

SOC 3811/5811:
BASIC SOCIAL STATISTICS

Numeric Representations of Distributions

Graphs versus Numbers

Using either graphs or numbers, we primarily want to describe the central tendency and spread of a distribution

Central tendency

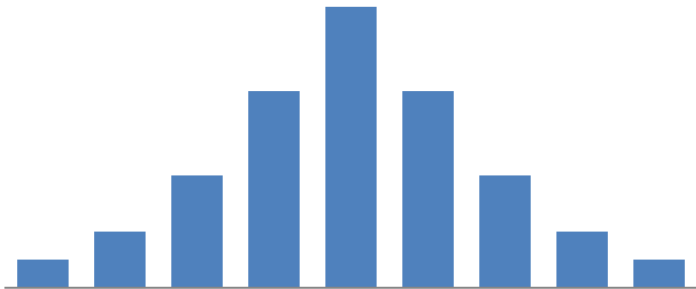
Where do the bulk of cases fall on the distribution?

Where is the middle of the distribution? (Is there one?)

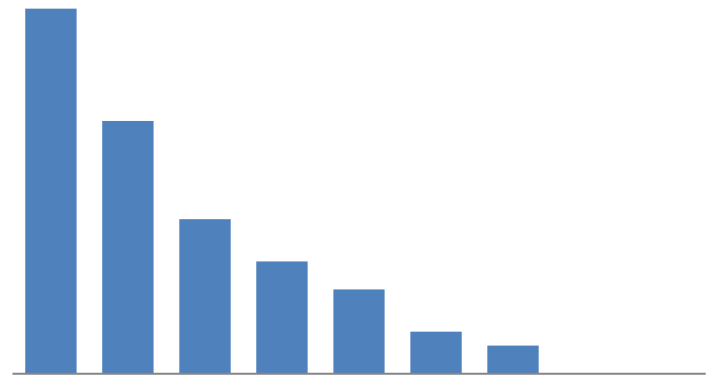
Spread (or Variation or Dispersion)

How tightly clumped together are cases, especially relative to the center of the distribution?

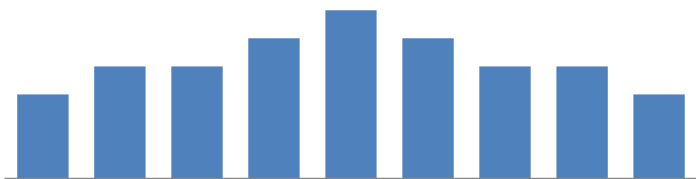
Graphs versus Numbers



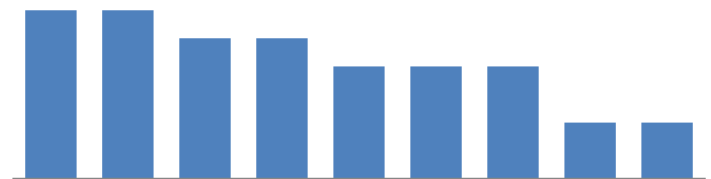
A



B

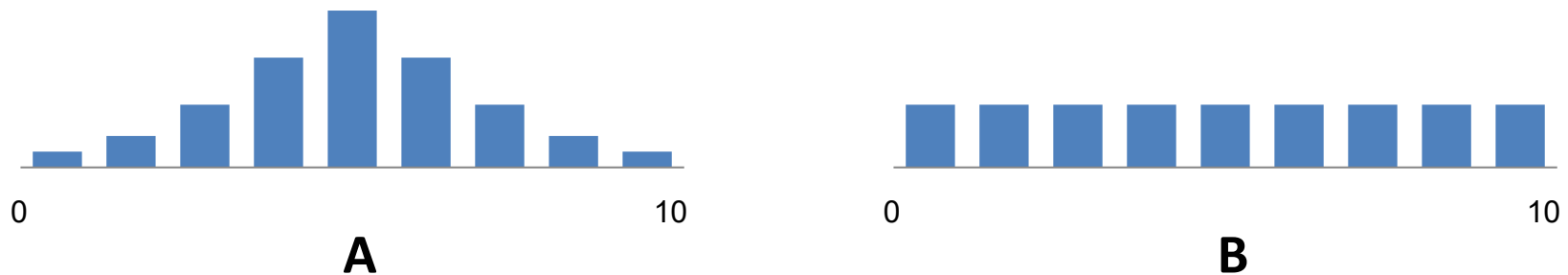


C



D

Worksheet



Which is True About Distributions A and B?

- a. Same center, same spread
- b. Same center, A has more spread
- c. Same center, B has more spread
- d. Different centers, same spread

Graphs versus Numbers

Unit of Observation: Presidents of the United States

Variable: Age when they took office

President	Age at Inauguration
George Washington	57
John Adams	61
...	
George W. Bush	54
Barack Obama	47
Donald Trump	70

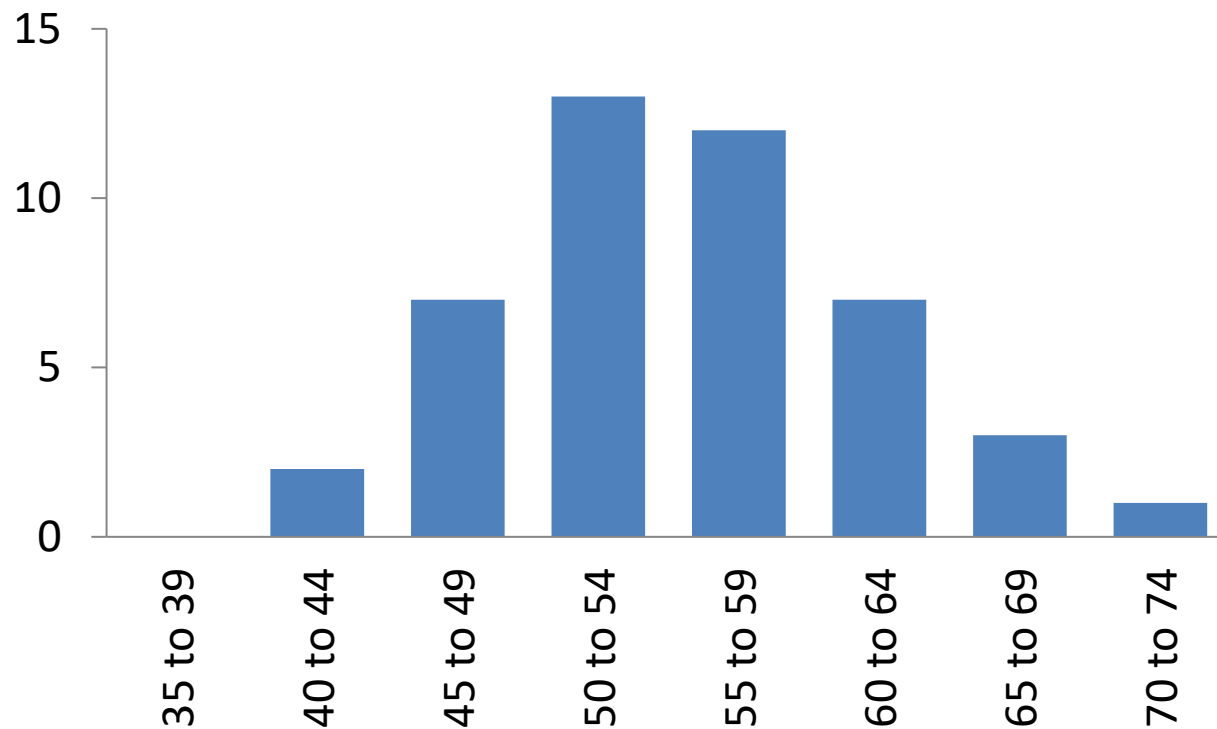


Grover Cleveland

Graphs versus Numbers

Always start by looking at a picture of the distribution

Ages of Presidents When They Took Office



Graphs versus Numbers

Many of the numeric summaries of central tendency and spread pertain only to *continuous* variables

