Biostat 602 Winter 2016

Lecture Set 1

Review of the Past

Introduction

In scientific research, an investigator often uses one of two types of reasoning, namely the *deductive reasoning* and the *inductive reasoning.*

Deductive Reasoning

- works from the general to specific; typically based on some general laws or rules which are then applied to a specific case.
- We make an assumption about a population and want specifics of a sample
- Suppose the lifetime of a particular bt=rand of car battery has an exponential distribution with a median of 7 years. We want to determine what percentage of these batteries that will last at least 10 years.
- Subject of **Biostat 601**

Inductive Reasoning

- generalizes the conclusion of findings observed from a specific.
- Suppose a particular supplier is providing batteries to a hardware manufacturer and it is intended to estimate the lifetime distribution of this particular brand and substantiate the manufacturer's claim that 90% of the batteries last over 700 hours.
- Typically one would select a random sample of batteries from the batch provided by the supplier and run a life-test on them
- Based on the findings from the sample estimate the distribution of the lifetime and test out the claim
- Subject of statistical inference (**Biostat 602**)

Review of Biostat 601

Probability

Let S be the sample space related to a random experiment. Probability is a set function with range in [0, 1] defined on all subsets of S satisfying:

i. $P(E) \geq 0$, for any event $E \subset S$.

ii. $P(S) = 1$.

iii. If $E_1, E_2, \ldots$ are mutually exclusive, (i.e. $E_i \cap E_j = \phi, \quad i \neq j$), then

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i) \qquad \textit{(countable additivity)}$$

Laws of Probability:

- *Addition Law*

  For any finite set of events $E_1, E_2, \ldots, E_n$,

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i) - \sum\sum\nolimits_{i<j} P(E_iE_j) + \sum\sum\sum\nolimits_{i<j<k} P(E_iE_jE_k) + \cdots + (-1)^{n+1} P(E_1E_2\cdots E_n).$$

- *Boole's Inequality*

$$P\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{i=1}^{n} P(E_i)$$

- *Bonferroni's Inequality*

$$P\left(\bigcap_{i=1}^{n} E_i\right) \geq \sum_{i=1}^{n} P(E_i) - (n-1)$$

- *Law of complementation*
- *Multiplication Law (Conditional Probability)*
- *Law of Independence*
- *Law of Total Probability*

  Suppose $A_1, A_2, \ldots, A_n$ are mutually exclusive and exhaustive events, i.e. $A_i \cap A_j = \phi, \;\; i \neq j$ and $S = \cup_{i=1}^{n} A_i$. Let B be any event in S. Then

$$P(B) = \sum_{i=1}^{n} P(A_i) P(B|A_i).$$

- *Bayes Theorem*

  Suppose $F_1, F_2, \ldots, F_n$ are **mutually exclusive** and **exhaustive** events, i.e. one and only one of them must occur. Suppose for some $j, j = 1, \ldots, n$, we are interested in the conditional probability of F_j given another conditioning event E, i.e. $P(F_j \mid E)$. Bayes' Theorem states that it can be obtained using the *reverse* conditional probability as

$$P(F_j \mid E) = \frac{P(E \mid F_j) P(F_j)}{\displaystyle\sum_{i=1}^{n} P(E \mid F_i) P(F_i)}.$$

Example 1

- The ELISA (Enzyme-Linked Immunosorbent Assay) test is used to detect antibodies in blood and can indicate the presence of the HIV virus.

- Approximately 5% of a population is HIV positive.

- Among those who have HIV virus, 96% test positive with ELISA (Sensitivity).

- Among those who do not have HIV virus, approximately 98% test negative with ELISA (Specificity).

- For a randomly chosen subject from this population if the test is positive, what is the probability that the subject has HIV virus?

Diagnostic Testing Nomenclature

	Test Results	
Disease	$+$	$-$
$+$	TP	FN
$-$	FP	TN

In any diagnostic test, there are four quantities which people are interested in:

Sensitivity: Probability of True Positives, i.e. probability of the test result being positive for a diseased individual ($TP/(TP + FN)$)

Specificity: Probability of True Negatives, i.e. probability of the test result showing negative finding for an individual w/o the disease ($TN/(FP + TN)$)

Positive Predictive Value: probability of the individual truly having the disease when the test result is positive ($TP/(TP + FP)$)

Negative Predictive Value: probability of the individual not having the disease when the test result is negative ($TN/(TN + FN)$)

Remarks

- High values of all four quantities are desirable for a diagnostic test.
- In designing the test, care is taken to maintain a reasonably high level of sensitivity and specificity. These two, along with the prevalence of the disease determine the predictive values.
- In our example, sensitivity, specificity are provided. We want to find the positive predictive value.

Back to AIDS example

- Let H = subject has HIV virus, and Pos = test result is positive.
- It is given that

$$P(H) = 0.05, \quad P(Pos \mid H) = 0.96, \quad P(Pos \mid H^c) = 0.02.$$

- Want to find $P(H \mid Pos)$.
- By definition of conditional probability

$$P(H \mid Pos) = \frac{P(H \cap Pos)}{P(Pos)}$$

- Now

$$P(H \ \cap \ Pos) = P(Pos \mid H)P(H) = 0.96 \times 0.05 = 0.048,$$

and

$$\begin{aligned} P(Pos) &= P(Pos \cap H) + P(Pos \cap H^c) \\ &= P(Pos \mid H)P(H) + P(Pos \mid H^c)P(H^c) \\ &= (0.96)(0.05) + (0.02)(0.95) = 0.067. \end{aligned}$$

- The required probability equals $0.048/0.067 = 0.716$.

Random Variables

A random variable Y is a real-valued function defined on a probability space.

- **Discrete Random Variables:** Probability mass function (pmf), Cumulative Distribution Function (cdf), Calculation of Expectation, Variance from a pmf

- **Continuous Random Variables:** Probability density function (pdf), Cumulative Distribution Function (cdf), Calculation of Expectation, Variance from a given pdf.

- Common Families of Discrete Distributions (Binomial, Poisson, Geometric, Negative Binomial)

- Common Families of Continuous Distributions (Normal, t, chi^2, F, Exponential, Gamma)

Example 2: A point is chosen at random on a line segment of length L. Find the probability that the ratio of the shorter to the longer segment is less than 1/4.

Solution: The given information tantamount to saying that a point randomly picked on the line segment has a length X which is has a *uniform* distribution on $(0, L)$. We are interested in

$$P\left(\frac{\min(x, L-x)}{\max(x, L-x)} \leq 1/4\right).$$

Now note that for $x < L/2$, $\min(x, L-x) = x$ and,

$$\frac{\min(x, L-x)}{\max(x, L-x)} \leq 1/4 \Rightarrow \frac{x}{L-x} \leq 1/4 \Rightarrow x \leq L/5.$$

For $x \geq L/2$, $\min(x, L-x) = L - x$ and,

$$\frac{\min(x, L-x)}{\max(x, L-x)} \leq 1/4 \Rightarrow \frac{L-x}{x} \leq 1/4 \Rightarrow x \geq 4L/5$$

So the required probability equals

$$\begin{aligned} P[X \leq L/5] + P[X \geq 4L/5] &= \int_0^{\frac{L}{5}} \frac{1}{L}\,dx + \int_{\frac{4L}{5}}^{L} \frac{1}{L}\,dx \\ &= \frac{1}{5} + \frac{1}{5} \\ &= \frac{2}{5}. \end{aligned}$$

Example 3: Suppose that the travel time from Adam's home to his office is a normally distributed random variable with mean = 40 minutes and standard deviation = 7 minutes.

(a) What proportion of time Adam reaches office within 38 and 45 minutes of leaving home?

Solution: Let X denote Adam's travel time. We need to find $P[38 < X < 45]$. Note

$$\begin{aligned} P[38 < X < 45] &= P\left[\frac{38-40}{7} < Z < \frac{45-40}{7}\right] \\ &= P\left[Z < \frac{45-40}{7}\right] - P\left[Z < \frac{38-40}{7}\right] \\ &= P[Z < 0.714] - P[Z < -0.286] \\ \\ &= \Phi(0.71) - \Phi(-0.29) = .7611 - .3859 = .3752. \end{aligned}$$

(b) If Adam wants to be 95% certain that he will not be late for an office appointment at 1 PM, what is the latest time he should leave home?

Solution: This falls under a class of problems involving *inverse transformation.* In these problems, one is interested in finding for a normal random variable X the $100p - th$ percentile x_p. So x_p satisfies the equation $P[X < x_p] = p$; it is the point to the left of which lies $100p\%$ of the distribution. One solves the problem in the following two steps.

Step 1: Calculate $100p - th$ percentile z_p of Z, that satisfies $P[Z < z_p] = p$.

Step 2: Find x_p using the formula $x_p = \mu + \sigma z_p$.

In our problem, we need to find the 95th percentile of the distribution of X. Using the *qnorm* function in R, $z_{0.95} = 1.645$, and

$$x_{0.95} = 40 + 7(1.645) = 51.515.$$

So, Adam needs to leave his home latest by 12:08 PM.

Multiple Random Variables

- Probability calculations from bivariate distributions
- Bivariate transformations, calculating jacobian, joint to marginal and conditional distribution
- Finding marginal distributions from a hierarchical structure
- Applying Conditional Expectation and Variance formula in Hierarchical Models

  $$E(Y) = E\left[E(Y|X)\right], \quad Var(Y) = E\left[Var(Y|X)\right] + Var\left[E(Y|X)\right].$$

- Applying variance and covariance formula for linear combinations

  $$Cov\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j Cov(X_i, Y_j),$$

  $$Var\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 Var(X_i) + 2\sum\sum\nolimits_{i<j} a_i a_j \; Cov(X_i, X_j).$$

- Chebyshev's Inequality

  If μ and σ are the mean and standard deviation of a random variable X, then for any positive constant k and $\sigma > 0$,

  $$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}.$$

- Jensen's Inequality

  For any random variable X, if $g(x)$ is a convex function, then

  $$E\left[g(X)\right] \geq g\left(E(X)\right).$$

Example 4: Let X, Y have joint pdf

$$f(x,y) = \begin{cases} cxy & 0 \leq x \leq y < 1 \\ \\ 0 & \text{otherwise} \end{cases}$$

(a) Find c.

(b) Find $P(X + Y \leq 1)$.

(c) Find $E(Y|X = x)$.

Example 5: Suppose X_1, X_2 have the joint pdf

$$f_{X_1,X_2}(x_1,x_2) = 16x_1^3x_2^3, \quad 0 \le x_1 \le 1,\ 0 \le x_2 \le 1.$$

Consider the transformation to $Y_1 = X_1\sqrt{X_2}$ and $Y_2 = X_2\sqrt{X_1}$. Find the joint density of Y_1 and Y_2. Are they independent?

Example 6: Drugs and HIV

$$\begin{aligned} N &= \text{No. of drug injections during specified time period} \\ X_i &= \begin{cases} 1 \text{ if needle is contaminated with HIV} \\ 0 \text{ otherwise} \end{cases} \\ S &= \text{No. of contaminated needles used in time period} \end{aligned}$$

$$S|N = n \sim Binomial(n, \theta), \quad N \sim Poisson(\lambda).$$

$$\begin{aligned} P(S = s) &= \sum_{n=0}^{\infty} P(S = s|N = n)P(N = n) \\ &= \sum_{n=s}^{\infty} \binom{n}{s} \theta^s (1-\theta)^{n-s} e^{-\lambda} \frac{\lambda^n}{n!} \\ &= e^{-\lambda} (\lambda\theta)^s \sum_{n=s}^{\infty} \binom{n}{s} \frac{\{\lambda(1-\theta)\}^{n-s}}{n!} \\ &= e^{-\lambda} \frac{(\lambda\theta)^s}{s!} \sum_{n=s}^{\infty} \frac{\{\lambda(1-\theta)\}^{n-s}}{(n-s)!} \\ &= e^{-\lambda} \frac{(\lambda\theta)^s}{s!} \sum_{n=0}^{\infty} \frac{\{\lambda(1-\theta)\}^{n}}{n!} \qquad \text{(change of index)} \\ &= e^{-\lambda} \cdot e^{\lambda(1-\theta)} \frac{(\lambda\theta)^s}{s!} \\ &= e^{-\lambda\theta} \frac{(\lambda\theta)^s}{s!} \end{aligned}$$

$S \sim Poisson(\lambda\theta)$.

Random Samples

- Basic objective in statistical inference is to estimate population parameters of interest, such as mean, median, sd, prevalence, odds.
- Inference on the population parameters is based on the corresponding measure derived from a sample. For example, the prevalence of a chronic condition in a certain population can be estimated on the basis of the proportion of individuals having this condition in a *random sample* drawn from the population.
- A random sample is a collection of random variables.
- A collection of random variables $X_1, X_2, \ldots, X_n$ is called a **random sample** of size n from a population with pdf/pmf $f(x)$ if
  1. $X_1, X_2, \ldots, X_n$ are mutually independent;
  2. The marginal pdf or pmf of X_i is the same as $f(x)$.
- Alternatively, we say $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables, expressed as

$$X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} f(x)$$

- The joint pdf or pmf of $X_1, X_2, \ldots X_n$ (also called the *likelihood function*) is

$$f(x_1, \ldots, x_n) = f(x_1) \times f(x_2) \times \ldots \times f(x_n) = \prod_{i=1}^{n} f(x_i)$$

Properties of sample mean and variance

Result: Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then

(a) $E(\overline{X}) = \mu$.

(b) $Var(\overline{X}) = \sigma^2/n$.

(c) $E(S^2) = \sigma^2$.

(d) $Var(S^2) = \left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right)/n$, where μ_4 is the fourth central moment of the population.

Properties of sample mean and variance from Normal population

Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution, and let

$$\overline{X} = \left(\sum_{i=1}^{n} X_i\right) \Big/ n \quad \text{and} \quad S^2 = \left\{\sum_{i=1}^{n} (X_i - \overline{X})^2\right\} \Big/ (n-1).$$

- **Result 1:** $\overline{X}$ and S^2 are independent random variables.
- **Result 2:**
$$\overline{X} \sim N(\mu, \sigma^2/n).$$
- **Result 3:**
$$(n-1)S^2/\sigma^2 \sim \chi^2(n-1).$$
- **Result 4:**
$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

- **Result 5:** Suppose $X_1, \ldots, X_n$ is a random sample from a $N(\mu_X, \sigma_X^2)$ population, and $Y_1, \ldots, Y_m$ is a random sample from an independent $N(\mu_Y, \sigma_Y^2)$ population. Then

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{n-1,m-1}.$$

- **Result 6:** Suppose $X_1, \ldots, X_n$ is a random sample from an arbitrary distribution F. Define $\overline{X}$ and S^2 as above. Then $\overline{X}$ and S^2 are *independently* distributed *if and only if* F is normal.

Order Statistics

Consider a continuous population. Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d with cdf and pdf $F_Y(y)$, $f_Y(y)$, respectively. The ordered observations

$$Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$$

are called order statistics. For example, the *minimum* is $Y_{(1)}$ and the *maximum* is $Y_{(n)}$. We are interested in finding the distribution of an arbitrary $Y_{(i)}$, as well as the joint distributions of sets of $Y_{(i)}$'s and $Y_{(j)}$'s.

I. Distribution of $Y_{(r)}$

Marginal pdf of the r-th order statistic is

$$f_{Y_{(r)}}(y) = \frac{n!}{(r-1)!(n-r)!} F(y)^{r-1}[1-F(y)]^{n-r} f(y)$$

II. Joint distribution of $Y_{(r)}, Y_{(s)}, \quad r < s$

Joint pdf of any pair of order statistics Y_r, Y_s is given by

$$\begin{aligned} f_{Y_{(r)},Y_{(s)}}(u,v) &= \frac{n!}{(r-1)!(s-r-1)!(n-s)!} F_Y(u)^{r-1} \\ & \quad \times \left[F_Y(v) - F_Y(u)\right]^{s-r-1} (1 - F_Y(v))^{n-s} f_Y(u) f_Y(v) \end{aligned}$$

III. Joint distribution of first r order statistics, $r < n$

Joint pdf of $Y_{(1)}, \ldots, Y_{(r)}$ from a sample of size n is

$$f_{Y_{(1)},\ldots,Y_{(r)}}(u_1, \ldots, u_r) = \frac{n!}{(n-r)!} \prod_{i=1}^{r} f_Y(u_i) \bigg(1 - F(u_r)\bigg)^{n-r}, \quad u_1 < u_2 < \ldots < u_r.$$

Large Sample Theory

Convergence of a sequence of random variables

A sequence of random variables $\{X_n\}$ is said to converge, as $n \longrightarrow \infty$,

(i) almost surely (or with probability 1) to a random variable X (Notation: $X_n \xrightarrow{a.s.} X$) if for any $\epsilon > 0$

$$P\left[\lim_{n\to\infty} |X_n - X| > \epsilon\right] = 0.$$

(ii) in probability to a random variable X (Notation: $X_n \xrightarrow{P} X$) if for any $\epsilon > 0$

$$\lim_{n\to\infty} P\Bigg[|X_n - X| > \epsilon\Bigg] = 0.$$

(iii) in distribution to a random variable X (Notation: $X_n \xrightarrow{d} X$) if

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

at all continuity points of $F_X(x)$.

(iv) in p-th mean to a random variable X (Notation: $X_n \xrightarrow{L_p} X$) if

$$\lim_{n\to\infty} E\Bigg[|X_n - X|^p\Bigg] = 0.$$

Example 7: Suppose $X_1, X_2, \ldots X_n$ be a random sample from a *lomax* distribution with parameter σ having pdf

$$f_X(x) = \frac{1}{\sigma\left(1 + \frac{x}{\sigma}\right)^2}, \qquad x > 0, \sigma > 0.$$

(a) Let $X_{(1)}$ be the minimum based on the random sample. Show that $nX_{(1)} \xrightarrow{d} Exp(\sigma)$ as $n \longrightarrow \infty$.

(b) Show that $X_{(1)} \xrightarrow{P} 0$ as $n \longrightarrow \infty$.

Proof:

Example 8: Suppose $X_1, X_2, \dots X_n$ be a random sample from a *lomax* distribution with parameter σ having pdf

$$f_X(x) = \frac{1}{\sigma\left(1+\frac{x}{\sigma}\right)^2}, \qquad x > 0, \sigma > 0.$$

(a) Let $X_{(1)}$ be the minimum based on the random sample. Show that $nX_{(1)} \xrightarrow{d} Exp(\sigma)$ as $n \longrightarrow \infty$.

(b) Show that $X_{(1)} \xrightarrow{P} 0$ as $n \longrightarrow \infty$.

Proof:

Slutsky's theorem

If $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{P} b$, and $Z_n \xrightarrow{P} a$, where a and b are constants, then

$$Z_n X_n + Y_n \xrightarrow{d} aX + b.$$

Weak law of large numbers

Suppose $Y_1, Y_2, \ldots, Y_n$ are i.i.d. with $E(Y_i) = m$ and $\mathrm{V}(Y_i) = \sigma^2$. Then $\overline{Y}_n = (Y_1 + \cdots + Y_n)/n \xrightarrow{P} m$

Strong law of large numbers

Let $Y_1, Y_2, \ldots, Y_n$ be a sequence of i.i.d. random variables with $E(Y_i) = m < \infty$. Then the Strong Law of Large Numbers states that $\overline{Y}_n \xrightarrow{a.s.} m$. In other words,

$$P\left\{\lim_{n\to\infty} \overline{Y}_n = m\right\} = 1.$$

Example 9: Let $X_n \sim F(n,n)$, a F distribution with n and n degrees of freedom. Show that as $n \longrightarrow \infty$,

$$X \xrightarrow{P} 1, \quad X \xrightarrow{a.s.} 1.$$

Central Limit Theorem (Laplace)

Let Y_i for $i = 1, 2, \ldots, n$, be i.i.d. each with finite mean $\mu < \infty$ and finite variance $\sigma^2 < \infty$. Then, the *Central Limit Theorem* states that

$$Z_n = \frac{(\overline{Y}_n - \mu)}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1).$$

This implies $\lim_{n\to\infty} P(Z_n \leq z) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{z} \exp(-x^2/2)dx.$

Example 10: Let $X_n \sim gamma(n, \beta)$.

(a) Show that $\frac{X_n}{n} \xrightarrow{P} \beta$.

(b) What is the limiting distribution of suitably scaled and centered X_n/n?

Delta Method

Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \xrightarrow{d} N(0, \sigma^2)$. For a given function g and a specific value of θ, suppose $g^{(1)}(\cdot)$ exists, continuous, and $g^{(1)}(\theta) \neq 0$. Then

$$\sqrt{n}\left[g(Y_n) - g(\theta)\right] \xrightarrow{d} N\left\{0, \sigma^2 \left[g^{(1)}(\theta)\right]^2\right\}$$

Example 11: Let $X_n \sim \textit{gamma}(n, \beta)$. Define $Y_n = X_n/n$.

(a) Obtain the limiting distribution of $\sqrt{n}(Y_n - \beta)$.

(b) Obtain the limiting distribution of $\sqrt{n}(\log(Y_n) - \log(\beta))$.

(c) What is the limiting distribution of (scaled and centered) Y_n^{-1}?

Example 12: Let $X_1, X_2, \ldots, X_n$ be a random sample from $Bernoulli(p)$. Consider the transformation function $g(x) = x(1-x)$. Find the large-sample distribution of suitably scaled and centered random variable $g(\overline{X}_n)$.

Biostat 602 Winter 2016

Lecture Set 2

Principles of Data Reduction

Premise

Reading: CB 6.1–6.2

We assume that the data was generated by a pdf (or pmf) that belongs to a class of pdfs (or pmfs).

$$\mathcal{P} = \{f_X(x|\theta), \theta \in \Omega \subset \mathbb{R}^p\}$$

For example $X \sim \text{Bernoulli}(\theta), \theta \in (0, 1) = \Omega \subset \mathbb{R}$.

We collect data in order to

- Estimate θ (point estimation)
- Perform tests of hypothesis about θ.
- Estimate confidence intervals for θ (interval estimation).
- Make predictions of future data.

Typical Questions

- What is the estimated probability of head given a series of observed coin tosses (H, H, T, T, T)? (**Point Estimation**)
- Given a series of coin tosses, can you tell whether the coin is biased or not? ($\theta = \frac{1}{2}$). (**Test of Hypothesis**)
- What is the plausible range of the true probability of head, given a series of coin tosses? (**Interval Estimation**)
- Given the series of coin tosses, can you predict what the outcome of the next coin toss? (**Prediction**)

Data Reduction

Data; $x_1, \cdots, x_n$: Realization of random variables $X_1, \cdots, X_n$. Often we deal with a random sample whereby $X_1, \cdots, X_n$ is i.i.d.

Define a function of data

$$T(\mathbf{X}) = T(x_1, \cdots, x_n) : \mathbb{R}^n \to \mathbb{R}^d$$

We wish this summary of data to

1. Be simpler than the original data, e.g. $d \leq n$.
2. Keep all the information about θ that is contained in the original data $x_1, \cdots, x_n$.

A **statistic** $T(\mathbf{X}) = T(X_1, \cdots, X_n)$ is a function of random variables $X_1, \cdots, X_n$. Clearly, $T(\mathbf{X})$ itself is a random variable.

Examples

- $T(\mathbf{X}) = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- $T(\mathbf{X}) = med(X_1, X_2, \ldots, X_n)$
- $T(\mathbf{X}) = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$
- $T(\mathbf{X}) = \max(X_1, X_2, \ldots, X_n)$

Data Reduction as Partition of Sample Space

Data reduction can be represented as a partition of the sample space $\mathcal{X}$ determined by a statistic $T(\mathbf{X})$

Domain of T: $\mathcal{X}$

Range of T : $\mathcal{T} = \{t : t = T(\mathbf{X}) \text{ for some } \mathbf{x} \in \mathcal{X}\}$

Partition of $\mathcal{X}$: $A_t = \{\mathbf{x} : T(\mathbf{X}) = t, t \in \mathcal{T}\}$

Example

Suppose $X_i \sim$ iid Bernoulli(p) for $i = 1, 2, 3$, and $0 < p < 1$. Define $T(X_1, X_2, X_3) = X_1 + X_2 + X_3$

- What is the domain and range of T?
- How is the sample space partitioned by T?

Partition	X_1	X_2	X_3	$T(\mathbf{X}) = X_1 + X_2 + X_3$
A_0	0	0	0	0
A_1	0	0	1	1
	0	1	0	1
	1	0	0	1
A_2	0	1	1	2
	1	0	1	2
	1	1	0	2
A_3	1	1	1	3

Partition of the sample space based on $T(\mathbf{X})$ is “coarser” than the original sample space.

- There are 8 elements in the sample space $\mathcal{X}$.
- They are partitioned into 4 subsets
- Thus, $T(\mathbf{X})$ is simpler (or coarser) than $\mathbf{X}$.

Therefore, a data reduction can be achieved by $T(\mathbf{X})$.

Sufficiency

- Making original data ”simpler” is one goal of ideal data reduction.
- The other goal is to make inference about an underlying parameter θ. Want a statistic that contains all information about θ. (**Sufficient statistic**)
- In the previous example, what is the parameter θ that $T(\mathbf{X})$ is trying to estimate?
- Does the proposed $T(\mathbf{X})$ keep the information about θ contained in $\mathbf{X}$ or not?

Sufficiency Principle

If $T(\mathbf{X})$ is sufficient for θ, then any inference about θ should depend on the sample $\mathbf{X}$ only through the value of $T(\mathbf{X})$. Thus, for any two sample points $\mathbf{x}$ and $\mathbf{y}$ such that $T(\mathbf{x}) = T(\mathbf{y})$, then the inference about θ should be the same whether $\mathbf{X} = \mathbf{x}$ or $\mathbf{X} = \mathbf{y}$ is observed.

Definition: A statistic $T(\mathbf{X})$ is sufficient for θ if the conditional distribution of the sample $\mathbf{X}$ given the value of $T(\mathbf{x})$ does not depend on θ.

Example 1: Let $X_1, \cdots, X_n$ be i.i.d. from a pdf f. Then the set of order statistics $T(\mathbf{X}) = (X_{(1)} < X_{(2)} < \cdots < X_{(n)})$ is sufficient since the joint pdf of the random sample can be written as

$$f(\mathbf{x}) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} f(x_{(i)}).$$

Theorem 6.2.2: Let $f_{\mathbf{X}}(\mathbf{x}|\theta)$ is a joint pdf or pmf of $\mathbf{X}$. Further let $q(t|\theta)$ be the pdf or pmf of $T(\mathbf{X})$. Then $T(\mathbf{X})$ is a sufficient statistic for θ if, for every $\mathbf{x} \in \mathcal{X}$, the ratio

$$\frac{f_{\mathbf{X}}(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)}$$

is constant as a function of θ.

Proof: (Discrete Case)

Assume that the ratio $f_{\mathbf{X}}(\mathbf{x}|\theta)/q(T(\mathbf{x})|\theta)$ is constant, then

$$\begin{aligned}
\Pr\left(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = t\right) &= \frac{\Pr\left(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = t\right)}{\Pr(T(\mathbf{X}) = t)} \\
&= \begin{cases} \dfrac{\Pr(\mathbf{X} = \mathbf{x})}{\Pr(T(\mathbf{X}) = t)} & \text{if } T(\mathbf{x}) = t \\ \\ 0 & \text{otherwise} \end{cases} \\
&= \begin{cases} \dfrac{f_{\mathbf{X}}(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} & \text{if } T(\mathbf{x}) = t \\ \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

which does not depend on θ by assumption. Therefore, $T(\mathbf{X})$ is a sufficient statistic for θ.

Example 2: Bernoulli Distribution

Let $X_1, \cdots, X_n \sim$ iid Bernoulli(p), $0 < p < 1$. Show that $T(\mathbf{X}) = \sum_{i=1}^n X_i$ is a sufficient statistic for p.

Proof: Let $x_1, \cdots, x_n$ be the realization corresponding to the random variables $X_1, \cdots, X_n$.

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|p) &= p^{x_1}(1-p)^{1-x_1} \times p^{x_2}(1-p)^{1-x_2} \times \cdots \times p^{x_n}(1-p)^{1-x_n} \\ &= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \\ T(\mathbf{X}) &= \sum_{i=1}^n X_i \sim \text{Binomial}(n,p) \\ q(t|p) &= \binom{n}{t} p^t (1-p)^{n-t} \\ \frac{f_{\mathbf{X}}(\mathbf{x}|p)}{q(T(\mathbf{x})|p)} &= \frac{p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}}{\binom{n}{\sum_{i=1}^n x_i} p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}} \\ &= \frac{1}{\binom{n}{\sum_{i=1}^n x_i}} = \frac{1}{\binom{n}{T(\mathbf{x})}} \end{aligned}$$

By Theorem 6.2.2. $T(\mathbf{X})$ is a sufficient statistic for p.

Example 3: Let Let $X_1, \cdots, X_n \sim$ iid Normal$(\mu, 1)$. Show that the sample mean $\overline{X} = (X_1 + \cdots + X_n)/n$ is sufficient for μ.

Proof:

Factorization Teorem – Theorem 6.2.6

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

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$$

Remarks

- θ can be vector valued and so can be T
- g is a function of $T(\mathbf{x})$ as well as of θ.
- h is a function of $\mathbf{x}$, but must be free of θ.

Proof for Discrete Distributions

only if part : sufficient $\Longrightarrow$ *factorization*

Suppose that $T(\mathbf{X})$ is a sufficient statistic

$$\begin{aligned}
f_{\mathbf{X}}(\mathbf{x}|\theta) &= \Pr(\mathbf{X} = \mathbf{x}|\theta) \\
&= \Pr(\mathbf{X} = \mathbf{x} \text{ and } T(\mathbf{X}) = T(\mathbf{x})|\theta) \\
&= \Pr(T(\mathbf{X}) = T(\mathbf{x})|\theta) \Pr(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}), \theta) \\
&= \Pr(T(\mathbf{X}) = T(\mathbf{x})|\theta) \Pr(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))
\end{aligned}$$

Choose $g(t|\theta) = \Pr(T(\mathbf{X}) = t|\theta)$, and $h(\mathbf{x}) = \Pr\left(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})\right)$, then

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$$

if part : factorization $\Longrightarrow$ sufficient

Assume that the factorization $f_{\mathbf{X}}(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$ holds and let $q(t|\theta)$ be the pmf of $T(\mathbf{X})$. Define $A_t = \{\mathbf{y} : T(\mathbf{y}) = t\}$. Then

$$q(t|\theta) = \Pr(T(\mathbf{X}) = t|\theta) = \sum_{\mathbf{y}\in A_t} f_{\mathbf{X}}(\mathbf{y}|\theta)$$

$$\begin{aligned}
\frac{f_{\mathbf{X}}(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} &= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{q(T(\mathbf{x})|\theta)} \\
&= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} f_{\mathbf{X}}(\mathbf{y}|\theta)} \\
&= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} g(T(\mathbf{y})|\theta)h(\mathbf{y})} \\
&= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{g(T(\mathbf{x})|\theta)\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} h(\mathbf{y})} \\
&= \frac{h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} h(\mathbf{y})}
\end{aligned}$$

which is free of θ and hence by Theorem 6.2.2, $T(\mathbf{X})$ is sufficient for θ.

Example 4 (Bernoulli): Let $X_1, \cdots, X_n \sim$ iid Bernoulli(p), $0 < p < 1$.

$$\begin{aligned}
f_{\mathbf{X}}(\mathbf{x}|p) &= p^{x_1}(1-p)^{1-x_1}\cdots p^{x_n}(1-p)^{1-x_n} \\
&= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \\
&= p^{T(\mathbf{x})}(1-p)^{n-T(\mathbf{x})} \;=\; g(T(\mathbf{x})|p)h(\mathbf{x}),
\end{aligned}$$

where $g(t|p) = p^t(1-p)^{n-t}, h(\mathbf{x}) = 1$. Then by Factorization Theorem $T(\mathbf{X}) = \sum_{i=1}^n X_i$ is a sufficient statistic for p.

Example 5: Normal Distribution with known variance
Let $X_1, \cdots, X_n$ iid $\mathcal{N}(\mu, \sigma^2)$ with σ^2 known.

$$f_{\mathbf{X}}(\mathbf{x}|\mu) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x})^2 - \frac{n}{2\sigma^2}(\overline{x} - \mu)^2\right) \tag{1}$$

Take

$$h(\mathbf{x}) \;\; = \;\; (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{2\sigma^2}\right)$$

and

$$g(t|\mu) \;\; = \;\; \Pr(T(\mathbf{X}) = t|\mu) = \exp\left(-\frac{n(t-\mu)^2}{2\sigma^2}\right)$$

Then $f_{\mathbf{X}}(\mathbf{x}|\mu) = h(\mathbf{x})g(T(\mathbf{x})|\mu)$ holds, and $T(\mathbf{X}) = \overline{X}$ is a sufficient statistic for μ.

Example 6: Normal Distribution with both parameters unknown

Both μ and σ^2 are unknown. The parameter is a vector : $\boldsymbol{\theta} = (\mu, \sigma^2)$. The problem is to use the Factorization Theorem to find a sufficient statistic for $\boldsymbol{\theta}$.

Since the parameter is two-dimensional it is natural to assume that the sufficient statistic is also two dimensional. Consider

$$\mathbf{T}(\mathbf{X}) = (T_1(\mathbf{X}), T_2(\mathbf{X})) \equiv \big(\frac{1}{n}\sum_{i=1}^{n} X_i, \sum_{i=1}^{n}(X_i - \overline{X})^2\big).$$

Take

$$\begin{aligned} h(\mathbf{x}) &= 1 \\ g(t_1, t_2|\mu, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} t_2 - \frac{n}{2\sigma^2}(t_1 - \mu)^2\right) \end{aligned}$$

Then, in view of (1)

$$f_{\mathbf{X}}(\mathbf{x}|\mu, \sigma^2) = g(T_1(\mathbf{x}), T_2(\mathbf{x})|\mu, \sigma^2) h(\mathbf{x})$$

Thus, $\mathbf{T}(\mathbf{X}) = (T_1(\mathbf{x}), T_2(\mathbf{x})) = \left(\overline{x}, \sum_{i=1}^n (x_i - \overline{x})^2\right)$ is sufficient for $\boldsymbol{\theta} = (\mu, \sigma^2)$.

Equivalently, $(\overline{x}, s^2)$ is also sufficient for $\boldsymbol{\theta}$, where $s^2 = (n-1)^{-1} T_2$ is the sample variance.

Example 7 (Discrete Uniform) Let $X_1, \cdots, X_n$ be iid observations uniformly drawn from $\{1, \cdots, \theta\}$, where θ is a positive integer. Find a sufficient statistic for θ.

The pmf of discrete uniform is given by

$$f_X(x|\theta) = \begin{cases} 1/\theta & x = 1, 2, \cdots, \theta \\ \\ 0 & \text{otherwise} \end{cases}$$

The joint pmf of $X_1, \cdots, X_n$ is

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \begin{cases} \theta^{-n} & x_i \in \{1, 2, \cdots, \theta\}, \quad i = 1, 2, \ldots, n \\ \\ 0 & \text{otherwise} \end{cases}$$

Question: How can you implement factorization theorem here?

Example 8: Assume $X_1, \cdots, X_n$ iid Uniform$(\theta, \theta + 1)$, $-\infty < \theta < \infty$. Find a sufficient statistic for θ.

Proof:

Biostat 602 Winter 2017

Lecture Set 3

Principles of Data Reduction (Minimal Sufficiency)

Minimal Sufficient Statistic

Reading: CB 6.2

- Sufficient statistics are not unique.
- $T(\mathbf{x}) = \mathbf{x}$: The random sample itself is a trivial sufficient statistic for any θ.
- The set of order statistics $T(\mathbf{X}) = (X_{(1)}, \cdots, X_{(n)})$ is always a sufficient statistic for θ, if $X_1, \cdots, X_n$ are iid.
- For any sufficient statistic $T(\mathbf{X})$, its one-to-one function $q(T(\mathbf{X}))$ is also a sufficient statistic for θ.

Question Can we find a sufficient statistic that achieves the maximum data reduction?

Definition 6.2.11

A sufficient statistic $T(\mathbf{X})$ is called a *minimal sufficient statistic* if, for any other sufficient statistic $T'(\mathbf{X})$, $T(\mathbf{X})$ is a function of $T'(\mathbf{X})$.

Remarks

- $T(\mathbf{X})$ is a function of $T'(\mathbf{X}) \implies$ if $T'(\mathbf{x}) = T'(\mathbf{y})$ then $T(\mathbf{x}) = T(\mathbf{y})$.
- The sample space $\mathcal{X}$ consists of every possible sample - *finest* partition
- Given $T(\mathbf{X})$, $\mathcal{X}$ can be partitioned into A_t where $t \in \mathcal{T} = \{t : t = T(\mathbf{X}) \text{ for some } \mathbf{x} \in \mathcal{X}\}$
- Maximum data reduction is achieved when cardinality of $\mathcal{T}$ is minimal.
- If size of $\mathcal{T}' = \{t : t = T'(\mathbf{x}) \text{ for some } \mathbf{x} \in \mathcal{X}\}$ is not less than that of $\mathcal{T}$, then $\mathcal{T}$ is a minimal sufficient statistic. In this case, the partition induced by $\mathcal{T}$ is the *coarsest* possible.

Question 1: If T is *minimal sufficient*, is a one-to-one function of T also *minimal sufficient*?

Question 2: Is there always a one-to-one function between any two *minimal sufficient* statistics?

Note that sufficiency is tied to the parameter under consideration. Consider a random sample $X_1, \ldots, X_n$ from a $N(\mu, \sigma^2)$ population, where σ^2 is **known**. We have seen earlier that in this case, $T(\mathbf{X}) = \overline{X}$ is sufficient for μ. Consider the statistic $\mathbf{T}'(\mathbf{X}) = (T_1(\mathbf{X}), T_2(\mathbf{X})) = (\overline{X}, S^2)$.

- $\mathbf{T}'$ is sufficient for μ (factorization theorem).

- T achieves a coarser data reduction than $\mathbf{T}'$.

- No additional information is gained about μ from $\mathbf{T}'$.

- When σ^2 is not known, T is **not sufficient** for (μ, σ^2). In this case, $\mathbf{T}' = (\overline{X}, S^2)$ is jointly sufficient for (μ, σ^2).

Question Is $(\overline{X}, S^2)$ *minimal sufficient* for (μ, σ^2) (how to check)?

Theorem 6.2.13

Suppose $f_{\mathbf{X}}(\mathbf{x}|\theta)$ be the pdf or pmf of a sample $\mathbf{X}$ parameterized by θ. Suppose there exists a function $T(\mathbf{x})$ such that, for any two sample points $\mathbf{x}$ and $\mathbf{y}$, the ratio $f_{\mathbf{X}}(\mathbf{x}|\theta)/f_{\mathbf{X}}(\mathbf{y}|\theta)$ is constant as a function of θ if and only if $T(\mathbf{x}) = T(\mathbf{y})$. Then $T(\mathbf{x})$ is *minimal sufficient* for θ.

In other words

- $f_{\mathbf{X}}(\mathbf{x}|\theta)/f_{\mathbf{X}}(\mathbf{y}|\theta)$ is constant as a function of $\theta \Longrightarrow T(\mathbf{x}) = T(\mathbf{y})$.
- $T(\mathbf{x}) = T(\mathbf{y}) \Longrightarrow f_{\mathbf{X}}(\mathbf{x}|\theta)/f_{\mathbf{X}}(\mathbf{y}|\theta)$ is constant as a function of θ

Proof:

Example 1: Let X_1, X_2, X_3 be i.i.d. Bernoulli(p). Consider

$$T_1(\mathbf{X}) = X_1 + X_2 + X_3.$$

(a) Is T_1 sufficient for p?

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|p) &= p^{x_1+x_2+x_3}(1-p)^{3-x_1-x_2-x_3} \\ &= \left(\frac{p}{1-p}\right)^{x_1+x_2+x_3} (1-p)^3 \end{aligned}$$

$$\begin{aligned} h(\mathbf{x}) &= 1 \\ g(t|p) &= \left(\frac{p}{1-p}\right)^{t} (1-p)^3 \end{aligned}$$

Since

$$f_{\mathbf{X}}(\mathbf{x}|p) = g(x_1 + x_2 + x_3|p)h(\mathbf{x}),$$

by factorization Theorem, T_1 is sufficient for p.

(b) Is T_1 minimal sufficient for p?

$$\begin{aligned} \frac{f_{\mathbf{X}}(\mathbf{x}|\theta)}{f_{\mathbf{X}}(\mathbf{y}|\theta)} &= \frac{p^{\sum x_i}(1-p)^{3-\sum x_i}}{p^{\sum y_i}(1-p)^{3-\sum y_i}} \\ &= \left(\frac{p}{1-p}\right)^{\sum x_i - \sum y_i} \end{aligned}$$

– If $T_1(\mathbf{x}) = T_1(\mathbf{y})$, i.e. $\sum x_i = \sum y_i$, then the ratio does not depend on p.

– The ratio above is constant as a function of p only if $\sum x_i = \sum y_i$, i.e. $T_1(\mathbf{x}) = T_1(\mathbf{y})$.

Therefore, $T_1(\mathbf{X}) = \sum X_i$ is a minimal sufficient statistic for p by Theorem 6.2.13.

Example 2: Same premise as in Example 1. Consider

$$\mathbf{T}_2(\mathbf{X}) = (X_1 + X_2, X_3).$$

(a) Is $\mathbf{T_2}$ sufficient for p?

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|p) &= p^{x_1+x_2+x_3}(1-p)^{3-x_1-x_2-x_3} \\ &= p^{x_1+x_2}(1-p)^{2-x_1-x_2}p^{x_3}(1-p)^{1-x_3} \end{aligned}$$

$$\begin{aligned} h(\mathbf{x}) &= 1 \\ g(t_1, t_2|p) &= p^{t_1}(1-p)^{2-t_1}p^{t_2}(1-p)^{1-t_2} \\ \text{and } \ f_{\mathbf{X}}(\mathbf{x}|p) &= g(x_1+x_2, x_3|p)h(\mathbf{x}) \end{aligned}$$

Hence $\mathbf{T}_2(\mathbf{X}) = (X_1 + X_2, X_3)$ is sufficient for p.

(b) Is $\mathbf{T_2}$ minimal sufficient for p?

Let $A(\mathbf{X}) = X_1 + X_2$, and $B(\mathbf{X}) = X_3$.

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|p) &= p^{x_1+x_2}(1-p)^{2-x_1-x_2}p^{x_3}(1-p)^{1-x_3} \\ &= p^{A(\mathbf{x})+B(\mathbf{x})}(1-p)^{3-A(\mathbf{x})-B(\mathbf{x})} \end{aligned}$$

$$\begin{aligned} \frac{f_{\mathbf{X}}(\mathbf{x}|\theta)}{f_{\mathbf{X}}(\mathbf{y}|\theta)} &= \frac{p^{A(\mathbf{x})+B(\mathbf{x})}(1-p)^{3-A(\mathbf{x})-B(\mathbf{x})}}{p^{A(\mathbf{y})+B(\mathbf{y})}(1-p)^{3-A(\mathbf{x})-B(\mathbf{y})}} \\ &= \left(\frac{p}{1-p}\right)^{A(\mathbf{x})+B(\mathbf{x})-A(\mathbf{y})-B(\mathbf{y})} \end{aligned}$$

- The ratio above is constant as a function of p if (but not only if) $A(\mathbf{x}) = A(\mathbf{y})$ and $B(\mathbf{x}) = B(\mathbf{y})$
- The ratio is still constant as long as $A(\mathbf{x}) + B(\mathbf{x}) = A(\mathbf{y}) + B(\mathbf{y})$, even though $A(\mathbf{x}) \neq A(\mathbf{y})$ and $B(\mathbf{x}) \neq B(\mathbf{y})$

Therefore, $\mathbf{T}_2(\mathbf{X}) = (A(\mathbf{X}), B(\mathbf{X})) = (X_1 + X_2, X_3)$ is not a minimal sufficient statistic for p by Theorem 6.2.13.

Partition of the Sample Space

X_1	X_2	X_3	$\mathbf{T_2}(X) = (X_1 + X_2, X_3)$	$T_1(\mathbf{X}) = X_1 + X_2 + X_3$
0	0	0	(0,0)	0
0	0	1	(0,1)	3*1
0	1	0	2*(1,0)	
1	0	0		
0	1	1	2*(1, 1)	3*2
1	0	1		
1	1	0	(2,0)	
1	1	1	(2,1)	3

Clearly the partition induced by T_1 is coarser than the one induced by $\mathbf{T_2}$.

Some Algebraic Results

Assume that $a, b, c, d, a_1, \cdots, a_n$ are constants.

1. $a\theta^2 + b\theta + c = 0$ for any $\theta \in \mathbb{R} \quad \Leftrightarrow a = b = c = 0$.

2. $\sum_{i=1}^{k} a_i \theta^i = c$ for any $\theta \in \mathbb{R} \quad \Leftrightarrow a_1 = \cdots = a_k = 0,\ c = 0$.

3. $a\theta_1 + b\theta_2 + c = 0$ for all $(\theta_1, \theta_2) \in \mathbb{R}^2 \quad \Leftrightarrow a = b = c = 0$.

4. The following equation is constant

$$\frac{1 + a_1\theta + a_2\theta^2 + \cdots + a_k\theta_k^k}{1 + b_1\theta + b_2\theta^2 + \cdots + b_k\theta_k^k}$$

$\Leftrightarrow a_1 = b_1, \cdots, a_k = b_k$.

Note that this does not hold without the constant 1, for example,

$$\frac{\theta + 2\theta^2}{2\theta + 4\theta^2} = \frac{1}{2}$$

5. $\dfrac{I(a < \theta < b)}{I(c < \theta < d)}$ is a constant function of $\theta \quad \Leftrightarrow a = c$, and $b = d$.

6. θ^t is constant function of $\theta \quad \Leftrightarrow t = 0$.

Example 3: Let $X_1, \cdots, X_n$ be iid Uniform$(\theta, \theta+1)$, where $-\infty < \theta < \infty$. Find a minimal sufficient statistic for θ.

Joint pdf of X

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \prod_{i=1}^{n} I(\theta < x_i < \theta + 1)$$

Hence,

$$\begin{aligned}
\frac{f_{\mathbf{X}}(\mathbf{x}|\theta)}{f_{\mathbf{X}}(\mathbf{y}|\theta)} &= \frac{\prod_{i=1}^{n} I(\theta < x_i < \theta + 1)}{\prod_{i=1}^{n} I(\theta < y_i < \theta + 1)} \\
&= \frac{I(\theta < x_1 < \theta + 1, \cdots, \theta < x_n < \theta + 1)}{I(\theta < y_1 < \theta + 1, \cdots, \theta < y_n < \theta + 1)} \\
&= \frac{I(\theta < x_{(1)} \ \text{ and } \ x_{(n)} < \theta + 1)}{I(\theta < y_{(1)} \ \text{ and } \ y_{(n)} < \theta + 1)} \\
&= \frac{I(x_{(n)} - 1 < \theta < x_{(1)})}{I(y_{(n)} - 1 < \theta < y_{(1)})}
\end{aligned}$$

The ratio above is constant if and only if $x_{(1)} = y_{(1)}$ and $x_{(n)} = y_{(n)}$. Therefore, $\mathbf{T}(\mathbf{X}) = (X_{(1)}, X_{(n)})$ is a minimal sufficient statistic for θ.

Example 4(a): Let $X_1, \cdots, X_n$ be iid $\mathcal{N}(\mu, \sigma^2)$, where both parameters are unknown. The parameter is a vector: $\boldsymbol{\theta} = (\mu, \sigma^2)$. The problem is to use find a minimal sufficient statistic for $\boldsymbol{\theta}$.

