2.$$

  with "tail probability" α_1 and α_2 satisfying $\alpha_1 + \alpha_2 = \alpha$.
- One can prove $\{\theta : \alpha_1 \leq F_T(t|\theta) \leq 1-\alpha_2\}$=$\{\theta : L(t) \leq \theta \leq U(t)\}$ (Theorem 9.2.12 in C&B).

Pivoting the CDF (cont'd)

- **Example** If $X_1, \ldots, X_n$ are iid with pdf $f(x|\mu) = e^{-(x-\mu)} I_{[\mu,\infty)}(x)$.
- Then, $Y = X_{(1)}$ is sufficient for μ with pdf

$$f_Y(y|\mu) = ne^{-n(y-\mu)} I_{[\mu,\infty)}(y).$$

- Since the CDF $F_Y(y|\mu) = 1 - e^{-n(y-\mu)}$, $\mu \le y < \infty$, is decreasing in μ, we can have

$$\int_{U(y)}^{y} ne^{-n(u-U(y))} du = \frac{\alpha}{2}, \quad \text{and} \quad \int_{y}^{\infty} ne^{-n(u-L(y))} du = \frac{\alpha}{2}.$$

- The solutions for $L(y)$ and $U(y)$ are

$$L(y) = y + \frac{1}{n}\log(\alpha/2), \quad \text{and} \quad U(y) = y + \frac{1}{n}\log(1-\alpha/2).$$

Pivoting the CDF (cont'd)

- The $1-\alpha$ confidence interval for μ is

$$C(y)=\left\{\mu : y+\frac{1}{n}\log(\alpha/2)\leq \mu \leq y+\frac{1}{n}\log(1-\alpha/2)\right\}.$$

- Can we invert the acceptance region of the LRT test to obtain the confidence interval?
- Can we use the pivotal quantity to obtain the confidence interval? What is the pivot?
- If these intervals are different, which one has a shorter length?
- Check Exercise 9.25 in C&B.

Evaluating Interval Estimators

- **Optimizing the length**: Minimization of $|a - b|$ is generally not easy.
- **(Theorem 9.3.2 in C&B)** For any unimodal density g with mode in $[a, b]$, subject to total tail area $\alpha_1 + \alpha_2 = \alpha$. Then $|a - b|$ is minimized by a and b with $g(a) = g(b)$.
- **Optimizing the expected length**: we have seen the example.
- Check Example 9.3.4 for which the application of Theorem 9.3.2 will not give the shortest confidence interval.

Exact versus Approximate Confidence Intervals

- **Exact confidence interval**:

$$P(L(X) < \theta < U(X)) = 1 - \alpha$$

- **Approximate confidence interval**:

$$P(L(X) < \theta < U(X)) \approx 1 - \alpha$$

- Let $X_1, \ldots, X_n$ be iid $\mathrm{N}(\mu, \sigma^2)$. The $1-\alpha$ confidence interval for μ could be $\bar{X} \pm t_{n-1,1-\alpha/2} S/\sqrt{n}$. Exact or approximate?
- The $1-\alpha$ confidence interval for σ^2 could be $((n-1)S^2/b, (n-1)S^2/a)$ for some a and b. Exact or approximate?

Exact versus Approximate CI (cont'd)

- Let $X_1, \ldots, X_n$ be iid Beroulli(θ). The MLE of θ is $\hat{\theta} = \bar{X}$. According to the CLT,
$$\sqrt{n}(\hat{\theta} - \theta) \to_d N(0, \sigma^2),$$
where
$$\sigma^2 = Var(X_1) = \theta(1 - \theta).$$
- With $\hat{\sigma}^2 = \bar{X}(1 - \bar{X})$, one can construct a $1 - \alpha$ **approximate** confidence interval
$$\bar{X} \pm z_{1-\alpha/2}\sqrt{\frac{\bar{X}(1 - \bar{X})}{n}}.$$

Exact versus Approximate CI (cont'd)

- In fact, an **exact** confidence interval can be constructed but may not have "exactly" $1 - \alpha$ confidence level.
- Check Example 9.2.11 in C&B for another binomial case.
- Check Example 9.2.15 in C&B for a Poisson case.

Large Sample ML-based Methods I

Feng-Chang Lin

Department of Biostatistics
University of North Carolina at Chapel Hill

flin@bios.unc.edu

(C&B §10)

Notations

- $X_1, \ldots, X_n$ be iid random variables from a family indexed by θ.
- Log-likelihood: $\ell(\theta) = \sum_{i=1}^n \ell_i(\theta|x_i)$, where

$$\ell_i(\theta|x_i) = \log f(x_i|\theta).$$

- Score function: $U(\theta) = \sum_{i=1}^n U_i(\theta|x_i)$, where

$$U_i(\theta|x_i) = (\partial/\partial\theta)\ell_i(\theta|x_i).$$

- Observed information: $J(\theta) = \sum_{i=1}^n J_i(\theta|x_i)$, where

$$J_i(\theta|x_i) = -(\partial^2/\partial\theta^2)\ell_i(\theta|x_i).$$

Notations (cont'd)

- Expected information: $I_n(\theta) = nI_1(\theta)$, where

$$I_1(\theta) = EJ_i(\theta|x_i) = E\{-(\partial^2/\partial\theta^2)\ell_i(\theta|x_i)\}.$$

- $\ell_1, \ldots, \ell_n$ are iid.
- $U_1, \ldots, U_n$ are iid mean 0 and variance $I_1(\theta)$.

