ISS 305:002
Evaluating Evidence:
Becoming a Smart Research Consumer

6. Problems of description

Reminder: Turn on your I<CLICKER

Problems of description

I. Descriptive Statistics
   a. Central Tendency
   b. Variability
II. Other distributional features
III. Proportions
IV. Graphic descriptions

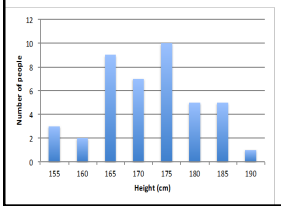
Descriptive Statistics

- Helps us get to know our data
- You will need to know:
  - Measures of Central Tendency = Mean, median, mode
  - Measures of Variability = Range, variance, standard deviation
- Look at these descriptive statistics on variables of interest before using inferential statistics

• Inferential statistics:
- Draw conclusions/make inferences that go beyond the scores from a research study
- Look at the relationship between two or more variables

Variables have distributions

- A variable is something that changes or has different values (e.g., heart rate, speed, etc.)
  - This is what you measure or manipulate in a study
- A distribution is a collection of measures, usually across people.



People have different heights

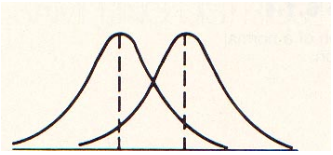
This is how height in a given sample of people is distributed

Descriptive Statistics

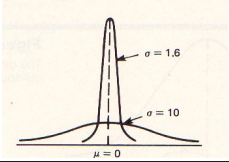
- Procedures for organizing, summarizing, and describing a large amounts of data with a few numbers
- Two Major Types:
  - Measures of Central tendency **("Averages")** – where do values tend to center?
    - Mean, median, mode
  - Measures of Variability – how spread out are the values?
    - Range, variance, standard deviation

Descriptive Statistics

- **Central Tendency ("averages")** refers to the middle of the distribution



- **Variability** is about the spread

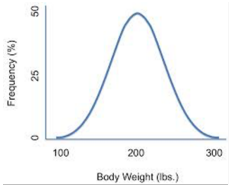


Descriptive Statistics

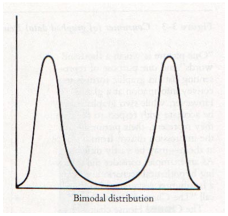
- Central tendency ("averages"): **What is the typical score or most representative value of a group of scores?**
  - The **mean** is the arithmetical average
  - The **median** is the score that divides a distribution in half so that half the scores fall above and half fall below this value
  - The **mode** is the most frequent score in the distribution

Central Tendency ("Averages"): The Mode

- The mode – the most frequently occurring score. Midpoint of most populous class interval. Can have bimodal and multimodal distributions.



Unimodal distribution



Bimodal distribution

…can also have multimodal

Central Tendency: The Mode

- Mode
  - Most common single number in a distribution
    - Value with the highest frequency
  - Good for describing the central tendency for a nominal variable
  - Find out mode for the following scores:
    2, 2, 4, 4, 5, 5, 5, 6, 6, 8
    Mode = 5 (unimodal distribution)
  - Find out mode for the following scores:
    3, 4, 4, 4, 5, 6, 6, 7, 7, 7, 8
    Mode = 4 and 7 (bimodal distribution)
  - When distribution is unimodal and symmetrical, mean = mode
  - Any other shape of distribution → mean and mode are not the same value

Central Tendency ("Averages"): The Median

- Score that separates top 50% from bottom 50% (is the middle most score)
- Even number of scores, median is half way between two middle scores.
  1 2 3 4 | 5 6 7 8 → Median is 4.5
- Odd number of scores, median is the middle number
  1 2 3 **4** 5 6 7 → Median is 4

