Descriptive Statistics

The farthest most people ever get

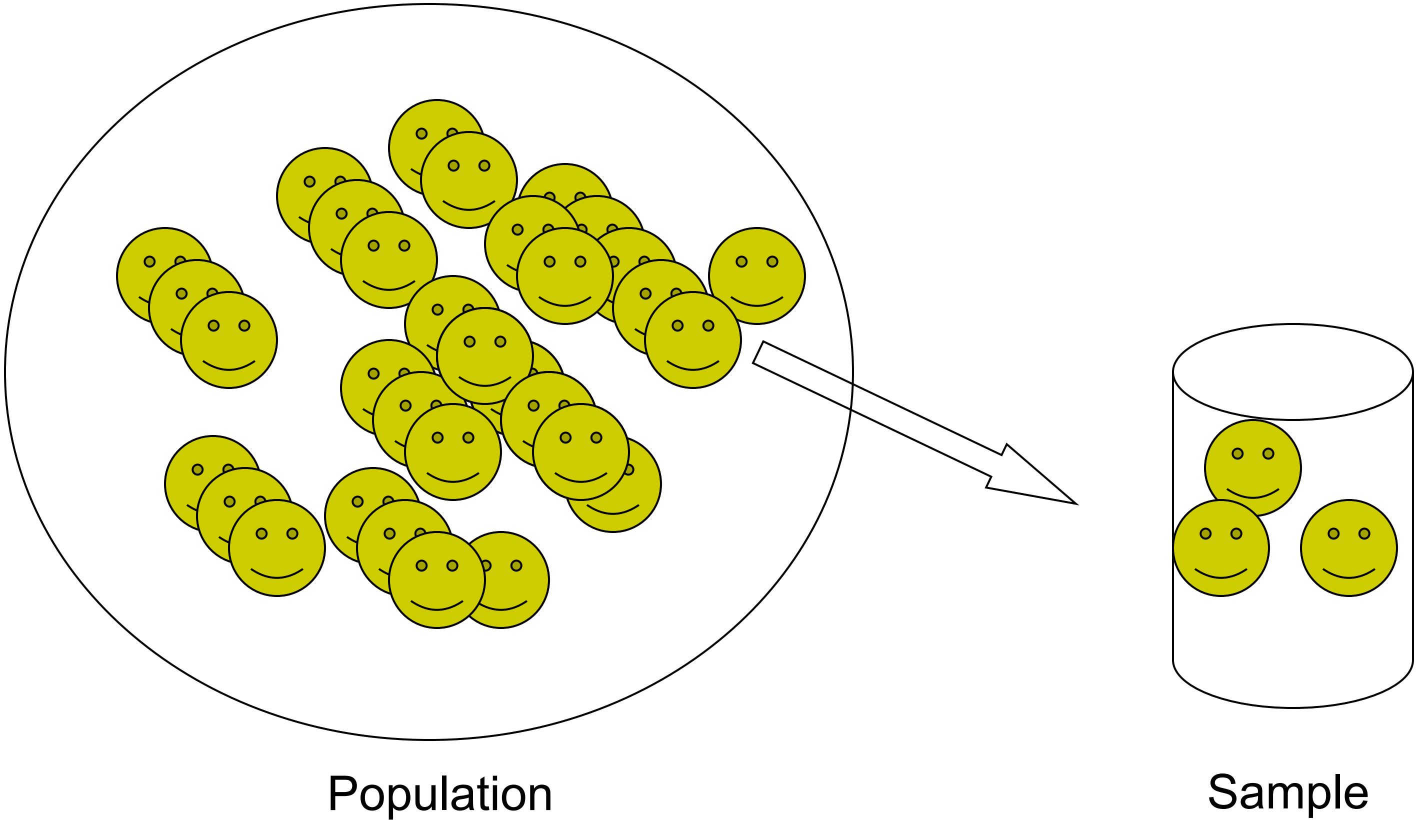
Descriptive Statistics

- Descriptive Statistics are Used by Researchers to Report on Populations <u>and</u> Samples

- In Sociology:
  Summary descriptions of measurements (variables) taken about a group of people

- By Summarizing Information, Descriptive Statistics Speed Up and Simplify Comprehension of a Group's Characteristics

Sample vs. Population



Descriptive Statistics

An Illustration:

Which Group is Smarter?

Class A--IQs of 13 Students		Class B--IQs of 13 Students	
102	115	127	162
128	109	131	103
131	89	96	111
98	106	80	109
140	119	93	87
93	97	120	105
110		109	

Each individual may be different. If you try to understand a group by remembering the qualities of each member, you become overwhelmed and fail to understand the group.

Descriptive **Statistics**

Which group is smarter now?

Class A--Average IQ	Class B--Average IQ
110.54	110.23

They're roughly the same!

With a summary descriptive statistic, it is much easier to answer our question.

Descriptive Statistics

Types of descriptive statistics:

- Organize Data
  - Tables
  - Graphs

- Summarize Data
  - Central Tendency
  - Variation

Descriptive Statistics

Types of descriptive statistics:

- Organize Data
  - Tables
    - Frequency Distributions
    - Relative Frequency Distributions
  - Graphs
    - Bar Chart or Histogram
    - Stem and Leaf Plot
    - Frequency Polygon

SPSS Output for Frequency Distribution

IQ

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	82.00	1	4.2	4.2	4.2
	87.00	1	4.2	4.2	8.3
	89.00	1	4.2	4.2	12.5
	93.00	2	8.3	8.3	20.8
	96.00	1	4.2	4.2	25.0
	97.00	1	4.2	4.2	29.2
	98.00	1	4.2	4.2	33.3
	102.00	1	4.2	4.2	37.5
	103.00	1	4.2	4.2	41.7
	105.00	1	4.2	4.2	45.8
	106.00	1	4.2	4.2	50.0
	107.00	1	4.2	4.2	54.2
	109.00	1	4.2	4.2	58.3
	111.00	1	4.2	4.2	62.5
	115.00	1	4.2	4.2	66.7
	119.00	1	4.2	4.2	70.8
	120.00	1	4.2	4.2	75.0
	127.00	1	4.2	4.2	79.2
	128.00	1	4.2	4.2	83.3
	131.00	2	8.3	8.3	91.7
	140.00	1	4.2	4.2	95.8
	162.00	1	4.2	4.2	100.0
	Total	24	100.0	100.0	

Frequency Distribution

Frequency Distribution of IQ for Two Classes

IQ	Frequency
82.00	1
87.00	1
89.00	1
93.00	2
96.00	1
97.00	1
98.00	1
102.00	1
103.00	1
105.00	1
106.00	1
107.00	1
109.00	1
111.00	1
115.00	1
119.00	1
120.00	1
127.00	1
128.00	1
131.00	2
140.00	1
162.00	1
Total	24

