SWK: 8408 Statistics I for Social Work

Gia Elise Barboza-Salerno, MA, MS, JD, PhD
barboza-salerno.1@osu.edu
The Ohio State University
Colleges of Public Health & Social Work

Autumn 2023

Class Website

Week 3: Measures of Central Tendency/Describing Data

Introduction

Today, things get harder
Reminder: Compare your knowledge today with last week, do not compare yourself to me!

Sampling techniques

Samples must be representative of the population!

Statistical sampling is the procedure by which we select a subset from a population of interest that should be *representative* of that population. The most common sampling methods are:

- Simple random sample
- Stratified sample
- Cluster sample
- Systematic sample
- Convenience sample

The sampling distribution is the single most important concept in statistics! (next week)

Simple random sample

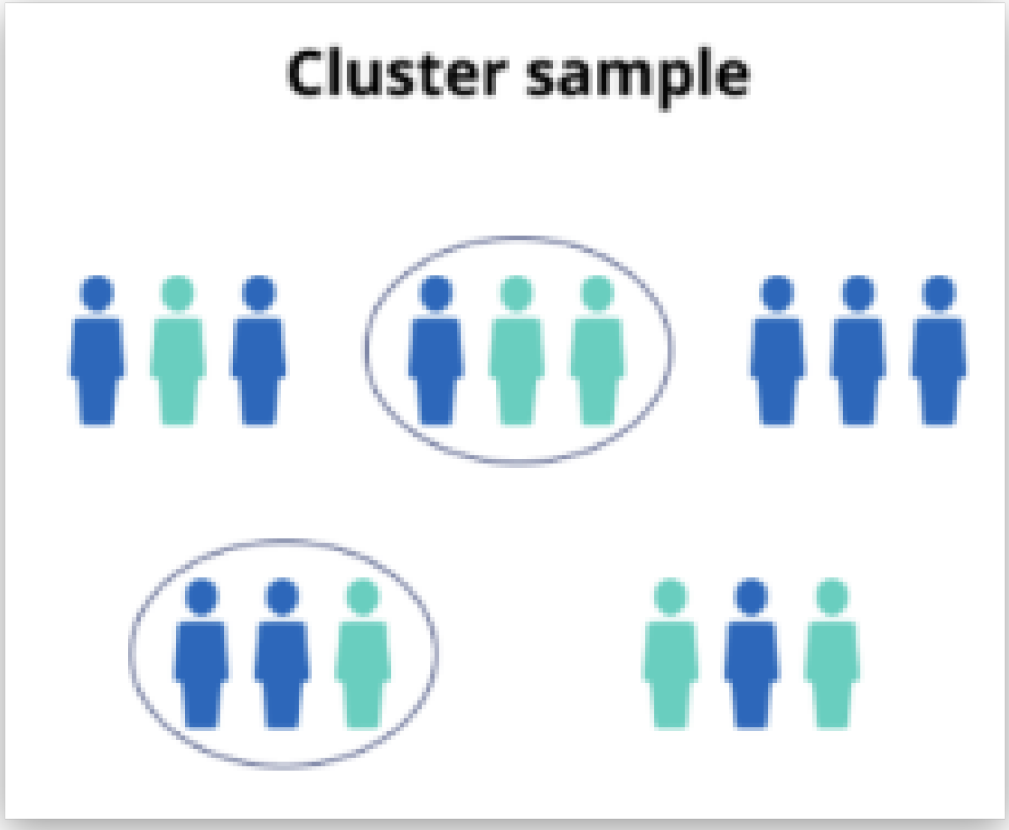
Simple random sampling (SRS): a sample selected such that each element in the population has the same probability of being selected

Stratified sample

Stratified sample: elements in the population are first divided into groups and a simple random sample is then taken from each group

