Exploring Data

Exploring Data - I

- mydata #Prints mydta
- str(mydata) # Provides the structure of the dataset
- Class(mydata) # Provides the class of an object
- names(mydata) # Lists variables in the dataset
- names(mydata) = normVarNames(names(mydata))
- head(mydata) # First 6 rows of dataset
- head(mydata, n=10)# First 10 rows of dataset
- head(mydata, n= -10)# All rows but the last 10
- tail(mydata) # Last 6 rows
- tail(mydata, n=10)# Last 10 rows
- tail(mydata, n= -10)# All rows but the first 10

Exploring Data - II

- levels(mydata$v1) # List levels of factor v1 in mydata
- dim(mydata) # Dimensions of an object
- mydata[1,1] # First observation of first variable
- mydata[1:10,] # First 10 observations of all variables
- mydata[,2] # All observations of 2nd variable
- mydata[1:10,1:3]# First 10 rows of data of the first 3 variables
- head(mydata[,3:4], 2) # First two observations of 3rd & 4th variables
- edit(mydata) # Open data editor
- summary(mydata) # provides summary of each variable

Descriptive Statistics

Summarizing Data:

- ✓ Central Tendency (or What is the centre of dataset?)
  - ✓ Mean
  - ✓ Median
  - ✓ Mode
- Variation (or What is the spread of dataset?)
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation

Mean

Most commonly called the “average” and is the centre of gravity or balancing point.

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

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

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

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		

$\Sigma Y_i = 1433$

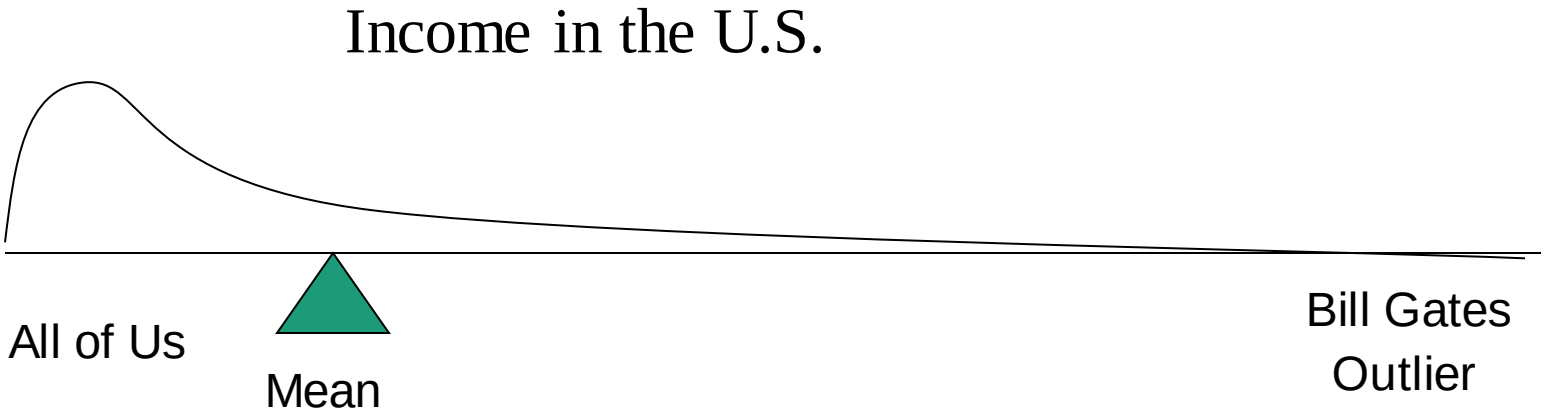
$$\text{Y-bar}_B = \frac{\Sigma Y_i}{n} = \frac{1433}{13} = 110.23$$

Pros of Mean

- crucial for inferential statistics

Cons of 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



Cons of Mean

Random sample of seven employees, obtaining the sample data (expressed in thousands of dollars).
24.8 22.8 24.6 192.5 25.2 18.5 23.7

the average income of employees at this corporation is $47,400" is surely misleading. It is approximately twice what six of the seven employees in the sample make and is nowhere near what any of them makes.