Relative Frequency Distribution

Relative Frequency Distribution of IQ for Two Classes

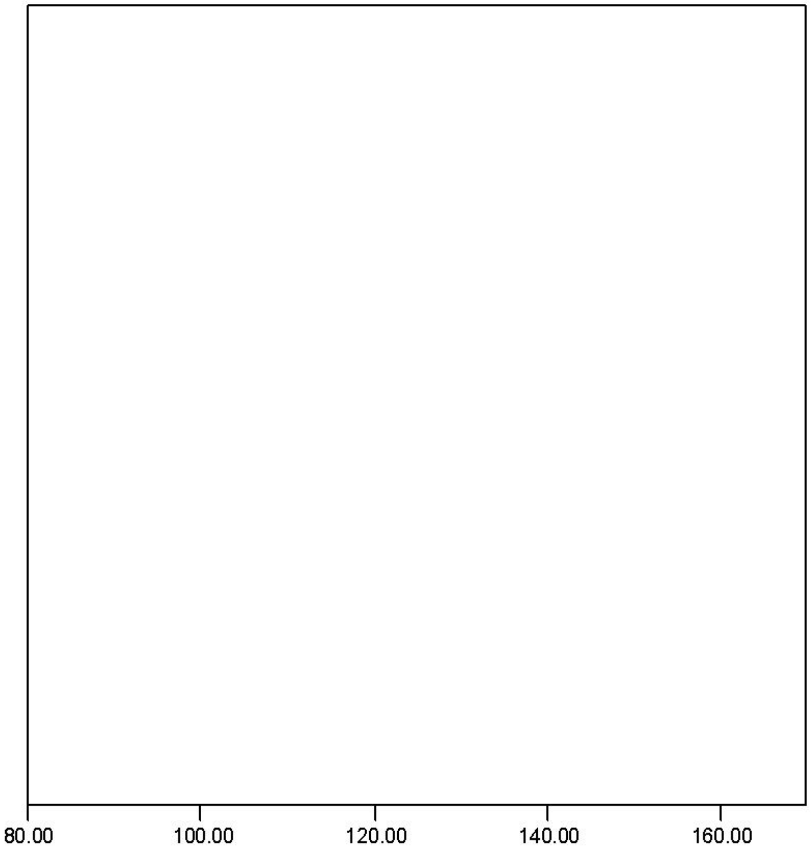
IQ	Frequency	Percent	Valid Percent	Cumulative Percent
82.00	1	4.2	4.2	4.2
87.00	1	4.2	4.2	8.3
89.00	1	4.2	4.2	12.5
93.00	2	8.3	8.3	20.8
96.00	1	4.2	4.2	25.0
97.00	1	4.2	4.2	29.2
98.00	1	4.2	4.2	33.3
102.00	1	4.2	4.2	37.5
103.00	1	4.2	4.2	41.7
105.00	1	4.2	4.2	45.8
106.00	1	4.2	4.2	50.0
107.00	1	4.2	4.2	54.2
109.00	1	4.2	4.2	58.3
111.00	1	4.2	4.2	62.5
115.00	1	4.2	4.2	66.7
119.00	1	4.2	4.2	70.8
120.00	1	4.2	4.2	75.0
127.00	1	4.2	4.2	79.2
128.00	1	4.2	4.2	83.3
131.00	2	8.3	8.3	91.7
140.00	1	4.2	4.2	95.8
162.00	1	4.2	4.2	100.0
Total	24	100.0	100.0	

Grouped Relative Frequency Distribution

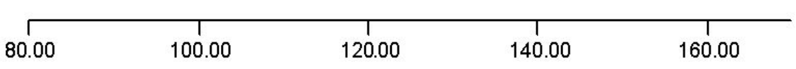
Relative Frequency Distribution of IQ for Two Classes

IQ	Frequency	Percent	Cumulative Percent
80 – 89	3	12.5	12.5
90 – 99	5	20.8	33.3
100 – 109	6	25.0	58.3
110 – 119	3	12.5	70.8
120 – 129	3	12.5	83.3
130 – 139	2	8.3	91.6
140 – 149	1	4.2	95.8
150 and over	1	4.2	100.0
Total	24	100.0	100.0

