COMPS NOTES

Table of contents

Probability Foundation

MGF-Moment Generate Function

- Use to find moment, if question asked to find second moment and you find yourself using pdf, usually wrong and much easier using MGF
- **Use MGF to prove converge in distribution**, see comps 2022

Theorem

Definition 2.3.6 Let X be a random variable with cdf F_X. The moment generating function (mgf) of X (or F_X), denoted by $M_X(t)$, is

$$M_X(t) = \mathrm{E}e^{tX}$$

provided that the expectation exists for t in some neighborhood of 0 . That is, there is an $h > 0$ such that, for all t in $-h < t < h, \mathrm{E}e^{tX}$ exists. If the expectation does not exist in a neighborhood of 0 , we say that the moment generating function does not exist. More explicitly, we can write the mgf of X as

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx \quad \text{if } X \text{ is continuous,}$$

or

$$M_X(t) = \sum_x e^{tx} P(X = x) \quad \text{if } X \text{ is discrete.}$$

It is very easy to see how the mgf generates moments. We summarize the result in the following theorem.

Theorem 2.3.7 If X has mgf $M_X(t)$, then

$$\mathrm{E}X^n = M_X^{(n)}(0)$$

where we define

$$M_X^{(n)}(0) = \frac{d^n}{dt^n} M_X(t)\Big|_{t=0}$$

That is, the nth moment is equal to the nth derivative of $M_X(t)$ evaluated at $t = 0$.

Theorem 2.3.12 (Convergence of mgfs) Suppose $\{X_i, i = 1, 2, ...\}$ is a sequence of random variables, each with mgf $M_{X_i}(t)$. Furthermore, suppose that

$$\lim_{i\to\infty} M_{X_i}(t) = M_X(t), \qquad \text{for all } t \text{ in a neighborhood of } 0,$$

and $M_X(t)$ is an mgf. Then there is a unique cdf F_X whose moments are determined by $M_X(t)$ and, for all x where $F_X(x)$ is continuous, we have

$$\lim_{i\to\infty} F_{X_i}(x) = F_X(x)$$

That is, convergence, for $|t| < h$, of mgfs to an mgf implies convergence of cdfs.

Theorem 2.3.15 For any constants a and b, the mgf of the random variable $aX + b$ is giten by

$$M_{aX+b}(t) = e^{bt} M_X(at)$$

Example: Use MGF to show poison Convergence

$$P(X = x) \approx P(Y = x)$$

for large n and small np. We now show that the mgfs converge, lending credence to this approximation. Recall that

$$M_X(t) = \left[pe^t + (1-p)\right]^n$$

For the Poisson(λ) distribution, we can calculate (see Exercise 2.33)

$$M_y(t) = e^{\lambda(e^t-1)}$$

and if we define $p = \lambda/n$, then $M_X(t) \to M_Y(t)$ as $n \to \infty$. The validity of the approximation in (2.3.9) will then follow from Theorem 2.3.12.

We first must digress a bit and mention an important limit result, one that has wide applicability in statistics. The proof of this lemma may be found in many standard calculus texts.

Lemma 2.3.14 Let $a_1, a_2, \ldots$ be a sequence of numbers converging to a, that is, $\lim_{n\to\infty} a_n = a$. Then

$$\lim_{n\to\infty} \left(1 + \frac{a_n}{n}\right)^n = e^a$$

Returning to the example, we have

$$M_x(t) = \left[pe^t + (1-p)\right]^n = \left[1 + \frac{1}{n}\left(e^t - 1\right)(np)\right]^n = \left[1 + \frac{1}{n}\left(e^t - 1\right)\lambda\right]^n,$$

because $\lambda = np$. Now set $a_n = a = \left(e^t - 1\right)\lambda$, and apply Lemms 2.3.14 to get

$$\lim_{n\to\infty} M_X(t) = e^{\lambda(e^t-1)} = M_Y(t)$$

the moment generating function of the Poisson.

The Poisson approximation can be quite good even for moderate p and n. In Figure 2.3.3 we show a binomial mass function along with its Poisson approximation, with $\lambda = np$. The approximation appears to be satisfactory.

Example: Rao's Blackwell 6.8

- *If the Blackwellization of W yields an estimator different from W, then the new estimator has a smaller MSE.*

Example 6.8 *Suppose $X_1, X_2, \ldots, X_n \sim$ iid Poisson (θ). Let $\tau(\theta) = P\{X = 0\} = e^{-\theta}$.*

- *Consider the statistic* $W = \mathbf{1}_{\{0\}}(X_1) \sim \text{Bernoulli}(\tau(\theta))$
- $\mathrm{E}_\theta[W] = \tau(\theta)$*, that is, W is unbiased for $\tau(\theta)$.*
- $\mathrm{Var}_\theta[W] = \tau(\theta)(1 - \tau(\theta))$
- *From previous work we know that $T = \sum X_i$ is sufficient for θ. Note that* $T \sim \text{Poisson}(n\theta)$.
- *Find a better unbiased estimator, in terms of the variance.*
- *Confirm this estimator is unbiased.*
- *Is this estimator UMVUE?*

$$\begin{aligned}
\Phi(s) = E[W|T = s] &= P(X_1 = 0|T = s) \\
&= \frac{P(X_1 = 0, \sum_1 X_i = s)}{P(T = s)} \\
&= \frac{P(X_1 = 0) * \underbrace{P(\sum_2 X_i = s)}_{Poi((n-1)\theta)}}{P(T = s)} \\
&= (\frac{n-1}{n})^s \\
\Phi(T) &= (\frac{n-1}{n})^{\sum X_i}
\end{aligned}$$

This is a better estimator by Rao-blackwell.

$$\begin{aligned}
E[\Phi(T)] = E[\frac{n-1}{n}^{\sum X_i}] &= E[exp(log(\frac{n-1}{n})\sum X_i)] \\
\text{by moment generate function}: M_T(t) &= exp(n\theta(e^t - 1)) \\
= M(log(\frac{n-1}{n})) &= exp[n\theta(\frac{n-1}{n} - 1)] \\
&= exp(-\theta)
\end{aligned}$$

This shows that Φ is unbiased.

35

Independent

- $f(x, y) = f(x)f(y)$
- $f(x|y) = f(x)$
- $E[xy] = E[x]E[y]$

Example: Show $\bar{X} \perp S^2$

Without loss of generality, we assume i.i.d sample $X_1, ..., X_n \sim N(0, 1)$

Let's look at the first case X_1:

Recall that $\sum_1 (X_i - \bar{X}) = n\frac{\sum_1 X_i}{n} - n\bar{X} = 0$, we have:

$$\sum_2 (X_i - \bar{X}) = -(X_1 - \bar{X})$$
$$\Rightarrow (X_1 - \bar{X})^2 = [\sum_2 (X_i - \bar{X})]^2$$

Replace it into sample variance we have:

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_1 (X_i - \bar{X})^2 \\ &= \frac{1}{n-1}[(X_1 - \bar{X})^2 + \sum_2 (X_i - \bar{X})^2] \\ &= \frac{1}{n-1}\{[\sum_2 (X_i - \bar{X})]^2 + \sum_2 (X_i - \bar{X})^2\} \\ &\text{is a function of } X_i - \bar{X}, i \in \{2, ..., n\} \end{aligned}$$

Next, we need to show jointly $\bar{X}$ and $X_i - \bar{X}, i \in \{2, ..., n\}$ are independent.

We do that by finding the joint pdf and

Now, let $Y_1 = \bar{X}, Y_2 = X_2 - \bar{X}, ..., Y_n = X_n - \bar{X}$, we have

$$\begin{aligned} \sum_2 Y_i &= \sum_2 X_i - \sum_2 \bar{X} \\ &= \sum_1 \bar{X} - X_1 - \sum_2 \bar{X} \\ &= \bar{X} - X_1 \\ &\Rightarrow \begin{cases} X_1 = Y_1 - \sum_2 Y_i \\ X_i = Y_i + Y_1 \end{cases} \end{aligned}$$

So the jacobian is

$$\begin{bmatrix} \frac{\partial X_1}{\partial Y_1} & ... & \frac{\partial X_n}{\partial Y_1} \\ ... & ... & ... \\ \frac{\partial X_1}{\partial Y_n} & ... & \frac{\partial X_n}{\partial Y_n} \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & ... & -1 \\ 1 & 1 & 0 & ... & 0 \\ 1 & 0 & 1 & ... & 0 \\ 1 & ... & ... & ... & 0 \\ 1 & 0 & 0 & ... & 1 \end{bmatrix}$$

We can show that $|J| = n$,

By i.i.d, we have

$$f_{X_1,...,X_n} = (2\pi)^{-n/2} exp\{-\frac{1}{2}\sum X_i^2\}$$

Replacing Y_i we have

$$f_{Y_1,...,Y_n} = n(2\pi)^{-n/2} exp\{-\frac{1}{2}[(Y_1 - \sum_2 Y_i)^2 + \sum_2 (Y_i + Y_1)^2]\}$$
$$= c * exp\{-0.5[(Y_1^2 - \cancel{2Y_1 \sum_2 Y_i} + (\sum_2 Y_i)^2) + \sum_2 (Y_1^2 + \cancel{2Y_1Y_i} + Y_i^2)]\}$$
$$= c * exp(-0.5nY_1^2) * exp(-0.5[(\sum_2 Y_i)^2 + \sum_2 Y_i^2])$$

We left out the indicator function since they don't have interaction indicator function

By factorization, $Y_1 = \bar{X} \perp \sum_2 Y_i$, since a function of independent R.V is also independent, we have $\bar{X} \perp S^2 \blacksquare$

Conditional

Definition

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- if $A \perp B$,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

- if $A \subset B$,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)}$$

Conditional Expectation & Variance ...

$$E[g(X)|Y] = \int g(x) * f_{X|Y}(x)dx \Rightarrow h(Y)$$

- Conditional something of Y should be a function of Y
- $E[X|X] = X$ itself:

Proof:

$$E[X|X] = \int xf(x|X)dx$$

where f(x|X) is the conditional probability density function of X given X.

However, since we already know the value of X, the conditional probability density function of X given X is just a Dirac delta function centered at X, which is defined as:

$$f(x|X) = \delta(x - X)$$

where $\delta(x - X)$ is the Dirac delta function, which is zero for all values of x except $x = X$, where it is infinite.

Substituting this into the expression for the conditional expectation, we get:

$$E[X|X] = \int x\delta(x - X)dx$$

Since the Dirac delta function is zero everywhere except at x = X, this integral evaluates to:

$$E[X|X] = X * \int \delta(x - X)dx$$

where the integral evaluates to 1 since the Dirac delta function integrates to 1 over its support. Therefore, we have:

$$E[X|X] = X$$

- $Var(X|X) = 0$, I know mind blowing

Proof:

$$Var(X|X) = E[(X - E[X|X])^2|X]$$

Since $E[X|X] = X$, we have:

$$Var(X|X) = E[(X - X)^2|X] = E[0|X] = 0$$

Total Variance

- You failed to show it in homework

$$Var(X) = Var(E[X|Y]) + E[Var(X|Y)]$$

Proof:

$$\begin{aligned} Var(X) &= E[X - E[X]]^2 \\ &= E\{X - E[X|Y] + E[X|Y] - E[X]\}^2 \\ &= E[X - E[E|Y]]^2 + E[E[X|Y] - E[X]]^2 \\ &\quad +2E[(X - E[X|Y])(E[X|Y] - E[X])] \end{aligned}$$

The last term in this expression is equal to 0 , however, which can easily be seen by iterating the expectation:

$$\mathrm{E}([X - \mathrm{E}(X \mid Y)][\mathrm{E}(X \mid Y) - \mathrm{E}X]) = \mathrm{E}(\mathrm{E}\{[X - \mathrm{E}(X \mid Y)][\mathrm{E}(X \mid Y) - \mathrm{E}X] \mid Y\})$$

In the conditional distribution $X \mid Y, X$ is the random variable. So in the expression

$$\mathrm{E}\{[X - \mathrm{E}(X \mid Y)][\mathrm{E}(X \mid Y) - \mathrm{E}X] \mid Y\}$$

$\mathrm{E}(X \mid Y)$ and $\mathrm{E}X$ are constants. Thus,

$$\begin{aligned} \mathrm{E}\{[X - \mathrm{E}(X \mid Y)][\mathrm{E}(X \mid Y) - \mathrm{E}X] \mid Y\} &= (\mathrm{E}(X \mid Y) - \mathrm{E}X)(\mathrm{E}\{[X - \mathrm{E}(X \mid Y)] \mid Y\}) \\ &= (\mathrm{E}(X \mid Y) - \mathrm{E}X)(\mathrm{E}(X \mid Y) - \mathrm{E}(X \mid Y)) \\ &= (\mathrm{E}(X \mid Y) - \mathrm{E}X)(0) \\ &= 0. \end{aligned}$$

Thus, from (4.4.6), we have that $\mathrm{E}((X - \mathrm{E}(X \mid Y))(\mathrm{E}(X \mid Y) - \mathrm{E}X)) = \mathrm{E}(0) = 0$. Referring back to equation (4.4.5), we see that

$$\begin{aligned} \mathrm{E}\left([X - \mathrm{E}(X \mid Y)]^2\right) &= \mathrm{E}\left(\mathrm{E}\left\{[X - \mathrm{E}(X \mid Y)]^2 \mid Y\right\}\right) \\ &= \mathrm{E}(\mathrm{Var}(X \mid Y)) \end{aligned}$$

and

$$\mathrm{E}\left([\mathrm{E}(X \mid Y) - \mathrm{E}X]^2\right) = \mathrm{Var}(\mathrm{E}(X \mid Y))$$

establishing (4.4.4).

Hierarchical Model

$$E[X] = E[E[X|Y]] = E[E[E[X|Y, Z]]] = ...$$

Example

$$\begin{aligned} X \mid Y &\sim \text{binomial}(Y, p) \\ Y \mid \Lambda &\sim \text{Poisson}(\Lambda) \\ \Lambda &\sim \text{exponential}(\beta) \end{aligned}$$

where the last stage of the hierarchy accounts for the variability across different mothers. The mean of X can easily be calculated as

$$\begin{aligned} \mathrm{E}X &= \mathrm{E}(\mathrm{E}(X \mid Y)) \\ &= \mathrm{E}(pY) \\ &= \mathrm{E}(\mathrm{E}(pY \mid \Lambda)) \\ &= \mathrm{E}(p\Lambda) && \text{(as before)} \\ &= p\beta, && \text{(exponential expectation)} \end{aligned}$$

Multivariate Distribution

Bivariate Normal Distribution

Let $\mathbf{x}_1$ be the first partition and $\mathbf{x}_2$ the second. Now define $\mathbf{z} = \mathbf{x}_1 + \mathbf{A}\mathbf{x}_2$ where $\mathbf{A} = -\Sigma_{12}\Sigma_{22}^{-1}$. Now we can write

$$\begin{aligned} \operatorname{cov}(\mathbf{z}, \mathbf{x}_2) &= \operatorname{cov}(\mathbf{x}_1, \mathbf{x}_2) + \operatorname{cov}(\mathbf{A}\mathbf{x}_2, \mathbf{x}_2) \\ &= \Sigma_{12} + \mathrm{A}\operatorname{var}(\mathbf{x}_2) \\ &= \Sigma_{12} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22} \\ &= 0 \end{aligned}$$

Therefore $\mathbf{z}$ and $\mathbf{x}_2$ are uncorrelated and, since they are jointly normal, they are independent. Now, clearly $E(\mathbf{z}) = \mu_1 + \mathbf{A}\mu_2$, therefore it follows that

$$\begin{aligned} E(\mathbf{x}_1 \mid \mathbf{x}_2) &= E(\mathbf{z} - \mathbf{A}\mathbf{x}_2 \mid \mathbf{x}_2) \\ &= E(\mathbf{z} \mid \mathbf{x}_2) - E(\mathbf{A}\mathbf{x}_2 \mid \mathbf{x}_2) \\ &= E(\mathbf{z}) - \mathbf{A}\mathbf{x}_2 \\ &= \mu_1 + \mathbf{A}(\mu_2 - \mathbf{x}_2) \\ &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2) \end{aligned}$$

which proves the first part. For the covariance matrix, note that

$$\begin{aligned} \operatorname{var}(\mathbf{x}_1 \mid \mathbf{x}_2) &= \operatorname{var}(\mathbf{z} - \mathbf{A}\mathbf{x}_2 \mid \mathbf{x}_2) \\ &= \operatorname{var}(\mathbf{z} \mid \mathbf{x}_2) + \operatorname{var}(\mathbf{A}\mathbf{x}_2 \mid \mathbf{x}_2) - \mathrm{A}\operatorname{cov}(\mathbf{z}, -\mathbf{x}_2) - \operatorname{cov}(\mathbf{z}, -\mathbf{x}_2)\mathbf{A}' \\ &= \operatorname{var}(\mathbf{z} \mid \mathbf{x}_2) \\ &= \operatorname{var}(\mathbf{z}) \end{aligned}$$

Now we're almost done:

$$\begin{aligned}\operatorname{var}\left(\mathbf{x}_1 \mid \mathbf{x}_2\right)=\operatorname{var}(\mathbf{z}) & =\operatorname{var}\left(\mathbf{x}_1+\mathbf{A} \mathbf{x}_2\right) \\ & =\operatorname{var}\left(\mathbf{x}_1\right)+\mathbf{A} \operatorname{var}\left(\mathbf{x}_2\right) \mathbf{A}^{\prime}+\mathbf{A} \mathbf{cov}\left(\mathbf{x}_1, \mathbf{x}_2\right)+\operatorname{cov}\left(\mathbf{x}_2, \mathbf{x}_1\right) \mathbf{A}^{\prime} \\ & =\Sigma_{11}+\Sigma_{12} \Sigma_{22}^{-1} \Sigma_{22} \Sigma_{22}^{-1} \Sigma_{21}-2 \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \\ & =\Sigma_{11}+\Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}-2 \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \\ & =\Sigma_{11}-\Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}\end{aligned}$$

So

$$f_{X,Y} = N(\begin{bmatrix}\mu_x\\ \mu_y\end{bmatrix}, \begin{bmatrix}\sigma_x^2 & \rho\sigma_x\sigma_y\\ \rho\sigma_x\sigma_y & \sigma_y^2\end{bmatrix})$$

And

$$f_{X|Y} = N\{\mu_x + \rho\sqrt{\frac{\sigma_x}{\sigma_y}}(Y-\mu_y), \sigma_x^2(1-\rho)\}$$

Example exercise 4.45

Show that if $(X, Y) \sim$ bivariate normal $\left(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho\right)$, then the following are true. (a) The marginal distribution of X is $\mathrm{n}\left(\mu_X, \sigma_X^2\right)$ and the marginal distribution of Y is $n\left(\mu_Y, \sigma_Y^2\right)$. (b) The conditional distribution of Y given $X=x$ is

$$\mathrm{n}\left(\mu_Y+\rho\left(\sigma_Y / \sigma_X\right)\left(x-\mu_X\right), \sigma_Y^2\left(1-\rho^2\right)\right)$$

(c) For any constants a and b, the distribution of $aX+bY$ is

$$\mathrm{n}\left(a \mu_X+b \mu_Y, a^2 \sigma_X^2+b^2 \sigma_Y^2+2 a b \rho \sigma_X \sigma_Y\right)$$

The mean is easy to check,

$$\mathrm{E}(aX+bY)=a \mathrm{E} X+b \mathrm{E} Y=a \mu_X+b \mu_Y$$

as is the variance,

$$\operatorname{Var}(aX+bY)=a^2 \operatorname{Var} X+b^2 \operatorname{Var} Y+2 a b \operatorname{Cov}(X, Y)=a^2 \sigma_X^2+b^2 \sigma_Y^2+2 a b \rho \sigma_X \sigma_Y$$

To show that $aX+bY$ is normal we have to do a bivariate transform. One possibility is $U=aX+bY, V=Y$, then get $f_{U,V}(u,v)$ and show that $f_U(u)$ is normal. We will do this in the standard case. Make the indicated transformation and write $x=\frac{1}{a}(u-bv), y=v$ and obtain

$$|J| = \begin{vmatrix} 1/a & -b/a \\ 0 & 1 \end{vmatrix} = \frac{1}{a}$$

Then

$$f_{UV}(u,v) = \frac{1}{2\pi a\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left[\frac{1}{a}(u-bv)\right]^2 - 2\frac{\rho}{a}(u-bv)+v^2\right]}.$$

Now factor the exponent to get a square in u. The result is

$$-\frac{1}{2\left(1-\rho^2\right)}\left[\frac{b^2+2\rho ab+a^2}{a^2}\right]\left[\frac{u^2}{b^2+2\rho ab+a^2} - 2\left(\frac{b+a\rho}{b^2+2\rho ab+a^2}\right)uv + v^2\right]$$

Note that this is joint bivariate normal form since $\mu_U = \mu_V = 0, \sigma_v^2 = 1, \sigma_u^2 = a^2+b^2+2ab\rho$ and

$$\rho^* = \frac{\operatorname{Cov}(U,V)}{\sigma_U\sigma_V} = \frac{\mathrm{E}\left(aXY+bY^2\right)}{\sigma_U\sigma_V} = \frac{a\rho+b}{\sqrt{a^2+b^2+ab\rho}}$$

thus

$$\left(1-\rho^{*2}\right) = 1 - \frac{a^2\rho^2+ab\rho+b^2}{a^2+b^2+2ab\rho} = \frac{\left(1-\rho^2\right)a^2}{a^2+b^2+2ab\rho} = \frac{\left(1-\rho^2\right)a^2}{\sigma_u^2}$$

where $a\sqrt{1-\rho^2} = \sigma_U\sqrt{1-\rho^{*2}}$. We can then write

$$f_{UV}(u,v) = \frac{1}{2\pi\sigma_U\sigma_V\sqrt{1-\rho^{*2}}}\exp\left[-\frac{1}{2\sqrt{1-\rho^{*2}}}\left(\frac{u^2}{\sigma_U^2} - 2\rho\frac{uv}{\sigma_U\sigma_V} + \frac{v^2}{\sigma_V^2}\right)\right]$$

which is in the exact form of a bivariate normal distribution. Thus, by part a), U is normal.

Example exercise 4.46

4.46 (*A* derivation of the bivariate normal distribution) Let Z_1 and Z_2 be independent $\mathrm{n}(0,1)$ random variables, and define new random variables X and Y by

$$X = a_X Z_1 + b_X Z_2 + c_X \text{ and } Y = a_Y Z_1 + b_Y Z_2 + c_Y$$

where a_X, b_X, c_X, a_Y, b_Y, and c_Y are constants. (a) Show that

$$\begin{aligned} \mathrm{E}X = c_X, &\quad \operatorname{Var} X = a_X^2 + b_X^2, \\ \mathrm{E}Y = c_Y, &\quad \operatorname{Var} Y = a_Y^2 + b_Y^2, \\ \operatorname{Cov}(X,Y) &= a_X a_Y + b_X b_Y. \end{aligned}$$

(b) If we define the constants a_X, b_X, c_X, a_Y, b_Y, and c_Y by

$$a_X = \sqrt{\frac{1+\rho}{2}}\sigma_X, \quad b_X = \sqrt{\frac{1-\rho}{2}}\sigma_X, \quad c_X = \mu_X,$$
$$a_Y = \sqrt{\frac{1+\rho}{2}}\sigma_Y, \quad b_Y = -\sqrt{\frac{1-\rho}{2}}\sigma_Y, \quad c_Y = \mu_Y,$$

where $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$, and ρ are constants, $-1 \le \rho \le 1$, then show that

$$\mathrm{E}X = \mu_X, \quad \operatorname{Var} X = \sigma_X^2,$$
$$\mathrm{E}Y = \mu_Y, \quad \operatorname{Var} Y = \sigma_Y^2,$$
$$\rho_{XY} = \rho.$$

(c) Show that (X, Y) has the bivariate normal pdf with parameters $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$, and ρ.

Let $D = a_X b_Y - a_Y b_X = -\sqrt{1-\rho^2}\sigma_X\sigma_Y$ and solve for Z_1 and Z_2,

$$Z_1 = \frac{b_Y(X - c_X) - b_X(Y - c_Y)}{D} = \frac{\sigma_Y(X-\mu_X) + \sigma_X(Y-\mu_Y)}{\sqrt{2(1+\rho)}\sigma_X\sigma_Y}$$
$$Z_2 = \frac{\sigma_Y(X-\mu_X) + \sigma_X(Y-\mu_Y)}{\sqrt{2(1-\rho)}\sigma_X\sigma_Y}.$$

Then the Jacobian is

$$J = \begin{pmatrix} \frac{\partial z_1}{\partial x_1} & \frac{\partial z_1}{\partial y} \\ \frac{\partial z_2}{\partial x} & \frac{\partial z_2}{\partial y} \end{pmatrix} = \begin{pmatrix} \frac{b_Y}{D} & \frac{-b_X}{D} \\ \frac{-a_Y}{D} & \frac{a_X}{D} \end{pmatrix} = \frac{a_X b_Y}{D^2} - \frac{a_Y b_X}{D^2} = \frac{1}{D} = \frac{1}{-\sqrt{1-\rho^2}\sigma_X\sigma_Y}$$

and we have that

$$f_{X,Y}(x,y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(\sigma_Y(x-\mu_X)+\sigma_X(y-\mu_Y))^2}{2(1+\rho)\sigma_X^2\sigma_Y^2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(\sigma_Y(x-\mu_X)+\sigma_X(y-\mu_Y))^2}{2(1-\rho)\sigma_X^2\sigma_Y^2}} \frac{1}{\sqrt{1-\rho^2}\sigma_X\sigma_Y}$$
$$= \left(2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}\right)^{-1} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right)$$
$$-2\rho\frac{x-\mu_X}{\sigma_X}\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2, \quad -\infty < x < \infty, -\infty < y < \infty,$$

a bivariate normal pdf.

Multinomial

Transformation/Location Scale Family

Univariate Transformation

Location Scale Family

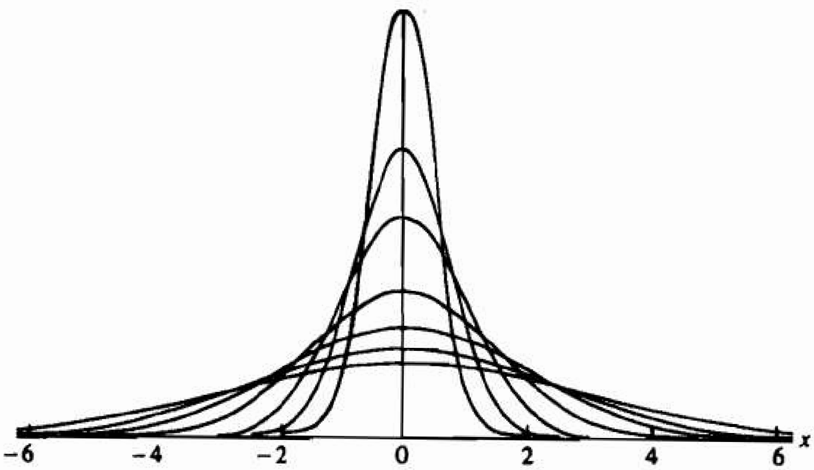

Figure 3.5.3. *Members of the same scale family*

The other two types of families to be discussed in this section are scale families and location–scale families.

Definition 3.5.4 Let $f(x)$ be any pdf. Then for any $\sigma > 0$, the family of pdfs $(1/\sigma)f(x/\sigma)$, indexed by the parameter σ, is called the *scale family with standard pdf* $f(x)$ and σ is called the *scale parameter* of the family.

The effect of introducing the scale parameter σ is either to stretch ($\sigma > 1$) or to contract ($\sigma < 1$) the graph of $f(x)$ while still maintaining the same basic shape of the graph. This is illustrated in Figure 3.5.3. Most often when scale parameters are used, $f(x)$ is either symmetric about 0 or positive only for $x > 0$. In these cases the stretching is either symmetric about 0 or only in the positive direction. But, in the definition, any pdf may be used as the standard.

Several of the families introduced in Section 3.3 either are scale families or have scale families as subfamilies. These are the gamma family if α is a fixed value and β is the scale parameter, the normal family if $\mu = 0$ and σ is the scale parameter, the exponential family, and the double exponential family if $\mu = 0$ and σ is the scale parameter. In each case the standard pdf is the pdf obtained by setting the scale parameter equal to 1. Then all other members of the family can be shown to be of the form in Definition 3.5.4.

Definition 3.5.5 Let $f(x)$ be any pdf. Then for any μ, $-\infty < \mu < \infty$, and any $\sigma > 0$, the family of pdfs $(1/\sigma)f((x-\mu)/\sigma)$, indexed by the parameter (μ, σ), is called the *location–scale family with standard pdf* $f(x)$; μ is called the *location parameter* and σ is called the *scale parameter*.

The effect of introducing both the location and scale parameters is to stretch ($\sigma > 1$) or contract ($\sigma < 1$) the graph with the scale parameter and then shift the graph so that the point that was above 0 is now above μ. Figure 3.5.4 illustrates this transformation of $f(x)$. The normal and double exponential families are examples of location–scale families. Exercise 3.39 presents the Cauchy as a location–scale family.

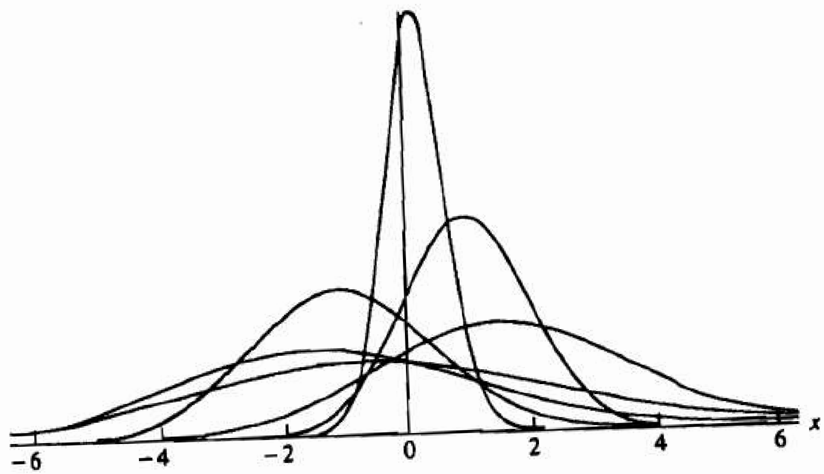

Figure 3.5.4. *Members of the same location–scale family*

The following theorem relates the transformation of the pdf $f(x)$ that defines a location–scale family to the transformation of a random variable Z with pdf $f(z)$. As mentioned earlier in the discussion of location families, the representation in terms of Z is a useful mathematical tool and can help us understand when a location–scale family might be appropriate in a modeling context. Setting $\sigma = 1$ in Theorem 3.5.6 yields a result for location (only) families, and setting $\mu = 0$ yields a result for scale (only) families.