Central Tendency	Continuous Variables	Discrete Variables	Spread	Continuous Variables	Discrete Variables
Mean	✓	✗	Range	✓	✗
Median	✓	✗	Interquartile Range	✓	✗
Mode	✓	✓	Variance	✓	✗
			Standard Deviation	✓	✗
			Index of Diversity	✗	✓
			Index of Qualitative Variation	✗	✓

Central Tendency I: Mean

The **mean** (or **average**) is computed by adding together (or summing) the value of the continuous variable across all individuals and then dividing by the number of individuals

Mathematically:

$$\bar{Y} = \frac{Y_1 + Y_2 + Y_3 + \dots + Y_N}{N} = \sum_{i=1}^N \frac{Y_i}{N}$$

where Y_i represents the value of the variable, Y , for individual i , “ Y -bar” is the mean, and N is the number of individuals in the data

$Y = 1 \ 3 \ 4 \ 8$

$$\bar{Y} = \frac{Y_1 + Y_2 + Y_3 + \dots + Y_N}{N} = \sum_{i=1}^N \frac{Y_i}{N}$$

Central Tendency I: Mean

Example: The **mean** age of U.S. Presidents when they took office equals:

$$\bar{Y} = \frac{57 + 61 + \dots + 54 + 47 + 70}{N} = \frac{2,475}{45} = 55.0$$

Central Tendency II: Median

Median

The point on the distribution at which exactly half of all individuals fall above that point and half fall below that point

The point on the distribution at which the cumulative percentage reaches 50%

To compute the median, or M , arrange the observations in order of size from lowest to highest

If the number of observations, N , is odd, then the median, M , is the middle observation

If the number of observations, N , is even, then the median, M , is the mean of the middle two observations

Central Tendency II: Median

Example: Ages of U.S. Presidents when they took office:

42 43 46 46 47 47 48 49 49 50 51 51 51 51 51 52 52
54 54 54 54 54 55 55 55 55 56 56 56 57 57 57 57 58
60 61 61 61 62 64 64 65 68 69 70

Because the number of observations (45) is an odd number,
the median is the 23rd case

So, $M = 55$

Central Tendency II: Median

Odd number of observation...

3 4 7

...median is 4

Even number of observations...

3 4 7 9

...median is $(4+7)/2$, or 5.5

Central Tendency III: Mode

Mode

The single value (or values) that appears most often in the distribution

A distribution with two modes is referred to as “bimodal”

For the ages of U.S. Presidents when they took office, the values 51 and 54 appear more frequently (5 times each) than any other value

42 43 46 46 47 47 48 49 49 50 51 51 51 51 51 52 52 54 54 54
54 54 55 55 55 55 56 56 56 57 57 57 57 58 60 61 61 61 62 64
64 65 68 69 70

Central Tendency: Outliers

Outliers ... or extreme values ... can strongly influence the mean, but not the mode or median

4 4 6 8

Median: 5

Mode: 4

Mean: 5.5

4 4 6 80

Median: 5

Mode: 4

Mean: 23.5

4 4 6 800

Median: 5

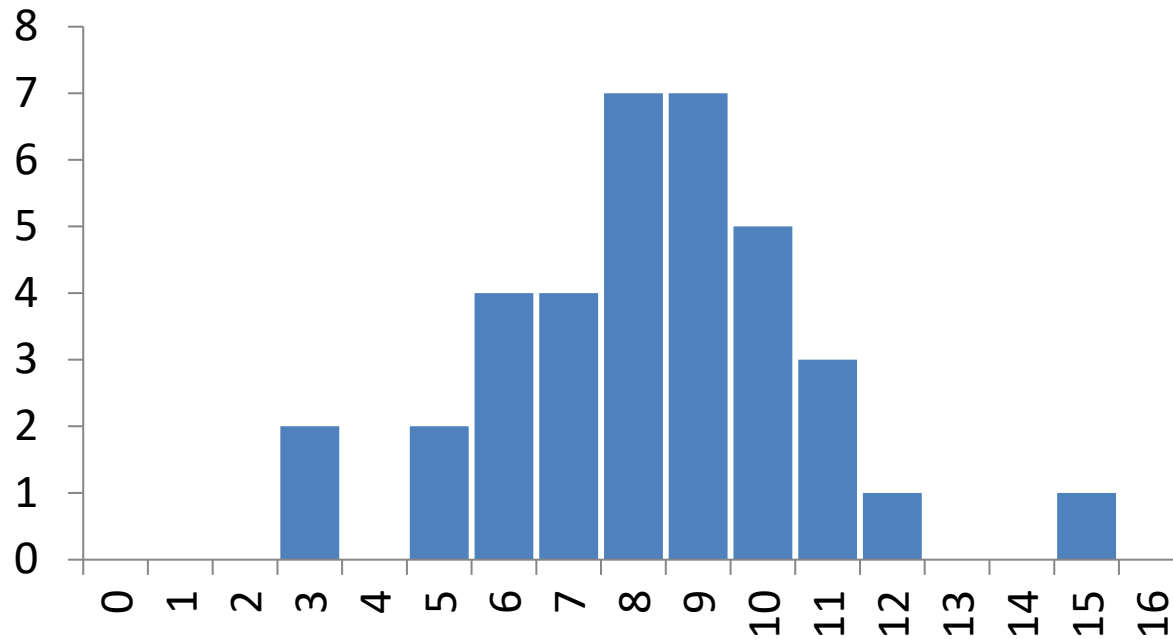
Mode: 4

Mean: 203.5

Measures of Central Tendency

Wins by the Minnesota Vikings

1978-2014 (except 1982)

