Chapter 7
Parameter Estimation

ESTIMATION PROBLEMS

Two types of Statistical Inference

→ Estimation of population parameters (Chapter 9)
ex: A manufacturer of a certain electronics component wishes to <u>estimate</u> the true life-time of components from a sample of size n.

→ Hypothesis Testing (Chapter 10)
ex: The same manufacturer claims that the life-time of the components have a mean 5 years. He takes a sample of size n to <u>test this hypothesis</u>.

———— o ————

<u>Defn</u>: A <u>point estimate</u> of some population parameter θ is a single value $\hat{\theta}$ of a statistic $\hat{\Theta}$.

Ex: the value $\bar{x}$ of the statistic $\bar{X}$, computed from a sample of size n is a point estimate of the parameter μ.

<u>Defn</u>: A statistic $\hat{\Theta}$ is said to be an <u>unbiased</u> estimator of the parameter θ if

$$\mu_{\hat{\Theta}} = E[\hat{\Theta}] = \theta$$

Example : $\mu_{\bar{X}} = \mu$ hence $\bar{X}$ is an unbiased estimator of μ.

Example : S^2 is an unbiased estimator of the population parameter σ^2

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

$$= \sum_{i=1}^{n} (X_i - \mu)^2 - 2(\bar{X} - \mu) \underbrace{\sum_{i=1}^{n} (X_i - \mu)}_{n(\bar{X} - \mu)} + n(\bar{X} - \mu)^2$$

$$= \sum_{i=1}^{n} (X_i - \mu)^2 - n(\bar{X} - \mu)^2 \quad (*)$$

$$E(S^2) = E\Big[\frac{1}{n-1}\underbrace{\sum_{i=1}^{n}(X_i-\bar{X})^2}_{\text{Substitute } \circledast \text{ for this}}\Big]$$

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

$$= \frac{1}{n-1}\Big(\sum_{i=1}^{n}\underbrace{E[(X_i-\mu)]^2}_{\sigma^2_{X_i}=\sigma^2} - n\,\underbrace{E[(\bar{X}-\mu)^2]}_{\sigma^2_{\bar{X}}=\sigma^2/n}\Big)$$

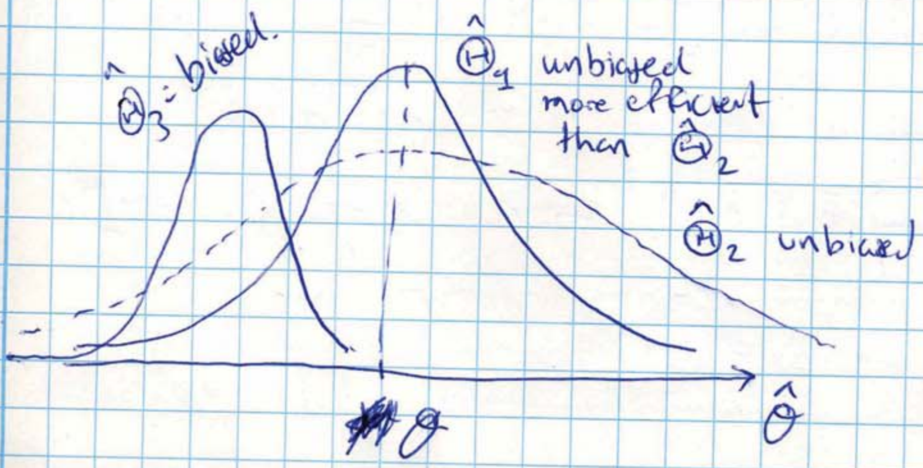
$$= \frac{1}{n-1}\Big(n\sigma^2 - \cancel{n}\frac{\sigma^2}{\cancel{n}}\Big) = \frac{1}{n-1}\big((n-1)\sigma^2\big) = \underline{\underline{\sigma^2}}$$

This is why the definition of S^2 has a division by $n-1$ instead of n!!

If $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are two unbiased estimators of the same population parameter θ, the one with the smaller variance is called the more efficient estimator.