Theorem 3.5.6 *Let $f(\cdot)$ be any pdf. Let μ be any real number, and let σ be any positive real number. Then X is a random variable with pdf $(1/\sigma)f((x-\mu)/\sigma)$ if and only if there exists a random variable Z with pdf $f(z)$ and $X = \sigma Z + \mu$.*

Proof: To prove the "if" part, define $g(z) = \sigma z + \mu$. Then $X = g(Z)$, g is a monotone function, $g^{-1}(x) = (x-\mu)/\sigma$, and $|(d/dx)g^{-1}(x)| = 1/\sigma$. Thus by Theorem 2.1.5, the pdf of X is

$$f_X(x) = f_Z(g^{-1}(x)) \left| \frac{d}{dx} g^{-1}(x) \right| = f\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma}.$$

To prove the "only if" part, define $g(x) = (x-\mu)/\sigma$ and let $Z = g(X)$. Theorem 2.1.5 again applies: $g^{-1}(z) = \sigma z + \mu$, $|(d/dz)g^{-1}(z)| = \sigma$, and the pdf of Z is

$$f_Z(z) = f_X(g^{-1}(z)) \left| \frac{d}{dz} g^{-1}(z) \right| = \frac{1}{\sigma} f\left(\frac{(\sigma z + \mu) - \mu}{\sigma}\right)\sigma = f(z).$$

Also,

$$\sigma Z + \mu = \sigma g(X) + \mu = \sigma\left(\frac{X-\mu}{\sigma}\right) + \mu = X. \qquad \square$$

An important fact to extract from Theorem 3.5.6 is that the random variable $Z = (X - \mu)/\sigma$ has pdf

$$f_Z(z) = \frac{1}{1} f\left(\frac{z-0}{1}\right) = f(z).$$

That is, the distribution of Z is that member of the location–scale family corresponding to $\mu = 0, \sigma = 1$. This was already proved for the special case of the normal family in Section 3.3.

Often, calculations can be carried out for the "standard" random variable Z with pdf $f(z)$ and then the corresponding result for the random variable X with pdf $(1/\sigma)f((x-\mu)/\sigma)$ can be easily derived. An example is given in the following, which is a generalization of a computation done in Section 3.3 for the normal family.

Theorem 3.5.7 *Let Z be a random variable with pdf $f(z)$. Suppose $\mathrm{E}Z$ and $\mathrm{Var}\, Z$ exist. If X is a random variable with pdf $(1/\sigma)f((x-\mu)/\sigma)$, then*

$$\mathrm{E}X = \sigma \mathrm{E}Z + \mu \quad \text{and} \quad \mathrm{Var}\, X = \sigma^2 \mathrm{Var}\, Z.$$

In particular, if $\mathrm{E}Z = 0$ and $\mathrm{Var}\, Z = 1$, then $\mathrm{E}X = \mu$ and $\mathrm{Var}\, X = \sigma^2$.

Proof: By Theorem 3.5.6, there is a random variable Z^* with pdf $f(z)$ and $X = \sigma Z^* + \mu$. So $\mathrm{E}X = \sigma\mathrm{E}Z^* + \mu = \sigma \mathrm{E}Z + \mu$ and $\mathrm{Var}\, X = \sigma^2 \mathrm{Var}\, Z^* = \sigma^2 \mathrm{Var}\, Z$. □

For any location–scale family with a finite mean and variance, the standard pdf $f(z)$ can be chosen in such a way that $\mathrm{E}Z = 0$ and $\mathrm{Var}\, Z = 1$. (The proof that this choice can be made is left as Exercise 3.40.) This results in the convenient interpretation of μ and σ^2 as the mean and variance of X, respectively. This is the case for the usual definition of the normal family as given in Section 3.3. However, this is not the choice for the usual definition of the double exponential family as given in Section 3.3. There, $\mathrm{Var}\, Z = 2$.

Probabilities for any member of a location–scale family may be computed in terms of the standard variable Z because

$$P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = P\left(Z \le \frac{x-\mu}{\sigma}\right).$$

Thus, if $P(Z \le z)$ is tabulated or easily calculable for the standard variable Z, then probabilities for X may be obtained. Calculations of normal probabilities using the standard normal table are examples of this.

3.6 Inequalities and Identities

Statistical theory is literally brimming with inequalities and identities—so many that entire books are devoted to the topic. The major work by Marshall and Olkin (1979) contains many inequalities using the concept of majorization. The older work by Hardy, Littlewood, and Polya (1952) is a compendium of classic inequalities. In this section and in Section 4.7 we will mix some old and some new, giving some idea of the

Multivariate Transformation

Year	Question	How it goes
2021 T	2	Bad, very important

Exercise CB 5.17

Order Statistics

If the kth order staistics is less than X, then it means that at least k-1 observations are less than X.

Based on this we can have the cdf for the kth order statistics:

Discrete Ordered Statistics

Given a random sample $X_1, X_2, \cdots, X_n$ from a discrete distribution with pmf $f(x)$,

The probability that there are exact k observations less than x is:

$$P(X_{(k)} = x) = \binom{n}{k} P(X_1 < x)^k (1 - P(X_1 < x))^{n-k}$$

The probablity that at least k observations are less than x is:

$$P(X_{(k)} \leq x) = \sum_{i=k}^{n} \binom{n}{i} F_X(x)^i (1 - F_X(x))^{n-i}$$

Continuous Ordered Statistics

Given a random sample $X_1, X_2, \cdots, X_n$ from a continuous distribution with cdf $F(x)$,

the pdf of the kth order statistics is:

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} F_X(x)^{k-1} (1 - F_X(x))^{n-k} f_X(x)$$

Just draw a line and you can write down the joint pdf

Joint Ordered Statistics

Theorem 5.4.6 Let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics of a random sample, $X_1, \ldots, X_n$, from a continuous population with cdf $F_X(x)$ and pdf $f_X(x)$. Then the joint pdf of $X_{(i)}$ and $X_{(j)}, 1 \leq i < j \leq n$, is

$$\begin{aligned} (5.4.7) \qquad f_{X_{(i)},X_{(j)}}(u,v) = & \frac{n!}{(i-1)!(j-1-i)!(n-j)!} f_X(u) f_X(v) \left[F_X(u)\right]^{i-1} \\ & \times \left[F_X(v) - F_X(u)\right]^{j-1-i} \left[1 - F_X(v)\right]^{n-j} \end{aligned}$$

for $-\infty < u < v < \infty$

Some Simple Examples

For general n, we have

$$\boxed{f_{X_{(1)},X_{(2)},\ldots,X_{(n)}}(x_1, x_2, \ldots, x_n)} \quad \text{“=”} \quad P(X_{(1)} = x_1, X_{(2)} = x_2, \ldots X_{(n)} = x_n)$$

$$= \quad \boxed{n! f(x_1) f(x_2) \cdots f(x_n)}$$

which holds for $\boxed{x_1 < x_2 < \cdots < x_n}$ with all x_i in the support for the original distribution. The joint pdf is zero otherwise.

The Formalities:

The joint cdf,

$$P(X_{(1)} \le x_1, X_{(2)} \le x_2, \ldots, X_{(n)} \le x_n),$$

is a little hard to work with. Instead, we consider something similar:

$$P(y_1 < X_{(1)} \le x_1, y_2 < X_{(2)} \le x_2, \ldots, y_n < X_{(n)} < x_n)$$

for values $y_1 < x_1 \le y_2 < x_2 \le y_3 < x_3 \le \cdots \le y_n < x_n$.

This can happen if

$$y_1 < X_1 \le x_1, \; y_2 < X_2 \le x_2, \; \ldots, \; y_n < X_n < x_n,$$

or if

$$y_1 < X_5 \le x_1, \; y_2 < X_3 \le x_2, \; \ldots, \; y_n < X_{n-2} < x_n,$$

or...

Because of the constraints on the x_i and y_i, these are disjoint events. So, we can add these $n!$ probabilities, which will all be the same, together to get

$$P(y_1 < X_{(1)} \le x_1, \ldots, y_n < X_{(n)} < x_n) = n!\, P(y_1 < X_1 \le x_1, \ldots, y_n < X_n < x_n).$$

Note that

$$P(y_1 < X_1 \le x_1, \ldots, y_n < X_n < x_n) \overset{indep}{=} \prod_{i=1}^{n} P(y_i < X_i \le x_i) = \prod_{i=1}^{n} [F(x_i) - F(y_i)].$$

So,

$$P(y_1 < X_{(1)} \le x_1, \ldots, y_n < X_{(n)} < x_n) = n! \prod_{i=1}^{n} [F(x_i) - F(y_i)] \tag{2}$$

The left-hand side is

$$\int_{y_n}^{x_n} \int_{y_{n-1}}^{x_{n-1}} \cdots \int_{y_1}^{x_1} f_{X_{(1)},X_{(2)},\ldots,X_{(n)}}(u_1, u_2, \ldots, u_n)\, du_1 du_2 \ldots, du_n.$$

Taking derivatives $\frac{d}{dx_1} \frac{d}{dx_2} \cdots \frac{d}{dx_n}$ gives

$$f_{X_{(1)},X_{(2)},\ldots,X_{(n)}}(x_1, x_2, \ldots, x_n)$$

Differentiating <u>both sides</u> of (2) with respect to $x_1, x_2, \ldots, x_n$ gives us

$$f_{X_{(1)},X_{(2)},\ldots,X_{(n)}}(x_1, x_2, \ldots, x_n) = n! f(x_1) f(x_2) \cdots f(x_n)$$

which holds for $x_1 < x_2 < \cdots, x_n$ and all x_i in the support of the original distribution. The pdf is zero otherwise.

Exercise CB 5.24

Show that $X_{(1)}/X_{(n)}$ and $X_{(n)}$ are independent, where $f_X(x) = 1/\theta$

Proof: Use $f_X(x) = 1/\theta, F_X(x) = x/\theta, 0 < x < \theta$. Let $Y = X_{(n)}, Z = X_{(1)}$. Then, from Theorem 5.4.6,

$$f_{Z,Y}(z,y) = \frac{n!}{0!(n-2)!0!}\frac{1}{\theta}\frac{1}{\theta}\left(\frac{z}{\theta}\right)^0\left(\frac{y-z}{\theta}\right)^{n-2}\left(1-\frac{y}{\theta}\right)^0 = \frac{n(n-1)}{\theta^n}(y-z)^{n-2}, 0 < z < y < \theta$$

Solutions Manual for Statistical Inference Now let $W = Z/Y, Q = Y$. Then $Y = Q, Z = WQ$, and $|J| = q$. Therefore

$$f_{W,Q}(w,q) = \frac{n(n-1)}{\theta^n}(q-wq)^{n-2}q = \frac{n(n-1)}{\theta^n}(1-w)^{n-2}q^{n-1}, 0 < w < 1, 0 < q < \theta.$$

The joint pdf factors into functions of w and q, and, hence, W and Q are independent.

Exercise CB 5.27

Exercise CB 5.42

Exercise CB 7.49

Example: 2021-T 3a-b

Asymptotic Probability

Necessary limits

Exponential

$$\lim_{n\to\infty}(1+\frac{a}{n})^n = e^a$$

$$\lim_{n\to\infty}(1-\frac{a}{n})^n = e^{-a}$$

Necessary Inequality

Markov's Inequality

$$P(X \geq s) \leq E(X)/s$$

Chebychev's Inequality

Let X be a random variable and let $g(z)$ be a nonnegative function. Then, for any $r > 0$,

$$P(g(X) \geq r) \leq \frac{\mathrm{E}g(X)}{r}.$$

Proof:

$$\begin{aligned} \mathrm{E}g g(X) &= \int_{-\infty}^{\infty} g(x) f_X(x) dx \\ &\geq \int_{\{x: g(x) \geq r\}} g(x) f_X(x) dx \qquad (g \text{ is nonnegative}) \\ &\geq r \int_{\{x: g(x) \geq r\}} f_X(x) dx \\ &= rP(g(X) \geq r). \quad \text{(definition)} \end{aligned}$$

Example

Holder's Inequality

Let X,Y be R.V and $q > 1, p > 1, \frac{1}{q} + \frac{1}{p} = 1$

$$|E[XY]| \leq E(|XY|) \leq [E(|X|^p)]^{\frac{1}{q}} [E(|Y|^q)]^{\frac{1}{p}}$$

$$\Rightarrow \sum |X_i Y_i| \leq [(\|X|^p)]^{\frac{1}{q}} [E(|Y|^q)]^{\frac{1}{p}}$$

Cauchy's Inequality

Extend from Holder,

$$|E(XY)| \leq E(|XY|) \leq E[(X^2)]^{\frac{1}{2}} E[(Y^2)]^{\frac{1}{2}}$$

Use this to show that $|cov(X, Y)| \leq \sigma_x \sigma_y$

$$\begin{aligned} \Rightarrow |cov(X, Y)| &= |E[(X - \mu_x)(Y - \mu_y)]| \\ &\leq E[|(X - \mu_x)(Y - \mu_y)|] \\ &\leq E[(X - \mu_x)^2]^{\frac{1}{2}} * E[(Y - \mu_y)^2]^{\frac{1}{2}} \\ &= \sigma_x \sigma_y \end{aligned}$$

$$\Rightarrow |\rho(x, y)| \leq 1$$

Jenson's Inequality

Theorem 4.7.7 (Jensen's Inequality)

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

$$Eg(X) \geq g(\mathrm{E}X)$$

Equality holds if and only if, for every line $a + bx$ that is tangent to $g(x)$ at $x = EX$, $P(g(X) = a + bX) = 1$

Proof:

To establish the inequality, let $l(x)$ be a tangent line to $g(x)$ at the point $g(\mathrm{E}X)$. (Recall that $\mathrm{E}X$ is a constant.) Write $l(x) = a + bx$ for some a and b. The situation is illustrated in Figure 4.7.2.

Now, by the convexity of g we have $g(x) \geq a + bx$. Since expectations preserve inequalities,

$$\begin{aligned}
\mathrm{E}g(X) &\geq \mathrm{E}(a + bX) \\
&= a + b\mathrm{E}X && \left(\begin{array}{c}\text{linearity of expectation,} \\ \text{Theorem 2.2.5}\end{array}\right) \\
&= l(\mathrm{E}X) && (\text{definition of } l(x)) \\
&= g(\mathrm{E}X), && (l \text{ is tangent at } \mathrm{E}X)
\end{aligned}$$

as was to be shown. If $g(x)$ is linear, equality follows from properties of expectations (Theorem 2.2.5). For the "only if" part see Exercise 4.62 .

One immediate application of Jensen's Inequality shows that $\mathrm{E}X^2 \geq (\mathrm{E}X)^2$, since $g(x) = x^2$ is convex. Also, if x is positive, then $1/x$ is convex; hence $\mathrm{E}(1/X) \geq 1/\mathrm{E}X$, another useful application.

To check convexity of a twice differentiable function is quite easy. The function $g(x)$ is convex if $g''(x) \geq 0$, for all x, and $g(x)$ is concave if $g''(x) \leq 0$, for all x. Jensen's Inequality applies to concave functions as well. If g is concave, then $\mathrm{E}g(X) \leq g(\mathrm{E}X)$.

Boferroni's Inequality

$$P(A \cap B) \geq P(A) + P(B) - 1$$

Convergence

Converge in Probability

Definition:

A sequence of random variables, $X_1, X_2, \ldots$, converges in probability to a random variable X if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(|X_n - X| \geq \epsilon\right) = 0 \quad \text{or, equivalently,} \quad \lim_{n\to\infty} P\left(|X_n - X| < \epsilon\right) = 1.$$

The $X_1, X_2, \ldots$ in Definition 5.5.1 (and the other definitions in this section) are typically not independent and identically distributed random variables, as in a random sample. The distribution of X_n changes as the subscript changes, and the convergence concepts discussed in this section describe different ways in which the distribution of X_n converges to some limiting distribution as the subscript becomes large.

Theorem 5.5.4

Suppose that $X_1, X_2, \ldots$ converges in probability to a random variable X and that h is a continuous function. Then $h\left(X_1\right), h\left(X_2\right), \ldots$ converges in probability to $h(X)$.

Exercise CB 5.32

Exercise CB 5.33

Converge in Distribution

We have already encountered the iden of convergence in distribution in Chapter 2. Remember the properties of moment generating functions (mgfs) and how their convergence implies convergence in distribution (Theorem 2.3.12).

Definition 5.5.10

A sequence of random variables, $X_1, X_2, \ldots$, converges in distri. bution to a random variable X if

$$\lim_{n\to\infty} F_{X_\infty}(x) = F_x(x)$$

st all points x where $F_X(x)$ is continuous.

Slutsky's Theorem

Theorem 5.5.17 (Slutsky's Theorem) If $X_n \to X$ in distribution and $Y \to a$, where a is constant, in probability, then

$$Y_n X_n \to^d a + X \text{ in distribution}$$

$$Y_n + X_n \to^d aX \text{ in distribution}$$

Continuous Mapping Theorem

Example 5.5.11 (Maximum of uniforms)

If $X_1, X_{21\ldots}$ are lid uniform(0,1) and $X_{(n)} = \max_{1\le i\le n} X_{in}$, let us examine if (and to where) $X_{(\mathrm{s})}$ converges in distribution. As $n \to \infty$, we expeet $X_{(n)}$ to get eloee to 1 and, as $X_{(n)}$ must necessarily be leas than 1 , we have for any $\varepsilon > 0$,

$$\begin{aligned} P\left(\left|X_{(n)}-1\right| \geq e\right) &= P\left(X_{(n)} \geq 1+\varepsilon\right)+P\left(X_{(n)} \leq 1-\varepsilon\right) \\ &= 0+P\left(X_{(n)} \leq 1-c\right). \end{aligned}$$

Next asing the fact that we have an tid sample, we can write

$$P\left(X_{(n)} \leq 1-c\right)=P\left(X_i \leq 1-\varepsilon_1 i=1, \ldots n\right)=(1-c)^n$$

which goes to 0 . So we have proved that $X_{(\mathrm{i})}$ converges to 1 in probability. However, if we take $c = t/n$, we then have

$$P\left(X_{(n)} \leq 1-t/n\right)=(1-t/n)^n \rightarrow e^{-t}$$

which, spon rearranging ylelds

$$P\left(n\left(1-X_{(n)}\right) \leq t\right) \rightarrow 1-e^{-r_1}$$

that is, the random variable $n\left(1-X_{(n)}\right)$ converges in distribution to an exponential(1) random varibble.

Exercise CB 5.18

Exercise CB 5.23

Converge Almost Surely (related to Strong Law of Large Number)

Definition 5.5.6

A sequence of random variables, $X_1, X_2, \ldots$, converges almost surely to a random variable X if, for every $\epsilon > 0$,

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

Notice the similarity in the statements of Definitions 5.5.1 and 5.5.6. Although they look similar, they are very different statements, with Definition 5.5.6 much stronger To understand almost sure convergence, we must recall the basic definition of a random variable as given in Definition 1.4.1. A random variable is a real-valued function defined on a sample space S. If a sample space S has elements denoted by s, then $X_n(s)$ and $X(s)$ are all functions defined on S. Definition 5.5.6 states that X_n converges to X almost surely if the functions $X_n(s)$ converge to $X(s)$ for all $s \in S$ except perhaps for $s \in N$, where $N \subset S$ and $P(N) = 0$.

Example 5.5.7 (Almost sure convergence)

Let the sample space S be the closed interval $[0,1]$ with the uniform probability distribution. Define random variables $X_n(s) = s + s^n$ and $X(s) = s$. For every $s \in [0,1), s^n \rightarrow 0$ as $n \rightarrow \infty$ and $X_n(s) \rightarrow s = X(s)$. However, $X_n(1) = 2$ for every n so $X_n(1)$ does not converge to $1 = X(1)$. But since the convergence occurs on the set $[0,1)$ and $P([0,1)) = 1, X_n$ converges to X almost surely.

Law of Large Number

Weak Law of Large Number

Definition

Let $X_1, X_2, \ldots$ be iid random variables with $\mathrm{E} X_i = \mu$ and $\operatorname{Var} X_i = \sigma^2 < \infty$. Define $\bar{X}_n = (1/n) \sum_{i=1}^n X_i$. Then, for every $\epsilon > 0$

$$\lim_{n \rightarrow \infty} P\left(\left|\vec{X}_n - \mu\right| < \epsilon\right) = 1$$

that is, $\bar{X}_n$ converges in probability to μ.

Proof: The proof is quite simple, being a straightforward application of Chebychev's Inequality. We have, for every $\epsilon > 0$,

$$P\left(\left|\bar{X}_n - \mu\right| \geq \epsilon\right) = P\left(\left(\bar{X}_n - \mu\right)^2 \geq \epsilon^2\right) \leq \frac{\mathrm{E}\left(\bar{X}_n - \mu\right)^2}{\epsilon^2} = \frac{\operatorname{Var} \bar{X}_{\frown}}{\epsilon^2} = \frac{\sigma^2}{n \epsilon^2}$$

Hence, $P\left(\left|\bar{X}_n - \mu\right| < \epsilon\right) = 1 - P\left(\left|\bar{X}_n - \mu\right| \geq \epsilon\right) \geq 1 - \sigma^2 / \left(n \epsilon^2\right) \rightarrow 1$, as $n \rightarrow \infty$.

The Weak Law of Large Numbers (WLLN) quite elegantly states that, under general conditions, the sample mean approaches the population mean as $n \rightarrow \infty$. In fact, there are more general versions of the WLLN, where we need assume only that the mean is finite. However, the version stated in Theorem 5.5.2 is applicable in most practical situations.

The property summarized by the WLLN, that a sequence of the "same" sample quantity approaches a constant as $n \rightarrow \infty$, is known as consistency.

Example 5.5.3 (Consistency of S^2)

Suppose we have a sequence $X_1, X_2, \ldots$ of iid random variables with $\mathrm{E} X_i = \mu$ and $\operatorname{Var} X_i = \sigma^2 < \infty$. If we define

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

can we prove a WLLN for S_n^2 ? Using Chebychev's Inequality, we have

$$P\left(\left|S_n^2-\sigma^2\right| \geq \epsilon\right) \leq \frac{\mathrm{E}\left(S_n^2-\sigma^2\right)^2}{\epsilon^2}=\frac{\operatorname{Var} S_n^2}{\epsilon^2}$$

and thus, a sufficient condition that S_n^2 converges in probability to σ^2 is that Var $S_n^2 \rightarrow 0$ as $n \rightarrow \infty$.

If S_n^2 is a consistent estimator of σ^2, then by Theorem 5.5.4, the sample standard deviation $S_n=\sqrt{S_n^2}=h\left(S_n^2\right)$ is a consistent estimator of σ. Note that S_n is, in fact, a biased estimator of σ (see Exercise 5.11), but the bias disappears asymptotically.

Strong Law of Large Number

Theorem 5.5.9 (Strong Law of Large Numbers) Let $X_1, X_2, \ldots$ be iid random variaWes with $\mathrm{E} X_i=\mu$ and Var $X_i=\sigma^2<\infty$, and define $X_n=(1/\mathrm{n}) \sum_{i=1}^n X_i$. Then, for enery $c>0$,

$$P\left(\lim _{n \rightarrow \infty}\left|\hat{X}_n-\mu\right|<e\right)=1$$

that is, X_n converges almost sturely to μ. For both the Weak and Strong Laws of Large Numbers we had the sssumption of a finite variance. Although such an assumption is trae (and desirable) in most applications, it is, in fact, a stronger assumption than is needed. Both the weak and strong laws hold without this assumption. The only moment condition needed is that E $\left|X_i\right|<\infty$ (see Resnick 1999, Chapter 7, or Billingsiey 1995, Section 22).

CLT

Theorem 5.5.14 (Central Limit Theorem)

Let $X_1, X_2, \ldots$ be a sequence of iid random variables whose mgfs exist in a neighborhood of O (that is, $M_{X_i}(t)$ exists for $|t|<h$, for some positive h). Let $\mathrm{E} X_i=\mu$ and Var $X_i=\sigma^2>0$. (Both μ and σ^2 are finite since the mgf exists.) Define $\bar{X}_n=(1/n) \sum_{i=1}^n X_i$. Let $G_n(x)$ denote the cdf of $\sqrt{n}\left(\bar{X}_n-\mu\right)/\sigma$. Then, for any $x,-\infty<x<\infty$,

$$\lim _{n \rightarrow \infty} G_n(x)=\int_{-\infty}^x \frac{1}{\sqrt{2 \pi}} e^{-y^2/2} d y ;$$

that is, $\sqrt{n}\left(\bar{X}_n-\mu\right)/\sigma$ has a limiting standard normal distribution.

Proof of CLT

Exercise CB 5.34

Delta Methods

- Application: how to use delta method to get the CI of non-linear regression
-

Taylar Series

Definition 5.5.20 If a function $g(x)$ has derivatives of order r, that is, $g^{(r)}(x) = \frac{d^-}{dx^+} g(x)$ exists, then for any constant a, the Taylor polynomial of order r about a is

$$T_r(x) = \sum_{i=0}^{r} \frac{g^{(i)}(a)}{i!}(x-a)^i.$$

Taylor's major theorem, which we will not prove here, is that the remainder from the approximation, $g(x) - T_r(x)$, always tends to 0 faster than the highest-order explicit term. Theorem 5.5.21 (Taylor) If $g^{(r)}(a) = \frac{d^r}{dx^r} g(x)\Big|_{x=a}$ exists, then

$$\lim_{x \to a} \frac{g(x) - T_r(x)}{(x-a)^r} = 0.$$

In general, we will not be concerned with the explicit form of the remainder. Since we are interested in approximations, we are just going to ignore the remainder. There are, however, many explicit forms, one useful one being

$$g(x) - T_r(x) = \int_a^x \frac{g^{(r+1)}(t)}{r!}(x-t)^r dt$$

For the statistical application of Taylor's Theorem, we are most concerned with the first-order Taylor series, that is, an approximation using just the first derivative (taking $r = 1$ in the above formulas). Furthermore, we will also find use for a multivariate Taylor series. Since the above detail is univariate, some of the following will have to be accepted on faith.

Let $T_1, \ldots, T_k$ be random variables with means $\theta_1, \ldots, \theta_k$, and define $\mathbf{T} = (T_1, \ldots, T_k)$ and $\theta = (\theta_1, \ldots, \theta_k)$. Suppose there is a differentiable function $g(T)$ (an estimator of some parameter) for which we want an approximate estimate of variance. Define

$$g_i'(\theta) = \frac{\partial}{\partial t_i} g(\mathbf{t})\Bigg|_{t_1=\theta_3, \ldots, t_k=\theta_k}$$

The first-order Taylor series expansion of g about θ is

$$g(\mathbf{t}) = g(\theta) + \sum_{i=1}^{k} g_i'(\theta)\left(t_i - \theta_i\right) + \text{ Remainder.}$$

For our statistical approximation we forget about the remainder and write

$$g(t) \approx g(\theta) + \sum_{i=1}^{k} g_i'(\theta)\left(t_i - \theta_i\right)$$

Now, take expectations on both sides of (5.5.7) to get (5.5.8)

$$\begin{aligned} \mathrm{E}_\theta g(\mathbf{T}) &\approx g(\theta)+\sum_{i=1}^{k} g_i'(\theta)\mathrm{E}_\theta\left(T_i-\theta_i\right) \\ &= g(\theta). \end{aligned}$$

(T_i has mean θ_i)

We can now approximate the variance of $g($ T$)$ by

$$\begin{aligned} \operatorname{Var}_\theta g(\mathbf{T}) &\approx \mathrm{E}_\theta\left([g(\mathbf{T})-g(\theta)]^2\right) \\ &\approx \mathrm{E}_\theta\left(\left(\sum_{i=1}^{k} g_i'(\theta)\left(T_i-\theta_i\right)\right)^2\right) \qquad \text{(using (5.5.8))} \\ &= \sum_{i=1}^{k}\left[g_i'(\theta)\right]^2 \operatorname{Var}_\theta T_i+2\sum_{i>j} g_i'(\theta)g_j'(\theta)\operatorname{Cov}_\theta\left(T_i,T_j\right), \end{aligned}$$

Delta Method for Univariate

Example 5.5.23 (Approximate mean and variance) Suppose X is a random variable with $\mathrm{E}_\mu X=\mu\neq 0$. If we want to estimate a function $g(\mu)$, a first-order approximation would give us

$$g(X)=g(\mu)+g'(\mu)(X-\mu)$$

If we use $g(X)$ as an estimator of $g(\mu)$, we can say that approximately

$$\begin{aligned} \mathrm{E}_\mu g(X) &\approx g(\mu), \\ \operatorname{Var}_\mu g(X) &\approx \left[g'(\mu)\right]^2 \operatorname{Var}_\mu X. \end{aligned}$$

For a specific example, take $g(\mu)=1/\mu$. We estimate $1/\mu$ with $1/X$, and we can say

$$\begin{aligned} \mathrm{E}_\mu\left(\frac{1}{X}\right) &\approx \frac{1}{\mu}, \\ \operatorname{Var}_\mu\left(\frac{1}{X}\right) &\approx \left(\frac{1}{\mu}\right)^4 \operatorname{Var}_\mu X. \end{aligned}$$

Using these Taylor series approximations for the mean and variance, we get the following useful generalization of the Central Limit Theorem, known as the Delta Mathod

Theorem 5.5.24 (Delta Method)

Let Y_n be a sequence of random variables that satisfies $\sqrt{n}\left(Y_n-\theta\right)\rightarrow \mathrm{n}\left(0,\sigma^2\right)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta)$ exists and is not

0 . Then (5.5.10) $\sqrt{n}\left[g\left(Y_n\right)-g(\theta)\right] \rightarrow \mathrm{n}\left(0, \sigma^2\left[g^{\prime}(\theta)\right]^2\right)$ in distribution. Proof: The Taylor expansion of $g\left(Y_n\right)$ around $Y_n=\theta$ is

$$g\left(Y_n\right)=g(\theta)+g^{\prime}(\theta)\left(Y_n-\theta\right)+\text { Remainder, }$$

where the remainder $\rightarrow 0$ as $Y_n \rightarrow \theta$. Since $Y_n \rightarrow \theta$ in probability it follows that the remainder $\rightarrow 0$ in probability. By applying Slutsky's Theorem (Theorem 5.5.17) to

$$\sqrt{n}\left[g\left(Y_n\right)-g(\theta)\right]=g^{\prime}(\theta) \sqrt{n}\left(Y_n-\theta\right),$$

the result now follows. See Exercise 5.43 for details.