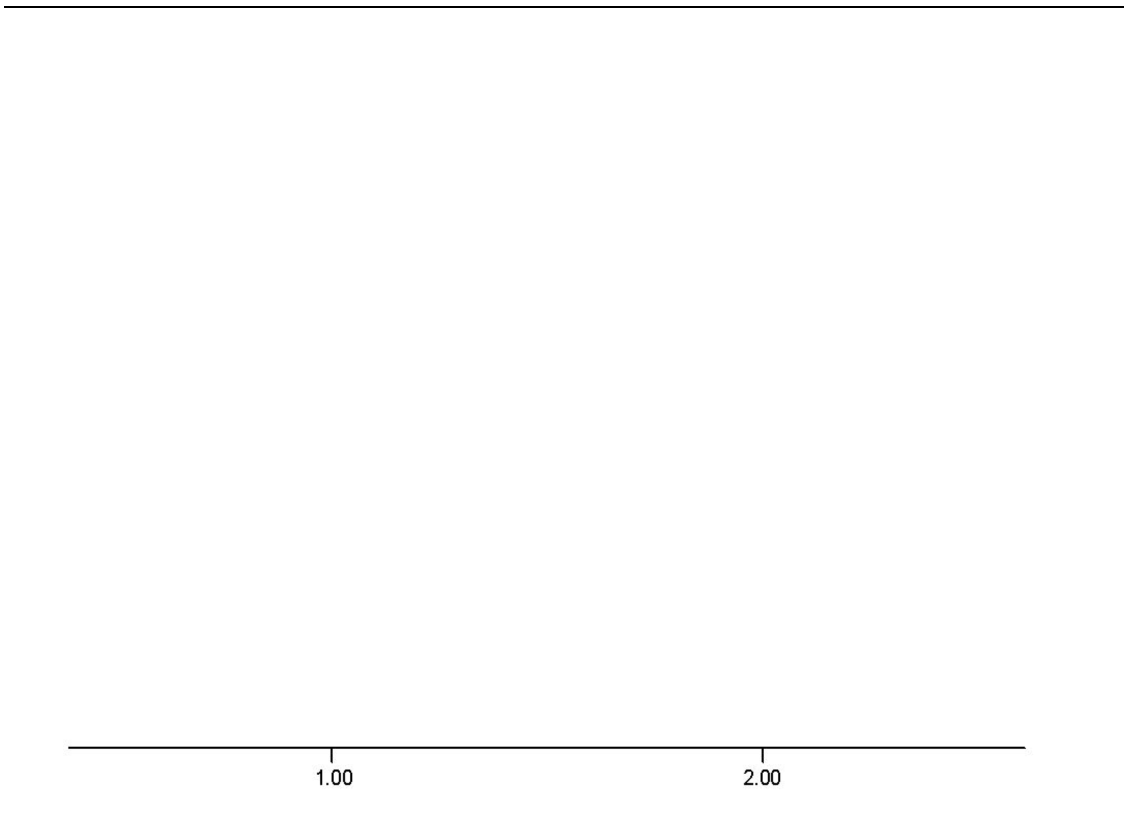
SPSS Output for Histogram



Histogram



Bar Graph



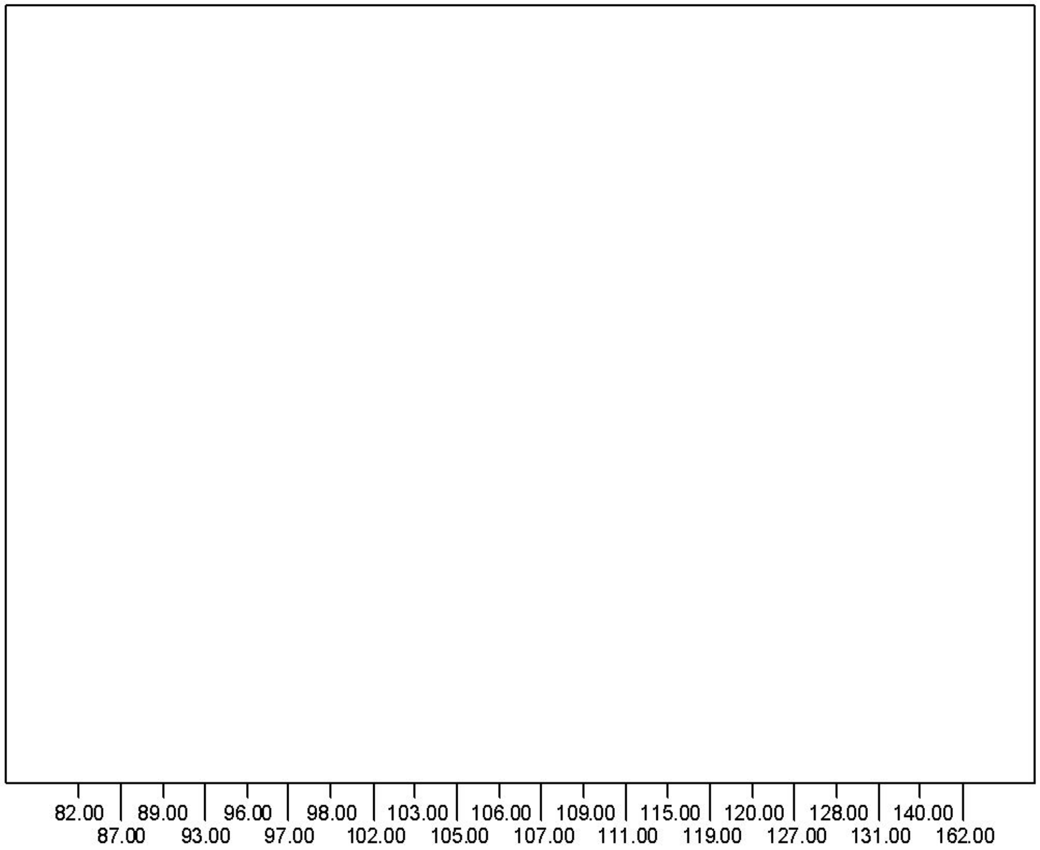
Stem and Leaf Plot

Stem and Leaf Plot of IQ for Two Classes

Stem	Leaf
8	2 7 9
9	3 6 7 8
10	2 3 5 6 7 9
11	1 5 9
12	0 7 8
13	1
14	0
15	
16	2

Note: SPSS does not do a good job of producing these.

SPSS Output of a Frequency Polygon



Descriptive Statistics

Summarizing Data:

- Central Tendency (or Groups' "Middle Values")
  - Mean
  - Median
  - Mode

- Variation (or Summary of Differences Within Groups)
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation

Mean

Most commonly called the "average."

Add up the values for each case and divide by the total number of cases.

$$\text{Y-bar} = \frac{(Y1 + Y2 + . . . + Yn)}{n}$$

$$\text{Y-bar} = \frac{\Sigma\, Y\mathit{i}}{n}$$

Mean

What's up with all those symbols, man?

$$\text{Y-bar} = \frac{(Y1 + Y2 + \ldots + Yn)}{n}$$

$$\text{Y-bar} = \frac{\Sigma Y\mathit{i}}{n}$$

Some Symbolic Conventions in this Class:

- Y = your variable (could be X or Q or ☺ or even "Glitter")
- "-bar" or line over symbol of your variable = mean of that variable
- Y1 = first case's value on variable Y
- ". . ." = ellipsis = continue sequentially
- Yn = last case's value on variable Y
- n = number of cases in your sample
- Σ = Greek letter "sigma" = sum or add up what follows
- *i* = a typical case or each case in the sample (1 through n)

Mean

Class A--IQs of 13 Students

102	115
128	109
131	89
98	106
140	119
93	97
110	

$\Sigma Y_i = 1437$

$$\text{Y-bar}_A = \frac{\Sigma Y_i}{n} = \frac{1437}{13} = 110.54$$

Class B--IQs of 13 Students

127	162
131	103
96	111
80	109
93	87
120	105
109	

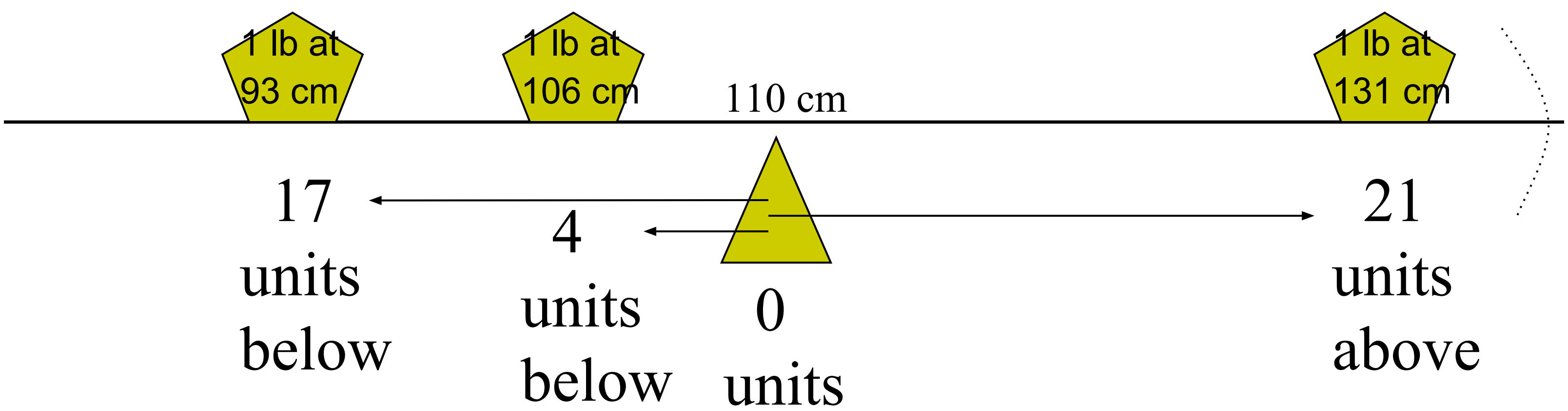
$\Sigma Y_i = 1433$

$$\text{Y-bar}_B = \frac{\Sigma Y_i}{n} = \frac{1433}{13} = 110.23$$

Mean

The mean is the "balance point."