__Ex:__ If $\sigma^2_{\hat{\theta}_1} < \sigma^2_{\hat{\theta}_2}$ $\hat{\Theta}_1$ is a more efficient estimator of θ than $\hat{\Theta}_2$ and is preferable.

__Defn:__ Of all the unbiased estimators of some parameter θ, the one with smallest variance is called the most efficient estimator of θ.



__Note:__ For a given estimator, increasing the sample size decreases the variance.

Defn: Interval Estimation

$$P(\hat{\Theta}_L < \theta < \hat{\Theta}_U) = 1-\alpha$$

This is called a $100(1-\alpha)\%$ confidence interval

Interpretation: Different samples yield different values of $\hat{\Theta}$. We find confidence limits $\hat{\Theta}_L$ and $\hat{\Theta}_U$ such that the true population parameter θ is within those limits with probability $1-\alpha$.

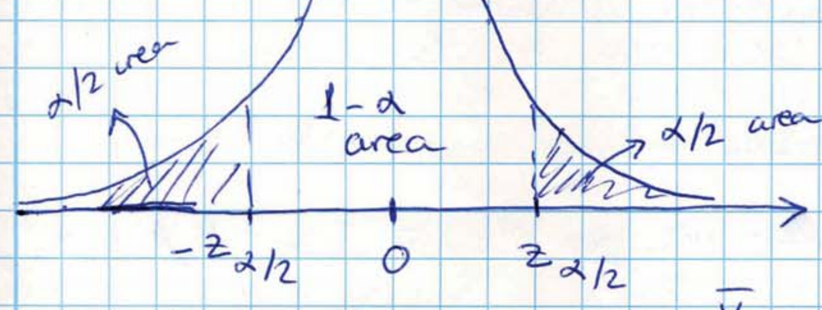
1. $\alpha = 0.05 \rightarrow 95\%$ confidence limits
2. $\alpha = 0.01 \rightarrow 99\%$ confidence "

Generally the range for the confidence limits in case 2 will be wider than the range in 1

ESTIMATING THE MEAN OF A SINGLE SAMPLE

We first study the simpler but unrealistic case where we are trying to estimate μ and σ is known.

The sampling distribution of $\bar{X}$ is centred at μ. Its variance is σ^2/n as we learned previously.



$z_{\alpha/2}$ is the value for which

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$$

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

We learned previously that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ has standard normal dist. for $n \geq 30$

$$\boxed{P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha}$$

multiply by $\sigma/\sqrt{n}$, subtract $\bar{X}$, multiply by -1 to get this.

Now we select a particular sample of size n and get a specific value of $\bar{x}$ then:

Confidence interval for μ; σ known

If $\bar{x}$ is the mean of a random sample of size n from a population with known variance σ^2, a $100(1-\alpha)\%$ confidence interval for μ is given by

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where $z_{\alpha/2}$ is the value from the standard normal distribution leaving an area of $\alpha/2$ to the right.

Note 1: $\hat{\Theta}_L = \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ $\quad \hat{\Theta}_U = \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

Note 2: We invoked the central limit theorem so we need the same assumptions.

Note 3: The larger n, the tighter the confidence interval.

Note 4. The smaller α, the wider " " ".

Example: A sample of 64 resistors from a production line are found to have a mean resistance of 206 Ohms. Find the 95% and 99% confidence intervals for the mean resistance of the population. Assume that the population standard deviation is 4 Ohms.

