Table of Contents (As covered from textbook)

Ch 1 Data and Decisions

What are data?

Data: recorded values whether numbers or labels, together with their context
Statistics: tools and associated reasoning to summarize, model and understand what data conveys

Data can be numerical (consisting only of numbers), alphabetic (consisting only of letters) or alphanumerical (mixed numbers and letters)

Understand data: who, what, when, where and possibly, why and how. Provides context for data

Generally rows of data table correspond to individual **cases** which some characteristic called **variables** have been recorded.
Case: individual about whom or which we have data

Variable: holds information about the same characteristic for many cases

Different names for cases depending on situation:

- **Respondents:** individuals who answer a survey
- **Subjects:** people on whom we experiment or **participants** (attempting to acknowledge the importance of their role)
- **Experimental units:** Animals, plants, websites and other inanimate subjects
- **Records**: name for rows in a database

- Column titles (variable names) tell *what* is recorded
- *Who* of the table commonly found on leftmost column, often an identifying variable

Metadata: typically contains information about variables in a database including e.g. *how, when, where, why* data was collected, *who* each case represents and definitions of all the variables

Spreadsheet is a data table however mainly effective for relatively small data sets

Other database architectures include:

- **Relational database**: two/more separate data tables linked together so that information can be merged across them. Each data table is a *relation* because it is about a specific set of cases with info about each of these cases for all of the variables

Variable Types

Categorical/qualitative variable when values of variables are the names of categories

- **Ordinal:** when there is an intrinsic order to the values
- **Nominal:** unordered categories

Quantitative variable when values of variables are measured numerical quantities with units

Some variables can be considered either categorical or quantitative depending on the question

- Age considered quantitative if responses are numerical with units
- Age considered categorical if lumped together into categories e.g child(12-), teen(13-19) etc
- Area codes may look quantitative but are categories
- ZIP codes are categories too but numbers contain information

Identifier variables: categorical variables whose only purpose is to assign a unique identifier code to each individual in the data set e.g. student ID number

Other data types: asking "how satisfied were you with the service?" 1) not satisfied; 2)somewhat satisfied etc….

- An *order* of perceived worth however values aren't strictly numbers, categorical ordinal

Time series: an ordered sequence of values of a single quantitative variable measured at regular intervals over time
Cross-sectional: several variables measured at same time point

Data sources: Where, How, When

- Values recorded at different times may differ in significance
- How data is collected can differ between insight and nonsense, inferences can only be made if data is valid e.g. designing a survey or performing an experiment where variables are manipulated
- Contextual location data collected from affects trends etc.

Ch 2 Displaying and Describing Categorical Data

Summarising a categorical variable: So as to display information to easily communicate results
Frequency table: records the counts (sometimes percentages) for each categories of the variable

Displaying a categorical variable
Three Rules of Data Analysis:

4. Make a picture - display of data will show trends and help plan your approach to the analysis
5. Make a picture - allow easier analysis and show important features and patterns
6. Make a picture - best way to report data to others

The area principle: the area occupied by a part of the graph should correspond to the magnitude of the value it represents

Bar charts: displays the **distribution** of a categorical variable, showing the counts for each category next to each other for easy comparison

- Should have small spaces between bars to indicate these are freestanding bars that could be rearranged into any order
- Variable name often subtitle for horizontal axis
- If draw attention to *proportion* of each category, can replace counts with percentages and use a **relative frequency bar chart**

Pie charts: show how a whole group breaks into several categories with areas proportional to fraction of cases in each category

Exploring two categorical variables: Contingency Tables
Contingency tables: show how individuals are distributed along each variable depending on/*contingent on* the value of the other variable

- Margins give totals
- At the margins of a contingency table, the frequency distribution of either one of the variables is called its **marginal distribution**
- Marginal dist. for variable in contingency table is same as frequency dist.
- Each **cell** of a contingency table (intersection of row and column of table) gives count for a combination of values of the two variables
- Most statistics programs offer choice of **total percent, row percent** or **column percent** for contingency tables

Conditional distributions: distribution of a variable for cases that satisfy a condition on another distribution of a variable restricting the *who* to consider only a smaller group of individuals

- In a contingency table, when distribution of one variable is same for all categories of another variable, those two variables are **independent**

Segmented bar charts and mosaic plots

Segmented bar charts: treats each bar as the "whole" and divides it proportionally into segments which correspond to further variable categories

Mosaic plot: similar to segmented bar chart however obeys area principle better by making bars proportional to sizes of the groups. Popular for contingency tables

Simpson's paradox

Be sure to combine only comparable measurements for comparable individuals. Especially careful when combining across different levels of a second variable. Usually better to compare percentages within each level rather than across levels.