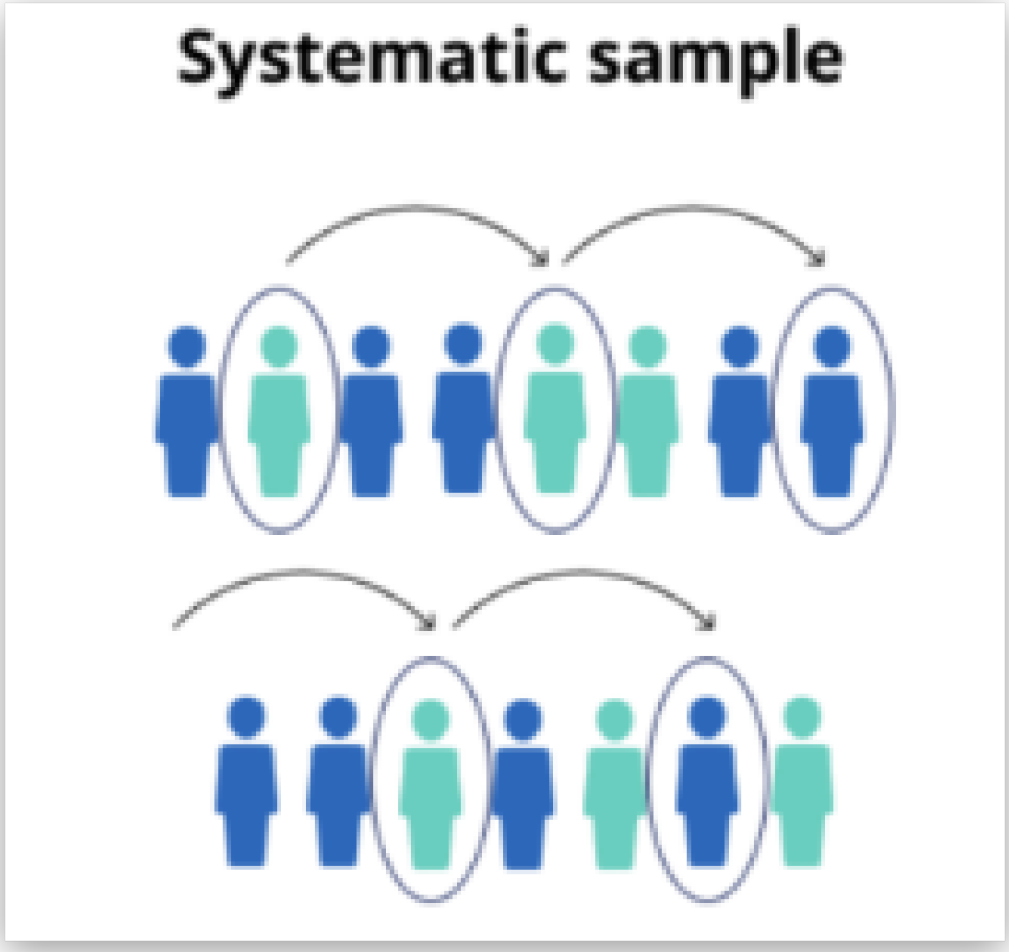
Cluster sample

Cluster sampling: the elements in the population are first divided into separate groups called clusters and then a simple random sample of the clusters is taken that all elements in a selected cluster are part of a sample



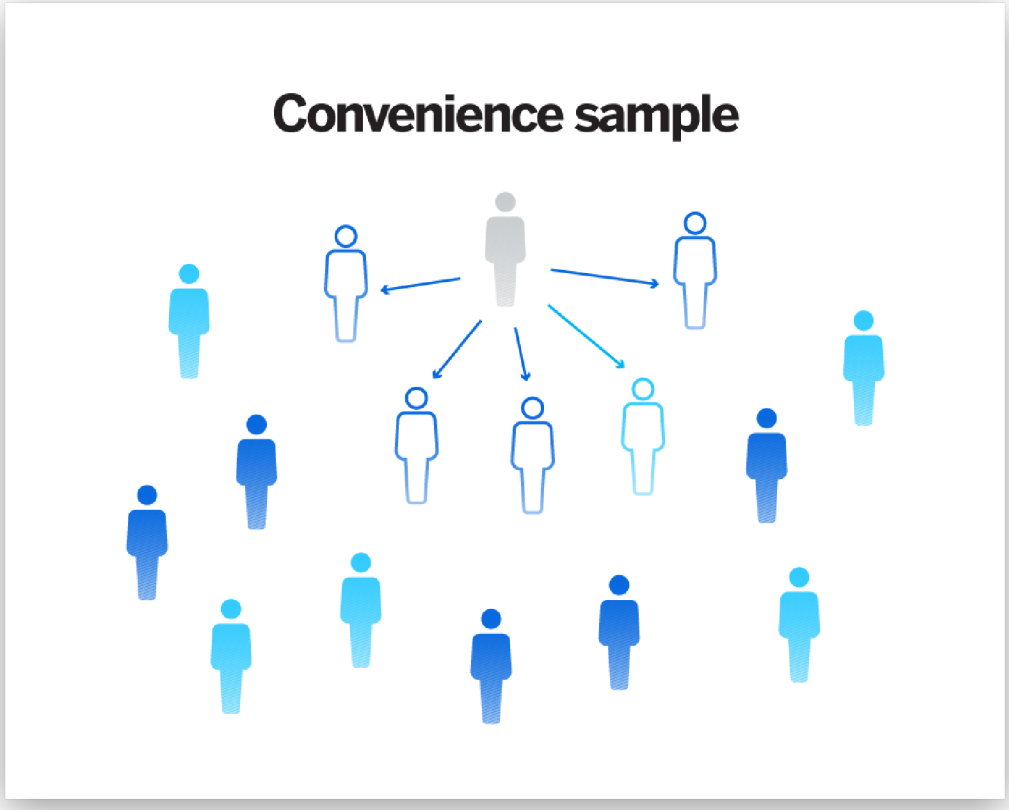
Systematic sample

Systematic sampling: randomly select one of the first k elements from the population and then every k^{th} element thereafter is picked



Convenience sample

Convenience sampling: elements selected from the population on the basis of convenience