Solution: The point estimate of μ is $\bar{x} = 206$

(1) 95%: $1-\alpha = 0.95$ $\frac{\alpha}{2} = 0.025$. From table A.3 $z_{0.025} = 1.96$

(remember $z_{\alpha/2}$ leaves an area $\alpha/2$ to the right)

$$206 - 1.96\frac{4}{\sqrt{64}} < \mu < 206 + 1.96\frac{4}{\sqrt{64}}$$

$$\boxed{95\% \text{ confidence interval: } 205.02 < \mu < 206.98}$$

② 99% : $1-\alpha = 0.99$ $\frac{\alpha}{2} = 0.005$ $z_{\alpha/2} = 2.575$ (Table A.3)

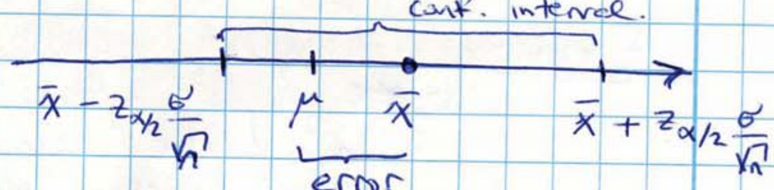
$$206 - 2.575\frac{4}{\sqrt{64}} < \mu < 206 + 2.575\frac{4}{\sqrt{64}}$$

$$\boxed{204.71 < \mu < 207.29}$$ 99% confidence interval

Notice this is wider than the 95% interval.

Also note if we want tighter confidence intervals we should increase n. (sample size)

Theorem: If $\bar{x}$ is used as an estimate of μ, we can be $100(1-\alpha)$% confident that the error will not exceed $z_{\alpha/2}\,\sigma/\sqrt{n}$.



Theorem: If $\bar{x}$ is used as an estimate of μ we can be $100(1-\alpha)$% confident that the error will not exceed a specified amount e when the sample size is

$$n = \left(\frac{z_{\alpha/2}\,\sigma}{e}\right)^2 \text{ rounded up.}$$

Example: How large a sample size is required if in our previous example we want to be 95% confident that our estimate of μ (mean resistance of population) is off by less than 0.1?

$z_{\alpha/2} = 1.96$ for 95% confidence interval ($\frac{\alpha}{2} = 0.025$)

So $n = \left(\frac{1.96 \times 4}{0.1}\right)^2 = 6146.56$ (4 ← pop σ)

round up $n = 6147$.

One-sided confidence bands:

Sometimes we are interested in questions of the form "What is the probability that the mean life-time of a component is at least 2 years?" (worst case scenarios)

$$\underbrace{P\left(\mu < \bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}}\right)}_{} = 1-\alpha$$

one-sided upper band for $100(1-\alpha)\%$ confidence for μ when $\bar{X}$ is the mean of a sample of size n from a population of variance σ^2

$$P\left(\underbrace{\bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}}} < \mu\right) = 1-\alpha$$

one-sided lower band.

Notice that z_α appears in the equation rather than $z_{\alpha/2}$.

<u>Example</u>: A quality control engineer takes a sample of 100 light bulbs from a production line and finds the sample mean life time to be 480 hours. The population standard deviation is known to be 25 hours. Find a lower band for the 95% confidence for the population mean.

$$\alpha = 0.05 \qquad z_\alpha = 1.645$$

$$\bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}} = 480 - 1.645\,\frac{25}{10} = 475.9$$

$$\boxed{P(475.9 < \mu) = 0.95}$$

<u>Example</u>: A clean room for chip manufacturing has to limit the number of particles found per volume. In an university clean room ~~sample~~ air samples are taken at 36 different time points and the mean number of particles per cubic foot is found as 105. Find an upper band for the 95% confidence for the population mean.

$$\alpha = 0.05 \qquad z_\alpha = 1.645$$

$\boxed{\text{Assume pop. } \sigma = 12.}$

$$\bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}} = 105 + 1.645\,\frac{12}{6} = 108.29$$

$$\boxed{P(\mu < 108.29) = 0.95}$$

Unknown σ

Usually when we are trying to estimate μ, σ is also unknown. From Chapter 8, if we have a random sample from a normal distribution, then the random variable

$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$ has a t-distribution with $n-1$ degrees of freedom.

σ (population standard dev) is unknown, but is replaced with S (sample standard dev)

Similar to before $P\left(-t_{\alpha/2} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2}\right) = 1-\alpha$

with $t_{\alpha/2}$ being the t-value (Table A.4) for $v = n-1$ degrees of freedom above which we can find an area of $\alpha/2$. The difference from before is the use of t-distribution (Table A-4) rather than the standard normal dist.

