POLS 203 DS 3

Stats Recap 12/11/2020

WEEK 1

- What Characterizes Scientific Inquiry?

1. Inference making with certainty intervals
2. Public procedure for replicability
3. Generalizability
4. Interval validity
5. External validity

hours of studying → score at exam

Independent Variable	Dependent Variable
(Cause)	(Effect)
(Explanatory Variable)	(Response Variable)

- If a study has many **confounding** variables, that study has low internal validity.
- **Internal validity:** Is the extent to which you can be confident that a cause-and-effect relationship established in a study cannot be explained by other factors. How good your IV measures your DV.
- **Confounding:** The situation in which a change in your DV is resulted by some third variable
- **External validity:** It is about how applicable a research is to the real world. How the results can be applied to other situations than your research.

- What Statisticians Do

1. Design: Planning data collection and best method to study it

2. Description: Summarizing the data

3. Inference: Making predictions using the data

- Fundamental Concepts

- **Data:** The observations gathered on the characteristics of interest are collectively called data
- **Population:** Total set of observations
- **Sample:** A subset of the population on which one collects data
- ***Parameter:** A numerical study of the population
- ***Statistic:** A numerical summary of the sample data

WEEK 2 VARIABLES AND MEASUREMENT

- Variable

- It is a characteristic that can vary in value among subjects in a sample or a population.
- Different subjects may have different values of a variable

  variable <- c(1,2,3,4,5)

  variable_2 <- c("Micheal", 'Pam", "Kevin")

- Measurement scale of variables

  - **Ordinal scale:** strongly agree, agree… (variables in an ordered fashion)
  - **Nominal scale:** hair color
  - **Interval scale:** income = [2500,3000] (quantitative)

- Variable types

  - **Quantitative variables:** Measures difference in quantity

- **Categorical variables:** Measures differences in quality

- Discrete and Continuous Variables
  - A variable is **discrete** if its possible values form a set of separate numbers like 1,2,3. For example, the money in your bank account
  - A variable is **continuous** if it can take an infinite continuum of possible real number values. If you try to count these kind of variables, it will take forever. e.g., stars in the universe

- Randomization
  - **Simple Random Sampling**
    - A simple random sample of n subjects from a population is one in which each possible sample of that size has the same probability of being selected
    - Randomization decreases bias, helping you to make more accurate inferences

- Experiments
  - Experiments are conducted to compare responses of subjects on some outcome measure under different conditions. Not very common method in social sciences
  - **Treatments:** Assigning treatments to subjects and planning experimental designs. In good experimental designs randomization is used to determine which treatment a subject receives.

- Observational Studies
  - In which individuals are observed. No experimental control.

- Sampling Error
  - The error that occurs when we use a statistic based on sample to predict the value of population parameter. How much statistics differ from the parameter it predicts because of the way results naturally exhibit variation from sample to sample.

- Bias
  - **Sampling bias:** Occurs from using nonprobability samples or the way you collect your data gives some groups lower or higher probability of being sampled.
  - **Response bias:** Incorrect answer from the subject or the way you ask your question changes the response of the subject.
  - **Missing data:** You do not have the data of sampled subjects maybe because they refuse to participate (also known as **nonresponse bias**) or do not answer your questions which were asked to measure specific variables.
- Methods of sampling
  - **Systematic random sampling:** n = sample size, N = population. N/n, population divided by sample size. Select a random number from the first x number of subjects and then select every xth subject listed after x. 2,22,42,62…
  - **Stratified random sampling:** Divides the population into separate groups called strata and select a simple random sample from each stratum. Stratifying according to the political party identification is an example.
  - **Cluster sampling:** Divide the population into a large number of clusters, such as city blocks or countries (heterogeneousity is an important concept here). Select a simple random sample of the clusters. Use these subjects as the sample. Think of the New York City and its blocks. You divide the city blocks into clusters to sample 2% of the families in the city. You select simple random samples from the clusters and sample every family on each block.
  - **What's the difference between a stratified sample and a cluster sample?**
    - A stratified sample uses *every* stratum. The strata are usually groups we want to compare. By contrast, a cluster sample uses a *sample* of the

