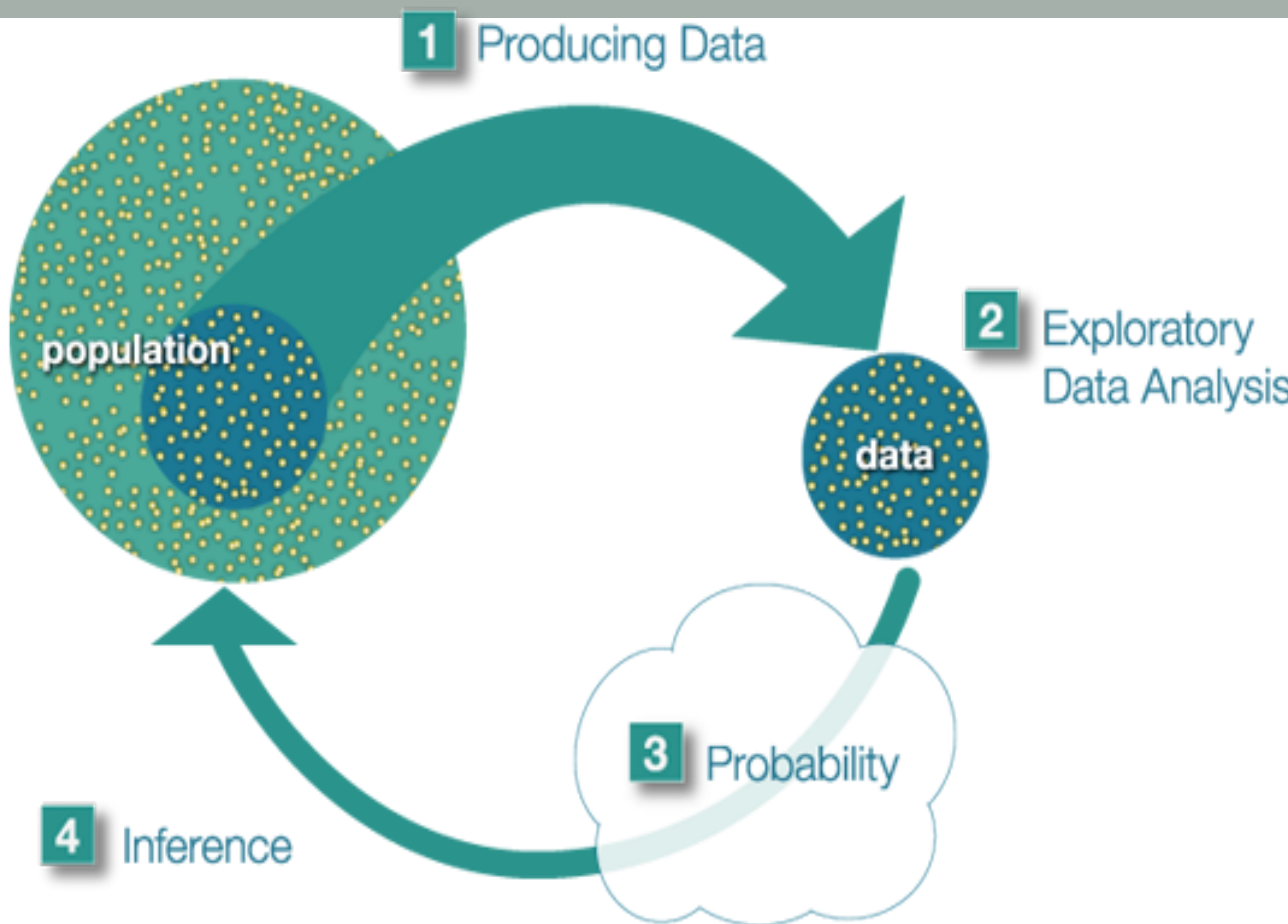
QUANTITATIVE METHODS

The New School
Aaron Hill


1 Producing Data
population
data
2 Exploratory Data Analysis
3 Probability
4 Inference

CENTRAL TENDENCY

Week Two

Mean

- The mean is the point of the distribution around which the variation of the scores is minimized. The mean is closer to all of the values than the other measures of central tendency.
- Least squares principle

Characteristics of the Mean

- **Least squares principle**: the mean is the point in a distribution around which the variation of the scores is minimized

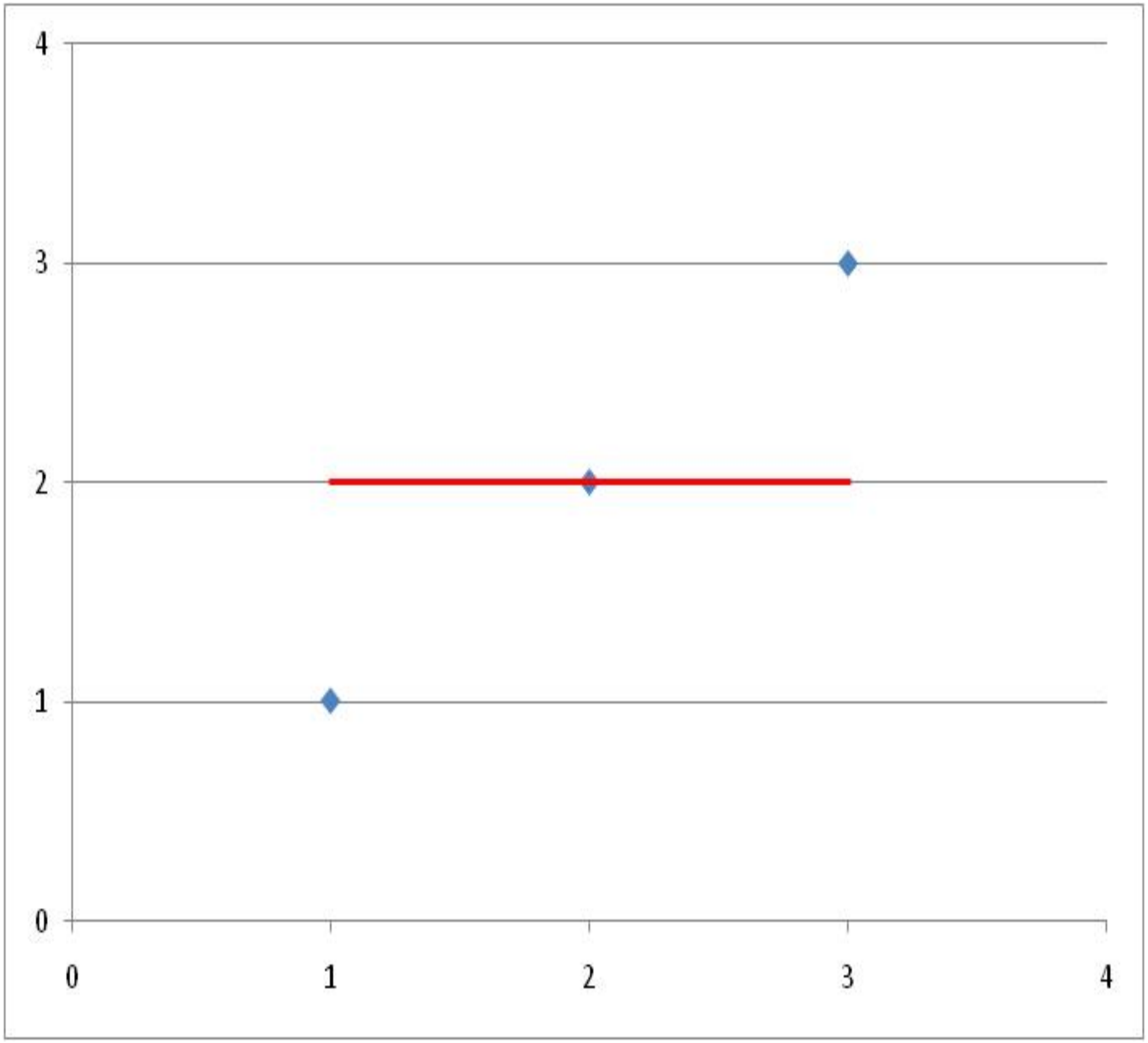
Variable X:

X	Mean	X - Mean
1	2	-1
2	2	0
3	2	1
Mean=2		**Sum=0**

"The average tax cut is $1,586."

White House estimate in 2003, on the then-proposed tax cuts

Characteristics of the Mean



Characteristics of the Mean

- Every score in the distribution affects it

X
1
2
3
794

"The average tax cut is $1,586."

White House estimate in 2003, on the then-proposed tax cuts

The median tax cut is $470.

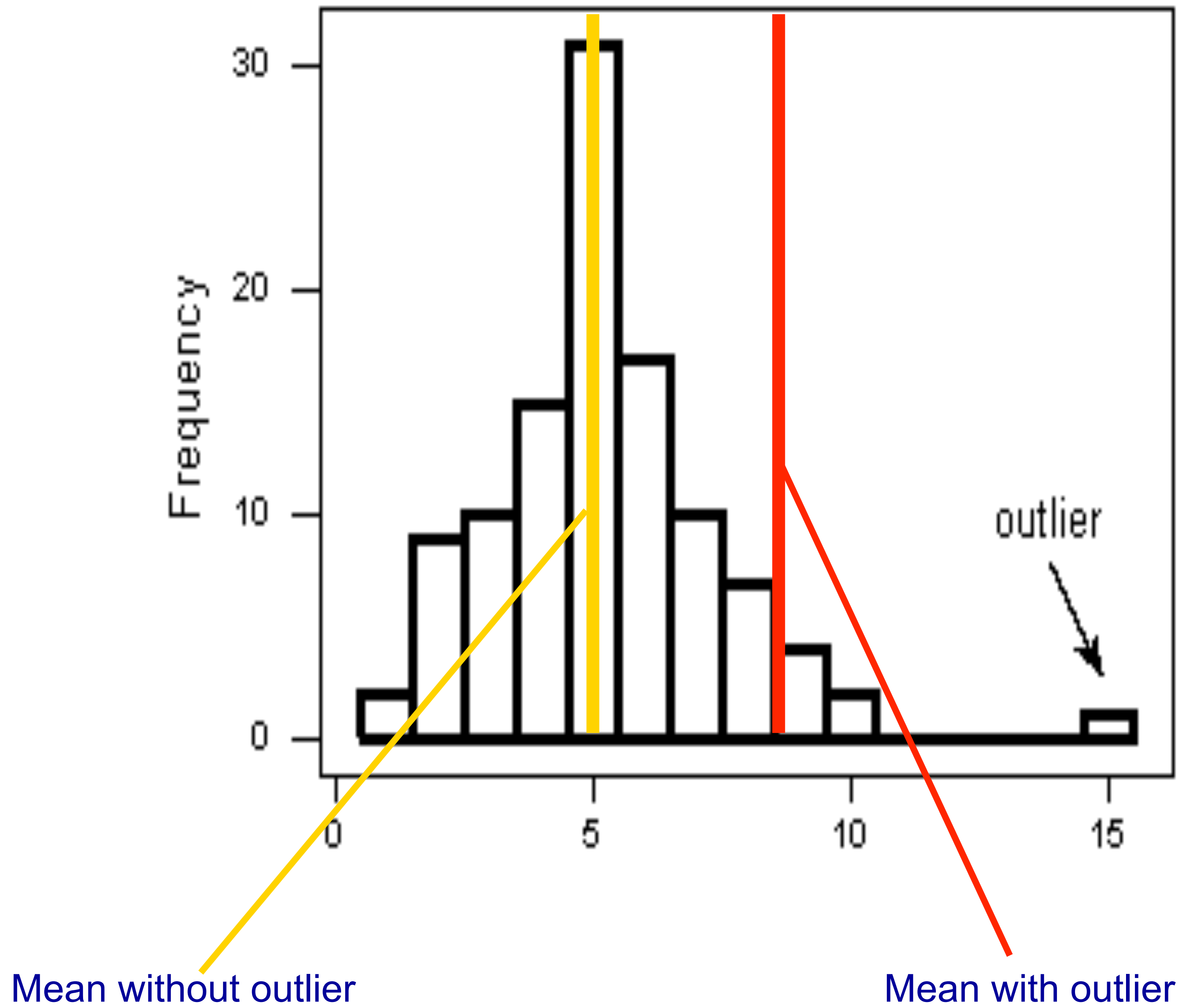
Combined Effect of 2003 Tax Cuts

Income (in thousands)	Percent of Households	Average Tax Change
Less than 10	23.7	-$8
10-20	16.6	-$307
20-30	13.3	-$638
30-40	9.7	-$825
40-50	7.6	-$1,012
50-75	13.0	-$1,403
75-100	6.8	-$2,543
100-200	6.6	-$3,710
200-500	1.6	-$7,173
500-1,000	0.3	-$22,485
More than 1,000	0.1	-$112,925