Confidence interval for μ; σ unknown

If $\bar{x}$ and s are the mean and standard deviation of a random sample of size n from a normal ~~distributed~~ population, a $100(1-\alpha)\%$ confidence interval for μ is

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

where $t_{\alpha/2}$ is the t-value with $v = n-1$ degrees of freedom leaving an area of $\alpha/2$ to the right.

One-sided $100(1-\alpha)\%$ bounds are:

$\bar{x} + t_{\alpha}\frac{s}{\sqrt{n}}$ upper bound

$\bar{x} - t_{\alpha}\frac{s}{\sqrt{n}}$ lower bound

Note t_α instead of $t_{\alpha/2}$.

Example: Assume that electrical potential measurements made at a particular node in a circuit are normally distributed (due to error in the measuremen
Ten measurements are made; finding:

8.95, 9.6, 10.7, 9.45, 10.5, 10.05, 10.7, 9.0, 9.9, 9.4

Find a 95% confidence interval for the true mean voltage.

Soln: $\bar{x} = 9.825 \quad s = 0.655 \quad \alpha = \frac{1-0.95}{2}$

From Table A-4 with $v = 10-1 = 9$ degrees of freedom $t_{0.025} = 2.262$

Then, the 95% confidence interval for μ is

$$9.825 - 2.262\frac{0.655}{\sqrt{10}} < \mu < 9.825 + 2.262\frac{0.655}{\sqrt{10}}$$

$$9.3565 < \mu < 10.2935$$

In other words $\boxed{P(9.3565 < \mu < 10.2935) = 0.95}$

Example: Assume that the internet connection speed at your house is normally distributed. You take a sample of 15 connection speeds at different times and find that the sample mean $\bar{x} = 2.3$ Mbps and $s = 0.5$ Mbps. Find the 99% lower-bound for the true mean.

Soln $\alpha = 0.01 \quad v = 15-1 = 14$ Table A-4 $t_{0.01} = 2.624$

$$\text{lower bound} = \bar{x} - t_\alpha \frac{s}{\sqrt{n}} = 2.3 - 2.624\frac{0.5}{\sqrt{15}} = 1.9612$$

In other words $\boxed{P(1.9612 < \mu) = 0.99}$

Estimating the Difference Between Two means

σ_1 and σ_2 known:

$\bar{X}_1 - \bar{X}_2$ is a point estimator of $\mu_1 - \mu_2$

Central Limit Theorem: $\bar{X}_1 - \bar{X}_2$ has normal distribution ~~standard deviation~~ with mean $\mu_1 - \mu_2$ and standard deviation $\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$ if both $n_1, n_2 \geq 30$ (or underlying population distributions normal)

$$P\left(-z_{\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < z_{\alpha/2}\right) = 1 - \alpha$$

<u>$100(1-\alpha)\%$ confidence interval</u> for $\mu_1 - \mu_2$

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where $z_{\alpha/2}$ is the z-value leaving an area $\alpha/2$ to the right.

Variances unknown, but known to be equal

Pooled estimate of variance $S_p^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}}$$

has a t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

If $\bar{x}_1$ and $\bar{x}_2$ are the means of independent random samples of sizes n_1 and n_2 from approximately normal populations with unknown but equal variances, the $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}\, s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}\, s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where s_p is the pooled estimate of the standard deviation and $t_{\alpha/2}$ is the t-value with $v = n_1 + n_2 - 2$ degrees of freedom leaving an area of $\alpha/2$ to the right.

Assume that the population variances are equal.

Example: Two manufacturing processes for an electrical component. Independent samples taken from both to asses the difference in life-time.

Sample 1 : $n_1 = 72$, $\bar{x}_1 = 3.4$, $s_1 = 0.5$

Sample 2 : $n_2 = 50$, $\bar{x}_2 = 3.8$, $s_2 = 0.6$

Find a 90 % confidence interval for $\mu_1 - \mu_2$, the difference of the population mean life-times.