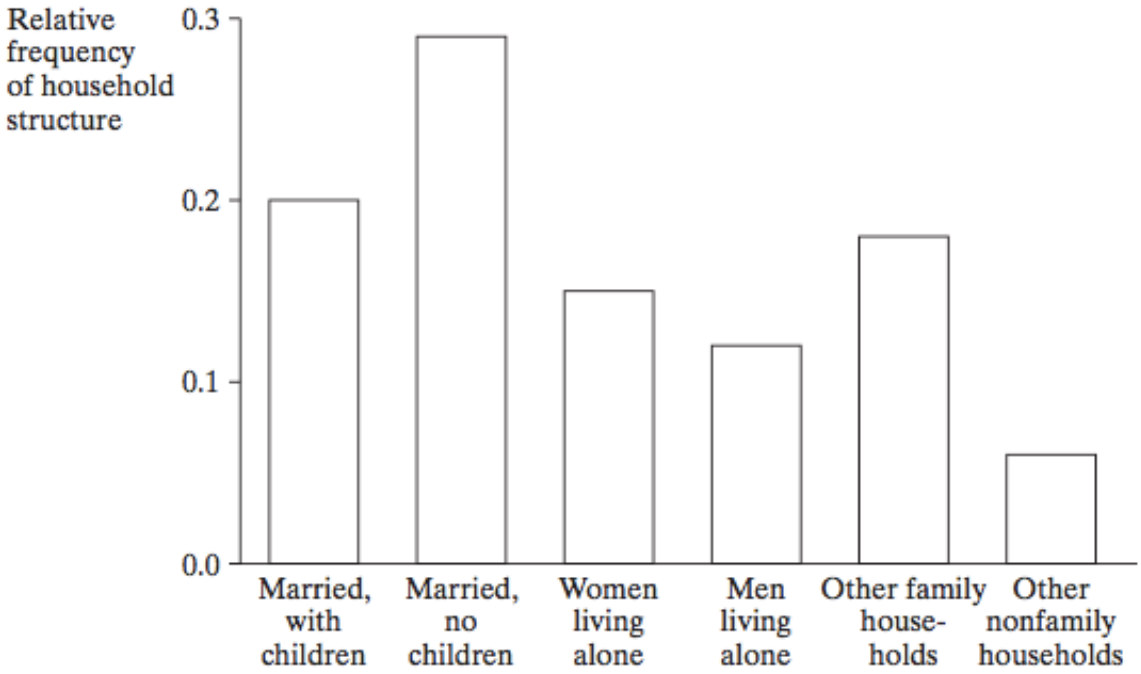
clusters, rather than all of them. In cluster sampling, clusters are merely ways of easily identifying groups of subjects. The goal is not to compare the clusters but rather to use them to obtain a sample.