What went wrong?
the presence of the one executive in the sample, whose salary is so large compared to everyone else's, caused the numerator in the formula for the sample mean to be far too large, pulling the mean far to the right of where we think that the average "ought" to be, namely around $24,000 or $25,000. The number 192.5 in our data set is called an **outlier**, a number that is far away from most or all of the remaining measurements.

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 50^{th} 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

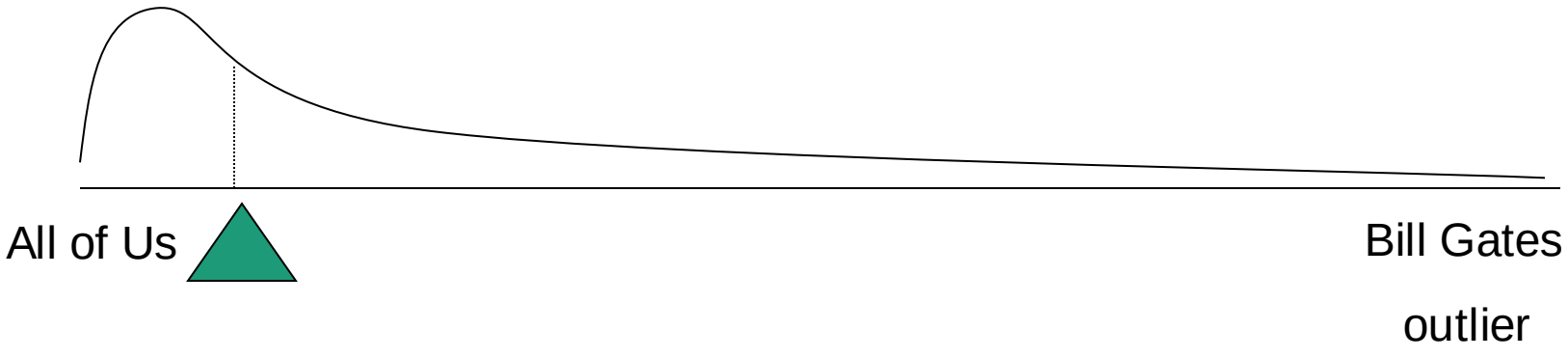
Median = 109.5

109 + 110 = 219/2 = 109.5

(six cases above, six below)