Example 5.5.25 (Continuation of Example 5.5.23) Suppose now that we have the mean of a random sample $\bar{X}$. For $\mu \neq 0$, we have

$$\sqrt{n}\left(\frac{1}{\bar{X}}-\frac{1}{\mu}\right) \rightarrow \mathrm{n}\left(0,\left(\frac{1}{\mu}\right)^4 \operatorname{Var}_\mu X_1\right)$$

in distribution. If we do not know the variance of X_1, to use the above approximation requires an estimate, say S^2. Moreover, there is the question of what to do with the $1 / \mu$ term, as we also do not know μ. We can estimate everything, which gives us the approximate variance

$$\widehat{\operatorname{Var}}\left(\frac{1}{\bar{X}}\right) \approx\left(\frac{1}{\bar{X}}\right)^4 S^2 .$$

Furthermore, as both $\bar{X}$ and S^2 are consistent estimators, we can again apply Slutsky's Theorem to conclude that for $\mu \neq 0$,

$$\frac{\sqrt{n}\left(\frac{1}{\bar{X}}-\frac{1}{\mu}\right)}{\left(\frac{1}{\bar{X}}\right)^2 S} \rightarrow \mathrm{n}(0,1)$$

Theorem: Second Order Delta Method

In some cases $\frac{d}{d\mu} g(\mu)=0$, ie. score function, we need to use second order delta method:

$$g(\hat{\theta})|_{\theta_0=\mu}=g(\mu)+\underbrace{g^{\prime}(\mu)}_{=0}(\hat{\theta}-\mu)+g^{\prime \prime}(\mu) \frac{(\hat{\theta}-\mu)^2}{2!}+\ldots$$

$$\Rightarrow g(\hat{\theta})-g(\mu) \approx \frac{g^{\prime \prime}(\mu)}{2}(\hat{\theta}-\mu)^2$$

By applying slutsky's theorem, where $\sqrt{n}(\hat{\theta}-\mu) \rightarrow N(0, \sigma_\theta^2)$ in distribution

$$\frac{2}{g^{\prime \prime}(\mu)} \frac{[g(\hat{\theta})-g(\mu)]^2}{\sigma^2 / n} \rightarrow N(0,1)^2 \sim \mathcal{X}_1^2$$

Delta Method for Multivariate

Given Taylor series:

$$h(B) \approx h(\beta) + \nabla h(\beta)^T \cdot (B - \beta)$$

We can have

$$\begin{aligned}
\mathrm{Var}(h(B)) &\approx \mathrm{Var}[h(\beta) + \nabla h(\beta)^T \cdot (B - \beta)] \\
&= \mathrm{Var}(h(\beta) + \nabla h(\beta)^T \cdot B - \nabla h(\beta)^T \cdot \beta(\beta)) \\
&= \mathrm{Var}(\nabla h(\beta)^T \cdot B) \\
&= \nabla h(\beta)^T \cdot cov(B) \cdot \nabla h(\beta) \\
&= \nabla h(\beta)^T \cdot \frac{\sum}{n} \cdot \nabla h(\beta)
\end{aligned}$$

Where in this case the scale parameter is $\alpha = n$,

So in general case we have:

$$\sqrt{\alpha}\,(h(B) - h(\beta)) \xrightarrow{D} N\left(0, \nabla h(\beta)^T \cdot \Sigma \cdot \nabla h(\beta)\right)$$

In regression case, we have scale parameter $\alpha = X^T X$ $(Var(\hat{\beta}) = X^T X^{-1}\Sigma)$

Statistics & Estimator

Sufficient Statistics

Definition

Factorization Theorem: How to find Sufficient Statistics

2.3 Factorization Theorem

Theorem 2.1 *Let $f_{\boldsymbol{X}}(\boldsymbol{x} \mid \boldsymbol{\theta})$ denote the joint pdf or pmf of a sample $\boldsymbol{X}$. A statistic $T(\boldsymbol{X})$ is a sufficient statistic for $\boldsymbol{\theta}$ if and only if there exist functions $g(T(\boldsymbol{x}) \mid \boldsymbol{\theta})$ and $h(\boldsymbol{x})$, where $h(\boldsymbol{x})$ does not depend on $\boldsymbol{\theta}$, such that*

$$f_{\boldsymbol{X}}(\boldsymbol{x} \mid \boldsymbol{\theta}) = h(\boldsymbol{x})g(T(\boldsymbol{x}) \mid \boldsymbol{\theta})$$

Factorization Theorem

Proof:

- Assume $T(X)$ is a sufficient statistics for θ (and prove factorization)

$$\begin{aligned} f(x|\theta) &= P_\theta(X = x) \\ &= P(X = x \text{and} T(X) = T(x)) \\ &= P(X = x|T(X) = T(x)) * P_\theta(T(X) = T(x)) \\ &= g(T(x)|\theta)h(x) \end{aligned}$$

- Assume $f_x(x|\theta) = h(x)g(T(x)|\theta)$ (and prove sufficient)

  Given the marginal distribution of $T(X)$ q:

$$q(T(x)|\theta) = \sum_{y=A_{T(x)}} g(T(y)|\theta)h(y)$$

where $A_{T(x)} = \{y : T(y) = T(x)\}$

$$\begin{aligned} \frac{f(x|\theta)}{q(T(x)|\theta)} &= \frac{g(T(x|\theta)h(x))}{\sum_{y=A_{T(x)}} g(T(y)|\theta)h(y)} \\ &= \frac{g(T(x|\theta)h(x))}{g(T(x)|\theta)\sum_{y=A_{T(x)}} h(y)} \\ &= \frac{h(x))}{\sum_{y=A_{T(x)}} h(y)} \leftarrow \text{independent of } \theta \end{aligned}$$

■

Example 2.1 *Given data from $N(\mu, \sigma^2)$, what if we want to estimate μ only(i.e given σ)*

$$f_X(\boldsymbol{x} \mid \mu, \sigma) = \underbrace{\left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{(n-1)s_X^2}{2\sigma^2}\right)}_{\textit{Does not depend on } \mu} \underbrace{\exp\left(-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right)}_{\textit{Depends only } \bar{x} \textit{ and } \mu}$$

So, $\bar{X}$ is a sufficient statistic for μ when σ is known.

If both μ' s and σ' s values are unknown then the model's parameter i $\boldsymbol{\theta} = (\mu, \sigma)$.$\bar{X}$ does not contain all of the information - in the joint distribution - about $\boldsymbol{\theta}$. However. $\left(\bar{X} \cdot S_{\mathrm{V}}^2\right)$ is jiointly sufficient for $\boldsymbol{\theta}$.

■

Example 2.2 *Given above, what if we only want to estimate σ?*

$$f_X(\boldsymbol{x} \mid \mu, \sigma) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

$\sum_{i=1}^{n}(x_i - \mu)^2$ *is sufficient for* σ■

■

Lemma 2.2 *If $T_1(\boldsymbol{X}), T_2(\boldsymbol{X}), \ldots, T_k(\boldsymbol{X})$ are a set of jointly sufficient statistics for $\boldsymbol{\theta}$, then any set of one-to-one functions, or transformations, of $(T_1, \ldots, T_k)$ is also jointly sufficient for $\boldsymbol{\theta}$.*

2.4 Exponential Family

Theorem 2.3 *Theorem: Let $X_1, \ldots, X_n$ be iid observations from a pdf or pmf $f_X(x \mid \boldsymbol{\theta})$ that belongs to an exponential family given by*

$$f_X(x \mid \boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x)\right)$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d), d \leq k$. Then

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

is a (jointly) sufficient statistic for $\boldsymbol{\theta}$.

Example 2.3

$$f_X(\boldsymbol{x} \mid \mu, \sigma) = \left(\left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right)\right)^n \exp\left(-\frac{1}{2\sigma^2}\left(\sum x_i^2 - 2\mu\sum x_i\right)\right)$$

- So $h(x) = 1, c(\mu, \sigma) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right)$,
- $w_1(\mu, \sigma) = -\frac{1}{2\sigma^2}, w_2(\mu, \sigma) = \frac{\mu}{\sigma^2}, T_1(\boldsymbol{x}) = \sum x_i^2$, and $T_2(\boldsymbol{x}) = \sum x_i$.
- Thus $(T_1(\boldsymbol{X}), T_2(\boldsymbol{X})) = \left(\sum X_i^2, \sum X_i\right)$ is sufficient for $\boldsymbol{\theta} = (\mu, \sigma)$.
- By Lemma 2.2 $(T_1(\boldsymbol{X}), T_2(\boldsymbol{X})) = \left(\frac{n}{n-1}\sum(X_i - \bar{X})^2, \sum \bar{X}_i\right)$ is also sufficient for $\boldsymbol{\theta} = (\mu, \sigma)$.

■

Minimum Sufficient Statistics and Complete Statistics

MSS

- The theorem only show MSS, but does not help finding MSS
- Complete statistics help finding MSS
-

3 Minimum Sufficient Statistics

Definition 3.1 *A sufficient statistic $T(X)$ is called a minimal sufficient statistic if, for any other sufficient statistic $T'(X)$, $T(X)$is a function of $T'(X)$.*

Some notes:

- $T(\boldsymbol{x})$ is a function of $T'(\boldsymbol{x})$ if and only if $T'(\boldsymbol{x}) = T'(\boldsymbol{y}) \Rightarrow T(\boldsymbol{x}) = T(\boldsymbol{y})$.

- The MSS is not unique(ie. $(\sum X, \sum X^2)$and $(\bar{x}, s^2)$ both MMS for Normal). However, the minimal sufficient partition is unique! An MSS will provide a partition that is as coarse as any other sufficient statistic.

- By coarse we mean few partition elements.

show proof

Theorem 3.1 *Let $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ be the pmf or pdf of the sample $\mathbf{X}$. Suppose there exists a function $T(\boldsymbol{x})$ such that for every two sample points $\boldsymbol{x}$ and $\boldsymbol{y}$ the ratio $\frac{f(\boldsymbol{x}|\boldsymbol{\theta})}{f(\boldsymbol{y}|\boldsymbol{\theta})}$ is a constant as a function of $\boldsymbol{\theta}$ if and only if $T(\boldsymbol{x}) = T(\boldsymbol{y})$. Then $T(\boldsymbol{X})$ is a minimal sufficient statistic for $\boldsymbol{\theta}$.*

MMS theorem

Proof:

To simplify the proof, we assume $f(\mathbf{x} \mid \theta) > 0$ for all $\mathbf{x} \in \mathcal{X}$ and θ. First we show that $T(\mathbf{X})$ is a sufficient statistic. Let $\mathcal{T} = \{t : t = T(\mathbf{x}) \text{ for some } \mathbf{x} \in \mathcal{X}\}$ be the image of $\mathcal{X}$ under $T(\mathbf{x})$. Define the partition sets induced by $T(\mathbf{x})$ as $A_t = \{\mathbf{x} : T(\mathbf{x}) = t\}$. For each A_t, choose and fix one element $\mathbf{x}_t \in A_t$. For any $\mathbf{x} \in \mathcal{X}$, $\mathbf{x}_{T(\mathbf{x})}$ is the fixed element that is in the same set, A_t, as $\mathbf{x}$. Since $\mathbf{x}$ and $\mathbf{x}_{T(\mathbf{x})}$ are in the same set $A_t, T(\mathbf{x}) = T\left(\mathbf{x}_{T(\mathbf{x})}\right)$ and, hence, $f(\mathbf{x} \mid \theta)/f\left(\mathbf{x}_{T(\mathbf{x})} \mid \theta\right)$ is constant as a function of θ. Thus, we can define a function on $\mathcal{X}$ by $h(\mathbf{x}) = f(\mathbf{x} \mid \theta)/f\left(\mathbf{x}_{T(\mathbf{x})} \mid \theta\right)$ and h does not depend on θ. Define a function on $\mathcal{T}$ by $g(t \mid \theta) = f\left(\mathbf{x}_t \mid \theta\right)$. Then it can be seen that

$$f(\mathbf{x} \mid \theta) = \frac{f\left(\mathbf{x}_{T(\mathbf{x})} \mid \theta\right) f(\mathbf{x} \mid \theta)}{f\left(\mathbf{x}_{T(\mathbf{x})} \mid \theta\right)} = g(T(\mathbf{x}) \mid \theta) h(\mathbf{x})$$

and, by the Factorization Theorem, $T(\mathbf{X})$ is a sufficient statistic for θ.

Now to show that $T(\mathbf{X})$ is minimal, let $T'(\mathbf{X})$ be any other sufficient statistic. By the Factorization Theorem, there exist functions g' and h' such that $f(\mathbf{x} \mid \theta) = g'\left(T'(\mathbf{x}) \mid \theta\right) h'(\mathbf{x})$. Let $\mathbf{x}$ and $\mathbf{y}$ be any two sample points with $T'(\mathbf{x}) = T'(\mathbf{y})$. Then

$$\frac{f(\mathbf{x} \mid \theta)}{f(\mathbf{y} \mid \theta)} = \frac{g'\left(T'(\mathbf{x}) \mid \theta\right) h'(\mathbf{x})}{g'\left(T'(\mathbf{y}) \mid \theta\right) h'(\mathbf{y})} = \frac{h'(\mathbf{x})}{h'(\mathbf{y})}$$

Since this ratio does not depend on θ, the assumptions of the theorem imply that $T(\mathbf{x}) = T(\mathbf{y})$. Thus, $T(\mathbf{x})$ is a function of $T'(\mathbf{x})$ and $T(\mathbf{x})$ is minimal.

Example 3.1 *Example 6.2.14 (Normal minimal sufficient statistic)*

Let $X_1, \ldots, X_n$ be iid $\mathrm{n}\left(\mu, \sigma^2\right)$, both μ and σ^2 unknown. Let $\mathbf{x}$ and $\mathbf{y}$ denote two sample points, and let $\left(\bar{x}, s_{\mathbf{x}}^2\right)$ and $\left(\bar{y}, s_{\mathbf{y}}^2\right)$ be the sample means and variances corresponding to the $\mathbf{x}$ and $\mathbf{y}$ samples, respectively. Then, using (6.2.5), we see that the ratio of densities is

$$\begin{aligned}\frac{f\left(\mathbf{x} \mid \mu, \sigma^2\right)}{f\left(\mathbf{y} \mid \mu, \sigma^2\right)} &= \frac{\left(2 \pi \sigma^2\right)^{-n / 2} \exp \left(-\left[n(\bar{x}-\mu)^2+(n-1) s_{\mathbf{x}}^2\right] /\left(2 \sigma^2\right)\right)}{\left(2 \pi \sigma^2\right)^{-n / 2} \exp \left(-\left[n(\bar{y}-\mu)^2+(n-1) s_{\mathbf{y}}^2\right] /\left(2 \sigma^2\right)\right)} \\ &= \exp \left(\left[-n\left(\bar{x}^2-\bar{y}^2\right)+2 n \mu(\bar{x}-\bar{y})-(n-1)\left(s_{\mathbf{x}}^2-s_{\mathbf{y}}^2\right)\right] /\left(2 \sigma^2\right)\right) .\end{aligned}$$

This ratio will be constant as a function of μ and σ^2 if and only if $\bar{x}=\bar{y}$ and $s_{\mathbf{x}}^2=s_{\mathbf{y}}^2$. Thus, by Theorem 6.2.13, $\left(\bar{X}, S^2\right)$ is a minimal sufficient statistic for $\left(\mu, \sigma^2\right)$.

Lemma 3.2 *If $f(\boldsymbol{x}|\theta)$ involve indicator function, we can use $f(\boldsymbol{x}|\theta) = h(\boldsymbol{x}, \boldsymbol{y})f(\boldsymbol{y}|\theta)$ instead of ratio and show that if and only if some function of $h(\boldsymbol{x}, \boldsymbol{y}) =$ constant existed.*

Example 3.2 *why collapsing the sufficient statistics will not be a new sufficient statistics?* ■

Where the fuck is this?

Example 3.3 *Show that $S = (-1)^{X_1}T$ is also a sufficient statistics*

If you combine or collapse sufficient statistics, it is not guaranteed that the new statistic will satisfy the factorization theorem, in which case it would not be a sufficient statistic. In some cases, combining sufficient statistics can lead to loss of information about the parameter of interest. ■

Example 3.4 *Location Exponential: Show that $X_{(1)}$ is minimum sufficient* ■

Example 3.5 *Two expressions of minimal sufficient statistic for Uniform with location shift:*

Example 6.2.15 (Uniform minimal sufficient statistic) Suppose $X_1, \ldots, X_n$ are iid uniform observations on the interval $(\theta, \theta+1), -\infty<\theta<\infty$. Then the joint pdf of $\mathbf{X}$ is

$$f(\mathbf{x} \mid \theta)= \begin{cases}1 & \theta<x_i<\theta+1, i=1, \ldots, n \\ 0 & \text{otherwise}\end{cases}$$

which can be written as

$$f(\mathbf{x} \mid \theta)= \begin{cases}1 & \max_i x_i-1<\theta<\min_i x_i \\ 0 & \text{otherwise.}\end{cases}$$

Thus, for two sample points $\mathbf{x}$ and $\mathbf{y}$, the numerator and denominator of the ratio $f(\mathbf{x} \mid \theta) / f(\mathbf{y} \mid \theta)$ will be positive for the same values of θ if and only if $\min_i x_i=\min_i y_i$ and $\max_i x_i=\max_i y_i$. And, if the minima and maxima are equal, then the ratio is constant and, in fact, equals 1 . Thus, letting $X_{(1)}=\min_i X_i$ and $X_{(n)}=\max_i X_i$, we have that $T(\mathbf{X})=\left(X_{(1)}, X_{(n)}\right)$ is a minimal sufficient statistic. This is a case in which the dimension of a minimal sufficient statistic does not match the dimension of the parameter. ■

Lemma 3.3 *If there is a 1-1 relationship, then the function of minimal sufficient is also a minimal sufficient, aka. minimal sufficient statistics is not unique(partition on the other hands is unique)*

3.0.1 MSS for a family of distributions

Theorem 3.4 *Let $\mathcal{P}_0$ be a finite family of pdf's or pmf's, $f_0, f_1, \ldots, f_k$, all having the same support. Then, the statistic*

$$T(\boldsymbol{X}) = \left(\frac{f_1(\boldsymbol{X})}{f_0(\boldsymbol{X})}, \frac{f_2(\boldsymbol{X})}{f_0(\boldsymbol{X})}, \ldots, \frac{f_k(\boldsymbol{X})}{f_0(\boldsymbol{X})}\right)$$

is minimal sufficient for $\mathcal{P}_0$.

Proof:
Sufficiency

- Let $g_i(T(\boldsymbol{X})) = T_i(\boldsymbol{X}) = f_i(\boldsymbol{X})/f_0(\boldsymbol{X})$ and $g_0(T(\boldsymbol{X})) = 1$.
- $f_i(\boldsymbol{x}) = g_i(T(\boldsymbol{x}))f_0(\boldsymbol{x})$ for $i = 1, 2, \ldots, k$
- By the factorization theorem, T is sufficient for $\mathcal{P}_0$.

Minimal Sufficiency

- Suppose $S(\boldsymbol{X})$ is another sufficient statistic.
- $f_i(\boldsymbol{x}) = \tilde{g}_i(S(\boldsymbol{x}))h(\boldsymbol{x})$ for $i = 1, 2, \ldots, k$
- $T_i(\boldsymbol{x}) = f_i(\boldsymbol{x})/f_0(\boldsymbol{x}) = \tilde{g}_i(S(\boldsymbol{x}))/\tilde{g}_0(S(\boldsymbol{x}))$ for $i = 1, 2, \ldots, k$
- $T(\boldsymbol{x})$ is a function of $S(\boldsymbol{x})$, thus $T(\boldsymbol{X})$ is minimal sufficient for $\mathcal{P}_0$.

Lemma 3.5 *Lemma 5.2: If $\mathcal{P}$ is a family of distributions with common support, T is minimal sufficient for $\mathcal{P}_0 \subset \mathcal{P}$, and T is sufficient for $\mathcal{P}$, then T is minimal sufficient for $\mathcal{P}$.*

Proof:
If U is sufficient for $\mathcal{P}$, it is also sufficient for $\mathcal{P}_0$, and hence T is a function of U. This implies T is minimal sufficient for $\mathcal{P}$.

Lemma 3.6 *MSS in exponential families*

Let $X_1, \ldots, X_n$ be iid observations from an exponential family: $f(x \mid \boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{k} w_j(\boldsymbol{\theta})t_j(x)\right)$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \in \Theta$. The vector of canonical parameters is given by $\boldsymbol{w_\theta}' = (w_1(\boldsymbol{\theta}), \ldots, w_k(\boldsymbol{\theta}))$.

- *Recall that* $T(\boldsymbol{X}) = \left(\sum_{i=1}^{n} t_1(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i)\right)' \in \mathbb{R}^k$ *is a sufficient for* $\mathcal{P} = \{f(x \mid \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ *and note that* $\sum_{j=1}^{k} w_j(\boldsymbol{\theta}) t_j(x) = \boldsymbol{w}'_{\boldsymbol{\theta}} T(\boldsymbol{x})$.

- *Let* $\Theta_0 = \{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k\} \subset \Theta$ *and* $\mathcal{P}_0 = \{f(x \mid \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_0\}$.

- *By Theorem 5.1,* $S(\boldsymbol{X}) = \left(\frac{f(\boldsymbol{X}|\boldsymbol{\theta}_1)}{f(\boldsymbol{X}|\boldsymbol{\theta}_0)}, \ldots, \frac{f(\boldsymbol{X}|\boldsymbol{\theta}_k)}{f(\boldsymbol{X}|\boldsymbol{\theta}_0)}\right)'$ *is MSS for* $\mathcal{P}_0$*, where*

$$\frac{f(\boldsymbol{X} \mid \boldsymbol{\theta}_i)}{f(\boldsymbol{X} \mid \boldsymbol{\theta}_0)} = \exp\left(\left(\boldsymbol{w}'_{\boldsymbol{\theta}_{\mathrm{i}}} - \boldsymbol{w}'_{\boldsymbol{\theta}_0}\right) T(\boldsymbol{X})\right)$$

- *Define vectors* $\boldsymbol{d}_i = \boldsymbol{w}_{\boldsymbol{\theta}_{\mathrm{i}}} - \boldsymbol{w}_{\boldsymbol{\theta}_0} \in \mathbb{R}^k, i = 1, 2, \ldots, k$ *and the matrix* $\boldsymbol{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_{\boldsymbol{k}}] \in \mathbb{R}^k \times \mathbb{R}^k$

- *Thus*

$$\log S(\boldsymbol{X}) = \boldsymbol{D} T(\boldsymbol{X})$$

- *If the matrix* $\boldsymbol{D}$ *is full rank, then there is a 1-1 relationship between* $S(\boldsymbol{X})$ *and* $T(\boldsymbol{X})$ *which means that* $T(\boldsymbol{X})$ *is also MSS for* $\mathcal{P}_0$.

- *The full rank property means that the* $\boldsymbol{d}_i$ *'s are linearly independent.*

- *Since* $T(\boldsymbol{X})$ *is a sufficient for* $\mathcal{P}$ *and is MSS for* $\mathcal{P}_0$*, then by Lemma 5.2,* $T(\boldsymbol{X})$ *is MSS for* $\mathcal{P}$.

Ancillary Statistics

4 Ancillary Statistics and Complete Statistics

4.1 Ancillary Statistic

Definition 4.1 *A statistic $S(\boldsymbol{X})$ with distribution that does not depend on the parameter $\boldsymbol{\theta}$ is called an ancillary statistic.*

More precisely, a statistic $S(\boldsymbol{X})$ is ancillary for Θ if its distribution is the same for all $\boldsymbol{\theta} \in \Theta$. That is, $P_{\boldsymbol{\theta}}\{S(\boldsymbol{X}) \in A\}$ is constant for $\boldsymbol{\theta} \in \Theta$ for any set A.

Example 4.1 *Revisit to $N(\mu, \sigma^2)$*

- We have previously showed that $\left(\bar{X}, S_X^2\right)$ is sufficient for estimating $\boldsymbol{\theta} = (\mu, \sigma)$. (Actually, it is MSS.)

- Note that the distribution of S_X^2 depends on σ but not on μ.

- S_X^2 is ancillary for $\Theta_1 = \left\{\left(\mu, \sigma^2\right) : \sigma^2 = \sigma_0^2\right\}$. - Here $\bar{X}$ is MSS and need not be paired with S_X^2 to be sufficient for Θ_1.

Why

- S_X^2 is not ancillary for $\Theta_2 = \left\{\left(\mu, \sigma^2\right) : \sigma^2 > 0\right\}$. ■

Complete Statistics

- The binomial example is useful to understand and proof completeness

4.2 Complete statistics and completeness

- - Suppose that $X_1, X_2, \ldots, X_n$ are iid from $\mathcal{F}(\theta)$, a family of distributions indexed by θ.
- - Let $T(\boldsymbol{X})$ be a statistic and $u(T)$ a real-valued function of T so that $E_\theta[u(T)] = \theta$. That is, $u(T)$ is unbiased for θ. [We will cover unbiasedness in more detail later.]
- - Under what conditions is $u(T)$ the only function of T which is unbiased?
- - Let $u_1(T)$ and $u_2(T)$ be unbiased for θ
- - Define $g(T) = u_1(T) - u_2(T)$. Then $E_\theta[g(T)] = 0$ for all θ.
- - If the only function $g(T)$ that satisfies $E_\theta[g(T)] = 0$ is $g(T) = 0$, then this implies $u_1(T) = u_2(T)$ and $u(T)$ is unique for all θ

Definition 4.2 *Let $f(t \mid \boldsymbol{\theta})$ be a family of pdfs or pmfs for a statistic $T(\boldsymbol{X})$. The family of probability distributions is called complete if $E_{\boldsymbol{\theta}}[g(T)] = 0$ for all $\boldsymbol{\theta}$ implies $P_{\boldsymbol{\theta}}\{g(T) = 0\} = 1$, that is, $g(T) \equiv 0$, for all $\boldsymbol{\theta}$. Equivalently, $T(\boldsymbol{X})$ is called a complete statistic.*

Example 4.2 *6.2.22 (Binomial complete sufficient statistic)*

Suppose that T has a binomial(n, p) distribution, $0 < p < 1$. Let g be a function such that$E[g(T)] = 0$. Then

$$
\begin{aligned}
0 = \mathrm{E}_p g(T) &= \sum_{t=0}^{n} g(t) \binom{n}{t} p^t (1-p)^{n-t} \\
&= (1-p)^n \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t
\end{aligned}
$$

for all $p, 0 < p < 1$. The factor $(1-p)^n$ is not 0 for any p in this range. Thus it must be that

$$
0 = \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t = \sum_{t=0}^{n} g(t) \binom{n}{t} r^t
$$

make a polynomial for g(t)

for all $r, 0 < r < \infty$. But the last expression is a polynomial of degree n in r, where the coefficient of r^t is $g(t) \binom{n}{t}$. For the polynomial to be 0 for all r, each coefficient must be 0 . Since none of the $\binom{n}{t}$ terms is 0 , this implies that $g(t) = 0$ for $t = 0, 1, \ldots, n$. Since T takes on the values $0, 1, \ldots, n$ with probability 1 , this yields that $P_p(g(T) = 0) = 1$ for all p, the desired conclusion. Hence, T is a complete statistic.

■

Theorem 4.1 *If a minimal sufficient statistic exists, then any complete sufficient statistic is also a minimal sufficient statistic.*

Proof

- Let S be a minimal sufficient statistic and T be any complete sufficient statistic.
- Define $g_1(S) = E[T \mid S]$. This is actually a function of T since S is a MSS, say $S = h(T)$.
- Define $g(T) = T - g_1(S) = T - g_1(h(T))$.
- Clearly $E[g(T)] = 0$ for all $\boldsymbol{\theta}$ since $E[g_1(S)] = E[E[T \mid S]] = E[T]$.
- Since T is complete, then $g(T) = 0$ for all $\boldsymbol{\theta}$.
- This implies that $T = g_1(S)$. I don't understand this step
- Since S is minimally sufficient, then it is a function of every other sufficient statistic. That is, if S^* is a sufficient statistic, then $S = g_*(S^*)$ for some function $g_*(\cdot)$.
- Hence $T = g_1(g_*(S^*))$ is a function of S^*. Since S^* is any sufficient statistic, then T is minimally sufficient.

■

Example 4.3 *Not all minimal sufficient statistics is complete*
Recall our earlier example where $\boldsymbol{X}$ *is a random sample from* $\text{Uniform}(\theta, \theta+1)$.
- $T(\boldsymbol{X}) = \left(X_{(1)}, X_{(n)}\right)$ *is minimally sufficient for* θ.
- $R(\boldsymbol{X}) = X_{(n)} - X_{(1)} \sim \text{Beta}(n-1, 2)$ *is ancillary.* $E[R] = \frac{(n-1)}{(n+1)}$.
- *Define* $g(T) = R - \frac{(n-1)}{(n+1)}$.
- *Then* $E[g(T)] = 0$ *for all* θ, *but* $g(T) \neq 0$ *for all* θ. ■

A key feature of the above example is the existence of an ancillary statistic. Suppose $A(\boldsymbol{X}) = g_1(T(\boldsymbol{X}))$ is ancillary and let $E[A(\boldsymbol{X})] = a$, a constant, independent of θ. Define $g(T) = g_1(T(\boldsymbol{X})) - a$. Then $E[g(T)] = 0$ for all θ, but $g(T) \neq 0$ for all θ.