Soln : $\bar{x}_1 - \bar{x}_2 = 3.4 - 3.8 = 0.4$

pooled variance $s_p^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$

$= \dfrac{71 \times 0.5^2 + 49 \times 0.6^2}{72 + 50 - 2}$

$= 0.2949$

$s_p = \sqrt{s_p^2} = 0.543$

90% confidence interval , $\alpha = 0.1$

$v = n_1 + n_2 - 2 = 72 + 50 - 2 = 120$

$t_{0.05} = \cancel{1.645}$ 1.658 (Table A.4)

$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

$-0.4 - \cancel{1.645}\ 1.658 \times 0.543 \times 0.184 < \mu_1 - \mu_2 < -0.4 + \cancel{1.645}\ 1.658 \times 0.543 \times 0.184$

-0.5656 $\cancel{0.2356} < \mu_1 - \mu_2 < \cancel{0.5644}$ -0.2343

with confidence 90 %

Estimating a single Sample variance

S^2 is a point estimator of σ^2.

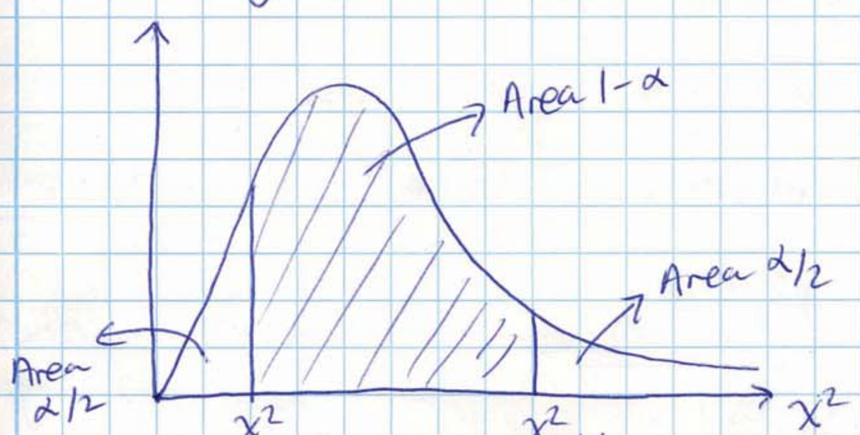
Let $X^2 = \dfrac{(n-1)S^2}{\sigma^2}$ Chi-squared distribution with $n-1$ degrees of freedom

$$P\left(\chi^2_{1-\alpha/2} < X^2 < \chi^2_{\alpha/2}\right) = 1-\alpha$$

$$P\left(\chi^2_{1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}\right) = 1-\alpha$$

$$P\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha$$

where $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ are the values of the chi-squared distribution with $n-1$ degrees of freedom leaving areas $1-\alpha/2$ and $\alpha/2$ to the right, respectively



If s^2 is the variance of a random sample of size n from a <u>normal population</u>, the $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

where $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ are the chi-squared values with $v = n-1$ degrees of freedom, leaving areas $\alpha/2$ and $1-\alpha/2$ to the right, respectively.

Example: A sample has the observations:

46.4, 46.1, 45.8, 47.0, 46.1, 45.9, 45.8, 46.9, 45.2 and 46.0.

Find a 95% confidence interval for the population variance σ^2.

Soln: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

or $s^2 = \frac{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}{n(n-1)}$

$n = 10 \quad s^2 = 0.2862$

95% confidence interval $\alpha = 0.05$

$v = 10 - 1 = 9$ degrees of freedom

From Table A.5 $\chi^2_{0.025} = 19.023$

$\chi^2_{0.975} = 2.7$

Note the lack of symmetry unlike the normal and t-distributions.

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

$$\frac{9 \times 0.286}{19.023} < \sigma^2 < \frac{9 \times 0.286}{2.7}$$

$$\boxed{0.135 < \sigma^2 < 0.953}$$ 95% confidence