Final Sheet

December 2021

1 Data

1.a Types of Variables

Qualitative/Categorical

- Outcomes fall into different categories
- Categories can be ordered

Quantitative

- Measured on a numeric scale

1.b Summarizing Data Visually

Qualitative/Categorical Data

- Frequency tables - displays all categories of a single categorical variable with associated frequencies
- Contingency tables - display two categorical variables simultaneously
- Marginal distributions - display distribution of one of the two variables only
- Conditional distributions - display distribution of one variable, satisfying a condition of the other variable
- Bar charts
- Pie charts

Quantitative Data

- **Graphically**
  - Histogram
  - Stem-and-leaf displays
  - Boxplots
- **Shape of the Distribution**
  - Modality (number of peaks):
    * unimodal
    * bimodal
    * multimodal
  - Symmetry of distribution:
    * unimodal

* skewed to right (long right tail)
* skewed to left (long left tail)
  - Presence of outliers

- **Numerically**

  - Measures of center:
    * mean
    * median
  - Measures of spread:
    * variance: $s^2 = \frac{\Sigma_{i=1}^{n}(y_1 - \overline{y})^2}{n-1}$
    * standard deviation: $s = \sqrt{\frac{\Sigma_{i=1}^{n}(y_1 - \overline{y})^2}{n-1}}$
    * interquartile range $IQR = Q3 - Q1$
  - Percentiles (also called quantiles)
  - 5-number summary:
    * minimum
    * first quartile (Q1)
    * second quartile (Q2)
    * third quartile (Q3)
    * maximum

- **Sensitivity to Outliers**

  - Sensitive to outliers:
    * mean
    * range, variance, standard deviation
  - Not sensitive to outliers
    * median
    * IQR

2   Normal Distribution

Characteristics of the Normal Model

- bell-shaped; unimodal

- perfectly symmetric about the mean

- spread of distribution determined by value of standard deviation

- mean μ and the standard deviation σ are parameters(numerical characteristics of a model)

- mean $\overline{y}$ and standard deviation s are statistics (numerical characteristics of a sample)

The 68-95-99.7 Rule

- 68% of data falls within 1 σ of μ

- 95% of data falls within 2 σ of μ

- 99.7% of data falls within 3 σ of μ

Finding Areas Under the Normal Model

Algorithm

- Identify the:
  μ - mean of the model
  σ - standard deviation of the model
  y - observed value

- Construct the normal model: N(μ,σ)

- Calculate the z-score (z): $z = \frac{y-\mu}{\sigma}$

- Using R compute the p-value:

  - Area below y: pnorm(z)

  - Area above y: pnorm(z , lower.tail = F)

  - Area in between y_1 and y_2 (where $y_1 > y_2$): pnorm(z_1) - pnorm(z_2)

- **Finding Z-Score from the Area Under the Normal Model**

  - Area above unknown y: qnorm(p, lower.tail = F)

  - Area below unknown y: qnorm(p)

3 Binomial Distribution

4 Correlation and Association

Scatterplots

- Direction:

  - Positive (x and (y) values tend to go in the same direction)

  - Negative (x and y values tend to go in the opposite direction)

- Form:

  - Linear

  - Non-linear

- Point relationship:

  - Strong relationship between points

  - weak or no relationship between points (randomly scattered)

- Outliers

Correlation (r)

- Positive correlation: large x values are linearly associated with large y values (r is positive)

- Negative correlation: large x values are linearly associated with small y values (r is negative)

- r has a value between 1 and -1, and has no units

- $r = \frac{\Sigma z_x * z_y}{n-1}$

Association vs Causality

- Association does not imply causation. There may be a lurking variable

5   Regression Analysis

The Regression Line

- Equation for regression line: $\hat{y} = intercept + (slope * x)$

- Equation for slope: $slope = r * \frac{s_y}{s_x}$
  (where s_y and s_x are the standard deviations of y and x respectively)

- Equation for intercept: $intercept = \overline{y} - (slope * \overline{x})$
  (where $\overline{y}$ and $\overline{x}$ are the mean y and x values respectively)

The Residuals

- The residual (e) is the difference between observed value y and the predicted value $\hat{y}$. Therefore:
  $e = y$ (from data) - $\hat{y}$ (from model)

- The sum of residuals is equal to zero

- Linear model is obtained by minimizing the sum of the squared residuals. Therefore, also referred to as the least squares regression line

- To assess appropriateness of regression model, we use the residual plot (plots residuals against explanatory variable data). If plot shows no pattern, model is appropriate.

