Summarizing/Describing Data

- Frequencies
- Pictures
- Descriptive Statistics
Summarizing Data with Descriptive Statistics: Types

- Measures of Central Tendency
  - Means
  - Medians
  - Modes

- Measures of Dispersion
  - Range
  - Standard Deviation
Measures of Dispersion

- Measures of Central Tendency don’t tell you about how much the data values differ from each other. *E.g.*, ages
  - 50 50 50 50 50 (Mean and Median = 50)
  - 10 20 50 80 90 (Mean and Median = 50)

- Measures of Dispersion or Variability attempt to quantify the spread of the distribution.

- Most common measures: **range**, **standard deviation**
The Range

- Simplest Measure of Dispersion. It is the difference between the smallest and largest values (Not for use with nominal variables since they can’t be meaningfully ordered!)
  - 50 50 50 50 50 Range: 50-50=0
  - 10 20 50 80 90 Range: 90-10=80

- A large value tells you that the largest and smallest values differ substantially but it doesn’t tell you much about the variability between the largest and smallest value.
  - 10 20 50 80 90 Range: 90-10=80
  - 10 11 12 13 90 Range: 90-10=80
Standard Deviation I

- Most commonly used measure for interval level data. It is very clever but, unfortunately, not the most intuitively obvious statistic.

- The basic idea: how different are the values in a distribution from the mean? The greater the dispersion, the less typical the mean.

- The problem is that the simple, intuitive approach doesn’t work: Sum of the Distances from the Mean/N of Case
  
  \[
  \frac{(28-75) + (29-75) + (30-75) + (98-75) + (190-75)}{5} = 0
  \]

- This will occur for any distribution because we have in effect defined the mean as precisely that point where the variation are balanced. But the notion that we ought be able to measure dispersion by comparing the closeness or remoteness of cases from the mean remains appealing…enter the standard deviation.
Standard Deviation 2

- Standard deviation builds on our intuition with a mathematical procedure that eliminates tendency of opposing distances to cancel each other out by squaring the distances (thereby eliminating the negative signs), averaging the squares around the mean, then taking the square root of the result so as to return the original units of distance among the values.

- Formula:

\[
\sqrt{\frac{\text{sum of squared distance from the mean of all cases}}{\text{number of cases} - 1}}
\]

- Example:

28 29 30 98 190 (Mean=75)

\[
\sqrt{\frac{(28-75)^2 + (29-75)^2 + (30-75)^2 + (98-75)^2 + (190-75)^2}{5 - 1}} = 70.89
\]
Standard Deviation 3

- The higher the standard deviation, the more spread out the values.

- But how do you use it (typically along with the mean)?
  1. Compare similar variables
  2. Compare different variables
  3. Compare values within a distribution
Standard Deviation 4

1. Compare Similar Variables
8th circuit: 28, 29, 30, 98, 190…  Mean=75, std. dev.= 70.89
7th circuit: 12, 35, 34, 150, 350…  Mean=75, std. dev.= 20
Mean is more representative for the 7th, than 8th.

2. Compare Different Variables: Coefficient of Variation
Age: Mean=46.56, std. dev.= 17.330
Education: Mean=13.23, std. dev.=2.895

Two variables not measured on the same units, plus age typically has higher standard deviation because more values. Can you compare them? Yes
2. Compare Different Variables: Coefficient of Variation

Age: Mean = 46.56, std. dev. = 17.330
Education: Mean = 13.23, std. dev. = 2.895

Two variables not measured on the same units, plus age typically has higher standard deviation because more values. Can you compare them? Yes via the coefficient of variation, which expresses the standard deviation as a percent of the mean.

\[
\text{Coefficient of Variation} = \frac{\text{standard deviation}}{|\text{mean}|} \times 100
\]

So you can compare the variability of different variables. Would = 100% if the standard deviation = mean.

Returning to the two examples above, the coefficient of var. for age = 37%; for education it is 22%. This means that, compared to their means, age varies more than education.
3. **Compare Values within a Distribution: Standard Score**

- Is a 65 on an exam a bad grade? Not if it's the best in the class. Is an 85 a good grade? Not if it's in the bottom 25%. But if all you know is your grade and the mean you can only tell if your score is less than, equal to, or greater than the mean. You can't say how far it is from the average unless you know the standard deviation.

- If the average score is 70 and the standard deviation is 5, a score of 80 is quite a bit better than the rest. It is two standard deviations about the mean. But if the standard deviation is 15, the score of 80 isn't very terrific. It's less than one standard deviation above the mean.

- How do I calculate this? value-mean/standard deviation.
  - You score 80 on an exam, with a mean of 70 and a standard deviation of 5:
    \[
    \frac{80-70}{5} = 2
    \]
  - You score 80 on an exam, with a mean of 70 and a standard deviation of 15:
    \[
    \frac{80-70}{15} = \frac{10}{15} = \frac{2}{3} < 1
    \]

- This is known as the standard score. It tells you how many standard deviations a case is above or below the mean. If the score=0, the value for the case=the mean. If the value=1, the case is one standard deviation above the mean, and so on.
3. Compare Values within a Distribution: Standard Score

- For many distributions, most values fall within +/- 2 standard deviations from their mean. For “bell curves”, 68.26% of all cases lie within +/- 1 standard deviation from the mean; 95.44%, within +/- 2 (95% is within +/- 1.96); about 99% (actually 99.7%), +/- 3.

- You score 80 on an exam, with a mean of 70 and a standard deviation of 5: 80-70/5 = 2. How do you stand?