$$\begin{aligned}
E(U_i) &= E\left\{\frac{\partial}{\partial\theta}\ell_i(\theta|x_i)\right\} = E\left\{\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right\} \\
&= E\left\{\frac{\frac{\partial}{\partial\theta}f(x_i|\theta)}{f(x_i|\theta)}\right\} = \int_{\mathcal{X}}\frac{\partial}{\partial\theta}f(x|\theta)dx = \frac{\partial}{\partial\theta}(1) = 0
\end{aligned}$$

Notations (cont'd)

- You may find the proof of $Var(U_i) = I_1(\theta)$ in Exercise 7.39 in C&B. Here are some outlines:

$$
\begin{aligned}
I_1(\theta) &= -E\left\{\frac{\partial^2}{\partial\theta^2}\log f(x_i|\theta)\right\} = -E\left[\frac{\partial}{\partial\theta}\left\{\frac{\partial}{\partial\theta}\log f(x_i|\theta)\right\}\right] \\
&= -E\left[\frac{\partial}{\partial\theta}\left\{\frac{\frac{\partial}{\partial\theta}f(x_i|\theta)}{f(x_i|\theta)}\right\}\right] = E\left\{\frac{\frac{\partial}{\partial\theta}f(x_i|\theta)}{f(x_i|\theta)}\right\}^2 \\
&= E\left\{\frac{\partial}{\partial\theta}\log f(x_i|\theta)\right\}^2 = Var(U_i)
\end{aligned}
$$

Notations (cont'd)

- $J_1, \ldots, J_n$ are iid mean $I_1(\theta)$.
- $I_1(\theta)$ is the expected (Fisher) information in one observation.
- We call $I_1(\theta)$ **information number**.
- $I_n(\theta) = nI_1(\theta)$ is the expected information in n observation.

Bernoulli Example

- Let $X_1, \ldots, X_n$ be iid Bernoulli(θ), $\theta \in (0, 1)$.
- The log-likelihood is $\ell(\theta) = \sum_{i=1}^{n} \ell_i(\theta|x_i)$, where

$$\ell_i(\theta|x_i) = x_i \log \frac{\theta}{1-\theta} + \log(1-\theta).$$

- The score function is $U(\theta) = \sum_{i=1}^{n} U_i(\theta|x_i)$, where

$$U_i(\theta|x_i) = \frac{x_i}{\theta(1-\theta)} - \frac{1}{1-\theta} = \frac{x_i - \theta}{\theta(1-\theta)}.$$

- The observed information is $J(\theta) = \sum_{i=1}^{n} J_i(\theta|x_i)$

$$J_i(\theta|x_i) = \frac{1}{\theta^2(1-\theta)^2}(x_i - 2x_i\theta + \theta^2).$$

Bernoulli Example (cont'd)

- The expected information is $nI_1(\theta)$, where

$$I_1(\theta) = EJ_i(\theta|x_i) = \frac{1}{\theta(1-\theta)}.$$

- Check: $E\{U_i(\theta|x_i)\} = 0$.
- Check: $Var\{U_i(\theta|x_i)\} = I_1(\theta)$.

Large Sample Properties of MLE

- When $\theta = \theta_0$ and $n \to \infty$,

$$\sqrt{n}\left\{\frac{1}{n}U(\theta_0) - 0\right\} = \frac{1}{\sqrt{n}}U(\theta_0) \to_d N\{0, I_1(\theta_0)\}.$$

- $n^{-1}J(\theta_0) \to_p I_1(\theta_0)$.
- Let $K(\theta_0) = \sum_{i=1}^{n} K_i(\theta_0|x_i)$, where $K_i(\theta|x_i) = (\partial^3/\partial\theta^3)\ell_i(\theta|x_i)$.
- $n^{-1}\sum_{i=1}^{n} K_i(\theta_0|x_i) \to_p E\{K_i(\theta_0)|x_i\}$.

Large Sample Properties of MLE (cont'd)

- Let $\hat{\theta}$ be MLE of θ based on n observations (also denoted by $\hat{\theta}$).
- **Theorem (Consistency)**:

$$\hat{\theta} \rightarrow_p \theta_0 \text{ as } n \rightarrow \infty.$$

- **Theorem (Asymptotic Normality)**:

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d N\{0, I_1(\theta_0)^{-1}\} \text{ as } n \rightarrow \infty.$$

- This implies: $\tau(\hat{\theta}) \rightarrow_p \tau(\theta_0)$, and

$$\sqrt{n}\left\{\tau(\hat{\theta}) - \tau(\theta_0)\right\} \rightarrow_d N\left[0, \frac{\{\tau'(\theta_0)\}^2}{I_1(\theta_0)}\right]$$

- What method did we use? It requires $\tau(\cdot)$ is a continuous function and $\tau'(\theta_0) \neq 0$.

Asymptotic Efficiency

- T_n is an “asymptotically efficient" estimator of $\tau(\theta)$ if
$$\sqrt{n}\{T_n - \tau(\theta)\} \to_d N(0, v(\theta)),$$
and
$$v(\theta) = \frac{\{\tau'(\theta_0)\}^2}{I_1(\theta_0)}.$$
- That means, asymptotic variance = CRLB
- MLE $\tau(\hat{\theta})$ is asymptotically efficient.