6   Experiments and Observational Studies

Types of Studies

- Observational Studies

  - Investigators have no control over either variable
  - No deliberate human intervention
  - Retrospective study: based on information from events that have taken place in the past
  - Prospective study: data and information is gathered in real time

- Experiments

  - Involves planned intervention on the exposure to a condition suspected of altering the response outcome
  - Most often control group(s) will be used

Randomized, Comparative Experiments

- Involves assessing the effect of an explanatory variable, called a factor, on a response variable

- Compares the response variable between different levels of the factor

- Experimenters control what type of treatment individuals receive, the treatment assignment is random

- Participants referred to as subjects or experimental units

- The treatment a subject receives will be a combination of the levels from different factors

Principles of Experimental Design

- Randomize

  - Treatments are randomly assigned to subjects

- Replicate

  - Comparison between different treatment groups will not be reliable unless more individuals receive each treatment

- Blocking

  - May be beneficial to control for variables that are not factors but are believed to have some influence on the response variable
  - Subjects are divided into blocks (ex. male and female groups). Treatment assignment and comparisons are done within each block separately

Blinding and Placebo

- Single Blind: either the subjects or the evaluators are blinded as to treatment assignment

- Double Blind: neither the subjects nor the evaluators knows the treatment assignments

- Blinding is usually done using a placebo which is designed to look like the treatment but has no real treatment value

7   Types of Sampling

Sampling Methods

- Simple Random Sampling

  - Consists of n individuals sampled at random from the population
  - Each individual has an equal chance of being selected
  - Each possible sample size n is equally likely

- Stratified Sampling

  - Population is divided into strata (a stratum is a subset of the population that shares a particular characteristic)
  - Simple random sample is drawn from each stratum
  - Stratified sample has smaller variability across samples and hence give more reliable results

- Cluster Sampling

  - Can be used when natural groups in a population exist
  - Population is divided into those groups/clusters
  - Simple random sample from all clusters is obtained
  - If all individuals in a selected cluster are included, final sample is a one-stage cluster sample
  - If additional simple random sample is drawn from selected clusters, final sample is a two-stage cluster sample
  - This method is used for the sake of convenience, practicality, and cost-efficiency

- Multistage Sampling

  - Involves more than one stage or more than one sampling procedure in obtaining a sample

- Systematic Sampling

  - Obtained by selecting every kth individual from the sampling frame
  - Method can be used as long as list being sampled from does not contain a hidden order

Bad Sampling Procedures and Biases

- Undercoverage

  – When sampling frame or sampling procedure excludes or under-represents certain types of individuals from the population

- Convenience Sampling

  – Selecting individuals from a population based on availability and access

- Voluntary Response Bias

  – If responses are voluntary, those with strong opinions tend to be over-represented

- Non-response Bias

  – Individuals who do not respond in a survey might differ from the respondents in certain aspects
  – Including only the respondents in a sample will result in non-response bias

- Response Bias

  – Subject's response is influenced by how the question was phrased or asked, or due to misunderstanding of a question, or unwillingness to disclose the truth