Example 4.4 *6.2.23 (Uniform complete sufficient statistic) Let* $X_1, \ldots, X_n$ *be iid uniform* $(0, \theta)$ *observations,* $0 < \theta < \infty$. *Using an argument similar to that in Example* 6.2.8, *we can see that* $T(\mathbf{X}) = \max_i X_i$ *is a sufficient statistic and, by Theorem* 5.4.4, *the pdf of* $T(\mathbf{X})$ *is*

$$f(t \mid \theta) = \begin{cases} n t^{n-1} \theta^{-n} & 0 < t < \theta \\ 0 & \textit{otherwise} \end{cases}$$

14

Suppose $g(t)$ is a function satisfying $\mathrm{E}_\theta g(T)=0$ for all θ. Since $\mathrm{E}_\theta g(T)$ is constant as a function of θ, its derivative with respect to θ is 0 . Thus we have that

$$\begin{aligned}
0=\frac{d}{d\theta}\mathrm{E}_\theta g(T)=\frac{d}{d\theta}\int_0^\theta g(t)nt^{n-1}\theta^{-n}dt& \\
&=\left(\theta^{-n}\right)\frac{d}{d\theta}\int_0^\theta ng(t)t^{n-1}dt+\left(\frac{d}{d\theta}\theta^{-n}\right)\int_0^\theta ng(t)t^{n-1}dt\\
&=\theta^{-n}ng(\theta)\theta^{n-1}+0\quad\left(\begin{array}{c}\text{applying the product}\\ \text{rule for differentiation}\end{array}\right)\\
&=\theta^{-1}ng(\theta)
\end{aligned}$$

The first term in the next to last line is the result of an application of the Fundamental Theorem of Calculus. The second term is 0 because the integral is, except for a constant, equal to $\mathrm{E}_\theta g(T)$, which is 0 . Since $\theta^{-1}ng(\theta)=0$ and $\theta^{-1}n\neq 0$, it must be that $g(\theta)=0$. This is true for every $\theta>0$; hence, T is a complete statistic. (On a somewhat pedantic note, realize that the Fundamental Theorem of Calculus does not apply to all functions, but only to functions that are Riemann-integrable. ■

Theorem 4.2 ***Basu'stheorem***
If $T(X)$ is a complete and minimal sufficient statistic, then $T(X)$ is independent of every ancillary statistic.

Proof:
- Define $g(t)=f_{S|T}(s\mid t)-f_S(s)$.

$$\begin{aligned}
\mathrm{E}_\theta[g(T)]&=\int_{-\infty}^{\infty}\left(f_{s|T}(s\mid t)-f_S(s)\right)f_T(t)dt\\
&=\int_{-\infty}^{\infty}f_{s|T}(s\mid t)f_T(t)dt-\int_{-\infty}^{\infty}f_S(s)f_T(t)dt\\
&=f_S(s)-f_S(s)\\
&=0
\end{aligned}$$

- Since T is complete, then $g(t)=0$ for all θ.
- Hence $f_{S|T}(s\mid t)=f_S(s)$ which implies that T and S are independent.

■

Theorem 4.3 *(Complete statistics in the exponential family)*
Let $X_1,\ldots,X_n$ be iid observations from an exponential family with pdf or pmf of the form

$$f(x\mid\boldsymbol{\theta})=h(x)c(\boldsymbol{\theta})\exp\left(\sum_{j=1}^k w\left(\theta_j\right)t_j(x)\right)$$

15

where $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$. Then the statistic

$$T(\mathbf{X}) = \left(\sum_{i=1}^{n} t_1(X_i), \sum_{i=1}^{n} t_2(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i)\right)$$

is complete as long as the parameter space Θ contains an open set in $\Re^k$.

The condition that the parameter space contain an open set is needed to avoid a situation like the following. The n $\left(\theta, \theta^2\right)$ distribution can be written in the form (6.2.7); however, the parameter space $\left(\theta, \theta^2\right)$ does not contain a two-dimensional open set, as it consists of only the points on a parabola. As a result, we can find a transformation of the statistic $T(\mathbf{X})$ that is an unbiased estimator of 0 (see Exercise 6.15). (Recall that exponential families such as the n $\left(\theta, \theta^2\right)$, where the parameter space is a lower-dimensional curve, are called curved exponential families; see Section 3.4.) The relationship between sufficiency, completeness, and minimality in exponential families is an interesting one. For a brief introduction, see Miscellanea 6.6.3.

see HW 2

Example 4.5 *Suppose X_1, X_2 are iid Exponential (β) and $Y = X_1 + X_2$.* - Exponential(β) *is an exponential family*

$$f_X(x) = \frac{1}{\beta} e^{\frac{-x}{\beta}} \mathbf{1}_{(0,\infty)}(x)$$

where $h(x) = \mathbf{1}_{(0,\infty)}(x), c(\beta) = \frac{1}{\beta}, w(\beta) = -\frac{1}{\beta}$, and $t(x) = x$. - Note that $w(\beta) = -\frac{1}{\beta}$, for $\beta > 0$, contains an open set in $\mathbb{R}$.

- Hence $Y = T(\boldsymbol{X}) = X_1 + X_2$ is a complete sufficient statistic.

Let $g(\boldsymbol{X}) = \frac{X_2}{Y}$. Find the value of $E_\beta[g(\boldsymbol{X})]$.

- $g(\boldsymbol{X})$ is an ancillary statistic: $\frac{X_i}{\beta} \sim$ Exponential(1), Thus $\frac{Y}{\beta} \sim$ Gamma$(2,1)$, Neither distribution dependent on β

the key is to find the distribution of the estimator

thus the distribution of

$$g(\boldsymbol{X}) = \frac{\left(\frac{X_2}{\beta}\right)}{\left(\frac{Y}{\beta}\right)} = \frac{X_2}{Y} \text{ does not depend on } \beta$$

- We know that $\beta = \mathrm{E}_\beta[X_2] = E_\beta[g(\boldsymbol{X})Y] = E_\beta[g(\boldsymbol{X})]E_\beta[Y] = 2\beta E_\beta[g(\boldsymbol{X})]$.

- Thus $E_\beta[g(\boldsymbol{X})] = \frac{1}{2}$

Basu's theorem: Y and $g(\boldsymbol{X})$ are independent

■

Estimator

MLE

Methods of Momment

Linear Regression

UMVUE

CRLB-Definition

- You always mess up how to take the expectation of fisher information
- Don't forget about the integral condition when takeing expectation
- You mess up the derivative of the bottom and the top
- fisher information **ONLY** applied to exponential family

6.2 Defining UMVUE

Example 6.2 - *Consider the family of estimators* $T_{\theta_0}(\boldsymbol{X}) \equiv \theta_0$ *for* $\theta_0 \in \Theta$.

- $\mathrm{MSE}_\theta\left[T_{\theta_0}\right]=\left(\theta_0-\theta\right)^2$, since $_{\mathrm{Var}_\theta}\left[T_{\theta_0}\right]=0$.

- *Hence* $\min_{\theta\in\Theta}\left(\mathrm{MSE}_\theta\left[T_{\theta_0}\right]\right)=0$, *since* $\mathrm{MSE}_{\theta_0}\left[T_{\theta_0}\right]=0$. - *If* $T^*(\boldsymbol{X})$ *is to be the uniformly minimum MSE estimator, then*

- $\mathrm{MSE}_\theta\left[T^*\right] \leq \min_{\theta\in\Theta}\left(\mathrm{MSE}_\theta\left[T_{\theta_0}\right]\right)=0$

- *That is,* $T^*(\boldsymbol{X})$ *must have zero variance and zero bias for all* $\theta \in \Theta$. *This is not possible, unless* $\Theta=\{\theta_0\}$.

So, if Θ *is not restricted, UBE may not exist* ■

Example 6.3 - *Define a class of estimators,* $\mathcal{C}$, *and find the estimator (or estimators) within* $\mathcal{C}$ *which has minimal MSE.*

- *Example: Suppose* $\boldsymbol{X}$ *is a sample point of size* n *from a* $N\left(\mu,\sigma^2\right)$ *family.*

- *Consider* $\mathcal{C}_{\sigma^2}=\left\{T_k(\boldsymbol{X})=\frac{\sum\left(X_i-\bar{X}\right)^2}{k}, k>0\right\}$, *a class of estimators of* σ^2.

$$
\begin{aligned}
T &= \frac{n-1}{k}S^2 \\
E[T] &= \frac{n-1}{k}\sigma^2 \\
Var[T] &= \frac{2(n-1)}{k^2}\sigma^4 \\
MSE_k[T] &= \frac{n-1-k}{k}^2\sigma^4+\frac{2(n-1)}{k^2}\sigma^4
\end{aligned}
$$

- *Argue that* $\underset{k}{\operatorname{argmin}}\left\{MSE_{\sigma^2}\left(T_k\right)\right\}=n+1$.

- *That is,* $\tilde{\sigma}^2=\frac{\sum^{n+1}_{(X_i-\bar{X})^2}}{n}$ *has the minimum MSE among estimators in* $\mathcal{C}_{\sigma^2}$.

■

Definition 6.3 - *Suppose* $\mathcal{C}_\theta=\{T(\boldsymbol{X}): \mathrm{E}_\theta[T(\boldsymbol{X})]=\theta$, *for all* $\theta\in\Theta\}$, *the class of all unbiased estimators of* θ.

- *Definition: An estimator* T^* *is a best unbiased estimator of* $\tau(\theta)$ *if it satisfies* $\mathrm{E}_\theta\left[T^*\right]=\tau(\theta)$ *for all* $\theta\in\Theta$ *and, for any other estimator* T *with* $\mathrm{E}_\theta[T]=\tau(\theta)$, *we have* $\mathrm{Var}_\theta\left[T^*\right] \leq \mathrm{Var}_\theta[T]$ *for all* θ. T^* *is also called a uniform minimum variance unbiased estimator (UMVUE) of* $\tau(\theta)$

6.2.1 CRLB

Definition 6.4 - *Let* $T=T(\boldsymbol{X})$ *be a statistic with* $\mathrm{E}_\theta[T]=g(\theta)$ *and* $\mathrm{Var}_\theta[T]<\infty$.

- For any random variable $W(\boldsymbol{X}, \theta)$ which has a finite second moment, the Cauchy-Schwarz inequality says that

$$(\operatorname{Cov}_\theta[T, W])^2 \leq \operatorname{Var}_\theta[T] \operatorname{Var}_\theta[W]$$

or equivalently,

$$\operatorname{Var}_\theta[T] \geq \frac{(\operatorname{Cov}_\theta[T, W])^2}{\operatorname{Var}_\theta[W]}$$

- This is a lower bound for the variance of $T(\boldsymbol{X})$.

*- **CRLB in general only apply to distribution which it's space doesn't dependent on the parameter***

Corollary 6.0.1 *The cleverness in this theorem follows from choosing X to be the estimator $W(\mathbf{X})$ and Y to be the quantity $\frac{\partial}{\partial \supsetneq} \log f(\mathbf{X} \mid \theta)$ and applying the Cauchy-Schwarz Inequality. First note that*

$$\begin{aligned}
\frac{d}{d\theta} \mathrm{E}_\theta W(\mathbf{X}) &= \int_X W(\mathbf{x}) \left[\frac{\partial}{\partial \theta} f(\mathbf{x} \mid \theta)\right] dx \\
&= \mathrm{E}_\theta \left[W(\mathbf{X}) \frac{\frac{\partial}{\partial \theta} f(\mathbf{X} \mid \theta)}{f(\mathbf{X} \mid \theta)}\right] \qquad \textit{(multiply by } f(\mathbf{X} \mid \theta)/f(\mathbf{X} \mid \theta)\textit{)}\Big) \\
&= \mathrm{E}_\theta \left[W(\mathbf{X}) \frac{\partial}{\partial \theta} \log f(\mathbf{X} \mid \theta)\right] \qquad \textit{(property of logs)}
\end{aligned}$$

which suggests a covariance between $W(\mathbf{X})$ and $\frac{\partial}{\partial \rho} \log f(\mathbf{X} \mid \theta)$. For it to be a covariance, we need to subtract the product of the expected values, so we calculate $\mathbf{E}_\theta \left(\frac{\partial}{\partial \partial} \log f(\mathbf{X} \mid \theta)\right)$. But if we apply (7.3.7) *with $W(\mathbf{x}) = 1$, we have*

$$\mathrm{E}_\theta \left(\frac{\partial}{\partial \theta} \log f(\mathbf{X} \mid \theta)\right) = \frac{d}{d\theta} \mathrm{E}_\theta[1] = 0.$$

- It can be shown that defining $W(\boldsymbol{X}, \theta)$ as

$$W(\boldsymbol{X}, \theta) = \frac{\partial}{\partial \theta} \log f(\boldsymbol{X} \mid \theta) = \frac{\frac{\partial}{\partial \theta} f(\boldsymbol{X} \mid \theta)}{f(\boldsymbol{X} \mid \theta)}$$

leads to the greatest lower bound.

$$\cdot \mathrm{E}_\theta[W] = \int W(\boldsymbol{x}, \theta) f(\boldsymbol{x} \mid \theta) d\boldsymbol{x} = \int \frac{\partial}{\partial \theta} f(\boldsymbol{x} \mid \theta) d\boldsymbol{x} = \frac{\partial}{\partial \theta} \int f(\boldsymbol{x} \mid \theta) d\boldsymbol{x} = \frac{\partial}{\partial \theta} 1 = 0$$

Therefore $\operatorname{Cov}_6 \left(W(\mathbf{X}), \frac{\partial}{\partial \theta} \log f(\mathbf{X} \mid \theta)\right)$ *is equal to the expectation of the product, and it follows from (7.3.7) and (7.3.8) that*

$$(7.3.9)\ \mathrm{Cov}_\theta\left(W(\mathbf{X}), \frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\right) = \mathrm{E}_\theta\left(W(\mathbf{X})\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\right) = \frac{d}{d\theta}\mathrm{E}_\theta W(\mathbf{X})$$

.

Also, since $\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\right) = 0$ *we have*

$$\mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\right) = \mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\right)^2\right) - 0^2.$$

Using the Cauchy-Schwarz Inequality together with (7.3.9) and (7.3.10), we obtain

$$\mathrm{Var}_\theta(W(\mathbf{X})) \geq \frac{\left(\frac{d}{\partial\hbar}\mathrm{E}_\theta W(\mathbf{X})\right)^2}{\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\right)^2\right)},$$

Theorem 6.1 *Crarmer-Rao Inequality*

Let $X_1, X_2, \ldots, X_n$ *be a sample from* $f(x\mid\theta)$ *and let* $T(\boldsymbol{X})$ *be any estimator satisfying*

$$\frac{\partial}{\partial\theta}\mathrm{E}_\theta[T(\boldsymbol{X})] = \int T(\boldsymbol{x})\frac{\partial}{\partial\theta}f(\boldsymbol{x}\mid\theta)d\boldsymbol{x}$$

and

$$\mathrm{Var}_\theta[T(\boldsymbol{X})] < \infty$$

Then

$$\mathrm{Var}_\theta[T] \geq \frac{\left(\frac{\partial}{\partial\theta}\mathrm{E}_\theta[T(\boldsymbol{X})]\right)^2}{\mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}\mid\theta)\right)^2\right]}$$

Corollary 6.1.1 *Corollary: If the conditions of the Cramér-Rao Inequality theorem are met and if* $X_1, X_2, \ldots, X_n$ *are iid with pdf* $f(x\mid\theta)$*, then*

$$\mathrm{Var}_\theta[T] \geq \frac{\left(\frac{\partial}{\partial\theta}\mathrm{E}_\theta[T(\boldsymbol{X})]\right)^2}{n\mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X\mid\theta)\right)^2\right]}$$

Proof:

$$\begin{aligned}
\mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\right)^2\right] &= \mathrm{E}_\theta\left[\left(\sum_i \frac{\partial}{\partial\theta}\log f\left(X_i\mid\theta\right)\right)^2\right] \\
&= \mathrm{E}_\theta\left[\sum_i\sum_j\left(\frac{\partial}{\partial\theta}\log f\left(X_i\mid\theta\right)\right)\left(\frac{\partial}{\partial\theta}\log f\left(X_j\mid\theta\right)\right)\right] \\
&= \sum_i\sum_j\ \mathrm{F}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f\left(X_i\mid\theta\right)\right)\left(\frac{\partial}{\partial\theta}\log f\left(X_j\mid\theta\right)\right)\right] \\
&= \sum_i \mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f\left(X_i\mid\theta\right)\right)^2\right] \\
&= n\mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X\mid\theta)\right)^2\right]
\end{aligned}$$

For $i \neq j$ we have

$$\begin{aligned}
\mathrm{E}_\theta\quad & \left(\tfrac{\partial}{\partial\theta}\log f\left(X_i\mid\theta\right)\tfrac{\partial}{\partial\theta}\log f\left(X_j\mid\theta\right)\right) \\
& = \mathrm{E}_\theta\left(\tfrac{\partial}{\partial\theta}\log f\left(X_i\mid\theta\right)\right)\mathrm{E}_\theta\left(\tfrac{\partial}{\partial\theta}\log f\left(X_j\mid\theta\right)\right) \qquad \text{(independence)} \\
& = 0. \qquad \text{(from (7.3.8))}
\end{aligned}$$

Example 6.4 *Example 7.3.13 (Unbiased estimator for the scale uniform)*
let $X_{1\ldots\ldots}$, X_n *be iid with pdf* $f(x\mid\theta) = 1/\theta, 0 < x < \theta$. *Since of* $\log f(x\mid\theta) = -1/\theta$, *we have*

$$E_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X\mid\theta)\right)^2\right) = \frac{1}{\theta^2}.$$

The Cramer-Rao Theorem would seem to indicate that if W *is any unbiased estimator of* θ_1

$$Var\ w \geq \frac{\theta^2}{n}$$

We would now like to find an unbiased estimator with small variance. As a firt goeos, consider the sufficient statistic $Y = \max\left(X_2\ldots, X_\infty\right)$, *the largest order statistic. The pdf of* Y *is* $fr(y\mid\theta) = ny^{n-1}/\theta^n, 0 < y < \theta$, *so*

$$E_0 Y = \int_0^\infty \frac{ny^n}{\theta^n}dy = \frac{n}{n+1}\theta_1$$

showing that $\frac{n+1}{n}Y$ *is an unbiased estimator of* θ*. We next calculate*

$$\begin{aligned}\operatorname{Var}_\theta\left(\frac{n+1}{n}Y\right) &= \left(\frac{n+1}{n}\right)^2 \operatorname{Var}_\theta Y \\ &= \left(\frac{n+1}{n}\right)^2 \left[\mathrm{E}_\varphi Y^2 - \left(\frac{n}{n+1}\theta\right)^2\right] \\ &= \left(\frac{n+1}{n}\right)^2 \left[\frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\theta\right)^2\right] \\ &= \frac{1}{n(n+2)}\theta^2\end{aligned}$$

which is uniformly smaller than θ^2/n*. This indicnten that the Cramer-Rao Theorem is not applicable to this pdf. To see that this is so, we can use Leibnitz's Rule (Section 2.4) to calculate*

$$\begin{aligned}\frac{d}{d\theta}\int_0^\theta h(x)f(x\mid\theta)dx &= \frac{d}{d\theta}\int_0^\theta h(x)\frac{1}{\theta}dx \\ &= \frac{h(\theta)}{\theta} + \int_0^\theta h(x)\frac{\partial}{\partial\theta}\left(\frac{1}{\theta}\right)dx \\ &\neq \int_0^\theta h(x)\frac{\partial}{\partial\theta}f(x\mid\theta)dx\end{aligned}$$

unless $h(\theta)/\theta = 0$ *for all* θ*. Hence, the Cramér-Rao Theorem does not apply. In general, if the range of the pdf depends on the parameter, the theorem will not be applicable.* ■

Example 6.5 *Estimate* β *for Exponential* (β) *family*

- *Assume we have an iid sample:* $f(x\mid\beta) = \frac{1}{\beta^n}\exp\left(-\sum x/\beta\right)$
- $\log f(x\mid\beta) = -n\log\beta - \sum x/\beta = -n\log\beta - n\bar{x}/\beta$
- *Score statistic:* $\frac{\partial[\log f(\boldsymbol{X}\mid\beta)]}{\partial\beta} = -\frac{n}{\beta} + \frac{n\bar{X}}{\beta^2}$*. Is the expectation equal to 0 ?*
- $\mathrm{E}_\beta\left[\left(\frac{\partial}{\partial\beta}\log f(\boldsymbol{X}\mid\beta)\right)^2\right] = \operatorname{Var}_\beta\left[-\frac{n}{\beta} + \frac{n\bar{X}}{\beta^2}\right] = \frac{n^2}{\beta^4}\operatorname{Var}_\beta[\bar{X}] = \frac{n^2}{\beta^4}\frac{\beta^2}{n} = \frac{n}{\beta^2}$
- $\frac{\partial}{\partial\beta}\mathrm{E}_\beta[\bar{X}] = \frac{\partial}{\partial\beta}\beta = 1$
- *So the CRLB is* β^2/n *which is equal to* $\operatorname{Var}_\beta[\bar{X}]$*. So* $\bar{X}$ *is UMVUE for* β*.*

■

Theorem 6.2 *Attainment*

Let $X_1, \ldots, X_n$ be iid $f(x \mid \theta)$, where $f(x \mid \theta)$ satisfies the conditions of the Cramér-Rao Theorem. Let $L(\theta \mid \mathbf{x}) = \prod_{i=1}^{n} f(x_i \mid \theta)$ denote the likelihood function. If $W(\mathbf{X}) = W(X_1, \ldots, X_n)$ is any unbiased estimator of $\tau(\theta)$, then $W(\mathbf{X})$ attains the Cramér-Rao Lower Bound if and only if

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

for some function $a(\theta)$.

Proof:

The Cramér-Rao Inequality, as given in (7.3.6), can be written as

$$\left[\operatorname{Cov}_\theta\left(W(\mathbf{X}), \frac{\partial}{\partial\theta} \log \prod_{i=1}^{n} f(X_i \mid \theta)\right)\right]^2 \leq \operatorname{Var}_\theta W(\mathbf{X}) \operatorname{Var}_\theta\left(\frac{\partial}{\partial\theta} \log \prod_{i=1}^{n} f(X_i \mid \theta)\right),$$

and, recalling that

$$\mathrm{E}_\theta W = \tau(\theta)$$

,

$$\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta} \log \prod_{i=1}^{n} f(X_i \mid \theta)\right) = 0$$

, and using the results of Theorem 4.5.7, we can have equality if and only if $W(\mathbf{x}) - \tau(\theta)$ is proportional to $\frac{\partial}{\partial\theta} \log \prod_{i=1}^{n} f(x_i \mid \theta)$. That is exactly what is expressed in (7.3.12).

*Theorem 4.5.7 For any random variables X and Y,

a. $-1 \leq \rho_{XY} \leq 1$.

b. $|\rho_{XY}| = 1$ if and only if there exist numbers $a \neq 0$ and b such that $P(Y = aX + b) = 1$. If $\rho_{XY} = 1$, then $a > 0$, and if $\rho_{XY} = -1$, then $a < 0$.

Example 6.6 *Example: Estimate β for Exponential (β) family*

- *Recall-Score statistic:* $\frac{\partial[\log f(x|\beta)]}{\partial\beta} = -\frac{n}{\beta} + \frac{n\bar{x}}{\beta^2}$
- *Define* $a(\beta) = \frac{n}{\beta^2}$
- *Then* $\frac{\partial}{\partial\theta_\beta} \log f(x \mid \beta) = a(\beta)(\bar{X} - \beta)$ *So* $\operatorname{Var}_\beta[\bar{X}]$ *attains the CRLB* $= \beta^2/n$ *and, again,* $\bar{X}$ *is UMVUE for* β

Corollary 6.2.1 *if the range of the pdf depends on the parameter, the theorem will not be applicable.*

Definition 6.5 *Fisher's Information*

The quantity

$$I(\theta) = \mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X} \mid \theta)\right)^2\right] \underbrace{= -\mathrm{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\log f(X \mid \theta)\right)}_{\textit{true for exponential family}}$$

is called the Fisher's Information. $I(\theta)$ depends on the particular parameterization of the model family. Suppose that $\theta = h(v)$, where $h(\cdot)$ is differentiable, then the information that $\boldsymbol{X}$ contains about v is

$$I^*(v) = I(h(v)) \cdot \left[h'(v)\right]^2$$

Example 6.7 *Example 7.3.14 (Normal variance bound) Let $X_1, \ldots, X_n$ be iid $\mathrm{n}\left(\mu, \sigma^2\right)$, and consider estimation of σ^2, where μ is unknown. The normal pdf satisfies the assumptions of the Cramér-Rao Theorem and Lemma 7.3.11, so we have*

$$\frac{\partial^2}{\partial\left(\sigma^2\right)^2}\log\left(\frac{1}{\left(2\pi\sigma^2\right)^{1/2}}e^{-(1/2)(x-\mu)^2/\sigma^2}\right) = \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6}$$

and

$$\begin{aligned}
-\mathrm{E}\left(\frac{\partial^2}{\partial\left(\sigma^2\right)^2}\log f\left(X \mid \mu, \sigma^2\right) \mid \mu, \sigma^2\right) &= -\mathrm{E}\left(\frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6} \mid \mu, \sigma^2\right) \\
&= -\frac{1}{2\sigma^4} + \frac{1}{\sigma^4} \\
&= \frac{1}{2\sigma^4}.
\end{aligned}$$

Thus, any unbiased estimator, W, of σ^2 must satisfy

$$\operatorname{Var}\left(W \mid \mu, \sigma^2\right) \geq \frac{2\sigma^4}{n}$$

■

33

Rao-Blackwell-Finding UMVUE(Must be Unbiased)

- Find by condition
- Find by proving it is unique/same
- How to find the conditional expectation?
- 2021 Theory 5a-c is good practice

6.3 Finding UMVUE

6.3.1 Rao-Blackwell theorem

Theorem 6.3 *Rao-Blackwell*

Let W be an estimator of $\tau(\boldsymbol{\theta})$ where $\mathrm{E}_{\boldsymbol{\theta}}[W] = \eta(\boldsymbol{\theta})$. Also let T be a sufficient statistic for $\boldsymbol{\theta}$, possibly vector valued. Define $\phi(T) = \mathrm{E}_{\boldsymbol{\theta}}[W \mid T]$. Then

- *$\phi(T)$ is a statistic which is a function of T.*
- $\mathrm{E}_{\boldsymbol{\theta}}[\phi(T)] = \eta(\boldsymbol{\theta})$,
- $\mathrm{MSE}_{\tau(\boldsymbol{\theta})}[\phi(T)] \leq \mathrm{MSE}_{\tau(\boldsymbol{\theta})}[W]$ *for all $\boldsymbol{\theta}$, and*
- $\mathrm{MSE}_{\tau(\boldsymbol{\theta})}[\phi(T)] < \mathrm{MSE}_{\tau(\boldsymbol{\theta})}[W]$ *for some $\boldsymbol{\theta}$ unless $\phi(T) = W$ with probability 1.*

Proof

1. Since T is sufficient then $f(\boldsymbol{x} \mid T, \boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$. W is a statistic that does not depend on $\boldsymbol{\theta}$, so $\mathrm{E}_{\boldsymbol{\theta}}[W \mid T]$ does not depend on $\boldsymbol{\theta}$.
2. $\mathrm{E}_{\boldsymbol{\theta}}[\phi] = \mathrm{E}_{\boldsymbol{\theta}}[\mathrm{E}[W \mid T]] = \mathrm{E}_{\boldsymbol{\theta}}[W] = \eta(\boldsymbol{\theta})$
3. $$\begin{aligned} \mathrm{MSE}_{\boldsymbol{\tau(\theta)}}[W] &= \mathrm{Var}_{\boldsymbol{\theta}}[W] + (\eta(\boldsymbol{\theta}) - \tau(\boldsymbol{\theta}))^2 \\ &= \mathrm{Var}_{\boldsymbol{\theta}}[\mathrm{E}[W \mid T]] + \mathrm{E}_{\boldsymbol{\theta}}[\mathrm{Var}[W \mid T]] + (\eta(\boldsymbol{\theta}) - \tau(\boldsymbol{\theta}))^2 \\ &\geq \mathrm{Var}_{\boldsymbol{\theta}}[\phi] + (\eta(\boldsymbol{\theta}) - \tau(\boldsymbol{\theta}))^2 \\ &= \mathrm{MSE}_{\boldsymbol{\tau(\theta)}}[\phi] \end{aligned}$$

Theorem 6.4 *Rao-Blackwell for Unbiased estimator(by proving equal)*

- *Suppose W is an unbiased estimator, that is,* $\mathrm{E}_{\boldsymbol{\theta}}[W] = \eta(\boldsymbol{\theta}) = \tau(\boldsymbol{\theta})$.
- *So, we can start with an unbiased estimator, W, and get a new estimator, $\phi(T) = \mathrm{E}_{\boldsymbol{\theta}}[W \mid T]$, that has variance that is no larger than the variance of W, and possibly smaller!*
- *Actually,* $\mathrm{MSE}_{\boldsymbol{\theta}}[W] = \mathrm{MSE}_{\boldsymbol{\theta}}[\phi]$ *for all $\boldsymbol{\theta} \in \Theta$ if and only if*
$$\mathrm{E}_{\boldsymbol{\theta}}[\mathrm{Var}[W \mid T]] = 0$$
as seen in the proof of the theorem.
- $\mathrm{E}_{\boldsymbol{\theta}}[\mathrm{Var}[W \mid T]] = 0 \Rightarrow \mathrm{Var}[W \mid T] = 0$, *since* $\mathrm{Var}[W \mid T] \geq 0$
- *Thus,* $\mathrm{E}_{\boldsymbol{\theta}}\left[(W - \phi)^2 \mid T\right] = 0 \Rightarrow W = \phi$.

