# 37262 Mathematical Statistics

Lecture 8

# Bias

- Let $T$ be a statistic used to estimate the value of a parameter $\theta$.

- The **bias** of $T$ in estimating $\theta$ is defined as $bias(T,\theta) = E(T) - \theta$.

- If the bias of an estimator for a parameter is zero, this is said to be **unbiased**. Otherwise, the estimator is said to be **biased**.

# Bias

- Let $X$ be a random variable with mean $\mu < \infty$ and variance $\sigma^2 < \infty$.

- Recall now the method of moments, which matches the moments of the distribution to those of a sample drawn from it.

- For a sample of independent realisations $\{x_1, x_2, \ldots, x_n\}$ of $X$ , the method gives estimates of the mean and variance

$$\hat{\mu}_{MM} = s_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \text{ and } \hat{\sigma}^2_{MM} = cs_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- We now show that one of these is an unbiased estimator, but the other is biased.

# Bias

- Trivially, we have that each $x_i$ is a realisation of $X$ for which $E(X) = \mu$ hence

$$E(s_1) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} x_i\right) = \frac{1}{n}(n\mu) = \mu.$$

- We therefore have that $E(\hat{\mu}_{MM}) = \mu$ so the sample mean is an unbiased estimator of the population mean.

# Bias

- Consider now $E(cs_2) = E\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right) = \frac{1}{n}E\left(\sum_{i=1}^{n}((x_i - \mu) - (\bar{x} - \mu))^2\right)$

$$= \frac{E\left(\sum_{i=1}^{n}(x_i - \mu)^2\right)}{n} + \frac{E\left(\sum_{i=1}^{n}(\bar{x} - \mu)^2\right)}{n} - \frac{2E\left(\sum_{i=1}^{n}(x_i - \mu)(\bar{x} - \mu)\right)}{n}.$$

- Now, $\bar{x} - \mu = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)$ hence

$$E(cs_2) = \frac{E\left(\sum_{i=1}^{n}(x_i - \mu)^2\right)}{n} + \frac{E\left(n(\bar{x} - \mu)^2\right)}{n} - 2E\left((\bar{x} - \mu)^2\right)$$

- $E(cs_2) = \frac{E\left(\sum_{i=1}^{n}(x_i - \mu)^2\right)}{n} - E\left((\bar{x} - \mu)^2\right)$

# Bias

- $E(cs_2) = \frac{E\left(\sum_{i=1}^{n}(x_i - \mu)^2\right)}{n} - E\left((\bar{x} - \mu)^2\right) \;= \sigma^2 - \frac{\sigma^2}{n}$

- The second (central) sample moment is therefore a biased estimator of the population variance, since $\sigma^2 - \frac{\sigma^2}{n} < \sigma^2$.

- When estimating a population variance from a sample, we often use $\hat{\sigma}^2 = \frac{n}{n-1} cs_2 = \frac{n}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ since this is unbiased.

# Comparing Unbiased Estimators

- Although it may seem, intuitively obvious that an estimator with small (or zero) bias is “better” than a biased estimator, there are some surprising examples which might suggest otherwise.

- There are several famous examples in which we might favour a biased estimator over an unbiased one.

- Keep in mind that an unbiased estimator is only giving the true value <u>on average</u>. Some unbiased estimators can provide huge overestimates or huge underestimates on any single realisation, but still be unbiased.

- In such cases, a more stable estimate with small bias may well be preferable.