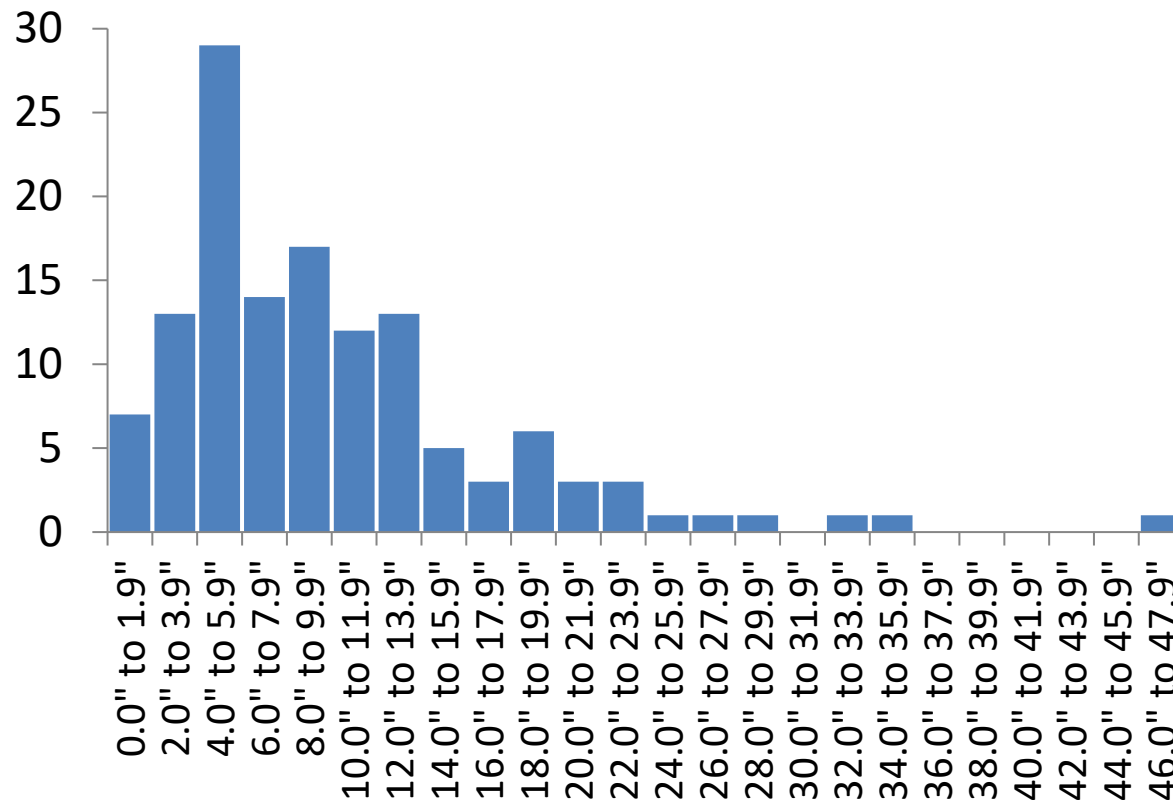


Mean: **8.3** Median: **8** Mode: **8 & 9**

Measures of Central Tendency

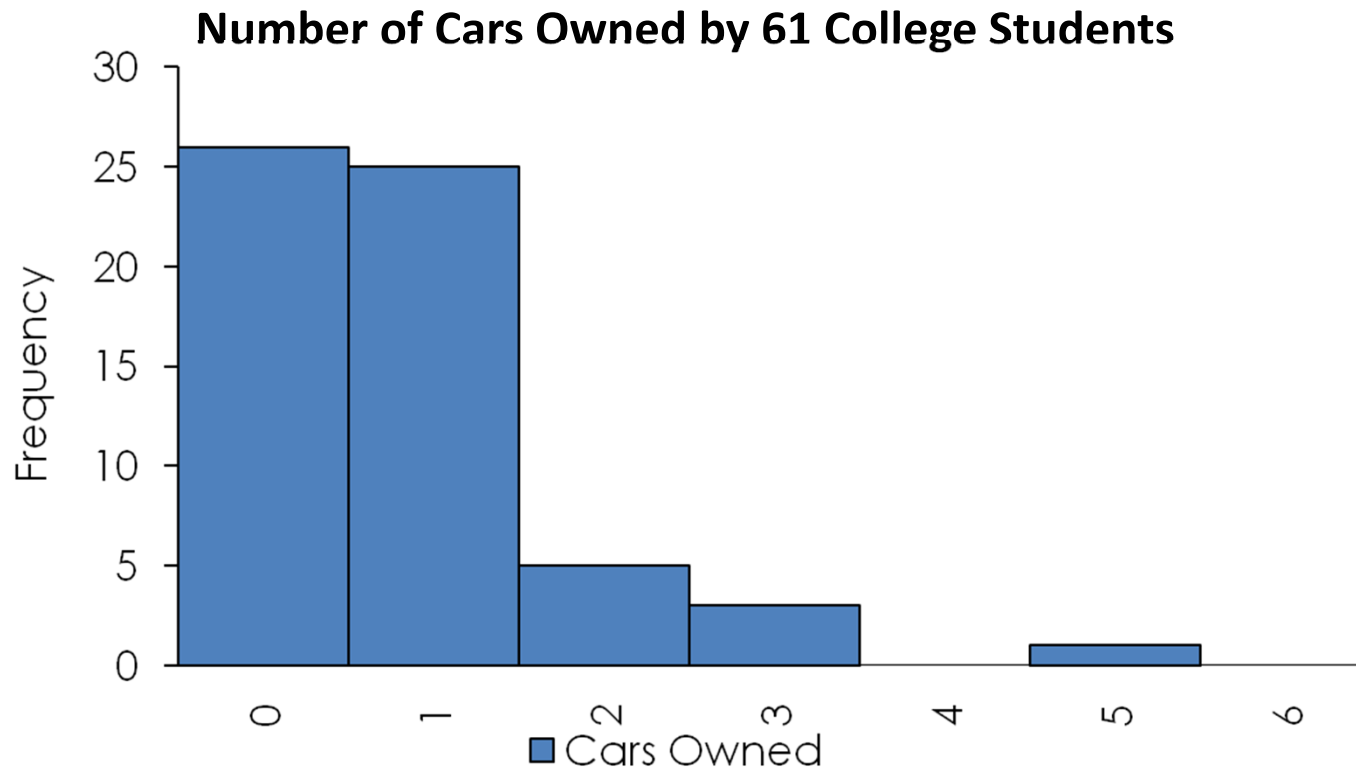
Inches of Snow in Minneapolis in January

1885-2014



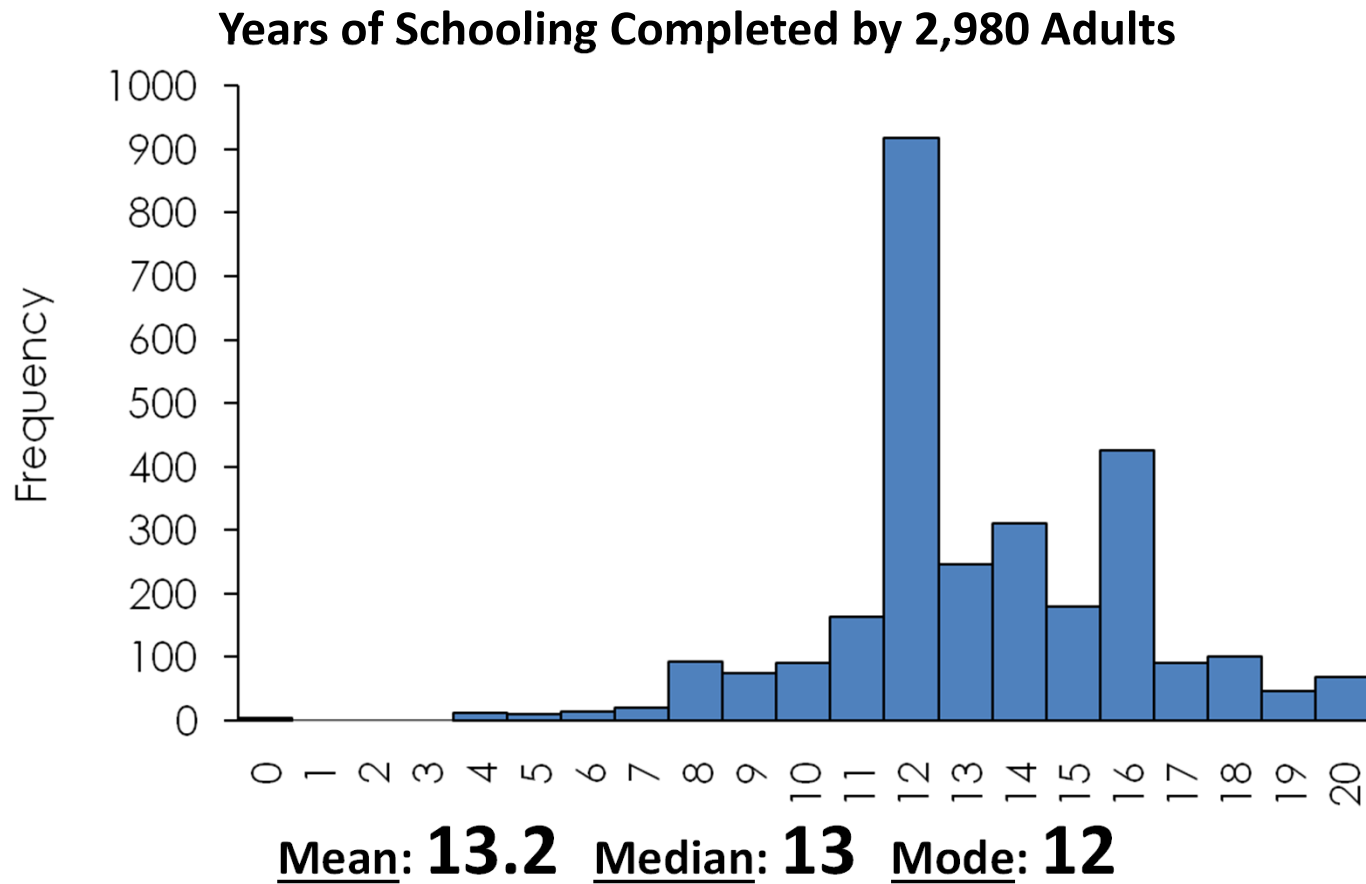
Mean: **10.0"** Median: **8.3"** Mode: **4.6"**

