Research Methodology UE18CS400SG
Unit 3

Aditeya Baral

March 2022

1 Hypothesis Testing

"Proposition or a set of proposition set forth as an explanation for the occurrence of some specified group of phenomena either asserted merely as a provisional conjecture to guide some investigation or accepted as highly probable in the light of established facts."

- An assumption or supposition to prove or disprove
- Suggests experiments and observations
- Principle Instrument of Research
- Used to decide if sample data can offer support for hypothesis to be generalized

1.1 Characteristics of Hypothesis

- Clear and precise, stated in simple terms, easily understandable
- States relationship between variables
- Limited in scope and specific
- Consistent with facts
- Possible to test within reasonable time
- Capable of being tested (a hypothesis is testable if other deductions can be made from it which in turn can be proved or disproved)

1.2 Basic Concepts

1.2.1 Null and Alternate Hypothesis

Null Hypothesis (H_0) compares two methods and both are equally good. **Alternate Hypothesis** (H_1) proves one method is better than the other.

		Actual	
		H_0 True	H_0 False
Predicted	Accept H_0	No Error Probability = 1 - α	Type 2 Error Probability = β
	Reject H_0	Type 1 Error Probability = α	No Error Probability = 1 - β

Table 1: Distribution of Type 1 and Type 2 Errors

1.2.2 Statistically Significant

- Test whether relationship between two categorical variables in a sample is also present between the two variables in the population
- **If the relationship is strong in the sample, it indicates the relationship in the population is real and as strong as (or even more then) the sample** – cannot be due to *chance*
- **Level of Significance (α)** – Probability of rejecting Null Hypothesis when it is actually true (conclude a difference exists when there is no difference)

1.2.3 Type 1 and Type 2 Errors

1. **Type 1**
   - **Null Hypothesis is rejected when it is true** (reject hypothesis when it should have been accepted)
   - Results in **False Positives**
   - α; related to p value
   - Probability distributed at tails of a normal curve
2. **Type 2**
   - **Null Hypothesis is accepted when it is untrue** (accept hypothesis when it should have been rejected)
   - Results in **False Negative**
   - β; related to *power*
   - Occurs when sample size is too small

1.3 One-Tailed and Two-Tailed Tests

- $\neq$ in H_1 – Two-Tailed
- $>$ in H_1 – Right-Tailed

- $<$ in H_1 – Left-Tailed
- If significance is α
  - Obtain Z test statistic

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma_p}{\sqrt{n}}} \tag{1}$$

  - Obtain probability P from Z-table
  - Reject H_0 (Statistically Significant) if
    - For One-Tail Tests, $P < \alpha$
    - For Two-Tail Tests, $P < \alpha/2$

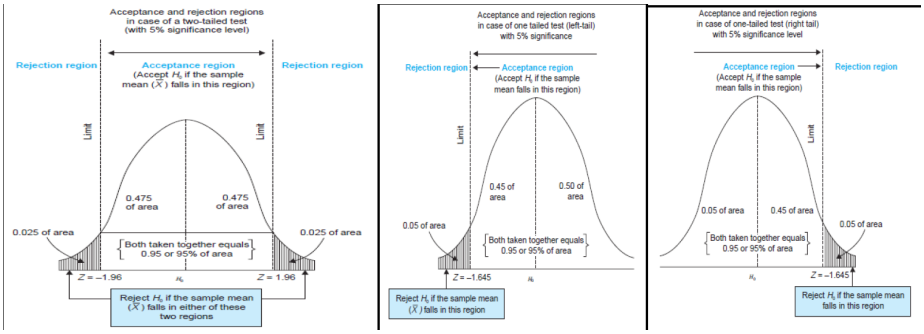

Figure 1: One-Tail and Two-Tail Acceptance and Rejection Regions

1.4 Steps in Hypothesis Testing

- State H_0 and H_1
- Specify α
- Decide sampling distribution
- Sample and obtain value from sample data
- Obtain probability that sample result will diverge from expectation if H_0 is true
- Accept or reject H_0 based on P and α

1.5 Power of Hypothesis Test

- $Power = 1 - \beta$
- **Power is the probability of rejecting the Null Hypothesis (or accepting Alternate Hypothesis) when the Alternate Hypothesis is true**
- The ability to test the existence of an effect
- High power is desired ($\geq$80%) – if power value is less, increase sample size

1.6 Types of Tests

1.6.1 Z-test

Normal, infinite population, **small or large sample size**, **known variance**, H_1 one or two-sided

