STAT1201
Analysis of Scientific Data
Summer Semester 2022

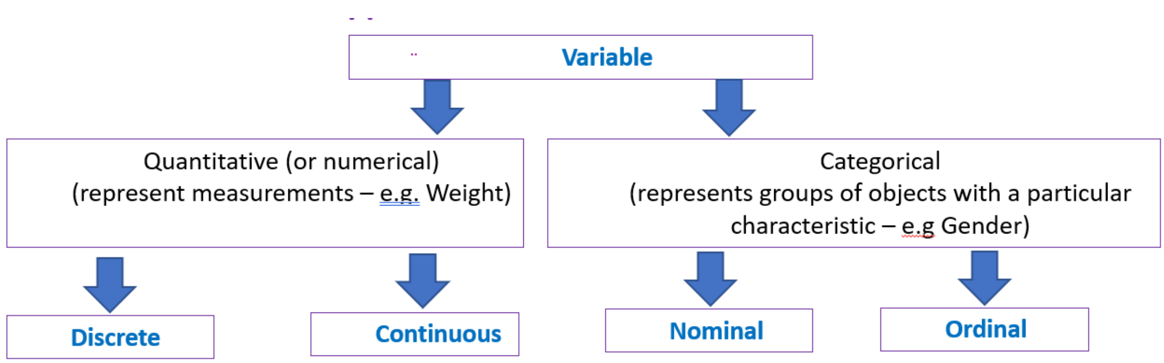
Lecture 14 – Revision

Dr. Wasanthi Thenuwara

1

Lecture 1 : We focused on

- Sources of variability (Natural variability; Measurement variability).
- Data and variable types.



2

Lecture 1 : We focused on

- Observational and Experimental studies. Experimental studies can be either a blind or a double blind study.
- The Language of Hypothesis Testing – The strength of evidence against the null hypothesis is determined by the p-value.

$p < 0.01$ – strong evidence against H_0

$0.01 \le p < 0.05$ – moderate evidence against H_0

$0.05 \le p < 0.1$ – weak evidence against H_0

$p \ge 0.1$ – no evidence against H_0

3

Lecture 2 : We focused on

- Visualising distributions of data and variables
  - ❑Categorical data – Bar charts, Tables
  - ❑Quantitative data – Histograms, Density plots, Boxplots
- Measures of Central tendency (Mode, Median, Mean)
- Measures of location (Percentiles and Quartiles)
- Measures of variability (Range, IQR, Variance and Standard Deviation)
  - ❑IQR = Q3 – Q1
  - ❑Detect outliers using IQR (Observation < Q1 – 1.5*IRQ; observation > Q3+1.5*IQR)
  - ❑Five number summary (min, Q1, Q2, Q3, Max) and boxplot (helps to identify the shape of a distribution). Symmetric, right skewed and left skewed distributions.
- Correlation (-1<=r<=+1) is used to measure the association between **two quantitative variables.**
- We used contingency tables to present two categorical variables.

4

Lecture 3 : We focused on

- Difference between population parameters and sample statistics.

Population Parameters	Sample Statistics
Population Size - N	Sample Size - n
Population Mean - μ	Sample Mean - $\overline{x}$
Population Variance - σ^2	Sample Variance - s^2
Population SD - σ	Sample SD - s
Population Proportion - p	Sample Proportion - $\hat{p}$

- Discrete and continuous probability distribution functions.
- Key probability concepts. What is the probability of seen 2 heads if you toss a coin a twice?
- Conditional probability P(A|B) = P(A and B)/P(B).

Discrete Random Variable: A random variable that has a countable number of possible values. E.g. Number of children in a family.

Continuous Random Variable: A random variable where the data can take infinitely many values. E.g. Height of the STAT1201 students

5

Lecture 3 : We focused on

Discrete Probability Distribution

The listing of all possible values of a discrete random variable X along with their associated probabilities.

E.g. Number of children (X) in a family and the associated probabilities from a random sample of families living in Brisbane.

X	P(X=x)
0	0.21
1	0.45
2	0.23
3	0.11

What is the probability that no more than two children in a family?

P($X \leq 2$) = P(X=0) + P(X=1) + P(X=2)

P($X \leq 2$) = 0.89

- Expected value and standard deviation of a discrete probability distribution.

6

Lecture 4 : We focused on

- Probability distributions and sampling distributions
- Binomial distribution is an example of a discrete probability distribution.
  X ~ Bin(n, p)
- Normal distribution is an example of a continuous probability distribution.

$$X \sim Normal\ (\mu, \sigma)$$

- Transform to a standard normal distribution

$$Z = \frac{X - \mu}{\sigma}$$

$$Z \sim N(0, 1)$$

- Probability calculations using binomial and normal distributions.
- Sampling distribution of the sample means ($\bar{X}$)
  $E(\bar{X}) = \mu$ and $sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

7

Lecture 4 : We focused on

- If the population is normally distributed, the sampling distribution of the sample means is normally distributed.
  If, $X \sim Normal\ (\mu, \sigma)$

$$\bar{X} \sim Normal\ \left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- Central Limit Theorem – As the sample size increases, the sampling distribution of the sample means is normally distributed regardless of the shape of the population variable distribution.

$$\bar{X} \sim Normal\ \left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- Transform to a standard normal distribution;