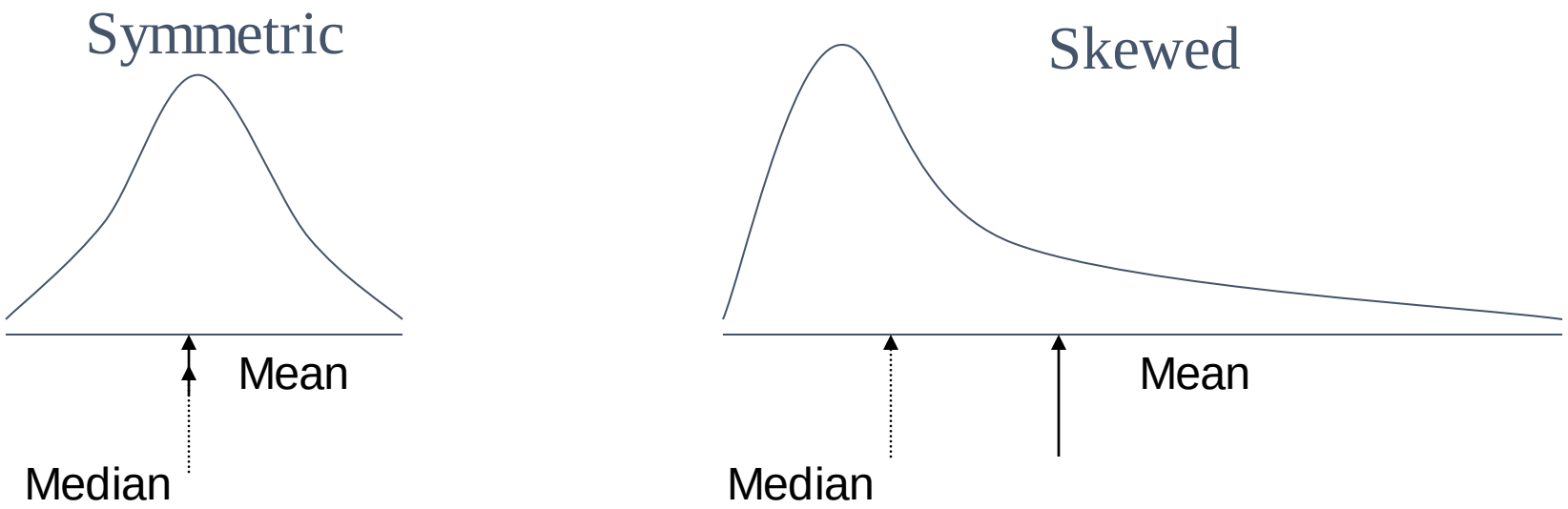
Median

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.



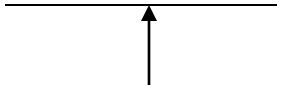
Mode

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