Source: Tax Policy Center table T03-0123

Some Flaws of Averages

- Outliers have strong influence on the mean
- Aggregated data is sometimes misleading

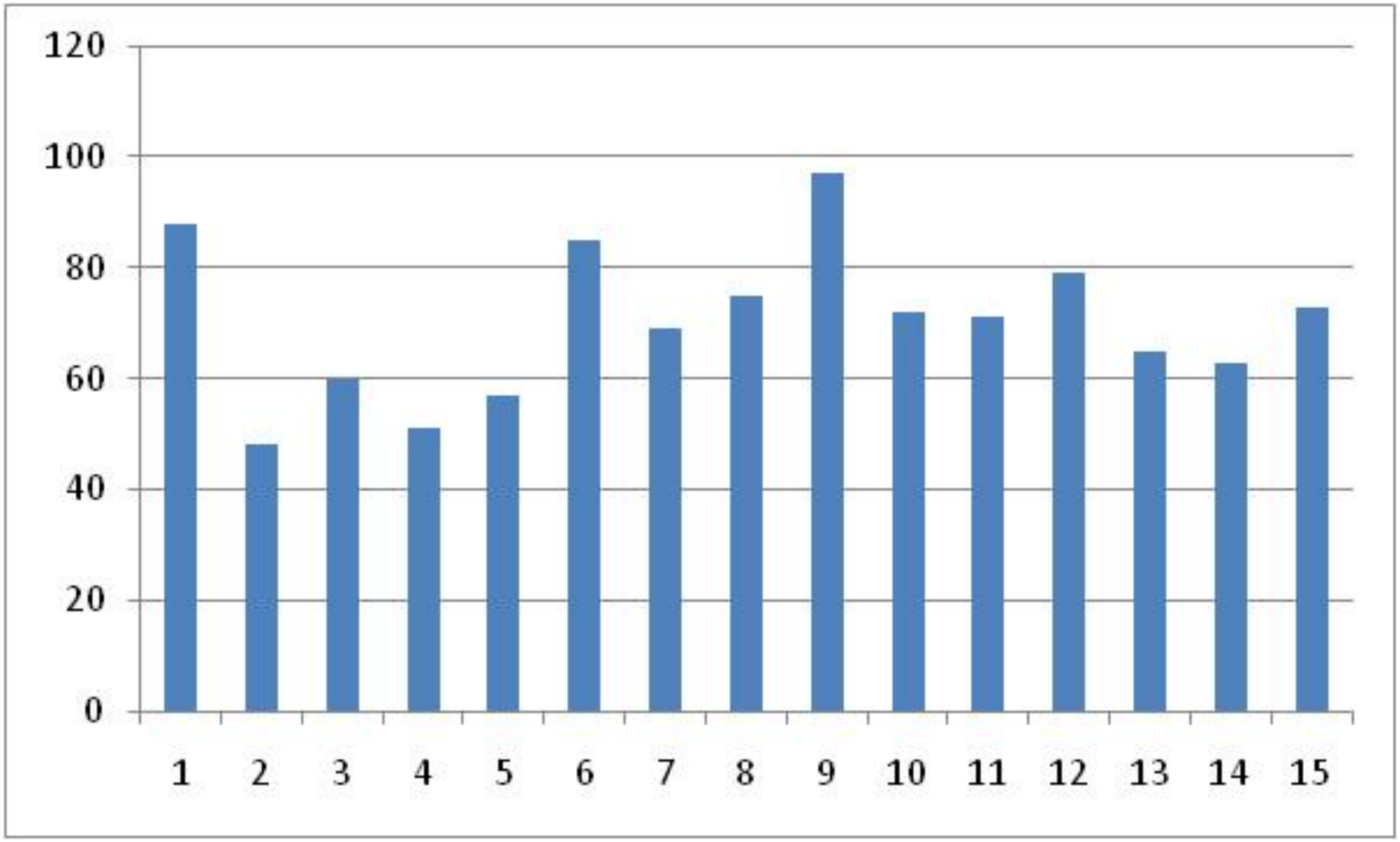

30
20
10
0
Frequency
0
5
10
15
outlier
Mean without outlier
Mean with outlier

DISTRIBUTIONS AND DISPERSION

Exam Scores

- Exam Scores
- 88, 48, 60, 51, 57, 85, 69, 75, 97, 72, 71, 79, 65, 63, 73

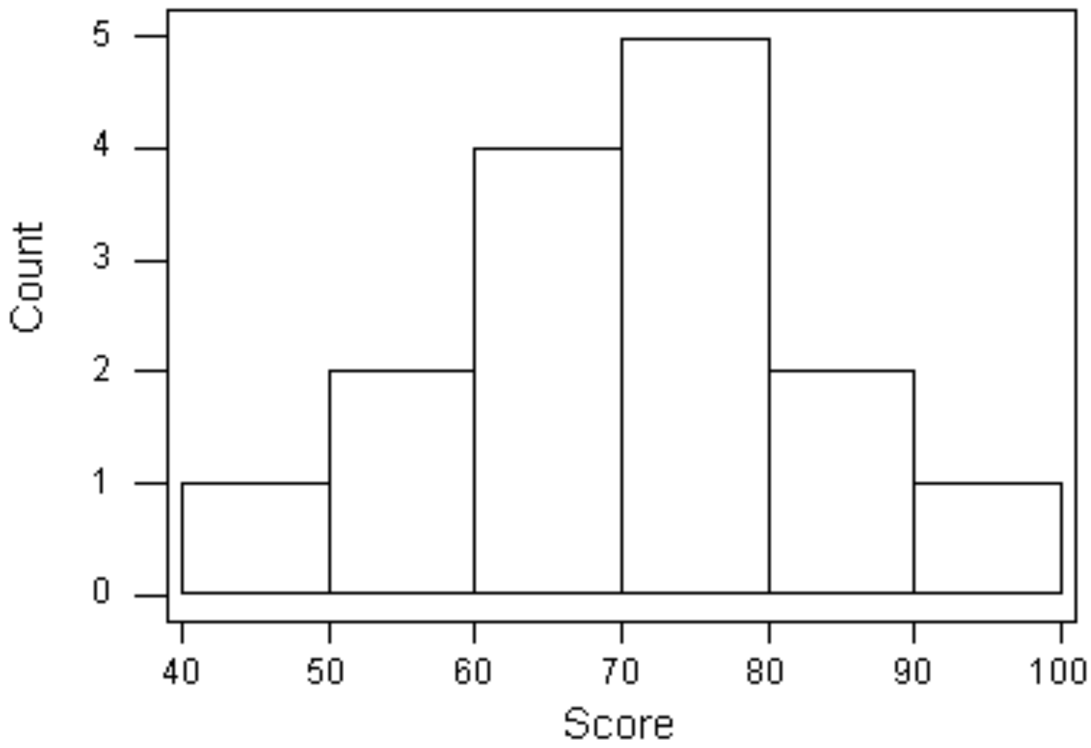
Bar Chart of Exam Scores



Histogram: A Better Graph for Interval-Ratio

Exam Grades

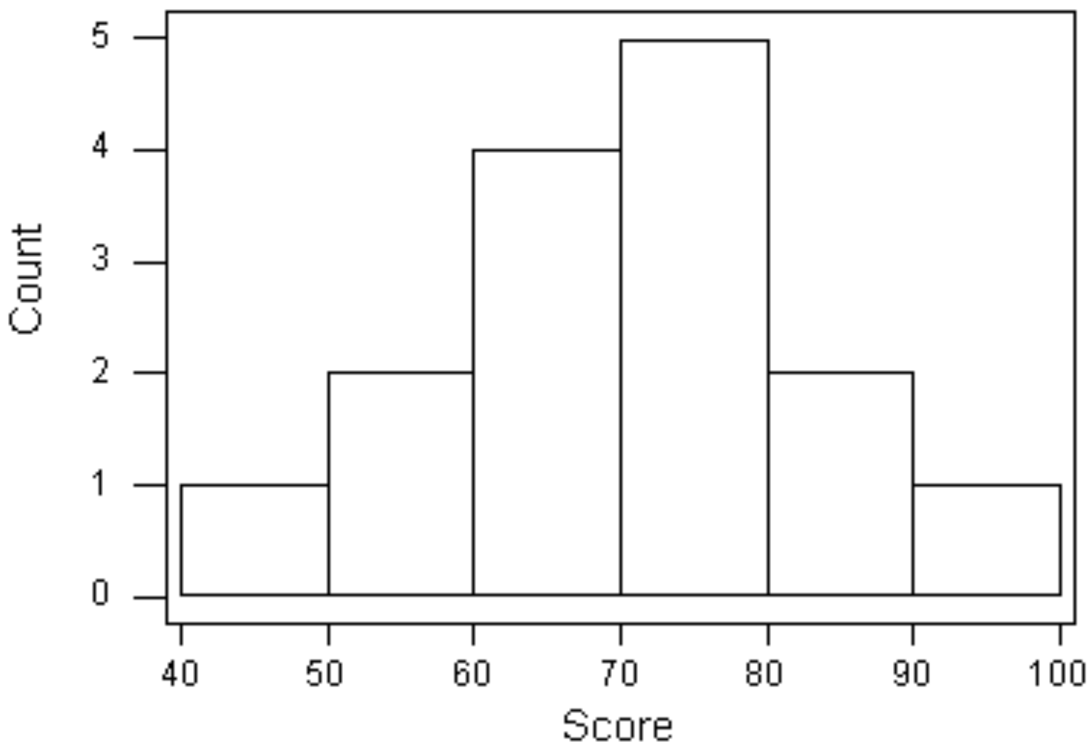
Score	**Count**
[40-50)	1
[50-60)	2
[60-70)	4
[70-80)	5
[80-90)	2
[90-100]	1



Histogram

- a graphical display of the distribution of a quantitative variable. It plots the number (count) of observations that fall in intervals of values.

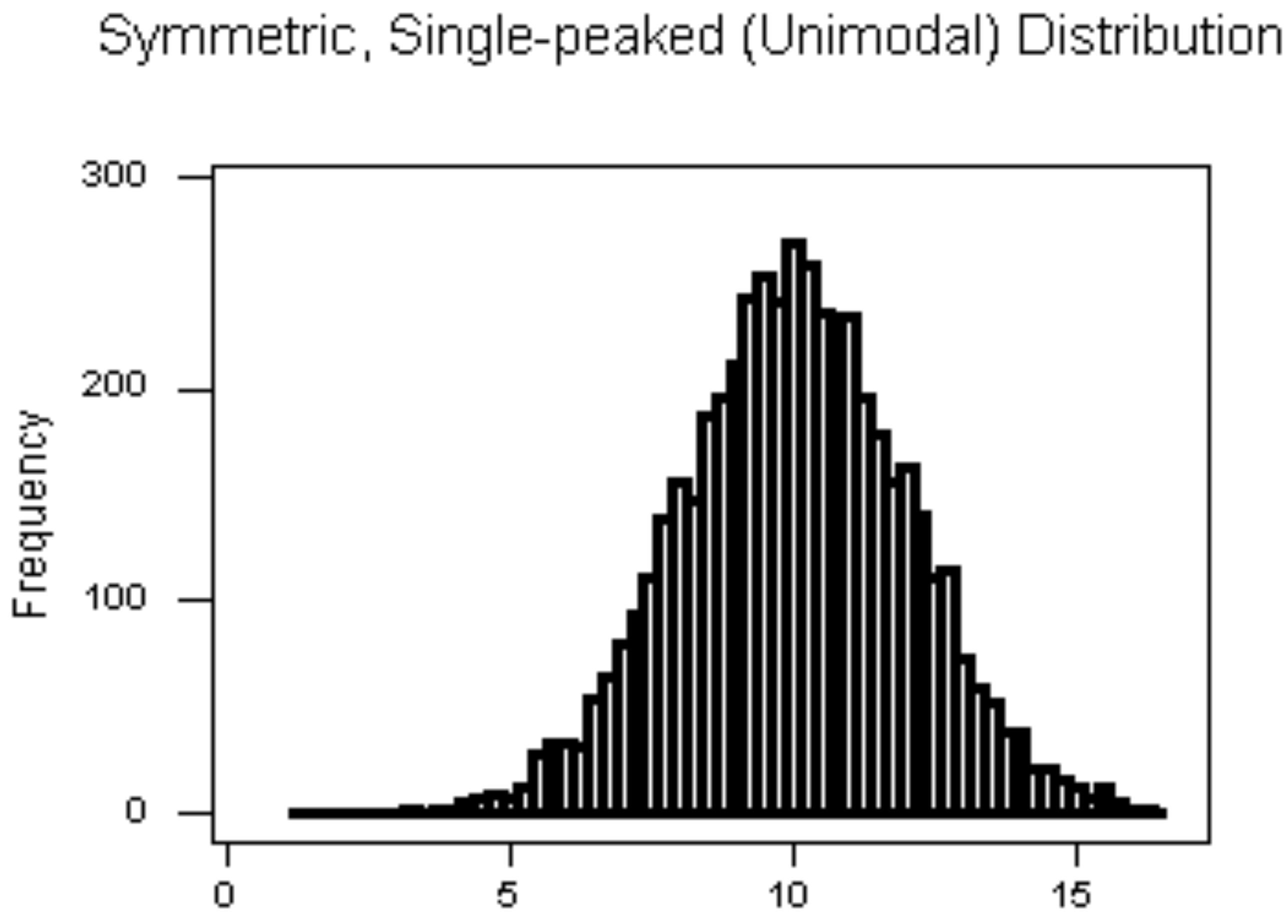
Histogram: A Better Graph for Interval-Ratio