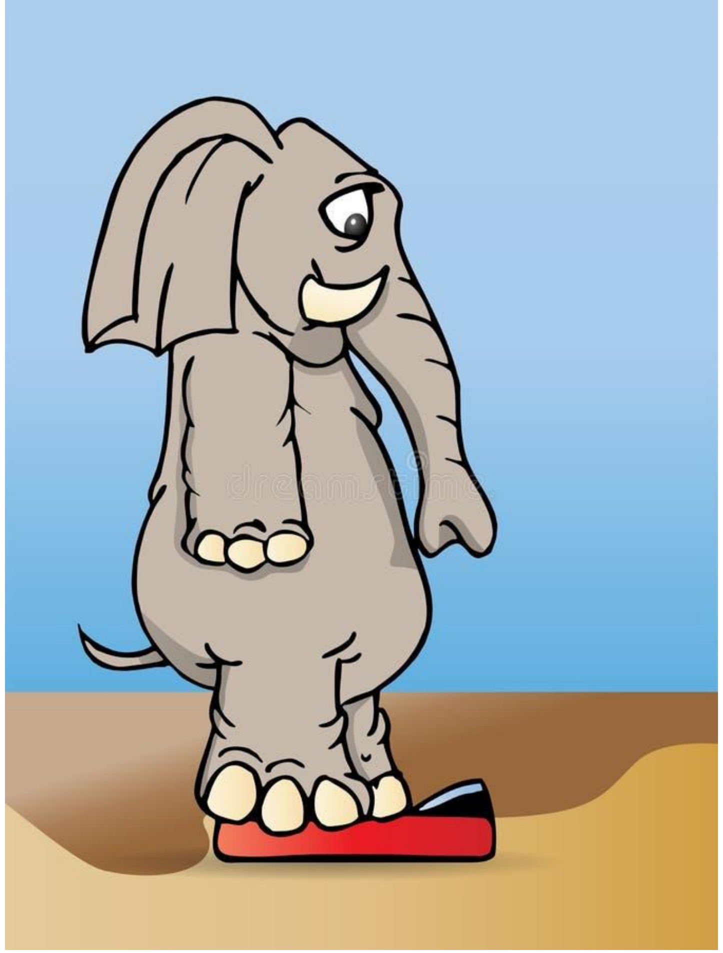
# Basu’s Elephant

- Maybe the most famous example of this is **Basu’s elephant**.

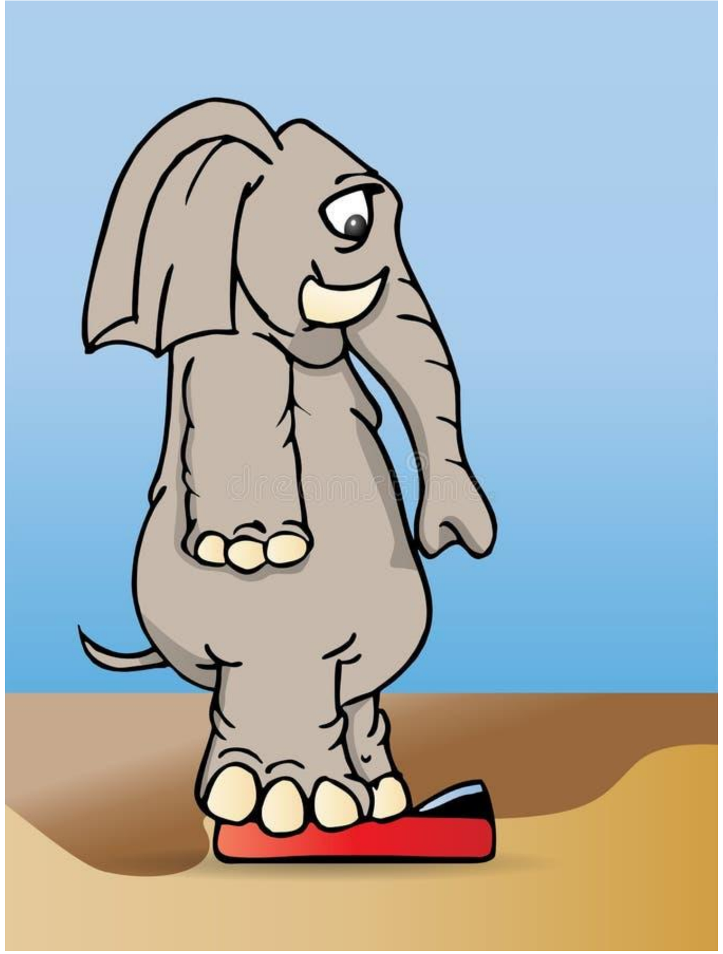
- This arises from a story whereby a circus which owns 50 elephants wishes to know the total mass of all of its elephants.

- When the circus owners conducted the same analysis previously, they noted that one elephant – Sambo - has a mass very close to the mean of all 50 elephants. They therefore propose to save time by simply measuring one elephant’s mass. Their initial idea is to measure only Sambo and estimate the total mass as $50\omega_{sambo}$ where $\omega_{sambo}$ is Sambo’s mass.

- This, in fact, is not guaranteed to be an unbiased estimator.