- *If the Blackwellization of W yields an estimator different from W, then the new estimator has a smaller MSE.*

Example 6.8 *Suppose $X_1, X_2, \ldots, X_n \sim$ iid Poisson (θ). Let $\tau(\theta) = P\{X = 0\} = e^{-\theta}$.*

- *Consider the statistic $W = \mathbf{1}_{\{0\}}(X_1) \sim \text{Bernoulli}(\tau(\theta))$*
- $\mathrm{E}_\theta[W] = \tau(\theta)$*, that is, W is unbiased for $\tau(\theta)$.*
- $\mathrm{Var}_\theta[W] = \tau(\theta)(1 - \tau(\theta))$
- *From previous work we know that $T = \sum X_i$ is sufficient for θ. Note that $T \sim \text{Poisson}(n\theta)$.*
- *Find a better unbiased estimator, in terms of the variance.*
- *Confirm this estimator is unbiased.*
- *Is this estimator UMVUE?*

$$
\begin{aligned}
\Phi(s) = E[W|T = s] &= P(X_1 = 0|T = s) \\
&= \frac{P(X_1 = 0, \sum_1 X_i = s)}{P(T = s)} \\
&= \frac{P(X_1 = 0) * \underbrace{P(\sum_2 X_i = s)}_{Poi((n-1)\theta)}}{P(T = s)} \\
&= (\frac{n-1}{n})^s \\
\Phi(T) &= (\frac{n-1}{n})^{\sum X_i}
\end{aligned}
$$

This is a better estimator by Rao-blackwell.

$$
\begin{aligned}
E[\Phi(T)] = E[\frac{n-1}{n}^{\sum X_i}] &= E[exp(log(\frac{n-1}{n})\sum X_i)] \\
by\ moment\ generate\ function: M_T(t) &= exp(n\theta(e^t - 1)) \\
= M(log(\frac{n-1}{n})) &= exp[n\theta(\frac{n-1}{n} - 1)] \\
&= exp(-\theta)
\end{aligned}
$$

This shows that Φ is unbiased.

$$
\begin{aligned}
Var(\Phi) &= \underbrace{E[\Phi^2]}_{E[\frac{n-1}{n}^{2T}]=M(2log(\frac{n-1}{n}))} - \underbrace{(E[\Phi])^2}_{exp(-2\theta)} \\
&= exp(-\theta) * exp(-\frac{n-1}{n}\theta) - exp(-2\theta) \\
&< exp(-\theta) * \underbrace{(1 - exp(-\theta))}_{\frac{n-1}{n}\theta > 0} \\
&= Var(W)
\end{aligned}
$$

This shows Rao-blackwell is true for this case and we have a better unbiased estimator. Finding the lower bound:

$$CRLB = \frac{\theta exp(-2\theta)}{n}$$

■

Theorem 6.5 *UMVUE is unique*
If W is a UMVUE of $\tau(\boldsymbol{\theta})$, then W is unique.

Proof: (A review of the proof in C&B, p. 343-344.) Suppose that both W and W' are UMVUE of $\tau(\boldsymbol{\theta})$. Then $W^* = \frac{(W+W')}{2}$ is unbiased for $\tau(\boldsymbol{\theta})$. By supposition $\mathrm{Var}_{\boldsymbol{\theta}} W' = \mathrm{Var}_{\boldsymbol{\theta}} W$. This implies

$$
\begin{aligned}
\mathrm{Var}_{\boldsymbol{\theta}} W^* &= \mathrm{Var}_{\boldsymbol{\theta}}\left[\frac{1}{2}W + \frac{1}{2}W'\right] \\
&= \frac{1}{4}\mathrm{Var}_{\boldsymbol{\theta}} W + \frac{1}{4}\mathrm{Var}_{\boldsymbol{\theta}} W' + \frac{1}{2}\mathrm{Cov}_{\boldsymbol{\theta}}\left[W, W'\right]
\end{aligned}
$$

$\mathrm{Cov}_{\boldsymbol{\theta}}\left[W, W'\right] \leq \sqrt{\mathrm{Var}_{\boldsymbol{\theta}} W \cdot \mathrm{Var}_{\theta} W'} \longrightarrow \leq \mathrm{Var}_{\boldsymbol{\theta}} W$

However, $\mathrm{Var}_{\boldsymbol{\theta}} W$ is the smallest variance among unbiased estimators, thus $\mathrm{Var}_{\boldsymbol{\theta}} W^* = \mathrm{Var}_{\boldsymbol{\theta}} W$.

The last result implies $\mathrm{Cov}_{\boldsymbol{\theta}}[W, W'] = \sqrt{\mathrm{Var}_{\boldsymbol{\theta}} W \cdot \mathrm{Var}_{\boldsymbol{\theta}} W'} = \mathrm{Var}_{\boldsymbol{\theta}} W$. We noted before that this implies $W' = aW + b, a \neq 0$, where a and b may depend on the parameters, but not the data. - $\mathrm{E}_{\boldsymbol{\theta}}[W'] = a\mathrm{E}_{\boldsymbol{\theta}}[W] + b = a\tau(\boldsymbol{\theta}) + b$. But $\mathrm{E}_{\boldsymbol{\theta}}[W'] = \tau(\boldsymbol{\theta})$. So

$$(a-1)\tau(\boldsymbol{\theta}) + b = 0$$

- $\mathrm{Var}_{\boldsymbol{\theta}}[W'] = a^2\,\mathrm{Var}_{\boldsymbol{\theta}}[W]$.

But $\mathrm{Var}_{\boldsymbol{\theta}}[W'] = \mathrm{Var}_{\boldsymbol{\theta}}[W]$, thus $a = 1$ and from above, $b = 0$.

if $a = -1$, $b = 2\tau(\theta)$

Thus $W' = W$ and the UMVUE is unique.

Theorem 6.6 *The unbiased estimator W is the UMVUE of $\tau(\boldsymbol{\theta})$ if and only if W is uncorrelated with all unbiased estimators of 0 .*

Proof: (This is an expansion of the proof in C&B, p. 344-345.)

($\Rightarrow$) Suppose W is the UMVUE of $\tau(\boldsymbol{\theta})$. Let $V = W + aU$ where $E_{\boldsymbol{\theta}}[U] = 0, \mathrm{Var}_{\boldsymbol{\theta}}[U] = 1$, and a is any non-zero real constant. Then

$$\mathrm{E}_{\boldsymbol{\theta}}[V] = \tau(\boldsymbol{\theta})$$

and

$$\mathrm{Var}_{\boldsymbol{\theta}}[V] = \mathrm{Var}_{\boldsymbol{\theta}}[W] + a^2 \,\mathrm{Var}_{\boldsymbol{\theta}}[U] + 2\,\mathrm{aCov}_{\boldsymbol{\theta}}[W, U] \geq \mathrm{Var}_{\boldsymbol{\theta}}[W]$$

So $a + 2\,\mathrm{Cov}_{\boldsymbol{\theta}}[W, U] > 0$ which implies $|\mathrm{Cov}_{\boldsymbol{\theta}}[W, U]| \leq \frac{|a|}{2}$ for all $a \neq 0$. This inequality holds as $|a| \to 0$, thus $\mathrm{Cov}_{\boldsymbol{\theta}}[W, U] = 0$ and $Var_\theta(V) \to Var_\theta(W)$.

($\Leftarrow$) Let V be an estimator such that $\mathrm{E}_{\boldsymbol{\theta}}[V] = \tau(\boldsymbol{\theta})$ Note that

$$V = W + (V - W) \text{ and } \mathrm{E}_{\boldsymbol{\theta}}[V - W] = 0$$

Thus, by supposition,

$$\mathrm{Cov}_{\boldsymbol{\theta}}[W, V - W] = 0$$

The variance of V is thus

$$\mathrm{Var}_{\boldsymbol{\theta}}[V] = \mathrm{Var}_{\boldsymbol{\theta}}[W] + \mathrm{Var}_{\boldsymbol{\theta}}[V - W] \geq \mathrm{Var}_{\boldsymbol{\theta}}[W]$$

Since this applies for any unbiased estimator V, W is the UMVUE of $\tau(\boldsymbol{\theta})$.

Theorem 6.7 *Suppose that the UMVUE exists, then it is unique and is a function of a sufficient statistic.*

Proof:

We already have that the UMVUE, if it exists, is unique. Suppose W is the UMVUE of $\tau(\boldsymbol{\theta})$ and T is a sufficient statistic for $\boldsymbol{\theta}$. Then by the Rao-Blackwell theorem, $\phi(T) = \mathrm{E}_{\boldsymbol{\theta}}[W \mid T]$ is unbiased and $\mathrm{Var}_\theta[\phi(T)] \leq \mathrm{Var}_{\boldsymbol{\theta}}[W]$. Since $\mathrm{Var}_{\boldsymbol{\theta}}[W]$ is the minimal achievable variance among unbiased estimators of $\tau(\boldsymbol{\theta})$, $\mathrm{Var}_{\boldsymbol{\theta}}[\phi(T)] = \mathrm{Var}_{\boldsymbol{\theta}}[W]$. This implies that $\phi(T)$ is a UMVUE of $\tau(\boldsymbol{\theta})$. Due to uniqueness,

$$\phi(T) = W.$$

Thus, the UMVUE is a function of a sufficient statistic.

Lehmann-scheffe-Finding unique UMVUE

- Completeness
- More example see *Mathematics Statistics*

6.3.2 Lehmann-scheffe

Theorem 6.8 *Lehmann-Scheffé Theorem*

Let T be any complete sufficient statistic for the parameter $\boldsymbol{\theta}$, and let $\phi(T)$ be any estimator based only on T. Then $\phi(T)$ is the unique UMVUE of $\mathrm{E}_{\boldsymbol{\theta}}[\phi(T)]$ (CRLB is not garenteed)

Theorem 6.9 *Theorem 3.1 (Lehmann-Scheffé theorem) From Mathematics Statistics.*

Suppose that there exists a sufficient and complete statistic $T(X)$ for $P \in \mathcal{P}$. If ϑ is estimable, then there is a unique unbiased estimator of ϑ that is of the form $h(T)$ with a Borel function h. (Two estimators that are equal a.s. $\mathcal{P}$ are treated as one estimator.) Furthermore, $h(T)$ is the unique UMVUE of ϑ.

Proof:

In the last theorem, we required T to be a sufficient statistic for $\boldsymbol{\theta}$. Suppose that T is also complete. Now consider $\phi(T)$ to be an unbiased estimator of $\tau(\boldsymbol{\theta})$

- Let $W(T)$ be any other unbiased estimator of $\tau(\boldsymbol{\theta})$ which is a function of T.
- $\mathrm{E}_{\boldsymbol{\theta}}[\phi(T) - W(T)] = 0$ for all $\boldsymbol{\theta} \in \Theta$
- By the completeness of $T, \phi(T) - W(T) = 0$ for all $\boldsymbol{\theta} \in \Theta$
- Hence $\phi(T) = W(T)$; that is, $\phi(T)$ is unique.
- So, there is at most one unbiased estimator of $\tau(\boldsymbol{\theta})$ which is a function of a complete sufficient statistic.
- This means for any unbiased estimator of $\tau(\boldsymbol{\theta})$, say W', we have $\phi(T) = \mathrm{E}_{\boldsymbol{\theta}}[W' \mid T]$ and $\mathrm{Var}_{\boldsymbol{\theta}}[\phi(T)] \leq \mathrm{Var}_{\boldsymbol{\theta}}[W']$
- Thus, $\phi(T)$ is the unique UMVUE!

This theorem is a consequence of Theorem 2.5(ii) (Rao-Blackwell theorem). One can easily extend this theorem to the case of the uniformly minimum risk unbiased estimator under any loss function $L(P, a)$ that is strictly convex in a. The uniqueness of the UMVUE follows from the completeness of $T(X)$

There are two typical ways to derive a UMVUE when a sufficient and complete statistic T is available. The first one is solving for h when the distribution of T is available. The following are two typical examples.

Example 6.9 *Example 3.1.*

Let $X_1, \ldots, X_n$ be i.i.d. from the uniform distribution on $(0, \theta), \theta > 0$. Let $\vartheta = g(\theta)$, where g is a differentiable function on $(0, \infty)$. Since the sufficient and complete statistic

$X_{(n)}$ has the Lebesgue p.d.f. $n\theta^{-n}x^{n-1}I_{(0,\theta)}(x)$, an unbiased estimator $h\left(X_{(n)}\right)$ of ϑ must satisfy

$$\theta^n g(\theta) = n\int_0^\theta h(x)x^{n-1}dx \quad \textit{for all } \theta > 0.$$

Differentiating both sizes of the previous equation and applying the result of differentiation of an integral (Royden (1968, §5.3)) lead to

$$n\theta^{n-1}g(\theta) + \theta^n g'(\theta) = nh(\theta)\theta^{n-1}.$$

Hence, the UMVUE of ϑ is $h\left(X_{(n)}\right) = g\left(X_{(n)}\right) + n^{-1}X_{(n)}g'\left(X_{(n)}\right)$. In particular, if $\vartheta = \theta$, then the UMVUE of θ is $\left(1+n^{-1}\right)X_{(n)}$

Example 6.10 • *For some $\tau(\boldsymbol{\theta})$ there is no unbiased estimator.*

- *Let $X \sim \text{Binomial}(m,\theta)$ and $\tau(\theta) = \ln\left(\frac{\theta}{1-\theta}\right)$, the log-odds.*
- *Suppose T is any estimator of $\tau(\theta)$.*
- $\mathrm{E}_\theta[T] = \sum_{k=0}^m T(k)\begin{pmatrix} m \\ k \end{pmatrix}\theta^k(1-\theta)^{m-k}$ *is an m^{th}-degree polynomial in θ.*
- $\tau(\theta) = \ln\left(\frac{\theta}{1-\theta}\right)$ *cannot be expressed as a finite-degree polynomial.*
- *Thus, T cannot be unbiased.*
- *Since T was an arbitrary estimator, no unbiased estimator of $\tau(\theta)$ exists.*
- *For some situations there is an unbiased estimator, but no UMVUE.* ■

Consistency, Asymptotic Variance, efficency of MLE

7.2 Large Sample Property (and CI)

Definition 7.5 *Consistency*

A sequence of estimators $W_n = W_n(\boldsymbol{X})$ is a consistent sequence of estimators of the parameter θ if, for every $\epsilon > 0$ and every $\theta \in \Theta$,

$$\lim_{n\to\infty} P_\theta\left\{|W_n - \theta| < \epsilon\right\} = 1$$

- *That is, $W_n \to \theta$ in probability.*
- *For large n, almost all possible values of $W_n(\boldsymbol{X})$ are close to θ.*
- *W_n is said to be a consistent estimator of θ.*
- *Equivalently, $\lim_{n\to\infty} P_\theta\left\{|W_n - \theta| \geq \epsilon\right\} = 0$.*

Example 7.9 -

Consider $X_1, \ldots, X_n$ as an iid random sample from Bernoulli(p).*Is $\hat{p} = \bar{X}$ is a consistent estimator of p ?*

We know that when n is large, $\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$ has an approximate $N(0,1)$ distribution; the larger the sample size, the better the approximation. Consider

$$P_p\{|\hat{p} - p| < \epsilon\} = P_p\left\{\left|\frac{\hat{p} - p}{\sqrt{p(1-p)/n}}\right| < \frac{\epsilon}{\sqrt{p(1-p)/n}}\right\} \approx P\{|Z| < \sqrt{n}K\}$$

where $K = \epsilon/\sqrt{p(1-p)}$ and Z is a standard normal variable. Thus

$$\lim_{n\to\infty} P_p\{|\hat{p} - p| < \epsilon\} \approx \lim_{n\to\infty} P\{|Z| < \sqrt{n}K\} = P\{|Z| < \infty\} = 1$$

Thus $\hat{p}$ is a consistent estimator of p.
■

Theorem 7.3 *If W_n is a sequence of estimators of a parameter θ satisfying*

$$\lim_{n\to\infty} \mathrm{Var}_\theta\left[W_n\right] = 0$$
$$\lim_{n\to\infty} \mathrm{Bias}_\theta\left[W_n\right] = 0$$

for every $\theta \in \Theta$, then W_n is a consistent sequence of estimators of θ.

- *By Chebychev's inequaiity, $P_\theta\left\{|W_n - \theta| \geq \epsilon\right\} \leq \frac{\mathrm{E}_\theta\left[(W_n-\theta)^2\right]}{\epsilon^2}$. The result follows given the sufficient conditions.*
- *These conditions are sufficient but may not be necessary.*

Example 7.10 -

- *We have shown that $\hat{p} = \bar{X}$ is an unbiased estimator of p, thus the bias is zero.*
- $\mathrm{Var}_p[\hat{p}] = \frac{p(1-p)}{n} \to 0$ *as $n \to \infty$.*
- *By the previous theorem, $\hat{p}$ is a consistent estimator of p.* ■

Example 7.11 - *In general, if $X_1, \ldots, X_n$ is an iid random sample from a distribution where the first two moments exist (thus the distribution has a finite variance) and $\mu = \mathrm{E}[X]$, then $\bar{X}$ is a consistent estimator of μ. Why?*

■

Theorem 7.4 *Let W_n be a consistent sequence of estimators of a parameter θ. Let $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$ be sequences of constants satisfying*

$$\lim_{n\to\infty} a_n = 1 \textit{(variance factor)}$$
$$\lim_{n\to\infty} b_n = 0 (bias)$$

Then the sequence $U_n = a_n W_n + b_n$ is a consistent sequence of estimators of θ.

Example 7.12 *Is S_n, the square root of the sample variance of a random sample of size n from a normal distribution, a consistent estimator of σ, the standard deviation of the distribution?*

It can be shown that $\mathrm{E}\left[S_n\right] = \sqrt{\frac{2}{n-1}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \sigma = c_n \sigma$ *and thus*

$\mathrm{Var}\left[S_n\right] = \sigma^2 - c_n^2 \sigma^2$ *and* $\mathrm{Bias}\left[S_n\right] = c_n \sigma - \sigma$. *For both the variance and the bias to go to 0 as $n \to \infty$, we need $c_n \to 1$. We'll make use of Stirling's approximation:*

$\Gamma(z) \approx \sqrt{\frac{2\pi}{z}} \left(\frac{z}{e}\right)^z$ *where z is defined as a complex number but used here as real.*

$$c_n = \sqrt{\frac{2}{n-1}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \approx \frac{1}{\sqrt{e}} \left(\frac{n}{n-1}\right)^{\frac{n-2}{2}} = \frac{1}{\sqrt{e}} \{ \frac{n}{n-1}^{-1} * [(1 - \frac{1}{n})^n]^{-\frac{1}{2}} \} \underset{n\to\infty}{\longrightarrow} 1$$

So $S_1, S_2, \ldots$ is a consistent sequence of estimators of σ. ■

Example 7.13 *The MLE of σ is a function of S_n, namely $T_n = a_n S_n$ where $a_n = \sqrt{\frac{n-1}{n}}$. Thus $a_1, a_2, \ldots$ is a sequence of constants such that $\lim_{n\to\infty} a_n = 1$.*

So $T_1, T_2, \ldots$ is a consistent sequence of estimators of σ. That is, the MLE, T_n, is a consistent estimator of σ.

Theorem 7.5 *Consistency of MLE*

Let $X_1, X_2, \ldots,$ be an iid random sample from a family of distributions indexed by $\theta \in \Theta$. Let $\hat{\theta}$ denote the MLE of θ and let $\tau(\theta)$ be a continuous function of θ. Under certain regularity conditions, for every $\epsilon > 0$ and every $\theta \in \Theta$,

$$\lim_{n\to\infty} P_\theta\{|\tau(\hat{\theta}) - \tau(\theta)| \geq \epsilon\} = 0$$

That is, $\tau(\hat{\theta})$ is a consistent estimator of $\tau(\theta)$.

Definition 7.6 *Asymptotic variance*

- First an example: We know that the sample mean, $\bar{X}_n$, from a $N\left(\mu, \sigma^2\right)$ distribution has variance σ^2/n. Thus, $\sigma_n^2 = \mathrm{Var}\left[\sqrt{n}\bar{X}_n\right] = \sigma^2$.

- In general, if $X_1, \ldots, X_n$ is an iid random sample from a distribution where the second moment exists (thus the distribution has a finite variance), then the central limit theorem says that $\frac{\bar{X}_n - \mathrm{E}\left[\bar{X}_n\right]}{\sqrt{\mathrm{Var}[X]/n}} \to N(0,1)$ in distribution.

- Define $\sigma_{\mathrm{n}}^2 = \mathrm{Var}\left[\sqrt{n}\bar{X}_n\right]$. Then $\sigma_{\mathrm{n}}^2 \to \mathrm{Var}[X] = \tau^2$. The constant τ^2 is the limiting variance or the limit of the variances.

- Definition: For an estimator T_n, suppose that $k_n\left(T_n - \tau(\theta)\right) \to N\left(0, \sigma^2\right)$ in distribution. The parameter σ^2 is called the asymptotic variance or variance of the limit distribution of T_n.

This is general Delta methods?

Definition 7.7 *Asymptotic Efficiency*

Definition: A sequence of estimators W_n is asymptotically efficient for a parameter $\tau(\theta)$ if

$$\sqrt{n}\left[W_n - \tau(\theta)\right] \to N(0, v(\theta))$$

in distribution and

$$v(\theta) = \frac{\left[\tau'(\theta)\right]^2}{\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\ln f(X \mid \theta)\right)^2\right)} = \frac{\left[\tau'(\theta)\right]^2}{-I(X)}$$

that is, the asymptotic variance of W_n achieves the Cramér-Rao Lower Bound.

Example 7.14 *:*

If $\hat{\theta}_{jn}$ satisfies tA.GTwith asymptotic covariance matrix $V_{jn}(\theta)$, $j = 1, 2$, and $V_{1,n}(\theta) \leq V_{2n}(\theta)$ (in the sense that $V_{2n}(\theta) - V_{1n}(\theta)$ is nonnegative definite for al $\theta \in \Theta$, then $\hat{\theta}_{1n}$ is said to be asymptoticlly more efficient than $\hat{\theta}_{2n}$

Theorem 7.6 *Asymptotic Efficiency of MLE*

Let $X_1, X_2, \ldots,$ be an iid random sample from a family of distributions indexed by $\theta \in \Theta$. Let $\hat{\theta}$ denote the MLE of θ and let $\tau(\theta)$ be a continuous function of θ. Under certain regularity conditions,

$$\sqrt{n}[\tau(\hat{\theta}) - \tau(\theta)] \to N(0, v(\theta))$$

where $v(\theta)$ is the Cramér-Rao Lower Bound. That is, $\tau(\hat{\theta})$ is a consistent and asymptotically efficient estimator of $\tau(\theta)$.

Thus, the bounds of an approximate $100(1-\alpha)\%$ confidence interval for θ is given as

$$\tau(\hat{\theta}) \pm z_{\alpha/2}\sqrt{v(\hat{\theta})/n} \text{ or } \tau(\hat{\theta}) \pm z_{\alpha/2}\sqrt{\hat{v}(\hat{\theta})/n}$$

Corollary 7.6.1 • *The quantity $v(\hat{\theta})$ may be found by applying the delta method to the transformation $\tau(\cdot)$ evaluated at $\theta = \hat{\theta}$ and using $\sigma^2 = I_1^{-1}(\theta)$; that is, $v(\hat{\theta}) = \left.\frac{[\tau'(\theta)]^2}{I_1(\theta)}\right|_{\theta=\hat{\theta}}$*

- *For a given n, the approximate (unobservable) variance of $\tau(\hat{\theta})$ is $v(\theta)$. To use this in practice, we must estimate the approximate variance.*
  - *First estimator: MLE of $v(\theta)$, given by $v(\hat{\theta})$.*
  - *This uses $I_1(\hat{\theta})$, the expectation of $-\frac{\partial^2 \ln(f(X_i|\theta))}{\partial\theta^2}$ evaluated at $\theta = \hat{\theta}$*
  - *We call this the expected information number.*
  - *Alternatively, we may estimate $I_1(\theta)$ with $-\frac{1}{n}\sum \frac{\partial^2 \ln(f(X_i|\theta))}{\partial\theta^2}$ evaluated at $\theta = \hat{\theta}$.*
  - *We call this the observed information number, denoted as $\hat{I}_1(\hat{\theta})$.*
  - *The variance estimator using the observed information number is denoted as $\hat{v}(\hat{\theta})$.*
  - *It has been shown that this provides a better estimator. (Efron & Hinkley, Biometrika, 1978)*
  - *If you can differentiate and evaluate the log-likelihood, you can calculate $\hat{v}(\hat{\theta})$.*

Example 7.15 *Suppose $X_1, \ldots, X_n$ is an iid random sample from Poisson (λ). The MLE of λ is $\hat{\lambda} = \bar{X}$.*

- *- Construct a* 95%Cl *for λ, where* n = 100 *and $\bar{x} = 0.5$.*

  - Large sample via the MLE

$$\begin{aligned}
&L(\lambda \mid x) = e^{-\lambda}\lambda^x/x! \\
&l(\lambda \mid x) = -\lambda + x\log\lambda - \log(x!) \\
&\frac{\partial l}{\partial\lambda} = -1 + \frac{x}{\lambda}; \frac{\partial^2 l}{\partial\lambda^2} = -\frac{x}{\lambda^2} \\
&I_1(\lambda) = -E\left[-\frac{x}{\lambda^2}\right] = \frac{\lambda}{\lambda^2} = \frac{-1}{\lambda} \\
&I_1(\hat{\lambda}) = \frac{1}{\bar{x}} \\
&v(\lambda) = \bar{x}
\end{aligned}$$

  - Pivot the cdf

  - Recall that $F_T(T \mid \theta) \sim \mathrm{Uniform}(0,1)$ if F_T is continuous in T. That is, $F_T(T \mid \theta)$ has the same distribution for all values of θ. Thus $F_T(T \mid \theta)$ is a pivotal quantity.

  - Hence $P\{\alpha_1 \leq F_T(t \mid \theta) \leq 1 - \alpha_2\} = 1 - \alpha_1 - \alpha_2$.

- *The following also works if F_T is not continuous in T.*

- *If $F_T(T \mid \theta)$ is non-increasing in θ then for given t* $\theta_L = \inf\{\theta : F_T(t \mid \theta) \leq 1-\alpha_2\}$ *and* $\theta_U = \sup\{\theta : F_T(t \mid \theta) \geq \alpha_1\}$.

- *If $F_T(T \mid \theta)$ is non-decreasing in θ then for given t* $\theta_L = \inf\{\theta : F_T(t \mid \theta) \geq \alpha_1\}$ *and* $\theta_U = \sup\{\theta : F_T(t \mid \theta) \leq 1-\alpha_2\}$.

Note $\sum_1^n x_i \sim \text{Porsson}(n\lambda)$

$$CLT \to \frac{\frac{1}{n}}{\sum_n} \frac{\sum x_i - n\lambda}{\sqrt{n\lambda}} = \sqrt{n}\left(\frac{\bar{x}-\lambda}{\sqrt{\lambda}}\right) \underset{\text{mis}_{dist.}}{\to} N(0,1)$$

$\Phi\left(\sqrt{n}\left(\frac{\bar{x}-\lambda}{\sqrt{\lambda}}\right)\right)$ *is a pivotal quantity.* Redo this

- - *Construct a* 95%Cl *for* $\theta = P_\lambda\{X=0\}$. - *Large sample via the MLE*

52

Statistical Inference

Hypothesis Testing

- What about testing $H_0 : g(\theta) = a$ instead of $H_0 : \theta = a$
-

Power/Size of the Test

update pages

8.2 Power, Size

Definition 8.4 *The power function of a hypothesis test with rejection region R is the function of θ defined by*

$$\beta(\theta) = P_\theta\{\mathbf{X} \in R\}$$

REVIEW THIS

- This is a function of θ

- Note: In some textbooks, β refers to the probability of a Type II error. While the power function can be used to measure this probability, $\beta(\theta)$ has a different meaning. Know the context in which β is used.

- The power function will be helpful in measuring and controlling the hypothesis testing errors. But first, an example.

Example 8.6 *:*

- Suppose we have $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$.

- Consider $\arg\max_{\theta \in \Theta_0} \beta(\theta) = \theta_0$. Then we may set α as an upper bound on the Type I error probability by constructing our test rule so that $\beta(\theta_0) \leq \alpha$

- For the previous example, we see that $\beta(\lambda)$ is maximal at $\lambda = 1$. So we find c so that $\beta(1) = P_{\lambda=1}\{\bar{T} \geq c\} = \alpha$. We find c from the quantile of the distribution of T

- When $\alpha = 0.05$ and $n = 25$, we have $c = 1.35$.

- See the figure on the next slide.

Definition 8.5 *A test with power function $\beta(\theta)$ is a size α test if*

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

for $0 \leq \alpha \leq 1$

(the maximum type-I Error can actually achieve)

Definition 8.6 *A test with power function $\beta(\theta)$ is a level α test if*

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

for $0 \leq \alpha \leq 1$. We say such a test is an alpha-level test.

(α is the upper bound of the type-I error rate)

Example 8.7 *:*

- *Redefine the hypotheses $H_0 : \lambda = \theta$ and $H_1 : \lambda \neq \theta$, where H_0 is a point*
- *We can show that the LRT has the rejection region $R = \{\boldsymbol{t} : \bar{T} \leq c_1 \text{ or } \bar{T} \geq c_2\}$ where $c_1 \leq 1$ and $c_2 \geq 1$ are constants.*

- *The power function is defined as* $\beta(\lambda) = P_\lambda \left\{ \bar{T} \leq c_1 \right\} + P_\lambda \left\{ \bar{T} \geq c_2 \right\}$.
- *To make this an* α*-level test (and a size* α *test), we may choose* c_1, c_2, α_1, *and* α_2 *so that* $P_{\lambda=1} \left\{ \bar{T} \leq c_1 \right\} = \alpha_1, P_{\lambda=1} \left\{ \bar{T} \geq c_2 \right\} = \alpha_2$, *and* $\alpha_1 + \alpha_2 = \alpha$

Definition 8.7 *Suppose we want the Type II error probability to be less than* $1 - \beta_*$ *when* β_* *is a real number and* $\theta = \theta_* \in \theta_1$. *We seek the smallest sample size,* n, *such that*

what is this?

$$\beta_n\left(\theta_*\right) \geq \beta_*$$

Example 8.8 *We may use the inverse cdf of the Gamma* $\left(n, \frac{\lambda}{n}\right)$ *distribution evaluated at* $\lambda = \lambda_*$ *and perform a binary search for* n *with, say, the bisection algorithm.*

Definition 8.8 *Unbiased Test:*

A test with power function $\beta(\theta)$ *is unbiased if*

$$\beta\left(\theta'\right) \geq \beta\left(\theta''\right)$$

for every $\theta' \in \Theta_1$ *and* $\theta'' \in \Theta_0$.