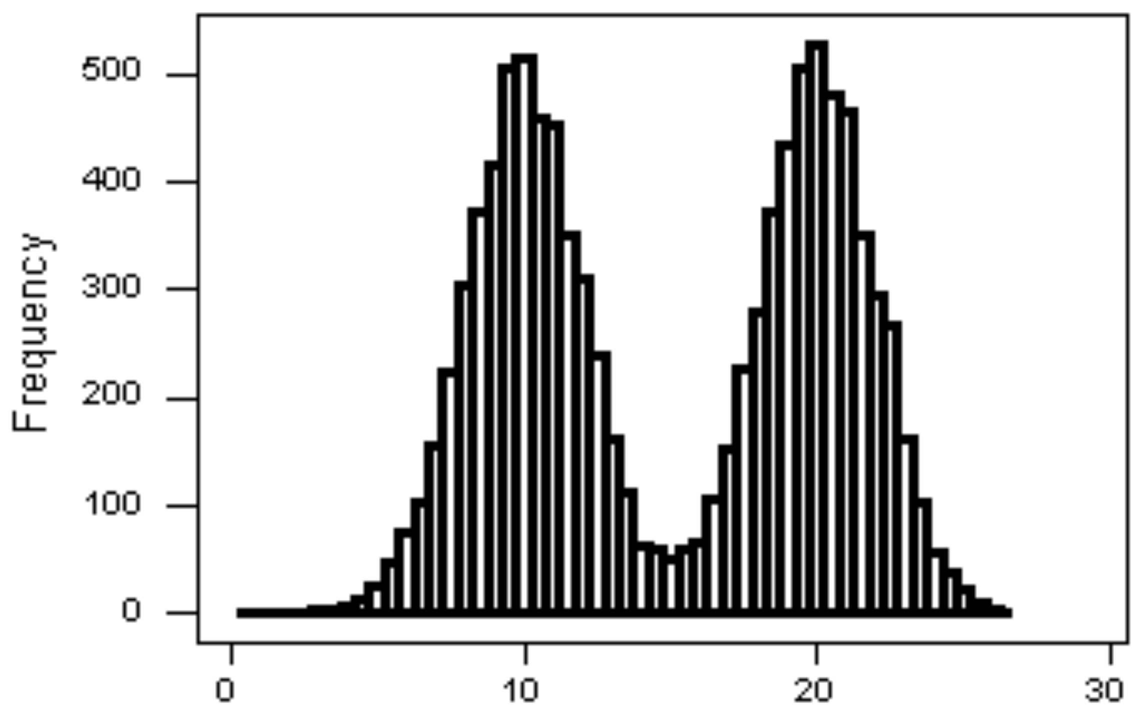
- Shape
- Center
- Spread
- Outliers

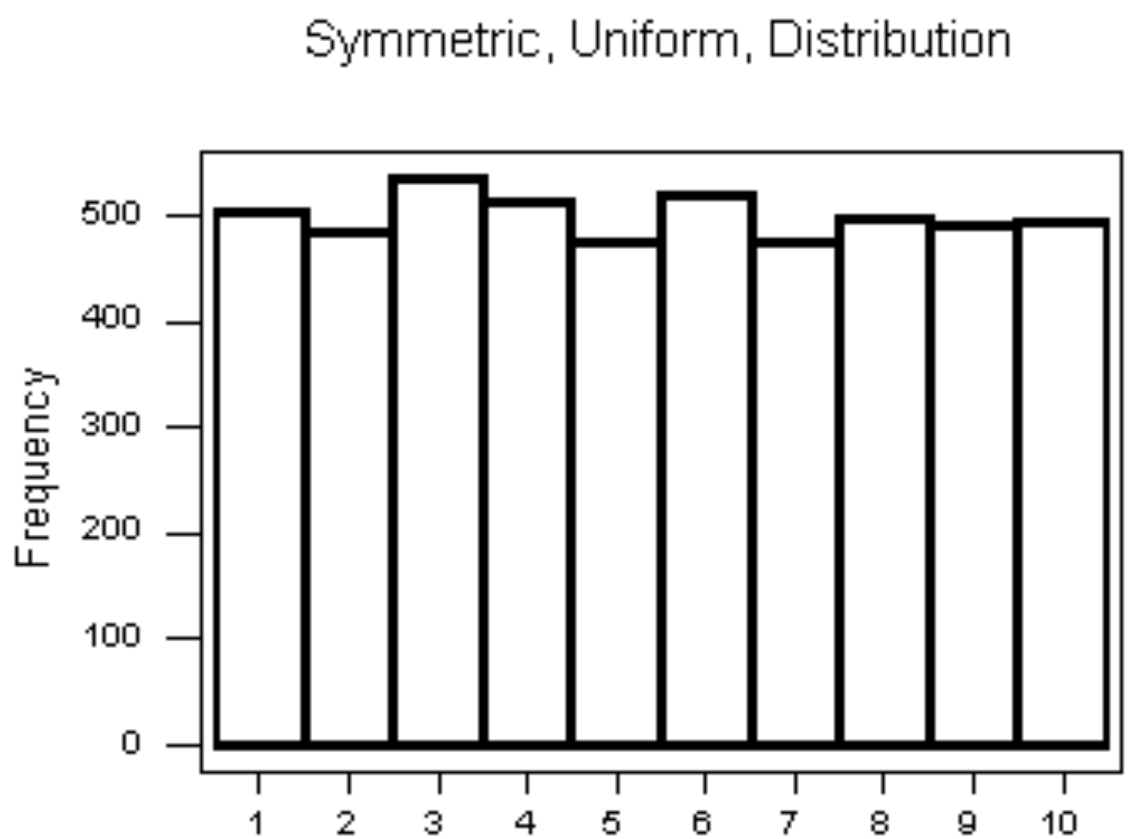
overall pattern

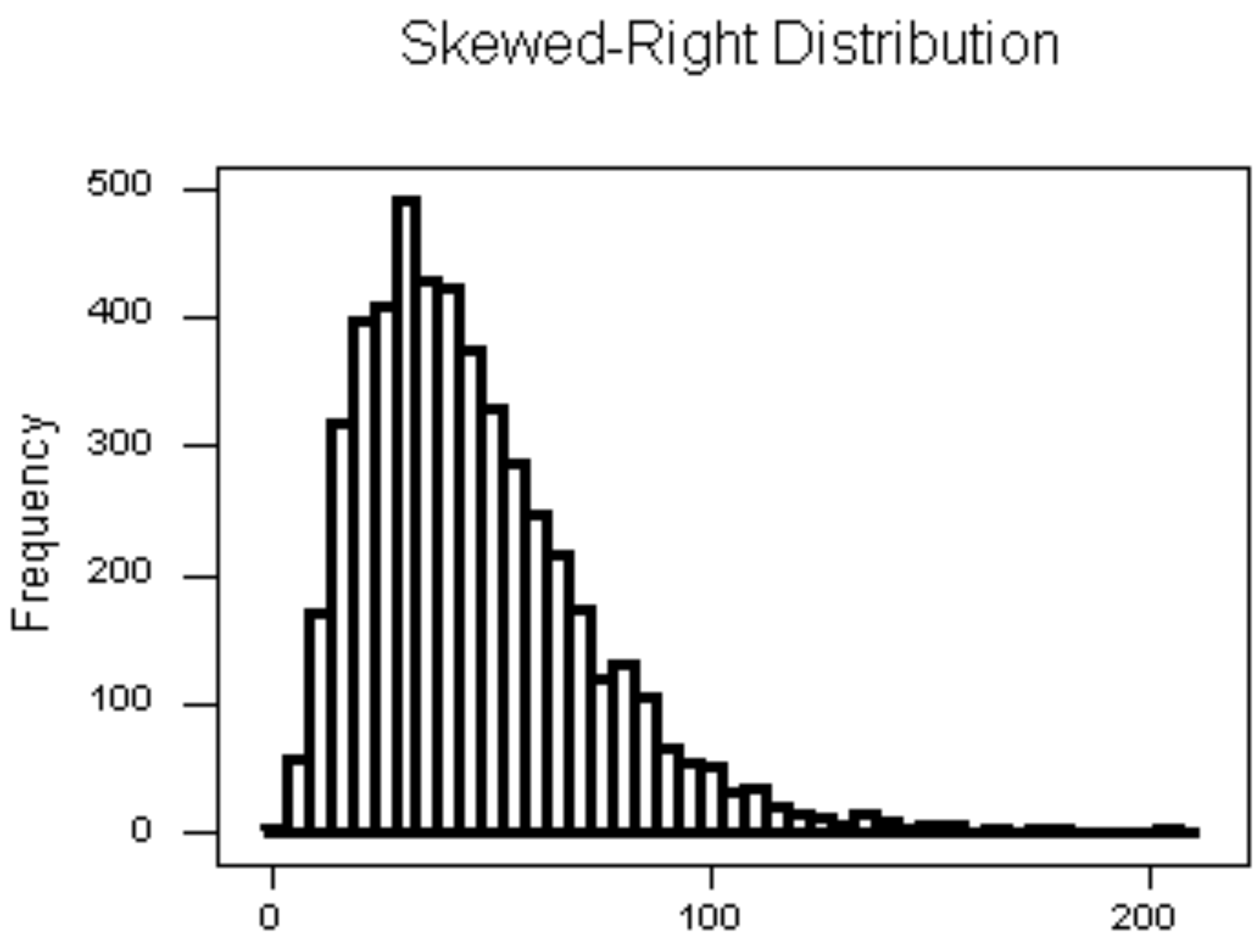
deviations from the pattern


Symmetric, Single-peaked (Unimodal) Distribution
300
200
100
0
Frequency
0
5
10
15

