Predictive Analytics

Statistical Variation & Regression Recap

Associate Professor Ole Maneesoonthorn

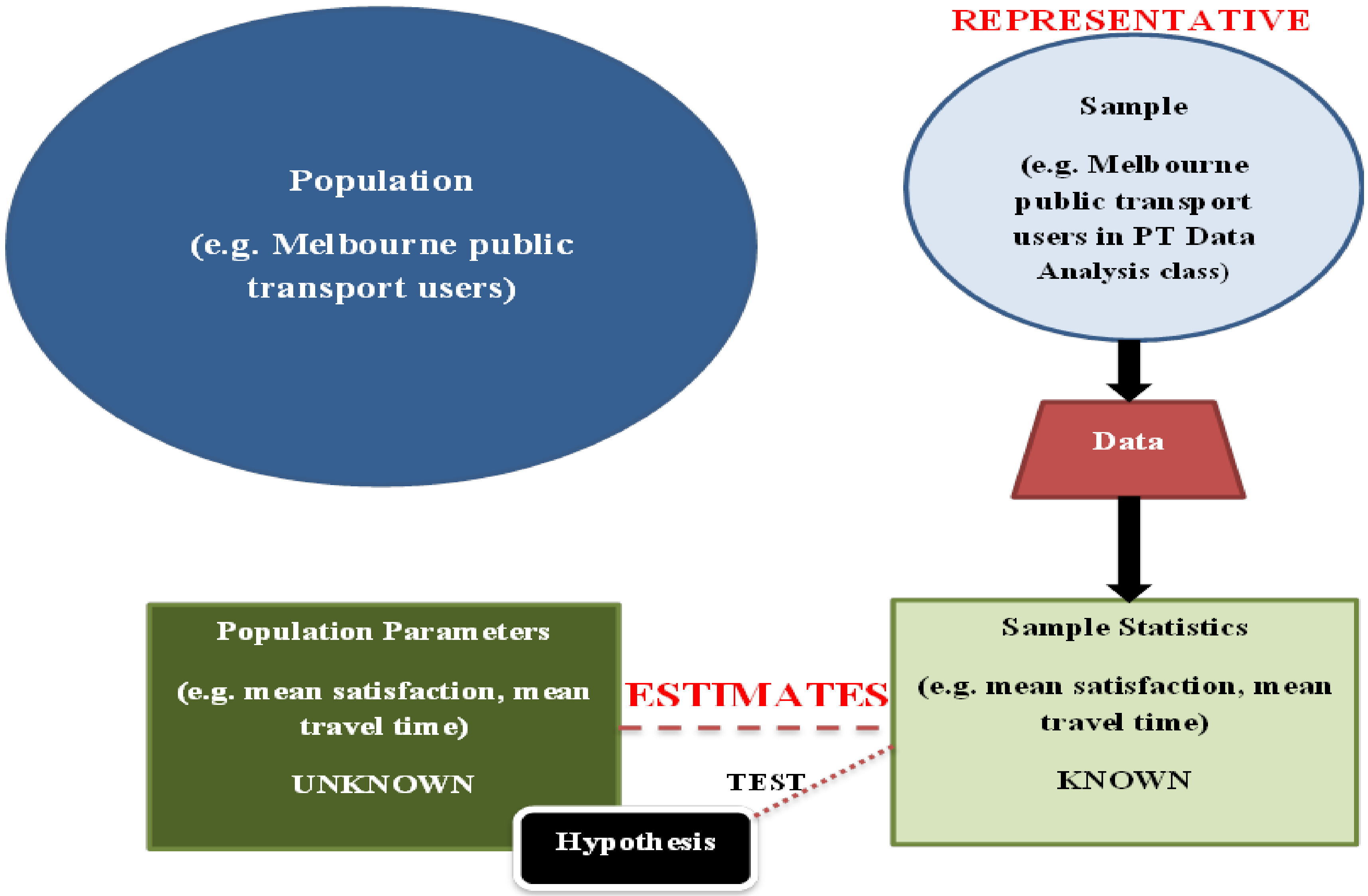
Associate Professor in Econometrics and Statistics