Most Powerful Test

Need to review This part

8.3 Most Powerful Test

A test for a hypothesis is a statistic $T(X)$ taking values in $[0, 1]$. When $X = x$ is observed, we reject H_0 with probability $T(x)$ and accept H_0 with probability $1 - T(x)$. If $T(X) = 1$ or 0 a.s. $\mathcal{P}$, then $T(X)$ is a nonrandomized test. Otherwise $T(X)$ is a randomized test. For a given test $T(X)$, the power function of $T(X)$ is defined to be

$$\beta_T(P) = E[T(X)], \quad P \in \mathcal{P},$$

which is the type I error probability of $T(X)$ when $P \in \mathcal{P}_0$ and one minus the type II error probability of $T(X)$ when $P \in \mathcal{P}_1$.

As we discussed in §2.4.2, with a sample of a fixed size, we are not able to minimize two error probabilities simultaneously. Our approach involves maximizing the power $\beta_T(P)$ over all $P \in \mathcal{P}_1$ (i.e., minimizing the type II error probability) and over all tests T satisfying

$$\sup_{P \in \mathcal{P}_0} \beta_T(P) \leq \alpha$$

where $\alpha \in [0, 1]$ is a given level of significance. Recall that the left-hand side of (6.2) is defined to be the size of T.

Definition 8.9 *Uniformly Most Powerful*

Let $\mathcal{C}$ be a class of tests for testing $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$. A test in class $\mathcal{C}$ with power function $\beta(\theta)$, is a uniformly most powerful (UMP) class $\mathcal{C}$ test if $\beta(\theta) \geq \beta'(\theta)$ for every $\theta \in \Theta_1$ and every other $\beta'(\theta)$ that is a power function of a test in class $\mathcal{C}$.

compared the definition with unbiased test:

- Typically, we choose a value of α and define $\mathcal{C}$ to be the class of all α-level tests.
- This may include many or perhaps an unbounded number of tests with test size being less than α.
- Tests with test size close to α may be more powerful. Why?
- UMP tests may not exist in some situations.

Definition 8.10 *Simple Hypothesis:*

- A simple hypothesis completely specifies all of the parameters that index the related known or assumed family of distributions.

- That is, a simple hypothesis specifies the population distribution completely.

Definition 8.11 *Composite Hypothesis:*

- A composite hypothesis is the union of a collection of simple hypotheses and specifies a subset of the related known or assumed family of distributions.

- That is, a composite hypothesis specifies possible population distributions but does not completely specify a single distribution.

Most Powerful Test

Lemma 8.2 *Neyman-Person lemme for two samples test:*

Consider testing the simple hypotheses $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$, where $\theta_0 \neq \theta_1$ and the pdf or pmf corresponding to θ_i is $f(\boldsymbol{x} \mid \theta_i), i = 0, 1$, using a test with rejection region R that satisfies

1. $\boldsymbol{x} \in R$ if $f(\boldsymbol{x} \mid \theta_1) > kf(\boldsymbol{x} \mid \theta_0)$ and $\boldsymbol{x} \in R^c$ if $f(\boldsymbol{x} \mid \theta_1) < kf(\boldsymbol{x} \mid \theta_0)$ for some $k \geq 0$, and

2. $\alpha = P_{\theta_0}(X \in R) = \beta(\theta_0)$

Then

- Sufficient Condition: Any test that satisfies (1) and (2) is a UMP level- α test.

UMP at this α level test

- Necessary Condition: If there exists a test satisfying (1) and (2) with $k > 0$, then every UMP level- α test is a size α test (satisfies (2)) and every UMP level- α test satisfies (1) except perhaps on a set A satisfying $P_{\theta_0}(\boldsymbol{X} \in A) = P_{\theta_1}(\boldsymbol{X} \in A) = 0$.

Proof

This proof is adapted from Wikipedia proof . Define the rejection region of the null hypothesis for the Neyman-Pearson (NP) test as

$$R_{\mathrm{NP}} = \left\{\boldsymbol{x} : \frac{f(\boldsymbol{x} \mid \theta_0)}{f(\boldsymbol{x} \mid \theta_1)} \leq \eta\right\} (8.2.1)$$

$$R_{\mathrm{NP}}^c = \left\{\boldsymbol{x} : \frac{f(\boldsymbol{x} \mid \theta_1)}{f(\boldsymbol{x} \mid \theta_0)} \leq \eta\right\} (8.2.2)$$

where η is chosen so that the power function $\beta_{NP}(\theta_0) = \alpha$. Any alternative test will have a different rejection region that we denote by R_{A}. The probability of the data falling within either region $R = R_{\mathrm{A}}$ or $R = R_{\mathrm{NP}}$ given parameter θ is

$$\beta_R(\theta) = P\{R \mid \theta\} = \int_R f(x \mid \theta) dx$$

For the test with critical region R_{A} to have significance level α, it must be true that the power function $\beta_A(\theta_0) \leq \alpha$, hence

$$\alpha = \beta_{NP}(\theta_0) \geq \beta_A(\theta_0).$$

It will be useful to break these down into integrals over distinct regions:

$$\beta_{NP}(\theta) = P\{R_{\mathrm{NP}} \mid \theta\} = P\left\{R_{\mathrm{NP}\} \cap R_{\mathrm{A}} \mid \theta}\right\} + P\{R_{\mathrm{NP}} \cap R_{\mathrm{A}}^c \mid \theta\}$$
$$\beta_A(\theta) = P\{R_{\mathrm{A}} \mid \theta\} = P\{R_{\mathrm{NP}} \cap R_{\mathrm{A}} \mid \theta\} + P\{R_{\mathrm{NP}}^c \cap R_{\mathrm{A}} \mid \theta\}$$

where $R^c \equiv \{x : x \notin R\}$ is the complement of region R. Setting $\theta = \theta_0$, these two expressions and the above inequality yield that

$$P\{R_{\mathrm{NP}} \cap R_{\mathrm{A}}^c \mid \theta_0\} \geq P\{R_{\mathrm{NP}}^c \cap R_{\mathrm{A}} \mid \theta_0\} (8.2.3)$$

Most Powerful Test

We would like to prove that

$$\beta_{NP}(\theta_1) \geq \beta_A(\theta_1)$$

As similarly shown above this is equivalent to

$$\mathrm{P}\left\{R_{\mathrm{NP}} \cap R_{\mathrm{A}}^c \mid \theta_1\right\} \geqslant \mathrm{P}\left\{R_{\mathrm{NP}}^c \cap R_{\mathrm{A}} \mid \theta_1\right\}$$

In what follows we show this inequality holds:

$$\begin{aligned}
P\left\{R_{\mathrm{NP}} \cap R_{\mathrm{A}}^c \mid \theta_1\right\} = & \int_{R_{\mathrm{NP}} \cap R_{\mathrm{A}}^c} f\left\{x \mid \theta_1\right) \mathrm{d}x \\
\geq & \frac{1}{\eta} \int_{R_{\mathrm{NP}} \cap R_{\mathrm{A}}^c} f\left(x \mid \theta_0\right) \mathrm{d}x \quad \text{by definition of } R_{\mathrm{NP}} \text{ this is true for its subset (8.2.1)} \\
= & \frac{1}{\eta} P\left\{R_{\mathrm{NP}} \cap R_{\mathrm{A}}^c \mid \theta_0\right\} \\
\geq & \frac{1}{\eta} P\left\{R_{\mathrm{NP}}^c \cap R_{\mathrm{A}} \mid \theta_0\right\} \text{(8.2.3)} \\
= & \frac{1}{\eta} \int_{R_{\mathrm{NP}}^c \cap R_{\mathrm{A}}} f\left(x \mid \theta_0\right) dx \\
\geq & \int_{R_{\mathrm{NP}}^c \cap R_{\mathrm{A}}} f\left(x \mid \theta_1\right) dx \text{(8.2.2)} \\
& \text{by definition of } R_{\mathrm{NP}} \text{ this is true for its complement and complement subsets;} \\
& \text{equality if } R_{\mathrm{NP}}^c \cap R_{\mathrm{A}} \text{ is empty} \\
= & P\left\{R_{\mathrm{NP}}^c \cap R_{\mathrm{A}} \mid \theta_1\right\}
\end{aligned}$$

■

Example 8.9 *Proof from* ***Mathematics Statistics***

(i) (Existence of a UMP test). For every α, there exists a UMP test of size α, which is equal to

$$T_*(X) = \begin{cases} 1 & f_1(X) > cf_0(X) \\ \gamma & f_1(X) = cf_0(X) \\ 0 & f_1(X) < cf_0(X) \end{cases}$$

where $\gamma \in [0,1]$ and $c \geq 0$ are some constants chosen so that $E\left[T_*(X)\right] = \alpha$ when $P = P_0(c = \infty$ is allowed).

(ii) (Uniqueness). If T_n is a UMP test of size α, then

$$T_{\ldots}(X) = \left\{ \begin{array}{ll} 1 & f_1(X) > cf_0(X) \\ 0 & f_1(X) < cf_0(X) \end{array} \right. \qquad a.s.\ P.$$

Proof.

64

Most Powerful Test

The proof for the case of $\alpha = 0$ or 1 is left as an exercise. Assume now that $0 < \alpha < 1$.

(i) We first show that there exist γ and c such that $E_0\left[T_*(X)\right] = \alpha$, where E_j is the expectation writ. P_j. Let $\gamma(t) = P_0\left(f_1(X) > tf_0(X)\right)$. Then $\gamma(t)$ is nonincreasing, $\gamma(0) = 1$, and $\gamma(\infty) = 0$ (why?). Thus, there exists a $c \in (0, \infty)$ such that $\gamma(c) \leq \alpha \leq \gamma(c-)$. Set

$$\gamma = \begin{cases} \frac{a-\gamma(c)}{\gamma(c-1-\gamma(0)} & \gamma(c-) \neq \gamma(c) \\ 0 & \gamma(c-) = \gamma(c) \end{cases}$$

Note that $\gamma(c-) - \gamma(c) = P\left(f_1(X) = cf_0(X)\right)$. Then

$$E_0\left[T_*(X)\right] = P_0\left(f_1(X) > cf_0(X)\right) + \gamma P_0\left(f_1(X) = cf_0(X)\right) = \alpha$$

Next, we show that T_* in (6.3) is a UMP test. Suppose that $T(X)$ is a test satisfying $E_0[T(X)] \leq \alpha$. If $T_*(x) - T(x) > 0$, then $T_*(x) > 0$ and, therefore, $f_1(x) \geq cf_0(x)$. If $T_*(x) - T(x) < 0$, then $T_*(x) < 1$ and, therefore, $f_1(x) \leq cf_0(x)$. In any case, $\left[T_*(x) - T(x)\right]\left[f_1(x) - cf_0(x)\right] \geq 0$ and, therefore.

$$\int \left[T_*(x) - T(x)\right]\left[f_1(x) - cf_0(x)\right] d\nu \geq 0$$

i.e..

$$\int \left[T_*(x) - T(x)\right] f_1(x) dv \geq c \int \left[T_*(x) - T(x)\right] f0(x) dv$$

The left-hand side of (6.5) is $E_1[T \cdot (X)] - E_1[T(X)]$ and the right-hand side of (6.5) is $c\left\{E_0\left[T_+(X)\right] - E_0[T(X)]\right\} = c\left\{\alpha - E_0[T(X)]\right\} \geq 0$. This proves the result in (i).

(ii) Let $\mathrm{T_n}(X)$ be a UMP test of size α. Define

$$A = \left\{x : T_*(x) \neq T_*(x), \quad f_1(x) \neq ef_0(x)\right\}$$

Then $\left\{T_*(x) - T_{**}(x)\right]\left[f_1(x) - cf_0(x)\right] > 0$ when $x \in A$ and $= 0$ when $x \in A^c$, and

$$\int \left[T_*(x) - T_*(x)\right]\left[f_1(x) - cf_0(x)\right] d\nu = 0,$$

since both T_* and $T_{\ldots}$ are UMP tests of size α. By Proposition 1.6(ii), $v(A) = 0$. This proves (6.4) ■

Corollary 8.2.1 *Base on theorem 8.1*

Consider testing the simple hypotheses $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$, where $\theta_0 \neq \theta_1$. Suppose $T(\boldsymbol{X})$ is a sufficient statistic for θ and $g\left(t \mid \theta_i\right)$ is the pdf or pmf of T corresponding to $\theta_i, i =$ 0,1 . Then any test based on T with rejection region R is a UMP level- α test if it satisfies

1. $t \in R$ if $g\left(t \mid \theta_1\right) > kg\left(t \mid \theta_0\right)$ and $t \in R^c$ if $g\left(t \mid \theta_1\right) < kg\left(t \mid \theta_0\right)$ for some $k \geq 0$, and

Most Powerful Test

2. $\alpha = P_{\theta_0}(T \in S) = \beta(\theta_0)$

The proof is a result of the Factorization Theorem which gives us $f(\boldsymbol{x} \mid \theta_i) = g(T(\boldsymbol{x}) \mid \theta_i)\, h(\boldsymbol{x})$. *Applying the N-P lemma, with* $h(\boldsymbol{x})$ *dividing out on both sides of the inequalities, we have the result of the corollary.*

so g(t) does not change the monontone of f(x)? since all > 0

Example 8.10 *Let* $X \sim \text{Binomial}(2, \theta)$. *We want to test* $H_0 : \theta = 1/2$ *versus* $H_1 : \theta = 3/4$. *Calculating the ratios of the pmfs gives*

$$\frac{f(0 \mid \theta = 3/4)}{f(0 \mid \theta = 1/2)} = \frac{1}{4}, \quad \frac{f(1 \mid \theta = 3/4)}{f(1 \mid \theta = 1/2)} = \frac{3}{4}, \quad \textit{and} \quad \frac{f(2 \mid \theta = 3/4)}{f(2 \mid \theta = 1/2)} = \frac{9}{4}$$

- If we choose $3/4 < k < 9/4$, *the Neyman-Pearson Lemma says that the test that rejects* H_0 *if* $X = 2$ *is the UMP level* $\alpha = P\{X = 2 \mid \theta = 1/2\} = 1/4$ *test.*

- If we choose $1/4 < k < 3/4$, *the Neyman-Pearson Lemma says that the test that rejects* H_0 *if* $X = 1$ *or 2 is the UMP level* $\alpha = P\{X = 1 \text{ or } 2 \mid \theta = 1/2\} = 3/4$ *test.*

- Choosing $k < 1/4$ *or* $k > 9/4$ *yields the UMP level* $\alpha = 1$ *or level* $\alpha = 0$ *test.* ∎

Example 8.11 *- Note that if* $k = 3/4$, *then the NP testing rule says we must reject* H_0 *for the sample point* $x = 2$ *and accept* H_0 *for* $x = 0$ *but leaves our action for* $x = 1$ *undetermined. But if we accept* H_0 *for* $x = 1$, *we get the UMP level* $\alpha = 1/_4$ *test as before. If we reject* H_0 *for* $x = 1$, *we get the UMP level* $\alpha = 3/4$ *test as before. Such is the case with discrete distributions. No problem with continuous ones.*

What is your opinion of this?:

- Suppose we want a α*-size test where* $\alpha < 1/4$.

- We accept H_0 *if* $x < 2$.

- If $x = 2$ *then we generate a uniform value* u *and reject* H_0 *if* $u \leq 4\alpha$. *Otherwise, we accept* H_0.

- Is this a size- α *test? Do we still have an UMP test?*∎

Definition 8.12 *Monotone Likelihood Ratio (MLR)*

A family of pdfs or pmfs $\{g(t \mid \theta) : \theta \in \Theta\}$ *for a univariate random variable* T *with real-valued parameter* θ *has a monotone likelihood ratio (MLR) if, for every* $\theta_2 > \theta_1$, $\frac{g(t|\theta_2)}{g(t|\theta_1)}$ *is a monotone (nonincreasing or nondecreasing) function of* t *on* $\{t : g(t \mid \theta_1) > 0 \ \text{or}\ g(t \mid \theta_2) > 0\}$

Theorem 8.3 *Karlin-Rubin Theorem*

Consider testing the composite hypotheses $H_0 : \theta \leq \theta_0$ *vs* $H_1 : \theta > \theta_0$. *Suppose that* T *is a sufficient statistic for* θ *and the family of pdfs or pmfs* $\mathcal{F} = \{g(t \mid \theta) : \theta \in \Theta\}$ *has a nondecreasing MLR. Then for any* t_0, *the test that rejects* H_0 *if and only if* $T > t_0$ *is a UMP level-* α *test, where* $\alpha = P_{\theta_0}(T > t_0) = \beta(\theta_0)$.

- The Karlin-Rubin Theorem is essentially the Neyman-Pearson Lemma for composite hypotheses.

Most Powerful Test

- Since $\mathcal{F}$ has a nondecreasing MLR, we can show the test's power function, $\beta(\theta)$, is nondecreasing. (Exercise) This leads to the theorem's proof (C&B, p. 391-392).

- Under conditions of the theorem, the test above is also UMP for $H_0 : \theta = \theta_0$ vs $H_1 : \theta > \theta_0$ (limited the parameter space into $\theta \in (\theta_0, \infty)$

Example 8.12 *- How would you restate the theorem for $H_0 : \theta \geq \theta_0$ vs $H_1 : \theta < \theta_0$?*

Proof:

Let $\beta(\theta) = P_\theta(T > t_0)$ be the power function of the test. Fix $\theta' > \theta_0$ and consider testing $H_0' : \theta = \theta_0$ versus $H_1' : \theta = \theta'$. Since the family of pdfs or pmfs of T has an MLR, $\beta(\theta)$ is nondecreasing (see Exercise 8.34), so

i. $\sup_{\theta \leq \theta_0} \beta(\theta) = \beta(\theta_0) = \alpha$, and this is a level α test.

ii. If we define

$$k' = \inf_{t \in \mathcal{T}} \frac{g(t \mid \theta')}{g(t \mid \theta_0)},$$

where $\mathcal{T} = \{t : t > t_0 \text{ and either } g(t \mid \theta') > 0 \text{ or } g(t \mid \theta_0) > 0\}$, it follows that

$$k > 0$$

and

$$T > t_0 \Leftrightarrow \frac{g(t \mid \theta')}{g(t \mid \theta_0)} > k'.$$

Together with Corollary 8.3.13, (i) and (ii) imply that this is an UMP and follows by UMP is unbias, $\beta(\theta') \geq \beta^*(\theta')$, where $\beta^*(\theta)$ is the power function for any other level α test of H_0', that is, any test satisfying $\beta(\theta_0) \leq \alpha$. However, any level α test of H_0 satisfies $\beta^*(\theta_0) \leq \sup_{\theta \in \theta_0} \beta^*(\theta) \leq \alpha$. Thus, $\beta(\theta') \geq \beta^*(\theta')$ for any level α test of H_0. Since θ' was arbitrary, the test is a UMP level α test. ■

is it because of monotone or Unbiased?

Example 8.13 *Define*

$$\varphi(t) = \begin{cases} 1 & t > t_0 \\ \gamma & t = t_0 \\ 0 & t < t_0 \end{cases}$$

- $\varphi(T)$ represents the test where we reject H_0 with probability 1 if $T > t_0$; with probability 0 if $T < t_0$ (we fail to reject); and with probability γ if $T = t_0$.

- The values t_0 and $\gamma \in [0, 1]$ are chosen so that

$$\beta(\theta_0) = \mathrm{E}_{\theta_0}[\varphi(T)] = P_{\theta_0}\{T > t_0\} + \gamma P_{\theta_0}\{T = t_0\} = \alpha.$$

This will allow us to construct a size α test.

- This use of randomization when $T = t_0$ is controversial but is not needed if T is a continuous random variable.■

Likelihood Ratio Test

- Learn how to construct Likelihood Ratio and how to find the cut off points

Table 2: Related Problem Sets

Year	Questions	Outcome
2021-T	3.d	Bad

8 Hypothesis Testing

8.1 classic hypothesis testing

Definition 8.1 *Statistical Hypothesis*

- A statistical hypothesis is an assertion or conjecture about the distribution of one or more random variables. Usually, we will know the distribution to be a member of a family of distributions indexed by $\theta \in \Theta$.

Definition 8.2 *Two complementary hypotheses*

The two complementary hypotheses in a hypothesis testing problem are called the null hypothesis and the alternative hypothesis. They are denoted by H_0 and H_1, respectively.

- Again, suppose we posit that the mean, μ, of a population is positive. By default, we may assume the complement is true, that is, the mean is zero or negative. Typically, the default hypothesis is the null hypotheses. Here we have $H_0 : \mu \leq 0$ and $H_1 : \mu > 0$

- In general, we specify the null hypothesis as $H_0 : \theta \in \Theta_0$ and the alternative hypothesis as $H_1 : \theta \in \Theta_1$ where $\Theta_0, \Theta_1 \subset \Theta, \Theta_0 \cup \Theta_1 = \Theta$, and $\Theta_0 \cap \Theta_1 = \emptyset$; that is, Θ_0 and Θ_1 are complementary subsets of Θ.

The subset of the sample space, $\mathcal{X}$, for which H_0 will be rejected is called the rejection region, R, or critical region. The complement of the rejection region is called the acceptance region, A.

In general, we denote the rejection region as $R \subset \mathcal{X}$ and we

- Reject H_0 and accept H_1 as true if $\boldsymbol{x} \in R$

- Accept H_0 as true if $\boldsymbol{x} \notin R$

- We do not support or accept the null hypothesis! We merely support our research hypothesis or fail to support it.

- The null hypothesis is a construct used solely for hypothesis testing. Failing to support the research hypothesis must not be confused with supporting the null hypothesis.

8.1.1 Likelihood Ratio Test

Definition 8.3 *The likelihood ratio test statistic for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, where $\Theta_1 = \Theta_0^c$, is*

$$\lambda(\mathbf{x}) = \frac{\sup_{\Theta_0} L(\theta \mid \mathbf{x})}{\sup_{\Theta} L(\theta \mid \mathbf{x})}$$

A likelihood ratio test (LRT) is any test that has a rejection region of the form $R = \{\mathbf{x} : \lambda(\mathbf{x}) \leq c\}$, where c is some known number satisfying $0 \leq c \leq 1$

- The denominator of the likelihood ratio statistic is the supremum of the likelihood function over all values of the parameter space. The numerator is the supremum of the likelihood function restricted to the subset of the parameter space corresponding to the null hypothesis. - Often, we are interested in the natural log of the LRT statistic.

Likelihood Ratio Test

Example 8.1 *Suppose* $\Theta = \{\theta_0, \theta_1\}$ *where* $\theta_0 \neq \theta_1$. *Further suppose that* $H_0 : \theta = \theta_0$. *Then the LRT statistic is*

$$\lambda(\boldsymbol{x}) = \frac{L(\theta_0 \mid \boldsymbol{x})}{\max\left(L(\theta_0 \mid \boldsymbol{x}), L(\theta_1 \mid \boldsymbol{x})\right)}$$

- Note that in most cases $\sup_\theta L(\theta \mid \boldsymbol{x}) = L(\hat{\theta} \mid \boldsymbol{x})$, where $\hat{\theta}$ is the overall MLE, and $\sup_{\theta_0} L(\theta \mid \boldsymbol{x}) = L\left(\hat{\theta}_0 \mid \boldsymbol{x}\right)$, where $\hat{\theta}_0$ is the MLE over Θ_0. In this case, the LRT statistic is given by

$$\lambda(\boldsymbol{x}) = \frac{L\left(\hat{\theta}_0 \mid \boldsymbol{x}\right)}{L(\hat{\theta} \mid \boldsymbol{x})}$$

Example 8.2 *Suppose we have an iid random sample from* $f(x \mid \theta) = \frac{1}{\theta} I_{[0,\theta]}(x)$, *where* $\Theta = (0, \infty)$. *We wish to test* $H_0 : \theta \geq 1$ *versus* $H_1 : \theta < 1$. *The likelihood function is* $L(\theta \mid x) = (\theta)^{-n} I_{[0,\theta]}\left(x_{(n)}\right)$.

- The overall observed MLE is $\hat{\theta} = X_{(n)}$ *and* $L(\hat{\theta} \mid x) = \left(x_{(n)}\right)^{-n}$.

- $\Theta_0 = [1, \infty)$. *If* $X_{(n)} \geq 1$ *then* $\hat{\theta}_0 = \hat{\theta} = X_{(n)}$. *Otherwise, if* $X_{(n)} < 1$ *then* $\hat{\theta}_0 = 1$ *(why?) and* $L(1 \mid x) = 1$.

- The LRT statistic is thus

$$L\left(\hat{\theta}_0 \mid \boldsymbol{x}\right) = \begin{cases} L(\hat{\theta} \mid \boldsymbol{x}) & \text{if } x_{(n)} \geq 1 \\ 1 & \text{if } x_{(n)} < 1 \end{cases} \text{ and } \lambda(\boldsymbol{x}) = \begin{cases} 1 & \text{if } x_{(n)} \geq 1 \\ \left(x_{(n)}\right)^n & \text{if } x_{(n)} < 1 \end{cases}$$

- The rejection region is given as $\{\boldsymbol{x} : \lambda(\boldsymbol{x}) \leq c\}$. *If we choose* $c = 1$, *then we will always reject* H_0. *If we choose* $0 \leq c < 1$, *then* $\lambda(\boldsymbol{x}) \leq c$ *only when* $x_{(n)} < 1$

- Hence the LRT is given as Reject $H_0 : \theta \geq 1$ *if* $\left(x_{(n)}\right)^n \leq c$ *which is equivalent to* $x_{(n)} \leq c^{1/n}$.

Theorem 8.1 *If* $T(\boldsymbol{X})$ *is a sufficient statistic for* θ *and* $\lambda^*(t)$ *and* $\lambda(\boldsymbol{x})$ *are the LRT statistics based on* T *and* $\boldsymbol{X}$, *respectively, then* $\lambda^*(T(\boldsymbol{x})) = \lambda(\boldsymbol{x})$ *for every* $\boldsymbol{x}$ *in the sample space,* $\mathcal{X}$.

Example 8.3 *Survival times for advanced-stage colon cancer patients are Assumed to be distributed as*

$$f(x \mid \alpha) = \alpha e^{-\alpha x} I_{(0,\infty)}(x); \alpha > 0.$$

Suppose we have two treatments which are proposed to affect survival times.

1. *Let* $X_1, \ldots, X_n$ *iid* $\sim f(x \mid \alpha)$ *denote a random sample of patients given treatment 1 .*
2. *Let* $Y_1, \ldots, Y_n$ *iid* $\sim f(y \mid \beta)$ *denote a random sample of patients given treatment 2 .*

The two samples are combined into one sample. Assume the two samples are independent. Note that $(\bar{X}, \bar{Y})$ are jointly sufficient.

Likelihood Function

$$\begin{aligned}
L(\alpha, \beta \mid \bar{x}, \bar{y}) &= L(\alpha \mid \bar{x}) \cdot L(\beta \mid \bar{y}), \text{ due to independence.} \\
&= \alpha^n e^{-n\alpha\bar{x}} \beta^n e^{-n\beta\bar{y}} \\
l(\alpha, \beta \mid \bar{x}, \bar{y}) &= n \ln \alpha - n\alpha\bar{x} + n \ln \beta - n\beta\bar{y}
\end{aligned}$$

MLE(General case without any constrain from the hypothesis)

$$\frac{\partial l}{\partial \alpha} = \frac{n}{\alpha} - n\bar{x} = 0 \Rightarrow \hat{\alpha} = \frac{1}{\bar{X}}$$

$$\left(\frac{\partial^2 l}{\partial \alpha^2} = -\frac{n}{\alpha^2} < 0, \text{ so } \hat{\alpha} \text{ provides a maximum} \right)$$

$$\text{Similarly, } \hat{\beta} = \frac{1}{\bar{Y}}$$

LRT for $H_0 : \alpha = \beta$ vs $H_1 : \alpha \neq \beta$

Restricted MLE(based on H_0)

$$\begin{aligned}
&\text{set } \alpha = \beta \\
l(\alpha, \alpha \mid \bar{x}, \bar{y}) &= 2n \ln \alpha - n\alpha(\bar{x} + \bar{y}) \\
\frac{\partial l}{\partial \alpha} &= \frac{2n}{\alpha} - n(\bar{x} + \bar{y}) \\
\hat{\alpha}_0 &= \frac{2}{\bar{x} + \bar{y}} = \hat{\beta}_0
\end{aligned}$$

LR test statistic

$$\lambda = \frac{L\left(\hat{\alpha}_0, \hat{\beta}_0 \mid \bar{x}, \bar{y}\right)}{L(\hat{\alpha}, \hat{\beta} \mid \bar{x}, \bar{y})} = \left[\frac{4\bar{x}\bar{y}}{(\bar{x} + \bar{y})^2} \right]^n$$

Let $U = \bar{X}/(\bar{X} + \bar{Y})$. Then (U here is the $T(.)$)

$$\lambda = [4u(1 - u)]^n.$$

LR test $R = \{\mathbf{x}, \mathbf{y} : \lambda \leq c\}$ for $c \in (0, 1)$. Reject H_0 (support H_1) if $(\mathbf{x}, \mathbf{y}) \in R$. Note that

$$\begin{aligned}
\lambda \leq c &\Rightarrow u(1 - u) \leq c^*, \text{ where } c^* = c^{1/n}/4 \\
&\Rightarrow u \leqslant \frac{1 - \sqrt{1 - 4c^*}}{2} \text{ or } u \geqslant \frac{1 + \sqrt{1 - 4c^*}}{2} \\
&\Rightarrow \left| u - \frac{1}{2} \right| \geqslant k, \text{ where } k = \sqrt{1 - 4c^x}
\end{aligned}$$

Likelihood Ratio Test

So equivalently, we may say reject H_0 (support H_1) if $\left|U-\frac{1}{2}\right| \geqslant k, k \in(0,1)$
■

Example 8.4 *(CONTINUE)*

LRT for $H_0: \alpha \geqslant \beta$ vs $H_1: \alpha<\beta$ The full parameter space is $\Theta=\{(\alpha, \beta): \alpha>0, \beta>0\}$. The restricted (null) parameter space is $\Theta_0=\{(\alpha, \beta): 0<\beta \leq \alpha\}$. We consider two cases.