Each person's score is like 1 pound placed at the score's position on a see-saw. Below, on a 200 cm see-saw, the mean equals 110, the place on the see-saw where a fulcrum finds balance:

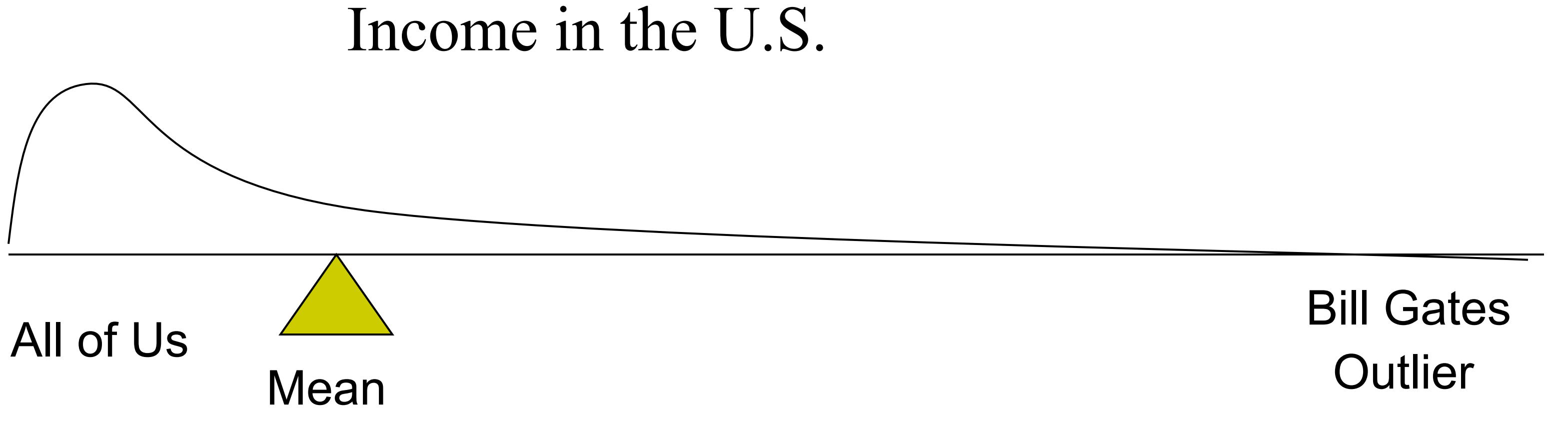


The scale is balanced because…

$$17 + 4 \text{ on the left} = 21 \text{ on the right}$$

Mean

1. Means can be badly affected by outliers (data points with extreme values unlike the rest)
2. Outliers can make the mean a bad measure of central tendency or common experience



Median

The middle value when a variable's values are ranked in order; the point that divides a distribution into two equal halves.

When data are listed in order, the median is the point at which 50% of the cases are above and 50% below it.

The 50th percentile.

Median

Class A--IQs of 13 Students

89
93
97
98
102
106
109
110
115
119
128
131 140

Median = 109

(six cases above, six below)

Median

If the first student were to drop out of Class A, there would be a new median:

89
93
97
98
102
106
109
110
115
119
128
131
140

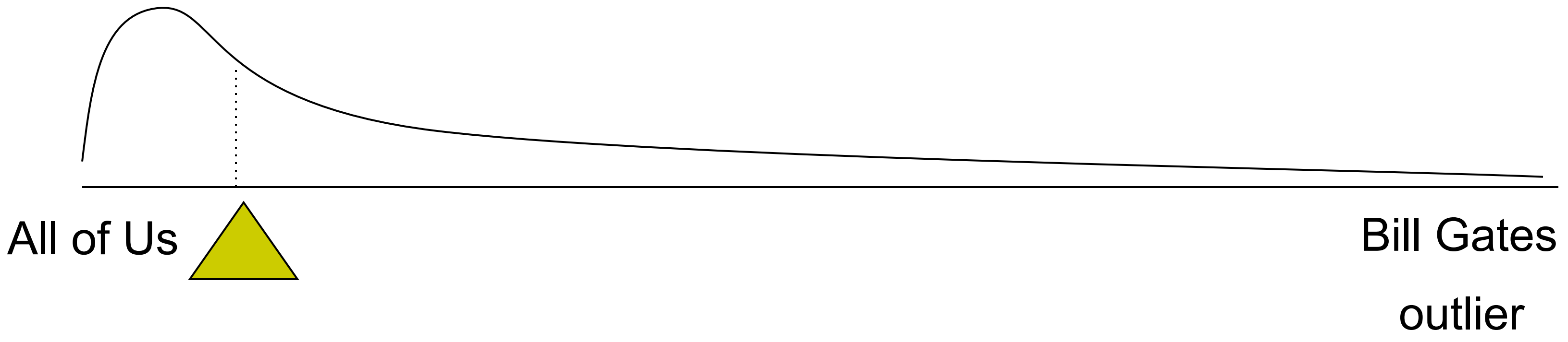
Median = 109.5

109 + 110 = 219/2 = 109.5

(six cases above, six below)

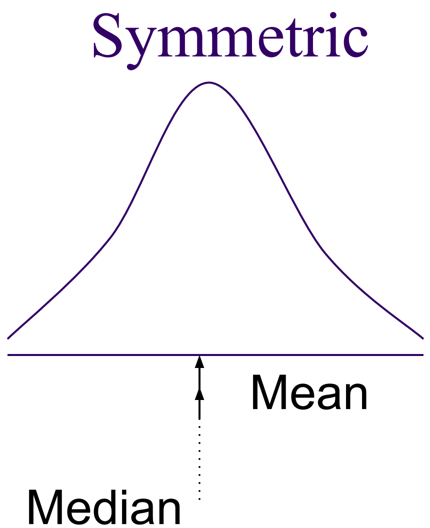
Median

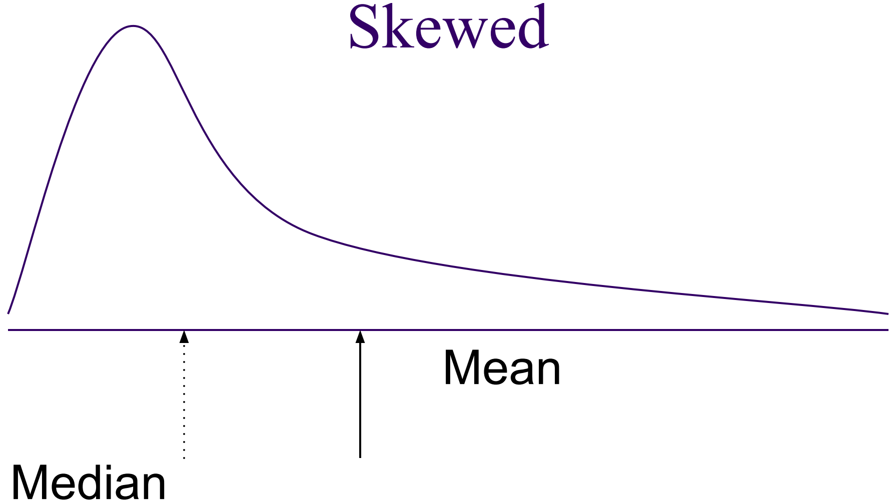
1. The median is unaffected by outliers, making it a better measure of central tendency, better describing the "typical person" than the mean when data are skewed.