Measures of Central Tendency



Mean: 0.75 Median: 1 Mode: 0

Measures of Central Tendency



Worksheet

I surveyed 20 graduate students and asked them how many pets they own. Eight of them had no pets at all. Six had just one pet, five had two pets, and one had four

1. What is the mean of this distribution?
2. What is the median?
3. What is the mode?

Measures of Spread I: Range

Range

The difference between the largest and the smallest observation

Example: Ages of U.S. Presidents when they took office:

42 43 46 46 47 47 48 49 49 50 51 51 51 51 51 52 52 54
54 54 54 54 55 55 55 55 56 56 56 57 57 57 57 58 60 61
61 61 62 64 64 65 68 69 70

In this example, the range is the difference between 42 and 70 ... so, $70 - 42 = 28$

Measures of Spread II: Inter-Quartile Range

The **inter-quartile range** is the range in the distribution that marks out the middle 50% of the distribution

25% of the distribution falls below the first quartile

25% of the distribution falls above the third quartile

To calculate the inter-quartile range:

1. Sort the value from lowest to highest and find the median, M
2. The 1st quartile (Q_1) is the median of the values to the left of M
3. The 3rd quartile (Q_3) is the median of the values to the right of M

The inter-quartile range (or IQR) equals Q_3 minus Q_1

Measures of Spread II:

Inter-Quartile Range

- *Example: Ages of U.S. Presidents when they took office:*

42 43 46 46 47 47 48 49 49 50 51 51 51 51 51 52 52 54
54 54 54 54 55 55 55 55 56 56 56 57 57 57 57 58 60 61
61 61 62 64 64 65 68 69 70

- The median, M , is 55
- The 1st quartile, Q_1 , is the median of the 22 values below the median ... so, $(51+51)/2 = 51$
- The 3rd quartile, Q_3 , is the median of the 22 values above the median ... so, $(58+60)/2 = 59.0$
- The inter-quartile range is $59.0 - 51.0 = 8.0$

Measures of Spread III: Variance & Standard Deviation

Variance (s^2)

Basically, the average of the squared differences between each observation and the mean

$$s_Y^2 = \frac{(Y_1 - \bar{Y})^2 + (Y_2 - \bar{Y})^2 + \dots + (Y_N - \bar{Y})^2}{N - 1} = \frac{\sum_{i=1}^N (Y_i - \bar{Y})^2}{N - 1}$$

Standard Deviation (s)

The square root of the variance

$$s_Y = \sqrt{s_Y^2} = \sqrt{\frac{\sum_{i=1}^N (Y_i - \bar{Y})^2}{N - 1}}$$

$$Y = 1 \ 3 \ 4 \ 8$$

$$s_Y^2 = \frac{(Y_1 - \bar{Y})^2 + (Y_2 - \bar{Y})^2 + \dots + (Y_N - \bar{Y})^2}{N - 1} = \frac{\sum_{i=1}^N (Y_i - \bar{Y})^2}{N - 1}$$

$$s_Y = \sqrt{s_Y^2} = \sqrt{\frac{\sum_{i=1}^N (Y_i - \bar{Y})^2}{N - 1}}$$

Measures of Spread III: Variance & Standard Deviation

Example: 4 4 6 8

Y	$Y - \bar{Y}$	$(Y - \bar{Y})^2$
4	$(4 - 5.5) = -1.500$	$(4 - 5.5)^2 = 2.250$
4	$(4 - 5.5) = -1.500$	$(4 - 5.5)^2 = 2.250$
6	$(6 - 5.5) = 0.500$	$(6 - 5.5)^2 = 0.250$
8	$(8 - 5.5) = 2.500$	$(8 - 5.5)^2 = 6.250$
		$\sum(Y - \bar{Y})^2 = 11.000$

so s^2 equals $11/(N-1) = 11/3 = 3.667$

Measures of Spread III: Variance & Standard Deviation

Example: Variance (s^2) of the age of U.S. Presidents

$$s_Y^2 = \frac{\sum_{i=1}^{44} (Y_i - 54.659)^2}{44 - 1} = \frac{1,683.886}{43} = 39.160$$

Example: Standard deviation (s) of age of U.S. Presidents

$$s_Y = \sqrt{\frac{\sum_{i=1}^N (Y_i - \bar{Y})^2}{N - 1}} = \sqrt{39.160} = 6.258$$

Measures of Spread III: Variance & Standard Deviation

Variance has no inherent unit of measurement; standard deviations are in the same unit of measurement as the original observations

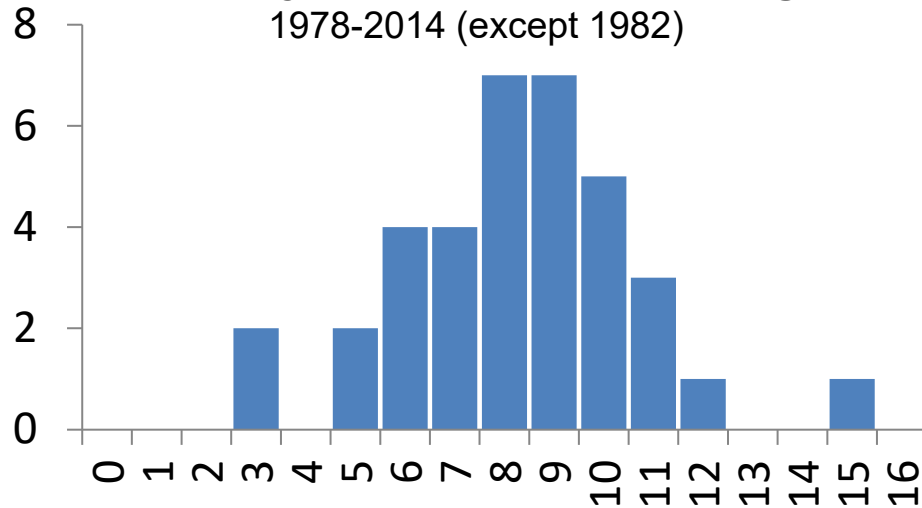
Variance and standard deviation measure spread around the mean, not spread around the median (or the mode)

Because means are strongly affected by outliers, and because s and s^2 are computed using the mean, s and s^2 are also strongly affected by outliers

Measures of Spread

Wins by the Minnesota Vikings

1978-2014 (except 1982)



Mean = 8.3

Range = 12 (3 to 15)

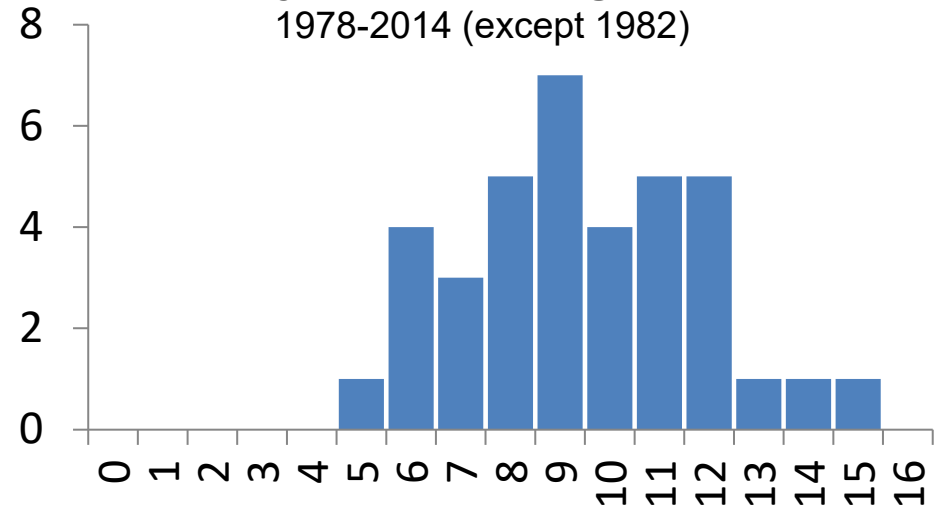
IQR = 3 (7 to 10)