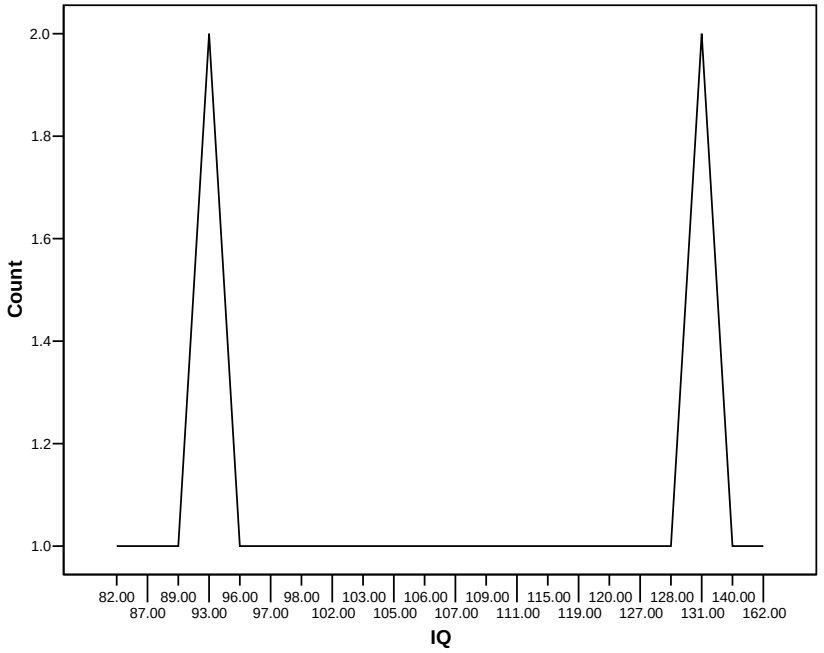
mode!!

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

Mode

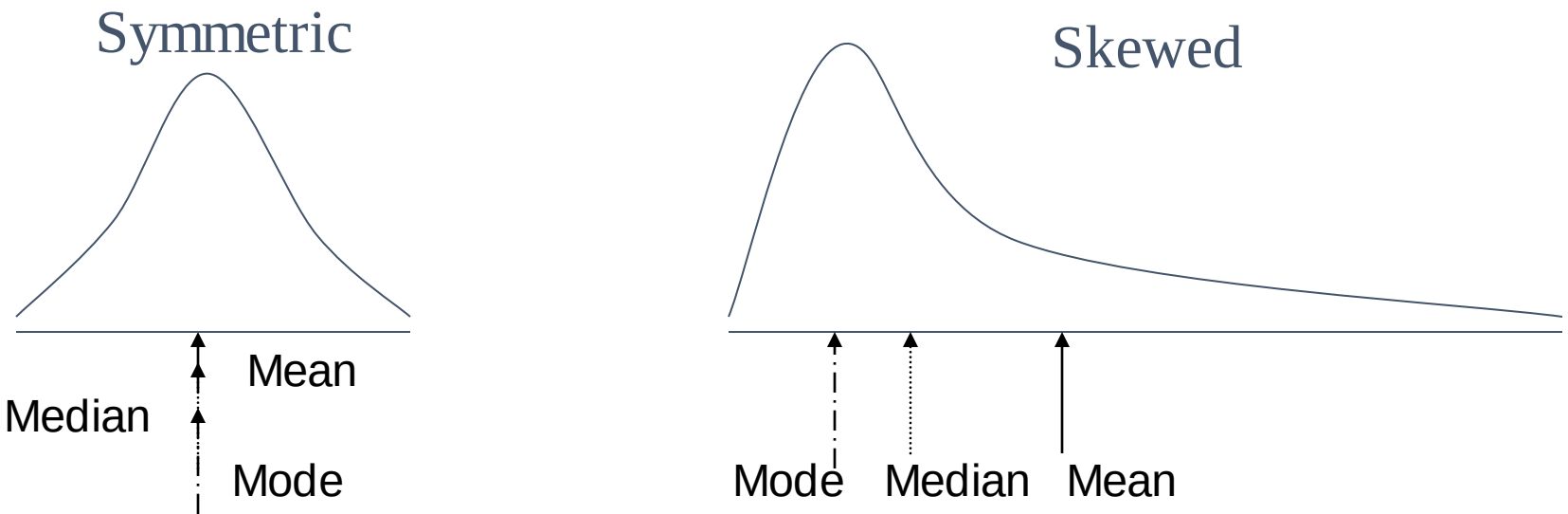
It may mot be at the center of a distribution.