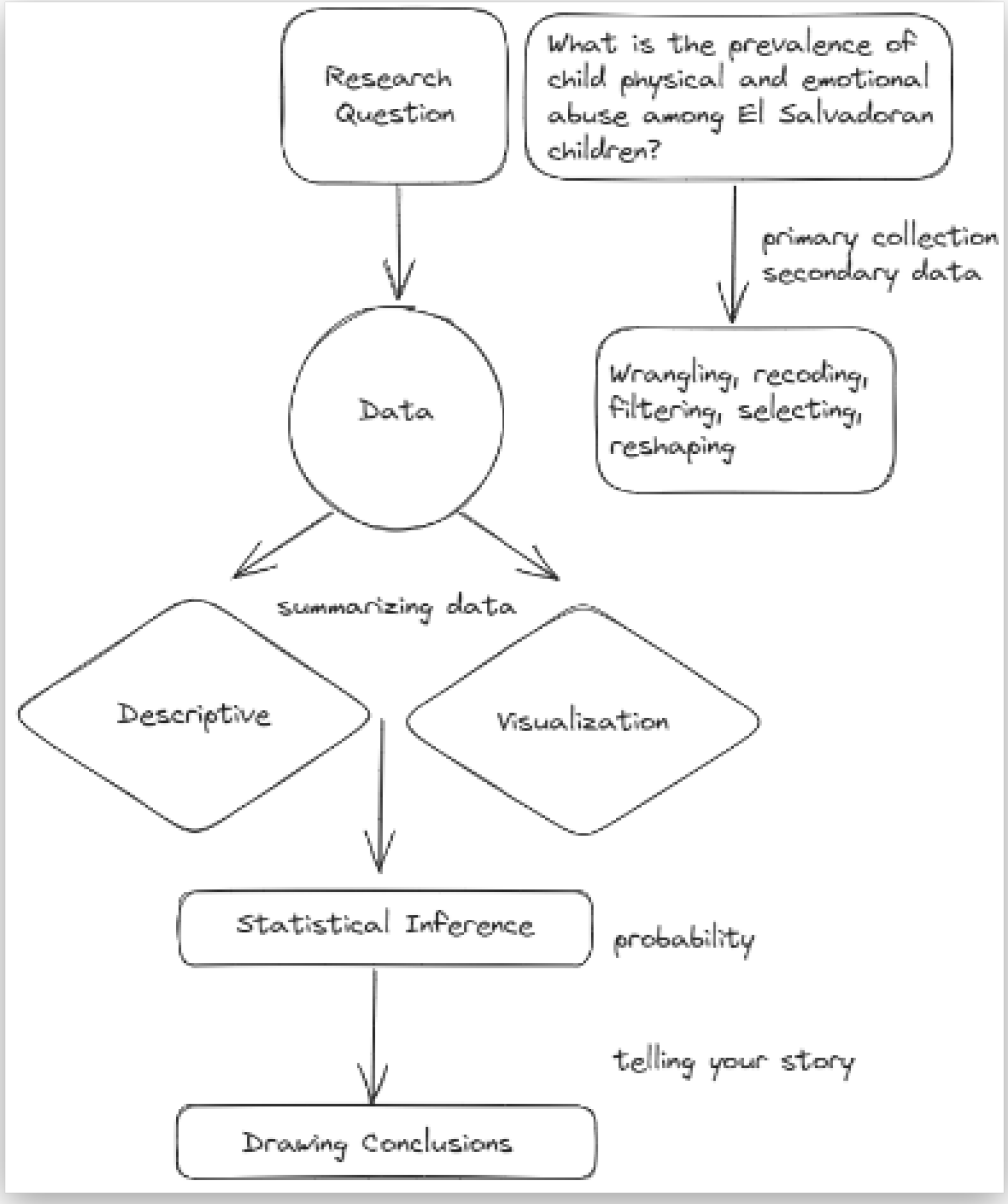
Lab 2

1. What is the prevalence of child physical and sexual abuse among El Salvadoran children?
2. Is there a relationship between abuse and internalizing symptoms?

- Background & Importance
- Data
- Summarizing data set
- Statistical analyses
- Interpretation

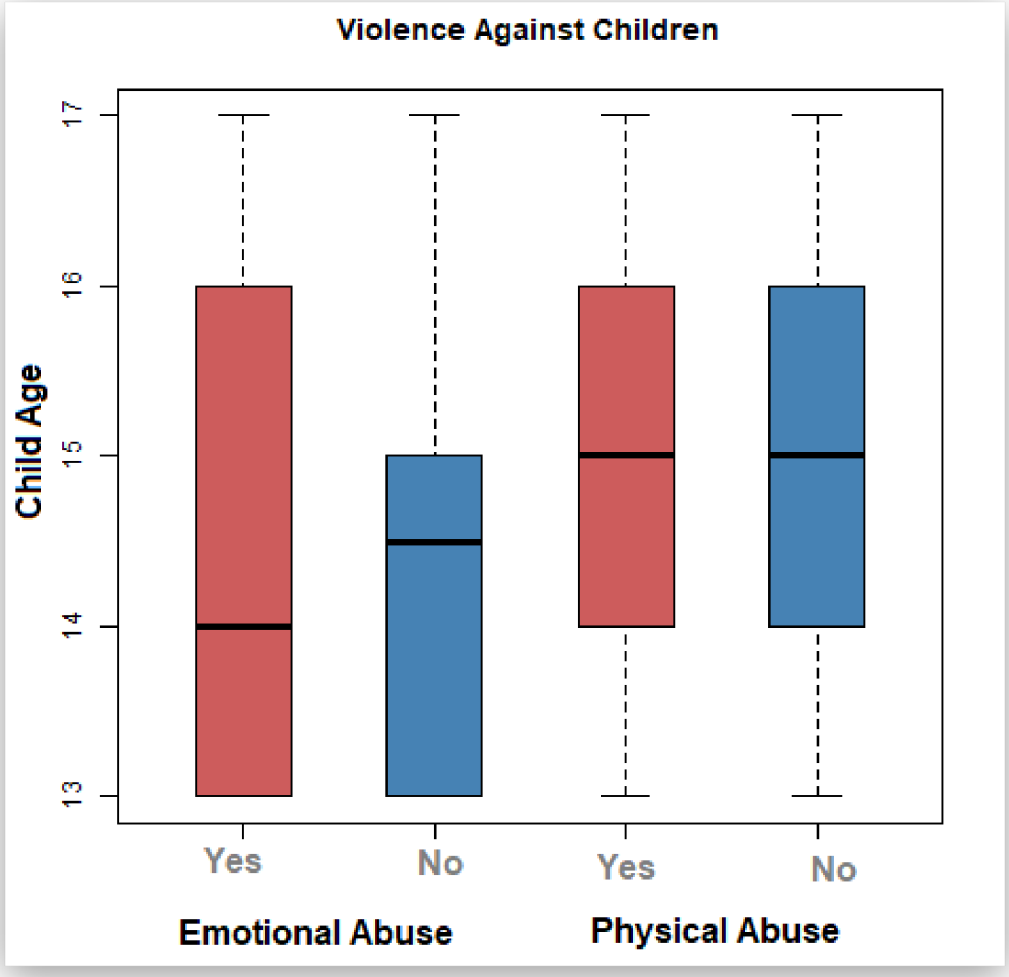
Overview

1. Importance of finding data
    - Violence against children
2. Look at codebook and identify key variables for El Salvador
3. Download data as SAS file (must request data)
    - I uploaded the data here
4. Open SAS file in SPSS
5. Locate and clean variables/make chart on next slide using R integration



