2013, Study Session # 3, Reading # 10

"SAMPLING & ESTIMATION"

SE = Standard Error
↑ = Rises
RV = Random Variable
CI = Class Interval

→ = Approaches to
df = Degrees Of Freedom
n = Sample Size

Sample

A subgroup of population.

Sample Statistic

- It describes the characteristic of a sample.
- Sample statistic itself is a random variable.

Methods of Sampling

Simple Random Sampling

- Each item of the population under study has equal probability of being selected.
- There is no guarantee of selection of items from a particular category.

Stratified Random Sampling

- Uses a classification system.
- Separates the population into strata (small groups) based on one or more distinguishing characteristics.
- Take random sample from each stratum.
- It guarantees the selection of items from a particular category.

Systematic Sampling

- Select every k^{th} number.
- Resulting sample should be approximately random

Sampling error

Sample – Corresponding Statistic Population Parameter.

Sampling Distribution

Probability distribution of all possible sample statistics computed from a set of equal size samples randomly drawn.

Standard Error (SE) of Sample Mean

- Standard deviation of the distribution of sample means.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- If σ is not known then;

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

- As n ↑; $\bar{x}$ approaches μ and S.E ↓.

Data	Time
Time series	Observations taken over equally spaced time interval
Cross- sectional	Single point estimate

Data	Observational Units	Characteristics
Longitudinal	Same	Multiple
Panel	Multiple	Same

Student's T-Distribution

- Bell shaped.
- Shape is defined by df
- df is based on 'sample size'.
- Symmetrical about it's mean.
- Less peaked than normal distribution.
- Has fatter tails.
- More probability in tails i.e., more observations are away from the centre of the distribution & more outliers.

Central Limit Theorem (CLT)

For a random sample of size 'n' with;

- population mean μ,
- finite variance (population variance divided by sample size) σ^2, the sampling distribution of sample mean $\overline{x}$ approaches a normal probability distribution with mean 'μ'& variance as 'n' becomes large.

Properties of CLT

- For n ≥ 30 ⇒ sampling distribution of mean is approx. normal.
- Mean of distribution of all possible samples = population mean 'μ'.
  - Variance of distribution $= \frac{\sigma^2}{n}$

CLT applies only when sample is random.

Point Estimate (PE)

- Single (sample) value used to estimate population parameter.

$$\bar{X} = \frac{\sum X}{n}$$

Estimator: Formula used to compute PE.

Confidence Interval (CI) Estimates

- Results in a range of values within which actual parameter value will fall.
- PE ±(reliability factor × SE).
- α= level of significance.
- 1- α= degree of confidence.

Desirable properties of an estimator

Unbiased

Expected value of estimator equals parameter e.g., $E(\overline{x})$ = μ i.e, sampling error is zero.

Efficient

If var $(\overline{x}_1)$ < var $(\overline{x}_2)$ of the same parameter then $\overline{x}_1$ is efficient than $\overline{x}_2$

Consistent

As n ↑, value of estimator approaches parameter & sample error approaches '0' e.g., As n → ∞ $\overline{x}$ → μ & SE → 0

Distribution		Variance		Sample		Test Statistic	
Normal	Non normal	Known	Unknown	Small (n<30)	Large (n≥30)	t	z
✓	×	✓	×	✓	×	×	✓
✓	×	✓	×	×	✓	×	✓
✓	×	×	✓	✓	×	✓	×
✓	×	×	✓	×	✓	✓*	×
×	✓	✓	×	✓	×	×	×
×	✓	✓	×	×	✓	×	✓
×	✓	×	✓	✓	×	×	×
×	✓	×	✓	×	✓	✓*	×

**The z-statistic is theoretically acceptable here, but use of the t-statistic is more conservative.*

Issues Regarding Selection of Appropriate Sample Size

As n ↑; s.e. ↓ & hence C.I is narrow.

Limitations of Large Sample Size

- Large sample may include observations from more than one population.
- Cost may increase more relative to an increase in precision.

Biases

Data Mining Bias

Statistical significance of the pattern is overestimated because the results were found through data mining.

Sample Selection Bias

- Systematically excluding some data from analysis.
- It makes the sample non-random.

Look –ahead Bias

Using sample data that wasn't available on the test date.

Time-period Bias

Time period over which the data is gathered is either too short or too long.

Data Mining

Using the same data to find patterns until the one that 'works' is discovered.

Survivorship Bias

- Most common form of sample selection bias.
- Excluding weak performances.
- Surviving sample is not random.

Warning Signs of Data Mining

- Evidence of testing many different, mostly unreported variables.
- Lack of economic theory consistent with empirical results.