WEEK 3 DESCRIPTIVE STATISTICS

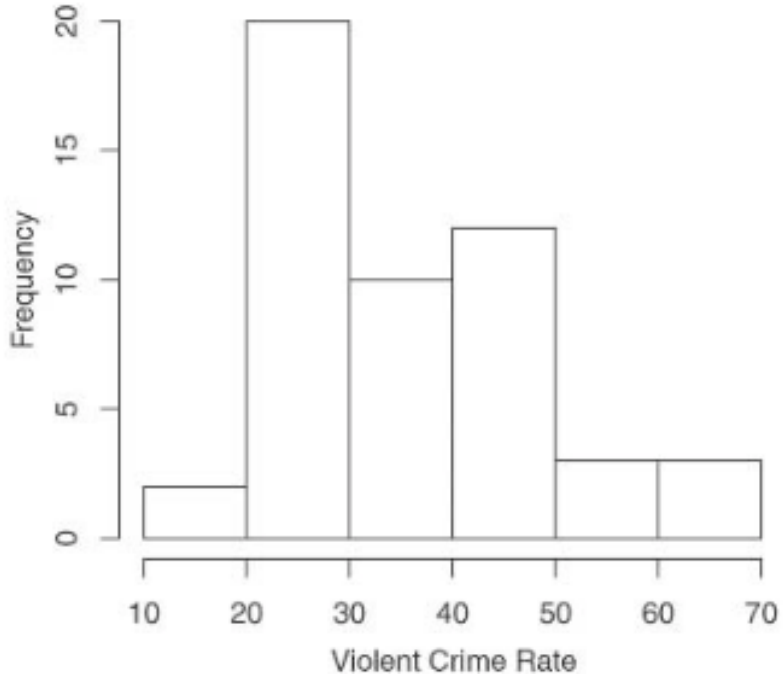
- Descriptive statistics involve:
  - Describing variables (categorical, quantitative)
  - Describing the shape of the distribution (bell-curve, skewed)
  - Describing characteristics of the data (mean, median, mode)
  - Measures of position (upper and lower quartiles)

DESCRIBING VARIABLES

- Categorical data
  - *Relative frequency* is the proportion or percentage of the observations that fall in that category. The proportion equals to number of observations in a category divided by total number of observations. Bar graph of a categorical data.

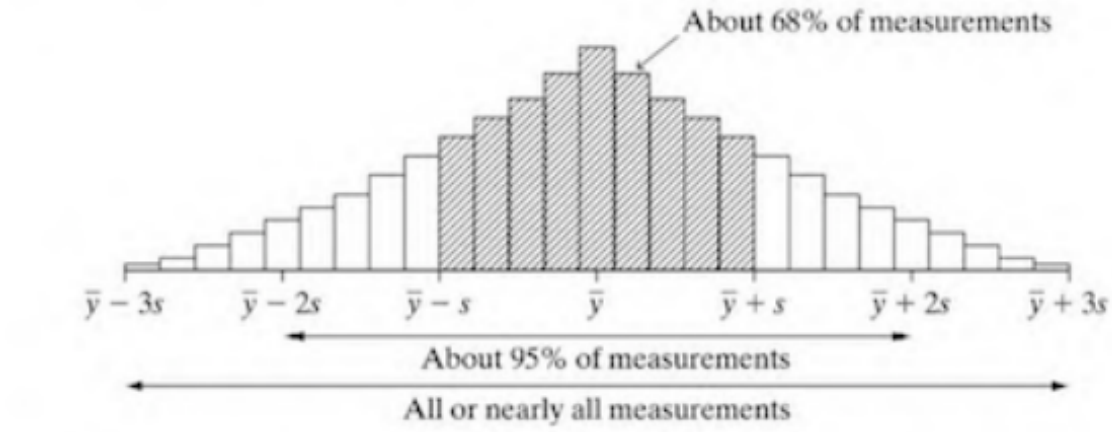


- Quantitative data
  - Category width equals to range divided by number of categories (generally 4 or 5). For instance, your range is 20 and you want to divide your sample into 4 intervals. 20/4 = 5, so the range of each category is going to be 5. {99-104}, {105-110}, {111-116}, {117-122}. Histogram of a quantitative data.

