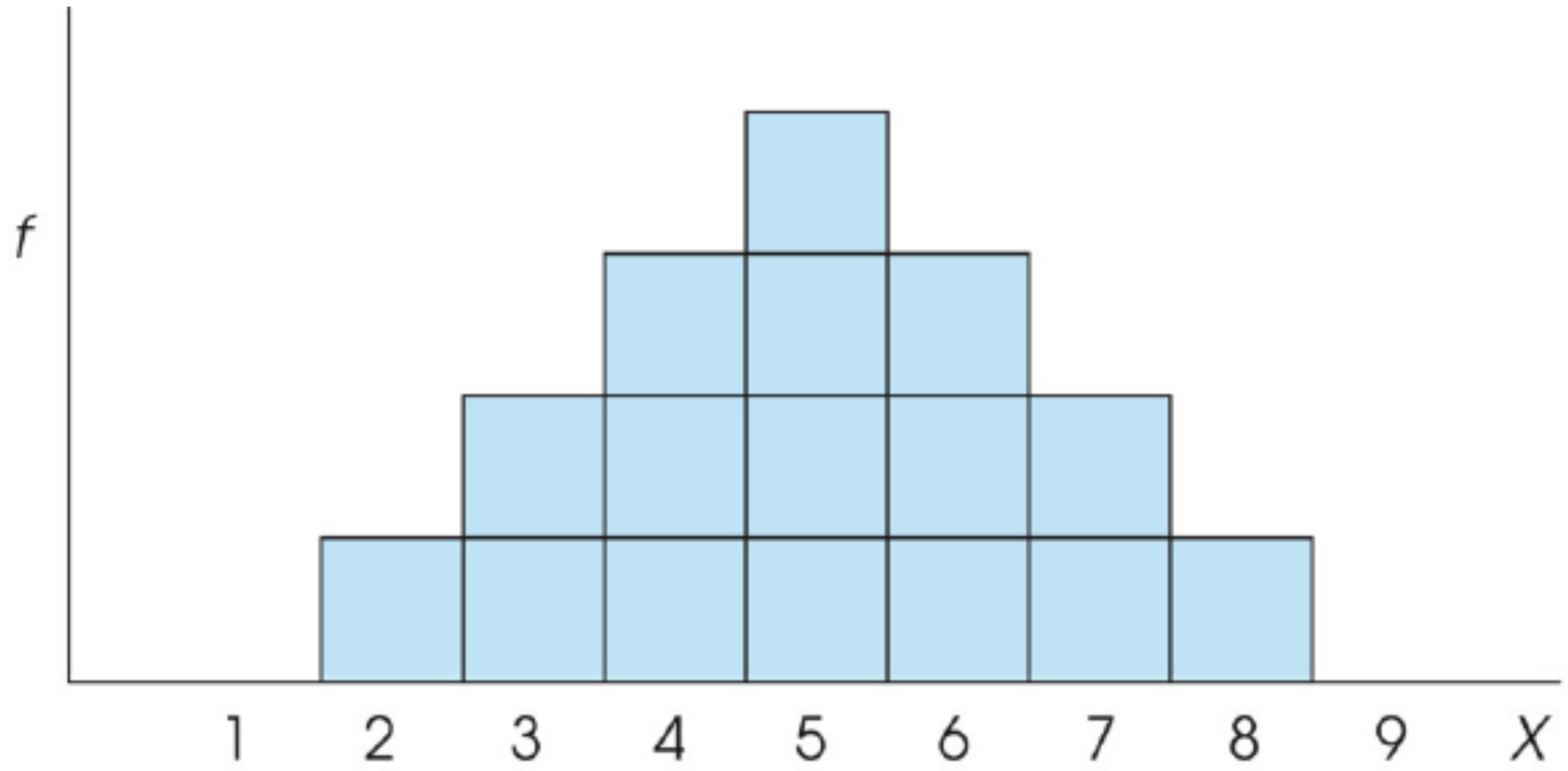
Descriptive Statistics — Quantitative

PSYC 203

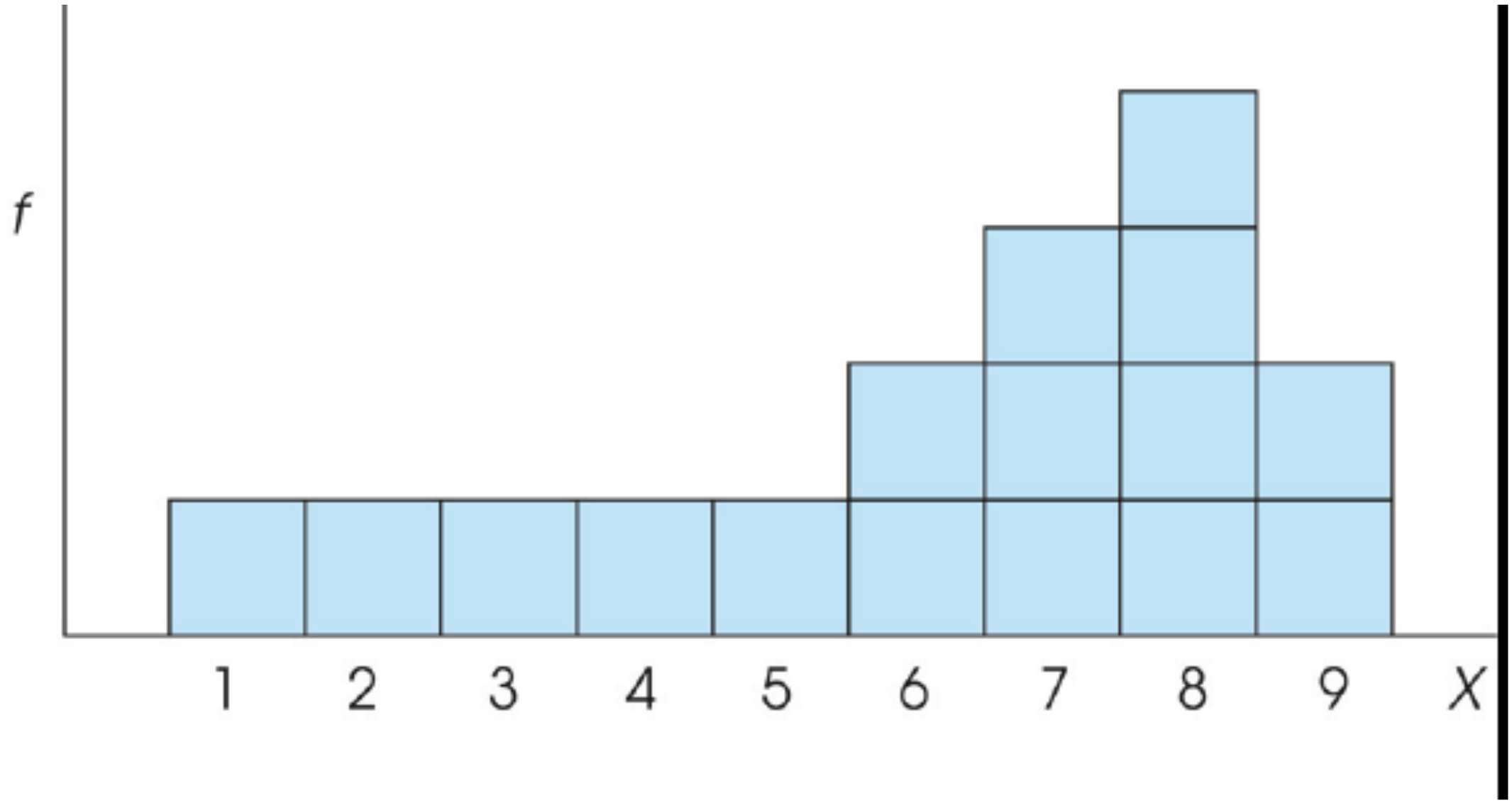
Agenda

- **Central Tendency**
  - **Mode, Median, and Mean**
  - **Sampling Stability**
- Dispersion
  - Range, Interquartile Range, Average Deviation, Variance, and Standard Deviation
  - Sampling Stability

