3Y03 - Probability and Statistics for Engineering

WS19 Lecture 21

Yesterday Fix parameter θ

Estimator $\hat{\Theta}$ of θ is

- **biased** if bias of $\hat{\Theta} = E(\hat{\Theta}) - \theta$ is not zero
- **unbiased** if no bias i.e. $E(\hat{\Theta}) = \theta$.

If $\hat{\Theta}$ unbiased:
- **standard error** is $\sigma_{\hat{\Theta}} = \sqrt{V(\hat{\Theta})}$ ← the one with the smallest = **M**inimum **V**ariance **U**nbiased **E**stimator for θ
- **estimated standard error**: same, with any unknown parameters involved estimated from data

For any estimator $\hat{\Theta}$ of θ, the **M**ean **S**quare **E**rror is $MSE(\hat{\Theta}) = E((\hat{\Theta}-\theta)^2)$

$= V(\hat{\Theta}) + (\text{bias})^2$

Relative efficiency: $\dfrac{MSE(\hat{\Theta})}{MSE(\hat{\Theta}_1)}$ ← $\hat{\Theta}$ **optimal** if <1 for all $\hat{\Theta}_1$

7_2 Central Limit Theorem

$\overline{X} = \frac{1}{n}(X_1 + \ldots + X_n)$ If $\{X_1, \ldots, X_n\}$ have mean μ, var. σ^2

↳ has mean μ, var. $\dfrac{\sigma^2}{n}$.

If X_i Normally distr., then so is $\overline{X}$.

What if X_i not (known to be) Norm. distr.?

Central Limit Theorem If $\{X_1, \ldots, X_n\}$ is random sample, mean μ, var. σ^2 (finite), then as n gets big (i.e. $n \longrightarrow \infty$) the limiting form of the distr. of

$$\frac{\bar{X}-\mu}{(\sigma/\sqrt{n})} \text{ is } N(0,1).$$

i.e. $\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$ as long as n big enough.

Big enough depends on assumptions but for practical purposes use $n \geqslant 30$.

<u>Example</u> Pipes manufactured with mean diameter 3.2cm, standard dev. 0.01 cm. Find the prob. that a random sample of 40 pipes will have sample mean between 3.199cm & 3.202 cm.

<u>Solution</u> Estimate with $Z = \frac{\bar{X}-\mu}{(\sigma/\sqrt{n})} \sim N(0,1)$

$\mu = 3.2$, $\sigma = 0.01$

$n = 40$

$\geqslant 30$

Want $P(3.199 < \bar{X} < 3.202)$

$$= P\left(\frac{3.199-3.2}{0.01/\sqrt{40}} < Z < \frac{3.202-3.2}{0.01/\sqrt{40}}\right)$$

$$= P(-0.63 < Z < 1.26)$$

$$= P(Z<1.26) - P(Z<-0.63)$$

$$= P(Z<1.26) - \left(1 - P(Z<0.63)\right)$$

(table) $= 0.896165 - (1 - 0.735653) = 0.631818.$

Difference of Sample Means

2 populations : means μ_1 , μ_2
variances σ_1^2 , σ_2^2

Random sample from each : $\{X_1, \ldots, X_{n_1}\}$
$\{Y_1, \ldots, Y_{n_2}\}$

If $n_1 \geqslant 30$, $\bar{X} \approx N(\mu_1, \frac{\sigma_1^2}{n_1})$

If $n_2 \geqslant 30$, $\bar{Y} \approx N(\mu_2, \frac{\sigma_2^2}{n_2})$

Now look at difference $\bar{Y} - \bar{X}$: has mean $\mu_2 - \mu_1$

& variance : $V(\bar{Y} - \bar{X}) = V(\bar{Y}) + V(\bar{X})$ if samples $\{X_1, \ldots, X_{n_1}\}$ & $\{Y_1, \ldots, Y_{n_2}\}$ are taken independently (so $\bar{X}$ & $\bar{Y}$ independent)

So assume that! So $V(\bar{Y} - \bar{X}) = \frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}$.

& $\bar{X}, \bar{Y}$ Norm distr. $\Rightarrow$ $\bar{Y} - \bar{X}$ Norm. distr. (see 5-4).