# Basu’s Elephant



- Instead, the circus calls in a statistician, who says they must perform a random sample to select the elephant to be measured.

- The statistician knows that an unbiased estimator for this is

  $\hat{\omega}_{total} = \dfrac{\omega_j}{p_j}$ where $\omega_j$ is the mass of selected elephant $j$

  and $p_j$ is the probability that elephant $j$ is the one selected.

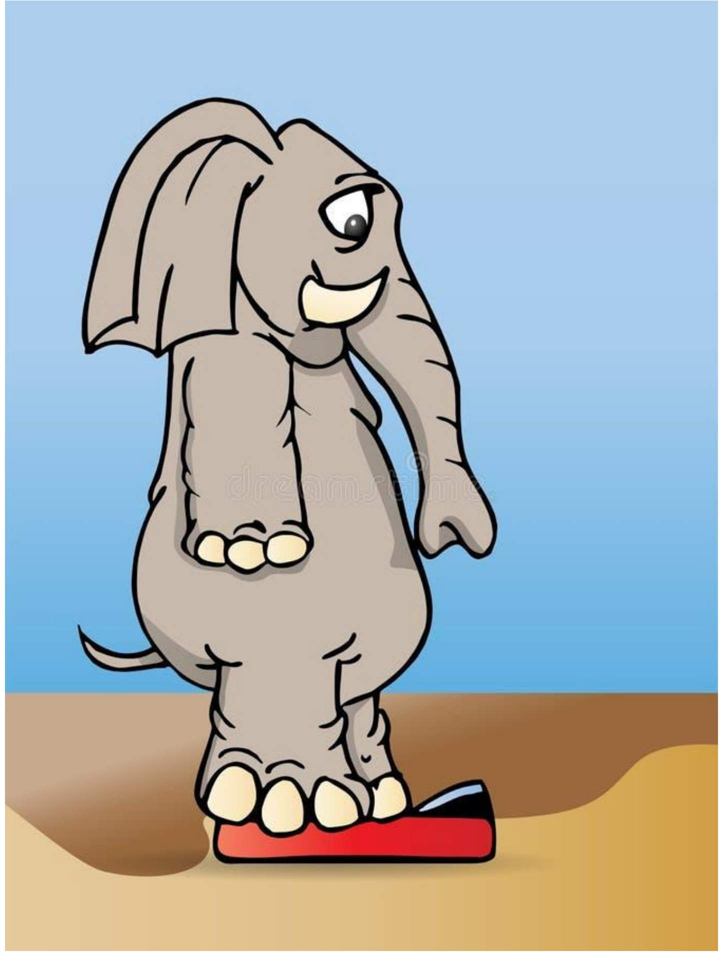
- The circus owners think that selecting Sambo alone is the best idea, so conduct a random sample which selects Sambo with probability 99% and each of the other 49 elephants with equal probability.

# Basu’s Elephant

- Introducing the notation $I_j = \begin{cases} 0 & \text{if elephant } j \text{ is not selected} \\ 1 & \text{if elephant } j \text{ is selected} \end{cases}$

  we can see that $E(\hat{\omega}_{total}) = E\left(\frac{\omega_i}{p_i}\right) = E\left(\frac{\omega_1 I_1}{p_1} + \frac{\omega_2 I_2}{p_2} + \ldots \frac{\omega_{50} I_{50}}{p_{50}}\right)$

  $$= E\left(\frac{\omega_1 p_1}{p_1} + \frac{\omega_2 p_2}{p_2} + \ldots \frac{\omega_{50} p_{50}}{p_{50}}\right) = E(\omega_{total})$$



- This tells us that the estimator is unbiased.

# Basu’s Elephant

- When this unbiased estimator is applied, the statistician estimates $\hat{\omega}_{total} = \frac{100}{99}\omega_{sambo}$ if Sambo is selected.
- Even more absurd is the case when the heaviest elephant is selected.
- If the mass of the heaviest elephant is $\omega_{jumbo}$ and this elephant is selected, then $\hat{\omega}_{total} = \left(\frac{49}{1}\right)\left(\frac{100}{1}\right)\omega_{jumbo} = 4900\omega_{jumbo}$
- The estimator is unbiased, but can be wildly inaccurate in either direction.