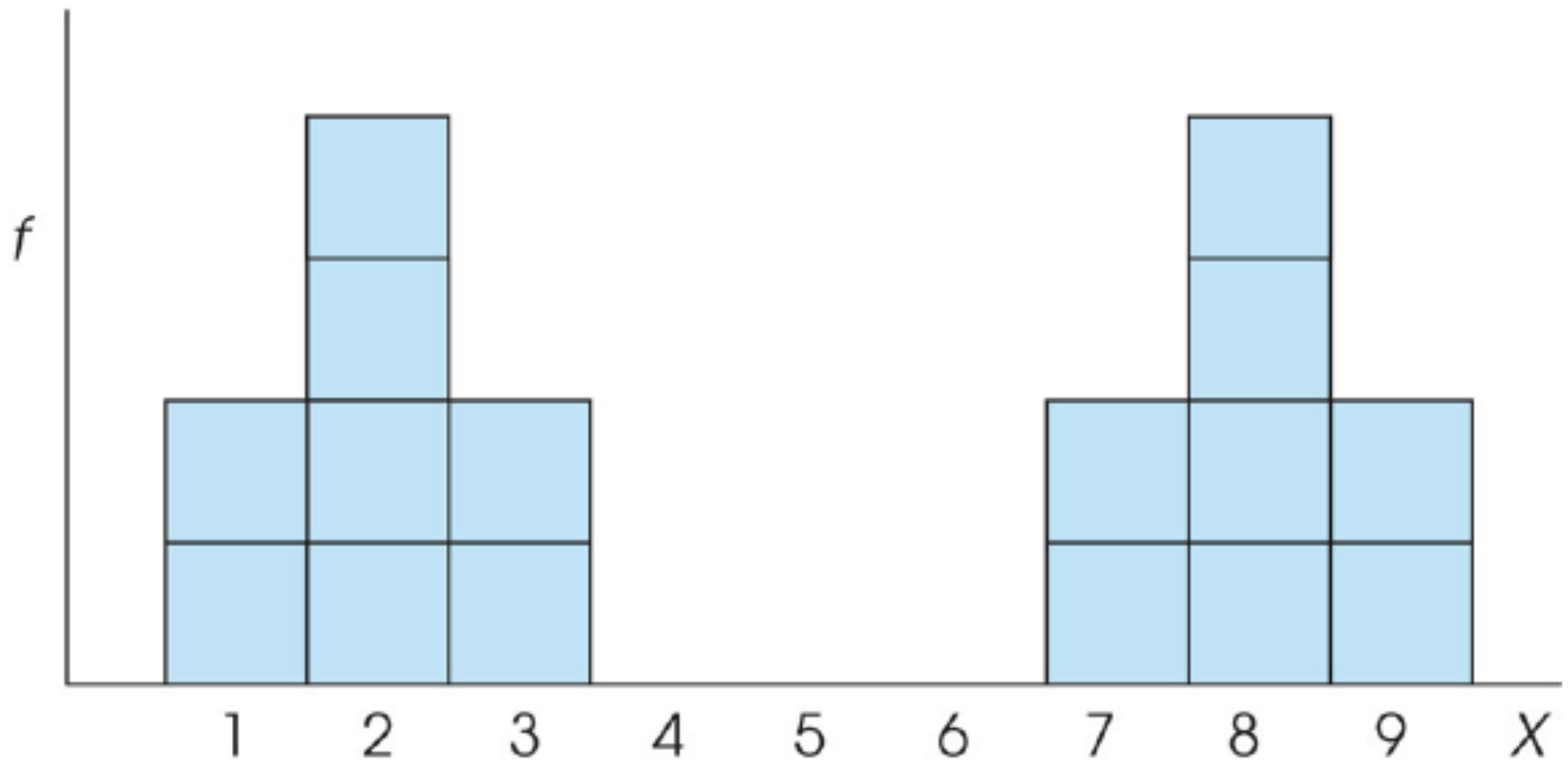
Where is the center?



Where is the center?



Where is the center?



Central Tendency

- Mode — most frequently occurring score(s)

Central Tendency

- In these data, the modes are
  - 41, 51

	Freq	%	% Cum.
26	1	1.00	1.00
30	1	1.00	2.00
31	1	1.00	3.00
32	1	1.00	4.00
34	1	1.00	5.00
36	3	3.00	8.00
38	3	3.00	11.00
39	4	4.00	15.00
40	3	3.00	18.00
41	7	7.00	25.00
42	5	5.00	30.00
43	3	3.00	33.00
44	1	1.00	34.00
45	4	4.00	38.00
47	1	1.00	39.00
48	1	1.00	40.00
49	4	4.00	44.00
50	6	6.00	50.00
51	7	7.00	57.00
52	6	6.00	63.00
53	3	3.00	66.00
54	3	3.00	69.00
55	1	1.00	70.00
56	2	2.00	72.00
58	3	3.00	75.00
59	2	2.00	77.00
60	3	3.00	80.00
61	1	1.00	81.00
62	6	6.00	87.00
63	3	3.00	90.00
65	1	1.00	91.00
66	2	2.00	93.00
67	3	3.00	96.00
72	1	1.00	97.00
73	1	1.00	98.00
80	1	1.00	99.00
88	1	1.00	100.00
Total	100	100.00	100.00