The joint pdf

$$f_{\mathbf{X}}(\mathbf{x}|\mu, \sigma^2) \ = \ (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right)$$

$$\begin{aligned}
\frac{f_{\mathbf{X}}(\mathbf{x}|\mu,\sigma^2)}{f_{\mathbf{X}}(\mathbf{y}|\mu,\sigma^2)} &= \exp\left(-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}\right) / \exp\left(-\frac{\sum_{i=1}^n (y_i-\mu)^2}{2\sigma^2}\right) \\
&= \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n (x_i^2 - 2\mu x_i + \mu^2) - \sum_{i=1}^n (y_i^2 - 2\mu y_i + \mu^2)\right)\right] \\
&= \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2 - \sum_{i=1}^n y_i^2\right) + \frac{\mu}{\sigma^2}\left(\sum_{i=1}^n x_i - \sum_{i=1}^n y_i\right)\right]
\end{aligned}$$

The ratio above will not depend on (μ, σ^2) if and only if

$$\begin{cases} \sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2 \\ \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \end{cases}$$

Therefore, $\mathbf{T}(\mathbf{X}) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is a minimal sufficient statistic for (μ, σ^2) by Theorem 6.2.13

Define $\mathbf{T}'(\mathbf{X}) = (\overline{X}, \sum (X_i - \overline{X})^2/(n-1)) = (\overline{X}, S^2)$. Then, there exist one-to-one functions such that

$$\begin{aligned}
\sum X_i &= g_1\left(\overline{X}, \sum (X_i - \overline{X})^2/(n-1)\right) \\
\sum X_i^2 &= g_2\left(\overline{X}, \sum (X_i - \overline{X})^2/(n-1)\right)
\end{aligned}$$

and

$$\begin{aligned}
\overline{X} &= h_1(\sum X_i, \sum X_i^2) \\
\sum (X_i - \overline{X})^2/(n-1) &= h_2(\sum X_i, \sum X_i^2)
\end{aligned}$$

Thus $\mathbf{T}'$ is minimal sufficient.

Example 4(b): Let $X_1, \cdots, X_n$ be iid $\mathcal{N}(\mu, \sigma^2)$. In each of the following cases, identify a minimal sufficient statistic for the parameter of interest.

- When $\sigma = \sqrt{\mu}$.
- When $\sigma = \mu$.

Example 5: Let $X_1, \cdots, X_n$ be a random sample from $Gamma(\alpha, \beta)$ with pdf

$$f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp(-x/\beta).$$

Define $T_1(\mathbf{x}) = \prod_{i=1}^n x_i, \;\; T_2(\mathbf{x}) = \sum_{i=1}^n x_i$. Show that (T_1, T_2) are jointly sufficient for (α, β). Are they minimal sufficient?

Biostat 602 Winter 2017

Lecture Set 4

Principles of Data Reduction

Ancillary Statistics, Completeness

Ancillary Statistic

Reading: CB 6.2

- Sufficient statistics contain all information about θ.
- At the other extreme is a statistic which does not contain any information on θ.

Definition 6.2.11

A statistic $S(\mathbf{X})$ is an *ancillary statistic* if its distribution does not depend on θ.

Question: Why then bother about an ancillary statistic when making an inference on θ?

Examples

1. $X_1, \cdots, X_n$ iid $\mathcal{N}(\mu, \sigma^2)$ where σ^2 is known.
   - $X_1 - X_2 \sim \mathcal{N}(0, 2\sigma^2)$ is ancillary.
   - $(X_1 + X_2)/2 - X_3 \sim \mathcal{N}(0, 1.5\sigma^2)$ is ancillary.
   - $s^2_{\mathbf{X}} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is ancillary.
   - $\frac{(n-1)s^2_{\mathbf{X}}}{\sigma^2} \sim \chi^2_{n-1}$ is ancillary.

2. $X_1, \cdots, X_n$ iid $\mathcal{N}(0, \sigma^2)$ where σ^2 is unknown.

   - X_1/X_2 is ancillary.

   - $\overline{X}/S_{\mathbf{X}}$ is ancillary.

   - Is $\overline{X}/\sigma$ ancillary?

3. Let $X_1, \cdots, X_n$ iid $Uniform(\theta,\ \theta+1)$. Show that the range statistic

$$R = X_{(n)} - X_{(1)}$$

is ancillary. What is its distribution?

Location-Scale Family of Distributions

Let $f(x)$ be any pdf *free of any parameter* and let $-\infty < \mu < \infty$ and $\sigma > 0$ be unknown constants. Then

$$g(x|\mu, \sigma) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$$

is a pdf.

Proof: Because $f(x)$ is a pdf, then $f(x) \geq 0$, and $g(x|\mu, \sigma) \geq 0$ for all x. Let $y = (x-\mu)/\sigma$, then $x = \sigma y + \mu$, and $dx/dy = \sigma$.

$$\int_{-\infty}^{\infty} \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) dx \;=\; \int_{-\infty}^{\infty} \frac{1}{\sigma} f(y) \sigma dy = \int_{-\infty}^{\infty} f(y) dy = 1$$

Therefore, $g(x|\mu, \sigma)$ is also a pdf.

- The pdf g corresponds to a **location-scale** family of distribution with location $= \mu$ and scale $= \sigma$.

- When $\mu = 0$, g is the pdf of a scale family with **scale** parameter σ.

- When $\sigma = 1$, g is the pdf of a **location** family with location parameter μ.

How do you show a pdf belongs to a location-scale family?

Use the transformation $Y = (X - \mu)/\sigma$. If Y has a **parameter-free** pdf, then the original pdf belongs to a location-scale family.

Examples

1. $X \sim N(\mu, \sigma^2)$

2. $X \sim Exp(\theta)$

3. $X \sim Cauchy(\theta, 1)$

4. $X \;\sim\; Uniform(0, \theta)$

5. $X \;\sim\; Uniform(\theta, 2\theta)$

Ancillary Statistic for Location Family

Let $X_1, \cdots, X_n$ be iid from a location family with pdf $f(x-\mu)$ where $-\infty < \mu < \infty$. Show that the range $R = X_{(n)} - X_{(1)}$ is an ancillary statistic.

Solution: Since the original population distribution belongs to a location family, $Z_1 = X_1 - \mu, \cdots, Z_n = X_n - \mu$ are iid observations from pdf $f(x)$ and cdf $F(x)$, which are free of the parameter μ. Then the cdf of the range statistic R becomes

$$\begin{aligned} F_R(r|\mu) &= \Pr(R \leq r|\mu) = \Pr(X_{(n)} - X_{(1)} \leq r|\mu) \\ &= \Pr(Z_{(n)} + \mu - Z_{(1)} - \mu \leq r|\mu) = \Pr(Z_{(n)} - Z_{(1)} \leq r|\mu) \end{aligned}$$

which does not depend on μ because $Z_1, \cdots, Z_n$ does not depend on μ. Therefore, R is an ancillary statistic.

Ancillary Statistic for Scale Family

Let $X_1, \cdots, X_n$ be iid from a scale family with pdf $f(x/\sigma)/\sigma$ where $\sigma > 0$. Show that the statistic

$$\mathbf{T}(\mathbf{X}) = (X_1/X_n, \cdots, X_{n-1}/X_n) \qquad \text{is ancillary.}$$

Solution: Let $Z_1 = X_1/\sigma, \cdots, Z_n = X_n/\sigma$ be iid observations from pdf $f(x)$. Then the joint cdf of $\mathbf{T}(\mathbf{X})$ is

$$\begin{aligned} F_{\mathbf{T}}(t_1, \cdots, t_{n-1}|\sigma) &= \Pr(X_1/X_n \leq t_1, \cdots, X_{n-1}/X_n \leq t_{n-1}|\sigma) \\ &= \Pr(\sigma Z_1/\sigma Z_n \leq t_1, \cdots, \sigma Z_{n-1}/\sigma Z_n \leq t_{n-1}|\sigma) \\ &= \Pr(Z_1/Z_n \leq t_1, \cdots, Z_{n-1}/Z_n \leq t_{n-1}|\sigma) \end{aligned}$$

Because $Z_1, \cdots, Z_n$ does not depend on σ, $\mathbf{T}(\mathbf{X})$ is an ancillary statistic.

Ancillary vs Minimal Sufficient Statistic

- Ancillary statistic is free of θ.
- Minimal sufficient statistic contains minimal information related to θ.
- Are ancillary statistics independent of minimal sufficient statistics?

Example: For $X_1, \cdots, X_n \sim \text{Uniform}(\theta, \theta+1)$, $R = X_{(n)} - X_{(1)}$ and $M = (X_{(n)} + X_{(1)})/2$ are jointly minimal sufficient statistic (why?)

But R is ancillary statistic, so ancillary statistics are not always independent of minimal sufficient statistic.

However, how does R give any information about θ?

- If $M = 1$, then $0 < \theta < 1$ (why?).
- Suppose now $R = 0.8$. By itself, it does not provide any information about θ.
- In combination with the fact that $M = 1$, it yields that $X_{(1)} = 0.6$ and $X_{(n)} = 1.4$, and so the possible range of θ is narrowed down to $0.4 < \theta < 0.6$.
- Combination of ancillary statistic and another statistic can be more informative jointly than the other statistic alone.
- Thus, an ancillary statistic can provide additional precision about the parameter when combined with another statistic.

Completeness

In statistical inference, the ulterior objective is to identify a statistic that is a *good* estimator for the parameter. **Sufficiency** helps us identify statistics that contain information on the parameter. While somewhat counter-intuitive, **Ancillary** statistics enhance that information, while working in conjunction with a sufficient statistics. The final piece of the puzzle is the concept of **completeness**. Together, these three principles provide enough structure for us to pursue our quest for an efficient estimator in a systematic way.

Definition: Let $\{f_T(t|\theta), \theta \in \Omega\}$ be a family of pdfs or pmfs for a statistic $T(\mathbf{X})$. This family of probability distributions is called *complete* if

$$E[g(T)|\theta] = 0 \textit{ for all } \theta \text{ implies } \Pr[g(T) = 0|\theta] = 1 \text{ for all } \theta.$$

Remarks

- In other words, $g(T) = 0$ almost surely, i.e. only the zero function of T can have a mean of zero for all parameter values.
- Loosely $T(\mathbf{X})$ is called a *complete statistic.* However, as we shall see soon, completeness is the property of the family of distributions induced by T, and not that of T itself.
- Completeness implies 'no unnecessary part' conceptually. There is no non-trivial $g(T)$ whose expectation (or distribution) does not depend on θ.
- If an ancillary statistic could be made out of $T(\mathbf{X})$, it is NOT complete.
- This is a more stringent requirement than that is needed for minimal sufficient statistics.

Example 1: Let $X_1, \cdots, X_n$ be a random sample from a $Bern(p)$ population. Show that $T = \sum_{i=1}^n X_i$ is complete.

Example 2: Let $X_1, \cdots, X_n$ be a random sample from a $Uniform(0, \theta)$ population. Show that $T = \max_i X_i$ is complete.

Example 3: Let $X_1, \cdots, X_n$ be a random sample from a $Uniform(\theta, \theta+1)$ population. We know $\mathbf{T} = (X_{(1)}, X_{(n)})$ is minimal sufficient. Is $\mathbf{T}$ complete?

Example 4: Let $X_1, \cdots, X_n$ be a random sample from a $Pois(\lambda)$. Show that $T = \sum_{i=1}^{n} X_i$ is complete.

Example 5: Let $T \sim Pois(\lambda)$, where the parameter space of λ is given by

$$\Omega = \{\lambda : \lambda = \{1, 2\}\}.$$

Show that the family of distributions induced by T is NOT complete.

Proof: We need to find a counter example which is a function g such that $E[g(T)|\lambda] = 0$ for $\lambda = 1, 2$ but $g(T) \neq 0$. The function g must satisfy

$$E[g(T)|\lambda] = \sum_{t=0}^{\infty} g(t) \frac{\lambda^t e^{-\lambda}}{t!} = 0$$

for $\lambda \in \{1, 2\}$. Thus,

$$\begin{cases} E[g(T)|\lambda = 1] & = & \sum_{t=0}^{\infty} g(t) \frac{1^t e^{-1}}{t!} = 0 \\ \\ E[g(T)|\lambda = 2] & = & \sum_{t=0}^{\infty} g(t) \frac{2^t e^{-2}}{t!} = 0 \end{cases}$$

The above equation can be rewritten as

$$\begin{cases} \sum_{t=0}^{\infty} g(t)/t! & = & 0 \\ \\ \sum_{t=0}^{\infty} 2^t g(t)/t! & = & 0 \end{cases}$$

Define $g(t)$ as

$$g(t) \;\; = \;\; \begin{cases} 2 & t = 0 \text{ and } t = 2 \\ -3 & t = 1 \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\begin{aligned} \sum_{t=0}^{\infty} g(t)/t! &= g(0)/0! + g(1)/1! + g(2)/2! = 2 - 3 + 2/2 = 0 \\ \sum_{t=0}^{\infty} 2^t g(t)/t! &= g(0)/0! + 2g(1)/1! + 2^2 g(2)/2! = 2 - 6 + 8/2 = 0 \end{aligned}$$

There exists a non-zero function g that satisfies $E[g(T)|\lambda] = 0$ for all $\lambda \in \Omega$. Therefore this family is NOT complete.

Question: Why is a complete statistic called 'complete'?

Note that requiring $g(T)$ to satisfy the definition of completeness puts a restriction on g. The larger the family of pdfs/pmfs, the greater is the restriction on g. When the family of pdfs/pmfs is augmented to the point that $E[g(T)] = 0$ for all θ rules out all g except for the trivial $g(T) = 0$, then the family is said to be complete. A common verbalization of this definition is that the family of distributions is complete if there is no *unbiased estimator* of zero except for the trivial estimator $g \equiv 0$.

As the Poisson example shows, 'completeness' is a property of the family of distributions rather than the random variable or its parametric form.

Ancillary and Complete Statistics

Fact 1: For a statistic $T(\mathbf{X})$, if a non-constant function of T, say $r(T)$ is ancillary, then $T(\mathbf{X})$ cannot be complete.

Proof: Define $g(T) = r(T) - E[r(T)]$, which does not depend on the parameter θ because $r(T)$ is ancillary. Then $E[g(T)|\theta] = 0$ for a non-zero function $g(T)$, and $T(\mathbf{X})$ is not a complete statistic.

Arbitrary Functions of Complete Statistics

Fact 2: If $T(\mathbf{X})$ is a complete statistic, then a non-constant function of T, say $T^* = r(T)$ is also complete.

Proof: We can write

$$E[g(T^*)|\theta] \;=\; E[g \circ r(T)|\theta]$$

Now assume that $E[g(T^*)|\theta] = 0$ for all θ. Then

$$E[g \circ r(T)|\theta] = 0$$

holds for all θ too. Since $T(\mathbf{X})$ is a complete statistic,

$$\Pr[g \circ r(T) = 0] = 1, \;\forall \theta \in \Omega.$$

Therefore $\Pr[g(T^*) = 0] = 1$, and T^* is a complete statistic.

Completeness and sufficiency

Theorem 6.2.28: If a minimal sufficient statistic exists, then any complete sufficient statistic is also a minimal sufficient statistic.

Proof: Known as *Bahadur's Theorem*, beyond the scope of the course. Book statement is inaccurate.

Remarks:

- With the exception of very unusual cases, under a mild assumption, minimal sufficient statistics always exist.
- The converse is NOT true. A minimal sufficient statistic is not necessarily complete. Recall the example of Uniform$(\theta, \theta + 1)$.

Basu's Theorem

If $T(\mathbf{X})$ is a complete sufficient statistic, then $T(\mathbf{X})$ is independent of every ancillary statistic.

Proof – for discrete case

Suppose that $S(\mathbf{X})$ is an ancillary statistic. We want to show that

$$\Pr(S(\mathbf{X}) = s|T(\mathbf{X}) = t) = \Pr(S(\mathbf{X}) = s), \ \forall t \in \mathcal{T} \qquad (*)$$

Now we have, using *law of total probability*,

$$\Pr(S(\mathbf{X}) = s|\theta) = \sum_{t\in\mathcal{T}} \Pr(S(\mathbf{X}) = s|T(\mathbf{X}) = t) \Pr(T(\mathbf{X}) = t|\theta) \qquad (1)$$

Since, $\sum_{t\in\mathcal{T}} \Pr(T(\mathbf{X}) = t|\theta) = 1$, we can write

$$\begin{aligned}\Pr(S(\mathbf{X}) = s|\theta) &= \Pr(S(\mathbf{X}) = s) \sum_{t\in\mathcal{T}} \Pr(T(\mathbf{X}) = t|\theta) \\ &= \sum_{t\in\mathcal{T}} \Pr(S(\mathbf{X}) = s) \Pr(T(\mathbf{X}) = t|\theta) \qquad (2)\end{aligned}$$

Define $g(t) = \Pr(S(\mathbf{X}) = s|T(\mathbf{X}) = t) - \Pr(S(\mathbf{X}) = s)$. Using (1) and (2),

$$\sum_{t\in\mathcal{T}} \big[\Pr(S(\mathbf{X}) = s|T(\mathbf{X}) = t) - \Pr(S(\mathbf{X}) = s)\big] \Pr(T(\mathbf{X}) = t|\theta) = 0$$

This implies

$$\sum_{t\in\mathcal{T}} g(t) \Pr(T(\mathbf{X}) = t|\theta) = E[g(T(\mathbf{X}))|\theta] = 0$$

$T(\mathbf{X})$ is complete, so $g(t) = 0$ almost surely for all possible $t \in \mathcal{T}$.

Therefore, $(*)$ is established and $S(\mathbf{X})$ is independent of $T(\mathbf{X})$.

Example 6: Let $X_1, X_2, \ldots, X_n$ be a random sample from a $Uniform(0, \theta)$ distribution. Let $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ be the corresponding order statistics.

(a) Show that $X_{(n)}$ and $X_{(1)}/X_{(n)}$ are independent random variables.

(b) Establish that

$$E\left[\frac{X_{(1)}}{X_{(n)}}\right] = \frac{E(X_{(1)})}{E(X_{(n)})} = \frac{1}{n}.$$

Example 7: Let $X_1, X_2, \ldots, X_n$ be a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution. Conclude that $\overline{X}$ and S^2 are independent.

Example 8: *Exercise 6.19 CB* The random variable X takes the values 0, 1, 2, according to one of the following distributions:

	$\Pr(X=0)$	$\Pr(X=1)$	$\Pr(X=2)$	
Distribution 1	p	$3p$	$1-4p$	$0<p<\frac{1}{4}$
Distribution 2	p	p^2	$1-p-p^2$	$0<p<\frac{1}{2}$

In each case, determine whether the family of distribution of X is complete.

Solution - Distribution 1

Suppose that there exist $g(\cdot)$ such that $E[g(X)|p]=0$ for all $0<p<\frac{1}{4}$.

$$\begin{aligned} f_X(x|p) &= p^{I(x=0)}(3p)^{I(x=1)}(1-4p)^{I(x=2)} \\ E[g(X)|p] &= \sum_{x\in\{0,1,2\}} g(x)f_X(x|p) \\ &= g(0)\cdot p+g(1)\cdot(3p)+g(2)\cdot(1-4p) \\ &= p[g(0)+3g(1)-4g(2)]+g(2)=0 \end{aligned}$$

Therefore, $g(2)=0$, $g(0)+3g(1)=0$ must hold, and it is possible that g is a nonzero function that makes $\Pr[g(X)=0]<1$. For example, $g(0)=3, g(1)=-1, g(2)=0$. Therefore the family of distributions of X is not complete.

Solution - Distribution 2

Suppose that there exist $g(\cdot)$ such that $E[g(X)|p] = 0$ for all $0 < p < \frac{1}{4}$.

$$\begin{aligned}
f_X(x|p) &= p^{I(x=0)}(p^2)^{I(x=1)}(1-p-p^2)^{I(x=2)} \\
E[g(X)|p] &= \sum_{x\in\{0,1,2\}} g(x)f_X(x|p) \\
&= g(0)\cdot p + g(1)\cdot p^2 + g(2)\cdot(1-p-p^2) \\
&= p^2[g(1)-g(2)] + p[g(0)-g(2)] + g(2) = 0
\end{aligned}$$

$g(0) = g(1) = g(2) = 0$ must hold in order to $E[g(X)|p] = 0$ for all p. Therefore the family of distributions of X is complete.

Example 9: Let $X_1, \cdots, X_n$ *i.i.d.* $Pois(\lambda)$, where $\lambda > 0$ is unknown. Let $\overline{X},\ S^2$ denote the sample mean and variance, respectively. Show that

$$E\left[S^2 \middle| \overline{X}\right] = \overline{X} \text{ almost surely.}$$

Biostat 602 Winter 2017

Lecture Set 5

Principles of Data Reduction

Exponential Family of Distributions

Exponential Family

Reading: CB 6.2

Definition 3.4.1: The random variable X belongs to an exponential family of distributions, if its pdf/pmf can be written in the form

$$f(x|\boldsymbol{\theta}) \;=\; h(x)c(\boldsymbol{\theta})\exp\left[\sum_{j=1}^{k} w_j(\boldsymbol{\theta})t_j(x)\right], \quad x \in A$$

where

- $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_d),\ d \leq k,$
- $w_j(\theta),\ j \in \{1, \cdots, k\}$ and $c(\boldsymbol{\theta}) \geq 0$ are real valued functions of $\boldsymbol{\theta}$ alone,
- $t_j(x)$ and $h(x) \geq 0$ only involve data,
- Support of X, i.e. the set $A = \{x : f(x|\boldsymbol{\theta}) > 0\}$ does not depend on $\boldsymbol{\theta}$.

Example 1: Show that a Poisson(λ) ($\lambda > 0$) belongs to the exponential family.

Proof: The pmf of X can be written as

$$\begin{aligned} f_X(x|\lambda) &= \frac{e^{-\lambda}\lambda^x}{x!} \\ &= \frac{1}{x!}e^{-\lambda}\exp\left(\log\lambda^x\right) \\ &= \frac{1}{x!}e^{-\lambda}\exp\left(x\log\lambda\right) \end{aligned}$$

Define $h(x) = 1/x!$, $c(\lambda) = e^{-\lambda}$, $w(\lambda) = \log\lambda$, and $t(x) = x$, then

$$f_X(x|\lambda) \;=\; h(x)c(\lambda)\exp\left[w(\lambda)t(x)\right]$$

Example 2: $\mathcal{N}(\mu, \sigma^2)$ belongs to an Exponential Family

Proof: The pdf of X is can be written as:

$$\begin{aligned} f_X(x|\boldsymbol{\theta} = (\mu, \sigma^2)) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{x^2}{2\sigma^2} + \frac{2\mu x}{2\sigma^2} - \frac{\mu^2}{2\sigma^2}\right] \end{aligned}$$

Here $k = 2$. Define

$$w_1(\boldsymbol{\theta}) = \frac{\mu}{\sigma^2}, \quad t_1(x) = x, \quad w_2(\boldsymbol{\theta}) = -\frac{1}{2\sigma^2}, \quad t_2(x) = x^2,$$

$$h(x) = 1, \quad c(\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\mu^2}{2\sigma^2}\right].$$

Then

$$f_X(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left[\sum_{j=1}^{2} w_j(\boldsymbol{\theta}) t_j(x)\right]$$

Example 3: Show that $Gamma(\alpha, \beta)$ belongs to an Exponential Family

Proof:

Example 4: $Unif(0, \theta)$ does not belong to an Exponential Family.

Alternative Parameterization of Exponential Families

An alternative parametrization of the exponential family of distributions in terms of "natural" or "canonical" parameters can be written as follows.

$$f_X(x|\boldsymbol{\eta}) \;=\; h(x)c^*(\boldsymbol{\eta})\exp\left[\sum_{j=1}^{k}\eta_j t_j(x)\right]$$

The alternative parametrization can be achieved by defining $\eta_j = w_j(\boldsymbol{\theta})$ from the following equation,

$$f_X(x|\boldsymbol{\theta}) \;=\; h(x)c(\boldsymbol{\theta})\exp\left[\sum_{j=1}^{k} w_j(\boldsymbol{\theta})t_j(x)\right]$$

where $c^*(\boldsymbol{\eta}) = c \circ w(\boldsymbol{\theta})$. This alternative parametrization is most often used in the context of GLM (Generalized Linear Model).

Example 5: In *Bern*(p) distribution, the canonical parameter is the logit function

$$\eta = \log\left(\frac{p}{1-p}\right),$$

since for $x = 0, 1$

$$\begin{aligned} f(x|p) &= p^x(1-p)^{1-x} \\ &= (1-p)\left(\frac{p}{1-p}\right)^x \\ &= (1+e^\eta)^{-1}\exp(\eta x) \end{aligned}$$

with $c(\eta) = (1+e^\eta)^{-1},\ k = 1,\ t(x) = x,\ h(x) = 1.$

Example 6 For $\mathcal{N}(\mu, \sigma^2)$ distribution,

$$\begin{aligned} f_X(x|\mu) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \exp\left(\frac{\mu}{\sigma^2}x\right) \\ &= h(x)c(\boldsymbol{\eta}) \exp\left[\eta_1 t_1(x) + \eta_2 t_2(x)\right] \end{aligned}$$

where

$$\begin{cases} \boldsymbol{\eta} &= (\eta_1, \eta_2) = \left(\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2}\right) \\ t_1(x) &= x,\ t_2(x) = -x^2 \\ h(x) &= 1/\sqrt{\pi} \\ c(\boldsymbol{\eta}) &= \sqrt{\eta_2} \exp\left[-\eta_1^2/(2\eta_2)\right] \end{cases}$$

Sufficient Statistics and Exponential Families

Theorem 6.2.10: Let $X_1, \cdots, X_n$ *i.i.d.* with pdf $f_X(x|\boldsymbol{\theta})$ that belongs to an exponential family given by

$$f_X(x|\boldsymbol{\theta}) \;=\; h(x)c(\boldsymbol{\theta})\exp\left[\sum_{j=1}^{k} w_j(\boldsymbol{\theta})t_j(x)\right]$$

where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_d),\ d \leq k$. Then the following $T(\mathbf{X})$ is a sufficient statistic for $\boldsymbol{\theta}$.

$$T(\mathbf{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \cdots, \sum_{j=1}^{n} t_k(X_j)\right)$$

Example 7: Let $X_1, \cdots, X_n$ *i.i.d.* $\mathcal{N}(\mu, \sigma^2)$, where both $\mu \in \mathbb{R}$, and $\sigma^2 > 0$ are unknown. Find sufficient statistics for μ and σ^2.

From the exponential family representation on Page 3,
$T_1(\mathbf{X}) = \sum_{j=1}^{n} X_j,\ T_2(\mathbf{X}) = \sum_{j=1}^{n} X_j^2$ are sufficient statistics for μ, σ^2 by Theorem 6.2.10.

Example 8: Let $X_1, \cdots, X_n$ *i.i.d.* $Pois(\lambda)$, where $\lambda > 0$ is unknown. Find sufficient statistic for λ.

Digression

Definition: Open Set A set A is open in $\mathbb{R}^k$ if for every $x \in A$, there exists a ϵ-ball $B(x, \epsilon)$ around x such that $B(x, \epsilon) \subset A$. Here

$$B(x, \epsilon) = \{y : ||y - x|| < \epsilon,\ y \in \mathbb{R}^k\}$$

where $||$ denotes a distance measure in $\mathbb{R}^k$.

Examples

- $A = (-1, 1)$: A is open in $\mathbb{R}$
- $A = (-\infty, 0) \times \mathbb{R}$: A is open in $\mathbb{R}^2$
- $A = (-1, 1]$: A is not open in $\mathbb{R}$
- $A = (-\infty, 0] \times \mathbb{R}$: A is not open in $\mathbb{R}^2$
- $A = \{(x, y) : x^2 + y^2 < 1\}$: A is open in $\mathbb{R}^2$
- $A = \{(x, y) : x \in (-1, 1), y = 0\}$: A is not open in $\mathbb{R}^2$
- $A = \{(x, y) : x \in \mathbb{R}, y = x^2\}$: A is not open in $\mathbb{R}^2$
- $A = \{1, 2, 3, \cdots, \}$: A is not open in $\mathbb{R}$

This is the only additional concept one needs to connect exponential families to completeness.

Completeness and Exponential Families

Theorem 5.2.11 & 6.2.25: Suppose $X_1, \cdots, X_n$ is a random sample from pdf or pmf $f_X(x|\theta)$ where

$$f_X(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left[\sum_{j=1}^{k} w_j(\boldsymbol{\theta})t_j(x)\right]$$

is a member of an exponential family. Define a statistic $T(\mathbf{X})$ by

$$\mathbf{T}(\mathbf{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \cdots, \sum_{j=1}^{n} t_k(X_j)\right)$$

If the set $\boldsymbol{\Theta} = \{w_1(\boldsymbol{\theta}), \cdots, w_k(\boldsymbol{\theta}),\ \forall \boldsymbol{\theta} \in \boldsymbol{\Omega}\}$ contains an open subset of $\mathbb{R}^k$, then the following are true.

(a) The distribution of $\mathbf{T}(\mathbf{X})$ is an exponential family of the form

$$f_T(u_1, \cdots, u_k|\boldsymbol{\theta}) = H(u_1, \cdots, u_k)[c(\boldsymbol{\theta})]^n \exp\left[\sum_{j=1}^{k} w_j(\boldsymbol{\theta})u_i\right]$$

(b) The family of distributions for the statistic $T(\mathbf{X})$

$$\mathbf{T}(\mathbf{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \cdots, \sum_{j=1}^{n} t_k(X_j)\right)$$

is complete.

Example 9: Let $X_1, X_2, \dots, X_n$ be a i.i.d. random sample from the following distributions. Identify complete, sufficient statistics:

(a) $\mathcal{N}(\mu, \sigma^2)$

(b) $Poisson(\lambda)$

(c) $Bernoulli(p)$

(d) $Beta(\alpha, \beta)$

(e) $Inverse \;\; Gaussian(\mu, \lambda)$ with pdf

$$f(x|\mu, \lambda) = \left(\frac{\lambda}{2\pi x^2}\right)^{1/2} \exp\left[-\lambda(x-\mu)^2/2\mu^2 x\right], \quad 0 < x < \infty, \; \mu > 0, \; \lambda > 0.$$

(f) $Negative \;\; Binomial(r, p)$ with pmf

$$\binom{r+x-1}{x} p^r (1-p)^x, \quad x = 0, 1, 2, \cdots, ; 0 < p < 1; \quad r \; known.$$

Curved and Full Exponential Families

For an exponential family, if $d = \dim(\boldsymbol{\theta}) < k$, then this exponential family is called *curved exponential family.* if $d = \dim(\boldsymbol{\theta}) = k$, then this exponential family is called *full exponential family.* The sufficiency and completeness results only hold for *full exponential families.*

- $\mathcal{N}(\mu, \mu^2), \mu \in \mathbb{R}$ is a curved exponential family

The parameter space no longer contains an open set in R^2.

Review

Basic Terminology

Model $\mathcal{P} = \{f_{\mathbf{X}}(\mathbf{x}|\theta), \theta \in \Omega\}$

Random Variables $\mathbf{X} = (X_1, \cdots, X_n)$ that can be generated from $f_{\mathbf{X}}(\mathbf{x}|\theta)$.

Data $\mathbf{x} = (x_1, \cdots, x_n)$ that is generated from $f_{\mathbf{X}}(\mathbf{x}|\theta)$.

Statistic A function of data or random variables $T(\mathbf{x})$ or $T(\mathbf{X})$.

Sample Space A set of possible values of random variables $\mathcal{X}$.

Partition $A_t = \{\mathbf{x} : T(\mathbf{x}) = t\} \subseteq \mathcal{X}$.

Data Reduction Partition of sample space in terms of particular statistic.

Sufficient Statistic

Concept The statistic contains all information about θ

Definition 6.2.1 $f_{\mathbf{X}}(\mathbf{x}|T(\mathbf{X}))$ does not depend on θ

Theorem 6.2.2 $f_{\mathbf{X}}(\mathbf{x}|\theta)/q(T(\mathbf{X})|\theta)$ does not depend on θ
$\Longrightarrow T(\mathbf{X})$ is sufficient.

Theorem 6.2.6 (Factorization) $f_{\mathbf{X}}(\mathbf{x}|\theta) = h(\mathbf{x})g(T(\mathbf{X})|\theta)$
$\Longleftrightarrow T(\mathbf{X})$ is sufficient.

Theorem 6.2.10 (Exponential Family) $(\sum_{i=1}^n t_1(X_i), \cdots, \sum_{i=1}^n t_k(X_i))$ is sufficient

Minimal Sufficient Statistic

Concept Sufficient statistic that achieves the maximum data reduction, or coarsest partition of the sample space.

Definition 6.2.11 T is sufficient and it is a function of every sufficient statistic.

Theorem 6.2.13 $T(\mathbf{X})$ is minimal sufficient if the following is true:
$f_{\mathbf{X}}(\mathbf{x}|\theta)/f_{\mathbf{X}}(\mathbf{y}|\theta)$ is constant as a function of θ
$\Longleftrightarrow T(\mathbf{x}) = T(\mathbf{y})$

Non-unique MSS Any one-to-one function of MSS is also a MSS (i.e. MSS is not unique).

Unique partition The partition created by any minimal sufficient statistic is unique.

Theorem 6.2.28 Any complete sufficient statistic is also minimal sufficient.

Ancillary Statistic

Concept A statistic that does not have any information about θ.

Definition 6.2.16 Its distribution is constant to θ.

Location Family For location family of θ, $\{f(x-\theta) : \theta \in R\}$, range statistic is an ancillary statistic of θ.

Scale Family For scale family of θ, $\left\{\frac{1}{\sigma}f\left(\frac{x}{\sigma}\right) : \sigma > 0\right\}$, median$(X)/\overline{X}$ or $X_{(1)}/X_{(n)}$ is an ancillary statistics for θ.

Theorem 6.2.24 (Basu) A complete sufficient statistic is independent of every ancillary statistic.

Complete Statistic

Concept Any non-zero function of the statistics cannot be ancillary (there is no unnecessary part).

Defined on family This family has to contain ”many” distributions in order to be complete.

Definition 6.2.21 $E[g(T)|\theta] = 0 \Longrightarrow g(T) = 0$ almost surely across all θ.

Theorem 6.2.24 (Basu) A complete and (minimal) sufficient statistic is independent of every ancillary statistics.

Theorem 6.2.25 (Exponential Family) If the parameter space $\Theta = \{(w_1(\theta), \ldots, w_k(\theta) : \theta \in \Omega\}$ contains an open subset of $\mathbb{R}^k$, then $(\sum_{i=1}^n t_1(X_i), \cdots, \sum_{i=1}^n t_k(X_i))$ is complete.

Theorem 6.2.28 Any complete sufficient statistic is also minimal sufficient.

Example 10: Let $X_1, \cdots, X_n$ be *i.i.d.* random sample from the following pdf

$$f_X(x|\theta) = e^{-(x-\theta)} \exp(-e^{-(x-\theta)}), \ -\infty < x < \infty, \quad -\infty < \theta < \infty$$

1. Does the distribution belong to an exponential family?
2. What are canonical parameters for the distribution?
3. What is a sufficient statistic from the distribution?
4. Is the sufficient statistic also complete and/or minimal sufficient?

Representing into an exponential distribution

$$\begin{aligned} f_X(x|\theta) &= e^{-(x-\theta)} \exp(-e^{-(x-\theta)}) \\ &= e^{-x} e^{\theta} \exp(-e^{-x} e^{\theta}) \\ &= h(x) c(\theta) \exp[w(\theta) t(x)] \qquad \text{if} \end{aligned}$$

$$h(x) = e^{-x,} \qquad c(\theta) = e^{\theta}, \qquad w(\theta) = -e^{\theta}, \qquad t(x) = e^{-x}$$

Representing into a canonical form

$$f_X(x|\theta) \;=\; h(x)c^*(\eta)\exp[\eta t(x)] \qquad \text{if}$$

$$h(x) = e^{-x,} \qquad c^*(\eta) = -\eta, \qquad t(x) = e^{-x}, \qquad (\eta < 0)$$

Sufficiency

$$T(\mathbf{X}) = \sum_{i=1}^{n} t(X_i) = \sum_{i=1}^{n} e^{-X_i}$$

Completeness and minimal sufficiency

The parameter space $\mathbf{\Theta} = \{w(\theta) = -e^{\theta} : \theta \in \mathbb{R}\}$ contains an open set in $\mathbb{R}$, so it is complete by Theorem 6.2.25, and minimal sufficient by Theorem 6.2.28.

Biostat 602 Winter 2017

Lecture Set 6

Point Estimation

Methods of Finding Estimators

Reading: CB 7.1–7.2

Point Estimation

Basic Premise

- Data: $\mathbf{x} = (x_1, \cdots, x_n)$ - realizations of random variables $(X_1, \cdots, X_n)$.
- $X_1, \cdots, X_n$ *i.i.d.* $f_X(x|\theta)$.
- Assume a model $\mathcal{P} = \{f_X(x|\theta) : \theta \in \Omega \subset \mathbb{R}^p\}$ where the functional form of $f_X(x|\theta)$ is known, but θ is unknown.
- Task is to use data $\mathbf{x}$ to make inference on θ

Definition If we use a function of sample $w(X_1, \cdots, X_n)$ as a "guess" of $\tau(\theta)$, where $\tau(\theta)$ is a function of true parameter θ. Then $w(\mathbf{X}) = w(X_1, \cdots, X_n)$ is called a *point estimator* of $\tau(\theta)$. The realization of the estimation, $w(\mathbf{x}) = w(x_1, \cdots, x_n)$ is called the *estimate* of $\tau(\theta)$.

Example 1: Let $X_1, \cdots, X_n$ *i.i.d.* $\mathcal{N}(\theta, 1)$, where $\theta \in \Omega \in \mathbb{R}$.

- Suppose $n = 6$, and $(x_1, \cdots, x_6) = (2.0,\ 2.1,\ 2.9,\ 2.6,\ 1.2,\ 1.8)$.
- Define the estimator $w_1(X_1, \cdots, X_n) = \frac{1}{n}\sum_{i=1}^n X_i = \overline{X}$. The estimate is 2.1.
- Define the estimator $w_2(X_1, \cdots, X_n) = X_{(1)}$. Its estimate is 1.2.

The estimator is a statistic that is constructed with an objective of making inference about a parameter. Thus, the specific structural form of the function of the parameter $\tau(\theta)$ is crucial in defining an estimator.

Clearly, the class of estimators for a given problem is infinite until we restrict our search to a given class.

We first explore different approaches to obtaining estimators.

Subsequently, we look at methods to evaluate these estimators and search for an optimal one using these criteria.

Method of Moments Estimation

The method of moments is a simple method of estimation that dates back to Karl Pearson, the Father of Statistics, in the late 1800s. It is a method to equate sample moments to population moments and solve the resulting equations for the parameters.

Sample moments	Population moments
$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\mu_1' = E[X\|\theta] = \mu_1'(\theta)$
$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$	$\mu_2' = E[X^2\|\theta] = \mu_2'(\theta)$
$m_3 = \frac{1}{n}\sum_{i=1}^{n} X_i^3$	$\mu_3' = E[X^3\|\theta] = \mu_3'(\theta)$
$\vdots$	$\vdots$

Point estimator of $\tau(\theta)$ is obtained by solving equations like this.

$$
\begin{aligned}
m_1 &= \mu_1'(\theta) \\
m_2 &= \mu_2'(\theta) \\
\vdots &\quad \vdots \\
m_k &= \mu_k'(\theta)
\end{aligned}
$$

Example 2: Let $X_1, \cdots, X_n$ be *i.i.d.* from $\mathcal{N}(\mu, \sigma^2)$ population. Find method of moments (MoM) estimator for μ, σ^2.

Solution: Note that

$$
\begin{aligned}
\mu_1' &= E(\mathbf{X}) = \mu = \overline{X} \\
\mu_2' &= E(\mathbf{X}^2) = [E(\mathbf{X})]^2 + \mathrm{Var}(\mathbf{X}) = \mu^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2
\end{aligned}
$$

The MoM estimators are obtained by setting up the equations

$$
\begin{cases}
\hat{\mu} = \overline{X} \\
\hat{\mu}^2 + \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2
\end{cases}
$$

Solving the two equations above, $\hat{\mu} = \overline{X}$, $\widehat{\sigma^2} = \sum_{i=1}^{n}(X_i - \overline{X})^2/n$, which are the required MoM estimators for μ, σ^2, respectively.

Example 3: Let $X_1, \cdots, X_n$ be *i.i.d.* from *Binomial*(k, p). Find a MoM estimator for k, p.

Remark: This application is somewhat unusual in the sense that we are interested here in estimating the parameter k which is treated as known in most applications. Examples of such application include (a) estimating the reporting rate of crimes that are typically under-reported such as domestic violence, and (b) estimating detection rate of bugs in a software code.

Solution: The pmf is given by

$$f_X(x|k,p) \;=\; \binom{k}{x} p^x (1-p)^{k-x} \qquad x \in \{0, 1, \cdots, k\}$$

Equating first two sample moments,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} X_i &= \overline{X} \approx \mu_1' = E(\mathbf{X}) = kp \\
\frac{1}{n}\sum_{i=1}^{n} X_i^2 &\approx \mu_2' = E[\mathbf{X}^2] = (E\mathbf{X})^2 + \mathrm{Var}(\mathbf{X}) = k^2p^2 + kp(1-p)
\end{aligned}$$

Solving these equations,

$$\begin{aligned}
\overline{X} &= \hat{k}\hat{p} \\
\frac{1}{n}\sum_{i=1}^{n} X_i^2 &= \hat{k}^2\hat{p}^2 + \hat{k}\hat{p}(1-\hat{p}) \\
&= \overline{X}^2 + \overline{X}(1-\hat{p}) \\
\hat{p} &= 1 - \frac{(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2)}{\overline{X}} \\
&= \frac{\overline{X} - \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}{\overline{X}} \\
\hat{k} &= \frac{\overline{X}}{\hat{p}} = \frac{\overline{X}^2}{\overline{X} - \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}
\end{aligned}$$

ARE THESE GOOD ESTIMATORS?

Example 4: Let $X_1, \cdots, X_n$ be i.i.d. Negative Binomial(r, p). Find method of moments estimator for (r, p).

Solution: The moment equations are

$$\begin{aligned} m_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i = E(\mathbf{X}) = \frac{r(1-p)}{p} \\ m_2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 = E(\mathbf{X}^2) = \left(\frac{r(1-p)}{p}\right)^2 + \frac{r(1-p)}{p^2} \end{aligned}$$

which gives

$$\hat{p} = \frac{m_1}{m_2 - m_1^2} = \frac{\overline{X}}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2}$$

$$\hat{r} = \frac{m_1 \hat{p}}{1-\hat{p}} = \frac{\overline{X}\hat{p}}{1-\hat{p}}$$

Example 5: Let $X_1, \cdots, X_n$ be *i.i.d.* $Unif(-\theta, \theta)$. What is the MoM estimator for θ?

Remarks

- MoM estimators are used to match sample moments to population moments, the latter of which is typically a function of the model parameters. The estimators for these model parameters are then obtained by solving equations. Thus, to estimate $\tau(\theta)$, one first solves $\overline{X} = \mu(\hat{\theta})$ to obtain MoM estimator $\hat{\theta}$ of θ and then use $\tau(\hat{\theta})$ as the MoM estimator of $\tau(\theta)$. For example,

$$\hat{\theta}_{MoM} = \exp(\overline{X}) \implies \hat{\theta}^{-1}_{MoM} = \exp(-\overline{X}).$$

- It is possible to have multiple moment equations estimating θ. For example, both $\overline{X}$ and $\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$ estimate the mean λ of a Poisson distribution. The custom in this case is to call the estimator involving lower order moments ($\overline{X}$ in the Poisson case) the MoM estimator.

- The MoM estimator is always calculated in the untransformed scale. For example, in the case of $X_1, \cdots, X_n$ *i.i.d.* from a $Unif(-\theta, \theta)$ population, we know that $|X|_1, \cdots, |X|_n$ is a random sample from $Unif(0, \theta)$. Yet, $\frac{2}{n}\sum_{i=1}^{n} |X|_i$ is not a MoM estimator of θ.

Maximum Likelihood Estimation

Likelihood Function

Definition: Let $X_1, \cdots, X_n \sim$ i.*i.d.* $f_X(x|\theta)$. The joint distribution of $\mathbf{X} = (X_1, \cdots, X_n)$ is

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \prod_{i=1}^{n} f_X(x_i|\theta)$$

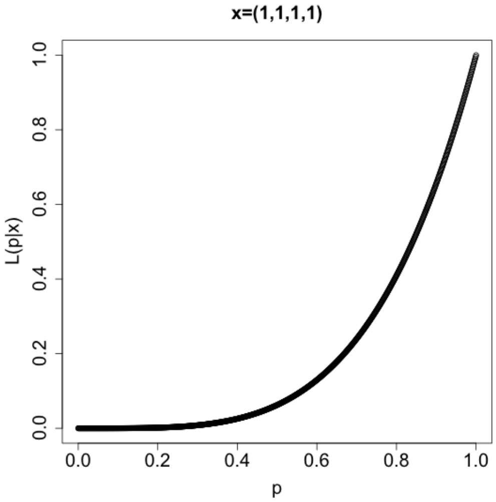
Given that $\mathbf{X} = \mathbf{x}$ is observed, the function of θ defined by

$$L(\theta|\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}|\theta)$$

is called the likelihood function.

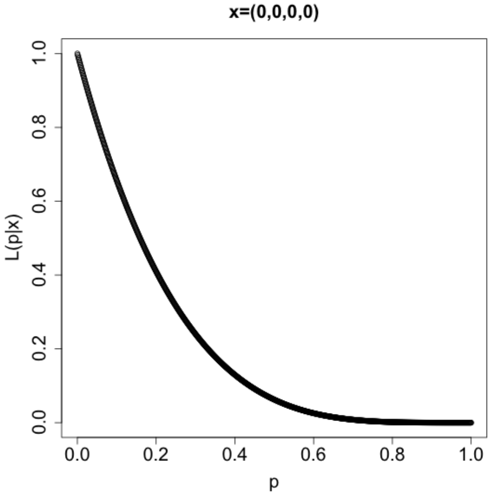
Example 5(a): Let X_1, X_2, X_3, X_4 be *i.i.d. Bernoulli*(p), $0 < p < 1$.

- $\mathbf{x} = (1, 1, 1, 1)^T$
- Intuitively, it is more likely that p is larger than smaller.
- $L(p|\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}|p) = \prod_{i=1}^{4} p^{x_i}(1-p)^{1-x_i} = p^4$.



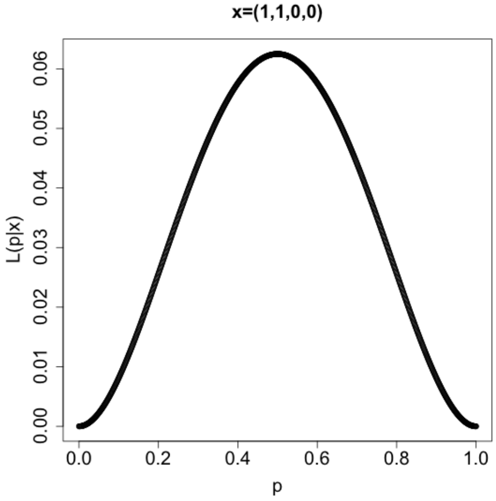
Example 5(b): Let X_1, X_2, X_3, X_4 be *i.i.d. Bernoulli*(p), $0 < p < 1$.

- $\mathbf{x} = (0,0,0,0)^T$
- Intuitively, it is more likely that p is smaller than larger.
- $L(p|\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}|p) = \prod_{i=1}^4 p^{x_i}(1-p)^{1-x_i} = (1-p)^4$.



Example 5(c): Let X_1, X_2, X_3, X_4 be *i.i.d. Bernoulli*(p), $0 < p < 1$.

- $\mathbf{x} = (1,1,0,0)^T$
- Intuitively, it is more likely that p is somewhere in the middle than in the extremes.
- $L(p|\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}|p) = \prod_{i=1}^4 p^{x_i}(1-p)^{1-x_i} = p^2(1-p)^2$.



Maximum Likelihood Estimator

Definition: For a given sample point $\mathbf{x} = (x_1, \cdots, x_n)$, let $\hat{\theta}(\mathbf{x})$ be the value such that $L(\theta|\mathbf{x})$ attains its maximum.

More formally,

$$L(\hat{\theta}(\mathbf{x})|\mathbf{x}) \geq L(\theta|\mathbf{x}) \quad , \forall \theta \in \Omega, \quad \hat{\theta}(\mathbf{x}) \in \Omega.$$

$\hat{\theta}(\mathbf{x})$ is called the *maximum likelihood estimate* of θ based on data $\mathbf{x}$,

$\hat{\theta}(\mathbf{X})$ is the *maximum likelihood estimator (MLE)* of θ.

Example 6: Let $X_1, \cdots, X_n$ be *i.i.d.* *Exp*(β). Find MLE of β.

Solution: The likelihood function is

$$\begin{aligned} L(\beta|\mathbf{x}) &= f_{\mathbf{X}}(\mathbf{x}|\theta) = \prod_{i=1}^{n} f_X(x_i|\theta) \\ &= \prod_{i=1}^{n} \left[\frac{1}{\beta} e^{-x_i/\beta} \right] = \frac{1}{\beta^n} \exp\left(-\sum_{i=1}^{n} \frac{x_i}{\beta} \right) \end{aligned}$$

where $\beta > 0$.