8   Hypothesis Testing
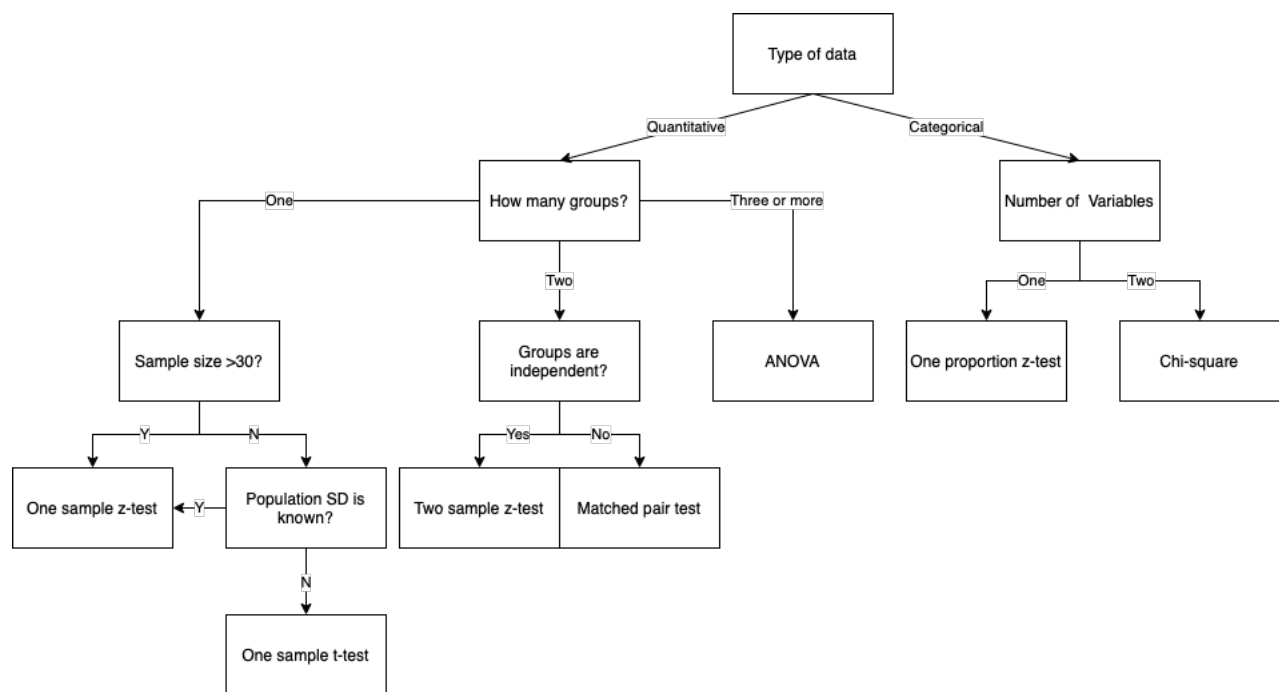
8.a   One sample z-test

Algorithm

- Idenitify parameter of interest. Find the null and alternative hypotheseses.
  s - The standard deviation of the sample.
  n - The sample size.
  μ - Hypothethised population mean.
  $\mathbf{SE}(\bar{y}) = \frac{s}{\sqrt{n}}$ - Standard error of the statistic.

- Construct the null-model: $\mathbf{N}(\mu, \frac{s}{\sqrt{n}})$

- Find the test-statistic(t): $\mathbf{Z} = \frac{x-\mu}{\mathbf{SE}(\bar{y})}$

- Using R compute the p-value:

  - One-sided hypothesis : pnorm(t)
  - Two-sided hypothesis : $2 \cdot$ pnorm(t)

- If the p-value is less than α - reject the null-hypothesis. Otherwise, you fail to reject the null-hypothesis.

8.b   One proportion z-test

Algorithm

- Idenitify parameter of interest. Find the null and alternative hypotheseses.
  n - The sample size.
  p_0 - Hypothethised proportion.
  $\mathbf{SD} = \sqrt{\frac{p_0(1-p_0)}{n}}$ - Standard error of the statistic.

- Construct the null-model: $\mathbf{N}(\mu, \sqrt{\frac{p_0(1-p_0)}{n}})$

- Find the test-statistic(t): $\mathbf{Z} = \frac{x-p_0}{\mathbf{SD}}$

- Using R compute the p-value:

  - One-sided hypothesis : pnorm(t)
  - Two-sided hypothesis : $2 \cdot$ pnorm(t)

- If the p-value is less than α - reject the null-hypothesis. Otherwise, you fail to reject the null-hypothesis.

8.c   Two sample z-test

Algorithm

- Idenitify parameter of interest. Find the null and alternative hypotheseses.
  s - The standard deviation of the sample.
  n - The sample size.
  μ - Hypothethised population mean.
  $\mathbf{SE}(\bar{y}) = \frac{s}{\sqrt{n}}$ - Standard error of the statistic.

- Construct the null-model: $\mathbf{N}(\mu, \frac{s}{\sqrt{n}})$

- Find the test-statistic(t): $\mathbf{Z} = \frac{x-p_0}{\mathbf{SE}(\bar{y})}$

- Using R compute the p-value:

  - One-sided hypothesis : pnorm(t)
  - Two-sided hypothesis : $2 \cdot$ pnorm(t)

- If the p-value is less than α - reject the null-hypothesis. Otherwise, you fail to reject the null-hypothesis.

8.d   Matched pair

8.e   One sample t-test

8.f   ANOVA