Chap1

Data arises from observations

Treatment group: individuals who receive the treatment of interest in an experiment

Control group: individuals NOT receiving treatment

Bias: the tendency to over/underestimate a population parameter due to a measurement process.

Random assignment helps minimize bias

Blinding: helps prevent bias by ensuring that the participants (and sometimes the researchers) do not know who is assigned to which study group.

Principals of Experimental Design

- Large sample size: This ensures that the study captures the full range of variability amongst the population and allows small differences to be noticed.
- Controlled and randomized: Random assignment of subjects to treatment or control groups to minimize bias.
- Double-blind: Neither subjects nor researchers know who is in which group.
- Placebo (if appropriate): This format controls for possible differences between groups that occur simply because subjects think their treatment is effective.

We can conclude causation if we use a controlled experiment, in which we deliberately assign the subjects to different treatments (usually at random)

Two way table displays counts of 2 categorical vars

Chap2

Bar chart displays the counts of each category, thus is for categorical data

Differences Between Bar Charts and Histograms

- A histogram displays numerical data. A bar chart displays categorical data.
- The bar widths of a histogram are meaningful and must all be the same size. The bar widths for a bar chart are meaningless.
- The bars of a histogram must touch each other. For a bar chart, there are gaps between bars.
- There is only one choice, ascending by x, for the order of the histogram, while there are many choices of order for a bar chart.

Review: Describing Distributions

The things to always describe when considering a distribution:

- Shape - examples: unimodal, symmetric
- Center - the "typical" value
  - Use the mean as the typical value for symmetric distribution
  - Use the median as the typical value for skewed distribution
- Spread - how spread out the data is
  - Use the standard deviation as the spread for symmetric distribution
  - Use the IQR as the spread for skewed distribution

Chap3

Remember to subtract 1 when calculating stdev!!

Q1, Q3

75% of data are below Q3

Q3-Q2<Q2-Q1 Easy Exam, clustered near high score, outliers are low

Q3-Q2=Q2-Q1 Mean is about the same as Median

Finding Q1, Q3, and IQR

- Let's again consider the final scores of 5 Stats 10 students: 79, 82, 94, 83, 92
- In order to find Q1 and Q3 we must first put the values in increasing order: 79, 82, 83, 92, 94

1. Find Q2. What is the median? Q2 = 83.
2. Find Q1, the median of the numbers less than 83.

$$Q1 = \frac{79 + 82}{2} = 80.5$$

3. Find Q3, the median of the numbers greater than 83.

$$Q3 = \frac{92 + 94}{2} = 93$$

5. Find IQR. IQR = Q3 - Q1 = 93 - 80.5 = 12.5

Least square line

The sum of the square of residuals around the regression line is minimum

Calculating IQR

If the data set has an odd number of values, we omit the median(centermost) value of the set.

If even, just split the set

1,2,**2**,3,4|4,7,**8**,9,11 IQR=8-2=6

21,25,27,31,32,37,43,45,49

Q1=26, Q3=44, IQR=44-26=18

Five number Summary: Min. Q1, Median, Q3, Max

Outlier

Left limit = Q1 - 1.5 x IQR

Right limit = Q3 + 1.5 x IQR

Outliers have NO effects on IQR, but increase stdev

Unusual

More than 2 stdev (Z-score exceeds +2 or below -2)

Use the mean and standard deviation when the distribution is symmetric and unimodal

Use the median and IQR when the distribution is left or right skewed

Boxplot

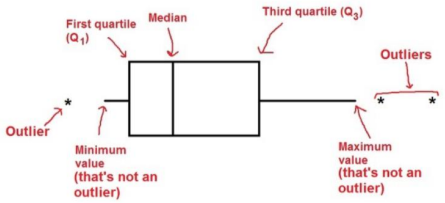
Should not use on very small data set(>=5)

Constructing a Boxplot

Calculate the five-number summary, right and left limits and outliers (if any)

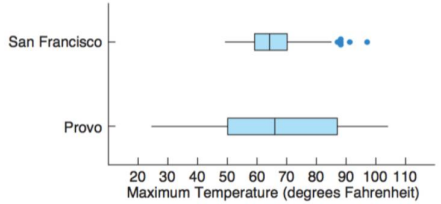
Draw a horizontal rectangle with a line segment in the middle. The short sides of the rectangle and the line segment in the middle should correspond to Q1, the median and Q3.

Sketch horizontal line segments (whiskers) on each side that extend to the most extreme values that are not potential outliers.



Comparing Distributions with Boxplots

- The median temperatures are about the same.
- The medians are in the center of the boxes and boxplots are fairly symmetric hence the distributions are fairly symmetric.
- The variation in daily temperatures in Provo is greater than in SF.
- The boxplot for SF shows some potential outliers.



Chap4

x-axis variable: predictor, independent, explanatory variable

y-axis variable: predicted, outcome, dependent, response variable

Correlation coefficient(r): measures the strength of linear relationship between 2 numerical variables.

Slope=r*(sy/sx)

Coefficient of determination (r^2): percentage of the variable of y is explained by x.

Interpretation of intercept: When x equals 0, we expect y to equal the intercept.

y-Intercept=y-bx

Extrapolation

do not make a prediction for a x value outside the range of the data - the linear model may no longer hold outside that range.

Chap5

Simulate randomness: coin toss, random number table, roll the dice

Law of Large Numbers: long-run relative frequency of repeated independent events gets closer to the true relative frequency as the number of trials increases.

E.g. If a coin is tossed many times, the overall percentage of heads should settle down to about 50% as the number of tosses increases

Dice Example

- Trial: Each die roll
- Outcome: A die has six sides
- Probability: Prob(rolling a side)=1/6

  Note: With a fair die the probability of getting each side is the same and is 1/6.
- Sample Space:

  Rolled once: S = {1, 2, 3, 4, 5, 6}
- Independence: The outcome of one die roll does not affect the outcome of the next die roll.

Disjoint events: Events that have no outcomes in common(thus cannot occur together P(A and B)=0

Disjoint events are NOT independent

Independent vs Disjoint

- Can two events be independent and disjoint?

  Remember: Independence means that the outcome of one trial doesn't influence the outcome of another. Disjoint means that two events can't happen at the same time.

  Disjoint events are NOT independent

  Example: Flipping heads or tails are two disjoint events. If we know that the outcome of a coin toss is heads, then we know that it is not tails. So whether or not the outcome is tails depends on whether or not the outcome is heads.

Non-disjoint events may/may not be independent

E.g. pick two random people in the class who scored 100% on the midterm. They could be complete strangers whose performance had nothing to do with each other, or they could be close friends who studied together a lot.

Complement Rule: Prob(A^c)=1-Prob(A)

Independence

- Independent events are variables or events that are not associated.
- Two events are independent if knowledge that one event has happened tells you nothing about whether or not the other event has happened.

  A and B are independent if the following occur:

$$P(A) \times P(B) = P(A \text{ and } B)$$

$$P(B|A) = P(B)$$

$$P(A|B) = P(A)$$

Sample space: collection of all possible outcomes of a trial. Two events have probabilities sum up to 1 **if they make up the sample space**.