3Y03 - 3J04 PROBABILITY & STATISTICS FOR (CIVIL) ENGINEERING

C01 - Lecture 21

Last time Parameters, Statistics, Estimators & Estimates

- **Parameters**: A quantity (perhaps associated with the distribution) e.g. proportion p, mean μ, variance σ^2
- **Statistics**: A function of the sample $h(X_1, \ldots, X_n)$ e.g. $\sum_{i=1}^{n} X_i^2$
- **Estimators**: A statistic used to estimate a parameter e.g. $\overline{X} = \frac{1}{n}(X_1 + \ldots + X_n)$
- **Estimates**: output of an estimator e.g. $\overline{x}$

Central Limit Theorem

If $\{X_1, \ldots, X_n\}$ is a random sample & underlying distribution has mean μ, variance σ^2, then for "large" n, $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$.

sample mean

If underlying distribution looks "almost" normal, say symmetric & "unimodal"

1 mode

then $n = 4, 5$ is large enough

But in general more skewed $\Rightarrow$ bigger n required but $n > 30$ almost always enough.

Example Pipes manufactured with mean diameter 3.2 cm & standard deviation 0.01 cm. Find the probability that a random sample of 10 pipes will have sample mean $\overline{X}$ between 3.199 cm and 3.202 cm.

Solution We can estimate this with $Z = \dfrac{\overline{X} - \mu}{(\sigma/\sqrt{n})} \sim N(0,1)$ because of CLT. We have $\mu = 3.2$, $\sigma = 0.01$, $n = 10$.

So $P(3.199 < \overline{X} < 3.202) =$

$$P\left(\frac{3.199-3.2}{\left(0.01/\sqrt{10}\right)} < Z < \frac{3.202-3.2}{\left(0.01/\sqrt{10}\right)}\right)$$

$$= P(-0.32 < Z < 0.63)$$ ... use Standard Normal Table.

We can use trick about linear combinations of r.v.s (§ 5.4) to compare two populations:

Two populations with means μ_1, μ_2, variances σ_1^2, σ_2^2.

Take random sample $\{X_1, \dots, X_{n_1}\}$ from pop^n 1.

" " " $\{Y_1, \dots, Y_{n_2}\}$ " " 2.

As n_1, n_2 get "big", (see above)

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)}} \sim N(0,1).$$

by CLT.

See "linear combinations of random variables"

variance of $\bar{X}-\bar{Y}$

$$= 1^2 V(\bar{X}) + (-1)^2 V(\bar{Y})$$

$$= V(\bar{X}) + V(\bar{Y})$$

$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

<u>Example</u> 2 sets of steelplates

First set manufactured via existing process

Second set " " new process.

& tested for strength.

Existing method $\{X_1, \dots, X_{30}\}$ has $\mu_1 = 11\ kg/cm^2$
$n_1 = 30$ $\sigma_1 = 1.5\ kg/cm^2$

New method $\{Y_1, \dots, Y_{35}\}$ has $\mu_2 = 15\ kg/cm^2$
$n_2 = 35$ $\sigma_2 = = 1\ kg/cm^2$

Find $P(\bar{Y} - \bar{X} > 3.5)$.

Solution From above $Z = \dfrac{(\bar{Y} - \bar{X}) - (\mu_2 - \mu_1)}{\sqrt{\dfrac{\sigma_2^2}{n_2} + \dfrac{\sigma_1^2}{n_1}}} \sim N(0,1)$

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

$$= P(Z > -1.55)$$ → use Standard Nomal Table.

7.3 Concepts of Point Estimation

How do we know how good an estimator $\hat{\Theta}$ is (of parameter θ)?

Bias $\hat{\Theta}$ is a function of $X_1, \dots, X_n$ so we can find /make sense of $E(\hat{\Theta})$, expected value.

If $E(\hat{\Theta}) = \theta$, then $\hat{\Theta}$ is an unbiased estimator for/ of θ.

"true" value of parameter

If $E(\hat{\Theta}) \neq \theta$, then $\hat{\Theta}$ is <u>biased</u> and

$$E(\hat{\Theta}) - \theta = \underline{\text{bias}} \text{ of } \hat{\Theta}.$$

<u>Examples</u> $\bar{X}$ is an unbiased estimator for μ (= mean of X_is) (So is the median.) S^2 is an unbiased estimator for σ^2 (= variance of X_is).

<u>Why?</u> $E(\bar{X}) = E\left(\frac{1}{n}(X_1 + \dots + X_n)\right) = \dots = \mu$ (see 5.4).

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

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} E(X_i^2) - n\,E(\bar{X}^2)\right) \qquad (5.4)$$

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

$$\boxed{\begin{aligned} V(Y) &= E(Y^2) - (E(Y))^2 \\ E(Y^2) &= V(Y) + (E(Y))^2 \end{aligned}}$$

Rearrange these & plug back in, with $Y = X_i$ or $\bar{X}$ as appropriate

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

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

<u>Non-Example</u> S <u>not</u> an unbiased estimator for σ.

<u>Example</u> (Trimmed/Truncated Mean) $\bar{X}_{tr(k)}$ is

the $k\%$ trimmed mean : calculate mean <u>after</u> discarding top $k\%$ and bottom $k\%$ of sample.

This is also an unbiased estimator for μ if the underlying distribution is symmetric.

<u>Comparison of Estimators</u> – which is the "best"?

e.g. which of unbiased estimators $\bar{X}$, $\bar{X}_{tr(k)}$ or median is best for μ?

Method/Idea (for unbiased estimators) is the one with the smallest variance is best – least likely to be far from expected value.

→ The smaller the variance, the more likely an estimator $\hat{\Theta}$ is to produce an estimate $\hat{\theta}$ close to the true value of θ

<u>Example</u> Could estimate μ with X_i or $\bar{X}$.

Yes, the ith observation, whatever it happens to be.

$$E(\bar{X}) = \mu \, , \, E(X_i) = \mu$$

$$V(\bar{X}) = \frac{\sigma^2}{n} \, , \, V(X_i) = \sigma^2$$

$$V(\bar{X}) < V(X_i) \quad \text{if } n \geq 2.$$

For a parameter θ, the unbiased estimator of θ with smallest variance among all unbiased estimators of θ is called MVUE = Minimum Variance Unbiased Estimator.

Important Example If random sample comes from Normal distribution mean μ the MVUE is $\bar{X}$.

Error / Precision Reframe question: given estimator $\hat{\Theta}$ for θ, how "good" is it?

Unbiased The standard error of $\hat{\Theta}$ is standard deviation of $\hat{\Theta}$ i.e. $\sigma_{\hat{\Theta}} = \sqrt{V(\hat{\Theta})}$

(also sometimes called "precision")

Problem: might be an expression that will involve other parameters! That we don't know!

Well, we can always estimate! TBC...