MATH324 (Statistics) – Lecture Notes

McGill University

Prof. Masoud Asgharian

Sam K.H.Targhi-Dunn

sam.targhi@mail.mcgill.ca

Fall 2019

Contents

1 Lecture 1

Overview of statistics

Point estimation:

Statistic and estimator + examples

Bias and Mean Square Error

Unbiasedness , Bias , $MSE(\hat{\theta})$, Decomposition of MSE (# 8.8 ($\hat{\theta}_1, \hat{\theta}_5$) , *page* 394 , # 8.6 , , *page* 394)

Common Unbiased Estimators

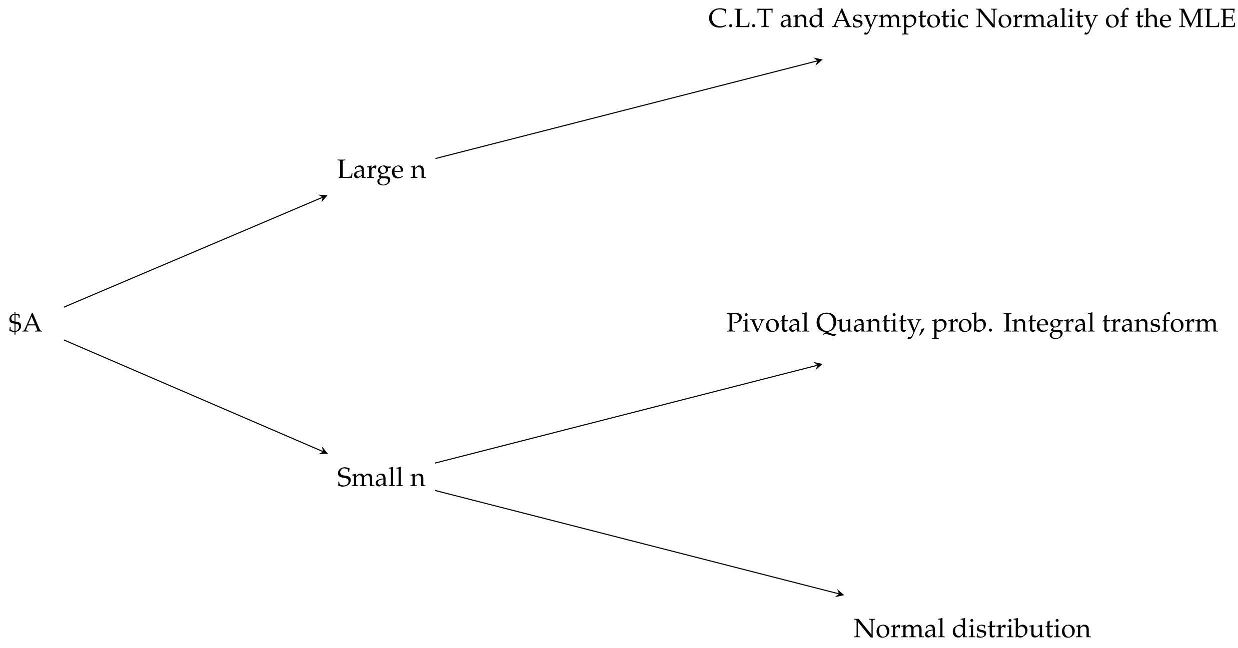
$\mu,\ p\ ,\ \mu_1 - \mu_2\ ,\ p_1 - p_2\ ,$ & σ^2 ($S^2_{n-1} \rightarrow S^2$ in the textbook)

Error of the Estimation

$\epsilon = |\hat{\theta} - \theta| \longrightarrow$ Tschebyscheff's Theorem if $\hat{\theta}$ is an unbiased estimator(example 8.2, page 401)

Confidence Interval

Pivotal Quantities



Pivotal

$$
\begin{aligned}
X_i \sim F_\theta(x) &\implies Y_i = F_\theta^{-1}(X_i) \ \sim \ Unif(0,1) \implies -\log Y_i \ \sim \ Exp(1) \\
&\implies \sum_{i=1}^{n} -\log F_\theta^{-1}(X_i) = \sum_{i=1}^{n} Y_i \sim G(n,1) \\
&\qquad Example: \quad \mathrm{X}_i \sim Exp(\lambda)
\end{aligned}
$$

Large n:

$\frac{\hat{\theta}_n - \theta}{\sigma_{\hat{\theta}}} \ \sim \ N(0,1) \quad for\ large\ n$ Example : $X_i \ \sim \ N(\mu, 1)$

Sample Size Determination:

Use the notes for 203.

1.1 Overview

- Statistics
  - Data Collection
    - Observational
    - Experimental
  - Modeling DGP
    - Parametric
    - Non-param.
    - Semi-Param.
  - Stat.Inf.
    - TPE
    - TSH
    - C.I
  - Com

In MATH324 we cover statistical Inference (Theory of Point Estimator (TPE) , Testting Statistical Hypothesis (TSH) , Confidence Interval (C.I) and Data Generating Process (DGP).

1.2 Parametric Models

Model is known up to finitely many unknown parameters.
E.g. $X_i \underset{i=1,2,\ldots,n}{\overset{iid}{\sim}} N(\mu, \sigma^2)$ where μ and σ^2 are unknown.
Note: "*iid*" means **I**ndependent **i**dentically **d**istributed

"$\sim$" means **Distributed according to**

1.3 Nonparametric Models

$X_i \overset{iid}{\sim} F_X(x)$ where the cdf F is completely unknown, but we may assume that F is smooth, for instance continuous or differentiable.
In the non-parametric setting $F_X(x)$ should be estimated for every x. Thus for a random variable X that can assume infinitely many values, we need to estimate $F(x)$ at infinitely many values of x. This is, particularly , the case when X is a continuous random variable. Recall that $F_X(x) = P(X \leq x)$. Then the sample counterpart of $F_X(x)$ is $\frac{\#X_i \leq x}{n}$ for a sample $X_1, \ldots, X_n$. Define:

$$\mathcal{E}(t) = \begin{cases} 1 & \text{if } t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Then $\frac{\#X_i \leq x}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{E}(x - X_i)$ and $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n}(x - X_i)$ is the Empricial Cumulative Distribution Function (ECDF) .

1.4 Point Estimator

Suppose $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, 1)$ where:

$$N(\mu,1) : f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}} \ , \ x \in \mathbb{R} \ , \ \mu \in \mathbb{R}$$

We want to have an estimate of μ ; i.e. a scientific guess, based on the observations , $X_1, \dots, X_n$. Recall that $\mathbb{E}(X_i) = \mu$, $\mu = 1, 2, \dots, n$, μ is the population mean.

What is an "estimate"?

Statistic: A function of observations that does not depend on any unknown parameter.

Estimator: An estimator is a statistic that aims at estimating a function of the population unknown parameters.

Example. $X_i \overset{iid}{\sim} N(\mu, 1)$

$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ is a statistic and as an estimator of μ

$(\bar{X}_n - \mu)$ is NOT a statistic since it depends on μ , which is an unknown parameter. $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$ is a statistic , but not an estimator of μ . Note that $dim(S^2) = (dim\ \mu)^2$. For instance, if X_is are returns of a fund and measured in dollars (\$) , then *dim* of μ is \$ while the *dim* of S^2 is $\$^2$. Besides, μ can be negative while S^2 is *always non-negative*.

1.5 Estimation Error

Going back to the example above ($X_i \overset{iid}{\sim} N(\mu, 1)$, $i = 1, 2, \dots, n$) and choosing $\bar{X}_n$ as the estimator of μ . We often want to study $\mathcal{E} = |\bar{X}_n - \mu|$ or a function of $\mathcal{E}$. Starting with $\mathcal{E}$ itself, the first thing that comes to mind is $P(\mathcal{E} \geq \delta)$ for a prespecified δ or perhaps $\mathbb{E}(\mathcal{E})$. A well known tool for studying the latter is *Tchbyshev's Inequality* .

2 Lecture 2

2.1 Markov's Inequality

Let X be a random variable and h be a **non-negative** function; ie:

$$h : R \to R^+ \cup \{0\} \quad = \quad [0, \infty)$$

Suppose $E(h(X)) < \infty$,then for some $\lambda > 0$, we have:

$$P(h(X) \geq \lambda) \leq \frac{E[h(X)]}{\lambda} \tag{1}$$

Proof. Suppose X is a continuous random variable:

$$\begin{aligned}
E[h(x)] &= \int_x h(x) f_X(x) dx \\
&= \Big(\int_{x:h(x)\geq\lambda} h(x) f_X(x) dx + \int_{x:h(x)<\lambda} h(x) f_X(x) dx \Big) \\
&\geq \int_{x:h(x)\geq\lambda} h(x) f_X(x) dx && \text{since } h \geq 0 \\
&\geq \lambda \int_{x:h(x\geq\lambda} f_X(x) dx = \lambda \, P(h(X) \geq \lambda)
\end{aligned}$$

$$\implies P(h(X) \geq \lambda) \leq \frac{E(h(X))}{\lambda}$$

The proof for the discrete case is similar. □

2.2 Tchebyshev's Inequality

Tchebyshev's Inequality is a special case of Markov's Inequality. Consider $h(x) = (x - \mu)^2$, then:

$$\begin{aligned}
P(|X - \mu| \geq \lambda) &= P((X - \mu)^2 \geq \lambda^2) \\
&\leq \frac{E[(X - \mu)^2]}{\lambda^2} && if \; E[(X - \mu)^2] < \infty
\end{aligned}$$

Let $\mu = E(X)$, then $E[(X-\mu)^2] = Var(X)$ denoted by σ_X^2. We therefore have:

$$P(|X - \mu_X| \geq \lambda) \leq \frac{\sigma_X^2}{\lambda^2} \qquad \textit{where } \mu_X = E(X) \tag{2}$$

Now consider $\lambda = K\sigma_x$ where K is a known number. Then:

$$P(|X - \mu_X| \geq K\sigma_X) \geq \frac{\sigma_X^2}{K^2\sigma_X^2} = \frac{1}{K^2} \tag{3}$$

This is called **Tchbyshev's Inequality**.

Example 2.1. *Suppose* $K = 3$*.*

$$P(|X - \mu_x| \geq 3\,\sigma_x) \leq \frac{1}{9}$$

In other words, at least 88% *of the observations are within* 3 *standard deviation from the population mean.*

Going back to the our example:

$$X_i \sim (\mu, 1) \quad , \quad \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

We want to study $P(\epsilon \geq \delta) = P(|\bar{X}_n - \mu| \geq \delta)$, first we note that:

$$E(X_i) = \mu \quad , \; i = 1, 2, \ldots, n$$

Then:

$$\begin{aligned} E(\bar{X}_n) &= E\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}(n\mu) = \frac{1}{\cancel{n}}.(\cancel{n}\mu) \\ &= \mu \qquad\qquad\qquad (*) \end{aligned}$$

Thus, using (2) we have:

$$P(|\bar{X}_n - \mu| \geq \delta) \leq \frac{Var(\bar{X}_n)}{\delta^2}$$

Now:

$$\begin{aligned}
Var(\bar{X}_n) &= Var\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n^2} Var\Big(\sum_{i=1}^{n} X_i\Big) \\
&= \frac{1}{n^2}\Big[\sum_{i=1}^{n} Var(X_i) + \sum_{1\leq i<j\leq n}\ \sum_{1\leq i<j\leq n} Cov(X_i, X_j)\Big] \quad \textit{using Thm } 5.12(b) - \textit{page } 271 \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \qquad\qquad \textit{since } \coprod_{1}^{n} X_i \\
&= \frac{1}{n^{\cancel{2}}}\cancel{n} Var(X) = \frac{Var(X)}{n} \qquad\qquad \textit{since } x_i s \textit{ are identically distributed} \\
&= \frac{\delta_X^2}{n} \qquad\qquad (**)
\end{aligned}$$

In our case $X \sim N(\mu, 1)$ so $Var(X) = \delta_X^2 = 1$. Thus $Var(\bar{X}_n) = \frac{1}{n}$

Remark. $X \coprod Y \implies Cov(X, Y) = 0$. *Note that:*

$$X \coprod Y \implies E\Big[g_1(X)g_2(Y)\Big] = E[g_1(X)].E[g_2(Y)]$$

in particular:

$$X \coprod Y \implies E\Big[XY\Big] = E[X].E[Y]$$

on the other hand:

$$Cov(X, Y) = E[XY] - E(X)E(Y)$$

thus:

$$X \coprod Y \implies Cov(X, Y) = 0.$$

recall that $X \coprod Y$ means X and Y are independent, i.e. $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ where $f_{X,Y}, f_X and f_Y$ represent respectively the joint and marginal distributions.

We therefore have:

$$P(|\bar{X}_n - \mu| \geq \delta) \leq \frac{1}{n\delta^2} \tag{4}$$

Using (4) and the sample size, n, we can find an upper bound for the proportion of deviations which are greater than a given threshold δ.

We can also use (4) for Sample Size Deterministic:

Suppose δ is given and we want $P(|\bar{X}_n - \mu| \geq \delta) \leq \beta$ where β is also given. Then setting $\frac{1}{n\delta^2} = \beta$, we can estimate $n \approx \frac{1}{\beta\delta^2}$.

2.3 Application to Voting

Define $X_i = \begin{cases} 1 & \text{NDP} \\ 0 & \text{otherwise} \end{cases}$. Associated to each eligible voter in Canada we have a binary variable X. Let $p = P(X = 1)$. So p represents the proportion of eligible voters who favor *NDP*. Of interest is often estimation of p. Suppose we have a sample of size n, $X_1, X_2, \ldots, X_n$.

$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample proportion; The counterpart of p which nat be denoted by $\hat{p}$. Note that:

$$\mu_X = E(X) = 1 \times P(X = 1) + 0 \times P(X = 0) = 1 - p + 0 \times (1 - p) = p$$

and:

$$E(X^2) = 1^2 \times P(X = 1) + 0^2 \times P(X = 0) = 1 - p + 0 \times (1 - p) = p$$

From (*) and (**) we find that :

$$E(\hat{p}_n) = E(\bar{X}_n)\mu_X = p$$

and:

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

Thus using (2), we have:

$$P(|\hat{p}_n - p| \geq \delta) \leq \frac{Var(\hat{p}_n)}{\delta^2} = \frac{p(1-p)}{n\delta^2}$$

Note that the above bound on the probability of derivation depends on p which is *unknown*. We however notice that $p(1-p) \leq \frac{1}{4}$.

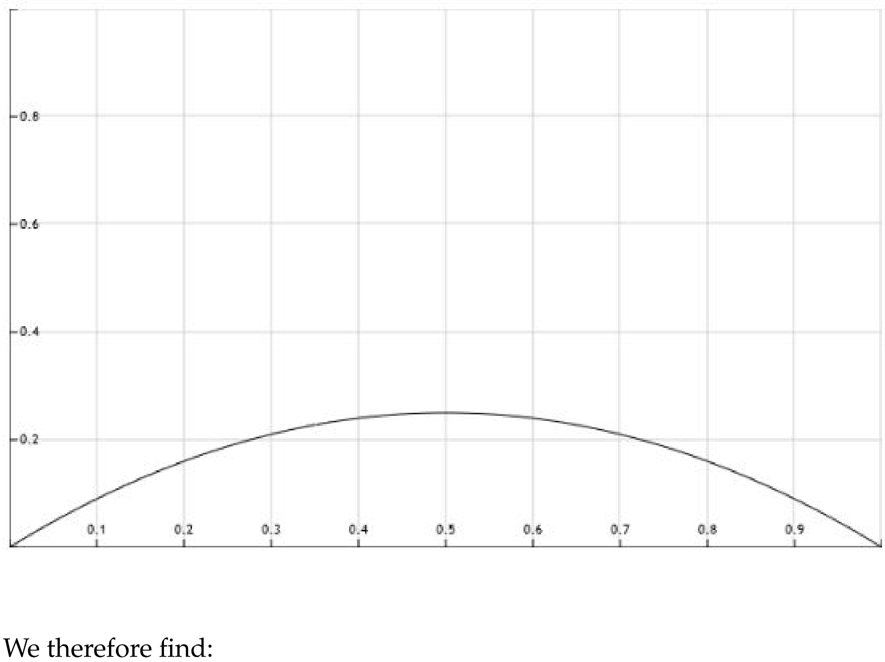
Define $\mathscr{C}(x) = x(1-x)\ for\ 0 < x < 1$. Then:

$$\mathscr{C}'(x) = 1 - 2x \implies \mathscr{C}'(x) = 0 \implies x = \frac{1}{2}$$

$$\mathscr{C}''(\frac{1}{2}) = -2 \implies x = \frac{1}{2} \qquad \text{which is a } \textbf{maximizer}$$

$$\mathscr{C}(\frac{1}{2}) = \frac{1}{2}(1 - \frac{1}{2}) = \frac{1}{4}$$

(Note that $\mathscr{C}''(x) = -2$ for all $0 < x < 1$)



We therefore find:

$$P(|\hat{p}_n - p| \geq \delta) \leq \frac{1}{4n\delta^2} \tag{5}$$

Using (5) and a given sample size n we can find an upper bound for the probability of derivation by δ and the amount for any given δ.
We can also use (5) for sample size deterministic for a size bound β and derivative δ as follows:

$$\frac{1}{4n\delta^2} = \beta \qquad \Longrightarrow \qquad n \geq \frac{1}{4\beta\delta^2}$$

This is of course conservative since $p(1-p) \leq \frac{1}{4}$.

3 Lecture 3

3.1 MSE

MSE: To study estimation error we started by studying $P(|\hat{\Theta}_n - \Theta| > \delta)$, deviation above a given threshold δ, by bounding this probability. One may take a different approach by studying average Euclidean distance, i.e. $E[\ |\hat{\Theta}_n - \Theta|^2]$, which denoted by **MSE($\hat{\Theta}_n$)**.
We note that if $\Theta = E(\hat{\Theta}_n)$, i.e. $\hat{\Theta}_n$ is an unbiased estimation of Θ, then:

$$MSE(\hat{\Theta}_n) = E[|\hat{\Theta}_n - \Theta|^2] = E[(\hat{\Theta}_n - \mu_{n_{\Theta_n}})^2] = Var(\hat{\Theta}_n)$$

Now recall that $Var(X) = 0 \implies P(X = \text{constant}) = 1$ which essentially means random variable X is a constant.
The same comment applies to MSE($\hat{\Theta}_n$). We want to find the closest estimator $\hat{\Theta}_n$ to Θ which means that we want to minimize $E[(\hat{\Theta}_n - \Theta)^2]$ over all possible estimators, ideally at least the above comment tells us that in real applications we cannot expect to find an estimator whose MSE is equal to zero. Let's try to understand the MSE a bit more:

$$\begin{aligned}
\text{MSE}(\hat{\Theta}_n) &= E[(\hat{\Theta}_n - \Theta)^2] \\
&= E\Big[\Big(\big(\hat{\Theta}_n - E(\hat{\Theta}_n)\big) + \big(E(\hat{\Theta}_n) - \Theta\big)\Big)^2\Big] \\
&= E\Big[\big(\hat{\Theta}_n - E(\hat{\Theta}_n)\big)^2\Big] + \big(E(\hat{\Theta}_n) - \Theta\big)^2 + 2 \cdot E\Big[\big(\hat{\Theta}_n - E(\hat{\Theta}_n)\big)\Big] \cdot \big(E(\hat{\Theta}_n) - \Theta\big) \\
&= E\Big[(\hat{\Theta}_n - E(\hat{\Theta}_n))^2\Big] + E\Big[\overbrace{(E(\hat{\Theta}_n) - \Theta)^2}^{\text{not a r.v.}}\Big] + 2 \cdot E\Big[\overbrace{\big(E(\hat{\Theta}_n) - \Theta)\big)}^{\text{not a r.v.}} \cdot \big(\hat{\Theta}_n - E(\hat{\Theta}_n)\big)\Big] \\
&= Var(\hat{\Theta}_n) + \Big[\overbrace{E(\hat{\Theta}_n) - \Theta}^{\text{Bias}(\hat{\Theta}_n)}\Big]^2 + 2 \cdot Bias(\hat{\Theta}_n) \cdot \overbrace{E[(\hat{\Theta}_n - E(\hat{\Theta}_n))]}^{E(\hat{\Theta}_n) - E(\hat{\Theta}_n) = 0} \\
&= Var(\hat{\Theta}_n) + Bias^2(\hat{\Theta}_n)
\end{aligned}$$

Roughly speaking, **bias** measures how far off the target we hit on the average while **variance** measures how much fluctuation our estimator may show from one sample to another.

3.2 Unbiased Estimators

In almost all real applications, the class of possible estimators for an **ESTIMANAL** is huge and the best estimator, i.e. the one that minimizes MSE no matter what the value of the **ESTIMANAL** is, almost never exists. Thus we try to reduce the class of potential estimators by improving a plausible restriction, for example Bias($\hat{\Theta}_n$) = 0.

Definition. *An estimator $\hat{\Theta}_n$ of an **ESTIMANAL** Θ is said to be **unbiased** if $E(\hat{\Theta}_n) = \Theta$, for all possible values of Θ.*

Example 3.1. $X_i \overset{iid}{\sim} N(\mu, \sigma^2) \qquad i = 1, 2, \ldots, n$
Suppose both μ and σ^2 are unknown. Consider $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

$$E(\bar{X}_n) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} \overbrace{E(X_i)}^{\mu} = \frac{1}{\cancel{n}} \cdot \cancel{n}\mu = \mu$$

Thus $\bar{X}_n$ is an unbiased estimator of μ. As for the MSE($\bar{X}_n$), we need to find Var($\bar{X}_n$).

$$\begin{aligned}
Var(\bar{X}_n) &= Var\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n^2} Var(\sum_{i=1}^{n} X_i) \\
&= \frac{1}{n^2}\Big[\sum_{i=1}^{n} Var(X_i) + 2 \cdot \sum\sum_{1 \le i < j \le n} \overbrace{Cov(X_i, X_j)}^{0}\Big] && \text{Theorem 5.12(b) - page 271} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) && \coprod_{i=1}^{n} X_i \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^{\cancel{2}}} \cdot \cancel{n}\sigma^2 = \frac{\sigma^2}{n} && \text{identically distributed}
\end{aligned}$$

$$\Longrightarrow MSE(\bar{X}_n) = Var(\bar{X}_n) + \overbrace{Bias^2(\bar{X}_n)}^{0} = Var(\bar{X}_n) = \frac{\sigma^2}{n}$$

An inspection of the above calculation shows that for unbiased μ we only require a common mean μ while for calculating the variance we would only require a common variance σ^2 and orthogonality, i.e:

$$Cov(X_i, X_j) = 0 \qquad \text{where } i \neq j$$

Suppose $X_1, \ldots, X_n$ have the same mean value μ. Then:

$$E(\bar{X}_n) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{\cancel{n}}\cancel{n}\mu = \mu$$

Suppose further that $X_1, \ldots, X_n$ have the same variance σ^2 and $Cov(X_i, X_j) = 0,\ i \neq j$.

Then:

$$\begin{aligned}
Var(\bar{X}_n) &= Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2} Var(\sum_{i=1}^{n} X_i) \\
&= \frac{1}{n^2}\Big[\sum_{i=1}^{n} Var(X_i) + 2\sum\sum_{1\leq i<j\leq n} Cov(X_i, X_j)\Big] && \text{Theorem 5.12(b) - Page 271} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) && \text{Orthogonality: i.e. } Cov(X_i, X_j) = 0 \ \ if\ i \neq j \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^{\cancel{2}}}\cancel{n}\sigma^2 = \frac{\sigma^2}{n} && \text{having the same variance} \\
&\implies MSE(\bar{X}_n) = Var(\bar{X}_n) = \frac{\sigma^2}{n}
\end{aligned}$$

If $X_1, \ldots, X_n$ have the same mean value and variance and they are orthogonal.

3.3 Stein's Paradox

We will learn later that if $X_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$ then $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ has many optimal properties. A paradox due to Charles Stein, however, shows that such a nice optimal properties are not preserved in higher dimensions. In fact if:

$$X_i \stackrel{iid}{\sim} N(\mu_X, 1),\ \ Y_i \stackrel{iid}{\sim} N(\mu_Y, 1) \text{ and } Z_i \stackrel{iid}{\sim} (\mu_Z, 1)$$

then, we can find the biased estimators of $\begin{pmatrix} \mu_X \\ \mu_Y \\ \mu_Z \end{pmatrix}$ which are closer to $\begin{pmatrix} \mu_X \\ \mu_Y \\ \mu_Z \end{pmatrix}$ than $\begin{pmatrix} \bar{X}_n \\ \bar{Y}_n \\ \bar{Z}_n \end{pmatrix}$ for any $\begin{pmatrix} \mu_X \\ \mu_Y \\ \mu_Z \end{pmatrix}$. We may then say that $\begin{pmatrix} \bar{X}_n \\ \bar{Y}_n \\ \bar{Z}_n \end{pmatrix}$ is an **inadmissible estimator** of $\begin{pmatrix} \mu_X \\ \mu_Y \\ \mu_Z \end{pmatrix}$.

3.4 Admissibility

An estimator $\hat{\Theta}$ is called admissible if there is no estimator $\tilde{\Theta}$ such that:

$$MSE(\tilde{\Theta}) \leq MSE(\hat{\Theta}) \qquad \text{for all possible values of } \Theta$$

and this inequality is strict for some values of Θ.