*Debabrata Basu*
*(1924 - 2001)*

# Comparing Unbiased Estimators

- Although it may seem, intuitively obvious that an estimator with small (or zero) bias is “better” than a strongly biased estimator, we should not consider that all unbiased estimators are equally good, especially for small sample sizes.

- Consider the case of independent samples $\boldsymbol{Q} = \{q_1, q_2, \ldots, q_n\}$ drawn from $Q \sim N(\mu, \sigma^2)$ where the population variance $\sigma^2$ is known and we estimate the population mean from samples.

- It is easily shown that $\hat{\mu}_1 = q_1$ is unbiased. So is $\hat{\mu}_2 = \frac{q_1 + q_2}{2}$. So is $\hat{\mu}_n = \frac{\sum_{i=1}^{n} q_i}{n}$.

- Other estimators such as $\hat{\mu}_{med} = median\{q_1, q_2, \ldots, q_n\}$ and $\hat{\mu}_w = \frac{q_1 + 2q_2 + 3q_3 + 4q_4}{10}$ are also unbiased.

# Comparing Unbiased Estimators

- Are any of these unbiased estimators “better” or “worse” than others?

- It seems reasonable to think that an estimator based on a larger sample of observations will give a better picture of the population than a smaller sample.

- For a normal variable, both the sample mean and the sample median can be calculated using the same number of observations, yet we may have reason to believe that one is a better estimator than the other.

- We first need to formalise what we mean here mathematically.

# Comparing Unbiased Estimators

- We can calculate the variance of an unbiased estimator around its true value. That is, for estimator $\hat{\theta}$ of $\theta$, we can calculate $Var(\hat{\theta})$.

- Intuition may tell us that we prefer lower variance estimators, since these will provide less uncertainty about the true parameter value.

- In all but the most trivial of cases, we will not be able to find estimators from finite samples which have zero variance.

- We can, though, put a lower bound on the possible variance of an estimator and look at how close to this bound an estimator is.

# Score Function

- Consider a single parameter random variable $X$ with probability density function $f(x\,|\,\theta)$ .

- Given an independent sample of observations $\boldsymbol{X} = \{x_1, x_2, ..., x_n\}$ , the associated loglikelihood function is $\ln L(\boldsymbol{X}\,|\,\theta) = \ell(\boldsymbol{X}\,|\,\theta)$.

- We call the derivative of the loglikelihood with respect to the parameter as the **score function.**

$$V = Score(\boldsymbol{X}\,|\,\theta) = \frac{\partial}{\partial\theta}\ell(\boldsymbol{X}\,|\,\theta)$$

- It is easily seen that the expected value of the score at the true parameter value of $\theta$ is zero.

# Score Function

- It is easily seen that the expected value of the score at the true parameter value of $\theta$ is zero.
- $E(V) = E\left(\frac{\partial}{\partial\theta}\ln L(\boldsymbol{X} \mid \theta)\right) = E\left(\frac{\frac{\partial}{\partial\theta}L(\boldsymbol{X} \mid \theta)}{L(\boldsymbol{X} \mid \theta)}\right) = \int \frac{\frac{\partial}{\partial\theta}L(\boldsymbol{X} \mid \theta)}{L(\boldsymbol{X} \mid \theta)} L(\boldsymbol{X} \mid \theta)d\boldsymbol{X} = \frac{\partial}{\partial\theta}\int L(\boldsymbol{X} \mid \theta)d\boldsymbol{X}$
- Since $\int L(\boldsymbol{X} \mid \theta)d\boldsymbol{X} = 1,\ E(V) = \frac{\partial}{\partial\theta}1 = 0.$
- In the above calculation, all integrals are over the full *n*-dimensional range of ***X*** and we have assumed some regularity conditions to interchange the order of the derivative and integral.
- Although shown for continuous variables, the same argument holds with probability mass functions for discrete variables.