Use the derivative to find potential MLE. Maximizing the likelihood function $L(\beta|\mathbf{x})$ is equivalent to maximize the log-likelihood function

$$\begin{aligned} l(\beta|\mathbf{x}) &= \log L(\beta|\mathbf{x}) = \log \left[\frac{1}{\beta^n} \exp\left(-\sum_{i=1}^{n} \frac{x_i}{\beta} \right) \right] \\ &= -\frac{\sum_{i=1}^{n} x_i}{\beta} - n \log \beta \end{aligned}$$

Setting the first derivative of the log-likelihood equal to zero, we get

$$\frac{\partial l}{\partial \beta} = \frac{\sum_{i=1}^n x_i}{\beta^2} - \frac{n}{\beta} = 0$$

that simplifies to

$$\sum_{i=1}^n x_i = n\beta$$

which yields the solution as

$$\hat{\beta} \;=\; \frac{\sum_{i=1}^n x_i}{n} = \overline{x}$$

Question: Is $\hat{\beta}$ the maximum likelihood estimator?

Use the double derivative to confirm local maximum.

$$\begin{aligned}
\left.\frac{\partial^2 l}{\partial \beta^2}\right|_{\beta=\overline{x}} &= \left.-2\frac{\sum_{i=1}^n x_i}{\beta^3} + \frac{n}{\beta^2}\right|_{\beta=\overline{x}} \\
&= \left.\frac{1}{\beta^2}\left(-\frac{2\sum_{i=1}^n x_i}{\beta} + n\right)\right|_{\beta=\overline{x}} \\
&= \frac{1}{\overline{x}^2}\left(-\frac{2n\overline{x}}{\overline{x}} + n\right) \\
&= \frac{1}{\overline{x}^2}(-n) < 0
\end{aligned}$$

Therefore, we can conclude that $\hat{\beta} = \overline{X}$ is unique local maximum on the interval $(0, \infty)$.

Check boundary and confirm global maximum

$\beta \in (0, \infty)$. If $\beta \to \infty$

$$
\begin{aligned}
l(\beta|\mathbf{x}) &= -\frac{\sum_{i=1}^n x_i}{\beta} - n \log \beta \to -\infty \\
L(\beta|\mathbf{x}) &\to 0
\end{aligned}
$$

If $\beta \to 0$, one can also show that $l(\beta|\mathbf{x}) \to -\infty$. This is harder to verify. Visualize this by plotting $l(\beta|\mathbf{x})$ against β.

Since at both ends, L dies off to zero, the local maximum at the interior is indeed the global maximum.

Putting Things Together

1. $\frac{\partial l}{\partial \beta} = 0$ at $\hat{\beta} = \overline{x}$
2. $\frac{\partial^2 l}{\partial \beta^2} < 0$ at $\hat{\beta} = \overline{x}$
3. $L(\beta|\mathbf{x}) \to 0$ (lowest bound) when β approaches the boundary

Therefore $l(\beta|\mathbf{x})$ and $L(\beta|\mathbf{x})$ attains the global maximum when $\hat{\beta} = \overline{x}$

$\hat{\beta}(\mathbf{X}) = \overline{X}$ is the MLE of β.

How do we find MLE?

If the function is differentiable with respect to θ

1. Find candidates that makes first order derivative to be zero
2. Check second-order derivative to check local maximum.
   - For one-dimensional parameter, $\frac{\partial^2 L(\theta)}{\partial \theta^2} < 0$ implies local maximum.
   - For two-dimensional parameter, we need to show
     (a) $\partial^2 L(\theta_1, \theta_2)/\partial\theta_1^2 < 0$ or $\partial^2 L(\theta_1, \theta_2)/\partial\theta_2^2 < 0$.
     (b) Determinant of second-order derivative is positive
3. Check whether boundary gives global maximum.
   - Or clearly justify that boundaries cannot be global maximum.

If the function is NOT differentiable with respect to θ

- Use numerical methods, or
- Directly maximize using inequalities or properties of the function.

Example 7: Let $X_1, \cdots, X_n$ be *i.i.d.* $Uniform(0, \theta)$, where $X_i \in (0, \theta)$ and $\theta > 0$. Find MLE of θ.

Example 8: Suppose n pairs of data $(X_1, Y_1), \cdots, (X_n, Y_n)$ where X_i is generated from an unknown distribution, and Y_i are generated conditionally on X_i.

$$Y_i | X_i \sim \mathcal{N}(\alpha + \beta X_i, \sigma^2)$$

Find the MLE of $(\alpha, \beta, \sigma^2)$.

Solution: The joint distribution of $(X_1, Y_1), \cdots, (X_n, Y_n)$ is

$$f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{X}}(\mathbf{x}) \prod_{i=1}^{n} f_{\mathbf{Y}}(y_i | x_i) = f_{\mathbf{X}}(\mathbf{x}) \prod_{i=1}^{n} \frac{1}{2\pi\sigma^2} \exp\left[-\frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2}\right]$$

The likelihood function is

$$L(\alpha, \beta, \sigma^2 | \mathbf{x}, \mathbf{y}) \;=\; f_{\mathbf{X}}(\mathbf{x})(2\pi\sigma^2)^{-n/2} \exp\left[-\frac{\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2}{2\sigma^2}\right]$$

The log-likelihood function can be simplied as

$$l(\alpha, \beta, \sigma^2) \;=\; C - \frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2}{2\sigma^2}$$

$$\begin{aligned} \frac{\partial l}{\partial \alpha} &= \frac{2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)}{2\sigma^2} = \frac{n\overline{y} - n\alpha - n\beta\overline{x}}{\sigma^2} = 0 \\ \hat{\alpha} &= \overline{y} - \hat{\beta}\overline{x} \end{aligned}$$

$$\frac{\partial l}{\partial \beta} = \frac{2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)x_i}{2\sigma^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\alpha\overline{x} - \beta\sum_{i=1}^{n} x_i^2}{\sigma^2} = 0$$

$$\sum_{i=1}^{n} x_i y_i - n\overline{x}(\overline{y} - \beta\overline{x}) - \beta\sum_{i=1}^{n} x_i^2 = 0$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}$$

$$\begin{aligned}
\frac{\partial l}{\partial \sigma^2} &= -\frac{n}{2}\frac{2\pi}{2\pi\sigma} + \frac{\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2}{2(\sigma^2)^2} = 0 \\
\hat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta} x_i)^2
\end{aligned}$$

Putting Things Together

Therefore, the MLE of $(\alpha, \beta, \sigma^2)$ is

$$\begin{aligned}
\hat{\alpha}(\mathbf{X}, \mathbf{Y}) &= \overline{Y} - \hat{\beta}\overline{X} \\
\hat{\beta}(\mathbf{X}, \mathbf{Y}) &= \frac{\sum_{i=1}^n X_i Y_i - n\overline{XY}}{\sum_{i=1}^n X_i^2 - n\overline{X}^2} \\
\hat{\sigma^2}(\mathbf{X}, \mathbf{Y}) &= \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2
\end{aligned}$$

Biostat 602 Winter 2017

Lecture Set 7

Point Estimation

Maximum Likelihood Estimation

Reading: CB 7.2

Maximum Likelihood Estimation

Recap

$X_1, \cdots, X_n$ *i.i.d.* $f_X(x|\theta)$. The joint distribution of $\mathbf{X} = (X_1, \cdots, X_n)$ is

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \prod_{i=1}^{n} f_X(x_i|\theta)$$

Given that $\mathbf{X} = \mathbf{x}$ is observed, the function of θ defined by $L(\theta|\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}|\theta)$ is called the **likelihood function**.

For a given sample point $\mathbf{x} = (x_1, \cdots, x_n)$, let $\hat{\theta}(\mathbf{x})$ be the value such that $L(\theta|\mathbf{x})$ attains its maximum. More formally,

$$L(\hat{\theta}(\mathbf{x})|\mathbf{x}) \geq L(\theta|\mathbf{x}) \quad , \forall \theta \in \Omega, \quad \text{where} \quad \hat{\theta}(\mathbf{x}) \in \Omega.$$

$\hat{\theta}(\mathbf{x})$ is called the *maximum likelihood estimate* of θ based on data $\mathbf{x}$, and

$\hat{\theta}(\mathbf{X})$ is the *maximum likelihood estimator (MLE)* of θ.

Strategies for finding MLE of θ

There are two situations.

If the function is differentiable with respect to θ

1. Find candidates that makes first order derivative to be zero
2. Check second-order derivative to check local maximum.
   - For one-dimensional parameter, $\frac{\partial^2 L(\theta)}{\partial\theta^2} < 0$ implies local maximum.
   - For two-dimensional parameter, we need to show
     (a) $\partial^2 L(\theta_1, \theta_2)/\partial\theta_1^2 < 0$ or $\partial^2 L(\theta_1, \theta_2)/\partial\theta_2^2 < 0$.
     (b) Determinant of second-order derivative is positive
3. Check whether boundary gives global maximum.
   - Or clearly justify that boundaries cannot be global maximum.

If the function is NOT differentiable with respect to θ

- Use numerical methods, or
- Directly maximize using inequalities or properties of the function.

In general, one is content with MLEs that are local maximum.

Example 1 – Normal MLEs, both parameters unknown

Let $X_1, \cdots, X_n$ be *i.i.d* observations from $\mathcal{N}(\mu, \sigma^2)$. Find MLE of (μ, σ^2).

Two possible approaches

- Use second-order partial derivatives and their Hessian to show global maximum
- Find a workaround to avoid complex calculations.

Common step : Calculate first-order derivatives

Likelihood Function

$$\begin{aligned} L(\mu, \sigma^2|\mathbf{x}) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}\right] \\ l(\mu, \sigma^2|\mathbf{x}) &= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2} \end{aligned}$$

Partial derivative with respect to μ

$$\begin{aligned} L(\mu, \sigma^2|\mathbf{x}) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}\right] \\ l(\mu, \sigma^2|\mathbf{x}) &= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2} \\ \frac{\partial l}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0 \\ \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n}x_i = \overline{x} \end{aligned}$$

partial derivative with respect to σ^2

$$\begin{aligned}\frac{\partial l}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \\ \hat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\end{aligned}$$

Checking second-order partial derivatives

With respect to μ

$$\frac{\partial^2 l}{\partial \mu^2} = -\frac{n}{\sigma^2} < 0$$

With respect to σ^2

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

With respect to both parameters

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

Calculate Hessian

$$\begin{aligned}
& \left|\begin{array}{cc} \frac{\partial^2 l}{\partial \mu^2} & \frac{\partial^2 l}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 l}{\partial \mu \partial \sigma^2} & \frac{\partial^2 l}{\partial (\sigma^2)^2} \end{array}\right|_{\mu=\hat{\mu},\sigma^2=\hat{\sigma^2}} \\
= & \left|\begin{array}{ll} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu) \\ -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i-\mu)^2 \end{array}\right|_{\mu=\overline{x},\sigma^2=\hat{\sigma^2}} \\
= & \frac{1}{\sigma^6}\left[-\frac{n^2}{2} + \frac{n}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(x_i-\mu)\right)^2\right]\Bigg|_{\mu=\overline{x},\sigma^2=\hat{\sigma^2}} \\
= & \frac{1}{\hat{\sigma^6}}\left[-\frac{n^2}{2} + \frac{n}{\hat{\sigma^2}}(n\hat{\sigma^2}) - \frac{1}{\hat{\sigma^2}}\left(\sum_{i=1}^{n}(x_i-\overline{x})\right)^2\right] = \frac{1}{\hat{\sigma^6}}\frac{n^2}{2} > 0
\end{aligned}$$

Thus, the conditions for local (interior) maximum is indeed found. Because this is a unique interior maximum, it is also a global maximum. Therefore, $(\hat{\mu}, \hat{\sigma^2}) = (\overline{x}, \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2)$ is an MLE.

A simpler workaround

First, fix one parameter, say σ^2.

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

If

$$\mu \neq \overline{x}, \quad \text{then} \quad \sum_{i=1}^n (x_i - \mu)^2 > \sum_{i=1}^n (x_i - \overline{x})^2$$

so $\hat{\mu} = \overline{x}$ must hold to maximize the log-likelihood.

Second, reduce the problem into one-parameter maximization
Given $\hat{\mu} = \overline{x}$, the log-likelihood is maximized at $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2$, because

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^4}(\sigma^2 - \hat{\sigma}^2)$$

is always positive when $\sigma^2 < \hat{\sigma}^2$ and always negative when $\sigma^2 > \hat{\sigma}^2$. Hence l as a function of σ^2 increases upto $\hat{\sigma}^2$ and then decreases.

Therefore, $(\hat{\mu}, \hat{\sigma^2}) = (\overline{x}, \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2)$ is an MLE.

Example 2 – Ranged Normal with Known Variance

Let $X_1, \cdots, X_n$ *i.i.d.* $\mathcal{N}(\mu, 1)$ where $\underline{\mu \geq 0}$. Find MLE of μ.

Solution:

$$\begin{aligned} L(\mu|\mathbf{x}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(x_i - \mu)^2}{2}\right] = (2\pi)^{-n/2} \exp\left[-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2}\right] \\ l(\mu|\mathbf{x}) &= \log L(\mu, \mathbf{x}) = C - \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2} \\ \frac{\partial l}{\partial \mu} &= \frac{2\sum_{i=1}^{n}(x_i - \mu)}{2} = 0, \qquad \frac{\partial^2 l}{\partial \mu^2} < 0 \\ \hat{\mu} &= \sum_{i=1}^{n} x_i / n = \overline{x} \end{aligned}$$

Question: ARE WE DONE?

The MLE parameter must be within the parameter space.

We need to check whether $\hat{\mu}$ is within the parameter space $[0, \infty)$.

- If $\overline{x} \geq 0$, $\hat{\mu} = \overline{x}$ falls into the parameter space.
- If $\overline{x} < 0$, $\hat{\mu} = \overline{x}$ does NOT fall into the parameter space.

When $\overline{x} < 0$

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{n}(x_i - \mu) = n(\overline{x} - \mu) < 0$$

for $\mu \geq 0$. Therefore, $l(\mu|\mathbf{x})$ is a decreasing function of μ. So $\hat{\mu} = 0$ when $\overline{x} < 0$.

Therefore, MLE is

$$\hat{\mu}(\mathbf{X}) = \max(\overline{X}, 0)$$

Example 3 – Binomial MLE, unknown number of trials

Let $X_1, \cdots, X_n$ be random sample from $Binomial(k, p)$ population, where p is known and k is unknown. Find the MLE of k.

Likelihood Function

$$L(k|\mathbf{x}, p) = \begin{cases} \prod_{i=1}^n \binom{k}{x_i} p^{x_i}(1-p)^{k-x_i} & (k \geq \max_i x_i) \\ 0 & (k < \max_i x_i) \end{cases}$$

The likelihood function is not differentiable with respect to k because k is an integer.

So how can we find MLE?

Idea: Instead of differentiating, take a ratio

We want to find k such that

$$\frac{L(k|\mathbf{x}, p)}{L(k-1|\mathbf{x}, p)} \geq 1 \qquad \text{and} \qquad \frac{L(k+1|\mathbf{x}, p)}{L(k|\mathbf{x}, p)} < 1$$

$$\begin{aligned} \frac{L(k, \mathbf{x}, p)}{L(k-1, \mathbf{x}, p)} &= \frac{\prod_{i=1}^n \binom{k}{x_i} p^{x_i}(1-p)^{k-x_i}}{\prod_{i=1}^n \binom{k-1}{x_i} p^{x_i}(1-p)^{k-1-x_i}} \\ &= \frac{\prod_{i=1}^n \frac{k!}{x_i!(k-x_i)!} p^{x_i}(1-p)^{k-x_i}}{\prod_{i=1}^n \frac{(k-1)!}{x_i!(k-1-x_i)!} p^{x_i}(1-p)^{k-1-x_i}} \\ &= \prod_{i=1}^n \frac{k(1-p)}{k-x_i} = \frac{k^n(1-p)^n}{\prod_{i=1}^n (k-x_i)} \end{aligned}$$

Finding MLE

Find maximum k such that $\frac{L(k|\mathbf{x},p)}{L(k-1|\mathbf{x},p)} \geq 1$

$$\begin{aligned} k^n(1-p)^n &\geq \prod_{i=1}^{n}(k-x_i) \\ (1-p)^n &\geq \prod_{i=1}^{n}\left(1-\frac{x_i}{k}\right) \qquad (k \geq \max_i x_i) \qquad (1) \end{aligned}$$

- The right-hand side is an increasing function of k
- The right-hand side is $0 < (1-p)^n$, when $k = \max_i x_i$.
- The right-hand side will converge to $1 > (1-p)^n$ as $k \to \infty$.
- Thus, there is a unique maximum for the likelihood function.
- $\hat{k}_{MLE}$ can be numerically solved as the maximum k satisfying (1).

Example 4 Let $X_1, \cdots, X_n$ be a random sample from a pdf

$$f_X(x|\theta) = \theta x^{\theta-1}, \quad 0 < x < 1, \quad 0 < \theta < \infty.$$

(a) Find method of moments estimator for θ.

(b) Find the MLE of θ.

Example 5 – Two-parameter Exponential

Let $X_1, \cdots, X_n$ be *i.i.d.* observations from a location-scale family of an exponential distribution with pdf

$$f_X(x|\theta) = \frac{1}{\sigma}\exp\left(-\frac{x-\mu}{\sigma}\right), \qquad x \geq \mu, \sigma > 0$$

(a) Find MLEs of μ and σ.

(b) Find MLE of $S(t) = \Pr(X > t)$ for a fixed t.

Invariance

MLE is invariant under monotonic transfotmation.

Question: If $\hat{\theta}$ is the MLE of θ, what is the MLE of $\tau(\theta)$?

Example 6: Let $X_1, \cdots, X_n$ be a random sample from $Bernoulli(p)$ where $0 < p < 1$.

1. What is the MLE of p?
2. What is the MLE of odds, defined by $\eta = p/(1-p)$?

MLE of p

$$\begin{aligned} L(p|\mathbf{x}) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i} \\ l(p|\mathbf{x}) &= \log p \sum_{i=1}^{n} x_i + \log(1-p)(n - \sum_{i=1}^{n} x_i) \\ \frac{\partial l}{\partial p} &= \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} = 0 \\ \hat{p} &= \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x} \end{aligned}$$

MLE of $\eta = \frac{p}{1-p}$

- $\eta = p/(1-p) = \tau(p)$
- $p = \eta/(1+\eta) = \tau^{-1}(\eta)$

$$\begin{aligned}
L^*(\eta|\mathbf{x}) &= p^{\sum x_i}(1-p)^{n-\sum x_i} \\
&= \frac{p}{1-p}^{\sum x_i}(1-p)^n = \frac{\eta^{\sum x_i}}{(1+\eta)^n} \\
l^*(\eta|\mathbf{x}) &= \sum_{i=1}^{n} x_i \log\eta - n\log(1+\eta) \\
\frac{\partial l^*}{\partial\eta} &= \frac{\sum_{i=1}^n x_i}{\eta} - \frac{n}{1+\eta} = 0 \\
\hat{\eta} &= \frac{\sum_{i=1}^n x_i/n}{1-\sum_{i=1}^n x_i/n} = \frac{\overline{x}}{1-\overline{x}} = \tau(\hat{p})
\end{aligned}$$

Another way to get MLE of $\eta = \frac{p}{1-p}$

$$L^*(\eta|\mathbf{x}) = \frac{\eta^{\sum x_i}}{(1+\eta)^n}$$

- From MLE of $\hat{p}$, we know $L^*(\eta|\mathbf{x})$ is maximized when $p = \eta/(1+\eta) = \hat{p}$.
- Equivalently, $L^*(\eta|\mathbf{x})$ is maximized when $\eta = \hat{p}/(1-\hat{p}) = \tau(\hat{p})$, because τ is a one-to-one function.
- Therefore $\hat{\eta} = \tau(\hat{p})$.

Result: Denote the MLE of θ by $\hat{\theta}$. If $\tau(\theta)$ is an one-to-one function of θ, then MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

Proof: The likelihood function in terms of $\tau(\theta) = \eta$ is

$$\begin{aligned} L^*(\tau(\theta)|\mathbf{x}) &= \prod_{i=1}^{n} f_X(x_i|\theta) = \prod_{i=1}^{n} f(x_i|\tau^{-1}(\eta)) \\ &= L(\tau^{-1}(\eta)|\mathbf{x}) \end{aligned}$$

We know this function is maximized when $\tau^{-1}(\eta) = \hat{\theta}$, or equivalently, when $\eta = \tau(\hat{\theta})$. Therefore, MLE of $\eta = \tau(\theta)$ is $\hat{\eta} = \tau(\hat{\theta})$.

Induced Likelihood Function

- Let $L(\theta|\mathbf{x})$ be the likelihood function for a given data $x_1, \cdots, x_n$,
- and let $\eta = \tau(\theta)$ be a (possibly not a one-to-one) function of θ.

We define the *induced likelihood function* L^* by

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

where $\tau^{-1}(\eta) = \{\theta : \tau(\theta) = \eta,\ \theta \in \Omega\}$.

- The value of η that maximize $L^*(\eta|\mathbf{x})$ is called the MLE of $\eta = \tau(\theta)$.

Theorem 7.2.10: If θ is the MLE of $\hat{\theta}$, then the MLE of $\eta = \tau(\theta)$ is $\tau(\hat{\theta})$, where $\tau(\theta)$ is any function of θ.

Proof - Using Induced Likelihood Function

$$\begin{aligned} L^*(\hat{\eta}|\mathbf{x}) &= \sup_{\eta} L^*(\eta|\mathbf{x}) = \sup_{\eta} \sup_{\theta \in \tau^{-1}(\eta)} L(\theta|\mathbf{x}) \\ &= \sup_{\theta} L(\theta|\mathbf{x}) = L(\hat{\theta}|\mathbf{x}) \\ L(\hat{\theta}|\mathbf{x}) &= \sup_{\theta \in \tau^{-1}(\tau(\hat{\theta}))} L(\theta|\mathbf{x}) = L^*[\tau(\hat{\theta})|\mathbf{x}] \end{aligned}$$

Hence, $L^*(\hat{\eta}|\mathbf{x}) = L^*[\tau(\hat{\theta})|\mathbf{x}]$ and $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$.

Properties of MLE

1. Optimal in some sense : We will study this later
2. By definition, MLE will always fall into the range of the parameter space.
3. Not always easy to obtain; may be hard to find the global maximum.
4. Heavily depends on the underlying distributional assumptions (i.e. not robust).

Biostat 602 Winter 2017

Lecture Set 8

Point Estimation

Methods for Evaluating Estimators

Reading: CB 7.3.1–7.3.2

Methods for Evaluating Estimators

Bias

Definition: Suppose $\hat{\theta}$ is an estimator for θ, then the bias of θ is defined as

$$\text{Bias}(\theta) = \text{E}(\hat{\theta}) - \theta$$

If the bias is equal to 0, then $\hat{\theta}$ is an unbiased estimator for θ.

Example 1: Let $X_1, \cdots, X_n$ be iid samples from a distribution with mean μ. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be an estimator of μ. Its bias is

$$\begin{aligned}\text{Bias}(\mu) &= \text{E}(\overline{X}) - \mu \\ &= \text{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu = \frac{1}{n}\sum_{i=1}^{n} \text{E}\left(X_i\right) - \mu = \mu - \mu = 0\end{aligned}$$

Therefore $\overline{X}$ is an unbiased estimator for μ.

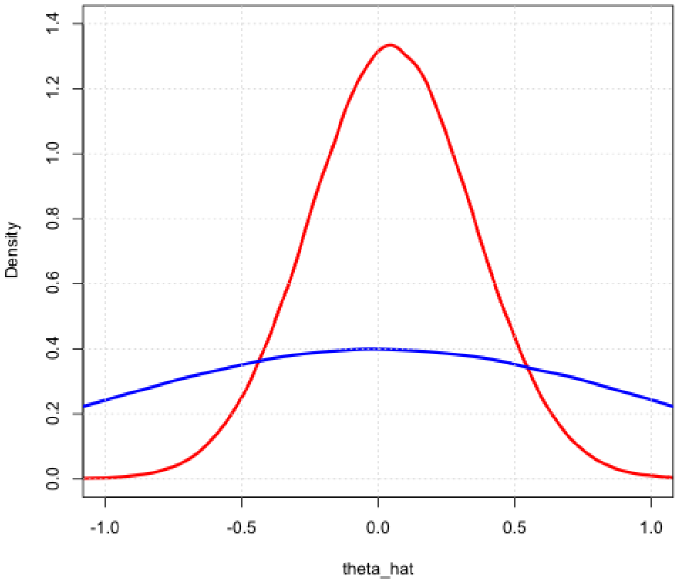
Example 2: Let $X_1, \cdots, X_n$ be iid samples from a distribution with mean μ and variance σ^2. Define

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \\ S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\end{aligned}$$

to be estimators of σ^2. Is either $\hat{\sigma}^2$ or S^2 an unbiased estimator? Which one?

How important is unbiasedness?

The Bias-Variance Trade-off



- $\hat{\theta}_1$ (blue) is unbiased but has a chance to be very far away from $\theta = 0$.
- $\hat{\theta}_2$ (red) is biased but more likely to be closer to the true θ than $\hat{\theta}_1$.

Mean Squared Error (MSE)

Definition: Mean Squared Error (MSE) of an estimator $\hat{\theta}$ is defined as

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}[(\hat{\theta} - \theta)]^2$$

Note that

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathrm{E}[\hat{\theta} - \mathrm{E}(\hat{\theta}) + \mathrm{E}(\hat{\theta}) - \theta]^2 \\
&= \mathrm{E}[\hat{\theta} - \mathrm{E}(\hat{\theta})^2] + \mathrm{E}[\mathrm{E}(\hat{\theta}) - \theta]^2 + 2\mathrm{E}[\hat{\theta} - \mathrm{E}(\hat{\theta})]\mathrm{E}[\mathrm{E}(\hat{\theta}) - \theta] \\
&= \mathrm{E}[\hat{\theta} - \mathrm{E}(\hat{\theta})^2] + [\mathrm{E}(\hat{\theta}) - \theta]^2 + 2[\mathrm{E}(\hat{\theta}) - \mathrm{E}(\hat{\theta})]\mathrm{E}[\mathrm{E}(\hat{\theta}) - \theta] \\
&= \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}^2(\theta)
\end{aligned}$$

MSE, as a measure of performance, combines both bias and variance. So looking for an estimator that minimizes MSE for all $\theta \in \Omega$ would tantamount to searching for one which is on target on an average without ever going too far away.

Question: Is it possible to find an estimator that uniformly minimizes the MSE?

Example 3: Let $X_1, \cdots, X_n$ be a random sample from $\mathcal{N}(\mu, 1)$ Define

$$\hat{\mu}_1 = 1, \quad \hat{\mu}_2 = \overline{X}.$$

$$\begin{aligned} \text{MSE}(\hat{\mu}_1) &= \text{E}(\hat{\mu}_1 - \mu)^2 = (1-\mu)^2 \\ \text{MSE}(\hat{\mu}_2) &= \text{E}(\overline{X} - \mu)^2 = \text{Var}(\overline{X}) = \frac{1}{n} \end{aligned}$$

- Suppose that the true $\mu = 1$, then $\text{MSE}(\mu_1) = 0 < \text{MSE}(\mu_2)$, and no estimator can beat μ_1 in terms of MSE when true $\mu = 1$.
- Therefore, we cannot find an estimator that is uniformly the best in terms of MSE across all $\theta \in \Omega$ among all estimators
- Restrict the class of estimators, and find the "best" estimator within the small class.

Example 4: Let $X_1, \cdots, X_n$ be an iid random sample from $\mathcal{N}(\mu, \sigma^2)$. Define

$$\begin{aligned} \hat{\sigma}^2_{MLE} &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \\ S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \end{aligned}$$

as estimators of σ^2. Which one has smaller MSE?

Solution: We shall use the following properties based on a i.i.d. random sample from a Normal Distribution (see pages 16-17 of Lecture Set 1; Theorem 5.3.1 C& B):

1. $E(S^2) = \sigma^2$.
2. $\overline{X}\text{and}S^2$ are independently distributed.
3. $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.
4. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

Example 5: Let $X_1, \cdots, X_n$ be iid $Poisson(\lambda)$. Let $\overline{X}$ and S^2 be the sample mean and variance, respectively. Since for Poisson distribution, mean = variance, both $\overline{X}$ and S^2 are unbiased for λ. Which estimator is better than the other (i.e. has smaller variance)?

Note that $Var(\overline{X}) = \lambda/n$, but $Var(S^2)$ is cumbersome and involves calculation of fourth moment.

Is there an alternative way to show which one is better?

We shall come back to this problem later.

Uniformly Minimum Variance Unbiased Estimator

Definition: $W^*(\mathbf{X})$ is the *best unbiased estimator*, or *uniformly minimum variance unbiased estimator (UMVUE)* of $\tau(\theta)$ if,

1. $\mathrm{E}[W^*(\mathbf{X})|\theta] = \tau(\theta)$ for all θ (unbiased)
2. and $\mathrm{Var}[W^*(\mathbf{X})|\theta] \leq \mathrm{Var}[W(\mathbf{X})|\theta]$ for all θ, where W is any other unbiased estimator of $\tau(\theta)$ (minimum variance).

First we develop some tools that will facilitate identification of best unbiased estimators. One of the key results towards that is an inequality called **Cramer-Rao (CR) Inequality** which we describe next.

Idea:

- CR inequality provides the lower bound of variances of any unbiased estimator of $\tau(\theta)$, say $B(\theta)$.
- If W^* is an unbiased estimator of $\tau(\theta)$ and satisfies $\mathrm{Var}[W^*(\mathbf{X})|\theta] = B(\theta)$, then W^* is the best unbiased estimator.

Some Terminology

Score and Fisher Information Number

Let $X_1, X_2, \ldots, X_n$ be random variables with joint pdf/pmf given by $f_{\mathbf{X}}(\mathbf{x}|\theta)$. The log-likelihood is defined by

$$l(\theta) = \log f_{\mathbf{X}}(\mathbf{x}|\theta).$$

Score Function: $\quad l'(\theta) = \frac{\partial}{\partial\theta} \log f_{\mathbf{X}}(\mathbf{x}|\theta)$

Fisher Information Number: $\quad I_n(\theta) = \mathrm{E}\left[\left\{\frac{\partial}{\partial\theta} \log f_{\mathbf{X}}(\mathbf{x}|\theta)\right\}^2\right]$

If $X_1, X_2, \ldots, X_n$ is a i.i.d. random sample from a well-behaved pdf/pmf (e.g. support of pdf/pmf does not depend on the parameter) then we have the following simplifications

$$\begin{aligned}
\mathrm{E}\left[l'(\theta)\right] &= \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f_X(x_i|\theta) = 0 \\
I_n(\theta) &= \mathrm{E}\left[\left\{\frac{\partial}{\partial\theta} \log f_{\mathbf{X}}(\mathbf{x}|\theta)\right\}^2\right] \qquad \leftarrow \text{based on all data} \\
&= n\mathrm{E}\left[\left\{\frac{\partial}{\partial\theta} \log f_X(x|\theta)\right\}^2\right] \\
&= nI(\theta) = n \times (\text{Information based on a single observation})
\end{aligned}$$

where $f_X(x|\theta)$ is the pdf/pmf based on a single observation.

Example 6: Let $X_1, X_2, \ldots, X_n$ be a random sample from $Bernoulli(p)$. Obtain the score function and the information number.

Example 7: Let $X_1, X_2, \ldots, X_n$ be a random sample from $Exponential(\theta)$. Obtain the score function and the information number.

Cramer-Rao inequality

Theorem 7.3.9: Let $X_1, \cdots, X_n$ be a collection of random variables with joint pdf/pmf of $f_{\mathbf{X}}(\mathbf{x}|\theta)$. Suppose $W(\mathbf{X})$ is an unbiased estimator of $\tau(\theta)$ with finite variance, i.e.

$$\mathrm{E}[W(\mathbf{X})|\theta] = \tau(\theta),\ \forall\theta \in \Omega, \quad \mathrm{Var}[W(\mathbf{X})|\theta] < \infty.$$

If

$$\frac{d}{d\theta}\mathrm{E}[\log f_{\mathbf{X}}(\mathbf{X}|\theta)] \;=\; \mathrm{E}\left(\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}(\mathbf{x}|\theta)\right) = 0$$

and

$$\frac{d}{d\theta}\mathrm{E}[W(\mathbf{X})|\theta] \;=\; \frac{d}{d\theta}\int_{x\in\mathcal{X}} W(\mathbf{x})f_{\mathbf{X}}(\mathbf{x}|\theta)d\mathbf{x} = \int_{x\in\mathcal{X}} W(\mathbf{x})\frac{\partial}{\partial\theta}f_{\mathbf{X}}(\mathbf{x}|\theta)d\mathbf{x}$$

Then, a lower bound of $\mathrm{Var}[W(\mathbf{X})|\theta]$ is

$$\mathrm{Var}[W(\mathbf{X})] \geq \frac{[\tau'(\theta)]^2}{\mathrm{E}\left[\{\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}(\mathbf{X}|\theta)\}^2\right]}$$

Corollary 7.3.10: If $X_1, \cdots, X_n$ are iid samples from pdf/pmf $f_X(x|\theta)$, and the assumptions in the above Cramer-Rao theorem hold, then the lower-bound of $\mathrm{Var}[W(\mathbf{X})]$ becomes

$$\mathrm{Var}[W(\mathbf{X})] \;\geq\; \frac{[\tau'(\theta)]^2}{n\mathrm{E}\left[\{\frac{\partial}{\partial\theta}\log f_X(X|\theta)\}^2\right]}$$

Remarks

1. The bound on the right of the inequality (called CRLB) is a uniform lower bound for the variance of all unbiased estimators of $\tau(\theta)$. Thus, if one finds an unbiased estimator whose variance satisfies CRLB, the search for best unbiased estimator is complete.

2. The proof of the CR inequality hinges on interchangeability of differentiation and integration. When we deal with pmf's for discrete distributions, integration is replaced by summation.

3. While the function f need not be differentiable with respect to $\mathbf{x}$ (e.g. pmf), it must be differentiable with respect to θ.

A Computational Tool

There is a computational simplification of Fisher Information number under mild regularity conditions.

Lemma 7.3.11: If $f_X(x|\theta)$ satisfies the two interchangeability conditions

$$\begin{aligned} \frac{d}{d\theta}\int_{x\in\mathcal{X}} f_X(x|\theta)dx &= \int_{x\in\mathcal{X}} \frac{\partial}{\partial\theta} f_X(x|\theta)dx \\ \frac{d}{d\theta}\int_{x\in\mathcal{X}} \frac{\partial}{\partial\theta} f_X(x|\theta)dx &= \int_{x\in\mathcal{X}} \frac{\partial^2}{\partial\theta^2} f_X(x|\theta)dx \end{aligned}$$

$\underline{\text{which are true for exponential family}}$, then

$$I(\theta) = \mathrm{E}\left[\left\{\frac{\partial}{\partial\theta}\log f_X(X|\theta)\right\}^2\right] = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f_X(X|\theta)\right]$$

Example 3 revisited: Let $X_1, \cdots, X_n$ be a iid random sample from Poisson(λ). Obtain the best unbiased estimator (if it exists) of λ. Define

$$\hat{\lambda}_1 = \overline{X}, \quad \hat{\lambda}_2 = s^2_{\mathbf{X}}.$$

Both $\hat{\lambda}_1,\ \hat{\lambda}_2$ are unbiased estimators of λ. In fact

$$\hat{\lambda}_a = a\overline{X} + (1-a)s^2_{\mathbf{X}}$$

is an unbiased estimator of λ for any $0 \le a \le 1$. Which one to choose? Since $\tau(\lambda) = \lambda$, the Cramer-Rao lower bound is $1/I_n(\lambda) = 1/\left[nI(\lambda)\right]$.

$$\begin{aligned} I(\lambda) &= \mathrm{E}\left[\left\{\frac{\partial}{\partial\lambda}\log f_X(X|\lambda)\right\}^2\right] = -\mathrm{E}\left[\frac{\partial^2}{\partial\lambda^2}\log f_X(X|\lambda)\right] \\ &= -\mathrm{E}\left[\frac{\partial^2}{\partial\lambda^2}\log\frac{e^{-\lambda}\lambda^X}{X!}\right] = -\mathrm{E}\left[\frac{\partial^2}{\partial\lambda^2}(-\lambda + X\log\lambda - \log X!)\right] \\ &= \mathrm{E}\left[\frac{X}{\lambda^2}\right] = \frac{1}{\lambda^2}\mathrm{E}(X) = \frac{1}{\lambda} \end{aligned}$$

Therefore, the Cramer-Rao lower bound is

$$\mathrm{Var}[W(\mathbf{X})] \ge \frac{1}{nI(\lambda)} = \frac{\lambda}{n}$$

where W is any unbiased estimator of λ.

$$\mathrm{Var}(\hat{\lambda}_1) \;=\; \mathrm{Var}(\overline{X}) = \frac{\mathrm{Var}(X)}{n} = \frac{\lambda}{n}$$

Therefore, $\hat{\lambda}_1 = \overline{X}$ is the best unbiased estimator of λ. With a lengthy calculation (need calculation of fourth moment), it is possible to show that

$$\mathrm{Var}(\hat{\lambda}_2) > \frac{\lambda}{n}$$

(details omitted), so $\hat{\lambda}_2$ is not the best unbiased estimator.

With and without Lemma 7.3.11

With Lemma 7.3.11

$$I(\lambda) = -\mathrm{E}\left[\tfrac{\partial^2}{\partial\lambda^2} \log f_X(X|\lambda)\right] = -\mathrm{E}\left[\tfrac{\partial^2}{\partial\lambda^2} \left(-\lambda + X\log\lambda - \log X!\right)\right] = \tfrac{1}{\lambda}$$

Without Lemma 7.3.11

$$\begin{aligned}
I(\lambda) &= \mathrm{E}\left[\left\{\frac{\partial}{\partial\lambda}\log f_X(X|\lambda)\right\}^2\right] \\
&= \mathrm{E}\left[\left\{\frac{\partial}{\partial\lambda}\left(-\lambda + X\log\lambda - \log X!\right)\right\}^2\right] \\
&= \mathrm{E}\left[\left\{-1 + \frac{X}{\lambda}\right\}^2\right] \\
&= \mathrm{E}\left[1 - 2\frac{X}{\lambda} + \frac{X^2}{\lambda^2}\right] \\
&= 1 - 2\frac{\mathrm{E}(X)}{\lambda} + \frac{\mathrm{E}(X^2)}{\lambda^2} \\
&= 1 - 2\frac{\mathrm{E}(X)}{\lambda} + \frac{\mathrm{Var}(X) + [\mathrm{E}(X)]^2}{\lambda^2} \\
&= 1 - 2\frac{\lambda}{\lambda} + \frac{\lambda + \lambda^2}{\lambda^2} = \frac{1}{\lambda}
\end{aligned}$$

Example 8: Let $X_1, \ldots, X_n$ be a random sample from $Bernoulli(p)$. Find the best unbiased estimator of p.

Example 9: Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, 1)$. Find the best unbiased estimator of μ.

Biostat 602 Winter 2017

Lecture Set 9

Point Estimation

Attainment of CRLB

Reading: CB 7.3.1–7.3.2

Attainment of CRLB

Question: How frequently can one find an unbiased estimator of $\tau(\theta)$ that attains Cramer Rao Lower bound?

Regularity condition for Cramer-Rao Theorem

$$\frac{d}{d\theta}\int_{x\in\mathcal{X}} h(\mathbf{x})f_{\mathbf{X}}(\mathbf{x}|\theta)d\mathbf{x} = \int_{x\in\mathcal{X}} h(\mathbf{x})\frac{\partial}{\partial\theta}f_{\mathbf{X}}(\mathbf{x}|\theta)d\mathbf{x}$$

for some function $h(x)$.

- This regularity condition holds for exponential family.
- How about non-exponential family?

Example 1: Let $X_1, \cdots, X_n$ be a random sample from $Uniform(0,\theta)$. Let us check the regularity condition. We use **Leibnitz's Rule** which states

$$\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} f(x|\theta)dx \;=\; f(b(\theta)|\theta)b'(\theta) - f(a(\theta)|\theta)a'(\theta) + \int_{a(\theta)}^{b(\theta)}\frac{\partial}{\partial\theta}f(x|\theta)dx$$

In our example of $Uniform(0,\theta)$

$$\begin{aligned}
f_X(x|\theta) &= 1/\theta \\
\frac{d}{d\theta}\int_0^\theta h(x)\left(\frac{1}{\theta}\right)dx &= \frac{h(\theta)}{\theta}\frac{d\theta}{d\theta} - h(0)f_X(0|\theta)\frac{d0}{d\theta} + \int_0^\theta \frac{\partial}{\partial\theta}h(x)\left(\frac{1}{\theta}\right)dx \\
&\neq \int_0^\theta \frac{\partial}{\partial\theta}h(x)\left(\frac{1}{\theta}\right)dx
\end{aligned}$$

Hence the interchangeability condition is not satisfied unless $h(\theta)/\theta = 0\ \forall\theta$. This raises the following questions

1. Is there an unbiased estimator of θ?
2. Does any unbiased estimator attain CRLB?
3. Is there a best estimator in the class of unbiased estimators?

Solution:

When is the Cramer-Rao Lower Bound Attainable?

It is possible that the value of Cramer-Rao bound may be strictly smaller than the variance of any unbiased estimator

Corollary 7.3.15: Let $X_1, \cdots, X_n$ be iid with pdf/pmf $f_X(x|\theta)$, where $f_X(x|\theta)$ satisfies the assumptions of the Cramer-Rao Theorem. Let

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

denote the likelihood function. If $W(\mathbf{X})$ is unbiased for $\tau(\theta)$, then $W(\mathbf{X})$ attains the Cramer-Rao lower bound if and only if

$$\frac{\partial}{\partial\theta} \log L(\theta|\mathbf{x}) = a(\theta)[W(\mathbf{x}) - \tau(\theta)]$$

for some function $a(\theta)$.

Proof:

Revisiting the Bernoulli Example

Example 2: Let $X_1, \cdots, X_n$ be i.i.d. $Bernoulli(p)$. Is $\overline{X}$ the best unbiased estimator of p? Does it attain the Cramer-Rao lower bound?

Method Using Corollary 7.3.15:

$$\begin{aligned} L(p|\mathbf{x}) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ \log L(p|\mathbf{x}) = \log \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} &= \sum_{i=1}^{n} \log[p^{x_i}(1-p)^{1-x_i}] \\ &= \sum_{i=1}^{n} [x_i \log p + (1-x_i)\log(1-p)] \\ &= \log p \sum_{i=1}^{n} x_i + \log(1-p)(n - \sum_{i=1}^{n} x_i) \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial p} \log L(p|\mathbf{x}) &= \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} \\ &= \frac{n\overline{x}}{p} - \frac{n(1-\overline{x})}{1-p} \\ &= \frac{(1-p)n\overline{x} - np(1-\overline{x})}{p(1-p)} \\ &= \frac{n(\overline{x}-p)}{p(1-p)} = \frac{n}{p(1-p)}(\overline{x}-p) \\ &= a(p)[W(\mathbf{x}) - \tau(p)] \end{aligned}$$

where $a(p) = \frac{n}{p(1-p)}$, $W(\mathbf{x}) = \overline{x}$, $\tau(p) = p$. Therefore, $\overline{X}$ is the best unbiased estimator for p and attains the Cramer-Rao lower bound.

Example 3: Let $X_1, \cdots, X_n$ be i.i.d. *Geometric*(p) with pmf

$$f_X(x|p) = (1-p)^{x-1}p, \quad 0 < p < 1, \; x = 1, 2, \ldots$$

Find a function $\tau(p)$ which admits an unbiased estimator that attains the CRLB.

Example 4: Let $X_1, \cdots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$. Consider estimating σ^2, assuming μ is known. Is Cramer-Rao bound attainable? What if μ is unknown?

Solution: Note that the information number equals

$$I(\sigma^2) = -\mathrm{E}\left[\frac{\partial^2}{\partial(\sigma^2)^2} \log f_X(x|\mu, \sigma)\right].$$

Now,

$$\begin{aligned}
f(x|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \\
\log f(x|\mu,\sigma^2) &= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2} \\
\frac{\partial}{\partial(\sigma^2)} \log f(x|\mu,\sigma^2) &= -\frac{1}{2}\frac{1}{\sigma^2} + \frac{(x-\mu)^2}{2(\sigma^2)^2} \\
\frac{\partial^2}{\partial(\sigma^2)^2} \log f(x|\mu,\sigma^2) &= \frac{1}{2}\frac{1}{(\sigma^2)^2} - \frac{2(x-\mu)^2}{2(\sigma^2)^3} \\
I(\sigma^2) &= -\mathrm{E}\left[\frac{1}{2\sigma^4} - \frac{2(x-\mu)^2}{2\sigma^6}\right] \\
&= -\frac{1}{2\sigma^4} + \frac{1}{\sigma^6}\mathrm{E}[(x-\mu)^2] = -\frac{1}{2\sigma^4} + \frac{1}{\sigma^6}\sigma^2 = \frac{1}{2\sigma^4}
\end{aligned}$$

Cramer-Rao lower bound is $\quad \dfrac{1}{nI(\sigma^2)} = \dfrac{2\sigma^4}{n}.$

The unbiased estimator of $\hat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2$, gives

$$\mathrm{Var}(\hat{\sigma^2}) = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n}$$

So, $\hat{\sigma^2}$ does not attain the Cramer-Rao lower-bound.

Is Cramer-Rao lower-bound for σ^2 attainable?

$$\begin{aligned}
L(\sigma^2|\mathbf{x}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] \\
\log L(\sigma^2|\mathbf{x}) &= -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} \\
\frac{\partial \log L(\sigma^2|\mathbf{x})}{\partial \sigma^2} &= -\frac{n}{2}\frac{2\pi}{2\pi\sigma^2} + \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2(\sigma^2)^2} \\
&= -\frac{n}{2\sigma^2} + \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^4} \\
&= \frac{n}{2\sigma^4}\left(\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n} - \sigma^2\right) \\
&= a(\sigma^2)(W(\mathbf{x}) - \sigma^2)
\end{aligned}$$

Therefore,

1. If μ is known, the best unbiased estimator for σ^2 is $\sum_{i=1}^{n}(x_i-\mu)^2/n$, and it attains the Cramer-Rao lower bound, i.e.

$$\mathrm{Var}\left[\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{n}\right] = \frac{2\sigma^4}{n}$$

2. If μ is not known, the Cramer-Rao lower-bound cannot be attained.

At this point, we do not know if $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$ is the best unbiased estimator for σ^2 or not.

Result for Exponential Family

Let $X_1, \cdots, X_n$ be iid from the one parameter exponential family with pdf/pmf

$$f_X(x|\theta) = c(\theta)h(x)\exp\left[w(\theta)t(x)\right].$$

Assume that $\mathrm{E}[t(X)] = \tau(\theta)$. Then $\frac{1}{n}\sum_{i=1}^{n} t(x_i)$, which is an unbiased estimator of $\tau(\theta)$, attains the Cramer-Rao lower-bound. That is,

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} t(X_i)\right) = \frac{[\tau'(\theta)]^2}{I_n(\theta)}$$

Proof:

$$\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} t(X_i)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[t(X_i)\right] = \frac{1}{n}\sum_{i=1}^{n}\tau(\theta) = \tau(\theta)$$

So, $\frac{1}{n}\sum_{i=1}^{n} t(x_i)$ is an unbiased estimator of $\tau(\theta)$.

$$\begin{aligned}\log L(\theta|\mathbf{x}) &= \sum_{i=1}^{n}\log f_X(x_i|\theta) \\ &= \sum_{i=1}^{n}\left[\log c(\theta) + \log h(x) + w(\theta)t(x_i)\right]\end{aligned}$$

$$\begin{aligned}\frac{\partial \log L(\theta|\mathbf{x})}{\partial\theta} &= \sum_{i=1}^{n}\left[\frac{c'(\theta)}{c(\theta)} + 0 + w'(\theta)t(x_i)\right] \\ &= nw'(\theta)\left[\frac{1}{n}\sum_{i=1}^{n} t(x_i) - \left\{-\frac{c'(\theta)}{c(\theta)w'(\theta)}\right\}\right] \qquad (1)\end{aligned}$$

Because $\mathrm{E}\left[\frac{\partial}{\partial\theta}\log L(\theta|\mathbf{x})\right] = 0$, from (1), we have $\tau(\theta) = -\frac{c'(\theta)}{c(\theta)w'(\theta)}$.

