Math 200

Mod 4-6 study guide

Collecting data (Mod 4)–

- Sampling**:**
  - Population and sampling frame
  - SRS- Random sampling provides every member of a population and equal chance of being selected thus creating a representative sample.
    - One of the important – and counterintuitive – consequences of random sampling g is that the size of the population doesn't significantly affect the sample size needed. A random sample of 1,000 will provide roughly the same accuracy for a population of 10,000 and one of 10,000,000!
  - Systematic
  - Cluster
  - Stratified – why might we might prefer stratified over SRS?
  - (Avoid) convenience sampling.
  - Bias (sampling that systematically favors a particular group or outcome).

The purpose of random sampling (selection) is to produce a representative sample which allows us to generalize results of a study from the sample to the general population from which it was taken.

Types of studies (Mod 5):

- Experiment:
  - Manipulates explanatory var., tests cause and effect vs observational: studies properties of population (summary OLI pg 25).
  - Elements of experimental design:
    - treatment groups
    - (possibly) control group
    - random assignment
    - blinding
    - direct control (making things the same for all participants – e.g. listening to the same music for the same length of time or washing hands with the same amount of soap.)
    - Placebo

Random assignment creates similar groups and mitigates the effects of confounding variables.

Direct control allows experimenters to eliminate or reduce the effects of confounding variables by making the actions or uniform across groups.

- Variables associated with studies:
  - Explanatory
  - Response
  - Confounding/Lurking

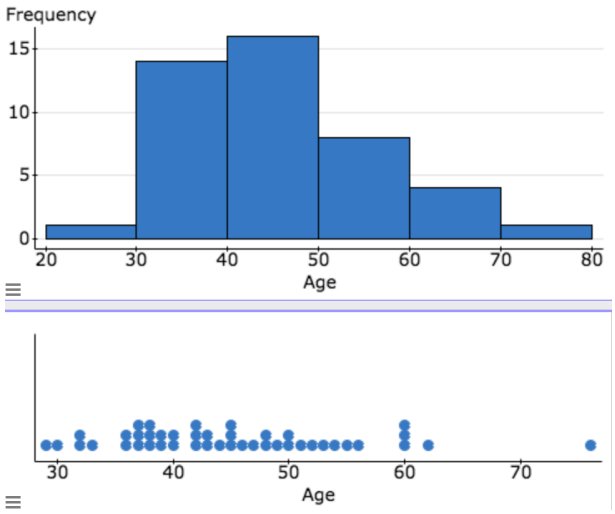
Experiments are the only consistent way to demonstrate cause and effect. Observational studies can be used to show this but the burden of eliminating confounding variables is often too costly or impractical to overcome.

Two important notes about sample size (yes, this is mentioned earlier – but it's important!)

1) When comparing two groups, the size of one group does not have to match the other. It's desirable to have the same sizes but not essential.
2) Sample size is largely independent of population size. You don't need to have a large sample in order to make inferences about a large population. The larger the sample, typically the more refined the estimate it provides but a random sample of size 1,000 will do a good job of estimating population parameters for populations that are 10,000 as well as 10,000,000.

Exploratory Data Analysis (Mod 6) –

- Pie and Bar (for categorical variables) – typically give percentages represented within a population e.g. proportion of left handers or smokers. Bar graphs allow for multiple groups to be compared – e.g. percentage of male smokers vs. percentage of female smokers.
- Stem and leaf
- Dotplot (detailed picture), Histogram, and Boxplot (big picture). e.g. The graphs below show a dotplot and histogram of ages of Best Actor Oscar award winners. Where the dotplot provides specific details, the histogram provides the general shape of the distribution.



Shape, Center, Spread, Outliers (SOCS)

- Shape: symmetric vs. skewed, uniform, bell or mound-shaped, unimodal, bimodal
- Center: Mean, Median, Mode
- Spread (variability): IQR measures the middle 50% of a distribution and is associated with median based statistics. Standard Deviation (and MAD) is associated with mean-based statistics and approximately 68% of the data occurs within one SD of the mean (OLI pg 57) Note that the Range measures the distance between the maximum and minimum data values. It is a poor measure of spread, however, since many different distributions all share the same range.
- Outliers: data that falls outside the pattern of the data. Sometimes due to variability, sometimes for exceptional reasons and sometimes from error.

Measures of center:

- Median – good with all distributions. The 50th percentile (half the data fall below and half above.
- Mean - use mean with symmetric distributions. Influenced by outliers, affected by skew. Mean is often described as the fair-share average

Rossman/Chance applet to compare relative locations of mean and median.

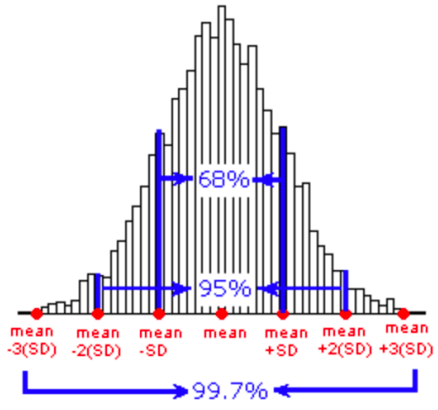
Measures of spread:

- Median: found by ordering n data values and computing the 50th percentile found at data point number: 0.5(1+n).
    - Quartiles Q_1 is the median of the lower half or the 25th percentile found at data point number: 0.25(1+n) and Q_3 is the median of the upper half of the data found at data point number: 0.75(1+n).
    - Percentiles are a generalization of quartiles. Any percentile, p, can be found at data point number: p/100(1+n). e.g. the 90th percentile is located at the 0.9(1+n) data point in a set.
    - 5 Number Summary (Low-Q_1-Med-Q_3-High)
    - IQR (single value but good to report as $Q_3 - Q_1$ as well)
    - Outliers (1.5 X IQR)

  Because we use with skewed sets, IQR not always symmetric about median.
  Anything below Q_1 is in the bottom 25%; anything above it is in the top 75% etc.

Mean:

- Standard Deviation – symmetric with respect to the mean so spread is often reported as $\mu \pm$ SD.
- Standard Deviation Rule (pg 57 OLI): 68-95-99.7. In a normally (or approximately normal) distributed data set about 68% of the data are within +/- 1 SD of the mean, approximately 95% of the data are within +/- 2 SD of the mean, and approximately 99.7% of the data are within +/- 3 SD of the mean. See human pregnancy example (OLI pg 58)



- ADM (MAD) good approximation to SD (smaller) and easier to calculate for rough estimate.

Sum of deviations from mean is zero.

While we use a formula for computing the standard deviation, the intuitive sense of SD as a measure of spread relative to the mean should give us enough of an idea to work from in order to estimate which distributions have greater spread. e.g. from OLI the example below shows student ratings of a professor for 3 different classes, which has the greatest spread (variability)?