Ch 3 Displaying and Describing Quantitative Data

Displaying quantitative variables:

No categories so usually place possible values into **bins**, then count no. of cases within each bin
Provides **distribution** of quantitative variable and the building blocks for the display of distribution

Histograms: uses adjacent bars to show the distribution of values in a quantitative variable. Each bar represents the frequency of values falling in an interval of values

- Plots the bin counts as the heights of bars
- **Gaps** indicate a region where there are no values
- Each category has its own bar
- For quantitative variables, choose the width of the bins

Relative frequency histogram: report the percentage of cases in each bin

- Shape of two histograms same, only vertical axis and label differ
- Displays percentage of cases in each bin instead of the count - faithful to area principle

Stem-and-leaf displays: similar to histograms but also show individual values

- Good for data sets that aren't too large
- Breaks each number into two parts

Check quantitative data condition before making a stem-and-leaf display/histogram
Quantitative data condition: data must be values of a quantitative variable whose units are known

Shape

Mode: single most frequent value, describes shape of distribution

- **Unimodal**: one main hump in a histogram
- **Bimodal:** two humps in a histogram
- **Multimodal:** three or more humps in a histogram
- Bimodal histograms often indicate two groups in data. Be aware fluctuations may be artefacts of where bin boundaries fall.
- True mode, humps should still be there when displaying histogram with slightly different bin widths
- **Uniform:** distribution where histogram doesn’t appear to have any mode

Symmetry: symmetric distribution if halves on either side of center look approx like mirror images

- **Tails:** thinner ends of distribution
- **Skewed** to the side of the longer tail if not symmetrical

Outliers: extreme values that don't appear to belong with the rest of the data

- Should point out any outliers as they can be informative or an error, discuss in conclusions

Using your judgement to characterise a distribution

<u>**Center**</u>

Mean: the average, x̄

x-generic variable

∑-sigma / "sum"

$$\bar{x} = \frac{Total}{n} = \frac{\sum x}{n}$$

- Can be misleading for skewed data or distributions with gaps or outliers

Median: value that splits the histogram into two equal *areas*

- commonly used for variables such as cost or income, which are likely to be skewed
- *Resistant* to unusual observations and shape of distribution
- If n is odd, median is middle value $= \frac{n+1}{2}$
- If n is even, two middle values, median is average of two values in positions $\frac{n}{2}$ and $\frac{n}{2} + 1$

<u>**Spread of the distribution**</u>

Range = max - min

- Measure of spread
- **Lower quartile Q1:** value for which one quarter of the data lie below it
- **Upper quartile Q3:** value for which one quarter of the data lie above it
- **Interquartile range(IQR)** summarises the spread by focusing on the middle half of the data
  IQR **= Q3 - Q1**

Finding quartiles by hand:

3. Tukey method: split sorted data at the median (if n is odd, include median with each half). Then find the median of each of these halves - use these as the quartiles
4. TI calculator method: same as Tukey method but *don't* include median with each half

Standard deviation: square root of average squared difference between each data value and the mean. Summary of choice for spread of unimodal, symmetric variables

Uses *deviations* of each data value from the mean, can be influenced by outlying observations

- Average of *squared* deviations called the **variance** denoted by s^2

$$s^2 = \frac{\sum(x - \bar{x})^2}{n-1}$$

- Variance plays as a measure of spread has a problem as it is in *squared* units
- Square root of the variance gives the **standard deviation**

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n-1}}$$

<u>**Shape, Center, Spread - Summary**</u>

- Used for quantitative variable
- If shape skewed, point that out and report median and IQR
  May want to include mean and standard deviation, explain why mean and median differ (may indicate skewed distribution), histogram may help

- If shape unimodal and symmetric, report mean and standard deviation, possibly median & IQR IQR usually bit larger than standard deviation. If false, look to make sure distribution isn't skewed or multimodal and no outliers
- If multiple modes, try understand why. If reason for separate modes identified, possibly split data into separate groups
- Point out any clearly unusual observations. If reporting mean and standard deviation, report them computed with and without the unusual observation
- Always pair median with IQR and mean with standard deviation. Should always report measure of center with corresponding measure of spread

Standardising variables: enables comparison of values from different distributions to see which is more unusual in context
z-score: tells how many standard deviations a value is from its mean

- shifts mean to 0
- changes standard deviation to 1
- does not change the shape
- removes the units so easier comparison

$$z = \frac{(x - \bar{x})}{s}$$

Five-number summary of a distribution reports its median, quartiles, and extremes (max and min)

- Provides good overall look at distribution

Boxplot displays the information from a five-number summary

- Highlights several features of the distribution of a variable
- Central box shows middle half of data between quartiles
- Top of box is upper quartile, bottom is lower quartile so height of box Q3-Q1 = IQR
- Median is displayed as a horizontal line
  If median roughly centred between quartiles, then middle half of data is roughly symmetric, otherwise skewed
- Whiskers reach out of box to most extreme values not considered outliers
- Boxplot nominates points as outliers if they fall farther than 1.5 IQRs beyond either quartile
- Outliers displayed individually to keep them from skewing data and also encourage special attention
- Useful for comparing several distributions side by side

Comparing groups: Boxplots usually do a better job of comparison, able compare centers & spreads

Identifying outliers: Once identified outliers, investigate them, may be errors or provide information

Time series plots: Display of values against time

- When no strong trend/change in variability we say it is **stationary**
- Histogram can provide useful summary of a stationary series
- Time series plot reveals patterns unable to be seen in either histogram or boxplot
- Show point to point variation, better to smooth this out
- Smooth trace highlights long-term patterns

Transforming skewed data

- Skewed distributions are difficult to summarise
- **Re-express/transform** a skewed distribution to make it more symmetric by applying a simple function to all the data values e.g. log/reciprocal/sqrt