Hence we have,

$$\frac{\partial \log L(\theta|\mathbf{x})}{\partial\theta} = nw'(\theta)\left[\frac{1}{n}\sum_{i=1}^{n} t(x_i) - \tau(\theta)\right]$$

Thus $\frac{1}{n}\sum_{i=1}^{n} t(x_i)$ attains the CRLB and is the best unbiased estimator of $\tau(\theta)$.

Remarks

- For exponential families, CRLB approach establishes $\frac{1}{n}\sum_{i=1}^{n} t(x_i)$ to be the best estimator for its expectation $\tau(\theta)$. If the parameter of interest is non-trivially different from $\tau(\theta)$ then CRLB cannot be used to obtain the best estimator.

- For non-exponential family, it is unlikely that Cramer-Rao Theorem can help finding the best unbiased estimator, but the bound still can be calculated to determine a loose lower bound of the variance of the best unbiased estimator (provided the regularity conditions hold).

Biostat 602 Winter 2017

Lecture Set 9

Point Estimation

Rao Blackwell Theorem, Lehman-Scheffe Theorem

Reading: CB 7.3.3

Important Facts

Let X and Y be two random variables. Then

- $\text{E}(X) = \text{E}[\text{E}(X|Y)]$ (Theorem 4.4.3)
- $\text{Var}(X) = \text{E}[\text{Var}(X|Y)] + \text{Var}[\text{E}(X|Y)]$ (Theorem 4.4.7)
- $\text{E}[g(X)|Y] = \int_{x \in \mathcal{X}} g(x) f(x|Y) dx$ is a function of Y.
- If X and Y are independent, $\text{E}[g(X)|Y] = \text{E}[g(X)]$.

Searching for a better unbiased estimator

Suppose $W(\mathbf{X})$ is an unbiased estimator of $\tau(\theta)$, i.e. $\text{E}[W(\mathbf{X})] = \tau(\theta)$. Further suppose $T(\mathbf{X})$ is any function of $\mathbf{X} = (X_1, \cdots, X_n)$.

Consider $\phi(T) = \text{E}[W(\mathbf{X})|T]$.

$$\begin{aligned}
\text{E}[\phi(T)] &= \text{E}[\text{E}(W(\mathbf{X})|T)] = \text{E}[W(\mathbf{X})] = \tau(\theta) \qquad \text{(unbiased for } \tau(\theta)\text{)} \\
\text{Var}(\phi(T)) &= \text{Var}[\text{E}(W|T)] \\
&= \text{Var}(W) - \text{E}[\text{Var}(W|T)] \\
&\leq \text{Var}(W) \qquad \text{(equal or smaller variance than } W\text{)}
\end{aligned}$$

Does this mean that $\phi(T)$ is a better estimator than $W(\mathbf{X})$?

1. If $\phi(T)$ is an estimator, then $\phi(T)$ is equal or better than $W(\mathbf{X})$.
2. $\phi(T) = \mathrm{E}[W|T] = \mathrm{E}[W|T, \theta]$.

$\phi(T)$ may depend on θ, which means that $\phi(T)$ may not be an estimator.

A Note about the notation $\mathbf{E}(\cdot)$, $\mathbf{E}_\theta(\cdot)$, and $\mathbf{E}(\cdot|\theta)$

To be explict that $\mathrm{E}[W|T]$ depends on θ, sometimes it is represented as $\mathrm{E}_\theta[W|T]$ as in the textbook. Note that most of $\mathrm{E}(\cdot)$ or $Var(\cdot)$ in this lecture note can be represented as $\mathrm{E}_\theta(\cdot) = \mathrm{E}(\cdot|\theta)$ or $Var_\theta(\cdot) = Var(\cdot|\theta)$

Example 1: Let $X_1, \cdots, X_n$ be an i.i.d. random sample from $\mathcal{N}(\theta, 1)$. Then $W(\mathbf{X}) = \frac{1}{2}(X_1 + X_2)$ is an unbiased estimator of θ. Consider conditioning it on $T(\mathbf{X}) = X_1$. Then

$$\begin{aligned}
\phi(T) &= \mathrm{E}[W|T] = \mathrm{E}\left[\frac{1}{2}(X_1 + X_2)|X_1\right] \\
&= \frac{1}{2}\mathrm{E}(X_1|X_1) + \frac{1}{2}\mathrm{E}(X_2|X_1) \\
&= \frac{1}{2}X_1 + \frac{1}{2}\mathrm{E}(X_2) \\
&= \frac{1}{2}X_1 + \frac{1}{2}\theta
\end{aligned}$$

- $\mathrm{E}[\phi(T)] = \frac{1}{2}\theta + \frac{1}{2}\theta = \theta$ (unbiased)
- $\mathrm{Var}[\phi(T)] = \frac{1}{4} < \mathrm{Var}(\frac{1}{2}(X_1 + X_2)) = \frac{1}{2}$
- But $\phi(T)$ is NOT an estimator.

Example 2: Consider again a random sample $X_1, \cdots, X_n$ from $\mathcal{N}(\theta, 1)$. Then $W(\mathbf{X}) = X_1$ is an unbiased estimator of θ. Consider conditioning it on $\overline{X}$.

$$\begin{aligned}
\phi(T) &= \mathrm{E}[W|T] = \mathrm{E}(X_1|\overline{X}) \\
&= \frac{\mathrm{E}(X_1|\overline{X}) + \mathrm{E}(X_2|\overline{X}) + \cdots + \mathrm{E}(X_n|\overline{X})}{n} \\
&= \frac{\mathrm{E}(X_1 + \cdots + X_n|\overline{X})}{n} \\
&= \frac{\mathrm{E}(n\overline{X}|\overline{X})}{n} = \frac{n\overline{X}}{n} = \overline{X}
\end{aligned}$$

- $\mathrm{E}[\phi(T)] = \theta$ (unbiased)
- $\mathrm{Var}[\phi(T)] = \frac{\mathrm{Var}(X)}{n} = \frac{1}{n} < \mathrm{Var}(W) = 1$
- $\phi(T)$ is an estimator, thus a better unbiased estimator.

Question: Why is $\mathrm{E}(X_1|\overline{X})$ free of θ?

Rao-Blackwell Theorem

Theorem 7.3.17: Let $W(\mathbf{X})$ be any unbiased estimator of $\tau(\theta)$, and T be a sufficient statistic for θ. Define $\phi(T) = \mathrm{E}[W|T]$. Then the following hold.

1. $\mathrm{E}[\phi(T)|\theta] = \tau(\theta)$
2. $\mathrm{Var}[\phi(T)|\theta] \leq \mathrm{Var}(W|\theta)$ for all θ.

That is, $\phi(T)$ is a uniformly better unbiased estimator of $\tau(\theta)$.

Proof: Using the properties of $E(\cdot)$ and $Var(\cdot)$,

1. $\mathrm{E}[\phi(T)] = \mathrm{E}[\mathrm{E}(W|T)] = \mathrm{E}(W) = \tau(\theta)$ (unbiased)
2. $\mathrm{Var}[\phi(T)] = \mathrm{Var}[\mathrm{E}(W|T)] = \mathrm{Var}(W) - \mathrm{E}[\mathrm{Var}(W|T)] \leq \mathrm{Var}(W)$ (better than W).
3. Need to show $\phi(T)$ is indeed an estimator.

$$\begin{aligned} \phi(T) &= \mathrm{E}(W|T) = \mathrm{E}[W(\mathbf{X})|T] \\ &= \int_{\mathbf{x}\in\mathcal{X}} W(\mathbf{x})f(\mathbf{x}|T)d\mathbf{x} \end{aligned}$$

Because T is a sufficient statistic, $f(\mathbf{x}|T)$ does not depend on θ.

Therefore, $\phi(T) = \int_{\mathbf{x}\in\mathcal{X}} W(\mathbf{x})f(\mathbf{x}|T)d\mathbf{x}$ does not depend on θ, and $\phi(T)$ is indeed an estimator of θ.

Uniqueness

Theorem 7.3.19 (Uniqueness of UMVUE) If W is a best unbiased estimator of $\tau(\theta)$, then W is unique.

Proof: Suppose W_1 and W_2 are two best unbiased estimators of $\tau(\theta)$. Since both are 'best', $Var(W_1) = Var(W_2)$. Consider estimator $W_3 = \frac{1}{2}(W_1 + W_2)$.

$$\begin{aligned}
\mathrm{E}(W_3) &= \mathrm{E}\left(\frac{1}{2}W_1 + \frac{1}{2}W_2\right) = \frac{1}{2}\tau(\theta) + \frac{1}{2}\tau(\theta) = \tau(\theta) \\
\mathrm{Var}(W_3) &= \mathrm{Var}\left(\tfrac{1}{2}W_1 + \tfrac{1}{2}W_2\right) \\
&= \frac{1}{4}\mathrm{Var}(W_1) + \frac{1}{4}\mathrm{Var}(W_2) + \frac{1}{2}\mathrm{Cov}(W_1, W_2) \\
&\leq \frac{1}{4}\mathrm{Var}(W_1) + \frac{1}{4}\mathrm{Var}(W_2) + \frac{1}{2}\sqrt{\mathrm{Var}(W_1)\mathrm{Var}(W_2)} \\
&= \mathrm{Var}(W_1) = \mathrm{Var}(W_2)
\end{aligned}$$

If strict inequality holds, W_3 is better than W_1 and W_2, which is contradictory to the assumption.

Therefore, the equality must hold, requiring

$$\frac{1}{2}\mathrm{Cov}(W_1, W_2) = \frac{1}{2}\sqrt{\mathrm{Var}(W_1)\mathrm{Var}(W_2)}$$

By Cauchy-Schwarz (correlation) inequality, this is true if and only if $W_2 = aW_1 + b$.

Since both W_1, W_2 are unbiased estimators for $\tau(\theta)$

$$\mathrm{E}(W_1) \;=\; \tau(\theta) = \mathrm{E}(W_2) = \mathrm{E}(aW_1 + b) = a\tau(\theta) + b$$

and so $a = 1, b = 0$ must hold, yielding $W_2 = W_1$. Therefore, the best unbiased estimator is unique.

Unbiased estimator of zero

Definition

If $U(\mathbf{X})$ satisfies $\mathrm{E}(U) = 0$ for all $\theta \in \Omega$, then we call U an unbiased estimator of 0.

Relationship with ancillary statistics

Let $S(\mathbf{X})$ is an ancillary statistic for θ. $U(\mathbf{X}) = S(\mathbf{X}) - \mathrm{E}[S(\mathbf{X})]$ is always an unbiased estimator of zero. However, in general, an unbiased estimator of zero simply requires the expectation to be zero and need not be ancillary.

Theorem 7.3.20: If $\mathrm{E}[W(\mathbf{X})] = \tau(\theta)$, W is the best unbiased estimator of $\tau(\theta)$ if an only if W is uncorrelated with all unbiased estimator of 0.

Does the Theorem share some similarity to Basu's Theorem?

Proof: Let W be an unbiased estimator of $\tau(\theta)$. Let $V = W + U$ and $U \in \mathcal{U}$, which is the class of unbiased estimators of 0.

By construction, V is an unbiased estimator of $\tau(\theta)$. Consider

$$\mathcal{V} \;=\; \{V_a = W + aU\}$$

where a is a constant.

$$\begin{aligned}
\mathrm{E}(V_a) &= \mathrm{E}(W + aU) = \mathrm{E}(W) + a\mathrm{E}(U) \\
&= \tau(\theta) + a \cdot 0 = \tau(\theta) \\
\mathrm{Var}(V_a) &= \mathrm{Var}(W + aU) \\
&= a^2\mathrm{Var}(U) + 2a\mathrm{Cov}(W, U) + \mathrm{Var}(W)
\end{aligned}$$

The variance is minimized when

$$a = \frac{-2\mathrm{Cov}(W, U)}{2\mathrm{Var}(U)} = -\frac{\mathrm{Cov}(W, U)}{\mathrm{Var}(U)}$$

The best unbiased estimator in this class is

$$W - \frac{\mathrm{Cov}(W, U)}{\mathrm{Var}(U)} U$$

W is the best unbiased estimator in this class if and only if $\mathrm{Cov}(W, U) = 0$.

Therefore W is the best among all unbiased estimators of $\tau(\theta)$ if and only if $\mathrm{Cov}(W, U) = 0$ for every $U \in \mathcal{U}$.

Example 3: Let X be an observation from a Uniform$(\theta, \theta+1)$ distribution.

1. Is $X - \frac{1}{2}$ an unbiased estimator of θ?
2. Find an example of an unbiased estimator of zero.
3. Is $X - \frac{1}{2}$ the best unbiased estimator of θ?

Bias of $X - \frac{1}{2}$

$$\mathrm{E}X = \int_{\theta}^{\theta+1} x dx = \theta + \frac{1}{2}$$

so $X - \frac{1}{2}$ is an unbiased estimator of θ.

Further note that $Var(X - \frac{1}{2}) = VarX = \frac{1}{12}$.

Finding unbiased estimators of zero

If $U(X)$ is an unbiased estimator of zero, then it has to satisfy

$$\int_{\theta}^{\theta+1} U(x)dx = 0, \qquad \forall\theta \in \mathbb{R}$$

then

$$\frac{d}{d\theta}\int_{\theta}^{\theta+1} U(x)dx = U(\theta+1) - U(\theta) = 0, \qquad \forall\theta \in \mathbb{R}$$

So a periodic function with frequency 1 qualifies for $U(X)$. For example,

$$U(X) = \sin(2\pi X)$$

is an unbiased estimator for zero.

Is $X - \frac{1}{2}$ the best unbiased estimator?

If we can identify an unbiased estimator of zero, say $U(X)$ such that $Cov(X - \frac{1}{2}, U(X)) = 0$, then $X - \frac{1}{2}$ is not the best estimator.

Define $U(X) = \sin(2\pi X)$,

$$\begin{aligned}
\operatorname{Cov}\left(X - \frac{1}{2}, \sin(2\pi X)\right) &= \operatorname{Cov}\left(X, \sin(2\pi X)\right) \\
&= \int_{\theta}^{\theta+1} x \sin(2\pi x) dx \\
&= \left[-\frac{x\cos(2\pi x)}{2\pi}\right]_{\theta}^{\theta+1} + \int_{\theta}^{\theta+1} \frac{\cos(2\pi x)}{2\pi} dx \\
&= -\frac{\cos(2\pi\theta)}{2\pi}
\end{aligned}$$

Hence, $X - \frac{1}{2}$ is correlated with an unbiased estimator of zero, and cannot be the best unbiased estimator of θ.

In fact, $Var\left(X - \frac{1}{2} + \frac{1}{2}\sin(2\pi X)\right) = 0.071 < \frac{1}{12}$. This provides an example of an unbiased estimator for θ which has smaller variance than $X - \frac{1}{2}$.

Lehmann-Scheffé Theorem

In searching for best unbiased estimators, we explored two approaches

CRLB

CRLB provides a loose lower bound for the variance of the unbiased estimators. But its use is limited due to the fact that this bound is not attained too frequently even if there are best estimators.

Rao-Blackwell Theorem

Rao Blackwell Theorem is useful but calculating $\phi(T) = \mathrm{E}[W|T]$ is not usually an easy task. Further, $\phi(T)$ is a 'better' unbiased estimator, but may not be the 'best'.

Question: Is there a way to easily obtain the 'best' unbiased estimator?

Theorem 7.3.23 - Lehmann-Scheffé: Let T be a complete sufficient statistic for parameter θ. Let $\phi(T)$ be any estimator based on T. Then $\phi(T)$ is the unique best unbiased estimator of its expected value.

Another version of Lehmann-Scheffé: Let T be a complete sufficient statistic for parameter θ, Then $\phi(T) = \mathrm{E}[W|T]$ is the unique best unbiased estimator of $\mathrm{E}(W)$.

Proof: Let $\phi(T)$ be any function of T such that $\mathrm{E}[\phi(T)] = \tau(\theta)$. Let W be any unbiased estimator of $\tau(\theta)$, i.e. $\mathrm{E}(W) = \tau(\theta)$.

1. Because T is sufficient, $\psi(T) = \mathrm{E}[W|T]$ is a legitimate estimator. Further it is unbiased for $\tau(\theta)$ since

$$\mathrm{E}[\psi(T)] = \mathrm{E}\left(\mathrm{E}[W|T]\right) = \mathrm{E}(W) = \tau(\theta).$$

2. By Rao-Blackwell Theorem, $Var[\psi(T)] \leq Var(W)$ for all θ.

3. Let $g(T) = \phi(T) - \psi(T)$, then because both $\phi(T)$ and $\psi(T)$ are unbiased estimators for $\tau(\theta)$,

   $$\mathrm{E}[g(T)|\theta] \;=\; \mathrm{E}[\phi(T) - \psi(T)|\theta] = 0$$

   Because T is a complete statistic, the above equation always implies

   $$\Pr(g(T) = 0|\theta) \;=\; \Pr(\phi(T) = \psi(T)|\theta) = 1$$

   for all θ, meaning that $\psi(T)$ and $\phi(T)$ are identical in distribution.

4. Because for any unbiased estimator of W,

   $$Var(W) \geq Var[\psi(T)] = Var[\phi(T)],$$

   so $\phi(T)$ is always the unique UMVUE for its expected value $\tau(\theta)$.

Note: When T is complete and sufficient, the Lehmann-Scheffé Theorem implies that there is at most one function of T thats unbiased for $\tau(\theta)$.

Application of Lehman-Scheffé

1. Find a complete sufficient statistic T.
2. If we can find an unbiased estimator $V(T)$, we have found the UMVUE.
3. Otherwise, find any unbiased estimator $W(X)$ and then compute $\phi(T) = \mathrm{E}[W|T]$.

Remarks

- From Rao-Blackwell Theorem, we can always improve an unbiased estimator by conditioning it on a sufficient statistics.
  - $W(\mathbf{X})$: unbiased for $\tau(\theta)$.
  - $T^*(\mathbf{X})$: sufficient statistic for θ.

  $\phi(T) = \mathrm{E}[W(\mathbf{X})|T^*(\mathbf{X})]$ is a better unbiased estimator of $\tau(\theta)$.

- Minimal sufficient statistics are more useful. In fact, we only need to consider functions of minimal sufficient statistics to find the best unbiased estimator.

  Let $T(\mathbf{X})$ be a minimal sufficient, and $T^*(\mathbf{X})$ be a sufficient statistic. Then by definition, there exists a function h that satisfies $T = h(T^*)$.

  $$\begin{aligned} \mathrm{E}[\phi(T)|T^*] &= \mathrm{E}\left[\phi\left\{h(T^*)\right\}|T^*\right] \\ &= \phi\left\{h(T^*)\right\} = \phi(T) \end{aligned}$$

  Therefore $\phi(T)$ remains the same after conditioning on any sufficient statistic T^*.

- Complete sufficient statistics is a very useful ingredient to obtain a UMVUE.

  - We need to limit our search to the class of minimal sufficient statistics T to find the best unbiased estimator.
  - Let $\phi(T)$ be unbiased for $\mathrm{E}[\phi(T)] = \tau(\theta)$.
  - Consider $\{\phi(T) + U(T) |\ U(T) \in \mathcal{U}\}$, where $\mathcal{U}$ is unbiased estimators of zero among the functions of T.
  - By Theorem 7.3.20, $\phi(T)$ is UMVUE if and only if

    $$\begin{aligned} \mathrm{Cov}(\phi(T), U(T)) &= \mathrm{E}[\phi(T)U(T)] - \mathrm{E}[\phi(T)]\mathrm{E}[U(T)] \\ &= \mathrm{E}[\phi(T)U(T)] = 0 \end{aligned}$$

  - If T is complete, $\phi(T)U(T) = 0$ almost surely, requiring $U(T) = 0$. Therefore, $\phi(T)$ is the best unbiased estimator of its expected value.

Example 4: Let $X_1, \cdots, X_n$ be i.i.d. from $\mathcal{N}(\mu, \sigma^2)$. Find the best unbiased estimator for (1) μ, (2) σ^2, (3) μ^2.

Solution:

- First, we need to find a complete and sufficient statistic for (μ, σ^2).
- We know that $\mathbf{T}(\mathbf{X}) = (\overline{X}, s^2_{\mathbf{X}})$ is complete, sufficient statistic for (μ, σ^2).
- Because $\mathrm{E}[\overline{X}] = \mu$, $\overline{X}$ is an unbiased estimator for μ, $\overline{X}$ is also a function of $\mathbf{T}(\mathbf{X})$.
- Therefore, $\overline{X}$ is the best unbiased estimator for μ.
- $\mathrm{E}(s^2_{\mathbf{X}}) = \sigma^2$
- $s^2_{\mathbf{X}}$ is a function of $\mathbf{T}$
- Therefore $s^2_{\mathbf{X}}$ is the best unbiased estimator of σ^2.

To obtain UMVUE for μ^2, we need a $\phi(\mathbf{T}) = \phi(\overline{X}, s^2_{\mathbf{X}})$ such that $\mathrm{E}[\phi(\mathbf{T})] = \mu^2$.

$$\begin{aligned}
\mathrm{E}(\overline{X}) &= \mu \\
\mathrm{E}((\overline{X})^2) &= \mathrm{Var}(\overline{X}) + \mathrm{E}[(\overline{X})]^2 = \frac{\sigma^2}{n} + \mu^2 \\
\mathrm{E}\left(\overline{X}^2 - \frac{\sigma^2}{n}\right) &= \mu^2 \\
\mathrm{E}\left(\overline{X}^2 - \frac{s^2_{\mathbf{X}}}{n}\right) &= \mu^2
\end{aligned}$$

- $\overline{X}^2 - s^2_{\mathbf{X}}/n$ is unbiased estimator for μ^2
- And it is a function of $(\overline{X}, s^2_{\mathbf{X}})$.
- Hence, $\overline{X}^2 - s^2_{\mathbf{X}}/n$ is the best unbiased estimator for μ^2.

Example 5: Let $X_1, \cdots, X_n$ be i.i.d. from $Bernoulli(p)$. Find the best unbiased estimator for (1) $p(1-p)$, (2) p^2.

Example 6: Let $X_1, \cdots, X_n$ be i.i.d. from $Poisson(\lambda)$. Find the best unbiased estimator for $\Pr(X_1 = 0) = \exp(-\lambda)$.

Biostat 602 Winter 2017

Lecture Set 11

Best Unbiased Estimation

Unbiasedness

- If there are at least two unbiased estimators, there are infinitely many. If T_1, T_2 are unbiased estimators of $\tau(\theta)$, then so is

$$\omega T_1 + (1-\omega)T_2$$

  for any $0 \leq \omega \leq 1$.

- If there is a best unbiased estimator, then it is unique.

Strategies for finding best unbiased estimator

Cramer-Rao Lower Bound

1. Calculate joint score function

$$u_n(\theta|\mathbf{x}) = \frac{\partial}{\partial\theta} \log L(\theta|\mathbf{x}).$$

2. Express u_n if possible as

$$u_n(\theta|\mathbf{x}) = a(\theta)[W(\mathbf{x}) - \tau(\theta)].$$

   Then $W(\mathbf{x})$ is the best unbiased estimator for $\tau(\theta)$ and attains CRLB.

- If "regularity conditions" are satisfied, then we have a Cramer-Rao bound for unbiased estimators of $\tau(\theta)$.
  - It helps to confirm an estimator is the best unbiased estimator of $\tau(\theta)$ if it happens to attain the CR-bound.
  - If an unbiased estimator of $\tau(\theta)$ has variance greater than the CR-bound, it does NOT mean that it is not the best unbiased estimator.

- When "regularity conditions" are not satisfied, $\frac{[\tau'(\theta)]^2}{I_n(\theta)}$ is no longer a valid lower bound.
  - There may be unbiased estimators of $\tau(\theta)$ that have variance smaller than $\frac{[\tau'(\theta)]^2}{I_n(\theta)}$.

Lehmann-Scheffé

Use complete sufficient statistic to find the best unbiased estimator for $\tau(\theta)$.

1. Find complete sufficient statistic T for θ.
2. Obtain $\phi(T)$, an unbiased estimator of $\tau(\theta)$ using either of the following two ways
   - Guess a function $\phi(T)$ such that $\mathrm{E}[\phi(T)] = \tau(\theta)$.
   - Guess an unbiased estimator $h(\mathbf{X})$ of $\tau(\theta)$. Construct $\phi(T) = \mathrm{E}[h(\mathbf{X})|T]$, then $\mathrm{E}[\phi(T)] = \mathrm{E}[h(\mathbf{X})] = \tau(\theta)$.

   In either case, $\phi(T)$ is the best unbiased estimator of $\tau(\theta)$.

Example 1: Let $X_1, \ldots, X_n$ be *i.i.d.* observations from the distribution with pdf

$$f_X(x|\theta) = e^{-(x-\theta)} \exp(-e^{-(x-\theta)}), \qquad -\infty < \theta < \infty, \ -\infty < x < \infty$$

(a) Find a Cramer-Rao lower bound to the variance of unbiased estimators of θ.

(b) Find a function $\tau(\theta)$ for which there exists an unbiased estimator whose variance attains the Cramer-Rao bound.

(c) Find the best unbiased estimator for $\tau(\theta)$ found in (b).

Solution: (a) The distribution belongs to an exponential family with $c(\theta) = e^{\theta}$, $h(x) = e^{-x}$, $w(\theta) = -e^{\theta}$, $t(x) = e^{-x}$. The Fisher information per observation can be calculated as

$$
\begin{aligned}
\log L(\theta|x) &= -(x-\theta) - \exp(-x+\theta) \\
u(\theta|x) &= \frac{\partial}{\partial\theta}\log L(\theta|x) = 1 - \exp(-x+\theta) \\
I(\theta) &= -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log L(\theta|X)\right] \\
&= -\mathrm{E}\left[\frac{\partial}{\partial\theta}u(\theta|X)\right] = \mathrm{E}\left[\exp(-X+\theta)\right] = 1
\end{aligned}
$$

The Cramer-Rao lower bound of θ is $\frac{1}{nI(\theta)} = \frac{1}{n}$.

(b, c) The score function of joint likelihood is

$$
\begin{aligned}
\log L(\theta|\mathbf{x}) &= -\left(\sum x_i - n\theta\right) - \exp\left(-\sum x_i + n\theta\right) \\
u_n(\theta|\mathbf{x}) &= \frac{\partial}{\partial\theta}\log L(\theta|\mathbf{x}) = n - n\exp\left(-\sum x_i + n\theta\right) \\
&= -n\exp(n\theta)\left[\exp\left(-\sum x_i\right) - \exp\left(-n\theta\right)\right] \\
&= a(\theta)[W(\mathbf{x}) - \tau(\theta)]
\end{aligned}
$$

so $\tau(\theta) = \exp\left(-n\theta\right)$ and $W(\mathbf{x}) = \exp\left(-\sum x_i\right)$ is its best unbiased estimator whose variance attains CRLB.

Example 2: Let $X_1, \cdots, X_n$ be i.i.d. Uniform$(0, \theta)$. Find the best unbiased estimator for (1) θ, (2) θ^2, (3) $1/\theta$.

Solution - UMVUE of θ: $T(\mathbf{X}) = X_{(n)}$ is a complete and sufficient statistic for θ.

- $f_T(t) = n\theta^{-n}t^{n-1}I(0 < t < \theta)$.
- $\mathrm{E}[T] = \mathrm{E}[X_{(n)}] = \int_0^\theta tn\theta^{-n}t^{n-1}dt = \frac{n}{n+1}\theta$ (biased)
- $\mathrm{E}[\phi(T)] = \mathrm{E}\left[\frac{n+1}{n}X_{(n)}\right] = \theta$.

$\frac{n+1}{n}X_{(n)}$ is the best unbiased estimator of θ.

Estimating θ^2

$$\begin{aligned} \mathrm{E}[X_{(n)}^2] &= \int_0^\theta t^2 n\theta^{-n}t^{n-1}dt \\ &= n\theta^{-n}\int_0^\theta t^{n+1}dt = n\theta^{-n} \times \frac{\theta^{n+2}}{n+2} = \frac{n}{n+2}\theta^2 \end{aligned}$$

So $\phi(X_{(n)}) = \frac{n+2}{n}X_{(n)}^2$ is the best unbiased estimator for θ^2.

Estimating $1/\theta$

$$\begin{aligned} \mathrm{E}[X_{(n)}^{-1}] &= \int_0^\theta t^{-1} n\theta^{-n}t^{n-1}dt \\ &= n\theta^{-n}\int_0^\theta t^{n-2}dt = n\theta^{-n} \times \frac{\theta^{n-1}}{n-1} = \frac{n}{n-1}\theta^{-1} \end{aligned}$$

So $\phi(X_{(n)}) = \frac{n-1}{n}X_{(n)}^{-1}$ is the best unbiased estimator for θ^{-1}.

Example 3: Let $X_1, \cdots, X_n$ i.i.d. Binomial(k, θ). Find the best unbiased estimator of the probability of exactly one success from a Binomial(k, θ).

Solution: The quantity we need to estimate is

$$\tau(\theta) = \Pr(X = 1|\theta) = k\theta(1-\theta)^{k-1}$$

We know that $T(\mathbf{X}) = \sum_{i=1}^{n} X_i \sim$ Binomial(kn, θ) and it is a complete sufficient statistic. So we need to find a $\phi(T)$ that satisfies $\mathrm{E}[\phi(T)] = \tau(\theta)$.

There is no immediately evident unbiased estimator of $\tau(\theta)$ as a function of T. Start with a simple-minded estimator

$$W(\mathbf{X}) = \begin{cases} 1 & X_1 = 1 \\ 0 & \text{otherwise} \end{cases}$$

The expectation of W is

$$\begin{aligned} \mathrm{E}[W] &= \sum_{x_1=0}^{k} W(x_1) \binom{k}{x_1} \theta^{x_1} (1-\theta)^{k-x_1} \\ &= k\theta(1-\theta)^{k-1} \end{aligned}$$

and hence it is an unbiased estimator of $\tau(\theta) = k\theta(1-\theta)^{k-1}$. The best unbiased estimator of $\tau(\theta)$ is

$$\phi(T) = \mathrm{E}[W|T] = \mathrm{E}\left[W(\mathbf{X}) \middle| T(\mathbf{X})\right]$$

$$\begin{aligned}
\phi(t) &= \mathrm{E}\left[W(\mathbf{X})\middle|\sum_{i=1}^{n} X_i = t\right] = \Pr\left[X_1 = 1\middle|\sum_{i=1}^{n} X_i = t\right] \\
&= \frac{\Pr(X_1 = 1, \sum_{i=1}^{n} X_i = t)}{\Pr(\sum_{i=1}^{n} X_i = t)} \\
&= \frac{\Pr(X_1 = 1, \sum_{i=2}^{n} X_i = t-1)}{\Pr(\sum_{i=1}^{n} X_i = t)} \\
&= \frac{\Pr(X_1 = 1)\Pr(\sum_{i=2}^{n} X_i = t-1)}{\Pr(\sum_{i=1}^{n} X_i = t)} \\
&= \frac{\left[k\theta(1-\theta)^{k-1}\right]\left[\binom{k(n-1)}{t-1}\theta^{t-1}(1-\theta)^{k(n-1)-t-1}\right]}{\binom{kn}{n}\theta^{t}(1-\theta)^{kn-t}} = k\frac{\binom{k(n-1)}{t-1}}{\binom{kn}{t}}
\end{aligned}$$

Therefore, the unbiased estimator of $k\theta(1-\theta)^{k-1}$ is

$$\phi\left(\sum_{i=1}^{n} X_i\right) = k\frac{\binom{k(n-1)}{\sum X_i - 1}}{\binom{kn}{\sum X_i}}$$

Example 4: Let X_1, X_2 be *i.i.d.* observations from the pdf

$$f_X(x|\lambda) = \lambda e^{-\lambda x}, \qquad x \geq 0,\ \lambda > 0.$$

1. Show that the distribution of X_1 conditional on $Z = z$ is Uniform$(0, z)$.
2. Prove the best unbiased estimators of $\Pr(X_1 > 1) = e^{-\lambda}$ is

$$T(X_1, X_2) = \begin{cases} 0, & \text{if } X_1 + X_2 \leq 1 \\ \\ \frac{X_1+X_2-1}{X_1+X_2}, & \text{if } X_1 + X_2 > 1 \end{cases}$$

Solution: Note that X_1, X_2 are i.i.d. $Exp(1/\lambda)$ and so the $Z = X_1 + X_2$ is distributed as a $Gamma(2, 1/\lambda)$ random variable with pdf

$$f_Z(z|\lambda) = \lambda^2 z e^{-\lambda z}, \qquad z > 0,\ \lambda > 0.$$

The conditional pdf of $X_1|Z = z$ is

$$f(x_1|z, \lambda) \;=\; \frac{f(x_1, z|\lambda)}{f_Z(z|\lambda)} = \frac{\lambda^2 e^{-\lambda z}}{\lambda^2 z e^{-\lambda z}} = \frac{1}{z}$$

when $0 < x_1 < z$. If $x_1 > z$, the pdf is zero. Therefore, the conditional pdf of X_1 given z is Uniform$(0, z)$.

1. A naive unbiased estimator of $\Pr(X_1 > 1)$ is $W = I(X_1 > 1)$.
2. We know that $Z = X_1 + X_2$ is a complete sufficient statistic.
3. By Theorem 7.3.23, the best unbiased estimator of $\Pr(X_1 > 1)$ can be obtained by conditional expectation $\mathrm{E}[W|Z]$.

Because $\Pr(X|Z)$ is uniformly distributed between 0 and Z,

$$\mathrm{E}[W|Z] \;=\; \Pr(X_1 > 1|Z) = \begin{cases} 0 & \text{if } Z \leq 1 \\ 1 - \frac{1}{Z} = \frac{Z-1}{Z} & \text{if } Z > 1 \end{cases}$$

Therefore $E[W|X_1 + X_2] = T(X_1, X_2)$ is the best unbiased estimator of $\Pr(X_1 > 1) = e^{-\lambda}$.

Biostat 602 Winter 2017

Lecture Set 12

Bayesian Estimation

Reading: CB 7.2.3

Frequentist Statistics

Ingredients for a Frequentist Framework

Random Variable $\mathbf{X} = (X_1, \cdots, X_n)$

Data $\mathbf{x} = (x_1, \cdots, x_n)$

Model $\mathcal{P} = \{f_{\mathbf{X}}(\mathbf{x}|\theta) : \theta \in \Omega\}$

Parameter $\theta \in \Omega$

Statistical Inference in a Frequentist Framework

Given $\mathcal{P} = \{f_{\mathbf{X}}(\mathbf{x}|\theta) : \theta \in \Omega\}$

Known $\mathbf{x} = (x_1, \cdots, x_n)$, generated from $f_{\mathbf{X}}(\mathbf{x}|\theta)$.

Unknown $\theta \in \Omega$.

There are no other assumptions of θ, which is an unknown, but a fixed value. Consequently, no probabilistic statement can be attached to the plausible values of θ.

Cancer Screening Example in a Frequentist Framework

Problem: Let $X \in \{0, 1\}$ be a random variable indicating whether an individual has a positive ($X = 1$) or negative ($X = 0$) outcome from a screening test for a particular cancer type. Let $\theta \in \{0, 1\}$ be a variable indicating whether the individual have the cancer ($\theta = 1$) or ($\theta = 0$) not at the time of screening. The distribution of X for each possible θ is given in the following table.

	$\Pr(X = 0\vert\theta)$	$\Pr(X = 1\vert\theta)$
$\theta = 0$	0.99	0.01
$\theta = 1$	0.05	0.95

- $\Pr(X = 1|\theta = 1)$ is called the **sensitivity** of the screening test.
- $\Pr(X = 0|\theta = 0)$ is called the **specificity** of the screening test.

Statistical Inference

1. Find the maximum likelihood estimator of θ.
2. What are the bias and MSE of the MLE?

Rephrasing the question

- If the individual does not have the cancer, there is 1% of chance of positive screening results.
- If the individual have the cancer, there is 95% of chance of positive screening results.
- Given X, what are the bias and MSE of MLE of θ?

Solution:

	$X = 0$	$X = 1$
$L(\theta = 0 \| X)$	0.99	0.01
$L(\theta = 1 \| X)$	0.05	0.95
$\hat{\theta}_{MLE}$	0	1

If the individual's screening result was positive, the MLE estimates that (s)he has the cancer. If the screening result was negative, the MLE estimates that (s)he does not have the cancer.

Bias of MLE

$$\begin{aligned}
\mathrm{E}\hat{\theta} &= \Pr(\hat{\theta} = 0) \cdot 0 + \Pr(\hat{\theta} = 1) \cdot 1 \\
&= 0.01\ I(\theta = 0) + 0.95\ I(\theta = 1) \\
&= (1 - \theta) \times 0.01 + \theta\ \ \times 0.95 = 0.94\ \theta + 0.01 \\
\mathrm{Bias}(\hat{\theta}) &= \mathrm{E}(\hat{\theta} - \theta) = 0.01 - 0.06\ \theta
\end{aligned}$$

MSE of MLE

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathrm{E}[(\hat{\theta} - \theta)^2] = \Pr(\hat{\theta} \neq \theta) \\
&= 0.05\ I(\theta = 1) + 0.01\ I(\theta = 0) = 0.05\ \theta + 0.01(1 - \theta) \\
&= 0.04\ \theta + 0.01
\end{aligned}$$

The MSE is 0.01 if the individual does not have the cancer ($\theta = 0$). If (s)he has the cancer ($\theta = 1$), MSE is 0.05.

If you're a patient with positive screening results, you may want to ask

- Do I have cancer or not?
- What is the chance that I have cancer now?

Possible Answers

Frequentist I do not know. You're asking a wrong question. Whether you have the cancer or not is not a random variable. It is a fixed value. Therefore, the phrase "chance that you have cancer now" does not make sense.

Bayesian I think you have . . . % chance of having the cancer (How?)

Bayesian Framework

The main distinction from the frequentist framework lies in the fact that

- Parameter θ is considered as a random quantity
- Distribution of θ can be described by probability distribution, referred to as *prior* distribution
- A sample is taken from a population indexed by θ, and the prior distribution is updated using information from the sample to get *posterior* distribution of θ given the sample.

Ingredients

- Prior distribution of θ : $\theta \sim \pi(\theta)$.
- Sample distribution of $\mathbf{X}$ given θ.

$$\mathbf{X}|\theta \sim f(\mathbf{x}|\theta)$$

- Joint distribution $\mathbf{X}$ and θ

$$f(\mathbf{x}, \theta) = \pi(\theta) f(\mathbf{x}|\theta)$$

- Marginal distribution of $\mathbf{X}$

$$m(\mathbf{x}) \;=\; \int_{\theta\in\Omega} f(\mathbf{x}, \theta) d\theta = \int_{\theta\in\Omega} f(\mathbf{x}|\theta)\pi(\theta) d\theta$$

- Posterior distribution of θ (conditional distribution of θ given $\mathbf{X}$)

$$\pi(\theta|\mathbf{x}) \;=\; \frac{f(\mathbf{x}, \theta)}{m(\mathbf{x})} = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})} \qquad \text{(Bayes' Rule)}$$

- All of the above have discrete counterparts in which integration is replaced by summation.

Inference Under Bayesian Framework

Leveraging Prior Information

Suppose that we know that the chance of the (rare) caner per individual in is 10^{-4} (**Prevalence**).

$$\Pr(\theta = 1|X = 1) = \Pr(X = 1|\theta = 1)\frac{\Pr(\theta = 1)}{\Pr(X = 1)} \qquad \text{(Bayes' rule)}$$

$$= \Pr(X = 1|\theta = 1)\frac{\Pr(\theta = 1)}{\Pr(\theta = 1, X = 1) + \Pr(\theta = 0, X = 1)}$$

$$= \frac{\Pr(X = 1|\theta = 1)\Pr(\theta = 1)}{\Pr(X = 1|\theta = 1)\Pr(\theta = 1) + \Pr(X = 1|\theta = 0)\Pr(\theta = 0)}$$

$$= \frac{0.95 \times 10^{-4}}{0.95 \times 10^{-4} + 0.01 \times (1 - 10^{-4})} \approx 0.0094$$

So, even if the screening results were positive, one can conclude that the patient have less than 1% of chance to have the cancer.

Sensitivity: $\Pr(X = 1|\theta = 1)$

Specificity: $\Pr(X = 0|\theta = 0)$

Positive Predictive Value: $\Pr(\theta = 1|X = 1)$

Negative Predictive Value: $\Pr(\theta = 0|X = 0)$

Predictive values are of interest to the patient. Predictive values are affected by the prevalence of the disease. Sensitivity and specificity are more intrinsic to the screening test.

Question: What if the prior information is misleading?

Suppose that, in fact, the cancer is highly heritable, and the patient has a parent died with the same cancer type. It is known that the chance that a child of an affected parent will also have the center is $\Pr(\theta = 1) = 0.1$.

$$
\begin{aligned}
& \Pr(\theta = 1 | X = 1) \\
= & \frac{\Pr(X = 1|\theta = 1) \Pr(\theta = 1)}{\Pr(X = 1|\theta = 1) \Pr(\theta = 1) + \Pr(X = 1|\theta = 0) \Pr(\theta = 0)} \\
= & \frac{0.95 \times 0.1}{0.95 \times 0.1 + 0.01 \times (1 - 0.1)} \approx 0.913
\end{aligned}
$$

Even though the patient has 91.3% chance of having cancer, if (s)he did not know that her/his biological parent died of the same type of cancer, (s)he may end up concluding that there are $> 99\%$ chance that this was a false alarm, and doing nothing to get a proper treatment in the early stage.

Advantages and Drawbacks of Bayesian Inference

Advantages over frequentist framework

- Allows making inference on the distribution of θ given data.
- Available information from prior experiment about θ can be utilized.
- Uncertainty of θ can be formally quantified.

Drawbacks of Bayesian Inference

- Bayesian inference is quite sensitive to prior choice. Misleading prior can result in misleading inference.
- Choice of prior distribution can be argued to be highly "subjective".
- Bayesian inference could be sometimes quite complex, requiring high dimensional integration.

Bayes Estimator

Bayes Estimator of θ is defined as the posterior mean of θ.

$$E(\theta|\mathbf{x}) = \int_{\theta\in\Omega} \theta\pi(\theta|\mathbf{x})d\theta$$

Example 1: Let $X_1, \cdots, X_n$ be i.i.d. Bernoulli(p) where $0 \leq p \leq 1$. Assume that the prior distribution of p is Beta(α, β). Find the posterior distribution of p and the Bayes estimator of p, assuming α and β are known.

Solution: Prior distribution of p is

$$\pi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

Sampling distribution of $\mathbf{X}$ given p is

$$f_{\mathbf{X}}(\mathbf{x}|p) = \prod_{i=1}^{n} \left\{p^{x_i}(1-p)^{1-x_i}\right\}$$

Joint distribution of $\mathbf{X}$ and p is

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}, p) &= f_{\mathbf{X}}(\mathbf{x}|p)\pi(p) \\ &= \prod_{i=1}^{n} \left\{p^{x_i}(1-p)^{1-x_i}\right\} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \end{aligned}$$

So the marginal distribution of $\mathbf{X}$ is

$$\begin{aligned}
m(\mathbf{x}) &= \int f(\mathbf{x}, p)dp = \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\sum_{i=1}^n x_i+\alpha-1}(1-p)^{n-\sum_{i=1}^n x_i+\beta-1}dp \\
&= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\sum x_i+\alpha)\Gamma(n-\sum x_i+\beta)}{\Gamma(\alpha+\beta+n)} \\
&\quad \times \frac{\Gamma(\sum x_i+\alpha+n-\sum x_i+\beta)}{\Gamma(\sum x_i+\alpha)\Gamma(n-\sum x_i+\beta)} p^{\sum x_i+\alpha-1}(1-p)^{n-\sum x_i+\beta-1}dp \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\sum_{i=1}^n x_i+\alpha)\Gamma(n-\sum_{i=1}^n x_i+\beta)}{\Gamma(\alpha+\beta+n)} \\
&\quad \times \int_0^1 f_{\text{Beta}(\sum x_i+\alpha, n-\sum x_i+\beta)}(p)dp \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\sum_{i=1}^n x_i+\alpha)\Gamma(n-\sum_{i=1}^n x_i+\beta)}{\Gamma(\alpha+\beta+n)}
\end{aligned}$$

The posterior distribution of $p|\mathbf{x}$:

$$\begin{aligned}
\pi(p|\mathbf{x}) &= \frac{f(\mathbf{x}, p)}{m(\mathbf{x})} \\
&= \frac{\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\sum x_i+\alpha-1}(1-p)^{n-\sum x_i+\beta-1}\right]}{\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\sum x_i+\alpha)\Gamma(n-\sum x_i+\beta)}{\Gamma(\alpha+\beta+n)}\right]} \\
&= \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\sum x_i+\alpha)\Gamma(n-\sum x_i+\beta)} p^{\sum x_i+\alpha-1}(1-p)^{n-\sum x_i+\beta-1} \\
&\sim \text{Beta}\left(\sum x_i+\alpha,\ n-\sum x_i+\beta\right)
\end{aligned}$$

The Bayes estimator of p is

$$\begin{aligned}
\hat{p} &= \frac{\sum_{i=1}^{n} x_i + \alpha}{\sum_{i=1}^{n} x_i + \alpha + n - \sum_{i=1}^{n} x_i + \beta} = \frac{\sum_{i=1}^{n} x_i + \alpha}{\alpha + \beta + n} \\
&= \frac{\sum_{i=1}^{n} x_i}{n} \times \frac{n}{\alpha + \beta + n} + \frac{\alpha}{\alpha + \beta} \times \frac{\alpha + \beta}{\alpha + \beta + n} \\
&= [\text{Guess about } p \text{ from data}] \cdot \text{weight}_1 \\
&\quad + [\text{Guess about } p \text{ from prior}] \cdot \text{weight}_2
\end{aligned}$$

Thus the Bayes estimator is a weighted average of the prior mean and sample mean (MLE) of p. As n increases, weight$_1 = \frac{n}{\alpha+\beta+n} = \frac{1}{\frac{\alpha+\beta}{n}+1}$ becomes bigger and bigger and approaches to 1. In other words, influence of data is increasing, and the influence of prior knowledge is decreasing.

Question: Is the Bayes estimator unbiased?

$$E\left[\frac{\sum_{i=1}^{n} X_i + \alpha}{\alpha + \beta + n}\right] = \frac{np + \alpha}{\alpha + \beta + n} \neq p$$

$$\text{Bias} = \frac{np + \alpha}{\alpha + \beta + n} - p = \frac{\alpha - (\alpha + \beta)p}{\alpha + \beta + n}$$

As n increases, the bias approaches to zero.

Sufficient statistic and posterior distribution

If $T(\mathbf{X})$ is a sufficient statistic, then the posterior distribution of θ given $\mathbf{X}$ is the same as the posterior distribution of θ given $T(\mathbf{X})$. In other words,

$$\pi(\theta|\mathbf{x}) \;=\; \pi(\theta|T(\mathbf{x}))$$

Proof: The result follows since

$$\begin{aligned}
\pi(\theta|\mathbf{x}) &= \frac{f(\mathbf{x},\theta)}{m(\mathbf{x})} \\
&= \frac{f(\mathbf{x},\theta)}{\int_{\theta\in\Omega} f(\mathbf{x},\theta)d\theta} \\
&= \frac{f(\mathbf{x},T(\mathbf{x}),\theta)}{\int_{\theta\in\Omega} f(\mathbf{x},T(\mathbf{x}),\theta)d\theta} \\
&= \frac{f(\mathbf{x}|T(\mathbf{x}),\theta)f(T(\mathbf{x})|\theta)\pi(\theta)}{\int_{\theta\in\Omega} f(\mathbf{x}|T(\mathbf{x}),\theta)f(T(\mathbf{x})|\theta)\pi(\theta)d\theta} \\
&= \frac{f(\mathbf{x}|T(\mathbf{x}))f(T(\mathbf{x})|\theta)\pi(\theta)}{f(\mathbf{x}|T(\mathbf{x}))\int_{\theta\in\Omega} f(T(\mathbf{x})|\theta)\pi(\theta)d\theta} \\
&= \frac{f(T(\mathbf{x})|\theta)\pi(\theta)}{m(T(\mathbf{x}))} \\
&= \pi(\theta|T(\mathbf{x})
\end{aligned}$$

Conjugate Family

Definition 7.2.15: Let $\mathcal{F}$ denote the class of pdfs or pmfs for $f(x|\theta)$. A class Π of prior distributions is a conjugate family of $\mathcal{F}$, if the posterior distribution is in the class Π for all $f \in \mathcal{F}$, and all priors in Π, and all $x \in \mathcal{X}$.