Asymptotic Relative Efficiency

- Definitions: If

$$\sqrt{n}(T_{1n} - \theta) \to_d N(0, \sigma_1^2), \text{ and}$$
$$\sqrt{n}(T_{2n} - \theta) \to_d N(0, \sigma_2^2), \text{ as } n \to \infty.$$

- The asymptotic relative efficiency of T_{1n} with respect to T_{2n} is

$$\text{ARE}(T_{1n}, T_{2n}) = \frac{\sigma_2^2}{\sigma_1^2}.$$

Asymptotic Relative Efficiency (cont'd)

- **Example**: $X_1, \ldots, X_n$ be iid logistic(θ) with $EX_i = \theta$ and $VarX_i = \pi^2/3$.
- We have

$$\sqrt{n}(\bar{X} - \theta) \rightarrow_d N(0, \pi^2/3), \text{ and}$$
$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow_d N(0, 3), \text{ as } n \rightarrow \infty.$$

  by CLT and asymptotic normality of the MLE, respectively.
- Note:

$$I_1(\theta) = \frac{1}{3} = -E\left\{\frac{\partial^2}{\partial\theta^2} \log f(x|\theta)\right\}.$$

- $\text{ARE}(\bar{X}, \hat{\theta}) = \frac{3}{\pi^2/3} = 9/\pi^2 \approx 0.91$.

Asymptotic Distribution of LRT

- The likelihood ratio statistic can be shown as

$$-2\log\lambda(\boldsymbol{x}) = 2\{\ell(\hat{\theta}) - \ell(\theta_0)\}.$$

- Taylor expansion of $\ell(\theta_0)$ around $\hat{\theta}$ leads to

$$\ell(\theta_0) = \ell(\hat{\theta}) + (\theta_0 - \hat{\theta})U(\hat{\theta}) - \frac{1}{2}(\theta_0 - \hat{\theta})^2 J(\hat{\theta}) + \frac{1}{6}(\theta_0 - \hat{\theta})^3 K(\theta^*).$$

- This implies

$$\begin{aligned}-2\log\lambda(\boldsymbol{x}) &= 2\{\ell(\hat{\theta}) - \ell(\theta_0)\}\\ &= \left\{\sqrt{n}(\hat{\theta} - \theta_0)\sqrt{\frac{J(\hat{\theta})}{n}}\right\}^2 + \frac{1}{3\sqrt{n}}\left\{\sqrt{n}(\hat{\theta} - \theta_0)\right\}^3 \frac{K(\theta^*)}{n}\end{aligned}$$

Asymptotic Distribution of LRT (cont'd)

- What is the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$?
- What does $\sqrt{J(\hat{\theta})/n}$ converge in probability to?
- What does the first term converge in distribution to?
- One can see that $\frac{1}{3\sqrt{n}}$ converges to 0 and $K(\theta^*)/n$ converges almost surely to $EK_1(\theta_0)$.
- What does the second term converge in probability to?
- Combining the convergence of both terms, we may prove

$$-2\log\lambda(\boldsymbol{x}) \to_d \chi_1^2 \ \text{ as } \ n \to \infty.$$

- One may have **Signed Likelihood Ratio Statistic**

$$\text{sign}(\hat{\theta} - \theta_0)\sqrt{-2\log\lambda(\boldsymbol{x})} \to_d N(0,1) \ \text{ as } \ n \to \infty.$$

Hypothesis Tests in Large Samples

- When testing $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$, we have
  (a) Likelihood ratio test: under H_0,
  $$2\{\ell(\hat{\theta}) - \ell(\theta_0)\} = -2\log\lambda(\boldsymbol{x}) \to_d \chi_1^2, \quad \text{as} \quad n \to \infty.$$
  (b) Score test: under H_0,
  $$\frac{U(\theta_0)}{\sqrt{nI_1(\theta_0)}} = \frac{U(\theta_0)}{\sqrt{I_n(\theta_0)}} \to_d N(0,1).$$
  (c) Wald test: under H_0, we have two options
  $$\sqrt{nI_1(\hat{\theta})}(\hat{\theta} - \theta_0) \to_d N(0,1), \quad \text{as} \quad n \to \infty,$$
  and
  $$\sqrt{J(\hat{\theta})}(\hat{\theta} - \theta_0) \to_d N(0,1), \quad \text{as} \quad n \to \infty,$$

Bernoulli Example

- Let $X_1, \dots, X_n$ be iid Bernoulli(θ), $\theta \in (0, 1)$.
- The log-likelihood is $\ell(\theta) = \sum_{i=1}^n \ell_i(\theta|x_i)$, where

$$\ell_i(\theta|x_i) = x_i \log \frac{\theta}{1-\theta} + \log(1-\theta).$$

- The score function is $U(\theta) = \sum_{i=1}^n U_i(\theta|x_i)$, where

$$U_i(\theta|x_i) = \frac{x_i}{\theta(1-\theta)} - \frac{1}{1-\theta} = \frac{x_i - \theta}{\theta(1-\theta)}.$$

- The observed information is $J(\theta) = \sum_{i=1}^n J_i(\theta|x_i)$

$$J_i(\theta|x_i) = \frac{1}{\theta^2(1-\theta)^2}(x_i - 2x_i\theta + \theta^2).$$

Bernoulli Example (cont'd)

- Information number: $I_1(\theta) = E\{J_1(\theta|x_1)\} = \theta^{-1}(1-\theta)^{-1}$.
- To test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$:
- Under the null hypothesis, we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \to_d N(0, I_1(\theta_0)^{-1}), \text{ as } n \to \infty.$$

- Hence, the Wald test statistic is

$$\frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sqrt{I_1(\hat{\theta})^{-1}}} = \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sqrt{\hat{\theta}(1-\hat{\theta})}}.$$

- Reject H_0 if

$$\left| \frac{\sqrt{n}(\bar{x} - \theta_0)}{\sqrt{\bar{x}(1-\bar{x})}} \right| \geq z_{1-\alpha/2}.$$

Bernoulli Example (cont'd)

- By the large sample normality of the score function, we have

$$n^{-1/2}U(\theta_0) \to_d N(0, I_1(\theta_0)).$$

- Hence, the score test statistic is

$$\frac{U(\theta_0)}{\sqrt{nI_1(\theta_0)}} = \frac{\sum_{i=1}^n (x_i - \theta_0)/\{\theta_0(1-\theta_0)\}}{\sqrt{n\theta_0^{-1}(1-\theta_0)^{-1}}} = \frac{\sqrt{n}(\bar{x} - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}}.$$