To Do

1. Look at codebook and find variables on child physical and emotional abuse
2. Rename variables, run analysis, compare to codebook
3. Identify issue, select cases, compare to codebook
4. Identify issue, etc.
5. Make R plot in SPSS

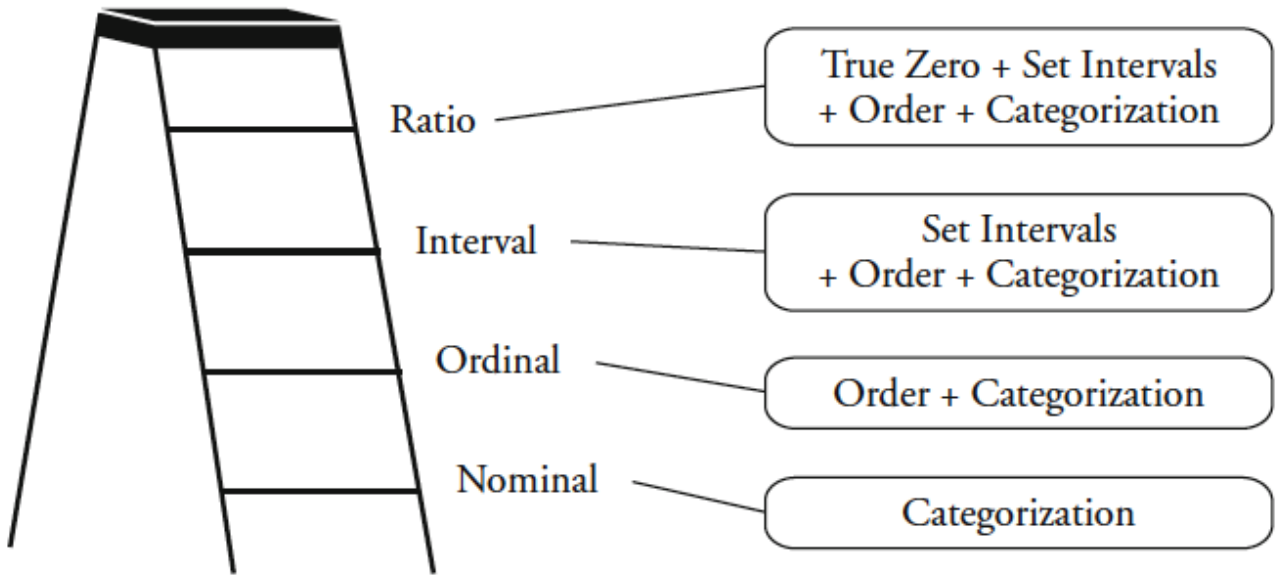


Overview

1. Measures of Central Tendency
2. Statistics in Practice
3. Measures of Dispersion
4. Measuring Dispersion: Range
5. Measuring Dispersion: Percentiles Interquartiles Range
6. Measuring Dispersion: Variance
7. Measuring Dispersion: Standard Deviation
8. Measuring Dispersion: Coefficient of Relative Variation
9. Summary and Wrap-up
10. Calculations Applications

Review Variable Measurement

Figure 2.1 *Ladder of Measurement*



Measures of Central Tendency

- The first step in summarizing research is to provide a basic portrait of the characteristics of a sample or population. What is the typical case?
- If you could choose one case to represent all others, which case would it be?
- Note that the ‘typical’ case is often very unrepresentative of the overall population
- Today we are focusing on describing data
  - Basic plotting of variables in R
  - Measures of central tendency

The Mode

Measure of central tendency for nominal variables

- If you have a variable that is nominal, how would you define the typical case?

Recall what a nominal variable is? Examples.

- It makes NO sense to describe the mean for such variables…
- Examples

How are perpetrators of DV represented in court

Category	Frequency (N)
No Attorney	20
Legal Aid	26
Court Appointed	92
Public Defender	153
Private Attorney	380
Total (**Σ**)	671

Example: How are victims of DV represented in court

Category	Frequency (N)
No Attorney	40
Legal Aid	7
Court Appointed	91
Public Defender	22
Private Attorney	70
Total (**Σ**)	230

A brief overview : Describing Representativeness

	Frequency	Percent
Black	231	21.8
Latine	195	18.4
White	621	58.6

Whites = 78.5%, Black = 6.5% and Latine = 17.6%

Click Here to download the data

Calculating the Median

- The median is a measure of central tendency that utilizes information on both the number of cases found in a particular category as well as the position of those categories
- For ordinal scales, it is the score directly in the middle of the distribution
- For interval scales, the median is the value that splits the distribution into half
- The median is referred to as the 50th percentile because it spits the distribution in half – 50% fall above the median and 50% fall below
- If you score in the 50th percentile on the SAT you didn't do well….

Starting Point: Quick Examples

- Arrange values from low to high
- Determine which score splits the distribution in half
- There are different calculations depending on whether the total number of cases is even or uneven

Example: Calculating the Median

Category	Frequency (N)	Cumulative N
Not serious at all	73	73
A bit serious	47	120
Somewhat serious	47	167
Serious	27	194
Very serious	26	220
Extremely serious	39	259
Most serious	22	281
N	281	

Median = $\frac{N+1}{2} = \frac{281+1}{2} = 141$

The Mean

- The mean takes into account not only the frequency of cases in a category and the positions of scores on a measure, but also the values of these scores.
- To calculate: add up the scores for all subjects and divide by the total

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N} i = 1 \ldots N$$

Example: Calculating the Mean

Example: calculate the mean of X and Y

i	X_i	Y_i
1	1	3
2	2	3
3	3	3
4	4	3
5	5	3

Calculating the Mean w/Aggregate Data

Total Number of ACEs	Frequency (N)	Cumulative (N)
0	4	4
1	1	5
2	2	7
4	3	10
5	3	13
7	4	17
8	2	19
10	1	20
Total	20	

Calculate the Mean?