Central Tendency

- Mode — most frequently occurring score(s)
- Median — 50/50 point $\left(\frac{n+1}{2}\right)^{th}$

Central Tendency: Median

- There are 100 cases, so average of 50th and 51st case.

$$\left(\frac{100+1}{2}\right)^{th} = 50.5$$

	Freq	%	% Cum.
26	1	1.00	1.00
30	1	1.00	2.00
31	1	1.00	3.00
32	1	1.00	4.00
34	1	1.00	5.00
36	3	3.00	8.00
38	3	3.00	11.00
39	4	4.00	15.00
40	3	3.00	18.00
41	7	7.00	25.00
42	5	5.00	30.00
43	3	3.00	33.00
44	1	1.00	34.00
45	4	4.00	38.00
47	1	1.00	39.00
48	1	1.00	40.00
49	4	4.00	44.00
50	6	6.00	50.00
51	7	7.00	57.00
52	6	6.00	63.00
53	3	3.00	66.00
54	3	3.00	69.00
55	1	1.00	70.00
56	2	2.00	72.00
58	3	3.00	75.00
59	2	2.00	77.00
60	3	3.00	80.00
61	1	1.00	81.00
62	6	6.00	87.00
63	3	3.00	90.00
65	1	1.00	91.00
66	2	2.00	93.00
67	3	3.00	96.00
72	1	1.00	97.00
73	1	1.00	98.00
80	1	1.00	99.00
88	1	1.00	100.00
Total	100	100.00	100.00

Central Tendency

- Mode — most frequently occurring score(s)
- Median — 50/50 point
- Mean — the arithmetic average $\bar{X} = \frac{\sum X}{n}$

Central Tendency — Mean

- Numerator is (26 + 30 + 31 + 32 + 34…88)
  - *n* is 100

$$\bar{X} = \frac{\sum X}{n} = 50.54$$

	Freq	%	% Cum.
26	1	1.00	1.00
30	1	1.00	2.00
31	1	1.00	3.00
32	1	1.00	4.00
34	1	1.00	5.00
36	3	3.00	8.00
38	3	3.00	11.00
39	4	4.00	15.00
40	3	3.00	18.00
41	7	7.00	25.00
42	5	5.00	30.00
43	3	3.00	33.00
44	1	1.00	34.00
45	4	4.00	38.00
47	1	1.00	39.00
48	1	1.00	40.00
49	4	4.00	44.00
50	6	6.00	50.00
51	7	7.00	57.00
52	6	6.00	63.00
53	3	3.00	66.00
54	3	3.00	69.00
55	1	1.00	70.00
56	2	2.00	72.00
58	3	3.00	75.00
59	2	2.00	77.00
60	3	3.00	80.00
61	1	1.00	81.00
62	6	6.00	87.00
63	3	3.00	90.00
65	1	1.00	91.00
66	2	2.00	93.00
67	3	3.00	96.00
72	1	1.00	97.00
73	1	1.00	98.00
80	1	1.00	99.00
88	1	1.00	100.00
Total	100	100.00	100.00

