

ESTIMATION PROBLEMS

Two types of Statistical Inference

Estimation of population parameters (Chapter 9)
ex: A manufacturer of a certain electronics component wishes to estimate 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 test this hypothesis.

Defn: A point estimate of some population parameter θ is a single value 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 .

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

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

Example: $E[\bar{x}] = \mu$ hence \bar{X} is an unbiased estimator of μ .

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

Proof:

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n [(x_i - \mu) - (\bar{x} - \mu)]^2 \\ &= \sum_{i=1}^n (x_i - \mu)^2 - 2(\bar{x} - \mu) \sum_{i=1}^n (x_i - \mu) + n(\bar{x} - \mu)^2 \\ &= \sum_{i=1}^n (x_i - \mu)^2 - n(\bar{x} - \mu)^2 \quad (*) \end{aligned}$$

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

Substitute \star for this

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

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

$\underbrace{\sigma_x^2}_{\sigma_x^2 = \sigma^2}$ $\underbrace{\sigma_{\bar{X}}^2}_{\sigma_{\bar{X}}^2 = \sigma^2/n}$

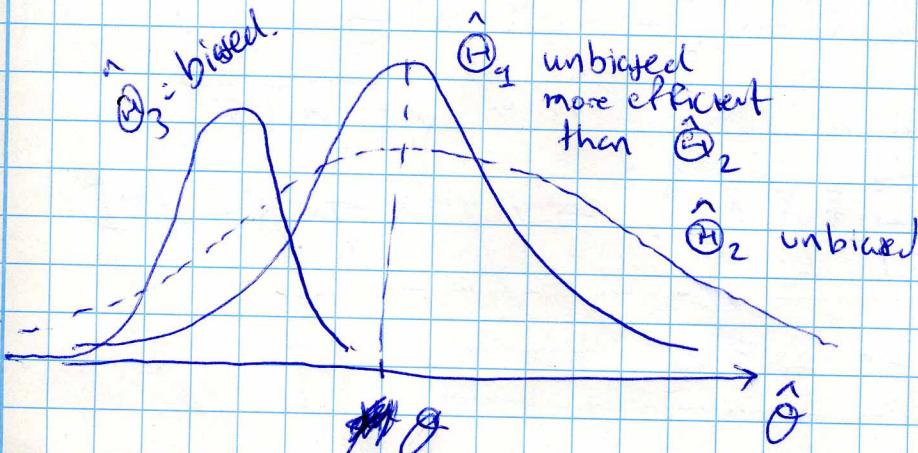
$$= \frac{1}{n-1} \left(n\sigma^2 - n \frac{\sigma^2}{n} \right) = \frac{1}{n-1} ((n-1)\sigma^2) = \underline{\underline{\sigma^2}}$$

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 $\hat{\Theta}_1^2 < \hat{\Theta}_2^2$ $\hat{\Theta}_1$ 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

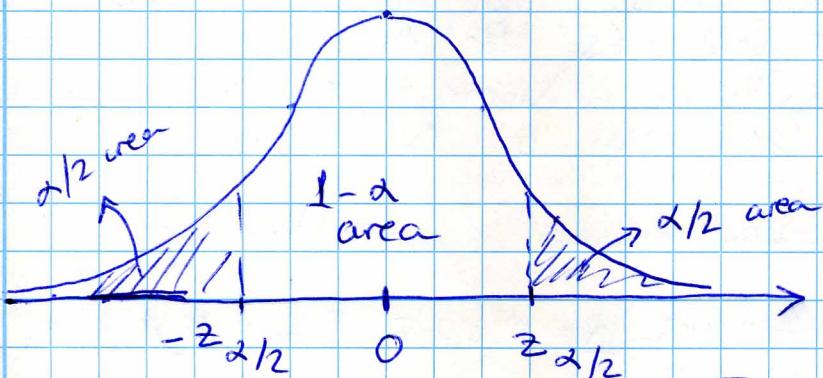
(2) $\alpha = 0.01 \rightarrow 99\%$ confidence "

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 $n \geq 30$

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

Multiply by σ/\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 : $\hat{H}_L = \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ $\hat{H}_U = \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

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)

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

95% confidence interval : $205.02 < \mu < 206.98$

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

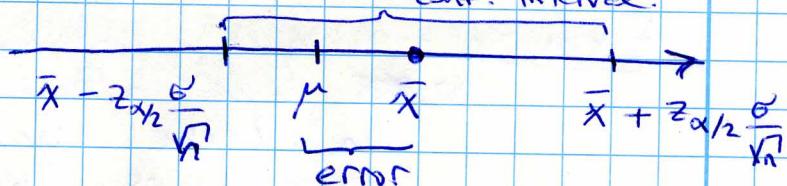
$$206 - 2.575 \frac{4}{\sqrt{64}} < \mu < 206 + 2.575 \frac{4}{\sqrt{64}}$$

$$204.71 < \mu < 207.29 \quad \xrightarrow{\text{99\% confidence interval}}$$

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 $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$.