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N} i = 1 \ldots N$$

What is N? **Answer**: 20

What are the X_i's? **Answer**: The Number of ACEs

Be careful, we are summing the number of ACEs in light of the overall frequency

- So there are four people with 0 arrests, 1 person with 1 arrest, 2 persons with 2 arrests and so on, in total there are 20 observations

Characteristics of the Mean

- Deviations or differences from the mean
- If we take each score in a distribution, subtract the mean from it, and sum the differences we will always get a result of 0
- $\sum_{i=1}^{N}(X_i - \bar{X}) = 0$
- This principle is illustrated below

Simple Example

Show that the sum of the values minus the mean for the sample is equal to 0

N	**Value** (X_i)	$(X_i - \bar{X})$
1	3	-2
2	4	-1
3	5	0
4	6	1
5	7	2
		Total $\sum = 0$

The Least Squares Property

- The least squares property is written in equation form as follows: $\sum_{i=1}^{N}(X_i - \bar{X})^2 = \textit{minimum}$
- If we then sum all of these values, the result we get will be smaller than the result we would have gotten if we had subtracted any other number besides the mean
- Try this for yourself and make sure you understand it

Using the Mean for Noninterval Scales

- The mean is ordinarily reserved for interval level scales.
- Sometimes the mean is used for ordinal level scales – is this wrong?
- Some ordinal scales have so many categories they begin to mimic interval-level measures

Using the mean for non interval scales

Is this a good idea?

Category	Frequency (N)	Cumulative N
Not serious at all	73	73
A bit serious	47	120
Somewhat serious	47	167
Serious	27	194
Very serious	26	220
Extremely serious	39	259
Most serious	22	281
Total	281	

Table: Student Views on Public Drunkenness

Statistics in Practice: Comparing the Mean and Median

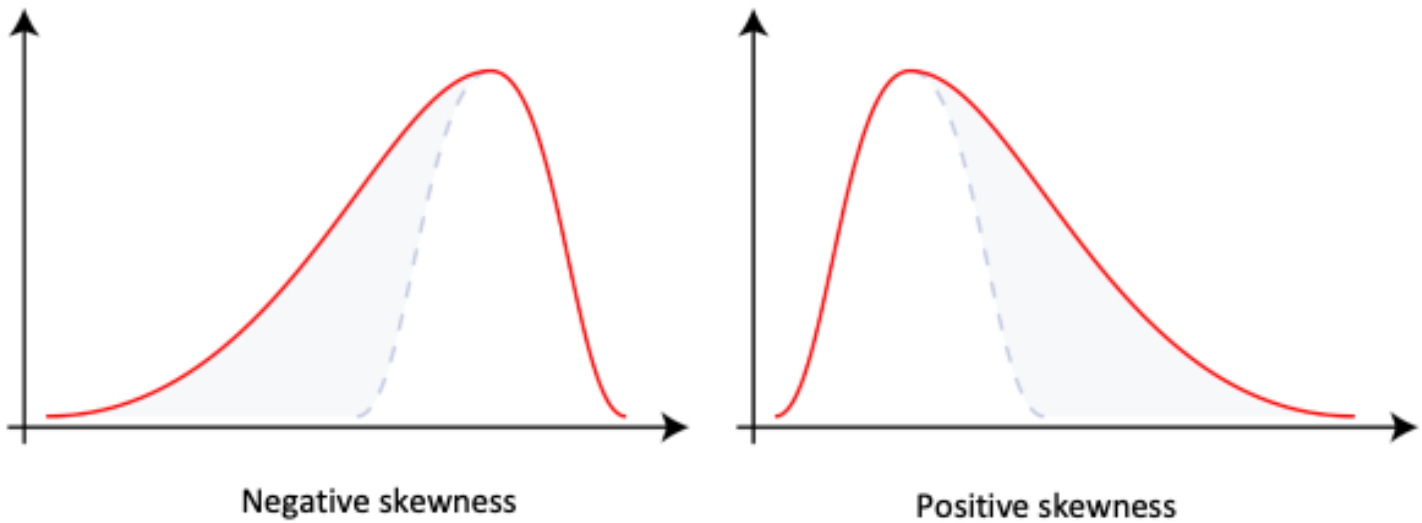
- GR: the mean provides the best measure of central tendency for an interval scale
- When we use the mean we account for both frequency of events in each category and their position as well as the values themselves
- By using more information, the mean is less likely than other measures of central tendency to be affected by changes in the nature of the sample
- The mean is generally preferred but when the data are skewed the **median** is preferred

Skewed Distributions

- A skewed distribution has scores that are weighted to one side and that frequencies of extreme values trail off in one direction away from the main cluster of cases
- A left-tailed distribution is negatively skewed
- A right-tailed distribution is positively skewed
- Examples