1.6.2 t-test

Normal, infinite population, **small sample size only**, **unknown variance**, H_1 one or two-sided

$$t = \frac{\overline{X} - \mu}{\frac{\sigma_s}{\sqrt{n}}} \tag{2}$$

1.6.3 Chi-Square test

- Determines if sample data matches population
- Obtain confidence interval estimate of *unknown population variance*
- **Non-parametric test** and **no assumptions about population**
- Used to test –
  - **Goodness of Fit** – how well does a theoretical distribution (binomial, poisson, normal, etc) fit the data
  - **Independence** – explain if two variables are dependent or associated
- **Conditions for chi-square test** –
  - Observations are random and independent
  - Each group's frequency is ≥ 10. Join groups if frequency is < 10
  - Overall number of observations is large (> 50)
  - Linear constrains
- **Degrees of Freedom** – Number of independent values which are assigned to a statistical distribution.

- If we have a set of n possible values, $df = n - 1$
- If we have an $r \times c$ matrix of values, $df = (r-1)(c-1)$

- **Steps in chi-square test** (refer this link for solved example)
  - H_0 states that the two variables are related. H_1 states that the two variables are unrelated
  - Find Expected table E using Observed table O

$$E_{ij} = \frac{row_i\ total \times column_j\ total}{Total\ observations} \tag{3}$$

    Or for a single event,

$$E_i = P(i) \times Total\ observations \tag{4}$$

  - Find $\chi^2_{calculated}$

$$\chi^2_{calculated} = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{5}$$

  - For the given degrees of freedom df and level of significance α, find value of χ^2 from the χ^2 table
  - If $\chi^2_{calculated} > \chi^2$, reject H_0

2 ANOVA – ANalysis Of VAriance

2.1 Hypothesis Testing with > 1 Samples or Populations

- It is used to find the **variability between sample means**
- Helps determine if *samples are from the same larger population*
- Multiple t-tests or z-tests are not the answer
  - *Increased* number of tests $\binom{n}{2}$
  - Error *compounds* with each test – if confidence level is 95% and n tests are conducted, the confidence is 0.95^n (large decrease), and hence $\alpha = 1 - 0.95^n$ (large increase)

2.1.1 Equations for Z and t statistic for 2 Samples or Populations

$$z = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{6}$$

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\sum_{i=1}^{n_1}(X_{1i} - \overline{X_1}) + \sum_{i=1}^{n_2}(X_{2i} - \overline{X_2})}{n_1 + n_2 - 2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{7}$$

2.2 Obtaining ANOVA Table

Overall Mean ($\overline{\overline{X}}$) is the mean of all possible data points ($x_i^{(j)}$) across populations or samples. If $N = n_1 + n_2 + n_3 + ... + n_P$, then

$$\overline{\overline{X}} = \frac{\sum_{i=1}^{N} x_i}{N}, \text{ or } \overline{\overline{X}} = \frac{\sum_{i=1}^{P} \sum_{j=1}^{n_i} x_i^{(j)}}{\sum_{i=1}^{P} n_i} \tag{8}$$

Total/Overall Sum of Squares (SST) or **SS for Total Variance** is the summation of the squares of the differences between every data point ($x_i^{(j)}$) and the overall mean ($\overline{\overline{X}}$). It is also the **sum of the variance within and variance between**.

$$SST = \sum_{i=1}^{P} \sum_{j=1}^{n_i} (x_i^{(j)} - \overline{\overline{X}})^2 \tag{9}$$

$$SST = SSC + SSE \tag{10}$$

Column/Between Sum of Squares (SSC) is the summation of the squares of the differences between the population or sample mean ($\overline{X_i}$) and overall mean ($\overline{\overline{X}}$)

$$SS\ Between = \sum_{i=1}^{P} n_i (\overline{X_i} - \overline{\overline{X}})^2 \tag{11}$$

Mean Squares Between (MS Between) is the ratio between SSC and its degree of freedom

$$MS\ Between = \frac{SSC}{P-1} \tag{12}$$

Error/Within Sum of Squares (SSE) is summation of the squares of the differences between every data point ($x_i^{(j)}$) and its population or sample mean ($\overline{X_i}$)

$$SSE = \sum_{i=1}^{P} \sum_{j=1}^{n_i} (x_i^{(j)} - \overline{X_i})^2 \tag{13}$$

Mean Squares Within (MS Within) is the ratio between SSE and its degree of freedom

$$MS\ Within = \frac{SSE}{\sum_{i=1}^{P} n_i - P} \tag{14}$$

***F*-ratio** is the ratio between MS Between and MS Within

$$F - ratio = \frac{MS\ Between}{MS\ Within} \tag{15}$$

Variation	SS	df	MS	F-ratio
Between	SSC	P - 1	$\frac{SSC}{P-1}$	$\frac{MS\ Between}{MS\ Within}$
Within	SSE	N - P	$\frac{SSE}{N-P}$	
Total	SST	N - 1		