What this example tells us is that by allowing a bit of bias we may be able to reduce variance considerably and hence find an estimator which is closer to the target than the most natural unbiased estimator. Note that this phenomena happens only when the dimension is at least 3.

4 Lecture 4

We now want to restrict the class of estimators even further. Suppose $X_1, \ldots, X_n$ have the same mean μ and variance σ^2 and they are orthogonal; i.e. $Cov(X_i, X_j) = 0\ ,\ i \neq j$. Consider $\tilde{X}_{n,\underset{\sim}{C}} = \sum_{i=1}^{n} C_i\, X_i$ and

$$\mathscr{C} = \left\{ \tilde{X}_{n,\underset{\sim}{C}} : \underset{\sim}{C} = (C_1, \ldots, C_n) \in \mathbf{R}^n,\ \sum_{i=1}^{n} C_i = 1 \right\}$$

Note that

$$\begin{aligned} E(\tilde{X}_{n,\underset{\sim}{C}}) &= E(\sum_{i=1}^{n} C_i\, X_i) = \sum_{i=1}^{n} C_i\, E(X_i) \\ &= \sum_{i=1}^{n} C_i\, \mu = \mu \overbrace{\sum_{i=1}^{n} C_i}^{1} = 1 \cdot \mu \\ &= \mu \end{aligned}$$

Thus $\tilde{X}_{n,\underset{\sim}{C}}$ is an unbiased estimator of μ fir any $\underset{\sim}{C} \in \mathbf{R}^n$ as long as $\sum_{i=1}^{n} C_i = 1$. Then $\mathscr{C}$ is the class of all unbiased linear estimators of μ. We want to find the best estimator with $\mathscr{C}$; i.e.:

$$\underset{\underset{\sim}{C} \in \mathbf{R}^n}{\text{Min}}\, MSE(\tilde{X}_{n,\underset{\sim}{C}}) \quad s.t\ \sum_{i=1}^{n} C_i = 1 \qquad (*)$$

First we note that $MSE(\tilde{X}_{n,\underset{\sim}{C}}) = Var(\tilde{X}_{n,\underset{\sim}{C}})$ since $\tilde{X}_{n,\underset{\sim}{C}}$ is an unbiased estimator of μ when $\sum_{i=1}^{n} C_i = 1$. On the other hand:

$$\begin{aligned}
Var(\tilde{X}_{n,\underset{\sim}{C}}) &= Var(\sum_{i=1}^{n} C_i\ X_i) \\
&= \sum_{i=1}^{n} C_i^{\,2}\ Var(X_i) + 2\sum\sum_{1\le i<j\le n} Cov(C_iX_i, C_j, X_j) && \text{Theorem 5.12 page 271} \\
&= \sum_{i=1}^{n} C_i^{\,2}\sigma^2 + \cancel{2\sum\sum_{1\le i<j\le n} C_iC_j\,\overbrace{Cov(X_i,X_j)}^{0}} \\
&= \sigma^2 \sum_{i=1}^{n} C_i^{\,2}
\end{aligned}$$

Thus (*) is equivalent to :

$$\underset{\underset{\sim}{C}\in\mathbf{R}^n}{\text{Min}}\ \sigma^2 \sum_{i=1}^{n} C_i^{\,2} \qquad\qquad (**)$$

Using the *Lagrange Theorem*, (**) is equivalent to:

$$\underset{\sim}{C} = (C_1, \overset{Min}{...}, C_n) \in \mathbb{R}^n\ \overbrace{\left\{\sigma^2 \sum_{i=1}^{n} C_i + \lambda(\sum_{i=1}^{n} C_i - 1)\right\}}^{\mathscr{C}_\lambda(\underset{\sim}{C})}.$$

Note that: $\qquad \frac{\partial\ \mathscr{C}_\lambda(\underset{\sim}{C})}{\partial C_i} = 2\ \sigma^2\ C_i + \lambda \quad , \ i = 1,2,3,\ldots$

$\frac{\partial}{\partial\lambda}\mathscr{C}_\lambda(\underset{\sim}{C}) = \sum_{i=1}^{n} C_i - 1$

$$\begin{cases}
\frac{\partial}{\partial C_i}\mathscr{C}_\lambda(\underset{\sim}{C}) = 2\ \sigma^2\ C_i + \lambda = 0\ \ ,\ \ i = 1,2,3,\ldots \\
\frac{\partial}{\partial\lambda}\mathscr{C}_\lambda = 0 \quad \Longrightarrow \quad \sum_{i=1}^{n} C_i = 1
\end{cases}$$

Thus $C_i = -\frac{\lambda}{2\ \sigma^2}\ \ ,\ \ i = 1,2,3,\ldots,n$ and using the last equation:

$$\sum_{i=1}^{n} -\frac{\lambda}{2\ \sigma^2} = 1 \implies \lambda = -\frac{2\ \sigma^2}{n}$$

and therefore:

$$C_i = -\frac{\lambda}{2\,\sigma^2} = -\frac{-\frac{2\,\sigma^2}{n}}{2\,\sigma^2} = \frac{1}{n} \qquad , \; i = 1, 2, 3, \ldots, n$$

We can further find:

$$\mathcal{H} = [\frac{\partial^2}{\partial C_i \partial C_j} \mathscr{C}_\lambda(\underset{\sim}{C})] \qquad , \; i, j = 1, 2, \ldots, n$$

and show that:

$$\begin{aligned} \underset{\sim}{x}^T \, \mathcal{H} \underset{\sim}{x} \; &\geq \; 0 \qquad \forall \underset{\sim}{x} \in \mathbb{R}^n \\ &= 0 \qquad \text{if and only if } \; \underset{\sim}{x} = 0 \end{aligned}$$

This then guarantees that $\underset{\sim}{C}^* = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$ is indeed a minimizer; in fact, the *unique minimizer*. To summarize:

$$\tilde{X}_{n,\underset{\sim}{C}^*} = \sum i = 1^n \frac{1}{n} X_i = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}_n$$

Thus $\bar{X}_n$ is the best unbiased linear estimator.

4.1 Estimating Variance

So far we confirmed ourselves to estimation of th population mean.

Now suppose we are interested in estimating variance from $X_1, \ldots, X_n$ where $X_i s$ have the same mean value μ , the same variance σ^2 and they are orthogonal, i.e. $Cov(X_i, X_j) = 0$, $i \neq j$, then a *natural estimator* of:

$$\sigma^2 = Var(X) = \mathbb{E}[(x - \mu)^2]$$

is its sample counterpart, i.e.

$$S^2_{n,*} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

Now the first question is if $S^2_{n,*}$ is an unbiased estimator of σ^2 , i.e. $\mathbb{E}(S^2_{n,*}) = \sigma^2$

$$\begin{aligned}
(X_i - \mu)^2 &= \Big[(X_i - \bar{X}_n) + (\bar{X}_n - \mu)\Big]^2 \\
&= (X_i - \bar{X}_n)^2 + (\bar{X}_n - \mu)^2 + 2 \cdot (X_i - \bar{X}_n)(\bar{X}_n - \mu) \\
\sum_{i=1}^{n}(X_i - \mu)^2 &= \sum_{i=1}^{n}(X_i - \bar{X}_n)^2 + n(\bar{X}_n - \mu)^2 + 2 \cdot (\bar{X}_n - \mu) \overbrace{\sum_{i=1}^{n}(X_i - \bar{X}_n)}^{0} \\
&= \sum_{i=1}^{n}(X_i - \bar{X}_n)^2 + n(\bar{X}_n - \mu)^2 \qquad (I)
\end{aligned}$$

Taking estimation we find:

$$\mathbb{E}\Big[\sum_{i=1}^{n}(X_i - \mu)^2\Big] = \mathbb{E}[n \cdot S^2_{n,*}] + \mathbb{E}[n(\bar{X}_n - \mu)^2] \qquad (II)$$

$$RHS = \sum_{i=1}^{n} \overbrace{\mathbb{E}(X_i - \mu)^2}^{\sigma^2} = n \cdot \sigma^2$$

Note that $\mathbb{E}(\bar{X}_n - \mu) = 0$, i.e. $\mathbb{E}(\bar{X}_n) = \mu$. Thus:

$$\mathbb{E}[n(\bar{X}_n - \mu)^2] = n\ \mathbb{E}[(\bar{X}_n - \mu)^2] = n\ Var(\bar{X}_n)\ .$$

On the other hand $Var(\bar{X}_n) = \frac{\sigma^2}{n}$. We therefore have:

$$\mathbb{E}[n(\bar{X}_n - \mu)^2] = n \cdot Var(\bar{X}_n) = n \cdot \frac{\sigma^2}{n} = \sigma^2$$

and hence from (II):

$$n\sigma^2 = \mathbb{E}(n\ S^2_{n,*}) + \sigma^2$$

which implies:

$$\implies \mathbb{E}(S^2_{n,*} = (\frac{n-1}{n})\sigma^2 = (1 - \frac{1}{n})\sigma^2$$

meaning that $S^2_{n,*}$ is **NOT** an unbiased estimator of σ^2.

Multiplying both sides of the last equation by the reciprocal of $(1 - \frac{1}{n})$ we find $\mathbb{E}(\frac{n}{n-1} S_{n,*}^2) = \sigma^2$. Note however that:

$$\frac{n}{n-1} S_{n,*}^2 = \frac{\cancel{n}}{n-1} \cdot \frac{1}{\cancel{n}} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

Thus $\boxed{S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2}$ is an unbiased estimator.

Question: Why $(n-1)$?

"$n-1$" is the dimension of $span \overbrace{\{X_i - \bar{X}_n : i = 1, 2, \ldots, n\}}^{\mathbb{V}}$.

$n-1 = dim(span\, V)$. Note however $dim(span\, W) = n$ where $W = X_i - \mu\ ,\ i = 1, 2, \ldots, n$.
We discuss these issues further in Chapter 11 where we learn about the regression.

So far we only considered sampling from one population. We may have samples from two or more populations and may want to make inference about differences between the populations.

Example 4.1.

Suppose we want to study the differences between the average salaries of men and women:

Men	Women
X_1	Y_1
$\vdots$	$\vdots$
X_m	Y_n

where $X_i s$ have the common mean μ_X and $Y_j s$ have the command mean μ_μ. We want to estimate $\mu_X - \mu_Y$. The natural estimator is $\bar{X}_m = \bar{Y}_n$. Show that:

$$\mathbb{E}[\bar{X}_m - \bar{Y}_n] = \mu_X - \mu_Y$$

Hence $\bar{X}_m - \bar{Y}_n$ is an unbiased estimator of $\mu_X - \mu_Y$.
Assume further that Xs and Ys are independent and Xs have common vari-

ance σ_X^2 and Ys have common variance σ_Y^2 and $Cov(X_i, X_j) = 0\ ,\quad i \neq j$ and $Cov(Y_i, Y_j) = 0\ ,\quad i \neq j$.

Find $Var(\bar{X}_m - \bar{Y}_n)$. Hint: use *Thm* 5.12.

The difference between two proportions can be treated similarly. Note that proportions are essentially means of binary variables.

5 Lecture 5 : Confidence Intervals

Definition. *Random Interval An interval whose endpoint(s) are random variables is called a* ***Random Interval****.*

5.1 Confidence Intervals

A $100(1-\alpha)\%$ confidence interval for a parameter θ is a *random interval* $\left(\hat{\Theta}_L(X), \hat{\Theta}_V(X)\right)$ such that:

$$P(\hat{\Theta}_L(X) < \Theta < \hat{\Theta}_V(X))$$

Pivotal Quantity : A function of the observation $X_1, \ldots, X_n$ and some unknown parameters, ideally just the parameter(s) of interest, whose distribution DOES NOT depend on any unknown parameter is called **Pivotal Quantity**.
Pivotal Quantities play a central role in theory confidence intervals.

Example. Let $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$ where σ^2 is <u>known</u>, but μ is <u>unknown</u>. We show that

$$\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n}) \qquad \textit{where} \qquad \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Recall that there are three methods for finding the distribution of a function of random variables:

1. **Method of Transformation:** This si essentially theorem of change of variables in calculus.

2. **Method of Distribution:** On this method we connect the *cdf* of the new variable to the *cdf* of the original variables.

   Example. *Suppose $X_i \overset{iid}{\sim} f$, $i = 1, 2, \ldots, n$ are continuous random variables with* pdf f *and* cdf F*. Define* $X_{(n)} = \max_{1 \le i \le n}$.

$$
\begin{aligned}
F_{x(n)}(t) = P(X_{(n)} \le t) = P(X_1 \le t, X_2 \le t, \dots, X_n \le t) \\
&= \prod_{i=1}^{n} P(X_i \le t) \qquad\qquad (\text{by } \coprod_{i=1}^{n} X_i) \\
&= \prod_{i=1}^{n} F_{X_i}(t) \\
&= \prod_{i=1}^{n} F(t) = F^n(t) \qquad\qquad \text{identically distributed}
\end{aligned}
$$

Thus

$$
\begin{aligned}
f_{X(n)}(t) &= \frac{d}{dt} F_{X(n)}(t) = \frac{d}{dt} F^n(t) \\
&= n\, f(t)\, F^{n-1}(t)
\end{aligned}
$$

3. **Method of Moment Generating Function(mgf):** This method is essentially based on the *mgf* of the new variable of the *mgf* if the original variables.

   Example. *Suppose $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2, \dots, n$ and X_is are independent. Define $S = \sigma_{i=1}^{n} X_i$.*

$$
\begin{aligned}
m_s(t) &= \mathbb{E}[e^{tS}] = \mathbb{E}[e^{t \sum_{i=1}^{n} X_i}] \\
&= \mathbb{E}\Big[\prod_{i=1}^{n} e^{tX_i}\Big] \qquad\qquad \text{using independence: } (\coprod) \\
&= \prod_{i=1}^{n} m_{X_i}(t) \\
&= \prod_{i=1}^{n} e^{\mu_i t + \frac{\sigma_i^2 t^2}{2}} \\
&= exp\Big\{t \sum_{i=1}^{n} \mu_i + \frac{t^2}{2} \sum_{i=1}^{n} \sigma_i^2\Big\} \\
\implies S &\sim N\Big(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\Big)
\end{aligned}
$$

If we further assume that $X_i s$ are identically distributed, then:

$$\mu_i = \mu \qquad \text{and} \qquad \sigma_i^2 = \sigma \qquad \forall i = 1, 2, \dots, n$$

Therefore we have:

$$m_S(t) = exp\Big\{n\mu t + \frac{n\sigma^2 t^2}{2}\Big\}$$

and hence: $\boxed{S \sim N(n\mu, n\sigma^2)}$ Then:

$$\begin{aligned}
& & m_{\bar{X}_n}(t) &= \mathbb{E}[e^{t\bar{X}_n}] = \mathbb{E}[e^{t\frac{1}{n}\sum_{i=1}^{n} X_i}] \\
\text{by } t^* = \frac{t}{n} & \implies & &= E[e^{t^* S}] \\
& & &= m_S(t^*) = e^{n\mu t^* + \frac{n\sigma^2 t^{*2}}{2}} \\
& & &= exp\Big\{n\mu t^* + \frac{n\sigma^2 t^{*2}}{2}\Big\}
\end{aligned}$$

$$\implies \boxed{\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})} \tag{1}$$

Note further that if $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$. We prove a general form of this. Let $X \sim N(\mu, \sigma^2)$; then:

$$aX + b \sim N(a\mu + b\ ,\ a^2\sigma^2) \qquad \text{for any constant } a, b$$

Let $V = ax + b$, then:

$$
\begin{aligned}
m_v(t) &= \mathbb{E}[e^{tV}] = \mathbb{E}[e^{t(ax+b)}] \\
&= \mathbb{E}[e^{taX+tb}] = \mathbb{E}[\underbrace{e^{tb}}_{constant} \cdot e^{\overbrace{taX}^{t^*}}] \\
&= e^{tb} \cdot \mathbb{E}[e^{t^*X}] \\
&= e^{tb} \cdot e^{\mu ta + \frac{\sigma^2 t^2 a^2}{2}} \\
&= exp\left\{t(a\mu + b) + \frac{t^2(a^2\sigma^2)}{2}\right\}
\end{aligned}
$$

Thus :

$$(ax + b) \sim N(a\mu + b \ , \ a^2\sigma^2)$$

Now :

$$
\begin{aligned}
Z &= \frac{X-\mu}{\sigma} = \frac{X}{\sigma} - \frac{\mu}{\sigma} \\
&= \frac{1}{\sigma}X - \frac{\mu}{\sigma} \\
&= aX + b \qquad\qquad \text{where } a = \frac{1}{\sigma} \text{ and } b = -\frac{\mu}{\sigma}
\end{aligned}
$$

Hence :

$$Z \sim N(\overbrace{\frac{1}{\sigma}\mu + (-\frac{\mu}{\sigma})}^{0} \ , \ \overbrace{(\frac{1}{\sigma})^2\sigma^2}^{1})$$

Thus :

$$\boxed{Z \sim N(0,1)} \qquad (2)$$

Using (1) and (2) :

$$\frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1).$$

This means that :

$$\boxed{\frac{\bar{X}_n - \mu}{\sqrt{h}} \text{ is a Pivotal Quantity}}$$

To summarize:

$$X_i \overset{iid}{\sim} N(\mu, \sigma^2) \qquad \Longrightarrow \qquad \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad \text{is a } \textbf{pivotal quantitity}$$

Notice that using the table for the normal distribution:

$$P\Big(\Big|\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}\Big| \le 1.96\Big) = 0.95$$

Equivalently:

$$P\Big(\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}} \ \le\ \mu\ \le \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\Big) = 0.95$$

This means that:

$$(\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}}\ ,\ \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}})$$

covers the true μ with 95% probability.

Thus a $100\,(1-\alpha)\%$ confidence interval for μ where $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ and σ^2 is known:

$$\bar{X}_n \pm \zeta_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

where:

$$P(Z > \zeta_{\frac{\alpha}{2}}) = \frac{\alpha}{2} \qquad , \qquad Z \sim N(0,1)$$

#MISSING GRAPH - (LECTURE 5 - P5)

Remark. In real applications we compute $\bar{X}_n$ and obtain an interval, say $(125, 135)$. Now either this interval covers the true μ or it does not. Then the question is what do we mean by a 95% #MISSING ?

Note that the $100(1-\alpha)\%$ confidence is the property of the procedure. It means that out of the all possible intervals of the form $(\bar{X}_n \pm 1.96\frac{\sigma}{\sqrt{n}})$ that we can make by taking samples of size n from $N(\mu, \sigma^2)$, 95% of them cover the true μ . Now this is a real application when we make one of the such intervals by taking a random sample of size n from $N(\mu, \sigma^2)$, it is like taking one of those intervals randomly. Since that 95% of them cover μ , my chance of selecting an interval that covers μ is 95% . Thus I can take a bet 19 to 1 that the interval I select covers μ.

5.2 Large Sample Confidence Interval

The derivation of the pivotal quantity in the above example totally hinges over the normality assumption, i.e. $X_i \sim N(\mu, \sigma^2)$.
What happens if we do not know the parametric for the population distribution?

Theorem (General Limit Theorem - GLT (baby version)**).** Suppose $X_1, \ldots, X_n$ are independent random variables with common μ and variance σ^2 . Then:

$$\frac{\bar{X}_n - \mu}{(\frac{\sigma}{\sqrt{n}})} \overset{app}{\sim} N(0,1) \qquad \text{when } n \text{ is large enough}$$

This is a powerful theorem that implies that $\frac{\bar{X}_n - \mu}{(\frac{\sigma}{\sqrt{n}})}$ is approximately a pivotal quantity distributed according to $N(0,1)$ for large enough n regardless of population distribution provided that the condition of the **GLT** are met.

σ^2

- σ^2 is known: $\implies$ $(\bar{X}_n \pm \zeta_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$ is a $100(1-\alpha)\%$ #MISSING for μ
- σ^2 is unknown: $\implies$ $(\bar{X}_n \pm \zeta_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$ is not <u>not useful</u>

Note: if σ^2 is unknown, $(\bar{X}_n \pm \zeta_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$ is still a $100(1-\alpha)\%$ #MISSING for μ but not useful.
We need to somehow get rid of the #MISSING parameter σ . We can replace σ by S_n where:

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

Justification

- Intuitive
- Formal

Intuitive : S_n^2 is the sample counterpart, almost, for σ^2 . Thus as n increases, greater portion of the population and hence our sample sets closer to the population.

Formal : The formal proof comprises three steps:

1. **GLT** of
$$\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

2. Consistency of S_n^2 for σ^2, i.e. $S_n^2 \xrightarrow{P} \sigma^2$ which we learn in *Chapter 9*. We then use a theorem called **Continuous Mapping Theorem** which says that if $S_n^2 \xrightarrow{P} \sigma^2$, then:
$$g(S_n^2) \xrightarrow{P} g(\sigma^2) \qquad\qquad \text{for any continuous function}$$
Considering $g(x) = \sqrt{X}$, we obtain $S_n \xrightarrow{P} \sigma$ and hence $\frac{\sigma}{S} \xrightarrow{P} 1$.

3. **Cramer's Theorem :**

   This result says that if $V_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} 1$, then $Y_n \cdot V_n \xrightarrow{D} X$:
$$\frac{\bar{X}_n - \mu}{\frac{S}{\sqrt{n}}} = \underbrace{\frac{\sigma}{S_n}}_{Y_n} \cdot \underbrace{\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}}_{V_n}$$

Note that GLT implies that $V_n \xrightarrow{D} 2$, i.e:
$$\overbrace{F_{V_n}(t)}^{\text{cdf of } V_n} \rightarrow F_Z(t) = \overbrace{\int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx}^{\Phi(t) \text{ cdf of } N(0,1)}$$

Using step (2), $Y_n = \frac{\sigma}{S_n} \xrightarrow{P} 1$ and application of Cramer's Theorem computes the proof. To summarize:

σ^2 — σ^2 is known $\implies$ $(\bar{X}_n \pm \zeta_{\frac{\alpha}{2}})$ is a $100(1-\alpha)\%$

σ^2 — σ^2 is unknown $\implies$ $(\bar{X}_n \pm \zeta_{\frac{\alpha}{2}})$ is a $100(1-\alpha)\%$

To be more precise, these confidence intervals are approximate $100(1-\alpha)\%$ confidence intervals for μ when n is large enough.

So far we focused on *C.I* for population mean. How can we make *C.I* for other estimates?

A common, perhaps the most common, method of estimation that we will learn about in Chapter 9 is the method of maximum likelihood. Suppose Θ is a parameter of interest. Suppose $\hat{\Theta}_n = \hat{\Theta}(X_1, \ldots, X_n)$ is the maximum likelihood estimate (*MLE*) of Θ based on $X_1, \ldots, X_n$. Then relatively several condition we have:

$$\frac{\hat{\Theta}_n - \Theta}{\sqrt{Var(\hat{\Theta}_n)}} \overset{app}{\sim} N(0,1) \text{ when } n \text{ is large enough} \qquad (*)$$

We therefore have a several recipe for confidence interval when the sample size n is large enough, namely:

$$\hat{\Theta}_n \pm \zeta_{\frac{\alpha}{2}} \sqrt{Var(\hat{\Theta}_n)} \qquad (\dagger)$$

that is a $100(1-\alpha)\%$ *C.I* for Q.

Example 5.1. $X_i \overset{iid}{\sim} N(\mu, \overbrace{\sigma^2}^{known})$, $i = 1, 2 \ldots, n$.

We show in chapter 9 that $\bar{X}_n$ in the *MLE* of μ. Note that $Var(\bar{X}_n) = \frac{\sigma^2}{n}$. Then using $(\dagger)$:

$$\bar{X}_n \pm \zeta_{\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \quad \text{is a } 100(1-\alpha)\% \text{ C.I for } \mu$$

Example 5.2. $X_i \overset{iid}{\sim} Bernoulli(p)$ $\forall\, i = 1, 2, \ldots, n$, *i.e:*

$$X_i = \begin{cases} 1 & p \\ 0 & 1-p \end{cases}$$

Then $\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the *MLE* of p. Thus using (†) :

$$\hat{p}_n \pm \zeta_{\frac{\alpha}{2}} \sqrt{Var(\hat{p}_n)} \quad \text{is a } 100(1-\alpha)\% \ C.I \text{ for } p \ .$$

Note that $Var(\hat{p}_n) = \frac{p(1-p)}{n}$. We have two choices:

1. replace p by $\hat{p}_n$ in $Var(\hat{p}_n)$:

$$\hat{p}_n \zeta_{\frac{\alpha}{2}} \frac{\sqrt{\hat{p}_n(1-\hat{p}_n)}}{\sqrt{n}}$$

2. replace $p(1-p)$ in $Var(\hat{p}_n)$ by $\frac{1}{4}$ to find a conservatively large *C.I* for p :

$$\hat{p}_n \pm \zeta_{\frac{\alpha}{2}} \frac{1}{2\sqrt{n}}$$

Example 5.3. *Suppose* $X_i \overset{iid}{\sim} Ber(p)$ *,* $i = 1, 2 \ldots, n$ *and we are interested in* $\Theta = p(1-p)$ *, the variance.*

An interesting property of *MLE* is the invariance , i.e. if $\hat{\Theta}_n$ if the *MLE* of Θ , then $h(\hat{\Theta}_n)$ is the *MLE* of $h(\Theta)$. The invariance property then implies that: $\hat{\Theta}_n = \hat{p}_n(1-\hat{p}_n)$ is the *MLE* of $p(1-p) = \Theta$.
The $100(1-\alpha)\%$ *C.I* for $\Theta = p(1-p)$ is $\hat{\Theta}_n \pm \zeta_{\frac{\alpha}{2}} \sqrt{Var(\hat{\Theta}_n)}$

5.3 Small Sample Confidence Intervals

Unlike the large sample case, there is no general recipe like (∗) using which we can find an approximate pivotal quantity. In fact, there is on the paper, but only gives #MISSING in special cases.
To summarize , small sample probabilities are solved mostly case by case. A case of particular importance is the *normal case* . We will learn about the importance of this case when we discuss regression and ANOVA (Analysis of Variance) .