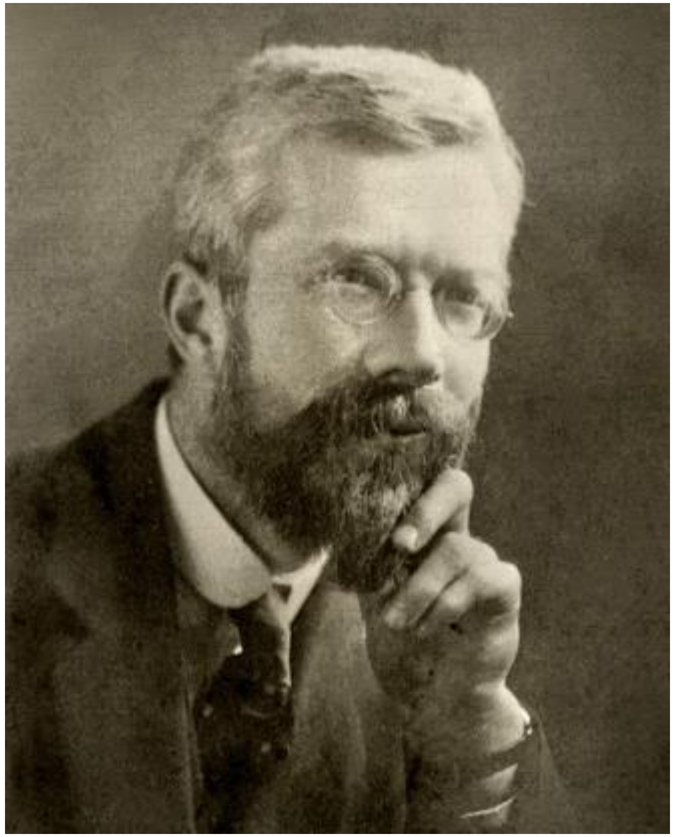
# Fisher Information

- The **Fisher Information** is defined as the variance of the score function.

- Alternatively, $Var(V) = -E\left(\frac{\partial^2}{\partial\theta^2}\ln L(\boldsymbol{X} \mid \theta)\right)$ (since $E(V) = 0$.)

- When the Fisher Information is larger, changes with respect to $\theta$ have larger changes in the likelihood of a sample.

- Informally, a larger Fisher Information tells us that our sample contains more information about the parameter value. Conversely, smaller Fisher Information implies a flatter likelihood function, hence less certainty about the parameter value.



*Ronald Fisher*
(*1890 - 1962*)

# Cramér-Rao Bound

- The Fisher Information provides a bound on the minimum possible variance which an estimator can have.

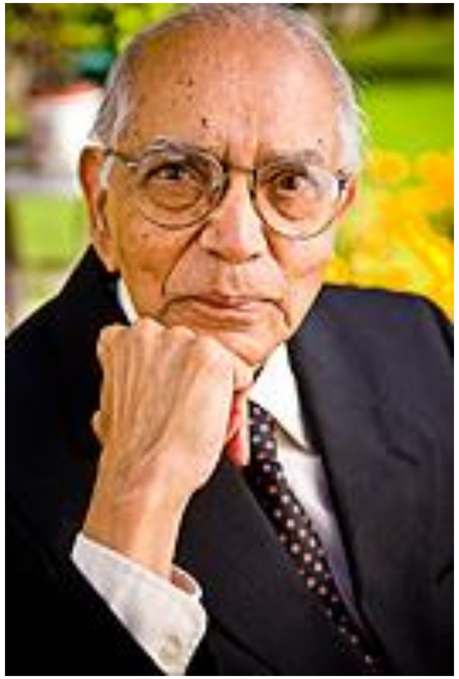
- It is beyond the scope of 37262 to prove this formally, but we have that the variance of an estimator cannot be less than the reciprocal of the Fisher Information.

- $Var(\hat{\theta}) \geq \frac{1}{I(\theta)}$ where $I(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2}\ln L(\boldsymbol{X} \mid \theta)\right)$ is the Fisher information

- This is the **Cramér-Rao bound**.

*Harald Cramér*

(*1893 - 1985*)



*Calyampudi Radhakrishna*

(*C.R.*) *Rao*

(*1920 -* )

# Efficiency

- We define the **efficiency** of an estimator by what proportion of the Cramér-Rao bound its variance obtains.
- $eff(\hat{\theta}) = \dfrac{1}{I(\theta)Var(\hat{\theta})}$
- Clearly the efficiency must be between 0 and 1.
- You may sometimes see **relative efficiency** of two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$.
- $eff(\hat{\theta}_1, \hat{\theta}_2) = \dfrac{Var(\hat{\theta}_1)}{Var(\hat{\theta}_2)}$ . Clearly this is non-negative but can take values above 1.