Example 1 (revisited):

Let $X_1, \cdots, X_n|p \sim \text{Bernoulli}(p)$, $\pi(p) \sim \text{Beta}(\alpha, \beta)$ where α, β are known.

The posterior distribution is

$$\pi(p|\mathbf{x}) \sim \text{Beta}\left(\sum_{i=1}^{n} x_i + \alpha,\ n - \sum_{i=1}^{n} x_i + \beta\right)$$

Example 2: Beta-Binomial conjugate

Let $X_1, \cdots, X_n|p \sim \text{Binomial}(m, p)$, $\pi(p) \sim \text{Beta}(\alpha, \beta)$ where m, α, β are known.

Then, following the steps of Example 1, the posterior distribution is shown to be

$$\pi(p|\mathbf{x}) \sim \text{Beta}\left(\sum_{i=1}^{n} x_i + \alpha,\ mn - \sum_{i=1}^{n} x_i + \beta\right)$$

Example 3: (Gamma-Poisson) Let $X_1, \cdots, X_n|\lambda \sim \text{Poisson}(\lambda)$, and let $\pi(\lambda) \sim \text{Gamma}(\alpha, \beta)$. Find Bayes estimator of λ.

Solution:

Prior:

$$\pi(\lambda) = \frac{1}{\Gamma(\alpha)\beta^\alpha}\lambda^{\alpha-1}e^{-\lambda/\beta}$$

Sampling distribution

$$\begin{aligned} \mathbf{X}|\lambda \quad i.i.d. \quad & \frac{e^{-\lambda}\lambda^x}{x!} \\ f_{\mathbf{X}}(\mathbf{x}|\lambda) \quad = \quad & \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \end{aligned}$$

Joint distribution of X and λ

$$\begin{aligned} f(\mathbf{x}|\lambda)\pi(\lambda) &= \left[\prod_{i=1}^n \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right]\frac{1}{\Gamma(\alpha)\beta^\alpha}\lambda^{\alpha-1}e^{-\lambda/\beta} \\ &= e^{-n\lambda-\lambda/\beta}\lambda^{\sum x_i+\alpha-1}\frac{1}{\prod_{i=1}^n x_i!}\frac{1}{\Gamma(\alpha)\beta^\alpha} \end{aligned}$$

Marginal distribution

$$m(\mathbf{x}) = \int f(\mathbf{x}|\lambda)\pi(\lambda)d\lambda = \frac{\Gamma(\sum x_i + \alpha)\left(\frac{1}{n+\frac{1}{\beta}}\right)^{\sum x_i+\alpha}}{\prod_{i=1}^n x_i!\Gamma(\alpha)\beta^\alpha}$$

Posterior distribution

$$
\begin{aligned}
\pi(\lambda|\mathbf{x}) &= \frac{f(\mathbf{x}|\lambda)\pi(\lambda)}{m(\mathbf{x})} \\
&= e^{-n\lambda-\lambda/\beta}\lambda^{\sum x_i+\alpha-1}\frac{1}{\Gamma(\sum x_i+\alpha)\left(\frac{1}{n+\frac{1}{\beta}}\right)^{\sum x_i+\alpha}}
\end{aligned}
$$

So, the posterior distribution is Gamma $\left(\sum x_i+\alpha, \left(n+\frac{1}{\beta}\right)^{-1}\right)$.

Bayes Estimator:

$$
\hat{\lambda}_B = \frac{\sum x_i+\alpha}{n+\frac{1}{\beta}} = \frac{n}{n+\frac{1}{\beta}}\cdot\frac{\sum x_i}{n} + \frac{1/\beta}{n+\frac{1}{\beta}}\cdot(\alpha\beta)
$$

Question: Is it necessary to calculate the marginal distribution?

- Marginal is a normalization constant and its evaluation is typically not needed to identify the posterior distribution if the posterior belongs to a standard distribution family.
- Posterior distribution is "proportional" to the joint distribution. Trick is to write the joint as proportional to the product of sampling and prior distribution omitting constants not depending on the parameter.
- Then identify the structure to a known family of distributions.

Example: Beta-Binomial

$$
\pi(p|\mathbf{x}) \propto p^{\sum_{i=1}^n x_i}(1-p)^{mn-\sum_{i=1}^n x_i} \times p^{\alpha-1}(1-p)^{\beta-1} = p^{\sum_{i=1}^n x_i+\alpha-1}(1-p)^{mn-\sum_{i=1}^n x_i+\beta-1}
$$

The structure matches that of a Beta distribution. Hence

$$
\pi(p|\mathbf{x}) \sim \text{Beta}\left(\sum_{i=1}^n x_i+\alpha,\ mn-\sum_{i=1}^n x_i+\beta\right)
$$

Example: Poisson-Gamma

Biostat 602 Winter 2017

Lecture Set 13

Loss Function

Reading: CB 7.3.4

Bayesian Inference – Recap

- Allows making inference on the distribution of θ given data.
- Available information (from prior experiments) about θ can be utilized.
- Uncertainty of θ can be formally quantified.
- Misleading prior can result in misleading inference.
- Bayesian inference (especially the prior formulation) can be highly "subjective".
- Bayesian inference can be computationally intensive.

Ingredients

- **Prior** of θ : $\theta \sim \pi(\theta)$.
- **Sampling distribution** of $\mathbf{X}$ given θ.

$$\mathbf{X}|\theta \sim f(\mathbf{x}|\theta)$$

- Marginal distribution of $\mathbf{X}$

$$m(\mathbf{x}) = \int_{\theta\in\Omega} f(\mathbf{x},\theta)d\theta = \int_{\theta\in\Omega} f(\mathbf{x}|\theta)\pi(\theta)d\theta$$

- Bayesian inference is based on **Posterior distribution** of θ (conditional distribution of θ given $\mathbf{X}$)

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x},\theta)}{m(\mathbf{x})} = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})} \qquad \text{(Bayes' Rule)}$$

Bayes Estimator

Bayes Estimator of θ is defined as the posterior mean of θ.

$$E(\theta|\mathbf{x}) = \int_{\theta \in \Omega} \theta \pi(\theta|\mathbf{x}) d\theta$$

We shall generalize this definition in this Lecture Set, but this is the most commonly accepted definition of Bayes estimator.

Conjugate Family

Definition 7.2.15: Let $\mathcal{F}$ denote the class of pdfs or pmfs for $f(x|\theta)$. A class Π of prior distributions is a conjugate family of $\mathcal{F}$, if the posterior distribution is in the class Π for all $f \in \mathcal{F}$, and all priors in Π, and all $x \in \mathcal{X}$.

Example 1: Normal Bayes Estimators Let $X \sim \mathcal{N}(\theta, \sigma^2)$ and suppose that the prior distribution of θ is $\mathcal{N}(\mu, \tau^2)$. Assuming that σ^2, μ^2, τ^2 are all known, it follows, that

$$\begin{aligned} \pi(\theta) &= \frac{1}{\sqrt{2\pi\tau^2}} \exp\left[-\frac{(\theta-\mu)^2}{2\tau^2}\right] \\ f(x|\theta) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\theta)^2}{2\sigma^2}\right] \end{aligned}$$

$$\begin{aligned}
\pi(\theta|x) &\propto \pi(\theta)f(x|\theta) \\
&\propto \exp\left[-\frac{(\theta-\mu)^2}{2\tau^2}-\frac{(x-\theta)^2}{2\sigma^2}\right] \\
&= \exp\left[-\frac{\sigma^2(\theta-\mu)^2+\tau^2(x-\theta)^2}{2\tau^2\sigma^2}\right] \\
&= \exp\left[-\frac{(\sigma^2+\tau^2)\theta^2-2(\sigma^2\mu+\tau^2x)\theta+\sigma^2\mu^2+\tau^2x^2}{2\tau^2\sigma^2}\right] \\
&= \\
&\propto
\end{aligned}$$

So $\theta|x$ also becomes normal, with mean and variance given by

$$\begin{aligned}
\mathrm{E}[\theta|x] &= \frac{\tau^2}{\sigma^2+\tau^2}x+\frac{\sigma^2}{\sigma^2+\tau^2}\mu \\
\mathrm{Var}(\theta|x) &= \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}
\end{aligned}$$

- The normal family is its own conjugate family.
- The Bayes estimator for θ is a weighted average of the prior and sample means.
- As the prior variance τ^2 approaches to infinity (prior information becomes more vague), the Bayes estimator tends towards sample mean.

Loss/Risk Function

A **Loss Function** associated with point estimation is a real-valued non-negative function of the estimate and estimator, that is typically an increasing function of the distance between the two.

Let $\hat{\theta}$ be an estimator of θ and let $L(\hat{\theta}, \theta)$ be a function of θ and $\hat{\theta}$. Following are some examples of loss functions.

Squared error loss

$$L(\hat{\theta}, \theta) \;=\; (\hat{\theta} - \theta)^2$$

Weighted squared error loss

$$L(\hat{\theta}, \theta) \;=\; \omega(\theta)(\hat{\theta} - \theta)^2$$

where $\omega(\theta) \geq 0$ is a weight function.

Absolute error loss

$$L(\hat{\theta}, \theta) \;=\; |\hat{\theta} - \theta|$$

Asymmetric loss function

$$L(\theta, \hat{\theta}) \;=\; (\hat{\theta} - \theta)^2 I(\hat{\theta} < \theta) + 10(\hat{\theta} - \theta)^2 I(\hat{\theta} \geq \theta)$$

A loss that penalties overestimation more than underestimation

Relative squared error loss

$$L(\theta, \hat{\theta}) = \frac{(\hat{\theta} - \theta)^2}{|\theta| + 1}$$

This is a special case of weighted squared error loss. This loss penalizes errors in estimation more if θ is near 0 than if $|\theta|$ is large.

Stein's loss in variance estimation

$$L(\sigma^2, \hat{\sigma}^2) = \frac{\hat{\sigma}^2}{\sigma^2} - 1 - \log\left(\frac{\hat{\sigma}^2}{\sigma^2}\right)$$

This loss is more complicated than squared error loss, but it has some reasonable properties. For any fixed value of σ^2, $L(\sigma^2, \hat{\sigma}^2) \to \infty$ as $\hat{\sigma}^2 \to 0$ or $\hat{\sigma}^2 \to \infty$. Thus, gross underestimation is penalized just as heavily as gross overestimation.

- All loss functions are non-negative
- The loss is zero when the estimator matches the parameter value

Risk Function

Definition: Risk function is expected loss of an estimator.

$$R(\theta, \hat{\theta}) \;=\; \mathrm{E}[L(\theta, \hat{\theta}(\mathbf{X}))|\theta]$$

Highlights on risk function

- If $L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$, $R(\theta, \hat{\theta})$ is MSE.
- Loss and risk functions are not restricted to the Bayesian framework. It can be applied to any estimators.
- For example, UMVUE minimizes the risk function for squared error loss among all unbiased estimators, across all θ.
- Across all possible estimators, uniformly minimizing risk function across all θ is extremely difficult and often impossible (e.g. MSE).
- However, under the Bayesian framework where the distribution of θ is given, finding the best estimator is possible.

Bayes Risk

Bayes risk is defined as the average risk across all values of θ given prior $\pi(\theta)$

$$\int_{\Omega} R(\theta, \hat{\theta})\pi(\theta)d\theta$$

The Bayes rule with respect to a prior π is the optimal estimator with respect to a Bayes risk, which is defined as the one that minimize the Bayes risk.

Alternative definition of Bayes Risk

$$
\begin{aligned}
\int_\Omega R(\theta, \hat{\theta})\pi(\theta)d\theta &= \int_\Omega \mathrm{E}[L(\theta, \hat{\theta}(\mathbf{X}))]\pi(\theta)d\theta \\
&= \int_\Omega \left[\int_\mathcal{X} f(\mathbf{x}|\theta)L(\theta, \hat{\theta}(\mathbf{x}))d\mathbf{x}\right]\pi(\theta)d\theta \\
&= \int_\Omega \left[\int_\mathcal{X} f(\mathbf{x}|\theta)L(\theta, \hat{\theta}(\mathbf{x}))\pi(\theta)d\mathbf{x}\right]d\theta \\
&= \int_\Omega \left[\int_\mathcal{X} \pi(\theta|\mathbf{x})m(\mathbf{x})L(\theta, \hat{\theta}(\mathbf{x}))d\mathbf{x}\right]d\theta \\
&= \int_\mathcal{X} \left[\int_\Omega \pi(\theta|\mathbf{x})L(\theta, \hat{\theta}(\mathbf{x}))d\theta\right]m(\mathbf{x})d\mathbf{x}
\end{aligned}
$$

The quantity in square brackets is a function of $\mathbf{x}$ only. Minimizing the Bayes risk is equivalent to minimizing for each given $\mathbf{x} \in \mathcal{X}$, the quantity inside the bracket, which is called the *posterior expected loss*.

Posterior Expected Loss

$$
\int_\Omega R(\theta, \hat{\theta})\pi(\theta)d\theta = \int_\mathcal{X} \left[\int_\Omega \pi(\theta|\mathbf{x})L(\theta, \hat{\theta}(\mathbf{x}))d\theta\right]m(\mathbf{x})d\mathbf{x}
$$

Posterior expected loss is defined as

$$
\mathrm{E}\left[L(\theta, \hat{\theta})|X = \mathbf{x}\right] = \int_\Omega \pi(\theta|\mathbf{x})L(\theta, \hat{\theta}(\mathbf{x}))d\theta
$$

Bayes estimator is the estimator that minimizes the posterior expected loss.

Bayes Estimator based on squared error loss

$$\begin{aligned} L(\hat{\theta}, \theta) &= (\hat{\theta} - \theta)^2 \\ \text{Posterior expected loss} &= \int_{\Omega} (\theta - \hat{\theta})^2 \pi(\theta|\mathbf{x}) d\theta \\ &= \mathrm{E}[(\theta - \hat{\theta})^2 | \mathbf{X} = \mathbf{x}] \end{aligned}$$

So, the goal is to minimize $\mathrm{E}[(\theta - \hat{\theta})^2 | \mathbf{X} = \mathbf{x}]$

$$\begin{aligned} \mathrm{E}\left[(\theta - \hat{\theta})^2 | \mathbf{X} = \mathbf{x}\right] &= \mathrm{E}\left[\left(\theta - \mathrm{E}(\theta|\mathbf{X}) + \mathrm{E}(\theta|\mathbf{X}) - \hat{\theta}\right)^2 \middle| \mathbf{X} = \mathbf{x}\right] \\ &= \mathrm{E}\left[(\theta - \mathrm{E}(\theta|\mathbf{X}))^2 \middle| \mathbf{X} = \mathbf{x}\right] + \mathrm{E}\left[\left(\mathrm{E}(\theta|\mathbf{X}) - \hat{\theta}\right)^2 \middle| \mathbf{X} = \mathbf{x}\right] \\ &= \mathrm{E}\left[(\theta - \mathrm{E}(\theta|\mathbf{X}))^2 \middle| \mathbf{X} = \mathbf{x}\right] + \left[\mathrm{E}(\theta|\mathbf{x}) - \hat{\theta}\right]^2 \end{aligned}$$

which is minimized when $\hat{\theta} = \mathrm{E}(\theta|\mathbf{x})$.

Example 2 - Binomial Bayes estimator Let $X_1, \cdots, X_n$ be i.i.d. *Bernoulli*(p), $p \sim \mathrm{Beta}(\alpha, \beta)$. Recall that

$$\hat{p} = \frac{\sum_{i=1}^n X_i}{n} \qquad \hat{p}_B = \frac{\sum_{i=1}^n X_i + \alpha}{\alpha + \beta + n}$$

are MLE and Bayes estimators of p, respectively. Assuming squared error loss,

1. What is the risk function of $\hat{p}$?
2. What is the risk function of $\hat{p}_B$?

3. Compare the Bayes risk between $\hat{p}$ and $\hat{p}_B$.
4. In the absence of good prior information about p, if we want to make risk function of $\hat{p}_B$ constant (based on squared error loss), what should be α and β?
5. Compare the risk functions between $\hat{p}$ and $\hat{p}_B$ from the previous problem, when $n = 4$ and $n = 400$.

Solution: For squared error loss, risk function is MSE. Now MSE of $\hat{p} = \overline{X}$ is

$$\mathrm{E}[\hat{p} - p]^2 \;=\; \mathrm{Var}(\overline{X}) = \frac{p(1-p)}{n}$$

On the other hand, risk function of $\hat{p}_B$ equals

$$\begin{aligned}
\mathrm{E}[\hat{p}_B - p]^2 &= \mathrm{Var}(\hat{p}_B) + [\mathrm{Bias}(\hat{p}_B)]^2 \\
&= \mathrm{Var}\left(\frac{\sum_{i=1}^n X_i + \alpha}{\alpha + \beta + n}\right) + \left[\mathrm{E}\left(\frac{\sum_{i=1}^n X_i + \alpha}{\alpha + \beta + n}\right) - p\right]^2 \\
&= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left[\frac{np + \alpha}{\alpha + \beta + n} - p\right]^2
\end{aligned}$$

Bayes Risk

For MLE $\hat{p}$

$$\begin{aligned}
R(\hat{p}, p) &= \mathrm{E}[\hat{p} - p]^2 = \mathrm{Var}(\overline{X}) = \frac{p(1-p)}{n} \\
\textstyle\int_0^1 R(\hat{p}, p)\pi(p)dp &= \int_0^1 \frac{p(1-p)}{n}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}dp \\
&= \tfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\tfrac{\Gamma(\alpha+1)\Gamma(\beta+1)}{n\Gamma(\alpha+\beta+2)}\textstyle\int_0^1 \tfrac{\Gamma(\alpha+\beta+2)}{\Gamma(\alpha+1)\Gamma(\beta+1)}p^{\alpha}(1-p)^{\beta}dp \\
&= \frac{\alpha\beta}{n(\alpha+\beta+1)(\alpha+\beta)}
\end{aligned}$$

For Bayes estimator $\hat{p}_B$

$$\begin{aligned}
R(\hat{p}_B, p) &= \mathrm{E}[\hat{p}_B - p]^2 \\
&= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left[\frac{np+\alpha}{\alpha+\beta+n} - p\right]^2 \\
&= \frac{np(1-p) + \alpha^2(1-p)^2 - 2\alpha\beta p(1-p) + \beta^2 p^2}{(\alpha+\beta+n)^2} \\
\mathrm{E}[R] &= \tfrac{\Gamma(\alpha+\beta)\big[(n-2\alpha\beta)\Gamma(\alpha+1)\Gamma(\beta+1)+\alpha^2\Gamma(\alpha)\Gamma(\beta+2)+\beta^2\Gamma(\alpha+2)\Gamma(\beta)\big]}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+2)(\alpha+\beta+n)^2} \\
&= \tfrac{\alpha\beta[n-2\alpha\beta+\alpha(\beta+1)+\beta(\alpha+1)]}{(\alpha+\beta+1)(\alpha+\beta)(\alpha+\beta+n)^2} \\
&= \frac{(n+\alpha+\beta)\alpha\beta}{(\alpha+\beta+n)^2(\alpha+\beta+1)(\alpha+\beta)} \\
&= \frac{\alpha\beta}{(\alpha+\beta+n)(\alpha+\beta+1)(\alpha+\beta)}
\end{aligned}$$

Comparing two Bayes risks

$$
\begin{aligned}
\int_0^1 R(\hat{p},p)\pi(p)dp &= \frac{\alpha\beta}{n(\alpha+\beta+1)(\alpha+\beta)} \\
\int_0^1 R(\hat{p}_B,p)\pi(p)dp &= \frac{\alpha\beta}{(\alpha+\beta+n)(\alpha+\beta+1)(\alpha+\beta)} \\
\frac{1}{(\alpha+\beta+n)} &\leq \frac{1}{n}
\end{aligned}
$$

$\hat{p}_B$ always has smaller Bayes risk than $\hat{p}$.

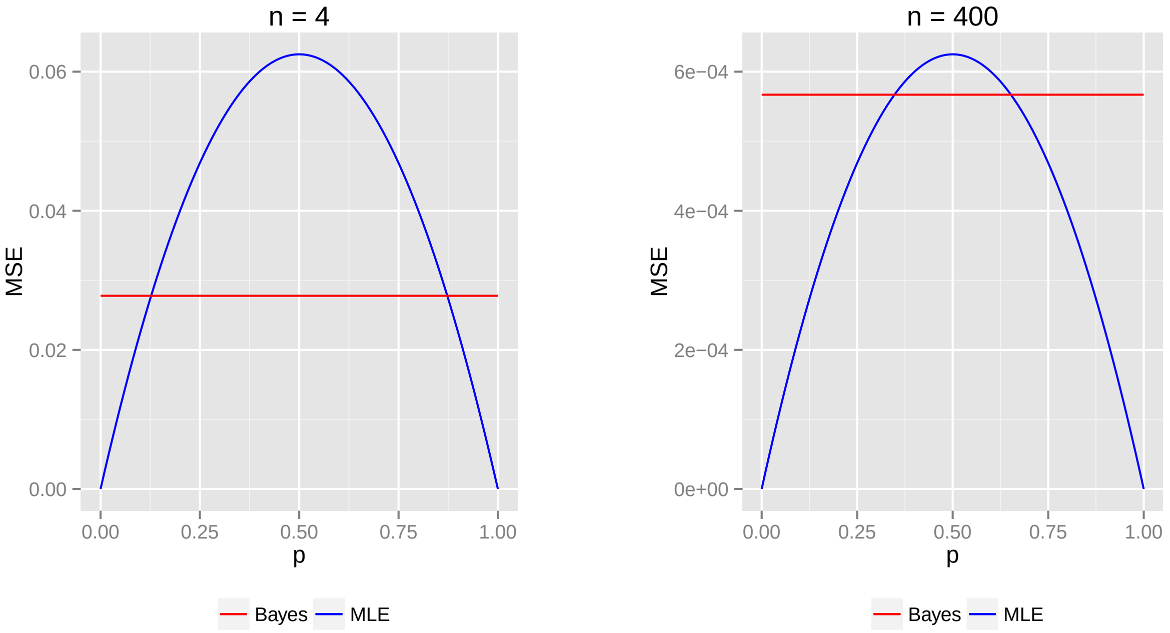
Condition for constant risk function

$$
\begin{aligned}
\mathrm{E}[\hat{p}_B - p]^2 &= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left[\frac{np+\alpha}{\alpha+\beta+n} - p\right]^2 \\
&= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left[\frac{\alpha-(\alpha+\beta)p}{\alpha+\beta+n}\right]^2 \\
&= \frac{[(\alpha+\beta)^2-n]p^2 + [n-2\alpha(\alpha+\beta)]p + \alpha^2}{(\alpha+\beta+n)^2}
\end{aligned}
$$

$$
\begin{aligned}
\alpha+\beta &= \sqrt{n} \\
\alpha &= \frac{n}{2(\alpha+\beta)} = \frac{1}{2}\sqrt{n} \\
\beta &= \sqrt{n} - \alpha = \frac{1}{2}\sqrt{n}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}[\hat{p} - p]^2 &= \frac{p(1-p)}{n} \\
\mathrm{E}[\hat{p}_B - p]^2 &= \frac{[(\alpha+\beta)^2-n]p^2 + [n-2\alpha(\alpha+\beta)]p + \alpha^2}{(\alpha+\beta+n)^2} \\
&= \frac{n}{4(n+\sqrt{n})^2}
\end{aligned}
$$

Comparing Risk functions



- There is no uniform winner. As p is closer to the boundaries of its domain, $\hat{p}$ is better than $\hat{p}_B$.
- As the sample size grows larger, there is a larger range of p for which $\hat{p}$ is superior to $\hat{p}_B$.

Different Bayes Estimators

Bayes estimators are minimizers of expected loss, and hence depend directly on the choice of loss function. Consider a point estimation problem for real-valued parameter θ.

Squared error loss

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$$

The posterior expected loss is

$$\int_{\Omega} (\theta - \hat{\theta})^2 \pi(\theta|\mathbf{x}) d\theta \;\;=\;\; \mathrm{E}[(\theta - \hat{\theta})^2 | \mathbf{X} = \mathbf{x}]$$

This expected value is minimized by $\hat{\theta}_B = \mathrm{E}(\theta|\mathbf{x})$. So the Bayes estimator is the mean of the posterior distribution.

Absolute error loss

$$L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$$

The posterior expected loss is

$$\begin{aligned} \mathrm{E}[L(\theta, \hat{\theta})|\mathbf{x}] &= \mathrm{E}[|\theta - \hat{\theta}||\mathbf{X} = \mathbf{x}] \\ &= \int_{\Omega} |\theta - \hat{\theta}(\mathbf{x})| \pi(\theta|\mathbf{x}) d\theta \\ &= \int_{-\infty}^{\hat{\theta}} -(\theta - \hat{\theta}) \pi(\theta|\mathbf{x}) d\theta + \int_{\hat{\theta}}^{\infty} (\theta - \hat{\theta}) \pi(\theta|\mathbf{x}) d\theta \end{aligned}$$

In order to minimize the posterior expected loss, we make use of Leibnitz's rule

$$\frac{\partial}{\partial \theta} \int_{a(\theta)}^{b(\theta)} f(x|\theta) dx = f(b(\theta)|\theta) b'(\theta) - f(a(\theta)|\theta) a'(\theta) + \int_{a(\theta)}^{b(\theta)} \frac{\partial}{\partial \theta} f(x|\theta) dx$$

where the formula includes $a(\theta) = -\infty,\ b(\theta) = \infty$. Taking derivative with respect to $\hat{\theta}$ and setting it equal to zero, we have (using Leibnitz's rule)

$$\begin{aligned} \frac{\partial}{\partial \hat{\theta}} \mathrm{E}[L(\theta, \hat{\theta}(\mathbf{x}))] &= -(\hat{\theta} - \hat{\theta}) \pi(\hat{\theta}|\mathbf{x}) + \int_{-\infty}^{\hat{\theta}} \pi(\theta|\mathbf{x}) d\theta \\ & \quad -(\hat{\theta} - \hat{\theta}) \pi(\hat{\theta}|\mathbf{x}) - \int_{\hat{\theta}}^{\infty} \pi(\theta|\mathbf{x}) d\theta = 0 \end{aligned}$$

The solution $\hat{\theta}_B$ satisfies

$$\int_{-\infty}^{\hat{\theta}} \pi(\theta|\mathbf{x})d\theta = \int_{\hat{\theta}}^{\infty} \pi(\theta|\mathbf{x})d\theta$$

Thus, $\hat{\theta}_B$ is the <u>posterior median</u>. That it is the unique minimizer is easily verified by observing

$$\frac{\partial}{\partial\hat{\theta}}\left[\int_{-\infty}^{\hat{\theta}} \pi(\theta|\mathbf{x})d\theta - \int_{\hat{\theta}}^{\infty} \pi(\theta|\mathbf{x})d\theta\right] = 2\pi(\hat{\theta}|\mathbf{x}) > 0$$

Example 3: Normal Bayes Estimators Let $X_1, \cdots, X_n \sim \mathcal{N}(\theta, \sigma^2)$ and suppose that the prior distribution of θ is $\mathcal{N}(\mu, \tau^2)$. Assuming that σ^2, μ^2, τ^2 are all known, what is the Bayes estimator based on (a) squared error loss and (b) the absolute error loss?

Solution: The posterior distributon of θ given $\mathbf{x}$ is normal with

$$\begin{aligned}
\mathrm{E}[\theta|\mathbf{x}] &= \frac{\tau^2}{\tau^2 + \frac{1}{n}\sigma^2}\overline{x} + \frac{\frac{1}{n}\sigma^2}{\tau^2 + \frac{1}{n}\sigma^2}\mu \\
\mathrm{Var}(\theta|\mathbf{x}) &= \frac{\frac{1}{n}\sigma^2\tau^2}{\tau^2 + \frac{1}{n}\sigma^2}
\end{aligned}$$

- For squared error loss, the Bayes estimator is $\hat{\theta} = \mathrm{E}[\theta|\mathbf{x}]$.
- For absolute error loss, the Bayes estimator is also $\hat{\theta} = \mathrm{E}[\theta|\mathbf{x}]$ (why?)

Example 4: Let $X_1, \cdots, X_n \sim \text{Bernoulli}(p)$ and $\pi(p) \sim \text{Beta}(\alpha, \beta)$. What is the Bayes estimator with respect to (a) squared error loss and (b) absolute error loss?

Solution:

- The posterior distribution follows $\text{Beta}(\sum x_i + \alpha, n - \sum x_i + \beta)$.
- Bayes estimator that minimizes posterior expected squared error loss is the posterior mean

$$\hat{p} = \frac{\sum x_i + \alpha}{\alpha + \beta + n}$$

Bayes estimator that minimizes posterior expected absolute error loss is the posterior median satisfying

$$\int_0^{\hat{\theta}} \frac{\Gamma(\alpha + \beta + n)}{\Gamma(\sum x_i + \alpha)\Gamma(n - \sum x_i + \beta)} p^{\sum x_i + \alpha - 1}(1-p)^{n - \sum x_i + \beta - 1} dp = \frac{1}{2}$$

There is no closed form solution for $\hat{\theta}$, but it can be represented in terms of incomplete beta function.

Example 5: Let $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. Consider an estimator of σ^2,

$$\sigma_b^2 \;=\; bs_{\mathbf{X}}^2 = \frac{b\sum_{i=1}^n (X_i - \overline{X})^2}{n-1},$$

i.e. consider an estimator in the class of scale multiples of the sample variance.

1. Using squared error loss, what is the b that minimizes Bayes risk?
2. Using Stein's loss function,

$$L(\sigma^2, \sigma_b^2) \;=\; \frac{\sigma_b^2}{\sigma^2} - 1 - \log \frac{\sigma_b^2}{\sigma^2}$$

   what is the b that minimizes Bayes risk?

Biostat 602 Winter 2017

Lecture Set 14

Topics on Bayesian Inference

Noninformative Prior, Empirical Bayes, Hierarchical Bayes, MCMC

Noninformative Prior

- A criticism about prior specification in Bayesian analysis is its subjectivity. Sometimes there is an objective basis (e.g. historical data) for specifying a prior distribution.
- Often, there is no such objective basis other than a knowledge of the domain of the variable. For example, for a $\mathcal{N}(\mu, 1)$ sampling distribution, $\mu \in \mathcal{R}$. So a prior stretched on the real line is appropriate.
- In such cases, a noninformative prior is often used. For the $\mathcal{N}(\mu, 1)$ example,
  $$\pi(\mu) = 1, \quad \mu \in \mathcal{R},$$
  is a noninformative prior.
- Most often noninformative priors are 'improper' in the sense that they do not integrate to 1. That is okay as long as the $\underline{\textbf{posterior is proper}}$.
- With non-informative priors, posterior estimates often match frequentist estimates.
- There are different considerations that lead to a class of candidates for a non-informative prior, mostly motivated from the perspective of matching frequentist results.

In the rest of the lecture, a distribution that will be used extensively is *Inverse Gamma*, a distribution you have seen in the last assignment. It will be useful to recap its definition.

Definition: A random variable X is said to follow an Inverse Gamma distribution with parameters α, β, to be denoted by $IG(\alpha, \beta)$ if

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

Further we have

$$\mathrm{E}(X) = \frac{\beta}{\alpha - 1}, \ \alpha > 1 \qquad \mathrm{Var}(X) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}, \ \alpha > 2.$$

Example 1: Consider $X_1, \ldots, X_n$ to be a i.i.d. random sample from $\mathcal{N}(\mu, \sigma^2)$, with σ^2 known. Assume the non-informative prior for μ as

$$\pi(\mu) = 1, \quad \mu \in \mathcal{R}.$$

Likelihood

$$\begin{aligned} f(\mathbf{x}|\mu) &= \left(\frac{1}{2\pi}\right)^{n/2} (\sigma^{-2})^{n/2} \exp\left[-\sum_{i=1}^{n}(x_i - \mu)^2/(2\sigma^2)\right] \\ &\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x})^2 - \frac{n}{2\sigma^2}(\overline{x} - \mu)^2\right] \end{aligned}$$

Posterior

$$\begin{aligned} \pi(\mu|\mathbf{x}) &\propto f(\mathbf{x}|\mu)\pi(\mu) \\ &\propto \exp\left[-\frac{n}{2\sigma^2}(\mu - \overline{x})^2\right] \quad \sim \mathcal{N}\left(\overline{x}, \frac{\sigma^2}{n}\right) \end{aligned}$$

Bayes estimator under squared error loss

$$\hat{\mu}_B = \mathrm{E}[\mu|\mathbf{x}] = \overline{x}.$$

Example 2: Same set up as Example 1, but σ^2 is unknown. Consider μ and σ^2 to be independent *apriori* with

$$\pi(\mu) = 1, \ \mu \in \mathcal{R}, \quad \pi(\sigma^2) = \sigma^{-2}, \ \sigma^2 > 0$$

leading to the joint prior

$$\pi(\mu, \sigma^2) = \sigma^{-2}, \quad \mu \in \mathcal{R}, \ \sigma^2 > 0.$$

Joint Posterior

$$
\begin{aligned}
\pi(\mu, \sigma^2|\mathbf{x}) &\propto f(\mathbf{x}|\mu, \sigma^2)\pi(\mu, \sigma^2) \\
&= \left(\frac{1}{2\pi}\right)^{n/2} (\sigma^2)^{-n/2} \exp\left[-\sum_{i=1}^{n}(x_i - \mu)^2/(2\sigma^2)\right] \times (\sigma^2)^{-1} \\
&\propto (\sigma^2)^{-(n/2)-1} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x})^2 - \frac{n}{2\sigma^2}(\overline{x} - \mu)^2\right]
\end{aligned}
$$

Marginal Posterior of σ^2

$$
\begin{aligned}
\pi(\sigma^2|\mathbf{x}) &= \int_{-\infty}^{\infty} \pi(\mu, \sigma^2|\mathbf{x}) \; d\mu \\
&\propto (\sigma^2)^{-(n/2)-1} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x})^2\right] \times \int_{-\infty}^{\infty} \exp\left[-\frac{n}{2\sigma^2}(\overline{x} - \mu)^2\right] \; d\mu \\
&= (\sigma^2)^{-(n/2)-1} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x})^2\right] \times \sqrt{2\pi}\left(\frac{\sigma^2}{n}\right)^{1/2} \\
&\propto (\sigma^2)^{-\frac{n-1}{2}-1} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x})^2\right] \\
&\sim IG\left(\frac{n-1}{2},\ \frac{1}{2}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)
\end{aligned}
$$

Bayes estimator under squared error loss:

$$
E[\sigma^2|\mathbf{x}] = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{2(\frac{n-1}{2} - 1)} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n - 3}.
$$

Marginal Posterior of μ

This is more complicated and yields a truncated distribution. But since the posterior is symmetric about $\overline{x}$, $E[\mu|\mathbf{x}] = \overline{x}$.

Empirical Bayes

- Bayesians introduce a hierarchy in the modeling framework by assuming the parameter to be random. The prior distribution has its own parameter(s) and in a single-stage hierarchy, these parameters are chosen to be known.
- Empirical Bayes (EB) strategy deviates from the usual Bayesian framework in that it estimates the prior parameters from the marginal distribution of data instead of assuming them to be known.
- Since prior specification depends on data, EB is not strictly a Bayesian procedure. However, EB is generally an effective technique of constructing estimators that perform well under both Bayesian and frequentist criteria.
- EB estimators tend to be more robust than the traditional Bayes estimators against misspecification of prior.

Model

$$\begin{aligned} f(x_i|\theta) &\sim f(x|\theta), \quad i = 1, 2, \ldots, n \\ \pi(\theta) &\sim g(\theta|\gamma) \end{aligned}$$

Estimate γ based on the marginal distribution

$$m(\mathbf{x}|\gamma) = \int \prod_{i=1}^{n} f(x_i|\theta) g(\theta|\gamma) \; d\theta.$$

It is most common to use $\hat{\gamma} = MLE(\gamma)$. In the final expression of the Bayes estimator, replace γ by $\hat{\gamma}$. Then

$$\hat{\theta}_{EB} = \min_{a(\mathbf{x})} \int L(\theta, a(\mathbf{x})) \pi(\theta|\mathbf{x}, \hat{\gamma}) \; d\theta.$$

Example 3: Consider independent random variables $X_1, X_2, \ldots, X_p$ such that $X_i|\theta_i \sim \mathcal{N}(\theta_i, \sigma^2), \quad i = 1, 2, \ldots, p$ where σ^2 is known. Assume θ_i to be i.i.d $\mathcal{N}(\mu, \tau^2)$.

This is like balanced one-way random effects model where X_i represents the mean of the i-th group. Assume Squared Error loss

$$L(\theta, \hat{\theta}) = \sum_{i=1}^{p} (\theta_i - \hat{\theta}_i)^2.$$

Posterior distribution is given by

$$\begin{aligned} \pi(\theta_i|x_i) &\propto \left(\frac{1}{2\pi\sigma^2}\right)^{p/2} \prod_{i=1}^{p} e^{-(x_i-\theta_i)^2/(2\sigma^2)} \frac{1}{\sqrt{2\pi}\tau} e^{-(\theta_i-\mu)^2/(2\tau^2)} \\ &\propto e^{-(x_i-\theta_i)^2/(2\sigma^2)} \times e^{-(\theta_i-\mu)^2/(2\tau^2)} \\ &\sim \mathcal{N}\left(\hat{\theta}_i^B, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right) \end{aligned}$$

where the posterior mean is the Bayes estimate for θ_i given by

$$\hat{\theta}_i^B = \frac{\sigma^2}{\sigma^2+\tau^2}\mu + \frac{\tau^2}{\sigma^2+\tau^2}X_i.$$

Unlike the single-stage Bayes estimation, μ and τ^2 are not assumed known. Instead they are estimated from the marginal distribution of X_i (unconditional on θ_i). It turns out that

$$m(X_i) \sim \mathcal{N}(\mu, \sigma^2 + \tau^2), \quad i = 1, 2, \ldots, p.$$

We shall prove this fact later. As a consequence of this distributional fact, we have

$$E(\overline{X}) = \mu, \quad E\left[\frac{(p-3)\sigma^2}{\sum_{i=1}^{p}(X_i - \overline{X})^2}\right] = \frac{\sigma^2}{\sigma^2+\tau^2}.$$

The second equation uses the fact that if $Y \sim \chi^2_k$ then $E(1/Y) = 1/(k-2)$.

Then, the EB estimator assumes the form

$$\hat{\theta}_i^{EB} = E[\theta_i|X_i] = \frac{(p-3)\sigma^2}{\sum_{i=1}^p (X_i - \overline{X})^2}\overline{X} + \left[1 - \frac{(p-3)\sigma^2}{\sum_{i=1}^p (X_i - \overline{X})^2}\right] X_i.$$

Interpretation: If X_i data has substantial variability compared to σ^2, then $\hat{\theta}_i^{EB}$ relies more on X_i. In the other case, $\hat{\theta}_i^{EB}$ should be shrunk more towards the mean $\overline{X}$.

The EB estimator has an appealing property: if $p \geq 4$, on an average it is always closer to θ_i than X_i. More specifically,

$$\mathrm{E}\left[\sum_{i=1}^p (\theta_i - \hat{\theta}_i^{EB})^2 \middle| \theta_i\right] < \mathrm{E}\left[\sum_{i=1}^p (\theta_i - X_i)^2 \middle| \theta_i\right]$$

Derivation of the Marginal Distribution

The marginal distribution of X_i can be derived by integrating out θ_i from the joint distribution of θ_i and X_i. This is possible but involves tedious manipulations. Instead one can use moment generating function to come up with an easy and elegant derivation. Towards that first note that if $Y \sim \mathcal{N}(m, \gamma^2)$, then its moment generating function is given by

$$E[e^{tY}] = \exp\left[mt + \frac{\gamma^2 t^2}{2}\right].$$

If we can show that the moment generating function of the marginal distribution of X_i corresponds to that of the desired normal, then we have established the result. Now,

$$\begin{aligned} E[e^{tX_i}] &= E\left[E[e^{tX_i}|\theta_i]\right] \\ &= E\left[\exp\left[\theta_i t + \frac{\sigma^2 t^2}{2}\right]\right] \\ &= \exp\left[\frac{\sigma^2 t^2}{2}\right] E\left[e^{\theta_i t}\right] \\ &= \exp\left[\frac{\sigma^2 t^2}{2}\right] \times \exp\left[\mu t + \frac{\tau^2 t^2}{2}\right] \\ &= \exp\left[\mu t + \frac{(\sigma^2 + \tau^2) t^2}{2}\right]. \end{aligned}$$

Hence we have our desired result.

Hierarchical Bayes

Hierarchical Bayes (HB) strategy is also motivated from introducing more robustness in prior specification. Typical HB procedure adds a second level of hierarchy by assuming the first-stage prior parameters to be random. Specifically, we assume

$$f(x_i|\theta) \sim f(x|\theta), \quad \theta|\gamma \sim \pi(\theta|\gamma), \quad \gamma \sim \pi(\gamma).$$

In the last specification, all parameters of γ are known. The posterior still is calculated as

$$f(\theta|\mathbf{x}) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)\, d\theta},$$

but $\pi(\theta)$ can no longer be specified as a single-stage prior. In order to evaluate $\pi(\theta)$ one has to integrate out over the uncertainty of γ. More specifically,

$$\pi(\theta) = \int \pi(\theta|\gamma)\pi(\gamma)\, d\gamma.$$

Example 4: Let $X_1, X_2, \ldots, X_n$ be i.i.d.ransom sample from $\mathcal{N}(\theta, \sigma^2)$ where σ^2 is known. Consider the following hierarchy.

$$f(x|\theta) \sim \mathcal{N}(\theta, \sigma^2), \quad \theta|\gamma^2 \sim N(0, \gamma^2), \quad \frac{1}{\gamma^2} \sim \mathit{Exponential}(1).$$

Find Bayes estimator of θ under squared error loss.

Solution:

Likelihood

$$
\begin{aligned}
f(\mathbf{x}|\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-(x_i-\theta)^2/(2\sigma^2)\right] \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\sum_{i=1}^{n}(x_i-\theta)^2/(2\sigma^2)\right] \\
&\propto \exp\left[-\frac{n}{2\sigma^2}(\overline{x}-\theta)^2\right]
\end{aligned}
$$

Prior

$$
\begin{aligned}
\pi(\theta) &= \int_0^\infty \pi(\theta|\gamma^2)\pi(\gamma^2)d\gamma^2 \\
&= \frac{1}{\sqrt{2\pi}} \int_0^\infty (\gamma^2)^{-1/2} e^{-\theta^2/(2\gamma^2)} (\gamma^2)^{-2} e^{-1/\gamma^2} d\gamma^2 \\
&= \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-(1+\frac{\theta^2}{2})/\gamma^2} (\gamma^2)^{-\frac{3}{2}-1} d\gamma^2 \\
&= \frac{\Gamma(3/2)}{\sqrt{2\pi}} \left(1+\frac{\theta^2}{2}\right)^{-3/2}
\end{aligned}
$$

Posterior

$$\begin{aligned}\pi(\theta|x) &\propto f(\mathbf{x}|\theta)\pi(\theta) \\ &\propto \left(1+\frac{\theta^2}{2}\right)^{-3/2} \exp\left[-\frac{n}{2\sigma^2}(\overline{x}-\theta)^2\right]\end{aligned}$$

Then

$$E(\theta|\mathbf{x}) = \frac{\int_{-\infty}^{\infty} \theta \left(1+\frac{\theta^2}{2}\right)^{-3/2} \exp\left[-\frac{n}{2\sigma^2}(\overline{x}-\theta)^2\right] d\theta}{\int_{-\infty}^{\infty} \left(1+\frac{\theta^2}{2}\right)^{-3/2} \exp\left[-\frac{n}{2\sigma^2}(\overline{x}-\theta)^2\right] d\theta}$$

which does not reduce further and has to be evaluated numerically. MSE or Bayes risk calculation is more arduous.

Computational complexity is a common problem of Bayesian inference. The simulational approach advanced in the early nineties revolutionized the area of Bayesian inference.

Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) methods refer to a class of algorithms for drawing samples from a probability distribution based on constructing a Markov chain.

In Bayesian computation, one has to frequently evaluate high dimensional integrals that are intractable. Hence, summary measures based on the joint posteriors often are difficult to obtain. The simulation based approach offers a viable alternative to high dimensional integration that generates good and reliable estimates of posterior quantities.

Gibbs Sampling

Gibbs Sampling refers to a special MCMC algorithm which iteratively draws from all possible conditional distributions to ultimately yield observations fromm a multivariate joint distribution. We shall explain Gibbs sampling in the context of approximating a multidimensional joint posterior. Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)$ be the parameter vector; $\mathbf{X}$ = Observed data, $\pi(\boldsymbol{\theta})$: Joint Prior. We seek to approximate the joint posterior

$$\pi(\theta_1, \theta_2, \ldots, \theta_d | \mathbf{X}).$$

Gibbs sampling proceeds through the following steps:

Step 1: Generate a sample point $(\theta_1^0, \theta_2^0, \ldots, \theta_d^0)$ from $\pi(\boldsymbol{\theta})$.

Step 2: Generate θ_1^1 from the full conditional distribution

$$\pi(\theta_1 | \theta_2^0, \ldots, \theta_d^0, \mathbf{X})$$

.

Step 3: Generate

$$\begin{aligned}
\theta_2^1 &\sim \pi(\theta_2 | \theta_1^1, \theta_3^0, \ldots, \theta_d^0, \mathbf{X}) \\
\theta_3^1 &\sim \pi(\theta_3 | \theta_1^1, \theta_2^1, \theta_3^0, \ldots, \theta_d^0, \mathbf{X}) \\
&\vdots \\
&\vdots \\
\theta_d^1 &\sim \pi(\theta_d | \theta_1^1, \theta_2^1, \ldots, \theta_{d-1}^1, \mathbf{X})
\end{aligned}$$

Step 4: Repeat Step 2–Step3 M times, with M typically in the order of 50,000 or more.

Then,

1. After an initial burn-in (of 5000, say), the random numbers $(\theta_1, \theta_2, \ldots \theta_d)$ constitute a sample from the joint posterior $\pi(\theta_1, \theta_2, \ldots, \theta_d | \mathbf{X})$. This follows from the stochastic behavior of the Markov Chain generated by the sequential random draws prescribed in Steps 2–3.

2. The marginal posterior distribution of any subset of parameters can be approximated by simply considering the samples for that subset.

3. Summary measures from the posterior are approximated by empirical estimates based on the MCMC samples.

- Gibbs sampling is particularly useful when the full conditionals yield easy to sample from distributions.

- If that is not the case for all conditionals, other sample generation algorithms are employed.

Poisson Hierarchy with Gibbs Sampling

Example 5: Consider

$$X|\lambda \sim Poisson(\lambda); \quad \lambda|b \sim Gamma(a, b),\ a\,\text{known}; \quad \frac{1}{b} \sim Gamma(\alpha, \beta),\ \ \alpha, \beta\,\text{known}.$$

In this hierarchy $\pi(\lambda|X)$ does not conform to a known distribution, neither is it expressible in a simple form. We shall attempt to use MCMC to simulate random samples from the targeted posterior.

Joint pdf

$$\begin{aligned} f(x, \lambda, b) &= f(x|\lambda)\pi(\lambda|b)\pi(b) \\ &\propto e^{-\lambda}\lambda^{x}\ \frac{1}{b^{a}}\lambda^{a-1}e^{-\lambda/b}\ b^{-\alpha-1}e^{-\beta/b} \end{aligned}$$

Full Conditionals

$$\lambda|x, b \ \propto\ \lambda^{x+a-1}e^{-\lambda(1+\frac{1}{b})}\ \sim\ Gamma\left(x+a, \frac{b}{b+1}\right) \tag{1}$$

$$b|x, \lambda \ \propto\ b^{-(a+\alpha)-1}e^{-(\lambda+\beta)/b}\ \sim\ InvGamma(a+\alpha, \lambda+\beta) \tag{2}$$

Iteratively generate random numbers from (1) and (2) for a large number of times. After throwing out an initial batch, the remaining samples of (λ, b) are assumed to come from the joint posterior distribution of (λ, b).

The λ-samples then correspond to a sample from the target posterior $\pi(\lambda|x)$. One can estimate summary measures of $\pi(\lambda|x)$ empirically using these MCMC samples. For example, the sample mean of these samples corresponds to $\mathrm{E}[\lambda|x]$. The distribution itself can be approximated by the histogram.

In the previous example, the full conditionals were distributions that were easy to sample from. Often one or more of the conditionals would not come from the standard list of distributions. One needs, at these times, an algorithm that allows sampling from these non-standard distributions in an efficient manner.