$s^2 = 5.85$

$s = 2.42$

Wins by the Pittsburgh Steelers

1978-2014 (except 1982)



Mean = 9.6

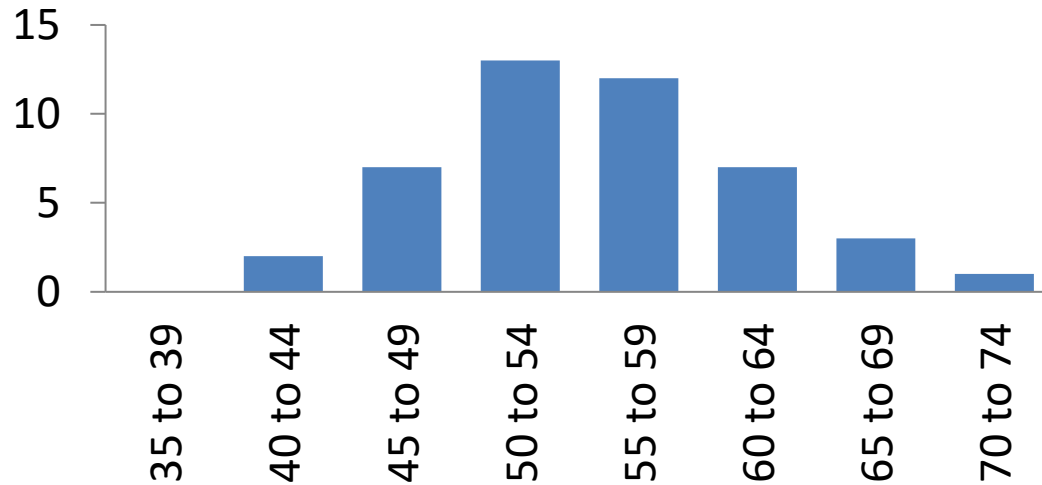
Range = 10 (5 to 15)

IQR = 3 (8 to 11)

$s^2 = 5.57$

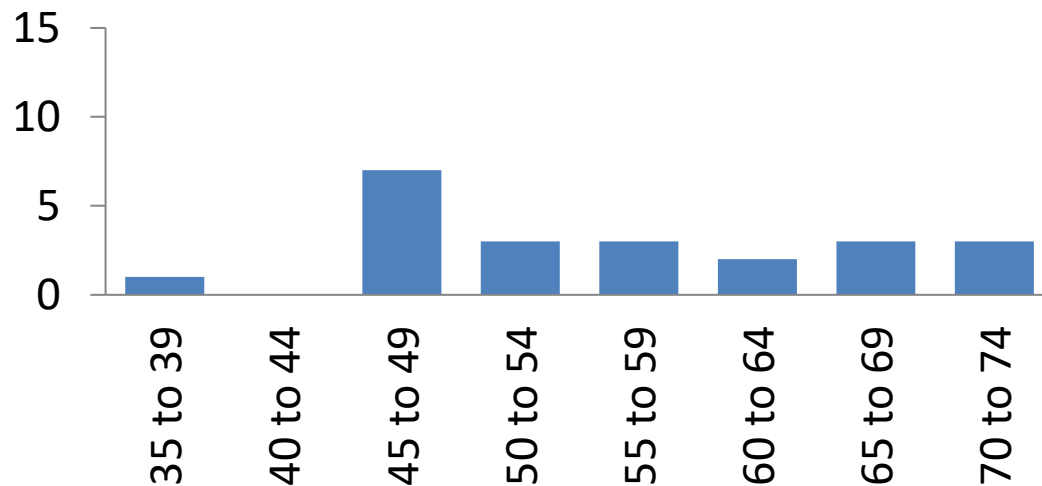
$s = 2.36$

Ages of **U.S. Presidents** When They Took Office



Mean: 55.0
Range: 28 (42 to 70)
IQR: 8 (51 to 59)
 s^2 : 42.5
s: 6.5

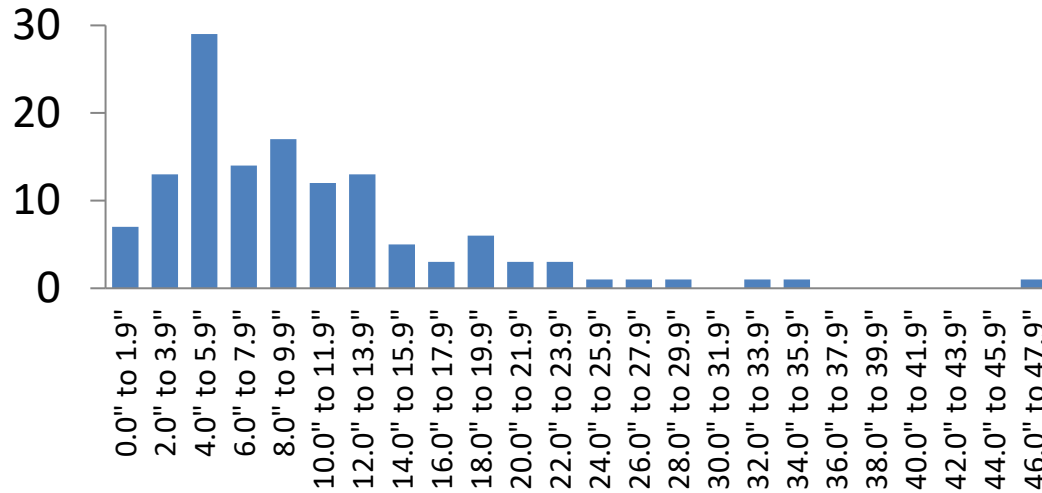
Ages of **Canadian Prime Ministers** When They Took Office



Mean: 55.6
Range: 35 (39 to 74)
IQR: 18 (47 to 65)
 s^2 : 95.6
s: 9.8

Snow in Minneapolis in **January**

1885-2014



Mean: 10.0

Range: 45.8 (0.6 to 46.4)

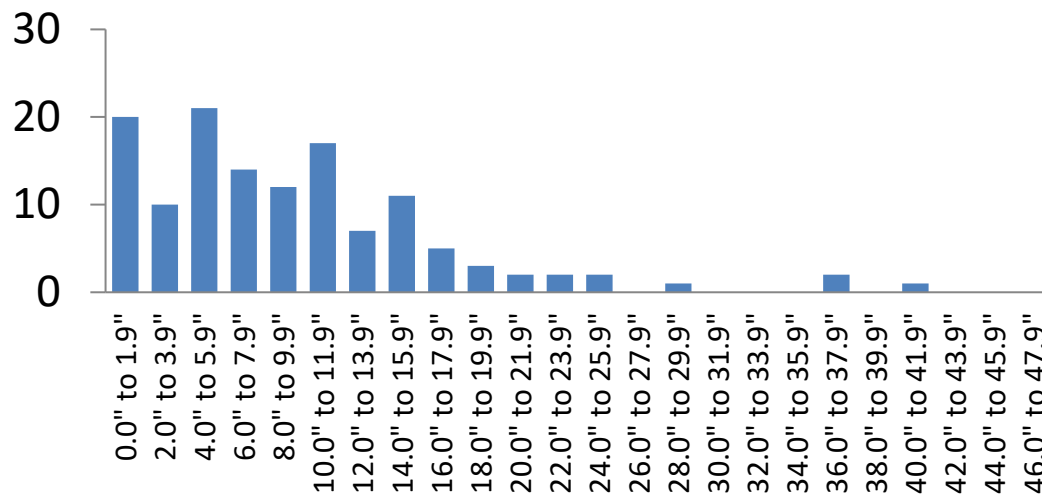
IQR: 7.9 (4.9 to 12.8)

s^2 : 55.4

s : 7.4

Snow in Minneapolis in **March**

1885-2014



Mean: 9.3

Range: 40 (0 to 40)