- Reject H_0 if

$$\left| \frac{\sqrt{n}(\bar{x} - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}} \right| \geq z_{1-\alpha/2}.$$

Bernoulli Example (cont'd)

- Based on LRT, we reject H_0 if $-2\log\lambda(\boldsymbol{x}) \geq \chi^2_{1,1-\alpha}$.
- Note that, in this example,

$$I_n(\hat{\theta}) = nI_1(\hat{\theta}) = \frac{n}{\bar{x}(1-\bar{x})},$$

and

$$J(\hat{\theta}) = \sum_{i=1}^{n} J_i(\theta|x_i) = \frac{1}{\bar{x}^2(1-\bar{x})^2}\sum_{i=1}^{n}(x_i - 2x_i\bar{x} + \bar{x}^2) = \frac{n}{\bar{x}(1-\bar{x})}.$$

- Here $I_n(\hat{\theta}) = J(\hat{\theta})$, not true in general.

Numerical Example

- Test $H_0 : \theta = 0.5$ versus $H_1 : \theta \neq 0.5$ given $\alpha = 0.05$.
- $n = 10$, $\sum x_i = 3$, $\hat{\theta} = \bar{x} = 0.3$.
- Likelihood ratio test:

$$-2\log\lambda(\boldsymbol{x}) = 2(10)\left(0.3\log\frac{0.3}{0.5} + 0.7\log\frac{0.7}{0.5}\right)$$
$$\approx 1.646 < \chi^2_{1,1-\alpha} = 3.84.$$

- Score test:

$$\left|\frac{\sqrt{10}(0.3 - 0.5)}{\sqrt{0.5(1 - 0.5)}}\right| \approx 1.265 < z_{1-\alpha/2} = 1.96$$

- Wald test:

$$\left|\frac{\sqrt{10}(0.3 - 0.5)}{\sqrt{0.3(1 - 0.3)}}\right| \approx 1.38 < z_{1-\alpha/2} = 1.96$$

Intervals

- How do we derive interval estimators?
- Inverting acceptance regions:

$$\{\theta_0 : \delta(\boldsymbol{X}, \theta_0, \alpha) = 0\},$$

where δ may be one of the three tests.

Intervals: Bernoulli Example

- Likelihood ratio:

$$\left\{\theta_0 : 20\left[0.3\log\frac{0.3}{\theta_0} + 0.7\log\frac{0.7}{1-\theta_0}\right] \leq 3.84\right\} = (0.085, 0.606).$$

- Score test:

$$\left\{\theta_0 : \left|\frac{\sqrt{10}(0.3-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}}\right| \leq 1.96\right\} = (0.108, 0.603).$$

- Wald test:

$$0.3 \pm 1.96\sqrt{\frac{0.3(1-0.3)}{10}} = (0.016, 0.584).$$

- These are large sample approximate 95% confidence intervals for θ. The “exact interval" (using CDF as a pivot) is (0.067,0.652).

Large Sample ML-based Methods II

Feng-Chang Lin

Department of Biostatistics
University of North Carolina at Chapel Hill

flin@bios.unc.edu

Generalized Likelihood Ratio Test

- To test the hypothesis $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, where θ is a vector (multivariate).
- Let $\Theta = \Theta_0 \cup \Theta_1$.
- A generalized likelihood ratio test (GLRT) is defined by

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)}.$$

- When $n \to \infty$, $-2 \log \lambda(x) \to_d \chi^2_r$, where $r = \mathrm{df}(\Theta) - \mathrm{df}(\Theta_0)$.
- Here, $\mathrm{df}(\Theta)$ means the degree of freedom under parameter space Θ, which is the number parameters needed to be estimated.

Multinomial Distribution

- Let $(X_1, \ldots, X_k)$ follow multinomial$(n, p_1, \ldots, p_k)$.
- To test $H_0 : p_i = p_{i0}$, $i = 1, \ldots, k$, versus $H_1 : H_0$ is not true.
- Show that the GLRT statistic is

$$\lambda(x) = n^n \prod_{i=1}^{k} \left(\frac{p_{i0}}{x_i} \right)^{x_i}.$$

- $\Theta_0 = \{(p_1, \ldots, p_k) | p_i = p_{i0}, i = 1, \ldots, k\}$.
- $\Theta = \{(p_1, \ldots, p_k) | 0 \le p_i \le 1, i = 1, \ldots, k\}$.
- The pdf of the multinomial distribution is

$$f(x_1, \ldots, x_k | p) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k},$$

where $\sum_{i=1}^{k} p_i = 1$ and $\sum_{i=1}^{k} x_i = n$.

Multinomial Distribution (cont'd)

- Under overall space Θ, $\hat{p}_i = x_i/n$.
- The GLRT statistic is

$$\lambda(x) = \frac{L(\Theta_0)}{L(\hat{\Theta})} = \frac{\prod_{i=1}^{k} p_{i0}^{x_i}}{\prod_{i=1}^{k} (\frac{x_i}{n})^{x_i}} = n^n \prod_{i=1}^{k} \left(\frac{p_{i0}}{x_i}\right)^{x_i}.$$

- Since $\mathrm{df}(\Theta) = k-1$ and $\mathrm{df}(\Theta_0) = 0$, when $n \to \infty$,

$$-2\log\lambda(x) \to_d \chi^2_{k-1}.$$

- One can show that, under null hypothesis H_0,

$$-2\log\lambda(x) \approx \sum_{i=1}^{k} \frac{(x_i - np_{i0})^2}{np_{i0}}.$$

Proof

- To prove the likelihood ratio test is asymptotically equivalent to the chi-square test, we re-write

$$-2\log\lambda(x) = -2\sum_{i=1}^{k} x_i \left\{\log p_{i0} - \log\left(\frac{x_i}{n}\right)\right\}.$$

- Using Taylor expansion on $\log p_{i0}$ around x_i/n, one has

$$\log p_{i0} = \log\left(\frac{x_i}{n}\right) + \frac{1}{x_i/n}\left(p_{i0} - \frac{x_i}{n}\right) - \frac{1}{2\xi^2}\left(p_{i0} - \frac{x_i}{n}\right)^2,$$

where $x_i/n < \xi < p_{i0}$.