Central Tendency

- **R code**
  - two different functions from two different packages

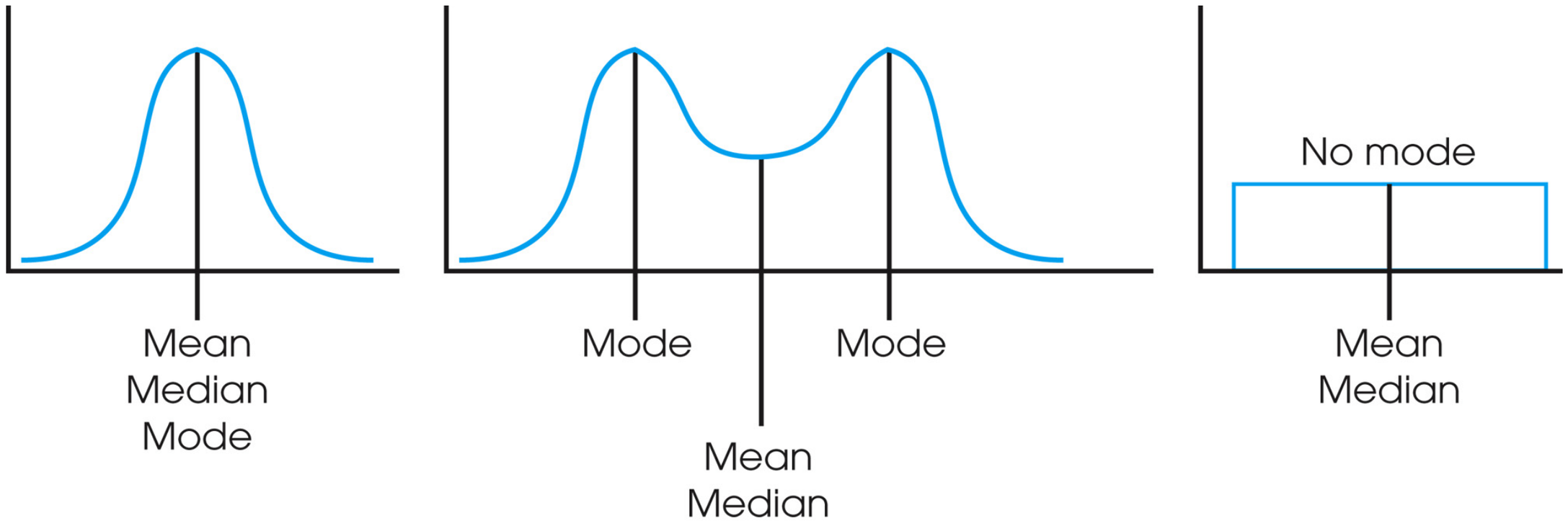
Symbols Matter

- What is the difference? $\bar{X}, \mu$
- Alternative expressions of the mean
  - include all data points $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
  - sum values of x starting at x4 and ending with x30 $\bar{X} = \frac{\sum_{i=4}^{30} X_i}{n}$