<u>Example</u> Random sample taken from each of 2 sets of steel plates

Set 1 : manufactured with old method

Set 2 : " " new "

Old: $n_1 = 30$, $\mu_1 = 11 \text{ kg/cm}^2$, $\sigma_1 = 1.5 \text{ kg/cm}^2$

New: $n_2 = 35$, $\mu_2 = 15 \text{ kg/cm}^2$, $\sigma_2 = 1 \text{ kg/cm}^2$.

If $\bar{X}$ = sample mean of old find $P(\bar{Y} - \bar{X} > 3.5)$.

$\bar{Y}$ = " " " new

<u>Solution</u> By above $\bar{Y} - \bar{X} \approx N\left(\mu_2 - \mu_1, \frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}\right)$ as $n_1, n_2 \geq 30$.

$$P(\bar{Y} - \bar{X} > 3.5) = P\left(Z > \frac{3.5 - (\mu_2 - \mu_1)}{\sqrt{\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}}}\right)$$

$$= P\left(Z > \frac{3.5 - (15 - 11)}{\sqrt{\frac{1^2}{35} + \frac{(1.5)^2}{30}}}\right)$$

$$= P(Z > -1.55) = P(Z < 1.55)$$

$$= 0.939429.$$

8 Confidence Intervals

You have $\hat{\Theta}$ for θ. Given data set $\{x_1, \ldots, x_n\}$
(estimator) (parameter)
$\longrightarrow$ get $\hat{\theta}$ estimate.

Q: How close to true value θ is estimate $\hat{\theta}$?

If you repeat & get a new data set, new value for $\hat{\theta}$ again & again, you could use values for $\hat{\theta}$ to build up an idea of true value

or range of plausible values for θ

↳ = likely . Can do all this

with one data set :

<u>Def^n</u> A <u>100(1−α)% confidence interval</u> (C.I.)

for θ is an interval $(L(x_1,\dots,x_n), R(x_1,\dots,x_n))$

s.t. $P(L \leq \theta \leq R) = 1-\alpha$

α = confidence level.

We say we're $100(1-\alpha)\%$ confident that the true value of θ lies in (L, R).

Usually want $90\%, 95\%, 99\%, \dots$ ← high percentages (= high confidence)

i.e. $\alpha = 0.1, 0.05, 0.01, \dots$ ← low α

To find a C.I., depends on distrib. & info. already known.

<u>C.I. for the mean of a Normal pop., Variance <u>known</u></u>

unrealistic!

But this gives basic ideas.

Use $\bar{X}$ to estimate μ, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

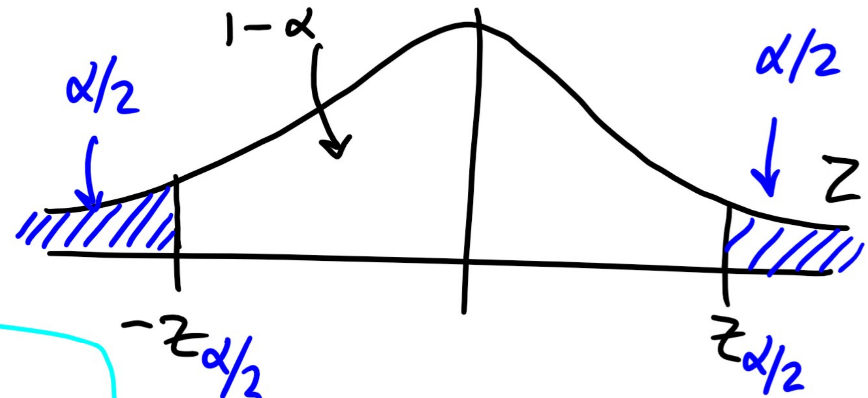
The recipe we use to find L, R with

$$P(L < \mu < R) = 1 - \alpha$$

is to make (L, R) symmetric with $\bar{X}$ in the middle.

i.e. Find z with $P(-z < Z < z) = 1 - \alpha$

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



$$\boxed{\begin{array}{l} z_\gamma = z\text{-value} \\ \quad \text{with } P(Z < z_\gamma) = 1 - \gamma \end{array}}$$

We want $z_{\alpha/2}$ with $P\left(-z_{\alpha/2} < \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\frac{\alpha}{2}}\right)$

$= 1 - \alpha$

So a $100(1-\alpha)\%$ C.I. for μ is given by

actual sample mean computed from data

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

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

$(\bar{x} - z_{\alpha/2}\,\sigma/\sqrt{n},\ \bar{x} + z_{\alpha/2}\,\sigma/\sqrt{n}).$