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad and\ Z \sim N(0, 1)$$

8

Lecture 4 : We focused on

- Sampling distribution of the sample proportions

$$\hat{p} = \frac{x}{n}$$

$$E(\hat{p}) = p \text{ and } sd(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Provided that *n* is large such that np>5 and n(1-p)>5;

$$\hat{p} \sim Normal(p, \sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$$

$$Z = \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \text{ and } Z \sim N(0,1)$$

9

Lecture 5 : We focused on

- Confidence interval estimation (sample statistic ± MOE). MOE depends on level of confidence and the standard error of the statistic.
- One sample t-test. $t_{stat} = \frac{\bar{x}-\mu}{s/\sqrt{n}}$. The t_{stat} has a *t*-distribution with df = *n*-1.
- Type I and Type II errors in hypothesis testing decisions.
- Probability of Type I error is also called the level of significance (α). This dos not depend on sample size. Decided by the researcher. Most common α=0.05.
- Power = 1-P(Type II Error). As the sample size increases power increases.

Possible Outcomes from Decisions

	Actual (reality) Situation	
Statistical Decision	H_0 True	H_0 False
Do Not Reject H_0	✓ $(1-\alpha)$	II (β)
Reject H_0	I (α)	✓ $(1-\beta)$

10

Lecture 6 : Focused on Ethics

No questions for the final exam related to the Ethics lecture

11

Lecture 7 : We focused on

Hypothesis test to compare two independent population means.

Two sample t-test

Two cases considered.

Case 1: $\sigma_1 \neq \sigma_2$. By hand or we used R to perform the Welch t-test.

By hand: $se(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ and $df = \min(n_1 - 1, n_2 - 1)$

Case 2: $\sigma_1 = \sigma_2$. Used pooled t-test. Pooled variance was used to find the $se(\bar{x}_1 - \bar{x}_2)$

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

$se(\bar{x}_1 - \bar{x}_2) = \sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})}$ and $df = (n_1 + n_2 - 2)$

12

Lecture 7 : We focused on

- Confidence interval for the difference in the means of two independent populations. We considered two cases as in the hypothesis tests in the previous slide.
- Hypothesis test to compare two independent population proportions using Z distribution.

$$z_{stat} = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

- Confidence interval for the difference in the proportions of the two independent populations.

13

Lecture 8 : We focused on

- Inferences for correlation.

$t_{stat} = \frac{r-\rho}{se(r)}$ where $se(r) = \sqrt{\frac{1-r^2}{n-2}}$. The t_{stat} has a t-distribution with (n-2) df

Simple Linear Regression.

- One dependent (or response) quantitative variable (Y)
- One independent (or explanatory) variable (X)
- Population regression equation: $Y = \beta_0 + \beta_1 X + U$
- Least square estimation method is used to estimate population regression coefficients.
- Four assumptions (Linearity, independent errors, errors are normally distributed, equal errors of the errors)
- Residual plots can be used to test assumptions

14

Lecture 8 : We focused on

Multiple Linear Regression.

- One dependent (or response) quantitative variable (Y)
- More than one independent (or explanatory) variables (X_1, X_2, …, X_k)
- Least square estimation method is used to estimate population regression coefficients.
- Four assumptions (Linearity, independent errors, errors are normally distributed, equal errors of the errors).
- Residual plots can be used to test assumptions.

Examples

Y – Breath holding time; X_1 – Height and X_2 – Weight

Y – Breath holding time; X_1 – Height and X_2 – Sex

Y – Breath holding time; X_1 – Height, X_2 - Weight and X_3 - Sex

15

Lecture 9 : We focused on

One-way ANOVA.

One quantitative response variable

- One independent categorical variable (has many categories or groups)
- Pairwise multiple comparisons can be performed using two sample t-test (pairwise.t.test() in R) or Tukey's Honestly Significant Difference (Tukey's HSD)

Statistical software usually summarises an analysis of variance in the form of an ANOVA table.

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Group (categorical Variable Name)	k - 1	SSG	MSG = SSG/(k-1)	MSG/MSR	P(F>=F*)
Residuals	n - k	SSR	MSR = SSR/(n-k)		
Total	n - 1	SST	MST = SST/(n-1)		

k = number of groups   n = sample size

16

Lecture 9 : We focused on

Two-way ANOVA.

- One quantitative response variable
- Two independent categorical variables.
- Two factors (or treatments) are simultaneously evaluated or examined. That is, interaction effect of the two categorical variables are examined.

17

Lecture 10: Focused on

Experimental design in practice; Replication; Randomisation; Blocking and multi-factor designs.

18

Lecture 11 : We focused on

Chi-square test.

- Chi-square test for independence (test whether two categorical variables are related or not). Use a contingency table.
- Chi-square goodness of fit test – to test whether observed data follows a specific probability distribution.

Logistic Regression.

- Binary (or dichotomous) response variable (Y).
- One or more independent variables.
- Used the concepts of odds and odds ratio.

Example: Investigate the risk factors for lung cancer

Response (or dependent) variable: Got a lung cancer (Yes/No)

Independent variables: No. of smokes per day, Sex, Age

19

Lecture 12 : We focused on

Non-parametric methods.

1. Sign test
2. Signed-Rank test (Wilcoxon Signed-Rank test)
3. Rank-sum test (Wilcoxon Rank-sum test/Mann Whitney test)

First and second tests are alternative to one-sample t-test or paired t-test.

Third test is an alternative to two-sample t-test.

20

Thank you.

Good Luck!

21