IQR: 9.3 (4.1 to 13.4)

s^2 : 57.0

s : 7.6

Worksheet

I surveyed 5 people and asked them how many brothers and sisters they have. The five people had zero, zero, two, three, and five siblings, respectively

1. What is the range of this distribution?
2. What is the variance?
3. What is the standard deviation?

Measuring Skew

For highly skewed distributions the mean may not accurately reflect central tendency; in these situations, many people prefer to use the median to summarize central tendency

How do you know when there is a “lot” of skew?

$$Skewness = \frac{3(\bar{Y} - Median)}{S_Y}$$

(Note: This is just one measure of skew. Lots of others exist.)

Measuring Skew

4 4 6 8

Median: 5

Mean: 5.5 s_y : 1.91

Skew: 0.78

4 4 6 800

Median: 5

Mean: 203.5 s_y : 397.67

Skew: 1.50

4 4 6 -800

Median: 4

Mean: -196.5 s_y : 402.33

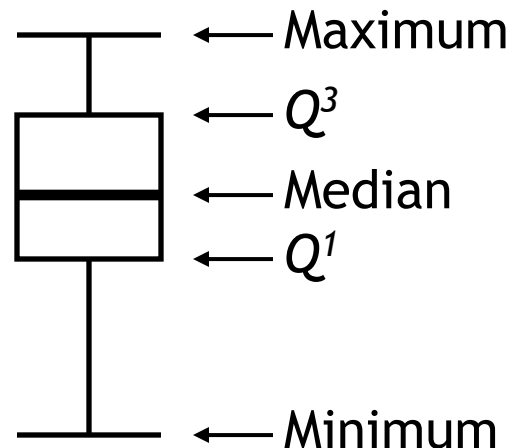
Skew: -1.50

Summarizing Distributions I: Five Number Summary and Box Plot

One convenient way to describe the central tendency and spread of a distribution is to use five numbers to make a “box plot”

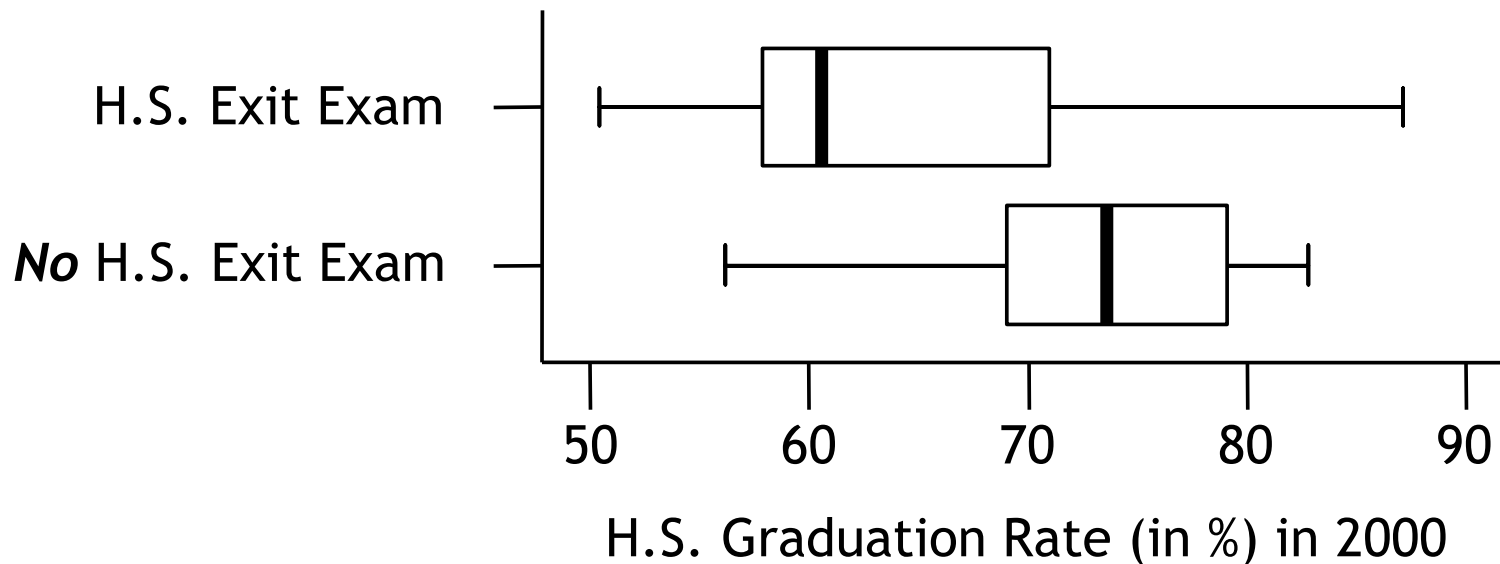
These numbers are the minimum value, Q_1 , M , Q_3 , and the maximum value

A box plot:



Summarizing Distributions I: Five Number Summary and Box Plot

Are state high school graduation rates related to whether states require students to pass “exit examinations” as a requirement for graduation? (All data as of 2000)