Normal Case:

Suppose $X_i \overset{N}{\sim} (\overbrace{\mu}^{\text{of interest}}, \underbrace{\sigma^2}_{\textit{nuisance}})$, $i = 1, 2, \ldots, n$ where n, the sample size is *NOT* large.

We learned that when $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$:

$$\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \overset{Exact}{\sim} N(0, 1) \qquad (\ddagger)$$

This by itself is not useful since σ is not known. We discussed in previous section at length why we can replace σ by S when n is large enough. The formal justification is **not** applicable now since n is small, the intuitive justification still stands though.

Replacing σ with (‡) changes the picture a bit. Given that S has the same spirit as σ , though in a small #MISSING the distribution of $T = \frac{\bar{X}_n - \mu}{\frac{S}{\sqrt{n}}}$ still has a bell curve shape. The tails of the distribution, however, die out much more slowly than those of normal distribution. Heavier tails mean much more variability and this should perhaps be expected since by replacing σ by S which can be crude estimate when n is small, can add suite a hit to the variability. This is, of course, a intuitive argument. Following we present the sketch of a formal argument:

step 1) $X_i \overset{iid}{\sim} N(\mu, \sigma^2) \implies \bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$

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

step 2) $X_i \overset{iid}{\sim} N(\mu, \sigma^2) \implies \frac{(n-1)S^2}{\sigma^2} \sim \mathcal{X}^2_{(n-1)}$

proof

$$\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}\left[(X_i-\bar{X}_n)+(\bar{X}_n-\mu)\right]^2$$

$$=\sum_{i=1}^{n}(X_i-\bar{X}_n)^2+n(\bar{X}_n-\mu)^2+2(\bar{X}_n-\mu)\overbrace{\sum_{i=1}^{n}(X_i-\bar{X}_n)}^{0}$$

$$=(n-1)S^2+n(\bar{X}_n-\mu)^2$$

by dividing both sides by σ^2 we obtain:

$$\underbrace{\sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma})^2}_{W}=\underbrace{\frac{(n-1)S^2}{\sigma^2}}_{U}+\underbrace{(\frac{\bar{X}_n-\mu}{\frac{\sigma}{\sqrt{n}}})^2}_{V}$$

Nownotethat :

$$X_i\overset{iid}{\sim}N(\mu,\sigma^2)\implies\frac{X_i-\mu}{\sigma}\sim N(0,1)$$
$$\implies(\frac{X_i-\mu}{\sigma})^2\sim\mathcal{X}_1^2$$
$$\implies\sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma})^2\sim\mathcal{X}_n^2$$

(Exercise: Theorem 7.2 , page 356)

Thus $W\sim\mathcal{X}_n^2$. On the other hand, using step 1:

$$(\frac{\bar{X}_n-\mu}{\frac{\sigma}{\sqrt{n}}})^2\sim\mathcal{X}_1^2$$

step 3) If $X_i\overset{iid}{\sim}N(\mu,\sigma^2)$ then $\bar{X}_n\coprod S^2$

step 4)

$$
\begin{aligned}
m_W(t) &= \mathbb{E}e^{tW} \\
&= \mathbb{E}[e^{t(U+V)}] \\
&= \mathbb{E}[e^{tU} \cdot e^{tV}] \\
&= \mathbb{E}[e^{tU}] \cdot \mathbb{E}[e^{tV}] \\
&= m_U(t) + m_V(t) \qquad\qquad U \coprod V \text{ using step 3} \\
\text{Thus} \qquad m_U(t) &= \frac{m_W(t)}{m_V(t)} \\
&= \frac{(1-2t)^{-\frac{n}{2}}}{(1-2t)^{-\frac{1}{2}}} \\
&= (1-2t)^{-\frac{n-1}{2}}
\end{aligned}
$$

which implies that $U \sim \mathcal{X}^2_{(n-1)}$

step 5) If $Z \sim N(0,1)$, $U \sim \mathcal{X}^2_V$ and $Z \coprod U$ then:

$$\frac{Z}{\sqrt{\frac{U}{V}}} \sim T_{n-1} \qquad\qquad \text{(Exercise 7.30 , page 367)}$$

step 6)

$$T_{n-1} = \frac{\bar{X}_n - \mu}{\frac{S}{\sqrt{n}}} = \frac{(\frac{\bar{X}_n-\mu}{\frac{\sigma}{\sqrt{n}}})}{\sqrt{\frac{(\frac{(n-1)S^2}{\sigma^2})}{(n-1)}}} = \frac{Z}{\sqrt{\frac{U}{V}}}$$

The pdf of T_v is :

$$f_{T_v}(t) = \frac{\Gamma\frac{v+1}{2}}{\Gamma(\frac{v}{2})\sqrt{v\pi}}(1+\frac{t^2}{v})^{-\frac{v+1}{2}} \qquad -\infty < t < +\infty$$

#MISSING GRAPH Lecture 5 - page 13

$$\mathbb{E}[T_v^r] = \begin{cases} 0 & \text{if } r < v \text{ and } r \text{ is odd} \\ v^{\frac{r}{2}} \cdot \frac{\Gamma(\frac{1+r}{2})\Gamma(\frac{v-r}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{v}{2})} & \text{if } r < v \text{ and } r \text{ is even} \end{cases}$$

Thus, if $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$ and μ and σ^2 are both <u>unknown</u> :

$$\bar{X}_n \pm t_{(n-1),\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

provides a $100(1-\alpha)\%$ $C.I$ for μ where $P(T_{(n-1)} > t_{(n-1),\frac{\alpha}{2}}) = \frac{\alpha}{2}$

5.4 Pivotal Quantity and Probability Integral Transform

Suppose X is a continuous random variable with $p.d.f$ f and cdf F. Then $F(X) \sim Uniform(0,1)$ (Exercise)

This result is referred to as the **Probability Integral Transform**. Now suppose $X_i \overset{iid}{\sim} F$. Then :

$$\begin{aligned} F(X_i) \sim Unif(0,1) &\implies -2\ln F(X_i) \sim \mathcal{X}^2_2 \\ &\implies -2\sum_{i=1}^{n} \ln F(X_i) \sim \mathcal{X}^2_{2n} \\ &\implies -2\sum_{i=1}^{n} \ln\left[1 - F(X_i)\right] \sim \mathcal{X}^2_{2n} \end{aligned}$$

There is hence a general recipe for finding a pivotal quantity when we have samples from continuous random variables. The usefulness of this pivotal quantity *depends* on the form of F , the cdf of X .

Suppose $X_i \overset{iid}{\sim} exp(\lambda)$, $i = 1, 2, \ldots, n$, i.e:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & o/w \end{cases}$$

Then:

$$F(x) = \int_0^x f(t)dt = 1 - e^{-\lambda x} \quad , \quad x > 0$$

and:

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & x > 0 \\ 0 & o/w \end{cases}$$

Using the above discussion:

$$-2\sum_{i=1}^{n}\ln F(X_i) \sim \mathcal{X}^2_{2n} \qquad \text{and} \qquad -2\sum_{i=1}^{n}\ln\left[1-F(X_i)\right] \sim \mathcal{X}^2_{2n}$$

for this example it is easier to work with the latter, i.e. :

$$\begin{aligned}2\sum_{i=1}^{n}\ln\left[1-F(X_i)\right] &= -2\sum_{i=1}^{n}\ln\left(e^{-\lambda X_i}\right)\\ &= 2\lambda\sum_{i=1}^{n}X_i = 2n\lambda\bar{X}_n\\ so \implies 2n\lambda\bar{X}_n &\sim \mathcal{X}^2_{2n}\end{aligned}$$

Using the $\mathcal{X}^2$ table (Application 3, page 850-851) , we can find $\mathcal{X}^2_{(2n),0.025}$ and $\mathcal{X}^2_{(2n),0.975}$ such that:

$$P(\mathcal{X}^2_{(2n),0.975} < 2n\lambda\bar{X}_n < \mathcal{X}^2_{(2n),0.025}) = 0.95$$

Thus:

$$\left(\frac{\mathcal{X}^2_{(2n),0.975}}{2n\bar{X}_n}, \frac{\mathcal{X}^2_{(2n),0.025}}{2n\bar{X}_n}\right)$$

provides that a 95% $C.I$ for λ . Note that $\mathcal{X}^2_{(2n),\alpha}$ is such that $P(\mathcal{X}^2_{(2n)} > \mathcal{X}^2_{(2n),\alpha}) = \alpha$

#MISSING GRAPH LECTURE 5 - PAGE 15

6 Lecture 6

6.1 Small Sample Confidence Interval(general case):

We learned in the last lecture how to find *C.I.* fir th population mean when the population distribution is normal. The two main pivotal quantities are:

$$(a) \quad \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{(n-1)} \qquad \& \qquad (b) \quad \frac{\bar{X}_n = \mu}{\frac{S}{\sqrt{n}}} \sim T_{(n-1)}$$

when $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$.
The first result can be used to make a *C.I* for σ^2 and σ which the latter is used for making a *C.I* for μ .
We now consider the general case.

6.2 Probability Integral Transform(PIT)

Suppose $X_i \overset{iid}{\sim} F_x$ and f is the pdf of X_is:

$$X \sim F_X \text{ , } Y = F_X(X)$$

$$\begin{aligned} F_Y(t) &= P(Y \leq t) = P(F_X(x) \leq t) \\ &= P(X \leq F_X^{-1}(t)) \\ &= F_X(F_X^{-1}(t)) = t \qquad \text{for } 0 \leq 1 \leq 1 \end{aligned}$$

$$\text{Thus} \quad F_Y(t) = \begin{cases} 0 & \text{if } t < 0 \\ t & \text{if } 0 \leq t < 1 \\ 1 & \text{if } 1 \leq t \end{cases}$$

and hence $Y \sim Unif(0.1)$. This is called **Probability Integral Transform(PIT)**.

Example 6.1. $X \sim Exp(\lambda)$, $f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$, $\lambda > 0$.

$$
\begin{aligned}
F_X(x) = P(X \le x) &= \int_{-\infty}^{x} f_X(t)dt = \int_0^x \lambda e^{-\lambda t}dt \\
&= -e^{-\lambda t}\Big|_0^x \\
&= 1 - e^{-\lambda x} \qquad\qquad\qquad\qquad (1)
\end{aligned}
$$

Now consider $Y = F_X(x) = 1 - e^{-\lambda x}$:

$$
\begin{aligned}
F_Y(t) &= P(Y \le t) = P(1 - e^{-\lambda x} \le t) \\
&= P(e^{-\lambda x} \ge 1 - t) = P(X \le -\frac{ln(1-t)}{\lambda}) \\
&= F_X(-\frac{ln(1-t)}{\lambda}) \\
&= 1 - e^{-\lambda(-\frac{ln(1-t)}{\lambda})} \qquad\qquad \text{using (1)} \\
&= 1 - e^{ln(1-t)} = 1 - (1-t) \\
&= t
\end{aligned}
$$

Thus $Y \sim Unif(0,1)$.

Remark. Using *PIT* we can essentially generate random numbers from any continuous distributions. In fact, suppose we want samples from cdf F. Then:

Step 1: Generate $U_i \overset{iid}{\sim} \text{Unif}(0,1)$, $i = 1, 2, \ldots, n$.

Step 2: $X_i = F^{-1}(U_i) \overset{iid}{\sim} F$, $i = 1, 2, \ldots, n$

This algorithm then works as long as we can generate uniform random numbers and F^{-1} can be explicitly found or well approximated.

Example. $f_X(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, $-\infty < x < +\infty$ $(X \sim N(0,1))$

$$
\text{Then:} \quad F_X(x) = \int_{-\infty}^{x} f_X(t)dt = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt \quad .
$$

In this case, F^{-1} does not have an explicit nice form, but it can be well approximated.

Remark. A simple and useful transformation:

$$X \sim F \implies \overbrace{Y = F(X) \sim Unif(0,1)}^{P.I.T} \implies -2 \cdot log(Y) \sim \chi^2_2$$

6.3 Pivotal Quantity

Suppose $X_i \overset{iid}{\sim} F$, $i = 1, 2, \ldots, n$. Define:

$$Y_i = F(X_i) \overset{iid}{\sim} Unif(0,1) \ , \ i = 1, 2, \ldots, n$$

Now consider:

$$V_i = -2 \cdot log(Y_i) \overset{iid}{\sim} \chi^2_{2n} \ , \ i = 1, 2, \ldots, n$$

Then:

$$\sum_{i=1}^{n} V_i \sim \chi^2_{2n} \qquad .$$

Having established the first two results, i.e. :

Step 1: $X_i \sim F \implies Y_i = F(X_i) \sim Unif(0,1)$ (PIT)

Step 2: $V_i = -2 \cdot log(Y_i) \sim chi^2_2$ (method of transformation)

The last result can be established using the method of moments:

$$\begin{aligned}
& m_{\sum_{i=1}^n V_i}(t) = \mathbb{E}[e^{-t\sum_{i=1}^n V_i}] = \mathbb{E}\Big[\prod_{i=1}^n e^{-tV_i}\Big] \\
\coprod_{i=1}^{n} V_i \implies \quad & = \prod_{i=1}^{n} \mathbb{E}[e^{-tV_i}] = \prod_{i=1}^{n} m_{V_i}(t) \\
\text{identically distributed} \implies \quad & = [m_Y(t)]^n = [(1-2t)^{-\frac{2}{2}}]^n \\
& = (1-2t)^{-\frac{2n}{2}} \implies \sum_{i=1}^{n} V_i \sim \chi^2_{2n}
\end{aligned}$$

Then a pivotal quantity based on $X_i \overset{iid}{\sim} F_\theta$, $i = 1, 2, \ldots, n$ is:

$$-2\sum_{i=1}^{n} log(F_\theta(X_i)) \sim chi^2_{2n} \qquad (1)$$

Example. $X_i \overset{iid}{\sim} Exp(\lambda)$, $f_x(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$

$$F_x(x) = \int_{-\infty}^{x} f_x(t)dt = \int_0^x \lambda e^{-\lambda t}dt = 1 - e^{-\lambda x}$$

Now we notice that $-2 \cdot log(F) = -2 \cdot log(1 - e^{-\lambda x})$ does not provide an useful form for the purpose of making a $C.I$ for λ. There is a dual to (1) that is useful in this case, however:

$$\sum_{i=1}^{n} W_i = \sum_{i=1}^{n} -2log(1 - F(X_i)) \sim \chi^2_{2n} \qquad (2)$$

This quickly follows from the fact that:

$$U \sim Unif(0,1) \implies 1 - U \sim Unif(0,1) \qquad .$$

Using (2) we have:

$$\sum_{i=1}^{n} -2log(1 - F(X_i)) = \sum_{i=1}^{n} -2log(e^{-\lambda X_i})$$
$$= 2\lambda \sum_{i=1}^{n} X_i = 2\lambda n\bar{X}_n \sim \chi^2_{2n}$$

Using the χ^2-table (App.3 m page 850-851) , we find $\chi^2_{2n,0.025}$ and $\chi^2_{2n,0.975}$ such that:

$$P(\chi^2_{2n,0.975} < 2\lambda n\bar{X}_n < \chi^2_{2n,0.025}) = 0.95$$

and hence:

$$P(\frac{\chi^2_{2n,0.975}}{2n\bar{X}_n} < \lambda < \frac{\chi^2_{2n,0.025}}{2n\bar{X}_n}) = 0.95$$

Thus:

$(\frac{\chi^2_{2n,0.975}}{2n\bar{X}_n}, \frac{\chi^2_{2n,0.025}}{2n\bar{X}_n})$ is a <u>95%</u> confidence interval for λ

#MISSING Graph Lecture 6 - page 36

6.4 Small Size Determination

Suppose we want to estimate the proportion of Canadian voters who are in favor of NDP and want our estimate to be one-percentage point from the actual population with 95% confidence. Define:

$$X = \begin{cases} 1 & \text{NDP} \\ 0 & \text{other parties} \end{cases} \qquad \text{associated to each potential voter.}$$

We learned that to estimate the proportion of interest $p = P(X = 1)$, we can use $\hat{P}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ from a random sample of size n. We further learned that if the sample size n is large enough, then:

$$\hat{p}_n \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

is a 95% confidence interval for p. Thus the margin of error is $\beta = 1.96\sqrt{\frac{p(1-p)}{n}}$ which is controlled by n , the sample size. We should therefore choose n such that:

$$0.01 = 1.96\sqrt{\frac{p(1-p)}{n}}$$

Given that p is unknown, we can either replace p by $\hat{p}_n$ or take a conservative approach and replace p by $\frac{1}{2}$ which maximizes $p(1-p)$. Thus we find:

$$n = \frac{p(1-p)\zeta^2_{\frac{\alpha}{2}}}{\beta^2} = \begin{cases} \frac{\hat{p}_n(1-\hat{p}_n)\zeta^2_{\frac{\alpha}{2}}}{\beta^2} & \text{replacing } p \text{ by } \hat{p}_n \\ \frac{\zeta^2_{\frac{\alpha}{2}}}{4\beta^2} & \text{replacing } p \text{ by } \frac{1}{2} \end{cases}$$

Taking the conservative approach, we have:

$$n = \frac{\zeta^2_{\frac{\alpha}{2}}}{4\beta^2} = \frac{(1.96)^2}{4(0.01)^2} = 9604$$

Likewise we can find the sample size formula for estimating the population mean with a given confidence $1-\alpha$ and margin of error β , we should in fact solve the following equation for n:

$$\beta = \zeta^2_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \qquad \text{where } \sigma^2 \text{ is the population variance.}$$

We then find $\boxed{n = \frac{\zeta^2_{\frac{\alpha}{2}} \sigma^2}{\beta^2}}$ where σ^2 should be estimated from a prior sample.

6.5 Sample Size Determination For Other Parameters

So far we only considered the population mean. Now consider a parameter θ . In chapter 9 we learn about different methods of estimation, among them there is a method called the method of maximum likelihood (ML). Suppose $\hat{\theta}_n$ is the maximum likelihood estimate (MLE) of θ. Then under some reasonable conditions for a considerably large class of parametric distributions, we have:

$$\frac{\hat{\theta}_n - \theta}{\sqrt{Var(\hat{\theta}_n)}} \overset{app}{\sim} N(0,1) \qquad , \text{ for large n}$$

Thus:

$$\hat{\theta} \pm \underbrace{\zeta_{\frac{\alpha}{2}} \sqrt{Var(\hat{\theta}_n)}}_{\beta} \qquad \text{is a } 100(1-\alpha)\% \; C.I \text{ for } \theta$$

Let $\beta = \zeta_{\frac{\alpha}{2}} \sqrt{Var(\hat{\theta}_n)}$. In many interesting cases, $Var(\hat{\theta}_n)$ is an explicit function of n and σ^2 , the variance in the target population , say $h(\sigma^2, n)$. Then the sample size can be determined by the solution of the following equation:

$$h(\sigma^2, n) = \frac{\beta^2}{\zeta^2_{\frac{\alpha}{2}}}$$

Recall that for Bernoulli case, i.e. $X_i = \begin{cases} 1 \\ 0 \end{cases} \quad , \qquad i = 1, 2, \ldots, n$:

$$h(\sigma^2, n) = Var(\hat{\theta}_n) = Var(\hat{p}_n) = \frac{\overbrace{p(1-p)}^{\sigma^2}}{n}$$

while for estimating the population mean:

$$h(\sigma^2, n) = Var(\hat{\theta}_n) = Var(\bar{X}_n) = \frac{\sigma^2}{n} \quad .$$

6.5.1 Sample Size Determination (Small Sample)

- Normal Case: We learned that if $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$:

  $$\frac{\bar{X}_n - \mu}{\frac{S}{\sqrt{n}}} \sim T_{n-1} \implies \bar{X}_n \pm t_{\frac{\alpha}{2},(n-1)} \frac{S}{\sqrt{n}} \qquad \text{is a } 100(1-\alpha)\% \ C.I \text{ for } \mu$$

  We can therefore find sample size from the following equation:

  $$\beta = t_{\frac{\alpha}{2},(n-1)} \frac{S}{\sqrt{n}} \implies \boxed{n = \frac{S^2 t^2_{\frac{\alpha}{2},(n-1)}}{\beta^2}}$$

  Now note that the sample size determination based on large sample and small sample in the normal case had to $N(0, 1)$ and $T_{(n-1)}$ respectively. These distributions are both symmetric. As such in the sample size determination we only deal with the half length of the confidence intervals when the sample size is small and the population from which the samples are taken is not normal, the pivotal quantities do not necessarily have asymmetric distribution and hence the confidence interval do not have the form of $\boxed{\hat{\theta}_n \pm \beta}$.

  In such cases we try to control the total length of the confidence interval.

  Example. $X_i \overset{iid}{\sim} Exp(\lambda)$, $i = 1, 2, \ldots, n$

We found that $(\frac{\chi^2_{2n,0.975}}{2n\bar{X}_n}, \frac{\chi^2_{2n,0.025}}{2n\bar{X}_n})$; let C be the desired length for $C.I$ for λ . Then:

$$C = \frac{\chi^2_{2n,\frac{\alpha}{x}} - \chi^2_{2n,1-\frac{\alpha}{2}}}{2n\bar{X}_n}$$

represents the length of a $C.I$ for λ based on a sample of size n with $100(1-\alpha)\%$ confidence.

6.5.2 Sample Size Determination(Two Sample Case)

So far we just confirmed ourselves to one population. We might, however, have two samples, $X_1, X_2, \ldots, X_m$ from the population of men with population mean $\mu_{_M}$, and $Y_1, Y_2, \ldots, Y_n$ from the population of women with population mean $\mu_{_W}$. Suppose the parameter of interest is $\theta = \mu_{_M} - \mu_{_W}$. Then the natural estimate of θ is $\hat{\theta} = \bar{X}_m - \bar{Y}_n$ and using the central limit theorem:

$$\frac{(\bar{X}_m - \bar{Y}_n) - (\mu_{_M} - \mu_{_W})}{\sqrt{Var(\bar{X}_m - \bar{Y}_n)}} \overset{app}{\sim} N(0,1) \qquad \text{for large } m \ \& \ n$$

Now:

$$Var(\bar{X}_m - \bar{Y}_n) = Var(\bar{X}_m) + Var(\bar{Y}_n) - 2Cov(\bar{X}_m, \bar{Y}_n)$$

Assuming that Xs and Ys are independent (i.e $Cov(\bar{X}_m, \bar{Y}_n) = 0$) :

$$Var(\bar{X}_m - \bar{Y}_n) = Var(\bar{X}_m) + Var(\bar{Y}_n) = \frac{\sigma^2_{_M}}{m} + \frac{\sigma^2_{_W}}{n}$$

Therefore:

$$(\bar{X}_m - \bar{Y}_n) \pm \zeta_{\frac{\alpha}{2}} \sqrt{\frac{\sigma^2_{_M}}{m} + \frac{\sigma^2_{_W}}{n}} \qquad \text{is a } 100(1-\alpha)\% \ C.I \text{ for } \mu_{_M} - \mu_{_W} .$$

To find the sample size we should solve:

$$\beta = \zeta_{\frac{\alpha}{2}} \sqrt{\frac{\sigma^2_{_M}}{m} + \frac{\sigma^2_{_W}}{n}}$$

We should assume that σ^2_M & σ^2_W are known or estimated from prior samples, we will have one equation with two unknowns, m & n. In order to have a unique solution we need another equation. We often consider $n = K \cdot m$, where K is a known value as the second equation. Suppose C_M & C_W represent respectively, the cost of taking a sample from population of men and women. Then $K \propto (\frac{C_W}{C_M})^{-1}$. In case that $C_W = C_M$, we choose $K = 1$. Now:

$$\begin{cases} \beta = \zeta_{\frac{\alpha}{2}} \sqrt{\frac{\sigma^2_M}{m} + \frac{\sigma^2_W}{n}} \\ n = Km \end{cases}$$

Solving the above system we find:

$$\boxed{m = (\frac{\zeta_{\frac{\alpha}{2}}}{\beta})^2 \cdot (\sigma^2_M + \frac{\sigma^2_W}{K})}$$

For proportions: $\sigma^2_M = p_M(1 - p_M)$ & $\sigma^2_W = p_W(1 - p_W)$ Taking the conservative approach and replacing both p_M & p_W by $\frac{1}{2}$ we find:

$$\boxed{m = \Big(\frac{\zeta_{\frac{\alpha}{2}}}{2\beta}\Big)^2 (1 + \frac{1}{K})}$$

7 Lecture 7

7.1 Chapter 9 - Relative Efficiency

Definition. The relative efficiency of two unbiased estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, is defined to be:

$$eff(\hat{\theta}_1, \hat{\theta}_2) = \frac{Var(\hat{\theta}_1)}{Var(\hat{\theta}_2)}$$

We learned that between the two unbiased estimators the one with smaller variance is closer to the target on the average, i.e. has smaller *MSE*.
We also learned that the length of confidence intervals for large sample size is controlled by the variance of the estimator; so, the smaller the variance, the shorter the confidence interval using that estimator is.
We now want to quantify the gain in using the estimator with smaller variance.