Central Tendency ("Averages"): The Median

- Median
  - The middle score (when distribution is lined up from lowest to highest)
  - When distribution is unimodal and symmetrical, mean = mode = median
  - Other shapes of distribution: mean, mode, and median may all be different values
  - Benefits of median
    - Unlike mean, it is less influenced by extreme values (outliers) and can sometimes be more representative of a group of scores when there is an extremely large or small score in the data

Central Tendency ("Averages"): The Median

- Median
  - What's the median of the following scores?
    8, 2, 5, 3, 5
    - Step 1: Line them up from lowest to highest
      2, 3, 5, 5, 8
    - Step 2: figure out the total number of scores and thus which score is the median
      Total number of score is 5; which is an odd number.
      In this case, the median will be 5/2 + .5 = 2.5 + .5 = 3rd score
    - Step 3: figure out the median
      3rd score in this distribution is 5; the median is 5

Central Tendency ("Averages"): The Median

- Median
  - What's the median of the following scores?

    8, 2, 5, 3, 5, 258
    - Step 1: Line them up from lowest to highest

      2, 3, 5, 5, 8, 258
    - Step 2: figure out the total number of scores and thus which score is the median

      Total number of score is 6; which is an even number.

      In this case, the median will be 6/2 = 3; the median is the average of 3rd and 4th score
    - Step 3: figure out the median

      3rd score in this distribution is 5, 4th distribution is 5, the median is (5+5)/2 = 5
  - **In this case, unlike mean (as we will see shortly), the median is not influenced by adding an extreme score, or an outlier**

Central Tendency: The Mean

- M = or $\bar{X} = \frac{\sum X}{N}$
- M = add up all the numbers in a set and divide by the total number of items in the set.
  - The perfect balancer: There is an equal weight of values above and below it
  - Best representation of what you can expect for a set of numbers
  - Best guess if you want to be closest to the individual values

Properties of the mean

Means...

- Are easy to compute
- Lend themselves to useful inferences
  - (because means have certain useful statistical properties)
- Are sensitive to **every** score, especially to extreme scores
  - Generally, changing a score will change the mean
  - Adding or subtracting a score will affect the mean
  - Adding or subtracting an extreme score can powerfully affect the mean
- Can (falsely) give an illusion of great precision

Ok that's all fine…

…but how do I know when to use the mean, median, or mode to describe my data?

What are the advantages and disadvantages of each measure of central tendency?

Comparison of mean, median and mode

- Mode
  - Good for nominal variables
  - Good if you need to know most frequent observation
  - Quick and easy
- Median
  - Good for "bad" distributions
  - Good for distributions with arbitrary ceiling or floor
  - Good for ordinal data

Comparison of mean, median and mode

What is the favorite color of MSU students?

How do we characterize the typical favorite color?

We can't take the mean or median…they are not numbers.

Mean color value? No, that's not really representative either. 

Mode is the only option! 

Can only use mode to characterize central tendency for nominal data

…remember NOIR?

Comparison of mean, median and mode

- Mean
  - Most reliable (normally) and commonly used
  - Only used for interval or ratio data
  - Used for inference as well as description; best estimator
  - Based on all data in the distribution (unlike the mode and median)
  - Generally preferred except for "bad" distributions. Most commonly used statistic for central tendency.
  - **Sensitive to every score (outliers!)**

When the mean is not the best

When your data are categorical

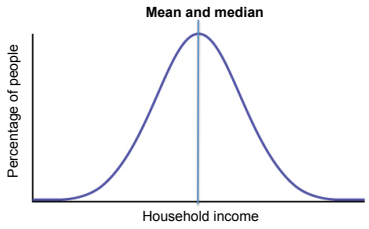
- Then the mean is not meaningful
- Use the mode instead