Median

2. If the recorded values for a variable form a symmetric distribution, the median and mean are identical.
3. In skewed data, the mean lies further toward the skew than the median.





Median

The middle score or measurement in a set of ranked scores or measurements; the point that divides a distribution into two equal halves.

Data are listed in order—the median is the point at which 50% of the cases are above and 50% below.

The 50th percentile.

Mode

The most common data point is called the mode.

The combined IQ scores for Classes A & B:

80 87 89 93 93 96 97 98 102 103 105 106 109 109 109 110 111 115 119 120 127 128 131 131 140 162

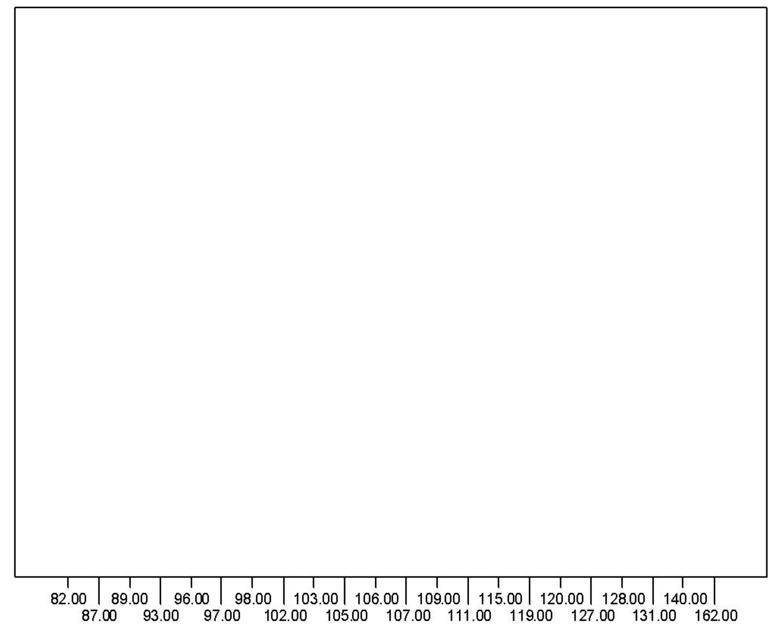
A la mode!!

BTW, It is possible to have more than one mode!

Mode

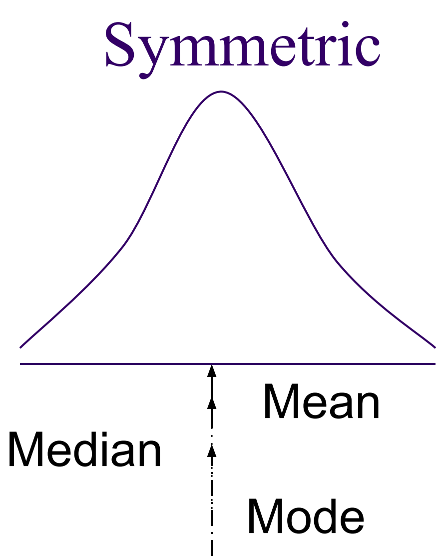
It may mot be at the center of a distribution.

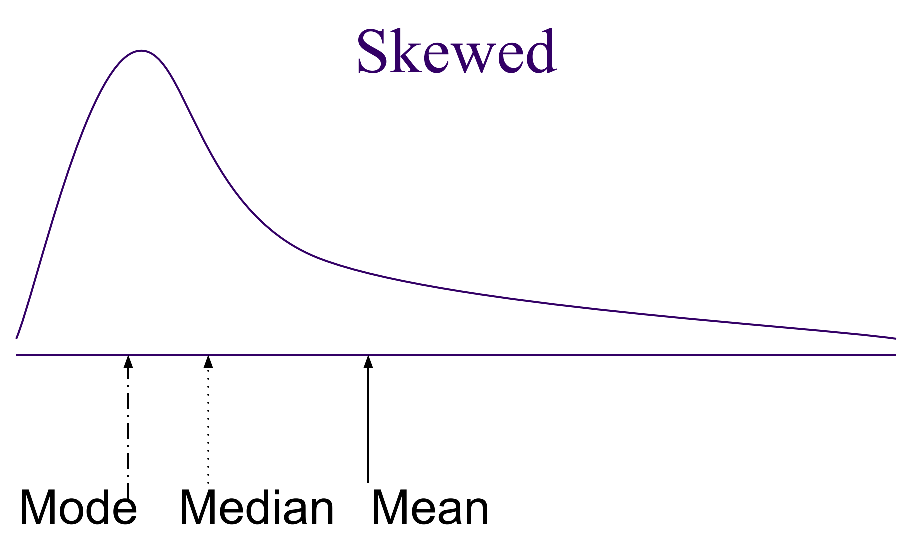
Data distribution on the right is “bimodal” (even statistics can be open-minded)



Mode

1. It may give you the most likely experience rather than the "typical" or "central" experience.
2. In symmetric distributions, the mean, median, and mode are the same.
3. In skewed data, the mean and median lie further toward the skew than the mode.





Descriptive Statistics

Summarizing Data:

- Central Tendency (or Groups' "Middle Values")
  - Mean
  - Median
  - Mode

- Variation (or Summary of Differences Within Groups)
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation

Range

The spread, or the distance, between the lowest and highest values of a variable.

To get the range for a variable, you subtract its lowest value from its highest value.