Example. *Suppose* $X_i \overset{iid}{\sim} f$, $i = 1, 2, \ldots, n$ *where* f *is a symmetric* pdf.

The mean and median of f are the same, say μ . Given that f is symmetric, we can use: $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ (the sample average) , or:

$$M_n = \begin{cases} X_{\frac{n+1}{2}} & \text{if n is odd} \\ \frac{1}{2}[X_{\frac{n}{2}} + X_{\frac{n}{2}+1}] & \text{if n is even} \end{cases} \qquad \text{where } X_{(1)} < X_{(2)} < \cdots < X_{(n)} \text{ are the order statistics.}$$

We learned that $Var(\hat{X}_n) = \frac{\sigma^2}{n}$ where σ^2 is the population variance, i.e. :

$$\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx$$

It can be shown (beyond the scope of this course), that:

$$Var(M_n) \approx \frac{1}{4 \cdot [f(\mu)]^2 n} \qquad \text{for large } n$$

For instance if f is Normal, i.e. $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, then $f(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$ and hence:

$$Var(M_n) = \frac{2\pi}{4} \cdot \frac{\sigma^2}{n}$$

Thus:

$$eff(\bar{X}_n, M_n) = \frac{Var(M_n)}{Var(\bar{X}_n)} = \frac{\frac{2\pi}{4} \cdot \frac{\sigma^2}{n}}{\frac{\sigma^2}{n}} = \frac{2\pi}{4} = 1.57$$

This then essentially means that if you can make a confidence interval of a given length using $\bar{X}_n$ with 100 observations, to make a confidence interval of the same length for μ using M_n , you need $100 \times 1.57 = 157$ observations.

Example (9.1 , page 446). $Y_i \overset{iid}{\sim} Unif(0, \theta) \qquad , i = 1, 2, \ldots, n \ , \theta > 0$ *and* θ *is unknown.*

Consider $\hat{\theta}_1 = 2\bar{Y}_n$ and $\hat{\theta}_2 = (\frac{n+1}{n})Y(n)$ where $Y(n) = \max_{1 \le i \le n} Y_i$

For $\hat{\theta}_1$:

$$\begin{aligned}
\mathbb{E}(\hat{\theta}_1) &= \mathbb{E}(2\bar{Y}_n) = 2\mathbb{E}(\bar{Y}_n) \\
&= 2\mathbb{E}(\tfrac{1}{n}\textstyle\sum_{i=1}^n Y_i) = \tfrac{2}{n}\textstyle\sum_{i=1}^n \mathbb{E}(Y_i) \\
&= \tfrac{2}{n}\textstyle\sum_{i=1}^n \tfrac{\theta}{2} = \tfrac{\cancel{2}}{\cancel{n}} \cdot \cancel{n}\tfrac{\theta}{\cancel{2}} \\
&= \theta
\end{aligned}$$

$$\begin{aligned}
Var(\hat{\theta}_1) &= Var(2\bar{Y}_n) = 4Var(\bar{Y}_n) \\
&= 4 \cdot \tfrac{Var(Y)}{n} = 4 \cdot \tfrac{\sigma_Y^2}{n} \\
&\sigma_Y^2 = Var(Y) = \mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2 \\
&\mathbb{E}(Y^2) = \textstyle\int_{-\infty}^{+\infty} y^2 f_Y(y)dy = \int_0^\theta y^2 \cdot \frac{dy}{\theta} \\
&\qquad = \tfrac{1}{3\theta} y^3\Big|_0^\theta = \tfrac{\theta^3}{3\theta} = \tfrac{\theta^2}{3} \\
&\sigma_Y^2 = \mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2 = \tfrac{\theta^2}{3} - [\tfrac{\theta}{2}]^2 \\
&\qquad = \tfrac{\theta^2}{3} - \tfrac{\theta^2}{4} = \tfrac{\theta^2}{12} , \qquad \text{then:} \\
&Var(\hat{\theta}_1) = \tfrac{4\sigma_Y^2}{n} = \tfrac{4\frac{\theta^2}{12}}{n} = \tfrac{\theta^2}{3n}
\end{aligned}$$

For $\hat{\theta}_2$:

$$F_{Y(n)}(t) = P(Y(n) \le t) = P(Y_1 \le t, Y_2 \le t, \cdots, Y_m \le t) \qquad \text{by } \textstyle\coprod_{i=1}^n Y_i :$$

$\implies \prod_{i=1}^{n} P(Y_i \leq t) = \prod_{i=1}^{n} F_{Y_i}(t)$

$= [F_Y(t)]^n \qquad \therefore$ identically distributed.

Thus $d_{Y(n)}(t) = \frac{d}{dt}F_{Y(n)}(t) = \frac{d}{dt}F_Y^n(t) = nf_Y(t)F_Y^{n-1}(t)$

$$f_{Y(n)}(t) = \begin{cases} n\frac{1}{\theta}(\frac{t}{\theta})^{n-1} & 0 < t < \theta \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned}
\mathbb{E}(\hat{\theta}_2) &= \mathbb{E}[(\tfrac{n+1}{n})Y(n)] = (\tfrac{n+1}{n})\mathbb{E}(Y(n)) \\
&= (\tfrac{n+1}{n})\int_0^\theta y \cdot n \cdot \tfrac{1}{\theta}(\tfrac{y}{\theta})^{n-1}dy \\
&= (\tfrac{n+1}{n}) \cdot \tfrac{n}{\theta^n}\int_0^\theta y^n dy \\
&= (\tfrac{n+1}{n})\tfrac{n}{\theta^n}\Big[\tfrac{1}{n+1}y^{n+1}\Big]_0^\theta \\
&= (\tfrac{\cancel{n+1}}{\cancel{n}})\tfrac{\cancel{n}}{\cancel{\theta^n}}\tfrac{\cancel{\theta^{n+1}}}{\cancel{n+1}} \\
&= \theta
\end{aligned}$$

$$\begin{aligned}
Var(\hat{\theta}_2) &= Var(\tfrac{n+1}{n}Y(n)) = \Big(\tfrac{n+1}{n})^2 Var(Y(n)\Big) \\
& Var(Y(n)) = \mathbb{E}(Y_{(2)}^2) - [\mathbb{E}(Y(n))]^2 \\
& E(Y_{(n)}^2) = \int_0^\theta y^2 \cdot n \cdot \tfrac{1}{\theta}(\tfrac{y}{\theta})^{n-1}dy \\
& \qquad = \tfrac{n}{\theta^n}\int_0^\theta y^{n+1}dy \\
& \qquad = \tfrac{n}{\theta^n} \cdot \tfrac{1}{n+2} \cdot \theta^{n+2} = \tfrac{n\theta^2}{n+2} \\
& Var(Y(n)) = \tfrac{n\theta^2}{n+2} - [\tfrac{n}{n+1}\theta]^2 \\
& \qquad = n\theta^2\Big(\tfrac{1}{n+2} - \tfrac{n}{(n+1)^2}\Big) \\
& \qquad = \tfrac{n\theta^2}{(n+2)(n+1)^2} \qquad \text{thus:} \\
& Var(\hat{\theta}_2) = (\tfrac{n+1}{n})^2 \cdot \tfrac{n\theta^2}{(n+2)(n+1)^2} = \tfrac{\theta^2}{n(n+2)}
\end{aligned}$$

Thus:

$$eff(\hat{\theta}_1, \hat{\theta}_2) = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)} = \frac{\frac{\cancel{\theta^2}}{\cancel{n}(n+2)}}{\frac{\cancel{\theta^2}}{3\cancel{n}}} = \frac{3}{n+2} \to 0 \ \ as \ \ n \to \infty$$

Note that $eff(\hat{\theta}_1, \hat{\theta}_2) < 1$ for $n \geq 2$. This means that $\hat{\theta}_2$ is more efficient than $\hat{\theta}_1$ for $n \geq 2$.

We also notice that the efficiency gap increases as the sample size n increases and for large values of n , the efficiency tends to <u>zero</u>.

7.2 Consistency

Definition (Consistent Estimator)**.** *We say $\hat{\theta}_n$ is a consistent estimator of θ if $\hat{\theta}_n \xrightarrow{p} \theta$ as $n \to \infty$; i.e:*

$$\lim_{n\to\infty} P\Big(|\hat{\theta}_n - \theta| > \epsilon\Big) = 0 \quad , \quad \forall\, \epsilon > 0 \qquad\qquad (\dagger)$$

Consistency essentially means "being right-headed". It essentially says that if we have all the population, our procedure , $\hat{\theta}_n$, sizes the target.
Note that (†) is equivalent to :

$$\lim_{n\to\infty} P\Big(|\hat{\theta}_n - \theta| leq \epsilon\Big) = 1 \quad , \quad \forall \epsilon > 0$$

Now this definition can be compared with the notion of the limit of a sequence of real numbers.

$$\lim_{n\to\infty} a_n = a \quad iff \quad \forall \epsilon > 0 \;\exists\; N(\epsilon) \ni \; |a_n - a| < \epsilon \quad if \;\; n \geq N(\epsilon)$$

Now since that $\hat{\theta}_n$ is a random variable no matter how large n is, there is always a chance that $|\hat{\theta}_n - \theta| > \epsilon$. This chance, however, tends to zero as $n \to \infty$.

Example. $X_i \overset{iid}{\sim} Bernoulli(p)$,

$$X_i = \begin{cases} 1 & \qquad\qquad p \\ 0 & \qquad\qquad 1-p \end{cases} \quad , \qquad\qquad i = 1, 2, \ldots, n$$

$\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. We want to show that:

$$\lim_{n\to\infty} P\Big(|\hat{p}_n - p| > \epsilon\Big) = 0\,, \;\; \forall \epsilon > 0\,.$$

Compare $P\Big(|\hat{p}_n - p| > \epsilon\Big)$ with Tchbyshev's Inequality:

$$P(|X - \mathbb{E}(X)| > K \overbrace{\;\sigma\;}^{\sqrt{Var(X)}}) \leq \frac{1}{K^2}$$

X is replaced by $\hat{p}_n$, $\mu_X = \mathbb{E}(X)$ by p and $K\sigma$ by ϵ.

Note that $\mathbb{E}(\hat{p}_n) = p$ so everything is in order for using Tchbyshev's Inequality.

Now $\epsilon = K\sigma_X$ implies that $K = (\frac{\sigma_X}{\epsilon})^{-1}$ and given that X is replaced by $\hat{p}_n$, we should have $K = (\frac{\sigma_{\hat{p}_n}}{\epsilon})^{-1}$. Thus:

$$P\Big(|\hat{p}_n - p| > \epsilon\Big) \leq \frac{1}{(\frac{\sigma_{\hat{p}_n}}{\epsilon})^{-2}} = \frac{\sigma^2_{\hat{p}_n}}{\epsilon^2}$$

$$P\Big(|\hat{p}_n - p| > \epsilon\Big) \leq \frac{Var(\hat{p}_n)}{\epsilon^2}$$

$$\Longrightarrow \quad Var(\hat{p}_n) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{p(1-p)}{n}$$

Therefore $$P\Big(|\hat{p}_n - p| > \epsilon\Big) \leq \frac{p(1-p)}{n\epsilon^2} \leq \frac{1}{4n\epsilon^2} \to 0 \quad as \quad n \to \infty \quad \because \; p(1-p) \leq \frac{1}{4} \qquad (\ddagger)$$

Thus $$\lim_{n\to\infty} P\Big(|\hat{p}_n - p| > \epsilon\Big) = 0 \;\; , \;\; \forall \epsilon > 0$$

Note further that we can let ϵ tend to zero as $n \to \infty$, i.e. ϵ_n depend on n and $\epsilon_n \to 0$ as $n \to \infty$. Using ($\ddagger$) :

$$\lim_{n\to\infty} P\Big(|\hat{p}_n - p| > \epsilon_n\Big) \leq \lim_{n\to\infty} \frac{1}{4n\epsilon_n^2}$$

Let $\epsilon_n = \frac{\log(n)}{\sqrt{n}}$, then:

$$\lim_{n\to\infty} P\Big(|\hat{p}_n - p| > \frac{\log(n)}{\sqrt{n}}\Big) \leq \lim_{n\to\infty} \frac{1}{4n(\frac{\log(n)}{\sqrt{n}})^2} = \lim_{n\to\infty} \frac{1}{4(\log(n))^2} = 0$$

This actually gives us an idea at what rate $|\hat{p}_n - p| \xrightarrow{P} p$. Note $\epsilon_n = \frac{\alpha n}{\sqrt{n}}$ as long as $\alpha_n \to \infty$, no matter how slow, we still have the same result. This then suggests that perhaps $|\hat{p}_n - p|$ tends to zero in probability at the same rate as $\frac{1}{\sqrt{n}}$.

Suppose $X_1, \ldots, X_n$ have the same mean μ and variance σ^2 . Suppose further that $Cov(X_i, X_j) = 0 \;\; i \neq j$. Then $\bar{X}_n \xrightarrow{P} \mu$, i.e. $\bar{X}_n$ is a consistent estimator of μ , the population mean. Like the previous case:

$$P\Big(|\bar{X}_n - \mu| > \epsilon\Big) \leq \frac{1}{(\frac{\epsilon}{\sqrt{\frac{\sigma^2}{n}}})^2} = \frac{\sigma^2}{n\epsilon^2} \to 0 \quad as \; n \to \infty \qquad (1)$$

Note that $\epsilon = K\sqrt{Var(\bar{X}_n)} = K\sqrt{\frac{\sigma^2}{n}}$ and hence $K = \frac{\epsilon}{\sqrt{\frac{\sigma^2}{n}}}$. Then using Tchbushev's Inequality we obtain (1) :

$$P\Big(|\bar{X}_n - \mu| > \epsilon\Big) \leq \frac{\sigma^2}{n\epsilon^2} \to 0 \quad as\ m \to \infty \ \ \forall\, \epsilon > 0$$

Thus:

$$\lim_{n\to\infty} P\Big(|\bar{X}_n - \mu| > \epsilon\Big) = 0 \quad , \ \ \forall\, \epsilon > 0$$

meaning that $\bar{X}_n \xrightarrow{P} \mu$, i.e. $\bar{X}_n$ is a consistent estimator of μ .
The same approach cannot be used to show that $\delta_n^2 \xrightarrow{P} \sigma^2$ (We need the law of large numbers(Kolmogorov's result)).

8 Lecture 8

8.1 Consistency

Consistency is the minimal property that an estimator is expected to possess. Consistency essentially means having right-headed ; in the sense that if "all" the population's information is available , the estimator produces the exact target. Recall once again:

Definition: Suppose $\hat{\theta}_n = \hat{\theta}(X_1, \ldots, X_n)$ is an estimator of θ . We say $\hat{\theta}_n$ is a consistent estimator of θ if $\hat{\theta}_n \xrightarrow{P} \theta$ as $n \to \infty$, i.e:

$$\lim_{n\to\infty} P\Big(|th\hat{e}ta_n - \theta| > \epsilon\Big) = 0 \quad , \quad \forall\, \epsilon > 0 \qquad .$$

In lecture 7 we used Tchbyshev's inequality to establish consistency.
Markov's Inequality is an important tool in establishing consistency. In fact , Tchbyshev's inequality is a special case of Markov's inequality. It is often more straight forward to use Markov's inequality.

8.2 Markov's Inequality

Let X be a random variable and g a non-negative function. Then:

$$P(g(X) \geq \lambda) \leq \frac{\mathbb{E}[g(X)]}{\lambda} \quad , \quad \forall\, \lambda > 0 \qquad .$$

Using Markov's inequality we have:

$$P\Big(|\hat{\theta}_n - \theta| > \epsilon\Big) \leq \frac{\mathbb{E}[|\hat{\theta}_n - \theta|]}{\epsilon} \qquad (\dagger)$$

To establish consistency it then suffices to show that the upper bound of the above inequality tends to zero as $n \to \infty$.
Note that ($\dagger$) follows from Markov's inequality if we define $g(x) = |x - \theta|$. Note also that our random variable is $\hat{\theta}_n$.

To apply (†) , we need to find $\mathbb{E}[|\hat{\theta}_n - \theta|]$ which is not always easy. We however have:

$$P\Big(|\hat{\theta}_n - \theta| > \epsilon\Big) = P\Big(|\hat{\theta}_n - \theta|^2 > \epsilon^2\Big) \overbrace{\leq}^{\text{Markov's Ineq.}} \frac{\mathbb{E}[|\hat{\theta}_n - \theta|^2]}{\epsilon^2}$$

and thus:

$$P\Big(|\hat{\theta}_n - \theta| > \epsilon\Big) \leq \frac{MSE(\hat{\theta}_n}{\epsilon^2} = \frac{Var(\hat{\theta}_n) + Bias^2(\hat{\theta}_n)}{\epsilon^2} \qquad (‡)$$

where

$$MSE(\hat{\theta}_n) = \mathbb{E}[(\hat{\theta}_n - \theta)^2] = Var(\hat{\theta}_n) + [\overbrace{\mathbb{E}(\hat{\theta}_n) - \theta}^{\text{Bias}(\hat{\theta}_n)}]^2$$

Now (‡) is often much easier to use since Variance and Bias of an estimator are often hard to find.

Theorem (Slight Generalization of Theorem 9.1 , P450)**.** *Suppose* $\hat{\theta}_n$ *is an estimator of* θ *. Then* $\hat{\theta}_n \xrightarrow{P} \theta$ *if* $MSE(\hat{\theta}_n) \to 0$ *as* $n \to \infty$*. In otherwords,* $\hat{\theta}_n$ *is a consistent estimator of* θ *if* $MSE(\hat{\theta}_n) \to 0$ *as* $n \to \infty$ *.*

Proof. Using (‡) we have:

$$P\Big(|\hat{\theta}_n - \theta| > \epsilon\Big) \leq \frac{MSE(\hat{\theta}_n}{\epsilon^2} \to 0 \qquad as\, n \to \infty\ \forall \epsilon > 0\ \because\ MSE(\hat{\theta}_n) \to 0\ as\ n \to \infty$$

□

Corollary 8.1. *Let* $\hat{\theta}_n$ *be an unbiased estimator of* θ*. Suppose* $Var(\hat{\theta}_n) \to 0$ *as* $n \to \infty$ *. Then* $\hat{\theta}_n \xrightarrow{P} \theta$ *, i.e.* $\hat{\theta}_n$ *is a consistent estimator of* θ *.*

Proof.

$$MSE(\hat{\theta}_n) = Var(\hat{\theta}_n) + Bias^2(\hat{\theta}_n) = Var(\hat{\theta}_n) \to 0\ \ as\ \ n \to \infty$$

Note that the $Bias(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta = 0$ if $\hat{\theta}_n$ is an unbiased estimator of θ , i.e. $\mathbb{E}(\hat{\theta}_n) = \theta$. □

Example 8.1. $X_i \overset{iid}{\sim} Bernoulli(p) \quad , \quad i = 1, 2, \ldots, n$

$$\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\mathbb{E}(\hat{p}_n) = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n} p = p$$

Then $\mathbb{E}(\hat{p}_n) = p$. i.e. $\hat{p}_n$ is an unbiased estimator of p .

$$Var(\hat{p}_n) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(\hat{X}_i) = \frac{1}{n^2}\sum_{i=1}^{n} p(1-p) = \frac{p(1-p)}{n}$$

Now: $Var(\hat{p}_n) = \dfrac{p(1-p)}{n} \rightarrow 0 \quad as\ n \rightarrow \infty$

Thus using corollary 8.1 , $\hat{p}_n \xrightarrow{P} p$, i.e. $\hat{p}_n$ is a consistent estimator of p .

Example 8.2. *Suppose $X_1, \ldots, X_n$ are independent and identically distributed random variables with the common mean value μ and common variance σ^2.*

Then:

$$\mathbb{E}(\bar{X}_n) = \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

$$Var(\bar{X}_n) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}$$

Thus: $$MSE(\bar{X}_n) = Var(\bar{X}_n) + \overbrace{Bias^2(\bar{X}_n)}^{0} = \frac{\sigma^2}{n} \rightarrow 0 \ \ as\ n \rightarrow \infty$$

and hence $\bar{X}_n \xrightarrow{\mu}$ using corollary 8.1 , i.e. $\bar{X}_n$ is a consistent estimator of μ.

Remark 8.1. *The conclusion of 2^{nd} example remains intact if the independence assumption is replaced by orthogonality , i.e. $Cov(X_i, X_j) \ \ , i \neq j$.*

Corollary 8.2. *Suppose $\hat{\theta}_n$ is an asymptotically unbiased estimator of θ , i.e. $\lim\limits_{n\to\infty}\mathbb{E}(\hat{\theta}_n) = \theta$. Suppose further that $Var(\hat{\theta}_n) \rightarrow 0$ as $n \rightarrow \infty$. Then $\hat{\theta}_n \xrightarrow{P} \theta$, i.e. $\hat{\theta}_n$ is a consistent estimator of θ*

Proof.

$$\lim_{n\to\infty} MSE(\hat{\theta}_n) = \lim_{n\to\infty} Var(\hat{\theta}_n) + \lim_{n\to\infty} Bias^2(\hat{\theta}_n) = 0 + [\overbrace{\lim_{n\to\infty}(\mathbb{E}(\hat{\theta}_n) - \theta)}^{0}]^2 = 0$$

The desired result then follows form the above theorem. ◻

Remark. *The above result tells us that unbiasedness is NOT necessary for consistency.*

Example 8.3. *Suppose $X_1, \dots, X_n$ from a random sample from a population with the mean μ and variance σ^2. We want to estimate σ^2.*

We showed that $\mathbb{E}(S_n^2) = \sigma^2$ where :

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

We want to show that $S_n^2 \xrightarrow{P} \sigma^2$, i.e. S_n^2 is a consistent estimator of σ^2 . Note that using Markov's inequality we have:

$$P\Big(|S_n^2 - \sigma^2| > \epsilon\Big) \leq \frac{Var(S_n^2)}{\epsilon^2}$$

We need to show that $Var(S_n^2) \to 0$ as $n \to \infty$. To do this, we need to find $Var(S_n^2)$ in terms of the moments of the population. We therefore require conditions on the 4^{th} moment of the population from which the samples were taken. Below we give a different approach that is much easier to apply and require lesser assumptions, but much more base.

8.3 Kolmogorov's Law of Large Numbers(LLN)

Suppose $X_{n_{n=1}}^{\infty}$ is a sequence of *iid* random variables with common mean μ. Then:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu = \mathbb{E}(X)$$

Remark. *Kolgomorov's theorem is actually much stronger than this. It established a*

stronger notion of convergence. The complete form has two sides. It also shows that if $\bar{X}_n$ converges to a constant, since C , in that stronger notion of convergence, then:
$\mathbb{E}(|X|) < \infty \quad \& \quad C = \mathbb{E}(X)$.

Corollary. *Suppose $\left\{X_n\right\}_{n=1}^{\infty}$ is a sequence of iid random variables with the common K^{th}-moment μ_K , i.e. $\mathbb{E}(X^K) = \mu_K$, for some $K \in \mathbb{N}$, then:*

$$\frac{1}{n}\sum_{i=1}^{n} X_i^K \xrightarrow{P} \mu_K = \mathbb{E}(X^K)$$

This corollary follows from Kolmogorov's theorem immediately upon defining $Y_i = X_i^K$.
Note that if $\mathbb{E}(X^K) < \infty$, then $\mathbb{E}(X^n) < \infty \ \ \forall 0 \le r \le K$. This then means that if $X_{n_{n=1}}^{\infty}$ is a sequence *iid* random variables with the common K^{th}-moment μ_K , then:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^r \xrightarrow{P} \mu_K = \mathbb{E}(X^r) \ \ \forall\ 0 \le r \le K$$

We also need the following theorem which is essentially theorem 9.2 , page 451 of the textbook.

Theorem 8.1 (Theorem 9.2, page 451)**.** *Suppose $\hat{\theta}_n = \hat{\theta}(X_1, \ldots, X_n)$ and $\hat{\mathscr{C}}_n = \hat{\mathscr{C}}(X_1, \ldots, X_n)$ are consistent estimators of θ and $\mathscr{C}$, respectively, i.e. $\hat{\theta}_n \xrightarrow{P} \theta$ and $\hat{\mathscr{C}}_n \xrightarrow{P} \mathscr{C}$.*

a) $\hat{\theta}_n \hat{\mathscr{C}}_n \xrightarrow{P} \theta\mathscr{C}$

b) $\hat{\theta}_n \pm \hat{\mathscr{C}}_n \rightarrow \hat{\theta}_n \pm \hat{\mathscr{C}}_n$

c) $\frac{\hat{\theta}_n}{\hat{\mathscr{C}}_n} \rightarrow \frac{\theta}{\mathscr{C}}$ *provided that* $\hat{\mathscr{C}} \neq 0$, $\mathscr{C} \neq 0$

d) $g(\hat{\theta}_n) \rightarrow g(\theta)$ *if* $g(.)$ *is a continuous function*

Part(d) of the above theorem is called **Continuous Mapping Theorem** .

Now to establish consistency of S_n^2 , we fist establish $\quad S_{n,*}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 \xrightarrow{P} \sigma^2$.