Political orientation
1=democrat, 2=republican, 3=independent

Mean=1.7 (what does that mean?)
Mode=2

Mean versus median

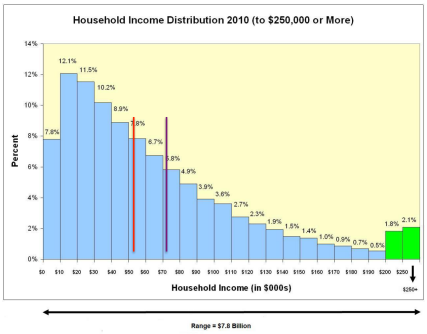
…let's illustrate with an example



When distributions are normal, mean and median will be very close…

…in these cases use the mean

Mean versus median



Median income: $52,000

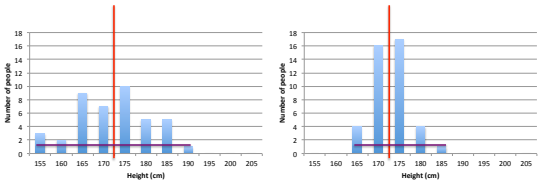
Mean income: $73,000

So if you want to know the most representative income, you'd chose the median…

Mean can be misleading when distributions are skewed or when there are outliers.

Variability: Influence of Distribution Shape

Spread of the distribution—how much do values fluctuate?



These two distributions have equal centers but different spreads

Descriptive Statistics

- **Variability**: How much do the scores fluctuate?
- How spread out the scores are around the mean?
- Indicates how similar or different is an individual score in the data to or from the mean?
  - The **range** is the difference between the largest and smallest value (can be too sensitive to extreme scores)
    - Range = high score minus low score

      12 14 14 16 16 18 20 = range = 20 – 12 = 8
  - The **variance** is the average squared deviation of each score from the mean
  - The **standard deviation** is the square root of the variance.

Variance and Standard Deviation

Tells you how far apart the data values are

- **Variance (s^2)** $$s^2 = \frac{\Sigma(X_i - M)^2}{n - 1}$$
  - mean squared deviation
  - Variance is average **squared** deviation from the mean.
  - To return to original, **unsquared** units, we just take the square root of the variance. This is the standard deviation.
- **Standard Deviation (SD)**
  - index of the average amount that scores in a group can be expected to differ from the mean of the group, or to vary from one another $$SD = \sqrt{s^2} = \sqrt{\frac{\Sigma(X_i - M)^2}{n - 1}}$$

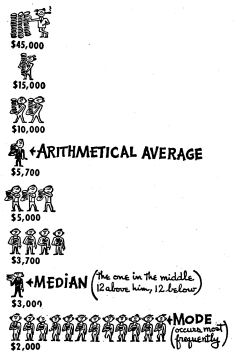
Mean vs. SD Example

	Group 1	Group 2
Person 1	6' 2"	6' 9"
Person 2	6' 1"	6' 5"
Person 3	5' 11"	5' 10"
Person 4	5' 10"	5' 0"
Mean	**6'**	**6'**
SD	**1.58**	**7.97**

Slippery "averages"

Slippery "averages"

- Q: Why would one choose the mean vs. the median?
- Who might prefer to use the mean?
- Who might prefer to use the mode?
- Who might prefer to use the median?
- Again, remember that if the distribution were symmetric, Mean=Median
  - Is the distribution symmetric here?



Slippery "Averages" – NFL Salary Example

- Are NFL players paid a lot?
- What's the "average" salary in the NFL?
- For the 2013-2014 season, here were the figures (using base salary):
  - N = 2454 players (Included players cut before the season).
  - Mean = arithmetic average = $1,064,704
  - Median = $555,000
  - Mode = $405,000
  - SD = $1,500,218
- Why so different?
- MORAL:

Slippery "Averages"

- A basic problem is trying to boil all the useful information in a set of observations down to a single summary number.
  - You often lose vital information
- For example,
  - For a diver deciding if its safe to dive:
    - Two swimming pools may have the same average depth (12 feet), but differ importantly in how safe they are to dive into.
  - For a retiree deciding where to live:
    - Two cities (e.g., Oklahoma City and San Diego) may have roughly the same average annual temperature, but differ importantly in how comfortable the climate is.
- Moral: We often need to know more than some central tendency statistic