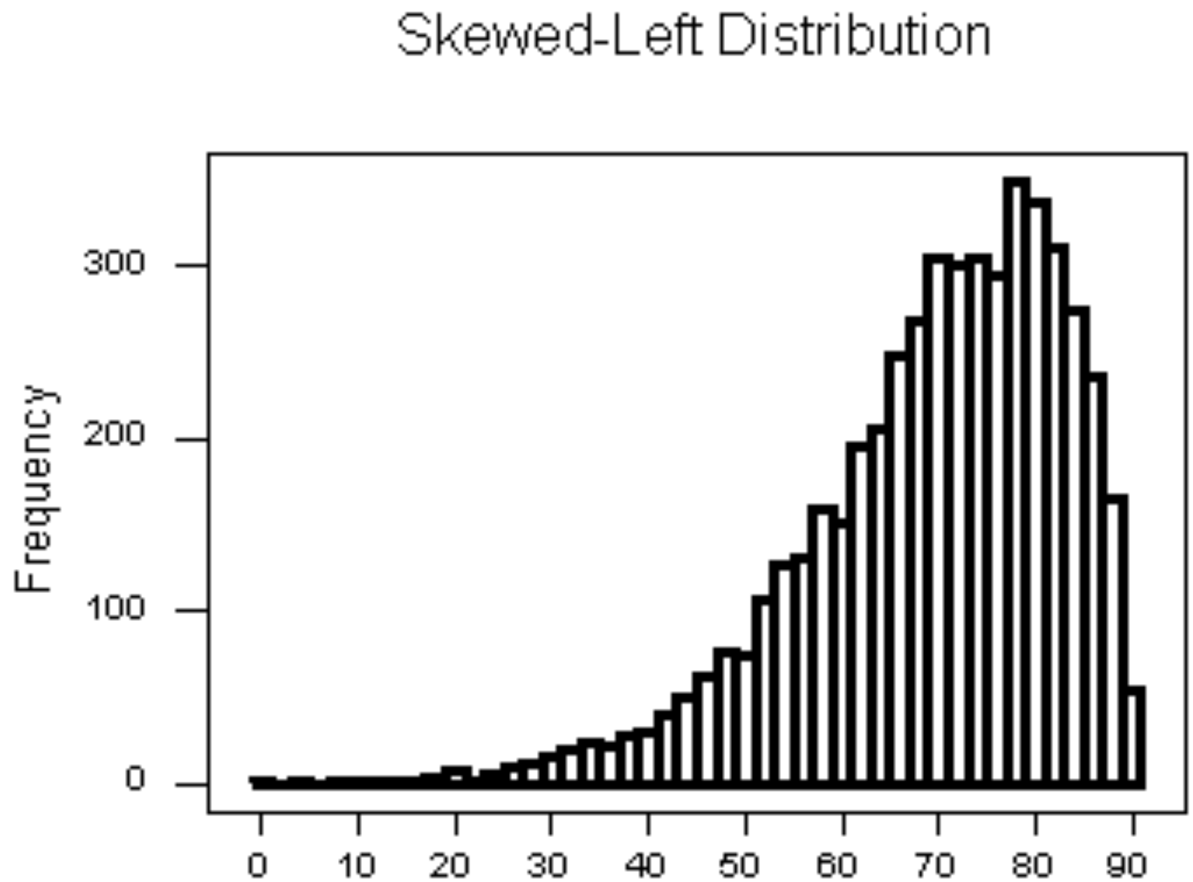
Symmetric, Double-peaked (Bimodal) Distribution

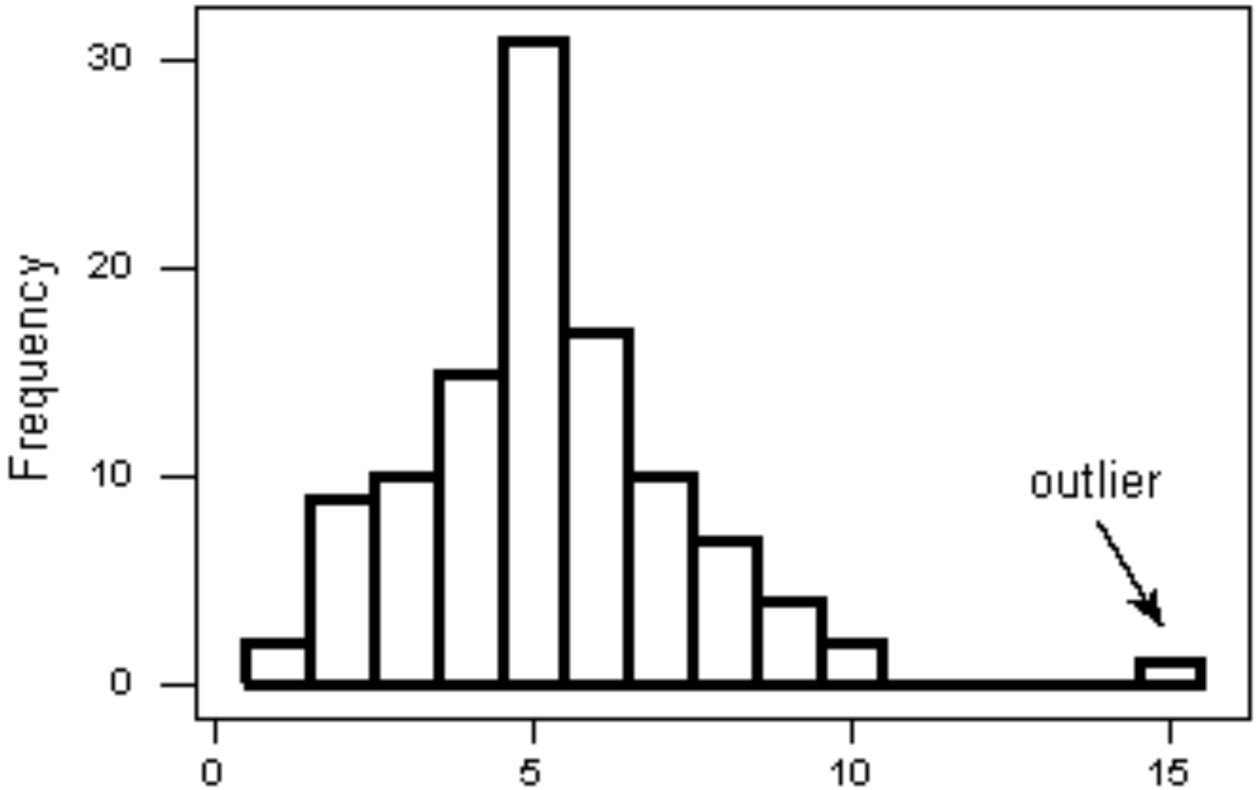



Symmetric, Uniform, Distribution
Frequency
0
100
200
300
400
500
1
2
3
4
5
6
7
8
9
10



Skewed-Right Distribution
Frequency
500
400
300
200
100
0
0
100
200


Skewed-Left Distribution
Frequency
300
200
100
0
0
10
20
30
40
50
60
70
80
90



Frequency
30
20
10
0
0
5
10
15
outlier

MEASURES OF DISPERSION

What does the mean tell us?

- The mean age is 30.

What does the mean tell us?

- The mean income is $60,000

Measures of Dispersion

- Tell us more about the distribution.
- Specifically, how consistently or inconsistently the data disperse around the mean.

Measures of Dispersion

- Would you walk across a river that is, on average, 3 feet deep?

The infamous tax cuts

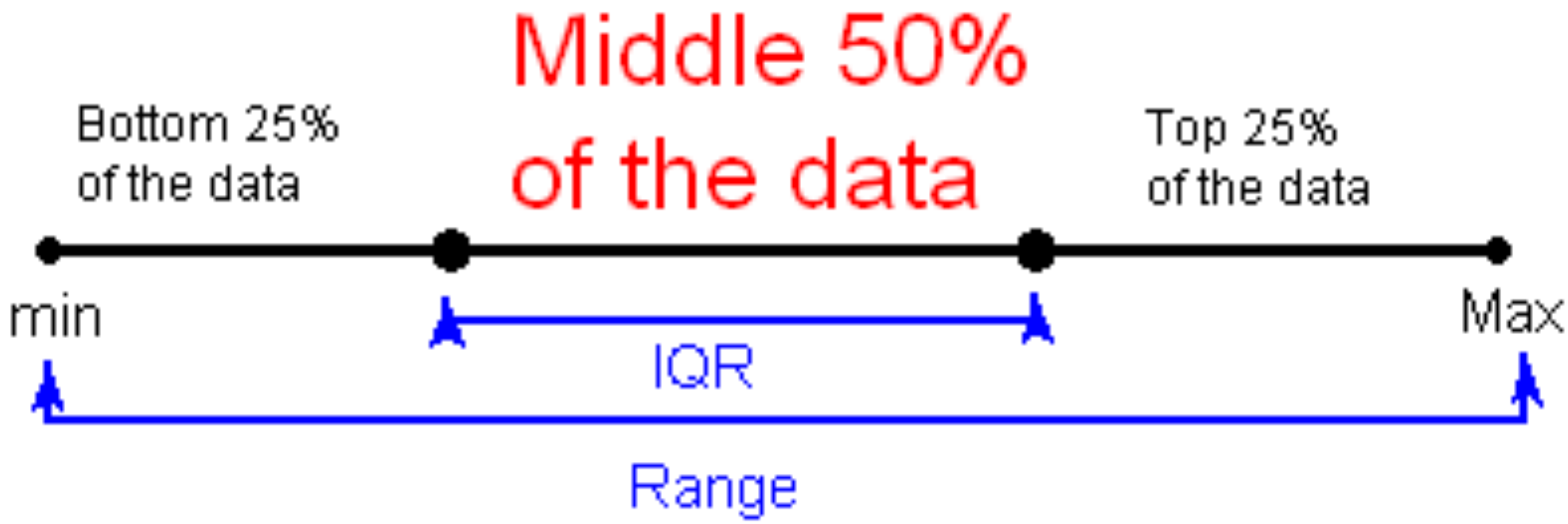
- The Bush tax cuts were justified with the following rationale:
  - The average family would save $1,586 in taxes.
- Technically correct, but misleading. The ***typical*** family would save less than $650.

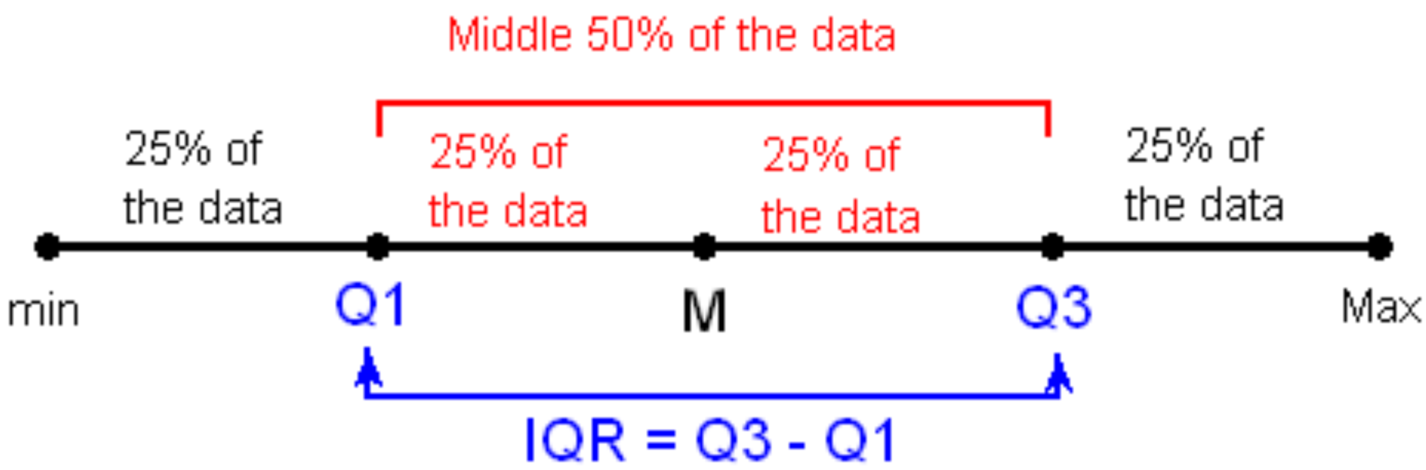
How do the distributions differ?