Step 1:

$$\begin{aligned}
S_{n,*}^2 &= \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n}[\sum_{i=1}^n X_i^2 - 2\bar{X}_n \overbrace{\sum_{i=1}^n X_i}^{n\bar{X}_n} + n\bar{X}_n^2] \\
&= \frac{1}{n}[\sum_{i=1}^n X_i^2 - 2\bar{X}_n(n\bar{X}_n) + n\bar{X}_n^2] \\
&= \frac{1}{n}[\sum_{i=1}^n X_i^2 - 2n\bar{X}_n^2 + n\bar{X}_n^2] \\
&= \frac{1}{n}[\sum_{i=1}^n X_i^2 - n\bar{X}_n^2] \\
&= \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2
\end{aligned}$$

Step 2:

Using Kolmogorov's Theorem:

a) $\frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{P} \mathbb{E}(X^2)$

b) $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{P} \mathbb{E}(X)$

Step 3:

Using Step 2 and continuous mapping theorem (Theorem 9.2 (d))

$$\bar{X}_n^2 \rightarrow [\mathbb{E}(X)]^2$$

Step 4:

Using Step 1,2,3 and Theorem 9.2 (b) we have:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2 \quad \xrightarrow{P} \quad \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = Var(X) = \sigma^2$$

Thus $S_{n,*}^2 \xrightarrow{P} \sigma^2$, i.e. $S_{n,*}^2$ is a consistent estimator of σ^2 .

Next we note that:

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

Using Theorem 9.2 (a):

$$\lim_{n\to\infty}(\frac{n}{n-1}) = 1 \quad \& \quad S_{n,*}^2 \xrightarrow{P} \sigma^2$$

Then Theorem 9.2 (a) implies that:

$$S_n^2 = (\frac{n}{n-1})S_{n,*}^2 \quad \xrightarrow{P} \quad 1 \cdot \sigma^2 = \sigma^2$$

i.e. S_n^2 is a consistent estimator of σ^2 .

Remark. *Suppose $P(X_n = C_n) = 1$, $n = 1, 2, \ldots$ where $\left\{C_n\right\}_{n=1}^{\infty}$ is a sequence of real numbers such that $\lim_{n\to\infty} C_n = C$. Then $X_n \xrightarrow{P} C$. The proof of this result is as follow:*

Proof.

$\lim_{n\to\infty} C_n = C$ i.e. $\forall\, \epsilon > 0\ \exists\, N(\epsilon) \in \mathbb{N} \ni |C_n - C| < \epsilon$, $\forall\, n \geq N(\epsilon)$

Now suppose $\epsilon > 0$ is given , then:

$P(|X_n - C| > \epsilon) = P(|C_n - C| > \epsilon) = 0 \qquad$ if $\quad n \geq N(\epsilon)$

Thus $\lim_{n\to\infty} P(|X_n - C| > \epsilon) = 0$, $\forall\, \epsilon > 0$ □

Question: Why couldn't we use Kolmogorov's theorem directly to establish consistency of $S_{n,*}^2$? In other words, coudln't we define $Z_i = (X_i - \bar{X}_n)^2$ and hence $S_{n,*}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} Z_i$ and then apply Kolgomorov's theorem? The answer is that Z_is are not independent . Note that $\sum_{i=1}^{n}(X_i - \bar{X}_n) = 0$.

8.4 Sufficiency

Sufficiency is essentially comparison. Sufficiency is one of the main pillars of the likelihood Inference .

As the following diagram shows the likelihood inference has three main components: the observable quantities, samples, the unobservable quantities, the unknown parameters to be estimated , and a parametric distribution that links the observables to unobservables .

Example: $f_\theta(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad where \quad \theta = (\mu, \sigma^2)$

#MISSING GRAPH Lecture 8 - Page 57

9 Lecture 9

9.1 Sufficiency

Suppose $T : \mathbb{R}^n \to \mathbb{R}^m$, $m < n$ is a map from $\mathbb{R}^n$ to $\mathbb{R}^m$ and $X = (X_1, X_2, \ldots, X_n)$ and $Y = (Y_1, Y_2, \ldots, Y_n)$ are two n-dimensional random vectors. Then $\underset{\sim}{X}$ & $\underset{\sim}{Y}$ are called T-similar if:

$$P_{\underset{\sim}{X}|T=t}(\underset{\sim}{u}|t, \theta) = P_{\underset{\sim}{Y}|T=t}(\underset{\sim}{u}|t, \theta) \qquad \forall\, u \text{ and } t$$

Definition 9.1 (T-similar). *A realization of $\underset{\sim}{X}$, say $\underset{\sim}{x}$, and a realization of $\underset{\sim}{Y}$ say $\underset{\sim}{y}$, are called T-similar if:*

1. *$\underset{\sim}{X}$ and $\underset{\sim}{Y}$ are T-similar .*
2. *$T(\underset{\sim}{x}) = T(\underset{\sim}{y})$*

What do we expect from a good comparison?

Suppose θ is the unknown parameter of interest. We want to estimate θ using $\underset{\sim}{X} = (X_1, \ldots, X_n)$. Now $T_n = T(\underset{\sim}{X}) = T(X_1, \ldots, X_n)$ is a good comparison if:

a) T_n can preserve all the pertinent "information" in $\underset{\sim}{X} = (X_1, \ldots, X_n) to \theta$

b) if $\underset{\sim}{X}^* = (X_1^*, \ldots, X_n^*)$ is the original sample, for any given value of T_n , say t, we can generate a T_n-similar sample of xs.

In order to generate a T_n-similar sample $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | T_n = t)$ should be free from any unknown parameter.

As per retaining pertinent information in the data to the unknown parameter $\underset{\sim}{\theta}$, given that the link between the data and the unknown parameter(s) is the joint distribution:

$$P\Big(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n; \underset{\sim}{\theta}\Big) = P_{\underset{\sim}{\theta}}\Big(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\Big)$$

any possible information in the data about θ should be in $P_{\underset{\sim}{\theta}}\Big(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\Big)$

.

As such $T_n = T(X_1, \ldots, X_n)$ can preserve all the pertinent information if $P_{\underset{\sim}{\theta}}\Big(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\Big)$ *The Fisher-Neyman Factorization Theorem* shows that this proportional is indeed a characterization of sufficient statistics. Let's dig into this a bit more. Note that:

$$P_{\underset{\sim}{\theta}}\Big(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\Big) = P_{\underset{\sim}{\theta}}\Big(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \Big| T_n = t\Big) P_\theta(T_n = t) \ .$$

If $P_{\underset{\sim}{\theta}}\Big(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \Big| T_n = t\Big)$ is actually free from $\underset{\sim}{\theta}$, then:

$$P_{\underset{\sim}{\theta}}\Big(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\Big) \ \propto \ P_\theta(T_n = t) \overset{def.}{=} g(t\ ;\theta)$$

where the proportionality constant is a function of $\underset{\sim}{x} = (x_1, x_2, \ldots, x_n)$, the observed sample. A formal definition then emerges.

Definition. *Let $X_1, \ldots, X_n$ be a random sample from a distribution with an unknown parameter $\underset{\sim}{\theta}$. A statistic $T_n = T(X_1, \ldots, X_n)$ is called **sufficient** for θ if the conditional distribution of $(X_1, \ldots, X_n)$ given T_n does not depend on $\underset{\sim}{\theta}$.*

Example. *$X_i \overset{iid}{\sim} Bernoulli(p)$, $i = 1, 2, \ldots, n$.*

$$P_p(X = x) = p^x (1-p)^{1-x} \ \ , \ \ x = 0, 1$$

Consider $T_n = \sum_{i=1}^{n} X_i$. Note that $T_n \sim Bin(n,p)$ Then:

$$P\Big(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \Big| T_n = t\Big) = \begin{cases} \frac{P(X_1=x_1, X_2=x_2, \ldots, X_n=x_n, T_n=t)}{P(T_n=t)} & \text{if } \sum_{i=1}^{n} x_i = t \\ 0 & \text{otherwise} \end{cases}$$

$$= \begin{cases} \frac{P(X_1=x_1, X_2=x_2, \ldots, X_n=x_n)}{P(T_n=t)} & \text{if } \sum_{i=1}^{n} x_i = t \\ 0 & \text{otherwise} \end{cases}$$

$$= \begin{cases} \frac{\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}}{\binom{n}{t} p^t (1-p)^{n-t}} & \text{if } \sum_{i=1}^{n} x_i = t \\ 0 & \text{otherwise} \end{cases}$$

$$= \begin{cases} \frac{\cancel{p^{\sum_{i=1}^{n} x_i}} \cancel{(1-p)^{n-\sum_{i=1}^{n} x_i}}}{\binom{n}{t} \cancel{p^t} \cancel{(1-p)^{n-t}}} & \text{if } \sum_{i=1}^{n} x_i = t \\ 0 & \text{otherwise} \end{cases}$$

$$\text{thus} \quad P\Big(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \Big| T_n = t\Big) = \begin{cases} \frac{1}{\binom{n}{t}} & \text{if } \sum_{i=1}^{n} x_i = t \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

and hence T_n is a sufficient statistic for p .

Remark 9.1. *To generate a T_n-similar sample when T_n is given, say $T_n = t$, we define:*

$$A_t = \Big\{(x_1, x_2, \ldots, x_n) : \sum_{i=1}^{n} x_i = t\Big\}$$

Note that $card(A_t) = \binom{n}{t}$. According to (1) we give equal weight , i.e. problem mass, to each element of A_t . We then choose one element of A_t randomly.

Example 9.1. $X_i \overset{iid}{\sim} P(\lambda) \quad , \quad i = 1, 2, \ldots, n$

$$P_\lambda(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad , \quad x = 0, 1, 2, \ldots$$

Consider $T_n = \sum_{i=1}^{n} x_i$. Note that $T_n \sim P_0(n\lambda)$ (Exercise)

$$
\begin{aligned}
P_\lambda(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \big| T_n = t) &= \begin{cases} \frac{P_\lambda(X_1=x_1, X_2=x_2, \ldots, X_n=x_n, T_n=t)}{P(T_n=t)} & \text{if } \sum_{i=1}^{n} x_i = t \\ 0 & \text{otherwise} \end{cases} \\
&= \begin{cases} \frac{\frac{\prod_{i=1}^{n} e^{-\lambda}\lambda^{x_i}}{x_i!}}{\frac{e^{-n\lambda}(n\lambda)^t}{t!}} & \text{if } \sum_{i=1}^{n} x_i = t \\ 0 & \text{otherwise} \end{cases} \\
&= \begin{cases} \frac{\frac{\cancel{e^{-n\lambda}}\,\cancel{\lambda^{\sum_{i=1}^{n} x_i}}}{\prod_{i=1}^{n} x_i!}}{\frac{\cancel{e^{-n\lambda}}\,(n\cancel{\lambda})^t}{t!}} & \text{if } \sum_{i=1}^{n} x_i = t \\ 0 & \text{otherwise} \end{cases} \\
&= \begin{cases} \frac{t!}{\prod_{i=1}^{n} x_i!} \cdot \frac{1}{n^t} & \text{if } \sum_{i=1}^{n} x_i = t \\ 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{2}
$$

Thus:

$$\boxed{\underset{\sim}{X}\,|_{T_n=t} \sim \text{Multinomial}(t, p_i = \frac{1}{n},\ i = 1, 2, \ldots, n)}$$

Recall that:

$(Y_1, Y_2, \ldots, Y_k) \sim \text{Multinomial}(n, p_1, p_2 \ldots, p_k)$ if $P(X_1 = x_1, \ldots, X_k = x_k) = \binom{n}{x_1, \ldots, x_k} \prod_{i=1}^{n} p_i^{x_i}$

where $\sum_{i=1}^{n} x_i = n$ & $\sum_{i=1}^{k} p_i = 1$ and $\binom{n}{x_1, x_2, \ldots, x_k} = \frac{n!}{x_1!, x_2! \ldots, x_k!}$

Again, to generate a T_n-similar sample we define:

$$A_t = \left\{(x_1, x_2, \ldots, x_n) \colon \sum_{i=1}^{n} x_i = t\right\}$$

The probability mass associated to elements of A_t is given by (2) . In other words, we choose an element of A_t using a multinomial distribution with $n = t$, $k = n$ *and* $p_i = \frac{1}{n}$, $i = 1, 2, \ldots, n$.

9.2 Likelihood

Definition. *Let* $\underset{\sim}{Y} = (Y_1, Y_2, \ldots, Y_n)$ *be a random vector whose joint pdf or pmf depends on* $\underset{\sim}{\theta}$ *, a vector of unknown parameters.* ***The Likelihood Function*** *. a function of* $\underset{\sim}{\theta}$ *, for a realization* $\underset{\sim}{y} = (y_1, y_2, \ldots, y_n)$ *is defined to be:*

$$\mathcal{L}(\underset{\sim}{\theta}; \underset{\sim}{y}) = \begin{cases} P_{\underset{\sim}{\theta}}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n) & \text{if } y_i\text{s are discrete random variables} \\ f_{\underset{\sim}{\theta}}(y_1, y_2, \ldots, y_n) & \text{if } y_i\text{s are continuous random variables} \end{cases}$$

Example. $X_i \overset{iid}{\sim} \textit{Bernoulli}(p)$, $i = 1, 2, \ldots, n$

$$P_p(X = x) = p^x (1-p)^{1-x} \;,\; i = 0, 1$$

$$\begin{aligned} \mathcal{L}(p; x_1, x_2, \ldots, x_n) &= P_p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ \implies &= P_p(X_1 = x_1) \ldots P_p(X_n = x_n) \qquad \text{Independenc and identically distributed} \\ &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum\limits_{i=1}^{n} x_i}(1-p)^{n - \sum\limits_{i=1}^{n} x_i} \qquad (B) \end{aligned}$$

Example. $X_i \overset{iid}{\sim} P_0(\lambda)$, $i = 1, 2, \ldots, n$

$$P_\lambda(X = x) = \tfrac{e^{-\lambda}\lambda^x}{x!} \;,\; x = 0, 1, 2, \ldots$$

$$\begin{aligned} \mathcal{L}(\lambda; x_1, x_2, \ldots, x_n) &= P_\lambda(X_1 = x_1, \ldots, X_n = x_n) \\ &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \qquad (P) \qquad \text{Independence and identically distributed} \\ &= \frac{e^{-n\lambda}\lambda^{\sum\limits_{i=1}^{n} x_i}}{\prod\limits_{i=1}^{n} x_i!} \end{aligned}$$

Example. $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < +\infty$$

$$\begin{aligned}
\mathscr{L}(\mu, \sigma^2; x_1, x_2, \ldots, x_n) &= f_{\mu,\sigma^2}(x_1, x_2, \ldots, x_n) \\
&= \prod_{i=1}^{n} f_{\mu,\sigma^2}(x_i) \qquad \text{Independent and identically distributed} \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\
&= (\frac{1}{\sqrt{2\pi\sigma^2}}) exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{2} (x_i - \mu)^2 \Big\} \qquad (N)
\end{aligned}$$

The examples we presented for sufficiency required first specifying a candidate statistic. The question then is how we come up with a sufficient statistic. The following theorem due to *Fisher* & *Neyman* tells us how to find sufficient statistic.

Theorem (Fisher-Neyman Factorization Theorem - Thm 9.4 , page 461)**.** *A statistic $T = T(Y_1, T_2, \ldots, T_n)$ for $\underset{\sim}{\theta}$ the parameter of the distribution of $Y_1, Y_2, \ldots, Y_n$ if and only if : $\mathscr{L}(\underset{\sim}{\theta}; y_1, y_2, \ldots, y_n) = g\ (t; \theta)\ h\ (y_1, y_2, \ldots, y_n)$ For any realization $(y_1, y_2, \ldots, y_n)$, where $t = T\ (y_1, y_2, \ldots, y_n)$.*

Example. $X_i \overset{iid}{\sim} Bernoulli(p)$, $i = 1, 2, \ldots, n$

Using (B) (the result above):

$$\mathscr{L}(p; x_1, x_2, \ldots, x_n) = p^{\sum_{i=1}^{n} x_i} (1-p)^{n - \sum_{i=1}^{n} x_i}$$

Then define $T = \sum_{i=1}^{n} x_i$, $g(t; p) = p^{\sum_{i=1}^{n} x_i} (1-p)^{n-\sum_{i=1}^{n} x_i} = p^t (1-p)^{n-t}$ and $h(x_1, x_2, \ldots, x_n) \equiv 1$. Using Fisher-Neyman theorem $T = \sum_{i=1}^{n} X_i$ is a sufficient statistic for p .

Example. $X_i \overset{iid}{\sim} P_0(\lambda)$, $i = 1, 2, \ldots, n$

Using (P) (the result above): $\mathscr{L}(\lambda; x_1, \ldots, x_n) = \frac{e^{-n\lambda} \lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$. Define $T = \sum_{i=1}^{n} X_i$ and $g(t; \lambda) = e^{-n\lambda}\lambda^t = e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}$ and $h(x_1, \ldots, x_n) = \frac{1}{\prod_{i=1}^{n} x_i!}$. Then using Fisher-

Neyman Theorem, $T = \sum i = 1^n X_i$ is a sufficient statistic for λ .

Example. $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$.

Note that now we have two unknown parameters, μ and σ^2 . Using (N) : $\mathscr{L}(\mu, \sigma^2; x_1, x_2, \ldots, x_n) = (\frac{1}{\sqrt{2\pi}\sigma})^n exp\Big\{-\frac{1}{2\sigma^2}\sum\limits_{i=1}^{n}(x_i - \mu)^2\Big\}$

Now note that: $\sum\limits_{i=1}^{n}(x_i - \mu)^2 = \sum\limits_{i=1}^{n} x_i^2 - 2\mu\sum\limits_{i=1}^{n} x_i + n\mu^2$

Thus: $\mathscr{L}(\mu, \sigma^2; x_1, x_2, \ldots, x_n) = (\frac{1}{\sqrt{2\pi}\sigma})^n exp\Big\{-\frac{\sum\limits_{i=1}^{n} x_i^2}{2\sigma^2} + \frac{\mu\sum\limits_{i=1}^{n} x_i}{\sigma^2} - \frac{n\mu^2}{2\sigma^2}\Big\}$

Define: $T = (\sum\limits_{i=1}^{n} X_i, \sum\limits_{i=1}^{n} X_i^2)$, $g(\underset{\sim}{t}; \underset{\sim}{\theta}) = (\frac{1}{\sqrt{2\pi}\sigma})^n exp\Big\{-\frac{\sum\limits_{i=1}^{n} x_i^2}{2\sigma^2} + \frac{\mu\sum\limits_{i=1}^{n} x_i}{\sigma^2} - \frac{n\mu^2}{2\sigma^2}\Big\}$ where $\underset{\sim}{\theta} = (\mu, \sigma^2)$ and $h(x_1, x_2, \ldots, x_n) \equiv 1$. Then using Fisher-Neyman Theorem $T = (\sum\limits_{i=1}^{n} X_i, \sum\limits_{i=1}^{n} X_i^2)$ is a sufficient statistic for $\underset{\sim}{\theta} = (\mu, \sigma^2)$.

Remark. *Note that do identify the sufficient statistic using Fisher-Neyman Theorem, you only need the part of the likelihood in which you cannot separate the unknown parameters from observations. This part is called* **kernel** *. In other words, you can write a likelihood as the product of a function of observations alone, a function of parameters alone and the kernel.* The sufficient statistic is in the kernel.

10 Lecture 10

10.1 The Rao-Blackwell Theorem

[Theorem 9.5 , page 464] An interesting and important application of sufficiency is in variance reduction. This application is formalized in a theorem due to Rao(C.R) and Blackwell (David) .

We first need to recall Theorem 5.14 (Page 286) and Theorem 5.15 (page 287):

Theorem 5.14 (page 286):

$$\mathbb{E}(X) = \mathbb{E}\left\{\mathbb{E}(X|Y)\right\}$$

Theorem 5.15 (page 287):

$$Var(X) = Var\left\{\mathbb{E}(X|Y)\right\} + \mathbb{E}\left\{Var(X|Y)\right\}$$

Theorem (The Rao-Blackwell Theorem - Thm 9.5 , page 464)**.** *Let $\hat{\theta}$ be an unbiased estimator for θ such that $V(\hat{\theta}) < \infty$. Suppose T is a sufficient statistic for θ . Define $\hat{\theta}^* = \mathbb{E}(\hat{\theta}\big|T)$. Then, for all θ :*

(*a*) $\mathbb{E}(\hat{\theta}^*) = \theta$

(*b*) $Var(\hat{\theta}^*) \leq Var(\hat{\theta})$.

Proof. First note that T is a sufficient statistic for θ , thus the distribution of $\hat{\theta}$ given T does not depend on θ . Therefore $\mathbb{E}(\hat{\theta}|T)$ is a statistic . This is when sufficiency plays its role.

To prove part (*a*) of formula, we use Theorem 5.14, page 286:

$$\mathbb{E}(\hat{\theta}^*) = \mathbb{E}\Big[\mathbb{E}(\hat{\theta}|T)\Big] = \mathbb{E}(\hat{\theta}) = \theta \qquad \forall\ \theta.$$

To prove part (b), we use Theorem 5.15, page 287:

$$Var(\hat{\theta}^*) = Var\Big\{\mathbb{E}(\hat{\theta}|T)\Big\} \le Var\Big\{\mathbb{E}(\hat{\theta}|T)\Big\} + \overbrace{\mathbb{E}\{\underbrace{Var(\hat{\theta}|T)}_{\ge 0}\}}^{\ge 0}$$

□

Remark (Completeness)**.** *A statistic T or its family of distribution $\{F_\theta : \theta \in \Theta\}$ where Θ is the set of all admissible values of θ, is called **complete** if for any reasonable g:*

$$\mathbb{E}_\theta[g(T)] = 0 \quad , \quad \forall\, \theta \in \Theta$$

implies that $g(t) = 0$ for all possible values of t. If a sufficient statistic T is also complete, then $\hat{\theta}^* = \mathbb{E}(\hat{\theta}|T)$ will be the Minimum Variance Unbiased Estimator (MVUE) . This often means that within the class of unbiased estimator $\hat{\theta}^*$ is the least; i.e. the closest in the MSE sense, to the unknown parameter θ. Recall that:

$$MSE_\theta(\hat{\theta}) = Var_\theta(\hat{\theta}) + Bias^2_\theta(\hat{\theta}) = Var_\theta(\hat{\theta})$$

if $Bias_\theta(\hat{\theta}) = 0$; i.e. if $\hat{\theta}$ is an unbiased estimator of θ.
The notion of Completeness is due to ***Lehmann(Eric Leo) and Scheffe'(Henry)***.

Then using the Rao-Blackwell and Lehmann-Schaffe' theorems we have an easy recipe for finding the $MVUE$.

Step 1: Using Fisher-Neyman theorem, find a sufficient statistic, say T, for θ .

Step 2: Find an unbiased estimator θ , say $\hat{\theta}$.

Step 3: Find $\hat{\theta}^* = \mathbb{E}(\hat{\theta}|T)$.

Remark. *For the examples and exercises in the course , the sufficient statistic you find in **Step 1** using Fisher-Neyman Theorem is also complete.*

Example (Ex. 9.6 , page 466)**.** $X_i \overset{iid}{\sim} Bernoulli(p) \;\; , \;\; i = 1, 2, \ldots, n$

Step 1:

$$
\begin{aligned}
\mathscr{L}(p; x_1, \ldots, x_n) &= P_p(X_1 = x_1, \ldots, X_n = x_n) \\
\text{Independence} \qquad &= \prod_{i=1}^{n} P_p(X_i = x_i) \\
\text{Identically distributed} \qquad &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\
&= p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}
\end{aligned}
$$

Then $T = \sum_{i=1}^{n} X_i$ is a sufficient statistic.

Step 2:

$$\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ is an unbiased estimator of } p$$

$$\mathbb{E}(\hat{p}_n) = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathbb{E}(X_i)}_{p} = \frac{1}{n} \cdot np = p$$

Step 3:

Note that $\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{T}{n}$ thus $\mathbb{E}(\hat{p}_n|T) = \frac{T}{n} = \hat{p}_n$.

Remark. What we observed in *step 3* of the above example tells us that *step 3* of our recipe is redundant when θ found in *step 2* is a function of the sufficient statistic found in *step 1*.

Example (Ex. 9.8 , page 467)**.** $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$

Step 1

$$\begin{aligned}\mathscr{L}(\mu,\sigma^2;x_1,\dots,x_n) &= f_{\mu,\sigma^2}(x_1,\dots,x_n)\\ &= \prod_{i=1}^{n} f_{\mu,\sigma^2}(x_i)\\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\end{aligned}$$

hence: $\mathscr{L}(\mu,\sigma^2;x_1,\dots,x_n) = (\frac{1}{\sqrt{2\pi}\sigma})^n\ exp\Big\{-\frac{1}{2\sigma^2}\Big[\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2\Big]\Big\}$

Thus: $$T = \Big(\sum_{i=1}^{n} X_i\ ,\ \sum_{i=1}^{n} X_i^2\Big)$$

Step 2: $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an unbiased estimator of μ and $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$ is an unbiased estimator of σ^2 .

Step 3:

$\mathbb{E}(\bar{X}_n\big|T) = \bar{X}_n$ since $\bar{X}_n$ is a

$S_n^2 = \frac{1}{n-1}\Big[\sum_{i=1}^{n} X_i^2 - n\bar{X}\Big]$ is also a funct

Thus $\mathbb{E}(S_n^2|T) = S_n^2$

Thus $\bar{X}_n$ is the $MVUE$ of μ and S_n^2 is the $MVUE$ of σ^2 .