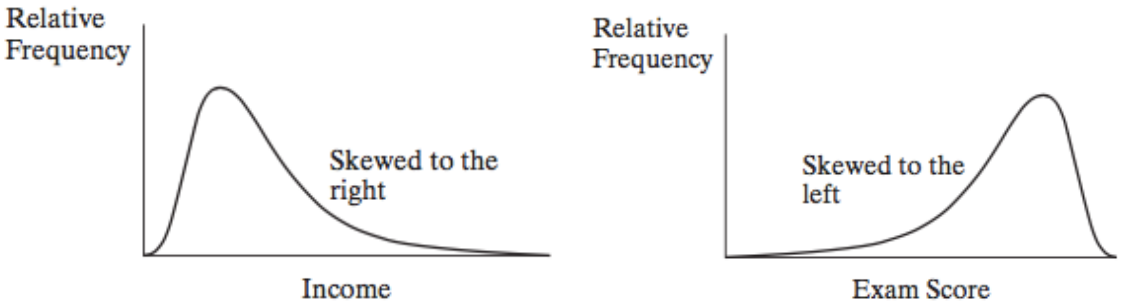


DESCRIBING THE SHAPE OF THE DISTRIBUTION

- Bell shaped distributions
  - Mean = 0, sd = 1. Symmetric distributions. Very rare and hard to achieve.



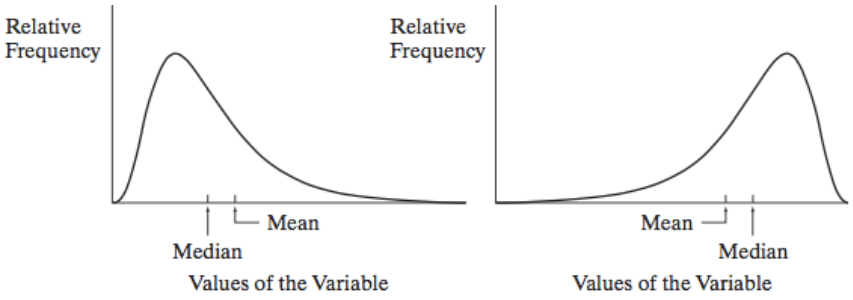
- Skewed distributions



DESCRIBING THE CHARACTERISTICS OF THE DATA

- Mean
  - Sum of the observations divided by the number of observations. Often called the *average.* $\bar{y} = \frac{\sum y_i}{n}$.
  - Valid for quantitative variables.
  - The mean can be highly influenced by an observation that falls well above or well below the bulk of the data, called an *outlier*.
  - The mean is pulled in the direction of the longer tail of a skewed distribution, relative to most of the data.

- Median
  - The median is the observation that falls in the middle of the *ordered* sample. When sample size is odd, a single observation is the mean. When the sample size is even, the median is the average of the two middle observations. Median score is (n+1)/2.
  - Valid for quantitative and ordinal variables. Cannot be used for nominal variables because there is no order in them.
  - For symmetric distributions such as the bell shaped distribution, mean = median. For *right skewed distribution*, mean > median and for *left skewed distribution,* mean < median.



- Median vs. Mean
  - The median is usually more appropriate than the mean when the distribution is highly skewed. The mean can be affected greatly by outliers, whereas the median is not.
  - If symmetric, you can use either because mean and median will be equal.
- Mode
  - It is the value that occurs most *frequently*.
  - The mode is appropriate for all kinds of data. Interval, nominal, ordinal.
  - A frequency distribution is called *bimodal* if two distinct mounds occur in the distribution.
  - In symmetric distributions, such as the bell shaped distribution, mean = median = mode.

VARIABILITY

- Range
  - It is the difference between the largest and smallest observations.
- Standard deviation
  - **Deviation:** The deviation of an observation y from the sample mean. The difference between an observation and the mean.

    The ***standard deviation*** s of n observations is

    $$s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{\text{sum of squared deviations}}{\text{sample size } -1}}.$$