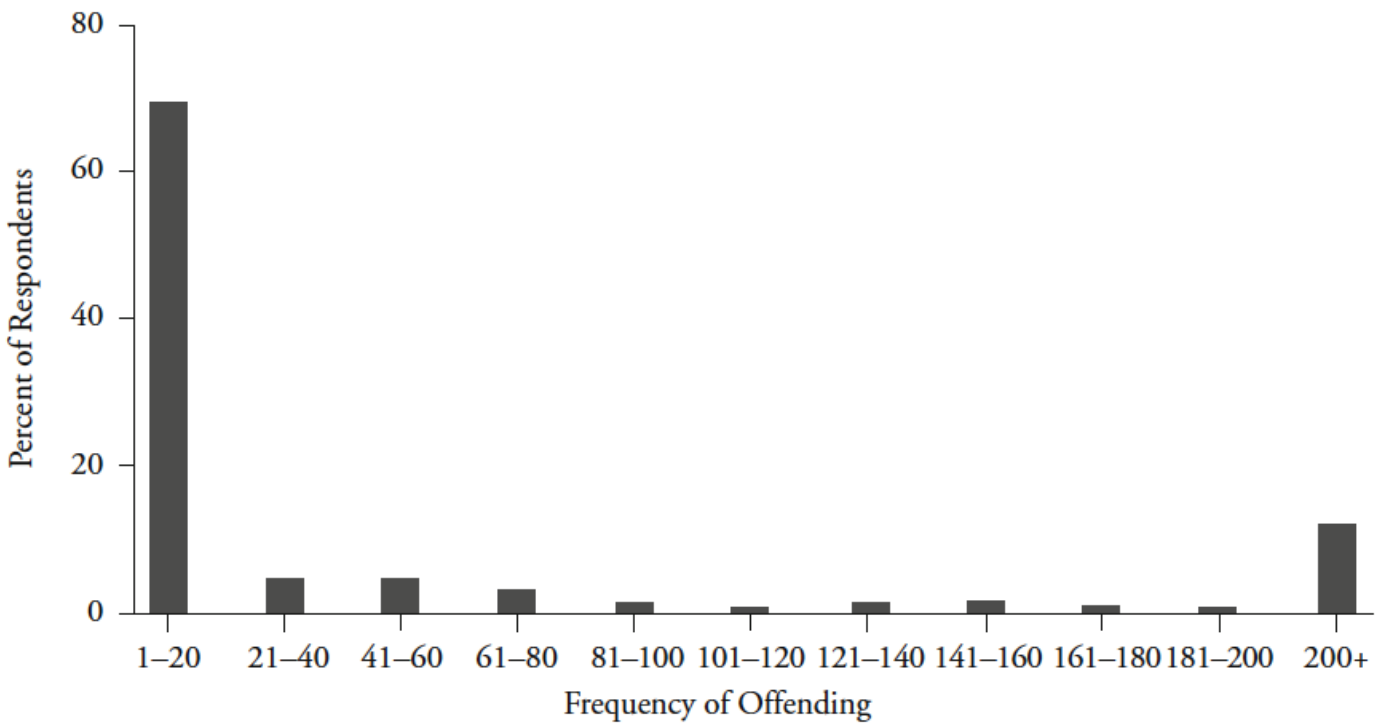
Figure

Skewness describes the shift of the distribution away from a normal distribution. Negative skewness shows that the mode moves to the right resulting in a dominant left tail. Positive skewness shows that the mode moves to the left resulting in a dominant right tail.



Example

A good example of skewed distributions are self-reports of behaviors (e.g., "offending") and income.



Handling Skewed Distributions

- Exclude outlier cases – but these are not really outliers
- Analyze outliers separately
- ** Use the median rather than the mean
- How should you decide when a distribution is so skewed that it is preferable to use the median as opposed to the mean?

Measures of Dispersion?

- Measures of Central Tendency provide a snapshot of the typical case
- The same statistic may be obtained from samples or populations that are very different
- Examples
  - Calculate the mean of these two distributions and compare: 3,3,3,3,3 and 1,2,3,4,5

Intro: Why are Measures of Dispersion Important?

- How typical is the typical case?
- Are most cases clustered closely around the average case?
- Is there a good deal of dispersion of cases both above and below the average

Measuring Dispersion in Interval Scales: The Range

- A common method of describing the spread of scores is to examine the **range** or difference between the highest and lowest scores
- Example: Calculate the mean and range for the Number of Calls in Hot Spots

Hot Spot Number	Number of Calls
1	2
2	9
3	11
4	13
5	20
6	20
7	20
8	24
9	27
10	29
11	31

Measuring Dispersion in Interval Scales: The Range

But, since the range bases its estimate of dispersion on just two observations, a change in one case can completely change the story.

- Example: Calculate the mean and range for the Number of Calls in Hot Spots

Hot Spot Number	Number of Calls
1	2
2	9
3	11
4	13
5	20
6	20
7	20
8	24
9	27
10	29
11	31

Other Methods of Dispersion

- One method for reducing the instability of the range is to examine cases that are not at the extremes of your distribution
- Examples
  - Examine the range between the 5th and 95th percentile or between the 20th and 80th percentiles
- Still relies on two scores

Interpreting and Calculating Percentiles

- When interpreting sets of important values, the raw number is not necessarily enough to relate the scores to one another
- Examples
  - You score a 156 on the LSAT exam. You have your score, great. How do you compare to others?
  - To get a full understanding of your score, relate it to others to determine if your score is higher or lower to the average for that test
- The percentile is a common method of comparison

Percentiles

- A percentile is a way of determining how a score compares to other scores in the same set
- Interpretation: The percentage of values that fall below a particular value in a set of scores
  - A male child age 12 with a weight of 130 pounds is at the 90th percentile of weight for males of that age, which indicates that he weighs more than 90 percent of other 12-year-old boys
- Percentiles can also be used to split a data set into portions to measure dispersion and identify central tendency

Useful Percentiles

- The Median: The 50th percentile
- Quartiles: Values that split the data into quarters based on percentiles
  - Q1 is the 25th percentile
  - Q2 is the 50th percentile also known as ...
  - Q3 is the 75th percentile
- The interquartile range, or IQR, used to measure dispersion and meant to show the middle half of the data. One quarter of the values is above this range and one quarter is below.
- The interquartile range is calculated as Q3 - Q1
  - Note the larger the IQR the greater the spread of the values

Calculating Percentiles

- Rank the data set from smallest to largest
- Calculate the index.
  - To find the *kth* percentile multiple $\frac{k}{100}$ by $N + 1$
  - k is the percentile you are trying to find, N is the number of cases or data points
  - $Rank = \frac{k}{100} \times (N + 1)$
- Round to the nearest whole number
- Count the values in your data set from smallest to largest until you reach the number you calculated, i.e. the Rank, as determined in the step immediately above.