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?

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

$$\text{so } n = \left(\frac{1.96 \times 4}{0.1} \right)^2 = 6146.56$$

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)

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

one-sided upper bound 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

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

one-sided lower bound.

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

Example: A quality control engineer takes a sample of 100 ~~light~~ 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 bound for the 95% confidence for the population mean.

$$\alpha = 0.05 \quad z_{\alpha} = 1.645$$

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

$$P(475.9 < \mu) = 0.95$$

Example: 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 bound for the 95% confidence for the population mean.

$$\alpha = 0.05 \quad z_{\alpha} = 1.645$$

Assume pop. $\sigma = 12$

$$\bar{X} + z_{\alpha} \frac{\sigma}{\sqrt{n}} = 105 + 1.645 \frac{12}{\sqrt{36}} = 108.29$$

$$P(\mu < 108.29) = 0.95$$

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)

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

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 distribution 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:

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

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

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 measurement). 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 $P(9.3565 < \mu < 10.2935) = 0.95$

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

$$\text{lower bound} = \bar{x} - t_\alpha \frac{s}{\sqrt{v}} = 2.3 - 2.624 \frac{0.5}{\sqrt{14}} = 1.9612$$

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

Estimating the Difference Between Two means

σ_1^2 and σ_2^2 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)

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

100(1 - α)% confidence interval 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 $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}$

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

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 assess 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.

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

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

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

$$= 0.2945$$

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

90% confidence interval, $\alpha = 0.1$

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

$$t_{0.05} = 1.645 \quad (\text{Table A-4})$$

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

$$-0.4 - 1.645 \times 0.543 \times 0.184 < \mu_1 - \mu_2 < -0.4 + 1.645 \times 0.543 \times 0.184$$

$$-0.5656 < \mu_1 - \mu_2 < 0.5644$$

with confidence 90%

$$-0.2343$$

Estimating a single Sample variance

S^2 is a point estimator of σ^2 .

$$\text{Let } X^2 = \frac{(n-1)S^2}{\sigma^2}$$

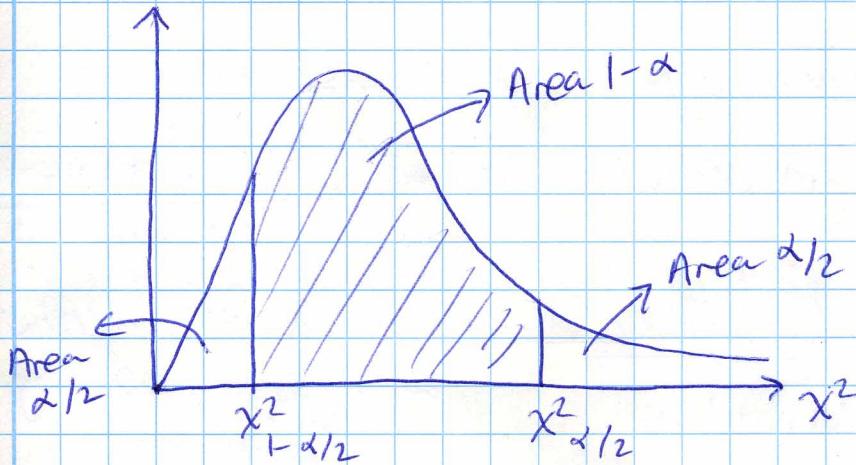
Chi-squared distribution with $n-1$ degrees of freedom

$$P(X_{1-\alpha/2}^2 < X^2 < X_{\alpha/2}^2) = 1 - \alpha$$

$$P(X_{1-\alpha/2}^2 < \frac{(n-1)S^2}{\sigma^2} < X_{\alpha/2}^2) = 1 - \alpha$$

$$P(\frac{(n-1)S^2}{X_{\alpha/2}^2} < \sigma^2 < \frac{(n-1)S^2}{X_{1-\alpha/2}^2}) = 1 - \alpha$$

where $X_{1-\alpha/2}^2$ and $X_{\alpha/2}^2$ 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 normal population, the $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\frac{(n-1)S^2}{X_{\alpha/2}^2} < \sigma^2 < \frac{(n-1)S^2}{X_{1-\alpha/2}^2}$$

where $X_{\alpha/2}^2$ and $X_{1-\alpha/2}^2$ 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 .

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

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

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

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.

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

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

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