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.



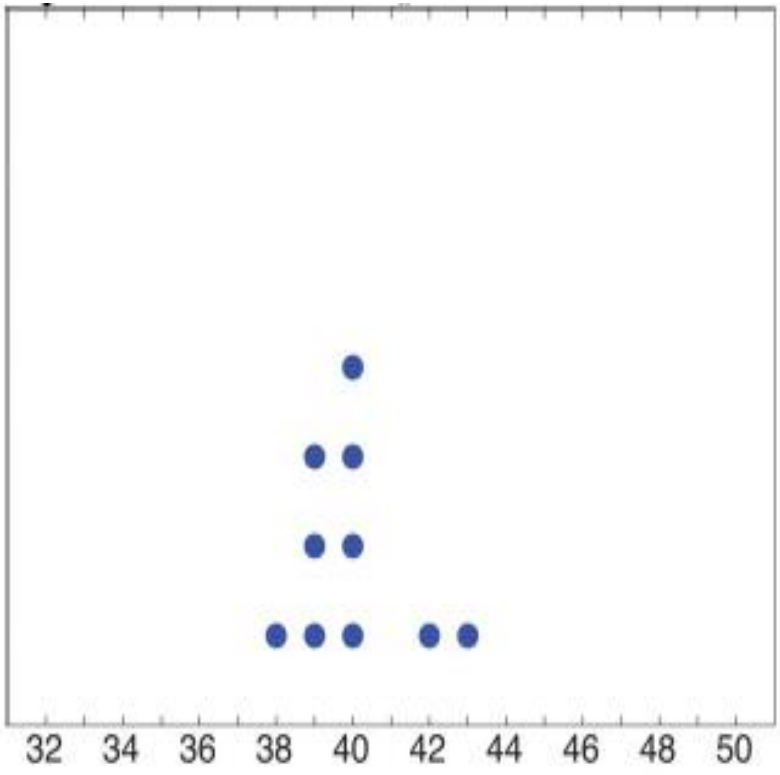
Mode

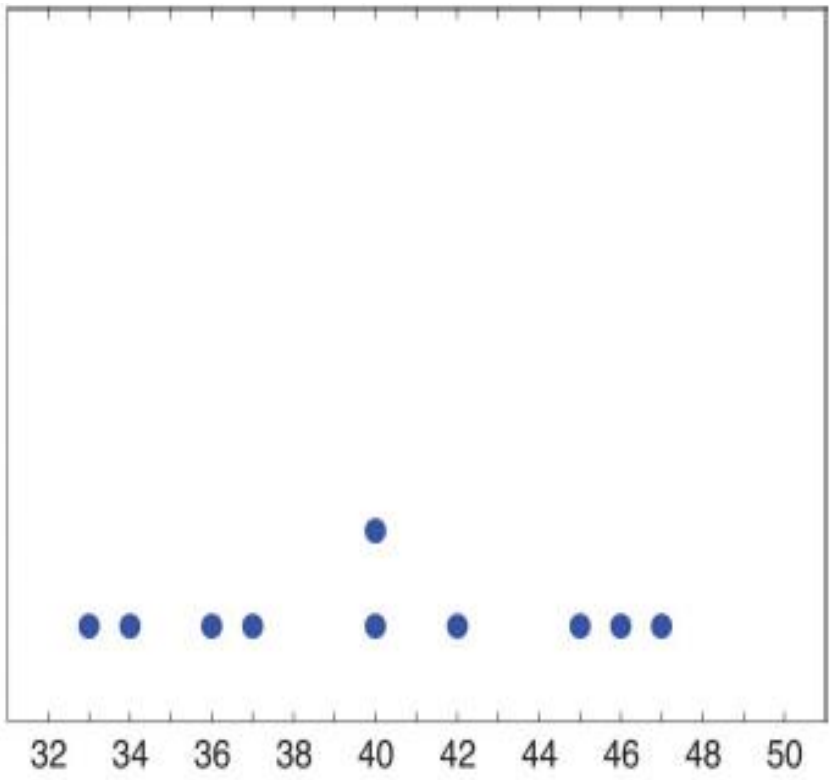
- The mode is a measure of central location since most real-life data sets have more observations near the center of the data range and fewer observations on the lower and upper ends. The value with the highest frequency is often in the middle of the data range.

Why do we need to measure variation or spread of data?

Data Set I:	40	38	42	40	39	39	43	40	39	40
Data Set II:	46	37	40	33	42	36	40	47	34	45

Why do we need to measure variation or spread of data?





Measures of spread

We have attached numbers to dataset to locate its center till now.

Associate to each data set numbers that measure how the data either scatter away from center or cluster close to it. These new quantities are called measure of spread or variability.

- ✓ 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	

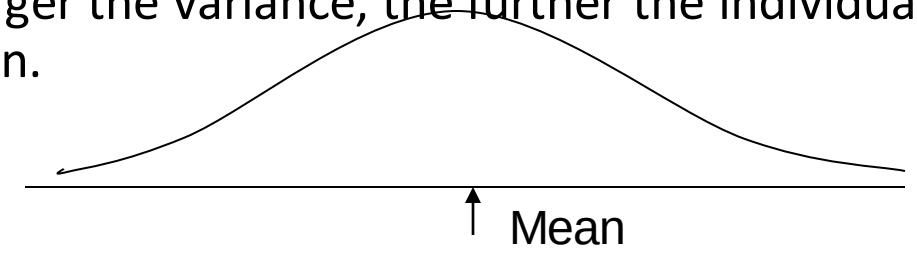
Class B Range = 162 - 80 = 82

- A smaller range indicates less variability (less dispersion) among the data, whereas a larger range indicates the opposite.

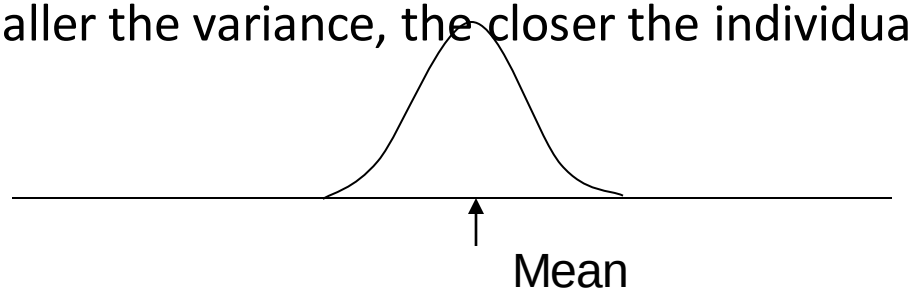
Variance

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.

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	

Mean ***= 110.54***

Variance

The deviation of 102 from 110.54 is?

102 - 110.54 = -8.54

Deviation of 115?

115 - 110.54 = 4.46

Class A--IQs of 13 Students

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

Mean ***= 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: $(Y_i - \text{Mean})^2$

Back to the IQ example,

A deviation squared for 102 is: $(102 - 110.54)^2 = (-8.54)^2 = 72.93$

of 115: $(115 - 110.54)^2 = (4.46)^2 = 19.89$

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) = $\Sigma\ (Yi - \text{Y-bar})^2$

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

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.