Example: Calculating Percentiles

- Calculate the 25th percentile for the following numbers.

Score	Rank
43	
56	
53	
30	
33	
72	
68	
67	

Step 1: Rank the Numbers from 1 to N

- Calculate the 25th percentile for the following numbers.

Rank

Score	Rank
43	3
56	5
53	4
30	1
33	2
72	8
68	7
67	6

Re-arrange

Score	Rank
30	1
33	2
43	3
53	4
56	5
67	6
68	7
72	8

Working it out

Select Score

Score	Rank
30	1
33	2
43	3
53	4
56	5
67	6
68	7
72	8

Now simply use the formula and plug in the numbers

$$k = 25, N = 8$$

$$Rank = \frac{k}{100} \times (N + 1)$$

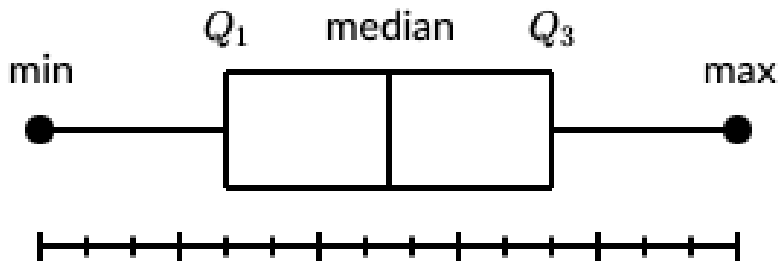
$$Rank = \frac{25}{100} \times (9)$$

$$Rank = 2.25 \Rightarrow 2$$

Let's calculate the range, 50th and 75th percentile ...

Do it yourself

The Box Plot



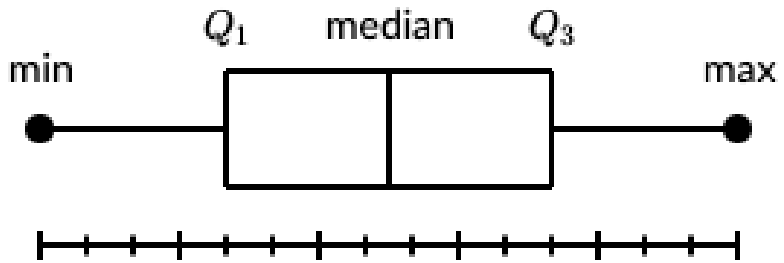
People's Ages in a Dataset

Age	Rank
25	1
28	2
29	3
29	4
30	5
34	6
35	7
35	8
37	9
38	10

Let's calculate the range, 50^{th} and 75^{th} percentile ... What percent of people were older than 29 years?

Try This

The Box Plot



Working it out

Other Methods of Dispersion

- Recall: we want to have an idea of dispersion, or spread, around the mean of a distribution that helps us understand how well the average represents the distribution
- Examples
  - Are scores clustered tightly around the mean or father away
- Still relies on two scores
- Question: Why not simply examine how much the average scores differ from the mean?

Other Methods of Dispersion

- Recall: When we add up the deviations above and below the mean, the positive and negative scores cancel each other out
- In order to use deviations from the mean as a basis for a measure of dispersion, we must have a way to account for the sign or direction of the statistic (whether the deviation is positive or negative)

Other Methods of Dispersion: The Variance

- The variance, denoted s^2 provides an estimate of dispersion

The formula for the variance is

$$s^2 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N-1}$$

Simple Example

Calculating the variance for these five numbers

- Example: Calculate the variance of the following scores

N	**Value**	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
1	3		
2	4		
3	5		
4	6		
5	7		
		Total $\sum = 0$	Total $\sum = 10$

Calculating Dispersion in Interval Scales

Calculating the variance for crime calls at hot spots in a year

- Example: Calculate the mean and range of the following scores

Hot Spot N	**No. of calls**	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
1	2	$2 - 21.5 = -19.5$	380.25
2	9	$9 - 21.5 = -12.5$	156.25
3	11	$11 - 21.5 = -10.5$	110.25
4	13	$13 - 21.5 = -8.5$	72.25
5	20	$20 - 21.5 = -1.5$	3.25
6	20	$20 - 21.5 = -1.5$	3.25
7	20	$20 - 21.5 = -1.5$	3.25
8	24	$24 - 21.5 = 2.5$	6.25
9	27	$27 - 21.5 = 5.5$	30.25
10	29	$29 - 21.5 = 7.5$	56.25
11	31	$31 - 21.5 = 9.5$	90.25
12	502	$52 - 21.5 = 30.5$	930.25
		Total $\sum = 0$	Total $\sum = 1,839.0$