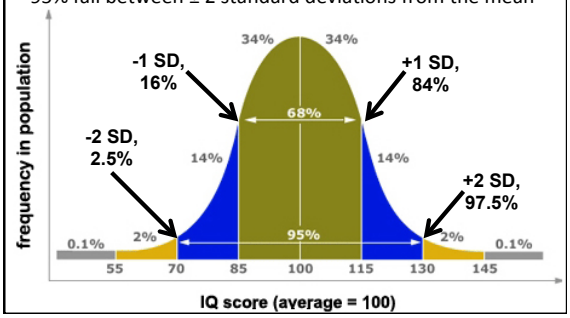
The Normal Distribution

- Many attributes (IQ, height, etc.) distributed "normally"= bell shaped curve
- Why?

- There are lots of bell shaped/normal curves
- However, for each, the standard deviation conveys the same useful information...

The Normal Distribution

- 68% fall between ±1 standard deviation from the mean
- 95% fall between ± 2 standard deviations from the mean

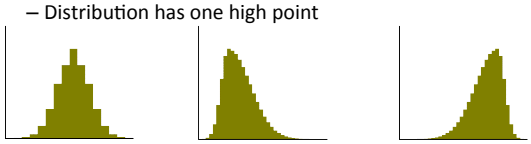


The Normal Distribution

- So, if one can assume a normal distribution, and knows the mean and standard deviation, one can make rough percentile estimates
  - e.g., assume Normal distribution with Mean = 12, standard deviation = 3
    - what is percentile rank of a score of 9?
    - of 18?
    - of 16?
- For an Exam – Mean = 42.61 and Standard Deviation = 9.8, so if normally distributed
    - -2 sd ~
    - -1 sd ~
    - 0 sd =
    - +1sd ~
    - +2sd ~
  - knowing this, one can approximately estimate the percentile rank of any score
    - e.g., score of 53 is near +1 sd, so something just over 84% (e.g., let's guesstimate around 85%. This student did better than ~85% of the students in the class.

Other useful distributional features

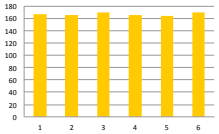
- Pattern or shape of how frequencies are spread out in the data
- Unimodal distribution
  - Distribution has one high point



- Bimodal distribution
  - Distribution has two fairly equal high points

Other useful distributional features

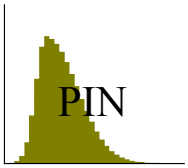
- Rectangular distribution
  - All values have about equal frequency



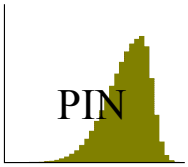
- Symmetrical distribution
  - About equal numbers on both sides of the middle

Other useful distributional features

- Skewed distribution
  - Asymmetrical distribution
  - Positively skewed distribution
    - Majority of the scores are at the lower end of the distribution
    - Floor effect



  - Negatively skewed distribution
    - Majority of the scores are at the higher end of the distribution
    - Ceiling effect



Proportions or Percentages

- Percentage makes a comparison between two numbers – a "base" number and a second number to be compared with that base.
- E.g.1: % of dog owners = (# dog owners) / (# pet owners)
- E.g.2: % of budget for interest = ($ for interest) / (all $ spent)

Problems with Proportions or Percentages

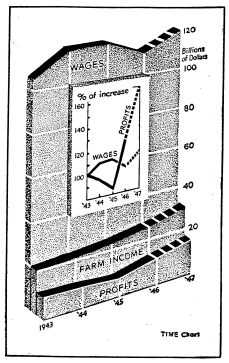
1. When the base is small.
- E.g., survey finds that 50% of those interviewed were in favor of a tax increase
- Small changes in numbers make big changes in the % when the base is small
- If base is small, usually more informative to see the numbers themselves rather than the %s