Table 2: ANOVA Table

***P*-value** can be obtained from the F-table. If P-value is less than α, then we reject H_0.

$$P - value = P(df_{between}, df_{within}) \tag{16}$$

3 Data Representation

Organization of data into tables, graphs or charts, so that logical and statistical conclusions can be derived from the collected measurements. Can be in the form of text, tables or graphs, help present and convey information

- Table – data organized orderly in rows and columns
- Visual/Figure – presentation like charts, graphs, diagrams, maps etc

3.1 Tabular Presentation

- Data presented orderly in rows and columns, based on characteristics
- 2D, each row/record represents an entity, each column is attribute and is unique
- Compact and self-explanatory, easier to comprehend and understand
- Numeric tables - store quantitative, or combination of quantitative and qualitative data (mostly numeric values)
- Table Types
  - **Purpose** – reference table, text table
  - **Content**
    - Simple – one characteristic per stub
    - Complex – double (2 characteristics per stub), triple (3 characteristics), multiple (each stub may be further divided into characteristics)
- Components – table number (to identify and reference tables), title, stub (row heading), caption (column heading), body, foot note, source note

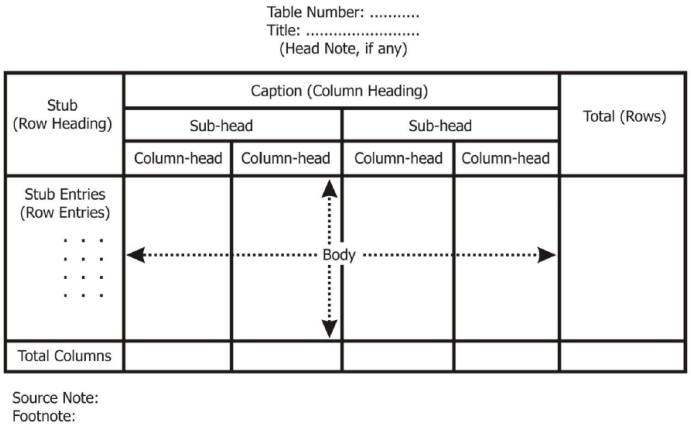

Figure 2: Components of a Table

3.1.1 Advantages of Table

- Represents large amount of data in an orderly manner
- Easy to read, organized in rows and columns
- Easy to construct
- Easy to add new records (as rows)

3.1.2 Features of a Good Table

- Attractive
- Simple, clear, easy to understand
- Avoid too many details, should not be too big or too small

3.2 Graphs and Diagrams

Used to represent data in visual form, supported by narration from presenter. Graphs can be either histogram, frequency curve ogive (cumulative frequency) or line graphs. Diagrams include bar plot and pie chart.

3.2.1 Rules for Drawing Graphs and Diagrams

- Choose appropriate form of visualization
- Title – information about graph
- Scale and units – neither too big or small, use right units

Graphs	Diagrams
Graph paper used	Plain paper used
Shows mathematical relationship between two variables	Does not show any relationship
Appropriate for frequency distributions and time series	Not suitable for such data
Not attractive	More attractive and suitable for publicity and propaganda
Adds meaning to data and aids analysis	Do not add to the meaning of data and does not help in analysis
Used by statisticians and research workers in analysis	Used to display data in different ways

Table 3: Difference between Graphs and Diagrams

- Neat, attractive, simple to interpret and convey meaning
- Original
- Economical – should not be laborious or costly to construct

3.2.2 Advantages of Graphs and Diagrams

- Attractive and long-lasting impression
- Bird's eye view of data
- Easy to understand and compare characteristics
- Graphs help visualize theorems and results of statistics

3.2.3 Limitations of Graphs and Diagrams

- Visual aids cannot replace numerical data
- Lack of mathematical rigour
- Not as accurate as tabular data
- Misrepresentation in visualization leads to misled observers and wrong interpretations

4 Results and Discussion – Data Interpretation

4.1 Results

Results are what you find in research, *without details, interpretation or references to literature.* Focuses on facts and presents a simple and clear account of what was achieved

4.2 Discussion

Discussion is the heart of the paper and commentary of results. It addresses the meaning of the findings in research.