1. $\bar{x} \leq \bar{y} \Rightarrow \frac{1}{\bar{x}} \geq \frac{1}{\bar{y}}$: In this case, $\hat{\alpha}=\frac{1}{X} \geq \hat{\beta}=\frac{1}{Y}$. So $\hat{\theta}=\hat{\theta}_0 \in \Theta_0$.

2. $\bar{x}>\bar{y} \Rightarrow \frac{1}{\bar{x}}<\frac{1}{\bar{y}}$: In this case, $\hat{\alpha}=\frac{1}{X}<\hat{\beta}=\frac{1}{Y}$. So $\hat{\theta} \in \Theta_0^c$. To maximize the likelihood function for $\theta \in \Theta_0$ the restricted MLE $\hat{\theta}_0$ should be close to $\hat{\theta}$ which puts it on the border of Θ_0 where $\alpha=\beta$. We've already found that $\hat{\alpha}_0=\hat{\beta}_0=\frac{2}{\bar{x}+\bar{y}}$ provides the maximum likelihood when $\alpha=\beta$.

LR Test Statistic

$$\lambda=\left\{\begin{array}{lll}1 & \text{for} & \bar{x} \leq \bar{y} \\ {[4 u(1-u)]^n} & \text{for} & \bar{x}>\bar{y}\end{array}\right.$$

LR Test

$$R=\{\mathbf{x}, \mathbf{y}: \lambda \leq c\} \text{ for } c \in(0,1).$$

Reject H_0 (support H_1) if $(\mathbf{x}, \mathbf{y}) \in R$.

Note that $\bar{x}>\bar{y}$ iff $u>1/2$. Thus,

$$\lambda \leqslant c \text{ when } u>\frac{1}{2}+k$$

So equivalently, we may say reject H_0 (support H_1) if $U>\frac{1}{2}+k$ ■

Example 8.5 *Suppose $X_1, \ldots, X_n$ are a random sample from a $\mathrm{N}\left(\mu, \sigma^2\right)$ distribution, and an experimenter is interested only in inferences about μ, such as testing $H_0: \mu \leq \mu_0$ versus $H_1: \mu>\mu_0$. Then the parameter σ^2 is a nuisance parameter. The LRT statistic is*

$$\lambda(\mathbf{x})=\frac{\max _{\left\{\mu, \sigma^2: \mu \leq \mu_0, \sigma^2 \geq 0\right\}} L\left(\mu, \sigma^2 \mid \mathbf{x}\right)}{\max _{\left\{\mu, \sigma^2:-\infty<\mu<\infty, \sigma^2 \geq 0\right\}} L\left(\mu, \sigma^2 \mid \mathbf{x}\right)}$$

Show that the LRT can be based on the Student's t statistic.

$$\begin{aligned} \lambda(\mathbf{x}) = & \frac{\max _{\left\{\mu, \sigma^2: \mu \leq \mu_0, \sigma^2 \geq 0\right\}} L\left(\mu, \sigma^2 \mid \mathbf{x}\right)}{\max _{\left\{\mu, \sigma^2:-\infty<\mu<\infty, \sigma^2 \geq 0\right\}} L\left(\mu, \sigma^2 \mid \mathbf{x}\right)} \\ & =\frac{L\left(\hat{\mu}_0, \hat{\sigma}_0^2 \mid \mathbf{x}\right)}{L\left(\hat{\mu}, \hat{\sigma}^2 \mid \mathbf{x}\right)}\end{aligned}$$

where

$$L\left(\mu, \sigma^2 \mid \mathbf{x}\right)=\left(\frac{1}{\sqrt{2 \pi \sigma^2}}\right)^n \exp \left\{-\frac{1}{2 \sigma^2} \sum_1^n\left(x_i-\mu\right)^2\right\}$$

Likelihood Ratio Test

The (unrestricted) observed MLEs of μ *and* σ^2 *are* $\hat{\mu} = \bar{x}, \hat{\sigma}^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$. *Plugging these in we get*

$$L\left(\hat{\mu}, \hat{\sigma}^2 \mid \mathbf{x}\right) = \left(\frac{1}{\sqrt{2\pi\hat{\sigma}^2}}\right)^n \exp\left\{-\frac{n}{2}\right\}$$

Now consider the restricted MLE where $\mu \leq \mu_0$. *If* $\bar{x} \leq \mu_0$ *then this restricted (observed) MLE is the same as the unrestricted MLE. Suppose* $\bar{x} > \mu_0$. *Recall that* $\sum(x_i - \mu)^2 = \sum(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$. *To maximize* L *we need to minimize* $(\bar{x} - \mu)^2$. *Thus we choose* μ *to be as close to* $\bar{x}$ *as we can get; that is,* $\hat{\mu}_0 = \mu_0$. *Thus*

$$L\left(\mu_0, \sigma^2 \mid \mathbf{x}\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum(x_i - \mu_0)^2\right\}$$

Maximizing this with respect to σ^2 *yields* $\hat{\sigma}_0^2 = \frac{1}{n}\sum(x_i - \mu_0)^2$ *and*

$$L\left(\mu_0, \hat{\sigma}_0^2 \mid \mathbf{x}\right) = \left(\frac{1}{\sqrt{2\pi\hat{\sigma}_0^2}}\right)^n \exp\left\{-\frac{n}{2}\right\}$$

Pulling it together, the LRT statistic is

$$\lambda(\mathbf{x}) = \begin{cases} 1 & \text{if } \bar{x} \leq \mu_0 \\ \frac{L\left(\mu_0, \hat{\sigma}_0^2 \mid \mathbf{x}\right)}{L\left(\hat{\mu}, \hat{\sigma}^2 \mid \mathbf{x}\right)} = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{\frac{n}{2}} & \text{if } \bar{x} > \mu_0 \end{cases}$$

The rejection region is

$$R = \left\{\mathbf{x} : \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{\frac{n}{2}} \leq c\right\}$$

To simplify, note that

$$\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} = \frac{\sum(x_i - \bar{x})^2}{\sum(x_i - \mu_0)^2} = \frac{\sum(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2} = \frac{1}{1 + \frac{n(\bar{x} - \mu_0)^2}{\sum(x_i - \bar{x})^2}}$$

An equivalent representation of the rejection region is thus

$$R = \left\{\mathrm{x} : \frac{n(\bar{x} - \mu_0)^2}{\sum(x_i - \bar{x})^2} \geq k^*\right\}$$

Recognizing that $\frac{n(\bar{x} - \mu_0)^2}{\sum(x_i - \bar{x})^2} = \frac{1}{n-1}\left(\frac{\bar{x} - \mu_0}{s/\sqrt{n}}\right)^2$ *we have the hypothesis test rule*

Reject H_0 *in favor of* H_1 *if* $\frac{\bar{x} - \mu_0}{s/\sqrt{n}} > k$ *for some* $k > \mu_0$.

We know that $\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim \text{t-}dist\ (df = n - 1)$. ■

59

Large Sample Test Method

8.4 Large Sample ML-based method

Set up:

For large sample size we may use approximate asymptotic methods for hypothesis testing. These are based on the maximum likelihood and thus depend on the assumed family of distributions

-We will assume that the family of distributions $\mathcal{T}$ is indexed by a p-dimensiona parameter vector $= \theta$

-Hypotheses: $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ where $\Theta_0 \in \Theta$ and $\Theta_1 \in \Theta_0^c$ is the complement of Θ_0 With respect to Θ

Given the Fisher's Information Matrix:

$$-\mathrm{E}_\theta \left[\frac{\partial^2 l(\theta|x)}{\partial\theta_i \partial\theta_j} \right]$$

We have

·Let $R_i(\theta) = 0, i = 1, 2, \ldots, r$, represent $r(< p)$ independent restrictions placed on the parameter vector $\emptyset$

·Consider $\Theta_0 = \{\theta; R_i(\theta) = 0, i = 1, 2, \ldots, r\}$.

.The composite hypotheses are $H_0 \colon \theta \in \Theta_0$ versus $H_1 \colon \theta \in \Theta_1$

·For example, with $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$, consider the $r = 3$ linearly independent linear restrictionis $R_1(\theta) = (\theta_1 - \theta_2) = 0$, $R_2(\theta) = (\theta_1 - \theta_3) = 0$ and $R_3(\theta) = (\theta_1 - \theta_4) = 0$ Then, the null hypothesis $H_0 : R(\theta) = 0 for i = 1, 2, 3$,is equivalent to the null hypothesis

$$H_0 \colon \theta_1 = \theta_2 = \theta_3 = \theta_4$$

OR We could just start with a specified $\mathfrak{G}_{\mathfrak{G}}$ and denote its dimension as d_{d}. The dimension of Θ is p.Then $r = p - d_{\mathrm{d}}$, the dimension of the space of restrictions

8.4.1 Likelihood Ratio Test

The likelihood ratio test statsitics is defined as:

$$\hat{\lambda} = \frac{\max_{\theta \in \Theta_0} L(\theta|x)}{\max_{\theta \in \Theta} L(\theta|x)} = \frac{L(\hat{\theta}_0|x)}{L(\hat{\theta}|x)}$$

Under certain regularity conditions, for large n and assuming $H_0 \colon \theta \in \Theta_0$,

$$-2\log\hat{\lambda} = 2\left(l(\hat{\theta}|x) - l(\hat{\theta}_0|x)\right) \to^d \chi_r^2\,;$$

Proof

First expand $\log L(\theta \mid \mathbf{x}) = l(\theta \mid \mathbf{x})$ in a Taylor series around $\hat{\theta}$, giving

$$l(\theta \mid \mathbf{x}) = l(\hat{\theta} \mid \mathbf{x}) + l'(\hat{\theta} \mid \mathbf{x})(\theta - \hat{\theta}) + l''(\hat{\theta} \mid \mathbf{x})\frac{(\theta - \hat{\theta})^2}{2!} + \cdots.$$

Now substitute the expansion for $l\left(\theta_0 \mid \mathbf{x}\right)$ in $-2 \log \lambda(\mathbf{x})=-2 l\left(\theta_0 \mid \mathbf{x}\right)+2 l(\hat{\theta} \mid \mathbf{x})$, and get

$$
\begin{aligned}
-2 \log \lambda(\mathbf{x}) &= -2 l(\hat{\theta}|x)+-2 \underbrace{l'(\hat{\theta}|x)}_{score=0}(\theta_0-\hat{\theta})+-2 l''(\hat{\theta}|x) \frac{(\theta_0-\hat{\theta})^2}{2!}+2 l(\hat{\theta}|x) \\
&= -2 l''(\hat{\theta}|x)(\theta_0-\hat{\theta})^2 \\
&\approx \frac{(\theta-\hat{\theta})^2}{I_n(\mathbf{x})},
\end{aligned}
$$

where we use the fact that $l'(\hat{\theta} \mid \mathbf{x})=0$. Since the denominator is the observed information $\hat{I}_n(\hat{\theta})$ and $\frac{1}{n} \hat{I}_n(\hat{\theta}) \rightarrow I\left(\theta_0\right)$ it follows from Theorem 10.1 .12 and Slutsky's Theorem (Theorem 5.5.17) that $-2 \log \lambda(\mathbf{X}) \rightarrow \chi_1^2$.

8.4.2 Wald Test

- Define the 1× r row vector $R(\theta)$ as

$$
R(\theta)=\left(R_1(\theta), R_2(\theta), \ldots, R_r(\theta)\right)
$$

- Let the r x p matrix $T'(\Theta)$ have its (i, j)th element defined as

$$
\frac{\partial R_i(\theta)}{\partial \theta_j}
$$

for i = 1,2, .r and j = 1,2. ...p
- Let the $r \times r$ matrix $A(\Theta)$ have the structure

$$
\Lambda(\theta)=T(\theta) J^{-1}(\theta) T'(\theta)
$$

.The Wald test statistic $\hat{W}>0$is defined as

$$
\tilde{W}=R(\bar{\theta}) \Lambda^{-1}(\bar{\theta}) R'(\bar{\theta})
$$

. Under certain regularity conditions, for large n and assuming $H_0: \theta \in \Theta_0$

$$
W \rightarrow^d X_r^2
$$

8.4.3 Score Test

- Define the $1 \times p$ row vector $\boldsymbol{S}(\boldsymbol{\theta})$ as

$$
\boldsymbol{S}(\boldsymbol{\theta})=\left(\frac{\partial l(\boldsymbol{\theta} \mid \boldsymbol{x})}{\partial \theta_1}, \frac{\partial l(\boldsymbol{\theta} \mid \boldsymbol{x})}{\partial \theta_2}, \ldots, \frac{\partial l(\boldsymbol{\theta} \mid \boldsymbol{x})}{\partial \theta_p}\right)
$$

- The score statistic $\hat{S}$ is defined as

$$\hat{S} = \boldsymbol{S}\left(\widehat{\boldsymbol{\theta}}_0\right)\boldsymbol{J}^{-1}\left(\widehat{\boldsymbol{\theta}}_0\right)\boldsymbol{S}'\left(\widehat{\boldsymbol{\theta}}_0\right)$$

- Under certain regularity conditions, for large n and assuming $H_0 : \boldsymbol{\theta} \in \Theta_0$,

$$\hat{S} \to \chi^2_r \text{ in distribution}$$

8.4.4 Large Sample Rejection Region

The large sample rejection region for an α-level LR test is

$$\left\{\boldsymbol{x} : -2\log\hat{\lambda} \geq \chi^2_{r,\alpha}\right\}$$

where $\chi^2_{r,\alpha}$ is the upper α-quantile of the chi-squared distribution with r degrees of freedom.

- Replace $-2\log\hat{\lambda}$ with $\widehat{W}$ for the Wald test rejection region.
- Replace $-2\log\hat{\lambda}$ with $\hat{S}$ for the Score test rejection region

Example 8.16 *- Let $X_1, X_2, \ldots, X_n$ constitute a random sample of size n from a $N\left(\mu, \sigma^2\right)$ population.*

- Consider testing the null hypothesis

$$R_l\left(\mu, \sigma^2\right) = \mu - \mu_0$$

$$H_0 : \mu = \mu_0, 0 < \sigma^2 < +\infty \textit{ versus } H_1 : \mu \neq \mu_0, 0 < \sigma^2 < +\infty$$

Note that this test is typically called a test of $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.

- The MLEs of μ and σ^2 are $\hat{\boldsymbol{\theta}} = \left(\hat{\mu}, \hat{\sigma}^2\right) = \left(\bar{X}, \left(\frac{n-1}{n}\right)S^2\right)$.

- We have that

$$T_{n-1} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \textit{ under } H_0 : \mu = \mu_0$$

Likelihood Ratio Test

It can be shown that

$$-2\log\hat{\lambda} = n\log\left(1 + \frac{T^2_{n-1}}{n-1}\right)$$

which for large n has an approximate χ^2_1 distribution.

$$-2\log\hat{\lambda} \geqslant k \Rightarrow \left|T_{n-1}\right| \geqslant k^*$$

Wald Test

$r = 1$

$$R_1 = \mu - \mu_0$$

It can be shown that

$$\hat{W} = \left(\frac{n}{n-1}\right) T_{n-1}^2$$

which for large n has an approximate χ_1^2 distribution.
Score Test
It can be shown that

$$\hat{S} = \left[\frac{\bar{X} - \mu_0}{\hat{\sigma}_0/\sqrt{n}}\right]^2$$

Test stat.
where

$$\hat{\sigma}_0^2 = n^{-1} \sum_{i=1}^{n} (X_i - \mu_0)^2$$

$$\left|\frac{\bar{x} - \mu_0}{\hat{\sigma}_0/\sqrt{n}}\right|$$

is the estimator of σ^2 under the null hypothesis $H_0 : \mu = \mu_0$.
For large $n, \hat{S}$ has an approximate χ_1^2 distribution.

- Both the LRT and the Wald test (for this example) have rejection regions that are asymptotically equivalent to the rejection region

$$\left\{|T_{n-1}| \geq z_{\alpha/2}\right\}$$

- The exact LRT has rejection region $\left\{|T_{n-1}| \geq t_{r,\alpha/2}\right\}$ where $t_{r,\alpha/2}$ is the upper $(\alpha/2)$-quantile of a t-distribution with r degrees of freedom.
- The score test is a modification of T_{n-1} that uses the MLE of σ^2 restricted to $H_0 : \mu = \mu_0, 0 < \sigma^2 < +\infty$. Its rejection region is equivalent to

$$\left\{\left|\frac{\bar{X} - \mu_0}{\hat{\sigma}_0/\sqrt{n}}\right| \geq z_{\alpha/2}\right\}$$

Confidence Interval

Confidence Interval: Definition, Coverage Probability, Coverage Coefficient and Length

You suck at this part

- Check Bob's 2023 final, you are really bad at it
- **Note that in a confidence set, which one is Random Variable and which one is not**
- How to find the coverage probablity

$X_{(n)}$ *has the Lebesgue p.d.f.* $n\theta^{-n}x^{n-1}I_{(0,\theta)}(x)$*, an unbiased estimator* $h\left(X_{(n)}\right)$ *of* ϑ *must satisfy*

$$\theta^n g(\theta) = n\int_0^\theta h(x)x^{n-1}dx \quad \text{for all } \theta > 0.$$

Differentiating both sizes of the previous equation and applying the result of differentiation of an integral (Royden $(1968, \S 5.3)$ *) lead to*

$$n\theta^{n-1}g(\theta) + \theta^n g'(\theta) = nh(\theta)\theta^{n-1}.$$

Hence, the UMVUE of ϑ *is* $h\left(X_{(n)}\right) = g\left(X_{(n)}\right) + n^{-1}X_{(n)}g'\left(X_{(n)}\right)$. *In particular, if* $\vartheta = \theta$*, then the UMVUE of* θ *is* $\left(1+n^{-1}\right)X_{(n)}$

Example 6.10 • *For some* $\tau(\boldsymbol{\theta})$ *there is no unbiased estimator.*

- *Let* $X \sim \text{Binomial}(m,\theta)$ *and* $\tau(\theta) = \ln\left(\frac{\theta}{1-\theta}\right)$*, the log-odds.*
- *Suppose* T *is any estimator of* $\tau(\theta)$.
- $\mathrm{E}_\theta[T] = \sum_{k=0}^m T(k)\begin{pmatrix} m \\ k \end{pmatrix}\theta^k(1-\theta)^{m-k}$ *is an* m^{th}*-degree polynomial in* θ.
- $\tau(\theta) = \ln\left(\frac{\theta}{1-\theta}\right)$ *cannot be expressed as a finite-degree polynomial.*
- *Thus,* T *cannot be unbiased.*
- *Since* T *was an arbitrary estimator, no unbiased estimator of* $\tau(\theta)$ *exists.*
- *For some situations there is an unbiased estimator, but no UMVUE.* ■

6.4 Loss

Definition 6.6 *Loss:*

Definition 6.7 *Risk:*

Example 6.11 *Quadratic Loss of σ^2*

Loss:

$$L(\sigma^2, bS^2) = (\sigma^2 - bS^2)^2 = b^2(S^2 - \sigma/b)^2$$

Risk:

$$R(\sigma^2, bS^2) = E[L(\sigma^2, bS^2)] = b^2(Var(S^2) + bias(S^2, \sigma^2/b)^2) = [\frac{4b}{n-1} + 2(b-1)]\sigma^4$$

Taking the derivative for Risk with regarding to b

$$\frac{dR}{db} \to b = \frac{n-1}{n+1}$$

Steins Loss: $L(\sigma^2, bS^2) = \frac{bS^2}{\sigma^2} - 1 - log(b) - log(\frac{S^2}{\sigma^2})$

Risk :

$$E[L] = bE[\frac{S^2}{\sigma^2}] - 1 - log(b) - E[log(\frac{S^2}{\sigma^2})] = b - 1 - log(b) - E[log(S^2/\sigma^2)]$$

Taking the derivative for Risk with regarding to b

$$\frac{dR}{db} \to b = 1$$

■

Definition 6.8 *One-number summaries of risk*

- Maximum risk

$$\bar{R}(\hat{\theta}) = \sup_{\theta \in \Theta} R(\theta, \hat{\theta})$$

- Bayes risk, under the prior distribution $\pi(\cdot)$

$$B_\pi(\hat{\theta}) = \int R(\theta, \hat{\theta})\pi(\theta)d\theta$$

The two summaries of the risk function suggest two different methods for devising estimators:

- Choosing $\hat{\theta}$ to minimize the maximum risk leads to minimax estimators.

-

$$\sup_{\theta} R(\theta, \widehat{\theta}) = \inf_{\tilde{\theta}} \sup_{\theta} R(\theta, \tilde{\theta})$$

40

, the infimum is over all estimators $\tilde{\theta}$.

- Choosing $\hat{\theta}$ to minimize the Bayes risk leads to Bayes estimators.

-

$$B_\pi(\widehat{\theta}) = \inf_{\widetilde{\theta}} B_\pi(\tilde{\theta})$$

Definition 6.9 *Bayes Risk Define the posterior risk of estimator $\hat{\theta}$ as*

$$R(\theta, \hat{\theta} \mid \boldsymbol{x}) = \int \mathcal{L}(\theta, \hat{\theta}(\boldsymbol{x}))\pi(\theta \mid \boldsymbol{x}) d\theta$$

Theorem: The Bayes risk $B_\pi(\widehat{\theta})$ satisfies

$$B_\pi(\widehat{\theta}) = \int R(\theta, \widehat{\theta} \mid \boldsymbol{x}) m(\boldsymbol{x}) d\boldsymbol{x}$$

Let $\hat{\theta}(\boldsymbol{x})$ be the value of θ that minimizes $R(\theta, \hat{\theta} \mid \boldsymbol{x})$. Then $\hat{\theta}$ is the Bayes estimator.

Unbiased CI

7 Interval Estimator

7.1 Confidence Intervals

7.1.1 Definition, Coverage Probablity, coefficient and length

Definition 7.1 *Confidence Interval:*

An interval estimate of a real-valued parameter θ is any pair of functions, $(L(\boldsymbol{x}), U(\boldsymbol{x}))$, of a sample point $\boldsymbol{x}$ that satisfies

$$L(\boldsymbol{x}) \leq U(\boldsymbol{x})$$

for all $\boldsymbol{x} \in \boldsymbol{X}$. The random interval $\underbrace{(L(\boldsymbol{X}), U(\boldsymbol{X}))}_{Random\ Variables}$ is called an interval estimator. Given the sample point $\boldsymbol{x}$, we infer that

$$L(\boldsymbol{x}) \leq \theta \leq U(\boldsymbol{x})$$

 # Not necessary center, but we will get to why center interval is good later

Although we are mainly concerned with confidence intervals, we occasionally will work with more general sets. When working in general, and not being quite sure of the exact form of our sets, we will speak of confidence sets. A confidence set with confidence coefficient equal to some value, say $1-\alpha$, is simply called a $1-\alpha$ confidence set.

Example 7.1 *Interpretation: In $100(2\Phi(c)-1)\%$ of all random samples if size n from $N\left(\mu, \sigma^2\right)$, the interval $\left(\bar{X} - c\frac{\sigma}{\sqrt{n}}, \bar{X} + c\frac{\sigma}{\sqrt{n}}\right)$ covers the true population mean μ.*

Definition 7.2 *Coverage Probability*

For an interval estimator $(L(\boldsymbol{X}), U(\boldsymbol{X}))$ of a parameter θ, the coverage probability of $(L(\boldsymbol{X}), U(\boldsymbol{X}))$ is the probability that the random interval $(L(\boldsymbol{X}), U(\boldsymbol{X}))$ covers the true parameter θ. The coverage probability is denoted by

$$P_\theta\{\theta \in (L(\boldsymbol{X}), U(\boldsymbol{X}))\}$$

or

$$P\{\theta \in (L(\boldsymbol{X}), U(\boldsymbol{X})) \mid \theta\}$$

or

$$P_\theta\{L(X) < \theta, U(X) > \theta\}$$

There are a number of things to be aware of in these definitions. One, it is important to keep in mind that the interval is the random quantity, not the parameter. Therefore, when we write probability statements such as $P_\theta(\theta \in [L(X), U(X)])$, these probability statements refer to X, not θ. In other words, think of $P_\theta(\theta \in [L(X), U(X)])$, which might look like a statement about a random θ, as the algebraically equivalent$P_\theta[L(X) < \theta, U(X) > \theta])$, a statement about a random X.

see Bob's Final 4(b)

Find CI by Pivoting

- More example in homework4, that shit hard
- Need more example
- Example 7.7 is good to learn how to find pivot function

7.1.3 Pivoting

Definition 7.4 *Pivotal Quantity*

A random variable $Q(\boldsymbol{X}, \theta)$ is a pivotal quantity if its distribution is independent of θ. That is $Q(\boldsymbol{X}, \theta)$ has the same distribution for all values of θ.

Lemma 7.2 *Suppose we have a statistic T with pdf $f_T(t \mid \theta)$ that may be written as*

$$f_T(t \mid \theta) = g(Q(t, \theta)) \left| \frac{d}{dt} Q(t, \theta) \right|$$

where $g(\cdot)$ does not depend on θ and $Q(t, \theta)$ is monotone in t. Let $y = Q(t, \theta)$, then $t = Q^{-1}(y, \theta)$. As long as $\frac{d}{dt}Q(t,\theta) \neq 0$, then $\frac{d}{dy}Q^{-1}(y,\theta) = 1/\frac{d}{dt}Q(t,\theta)$. So, making a change of variables,

This essentially transformation of random variable

$$\begin{aligned} f_Y(y \mid \theta) &= f_T\left(Q^{-1}(y,\theta) \mid \theta\right) \left| \frac{d}{dy} Q^{-1}(y, \theta) \right| \\ &= g\left(Q\left(Q^{-1}(y,\theta), \theta\right)\right) \left| \frac{d}{dt} Q(t,\theta) \right| \left| 1/\frac{d}{dt} Q(t, \theta) \right| \\ &= g(y) \end{aligned}$$

what is this???

Example 7.7

- *$X \sim$ Uniform $\left(\mu - \frac{1}{2}, \mu + \frac{1}{2}\right) \rightarrow f_X(x \mid \mu) = I_{\left(\mu-\frac{1}{2},\mu+\frac{1}{2}\right)}(x) = I_{\left(-\frac{1}{2}\frac{}{2}, \frac{1}{2}\right)}(x - \mu) = f(x - \mu)$, where $f(x - \mu)$ does not depend on μ. Thus $f(x - \mu)$ is a location pdf and $\bar{X} - \mu$ is a pivotal quantity.*
- *The previous Exponential example is of the form $f(x/\lambda)$, a scale pdf, and $\frac{\bar{X}}{\lambda}$ is a pivotal quantity.*
- *$X \sim N\left(\mu, \sigma^2\right) \rightarrow f_X\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$ where $f(t) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2}\right\}$ is the standard normal distribution. Thus $f\left(\frac{\bar{X}-\mu}{\sigma}\right)$ is a location-scale pdf and $\frac{\bar{X}-\mu}{\sigma}$ is a pivotal quantity.*

Example 7.8 *In one of our previous examples, we had $\bar{X}/\lambda \sim \mathrm{Gamma}\left(n, \frac{1}{n}\right)$. While the quantity $\bar{X}/\lambda$ depends on λ, its distribution does not. Using this distribution, we can find t_1 and t_2 such that*

$$\delta = P\left\{t_1 \leq \bar{X}/\lambda \leq t_2\right\} = P\left\{\frac{\bar{X}}{t_2} \leq \lambda \leq \frac{\bar{X}}{t_1}\right\}$$

So $\left(\frac{\bar{x}}{t_2}, \frac{\bar{x}}{t_1}\right)$ is a $100\delta\%$ confidence interval for λ. The expected length is $\left(\frac{1}{t_1} - \frac{1}{t_2}\right)\lambda$. How does this compare to $2c\frac{\lambda}{\sqrt{n}}$ from our previous example?

So $\bar{X}/\lambda$ is a pivotal quantity.
The idea is we can find q_1 and q_2 such that

read Mathematics Statistics

$$\delta = P\{q_1 \leq Q(\boldsymbol{X}, \theta) \leq q_2\}$$

and construct the confidence set

$$\{\theta : q_1 \leq Q(\boldsymbol{X}, \theta) \leq q_2\}$$

- This set is not always an interval, but it is when $Q(\boldsymbol{X}, \theta)$ is monotonic in θ.

