MATH324 (Statistics) – Lecture Notes

McGill University

Prof. Masoud Asgharian

Sam K.H.Targhi-Dunn

sam.targhi@mail.mcgill.ca

Winter 2019

Contents

1 Lecture 1

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:

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

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

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:

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

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:

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

Thus, using (2) we have:

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

Now:

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

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

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:

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

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 $\text{Bias}(\hat{\Theta}_n) = 0$.

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 $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

$$E(\bar{X}_n) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} \overbrace{E(X_i)}^{\mu} = \frac{1}{\not{n}} \cdot \not{n}\mu = \mu$$

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

$$\begin{aligned}
Var(\bar{X}_n) &= Var\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n^2} Var(\sum_{i=1}^{n} X_i) \\
&= \frac{1}{n^2}\Big[\sum_{i=1}^{n} Var(X_i) + 2 \cdot \sum\sum_{1 \le i < j \le n} \overbrace{Cov(X_i, X_j)}^{0}\Big] && \text{Theorem 5.12(b) - page 271} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) && \coprod_{i=1}^{n} X_i \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n\not{n}} \cdot \not{n}\sigma^2 = \frac{\sigma^2}{n} && \text{identically distributed} \\
&\implies MSE(\bar{X}_n) = Var(\bar{X}_n) + \overbrace{Bias^2(\bar{X}_n)}^{0} = Var(\bar{X}_n) = \frac{\sigma^2}{n}
\end{aligned}$$

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:

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

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 $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ 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:

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

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.:

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

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:

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

Thus (*) is equivalent to :

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

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

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

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

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

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

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 $\mathbb{E}(\frac{n}{n-1} S^2_{n,*}) = \sigma^2$. Note however that:

$$\frac{n}{n-1} S^2_{n,*} = \frac{\not{n}}{n-1} \cdot \frac{1}{\not{n}} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

Thus $\boxed{S^2_n = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2}$ 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_is have the common mean μ_X and Y_js 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 known, but μ is unknown. 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* $X_{(n)} = \max_{1 \le i \le n}$.

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

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

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

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

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

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:

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

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

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 :

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

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:

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

Equivalently:

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

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.

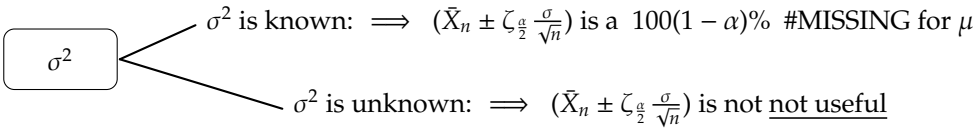
5.2 Large Sample Confidence Interval

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:

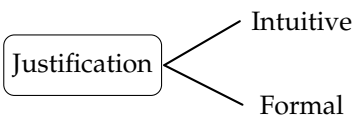
$$\frac{\bar{X}_n - \mu}{(\frac{\sigma}{\sqrt{n}})} \overset{app}{\sim} N(0,1) \qquad \text{when } n \text{ is large enough}$$

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.



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:

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



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:

$$g(S_n^2) \xrightarrow{P} g(\sigma^2) \qquad \text{for any continuous function}$$

Considering $g(x) = \sqrt{X}$, we obtain $S_n \xrightarrow{P} \sigma$ and hence $\frac{\sigma}{S} \xrightarrow{P} 1$.

3. **Cramer's Theorem :**

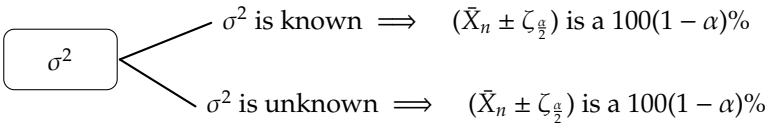
This result says that if $V_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} 1$, then $Y_n \cdot V_n \xrightarrow{D} X$:

$$\frac{\bar{X}_n - \mu}{\frac{S}{\sqrt{n}}} = \underbrace{\frac{\sigma}{S_n}}_{Y_n} \cdot \underbrace{\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}}_{V_n}$$

Note that GLT implies that $V_n \xrightarrow{D} 2$, i.e:

$$\overbrace{F_{V_n}(t)}^{\text{cdf of } V_n} \to F_Z(t) = \overbrace{\int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx}^{\Phi(t) \text{ cdf of } N(0,1)}$$

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



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

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

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

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

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)

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

$$\text{Thus} \qquad \begin{aligned} m_U(t) &= \frac{m_W(t)}{m_V(t)} \\ &= \frac{(1-2t)^{-\frac{n}{2}}}{(1-2t)^{-\frac{1}{2}}} \\ &= (1-2t)^{-\frac{n-1}{2}} \end{aligned}$$

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 unknown :

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

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 :

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

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:

$$X \sim F_X \, , Y = F_X(X)$$

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

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

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

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

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

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

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

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:

$$m_{\sum_{i=1}^n V_i}(t) = \mathbb{E}[e^{-t\sum_{i=1}^n V_i}] = \mathbb{E}\Big[\prod_{i=1}^{n} e^{-tV_i}\Big]$$

$$\coprod_{i=1}^{n} V_i \implies = \prod_{i=1}^{n} \mathbb{E}[e^{-tV_i}] = \prod_{i=1}^{n} m_{V_i}(t)$$

$$\text{identically distributed} \implies = [m_Y(t)]^n = [(1-2t)^{-\frac{2}{2}}]^n$$

$$= (1-2t)^{-\frac{2n}{2}} \implies \sum_{i=1}^{n} V_i \sim \chi^2_{2n}$$

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:

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

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 95% 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:

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

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:

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

We then find $\boxed{n = \frac{\zeta_{\frac{\alpha}{2}}^2 \sigma^2}{\beta^2}}$ 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 likehood (ML). Suppose $\hat{\theta}_n$ is the maximum likehood 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:

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

7 Lecture 7

8 Lecture 8

9 Lecture 9

10 Lecture 10

11 Lecture 11

12 Lecture 12