- **Purpose** – state interpretations, opinions, implications of findings and suggestions for future research
- **Function** – answer questions in introduction, explain how results support answers and how answers fit with existing knowledge. If results are not in the desired direction, explain why (sampling, measurement, procedure, confounding variables etc)
  - Summarize previous findings and then your findings
  - Provide meaningful answers
  - Interpret objectively and subjectively
  - Include references to others' interpretations
  - Every conclusion should be defensible
- **Include** – unexpected results, references to previous research, explanations, exemplifications, deductions, hypothesis and recommendations

4.2.1 Technique of Discussion

1. **Organize Discussion** – specific to general, $findings \rightarrow literature \rightarrow theory \rightarrow practice$
2. **Re-state Hypothesis** and answer questions in introduction
3. **Explain results** – are they expected, why are they acceptable, how are they consistent with published knowledge
4. **Address all results** – regardless of statistical significance
5. **Describe patterns, principles and relationships** – State answer, then support with results, then cite other work
6. **Defend answers** – why is the work satisfactory, why is others' work not suitable, give convincing reasons
7. **Discuss and evaluate conflicting explanations**
8. **Discuss unexpected findings** – mention finding and then describe
9. **Identify limitations and weaknesses** – are they important to interpretation of results, do they affect validity of findings, no apologetic tone
10. **Summarize** – concise, brief, specific, importance and implication, recommendations for future research

Do's	Don't's
Provide context and explanation	Rehash results
Emphasize positives	Exaggerate positives
Look towards the future	End with it

Table 4: Do's and Don't's of writing a Discussion

5 Summary, Conclusion and Recommendation

5.1 Summary

Summary is a restatement of findings under each factor.

- Summarizes findings
- May be the conclusion too
- Limitations and future work

5.2 Conclusion

Conclusion is an *interpretation (not repetition)* of facts discussed. It is a take-away from the study, bounded by what is shown in data. Highlights the value and point of the study.

5.2.1 Purpose of Conclusion

- Integrate issues raised in discussion
- Reflect on introduction, problem statement and objectives
- Answer research questions
- Identify theoretical implications of study
- Highlight limitations
- Direction for future research

5.2.2 Content of a Good Conclusion

- Logical ending, integrates what was discussed
- No new or undiscussed information
- Pulls together all parts of the argument
- Refers the reader to the focus or central topic of the study
- Systematic and brief

- Adds to overall quality and impact of study
- Consists of the following
  - Restate research questions – reinforce importance
  - Empirical findings
  - Theoretical implications
  - Recommendations for future research
  - Limitations

6 References

A **referencing style** is a specific format for presenting in-text references (footnotes or endnotes) and bibliography. A **reference** is the action of mentioning or alluding a source of information to ascertain something.

The **types of references** are **journal, book and internet**. **Reference elements** include **Authors' names, title, journal name, year, volume and page numbers**

6.1 Reasons for Referencing

- Proves research has been done to support analysis
- Easy to follow up on work
- Credit to others' work
- Avoids plagiarism charges
- Needed to support all significant statements
- Indicates origin of material and source of research (for further reading)

6.2 Styles of Referencing

[1]

6.2.1 Harvard Style

1. Author's name followed by its initials
2. Year of publication
3. Article title with single quotation mark followed by full stop
4. Name of Journal in italic form

[1] Not needed for ISA. Memorize before ESA

5. Volume followed by a comma
6. Issue no. in bracket
7. Page no

6.2.2 Vancouver Style

1. Author Surname followed by Initials
2. Title of article followed by double quotation
3. Title of journal (abbreviated)
4. Date of Publication followed by semicolon
5. Volume Number
6. Issue Number in bracket
7. Page Number

6.2.3 MLA Citation Style (Modern language Association)

1. Authors name
2. Title of article
3. Name of journal
4. Volume number followed by decimal issue no
5. Year of publication
6. Page numbers
7. Medium of publication

6.2.4 American Psychological Association Style (APA)

1. Author's name followed by its initials
2. Year of publication
3. Article title followed by full stop
4. Name of Journal in italic form
5. Volume followed by a comma
6. Page no

6.2.5 Chicago Manual Style

1. Name of author
2. Article title in double quotation mark
3. Title of journal in italic
4. Volume
5. Year of publication
6. Page no

6.2.6 Royal Society of Chemistry

1. Initials followed by author's surname
2. Title of journal (abbreviated)
3. Year of publication
4. Volume number
5. Pages no

6.3 Difference between Reference List and Bibliography

Reference List	**Bibliography**
Sources cited in text	Sources consulted for study but may not be cited
Arranged in citation order	Arranged in alphabetical order of author surnames
Placed at the end of the paper, or as a footnote or endnote	Placed at the end of the paper

Table 5: Difference between Reference List and Bibliography