# Efficiency

- Consider the example of $X \sim Bin(n,p)$ where *n* is known, but *p* is to be estimated.

- The loglikelihood of a sample $\boldsymbol{X} = \{x_1, x_2, \ldots, x_k\}$ is then

$$\ell(\boldsymbol{X} \mid p) = \sum_{i=1}^{k} \ln\left(\frac{n!}{x_i!(n-x_i)!} p^{x_i} (1-p)^{n-x_1}\right) = \ln(p)\sum_{i=1}^{k} x_i + \ln(1-p)\sum_{i=1}^{k}(n-x_i) + \sum_{i=1}^{k} \ln\left(\frac{n!}{x_i!(n-x_i)!}\right).$$

- Differentiating gives $\dfrac{\partial}{\partial p}\ell(\boldsymbol{X} \mid p) = \dfrac{\sum_{i=1}^{k} x_i}{p} - \dfrac{\sum_{i=1}^{k}(n-x_i)}{(1-p)}.$

- The maximum likelihood is $\hat{p} = \dfrac{\sum_{i=1}^{k} x_i}{kn}.$

# Efficiency

- We find the variance of the estimator $\hat{p}=\dfrac{\sum_{i=1}^{k}x_i}{kn}$.
- $Var(\hat{p})=\dfrac{Var\left(\sum_{i=1}^{k}x_i\right)}{k^2n^2}=\dfrac{npk(1-p)}{k^2n^2}=\dfrac{p(1-p)}{kn}$.
- The Fisher Information is $-E\left(\dfrac{\partial^2}{\partial p^2}\ell(\boldsymbol{X}\mid p)\right)=E\left(\dfrac{\partial}{\partial p}\left(-\dfrac{\sum_{i=1}^{k}x_i}{p}+\dfrac{\sum_{i=1}^{k}(n-x_i)}{(1-p)}\right)\right)=E\left(\dfrac{\sum_{i=1}^{k}x_i}{p^2}+\dfrac{\sum_{i=1}^{k}(n-x_i)}{(1-p)^2}\right)$
- $I(p)=\dfrac{npk}{p^2}+\dfrac{nk(1-p)}{(1-p)^2}=\dfrac{nk}{p}+\dfrac{nk}{(1-p)}=\dfrac{nk}{p(1-p)}$.
- The maximum likelihood estimator here therefore has efficiency of 1.

# Ordinary Least Squares

- Recall that, for a simple linear regression $y_i = \alpha + \beta x_i + \varepsilon_i$ , the estimates of the intercept and slope parameters are

$$\hat{\alpha} = \bar{y} - \beta\bar{x} \text{ and } \hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \text{ where } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i .$$

- Consider having a set of observations of $X$, $\boldsymbol{X} = \{x_1, x_2, \ldots, x_n\}$.

- $$E(\hat{\beta} \mid \boldsymbol{X}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(E(y_i) - E(\bar{y}))}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\alpha + \beta x_i - \alpha - \beta\bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta$$

# Ordinary Least Squares

- Similarly, $E(\hat{\alpha} \mid \boldsymbol{X}) = E(\bar{y}) - E(\hat{\beta})\bar{x} \; = \alpha + \hat{\beta}\bar{x} - \hat{\beta}\bar{x} = \alpha.$

- The estimates of the slope and intercept parameters are therefore both unbiased.

- Note that no assumptions have been made here beyond that the residuals have zero expectation. No further distributional assumptions have been made.

# Ordinary Least Squares

- We now consider the variance of these estimates around the true unknown values.
- $Var(\hat{\beta} \mid \boldsymbol{X}) = Var\left( \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right)$
- Now, in the numerator, $\bar{y}\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ and, conditional on knowledge of $\boldsymbol{X}$ $Var\sum_{i=1}^{n}(x_i - \bar{x})(\alpha + \beta x_i) = 0.$
- $Var(\hat{\beta} \mid \boldsymbol{X}) = Var\left( \frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 Var(\varepsilon_i)}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$

# Ordinary Least Squares

- Note that the simplification of $Var\left(\sum_{i=1}^{n}(x_i-\bar{x})\varepsilon_i\right)=\sum_{i=1}^{n}(x_i-\bar{x})^2 Var(\varepsilon_i)$ relies not only on the fact that, for a constant *k* and variable *Z*, $Var(kZ)=k^2 Var(Z)$ but also on all of the residuals being uncorrelated.