Proof (cont'd)

- Bringing the expansion back to the formula, one has

$$
\begin{aligned}
-2\log\lambda(x) &= \sum_{i=1}^{k}(-2)x_i\left\{\frac{1}{x_i/n}\left(p_{i0}-\frac{x_i}{n}\right)-\frac{1}{2\xi^2}\left(p_{i0}-\frac{x_i}{n}\right)^2\right\} \\
&= \sum_{i=1}^{k}\frac{x_i}{\xi^2}\left(p_{i0}-\frac{x_i}{n}\right)^2=\sum_{i=1}^{k}\frac{x_i(x_i-np_{i0})^2}{n^2\xi^2}.
\end{aligned}
$$

- Since $x_i/n \to_p p_{i0}$ and $\xi \to_p p_{i0}$ under the null hypothesis, we have

$$
-2\log\lambda(x)\approx\sum_{i=1}^{k}\frac{(x_i-np_{i0})^2}{np_{i0}}.
$$

Example 1: Goodness-of-fit Test

- $H_0 : p_i = p_{i0}(\theta)$, $i = 1, \ldots, k$, where $\theta = (\theta_1, \ldots, \theta_r)$, versus $H_1 : H_0$ is not true.
- The GLRT statistic is

$$\lambda(x) = n^n \prod_{i=1}^{k} \left(\frac{p_{i0}(\hat{\theta})}{x_i} \right)^{x_i}.$$

- $\Theta_0 = \{(\theta_1, \ldots, \theta_r) | p_i = p_{i0}(\theta), i = 1, \ldots, k\}$.
- $\Theta = \{(p_1, \ldots, p_k) | 0 \le p_i \le 1, i = 1, \ldots, k\}$.
- Since $\mathrm{df}(\Theta) = k - 1$ and $\mathrm{df}(\Theta_0) = r$, we know

$$-2 \log \lambda(x) \approx \sum_{i=1}^{k} \frac{(x_i - np_{i0}(\hat{\theta}))^2}{np_{i0}(\hat{\theta})} \to_d \chi^2_{k-1-r},$$

when $n \to \infty$.

Poisson Distribution

- The number of automobile accidents occurring per day in a particular city is believed to follow Poisson distribution.
- A sample of 80 days during the year gives the data shown as follows.

Number of accidents	0	1	2	3	4
Observed frequency	34	25	11	7	3

- Does the data support the belief that the number of accidents per day has a Poisson distribution averaging one accident per day, i.e. $\theta = 1$?

Poisson Distribution (cont'd)

Number of accidence	0	1	2	3	4
Observed frequency	34	25	11	7	3
$p_{i0}(\theta)$	$e^{-\theta}$	$\theta e^{-\theta}$	$\theta^2 e^{-\theta}/2$	$\theta^3 e^{-\theta}/6$	rem.
$p_{i0}(1)$	0.368	0.368	0.184	0.061	0.019
Expected frequency	29.4	29.4	14.7	4.9	1.52

- The chi-square statistic, combining the last two columns, is

$$Q = \sum_{i=0}^{3} \frac{(x_i - np_{i0}(\hat{\theta}))^2}{np_{i0}(\hat{\theta})} = 4.3 < \chi^2_{3,0.05} = 7.81,$$

  where x_i is the observed frequency.
- What if the distribution is with any arbitrary mean?

Example 2: Hardy-Weinberg Equilibrium

- Punnett square is a 2×2 contingency table

		Females		
		$A(\theta)$	$a(1-\theta)$	
Males	$A(\theta)$	$n_{11}(\pi_{11})$	$n_{12}(\pi_{12})$	$n_{1\cdot}$
	$a(1-\theta)$	$n_{21}(\pi_{21})$	$n_{22}(\pi_{22})$	$n_{2\cdot}$
		$n_{\cdot 1}$	$n_{\cdot 2}$	n

- Null hypothesis H_0: $\pi_{11} = \theta^2$, $\pi_{12} = \pi_{21} = \theta(1-\theta)$, $\pi_{22} = (1-\theta)^2$.
- The GLRT is

$$Q = \frac{(n_{11} - n\hat{\pi}_{11})^2}{n\hat{\pi}_{11}} + \frac{(n_{12} + n_{21} - 2n\hat{\pi}_{12})^2}{2n\hat{\pi}_{12}} + \frac{(n_{22} - n\hat{\pi}_{22})^2}{n\hat{\pi}_{22}},$$

where $\hat{\pi}_{11} = \hat{\theta}^2$, $\hat{\pi}_{21} = \hat{\theta}(1-\hat{\theta})$, and $\hat{\pi}_{22} = (1-\hat{\theta})^2$.

Example 3: McNemar Test

- Responses of subjects are collected before and after an intervention.
- The 2×2 contingency table is formatted as

		After	
		S	F
Before	S	$O_{11}(p_{11})$	$O_{12}(p_{12})$
	F	$O_{21}(p_{21})$	$O_{22}(p_{22})$

- Null hypothesis H_0: $p_{11} + p_{12} = p_{11} + p_{21}$, i.e., $p_{12} = p_{21}$
- Show that the GLRT is

$$Q = \frac{(O_{12} - O_{21})^2}{O_{12} + O_{21}} \sim \chi_1^2,$$

which is the test statistic of McNemar Test.