How to construct a pivot interval

$T = \sum_{i=1}^{n} X_i/n$ is sufficient for the Exponential λ. We can show that $T \sim \mathrm{Gamma}(n, \lambda/n)$, where the form of the distribution is

$$f(t \mid n, \lambda/n) = \frac{\mathrm{n}^{\mathrm{n}}}{\Gamma(n)\lambda}\left(\frac{t}{\lambda}\right)^{n-1} e^{-\frac{nt}{\lambda}}, 0 \leq x < \infty, \lambda > 0$$

Let $Q(T, \lambda) = T/\lambda$ and thus $\frac{d}{dt}Q(t, \lambda) = 1/\lambda$. So if

$$g(y) = \frac{\mathrm{n}^{\mathrm{n}}}{\Gamma(n)}(y)^{n-1}e^{-ny}$$

then

$$f(t \mid n, \lambda/n) = g(Q(t, \lambda))\left|\frac{d}{dt}Q(t, \lambda)\right|$$

I am so confused what's the goal of this proof

■

7.1.4 Wald CI

7.1.5 Wilson's Score CI

7.1.6 Exact CI- Clopper Person

Find CI by MLE asymptotic

- See How in [Consistency, Asymptotic Variance, efficiency of MLE]
- Not a lot of example for this topics
- Not familiar with this topics

Definition 7.6 *Asymptotic variance*

- *First an example: We know that the sample mean,* $\bar{X}_n$*, from a* $N\left(\mu, \sigma^2\right)$ *distribution has variance* σ^2/n*. Thus,* $\sigma_n^2 = \operatorname{Var}\left[\sqrt{n}\bar{X}_n\right] = \sigma^2$*.*

- *In general, if* $X_1, \ldots, X_n$ *is an iid random sample from a distribution where the second moment exists (thus the distribution has a finite variance), then the central limit theorem says that* $\frac{\bar{X}_n - \mathrm{E}\left[\bar{X}_n\right]}{\sqrt{\operatorname{Var}[X]/n}} \to N(0,1)$ *in distribution.*

- *Define* $\sigma_{\mathrm{n}}^2 = \operatorname{Var}\left[\sqrt{n}\bar{X}_n\right]$*. Then* $\sigma_{\mathrm{n}}^2 \to \operatorname{Var}[X] = \tau^2$*. The constant* τ^2 *is the limiting variance or the limit of the variances.*

- *Definition: For an estimator* T_n*, suppose that* $k_n\left(T_n - \tau(\theta)\right) \to N\left(0, \sigma^2\right)$ *in distribution. The parameter* σ^2 *is called the asymptotic variance or variance of the limit distribution of* T_n*.*

This is general Delta methods?

Definition 7.7 *Asymptotic Efficiency*

Definition: A sequence of estimators W_n *is asymptotically efficient for a parameter* $\tau(\theta)$ *if*

$$\sqrt{n}\left[W_n - \tau(\theta)\right] \to N(0, v(\theta))$$

in distribution and

$$v(\theta) = \frac{\left[\tau'(\theta)\right]^2}{\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\ln f(X \mid \theta)\right)^2\right)} = \frac{\left[\tau'(\theta)\right]^2}{-I(X)}$$

that is, the asymptotic variance of W_n *achieves the Cramér-Rao Lower Bound.*

Example 7.14 *:*

If $\hat{\theta}_{jn}$ *satisfies tA.GTwith asymptotic covariance matrix* $V_{jn}(\theta)$ *,* $j = 1, 2$ *, and* $V_{1,n}(\theta) \leq V_{2n}(\theta)$ *(in the sense that* $V_{2n}(\theta) - V_{1n}(\theta)$ *is nonnegative definite for al* $\theta \in \Theta$ *, then* $\hat{\theta}_{1n}$ *is said to be asymptoticlly more efficient than* $\hat{\theta}_{2n}$

Theorem 7.6 *Asymptotic Efficiency of MLE*

Let $X_1, X_2, \ldots,$ *be an iid random sample from a family of distributions indexed by* $\theta \in \Theta$*. Let* $\hat{\theta}$ *denote the MLE of* θ *and let* $\tau(\theta)$ *be a continuous function of* θ*. Under certain regularity conditions,*

$$\sqrt{n}[\tau(\hat{\theta}) - \tau(\theta)] \to N(0, v(\theta))$$

where $v(\theta)$ *is the Cramér-Rao Lower Bound. That is,* $\tau(\hat{\theta})$ *is a consistent and asymptotically efficient estimator of* $\tau(\theta)$*.*

Thus, the bounds of an approximate $100(1-\alpha)\%$ *confidence interval for* θ *is given as*

$$\tau(\hat{\theta}) \pm z_{\alpha/2}\sqrt{v(\hat{\theta})/n} \text{ or } \tau(\hat{\theta}) \pm z_{\alpha/2}\sqrt{\hat{v}(\hat{\theta})/n}$$

Corollary 7.6.1 • *The quantity $v(\hat{\theta})$ may be found by applying the delta method to the transformation $\tau(\cdot)$ evaluated at $\theta = \hat{\theta}$ and using $\sigma^2 = I_1^{-1}(\theta)$; that is, $v(\hat{\theta}) = \left.\frac{[\tau'(\theta)]^2}{I_1(\theta)}\right|_{\theta=\hat{\theta}}$*

- *For a given n, the approximate (unobservable) variance of $\tau(\hat{\theta})$ is $v(\theta)$. To use this in practice, we must estimate the approximate variance.*
  - *First estimator: MLE of $v(\theta)$, given by $v(\hat{\theta})$.*
  - *This uses $I_1(\hat{\theta})$, the expectation of $-\frac{\partial^2 \ln(f(X_i|\theta))}{\partial\theta^2}$ evaluated at $\theta = \hat{\theta}$*
  - *We call this the expected information number.*
  - *Alternatively, we may estimate $I_1(\theta)$ with $-\frac{1}{n}\sum \frac{\partial^2 \ln(f(X_i|\theta))}{\partial\theta^2}$ evaluated at $\theta = \hat{\theta}$.*
  - *We call this the observed information number, denoted as $\hat{I}_1(\hat{\theta})$.*
  - *The variance estimator using the observed information number is denoted as $\hat{v}(\hat{\theta})$.*
  - *It has been shown that this provides a better estimator. (Efron & Hinkley, Biometrika, 1978)*
  - *If you can differentiate and evaluate the log-likelihood, you can calculate $\hat{v}(\hat{\theta})$.*

Example 7.15 *Suppose $X_1, \ldots, X_n$ is an iid random sample from Poisson (λ). The MLE of λ is $\hat{\lambda} = \bar{X}$.*

- - *Construct a* 95%Cl *for λ, where* n = 100 *and $\bar{x} = 0.5$.*

  - *Large sample via the MLE*

$$\begin{aligned}
&L(\lambda \mid x) = e^{-\lambda}\lambda^x/x! \\
&l(\lambda \mid x) = -\lambda + x\log\lambda - \log(x!) \\
&\frac{\partial l}{\partial\lambda} = -1 + \frac{x}{\lambda}; \frac{\partial^2 l}{\partial\lambda^2} = -\frac{x}{\lambda^2} \\
&I_1(\lambda) = -E\left[-\frac{x}{\lambda^2}\right] = \frac{\lambda}{\lambda^2} = \frac{-1}{\lambda} \\
&I_1(\hat{\lambda}) = \frac{1}{\bar{x}} \\
&v(\lambda) = \bar{x}
\end{aligned}$$

  - *Pivot the cdf*

  - *Recall that $F_T(T \mid \theta) \sim \mathrm{Uniform}(0,1)$ if F_T is continuous in T. That is, $F_T(T \mid \theta)$ has the same distribution for all values of θ. Thus $F_T(T \mid \theta)$ is a pivotal quantity.*

  - *Hence $P\{\alpha_1 \leq F_T(t \mid \theta) \leq 1 - \alpha_2\} = 1 - \alpha_1 - \alpha_2$.*

- *The following also works if F_T is not continuous in T.*

- *If $F_T(T \mid \theta)$ is non-increasing in θ then for given t* $\theta_L = \inf\{\theta : F_T(t \mid \theta) \leq 1-\alpha_2\}$ *and* $\theta_U = \sup\{\theta : F_T(t \mid \theta) \geq \alpha_1\}$.

- *If $F_T(T \mid \theta)$ is non-decreasing in θ then for given t* $\theta_L = \inf\{\theta : F_T(t \mid \theta) \geq \alpha_1\}$ *and* $\theta_U = \sup\{\theta : F_T(t \mid \theta) \leq 1-\alpha_2\}$.

Note $\sum_1^n x_i \sim \text{Porsson}(n\lambda)$

$$CLT \to \frac{\frac{1}{n}}{\sum_n} \frac{\sum x_i - n\lambda}{\sqrt{n\lambda}} = \sqrt{n}\left(\frac{\bar{x}-\lambda}{\sqrt{\lambda}}\right) \underset{\text{mis}_{dist.}}{\to} N(0,1)$$

$\Phi\left(\sqrt{n}\left(\frac{\bar{x}-\lambda}{\sqrt{\lambda}}\right)\right)$ *is a pivotal quantity.* Redo this

- - *Construct a* 95%Cl *for* $\theta = P_\lambda\{X=0\}$. - *Large sample via the MLE*

Find CI by Inverting Test

- More Example on Mathematics Statistics
- This is related to [Delta Method]

9 Hypothesis and CI Inverting

Example 9.1 *Let $X_1, X_2, \ldots, X_n \sim$ iid $N\left(\mu, \sigma^2\right)$ with $\left(\mu, \sigma^2\right)$ unknown.*

- Find the level α LRT for $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 \neq \sigma_0^2$.

- The unrestricted MLE of σ^2 is given by $\hat{\sigma}^2 = \frac{n-1}{n} S^2$ while the restricted MLE is $\hat{\sigma}_0^2 = \sigma_0^2$. In both cases, $\hat{\mu} = \bar{X}$. The likelihood ratio is given by

$$
\begin{aligned}
\lambda(\mathbf{x}) = & \frac{L\left(\hat{\mu}, \sigma_0^2\right)}{L\left(\hat{\mu}, \hat{\sigma}^2\right)} = \frac{[2\pi]^{-n/2}\left[\sigma_0^2\right]^{-n/2} \exp\left\{-\frac{1}{2\sigma_0^2} \Sigma\left(x_i - \bar{x}\right)^2\right\}}{[2\pi]^{-n/2}\left[\frac{n-1}{n} S^2\right]^{-n/2} \exp\left\{-\frac{n}{2}\right\}} \\
= & \left[\frac{n-1}{\sigma_0^2} S^2\right]^{n/2} \exp\left\{-\frac{1}{2} \frac{n-1}{\sigma_0^2} S^2\right\}\left[\frac{1}{n}\right]^{n/2} \exp\left\{\frac{n}{2}\right\}
\end{aligned}
$$

- Let $T = T_{\sigma_0^2}(\mathbf{X}) = \frac{n-1}{\sigma_0^2} S^2$

- Then

$$
\begin{aligned}
\lambda(\mathbf{X}) &= T^{n/2} \exp\left\{-\frac{1}{2} T\right\}\left[\frac{1}{n}\right]^{n/2} \exp\left\{\frac{n}{2}\right\} \\
\log \lambda(\mathbf{X}) &= \frac{n}{2}\left(\log T - \frac{1}{n} T - \log n + 1\right)
\end{aligned}
$$

- We reject the null hypothesis and support the research hypothesis if $\lambda(\mathbf{X}) < k$ or $\log \lambda(\mathbf{X}) < k^*$.

- Given $\boldsymbol{x}$, let $g(t) = \log \lambda(\boldsymbol{x})$. When $g(t) < k^*, T < t_1$ or $T > t_2$, as illustrated in the figure (here, $n = 25$).

- How might we use this test to construct an interval estimator for σ^2 ?

graph

Theorem 9.1 *Given the rejection region R of a hypotheses test, the acceptance region is $A = R^c$.*

For each $\theta_0 \in \Theta$, let $A\left(\theta_0\right)$ be the acceptance region of a level α test of $H_0 : \theta = \theta_0$. For each $\boldsymbol{x} \in \mathcal{X}$, define a set $\mathcal{C}(\boldsymbol{x}) \subset \Theta$ by

$$
\mathcal{C}(\boldsymbol{x}) = \left\{\theta_0 : \boldsymbol{x} \in A\left(\theta_0\right)\right\}
$$

Then the random set $\mathcal{C}(\mathbf{X})$ is a $1-\alpha$ confidence set.

Conversely, let $\mathcal{C}(\mathbf{X})$ be a $1-\alpha$ confidence set. For any $\theta_0 \in \Theta$, define $A\left(\theta_0\right) = \left\{\mathbf{x} : \theta_0 \in \mathcal{C}(\mathbf{x})\right\}$. Then $A\left(\theta_0\right)$ is the acceptance region of a level α test of $H_0 : \theta = \theta_0$.

Proof. We prove the first assertion only. The proof for the second assertion is similar. Under the given condition, assuming an arbitrary $\theta_0 \in \Theta$

$$
\sup_{\theta=\theta_0} P\left(X \notin A\left(\theta_0\right)\right) = \sup_{\theta=\theta_0} P\left(T_{\theta_0} = 1\right) \leq \alpha,
$$

which is the same as

$$1-\alpha \leq \inf_{\theta=\theta_0} P\left(X \in A\left(\theta_0\right)\right)=\inf_{\theta=\theta_0} P\left(\theta_0 \in C(X)\right).$$

Since this holds for all θ_0, the result follows from

$$\inf_{P \in \mathcal{P}} P(\theta \in C(X))=\inf_{\theta_0 \in \Theta} \inf_{\theta=\theta_0} P\left(\theta_0 \in C(X)\right) \geq 1-\alpha. \quad .$$

The converse of Theorem 7.2 is partially true, which is stated in the next result whose proof is left as an exercise.

Corollary 9.1.1 *Proposition 7.2. Let $C(X)$ be a confidence set for θ with significance level (or confidence coefficient) $1-\alpha$. For any $\theta_0 \in \Theta$, define a region $A\left(\theta_0\right)=\left\{x: \theta_0 \in C(x)\right\}$. Then the test $T(X)=1-I_{A\left(\theta_0\right)}(X)$ has significance level α for testing $H_0: \theta=\theta_0$ versus some H_1.*

Example 9.2 *CONTINUE:*

- *The rejection region is*

$$R\left(\sigma_0^2\right)=\left\{\boldsymbol{x}: T_{\sigma_0^2}(\boldsymbol{x})<t_1 \text{ or } T_{\sigma_0^2}(\boldsymbol{x})>t_2\right\}$$

thus, the acceptance region is

$$A\left(\sigma_o^2\right)=\left\{\mathbf{x}: t_1 \leq T_{\sigma_0^2}(\boldsymbol{x}) \leq t_2\right\}$$

- *Define $\mathcal{C}(\boldsymbol{x})=\left\{\sigma_0^2: \boldsymbol{x} \in A\left(\sigma_o^2\right), \sigma_0^2>0\right\}$. Then $\mathcal{C}(\mathbf{X})$ is a $1-\alpha$ confidence set.*
- *Since $A\left(\sigma_0^2\right)=\left\{\boldsymbol{x}: t_1 \leq \frac{n-1}{\sigma_0^2} S^2 \leq t_2\right\}$, then*

$$\mathcal{C}(\mathbf{x})=\left\{\sigma_0^2: \frac{n-1}{t_2} S^2 \leq \sigma_0^2 \leq \frac{n-1}{t_1} S^2\right\}$$

- *For proper choice of (t_1, t_2), we have $\left(\frac{n-1}{t_2} S^2, \frac{n-1}{t_1} S^2\right)$ is a $100(1-\alpha)\%$ Cl for σ^2.*

Example 9.3 *CONTINUE*

We choose t_1 and t_2 so that the test is size α.

- *Values of t_1 and t_2 must be such that $g\left(t_1\right)=g\left(t_2\right)=k^*$ (refer to earlier figure). Thus $g\left(t_1\right)-g\left(t_2\right)=0$.*
- *$T_{\sigma_0^2}(\mathbf{X}) \sim \chi_{n-1}^2$, central chi-squared, when $H_0: \sigma^2=\sigma_0^2$ is true.*
- *Let F denote the cdf of a χ_{n-1}^2 distribution. Then t_1 and t_2 must satisfy $F\left(t_1\right)+1-F\left(t_2\right)=\alpha$*
- *Define*
- *$h_1\left(t_1, t_2\right)=g\left(t_1\right)-g\left(t_2\right)$*

Two Sample Testing

Assumption

Table 3: Hypothesis Testing

	Assumption	Normal	T-Distribution
A	1. Independent 2. Finite Variance	Approx by CLT, Lindeberg-Feller	NA
B	1. Independent 2. Finite Variance 3. Identical	Approx by CLT	Approx By CLT
C	1. Independent 2. Unknown Finite Variance 3. Identical 4. X_i Normality	Approx by LLN	Exact

Simon's Note Assumption Simon's Note hypothesis testing

Two sample t-test

unequal variance

Given $E[\bar{X}_1 - \bar{X}_2] = \mu_1 - \mu_2$

by independent

$$
\begin{aligned}
Var(\bar{X}_1 - \bar{X}_2) &= \frac{\sum Var(X_1)}{n_1^2} + \frac{\sum Var(X_1)}{n_1^2} \\
&= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \\
&\text{for unknown variance, use sample variance approx} \\
&\approx \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}
\end{aligned}
$$

Test statistics:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Under Null hypothesis $\mu_1 = \mu_2$, T follow t-distribution with degrees of freedom: $\frac{(s_1^2/n_1+s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1)+(s_2^2/n_2)^2/(n_2-1)}$

equal variance

Given $E[\bar{X}_1 - \bar{X}_2] = \mu_1 - \mu_2$

by independent

$$Var(\bar{X}_1 - \bar{X}_2) = \frac{\sum Var(X_1)}{n_1^2} + \frac{\sum Var(X_1)}{n_1^2}$$
$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$
$$\text{under null, use pooled sample variance approx}$$

because sample size can be different, using sample-size adjusted weighted estimator for the pooled sample variance

$$s_{pool}^2 = \frac{n_1 - 1}{n_1 + n_2 - 2} s_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} s_2^2$$

Test statistics:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

where s_p is the pooled standard deviation:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

with degrees of freedom: $n_1 + n_2 - 2$

Bayesian

Posterior

Given a parameter θ, which has a prior of $P_\Theta(\theta)$, observed data X_i, which has a likelihood function of $P(X_i|\theta) = L(\theta|X_i)$,

The posterior probability is

$$\begin{aligned} P(\hat{\theta}|X) &= \frac{P(\theta, X)}{P(X)} \\ &= \frac{P(X|\theta)P(\theta)}{P(X)} \\ &= \frac{L(\theta|X)P(\theta)}{\int_{\Theta|X} L(\theta|X)P(\theta)d\theta} \\ &\propto L(\theta|X)P(\theta) \end{aligned}$$

Credible Interval

They characterize the range of θ that are most believable on the basis of the data (and the prior belief regarding θ).

In Bayesian world, we say there is a 95% probability that θ lies between the endpoints of a 95% credible interval (but the word "chance" in place of "probability" is still not ideal).

Keep in mind: Talking about probability in the subjective sense, and not in the long-term frequency sense.

Regression

Linear Regression

Projection

Assumption

For inference β - i.i.d - We don't need *Linear* if N sufficient large and X random - We don't need *Normality* since we have CLT - We don't need *Homoscadecity aka. equal variance* since we have robost variance estimator and N sufficient large

For sub-group specific mean $E[Y|X]$ - i.i.d - Linear - We don't need *normality* - We don't need *Homoscadecity aka. equal variance* since we have robust variance estimator and N sufficient large

For prediction $E[\hat{Y}|X]$ - i.i.d - Linear - normality - Homoscadecity

Algreba Form

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- $\hat{\beta}$ is unbiased

Assuming $E[y|X] = X\beta$:

$$\begin{aligned} E[\hat{\beta}] &= E[E[\hat{\beta}|X]] \\ &= E[(X^T X)^{-1} X^T y|X] \\ &= (X^T X)^{-1} X^T E[y|X] = (X^T X)^{-1} X^T X\beta \\ &= \beta \end{aligned}$$

- Variance of $\hat{\beta}$

Assumption: 1. error terms are pairwise independent 2. $Var[Y|X=x] = \sigma^2$ for all x (homoscedacity), this assumption can be elimated with robust estimator

$$\begin{aligned} Var(\hat{\beta}) &= Var[E[\hat{\beta}|X]] + E[Var(\hat{\beta}|X)] && \text{total variance} \\ &= \cancel{Var(\beta)} + E[Var(\hat{\beta}|X)] && \text{constant} \\ &= E[Var((X^T X)^{-1} X^T y|X)] && \text{Linearity} \\ &= E[(X^T X)^{-1} X^T * Var(y|X) * ((X^T X)^{-1} X^T)^T] && \text{algreba} \\ &= E[(X^T X)^{-1} X^T * (\sigma^2 I) * ((X^T X)^{-1} X^T)^T] && \text{homoscedacity} \\ &= E[\sigma^2 (X^T X)^{-1}] \end{aligned}$$

Where σ^2 is estimated with MSE under homoscedasticity. Under heteroscedasticity we use sandwich variance. If there is heteroscedasticity, then the MSE can instead be interpreted as estimating the average within-group variance (i.e., averaged over the values of X).

Interpretation

β_0:

- β_0 corresponds to a theoretical subgroup that is not well represented by the data
- If the predictor of interest, X, were something like height, then the intercept (β_0) would mark the mean HbA1c among those of height zero.
- Here, β_0 does not even carry a real-world interpretation, let alone the fact that our data cannot reasonably be used to reliably estimate it.
- by removing the intercept from the model we've actually made a very huge, untestable, and very often untrue assumption that the line goes through the origin! That is, by the above model, $E[Y|X=0] = 0$ necessarily.

β_1**:**

For $E[Y|X] = \beta_0 + \beta_1 x$

$c\beta_1$ denotes difference in mean Y between subgroups differing in X by c units.

For $E[log(Y)|X] = \beta_0 + \beta_1(x)$

$\beta_1 log(1+q) \Rightarrow$ Difference in mean Y between subgroups differing in X by (100 × q)%

For $E[log(Y)|X] = \beta_0 + \beta_1 x$

e^{β_1} denotes Geometric mean ratio, comparing subgroups differing in X by one unit.

Right-skewness is again not a justifiable reason on its own to log-transform an outcome.

Example:

Consider following example:

$$E[log(Y)|X] = \beta_0 + \beta_1(log(x) - 5)$$

- $exp(\beta_0) \Rightarrow$ Geometric mean Y among subgroup $log(x) - 5 = 0 \rightarrow x = exp(5)$.
- $1.6^{\beta_1} \Rightarrow$ Geometric mean ratio of Y compareing subgroup differing in X by $exp(log(1.6)\beta_1) \rightarrow log(1.6)$ unit

Multivariate Regression

Spline

Example: B5

The most general form for f is as follows:

$$f(x) = \begin{cases} \alpha_0 & \text{if } 0 < x \leq c \\ \gamma_0 + \gamma_1 x + \gamma_2 x^2 & \text{if } x > c' \end{cases}$$

for real numbers $\alpha_0, \gamma_0, \gamma_1$, and γ_2. First, let's impose the continuity constraint, which is that:

$$\lim_{x \to c^-} f(x) = \lim_{x \to c^+} f(x) \Rightarrow \alpha_0 = \gamma_0 + \gamma_1 c + \gamma_2 c^2.$$

Updating the function so far, it takes the following form:

$$f(x) = \begin{cases} \gamma_0 + \gamma_1 c + \gamma_2 c^2 & \text{if } 0 < x \leq c \\ \gamma_0 + \gamma_1 x + \gamma_2 x^2 & \text{if } x > c' \end{cases}$$

for real numbers γ_0, γ_1, and γ_2. Next, let's impose the differentiability constraint, which is that:

$$\lim_{x \to c^-} f'(x) = \lim_{x \to c^+} f'(x) \Rightarrow 0 = \gamma_1 + 2\gamma_2 c \Rightarrow \gamma_1 = -2\gamma_2 c$$

Updating the function so far, it takes the following form:

$$f(x) = \begin{cases} \gamma_0 - 2\gamma_2 c^2 + \gamma_2 c^2 & \text{if } 0 < x \le c \\ \gamma_0 - 2\gamma_2 cx + \gamma_2 x^2 & \text{if } x > c \end{cases} = \begin{cases} \gamma_0 + \gamma_2 \left(-c^2\right) & \text{if } 0 < x \le c \\ \gamma_0 + \gamma_2 \left(x^2 - 2cx\right) & \text{if } x > c \end{cases}$$

for real numbers γ_0 and γ_2. Re-expressing as a basis expansion,

$$f(x) = \beta_0 + \beta_1 \left[\left(x^2 - 2cx\right) 1_{(c,\infty)}(x) - c^2 1_{(0,c]}(x)\right], \quad (\beta_0, \beta_1) \in R^2$$

The basis function of interest is given by $h_c(x) = \left(x^2 - 2cx\right) 1_{(c,\infty)}(x) - c^2 1_{(0,c]}(x)$. Now, $f'(x) = h'_c(x) = \beta_1(2x - 2c)1_{(c,\infty)}(x)$, which is continuous; further, $\lim_{x \to c^-} h''_c(x) = 0$ and $\lim_{x \to c^+} h''_c(x) = 2\beta_1$, so $f''(x)$ clearly not defined at $x = c$.

The unconstrained model uses four degrees of freedom, which makes sense because it involves an unrestricted constant (one degree of freedom) and an unrestricted quadratic function (three degrees of freedom). One degree of freedom is taken away by imposing the continuity constraint, and another is taken away by imposing the differentiability constraint. Therefore, the total number of degrees of freedom is given by 2, matching the number of basis functions.

Weighted Regression

Gauss-Markov Theorem

Gauss-Markov Theorem: Correct specified weighted least square estimator has the minumum variance ammong all **unbiased linear estimator** of β, if normality, linearity, no collinearity and finite variance assumptions are satisfied.

aka. Best Linear Unbiased Estimator.

- The OLS estimator is the best (i.e., minimum variance) linear (in y) unbiased (for β) estimator of β under the assumptions of the Gauss-Markov theoerm if the errors are homoscedastic.
- The OLS estimator is unbiased even if homoscedasticity does not hold.
- The OLS estimator is less efficient than WLS (with $W \propto V^{-1}$) if homoscedasticity does not hold.
- The sandwich variance is still valid even if $W \not\propto V^{-1}$, but WLS estimator is no longer BLUE.

Saturated Model

The idea of saturated model is that no information is borrowed across subgroup, such that we can estimate sub-group specific mean/odd/...

- A model is saturated if the number of parameters and constrains add up to the number of groups
- *Example: Comps Applied 2022.3*

	Y = 1	Y = 2	Y = 3
X = 0	a	d	g
X = 1	b	e	h
X = 2	c	f	i

$$logit(P(Y \geq 1|X = x)) = \alpha_0 + \alpha_1 I(x \geq 1)$$

- This is a logistic model, and we have 4 groups ($\{x = 0, x \geq 1\} \cap \{Y = 0, Y \geq 1\}$), and we have 2 paramteres and 2 constrains ($P(Y \geq 1|X = x) + P(Y = 0|X = x) = 1$) so the model is saturated.

$$log(\frac{P(Y = k|x = x)}{P(Y = 0|X = x)}) = \beta_{0k} + \beta_{1k} I(x = 1) + \beta_{2k} I(x = 2), k = 1, 2$$

- This is a multinomial model, we have 9 groups(3*3), and the model have 6 parameter(β_{ij}) and 3 constrains $\sum_k P(Y = k|X = x) = 1$, so the model is saturated.

$$logit(P(Y \leq k|X = x)) = \gamma_{0k} - \gamma_1 I(x = 1) - \gamma_2 I(x = 2), k = 0, 1$$

- This is a proportional odds model, this model cannot be saturated because it's borrowing information from the left-side of the equation.

Multivariate Regression

Logistic Regression

Ordinal Regression

Poisson Regression

Longitudinal Regressoin

Survival Regression

Integral

Subsitution

Distribution

Discribe

Poisson

Note: $Y \sim Poi$, $aY \nsim Poi$ because Y is discrete.

Simulation

COMPS Practice

Table 5: Comps Practice

Year	Question	Related field	Time spend	Turn out
2021-T	1a-c	1. pdf, cdf 2. E, Var	short	good
	1 d-f	1. [Asymptotic Probability]	long	don't know how to come up with the thing
	2	1. Conditional Probability 2. [Multivariate Transformation] 3. [Binormal Distribution]	3h	Can't do Binormal pdf, Expectation and variance Can't do conditional expectation
	3a	Transformation	short	good
	3b	1. Order Statistics 2.	Long	didn't finish
	3c-d	1. [Hypothesis Testing]	Long	Didn't finish
	4a-d	• CLT • Delta		
	4e	• Converge • Taylor Series • Second Order Delta		Can't do it Very easy if you think of second order Delta
	5a-c	• MOM • UMVUE • Rao-blackwell	Long	Fisher info can only applied to exponential family CRLB only apply for family distribution that support doesn't depend on paramter
	5d-f	• Bayesian		

Year	Question	Related field	Time spend	Turn out
2021-A	6	• Gal-Markov • lagrange multiplier • Order Statsitics		Don't know how to get the order statistics distribution
2022-A	1c	[T-test]	Long	Didn't Finish, still don't understand equal variance vs. unequal variance
	1g	[T-test]	Long	Also don't understand the assumption to get a exact t-test, approx. t-test.
	2	1. Log-transform 2. Logistic	medium	Not familiar with log transform value intepretation Not familiar with the idea of Risk and Odd Not familiar with Logisitcs
	3a	Saturated	Long	Don't understand how a model can be saturated Don't understand how constrains works on the saturated
	3b	1. logistic 2. Multiordinal 3. Log-odd	Long	Not familiar with the forms of the model Don't know the proportional odd assumption Don't know how to hand calculated logistic model Don't know how to hand calcualted Log-odd Model
	4	Baysian	Okayish	I don't know
	5	1. Simulation 2. Bootstrap CI	Okayish	Not sure how to evaluate CI performance Not sure how boostrap CI
	6	• Non-linear model • Delta method	Long	Not familiar with multivariate delta Methods
2022-T	1b	• Geometric • MGF		Don't know the MGF of Geometric Don't know the trick to find MGF
	1j	AUC		Don't understand the upper bound of AUC
	2c			

Year	Question	Related field	Time spend	Turn out
	2d	Transformation		Transformation with Jacobian apply to pdf Non-independent b/c of Indicator function
	2g	• Slusky • LLN		good, but worth reviewing
	2h	• CLT • Delta Method		Confused at which variance should use in delta method
	3c	• UMVE • Complete		Completely forget about Lehmann-scheffe
	3d			Don't know how to prove
	3e			Completely no idea
	4a-b	• MGF		Bad, not thinking of MGF and got stuck
	4c	Method of Moment		Still don't know how to get a MOM estimator
	4d	Asymptotic Distribution		How to determine asymptotic distribution? converge in distribution?
	5			
	6a	MLE	fast	Like bob's
	6b-c	bayesian		
	6d-f	Unbiased	Bad	F