A Rejection Algorithm

Consider a target pdf (without the normalization factor) $f(\theta)$ which is hard to sample from. Thus, $f(\theta)$ is simply a positive function that is integrable. Let $g(\theta)$ be a pdf which has the same support as f, but is easy to sample from. Consider the situation when there is a constant $M > 0$ such that

$$\frac{f(\theta)}{g(\theta)} \leq M \quad \text{for all } \theta.$$

To generate a sample from f, follow the steps:

1. Generate θ from $g(\theta)$.
2. Generate u from $Uniform(0,1)$.
3. If $u \leq \frac{f(\theta)}{M\ g(\theta)}$, then accept θ. Otherwise repeat steps 1–3.

Any accepted θ is a random observation from the (normalized) f.

Remarks

- Easy to program algorithm
- One has to be careful in choosing M. M may not be readily available. Even if it is, if M is too large, the acceptance probability will be low, thereby considerably slowing down the sample generation.

Weighted Bootstrap

In cases where M is not readily available, we can carry out a weighted bootstrap using the following steps:

1. Draw $\theta_i,\ i = 1, 2, \ldots, k$, a sample from $g(\theta)$.
2. Calculate, for $i = 1, 2, \ldots, k$,

$$\begin{aligned} \omega_i &= \frac{f(\theta_i)}{g(\theta_i)} \\ q_i &= \frac{\omega_i}{\sum_{j=1}^{k} \omega_j} \end{aligned}$$

3. Draw θ^* from the discrete distribution over $\{\theta_1, \theta_2, \ldots, \theta_k\}$ placing mass q_i on θ_i.

Then θ^* is approximately distributed with a pdf equaling (normalized) f. The improvement increases as k increases.

Example 6: $X_1, \ldots, X_n \sim Weibull(\gamma, \beta)$ with pdf

$$h(x|\gamma, \beta) = \frac{\gamma}{\beta} x^{\gamma-1} \exp(-x^{\gamma}/\beta), \quad x > 0, \gamma > 0, \beta > 0.$$

Further β and γ are assumed to be independent random variables with

$$\beta \sim InvGamma(a, b), \quad \gamma \sim Gamma(c, d).$$

Joint posterior is not in tractable form. So we shall use MCMC.

Likelihood

$$h(\mathbf{x}|\gamma,\beta) \;=\; \left(\frac{\gamma}{\beta}\right)^n \left(\prod_{i=1}^n x_i\right)^{\gamma-1} \exp(-\sum_{i=1}^n x_i^\gamma/\beta)$$

Joint Prior

$$\pi(\gamma,\beta) \;=\; \frac{1}{\Gamma(c)d^c}\gamma^{c-1}e^{-\gamma/d} \times \frac{b^a}{\Gamma(a)}\beta^{-a-1}e^{-b/\beta}$$

Joint Posterior

$$\begin{aligned}\pi(\gamma,\beta|\mathbf{x}) \;&\propto\; \gamma^{n+c-1}\exp\left[-\gamma\left(\sum_{i=1}^n \log(1/x_i)+d\right)\right]\\ &\qquad\times\,\beta^{-(n+a)-1}\exp\left[-\left(b+\sum_{i=1}^n x_i^\gamma\right)/\beta\right]\end{aligned}$$

Full Conditional of β

$$\pi(\beta|\gamma,\mathbf{x}) \;\propto\; \beta^{-(n+a)-1}\exp\left[-\left(b+\sum_{i=1}^n x_i^\gamma\right)/\beta\right] \;\;\sim\;\; InvGamma(n+a,\sum_{i=1}^n x_i^\gamma + b)$$

Full Conditional of γ

$$\pi(\gamma|\beta,\mathbf{x}) \;\propto\; \gamma^{n+c-1}\exp\left[-\gamma\left(\sum_{i=1}^n \log(1/x_i)+d\right)\right] \times e^{-\sum_{i=1}^n x_i^\gamma/\beta} \;\equiv\; f(\gamma)$$

Take g to be the pdf of $Gamma\left(n+c, \left(\sum_{i=1}^n \log(1/x_i)+d\right)^{-1}\right)$. Then

$$\begin{aligned}\frac{f(\gamma)}{g(\gamma)} \;&=\; \frac{\Gamma(n+c)}{\left(\sum_{i=1}^n \log(1/x_i)+d\right)^{n+c}} \times e^{-\sum_{i=1}^n x_i^\gamma/\beta}\\ &\le\; \frac{\Gamma(n+c)}{\left(\sum_{i=1}^n \log(1/x_i)+d\right)^{n+c}} \;\equiv\; M\end{aligned}$$

R code for a Gibbs Example

Example Consider $X_1, X_2, \ldots, X_n$ a random sample from $Poisson(\lambda)$. Consider the setup,

$$X_i|\lambda \sim Poisson(\lambda);\quad \lambda|b \sim Gamma(a,b),\ a\,\text{known};\quad \frac{1}{b} \sim Gamma(\alpha,\beta),\ \ \alpha,\beta\,\text{known}.$$

Full Conditionals

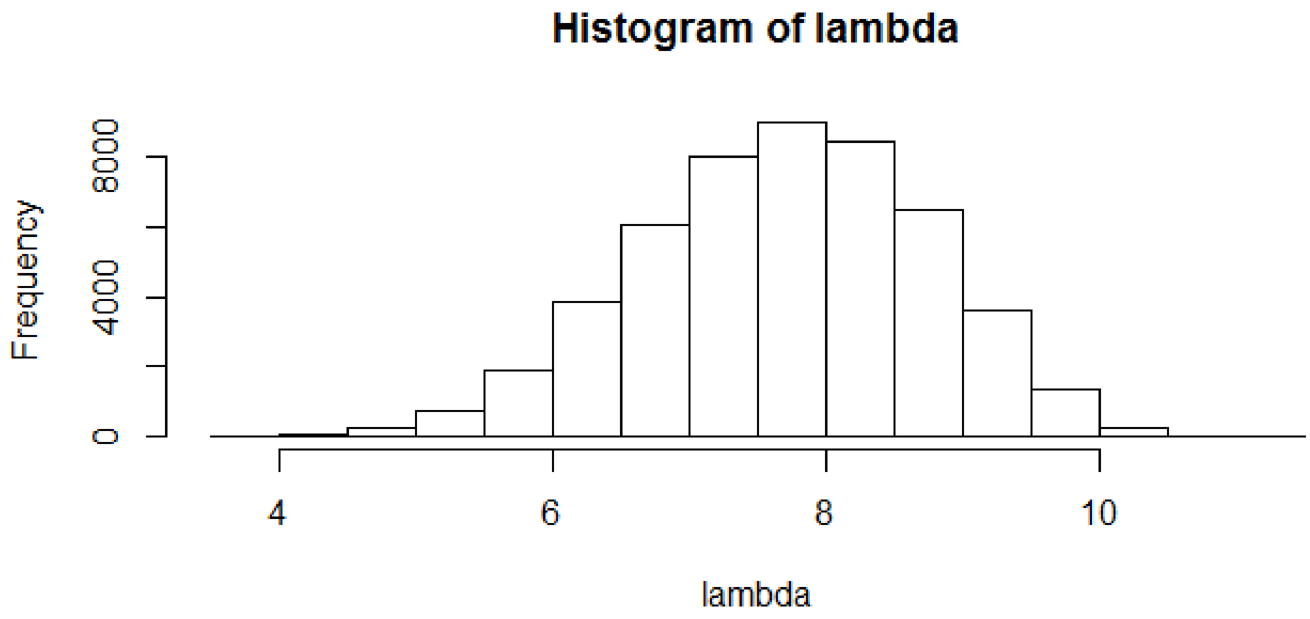
$$\lambda|x,b \ \propto\ \lambda^{\sum_{i=1}^{n} x_i+a-1} e^{-\lambda(n+\frac{1}{b})} \ \sim\ Gamma\left(x+a, \frac{b}{nb+1}\right) \tag{3}$$

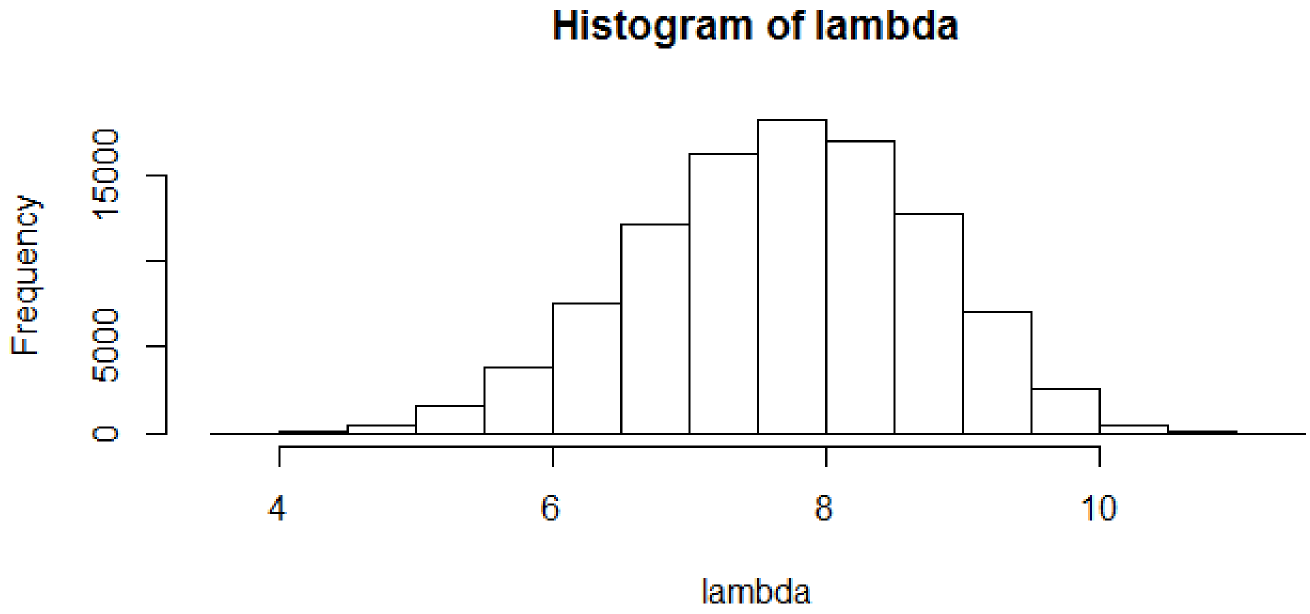
$$b|x,\lambda \ \propto\ b^{-(a+\alpha)-1} e^{-(\lambda+\beta)/b} \ \sim\ InvGamma(a+\alpha, \lambda+\beta) \tag{4}$$

Choose a = 2, alpha = 2, beta = 2, n = 100. Consider the sampling distribution to be Poisson(10).

R Code

```
par(mfrow=c(2,1))
lambda=NULL
b=NULL
a=2
b[1]=1
alpha=2
beta=2
n=100
x=sum(rpois(n,10))

m=50000
for(i in 2:m){
lambda[i]=rgamma(1,shape=x+a,scale=(b[i-1]/(1+(n*b[i-1]))))
b[i] = 1/(rgamma(1,shape=alpha+a,scale=(beta + lambda[i])))
}
hist(lambda)[2001:m]
summary(lambda)

m=100000
for(i in 2:m){
lambda[i]=rgamma(1,shape=x+a,scale=(b[i-1]/(1+(n*b[i-1]))))
b[i] = 1/(rgamma(1,shape=alpha+a,scale=(beta + lambda[i])))
}
hist(lambda)[2001:m]
summary(lambda)

OUTPUT
------
m = 50000

summary(lambda)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.759   6.985   7.730   7.692   8.443  10.940

m = 100000

 summary(lambda)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.563   6.982   7.727   7.687   8.435  11.060
```


Histogram of lambda
Frequency
0
4000
8000
4
6
8
10
lambda


Histogram of lambda
Frequency
0
5000
15000
4
6
8
10
lambda

Biostat 602 Winter 2017

Lecture Set 15

Asymptotic Evaluation of Estimators

Reading: CB 10.1

Consistency

Asymptotic Evaluation of Point Estimators

When the sample size n approaches infinity, the behaviors of an estimator are unknown as its *asymptotic* properties.

Definition - Consistency: Let $W_n = W_n(X_1, \cdots, X_n) = W_n(\mathbf{X})$ be a sequence of estimators for $\tau(\theta)$. We say W_n is consistent for estimating $\tau(\theta)$ if $W_n \xrightarrow{P} \tau(\theta)$ under P_θ for every $\theta \in \Omega$.

$W_n \xrightarrow{P} \tau(\theta)$ (converges in probability to $\tau(\theta)$) means that, given any $\epsilon > 0$.

$$\lim_{n\to\infty} \Pr(|W_n - \tau(\theta)| \geq \epsilon) = 0$$

$$\lim_{n\to\infty} \Pr(|W_n - \tau(\theta)| < \epsilon) = 1$$

Consistency implies that the probability of W_n being close to $\tau(\theta)$ approaches to 1 as n goes to ∞.

Tools for proving consistency

- Use definition (complicated)
- Chebychev's Inequality

$$\begin{aligned}
\Pr(|W_n - \tau(\theta)| \geq \epsilon) &= \Pr((W_n - \tau(\theta))^2 \geq \epsilon^2) \\
&\leq \frac{\mathrm{E}[W_n - \tau(\theta)]^2}{\epsilon^2} \\
&= \frac{\mathrm{MSE}(W_n)}{\epsilon^2} = \frac{\mathrm{Bias}^2(W_n) + \mathrm{Var}(W_n)}{\epsilon^2}
\end{aligned}$$

Need to show that both Bias(W_n) and Var(W_n) converge to zero

Theorem 10.1.3: If W_n is a sequence of estimators of $\tau(\theta)$ satisfying

- $\lim_{n->\infty} \text{Bias}(W_n) = 0$.
- $\lim_{n->\infty} \text{Var}(W_n) = 0$.

for all θ, then W_n is consistent for $\tau(\theta)$

Consistency of $\overline{X}$

Theorem 5.5.2 - Weak Law of Large Numbers: Let $X_1, \cdots, X_n$ be iid random variables with $\text{E}(X) = \mu$ and $\text{Var}(X) = \sigma^2 < \infty$. Then $\overline{X}$ converges in probability to μ, i.e. $\overline{X} \xrightarrow{P} \mu$.

Consistent sequence of estimators

Theorem 10.1.5: Let W_n is a consistent sequence of estimators of $\tau(\theta)$. Let a_n, b_n be sequences of constants satisfying

1. $\lim_{n\to\infty} a_n = 1$
2. $\lim_{n\to\infty} b_n = 0$.

Then $U_n = a_n W_n + b_n$ is also a consistent sequence of estimators of $\tau(\theta)$.

Continuous Mapping Theorem - Theorem 5.5.4: If W_n is consistent for θ ($W_n \xrightarrow{P} \theta$) and g is a continuous function, then $g(W_n)$ is consistent for $g(\theta)$ ($g(W_n) \xrightarrow{P} g(\theta)$).

Example 1: $X_1, \cdots, X_n$ are iid samples from a distribution with mean μ and variance $\sigma^2 < \infty$.

1. Show that $\overline{X}$ is consistent for μ.
2. Show that $\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$ is consistent for σ^2.
3. Show that $\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ is consistent for σ^2.

Proof: By law of large numbers, $\overline{X}$ is consistent for μ.

Also

- $\text{Bias}(\overline{X}) = \text{E}(\overline{X}) - \mu = \mu - \mu = 0$.
- $\text{Var}(\overline{X}) = \text{Var}\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^n \text{Var}(X_i) = \sigma^2/n$.
- $\lim_{n\to\infty} \text{Var}(\overline{X}) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$.

By Theorem 10.1.3. $\overline{X}$ is consistent for μ.

Solution - consistency for σ^2

$$\begin{aligned}
\frac{\sum(X_i - \overline{X})^2}{n} &= \frac{\sum(X_i^2 + \overline{X}^2 - 2X_i\overline{X})}{n} \\
&= \frac{\sum X_i^2 + n\overline{X}^2 - 2\overline{X}\sum_{i=1}^n X_i}{n} = \frac{\sum X_i^2}{n} - \overline{X}^2
\end{aligned}$$

By law of large numbers,

$$\frac{1}{n}\sum X_i^2 \xrightarrow{P} \text{E}X^2 = \mu^2 + \sigma^2$$

Note that $\overline{X}^2$ is a function of $\overline{X}$. Define $g(x) = x^2$, which is a continuous function. Then $\overline{X}^2 = g(\overline{X})$ is consistent for μ^2. Therefore,

$$\frac{\sum(X_i - \overline{X})^2}{n} = \frac{\sum X_i^2}{n} - \overline{X}^2 \xrightarrow{P} (\mu^2 + \sigma^2) - \mu^2 = \sigma^2$$

So, $\sum(X_i - \overline{X})^2/n$ is consistent for σ^2.

Define $S_n^2 = \frac{1}{n-1}\sum(X_i - \overline{X})^2$, and $(S_n^*)^2 = \frac{1}{n}\sum(X_i - \overline{X})^2$.

$$S_n^2 = \frac{1}{n-1}\sum(X_i - \overline{X})^2 = (S_n^*)^2 \cdot \frac{n}{n-1}$$

Because $(S_n^*)^2$ was shown to be consistent for σ^2 previously, and $a_n = \frac{n}{n-1} \to 1$ as $n \to \infty$, by Theorem 10.1.5, S_n^2 is also consistent for σ^2.

Example 2 - Exponential Suppose $X_1, \cdots, X_n$ *iid* Exponential(β).

1. Propose a consistent estimator of the population median.
2. Propose a consistent estimator of $\Pr(X \leq c)$ where c is constant.

Consistent estimator for the median

First, we need to express the median in terms of the parameter β.

$$\begin{aligned}
\int_0^m \frac{1}{\beta} e^{-x/\beta} dx &= \frac{1}{2} \\
-e^{-x/\beta}\Big|_0^m &= \frac{1}{2} \\
1 - e^{-m/\beta} &= \frac{1}{2} \\
\text{median} &= m = \beta \log 2
\end{aligned}$$

By law of large numbers, $\overline{X}$ is consistent for $\mathrm{E}(X) = \beta$. Applying continuous mapping Theorem to $g(x) = x \log 2$, $g(\overline{X}) = \overline{X} \log 2$ is consistent for $g(\beta) = \beta \log 2$ (median).

Consistent estimator of $\Pr(X \leq c)$

$$\begin{aligned} \Pr(X \leq c) &= \int_0^c \frac{1}{\beta} e^{-x/\beta} dx \\ &= 1 - e^{-c/\beta} \end{aligned}$$

As $\overline{X}$ is consistent for β, $1 - e^{-c/\beta}$ is continuous function of β. By continuous mapping Theorem, $g(\overline{X}) = 1 - e^{-c/\overline{X}}$ is consistent for $\Pr(X \leq c) = 1 - e^{-c/\beta} = g(\beta)$

Consistent estimator of $\Pr(X \leq c)$ **- Alternative Method**

Define $Y_i = I(X_i \leq c)$. Then Y_i *iid* Bernoulli(p) where $p = \Pr(X \leq c)$.

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq c)$$

is consistent for p by the Weak Law of Large Numbers.

Theorem 10.1.6 - Consistency of MLEs

Suppose X_i *iid* $f(x|\theta)$. Let $\hat{\theta}$ be the MLE of θ, and $\tau(\theta)$ be a continuous function of θ. Then under "regularity conditions" on $f(x|\theta)$, the MLE of $\tau(\theta)$ (i.e. $\tau(\hat{\theta})$) is consistent for $\tau(\theta)$. The regularity conditions, described in 10.6.2, include iid, identifiability, differentiability, parameter space containing open set.

Asymptotic Normality

Definition: Asymptotic Normality A statistic (or an estimator) $W_n(\mathbf{X})$ is *asymptotically normal* if

$$\sqrt{n}(W_n - \tau(\theta)) \xrightarrow{d} \mathcal{N}(0, \nu(\theta))$$

for all θ
where $\xrightarrow{d}$ stands for "converge in distribution"

- $\tau(\theta)$: "asymptotic mean"
- $\nu(\theta)$: "asymptotic variance"

We denote $W_n \sim \mathcal{AN}\left(\tau(\theta), \frac{\nu(\theta)}{n}\right)$.

A general definition of asymptotic normality is defined as
$k_n(T_n - \tau(\theta)) \xrightarrow{d} N(0, \nu(\theta))$. The definition above is when $k_n = \sqrt{n}$.

Given a statistic $W_n(\mathbf{X})$, for example $\overline{X}$, $s^2_{\mathbf{X}}$, $e^{-\overline{X}}$

$$\sqrt{n}(W_n - \tau(\theta)) \quad \xrightarrow{d} \quad \mathcal{N}(0, \nu(\theta)) \qquad \text{for all } \theta$$

$$\iff W_n \quad \sim \quad \mathcal{AN}\left(\tau(\theta), \frac{\nu(\theta)}{n}\right)$$

Tools to show asymptotic normality

1. Central Limit Theorem
2. Slutsky Theorem
3. Delta Method (Theorem 5.5.24)

Central Limit Theorem 5.5.14

Assume X_i *iid* $f(x|\theta)$ with finite mean $\mu(\theta)$ and variance $\sigma^2(\theta)$. Then

$$
\begin{aligned}
\overline{X} &\sim \mathcal{AN}\left(\mu(\theta), \frac{\sigma^2(\theta)}{n}\right) \\
\Leftrightarrow \sqrt{n}\left(\overline{X} - \mu(\theta)\right) &\xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta))
\end{aligned}
$$

Theorem 5.5.17 - Slutsky's Theorem If $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{P} a$, where a is a constant,

1. $Y_n \cdot X_n \xrightarrow{d} aX$
2. $X_n + Y_n \xrightarrow{d} X + a$

Theorem 5.5.24 - Delta Method Assume $W_n \sim \mathcal{AN}\left(\theta, \frac{\nu(\theta)}{n}\right)$. If a function g satisfies $g'(\theta) \neq 0$, then

$$
g(W_n) \sim \mathcal{AN}\left(g(\theta), [g'(\theta)]^2 \frac{\nu(\theta)}{n}\right)
$$

Example 3 - Estimator of $\Pr(X \leq c)$ Define $Y_i = I(X_i \leq c)$. Then Y_i *iid* Bernoulli(p) where $p = \Pr(X \leq c)$.

$$
\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq c)
$$

is consistent for p. Therefore,

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} I(X_i \leq c) &\sim \mathcal{AN}\left(\mathrm{E}(Y), \frac{\mathrm{Var}(Y)}{n}\right) \\
&= \mathcal{AN}\left(p, \frac{p(1-p)}{n}\right)
\end{aligned}
$$

Example 4: Let $X_1, \cdots, X_n$ be iid samples with finite mean μ and variance σ^2. Define

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

By Central Limit Theorem,

$$\begin{aligned} \overline{X} \quad &\sim \quad \mathcal{AN}\left(\mu, \frac{\sigma^2}{n}\right) \\ \Leftrightarrow \sqrt{n}(\overline{X}-\mu) \quad &\xrightarrow{d} \quad \mathcal{N}(0,\sigma^2) \\ \Leftrightarrow \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \quad &\xrightarrow{d} \quad \mathcal{N}(0,1) \end{aligned}$$

$$\frac{\sqrt{n}(\overline{X}-\mu)}{S_n} \quad = \quad \frac{\sigma}{S_n}\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}$$

We showed previously $S_n^2 \xrightarrow{P} \sigma^2 \Rightarrow S_n \xrightarrow{P} \sigma \Rightarrow \sigma/S_n \xrightarrow{P} 1$. Therefore, By Slutsky's Theorem $\frac{\sqrt{n}(\overline{X}-\mu)}{S_n} \xrightarrow{d} \mathcal{N}(0,1)$.

Example 5: Delta Method $X_1, \cdots, X_n$ *iid* Bernoulli(p) where $p \neq \frac{1}{2}$, we want to know the asymptotic distribution of $\overline{X}(1-\overline{X})$. By central limit Theorem,

$$\begin{aligned} \frac{\sqrt{n}(\overline{X}-p)}{\sqrt{p(1-p)}} \quad &\xrightarrow{d} \quad \mathcal{N}(0,1) \\ \Leftrightarrow \overline{X} \quad &\sim \quad \mathcal{AN}\left(p, \frac{p(1-p)}{n}\right) \end{aligned}$$

Define $g(y) = y(1-y)$, then $\overline{X}(1-\overline{X}) = g(\overline{X})$.

$$g'(y) = (y - y^2)' = 1 - 2y$$

By Delta Method,

$$\begin{aligned} g(\overline{X}) = \overline{X}(1-\overline{X}) &\sim \mathcal{AN}\left(g(p), [g'(p)]^2 \frac{p(1-p)}{n}\right) \\ &= \mathcal{AN}\left(p(1-p), (1-2p)^2 \frac{p(1-p)}{n}\right) \end{aligned}$$

Example 6 - Normal MLE Let $X_1, \cdots, X_n$ *iid* $\mathcal{N}(\mu, \sigma^2) \qquad \mu \neq 0$. Find the asymptotic distribution of MLE of μ^2.

Solution:

1. It can be easily shown that MLE of μ is $\overline{X}$.
2. By the invariance property of MLE, MLE of μ^2 is $\overline{X}^2$.
3. By central limit theorem, we know that

$$\overline{X} \sim \mathcal{AN}\left(\mu, \frac{\sigma^2}{n}\right)$$

4. Define $g(y) = y^2$, and apply Delta Method.

$$\begin{aligned} g'(y) &= 2y \\ \overline{X}^2 &\sim \mathcal{AN}\left(g(\mu), [g'(\mu)]^2 \frac{\sigma^2}{n}\right) \\ &\sim \mathcal{AN}\left(\mu^2, (2\mu)^2 \frac{\sigma^2}{n}\right) \end{aligned}$$

Asymptotic Efficiency

If both estimators are consistent and asymptotic normal, we can compare their asymptotic variance.

Definition 10.1.16 : Asymptotic Relative Efficiency If two estimators W_n and V_n satisfy

$$\sqrt{n}[W_n - \tau(\theta)] \xrightarrow{d} \mathcal{N}(0, \sigma_W^2)$$

$$\sqrt{n}[V_n - \tau(\theta)] \xrightarrow{d} \mathcal{N}(0, \sigma_V^2)$$

The asymptotic relative efficiency (ARE) of V_n with respect to W_n is

$$ARE(V_n, W_n) = \frac{\sigma_W^2}{\sigma_V^2}$$

If $ARE(V_n, W_n) \geq 1$ for every $\theta \in \Omega$, then V_n is asymptotically more efficient than W_n.

Example 7: Let X_i *iid Poisson*(λ). Consider estimating

$$\Pr(X = 0) = e^{-\lambda}$$

Our estimators are

$$\begin{aligned} W_n &= \frac{1}{n}\sum_{i=1}^{n} I(X_i = 0) \\ V_n &= e^{-\overline{X}} \end{aligned}$$

Determine which one is more asymptotically efficient estimator.

Solution: Asymptotic Distribution of V_n

$V_n(\mathbf{X}) = e^{-\overline{X}}$. By CLT

$$\overline{X} \sim \mathcal{AN}(\mathrm{E}X, \mathrm{Var}X/n) \sim \mathcal{AN}(\lambda, \lambda/n)$$

Define $g(y) = e^{-y}$, then $V_n = g(\overline{X})$ and $g'(y) = -e^{-y}$. By Delta Method

$$\begin{aligned} V_n = e^{-\overline{X}} &\sim \mathcal{AN}\left(g(\lambda), [g'(\lambda)]^2\frac{\lambda}{n}\right) \\ &\sim \mathcal{AN}\left(e^{-\lambda}, e^{-2\lambda}\frac{\lambda}{n}\right) \end{aligned}$$

Asymptotic Distribution of W_n

Define $Z_i = I(X_i = 0)$

$$\begin{aligned} W_n &= \frac{1}{n}\sum_{i=1}^{n} I(X_i = 0) = \overline{Z} \\ Z_i &\sim \mathit{Bernoulli}(\mathrm{E}(Z)) \\ \mathrm{E}(Z) &= \Pr(X = 0) = e^{-\lambda} \\ \mathrm{Var}(Z) &= e^{-\lambda}(1 - e^{-\lambda}) \end{aligned}$$

By CLT,

$$\begin{aligned} W_n = \overline{Z} &\sim \mathcal{AN}\left(\mathrm{E}(Z), \mathrm{Var}(Z)/n\right) \\ &\sim \mathcal{AN}\left(e^{-\lambda}, \frac{e^{-\lambda}(1 - e^{-\lambda})}{n}\right) \end{aligned}$$

Calculating ARE

$$
\begin{aligned}
ARE(W_n, V_n) &= \frac{e^{-2\lambda}\lambda/n}{e^{-\lambda}(1-e^{-\lambda})/n} \\
&= \frac{\lambda}{e^{\lambda}(1-e^{-\lambda})} \\
&= \frac{\lambda}{e^{\lambda}-1} \\
&= \frac{\lambda}{\left(1+\lambda+\frac{\lambda^2}{2}+\frac{\lambda^3}{3!}+\cdots\right)-1} \\
&\leq 1 \qquad (\forall \lambda \geq 0)
\end{aligned}
$$

Therefore $W_n = \frac{1}{n}\sum I(X_i = 0)$ is less efficient than V_n (MLE), and ARE attains maximum at $\lambda = 0$.

Asymptotic Efficiency

Asymptotic Efficiency for iid samples

A sequence of estimators W_n is asymptotically efficient for $\tau(\theta)$ if for all $\theta \in \Omega$,

$$
\begin{aligned}
\sqrt{n}(W_n - \tau(\theta)) &\xrightarrow{d} \mathcal{N}\left(0, \frac{[\tau'(\theta)]^2}{I(\theta)}\right) \\
\Longleftrightarrow W_n &\sim \mathcal{AN}\left(\tau(\theta), \frac{[\tau'(\theta)]^2}{nI(\theta)}\right) \\
I(\theta) &= \mathrm{E}\left[\left\{\frac{\partial}{\partial\theta}\log f(X|\theta)\right\}^2\right] \\
&= -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right] \text{ (with Lemma 7.3.11)}
\end{aligned}
$$

Note: $\frac{[\tau'(\theta)]^2}{nI(\theta)}$ is the C-R bound for unbiased estimators of $\tau(\theta)$.

Theorem 10.1.12 Let $X_1, \cdots, X_n$ be iid samples from $f(x|\theta)$. Let $\hat{\theta}$ denote the MLE of θ. Under regularity conditions, $\hat{\theta}$ is consistent and asymptotically normal for θ, i.e.

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I(\theta)}\right) \text{ for every } \theta \in \Omega$$

And if $\tau(\theta)$ is continuous and differentiable in θ, then

$$\sqrt{n}(\tau(\hat{\theta}) - \tau(\theta)) \xrightarrow{d} \mathcal{N}\left(0, \frac{[\tau'(\theta)]^2}{I(\theta)}\right)$$

- The regularity condition includes the ones in 10.1.6, plus finite three-times differentiabilble log-likelihood functions (See 10.6.2)
- Note that the asymptotic variance of $\tau(\hat{\theta})$ is Cramer-Rao lower bound for unbiased estimators of $\tau(\theta)$.

Example 8: Suppose $X_1, \cdots, X_n$ *iid* Exponential(β) and let $\tau(\beta) = \Pr(X \leq c)$ for a known constant c. Consider the two consistent estimators

1. Is $W(\mathbf{X}) = 1 - e^{-\frac{c}{\overline{X}}}$ asymptotically efficient for $\tau(\beta)$?
2. Is $U(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq c)$ asymptotically efficient for $\tau(\beta)$?

Solution 1: By invariance property, $W(\mathbf{X}) = \tau(\hat{\beta})$ is MLE. So it is asymptotically efficient.

Solution 2: Let $Y_i = I(X_i \leq c) \sim Bernoulli(1 - e^{-\frac{c}{\beta}})$.

$$\begin{aligned}
\mathrm{E}Y_i &= 1 - e^{-\frac{c}{\beta}} \\
\mathrm{Var}Y_i &= e^{-\frac{c}{\beta}}(1 - e^{-\frac{c}{\beta}}) \\
U(\mathbf{X}) &= \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq c) = \overline{Y} \sim \mathcal{AN}\left(1 - e^{-\frac{c}{\beta}}, \frac{e^{-\frac{c}{\beta}}(1 - e^{-\frac{c}{\beta}})}{n}\right) \\
I(\beta) &= \frac{1}{\beta^2} \\
\frac{[\tau'(\beta)]^2}{nI(\beta)} &= \frac{c^2 e^{-\frac{2c}{\beta}}}{n\beta^2} \leq \frac{e^{-\frac{c}{\beta}}(1 - e^{-\frac{c}{\beta}})}{n}
\end{aligned}$$

So $U(\mathbf{X})$ is not asymptotically efficient.

Summary - Consistency and Efficiency

Consistency

- $W_n(\mathbf{X}) \longrightarrow \mathrm{P}\tau(\theta)$.
- Use W.L.L.N., Theorem 10.1.3, Continuous Mapping Theorem.

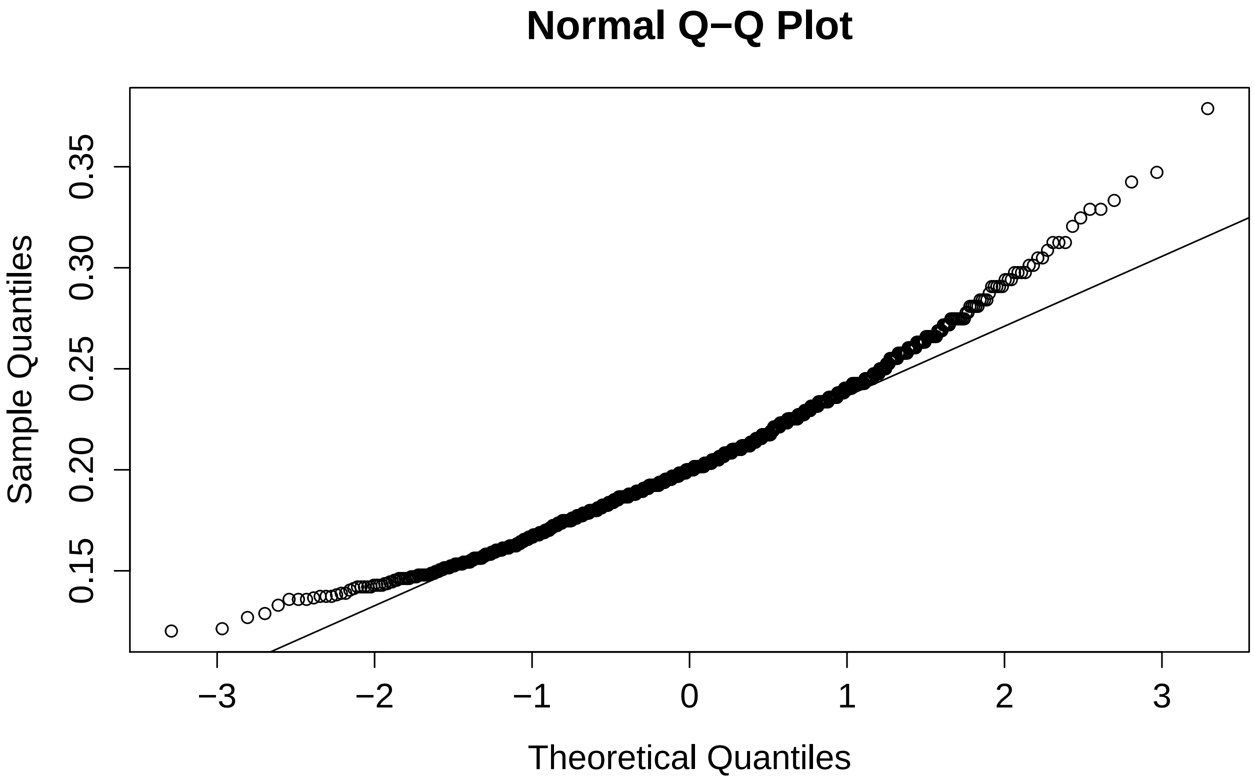
Asymptotic Normality and Efficiency

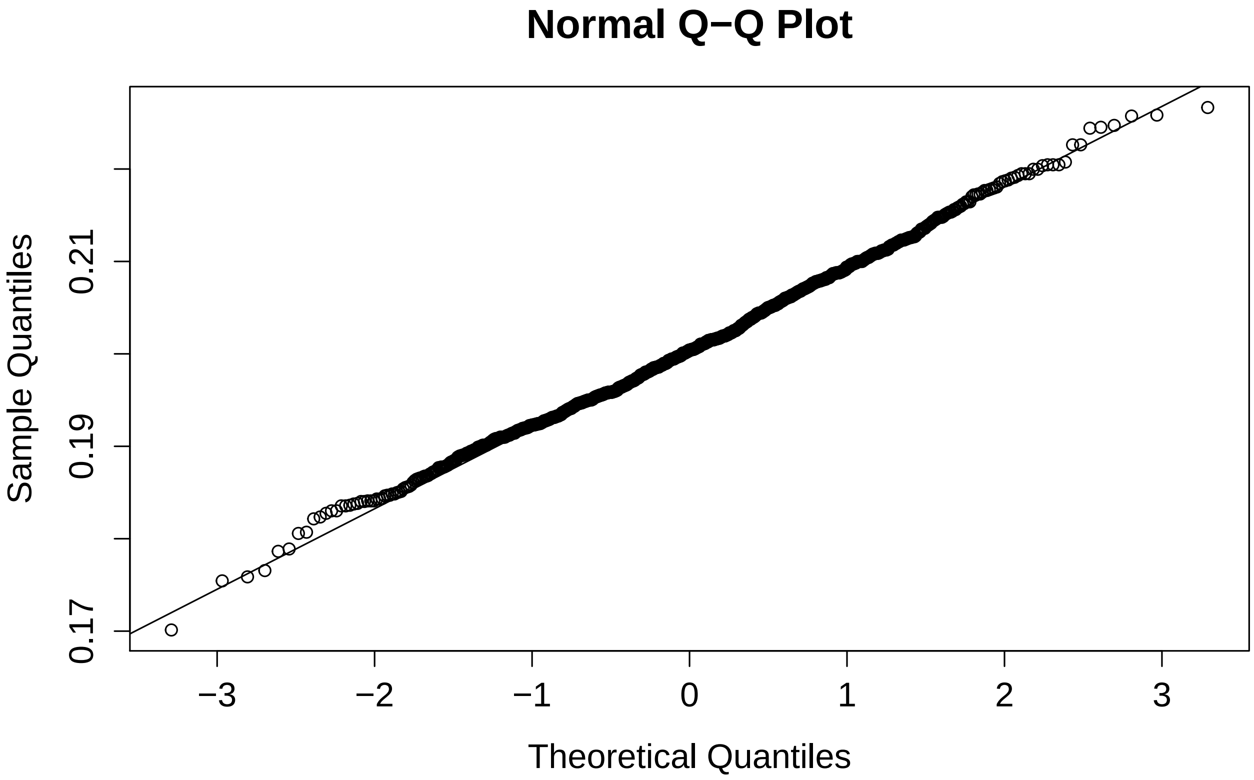
- Asymptotic behavior of mean (consistency) and variance (efficiency).
- Useful tools are C.L.T, Slutsky's Theorem, and Delta Method.
- Asymptotic Relative Efficiency (ARE) allows to compare the efficiency between two consistent and asymptotically normal estimators.
- Asymptotically efficient : asymptotic variance approaches CR-bound.
- MLE is always asymptotically efficient (under mild condition).

Example 9: Let $X_1, X_2, \ldots, X_n$ be a i.i.d. random sample from Negative Binomial (r, p). The following R code shows how close the MLEs resemble their large-sample properties. For this example, we take $r = 5, p = 0.2$. We generate a random sample of size n from this random sample and calculate the MLE of p. We repeat this process k times.

```
n=5
r=5
p=0.2
k=1000
mle=NULL
for(i in 1:k){
  x=rnbinom(n,size=r,p)
  mle[i] = (n*r)/((n*r) + sum(x))
}
mean(mle) # mle is consistent for p = 0.2

[1] 0.2072064

n*(var(mle)) # CRLB = (1-p)*p^2/r = 0.0064

[1] 0.007368217
qqnorm(mle)
qqline(mle)
```

We then repeat the process with n = 100. mean(MLE) = .2006, asymptotic variance = .00731.


Normal Q−Q Plot
Sample Quantiles
0.15
0.20
0.25
0.30
0.35
Theoretical Quantiles
−3
−2
−1
0
1
2
3


Normal Q−Q Plot
Sample Quantiles
0.17
0.19
0.21
Theoretical Quantiles
−3
−2
−1
0
1
2
3

Central Limit Theorem for Sample Median

Result: Let $X_1, X_2, \ldots, X_n$ be a i.i.d. random sample from a pdf/pmf f that is differentiable. Define μ to be the *median* of the population, i.e. $P(X_i \leq \mu) = 1/2$. Let M_n be the sample median. Then, as $n \longrightarrow \infty$,

$$\sqrt{n}(M_n - \mu) \xrightarrow{d} \mathcal{N}(0, 1/\left[2f(\mu)\right]^2).$$

Example 10: Suppose f denotes the pdf of Cauchy with median θ, i.e.

$$f(x) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}, \qquad -\infty < x < \infty.$$

Thus, $f(\theta) = 1/\pi$. With M_n denoting the sample median based on a random sample from Cauchy,

$$\sqrt{n}(M_n - \mu) \xrightarrow{d} \mathcal{N}(0, \pi^2/4).$$

Example 11: Let $X_1, X_2, \ldots, X_n$ be a i.i.d. random sample from $\mathcal{N}(\mu, \sigma^2)$. Consider the sample mean $\overline{X}$ and the sample median M_n to be competing estimators for μ which is both the mean and the median. Note that $f(\mu) = 1/(\sqrt{2\pi}\sigma)$ and so

$$\sqrt{n}(M_n - \mu) \xrightarrow{d} \mathcal{N}(0, \pi\sigma^2/2).$$

Since $\sqrt{n}(\overline{X} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, hence

$$ARE(\overline{X}, M_n) = 2/\pi = 0.64.$$

Example 12: Let $X_1, X_2, \ldots, X_n$ be a i.i.d. random sample from $Gamma(3, \beta)$. Consider

$$T_1(\mathbf{X}) = \overline{X}, \quad \text{and} \quad T_2(\mathbf{X}) = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}}$$

to be the sample arithmetic mean and sample harmonic mean respectively.

(a) Show that $W = T_1/3$ and $V = T_2/2$ are both consistent estimators of β.

(b) Prove the asymptotic normality results for both W and V.

(c) Find the ARE of V with respect to W.

Note: If $X \sim Gamma(\alpha, \beta)$ then $Y = 1/X \sim Inverse\ Gamma(\alpha, \beta^{-1})$ with

$$E(Y) = \frac{1}{\beta(\alpha - 1)}, \quad \alpha > 1, \quad Var(Y) = \frac{1}{\beta^2(\alpha - 1)^2(\alpha - 2)}, \quad \alpha > 2.$$

Biostat 602 Winter 2017

Lecture Set 16

Hypothesis Testing

Reading: CB Chapter 8

Hypothesis Testing

A *hypothesis* is a statement about a population parameter

Two complementary statements about θ:

- Null hypothesis : $H_0 : \theta \in \Omega_0$
- Alternative hypothesis : $H_1 : \theta \in \Omega_0^c$

$\theta \in \Omega = \Omega \cup \Omega^c$.

Simple and composite hypothesis

Simple hypothesis

Both H_0 and H_1 consist of only one parameter value.

- $H_0 : \theta = \theta_0 \in \Omega_0$
- $H_1 : \theta = \theta_1 \in \Omega_0^c$

Composite hypothesis

One or both of H_0 and H_1 consist more than one parameter values.

- One-sided hypothesis: $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$.
- One-sided hypothesis: $H_0 : \theta \geq \theta_0$ vs $H_1 : \theta < \theta_0$.
- Two-sided hypothesis: $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$.

An Example of Hypothesis

$$X_1, \cdots, X_n \quad iid \quad \mathcal{N}(\theta, 1)$$

Let X_i denote the change in blood pressure after a treatment.

$$\begin{aligned} H_0 &: \theta = 0 \qquad \text{(no effect)} \\ H_1 &: \theta \neq 0 \qquad \text{(some effect)} \end{aligned}$$

Two-sided composite hypothesis.

Another Example of Hypothesis

- Let θ denote the proportion of defective items from a machine.
- One may want the proportion to be less than a specified maximum acceptable proportion θ_0.
- We want to test whether the products produced by the machine is acceptable.

$$\begin{aligned} H_0 &: \theta \leq \theta_0 \qquad \text{(acceptable)} \\ \\ H_1 &: \theta > \theta_0 \qquad \text{(unacceptable)} \end{aligned}$$

Hypothesis Testing Procedure

A hypothesis testing procedure is a rule that specifies:

1. For which sample points H_0 is accepted as true (the subset of the sample space for which H_0 is accepted is called the acceptable region).
2. For which sample points H_0 is rejected and H_1 is accepted as true (the subset of sample space for which H_0 is rejected is called the rejection region or critical region).

Rejection region (R) on a hypothesis is usually defined through a test statistic $W(\mathbf{X})$. For example,

$$\begin{aligned} R_1 &= \{\mathbf{x} : W(\mathbf{x}) > c, \mathbf{x} \in \mathcal{X}\} \\ R_2 &= \{\mathbf{x} : W(\mathbf{x}) \leq c, \mathbf{x} \in \mathcal{X}\} \end{aligned}$$

Example of hypothesis testing

X_1, X_2, X_3 i.i.d. *Bernoulli*(p). Consider hypothesis tests

$$\begin{aligned} H_0 &: p \leq 0.5 \\ H_1 &: p > 0.5 \end{aligned}$$

- Test 1 : Reject H_0 if $\mathbf{x} \in \{(1,1,1)\}$
  $\iff$ rejection region $= \{(1,1,1)\}$
  $\iff$ rejection region $= \{\mathbf{x} : \sum x_i > 2\}$
- Test 2 : Reject H_0 if $\mathbf{x} \in \{(1,1,0),(1,0,1),(0,1,1),(1,1,1)\}$
  $\iff$ rejection region $= \{(1,1,0),(1,0,1),(0,1,1),(1,1,1)\}$
  $\iff$ rejection region $= \{\mathbf{x} : \sum x_i > 1\}$

Example Let $X_1, \cdots, X_n$ be change in blood pressure after a treatment.

$$\begin{aligned} H_0 &: \theta = 0 \\ H_1 &: \theta \neq 0 \end{aligned}$$

An example rejection region $R = \left\{ \mathbf{x} : \frac{\overline{x}}{s_{\mathbf{x}}/\sqrt{n}} > 3 \right\}$.

Decision

		Accept H_0	Reject H_0
Truth	H_0	Correct Decision	Type I error
	H_1	Type II error	Correct Decision

Type I and Type II error

Type I error

If $\theta \in \Omega_0$ (if the null hypothesis is true), the probability of making a type I error is

$$\Pr(\mathbf{X} \in R | \theta)$$

Type II error

If $\theta \in \Omega_0^c$ (if the alternative hypothesis is true), the probability of making a type II error is

$$\Pr(\mathbf{X} \notin R | \theta) = 1 - \Pr(\mathbf{X} \in R | \theta)$$

Power function

Definition: The power function of a hypothesis test with rejection region R is the function of θ defined by

$$\beta(\theta) = \Pr(\mathbf{X} \in R | \theta) = \Pr(\text{reject } H_0 | \theta)$$

If $\theta \in \Omega_0^c$ (alternative is true), the probability of rejecting H_0 is called the power of test for this particular value of θ.

- Probability of type I error $= \beta(\theta)$ if $\theta \in \Omega_0$.
- Probability of type II error $= 1 - \beta(\theta)$ if $\theta \in \Omega_0^c$.

An ideal test should have power function satisfying $\beta(\theta) = 0$ for all $\theta \in \Omega_0$, $\beta(\theta) = 1$ for all $\theta \in \Omega_0^c$, which is typically not possible in practice.

Example 1: Let $X_1, X_2, \cdots, X_n$ i.i.d. *Bernoulli*(θ) where $n = 5$.

$$\begin{aligned} H_0 &: \theta \leq 0.5 \\ H_1 &: \theta > 0.5 \end{aligned}$$

Test 1 rejects H_0 if and only if all ”success” are observed. i.e.

$$\begin{aligned} R &= \{\mathbf{x} : \mathbf{x} = (1,1,1,1,1)\} \\ &= \{\mathbf{x} : \sum_{i=1}^{5} x_i = 5\} \end{aligned}$$

1. Compute the power function
2. What is the maximum probability of making type I error?
3. What is the probability of making type II error if $\theta = 2/3$?