Derivation of McNemar Test

- $\Theta_0 = \{(p_{11}, p_{12}, p_{21}, p_{22}) | p_{12} = p_{21}, \sum_{i=1}^{2} \sum_{j=1}^{2} p_{ij} = 1\}$.
- $\Theta = \{(p_{11}, p_{12}, p_{21}, p_{22}) | \sum_{i=1}^{2} \sum_{j=1}^{2} p_{ij} = 1, 0 \leq p_{ij} \leq 1, i, j = 1, 2\}$.
- Under Θ_0, $\hat{p}_{110} = O_{11}/n$, $\hat{p}_{120} = \hat{p}_{210} = (O_{12} + O_{21})/(2n)$, and $\hat{p}_{220} = O_{22}/n$.
- The GLRT is

$$\begin{aligned} Q &= \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(O_{ij} - n\hat{p}_{ij0})^2}{n\hat{p}_{ij0}} \\ &= \frac{\left\{\frac{1}{2}(O_{12} + O_{21}) - O_{12}\right\}^2}{\frac{1}{2}(O_{12} + O_{21})} + \frac{\left\{\frac{1}{2}(O_{12} + O_{21}) - O_{21}\right\}^2}{\frac{1}{2}(O_{12} + O_{21})} \\ &= \frac{(O_{12} - O_{21})^2}{O_{12} + O_{21}} \sim \chi_1^2. \end{aligned}$$

Nursing Home Trial

- Feeding problems are common in advanced dementia.
- The decision aids (intervention) is to reduce the expectation of benefit from tube feeding.
- The same question was asked before and after intervention.

	Prev	%	Post	%
Complete nutrition	76	60.3	100	79.4
Survival	31	24.6	11	8.7
Less/no choking	10	7.9	12	9.5
Total	126		126	

Nursing Home Trial

- Taking survival in advantages of tube feeding for example:

	Post no	Post yes	Prev total
Prev no	92	3	95
Prev yes	23	8	31 (24.6%)
Post total	115	11 (8.7%)	126

- McNemar test: $Q = \frac{(3-23)^2}{3+23} = 15.38 > \chi^2_{1,0.05} = 3.84$.

Introduction to Bayesian Statistics

Feng-Chang Lin

Department of Biostatistics
University of North Carolina at Chapel Hill

flin@bios.unc.edu

Introduction

- What we learned in this semester conceptually is called *frequentist* approach.
- In the frequentist approach, parameters are treated as unknown non-random constants.
- Probability statements are about observable random variables.
- For example, if the 95% CI is denoted as

$$P(L(X) < \theta < U(X)) = 0.95,$$

  the probability measure P is about X, not θ.
- Give it a try: How do we interpret the confidence interval?

Introduction (cont'd)

- In Bayesian approach, parameters such as θ are conceptualized as random variables.
- Their distribution is called *prior distribution*.
- The prior distribution can be interpreted as our belief or knowledge about θ before observing X.
- It can also be interpreted as a *plausibility* function.
- The interpretation of the prior are key differences between different Bayesian schools.

Notations

- Prior, pdf or pmf: $\pi(\theta)$, completely known and specified in advance.
- Likelihood: $f(x|\theta)$, conditional distribution of X given θ.
- Posterior: $\pi(\theta|x)$, conditional distribution of θ given X. It can be expressed as

$$\pi(\theta|x) = \pi(x)f(x|\theta)/m(x),$$

  where $m(x)$ is the marginal distribution of X,

$$m(x) = \int_\Theta \pi(\theta)f(x|\theta)d\theta.$$

- The integral is replaced by a summation if θ is discrete.

Binomial Bayes Estimation

- Let $X_1, \ldots, X_n$ be iid Bernoulli(p). Then $Y = \sum_{i=1}^n X_i$ is binomial(n, p).
- We assume that the prior distribution on p is beta(α, β).
- The joint distribution of Y and p is

$$f(y, p) = f(y|p)\pi(p)$$

- The marginal distribution of Y is

$$m(y) = \int_0^1 f(y, p) dp$$

- The posterior distribution is

$$\pi(p|y) = \frac{f(y, p)}{m(y)}$$

Binomial Bayes Estimation (cont'd)

- The posterior is beta$(y+\alpha, n-y+\beta)$.
- How to estimate p?
- The mean of the posterior is

$$\hat{p}_B = \frac{y+\alpha}{\alpha+\beta+n},$$

which can be written as

$$\hat{p}_B = \left(\frac{n}{\alpha+\beta+n}\right)\left(\frac{y}{n}\right) + \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\left(\frac{\alpha}{\alpha+\beta}\right).$$

- This is a weighted mean between sample mean and prior mean.
- Notice that, one can have different options for the prior.

MSE of Binomial Bayes Estimator

- The MSE of $\hat{p}$, the MLE, as an estimator of p, is

$$E(\hat{p} - p)^2 = \text{Var}\bar{X} = \frac{p(1-p)}{n}.$$

- The MSE of the Bayes estimator of p is

$$E(\hat{p}_B - p)^2 = \text{Var}\left(\frac{\sum_{i=1}^n X_i + \alpha}{\alpha + \beta + n}\right) + \left\{E\left(\frac{\sum_{i=1}^n X_i + \alpha}{\alpha + \beta + n}\right) - p\right\}^2$$

- Choose $\alpha = \beta = \sqrt{n/4}$ yields

$$E(\hat{p}_B - p)^2 = \frac{n}{4(n + \sqrt{n})^2}.$$

- For small n, $\hat{p}_B$ is the better choice; for large n, $\hat{p}$ is the better choice.