Summarizing Distributions I: Five Number Summary and Box Plot

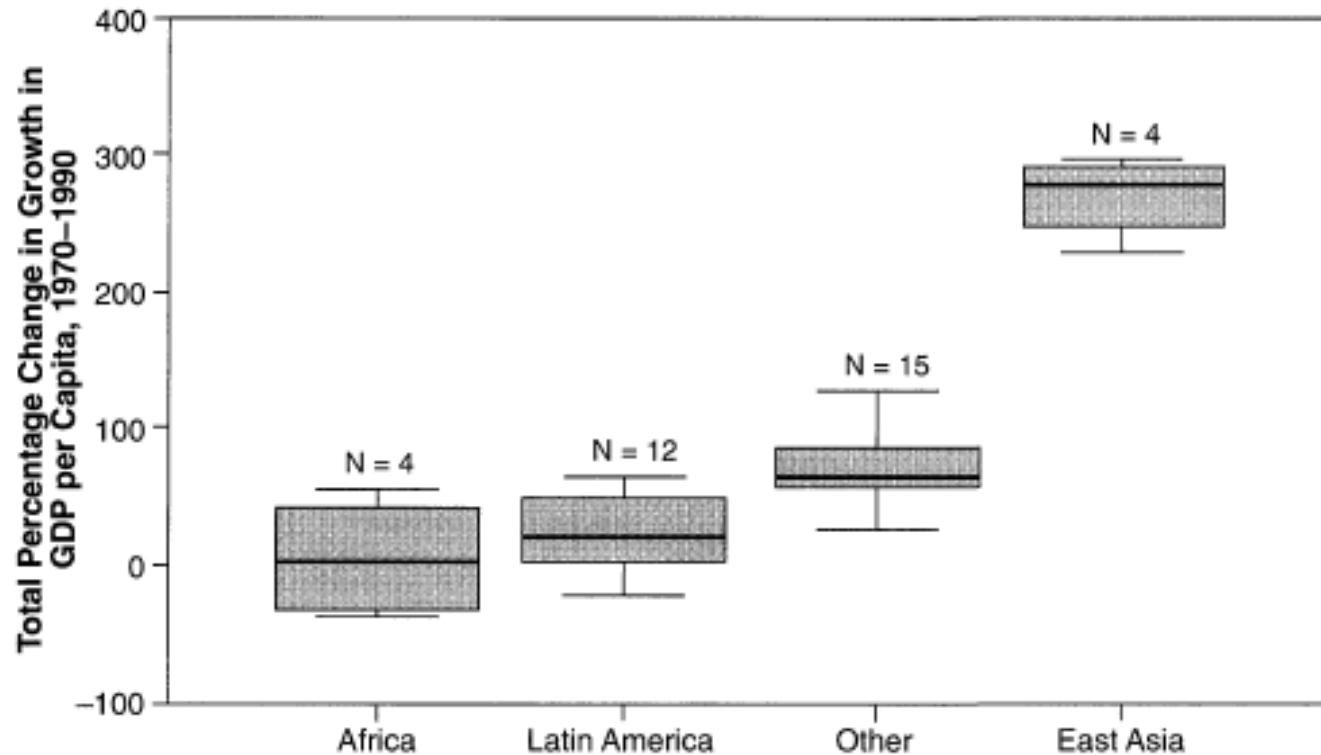
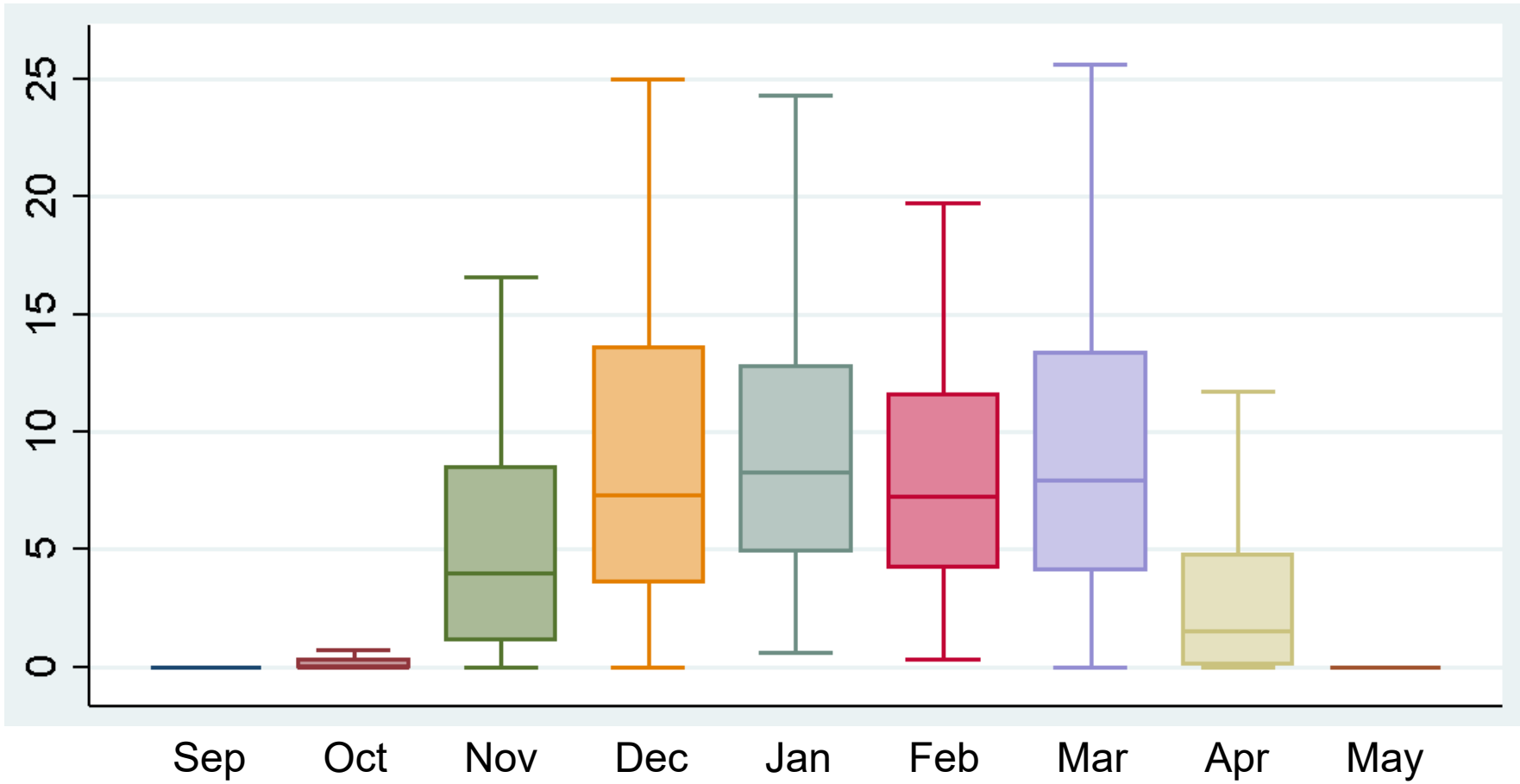


Figure 2. Boxplot of Growth in GDP per Capita by Region, 1970 to 1990

Source: Evans, Peter Evans and James E. Rauch. 1999. "Bureaucracy and Growth: A Cross-National Analysis of the Effects of 'Weberian' State Structures on Economic Growth Bureaucracy and Growth: A Cross-National Analysis of the Effects of "Weberian" State Structures on Economic Growth." *American Sociological Review* 64: 748-765.

Snow in Minneapolis, by Month

1885-2014



Summarizing Distributions II: Mean and Standard Deviation

By far the most common way to summarize the central tendency and spread in a distribution is to use the **mean** and **standard deviation**

However, this way of describing a distribution only works well if the distribution is fairly symmetrical and if there are no outliers (or if outliers have been removed) ... this is because the mean and standard deviation are heavily influenced by outliers and skew

Summarizing Distributions

Many of the numeric summaries of central tendency and spread pertain only to *continuous* variables

	Continuous Variables	Discrete Variables		Continuous Variables	Discrete Variables
Central Tendency			Spread		
Mean	✓	✗	Range	✓	✗
Median	✓	✗	Interquartile Range	✓	✗
Mode	✓	✓	Variance	✓	✗
			Standard Deviation	✓	✗
			Index of Diversity	✗	✓
			Index of Qualitative Variation	✗	✓

Summarizing Distributions

You interview 10 fellow students and ask them:

What is your home state?