Solution for Test 1

$$\begin{aligned} \beta(\theta) &= \Pr(\text{reject } H_0|\theta) = \Pr(\mathbf{X} \in R|\theta) \\ &= \Pr\left(\sum X_i = 5|\theta\right) \end{aligned}$$

Because $\sum X_i \sim Binomial(5, \theta)$, $\beta(\theta) = \theta^5$.

Maximum type I error
When $\theta \in \Omega_0 = (0, 0.5]$, the power function $\beta(\theta)$ is Type I error.

$$\max_{\theta \in (0,0.5]} \beta(\theta) = \max_{\theta \in (0,0.5]} \theta^5 = 0.5^5 = 1/32 \approx 0.031$$

Type II error when $\theta = 2/3$

$$1 - \beta(\theta)|_{\theta=\frac{2}{3}} = 1 - \theta^5\Big|_{\theta=\frac{2}{3}} = 1 - (2/3)^5 = 211/243 \approx 0.868$$

Another Example

Example 2: $X_1, X_2, \cdots, X_n$ i.i.d. $Bernoulli(\theta)$ where $n = 5$.

$$H_0 \;:\; \theta \leq 0.5$$

$$H_1 \;:\; \theta > 0.5$$

Test 2 rejects H_0 if and only if 3 or more "success" are observed. i.e.

$$R \;=\; \{\mathbf{x} : \sum_{i=1}^{5} x_i \geq 3\}$$

1. Compute the power function
2. What is the maximum probability of making type I error?
3. What is the probability of making type II error if $\theta = 2/3$?

Solution for Test 2

Power function

$$\begin{aligned}\beta(\theta) &= \Pr(\sum X_i \geq 3|\theta) = \binom{5}{3}\theta^3(1-\theta)^2 + \binom{5}{4}\theta^4(1-\theta) + \binom{5}{5}\theta^5 \\ &= \theta^3(6\theta^2 - 15\theta + 10)\end{aligned}$$

Maximum type I error

We need to find the maximum of $\beta(\theta)$ for $\theta \in \Omega_0 = (0, 0.5]$

$$\beta'(\theta) = 30\theta^2(\theta-1)^2 > 0$$

$\beta(\theta)$ is increasing in $\theta \in (0, 1)$. Maximum type I error is $\beta(0.5) = 0.5$

Type II error when $\theta = 2/3$

$$1 - \beta(\theta)|_{\theta=\frac{2}{3}} = 1 - \theta^3(6\theta^2 - 15\theta + 10)\Big|_{\theta=\frac{2}{3}} \approx 0.21$$

Sizes and Levels of Tests

Size α test

A test with power function $\beta(\theta)$ is a size α test if

$$\sup_{\theta\in\Omega_0} \beta(\theta) = \alpha$$

In other words, the maximum probability of making a type I error is α.

Level α test

A test with power function $\beta(\theta)$ is a level α test if

$$\sup_{\theta \in \Omega_0} \beta(\theta) \leq \alpha$$

In other words, the maximum probability of making a type I error is equal or less than α.

Any size α test is also a level α test

Revisiting Previous Examples
Test 1

$$\sup_{\theta \in \Omega_0} \beta(\theta) = \sup_{\theta \in \Omega_0} \theta^5 = 0.5^5 = 0.03125$$

The size is 0.03125, and this is a level 0.05 test, or a level 0.1 test, but not a level 0.01 test.

Test 2

$$\sup_{\theta \in \Omega_0} \beta(\theta) = 0.5$$

The size is 0.5

Constructing a good test

1. Construct all the level α test.
2. Within this level of tests, we search for the test with Type II error probability as small as possible; equivalently, we want the test with the largest power if $\theta \in \Omega_0^c$.

Review on standard normal and t distribution

Quantile of standard normal distribution

Let $Z \sim \mathcal{N}(0,1)$ with pdf $f_Z(z)$ and cdf $F_Z(z)$. The α-th quantile z_α or $(1-\alpha)$-th quantile $z_{1-\alpha}$ of the standard distribution satisfy

$$\Pr(Z \geq z_\alpha) = \alpha \quad \text{or} \quad z_\alpha = F_Z^{-1}(1-\alpha)$$

$$\Pr(Z \leq z_{1-\alpha}) = \alpha \quad \text{or} \quad z_{1-\alpha} = F_Z^{-1}(\alpha)$$

$$z_{1-\alpha} \;=\; -z_\alpha$$

Quantile of t distribution

Let $T \sim t_{n-1}$ with pdf $f_{T,n-1}(t)$ and cdf $F_{T,n-1}(t)$. The α-th quantile $t_{n-1,\alpha}$ or $(1-\alpha)$-th quantile $t_{n-1,1-\alpha}$ of the standard distribution satisfy

$$\Pr(T \geq t_{n-1,\alpha}) = \alpha \quad \text{or} \quad t_{n-1\alpha} = F_{T,n-1}^{-1}(1-\alpha)$$

$$\Pr(T \leq t_{n-1,1-\alpha}) = \alpha \quad \text{or} \quad t_{n-1,1-\alpha} = F_{T,n-1}^{-1}(\alpha)$$

$$t_{n-1,1-\alpha} \;=\; -t_{n-1,\alpha}$$

Likelihood Ratio Tests (LRT)

Definition Let $L(\theta|\mathbf{x})$ be the likelihood function of θ. The likelihood ratio test statistic for testing $H_0 : \theta \in \Omega_0$ vs. $H_1 : \theta \in \Omega_0^c$ is

$$\lambda(\mathbf{x}) = \frac{\sup_{\theta\in\Omega_0} L(\theta|\mathbf{x})}{\sup_{\theta\in\Omega} L(\theta|\mathbf{x})} = \frac{L(\hat{\theta}_0|\mathbf{x})}{L(\hat{\theta}|\mathbf{x})}$$

where $\hat{\theta}$ is the MLE of θ over $\theta \in \Omega$, and $\hat{\theta}_0$ is the MLE of θ over $\theta \in \Omega_0$ (restricted MLE).

The *likelihood ratio test* is a test that rejects H_0 if and only if $\lambda(\mathbf{x}) \leq c$ where $0 \leq c \leq 1$.

Example of LRT

Example 3: Consider $X_1, \cdots, X_n$ iid $\mathcal{N}(\theta, \sigma^2)$ where σ^2 is known.

$$H_0 : \theta \leq \theta_0$$

$$H_1 : \theta > \theta_0$$

Find the LRT test and its power function

Solution:

$$\begin{aligned} L(\theta|\mathbf{x}) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i-\theta)^2}{2\sigma^2}\right] \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{\sum_{i=1}^n (x_i-\theta)^2}{2\sigma^2}\right] \end{aligned}$$

We need to find MLE of θ over $\Omega = (-\infty, \infty)$ and $\Omega_0 = (-\infty, \theta_0]$.

MLE of θ over $\Omega = (-\infty, \infty)$

To maximize $L(\theta|\mathbf{x})$, we need to maximize $\exp\left[-\frac{\sum_{i=1}^n (x_i-\theta)^2}{2\sigma^2}\right]$, or equivalently to minimize $\sum_{i=1}^n (x_i - \theta)^2$.

$$\begin{aligned}\sum_{i=1}^n (x_i-\theta)^2 &= \sum_{i=1}^n (x_i^2 + \theta^2 - 2\theta x_i) \\ &= n\theta^2 - 2\theta\sum_{i=1}^n x_i + \sum_{i=1}^n x_i^2\end{aligned}$$

The equation above is minimized when $\theta = \hat{\theta} = \frac{\sum_{i=1}^n x_i}{n} = \overline{x}$.

- $L(\theta|\mathbf{x})$ is maximized at $\theta = \frac{\sum_{i=1}^n x_i}{n} = \overline{x}$ if $\overline{x} \leq \theta_0$.
- However, if $\overline{x} \geq \theta_0$, $\overline{x}$ does not fall into a valid range of $\hat{\theta}_0$, and $\theta \leq \theta_0$, the likelihood function will be an increasing function. Therefore $\hat{\theta}_0 = \theta_0$.

To summarize,

$$\hat{\theta}_0 = \begin{cases} \overline{X} & \text{if } \overline{X} \leq \theta_0 \\ \theta_0 & \text{if } \overline{X} > \theta_0 \end{cases}$$

$$\begin{aligned}\lambda(\mathbf{x}) = \frac{L(\hat{\theta}_0|\mathbf{x})}{L(\hat{\theta}|\mathbf{x})} &= \begin{cases} 1 & \text{if } \overline{X} \leq \theta_0 \\ \frac{\exp\left[-\frac{\sum_{i=1}^n (x_i-\theta_0)^2}{2\sigma^2}\right]}{\exp\left[-\frac{\sum_{i=1}^n (x_i-\overline{x})^2}{2\sigma^2}\right]} & \text{if } \overline{X} > \theta_0 \end{cases} \\ &= \begin{cases} 1 & \text{if } \overline{X} \leq \theta_0 \\ \exp\left[-\frac{n(\overline{x}-\theta_0)^2}{2\sigma^2}\right] & \text{if } \overline{X} > \theta_0 \end{cases}\end{aligned}$$

Therefore, the likelihood test rejects the null hypothesis if and only if

$$\exp\left[-\frac{n(\overline{x}-\theta_0)^2}{2\sigma^2}\right] \leq c$$

and $\overline{x} \geq \theta_0$.

Specifying c

$$\begin{aligned}
\exp\left[-\frac{n(\overline{x}-\theta_0)^2}{2\sigma^2}\right] &\leq c \\
\Longleftrightarrow -\frac{n(\overline{x}-\theta_0)^2}{2\sigma^2} &\leq \log c \\
\Longleftrightarrow (\overline{x}-\theta_0)^2 &\geq -\frac{2\sigma^2\log c}{n} \\
\Longleftrightarrow \overline{x}-\theta_0 &\geq \sqrt{-\frac{2\sigma^2\log c}{n}} \qquad (\because \overline{x} > \theta_0)
\end{aligned}$$

So, LRT rejects H_0 if and only if

$$\begin{aligned}
\overline{x}-\theta_0 &\geq \sqrt{-\frac{2\sigma^2\log c}{n}} \\
\Longleftrightarrow \frac{\overline{x}-\theta_0}{\sigma/\sqrt{n}} &\geq \frac{\sqrt{-\frac{2\sigma^2\log c}{n}}}{\sigma/\sqrt{n}} = c^*
\end{aligned}$$

Therefore, the rejection region is

$$\left\{\mathbf{x} : \frac{\overline{x}-\theta_0}{\sigma/\sqrt{n}} \geq c^*\right\}$$

Power function

$$\begin{aligned}
\beta(\theta) &= \Pr(\text{reject } H_0) = \Pr\left(\frac{\overline{X}-\theta_0}{\sigma/\sqrt{n}} \geq c^*\right) \\
&= \Pr\left(\frac{\overline{X}-\theta+\theta-\theta_0}{\sigma/\sqrt{n}} \geq c^*\right) \\
&= \Pr\left(\frac{\overline{X}-\theta}{\sigma/\sqrt{n}} \geq \frac{\theta_0-\theta}{\sigma/\sqrt{n}} + c^*\right)
\end{aligned}$$

Since $X_1, \cdots, X_n$ i.i.d. $\mathcal{N}(\theta, \sigma^2)$, $\overline{X} \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{n}\right)$. Therefore,

$$\begin{aligned}
\frac{\overline{X} - \theta}{\sigma/\sqrt{n}} &\sim \mathcal{N}(0,1) \\
\Longrightarrow \beta(\theta) &= \Pr\left(Z \geq \frac{\theta_0 - \theta}{\sigma/\sqrt{n}} + c^*\right)
\end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$.

Making size α LRT

To make a size α test,

$$\begin{aligned}
\sup_{\theta \in \Omega_0} \beta(\theta) &= \alpha \\
\sup_{\theta \leq \theta_0} \Pr\left(Z \geq \frac{\theta_0 - \theta}{\sigma/\sqrt{n}} + c^*\right) &= \alpha \\
\Pr\left(Z \geq c^*\right) &= \alpha \\
c^* &= z_\alpha
\end{aligned}$$

Note that $\Pr\left(Z \geq \frac{\theta_0 - \theta}{\sigma/\sqrt{n}} + c^*\right)$ is maximized when θ is maximum (i.e. $\theta = \theta_0$).

Therefore, size α LRT test rejects H_0 if and only if $\frac{\overline{x} - \theta_0}{\sigma/\sqrt{n}} \geq z_\alpha$.

Another Example of LRT

Example 4: Let $X_1, \cdots, X_n$ i.i.d. from $f(x|\theta) = e^{-(x-\theta)}$ where $x \geq \theta$ and $-\infty < \theta < \infty$. Find a LRT testing the following one-sided hypothesis.

$$\begin{aligned} H_0 &: \theta \leq \theta_0 \\ H_1 &: \theta > \theta_0 \end{aligned}$$

Solution:

$$\begin{aligned} L(\theta|\mathbf{x}) &= \prod_{i=1}^{n} e^{-(x_i-\theta)} I(x_i \geq \theta) \\ &= e^{-\sum x_i + n\theta} I(\theta \leq x_{(1)}) \end{aligned}$$

The likelihood function is a increasing function of θ, bounded by $\theta \leq x_{(1)}$. Therefore, when $\theta \in \Omega = \mathbb{R}$, $L(\theta|\mathbf{x})$ is maximized when $\theta = \hat{\theta} = x_{(1)}$.

When $\theta \in \Omega_0^c$, the likelihood is still an increasing function, but bounded by $\theta \leq \min(x_{(1)}, \theta_0)$. Therefore, the likelihood is maximized when $\theta = \hat{\theta}_0 = \min(x_{(1)}, \theta_0)$. The likelihood ratio test statistic is

$$\begin{aligned} \lambda(\mathbf{x}) &= \begin{cases} \frac{e^{-\sum x_i + n\theta_0}}{e^{-\sum x_i + n x_{(1)}}} & \text{if } \theta_0 < x_{(1)} \\ 1 & \text{if } \theta_0 \geq x_{(1)} \end{cases} \\ &= \begin{cases} e^{n(\theta_0 - x_{(1)})} & \text{if } \theta_0 < x_{(1)} \\ 1 & \text{if } \theta_0 \geq x_{(1)} \end{cases} \end{aligned}$$

The LRT rejects H_0 if and only if

$$\begin{aligned} e^{n(\theta_0 - x_{(1)})} &\leq c \qquad (\text{and } \theta_0 < x_{(1)}) \\ \Longleftrightarrow \theta_0 - x_{(1)} &\leq \frac{\log c}{n} \Longleftrightarrow x_{(1)} \geq \theta_0 - \frac{\log c}{n} \end{aligned}$$

So, LRT reject H_0 is $x_{(1)} \geq \theta_0 - \frac{\log c}{n}$ and $x_{(1)} > \theta_0$. The power function is

$$\begin{aligned}\beta(\theta) &= \Pr\left(X_{(1)} \geq \theta_0 - \frac{\log c}{n} \wedge X_{(1)} > \theta_0\right) \\ &= \Pr\left(X_{(1)} \geq \theta_0 - \frac{\log c}{n}\right)\end{aligned}$$

To find size α test, we need to find c satisfying the condition

$$\sup_{\theta \leq \theta_0} \beta(\theta) = \alpha$$

Constructing size α test

$$\begin{aligned}\beta(\theta) &= \Pr\left(X_{(1)} \geq \theta_0 - \frac{\log c}{n}\right) = \prod_{i=1}^{n} \Pr\left(X_i \geq \theta_0 - \frac{\log c}{n}\right) \\ &= \prod_{i=1}^{n} \Pr\left(X_i - \theta \geq \theta_0 - \theta - \frac{\log c}{n}\right) \\ &= \prod_{i=1}^{n} \exp\left[-\theta_0 + \theta + \frac{\log c}{n}\right] = \left[\exp\left(-\theta_0 + \theta + \frac{\log c}{n}\right)\right]^n\end{aligned}$$

which is increasing in θ. Hence

$$\sup_{\theta \leq \theta_0} \beta(\theta) = \left[\exp\left(\frac{\log c}{n}\right)\right]^n = \alpha$$

Therefore, $\frac{\log c}{n} = \frac{1}{n}\log\alpha$, and the rejection region of the size α test is

$$R = \left\{\mathbf{X} : X_{(1)} \geq \theta_0 - \tfrac{1}{n}\log\alpha\right\}$$

Biostat 602 Winter 2017

Lecture Set 17

Hypothesis Testing
Likelihood Ratio Test

Reading: CB 8.2

Likelihood Ratio Tests (LRT)

Definition Let $L(\theta|\mathbf{x})$ be the likelihood function of θ. The likelihood ratio test statistic for testing $H_0 : \theta \in \Omega_0$ vs. $H_1 : \theta \in \Omega_0^c$ is

$$\lambda(\mathbf{x}) \;=\; \frac{\sup_{\theta\in\Omega_0} L(\theta|\mathbf{x})}{\sup_{\theta\in\Omega} L(\theta|\mathbf{x})} = \frac{L(\hat{\theta}_0|\mathbf{x})}{L(\hat{\theta}|\mathbf{x})}$$

where $\hat{\theta}$ is the MLE of θ over $\theta \in \Omega$, and $\hat{\theta}_0$ is the MLE of θ over $\theta \in \Omega_0$ (restricted MLE).

The *likelihood ratio test* is a test that rejects H_0 if and only if $\lambda(\mathbf{x}) \leq c$ where $0 \leq c \leq 1$.

c is obtained from the size condition of the test, namely

$$\sup_{\theta\in\Omega_0} \beta(\theta) = \alpha$$

where $\beta(\theta) = \Pr(\mathbf{X} \in R|\theta) = \Pr(\text{reject } H_0|\theta)$ is the power function of the test.

LRT based on sufficient statistics

Theorem 8.2.4: If $T(\mathbf{X})$ is a sufficient statistic for θ, $\lambda^*(t)$ is the LRT statistic based on T, and $\lambda(\mathbf{x})$ is the LRT statistic based on $\mathbf{x}$ then

$$\lambda^*[T(\mathbf{x})] = \lambda(\mathbf{x})$$

for every $\mathbf{x}$ in the sample space.

Proof: By Factorization Theorem, the joint pdf of $\mathbf{x}$ can be written as

$$f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$$

and we can choose $g(t|\theta)$ to be the pdf or pmf of $T(\mathbf{x})$. Then, the LRT statistic based on $T(\mathbf{X})$ is defined as

$$\lambda^*(t) = \frac{\sup_{\theta\in\Omega_0} L(\theta|T(\mathbf{x}) = t)}{\sup_{\theta\in\Omega} L(\theta|T(\mathbf{x}) = t)} = \frac{\sup_{\theta\in\Omega_0} g(t|\theta)}{\sup_{\theta\in\Omega} g(t|\theta)}$$

LRT statistic based on $\mathbf{X}$ is

$$\begin{aligned}\lambda(\mathbf{x}) &= \frac{\sup_{\theta\in\Omega_0} L(\theta|\mathbf{x})}{\sup_{\theta\in\Omega} L(\theta|\mathbf{x})} \\ &= \frac{\sup_{\theta\in\Omega_0} f(\mathbf{x}|\theta)}{\sup_{\theta\in\Omega} f(\mathbf{x}|\theta)} \\ &= \frac{\sup_{\theta\in\Omega_0} g(T(\mathbf{x})|\theta)h(\mathbf{x})}{\sup_{\theta\in\Omega} g(T(\mathbf{x})|\theta)h(\mathbf{x})} \\ &= \frac{\sup_{\theta\in\Omega_0} g(T(\mathbf{x})|\theta)}{\sup_{\theta\in\Omega} g(T(\mathbf{x})|\theta)} = \lambda^*(T(\mathbf{x}))\end{aligned}$$

The simplified expression of $\lambda(\mathbf{x})$ should depend on $\mathbf{x}$ only through $T(\mathbf{x})$, where $T(\mathbf{x})$ is a sufficient statistic for θ.

Example 1: Consider $X_1, \cdots, X_n$ i.i.d. $\mathcal{N}(\theta, \sigma^2)$ where σ^2 is known.

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0 \end{aligned}$$

Find a size α LRT.

Solution - Using sufficient statistics: Note that in this case, $T(\mathbf{X}) = \overline{X}$ is a sufficient statistic for θ.

$$\begin{aligned} T &\sim \mathcal{N}\left(\theta, \frac{\sigma^2}{n}\right) \\ \lambda(t) &= \frac{\sup_{\theta \in \Omega_0} L(\theta|t)}{\sup_{\theta \in \Omega} L(\theta|t)} = \frac{\sqrt{\frac{1}{2\pi\sigma^2/n}}\ \exp\left[-\frac{(t-\theta_0)^2}{2\sigma^2/n}\right]}{\sup_{\theta \in \Omega} \sqrt{\frac{1}{2\pi\sigma^2/n}}\ \exp\left[-\frac{(t-\theta)^2}{2\sigma^2/n}\right]} \end{aligned}$$

The numerator is fixed, and MLE in the denominator is $\hat{\theta} = t$. Therefore the LRT statistic is

$$\lambda(t) = \exp\left[-\frac{n(t-\theta_0)^2}{2\sigma^2}\right]$$

LRT rejects H_0 if and only if

$$\lambda(t) = \exp\left[-\frac{n(t-\theta_0)^2}{2\sigma^2}\right] \ \leq \ c$$

$$\Longrightarrow \left|\frac{t-\theta_0}{\sigma/\sqrt{n}}\right| \geq \sqrt{-2\log c} = c^*$$

Note that

$$\begin{aligned} T = \overline{X} &\sim \mathcal{N}\left(\theta, \frac{\sigma^2}{n}\right) \\ \frac{T-\theta_0}{\sigma/\sqrt{n}} &\sim \mathcal{N}(0,1) \end{aligned}$$

A size α test satisfies

$$\begin{aligned} \sup_{\theta\in\Omega_0} \Pr\left(\left|\frac{T-\theta}{\sigma/\sqrt{n}}\right| \geq c^*\right) &= \alpha \\ \Pr\left(\left|\frac{T-\theta_0}{\sigma/\sqrt{n}}\right| \geq c^*\right) &= \alpha \\ \Pr\left(|Z| \geq c^*\right) &= \alpha \\ \Pr(Z \geq c^*) + \Pr(Z \leq -c^*) &= \alpha \\ |Z| = \left|\frac{T-\theta}{\sigma/\sqrt{n}}\right| \geq z_{\alpha/2} & \end{aligned}$$

LRT with nuisance parameters

Example 2: Let $X_1, \cdots, X_n$ be i.i.d $\mathcal{N}(\theta, \sigma^2)$ where both θ and σ^2 are unknown. Obtain a LRT for testing $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$.

1. Specify Ω and Ω_0
2. Find size α LRT.

Solution - Ω and Ω_0

$$\begin{aligned}\Omega &= \{(\theta, \sigma^2) : \theta \in \mathbb{R}, \sigma^2 > 0\} \\ \Omega_0 &= \{(\theta, \sigma^2) : \theta \leq \theta_0, \sigma^2 > 0\}\end{aligned}$$

Size α LRT

$$\lambda(\mathbf{x}) = \frac{\sup_{\{(\theta,\sigma^2):\theta\leq\theta_0,\sigma^2>0\}} L(\theta, \sigma^2|\mathbf{x})}{\sup_{\{(\theta,\sigma^2):\theta\in\mathbb{R},\sigma^2>0\}} L(\theta, \sigma^2|\mathbf{x})}$$

For the denominator, the MLE of θ and σ^2 are

$$\begin{cases} \hat{\theta} = \overline{X} \\ \hat{\sigma}^2 = \frac{\sum(X_i - \overline{X})^2}{n} = \frac{n-1}{n} s^2_{\mathbf{X}} \end{cases}$$

For numerator, we need to maximize $L(\theta, \sigma^2|\mathbf{x})$ over the region $\theta \leq \theta_0$ and $\sigma^2 > 0$.

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

Maximizing Numerator

Step 1: fix σ^2, likelihood is maximized when $\sum_{i=1}^n (x_i - \theta)^2$ is minimized over $\theta \leq \theta_0$.

$$\hat{\theta}_0 = \begin{cases} \overline{x} & \text{if } \overline{x} \leq \theta_0 \\ \theta_0 & \text{if } \overline{x} > \theta_0 \end{cases}$$

Step 2: Now, we need to maximize likelihood (or log-likelihood) with respect to σ^2 and we substitute $\hat{\theta}_0$ for θ.

$$\begin{aligned} l(\hat{\theta}, \sigma^2|\mathbf{x}) &= -\frac{n}{2}\left(\log 2\pi + \log \sigma^2\right) - \frac{\sum (x_i - \hat{\theta}_0)^2}{2\sigma^2} \\ \frac{\partial \log l}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{\sum (x_i - \hat{\theta}_0)^2}{2(\sigma^2)^2} = 0 \\ \hat{\sigma}_0^2 &= \frac{\sum_{i=1}^n (x_i - \hat{\theta}_0)^2}{n} \end{aligned}$$

Combining the results together

$$\lambda(\mathbf{x}) = \begin{cases} 1 & \text{if } \overline{x} \leq \theta_0 \\ \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{n/2} & \text{if } \overline{x} > \theta_0 \end{cases}$$

Constructing LRT

LRT test rejects H_0 if and only if $\overline{x} > \theta_0$ and

$$\begin{aligned}
\left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{n/2} &\leq c \\
\left(\frac{\sum(x_i - \overline{x})^2/n}{\sum(x_i - \theta_0)^2/n}\right)^{n/2} &\leq c \\
\frac{\sum(x_i - \overline{x})^2}{\sum(x_i - \theta_0)^2} &\leq c^* \\
\frac{\sum(x_i - \overline{x})^2}{\sum(x_i - \overline{x})^2 + n(\overline{x} - \theta_0)^2} &\leq c^* \\
\frac{1}{1 + \frac{n(\overline{x} - \theta_0)^2}{\sum(x_i - \overline{x})^2}} &\leq c^*
\end{aligned}$$

$$\begin{aligned}
\frac{n(\overline{x} - \theta_0)^2}{\sum(x_i - \overline{x})^2} &\geq c^{**} \\
\frac{\overline{x} - \theta_0}{s_{\mathbf{X}}/\sqrt{n}} &\geq c^{***}
\end{aligned}$$

LRT test rejects H_0 if

$$\frac{\overline{x} - \theta_0}{s_{\mathbf{X}}/\sqrt{n}} \geq c^{***}$$

The next step is to specify c^{***} to get size α test (can you figure out?).

Unbiased Test

Definition: If a test always satisfies

$$\Pr(\text{reject } H_0 \text{ when } H_0 \text{ is false }) \geq \Pr(\text{reject } H_0 \text{ when } H_0 \text{ is true })$$

Then the test is said to be unbiased.

Alternative Definition: Recall that $\beta(\theta) = \Pr(\text{reject } H_0)$. A test is unbiased if

$$\beta(\theta') \geq \beta(\theta)$$

for every $\theta' \in \Omega_0^c$ and $\theta \in \Omega_0$.

Example 3: Let $X_1, \cdots, X_n$ be i.i.d. $\mathcal{N}(\theta, \sigma^2)$ where σ^2 is known, testing $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$. LRT test rejects H_0 if

$$\frac{\overline{x} - \theta_0}{\sigma/\sqrt{n}} > c.$$

$$\begin{aligned}
\beta(\theta) &= \Pr\left(\frac{\overline{X} - \theta_0}{\sigma/\sqrt{n}} > c\right) \\
&= \Pr\left(\frac{\overline{X} - \theta + \theta - \theta_0}{\sigma/\sqrt{n}} > c\right) \\
&= \Pr\left(\frac{\overline{X} - \theta}{\sigma/\sqrt{n}} + \frac{\theta - \theta_0}{\sigma/\sqrt{n}} > c\right) \\
&= \Pr\left(\frac{\overline{X} - \theta}{\sigma/\sqrt{n}} > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right)
\end{aligned}$$

Note that $X_i \sim \mathcal{N}(\theta, \sigma^2)$, $\overline{X} \sim \mathcal{N}(\theta, \sigma^2/n)$, and $\frac{\overline{X}-\theta}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$. Therefore, for $Z \sim \mathcal{N}(0,1)$

$$\beta(\theta) \ = \ \Pr\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right)$$

Because the power function is increasing function in θ,

$$\beta(\theta') \geq \beta(\theta)$$

always holds when $\theta \leq \theta_0 < \theta'$. Therefore the LRTs are unbiased.

Question: Can the same test be biased when hypotheses change?

Example 4: Same framework as before.

- New hypotheses : $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$.
- Same test : $R = \left\{\frac{\overline{x}-\theta_0}{\sigma/\sqrt{n}} > c\right\}$.

Testing unbiasedness

The power function $\beta(\theta)$ is still an increasing function. Therefore, if $\theta_+ > \theta_0 > \theta_-$, then

$$\beta(\theta_+) > \beta(\theta_0) > \beta(\theta_-)$$

where both $\beta(\theta_+)$ and $\beta(\theta_-)$ are power but $\beta(\theta_0)$ is Type I error.

Hence, power can be smaller than the Type I error when $\theta < \theta_0$, so the test is biased.

Uniformly Most Powerful Test (UMP)

Definition: Let $\mathcal{C}$ be a class of tests between $H_0 : \theta \in \Omega$ vs $H_1 : \theta \in \Omega_0^c$. A test in $\mathcal{C}$, with power function $\beta(\theta)$ is *uniformly most powerful (UMP) test* in class $\mathcal{C}$ if $\beta(\theta) \geq \beta'(\theta)$ for every $\theta \in \Omega_0^c$ and every $\beta'(\theta)$, which is a power function of another test in $\mathcal{C}$.

UMP level α test

Consider $\mathcal{C}$ to be the set of all the level α test. The UMP test in this class is called a UMP level α test.

UMP level α test has the smallest type II error probability for any $\theta \in \Omega_0^c$ in this class.

- A UMP test is "uniform" in the sense that it is most powerful for every $\theta \in \Omega_0^c$.
- For simple hypothesis such as $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$, UMP level α test always exists.

Neyman-Pearson Lemma

Theorem 8.3.12 - Neyman-Pearson Lemma: Consider testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ where the pdf or pmf corresponding to θ_i is $f(\mathbf{x}|\theta_i)$, $i = 0, 1$, using a test with rejection region R that satisfies

$$\begin{aligned} \mathbf{x} \in R \qquad & \text{if } f(\mathbf{x}|\theta_1) > kf(\mathbf{x}|\theta_0) \\ \mathbf{x} \in R^c \qquad & \text{if } f(\mathbf{x}|\theta_1) < kf(\mathbf{x}|\theta_0) \end{aligned} \qquad (8.3.1)$$

for some $k \geq 0$ and

$$\alpha = \Pr(\mathbf{X} \in R|\theta_0). \qquad (8.3.2)$$

Then,

- (Sufficiency) Any test that satisfies 8.3.1 and 8.3.2 is a UMP level α test
- (Necessity) If there exist a test satisfying 8.3.1 and 8.3.2 with $k > 0$, then every UMP level α test is a size α test (satisfies 8.3.2), and every UMP level α test satisfies 8.3.1 except perhaps on a set A satisfying $\Pr(\mathbf{X} \in A|\theta_0) = \Pr(\mathbf{X} \in A|\theta_1) = 0$.

Example 5: Let $X \sim \mathit{Binomial}(2, \theta)$, and consider testing $H_0 : \theta = \theta_0 = 1/2$ vs. $H_1 : \theta = \theta_1 = 3/4$.

Calculating the ratios of the pmfs given,

$$\frac{f(0|\theta_1)}{f(0|\theta_0)} = \frac{1}{4}, \qquad \frac{f(1|\theta_1)}{f(1|\theta_0)} = \frac{3}{4}, \qquad \frac{f(2|\theta_1)}{f(2|\theta_0)} = \frac{9}{4}$$

- Suppose that $k < 1/4$, then the rejection region $R = \{0, 1, 2\}$, and UMP level α test always rejects H_0. Therefore

$$\alpha = \Pr(\text{reject } H_0|\theta = \theta_0 = 1/2) = 1.$$

- Suppose that $1/4 < k < 3/4$, then $R = \{1, 2\}$, and UMP level α test rejects H_0 if $x = 1$ or $x = 2$.

$$\alpha = \Pr(\text{reject } H_0|\theta = \frac{1}{2}) = \Pr(x \neq 0|\theta = 1/2) = \frac{3}{4}$$

- Suppose that $3/4 < k < 9/4$, then UMP level α test rejects H_0 if $x = 2$

$$\alpha = \Pr(\text{reject} H_0|\theta = 1/2) = \Pr(x = 2|\theta = 1/2) = \frac{1}{4}$$

- If $k > 9/4$ the UMP level α test will always not reject H_0, and $\alpha = 0$

Example 6 – Normal Distribution: $X_i \sim \mathcal{N}(\theta, \sigma^2)$ where σ^2 is known. Consider testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ where $\theta_1 > \theta_0$.

$$\begin{aligned}
f(\mathbf{x}|\theta) &= \prod_{i=1}^{n} \left[\frac{1}{2\pi\sigma^2} \exp\left\{ -\frac{(x_i - \theta)^2}{2\sigma^2} \right\} \right] \\
\frac{f(\mathbf{x}|\theta_1)}{f(\mathbf{x}|\theta_0)} &= \frac{\exp\left\{ -\frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{2\sigma^2} \right\}}{\exp\left\{ -\frac{\sum_{i=1}^{n}(x_i-\theta_0)^2}{2\sigma^2} \right\}} \\
&= \exp\left[-\frac{\sum_{i=1}^{n}(x_i - \theta_1)^2}{2\sigma^2} + \frac{\sum_{i=1}^{n}(x_i - \theta_0)^2}{2\sigma^2} \right] \\
&= \exp\left[\frac{\sum_{i=1}^{n}(x_i - \theta_0)^2 - \sum_{i=1}^{n}(x_i - \theta_1)^2}{2\sigma^2} \right] \\
&= \exp\left[\frac{n(\theta_0^2 - \theta_1^2) + 2\sum_{i=1}^{n} x_i(\theta_1 - \theta_0)}{2\sigma^2} \right]
\end{aligned}$$

UMP level α test rejects H_0 if

$$\begin{aligned}
& \exp\left[\frac{n(\theta_0^2-\theta_1^2)+2\sum_{i=1}^n x_i(\theta_1-\theta_0)}{2\sigma^2}\right] > k \\
\iff \quad & \frac{n(\theta_0^2-\theta_1^2)+2\sum_{i=1}^n x_i(\theta_1-\theta_0)}{2\sigma^2} > \log k \\
\iff \quad & \sum_{i=1}^n x_i > k^*
\end{aligned}$$

$$\alpha \;=\; \Pr\left(\sum_{i=1}^n X_i > k^*|\theta_0\right)$$

Under H_0,

$$\begin{aligned}
X_i &\sim \mathcal{N}(\theta_0, \sigma^2) \\
\overline{X} &\sim \mathcal{N}(\theta_0, \sigma^2/n) \\
\frac{\overline{X}-\theta_0}{\sigma/\sqrt{n}} &\sim \mathcal{N}(0,1)
\end{aligned}$$

$$\begin{aligned}
\alpha &= \Pr\left(\sum_{i=1}^n X_i > k^*|\theta_0\right) \\
&= \Pr\left(Z > \frac{k^*/n-\theta_0}{\sigma/\sqrt{n}}\right)
\end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$.

$$\begin{aligned}
\frac{k^*/n-\theta_0}{\sigma/\sqrt{n}} &= z_\alpha \\
k^* &= n\left(\theta_0 + z_\alpha\frac{\sigma}{\sqrt{n}}\right)
\end{aligned}$$

Thus, the UMP level α test rejects H_0 if $\sum X_i > k^*$, or equivalently, rejects H_0 if $\overline{X} > k^*/n = \theta_0 + z_\alpha \sigma/\sqrt{n}$

Neyman-Pearson Lemma on Sufficient Statistics

Corollary 8.3.13: Consider $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$. Suppose $T(\mathbf{X})$ is a sufficient statistic for θ and $g(t|\theta_i)$ is the pdf or pmf of T. Corresponding $\theta_i, i \in \{0, 1\}$. Then any test based on T with rejection region S is a UMP level α test if it satisfies

$$\begin{aligned} t \in S & \qquad \text{if } g(t|\theta_1) > k \cdot g(t|\theta_0) \text{ and} \\ t \in S^c & \qquad \text{if } g(t|\theta_1) < k \cdot g(t|\theta_0) \end{aligned}$$

for some $k > 0$ and $\alpha = \Pr(T \in S|\theta_0)$

Proof: The rejection region in the sample space is

$$\begin{aligned} R &= \{\mathbf{x} : T(\mathbf{x}) = t \in S\} \\ &= \{\mathbf{x} : g(T(\mathbf{x})|\theta_1) > kg(T(\mathbf{x})|\theta_0)\} \end{aligned}$$

By Factorization Theorem:

$$\begin{aligned} f(\mathbf{x}|\theta_i) &= h(\mathbf{x})g(T(\mathbf{x})|\theta_i) \\ R &= \{\mathbf{x} : g(T(\mathbf{x})|\theta_1)h(x) > kg(T(\mathbf{x})|\theta_0)h(x)\} \\ &= \{\mathbf{x} : f(\mathbf{x}|\theta_1) > kf(\mathbf{x}|\theta_0)\} \end{aligned}$$

By Neyman-Pearson Lemma, this test is the UMP level α test, and

$$\alpha \;=\; \Pr(\mathbf{X} \in R) = \Pr(T(\mathbf{X}) \in S|\theta_0)$$

Revisiting the Example of Normal Distribution

$X_i \sim \mathcal{N}(\theta, \sigma^2)$ where σ^2 is known. Consider testing

$$H_0 : \theta = \theta_0 \;\; vs. \;\; H_1 : \theta = \theta_1, \quad where \;\; \theta_1 > \theta_0.$$

It is known that $T = \overline{X}$ is a sufficient statistic for θ, where $T \sim \mathcal{N}(\theta, \sigma^2/n)$.

$$\begin{aligned}
g(t|\theta_i) &= \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left\{-\frac{(t-\theta_i)^2}{2\sigma^2/n}\right\} \\
\frac{g(t|\theta_1)}{g(t|\theta_0)} &= \frac{\exp\left\{-\frac{(t-\theta_1)^2}{2\sigma^2/n}\right\}}{\exp\left\{-\frac{(t-\theta_0)^2}{2\sigma^2/n}\right\}} \\
&= \exp\left\{-\frac{1}{2\sigma^2/n}\left[(t-\theta_1)^2 - (t-\theta_0)^2\right]\right\} \\
&= \exp\left\{-\frac{1}{2\sigma^2/n}\left[\theta_1^2 - \theta_0^2 - 2t(\theta_1 - \theta_0)\right]\right\}
\end{aligned}$$

UMP level α test rejects H_0 if

$$\begin{aligned}
& \exp\left\{-\frac{1}{2\sigma^2/n}\left[\theta_1^2 - \theta_0^2 - 2t(\theta_1 - \theta_0)\right]\right\} > k \\
\Longleftrightarrow \;\; & \frac{1}{2\sigma^2/n}\left[-(\theta_1^2 - \theta_0^2) + 2t(\theta_1 - \theta_0)\right] > \log k \\
\Longleftrightarrow \;\; & \overline{X} = t > k^*
\end{aligned}$$

Under H_0, $\overline{X} \sim \mathcal{N}(\theta_0, \sigma^2/n)$. Now,

$$\begin{aligned}
\Pr(\text{reject } H_0|\theta_0) &= \alpha \\
\alpha &= \Pr(\overline{X} > k^*|\theta_0) \\
&= \Pr\left(\frac{\overline{X} - \theta_0}{\sigma/\sqrt{n}} > \frac{k^* - \theta_0}{\sigma/\sqrt{n}}\right) \\
&= \Pr\left(Z > \frac{k^* - \theta_0}{\sigma/\sqrt{n}}\right) \\
\frac{k^* - \theta_0}{\sigma/\sqrt{n}} &= z_\alpha \\
k^* &= \theta_0 + z_\alpha \frac{\sigma}{\sqrt{n}}
\end{aligned}$$

Monotone Likelihood Ratio (Karlin-Rubin)

Definition: A family of pdfs or pmfs $\{g(t|\theta) : \theta \in \Omega\}$ for a univariate random variable T with real-valued parameter θ have a monotone likelihood ratio if $\frac{g(t|\theta_2)}{g(t|\theta_1)}$ is an increasing (or non-decreasing) function of t for every $\theta_2 > \theta_1$ on $\{t : g(t|\theta_1) > 0 \text{ or } g(t|\theta_2) > 0\}$.

Note: we may define MLR using decreasing function of t. But all following theorems are stated according to the definition.

Examples of Monotone Likelihood Ratio

- Normal, Poisson, Binomial have the MLR Property (Exercise 8.25)
- If T is from an exponential family with the pdf or pmf

$$g(t|\theta) \;=\; h(t)c(\theta)\exp[w(\theta)\cdot t]$$

Then T has an MLR if $w(\theta)$ is a non-decreasing function of θ.

Proof: Suppose that $\theta_2 > \theta_1$.

$$\begin{aligned}\frac{g(t|\theta_2)}{g(t|\theta_1)} &= \frac{h(t)c(\theta_2)\exp[w(\theta_2)t]}{h(t)c(\theta_1)\exp[w(\theta_1)t]} \\ &= \frac{c(\theta_2)}{c(\theta_1)}\exp[\{w(\theta_2) - w(\theta_1)\}t]\end{aligned}$$

If $w(\theta)$ is a non-decreasing function of θ, then

1. $w(\theta_2) - w(\theta_1) \geq 0$ and
2. $\exp[\{w(\theta_2) - w(\theta_1)\}t]$ is an increasing function of t.

Therefore, $\frac{g(t|\theta_2)}{g(t|\theta_1)}$ is a non-decreasing function of t, and T has MLR if $w(\theta)$ is a non-decreasing function of θ.

Karlin-Rubin Theorem

Theorem 8.3.17: Suppose $T(\mathbf{X})$ is a sufficient statistic for θ and the family $\{g(t|\theta) : \theta \in \Omega\}$ is an MLR family. Then

1. For testing $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$, the UMP level α test is given by rejecting H_0 is and only if $T > t_0$ where $\alpha = \Pr(T > t_0|\theta_0)$.
2. For testing $H_0 : \theta \geq \theta_0$ vs $H_1 : \theta < \theta_0$, the UMP level α test is given by rejecting H_0 if and only if $T < t_0$ where $\alpha = \Pr(T < t_0|\theta_0)$.

Example 7: Let $X_i \sim \mathcal{N}(\theta, \sigma^2)$ where σ^2 is known, Find the UMP level α test for $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$.

Solution: Here $T(\mathbf{X}) = \overline{X}$ is a sufficient statistic for θ, and $T \sim \mathcal{N}(\theta, \sigma^2/n)$. Therefore

$$\begin{aligned}
g(t|\theta) &= \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left\{-\frac{(t-\theta)^2}{2\sigma^2/n}\right\} \\
&= \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left\{-\frac{t^2+\theta^2-2t\theta}{2\sigma^2/n}\right\} \\
&= \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left\{-\frac{t^2}{2\sigma^2/n}\right\} \exp\left\{-\frac{\theta^2}{2\sigma^2/n}\right\} \exp\left\{\frac{t\theta}{\sigma^2/n}\right\} \\
&= h(t)c(\theta)\exp[w(\theta)t]
\end{aligned}$$

where $w(\theta) = \frac{\theta}{\sigma^2/n}$ is an increasing function in θ. Therefore T has an MLR property.

Finding a UMP level α test

By Karlin-Rubin Theorem, UMP level α test rejects H_0 iff $T > t_0$

$$
\begin{aligned}
\alpha &= \Pr(T > t_0|\theta_0) \\
&= \Pr\left(\frac{T-\theta_0}{\sigma/\sqrt{n}} > \frac{t_0-\theta_0}{\sigma/\sqrt{n}}\middle|\,\theta_0\right) \\
&= \Pr\left(Z > \frac{t_0-\theta_0}{\sigma/\sqrt{n}}\right)
\end{aligned}
$$

where $Z \sim \mathcal{N}(0,1)$.

$$
\begin{aligned}
\frac{t_0-\theta_0}{\sigma/\sqrt{n}} &= z_\alpha \\
\Rightarrow t_0 &= \theta_0 + \frac{\sigma}{\sqrt{n}} z_\alpha
\end{aligned}
$$

UMP level α test rejects H_0 if $T = \overline{X} > \theta_0 + \frac{\sigma}{\sqrt{n}} z_\alpha$.

Testing $H_0 : \theta \geq \theta_0$ **vs.** $H_1 : \theta < \theta_0$

UMP level α test rejects H_0 if $T < t_0$ where

$$
\begin{aligned}
\alpha &= \Pr(T < t_0|\theta_0) = \Pr\left(\frac{T-\theta_0}{\sigma/\sqrt{n}} < \frac{t_0-\theta_0}{\sigma/\sqrt{n}}\middle|\,\theta_0\right) \\
&= \Pr\left(Z < \frac{t_0-\theta_0}{\sigma/\sqrt{n}}\right) \\
1-\alpha &= \Pr\left(Z \geq \frac{t_0-\theta_0}{\sigma/\sqrt{n}}\right) \\
\frac{t_0-\theta_0}{\sigma/\sqrt{n}} &= z_{1-\alpha} \\
t_0 &= \theta_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha} = \theta_0 - \frac{\sigma}{\sqrt{n}} z_\alpha
\end{aligned}
$$

Therefore, the test rejects H_0 if $T < t_0 = \theta_0 - \frac{\sigma}{\sqrt{n}} z_\alpha$

Example 8: Normal Example with Known Mean Let $X_i \sim \mathcal{N}(\mu_0, \sigma^2)$ where σ^2 is unknown and μ_0 is known. Find the UMP level α test for testing $H_0 : \sigma^2 \leq \sigma_0^2$ vs. $H_1 : \sigma^2 > \sigma_0^2$. Let $T = \sum_{i=1}^n (X_i - \mu_0)^2$ is sufficient for σ^2.

To check whether T has MLR property, we need to find $g(t|\sigma^2)$.

$$\begin{aligned}
\frac{X_i - \mu_0}{\sigma} &\sim \mathcal{N}(0,1) \\
\left(\frac{X_i - \mu_0}{\sigma}\right)^2 &\sim \chi_1^2 \\
Y &= T/\sigma^2 = \sum_{i=1}^n \left(\frac{X_i - \mu_0}{\sigma}\right)^2 \sim \chi_n^2 \\
f_Y(y) &= \frac{1}{\Gamma\left(\frac{n}{2}\right) 2^{n/2}} y^{\frac{n}{2}-1} e^{-\frac{y}{2}}
\end{aligned}$$

$$\begin{aligned}
f_T(t) &= \frac{1}{\Gamma\left(\frac{n}{2}\right) 2^{n/2}} \left(\frac{t}{\sigma^2}\right)^{\frac{n}{2}-1} e^{-\frac{t}{2\sigma^2}} \left|\frac{dy}{dt}\right| \\
&= \frac{1}{\Gamma\left(\frac{n}{2}\right) 2^{n/2}} \left(\frac{t}{\sigma^2}\right)^{\frac{n}{2}-1} e^{-\frac{t}{2\sigma^2}} \frac{1}{\sigma^2} \\
&= \frac{t^{\frac{n}{2}-1}}{\Gamma\left(\frac{n}{2}\right) 2^{n/2}} \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{t}{2\sigma^2}} \\
&= h(t)c(\sigma^2)\exp[w(\sigma^2)t]
\end{aligned}$$

where $w(\sigma^2) = -\frac{1}{2\sigma^2}$ is an increasing function in σ^2. Therefore,
$T = \sum_{i=1}^n (X_i - \mu_0)^2$ has the MLR property.

By Karlin-Rubin Theorem, UMP level α rejects s H_0 if and only if $T > t_0$ where t_0 is chosen such that $\alpha = \Pr(T > t_0 | \sigma_0^2)$. Note that $\frac{T}{\sigma^2} \sim \chi_n^2$. Hence

$$\begin{aligned}\Pr(T > t_0|\sigma_0^2) &= \Pr\left(\frac{T}{\sigma_0^2} > \frac{t_0}{\sigma_0^2}\right|\left.\sigma_0^2\right)\\ \frac{T}{\sigma_0^2} &\sim \chi_n^2\\ \Pr\left(\chi_n^2 > \frac{t_0}{\sigma_0^2}\right) &= \alpha\\ \frac{t_0}{\sigma_0^2} &= \chi_{n,\alpha}^2\\ t_0 &= \sigma_0^2\chi_{n,\alpha}^2\end{aligned}$$

where $\chi_{n,\alpha}^2$ satisfies $\int_{\chi_{n,\alpha}^2}^{\infty} f_{\chi_n^2}(x)dx = \alpha$.