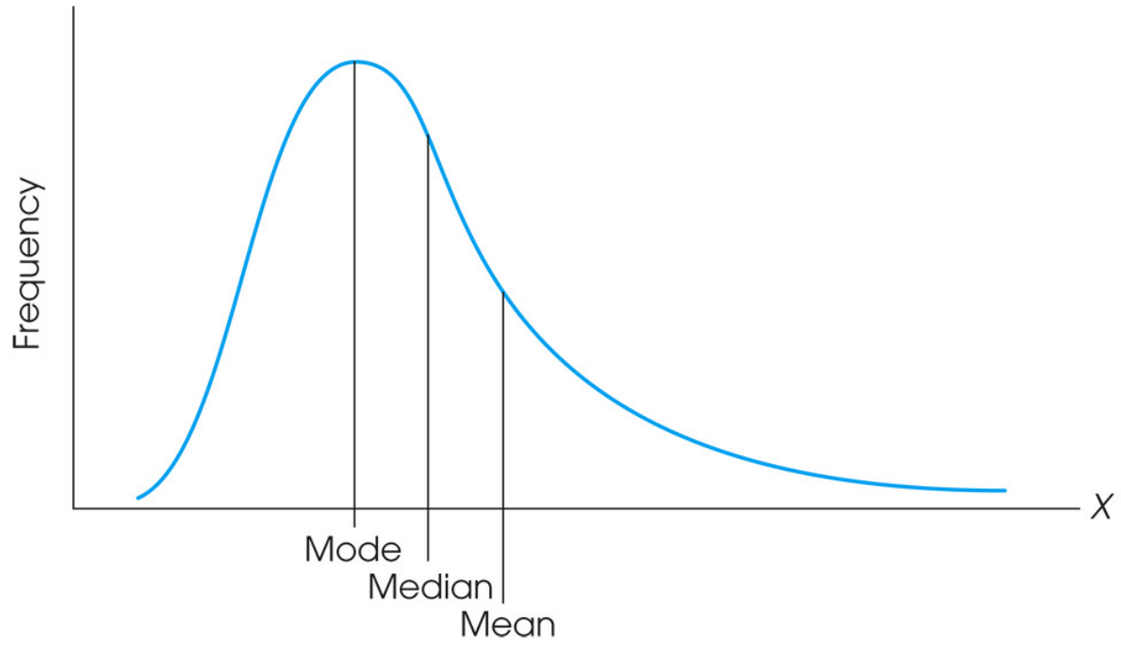
Which Measure to Use

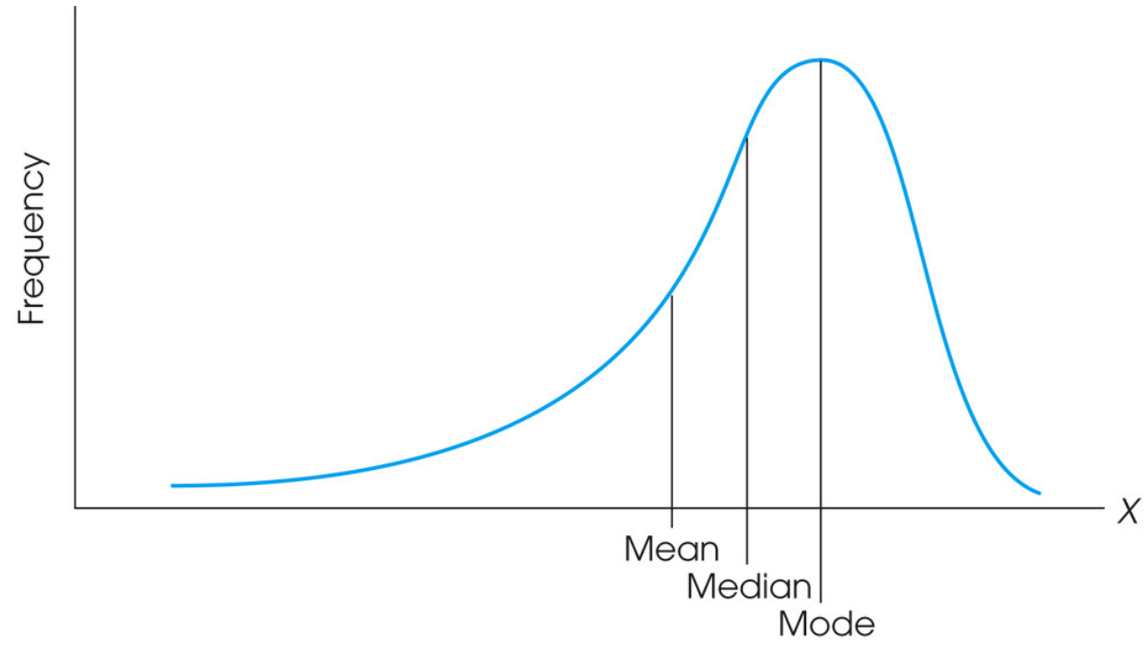
- **Mode**
  - Only appropriate CT measure for nominal data
  - Must be an observed score
  - Rather simple measure that can create interpretation problems (relatively unstable)
  - Not useful for any advanced analyses
- **Median**
  - Not influenced by extreme scores (both positive and negative)
  - Appropriate for ordinal data (no assumptions about units) or higher
  - Often useful in a descriptive sense, but (like the mode) is not useful for any advanced analyses
- **Mean**
  - Only appropriate with interval/ratio data
  - A distribution's "center of gravity"
  - Mathematically advantageous for operations
  - Better estimator of population parameter than mode or median
  - Heavily influenced by extreme scores

Distributions & CT



Distributions & CT





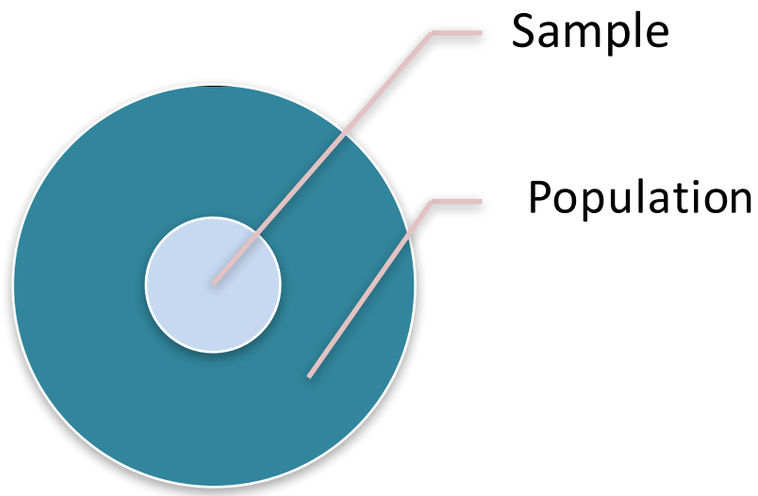
Measures of CT and Levels of Measurement

	Mode	Median	Mean
Nominal	YES	NO	NO
Ordinal	YES	YES	NO
Interval/Ratio	YES	YES	YES

Sampling Stability and Central Tendency

- Our sample **statistics** provide estimates of population **parameters**

- We have a population and draw a random sample of 5 cases.