    This is the positive square root of the ***variance*** s^2, which is

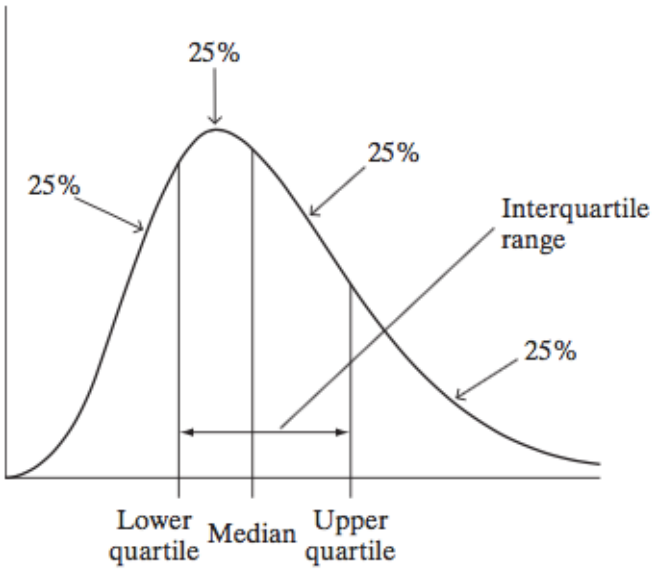
    $$s^2 = \frac{\sum (y_i - \bar{y})^2}{n-1} = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n-1}.$$

  - n-1 is used because we want our sample to be as close as possible to the population. n-1 is used for sample and n for population.
  - sd ≥ 0
  - s = 0 when all observations have the same value. For example, ages for a sample of five students are 20, 20, 20, 20, 20. The sample mean equals to 20 and each of the five deviations are equal to 0.
  - The greater the variability about the mean, the larger is the value of standard deviation.

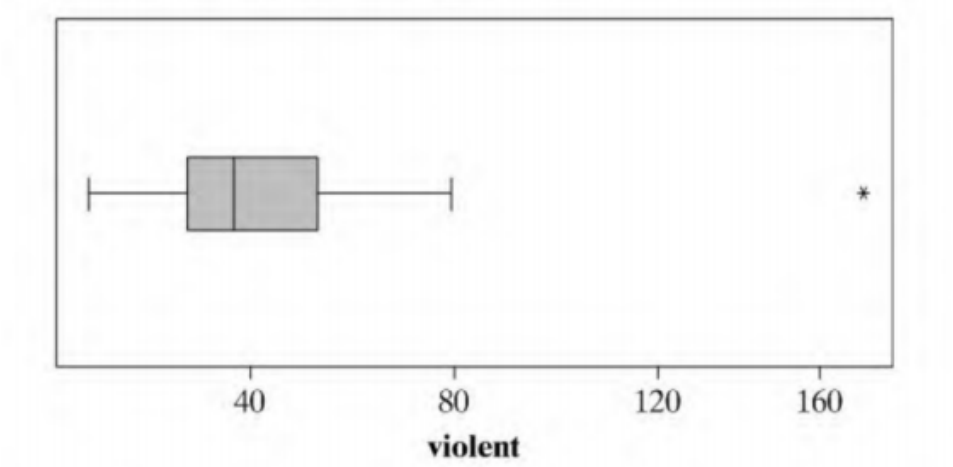
MEASURES OF POSITION

- Percentile
  - The pth percentile is the point such that p% of the observations fall below or at that point and (100-p)% for above it.
  - Substituting p = 50 gives the 50^{th} percentile. This is the *median*. In proportion terms, a percentile is called a *quantile.* The 50^{th} percentile is the 0.50 quantile.

- The 25th percentile is called the *lower quartile.*
- The 75th percentile is called the *upper quartile.*
- One quarter of the data fall below the lower quartile. One quarter fall above the upper quartile.
- The lower quartile is the median for the observations that fall below the median, that is, the bottom half of the data. The upper quartile is the median for the observations that fall above the median, that is, the upper half of the data.
- Using summary() function in R, you can access upper and lower quartiles. They are labeled as 1st Qu. and 3rd Qu.



- Box plots



- Box plot summarizes center and variability.
- Box contains 50% of observations.
- Line in the box denotes median.
- Beginning and the end of the graph show maximum and minimum values.
- Left whisker represents the bottom 25% of the data.
- Right whisker represents the upper 75% of the data.
- Star on the far right represents an outlier value.
- This is a left skewed distribution. Remember mean tends to be pulled to the direction of the longer tail. More observations of lower value and some outliers of larger values.