Problems with Proportions or Percentages

2. Combining percentages when the bases are different.
  - E.g., "Wages have gone up 5%, but taxes have also gone up 5%, so citizens have nothing to show for their raises."
    - But, if average wage is $20,000, and average tax liability is $5000, wages have gone up $1000, while taxes have only gone up $250.
    - If average wage is $10,000, and average tax liability is $15,000 (including taxes on unearned income), then wages have gone up only $500 while taxes have gone up $750.
  - E.g. "Last year we took a 10% wage cut, and this year we caught up with a 10% wage increase"
    - If last year's wage was $10,000, the cut resulted in $1000 lost. But the 10% increase is on the new, lower salary ($9000), so one is still behind (by $100)
  - "We cut prices by 50%. Now we're closing out and taking another 20% off, so you save a total of 70%.
    - If original cost was $100, first cut resulted in $50 off or new price of $50. Second cut was on the $50 base, and was $10, resulting in final cost of $40. Net reduction is 60%, not 70%.

Problems with Proportions or Percentages

3. Comparisons of <u>increases/ decreases</u> in %s.
- (e.g., see p. 119, in Huff)
  - % increases in wages < % increases in profits
  - trying to imply that wage increases lag far behind profit increases
- Need to know what the bases are to meaningfully compare %s
- If they're not the same, comparisons can be misleading



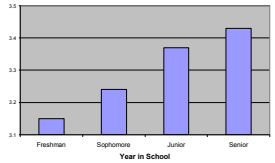
Redrawn with the kind permission of TIME *magazine as an example of a non-lying chart.*

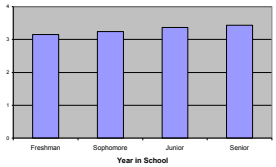
Problems with Proportions or Percentages

4. When the base is unclear or arbitrarily chosen
- E.g., % increases in foreclosures
  - what's the base?
    - last year vs.
    - last month
    - what?
- E.g., Using old vs. new price.
  - Suppose old price was $100 and new price is $50. What percent is saved?
    - if you use the old price as a base, then it's $50/$100 =.5 or 50%
    - if you use the new price as the base, its $50/$50 = 1.0 or 100%
- Again, you need to know what the base is (and judge whether it is reasonable).

Graphical illusions/deceptions

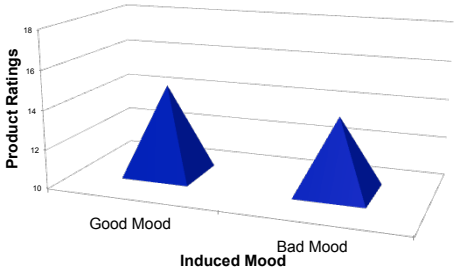
- Suppose I'm a freshman at MSU, and I don't do well my 1st year.
- My parents are concerned, but I tell them, "don't worry, everyone struggles their first year, but things get better later"
- And I present them with graphical evidence on mean GPA at MSU by year of study
- What's wrong with my argument?





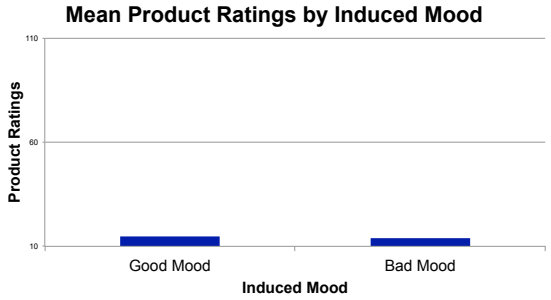
The 2-DRL to Figures!

- Use 2-D. DON'T use 3D figures! The 3rd dimension is useless and misleading.



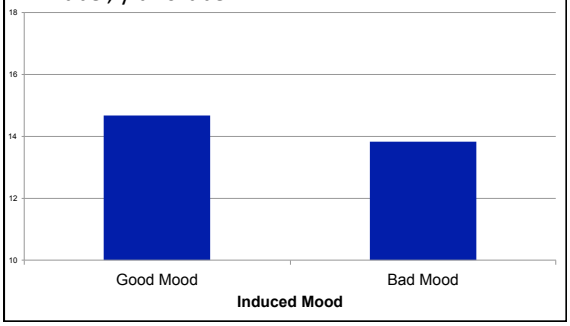
The 2-DRL to Figures!