- 50 50 50 50 50 (mean=50)
- 10 20 50 80 90 (mean=50)

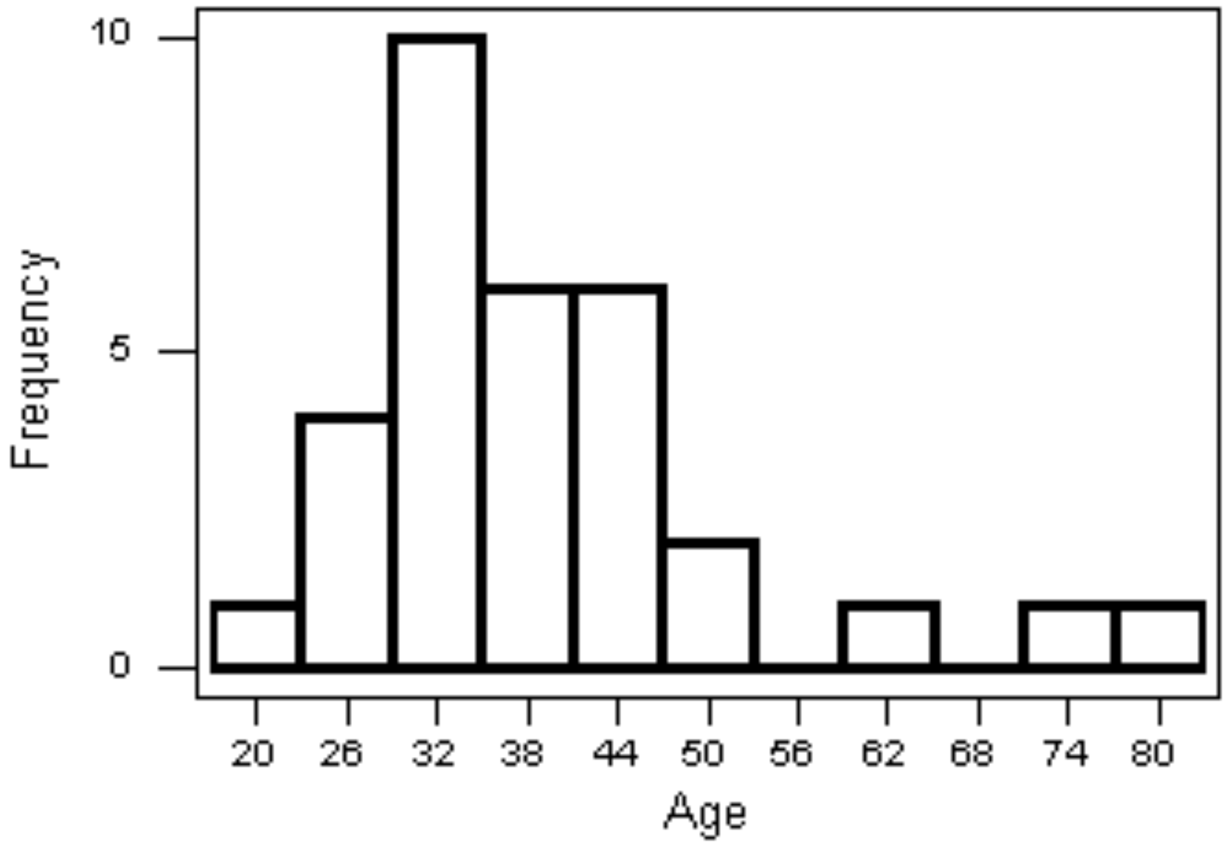
Measures of Dispersion

- Min & Max – the minimum and maximum values
- Range – the minimum value subtracted from the maximum value
- Interquartile Range – the first quartile subtracted from the third quartile
- Standard Deviation – a measure of deviation from the mean


Middle 50%
of the data
Bottom 25%
of the data
Top 25%
of the data
min
Max
IQR
Range


Middle 50% of the data
25% of the data
25% of the data
25% of the data
25% of the data
min
Q1
M
Q3
Max
IQR = Q3 - Q1