- That is, we require the assumption that $Cov(\varepsilon_i,\varepsilon_j)=\begin{cases}\sigma^2 & i=j\\ 0 & i\neq j\end{cases}$ .

- Similarly, $Var(\hat{\alpha}\,|\,\boldsymbol{X})=Var\left(\bar{y}-\hat{\beta}\bar{x}\right)\;=\frac{\sigma^2}{n}+Var\left(\bar{x}\hat{\beta}\right)=\frac{\sigma^2}{n}+\frac{\bar{x}^2\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$.

# Confidence Intervals

- A **confidence interval** is a range of values for the estimate of an unknown parameter.

- It is associated with a **confidence level**. Which represents the long-run proportion of confidence intervals generated which would contain the true value of the parameter.

- For example, if a sampling procedure is developed and associated 95% confidence intervals calculated, then if repeated independent samples are generated under the same procedure, then 95% of the resulting intervals will contain the true parameter value.

- Although widely used, confidence intervals are easily and often misunderstood.

# Confidence Intervals

- A confidence level of 95% does not mean that 95% of sample data will lie in the interval.

- A 95% confidence interval for a parameter does not imply that there is a 95% chance that the true value parameter for which the interval is calculated lies in the interval.

# Confidence Intervals

- Most commonly, we apply the central limit theorem to construct confidence intervals.

- The central limit theorem states that, if $x_1, x_2, \ldots, x_n$ are independent realisations of the same variable $X$, with mean $\mu < \infty$ and variance $\sigma^2 < \infty$, then the variable

$$z = \lim_{n \to \infty} \left( \frac{\bar{x}_n - \mu}{\left( \frac{\sigma}{\sqrt{n}} \right)} \right) \sim N(0,1)$$ where $\bar{x}_n$ is the sample mean $\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$

- A $(100 - \gamma)\%$ symmetric confidence interval for $\mu$ (assuming we know $\sigma^2$) is then given by

$\bar{x}_n \pm z_{\gamma/2} \sqrt{\frac{\sigma^2}{n}}$ where $z_{\gamma/2}$ is the $(\gamma/2)^{th}$ percentile of the standard normal distribution.

# Confidence Intervals

- In practice, we often do not know the population standard deviation $\sigma^2$ and instead have to estimate it from the same sample which is being used to estimate the mean.

- Consider the expectation of $cs_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$
- $E(cs_2) = \frac{1}{n}E\left(\sum_{i=1}^{n}((x_i - \mu) - (\overline{x} - \mu))^2\right) = \frac{1}{n}\sum_{i=1}^{n}Var(x_i) - Var(\overline{x}) = \sigma^2 - \frac{\sigma^2}{n} = \sigma^2\left(\frac{n-1}{n}\right)$
- $\left(\frac{n}{n-1}\right)cs_2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ is therefore an unbiased estimator of $\sigma^2$.

# Confidence Intervals

- Recall that, for a regression model $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$ , assuming that residuals are independent and normally distributed with equal variance $\sigma^2$.

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \sim \sigma^2 \chi^2 (n-1)$$

- Similarly, we have that $\sum_{i=1}^{n} (x_i - \bar{x})^2 \sim \sigma^2 \chi^2 (n-1)$

- If we are estimating $\sqrt{\sigma^2}$ with $\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$ , we now note that

$z = \lim_{n \to \infty} \left( \frac{\bar{x}_n - \mu}{\left( \frac{\sigma}{\sqrt{n}} \right)} \right) \sim N(0,1)$ now has the square root of a chi-squared variable in its denominator.

# Confidence Intervals

- The ratio of a normal variable to the square root of a $\chi^2(n-1)$ variable over its degrees of freedom follows a $t_{n-1}$ distribution.

- If the population variance is unknown, our confidence interval for the mean, based on a sample is therefore $\bar{x}_n \pm t_{n-1,(\gamma/2)} \dfrac{s}{\sqrt{n}}$ where $t_{n-1,(\gamma/2)}$ is the $(\gamma/2)^{th}$ percentile of the $t$ distribution with $n-1$ degrees of freedom, where $s = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

# Confidence Intervals

- Recall that, for a regression model $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$ , assuming that residuals are independent and normally distributed with equal variance $\sigma^2$.