Melbourne Business School

O.Maneesoonthorn@mbs.edu

Statistical Variation

Population vs Sample



Sample Statistics

- The **population** mean μ is unknown
- From our **sample** of data $X_1, X_2, \dots, X_n$, we can compute an **estimate** **(statistic)**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- But you also know that if you are to collect another sample, you will get different data
- **→ $\bar{X}$ varies with the sample**

Sample Statistics

- Since $\bar{X}$ varies with the sample, we would like to know **how variable it is**
- This is quantified by the **standard error** of the sample mean

$$StdErr(\bar{X}) = \frac{s}{\sqrt{n}}$$

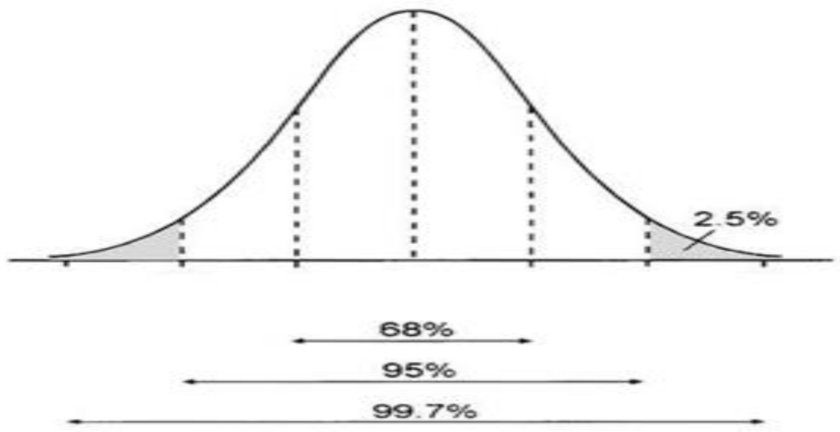
- Here s is the standard deviation of the data
- The **more variable the data, the more variable the sample mean**
- The more data points you have (n) the less variable the sample mean
- **More data → more accurate estimate**

Sampling Distributions

- In order to do **inference**, we need to know the **distribution** of the sample statistic

- Theoretically, the sample mean is **normally distributed** (if $n > 30$)
  - Its mean is located at the population mean μ
  - Its variation is defined by the standard error of the sample mean

- Inference?
  - **Confidence intervals** – gives you likely values of the population mean
  - **Hypothesis test** – establishing concrete evidence in favour of a certain hypothesis



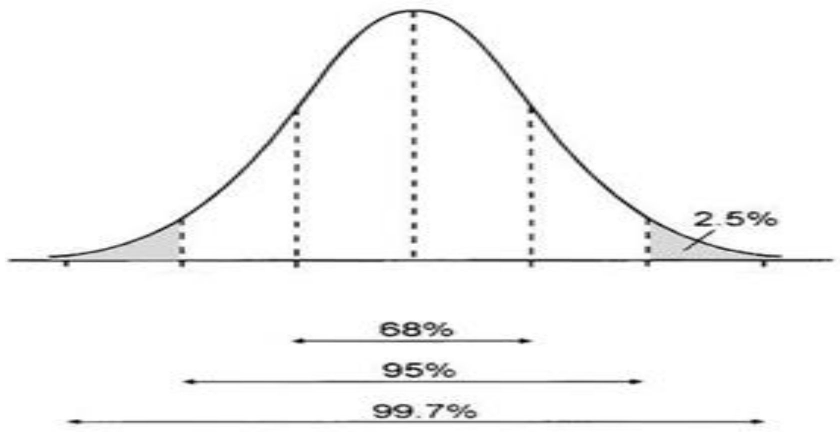
Statistical Inference

Confidence Intervals 5 Easy Steps:

1. Compute the sample mean
2. Compute the sample standard deviation
3. Compute the standard error of the sample mean
4. Choose the probability level (90%, 95%, 99%)
5. Compute the lower and upper bounds of the confidence interval

$$\text{Estimate} \pm Q \times \text{Std.error}$$

Coverage	90%	95%	99%
Q	1.65	1.96	2.58



Statistical Inference

Hypothesis testing - 5 easy steps:

1. Form the null hypothesis (“status quo”)
2. Form the alternative hypothesis (“suspected relation”)
3. Compute the sample mean, standard deviation and standard error
4. Compute the test statistic & p-value
5. Make a decision regarding the hypothesis

Statistical Inference

Hypothesis testing - the test statistic:

$$T = \frac{\bar{X} - \mu_0}{SE(\bar{X})}$$

The p-value depends on the definition of your *alternative* hypothesis

In the case of the alternative being $\mathrm{H}_a: \mu < \mu_0$

$$p - value = Pr(T < Tstat)$$

In the case of the alternative being $\mathrm{H}_a: \mu > \mu_0$

$$p - value = Pr(T > Tstat)$$

Statistical Inference

Hypothesis testing – decision rules

- Set your tolerance (significance) level
  - 5% is a typical value
  - (There is a 5% chance you will reject a true null hypothesis)
  - Decrease this number if you wish to be more conservative

- Compare your p-value to the tolerance level
  - If p-value is smaller than the tolerance level → reject the null hypothesis. There is **strong evidence in favour of** the alternative hypothesis.
  - If p-value is larger than the tolerance level → fail to reject the null hypothesis. There is **not enough evidence to support** the alternative hypothesis.

- **Notice: our final conclusion is about the alternative!**

Statistical Inference

- **Sample mean used here as a conceptual tool**
- **Statistical inference applies to any type of statistics**
- **Need:**
  - Sample estimate (statistic)
  - Standard error
  - Reasonable sample size
- **Normal distribution generally holds for most statistics given large sample size**
- **You can apply confidence intervals, hypothesis testing using the same set of rules**

Regression

Regression Recap

- The **multiple linear regression**

$$Y = c + b_1X_1 + b_2X_2 + \cdots + b_kX_k + error$$

- Captures the **linear** relationship between the dependent variable Y and k potential explanatory variables
- Focus so far has been on the **diagnostic** relationship
- i.e. the **interpretation** of the slope coefficients $b_1, b_2, \ldots, b_k$

Regression Recap

- The **multiple linear regression – interpretation**

$$Y = c + b_1X_1 + b_2X_2 + \cdots + b_kX_k + error$$

- **Given all other variables held fixed**, an increase in 1 unit of X_1 is expected to shift Y by b_1 units

Regression Recap

- The **multiple linear regression – accounting for nonlinear relations**

$$Y = c + b_1X_1 + b_2X_2 + \cdots + b_kX_k + error$$

- The explanatory variables can include
  - Quadratic/polynomial terms – to capture **curvature** effects
  - Dummy variables – capture the effect of **categorical/qualitative** variables
  - **Log** transforms – percentage interpretations
  - **Interaction** terms to account for varying effects

Regression Recap

- The **multiple linear regression – model statistics**

$$Y = c + b_1X_1 + b_2X_2 + \cdots + b_kX_k + error$$

- **Adjusted R-squared** – proportion of variation of Y explained by model
- **Residual standard error**/model error – smaller error = larger R-squared
- **Statistical significance** of coefficients – judged by the p-value of each coefficient

Model selection

“ESSENTIALLY, ALL MODELS ARE WRONG, BUT SOME ARE USEFUL.”

George E. Box (1987)
(Famous Statistician)

Model selection

- **Are all predictors useful?** Which one(s) should we include?
- Example: one predictor
  - Two options – include it or not
  - → Two possible models
- Example: two predictors
  - Four possible models – include X1 AND X2; include X1 only; include X2 only; do not include both.

Model selection

- Generally, if you have ‘j’ predictors
- ➔ you have 2^j possible models
- ➔ 10 predictors gives 1024 possible models

How do we choose the “best” model?

Model selection

Logical reasoning – remove irrelevant (nonsense) predictors.

- Not every bit of the data set need to be used
- Exclude any exact relationships, e.g. profit=revenue-cost
- Highest adj. R-squared does not mean best model
- Ask yourself: Does the model make logical sense?

Using statistics!

- Backward procedure: Big → small
- Forward procedure: Small → big (subject to ordering)
- Stepwise procedure: Certain algorithm rules determine the pathway
- Variable removed based on statistical criteria – typically the p-value

Combination of both – involve judgement!