Best Actress Oscar Winners 1970-2001

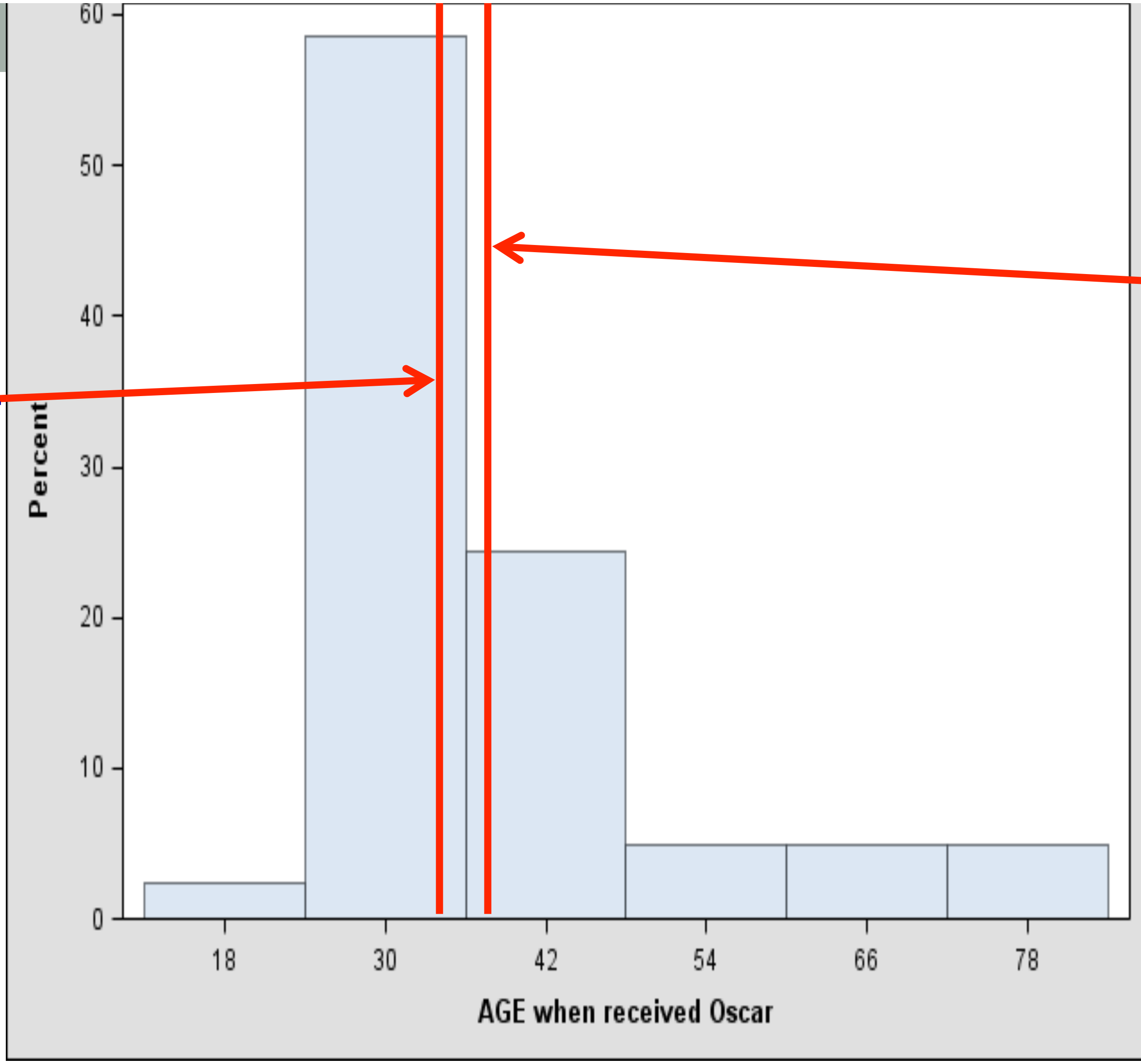


Academy Award for Best Actress

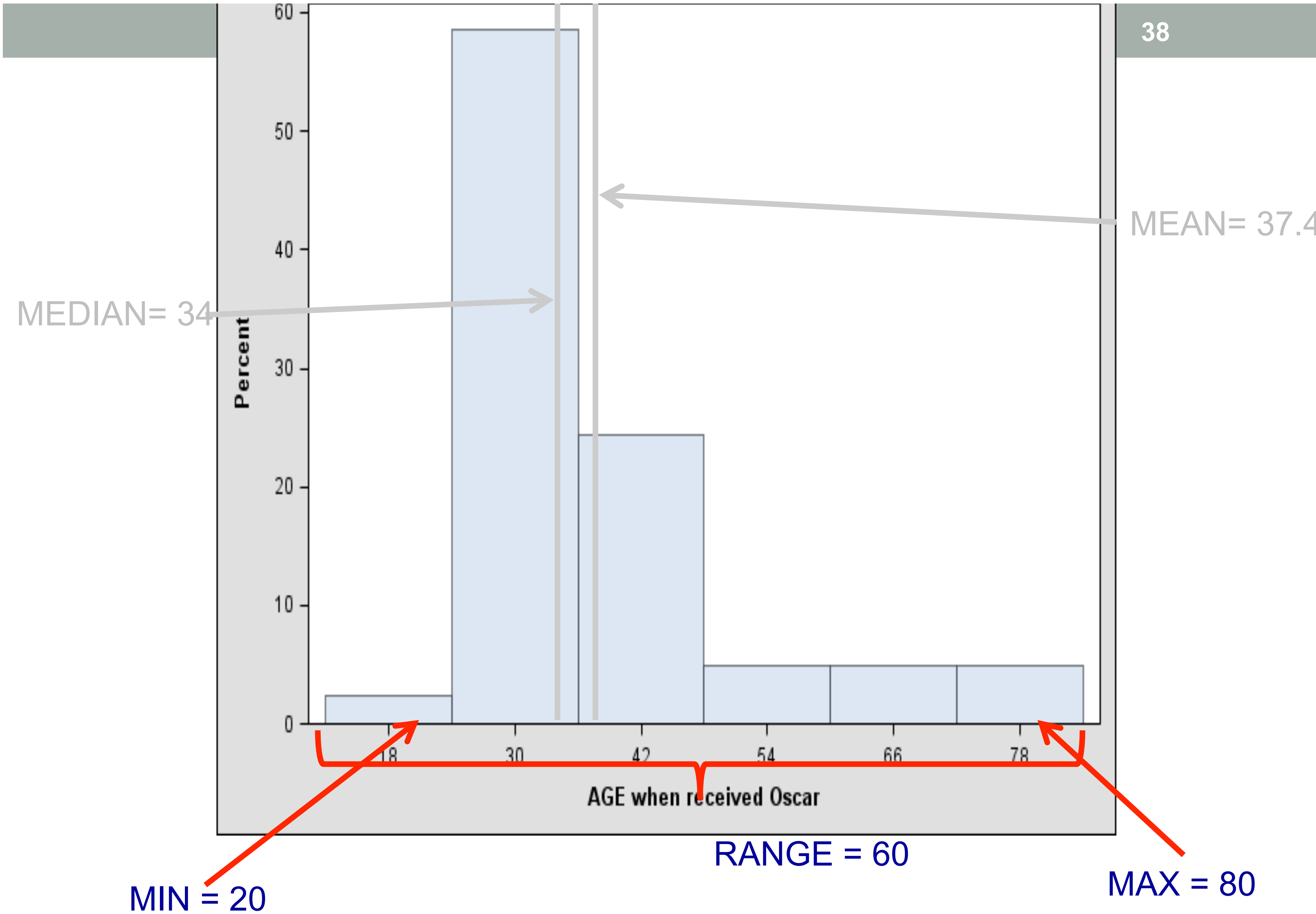
1970 – 2010

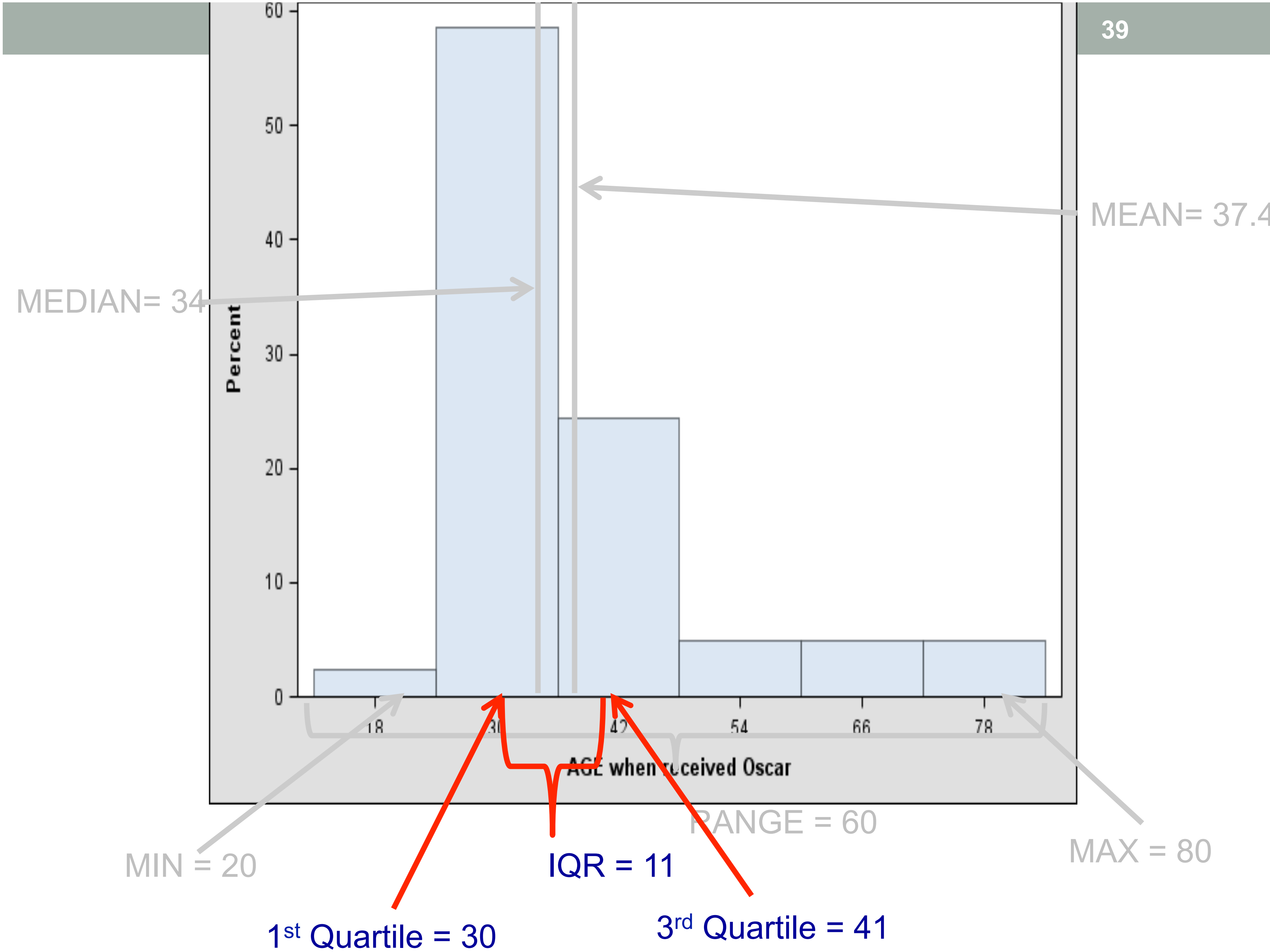
- Average age: 37.4

MEDIAN= 34

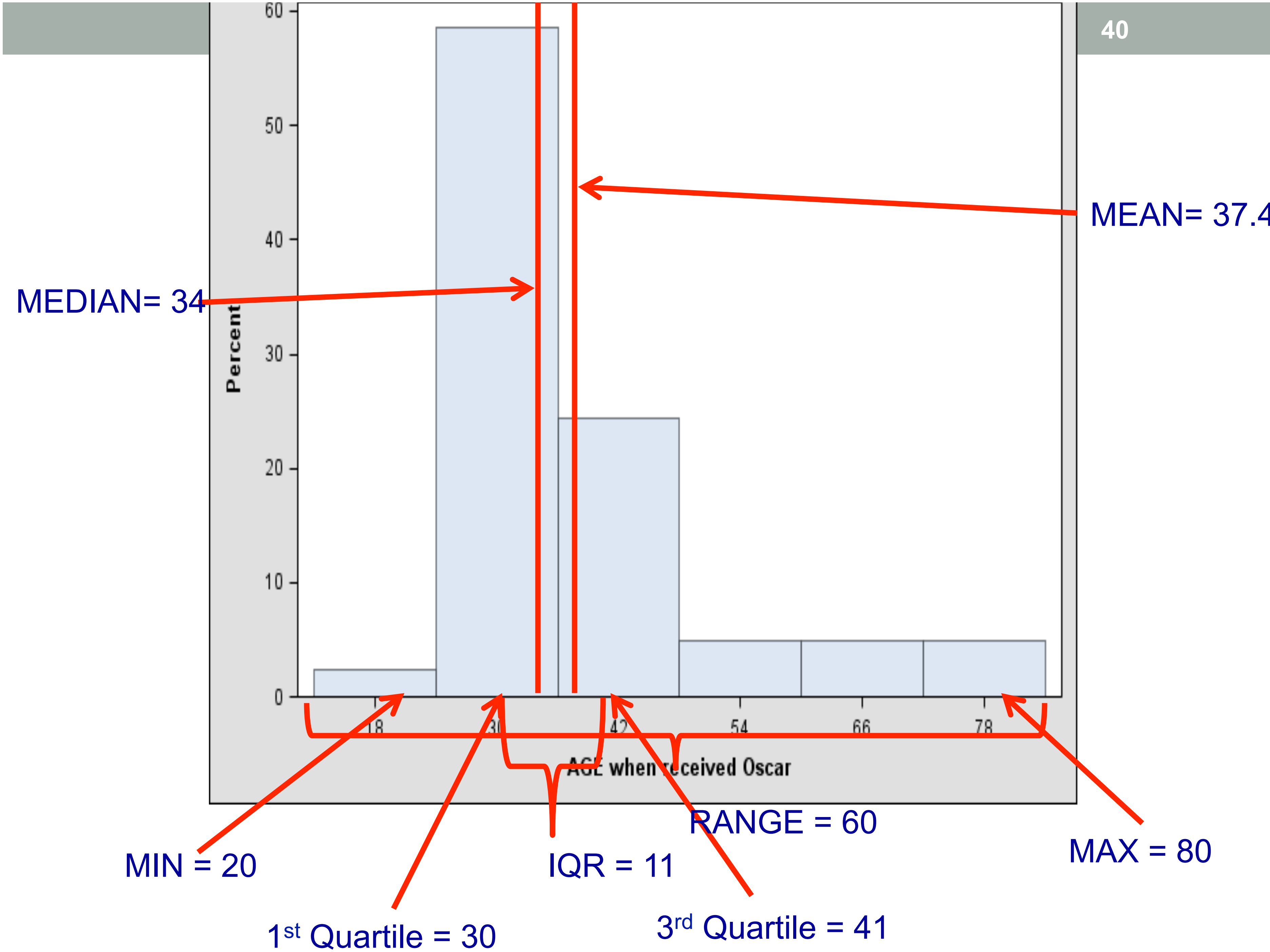


MEAN= 37.4

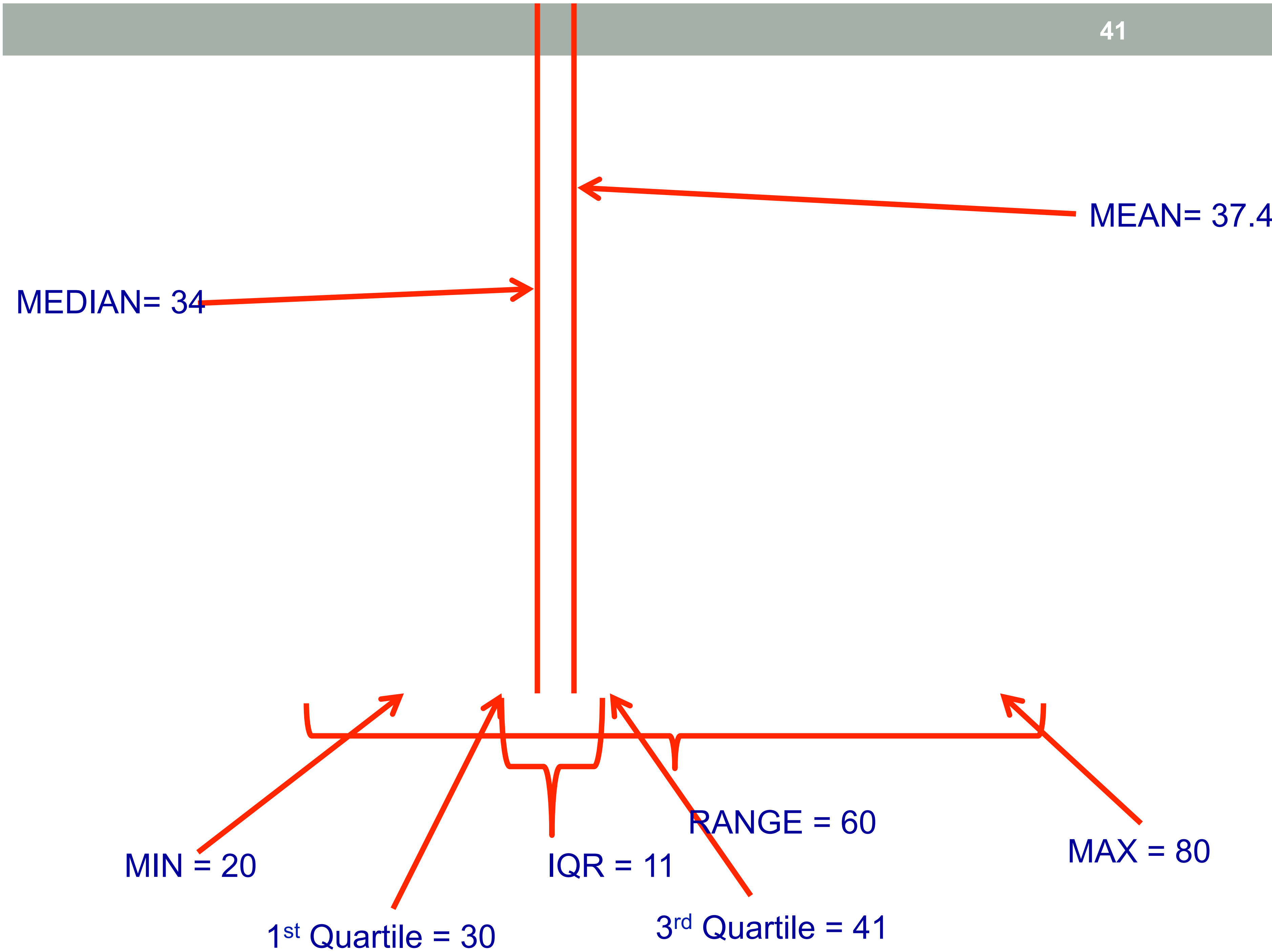

MEAN= 37.4
MEDIAN= 34
Percent
60
50
40
30
20
10
0
18
30
42
54
66
78
AGE when received Oscar
RANGE = 60
MIN = 20
MAX = 80



MEAN= 37.4
MEDIAN= 34
Percent
60
50
40
30
20
10
0
18
54
66
78
AGE when received Oscar
RANGE = 60
MIN = 20
MAX = 80
IQR = 11
1st Quartile = 30
3rd Quartile = 41



60
50
40
30
20
10
0
Percent
18
30
42
54
66
78
AGE when received Oscar
MEAN= 37.4
MEDIAN= 34
MIN = 20
1st Quartile = 30
IQR = 11
RANGE = 60
3rd Quartile = 41
MAX = 80


MEAN= 37.4
MEDIAN= 34
RANGE = 60
MAX = 80
MIN = 20
IQR = 11
1st Quartile = 30
3rd Quartile = 41

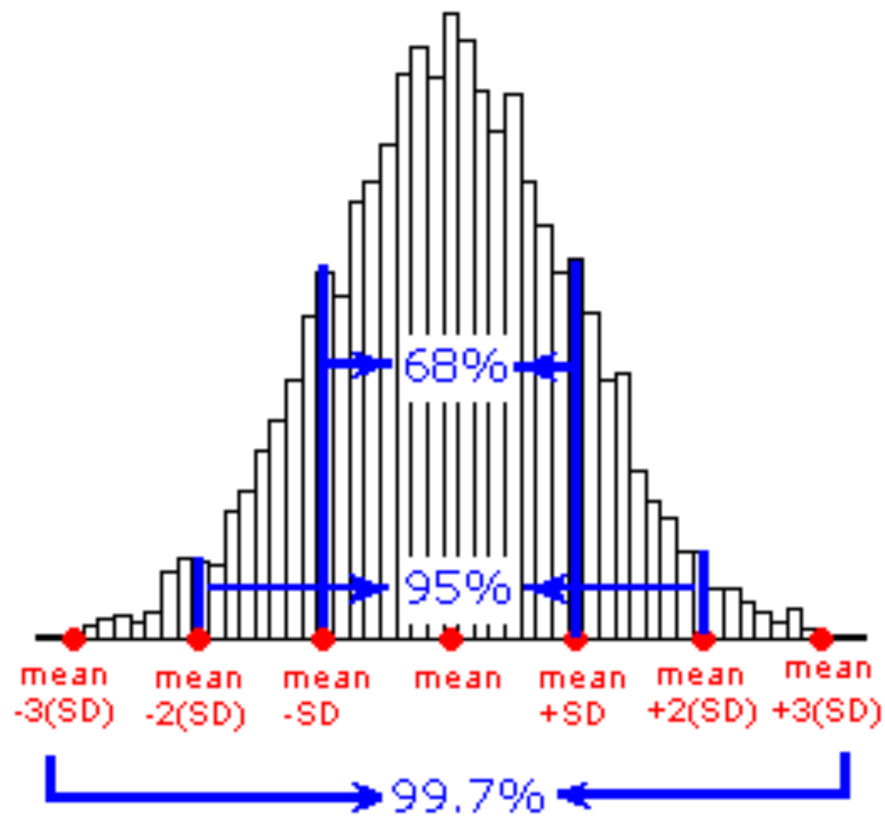
Standard Deviation

- Quantifies the spread of a distribution by measuring how far the observations are from their mean. The standard deviation gives the average (or typical distance) between a data point and the mean.