Conjugate Priors

- In the binomial example above, both the prior and the posterior distributions were in the beta family.
- A family of priors that leads to posteriors in the same family is called a *conjugate family*.
- Such priors are called *conjugate priors*.
- **Example (Normal Bayes estimators)** Let $X \sim n(\theta, \sigma^2)$, and suppose that the prior distribution of θ is $n(\mu, \tau^2)$.
- The posterior distribution of θ is also normal, with mean and variance given by

$$E(\theta|x) = \frac{\tau^2}{\tau^2 + \sigma^2} x + \frac{\sigma^2}{\tau^2 + \sigma^2} \mu$$

and

$$\mathrm{Var}(\theta|x) = \frac{\tau^2 \sigma^2}{\tau^2 + \sigma^2}.$$

Conjugate Priors (cont'd)

- If the random sample is extended to $X_1, \ldots, X_n$, the posterior mean and variance become

$$E(\theta|x) = \frac{\tau^2}{\tau^2 + \sigma^2/n}\bar{x} + \frac{\sigma^2/n}{\tau^2 + \sigma^2/n}\mu$$

  and

$$\mathrm{Var}(\theta|x) = \frac{\tau^2\sigma^2/n}{\tau^2 + \sigma^2/n},$$

  respectively.
- What do we learn from the binomial and normal Bayes estimation?

Hypothesis Testing

- Suppose we want to test $H_0 : p \in A$ versus $H_1 : p \in A^c$.
- We can use the posterior $\pi(p|y)$ to compute the probability

$$a_0 = P(p \in A|y)$$

- Reject H_0 when $a_0 > 1/2$.
- **(Normal Bayesian Test)** Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. We will reject H_0 if and only if

$$P(\theta \leq \theta_0|\boldsymbol{x}) > 1/2.$$

- Since $\pi(\theta|\boldsymbol{x})$ is symmetric, H_0 will be rejected if $E(\theta|\boldsymbol{x}) > \theta_0$, i.e.,

$$\bar{X} > \theta_0 + \frac{\sigma^2(\theta_0 - \mu)}{n\tau^2}.$$

Hypothesis Testing (cont'd)

- If the type I error is considered more serious than the type II error we may change our cutoff "1/2" to a smaller number.
- **(Bayes Factor)** A Bayesian measure of evidence against the null hypothesis (H_0), and in favor of an alternative hypothesis (H_1), is called *Bayes Factor*, which is defined by

$$\mathrm{BF} = \frac{P(H_1|\boldsymbol{x})/P(H_0|\boldsymbol{x})}{P(H_1)/P(H_0)}.$$

- This factor can be interpreted as the ratio of posterior odds of H_1 against the prior odds of H_1.

Hypothesis Testing (cont'd)

- According Kass and Raftery (1995), $1 < \text{BF} \leq 3$ provides "weak" evidence, $3 < \text{BF} \leq 20$ provides "positive" evidence, $20 < \text{BF} \leq 150$ provides "strong" evidence, and $\text{BF} > 150$ provides "very strong" evidence in favor of H_1.

Bayes Factor: An Example

- Assume the survival time for advanced-stage colorectal cancer follows

$$f(x|\lambda) = \lambda e^{-\lambda x}, \;\; x > 0, \;\; \lambda > 0.$$

- For unknown λ, one is willing to assume the prior of λ follows

$$\pi(\beta) = \beta e^{-\beta\pi}, \;\; x > 0, \;\; \beta > 0.$$

- If $\beta = 1$ and $x = 3$, what is the Bayes factor for testing $H_0 : \lambda \geq 1$ versus $H_1 : \lambda < 1$?
- What is the strength of evidence in favor of H_1 according to the scale proposed by Kass and Raftery (1995)?

Bayes Factor: An Example (cont'd)

- It may be easier to get the marginal CDF of X, which equals

$$E\{F(x|\lambda)\} = \int_0^\infty F(x|\lambda)\pi(\lambda)d\lambda = 1 - \left(1 + \frac{x}{\beta}\right)^{-1}.$$

- The posterior distribution hence is

$$\pi(\lambda|x) = \frac{f(x|\lambda)\pi(\lambda)}{m(x)} = \frac{\lambda\beta e^{-(x+\beta)\lambda}}{\beta^{-1}(1 + x/\beta)^{-2}} = \lambda(x+\beta)^2 e^{-(x+\beta)\lambda},$$

which is Gamma$(2, (x+\beta)^{-1})$ ($\pi(\lambda)$ conjugate prior?).

- Since $P(\lambda < \lambda^*|x) = 1 - \{\lambda^*(x+\beta)+1\}e^{-(x+\beta)\lambda^*}$, one can have

$$\text{BF} = \frac{P(H_1|\boldsymbol{x})P(H_0)}{P(H_0|\boldsymbol{x})P(H_1)} = \frac{(1-5e^{-4})e^{-1}}{5e^{-4}(1-e^{-1})} = 5.77.$$

Interval Estimation

- Quantile of $\pi(p|x)$ can be used to compute interval estimators.
- The resulting intervals are called *Bayesian credible intervals* or *credible set* (C&B).
- For a $1 - \alpha$ credible interval, we choose $\alpha_1 \geq 0$ and $\alpha_2 \geq 0$ with $\alpha = \alpha_1 + \alpha_2$.
- Define $L(x)$ and $U(x)$ to be α_1 and $1 - \alpha_2$ quantiles of $\pi(p|x)$, respectively.
- The intervals can be one-sided by taking either α_1 or α_2 to be zero.