Example (Ex. 9.7 , page 466-467). $Y_i \overset{iid}{\sim} Weibull(m=2,\theta)$, $i = 1,2,\dots,n$

$$f_\theta(y) = \begin{cases} (\frac{2y}{\theta})e^{-\frac{y^n}{\theta}} & y > 0\\ 0 & \text{otherwise}\end{cases}$$

Step 1:

$$\begin{aligned}\mathscr{L}(\theta;y_1,y_2,\dots,y_n) &= \prod_{i=1}^{n}(\frac{2y_i}{\theta})e^{-\frac{y_i^2}{\theta}}\\ &= (\frac{2}{\theta})^n\ e^{-\frac{1}{\theta}\sum_{i=1}^{n} y_i^2}\prod_{i=1}^{n} y_i\end{aligned}$$

Thus $T = \sum_{i=1}^{n} Y_i^2$ is a sufficient statistic for θ .

Note that the Kernel is $exp\Big\{ -\frac{1}{\theta} \sum_{i=1}^{n} Y_i^2 \Big\}$. We can also see this through Fisher-Neyman by choosing:

$$g(t;\theta) = (\frac{2}{\theta})^n \, e^{-\frac{t}{\theta}} \quad \text{and} \quad h(y_1,\ldots,y_n) = \prod_{i=1}^{n} y_i \quad \text{where } t = \sum_{i=1}^{n} y_i^2$$

Step 2: Define $W_i = Y_i^2$, $i = 1,2,\ldots,n$. Note that:

$$\begin{aligned} f_W(w) &= f_Y(\sqrt{w}) \Big|\frac{d\sqrt{w}}{dw}\Big| \qquad \text{using } \textit{Transformation Method} \\ &= \begin{cases} (\frac{2\sqrt{w}}{\theta})\, e^{-\frac{(\sqrt{w})^2}{\theta}} \cdot \frac{1}{2\sqrt{w}} & \text{if } w > 0 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{1}{\theta}\, e^{-\frac{w}{\theta}} & \text{if } w > 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Thus $W \sim Exp(\theta)$ and therefore

$$\mathbb{E}(T) = \mathbb{E}(\sum_{i=1}^{n} Y_i^2) = \sum_{i=1}^{n} \mathbb{E}(Y_i^2) = \sum_{i=1}^{n} \mathbb{E}(W_i) = n\theta \implies \mathbb{E}(\frac{T}{n}) = \theta$$

Step 3

$$\mathbb{E}(\frac{T}{n} \,\Big|\, T) = \frac{T}{n} \quad \text{therefore} \quad \frac{T}{n} = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 \quad \text{is the } MVUE \text{ of } \theta \ .$$

Remark. Sufficient statistics can often be used to make a pivotal quantity , in the above example for instance,

$$\frac{2}{\theta} W \sim \chi^2_{(2)} \qquad \text{(Exercise)}$$

and hence

$$\frac{2}{\theta}\sum_{i=1}^{n} W_i = \frac{2}{\theta}\sum_{i=1}^{n} Y_i^2 \sim \chi^2_{(2n)} \qquad \text{i.e.} \qquad \frac{2T}{\theta} \sim \chi^2_{2n}$$

This pivotal quantity can therefore be used to make *exact confidence interval* for θ . See example 9.10 , page 468, confidence interval made using sufficient statistic based on pivotal quantities often have the shortest possible length for a given confidence level .

11 Lecture 11

Methods of Estimation: A) Method of Maximum Likelihood (ML)

B) Method of Moments

11.1 Method of Maximum Likelihood (ML)

Definition. *The **Maximum Likelihood Estimation (MLE)** of a parameter θ based on the realized values $(y_1, y_2, \ldots, y_n)$ of a sample $Y_1, Y_2, \ldots, Y_n$ is:*

$$\hat{\theta}_{ML} = argmax\ \mathscr{L}(\theta; y_2, \ldots, y_n)$$

Then we have a two step procedure for finding the $\hat{\theta}_{ML}$:

Step 1: Set up the likelihood $\mathscr{L}(\theta; y_1, \ldots, y_n)$

Step 2: Find the maximizer of the likelihood when we have a random sample, which is the case in this course:

$$\mathscr{L}(\theta; y_1, \ldots, y_n) = \prod_{i=1}^{n} f_\theta(x_i)$$

To find $\hat{\theta}_{ML}$ is often easier to work with $\mathscr{L}(\theta; y_1, \ldots, y_n) = \log(\mathscr{L}(\theta; y_1, \ldots, y_n)) = \sum_{i=1}^{n} \log(f_\theta(x_i))$. Note that log is a monotone increasing function, Thus:

$$\text{argmax}\ l\,(\theta; y_1, \ldots, y_n) = \text{argmax}\ \mathscr{L}(\theta; y_1, \ldots, y_n)$$

Example. $X_i \overset{iid}{\sim} Bernoulli(p)\ \ ,\ \ i = 1, 2, \ldots, n$

$$P(X = x) = p^x(1-p)^{1-x}\ \ ,\ \ x = 0, 1$$

Step 1: $\mathscr{L}(p; x_1, \ldots, x_n) = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$

Step 2:

$$l(p;x_1,\dots,x_n) = \log\,(\,\mathscr{L}(p;x_1,\dots,x_n)\,)$$
$$= (\sum_{i=1}^{n} x_i)\log p + (n - \sum_{i=1}^{n})\ \log(1-p)$$
$$\frac{\partial}{\partial p} l(p;x_1,\dots,x_n) = \frac{\sum\limits_{i=1}^{n} x_i}{p} - \frac{n - \sum\limits_{i=1}^{n} x_i}{1-p}$$

Let $t = \sum\limits_{i=1}^{n} x_i$. The $\hat{p}_{ML}$ is then the solution to

$$\frac{t}{\hat{p}_{ML}} - \frac{n-t}{1-\hat{p}_{ML}} = 0 \iff \frac{t}{n-t} = \frac{\hat{p}_{ML}}{1-\hat{p}_{ML}} \iff \hat{p}_{ML} = \frac{t}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \hat{p}_n$$

Note that

$$\frac{\partial^2 l}{\partial p^2} = -\frac{\sum\limits_{i=1}^{n} x_i}{p^2} - \frac{n - \sum\limits_{i=1}^{n} x_i}{(1-p)^2} < 0$$

Thus $\hat{p}_{_{ML}}$ is the maximizer of $l\,(p;x_1,\dots,x_n)$.

Example. $X_i \overset{iid}{\sim} N(\mu,\sigma^2)$, $i = 1,2,\dots,n$

Step 1:

$$\mathscr{L}(\mu,\sigma^2;x_1,\dots,x_n) = (\frac{1}{\sqrt{2\pi}})^n\ exp\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\Big\}$$

hence $l\,(\mu,\sigma^2,x_1,\dots,x_n) = -\frac{n}{2}\log\,\sqrt{2\pi} - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum\limits_{i=1}^{n}(x_i-\mu)^2$

Step 2:

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

the MLE , $\hat{\mu}_{ML}$ & $\hat{\sigma^2}_{ML}$ are therefore solutions to:

$$\begin{cases} -\frac{1}{2\,\hat{\sigma}^2_{ML}}\sum_{i=1}^{n} -2(x_i-\hat{\mu}_{ML}) = 0 \\ -\frac{n}{2\,\hat{\sigma}^2_{ML}} + \frac{1}{2\,\hat{\sigma}^4_{ML}}\sum_{i=1}^{n}(x_i-\hat{\mu}_{ML}) = 0 \end{cases}$$

From the $1^{st} equation we find$:

$$\sum_{i=1}^{n}(x_i-\hat{\mu}_{ML}) = 0 \implies \sum x_i = n\hat{\mu}_{ML} \implies \boxed{\hat{\mu}_{ML} = \bar{x}_n}$$

If we plug in for $\hat{\mu}_{ML}$ in the second equation, we find

$$-\frac{n}{\cancel{2}\cancel{\hat{\sigma}^2_{ML}}} + \frac{1}{\cancel{2}\hat{\sigma}^{\cancel{4}}_{ML}}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2 = 0$$
$$\implies \qquad \frac{1}{\hat{\sigma}^2_{ML}}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2 = n$$
$$\implies \qquad \boxed{\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2}$$

To show that this is a maximizer we should check that:

$$\begin{bmatrix} \frac{\partial^2 l}{\partial\mu^2} & \frac{\partial^2 l}{\partial\mu\partial\sigma^2} \\ \frac{\partial^2 l}{\partial\sigma^2\partial\,\mu} & \frac{\partial^2 l}{\partial(\sigma^2)^2} \end{bmatrix} \text{ is a negative-definite matrix .}$$

This is not a hard task, but it is not required in this course. We only check the 2^{nd} derive for cases that there is only one unknown parameter.

Example. $X_i \overset{iid}{\sim} Unif(0,\theta)$, $i = 1,2,\ldots,n$

$$f_\theta(x) = \begin{cases} \frac{1}{\theta} & 0 < x < \theta \\ 0 & \text{otherwise} \end{cases}$$

In other words, $f_\theta(x) = \frac{1}{\theta} \cdot I_{[0,\theta]}(x)$ where $I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$ **Step 1:**

$$\mathscr{L}(\theta; x_1 1, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i) = \frac{1}{\theta^n} \prod_{i=1}^{n} I_{[0,\theta]}(x_i)$$

Now note that

$$\prod_{i=1}^{n} I_{[0,\theta]}(x_i) = I_{[0,\theta]}(\max_{1\le i\le n} x_i)$$

Since $0 \le x_i \le \theta$, $i = 1, 2, \ldots, n$ if and only if $0 \le \max\limits_{1\le i\le n} x_i \le \theta$. Therefore:

$$\mathscr{L}(\theta; x_1, \ldots, x_n) = \frac{1}{\theta^n} I_{[0,\theta]}(\max_{1\le i\le n} x_i)$$

Step 2: We note that the likelihood is a monotone decreasing function of θ. That is why, the max of $\mathscr{L}(\theta; x_1, \ldots, x_n)$ happens when θ takes its smallest possible value. Since that $0 \le \max\limits_{1\le i\le n} x_i \le \theta$, the smallest value for θ is $\mathop{max}\limits_{1\le i\le n} x_i$, thus $\hat{\theta}_{ML} = \mathop{max}\limits_{1\le i\le n} x_i$.

The method of *ML* has both the intuitive and theoretical appeal.

Intuitive Appeal: The *ML* method is based on the idea that *"What I have observed is what should have expected to observe"*. In other words, we observe the most likely scenario. Now given a sample, we choose the unknown parameter such that what we have observed has its maximum possible chance.

Theoretical Appeal:

* **Consistency:** *MLE* is a consistent estimator under rather mild conditions.

* **Asymptotic Normality:** $\sqrt{n}(\hat{\theta}_{ML} - \theta) \overset{app}{\sim} N(0, I^{-1}(\theta))$ for large n where $I(\theta) = \mathbb{E}\left\{[\frac{\partial}{\partial\theta} \log f_\theta(X)]^2\right\}$, the Fisher information.

  We can therefore make confidence interval for θ easily if we use $\hat{\theta}_n = \hat{\theta}_{ML}$ as the estimator.

* **Asymptotic Efficiency:** *MLE* is the most concentrated estimator about its estimand among a considerably large class of reasonable estimators called

"regular estimators".

* **Invariance:** If $\hat{\theta}_{ML}$ is the *MLE* of θ , then $g(\hat{\theta}_{ML})$ is the *MLE* of $g(\theta)$.

  This property simplifies life a bit. Since if we find the *MLE* of θ , then we have found *MLE* of any function of θ.

Example. $X_i \overset{iid}{\sim} Bernoulli(p)$, $i = 1, 2, \ldots, n$

We showed that $\hat{p}_{ML} = \hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a *MLE* of p. Now suppose that we want to find the *MLE* of $p(1-p)$, which is the variance of Bernoulli(p) . Using the Invariance, the *MLE* of $p(1-p)$ is $\hat{p}_{ML}(1-\hat{p}_{ML}) = (\frac{\sum_{i=1}^{n} X_i}{n})(1 - \frac{\sum_{i=1}^{n} X_i}{n})$

11.2 Method of Moments

Let $\mu_K = \mathbb{E}(X^K)$ and $m_K = \frac{1}{n}\sum_{i=1}^{n} X_i^K$. Now suppose $X_i \overset{iid}{\sim} f_{\underset{\sim}{\theta}}$, $i = 1, 2 \ldots, n$ where $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_r)$. Clearly μ_K is going to be a function of $\underset{\sim}{\theta}$. Suppose μ_K exists. Then the method of moments estimators are the solution to the following equation:

$$\mu_K(\underset{\sim}{\theta}) = m_K \quad , \quad K = 1, 2, \ldots, r \quad .$$

Example. $X_i \overset{iid}{\sim} \Gamma(\alpha, \beta)$, $i = 1, 2, \ldots, n$

$$\mathbb{E}(X) = \alpha\beta \qquad , \qquad Var(X) = \alpha\beta^2$$

Thus $\mathbb{E}(X^2) = Var(X) + [\mathbb{E}(X)]^2 = \alpha\beta^2 + \alpha^2\beta^2 = \alpha\beta(\beta + \alpha\beta)$

Now: $\begin{cases} \alpha\beta = \bar{X}_n \\ \alpha\beta(\beta + \alpha\beta) = \frac{1}{n}\sum_{i=1}^{n} X)i^2 \end{cases}$

Plugging-in from the 1^{st} equation into the 2^{nd} equation we find:

$$\bar{X}_n(\beta + \bar{X}_n) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \qquad \text{hence} \qquad \boxed{\beta = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2}{\bar{X}_n} - \bar{X}_n}$$

Using the first equation:

$$\boxed{\alpha = \frac{\bar{X}_n}{\beta} = \bar{X}_n[\frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2}{\bar{X}_n} - \bar{X}_n]^{-1}}$$

12 Lecture 12

12.1 Section 9.8 - Large Sample Property of the MLEs

Let $X \sim f_\theta(x)$. Suppose we have two observations from f_θ , $x_1 = 2$ and $x_2 = 5$. We want to see how we can quantify the amount of information in each of these two observations about θ. Note that our only link between the observations, i.e. X, and unobservable , .i.e. θ , is $f_\theta(x)$. So this is the channel through which information are transmitted. Now suppose the following figures depict the graph of $f_\theta(2)$ and $f_\theta(5)$:

#MISSING GRAPHS - Lecture 12 P1

Since that $f_\theta(2)$ is constant, it is not sensitive w.r.t changes in θ . In contrast $f_\theta(5)$ seems to respond to the changes in θ values. As such 5 is much more information about θ than 2 . In fact, 2 does not have any information about θ . Such sensitivity can be measured by derivative, i.e. $\frac{\partial}{\partial\theta} f_\theta$. Now Fisher would believe that information should increase linearly with the sample size n . As such, he would work with $\log f_\theta$, rather that f_θ . This then leads us to $\frac{\partial}{\partial\theta} \log f_\theta$. Now to define the information in the random variable X about θ , we can look at the (Eucledean) length of $\frac{\partial}{\partial\theta} \log_\theta(x)$ which is $\mathbb{E}\Big\{[\frac{\partial}{\partial\theta} \log f_\theta(x)]^2\Big\}$

$$I(\theta) = \mathbb{E}\Big\{[\frac{\partial}{\partial\theta} \log f_\theta(x)]^2\Big\} = \int_\theta [\frac{\partial}{\partial\theta} \log f_\theta]^2 \, f_\theta(x)\, dx$$

is called the *Fisher Information Amount* . Note that :

$$\mathbb{E}\Big\{[\frac{\partial}{\partial\theta}\log f_\theta(x)]\Big\} = \int_x \frac{\partial}{\partial\theta}\log f_\theta(x)\cdot f_\theta(x)\,dx$$

$$= \int_x \frac{\dot{f}_\theta(x)}{\cancel{f_\theta(x)}}\cdot \cancel{f_\theta(x)}\,dx \qquad \text{s.t} \qquad \dot{f}_\theta(x) = \frac{\partial}{\partial\theta}f_\theta(x)$$

$$= \frac{\partial}{\partial\theta}\int_x \overbrace{f_\theta(x)dx}^{1} = 0$$

Provided that $\quad \displaystyle\frac{\partial}{\partial\theta}\int_x f_\theta(x)dx = \int_x \frac{\partial}{\partial\theta}f_\theta(x)dx \qquad (1)$

Thus: $\quad \displaystyle I(\theta) = Var\{\frac{\partial}{\partial\theta}\log f_\theta(x)\}$

Taking the 2^{nd} derivative we have:

$$\frac{\partial^2}{\partial\theta^2}\log f_\theta(x) = \frac{\partial}{\partial\theta}\Big[\frac{\dot{f}_\theta(x)}{f_\theta(x)}\Big] = \frac{\dot{f}_\theta(x)f_\theta(x) - [\ddot{f}_\theta(x)]^2}{[f_\theta(x)]^2}$$

where $\quad \displaystyle\ddot{f}_\theta(x) = \frac{\partial^2}{\partial\theta^2}f_\theta(x)$. Thus

$$\frac{\partial^2}{\partial\theta^2}\log f_\theta(x) = \frac{\ddot{f}_\theta(x)}{f_\theta(x)} - \Big[\frac{\dot{f}_\theta(x)}{f_\theta(x)}\Big]^2$$

$$= \frac{\ddot{f}_\theta(x)}{f_\theta(x)} - \Big[\frac{\partial}{\partial\theta}\log f_\theta(x)\Big]^2$$

Now taking $\mathbb{E}$ from both sides, we have:

$$\mathbb{E}\Big[\frac{\partial^2}{\partial\theta^2}\log f_\theta(X)\Big] = \mathbb{E}\Big[\frac{\ddot{f}_\theta(X)}{f_\theta(X)} - I(\theta)\Big] \qquad (\dagger)$$

Note that:

$$\mathbb{E}\Big[\frac{\ddot{f}_\theta(X)}{f_\theta(X)}\Big] = \int_x \frac{\ddot{f}_\theta(x)}{\cancel{f_\theta(x)}}\cdot\cancel{f_\theta(x)}dx = \int_x \ddot{f}_\theta(x)dx$$

$$= \int_x \frac{\partial^2}{\partial\theta^2}f_\theta(x)dx$$

Now if: $\quad \displaystyle\frac{\partial^2}{\partial\theta^2}\int_x f_\theta(x)dx = \int_x \frac{\partial^2}{\partial\theta^2}f_\theta(x)dx \qquad (2)$

We obtain: $\quad \displaystyle\mathbb{E}\Big[\frac{\ddot{f}_\theta(x)}{f_\theta(x)}\Big] = \int_x \frac{\partial^2}{\partial\theta^2}f_\theta(x)dx = \frac{\partial^2}{\partial\theta}\overbrace{\int_x f_\theta(x)dx}^{1} = 0$

Using ($\dagger$) we therefore have:

$$\mathbb{E}\Big[\frac{\partial^2}{\partial\theta^2}\log f_\theta(X)\Big] = -I(\theta)$$

and hence: $\quad \boxed{I(\theta) = \mathbb{E}\Big[\Big\{\frac{\partial}{\partial\theta}\log f_\theta(X)\Big\}^2\Big] = -\mathbb{E}\Big[\frac{\partial^2}{\partial\theta^2}\log f_\theta(X)\Big]}$

Provided that (2) holds. Note that a necessary condition for (2) is that supp $f_\theta(x) = \{x : f_\theta(x) > 0\}$ does not depend on θ . This is the case for most distributions you have seen, i.e. Bernoulli, Binomial, Poisson, Exponential, Gamma, χ^2, Normal. This is **is not**, however the case for **uniform distribution** .

12.2 Asymptotic Distribution of *MLE* and Approximate Pivotal Quantity

We discussed properties of the *MLE*s . We leaned the invariance property of *MLE* which says that if $\hat{\theta}_{ML}$ is the *MLE* of θ , then $\tau(\hat{\theta}_{ML})$ is the *MLE* of $\tau(\theta)$. Under some mild conditions:

$$\sqrt{n}\Big(\tau(\hat{\theta}_{ML}) - \tau(\theta)\Big) \overset{app}{\sim} N\Big(0, \Big[\frac{\partial}{\partial\theta}\tau(\theta)\Big]^2 I^{-1}(\theta)\Big) \qquad \text{for large } n \qquad (*)$$

In particular when $\tau(\theta) = \theta$, we have:

$$\sqrt{n}\Big(\tau(\hat{\theta}_{ML}) - \theta\Big) \overset{app}{\sim} N\Big(0, I^{-1}(\theta)\Big) \qquad (**)$$

Note that $\Big[\frac{\partial}{\partial\theta}\tau(\theta)\Big]^2 I^{-1}(\theta)$ is the amount of information in variable X about $\eta = \tau(\theta)$. In fact:

$$\begin{aligned} I(\eta) &= \mathbb{E}\Big[\Big\{\frac{\partial}{\partial\theta}\log f_\theta(X)\Big\}^2\Big] = \mathbb{E}\Big[\Big\{\frac{\partial}{\partial\theta}\log f_\theta(X)\cdot\frac{\partial\theta}{\partial\eta}\Big\}^2\Big] \\ &= \Big[\frac{\partial\theta}{\partial\eta}\Big]^2 \cdot \mathbb{E}\Big[\Big\{\frac{\partial}{\partial\theta}\log f_\theta(X)\Big\}^2\Big] \\ &= \Big[\frac{\partial\theta}{\partial\eta}\Big]^2 \cdot I(\theta) \end{aligned}$$

Since $\frac{\partial\theta}{\partial\eta}$ is not a function of X . Thus:

$$I^{-1}(\eta) = \Big[\frac{\partial\eta}{\partial\theta}\Big]^2 I^{-1}(\theta) = \Big[\frac{\partial\tau(\theta)}{\theta}\Big]^2 I^{-1}(\theta)$$

using the above result:

$$\frac{\sqrt{n}\Big(\tau(\hat{\theta}_{_{ML}} - \tau(\theta))\Big)}{\sqrt{\Big[\frac{\partial}{\partial\theta}\tau(\theta)\Big]^2 I^{-1}(\theta)}} \overset{app}{\sim} N(0,1)\ ,\ (\mathscr{L}) \text{ for large } n$$

For $(*)$ **and** $(**)$: When we develop theory for the maximum likelihood estimators, we show that under same mild condition:

$$\sqrt{n}\Big(\hat{\theta}_{_{ML}} - \theta\Big) \overset{app}{\sim} N\Big(0, I^{-1}(\theta)\Big)$$

Now for $\eta = \tau(\theta)$ we have the same result, i.e.:

$$\sqrt{n}\Big(\hat{\eta}_{_{ML}} - \eta\Big) \overset{app}{\sim} N\Big(0, I^{-1}(\eta)\Big)\ .$$

Then we use the invariance property to conclude that $\hat{\eta}_{_{ML}} = \tau(\hat{\theta}_{_{ML}})$, if further, η is continuously differentiable, then:

$$I^{-1}(\eta) = \Big[\frac{\partial}{\partial\theta}\tau(\theta)\Big]^2 I^{-1}(\theta)$$

Thus $\tau(\hat{\theta}_{_{ML}}) \pm \zeta_{\frac{\alpha}{2}}\sqrt{\frac{[\dot{\tau}(\theta)]^2\ I^{-1}(\theta)}{n}}$ is a $100(1-\alpha)\%$ confidence interval for $\tau(\theta)$, where $\dot{\tau}(\theta) = \frac{\partial}{\partial\theta}\tau(\theta)$. For practical purposes the margin of error should be estimated.

As long as the estimator is consistent , the asymptotic confidence interval is still valid. Now if τ is continuously differentiable, $\dot{\tau}(\hat{\theta}_{_{ML}})$ is a consistent estimator of $\dot{\tau}(\theta)$ by continuous mapping theorem. Under mild conditions $I(\theta)$ is a continuous function of θ , and hence $I(\hat{\theta}_{_{ML}})$ is a consistent estimator of $I(\theta)$ by the continuous mapping theorem. Thus:

$$\tau(\hat{\theta}_{_{ML}}) \pm \zeta_{\frac{\alpha}{2}}\ \sqrt{\frac{\dot{\tau}(\hat{\theta}_{_{ML}})I^{-1}(\hat{\theta}_{_{ML}})}{n}} \qquad (\$)$$

is an approximate $100(1-\alpha)\%$ confidence interval for $\tau(\theta)$, where $\zeta_{\frac{\alpha}{2}}$ is chosen such that $P(Z > \zeta_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$, $Z \sim N(0,1)$.