Class A--IQs of 13 Students

102	115
128	109
131	89
98	106
140	119
93	97
110	

Class A Range = 140 - 89 = 51

Class B--IQs of 13 Students

127	162
131	103
96	111
80	109
93	87
120	105
109	

Class B Range = 162 - 80 = 82

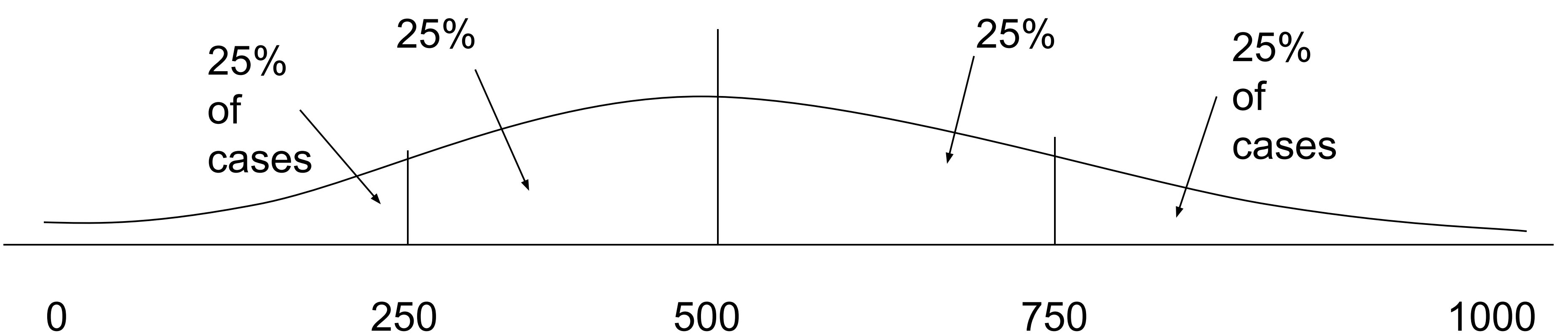
Interquartile Range

A quartile is the value that marks one of the divisions that breaks a series of values into four equal parts.

The median is a quartile and divides the cases in half.

25^{th} percentile is a quartile that divides the first ¼ of cases from the latter ¾.
75^{th} percentile is a quartile that divides the first ¾ of cases from the latter ¼.

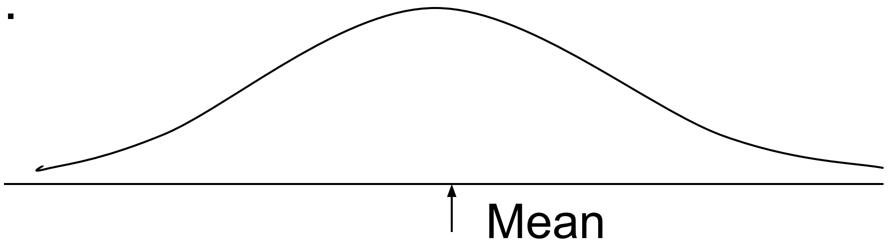
The interquartile range is the distance or range between the 25^{th} percentile and the 75^{th} percentile. Below, what is the interquartile range?



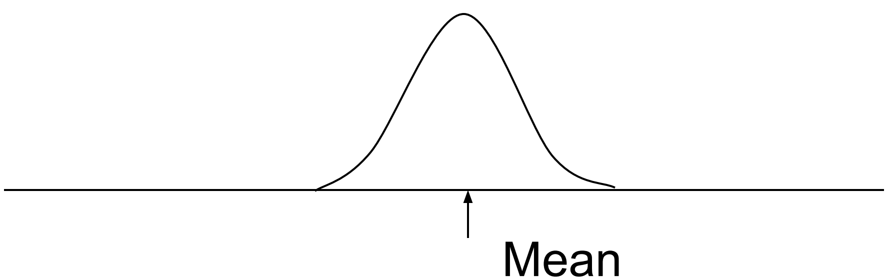
Variance

A measure of the spread of the recorded values on a variable. A measure of dispersion.

The larger the variance, the further the individual cases are from the mean.



The smaller the variance, the closer the individual scores are to the mean.



Variance

Variance is a number that at first seems complex to calculate.

Calculating variance starts with a "deviation."

A deviation is the distance away from the mean of a case's score.

Y_i – Y-bar

If the average person's car costs $20,000, my deviation from the mean is - $14,000!

6K - 20K = -14K

Variance

The deviation of 102 from 110.54 is? Deviation of 115?

Class A--IQs of 13 Students

102	115
128	109
131	89
98	106
140	119
93	97
110	

Y-bar$_A$ = *110.54*

Variance

The deviation of 102 from 110.54 is? Deviation of 115?

102 - 110.54 = -8.54   115 - 110.54 = 4.46

Class A--IQs of 13 Students

102	115
128	109
131	89
98	106
140	119
93	97
110	

Y-bar$_A$ = *110.54*

Variance

- We want to add these to get total deviations, but if we were to do that, we would get zero every time. Why?
- We need a way to eliminate negative signs.

Squaring the deviations will eliminate negative signs...

A Deviation Squared: $(Y_i - \text{Y-bar})^2$

Back to the IQ example,

A deviation squared for 102 is: of 115:

$(102 - 110.54)^2 = (-8.54)^2 = 72.93$ $(115 - 110.54)^2 = (4.46)^2 = 19.89$

Variance

If you were to add all the squared deviations together, you'd get what we call the "Sum of Squares."

Sum of Squares (SS) = $\Sigma\,(Y\mathit{i} - Y\text{-bar})^2$

$SS = (Y1 - Y\text{-bar})^2 + (Y2 - Y\text{-bar})^2 + \ldots + (Yn - Y\text{-bar})^2$

Variance

Class A, sum of squares:

$(102 – 110.54)^2 + (115 – 110.54)^2 +$
$(126 – 110.54)^2 + (109 – 110.54)^2 +$
$(131 – 110.54)^2 + (89 – 110.54)^2 +$
$(98 – 110.54)^2 + (106 – 110.54)^2 +$
$(140 – 110.54)^2 + (119 – 110.54)^2 +$
$(93 – 110.54)^2 + (97 – 110.54)^2 +$
$(110 – 110.54) = SS = 2825.39$

Class A--IQs of 13 Students

102	115
128	109
131	89
98	106
140	119
93	97
110	

Y-bar = 110.54

Variance

The last step…

The approximate average sum of squares is the variance.

SS/N = Variance for a population.

SS/n-1 = Variance for a sample.

Variance = $\Sigma(Y\textit{i} – \text{Y-bar})^2$ / n – 1

Variance

For Class A, Variance = 2825.39 / n - 1

= 2825.39 / 12 = 235.45

How helpful is that???

Standard Deviation

To convert variance into something of meaning, let's create standard deviation.

The square root of the variance reveals the average deviation of the observations from the mean.

$$\text{s.d.} = \sqrt{\frac{\Sigma(Y_i - \text{Y-bar})^2}{n - 1}}$$

Standard Deviation

For Class A, the standard deviation is:

$$\sqrt{235.45} = 15.34$$

The average of persons' deviation from the mean IQ of 110.54 is 15.34 IQ points.