- Scale the y-axis to include the **Range** of scores of your DV in your study.

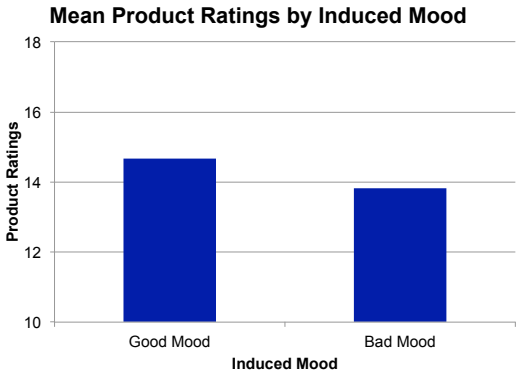


The 2-DRL to Figures!

- Everything gets a **Label**: Figure name, x-axis label, y-axis label.

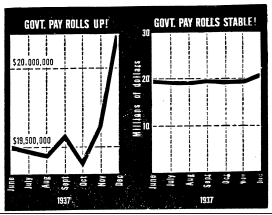


Perfect!



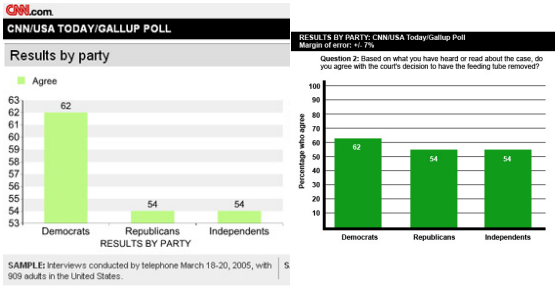
Graphical illusions/deceptions: The "Gee whiz" graph

- Creating an impression of larger vs. smaller differences by
  - stretching/compressing/zooming in on the axes of a graph
    - usually the vertical/Y axis
    - but sometimes the horizontal/X axis
  - and/or breaking the Y axis
    - i.e., leave out some or most of the full range of the Y axis



Graphical illusions/deceptions: Gee Whiz graphs

- People were asked – Based on what you have heard or read about the case, do you agree with the court's decision to have the feeding tube removed?



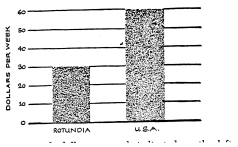
Graphical illusions/deceptions: Gee Whiz graphs

- In looking at graphs, you need to
  - Make sure all axes are clearly labeled
  - Make sure that either there are no breaks in an axis or that one is fully aware of the break (and can correct for it)
    - note: sometime the break is only on the Y-axis, not the bars or full chart
  - Make sure the range through which the graph varies is a meaningful one.

Graphical illusions/deceptions: The "One-dimensional Picture"

Using two (or even three) dimensional objects comparing two groups or conditions which are actually only being compared on a single dimension.

- Illustration
  - simple bar graph accurately conveys the difference in incomes in the two countries
  - but graphic does not. Why?



Graphical illusions/deceptions: Comparing apples and oranges

- Focusing on one feature of a graph when it is another feature which is actually being depicted
- For example, Huff's "Darkening Shadow" (see p. 105 of Huff)
  - Graph shows how much of total income is spent by Federal Government in 1954
    - Clear impression is that most income is spent by Feds
  - A second graph depicts the exact same data
    - It's impression is that very little income is spent by Feds
  - What's going on?
    - Graphic is based on the population of the states (relative to total US population), not state's land area
    - But this is not evident, and the viewer is likely to focus only on land area
- More modern examples
  - http://www-personal.umich.edu/~mejn/election/2012/
  - http://popvssoda.com/