Working it out

Now simply use the formula and plug in the numbers

$$s^2 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N-1}$$

$$= \frac{\sum_{i=1}^{N}(X_i - 21.5)}{211}$$

$$= \frac{1{,}839}{11} = 167.18$$

Example 2

Variance Calculations

Variance for Bail Amounts for a Sample of Persons Arrested for Felonies

DEFENDANT	BAIL AMOUNT	$(X_i - \overline{X})$	$(X_i - \overline{X})^2$
1	500	−2,763.33	7,635,992.69
2	1,000	−2,263.33	5,122,662.69
3	1,000	−2,263.33	5,122,662.69
4	1,000	−2,263.33	5,122,662.69
5	1,200	−2,063.33	4,257,330.69
6	1,500	−1,763.33	3,109,332.69
7	2,500	−763.33	582,672.69
8	2,500	−763.33	582,672.69
9	2,500	−763.33	582,672.69
10	2,750	−513.33	263,507.69
11	5,000	1,736.67	3,016,022.69
12	5,000	1,736.67	3,016,022.69
13	5,000	1,736.67	3,016,022.69
14	7,500	4,236.67	17,949,372.69
15	10,000	6,736.67	45,382,722.69
		Total (Σ) = 0.05	**Total (Σ) = 104,762,333.33**

Working it out

Plug in numbers

$$s^2 = \frac{\sum_{i=1}^{N}(X_i - 3{,}263.33)^2}{N-1}$$

$$= \frac{104{,}762{,}333.33}{14}$$

$$= 7,483,0235.81$$

The Standard Deviation

- Another measure of dispersion that is based on the variance is the standard deviation, denoted s
- Easily calculated as the square root of the variance
- This measure solves many problems most importantly the estimate is in units that are similar to the original scores

The formula for the standard deviation is

$$s = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N-1}}$$

Characteristics of Standard Deviations

- A standard deviation of zero means the data has no variability. When will this be the case?
- When cases are spread widely from the mean, there is more dispersion and the standard deviation will be larger. Also, the converse.
- The size of the standard deviation (and the variance) is dependent on both the amount of dispersion in the measure and the units of analysis that are used. When the data are in units that are large the standard deviation will reflect large units
- Extreme deviations from the mean have the greatest weight in constructing the standard deviation. Beware of outliers in the interpretation. In this case the outliers affect BOTH the calculation of the mean and the deviations that go into calculating the standard deviation.

Example

The standard deviation is a useful statistic for comparing the extent to which characteristics are clustered or dispersed around the mean in different samples.

- Example: Duncan SEI by Criminal Justice System contact
- A sample of individuals without institutional contact is compared to a sample of offenders w/institutional contact on a measure of social status, i.e., the Duncan socioeconomic index (SEI)

Category	**N**	$\bar{X}$	s
Contact	83	59.27	19.45
No Contact	112	61.05	11.13
Total $\sum$	= 195		

Coefficient of Relative Variation

- When the means of two groups are similar a direct comparison of standard deviations provides a good view of differences
- There are situations in which comparing standard deviations may not work
  - When the means of two groups are very different
  - When the data are measured in different units (e.g. age in years and income in dollars)
- The solution is to use the coefficient if relative variation (CRV) to explore the relative size of the standard deviation relative to its mean

Working it out

To calculate the CRV for antitrust offenders, simply plug in the numbers

$$CRV = \frac{s}{\bar{X}}$$

$$= \frac{11.13}{61.05}$$

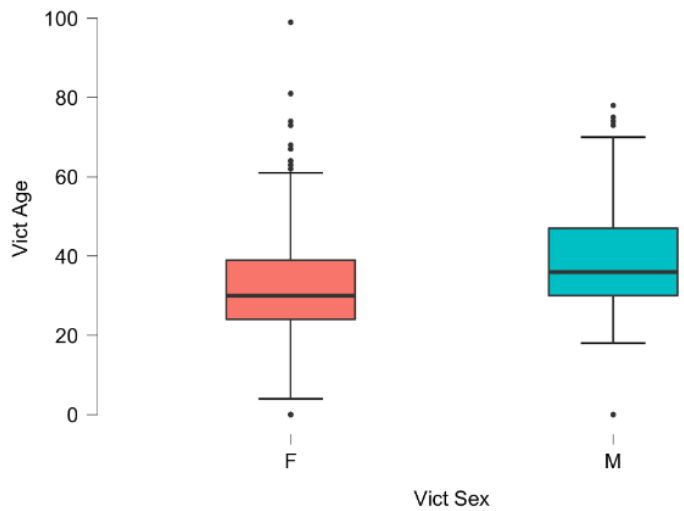
$$= 0.1823$$

Calculate the Standard Deviation and Variance in JASP

- Calculate the standard deviation and variance for the age of victims of IPV in LA

Calculating Quartiles and Making Boxplots in JASP

Box Plot of Victim Age By Gender



Working it out

- Open Jasp and import the IPV data from LA city
- Click on "Descriptives" and move "Vict Age" to the Variables Box and "Vict Sex" to the Split Box
- In the "Stats" tab check "Quartiles" and under "Dispersion" click "IQR" "Maximum" and "Minimum"
- In the "Plots" tab click "Boxplots" and "Boxplot Element" and "Use Color Palette" (!)

Summary

Concepts

1. Deviation from the mean
2. Least squares property
3. mean
4. median
5. mode
6. outliers
7. skewed (left and right)

Symbols

1. X
2. $\bar{X}$
3. N
4. $\sum$

Formulas

1. $Mdn = \frac{N+1}{2}$
2. $\bar{X} = \sum_{i=1}^{N} X_i$
3. $\sum_{i=1}^{N}(X_i - \bar{X}) = 0$
4. $\sum_{i=1}^{N}(X_i - \bar{X})^2 = minimum$

Summary, cont'd

Concepts

1. Coefficient of Relative Variation
2. Range
3. Standard Deviation (s)
4. Variance (s^2)

Symbols

1. s
2. s^2
3. $\sum$

Symbols

1. $s = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N-1}}$
2. $s^2 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N-1}$
3. $CRV = \frac{s}{\bar{X}}$