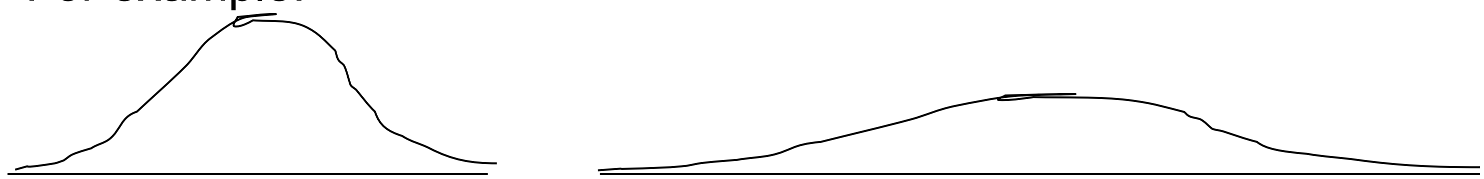
Review:

1. Deviation
2. Deviation squared
3. Sum of squares
4. Variance
5. Standard deviation

Standard Deviation

1. Larger s.d. = greater amounts of variation around the mean.

For example:



19 25 31 13 25 37

$\bar{Y} = 25$ $\bar{Y} = 25$

s.d. = 3 s.d. = 6

2. s.d. = 0 only when all values are the same (only when you have a constant and not a "variable")
3. If you were to "rescale" a variable, the s.d. would change by the same magnitude—if we changed units above so the mean equaled 250, the s.d. on the left would be 30, and on the right, 60
4. Like the mean, the s.d. will be inflated by an outlier case value.

Standard Deviation

- Note about computational formulas:
  - Your book provides a useful short-cut formula for computing the variance and standard deviation.
  - This is intended to make hand calculations as quick as possible.
  - They obscure the conceptual understanding of our statistics.
  - SPSS and the computer are "computational formulas" now.

Practical Application for Understanding Variance and Standard Deviation

Even though we live in a world where we pay real dollars for goods and services (not percentages of income), most American employers issue raises based on percent of salary.

Why do supervisors think the most fair raise is a percentage raise?

Answer: 1) Because higher paid persons win the most money.

2) The easiest thing to do is raise everyone's salary by a fixed percent.

If your budget went up by 5%, salaries can go up by 5%.

The problem is that the flat percent raise gives unequal increased rewards. . .

Practical Application for Understanding Variance and Standard Deviation

Acme Toilet Cleaning Services

Salary Pool: $200,000

Incomes:

President: $100K; Manager: 50K; Secretary: 40K; and Toilet Cleaner: 10K

Mean: $50K

Range: $90K

Variance: $1,050,000,000

Standard Deviation: $32.4K

These can be considered "measures of inequality"

Now, let's apply a 5% raise.

Practical Application for Understanding Variance and Standard Deviation

After a 5% raise, the pool of money increases by $10K to $210,000

Incomes:

President: $105K; Manager: 52.5K; Secretary: 42K; and Toilet Cleaner: 10.5K

Mean: $52.5K – went up by 5%

Range: $94.5K – went up by 5%

Variance: $1,157,625,000

Standard Deviation: $34K –went up by 5%

Measures of Inequality

The flat percentage raise increased inequality. The top earner got 50% of the new money. The bottom earner got 5% of the new money. Measures of inequality went up by 5%.

Last year's statistics:

Acme Toilet Cleaning Services annual payroll of $200K

Incomes:

$100K, 50K, 40K, and 10K

Mean: $50K

Range: $90K; Variance: $1,050,000,000; Standard Deviation: $32.4K

Practical Application for Understanding Variance and Standard Deviation

The flat percentage raise increased inequality. The top earner got 50% of the new money. The bottom earner got 5% of the new money. Inequality increased by 5%.

Since we pay for goods and services in real dollars, not in percentages, there are substantially more new things the top earners can purchase compared with the bottom earner for the rest of their employment years.

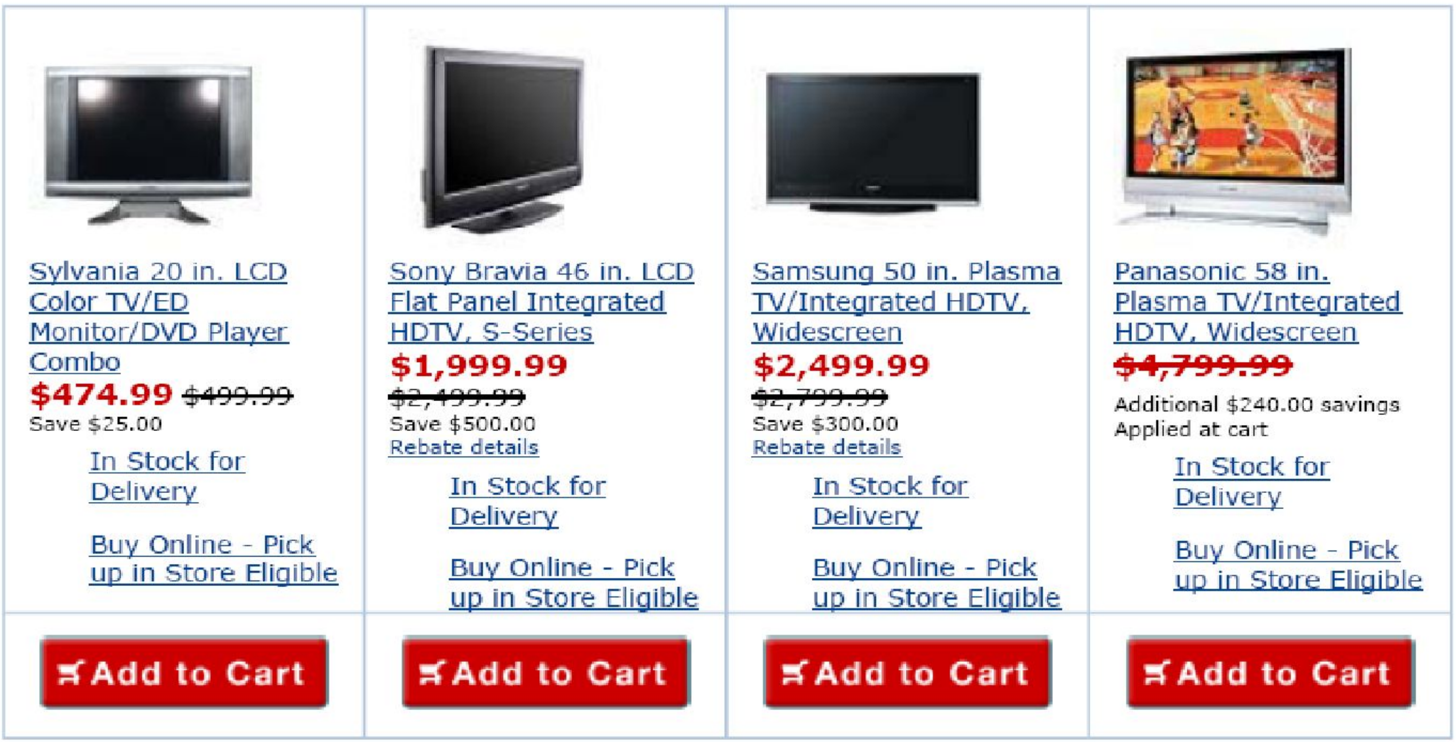
Acme Toilet Cleaning Services is giving the earners $5,000, $2,500, $2,000, and $500 more respectively ***each and every year forever***.