Example 12.1 (# 9.14 - page 484).

$$Y_i \overset{iid}{\sim} Ber(p)\ ,\ i = 1,2,\ldots,n\ ,\ \tau(p) = p(1-p)\ ,\ \dot{\tau}(p) = \frac{\partial}{\partial p}\tau(p) = 1-2p \text{ and } P_p(Y = y) = p^y(1-p)^{1-y}\ ,\ y = 0,1$$

$$\begin{aligned}
\log P_p(Y = y) &= y \log p + (1-y)\log(1-p) \\
\frac{\partial}{\partial p}\log P_p(Y = y) &= \frac{y}{p} - \frac{1-y}{1-p} \\
\frac{\partial^2}{\partial p^2}\log P_p(Y = y) &= -\frac{y}{p^2} - \frac{1-y}{(1-p)^2}
\end{aligned}$$

Now we have:

$$\mathbb{E}\Big\{-\frac{\partial^2}{\partial p^2}\log P_p(Y = y)\Big\} = I(p)$$

$$\begin{aligned}
I(p) &= \mathbb{E}\Big\{\frac{Y}{p^2} + \frac{1-Y}{(1-p)^2}\Big\} = \frac{\mathbb{E}(Y)}{p^2} + \frac{\mathbb{E}(1-Y)}{(1-p)^2} \\
&= \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}
\end{aligned}$$

Thus: $\quad I^{-1}(p) = p(1-p)\quad ,\ \hat{p}_{ML} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

Using ($) , we have:

$$\hat{p}_{ML}(1-\hat{p}_{ML}) \pm \zeta_{\frac{\alpha}{2}}\sqrt{\frac{(1-2\hat{p}_{ML})^2\,\hat{p}_{ML}(1-\hat{p}_{ML})}{n}}$$

is a $100(1-\alpha)\%$ confidence interval for p(1-p) .

12.3 Chapter 10 - Testing Statistical Hypothesis

A statistical hypothesis is a statement about the parameter(s) of a population that specifies , partially or totally , structure of the population . A statistical hypothesis is tested against observations from the population that the hypothesis describes.

Remark. The formation of a hypothesis should be independent from the data used to test the hypothesis. We should not use the same data for both *exploration* (formation of a hypothesis) and *validation* of a theory or hypothesis.

The essential elements of TSH are:

* A null hypothesis ($\mathcal{H}_0$)

* An alternative hypothesis ($\mathcal{H}_A$)

* A test statistic

* A rejection region

When testing a hypothesis $\mathcal{H}_0$ two types of errors may happen, Type I and Type II . The following table shows thus errors:

$\mathcal{H}_0$:		$\mathcal{H}_0$: True	$\mathcal{H}_0$: False
	Accept	✓	Type II error
	Reject	Type I error	✓

Probability of Type I & Type II error are respectively denoted by α and β , i.e. :

$$\begin{aligned}\alpha &= P(\text{Type I error})\\ &= P(\text{Rejecting } \mathcal{H}_0 \text{ when } \mathcal{H}_0 \text{ is true})\\ &= P_{\mathcal{H}_0}(\text{Rejecting } \mathcal{H}_0)\end{aligned}$$

and:

$$\begin{aligned}\beta &= P(\text{Type II error})\\ &= P(\text{Accepting } \mathcal{H}_0 \text{ when } \mathcal{H}_0 \text{ is false})\\ &= P_{\mathcal{H}_A}(\text{Accepting } \mathcal{H}_0)\end{aligned}$$

Example 12.2 (# 10.1 , page 491). $X_i \overset{iid}{\sim} Ber(p)$, $i = 1, 2, \ldots, \overbrace{15}^{n}$, $Y = \sum_{i=1}^{15} X_i$

$\mathcal{H}_0 : p = 0.5$ vs $\mathcal{H}_A : p > 0.5$

Rejection Region: $(RR) = \{Y \leq 2\}$

$\alpha = P(Type\ I\ error) = P_{p=0.5}(Y \leq 2)$

$= \sum_{y=0}^{2} \binom{15}{y}(\frac{1}{2})^y(1-\frac{1}{2})^{15-y} \approx 0.004$ *(using table* 1 *in App* 3*)*

Example 12.3 (example 10.2 , page 492). $X_i \overset{iid}{\sim} Ber(p)$, $i = 1, 2, \ldots, \overbrace{15}^{n:\ sample\ size}$

$\mathcal{H}_0 : p = 0.5 \qquad vs \qquad \mathcal{H}_A : p = 0.3$

$RR = \{Y \leq 2\}$

$\beta = P(Type\ II\ error) = P_{\mathcal{H}_0}(Y > 2)$

$= P_{p=0.3}(Y > 2) = \sum_{y=3}^{15} \binom{15}{y} (0.3)^y (1 - 0.3)^{15-y}$

$= 1 - \sum_{y=0}^{2} \binom{15}{y} (0.3)^y (1 - 0.3)^{15-y} \approx 0.873 \qquad (using\ Table\ 1\ in\ App\ 3)$

For $\mathcal{H}_A : p = 0.1$ we find $\beta \approx 0.184$ using similar calculation.

13 Lecture 13

Question: How do we find an appropriate test statistic and specify an rejection region?

As mentioned in the previous lecture, statistical hypotheses are often statements about the parameters of the population of interest, i.e. θ equal something or belong to some set. As test statistic we choose an estimator of θ , often the *MLE* . As for the rejection region (RR) or set, the form of the set is determined by $\mathcal{H}_A$. For example for testing $\mathcal{H}_0 : \theta = \theta_0$ versus $\mathcal{H}_A : \theta > \theta_0$, the rejection region (or rejection set) is $\left\{(x_1, \dots, x_n) : \hat{\theta}_{ML} = \hat{\theta}(x_1, \dots, x_n) > K\right\}$ when K is determined in such a way that:

$$\alpha = P_{\mathcal{H}_0}(\text{Type I error}) = P_{\theta=\theta_0}(\hat{\theta}_{ML} > K)$$

The following two tables summarize the the above discussion for testing $\mathcal{H}_0 : \theta = \theta_0$, for most commonly encountered parameters and most common forms of $\mathcal{H}_A$:

Table I :

θ	μ	p	σ^2	$\mu_1 - \mu_2$	$\frac{\sigma_1^2}{\sigma_2^2}$
$\hat{\theta}$	$\bar{X}_n$	$\hat{p}_n$	S^2	$\bar{X}_1 - \bar{X}_2$	$\frac{S_1^2}{S_2^2}$

and

Table II :

$\mathcal{H}_A$	$\theta > \theta_0$	$\theta < \theta_0$	$\theta \neq \theta_0$
RR	$\hat{\theta} > K$	$\hat{\theta} < K$	$\lvert\hat{\theta}\rvert > K$

where K is chosen such that $\alpha = P_{\theta=\theta_0}(\text{Type I error})$, α is the significance level .

How to determine K:

As we said above, K is determined such that $P_{\theta=\theta_0}(\text{Type I error}) = \alpha$, thus:

$\mathcal{H}_A : \theta > \theta_0 \rightarrow P_{\theta=\theta_0}(\hat{\theta} > K) = \alpha$

$\mathcal{H}_A : \theta < \theta_0 \rightarrow P_{\theta=\theta_0}(\hat{\theta} < K) = \alpha$

$\mathcal{H}_A : \theta \neq \theta_0 \rightarrow P_{\theta=\theta_0}(\lvert\hat{\theta}\rvert > K) = \alpha$

We therefore need the distribution of $\hat{\theta}$. Now we have two cases:

A: We can find the exact distribution of $\hat{\theta}$

B: The exact distribution cannot be found

As we will see soon, if θ is equal to μ ,σ^2 , $\mu_1 - \mu_2$ or $\frac{\sigma_1^2}{\sigma_2^2}$ and observations are coming from a normal distribution , then the exact distribution of $\hat{\theta}$ can be found , under some additional conditions in *Two-Sample* cases , i.e. $\mu_1 - \mu_2$ or $\frac{\sigma_1^2}{\sigma_2^2}$.

When the exact distribution of $\hat{\theta}$ is not available, we try to approximate it. Our main tools in such cases are ($\mathscr{L}$) on page 81 that presents the asymptotic distribution of functions of *MLE* and the General Limit Theorem (*GLT*).

$$\frac{\sqrt{n}\Big(\tau(\hat{\theta}_{_{ML}} - \tau(\theta))\Big)}{\sqrt{\Big[\frac{\partial}{\partial\theta}\tau(\theta)\Big]^2 I^{-1}(\theta)}} \overset{app}{\sim} N(0,1) \ : \qquad (\mathscr{L}) \text{ for large } n$$

13.1 Large Sample Tests

Example 13.1 (#10.5 , page 497)**.**

$n = 36$, $\bar{X}_{36} = 17$, $S^2_{36-1} = 9$, $\mathcal{H}_{_0} : \mu = 15$, $\mathcal{H}_{_A} : \mu > 15$,$\alpha = 0.05$

- Using Table I, our test statistic is $\bar{X}_n$

- Using Table II, our $RR = \{\bar{X}_n > K\}$

We should choose K such that :

$$\begin{aligned} P_{\mu=15}(\bar{X}_{36} > K) &= 0.05 \qquad \text{then:} \\ &= P_{\mu=15}\Big(\frac{\bar{X}_{36} - 15}{\frac{\sigma}{\sqrt{36}}} > \frac{K-15}{\frac{\sigma}{\sqrt{6}}}\Big) \qquad \text{replace } \sigma^2 \text{ by } S^2 = 9 \\ &= P_{\mu=15}\Big(\frac{\bar{X}_{36} - 15}{\frac{3}{\sqrt{36}}} > \frac{K-15}{\frac{3}{\sqrt{36}}}\Big) \\ &= P_{\mu=15}(Z > 2(K-15)) \end{aligned}$$

Under $\mathcal{H}_{_0} : \mu = 15$, $Z = \frac{\bar{X}_{36}-15}{\frac{3}{\sqrt{36}}} \overset{app}{\sim} N(0,1)$. Using Table IV , App3 on page 848 , we have : $P(z > 1.645) \approx 0.05$ #MISSING GRAPH page 87

Thus $2(K-15) = 1.645$ and hence $\boxed{K \approx 15.82}$. The *RR* is then $RR = \{\bar{X}_{36} > 15.82\}$. In our example, the observed value of $\bar{X}_{36}$ is 17 which is greater than 15.82 . Thus the data do **NOT** support $\mathcal{H}_{_0}$ and hence $\mathcal{H}_{_0}$ is **rejected** .

Example 13.2 (#10.6 , page 498).

$\mathcal{H}_0 : p = 0.1$, $\mathcal{H}_A : p > 0.1$, $\alpha = 0.01$, $n = 100$, $\hat{p}_n = \hat{p}_{ML} = \frac{1}{n}\sum_{1}^{n} X_i = \frac{1}{100}\sum_{1}^{100} = \frac{15}{100} = 0.15$

where: $X_i = \begin{cases} 1 & \textit{if the } i^{th} \textit{ item is defective} \\ 0 & \textit{otherwise} \end{cases}$

$X_i \overset{iid}{\sim} Bern(p)$, $\imath = 1, 2 \ldots, n$ *where* $n = 100$

Using Table I , our test statistic is $\hat{p}_{ML}$ and using Table II $RR = \{\hat{p}_{ML} > K\}$.

Now:

$$
\begin{aligned}
P_{\mu_0}(\hat{p}_{ML} < K) &= P_{\mathcal{H}_0}\Big(\frac{\hat{p}_{ML} > 0.1}{\sigma_{\hat{p}_{ML}}} > \frac{K - 0.1}{\sigma_{\hat{p}_{ML}}}\Big) \\
&= P_{\mathcal{H}_0 : p=0.1}\Big(\frac{\hat{p}_{ML} > 0.1}{\sqrt{\frac{p(1-p)}{n}}} > \frac{K - 0.1}{\sqrt{\frac{0.1(1-0.1)}{100}}}\Big) \\
&\approx P_{p=0.1}\Big(Z > \frac{K - 0.1}{\sqrt{\frac{0.1(1-0.1)}{100}}}\Big)
\end{aligned}
$$

where $Z \sim N(0, 1)$. Using Table IV in App3 , we have $P(Z > 2.33) \approx 0.01$.

Thus:

$$
\frac{K - 0.1}{\sqrt{\frac{0.1(1-0.1)}{100}}} = 2.33 = \boxed{K = 0.1699}
$$

Now the observed value of $\hat{p}_{ML}$ is 0.15 which is smaller than $K(= 0.1699)$, hence $\mathcal{H}_0$ cannot be rejected.

Example 13.3 (#10.8 , page 498).

$n = 36$, $\bar{X}_{36} = 17$, $S^2 = 9$, $\mathcal{H}_0 : \mu = 15$, $\alpha = 0.05$

From example #10.5 $RR = \{\bar{X}_{36} > 15.82\}$. Now:

$$
\begin{aligned}
\beta &= P_{H_A}\Big(\frac{\bar{X}_{36} - \overbrace{16}^{\mu_A}}{\frac{3}{\sqrt{36}}} \leq \frac{15.82 - 16}{\frac{3}{\sqrt{36}}}\Big) \\
&\approx P_{\mu=16}\big(Z \leq -0.36\big) \qquad \text{, where } Z \sim N(0, 1) \\
&\approx 0.36
\end{aligned}
$$

13.2 Sample Size Calculation using Type I & Type II errors

We can determine sample size to control both types of error for testing $\mathcal{H}_A : \mu = \mu_0$ versus $\mathcal{H}_A : \mu = \mu_A$.

$$-\alpha = P_{\mu_0}(\bar{X}_n > K) = P_{\mu_0}\Big(\frac{\bar{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} > \frac{K - \mu_0}{\frac{\sigma}{\sqrt{n}}}\Big)$$

$$\approx P_{\mu_0}(Z > \zeta_\alpha)\ ,\ Z \sim N(0,1) \qquad \text{(for large } n\text{)}$$

Likewise

$$-\beta = P_{\mu_A}(\bar{X}_n \leq K) = P_{\mu_A}\Big(\frac{\bar{X}_n - \mu_A}{\frac{\sigma}{\sqrt{n}}} \leq \frac{K - \mu_A}{\frac{\sigma}{\sqrt{n}}}\Big)$$

$$\approx P_{\mu_A}(Z < -\zeta_\beta)\ ,\ Z \sim N(0,1) \qquad \text{(for large } n\text{)}$$

Thus: $$\begin{cases} \frac{K-\mu_0}{\frac{\sigma}{\sqrt{n}}} = \zeta_\alpha \\ \frac{K-\mu_A}{\frac{\sigma}{\sqrt{n}}} = -\zeta_\beta \end{cases} \Longrightarrow \qquad K = \mu_0 + \zeta_\alpha \frac{\sigma}{\sqrt{n}} = \mu_A - \zeta_\beta \frac{\sigma}{\sqrt{n}}$$

and hence $$\mu_A - \mu_0 = \frac{\sigma}{\sqrt{n}}(\zeta_\alpha + \zeta_\beta)$$

therefore $$\boxed{n = \frac{(\zeta_\alpha + \zeta_\beta)^2\sigma^2}{(\mu_A - \mu_0)^2}} \qquad (\eta)$$

Example 13.4 (#10.9 , page 510)**.**

$\mu_0 : \mu = \overbrace{15}^{\mu_0}\ , \mathcal{H}_A : \mu = \overbrace{16}^{\mu_A}\ , \alpha = \beta = 0.05\ ,\ \sigma^2 = 9$

Thus $\zeta_\alpha = \zeta_\beta = 1.645$, $n = \frac{(1.645+1.645)^2 \cdot 9}{(16-15)^2} = 97.4 \to 98$

13.3 p-value (Fisher's Presentation)

What we learned about *TSH* so far is Nyman-Pearson approach to *TSH* . Fisher has a different perspective.

For Fisher, a hypothesis is a theory and to test the theory, we weigh the theory against the observations. To this end, we can, for example see how likely it is to see what we have observed if the hypothesis is valid. There are two technical issues here:**First** , when we are dealing with continuous random variables, the probability of any set of observations $\underset{\sim}{x} = \{x_1, \ldots, x_n\}$ is zero.**Second** , this idea can be futile as the following example shows:

example $X_i \overset{iid}{\sim} \text{Ber}(p), i = 1, 2, \ldots, n$ and $\mathcal{H} : p = \frac{1}{2}$

$$P_{\mathcal{H}:p=\frac{1}{2}}(X_i = x_i, i = 1, 2, \ldots, n) = \prod_{i=1}^{n} (\frac{1}{2})^{x_i} (1 - \frac{n}{2})^{1-x_i} = (\frac{1}{2})^n$$

Thus all the observations are equally likely. Note that if $\mathcal{H}_0$ is true, we expect to observe 50% $1's$ and 50% $0's$, more or less, at least for large n, i.e. the chance od such observation should be high.

Fisher then suggested to compute the probability of observations which are as bad as what we have observed for the hypothesis that is to be tested. This lid to the notation of *p-value*:

p-value: The p-value or attained significance level, is the smallest level of significance α for which the obtained data indicate that the null hypothesis should be rejected.

Example 13.5 (#10.7 , page 500 and #10.11 , page 515)**.**

$\alpha = 0.05, n_1 = 50, n_2 = 50, \bar{Y}_1 = 3.6\, sec, \bar{Y}_2 = 3.8\, sec, S_1^2 = 0.18, S_2^2 = 0.14, \mathcal{H}_0 : \mu_1 - \mu_2 = 0, \mathcal{H}_A : \mu_1 - \mu_2 \neq 0$

Using Table I , the test statistic is $\bar{Y}_1 - \bar{Y}_2$

Using Table II, $RR = \left\{|\bar{Y}_1 - \bar{Y}_2| > K\right\}$

$$
\begin{aligned}
0.05 = \alpha &= P_{\mathcal{H}_0}\Big(|\bar{Y}_1 - \bar{Y}_2| > K\Big) \\
&= P_{\mathcal{H}_0:\mu_1-\mu_2=0}\Big(\Big|\frac{(\bar{Y}_1 - \bar{Y}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\Big| > \frac{K}{\sqrt{\frac{0.18}{50} + \frac{0.14}{50}}}\Big) \\
&\approx P\Big(|Z| > \frac{K}{\sqrt{\frac{0.18}{50} + \frac{0.14}{50}}}\Big) \qquad , \qquad Z \sim N(0,1)
\end{aligned}
$$

Thus:

$$\frac{K}{\sqrt{\frac{0.18}{50} + \frac{0.14}{50}}} = 1.96$$ Using Table IV in App3 and hence: $\boxed{K = 0.1568}$

$RR = \Big\{|\bar{Y}_1 - \bar{Y}_2| > 0.1568\Big\}$. Given that $\bar{y}_1 - \bar{y}_2 = 0.2$, $\mathcal{H}_0$ is rejected.

Now to computer the p-value :

$$
\begin{aligned}
p - value &= P_{\mathcal{H}_0}(|\bar{Y}_1 - \bar{Y}_2| > |3.6 - 3.8|) \\
&= P_{\mathcal{H}_0}(|\bar{Y}_1 - \bar{Y}_2| > 2) \\
&= P_{\mathcal{H}_0}\Big(\Big|\frac{(\bar{Y}_1 - \bar{Y}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\Big| > \frac{2}{\sqrt{\frac{0.18}{50} + \frac{0.14}{50}}}\Big) \\
&\approx P_{\mathcal{H}_0}(|Z| > 2.5) \qquad , \qquad Z \sim N(0,1) \\
&= 2P(Z > 2.5) = 2(0.0062) \\
&= 0.0124
\end{aligned}
$$

14 Lecture 14

14.1 Relationship between *TSA* and *Confidence Intervals*

A $100\%(1-\alpha)$ confidence interval $\hat{\theta}_{ML} \pm \zeta_{\frac{\alpha}{2}} \sigma_{\hat{\theta}_{ML}}$ is essentially a $100\%(1-\alpha)$ acceptance region. Thus:

$$\overline{RR}_\alpha = \left\{ \hat{\theta}_{ML} : \quad \left| \frac{\hat{\theta}_{ML} - \theta_0}{\sigma_{\hat{\theta}_{ML}}} \right| \leq \zeta_{\frac{\alpha}{2}} \right\}$$

is the complement of an α-level RR , i.e:

$$RR_\alpha = \left\{ \hat{\theta}_{ML} : \quad \left| \frac{\hat{\theta}_{ML} - \theta_0}{\sigma_{\hat{\theta}_{ML}}} \right| > \zeta_{\frac{\alpha}{2}} \right\}$$

This means that RR_α is a rejection region for testing $\mathcal{H}_0 : \theta = \theta_0$ versus $\mathcal{H}_A : \theta \neq \theta_0$ at the α significance level.

This relationship is two-sided . In other words, any $100\%(1-\alpha)$ confidence interval can be translated into an acceptance region and hence its complement can provide an α-level RR . On the other hand, the complement of α-level RR can provide a $100\%(1-\alpha)$ C.I.

14.2 Small Sample Tests

- For μ (one sample case):

  Assumptions: $Y_1, \ldots, Y_n$ are *iid*

  $$Y_i \sim N(\mu, \sigma^2) \quad , \quad i = 1, 2, \ldots, n$$

  $$\mathcal{H}_0 : \mu - \mu_0 \qquad \text{versus} \qquad \mathcal{H}_0 : \begin{cases} \mu > \mu_0 & : \quad \text{(upper tail)} \\ \mu < \mu_0 & : \quad \text{(lower tail)} \\ \mu \neq \mu_0 & : \quad \text{(two-sided)} \end{cases}$$

  Test statistic $T = \frac{\bar{Y} - \mu_0}{\frac{s}{\sqrt{n}}} \sim T_{(n-1)}$ under $\mathcal{H}_0 : \mu = \mu_0$

  $$RR_\alpha : \begin{cases} t > t_{\alpha, n-1} & : \quad \text{(upper tail)} \\ t < -t_{\alpha, n-1} & : \quad \text{(lower tail)} \\ |\, t \,| > t_{\frac{\alpha}{2}, n-1} & : \quad \text{(two-sided)} \end{cases} \qquad t_\alpha \text{ is chosen such that } P(T > t_\alpha) = \alpha$$

Value of $t_{\alpha,n-1}$ are given in Table VI, App.3, page 849.

p-value $= P\Big(|\, T_{n-1} \,| > |\, t_{\text{obs.}} \,|\Big)$ for two sided $\mathcal{H}_A$ where $t_{\text{obs.}} = \frac{\bar{y}-\mu_0}{\frac{s}{\sqrt{n}}}$, $\bar{y}$ is the observed sample average.

- For $\mu_1 - \mu_2$ (two sample case)

  Assumption: X's and Y's are independent

  $$\begin{cases} X_i \overset{iid}{\sim} N(\mu_1, \sigma_1^2) & , \quad i = 1, 2, \ldots, n_1 \\ Y_j \overset{iid}{\sim} N(\mu_2, \sigma_2^2) & , \quad j = 1, 2, \ldots, n_2 \end{cases}$$
  $$\sigma_1^2 = \sigma_2^2$$

  Note that we need **between** and **within** independence both, i.e X-samples are independent from the Y-samples and from each other.

  $$\mathcal{H}_0 : \mu_1 - \mu_2 = D_0 \qquad \text{versus} \qquad \mathcal{H}_A : \begin{cases} \mu_1 - \mu_2 > D_0 & : \text{(upper tail)} \\ \mu_1 - \mu_2 < D_0 & : \text{(lower tail)} \\ \mu_1 - \mu_2 \neq D_0 & : \text{(two-sided)} \end{cases}$$

  Test statistic: $T = \frac{(\bar{X}_{n_1} - \bar{Y}_{n_2}) - D_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ where:

  $$s_p = \sqrt{\frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{n_1 + n_2 - 2}} \;,\; S_X^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X}_{n_1})^2 \;\text{ and }\; S_Y^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (Y_j - \bar{Y}_{n_2})^2$$

  $$\text{Under } \mathcal{H}_0 \;,\; T \sim T_{n_1+n_2-2} \text{ and } RR_\alpha : \begin{cases} t > t_{\alpha, n_1+n_2-2} & : \text{(upper tail)} \\ t < -t_{\alpha, n_1+n_2-2} & : \text{(lower tail)} \\ |\, t \,| > t_{\frac{\alpha}{2}, n_1+n_2-2} & : \text{(upper tail)} \end{cases}$$

  Values of t_α , $n_1 + n_2 - 2$ are given in Table VI , App.3 , page 849.

Example 14.1 (#10.14 & #10.15 , page 524)**.**

$n_1 = 9$, $\bar{y}_1 = 35.22sec$, $\sum_{i=1}^{9} (y_{i_1} - \bar{y}_1)^2 = 195.56$

$n_2 = 9$, $\bar{y}_2 = 31.56sec$, $\sum_{i=1}^{9} (y_{i_2} - \bar{y}_2)^2 = 160.22$

$\alpha = 0.05$, $\mathcal{H}_0 : \mu_1 - \mu_2 = 0$ *vs.* $\mathcal{H}_A : \mu_1 - \mu_2 \neq 0$

$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - D_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, $\quad S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{195.56+160.22}{9+9-2} \approx 22.24 \implies S_p \approx 4.716$

The observed t is: $\quad t_{obs} = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{4.716 \sqrt{\frac{1}{9} + \frac{1}{9}}} = \frac{35.22 - 31.56}{4.716 \sqrt{\frac{2}{9}}} \approx 1.56$

$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - D_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim T_\nu \qquad$ where $\nu = n_1 + n_2 - 2 = 9 + 9 - 2 = 16$

Using Table 5, App.3, page 849:

$t_{\frac{\alpha}{2},\nu} = t_{0.025,16} = 2.12$ #MISSING GRAPH - LECTURE 14 PAGE 95

$-2.12 < 1.56 < 2.12 \implies \mathcal{H}_0$ is <u>NOT</u> rejected.