$$\bar{x}$$

Standard Deviation

- The Standard Deviation Rule:
  - Approximately 68% of the observations fall within 1 standard deviation of the mean
  - Approximately 95% of the observations fall within 2 standard deviations of the mean
  - Approximately 99.7% (or virtually all) of the observations fall within 3 standard deviations of the mean



Calculating Standard Deviation

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

$$\frac{\sum_{i=1}^{n} x_i}{n}$$

$$\text{Mean}\,(\bar{X}) = \frac{\sum_{i=1}^{n} x_i}{n}$$

The mean: A Perfect Fit

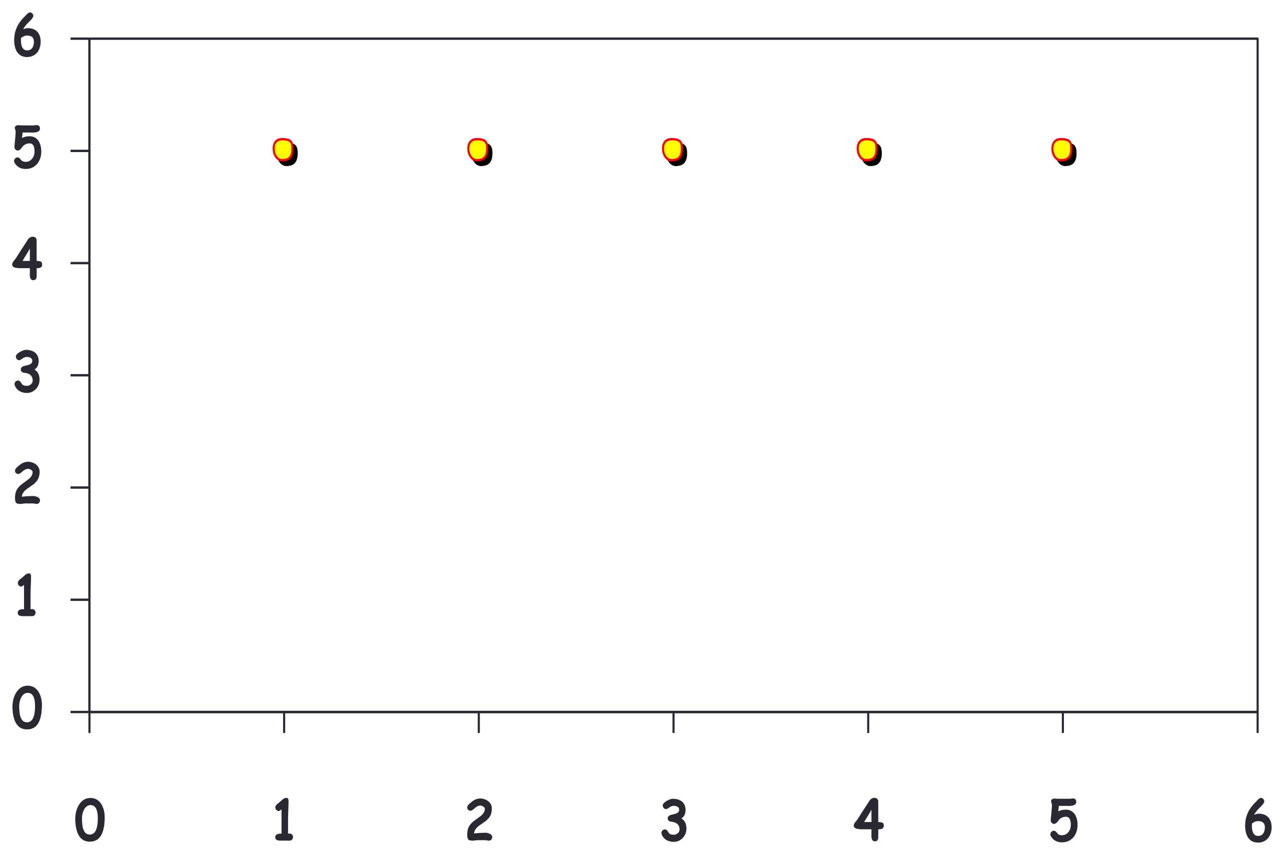
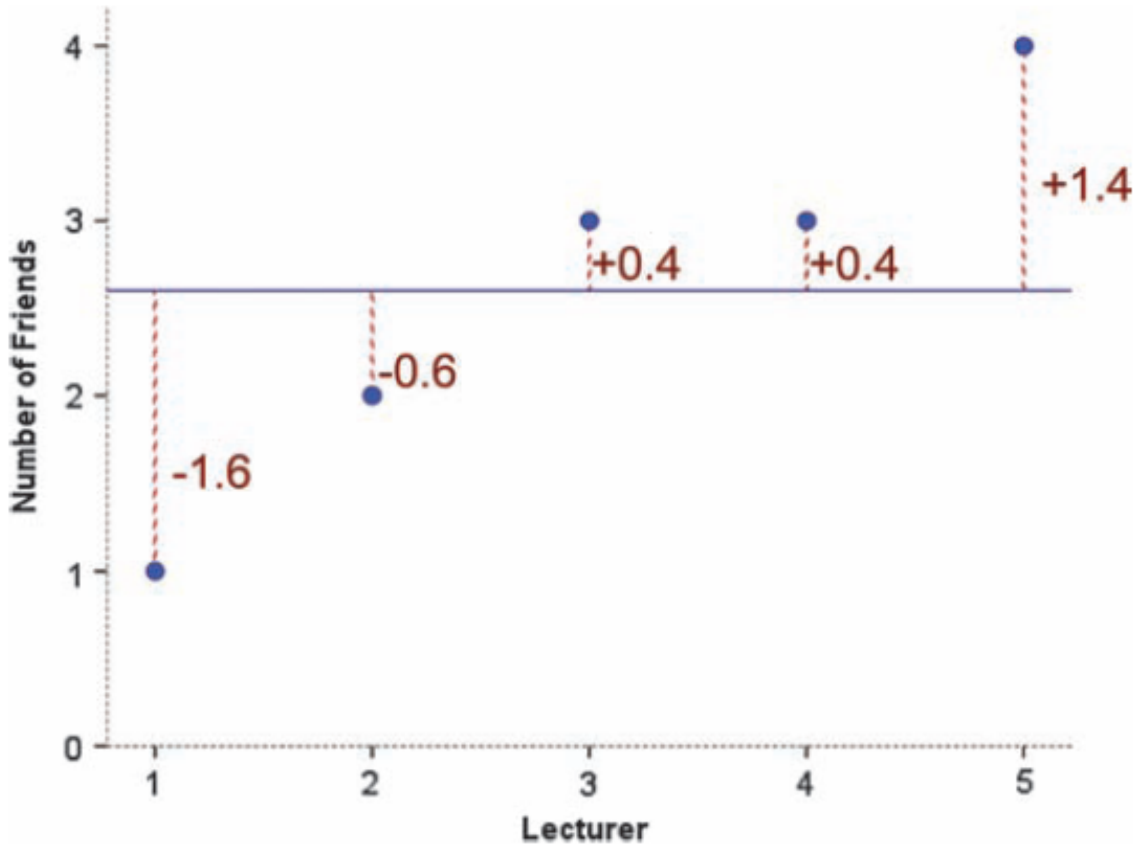


FIGURE 2.4
Graph showing the difference between the observed number of friends that each statistics lecturer had, and the mean number of friends



Xbar = 2.6

Calculating 'Error'

- A deviation is the difference between the mean and an actual data point.
- Deviations can be calculated by taking each score and subtracting the mean from it:

$$\text{Deviation} = x_i - \bar{x}$$

Use the Total Error?

- We could just take the error between the mean and the data and add them.

Score	Mean	Deviation
1	2.6	-1.6
2	2.6	-0.6
3	2.6	0.4
3	2.6	0.4
4	2.6	1.4
	Total =	0

$$\boxed{\sum(X - \bar{X}) = 0}$$

Sum of Squared Errors

- We could add the deviations to find out the total error.
- Deviations cancel out because some are positive and others negative.
- Therefore, we square each deviation.
- If we add these squared deviations we get the Sum of Squared Errors (*SS*).

Score	Mean	Deviation	Squared Deviation
1	2.6	-1.6	2.56
2	2.6	-0.6	0.36
3	2.6	0.4	0.16
3	2.6	0.4	0.16
4	2.6	1.4	1.96
		Total	5.20

$$SS = \sum (X - \bar{X})^2 = 5.20$$

Variance

- The sum of squares is a good measure of overall variability, but is dependent on the number of scores.
- We calculate the average variability by dividing by the number of scores (*n*).
- This value is called the variance (s^2).

$$\text{variance } (s^2) = \frac{\text{SS}}{N-1} = \frac{\sum(x_i - \bar{x})^2}{N-1} = \frac{5.20}{4} = 1.3$$

Standard Deviation

- The variance has one problem: it is measured in units squared.
- This isn't a very meaningful metric so we take the square root value.
- This is the Standard Deviation (s).

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}} = \sqrt{\frac{5.20}{4}} = 1.14$$

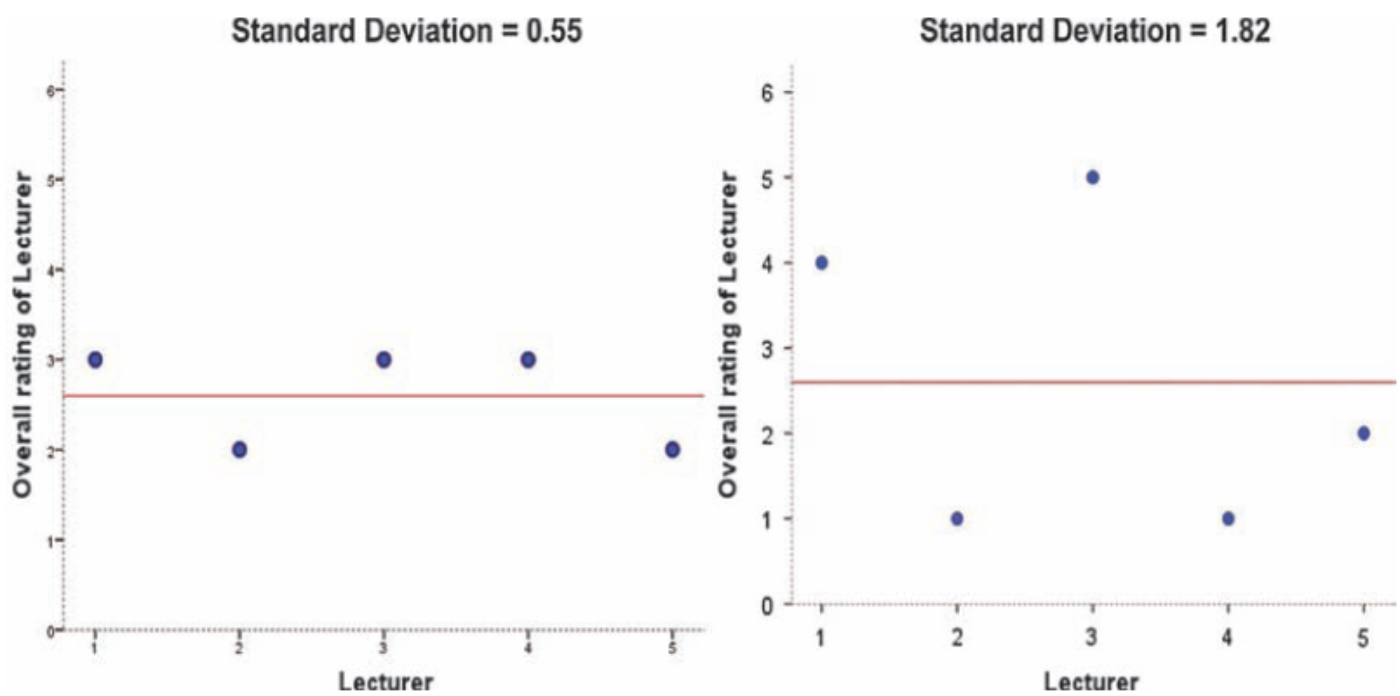
Important Things to Remember

- The Sum of Squares, Variance, and Standard Deviation represent the same thing:
  - The ‘Fit’ of the mean to the data
  - The variability in the data
  - How well the mean represents the observed data
  - Error