What does this mean in terms of compounding raises?

Acme is essentially saying: “Each year we’ll buy you a new TV, in addition to everything else you buy, here’s what you’ll get:”

Practical Application for Understanding Variance and Standard Deviation

Toilet Cleaner Secretary Manager President



The gap between the rich and poor expands.

This is why some progressive organizations give a percentage raise with a flat increase for lowest wage earners. For example, 5% or $1,000, whichever is greater.

Descriptive Statistics

Summarizing Data:

- Central Tendency (or Groups' "Middle Values")
  - Mean
  - Median
  - Mode

- Variation (or Summary of Differences Within Groups)
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation

- …Wait! There's more

Box-Plots

A way to graphically portray almost all the descriptive statistics at once is the box-plot.

A box-plot shows: Upper and lower quartiles

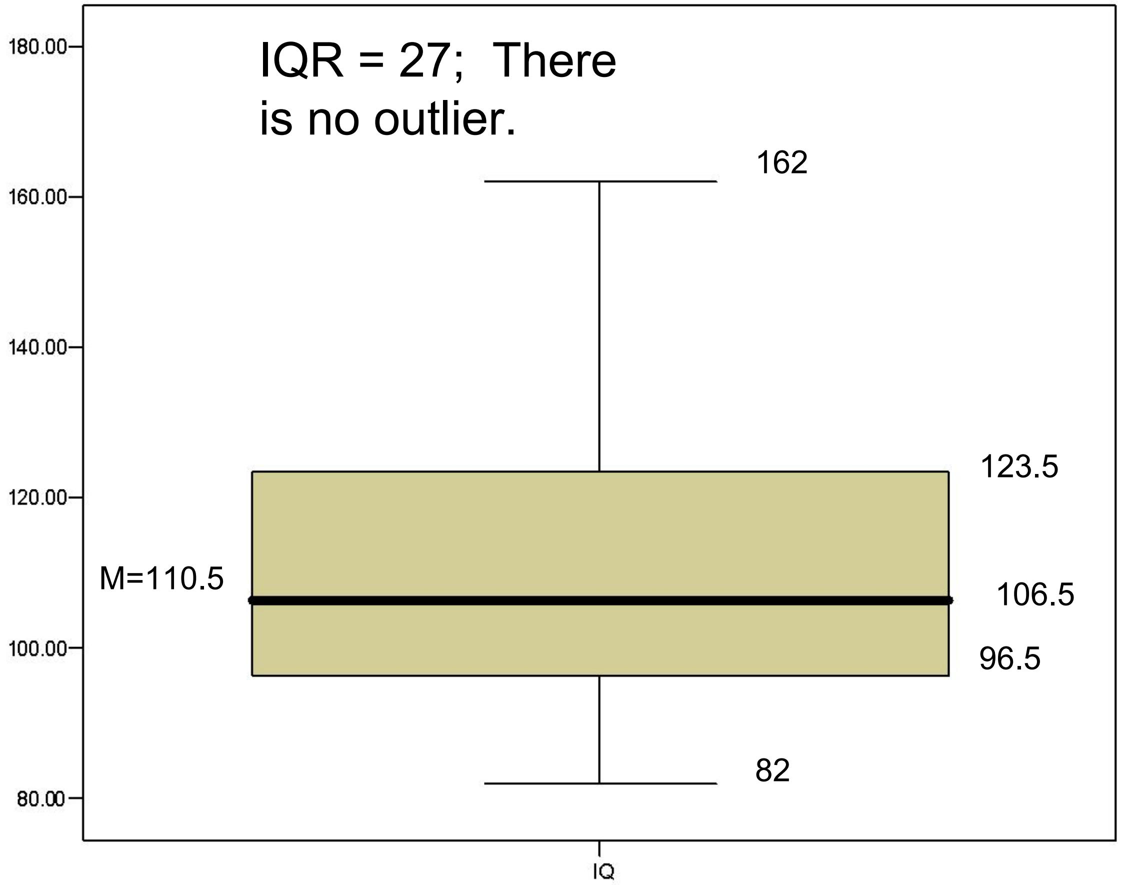
Mean

Median

Range

Outliers (1.5 IQR)

Box-Plots


180.00
160.00
140.00
120.00
100.00
80.00
IQR = 27; There
is no outlier.
162
123.5
M=110.5
106.5
96.5
82
IQ

IQV—Index of Qualitative Variation

- For nominal variables
- Statistic for determining the dispersion of cases across categories of a variable.
- Ranges from 0 (no dispersion or variety) to 1 (maximum dispersion or variety)
- 1 refers to even numbers of cases in all categories, NOT that cases are distributed like population proportions
- IQV is affected by the number of categories

IQV—Index of Qualitative Variation

To calculate:

$$\mathrm{IQV} = \frac{K(100^2 - \Sigma\ \mathrm{cat.}\%^2)}{100^2(K - 1)}$$

K=# of categories

Cat.% = percentage in each category

IQV—Index of Qualitative Variation

Problem: Is SJSU more diverse than UC Berkeley?

Solution: Calculate IQV for each campus to determine which is higher.

SJSU:		UC Berkeley:	
Percent	Category	Percent	Category
00.6	Native American	00.6	Native American
06.1	Black	03.9	Black
39.3	Asian/PI	47.0	Asian/PI
19.5	Latino	13.0	Latino
34.5	White	35.5	White

What can we say before calculating? Which campus is more evenly distributed?

$$\text{IQV} = \frac{K\,(100^2 - \Sigma\ \text{cat.}\%^2)}{100^2(K-1)}$$

IQV—Index of Qualitative Variation

Problem: Is SJSU more diverse than UC Berkeley? YES

Solution: Calculate IQV for each campus to determine which is higher.

SJSU:

Percent	Category	$\%^2$
00.6	Native American	0.36
06.1	Black	37.21
39.3	Asian/PI	1544.49
19.5	Latino	380.25
34.5	White	1190.25

UC Berkeley:

Percent	Category	$\%^2$
00.6	Native American	0.36
03.9	Black	15.21
47.0	Asian/PI	2209.00
13.0	Latino	169.00
35.5	White	1260.25

SJSU: K = 5 $\quad \Sigma\ \text{cat.}\%^2 = 3152.56$

UC Berkeley: k = 5 $\quad \Sigma\ \text{cat.}\%^2 = 3653.82$

$100^2 = 10000$

$$IQV = \frac{K\,(100^2 - \Sigma\ \text{cat.}\%^2)}{100^2(K-1)}$$

SJSU:

5(10000 – 3152.56) = 34237.2

10000(5 – 1) = 40000 SJSU IQV = .856

UC Berkeley:

5(10000 – 3653.82) = 31730.9

10000(5 – 1) = 40000 UCB IQV =.793

Descriptive Statistics

- Now you are qualified use descriptive statistics!
- Questions?