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

Variance

For Class A, Variance = 2825.39 / n - 1

= 2825.39 / 12 = 235.45

How helpful is that???

The variance has different units from the data/mean. For example, if the units in the data set were inches, the new units would be inches squared, or square inches. It is thus primarily of theoretical importance and will not be useful in practice

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.

$$\text{s.d.} = \sqrt{\frac{\Sigma(Yi - \text{Mean})^2}{n - 1}}$$

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.

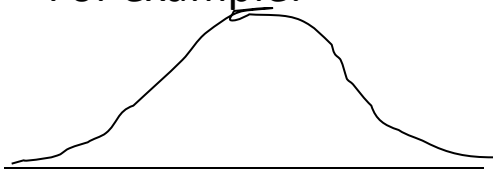
Review:

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

Y = 25

s.d. = 3

13 25 37

Y = 25

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.

Interquartile Range

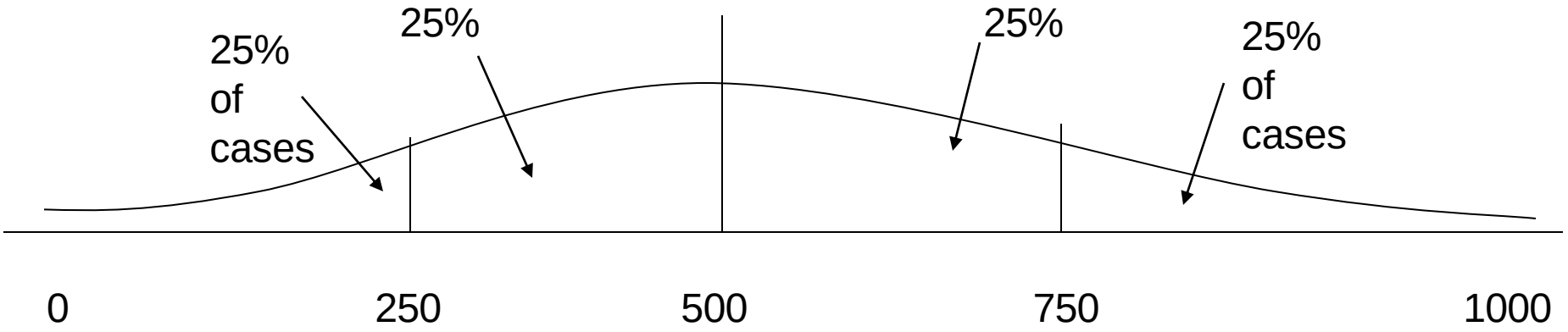
A quartile is the value that marks one of the divisions that breaks a series of values into four equal parts.

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

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

75th 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 25th percentile and the 75th percentile. Below, what is the interquartile range?



```
mean(mydata$SAT)
with(mydata, mean(SAT))
median(mydata$SAT)
table(mydata$Country)# Mode by frequencies -> max(table(mydata$Country)) /
names(sort(-table(mydata$Country)))[1]
var(mydata$SAT) # Variance
sd(mydata$SAT) # Standarddeviation
max(mydata$SAT) # Max value
min(mydata$SAT) # Min value
range(mydata$SAT) # Range
quantile(mydata$SAT)
quantile(mydata$SAT, c(.3,.6,.9))
fivenum(mydata$SAT)# Boxplotelements. From help: "Returns Tukey'sfive number
summary (minimum, lower-hinge, median, upper-hinge, maximum) for the input data ~
boxplot"
length(mydata$SAT)# Num of observations when a variable is specify
length(mydata)# Number of variables when a dataset is specify
which.max(mydata$SAT)# From help: "Determines the location, i.e., index of the (first)
minimum or maximum of a numeric vector"
which.min(mydata$SAT)# From help: "Determines the location, i.e., index of the (first)
minimum or maximum of a numeric vector”
stderr<-function(x) sqrt(var(x)/length(x))
incster<-tapply(incomes, statef, stderr)
```