<u>**For** σ^2 **:**</u> Assumption: $Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$

$$\mathcal{H}_0 : \sigma^2 = \sigma_0^2 \quad \text{versus} \quad \mathcal{H}_A : \begin{cases} \sigma^2 > \sigma_0^2 \\ \sigma^2 < \sigma_0^2 \\ \sigma^2 \neq \sigma_0^2 \end{cases}$$

$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{(n-1)}$ under $\mathcal{H}_0 : \sigma^2 = \sigma_0^2$: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$

$$RR_\alpha : \begin{cases} \chi^2 > \chi^2_{\alpha,\nu} \\ \chi^2 < \chi^2_{1-\alpha,\nu} \\ \chi^2 > \chi^2_{\frac{\alpha}{2},\nu} \quad \text{or} \quad \chi^2 > \chi^2_{1-\frac{\alpha}{2},\nu} \end{cases}$$

$\chi^2_{\alpha,\nu}$ is chosen such that $P(\chi^2 > \chi^2_{\alpha,\nu}) = \alpha$, $\nu = n - 1$

Value of $\chi^2_{\alpha,\nu}$ is given in Table 6, App.3, page 850. #MISSING TABLE - LECTURE 14 PAGE 96.

Example 14.2 (#10.16 , page 532)**.**

$n = 10$, $S^2 = 0.0003$, $\mathcal{H}_0 : \sigma^2 = 0.0002$, $\alpha = 0.05$, $\mathcal{H}_A : \sigma^2 > 0.0002$

$\chi^2_{obs} = \frac{(10-1)(0.0003)}{0.0002} = 13.5$

$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_9$, $\chi^2_{0.05,9} = 16.919$

$13.5 = \chi^2_{obs} < \chi^2_{0.05,9} = 16.919 \implies$ Thus $\mathcal{H}_0$ is NOT rejected.

15 Lecture 15

15.1 Small Sample Test : cont'd

- **For $\frac{\sigma_1^2}{\sigma_2^2}$ (Two sample case)**

  Assumption:

  - X's and Y's are independent
  - $X_i \overset{iid}{\sim} N(\mu_1, \sigma_1^2)\ \ i = 1, \ldots, n\ ,\ \ Y_j \overset{iid}{\sim} N(\mu_2, \sigma_2^2)\ \ ,\ \ j = 1, \ldots, n$

$$\mathcal{H}_0 : \sigma_1^2 = \sigma_1^2 \qquad vs \qquad \mathcal{H}_A : \begin{cases} \sigma_1^2 > \sigma_2^2 \\ \sigma_1^2 < \sigma_2^2 \\ \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

Note that we need both "between" and "within" independence ; i.e. $X_1, \ldots, X_{n_1}$ are independent of each other and also they are independent of $Y_1, \ldots, Y_{n_2}$.

The test statistic is:

$$F = \frac{S_X^2}{S_Y^2} = \frac{\frac{\frac{(n_1-1)S_X^2}{\sigma^2}}{(n_1-1)}}{\frac{\frac{(n_2-1)S_Y^2}{\sigma^2}}{(n_2-1)}} \sim F_{n_1-1, n_2-1} \qquad \text{where } \sigma^2 = \sigma_1^2 = \sigma_2^2 \text{ under } \mathcal{H}_0.$$

$$RR_\alpha : \begin{cases} F > F_{\alpha, v_1, v_2} \\ F < F_{1-\alpha, v_1, v_2} \\ F > F_{\frac{\alpha}{2}, v_1, v_2} \text{ or } F < F_{1-\alpha, v_1, v_2} \end{cases} \qquad v_i = n_i - 1\ \ ,\ \ i = 1, 2$$

where F_{α, v_1, v_2} is chosen such that: $P(F > F_{\alpha, v_1, v_2}) = \alpha$ and values of F_{α, v_1, v_2} are given in Table 7, App.3.

Example 15.1 (#10.21 , Page 536)**.**

Gender	n	$\bar{y}$	S^2
Men	14	16.2	12.7
Women	10	14.9	26.4

$$\begin{cases} \mathcal{H}_0 : \sigma_W^2 = \sigma_M^2 \\ \mathcal{H}_A : \sigma_W^2 \neq \sigma_M^2 \end{cases} \quad , \quad \alpha = 0.1$$

$F_{obs} = \frac{26.4}{12.7} \approx 2.08$

$F_{0.05, v_1, v_2} = 2.71\ ,\ v_1 = 10 - 1 = 9\ \ ,\ v_2 = 14 - 1 = 13$

$F_{0.95,9,13} < F_{obs} < F_{0.05,9,13}$

We cannot reject $\mathcal{H}_0$. #MISSING GRAPH - LECTURE 15 - Page 98

Remark: Notice that χ^2-distribution and F-distributions are not symmetric; Thus calculating p-values for two-sided alternative is tricky.

* χ^2

In One-Sample-case, i.e. χ^2 , we can find p-values for one side. To account for the other side, we multiply it by 2. The rational behind this approach is that we want to account for comparable evidence on the other side. In the symmetric case, the "comparability" has a clear interpretation. It is interpreted using the observed value of the test statistic; i.e. $|T| > t$. In the asymmetric cases, such as χ^2 , the interpretation is in terms of the same chance of observing evidence against the null hypothesis, $\mathcal{H}_0$.

** F

Following the same logic as above we multiply by 2 when we work with the F-distributions. In the F-distribution case we can also approach the p-value differently. Since that Table 7,App.3 provides the upper tail values, we usually put the larger of S_1^2 & S_2^2 in the numerator. Now using the fact that F^{-1} is also distributed accordingly to F , i.e. $F \sim F_{v_1,v_2} \implies F^{-1} \sim F_{v_2,v_1}$, we can also consider putting the smaller value of $\sigma_1^2 \& \sigma_2^2$ in the numerator and find an approximate p-value for the lower tail.

15.2 The Neyman-Pearson Lemma

Suppose that we wish to test the simple null hypothesis $\mathcal{H}_0 : \theta = \theta_0$ versus the simple alternative hypothesis $\mathcal{H}_A : \theta = \theta_A$, based on a random sample $Y_1, \ldots, Y_n$ for a distribution with parameter θ. Let $L(\theta)$ denote the likelihood of the sample when the value of the parameter is θ . Then, for a given α , the test that minimizes the probability of Type II error at $\theta = \theta_A$ has a rejection region , RR , determined by $\frac{L(\theta_0)}{L(\theta_A)} < K$.

The value of K is chosen so that the test has the desired probability of Type I error α , i.e. K is chosen such that:

$$P_{\mathcal{H}_0:\theta=\theta_0}\Big(\underset{\sim}{X} : \frac{L(\theta_0;\underset{\sim}{Y})}{L(\theta_A;\underset{\sim}{Y})} < K\Big) = \alpha \qquad \text{where } \underset{\sim}{Y} = (Y_1,\ldots,Y_n)$$

Remark: When Y_i's are continuous random variables for any α , we can find a K . This is not, however, the case in general for discrete random variables. In the general for of the Neyman-Pearson lemma $\mathcal{H}_0 : \theta = \theta_0$ is rejected if $\underset{\sim}{y}_{obs} = (y_1,\ldots,y_n)$ is such that:

$$\frac{\mathscr{L}(\theta_0,\underset{\sim}{y}_{obs})}{\mathscr{L}(\theta_A,\underset{\sim}{y}_{obs})} < K \qquad \text{or if} \qquad \frac{\mathscr{L}(\theta_0,\underset{\sim}{y}_{obs})}{\mathscr{L}(\theta_A,\underset{\sim}{y}_{obs})} = K$$

we reject $\mathcal{H}_0 : \theta = \theta_0$ with probability λ where K and λ are chosen as follows:

Step 1: Choose K for a given α such that:

$$P_{\mathcal{H}_0}\Big(\frac{\mathscr{L}(\theta_0,\underset{\sim}{Y})}{\mathscr{L}(\theta_A,\underset{\sim}{Y})} < K\Big) \le \alpha < P_{\mathcal{H}_0}\Big(\frac{\mathscr{L}(\theta_0,\underset{\sim}{Y})}{\mathscr{L}(\theta_A,\underset{\sim}{Y})} \le K\Big) \qquad (\dagger)$$

Step 2: Choose λ such that:

$$\gamma = \frac{\alpha - P_{\mathcal{H}_0}\Big(\frac{\mathscr{L}(\theta_0,\underset{\sim}{Y})}{\mathscr{L}(\theta_A,\underset{\sim}{Y})} < K\Big)}{P_{\mathcal{H}_0}\Big(\frac{\mathscr{L}(\theta_0,\underset{\sim}{Y})}{\mathscr{L}(\theta_A,\underset{\sim}{Y})} = K\Big)}$$

Note that $\gamma = 0$ if the chosen K in *step 1* satisfies the equality in the lower bound of ($\dagger$) .

Remark: We note that to find K for a given α and γ when needed, we need to have the distribution of $\frac{\mathscr{L}(\theta_0,\underset{\sim}{Y})}{\mathscr{L}(\theta_A,\underset{\sim}{Y})}$. However, if $\frac{\mathscr{L}(\theta_0,\underset{\sim}{Y})}{\mathscr{L}(\theta_A,\underset{\sim}{Y})}$ is a monotonic function of

$T = T(\underset{\sim}{Y})$, then:

$$RR_{N-P} = \Big\{ \underset{\sim}{y} = (y_1, \ldots, y_n) : \frac{\mathscr{L}(\theta_0, \underset{\sim}{Y})}{\mathscr{L}(\theta_A, \underset{\sim}{Y})} < K \Big\}$$

$$= \begin{cases} \{\underset{\sim}{y} : T(\underset{\sim}{y}) < K_1\} & \text{if } \frac{\mathscr{L}(\theta_0, \underset{\sim}{Y})}{\mathscr{L}(\theta_A, \underset{\sim}{Y})} \text{ is a monotonic } \underline{\text{increasing}} \text{ function of } T \\ \{\underset{\sim}{y} : T(\underset{\sim}{y}) > K_2\} & \text{if } \frac{\mathscr{L}(\theta_0, \underset{\sim}{Y})}{\mathscr{L}(\theta_A, \underset{\sim}{Y})} \text{ is a monotonic } \underline{\text{decreasing}} \text{ function of } T \end{cases}$$

Example 15.2. $X_i \overset{iid}{\sim} Ber(p)$, $i = 1, 2, \ldots, n$

$\mathcal{H}_0 : p = p_0$, $\mathcal{H}_A : p = p_A$ $(p_A < p_0)$

Step 1: $\mathscr{L}(p; x_1, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^t(1-p)^{n-t}$ where $t = \sum_{i=1}^{n} x_i$

Step 2: $\frac{\mathscr{L}(p_0; \underset{\sim}{x})}{\mathscr{L}(p_A; \underset{\sim}{x})} = \frac{p_0^t(1-p_0)^{n-t}}{p_A^t(1-p_A)^{n-t}} = (\frac{p_0}{p_A})^t(\frac{1-p_0}{1-p_A})^{n-t} = \phi(t)$

Now we show that $\phi(t) \uparrow t$ if $p_0 > p_A$.

First note that $\phi(t)$ is a monotone increasing function of t , i.e. $\phi(t) \uparrow t$, if and only if $\log \phi(t) \uparrow t$:

$$\begin{aligned} \log \phi(t) &= t \log \frac{p_0}{p_A} + (n-t) \log \frac{1-p_0}{1-p_A} \\ &= t \log \Big[\frac{\frac{p_0}{(1-p_0)}}{\frac{p_A}{(1-p_A)}} \Big] + n \log(\frac{1-p_0}{1-p_A}) \\ &= at + b \end{aligned}$$

$$\text{where: } a = \log \Big[\frac{\frac{p_0}{1-p_0}}{\frac{p_A}{1-p_A}} \Big]$$

$$b = n \log(\frac{1-p_0}{1-p_A})$$

Thus: $\log \phi(t) \uparrow t \Leftrightarrow a > 0.$

Now notice that $\underbrace{p_0 > p_A}_{(1)} \Leftrightarrow (1-p_0) < (1-p_A) \Leftrightarrow \underbrace{\frac{1}{(1-p_0)} > \frac{1}{(1-p_A)}}_{(2)}$

Using (1) & (2): $\frac{p_0}{1-p_0} > \frac{p_A}{1-p_A} \Leftrightarrow \frac{\frac{p_0}{1-p_0}}{\frac{p_A}{1-p_A}} > 1$

Hence $\log \phi(t) \uparrow t$ if $p_{0} > p_{A}$. Thus:

$$RR = \left\{ \underset{\sim}{x} : \frac{\mathscr{L}(p_0; \underset{\sim}{x})}{\mathscr{L}(p_A; \underset{\sim}{x})} < K \right\} \quad = \quad \left\{ \underset{\sim}{x} : T(\underset{\sim}{x}) = \sum_{i=1}^{n} x_i < K_1 \right\}$$

Now under $\mathcal{H}_0 : p = p_0$, $T(\underset{\sim}{X}) = \sum_{i=1}^{n} X_i \sim \mathrm{Bin}(n, p_0)$. Thus we choose K_1 such that $P_{\mathcal{H}_0 : p = p_0}(T < K_1) = \alpha$ using Table 1, App.3 .

16 Lecture 16

16.1 Likelihood Ratio Tests

Let $F = \{ f_\theta : \theta \in \Theta\}$ be a family of parametric distributions on Θ , the parameter space, is the set of all admissible parameter values, i.e. all θ such that $f_\theta(x) \leq 0$ for all x and $\int_{-\infty}^{+\infty} f_\theta(x)dx = 1$.

Example $F = \left\{N(\mu, \sigma^2) : \underset{\sim}{\theta} = (\mu, \sigma^2) \text{ , } \mu \in \mathbb{R} \text{ , } \sigma^2 > 0\right\}$

Thus $\Theta = \{\underset{\sim}{\theta} = (\mu, \sigma^2) : \mu \in \mathbb{R} \text{ , } \sigma^2 \in \mathbb{R}^+\}$ is the parameter space in this case.

Suppose $\Theta_0 \subseteq \Theta$ and we want to test $\mathcal{H} : \theta \in \Theta_0$, i.e. the true θ belongs to Θ_0 , a subset of Θ , the parameter of #MISSING

Fischer suggested the following ratio, known as the *likelihood ratio test* , as a test statistic for testing $\mathcal{H}_0 : \theta \in \Theta_0$, $\lambda(x_1, \ldots, x_n) = \frac{\underset{\underset{\sim}{\theta} \in \Theta_0}{max} L(\theta; x_1, \ldots, x_n)}{\underset{\underset{\sim}{\theta} \in \Theta}{max} L(\theta; x_1, \ldots, x_n)}$ where $\underset{\sim}{x} = (x_1, \ldots, x_n)$ is the realization of $\underset{\sim}{X} = (X_1, \ldots, X_n)$.

16.2 The rational behind the Likelihood Ratio Tests (*LRT*)

Note that *LRT* is aligned with the idea of *MLE* . We #MISSING-L16P103 to compare the best that $\mathcal{H}_0$ can offer with the truth. The latter, because , is not available. The *MLE* has the theoretical intuitive appeal both. That is why we have:

$$L(\hat{\theta}_{ML}; x_1, \ldots, x_n) = \underset{\underset{\sim}{\theta} \in \Theta}{max} L(\underset{\sim}{\theta}; x_1, \ldots, x_n)$$

in the denominator. As for the numerator, we consider the best scenario under the $\mathcal{H}_0$, i.e. $\underset{\underset{\sim}{\theta} \in \Theta_0}{max} L(\underset{\sim}{\theta}; x_1, \ldots, x_n)$

This then essentially means that we do not want to reject $\mathcal{H}_0$ easily . Now if $\mathcal{H}_0$ is true , $\lambda(x_1, \ldots, x_n)$ will be close to 1 for large sample sizes n since the *MLE* is a consistent estimator of the true parameter value. Thus small values of λ are against $\mathcal{H}_0$, and hence:

$RR = \left\{\underset{\sim}{x} = (x_1, \ldots, x_n) \ : \ \lambda(x_1, \ldots, x_n) < K\right\}$ where K is chosen such that:

$P_{\mathcal{H}_0}\left\{\underset{\sim}{X} = (X_1, \ldots, X_n) \ : \ \lambda(\underset{\sim}{X}) < K\right\} = \alpha$ where α is given.

Remark: This is Neyman's view on testing. Fisher would provide the p-value:

p-value $= P_{\mathcal{H}_0}\left\{\underset{\sim}{X} = (X_1, \ldots, X_n) \ : \ \lambda(\underset{\sim}{X}) < \lambda_{obs}\right\}$ where

$\lambda_{obs} = \lambda(\underset{\sim}{x}_{obs} = (x_1, \ldots, x_n)$ is the observation sample .

Example 16.1. $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $i = 1, \ldots, n$

$\mathcal{H}_0 : \mu = \mu_0$ *where* μ *and* μ_0 *are known values*

Note that $\mathcal{H}_0$ is ***NOT*** a simple hypothesis, since σ^2 is left unspecified. It is essentially $\mathcal{H}_0 : \mu = \mu_0, \ \sigma^2 > 0$. This is a **composite hypothesis**:

$$\underset{\sim}{\theta} = (\mu, \sigma^2) : \mu \in \mathbb{R} \ , \ \sigma^2 \in \mathbb{R}^+$$
$$\underset{\sim}{\theta} = (\mu_0, \sigma^2) : \sigma^2 \in \mathbb{R}^+$$

Step 1: Numerator:

$$\max_{\underset{\sim}{\theta} \in \Theta_0} L(\underset{\sim}{\theta}; x_1, \ldots, x_n) \ = \ \max_{\underset{\sim}{\theta} \in \Theta_0} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu_0)^2}{2\sigma^2}} \ = \ \max_{\sigma^2} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_0)^2}$$

Taking the logarithm:

$$= \max_{\sigma^2} \left[-\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n}\frac{(x_i-\mu_0)}{2\sigma^2}\right]$$

Taking $\frac{d}{d\sigma^2}$ and equating the derivative to zero, we find:

$$0 = -\frac{n}{2}\cdot\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu_0)^2$$

and hence: $$\boxed{\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2}$$

Thus: $$\max_{\theta \in \Theta_0} L(\underset{\sim}{\theta}; x_1, \ldots, x_n) = L(\mu_0, \hat{\sigma}_0^2; x_1, \ldots, x_n)$$
$$= \left(\frac{1}{\sqrt{2\pi}\hat{\sigma}_0}\right)^n e^{-\frac{1}{2\hat{\sigma}_0^2}\sum_{i=1}^{n}(x_i-\mu_0)^2}$$
$$= \left(\frac{1}{\sqrt{2\pi}\hat{\sigma}_0}\right)^n e^{-\frac{n}{2}} \qquad (1)$$

Step 2: Denominator:

$$\max_{\underset{\sim}{\theta} \in \Theta} L(\underset{\sim}{\theta}; x_1, \ldots, x_n) = \max_{(\mu,\sigma^2)} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

We found the *MLE* of (μ, σ^2) in previous lecture:

$$\hat{\mu}_{ML} = \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

$$\max_{\underset{\sim}{\theta}\in\Theta} L(\underset{\sim}{\theta}; x_1,\ldots,x_n) = L(\bar{x}_n, \hat{\sigma}^2_{ML}; x_1,\ldots,x_n)$$
$$= (\frac{1}{\sqrt{2\pi}\hat{\sigma}_{ML}})^n e^{-\frac{1}{2\hat{\sigma}^2_{ML}}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2}$$
$$= (\frac{1}{\sqrt{2\pi}\hat{\sigma}_{ML}})^n e^{-\frac{n}{2}} \qquad (2)$$

Step 3: *LRT* and *RR*

$\underset{\sim}{x} = (x_1,\ldots,x_n)$

$$\lambda(\underset{\sim}{x}) = \frac{(\frac{1}{\sqrt{2\pi}\hat{\sigma}_0})^n \cancel{e^{-\frac{n}{2}}}}{(\frac{1}{\sqrt{2\pi}\hat{\sigma}_{ML}})^n \cancel{e^{-\frac{n}{2}}}} = (\frac{\hat{\sigma}^2_{ML}}{\hat{\sigma}^2_{ML}})^{\frac{n}{2}} \qquad (\dagger)$$

Now, note that:

$$\sum_{i=1}^{n}(x_i-\mu_0)^2 = \sum_{i=1}^{n}\left[(x_i-\bar{x}_n)+(\bar{x}_n-\mu_0)\right]^2$$
$$= \sum_{i=1}^{n}(x_i-\bar{x}_n)^2 + n(\bar{x}_n-\mu_0)^2 + 2(\bar{x}_n-\mu_0)\underbrace{\sum_{i=1}^{n}(x_i-\bar{x}_n)}_{0}$$
$$= \sum_{i=1}^{n}(x_i-\bar{x}_n)^2 + n(\bar{x}_n-\mu_0)^2$$

Thus $\quad \hat{\sigma}^2_0 = \hat{\sigma}^2_{ML} + (\bar{x}_n-\mu_0)^2 \qquad (\ddagger)$

$$\lambda(\underset{\sim}{x}) = \left(\frac{\hat{\sigma}^2_{ML}}{\hat{\sigma}^2_0}\right)^{\frac{n}{2}} = \left(\frac{\hat{\sigma}^2_{ML}}{\hat{\sigma}^2_{ML}+(\bar{x}_n-\mu_0)^2}\right)^{\frac{n}{2}}$$
$$= \left(\frac{1}{1+\left[\frac{(\bar{x}_n-\mu_0)^2}{\hat{\sigma}^2_{ML}}\right]}\right)^{\frac{n}{2}} = \left(\frac{1}{1+\mathrm{U}}\right)^{\frac{n}{2}}$$

Note that $\lambda(\underset{\sim}{x})$ is a *monotone decreasing function* of U. Thus for any K , we can find K' such that:

$$RR = \{\underset{\sim}{x} = (x_1,\ldots,x_n) : \lambda(\underset{\sim}{x}) < K\}$$
$$= \{\underset{\sim}{x} = (x_1,\ldots,x_n) : \mathrm{U} > K'\}$$

Step 4: Finding K' for a given α :

In order to find K' , we need the distribution of U :

$$\mathrm{U} = \frac{(\bar{X}_n-\mu_0)^2}{\hat{\sigma}^2_{ML}} = \frac{n(\bar{X}_n-\mu_0)^2}{\sum_{i=1}^{n}(X_i-\bar{X}_n)^2} = \frac{\frac{n(\bar{X}_n-\mu_0)^2}{\sigma_0^2}}{\frac{(n-1)S^2}{\sigma_0^2}}$$

where σ_0^2 is the true variance . Now note that:

$$\frac{\bar{X}_n-\mu_0}{\frac{\sigma_0}{\sqrt{n}}} \sim N(0,1) \implies \left[\frac{\bar{X}_n-\mu_0}{\frac{\sigma_0}{\sqrt{n}}}\right]^2 \sim \chi^2_{(1)}$$

Thus the numerator of U is a $\chi^2_{(1)}$. We also learned in previous lectures that the denominator of U , $\frac{(n-1)S^2}{\sigma_0^2}$, is a $\chi^2_{(n-1)}$.

Theorem. *If* $X_i \overset{iid}{\sim} N(\mu,\sigma^2)$ *, then* $\bar{X}_n \coprod S^2$.

The above theorem implies that the numerator and denominator of U are independent, Thus:

$$(n-1)\mathrm{U} = \frac{\frac{\chi^2_{(1)}}{1}}{\frac{\chi^2_{(n-1)}}{n-1}} \sim F_{1,n-1}$$

Thus, for a given α , $K^{'}$ is chosen such that:

$$\alpha = P_{\mathcal{H}_0:\mu=\mu_0}\{\mathrm{U} > K^{'}\} = P_{\mathcal{H}_0:\mu=\mu_0}\{F_{1,n-1} > (n-1)K^{'}\} .$$

The above calculation heavily hings upon normality assumption. If the samples are not coming from a normal distribution, we may not be able to find the distribution of $\lambda(\underset{\sim}{X})$ or a function of $\lambda(\underset{\sim}{X})$ for any given sample size n. The following theorem (Theorem 10.2, page 553 from the textbook) provides a useful result when n is large.

Theorem. Let $Y_1, \ldots, Y_n$ have joint likelihood function $L(\theta)$ and r_0 denote the number of free parameters that are specified by $\mathcal{H}_0 : \theta \in \Theta_0$ and r denote the number of free parameters specified by the statement $\theta \in \Theta$. Then, for a large n:

$$-2\ln(\lambda(\underset{\sim}{X})) \overset{app}{\sim} \chi^2_{r-r_0}$$

17 Note on Relative Efficiency

Suppose $X_i \overset{iid}{\sim} N(0, \sigma^2), i = 1, \ldots, n$ where σ^2 is unknown. Consider the following estimators of σ :

$$\hat{\sigma}_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2} \qquad \text{and} \qquad \hat{\sigma}_n = \frac{\sqrt{\frac{\pi}{2}}}{n}\sum_{i=1}^{n} |X_i|$$

Define $Y_i = X_i^2 \implies \frac{Y_i}{\sigma^2} \overset{iid}{\sim} \chi^2_{(i)} \implies \underbrace{\sum_{i=1}^{n} \frac{Y_i}{\sigma^2}}_{\mathrm{U}_n} \sim \chi^2_{(n)}$.

Let $\mathcal{Z}_i = |X_i|$, then:

$P(\mathcal{Z} \le \zeta) = P(|X| \le \zeta) = P(-\zeta \le X \le \zeta)$

$$= F_X(\zeta) - F_X(-\zeta)$$

which implies: $f_Z(\zeta) = f_X(\zeta) + f_X(-\zeta)$