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \sim \sigma^2 \chi^2 (n-1)$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta_1 x_{1i} - \beta_2 x_{2i} - \ldots - \beta_k x_{ki})^2 \sim \sigma^2 \chi^2 (n-k-1)$$

- Considering now only the simple linear regression case, $y_i = \alpha + \beta x_i + \varepsilon_i$ (i.e. $k=1$) we have

that $SSE = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \sim \sigma^2 \chi^2 (n-2)$

- Taking the expectation of both sides, we obtain $E\left( \sum_{i=1}^{n} (y_i - \hat{y})^2 \right) = (n-2)\sigma^2$.

# Confidence Intervals

- $E\left(\sum_{i=1}^{n}(y_i-\hat{y})^2\right)=(n-2)\sigma^2$ hence $\hat{\sigma}^2=\dfrac{\sum_{i=1}^{n}(y_i-\hat{y})^2}{n-2}=\dfrac{\sum_{i=1}^{n}\varepsilon_i{}^2}{n-2}$ is an unbiased estimator of $\sigma^2$.

- We saw that $Var(\hat{\beta}\,|\,\boldsymbol{X})=\dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ and $Var(\hat{\alpha}\,|\,\boldsymbol{X})=\dfrac{\sigma^2}{n}+\dfrac{\bar{x}^2\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$.

- Our confidence intervals for these estimates are therefore $\hat{\beta}\pm t_{n-2,(\gamma/2)}s_\beta$ and $\hat{\alpha}\pm t_{n-2,(\gamma/2)}s_\alpha$

where $s_\beta=\sqrt{\dfrac{\sum_{i=1}^{n}(y_i-\hat{y})^2}{(n-2)\sum_{i=1}^{n}(x_i-\bar{x})^2}}$ and $s_\alpha=\sqrt{\dfrac{\sum_{i=1}^{n}(y_i-\hat{y})^2}{n(n-2)}+\dfrac{\bar{x}^2\sum_{i=1}^{n}(y_i-\hat{y})^2}{(n-2)\sum_{i=1}^{n}(x_i-\bar{x})^2}}$.

# Prediction Intervals

- As we have already said, there are a number of common misinterpretations of confidence intervals. One of these is to confuse them with prediction intervals.

- A **prediction interval** is an estimate of an interval in which a future observation drawn from a known (or estimated) distribution may lie.

# Prediction Intervals

- Consider a regression model $y_i = \alpha + \beta x_i + \varepsilon_i$ and observed predictors $\{x_1, x_2, \ldots, x_n\}$ with corresponding responses $\{y_1, y_2, \ldots, y_n\}$ respectively.

- Given these observations, we can construct a prediction interval for the response value $y_k$ corresponding to predictor value $x_k$ $(k \notin \{1, 2, \ldots, n\})$

- Clearly, the prediction interval should be centred on $E(y_k \mid x_k) = \alpha + \beta x_k$.

- We can simply evaluate this with our unbiased estimates of the two regression parameters.

- The variance is less straightforward.

# Prediction Intervals

- $\hat{y}_k = \hat{\alpha} + \hat{\beta} x_k$ hence $Var(\hat{y}_k) = Var(\hat{\alpha} + \hat{\beta} x_k) = Var(\bar{y} - \hat{\beta}(\bar{x} - x_k)) = \dfrac{\sigma^2}{n} + \dfrac{(\bar{x}^2 - x_k)^2 \sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$.

- The variance of the future observation is $Var(\hat{y}_k + \varepsilon_k) = \dfrac{\sigma^2}{n} + \dfrac{(\bar{x}^2 - x_k)^2 \sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} + \sigma^2$.

- Again, assuming that the population variance is not known, we can replace it with an unbiased estimator $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n-2} = \dfrac{\sum_{i=1}^{n} \varepsilon_i{}^2}{n-2}$ and again obtain the interval through a $t$-distribution with $n-2$ degrees of freedom.

# Prediction Intervals vs Confidence Intervals

- In general, prediction intervals will always be wider than confidence intervals, since the confidence interval simply describes our uncertainty in estimating the parameters.

- On the other hand, prediction intervals capture our uncertainty in estimating the parameters and the random variation between realisations of the resulting distribution.