1 = Minnesota

2 = Wisconsin

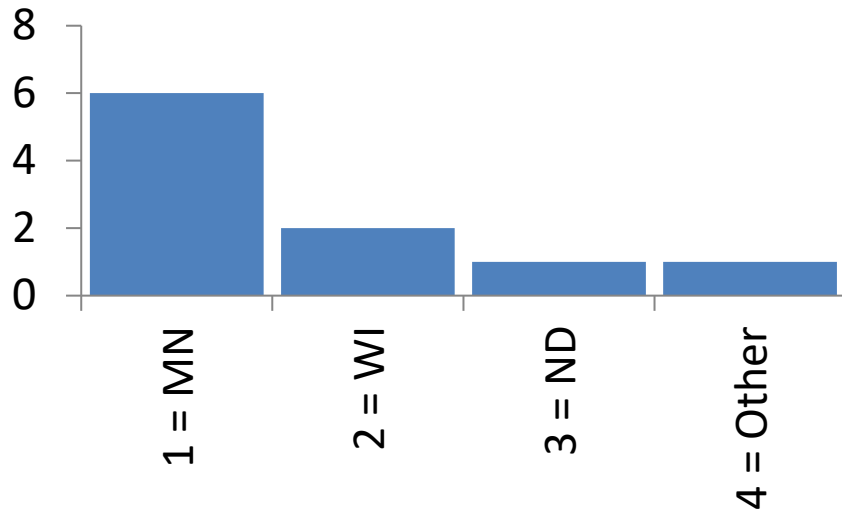
3 = North Dakota

4 = Someplace else

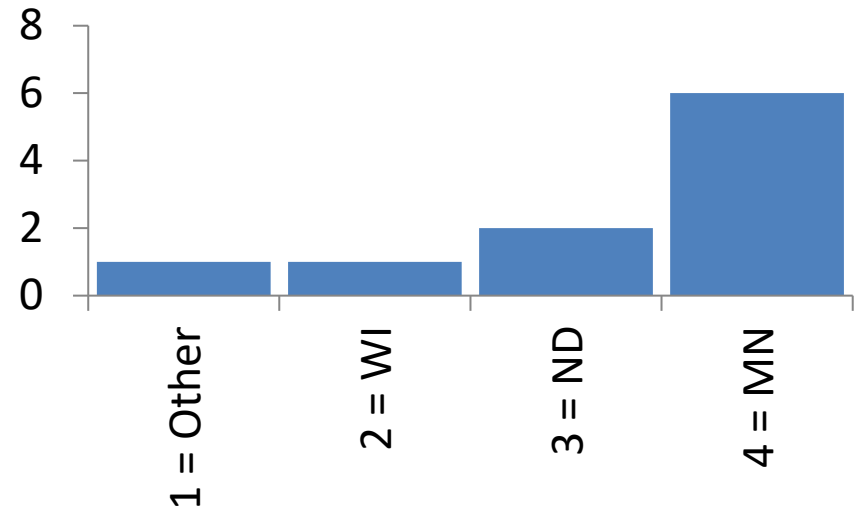
Here are the responses you obtained:

1 1 1 1 1 1 2 2 3 4

Summarizing Distributions



Mean: 1.7
Range: 3 (1 to 4)
IQR: 1 (1 to 2)
 s^2 : 1.122
s: 1.059



Mean: 3.2
Range: 3 (1 to 4)
IQR: 2 (2 to 4)
 s^2 : 1.289
s: 1.135

Summarizing Distributions

You interview 10 fellow students and ask them:

How happy are you?

1 = Very happy

2 = Somewhat happy

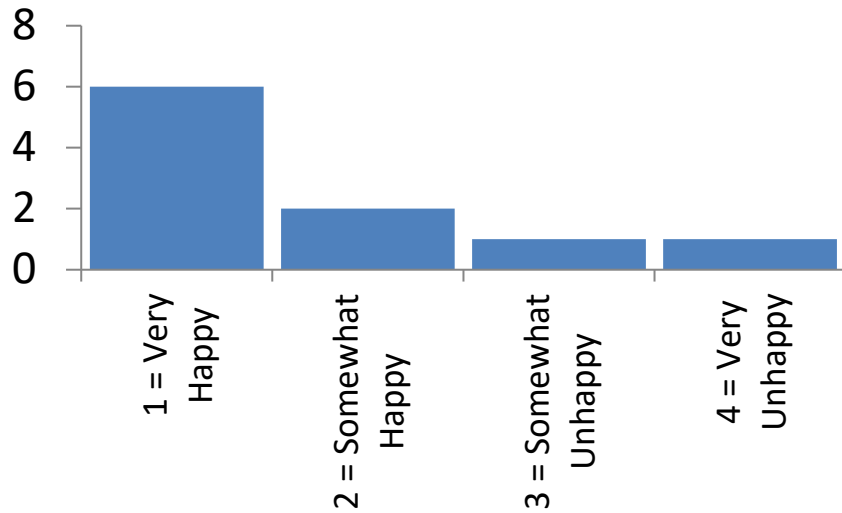
3 = Somewhat unhappy

4 = Very unhappy

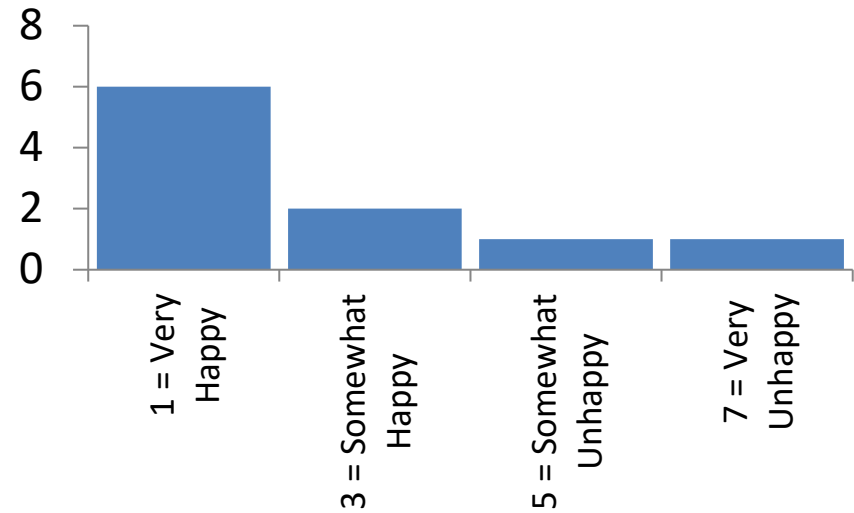
Here are the responses you obtained:

1 1 1 1 1 1 2 2 3 4

Summarizing Distributions



Mean: 1.7
Range: 3 (1 to 4)
IQR: 1 (1 to 2)
 s^2 : 1.122
s: 1.059



Mean: 2.4
Range: 6 (1 to 7)
IQR: 2 (1 to 3)
 s^2 : 4.489
s: 2.119

Measures of Spread for Discrete Variables

Index of Diversity (D)

Measure of spread (or dispersion) for discrete variables

Measures whether two observations selected randomly from a population are likely to fall into the same or different categories

$$D_Y = 1 - \sum_{i=1}^K p_i^2$$

where K indexes the categories in discrete variable Y and p is the proportion of observation in the i^{th} category

Measures of Spread for Discrete Variables

	Cat. 1	Cat. 2	Cat. 3	D
Y	10	10	10	0.667
Z	20	5	5	0.500

The higher the value of D , the more dispersed the cases across categories (and thus the less variability there is); $D = 0$ when all cases are in one category

The maximum value of D depends on how many categories there are ... thus D cannot be compared across discrete variable with different numbers of categories

Measures of Spread for Discrete Variables

D doesn't work when variables have different numbers of categories

	Cat. 1	Cat. 2	Cat. 3	Cat. 4
Y	10	10	10	10
Z	10	10	10	

$$D = 1 - \sum_{i=1}^K p_i^2$$

$$D_Y = 1 - [(0.25^2) + (0.25^2) + (0.25^2) + (0.25^2)] = 0.75$$

$$D_Z = 1 - [(0.33^2) + (0.33^2) + (0.33^2)] = 0.67$$

But they are both uniformly distributed...

Measures of Spread for Discrete Variables

Index of Qualitative Variation (IQV)

A measure of dispersion in the distribution of a discrete variable that is standardized such that it can be compared across discrete variables with different numbers of categories

$$IQV = [K/(K - 1)](D)$$

The maximum value of IQV is always 1.0 (when the cases are distributed evenly across categories)

The minimum value of IQV is always 0.0 (when all of the cases fall into one category)

Thus IQV can be compared across discrete variables that differ with respect to how many categories they have

Measures of Spread for Discrete Variables

D doesn't work when variables have different numbers of categories

	Cat. 1	Cat. 2	Cat. 3	Cat. 4
Y	10	10	10	10
Z	10	10	10	

$$IQV = [K/(K - 1)](D)$$

$$IQV_Y = [4/(4-1)] \times 0.75 = 1.00$$

$$IQV_Z = [3/(3-1)] \times 0.67 = 1.00$$

Want More?

David Lane's Text

http://onlinestatbook.com/2/summarizing_distributions/summarizing_distributions.html

Richard Lowry's Text

<http://vassarstats.net/textbook/ch2pt1.html> and
<http://vassarstats.net/textbook/ch2pt2.html>

Gerard Dallal's Text

<http://www.jerrydallal.com/LHSP/summary.htm>