- How close will the mean of the sample be to the population?

- Let's simulate

- How about if we draw a sample of 10, 20, or more?

Agenda

- Central Tendency
  - Mode, Median, and Mean
  - Sampling Stability
- **Dispersion**
  - **Range, Interquartile Range, Average Deviation, Variance, and Standard Deviation**
  - **Sampling Stability**

Variability / Dispersion

- Fundamentally important for our research
  - individual differences
- Do the mean, median, or mode tell us anything about individual differences?
  - NO!!!
- We want a measure that quantifies the extent to which cases vary.

Variability / Dispersion

- Range
  - Difference of lowest and highest score in a distribution
  - PROS: Easy to compute and understand
  - CONS: Often related to sample size and tends to be unstable
- Interquartile Range
  - Range between 25th and 75th percentile
  - PROS: More stable than Range
  - CONS: Not a particularly rich statistic in terms of inferences
- Average Deviation
  - ‘simple’ measure of dispersion

$$AD = \frac{\sum \left| X_i - \bar{X} \right|}{n}$$

Dispersion is

- Range
  - 88-26 = 62
- IQR
  - 58-41 = 8
- AD
  - 8.7

	Freq	%	% Cum.
26	1	1.00	1.00
30	1	1.00	2.00
31	1	1.00	3.00
32	1	1.00	4.00
34	1	1.00	5.00
36	3	3.00	8.00
38	3	3.00	11.00
39	4	4.00	15.00
40	3	3.00	18.00
41	7	7.00	25.00
42	5	5.00	30.00
43	3	3.00	33.00
44	1	1.00	34.00
45	4	4.00	38.00
47	1	1.00	39.00
48	1	1.00	40.00
49	4	4.00	44.00
50	6	6.00	50.00
51	7	7.00	57.00
52	6	6.00	63.00
53	3	3.00	66.00
54	3	3.00	69.00
55	1	1.00	70.00
56	2	2.00	72.00
58	3	3.00	75.00
59	2	2.00	77.00
60	3	3.00	80.00
61	1	1.00	81.00
62	6	6.00	87.00
63	3	3.00	90.00
65	1	1.00	91.00
66	2	2.00	93.00
67	3	3.00	96.00
72	1	1.00	97.00
73	1	1.00	98.00
80	1	1.00	99.00
88	1	1.00	100.00
Total	100	100.00	100.00

Variability/Dispersion: Variance

- Summarizes the magnitude of individual differences
  - population - $$\sigma^2 = \frac{\sum\left(X-\mu\right)^2}{N}$$
  - sample – $$s^2 = \frac{\sum\left(X-\bar{X}\right)^2}{n-1}$$

- Why two different formulas?
- Notation

Dispersion

- Variance (Population)
  - 121.88
- Variance (Sample)
  - 123.11

	Freq	%	% Cum.
26	1	1.00	1.00
30	1	1.00	2.00
31	1	1.00	3.00
32	1	1.00	4.00
34	1	1.00	5.00
36	3	3.00	8.00
38	3	3.00	11.00
39	4	4.00	15.00
40	3	3.00	18.00
41	7	7.00	25.00
42	5	5.00	30.00
43	3	3.00	33.00
44	1	1.00	34.00
45	4	4.00	38.00
47	1	1.00	39.00
48	1	1.00	40.00
49	4	4.00	44.00
50	6	6.00	50.00
51	7	7.00	57.00
52	6	6.00	63.00
53	3	3.00	66.00
54	3	3.00	69.00
55	1	1.00	70.00
56	2	2.00	72.00
58	3	3.00	75.00
59	2	2.00	77.00
60	3	3.00	80.00
61	1	1.00	81.00
62	6	6.00	87.00
63	3	3.00	90.00
65	1	1.00	91.00
66	2	2.00	93.00
67	3	3.00	96.00
72	1	1.00	97.00
73	1	1.00	98.00
80	1	1.00	99.00
88	1	1.00	100.00
Total	100	100.00	100.00