Example 9: Let $X_1, \cdots, X_n \sim \mathcal{N}(\theta, \sigma^2)$ where σ^2 is known.
Consider testing $H_0 : \theta = \theta_0$ versus an alternative hypothesis.

1. When the alternative hypothesis is $H_1 : \theta_1 < \theta_0$, does UMP level α test exist? If yes, what is it?

2. When the alternative hypothesis is $H_1 : \theta_1 > \theta_0$, does UMP level α test exist? If yes, what is it?

3. When the alternative hypothesis is $H_1 : \theta_1 \neq \theta_0$, does UMP level α test exist? If yes, what is it?

4. Are the tests above unbiased?

$H_1 : \theta < \theta_0$

A level α test should satisfy $\Pr(\mathbf{X} \in R|\theta_0) \leq \alpha$.

As $\overline{X}$ is sufficient and its distribution has an MLR as shown in the previous example, by Karlin-Rubin Theorem, the rejection region of UMP level α test is

$$\overline{X} < -\frac{\sigma z_\alpha}{\sqrt{n}} + \theta_0$$

$H_1 : \theta > \theta_0$

As $\overline{X}$ is sufficient and its distribution has an MLR as shown in the previous example, by Karlin-Rubin Theorem, the rejection region of UMP level α test is

$$\overline{X} > \frac{\sigma z_\alpha}{\sqrt{n}} + \theta_0$$

$H_1 : \theta \neq \theta_0$

1. When $\theta < \theta_0$, $\beta_1(\theta) = \Pr(\mathbf{X} \in R_1) = \Pr\left(\overline{X} < -\frac{\sigma z_\alpha}{\sqrt{n}} + \theta_0\right)$ is the largest among level α tests.
2. If UMP level α test exists, the rejection region must be R_1 by the necessity condition of Neyman-Pearson Lemma.
3. When $\theta > \theta_0$, $\beta_2(\theta) = \Pr(\mathbf{X} \in R_2) = \Pr\left(\overline{X} > \frac{\sigma z_\alpha}{\sqrt{n}} + \theta_0\right)$ is the largest among level α tests.
4. Accordingly, $\beta_1(\theta)$ is not the power function of a UMP level α test.
5. Therefore, UMP level α test does not exist.

Are these tests unbiased?

Test based on $\overline{X} < -\frac{\sigma z_\alpha}{\sqrt{n}} + \theta_0$

1. When $\theta < \theta_0$, $\beta_1(\theta) > \beta_1(\theta_0)$.
2. When $\theta > \theta_0$, $\beta_1(\theta) < \beta_1(\theta_0)$.
3. Therefore, the test is not unbiased.

Test based on $\overline{X} > \frac{\sigma z_\alpha}{\sqrt{n}} + \theta_0$

1. When $\theta > \theta_0$, $\beta_2(\theta) > \beta_2(\theta_0)$.
2. When $\theta < \theta_0$, $\beta_2(\theta) < \beta_2(\theta_0)$.
3. Therefore, the test is not unbiased.

UMPU test

What is the optimal test for the two-sided test?

Consider a class of unbiased tests. Define a rejection region

$$|\overline{X} - \theta_0| > \frac{\sigma z_{\alpha/2}}{\sqrt{n}}$$

1. The test is unbiased. $\beta_3(\theta) > \beta_3(\theta_0)$ for all $\theta \neq \theta_0$.
2. The test is indeed the UMP test in the class of unbiased level α test.
3. This test is called a UMPU level α test.
4. Proving that the test is UMPU level α test is a little more complicated than UMP.

Example 8: Let $X_1, X_2, \ldots, X_n \sim Uniform(0, \theta)$. Consider testing

$$H_0 : \theta \leq \theta_0 \ vs. \ H_1 : \theta > \theta_0.$$

(a) Show that the family of $Uniform(0, \theta)$ has MLR in $X_{(n)}$.

(b) Find a size α UMP test for the above testing problem.

Example 9: Suppose that $X_1, \cdots, X_n$ are *i.i.d.* observations from Exponential(θ), and $Y_1, \cdots, Y_m$ are *i.i.d.* observations from Exponential(μ). Assume that $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_m$ are independent between them.

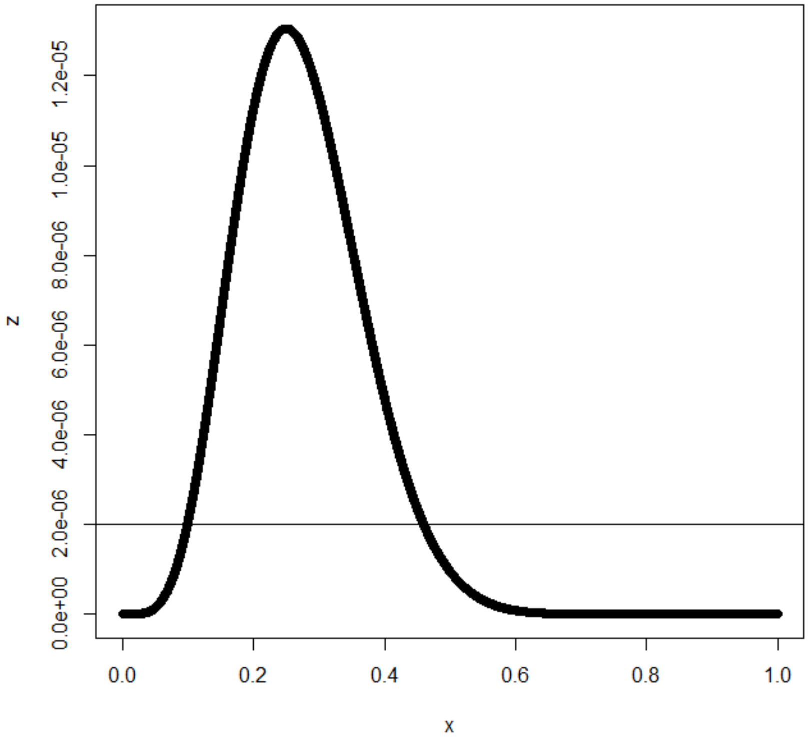
(a) Find the LRT statistic of $H_0 : \theta = \mu$ versus $H_1 : \theta \neq \mu$

(b) Show that the LRT from part (a) can be represented as a function of the following statistic T.

$$T = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i + \sum_{i=1}^{m} Y_i}$$

(Note that it is possible to construct a size α LRT using the fact that T follows a beta distribution under the null hypothesis.)

```
x=c(seq(0,1,by=0.0001))
z=(x^5)*((1-x)^15)
plot(x,z)
abline(h=.000002)
```



Biostat 602 Winter 2017

Lecture Set 18

Hypothesis Testing

Large-Sample Tests

Reading: CB 10.3

Large-sample Results for LRT

Question: Why do we need this?

We have a seen a few examples where the LRT rejection region is equivalent to a rejection region based on a statistic whose distribution is known, at least under H_0, so that the critical (a.k.a. rejection) region could be formed. However, these scenarios are quite limited to some standard distribution examples. In cases where such distributions are not available, one needs to take recourse some approximate method to construct a critical region. The large-sample result of LRT addresses this issue through a general asymptotic (valid for large n) result that applies to a large class of distributions.

Theorem 10.3.1: Consider testing $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$. Suppose $X_1, \cdots, X_n$ are iid samples from $f(x|\theta)$, and $\hat{\theta}$ is the MLE of θ, and $f(x|\theta)$ satisfies certain "regularity conditions" (e.g. see misc 10.6.2), then under H_0:

$$-2\log\lambda(\mathbf{X}) \xrightarrow{d} \chi^2_1$$

as $n \to \infty$.

Proof of Theorem 10.3.1: Note that

$$\begin{aligned}
\lambda(\mathbf{x}) &= \frac{\sup_{\theta\in\Omega_0} L(\theta|\mathbf{x})}{\sup_{\theta\in\Omega} L(\theta|\mathbf{x})} = \frac{L(\theta_0|\mathbf{x})}{L(\hat{\theta}|\mathbf{x})} \\
-2\log\lambda(\mathbf{x}) &= -2\log\left(\frac{L(\theta_0|\mathbf{x})}{L(\hat{\theta}|\mathbf{x})}\right) \\
&= -2\log L(\theta_0|\mathbf{x}) + 2\log L(\hat{\theta}|\mathbf{x}) \\
&= -2l(\theta_0|\mathbf{x}) + 2l(\hat{\theta}|\mathbf{x})
\end{aligned}$$

Expanding $l(\theta|\mathbf{x})$ around $\hat{\theta}$,

$$\begin{aligned}
l(\theta|\mathbf{x}) &= l(\hat{\theta}|\mathbf{x}) + l'(\hat{\theta}|\mathbf{x})(\theta - \hat{\theta}) + l''(\hat{\theta}|\mathbf{x})\frac{(\theta - \hat{\theta})^2}{2} + \cdots \\
l'(\hat{\theta}|\mathbf{x}) &= 0 \qquad \text{(assuming regularity condition)} \\
l(\theta_0|\mathbf{x}) &\approx l(\hat{\theta}|\mathbf{x}) + l''(\hat{\theta}|\mathbf{x})\frac{(\theta_0 - \hat{\theta})^2}{2} \\
-2\log\lambda(\mathbf{x}) &= -2l(\theta_0|\mathbf{x}) + 2l(\hat{\theta}|\mathbf{x}) \\
&\approx -(\theta_0 - \hat{\theta})^2 l''(\hat{\theta}|\mathbf{x})
\end{aligned}$$

Because $\hat{\theta}$ is MLE, under H_0,

$$\begin{aligned}
\hat{\theta} &\sim \mathcal{AN}\left(\theta_0, \frac{1}{I_n(\theta_0)}\right) \\
(\hat{\theta} - \theta_0)\sqrt{I_n(\theta_0)} &\xrightarrow{d} \mathcal{N}(0,1) \\
(\hat{\theta} - \theta_0)^2 I_n(\theta_0) &\xrightarrow{d} \chi^2_1
\end{aligned}$$

Therefore,

$$\begin{aligned}
-2\log\lambda(\mathbf{x}) &\approx -(\theta_0 - \hat{\theta})^2 l''(\hat{\theta}|\mathbf{x}) \\
&= (\hat{\theta} - \theta_0)^2 I_n(\theta_0) \frac{-\frac{1}{n} l''(\hat{\theta}|\mathbf{x})}{\frac{1}{n} I_n(\theta_0)}
\end{aligned}$$

$$\begin{aligned} -\frac{1}{n} l''(\hat{\theta}|\mathbf{x}) &= -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log f(x_i|\theta)\bigg|_{\theta=\hat{\theta}} \\ &\xrightarrow{P} -\mathrm{E}\left(\frac{\partial^2}{\partial\theta^2} \log f(x|\theta)\right)\bigg|_{\theta=\theta_0} = I(\theta_0) \quad \text{(by WLLN)} \end{aligned}$$

$$\frac{-\frac{1}{n} l''(\hat{\theta}|\mathbf{x})}{\frac{1}{n} I_n(\theta_0)} = \frac{-\frac{1}{n} l''(\hat{\theta}|\mathbf{x})}{I(\theta_0)} \xrightarrow{P} 1$$

By Slutsky's Theorem, under H_0

$$-(\hat{\theta}-\theta_0)^2 l''(\hat{\theta}|\mathbf{X}) \xrightarrow{d} \chi_1^2$$

$$-2\log\lambda(\mathbf{X}) \xrightarrow{d} \chi_1^2$$

The following result is the version of large-sample LRT result that generalizes the above to one with nuisance parameters.

Theorem 10.3.3: Let $X_1, \ldots, X_n$ be a random sample from a pdf or pmf $f(x|\theta)$. (Under the regulatory condition in 10.6.2), if $\theta \in \Omega_0$:

$$-2\log\lambda(\mathbf{X}) \xrightarrow{d} \chi_{q-p}^2$$

if the number of **free** parameters specified by $H_0 : \theta \in \Omega_0$ and $H_1 : \theta \in \Omega$ are p and q, respectively.

Example 1: Let $X_i \sim Poisson(\lambda)$. Consider testing $H_0 : \lambda = \lambda_0$ vs $H_1 : \lambda \neq \lambda_0$.

Using LRT,

$$\lambda(\mathbf{x}) \;=\; \frac{L(\lambda_0|\mathbf{x})}{\sup_\lambda L(\lambda|\mathbf{x})}$$

MLE of λ is $\hat{\lambda} = \overline{X} = \frac{1}{n}\sum X_i$.

$$\lambda(\mathbf{x}) \;=\; \frac{\prod_{i=1}^n \frac{e^{-\lambda_0}\lambda_0^{x_i}}{x_i!}}{\prod_{i=1}^n \frac{e^{-\overline{x}}\ \overline{x}^{x_i}}{x_i!}} \;=\; \frac{e^{-n\lambda_0}\lambda_0^{\sum x_i}}{e^{-n\overline{x}}\ \overline{x}^{\sum x_i}} \;=\; e^{-n(\lambda_0-\overline{x})}\left(\frac{\lambda_0}{\overline{x}}\right)^{\sum x_i}$$

LRT size α is to reject H_0 when $\lambda(\mathbf{x}) \leq c$.

$$\alpha \;=\; \Pr(\lambda(\mathbf{X}) \leq c|\lambda_0)$$

$$\begin{aligned}
-2\log\lambda(\mathbf{X}) \;&=\; -2\left[-n(\lambda_0 - \overline{X}) + \sum X_i(\log\lambda_0 - \log\overline{X})\right] \\
&=\; 2n\left(\lambda_0 - \overline{X} - \overline{X}\log\left(\frac{\lambda_0}{\overline{X}}\right)\right) \xrightarrow{d} \chi_1^2
\end{aligned}$$

under H_0, (by Theorem 10.3.1).

Therefore, asymptotic size α test is given by

$$\begin{aligned}
\Pr(\lambda(\mathbf{X}) \leq c|\lambda_0) \;&=\; \alpha \\
\Pr(-2\log\lambda(\mathbf{X}) \geq c^*|\lambda_0) \;&=\; \alpha \\
\Pr(\chi_1^2 \geq c^*) \;&\approx\; \alpha \\
c^* \;&=\; \chi_{1,\alpha}^2
\end{aligned}$$

which rejects H_0 if and only if $-2\log\lambda(\mathbf{x}) \geq \chi_{1,\alpha}^2$

Wald Test

Wald test relates point estimator of θ to hypothesis testing about θ.

Definition: Suppose W_n is an estimator of θ and $W_n \sim \mathcal{AN}(\theta, \sigma^2_{W_n})$. Then Wald test statistic is defined as

$$Z_n = \frac{W_n - \theta_0}{S_n}$$

where θ_0 is the value of θ under H_0 and S_n is a consistent estimator of σ_{W_n}

Two-sided Wald Test:

For testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$, Wald asymptotic level α test is to reject H_0 if and only if

$$|Z_n| > z_{\alpha/2}$$

One-sided Wald Test:

For testing $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$, Wald asymptotic level α test is to reject H_0 if and only if

$$Z_n > z_\alpha$$

Remarks:

- Different estimators of θ leads to different testing procedures.
- One choice of W_n is MLE and we may choose $S_n = \sqrt{\frac{1}{I_n(W_n)}}$ or $\sqrt{\frac{1}{I_n(\hat{\theta})}}$ (observed information number) when $\sigma^2_{W_n} = \frac{1}{I_n(\theta)}$.

Example 2: Suppose $X_i \sim Bernoulli(p)$, and consider testing $H_0 : p = p_0$ vs $H_1 : p \neq p_0$.

MLE of p is $\overline{X}$, which follows

$$\overline{X} \sim \mathcal{AN}\left(p, \frac{p(1-p)}{n}\right)$$

by the Central Limit Theorem. So the Wald test statistic is

$$Z_n \;=\; \frac{\overline{X} - p_0}{S_n}$$

where S_n is a consistent estimator of $\sqrt{\frac{p(1-p)}{n}}$, given by

$$S_n \;=\; \sqrt{\frac{\overline{X}(1-\overline{X})}{n}}$$

which is the MLE of $\sqrt{\frac{p(1-p)}{n}}$ by the invariance property of MLE.

The Wald statistic is

$$Z_n \;=\; \frac{\overline{X} - p_0}{\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}}$$

An asymptotic level α Wald test rejects H_0 if and only if

$$\left| \frac{\overline{X} - p_0}{\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}} \right| > z_{\alpha/2}$$

Score Test

Definition: Let $S(\theta) = \frac{\partial}{\partial\theta} \log L(\theta|\mathbf{x})$ be a score function. Then the variance of the score function is

$$\mathrm{Var}\left[S(\theta)\right] = \mathrm{E}\left[S^2(\theta)\right] = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2} \log L(\theta|\mathbf{x})\right] = I_n(\theta)$$

if the interchangeability condition holds. The test statistic for score test of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ is

$$Z_S = \frac{S(\theta_0)}{\sqrt{I_n(\theta_0)}}$$

If H_0 is true

- Z_S has mean 0 and variance 1.
- $Z_S \xrightarrow{d} \mathcal{N}(0,1)$.

Example 3: Let $X_i \sim \mathit{Bernoulli}(p)$. Consider testing $H_0 : p = p_0$ vs $H_1 : p \neq p_0$.

The likelihood and score function is

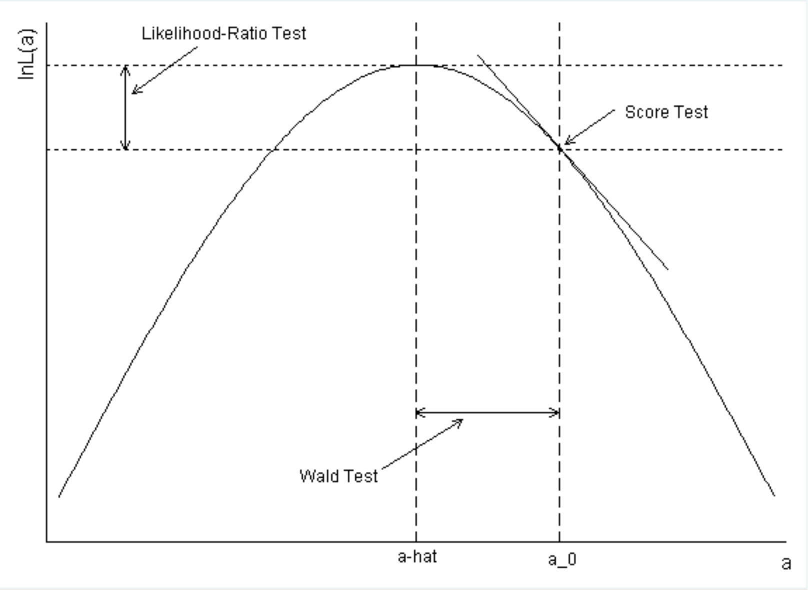
$$\begin{aligned} \log L(p|\mathbf{x}) &= \sum x_i \log p + (n - \sum x_i) \log(1-p) \\ S(p) &= \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p} = \frac{\overline{x} - p}{p(1-p)/n} \\ I(p) &= \frac{1}{p(1-p)} \end{aligned}$$

An asymptotic level α score test rejects H_0 if and only if

$$|Z_S| = \left|\frac{S(p_0)}{\sqrt{I_n(p_0)}}\right| = \left|\frac{\overline{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}\right| > z_{\alpha/2}$$

Comparison of the Three Tests

- The three tests are approximately equivalent in terms of asymptotic power.
- For likelihood functions that are not well-behaved, LRT has the best small-sample properties.



Example 4: Let $X_1, \ldots, X_n$ be *i.i.d.* random variables from Exponential$\,(\theta)$ distribution with pdf

$$f_X(x) = \frac{1}{\theta} \exp\left(-\frac{x}{\theta}\right) I(x > 0), \qquad \theta > 0$$

(a) Construct a large-sample (asymptotic) size α Wald test for testing $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$ for an arbitrary $\theta_0 > 0$.

(b) Consider a test for testing $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$ given by the following rejection region:

$$R = \left\{ \mathbf{X} : \frac{\sqrt{n}(\overline{X} - \theta_0)}{\theta_0} > z_\alpha \right\}$$

where z_α is upper α-quantile of $N(0, 1)$. Is the test defined above always more powerful than the Wald test defined in part (a)? Justify your answer.

Biostat 602 Winter 2017

Lecture Set 19

Interval Estimation

Reading: CB Chapter 9

Interval Estimation

In Chapter 7, we have focused on $\hat{\theta}(\mathbf{X})$, which is a *point estimator* of θ, i.e. a single value as a guess for the unknown parameter. Such an estimator does not incorporate any margin of error in the estimation. This motivates interval estimation which provides a set of values as possible values for the sample space and has the capability of incorporating the error in estimation.

Interval Estimator

Let $[L(\mathbf{X}), U(\mathbf{X})]$, where $L(\mathbf{X})$ and $U(\mathbf{X})$ are functions of sample $\mathbf{X}$ and $L(\mathbf{X}) \leq U(\mathbf{X})$. Based on the observed sample $\mathbf{x}$, we can make an inference that

$$\theta \in [L(\mathbf{X}), U(\mathbf{X})]$$

Then we call $[L(\mathbf{X}), U(\mathbf{X})]$ an interval estimator of θ.

Three types of intervals

- Two-sided interval $[L(\mathbf{X}), U(\mathbf{X})]$
- One-sided (with lower-bound) interval $[L(\mathbf{X}), \infty)$
- One-sided (with upper-bound) interval $(-\infty, U(\mathbf{X})]$

Example 1: Let $X_i \sim \mathcal{N}(\mu, 1)$. Define

1. A point estimator of $\mu : \overline{X}$

$$\Pr(\overline{X} = \mu) = 0$$

2. An interval estimator of $\mu : [\overline{X} - 1, \overline{X} + 1]$

$$\begin{aligned}
\Pr(\mu \in [\overline{X} - 1, \overline{X} + 1]) &= \Pr(\overline{X} - 1 \leq \mu \leq \overline{X} + 1) \\
&= \Pr(\mu - 1 \leq \overline{X} \leq \mu + 1) \\
&= \Pr(-\sqrt{n} \leq \sqrt{n}(\overline{X} - \mu) \leq \sqrt{n}) \\
&= \Pr(-\sqrt{n} \leq Z \leq \sqrt{n}) \longrightarrow 1
\end{aligned}$$

as $n \to \infty$, where $Z \sim \mathcal{N}(0, 1)$.

For specific values of n, there is a positive probability content. For example, with $n = 4$, the above probability equals

$$\Pr(-2 \leq Z \leq 2) = .9544.$$

Thus we have over a 95% chance of covering the unknown parameter μ with the interval estimator. In moving from a point to an interval estimator resulted in increased confidence in our estimation.

Some Definitions

Coverage Probability: Given an interval estimator $[L(\mathbf{X}), U(\mathbf{X})]$ of θ, its *coverage probability* is defined as

$$\Pr(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$$

In other words, it is the probability that the random interval $[L(\mathbf{X}), U(\mathbf{X})]$ covers the parameter θ.

Confidence Coefficient: The *confidence coefficient* associated with an interval estimator is defined as

$$\inf_{\theta \in \Omega} \Pr(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$$

Confidence Interval: Given an interval estimator $[L(\mathbf{X}), U(\mathbf{X})]$ of θ, if its confidence coefficient is $1 - \alpha$, we call it a $(1 - \alpha)$ *confidence interval*

Confidence Set: If a set of estimators has confidence coefficient is $1 - \alpha$, we call it a $(1 - \alpha)$ *confidence set*

Expected Length: Given an interval estimator $[L(\mathbf{X}), U(\mathbf{X})]$ of θ, its *expected length* is defined as

$$\mathrm{E}[U(\mathbf{X}) - L(\mathbf{X})]$$

where $\mathbf{X}$ are random samples from $f_{\mathbf{X}}(\mathbf{x}|\theta)$. In other words, it is the average length of the interval estimator.

How to construct confidence interval?

A confidence interval can be obtained by inverting the acceptance region of a test. There is a one-to-one correspondence between tests and confidence intervals (or confidence sets).

Example 2: $X_i \sim \mathcal{N}(\theta, \sigma^2)$ where σ^2 is known. Consider $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$. As previously shown, level α LRT test reject H_0 if and only if

$$\left| \frac{\overline{X} - \theta_0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2}$$

Equivalently, we accept H_0 if $\left| \frac{\overline{X}-\theta_0}{\sigma/\sqrt{n}} \right| \leq z_{\alpha/2}$. Accepting $H_0 : \theta = \theta_0$ implies we believe our data "agrees with" the hypothesis $\theta = \theta_0$.

$$\begin{aligned} -z_{\alpha/2} \leq & \; \frac{\overline{X}-\theta_0}{\sigma/\sqrt{n}} \; \leq z_{\alpha/2} \\ \theta_0 - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq & \quad \overline{X} \quad \leq \theta_0 + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \end{aligned}$$

The *Acceptance region* is

$$\left\{ \mathbf{x} : \theta_0 - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \overline{x} \leq \theta_0 + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right\}.$$

As this is size α test, the probability of accepting H_0 is $1 - \alpha$.

$$\begin{aligned} 1 - \alpha &= \Pr\left(\theta_0 - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \overline{X} \leq \theta_0 + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right) \\ &= \Pr\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \theta_0 \leq \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right) \end{aligned}$$

Since θ_0 is arbitrary,

$$1 - \alpha = \Pr\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \theta \leq \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right)$$

Therefore, $[\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}]$ is $(1 - \alpha)$ confidence interval (CI).

Confidence intervals and level α test

Theorem 9.2.2

1. For each $\theta_0 \in \Omega$, let $A(\theta_0)$ be the acceptance region of a level α test of $H_0 : \theta = \theta_0$ Define a set $C(\mathbf{x}) = \{\theta : \mathbf{x} \in A(\theta)\}$, then the random set $C(\mathbf{x})$ is a $1-\alpha$ confidence set.
2. Conversely, if $C(\mathbf{x})$ is a $(1-\alpha)$ confidence set for θ, for any θ_0, define the acceptance region of a test for the hypothesis $H_0 : \theta = \theta_0$ by $A(\theta_0) = \{\mathbf{x} : \theta_0 \in C(\mathbf{x})\}$. Then the test has level α.

In other words, if we invert the acceptance region of the test statistic, we can obtain confidence interval, and vice versa.

Example 3: For $X_i \sim \mathcal{N}(\theta, \sigma^2)$, the acceptance region $A(\theta_0)$ is a subset of the sample space

$$A(\theta_0) = \left\{\mathbf{x} : \theta_0 - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \le \overline{x} \le \theta_0 + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right\}$$

The confidence set $C(\mathbf{x})$ is a subset of the parameter space

$$\begin{aligned} C(\mathbf{x}) &= \left\{\theta : \theta - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \le \overline{x} \le \theta + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right\} \\ &= \left\{\theta : \overline{x} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \le \theta \le \overline{x} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right\} \end{aligned}$$

There is no guarantee that the confidence set obtained from Theorem 9.2.2 is an interval, but it is so quite often

1. To obtain $(1-\alpha)$ two-sided CI $[L(\mathbf{X}), U(\mathbf{X})]$, we invert the acceptance region of a level α test for $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$
2. To obtain a lower-bounded CI $[L(\mathbf{X}), \infty)$, then we invert the acceptance region of a test for $H_0 : \theta = \theta_0$ vs. $H_1 : \theta > \theta_0$, where $\Omega = \{\theta : \theta \geq \theta_0\}$.
3. To obtain a upper-bounded CI $(-\infty, U(\mathbf{X})]$, then we invert the acceptance region of a test for $H_0 : \theta = \theta_0$ vs. $H_1 : \theta < \theta_0$, where $\Omega = \{\theta : \theta \leq \theta_0\}$.

Example 4: Let $X_i \sim \mathcal{N}(\mu, \sigma^2)$ where both parameters are unknown.

1. Find $1-\alpha$ two-sided CI for μ
2. Find $1-\alpha$ upper bound for μ

Solution - Two-sided CI

The testing problem is $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$. The LRT test rejects if and only if

$$\left| \frac{\overline{X} - \mu_0}{s_{\mathbf{X}}/\sqrt{n}} \right| > t_{n-1,\alpha/2}$$

The acceptance region is

$$A(\mu_0) = \left\{ \mathbf{x} : \left| \frac{\overline{x} - \mu_0}{s_{\mathbf{x}}/\sqrt{n}} \right| \leq t_{n-1,\alpha/2} \right\}$$

The confidence set is

$$\begin{aligned} C(\mathbf{x}) &= \left\{\mu : \left|\frac{\overline{x}-\mu}{s_{\mathbf{x}}/\sqrt{n}}\right| \leq t_{n-1,\alpha/2}\right\} \\ &= \left\{\mu : -t_{n-1,\alpha/2} \leq \frac{\overline{x}-\mu}{s_{\mathbf{x}}/\sqrt{n}} \leq t_{n-1,\alpha/2}\right\} \\ &= \left\{\mu : \overline{x} - \frac{s_{\mathbf{x}}}{\sqrt{n}} t_{n-1,\alpha/2} \leq \mu \leq \overline{x} + \frac{s_{\mathbf{x}}}{\sqrt{n}} t_{n-1,\alpha/2}\right\} \end{aligned}$$

Solution - upper-bounded CI

The CI is $(-\infty, U(\mathbf{X})]$. We need to invert a testing procedure for $H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$.

$$\Omega_0 = \{(\mu, \sigma^2) : \mu = \mu_0, \sigma^2 > 0\}$$

$$\Omega = \{(\mu, \sigma^2) : \mu \leq \mu_0, \sigma^2 > 0\}$$

LRT statistic is

$$\lambda(\mathbf{x}) = \frac{L(\hat{\mu}_0, \hat{\sigma}_0^2|\mathbf{x})}{L(\hat{\mu}, \hat{\sigma}^2|\mathbf{x})}$$

where $(\hat{\mu}_0, \hat{\sigma}_0^2)$ is the MLE restricted to Ω_0, and$(\hat{\mu}, \hat{\sigma}^2)$ is the MLE restricted to Ω, and

within Ω_0, $\hat{\mu}_0 = \mu_0$, and $\hat{\sigma}_0^2 = \frac{\sum_{i=1}^n (X_i-\mu_0)^2}{n}$

Within Ω, the MLE is

$$\begin{cases} \hat{\mu} = \overline{X} \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i-\overline{X})^2}{n} & \text{if } \overline{X} \leq \mu_0 \\ \hat{\mu} = \mu_0 \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i-\mu_0)^2}{n} & \text{if } \overline{X} > \mu_0 \end{cases}$$

$$\lambda(\mathbf{x}) = \begin{cases} 1 & \text{if } \overline{X} > \mu_0 \\ \dfrac{\left(\frac{1}{\sqrt{2\pi\hat{\sigma}_0^2}}\right)^n \exp\left\{-\frac{\sum_{i=1}^n (X_i-\mu_0)^2}{2\hat{\sigma}_0^2}\right\}}{\left(\frac{1}{\sqrt{2\pi\hat{\sigma}^2}}\right)^n \exp\left\{-\frac{\sum_{i=1}^n (X_i-\overline{X})^2}{2\hat{\sigma}^2}\right\}} & \text{if } \overline{X} \le \mu_0 \end{cases}$$

$$= \begin{cases} 1 & \text{if } \overline{X} > \mu_0 \\ \left(\frac{\frac{n-1}{n}s_{\mathbf{X}}^2}{\frac{n-1}{n}s_{\mathbf{X}}^2 + (\overline{X}-\mu_0)^2}\right)^{\frac{n}{2}} & \text{if } \overline{X} \le \mu_0 \end{cases}$$

For $0 < c < 1$, LRT test rejects H_0 if $\overline{X} \le \mu_0$ and

$$\left(\frac{\frac{n-1}{n}s_{\mathbf{X}}^2}{\frac{n-1}{n}s_{\mathbf{X}}^2 + (\overline{X}-\mu_0)^2}\right)^{\frac{n}{2}} < c$$

$$\left(\frac{\frac{n-1}{n}}{\frac{n-1}{n} + \frac{(\overline{X}-\mu_0)^2}{s_{\mathbf{x}}^2}}\right)^{\frac{n}{2}} < c$$

$$\frac{(\overline{X}-\mu_0)^2}{s_{\mathbf{X}}^2} > c^*$$

$$\frac{\mu_0 - \overline{X}}{s_{\mathbf{X}}/\sqrt{n}} > c^{**}$$

c^{**} is chosen to satisfy

$$\begin{aligned}
\alpha &= \Pr(\text{reject } H_0|\mu_0) \\
&= \Pr\left(\frac{\mu_0 - \overline{X}}{s_{\mathbf{X}}/\sqrt{n}} > c^{**}\right) \\
&= \Pr\left(\frac{\overline{X} - \mu_0}{s_{\mathbf{X}}/\sqrt{n}} < -c^{**}\right) \\
&= \Pr(T_{n-1} < -c^{**}) \\
1 - \alpha &= \Pr(T_{n-1} > -c^{**}) \\
c^{**} &= -t_{n-1,1-\alpha} = t_{n-1,\alpha}
\end{aligned}$$

Therefore, LRT level α test reject H_0 if

$$\frac{\overline{X} - \mu_0}{s_{\mathbf{X}}/\sqrt{n}} \;<\; -t_{n-1,\alpha}$$

Acceptance region is

$$A(\mu_0) = \left\{\mathbf{x} : \frac{\overline{X} - \mu_0}{s_{\mathbf{X}}/\sqrt{n}} \geq -t_{n-1,\alpha}\right\}$$

Inverting the above to get CI

$$\begin{aligned} C(\mathbf{X}) &= \{\mu : \mathbf{X} \in A(\mu)\} \\ &= \left\{\mu : \frac{\overline{X} - \mu}{s_{\mathbf{X}}/\sqrt{n}} \geq -t_{n-1,\alpha}\right\} \\ &= \left\{\mu : \overline{X} - \mu \geq -\frac{s_{\mathbf{X}}}{\sqrt{n}} t_{n-1,\alpha}\right\} \\ &= \left\{\mu : \mu \leq \overline{X} + \frac{s_{\mathbf{X}}}{\sqrt{n}} t_{n-1,\alpha}\right\} \\ &= \left(-\infty, \overline{X} + \frac{s_{\mathbf{X}}}{\sqrt{n}} t_{n-1,\alpha}\right] \end{aligned}$$

Solution - lower-bounded CI

LRT level α test reject H_0 if and only if

$$\frac{\overline{X} - \mu_0}{s_{\mathbf{X}}/\sqrt{n}} \ > \ t_{n-1,\alpha}$$

Acceptance region is

$$A(\mu_0) = \left\{\mathbf{x} : \frac{\overline{X} - \mu_0}{s_{\mathbf{X}}/\sqrt{n}} \leq t_{n-1,\alpha}\right\}$$

Confidence interval is

$$\begin{aligned} C(\mathbf{X}) &= \{\mu : \mathbf{X} \in A(\mu)\} = \left\{\mu : \frac{\mathbf{X} - \mu}{s_{\mathbf{X}}/\sqrt{n}} \leq t_{n-1,\alpha}\right\} \\ &= \left\{\mu : \mu \geq \overline{X} - \frac{s_{\mathbf{X}}}{\sqrt{n}} t_{n-1,\alpha}\right\} \\ &= \left[\overline{X} - \frac{s_{\mathbf{X}}}{\sqrt{n}} t_{n-1,\alpha}, \infty\right) \end{aligned}$$

Example 4: Let $X_1, \cdots, X_n$ be iid sample from exponential distribution with mean θ. What is a $1-\alpha$ confidence interval for the estimator of θ?

Solution: We can use LRT test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

The LRT statistic is given by

$$\begin{aligned}\lambda(\mathbf{x}) &= \frac{\frac{1}{\theta_0^n} e^{-\sum x_i/\theta_0}}{\sup_\theta \frac{1}{\theta^n} e^{-\sum x_i/\theta}} \\ &= \frac{\frac{1}{\theta_0^n} e^{-\sum x_i/\theta_0}}{\frac{1}{(\sum x_i/n)^n} e^{-n}} \\ &= \left(\frac{\sum x_i}{n\theta_0}\right)^n e^{n-\sum x_i/\theta_0}\end{aligned}$$

The acceptance region is given by

$$A(\theta_0) = \left\{\mathbf{x} : \left(\frac{\sum x_i}{\theta_0}\right)^n e^{-\sum x_i/\theta_0} \geq k\right\}$$

where k is chosen to be $\Pr(\mathbf{X} \in A(\theta_0)|\theta_0) = 1-\alpha$. Inverting this acceptance region gives the $1-\alpha$ confidence set

$$\begin{aligned}C(\mathbf{x}) &= \left\{\theta : \left(\frac{\sum x_i}{\theta}\right)^n e^{-\sum x_i/\theta} \geq k\right\} \\ &= \left\{\theta : L\left(\sum x_i\right) \leq \theta \leq U\left(\sum x_i\right)\right\}\end{aligned}$$

where L and U are functions satisfying

$$\left(\frac{\sum x_i}{L(\sum x_i)}\right)^n e^{-\sum x_i/L(\sum x_i)} = \left(\frac{\sum x_i}{U(\sum x_i)}\right)^n e^{-\sum x_i/U(\sum x_i)} = k$$

Finally,

$$\frac{\sum x_i}{L(\sum x_i)} = a \qquad \frac{\sum x_i}{U(\sum x_i)} = b \qquad (a > b)$$

where a, b satisisfy the following two conditions

$$a^n e^{-a} = b^n e^{-b} \tag{1}$$

$$\Pr\left(\tfrac{1}{a}\sum X_i \le \theta < \tfrac{1}{b}\sum X_i\right) = \Pr\left(b \le \tfrac{\sum X_i}{\theta} \le a\right) = 1 - \alpha \tag{2}$$

The fact that $\frac{2\sum X_i}{\theta} \sim \chi^2_{2n}$ can be used to select a, b.

Example of asymptotic confidence interval

Example 5: Let $X_1, \cdots, X_n$ be iid from a distribution with mean μ and finite variance σ^2. Construct asymptotic $(1-\alpha)$ two-sided interval for μ

Solution: Recall that $\overline{X}$ is the method of moment estimator for μ.
By law of large number, $\overline{X}$ is consistent for μ, and by central limit theorem,

$$\overline{X} \sim \mathcal{AN}\left(\mu, \frac{\sigma^2}{n}\right)$$

Consider testing $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$. The Wald statistic

$$Z_n = \frac{\overline{X} - \mu_0}{S_n}$$

where

$$S_n = \sqrt{\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{(n-1)n}}$$

is chosen as a consistent estimator of $\sigma/\sqrt{n}$. From previous lectures, we know that

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2 \xrightarrow{P} \sigma^2$$

$$\sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{(n-1)n}} \xrightarrow{P} \frac{\sigma}{\sqrt{n}}$$

The Wald level α test is

$$\left|\frac{(\overline{X}-\mu_0)\sqrt{n}}{\sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}}}\right| > z_{\alpha/2}$$

The acceptance region is

$$A(\mu_0) = \left\{\mathbf{x} : \left|\frac{(\overline{x}-\mu_0)\sqrt{n}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}}\right| \le z_{\alpha/2}\right\}$$

and so the $(1-\alpha)$ CI is

$$\begin{aligned}
C(\mathbf{x}) &= \{\mu : \mathbf{x} \in A(\mu)\} \\
&= \left\{\mu : \left|\frac{(\overline{x}-\mu)\sqrt{n}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}}\right| \le z_{\alpha/2}\right\} \\
&= \left[\overline{x} - \tfrac{1}{\sqrt{n}}\sqrt{\tfrac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}\,z_{\alpha/2},\ \overline{x} + \tfrac{1}{\sqrt{n}}\sqrt{\tfrac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}\,z_{\alpha/2}\right]
\end{aligned}$$

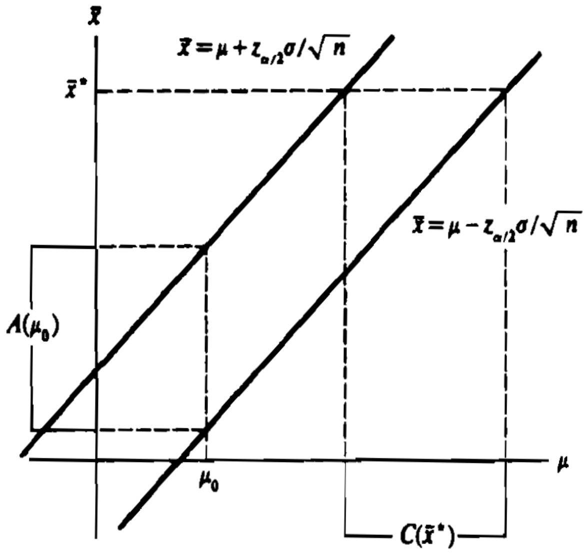

$\bar{x}$
$\bar{x} = \mu + z_{\alpha/2}\sigma/\sqrt{n}$
$\bar{x}^*$
$\bar{x} = \mu - z_{\alpha/2}\sigma/\sqrt{n}$
$A(\mu_0)$
μ_0
μ
$C(\bar{x}^*)$

Figure 9.2.1. *Relationship between confidence intervals and acceptance regions for tests. The upper line is* $\bar{x} = \mu + z_{\alpha/2}\sigma/\sqrt{n}$ *and the lower line is* $\bar{x} = \mu - z_{\alpha/2}\sigma/\sqrt{n}$.

Discrete Distributions

Typically for discrete distributions, it is quite hard to get an explicit interval.

Example 6: Let $X_1, \cdots, X_n$ be iid $Bernoulli(p)$ an consider testing

$$H_0 : p = p_0 \ vs \ H_1 : p > p_0.$$

In this problem, $T = \sum_{i=1}^n X_i$ is a sufficient statistic. Since

$$\begin{aligned} f(\mathbf{x}|p) &= \prod_{i=1}^n p^{x_i}(1-p)^{x_i} \\ &= \left(\frac{p}{1-p}\right)^{\sum_{i=1}^n x_i} (1-p)^n \\ &= (1-p)^n \exp\left[\log\left(\frac{p}{1-p}\right) \sum_{i=1}^n x_i\right] \end{aligned}$$

conforms to an exponential family with $\omega(p) = \log\left(\frac{p}{1-p}\right)$ an increasing function of p, the family of pmf's has MLR in $T = \sum_{i=1}^n X_i$. So by Karlin-Rubin Theorem, the test that

$$\text{rejects } H_0 \text{ if } T > k(p_0)$$

is the UMP test of its size. We cannot get the size of the test to be exactly α, except for certain values of p_0, because of the discreteness of T.

The cut-off $k(p_0)$ is the integer between 0 and n that satisfies

$$\sum_{y=0}^{k(p_0)} \binom{n}{y} p_0^y (1-p_0)^{n-y} \geq 1-\alpha, \quad \sum_{y=0}^{k(p_0)-1} \binom{n}{y} p_0^y (1-p_0)^{n-y} < 1-\alpha.$$

For each p_0, the acceptance region is given by

$$A(p_0) = \{t : t \leq k(p_0)\}.$$

Correspondingly, for each value of t, the confidence set is

$$C(t) = \{p_0 : t \leq k(p_0)\}.$$

While this is formally correct, this is not explicit. The $(1-\alpha)$ lower confidence bound can be shown to be given by

$$C(t) = \left\{ p_0 : p_0 > \sup_p \left\{ \sum_{y=0}^{t-1} \binom{n}{y} p^y (1-p)^{n-y} \ \geq \ 1-\alpha \right\} \right\}.$$

Pivotal Quantities

Pivotal quantities are quite useful in constructing confidence intervals.

Definition 9.2.6: A random variable $Q(\mathbf{X}; \theta) = Q(X_1, \ldots, X_n; \theta)$ is a pivotal quantity if the distribution of $Q(\mathbf{X}, \theta)$ is free of all parameters.

$Q(\mathbf{X}; \theta)$ contains both parameters and statistics, but its distribution is free of θ. Note that a pivotal quantity is different from an ancillary statistic.

Examples

1. Consider $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$; σ^2 known.

$$Q(\mathbf{X}; \mu) = \overline{X} - \mu$$

2. Consider $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$; both parameters unknown.

   - $Q_1(\mathbf{X}; \mu, \sigma^2) = \frac{S^2_{\mathbf{X}}}{\sigma^2}$.
   - $Q_2(\mathbf{X}; \mu, \sigma^2) = \frac{\overline{X} - \mu}{S_{\mathbf{X}}}$.

3. Consider $X_1, \cdots, X_n \sim Exp(\theta)$.

$$Q(\mathbf{X};\theta) = \frac{\sum_{i=1}^n X_i}{\theta}$$

4. Consider $X_1, \cdots, X_n \sim Uniform(\theta, \theta+1)$.

$$Q(\mathbf{X};\theta) = X_{(n)} - \theta$$

Pivotal quantity and location-scale family

Let $X_1, \cdots, X_n$ be a random sample from $f(x|\theta)$.

Location Family

$$f(x|\theta) \sim f_0(x-\theta) \ \text{ where } \ f_0 \text{ is parameter free.}$$

Then

$$Q(\mathbf{X};\theta) = (\hat{\theta}_{MLE} - \theta) \ \text{ is a pivotal.}$$

Scale Family

$$f(x|\theta) \sim \frac{1}{\theta} f_0\left(\frac{x}{\theta}\right) \ \text{ where } \ f_0 \text{ is parameter free.}$$

Then

$$Q(\mathbf{X};\theta) = \frac{\hat{\theta}_{MLE}}{\theta} \ \text{ is a pivotal.}$$

Location-Scale Family

$$f(x|\mu,\sigma) \sim \frac{1}{\sigma} f_0\left(\frac{x-\mu}{\sigma}\right) \ \text{ where } \ f_0 \text{ is parameter free.}$$

Then

$$Q(\mathbf{X};\mu,\sigma) = \frac{\hat{\mu}_{MLE} - \mu}{\hat{\sigma}_{MLE}} \ \text{ is a pivotal.}$$

Once we have a pivotal quantity, then for any specified α, we can find numbers a and b, which do not depend on θ, and satisfy

$$\Pr_{\theta}\left[a \leq Q(\mathbf{X};\theta) \leq b\right] \;\geq\; 1-\alpha.$$

So a $1-\alpha$ confidence set for θ is given by

$$C(\mathbf{x}) = \{\theta_0 : a \leq Q(\mathbf{X};\theta_{\mathbf{0}}) \leq b.\}$$

If θ is a real-valued parameter, and if for each $\mathbf{x} \in \mathcal{X}$, the pivotal $Q(\mathbf{X};\theta)$ is a monotone function of θ, then $C(\mathbf{x})$ will be an interval.

Example 7: Let $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ where both parameters are unknown. Since $\mathcal{N}(\mu,\sigma^2)$ is a location-scale family, and

$$\hat{\mu}_{MLE} = \overline{X}, \quad \hat{\sigma}^2_{MLE} = \frac{(n-1)S^2_{\mathbf{X}}}{n},$$

$$Q(\mathbf{X};\mu,\sigma^2) = \frac{\overline{x}-\mu}{\sqrt{(n-1)S^2_{\mathbf{X}}/n}} \quad \text{is a pivot.}$$

Note that $T = \sqrt{n-1}\; Q = \frac{\overline{X}-\mu}{S_{\mathbf{X}}/\sqrt{n}} \;\sim\; t_{(n-1)}$ and hence

$$\Pr[a \leq T \leq b] = 1-\alpha$$

for specific percentiles of $t_{(n-1)}$. Making an equal tailed choice

$$a = -t_{(n-1),\alpha/2}, \;\; b = t_{(n-1),\alpha/2}$$

and so a $1-\alpha$ confidence interval for μ is the familiar one

$$C(\mathbf{x}) = \left\{\mu : \overline{x} - t_{(n-1),\alpha/2}\,\frac{s_{\mathbf{x}}}{\sqrt{n}} \;\leq\; \mu \;\leq\; \overline{x} + t_{(n-1),\alpha/2}\,\frac{s_{\mathbf{x}}}{\sqrt{n}}\right\}.$$

Example 8: Let $X_1, \cdots, X_n \sim Exp(\theta)$. Find a pivotal and construct a equal-tailed $1 - \alpha$ confidence interval for θ based on the pivotal.

Example 9: Let $X_1, \cdots, X_n \sim Exp(\mu, 1)$ with pdf

$$f(x|\mu) = e^{-(x-\mu)}\ I(x > \mu), \quad -\infty < \mu < \infty.$$

Find a pivotal and construct a equal-tailed $1-\alpha$ confidence interval for μ based on the pivotal.