Normal Credible Set

- In the previous normal example,

$$\pi(\theta|\bar{x}) = n(\delta^B(\bar{x}), \sigma^2(\theta|\bar{x})).$$

- The $1-\alpha$ credible set for θ is given by

$$1-\alpha = P\left(\delta^B(\bar{x}) - z_{\alpha/2}\sigma(\theta|\bar{x}) \le \theta \le \delta^B(\bar{x}) + z_{\alpha/2}\sigma(\theta|\bar{x})\right).$$

- How about the coverage probability of this region in frequentist sense? One can have

$$\begin{aligned}
&P(|\theta - \delta^B(\bar{x})| \le z_{\alpha/2}\sigma(\theta|\bar{x})) \\
&\quad = P\left(-\sqrt{1+\gamma}z_{\alpha/2} + \frac{\gamma(\theta-\mu)}{\sigma/\sqrt{n}} \le Z \le \sqrt{1+\gamma}z_{\alpha/2} + \frac{\gamma(\theta-\mu)}{\sigma/\sqrt{n}}\right),
\end{aligned}$$

where $\gamma = \sigma^2/(n\tau^2)$ and $Z = \sqrt{n}(\bar{X}-\theta)/\sigma$.

Bayesian Optimality

- One can obtain the smallest credible interval with a specific coverage probability.
- We would like to find the set $C(x)$ that satisfies
  (a) $\int_{C(x)} \pi(\theta|x)dx = 1 - \alpha$,
  (b) $\text{size}(C(x)) \leq \text{size}(C'(x))$,

  for any set $C'(x)$ satisfying $\int_{C'(x)} \pi(\theta|x)dx \geq 1 - \alpha$.
- Using Theorem 9.3.2 in C&B, we can conclude if the posterior density $\pi(\theta|x)$ is unimodal, then for a given value of α, the shortest credible interval for θ is given by

$$\{\theta : \pi(\theta|x) \geq k\}, \text{ where } \int_{\{\theta:\pi(\theta|x)\geq k\}} \pi(\theta|x)d\theta = 1 - \alpha.$$

- We call this *highest posterior density* (HPD) region.

Decision Theory

- Estimating θ can be viewed as a decision or an action.
- A loss function $L(\theta, \hat{\theta})$ quantifies the penalty for choosing $\hat{\theta}$ when the true value is θ.
- Two types of loss functions: *squared-error loss*

$$L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2,$$

and *absolute-error loss*

$$L(\theta, \hat{\theta}) = |\hat{\theta} - \theta|.$$

- One may also consider *weighted* loss function

$$L(\theta, \hat{\theta}) = \omega(\theta)|\hat{\theta} - \theta|^r,$$

for some $\omega(\theta) \geq 0$ and $r > 0$.

Risk Function

- Formally, let $X_1, \ldots, X_n$ be a random sample from distribution $f(x|\theta)$, $\theta \in \Theta \subseteq \Re$, and let $\delta(x)$ be an estimator of θ.
- The loss function $L(\theta, \delta(x)) \geq 0$ is defined over $\Theta \times D \rightarrow \Re^+$.
- The *risk function* represents the expected loss over the sample space D, which is defined by

$$R(\theta, \delta(x)) = E_X\{L(\theta, \delta(x))\} = \int_{\mathcal{X}} L(\theta, \delta(x)) f(x|\theta) dx.$$

- Given two decision rules $\delta^*(x)$ and $\delta(x)$, if $R(\theta, \delta^*) \leq R(\theta, \delta(x))$ $\forall \theta \in \Theta$, and $R(\theta, \delta^*) < R(\theta, \delta(x))$ at at least one $\theta \in \Theta$, we will call $\delta^*(x)$ is better than $\delta(x)$.

Minimax Decision Rule

- We will call a decision rule $\delta^*(x)$ a *minimax decision rule* if

$$\sup_{\theta} R(\theta, \delta^*(x)) = \inf_{\delta} \sup_{\theta} R(\theta, \delta(x))$$

- **Example** On a rainy day a teacher has three choices: (a_1) to take an umbrella and face the possible prospect of carrying it around the sunshine; (a_2) to leave the umbrella at home and perhaps get drenched; (a_3) to just give up the lecture and stay at home.
- Let $\Theta = \{\theta_1, \theta_2\}$ and θ_1 corresponds to rain, and θ_2 to no rain.
- The following table give the losses for the decision problem:
- The weather report that depends on θ as follows:
- Find the minimax rule to help the teacher make a decision.

Minimax Decision Rule (cont'd)

- There are 9 decisions when you saw the weather outside.
- The risk function:

$$R(\theta_j, \delta_i) = E\{L(\theta_j, \delta_i)\} = \sum_{k=1}^{2} L(\theta_j, \delta_{ik}) P(W_k | \theta_j),$$

  where δ_{ik} is the action $\{a_1, a_2, a_3\}$ you take when you saw W_1 (rain) or W_2 (shine).
- The conclusion: Bring the umbrella no matter rain or shine.

Bayes Decision Rule

- Add a probability measure on the parameter, $\pi(\theta)$.
- Bayes risk with respect to $\pi(\theta)$:

$$r^B(d) = E\{R(\theta, \delta(x))\} = \int_\Theta R(\theta, \delta(x))\pi(\theta)d\theta.$$

- If there exists a decision function $\delta^*(x)$, satisfying

$$r^B(\delta^*) = \inf_\delta r^B(\delta),$$

  we will call $\delta^*(x)$ is Bayes decision function with respect to $\pi(\theta)$
- Show that $\inf_\delta r^B(\delta)$ has the same solution as

$$\inf_\delta \int_\Theta L(\theta, \delta(x))\pi(\theta|x)d\theta.$$