Simple Example

Test Score A
90
90
89
87
85
84
83
Mode = 90
Median = 87
Mean = 86.86
Sample Variance = 8.48
Population Variance = 7.27

Test Score B
90
90
89
87
85
84
20
Mode = 90
Median = 87
Mean = 77.86
Sample Variance = 656.48
Population Variance = 562.69

Variability/Dispersion: Standard Deviation

- statistic that indicates distance of any single score from the mean
  - population $$\sigma = \sqrt{\frac{\sum\left(X-\mu\right)^2}{N}}$$
  - sample $$s = \sqrt{\frac{\sum\left(X-\bar{X}\right)^2}{n-1}}$$
- square root of the variance

Dispersion

- Standard Deviation (Population)
  - 11.04
- Standard Deviation (Sample)
  - 11.09

	Freq	%	% Cum.
26	1	1.00	1.00
30	1	1.00	2.00
31	1	1.00	3.00
32	1	1.00	4.00
34	1	1.00	5.00
36	3	3.00	8.00
38	3	3.00	11.00
39	4	4.00	15.00
40	3	3.00	18.00
41	7	7.00	25.00
42	5	5.00	30.00
43	3	3.00	33.00
44	1	1.00	34.00
45	4	4.00	38.00
47	1	1.00	39.00
48	1	1.00	40.00
49	4	4.00	44.00
50	6	6.00	50.00
51	7	7.00	57.00
52	6	6.00	63.00
53	3	3.00	66.00
54	3	3.00	69.00
55	1	1.00	70.00
56	2	2.00	72.00
58	3	3.00	75.00
59	2	2.00	77.00
60	3	3.00	80.00
61	1	1.00	81.00
62	6	6.00	87.00
63	3	3.00	90.00
65	1	1.00	91.00
66	2	2.00	93.00
67	3	3.00	96.00
72	1	1.00	97.00
73	1	1.00	98.00
80	1	1.00	99.00
88	1	1.00	100.00
Total	100	100.00	100.00

Example continued

Test Score A
90
90
89
87
85
84
83
Sample SD = 2.91
Population SD = 2.70

Test Score B
90
90
89
87
85
84
20
Sample SD = 25.62
Population SD = 23.72

What Does Variance Mean?

- **Case A**
    - 10-item T/F test
    - Mel answers 10 correct and Chris answers 0 correct
    - Mean = 5, Variance = 25, SD = 5
- **Case B**
    - 10-item T/F test
    - Mel answers 6 correct and Chris answers 4 correct
    - Mean = 5, Variance = 1, SD = 1
- **Case C**
    - 10-item T/F test
    - Mel answers 6 correct and Chris answers 4 correct (each item is worth 10 points)
    - Mean = 50, Variance = 100, SD = 10

Comparing Variability

- Coefficient of Variation — comparison across contexts
- Cases on previous slides

$$CV_A = \left(\frac{s}{\bar{X}}\right)100 = \left(\frac{5}{5}\right)100 = 100$$

$$CV_B = \left(\frac{s}{\bar{X}}\right)100 = \left(\frac{1}{5}\right)100 = 20$$

$$CV_C = \left(\frac{s}{\bar{X}}\right)100 = \left(\frac{10}{50}\right)100 = 20$$

Sampling Stability and Variability

- We again have a population and draw a random sample of 5 cases.
    - How close will the variance of the sample be to the population?
    - Let's simulate
    - How about if we draw a sample of 10, 20, or more?

Symbols Matter Here Too..

TABLE 4-2. Variance and Standard Deviation in Symbols

The variance or standard deviation of a sample is an example of a statistic, whereas the variance or standard deviation of a population is an example of a parameter. The symbols we use depend on whether we are referring to the spread of a sample or a population.

Number	Used for . . .	Standard Deviation Symbol	Pronounced	Variance Symbol	Pronounced
Statistic	Sample	*SD* or *s*	As written	SD^2, s^2, or *MS*	Letters as written; if superscript 2, then followed by "squared" (e.g., "ess squared")
Parameter	Population	σ	"Sigma"	σ^2	"Sigma squared"

Statistics as Estimators of Parameters

- Sufficient
  - extent to which a statistic uses available information
- Unbiased
  - extent to which statistic's expected value is equal to the parameter (not systematically too large or small)
- Efficient
  - speaks to the variability of a statistic's value relative to the population parameter
- Resistant
  - extent to which a statistic is influenced by outliers