	A	B	C
1		**Evaluation Scores**	
2		Course 1	Course 2
8	Mean	2.6	2.6
9	Median	3	2
10	Mode	3	1
11	N	5	5
12	StDev	0.55	1.82

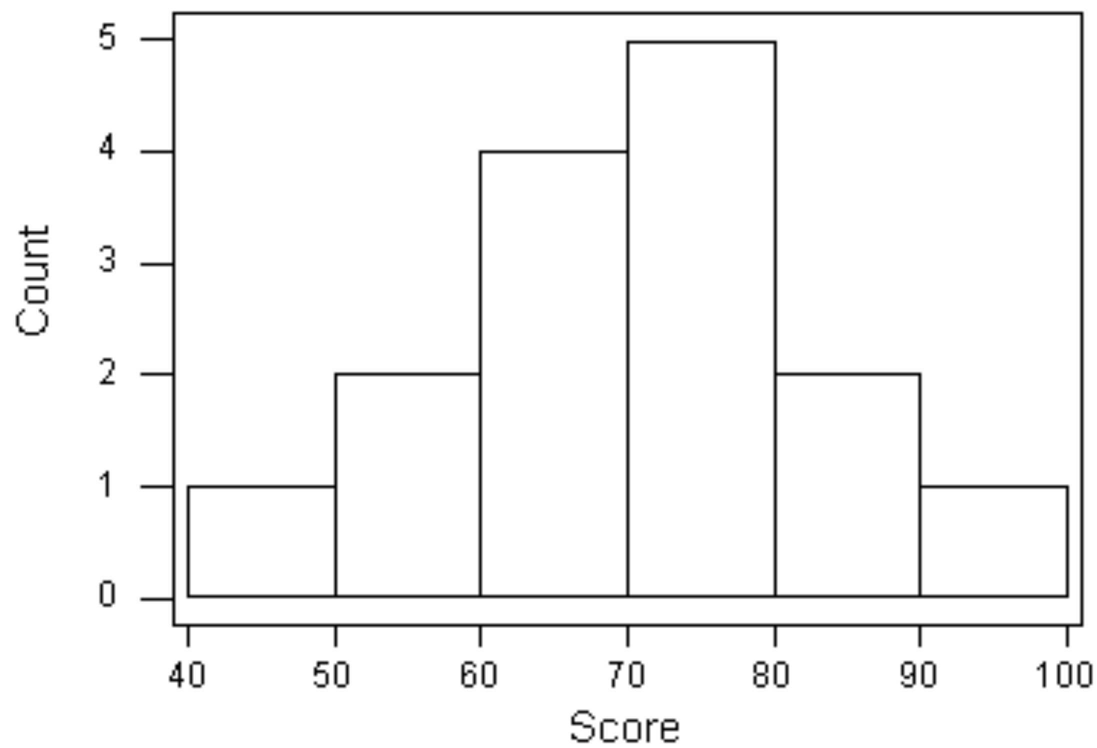
Same Mean, Different SD

FIGURE 2.5
Graphs illustrating data that have the same mean but different standard deviations



Exam Scores

- 88, 48, 60, 51, 57, 85, 69, 75, 97, 72, 71, 79, 65, 63, 73



Which ambulance service do you prefer?

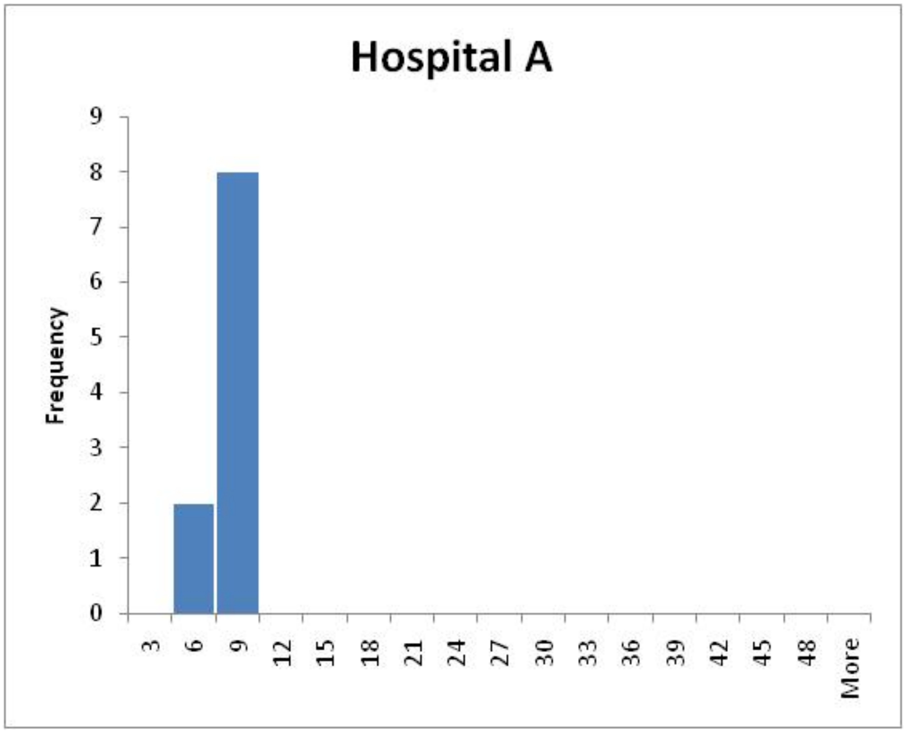
- Variable: number of minutes between 911 call and arrival of ambulance

Ambulance Service	A	B
Mean	7.1	7.1

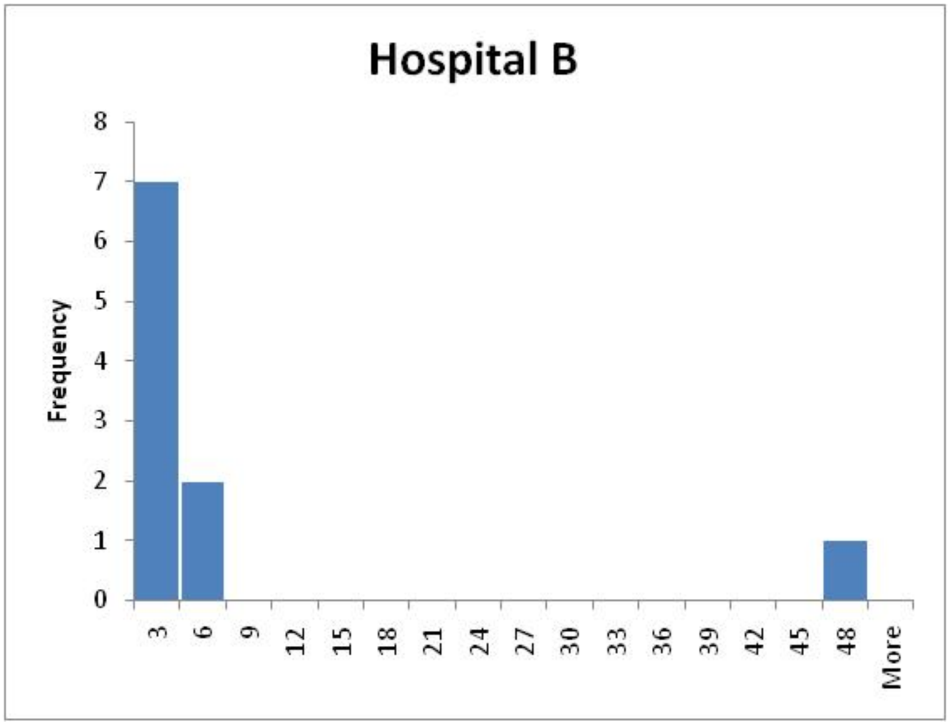
Ambulance Response Time at Two Hospitals

A	**B**
7	2
8	3
9	3
7	1
7	47
7	4
7	5
6	2
6	3
7	1

Histograms



Hospital A
Frequency
9
8
7
6
5
4
3
2
1
0
3
6
9
12
15
18
21
24
27
30
33
36
39
42
45
48
More



Hospital B
Frequency
8
7
6
5
4
3
2
1
0
3
6
9
12
15
18
21
24
27
30
33
36
39
42
45
48
More

Statistical humor

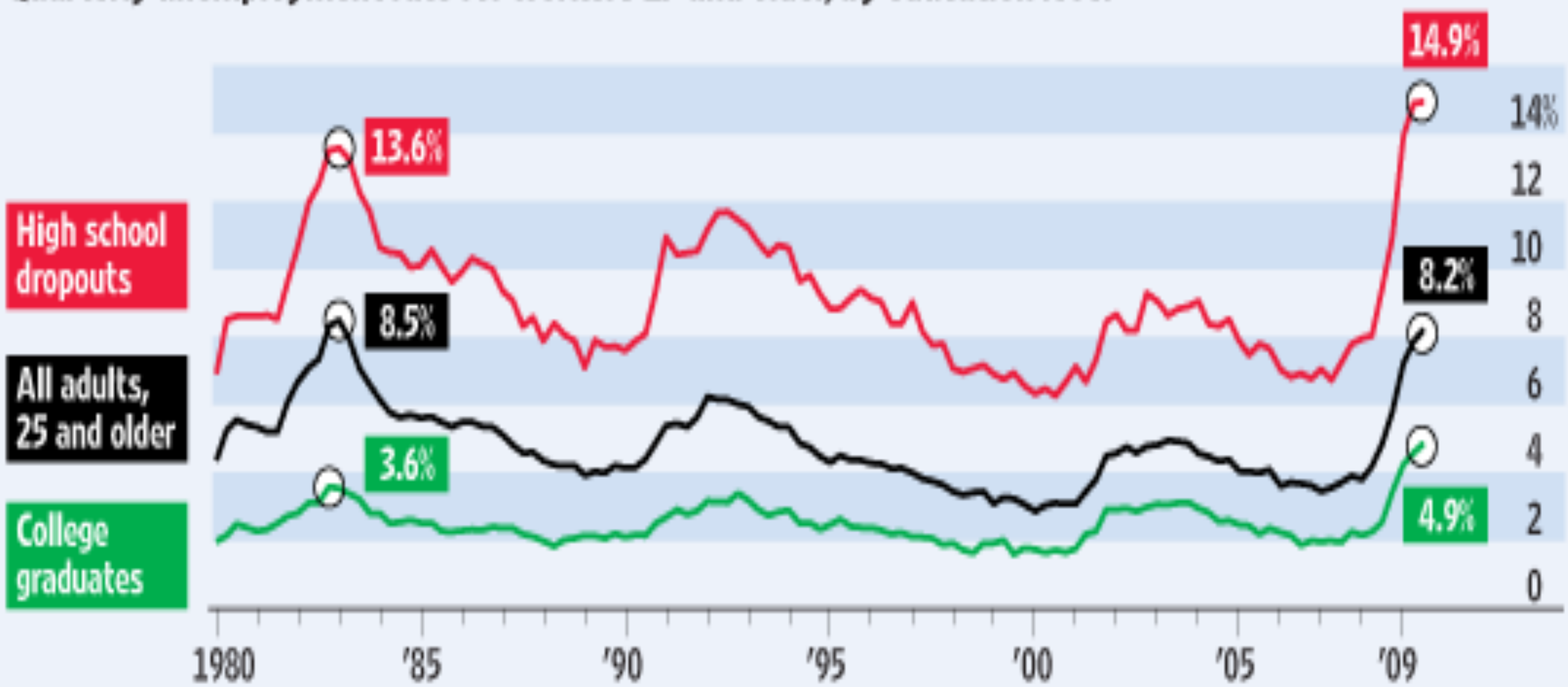
Aggregated Averages

- In October 2009, the jobless rate was 10.2%
- In 1982 the jobless rate was 10.8%
- ***Which recession was worse?***

Adding Up

Jobless rates for workers at each education level have passed peaks reached in the recession of the early 1980s. But the combined unemployment rate hasn't eclipsed its prior peak. The work force now includes more workers with college diplomas, who have the lowest unemployment of any group.

Quarterly unemployment rate for workers 25 and older, by education level



Source: Henry Farber of Princeton University, analyzing Bureau of Labor Statistics data

The SD and the Shape of a Distribution

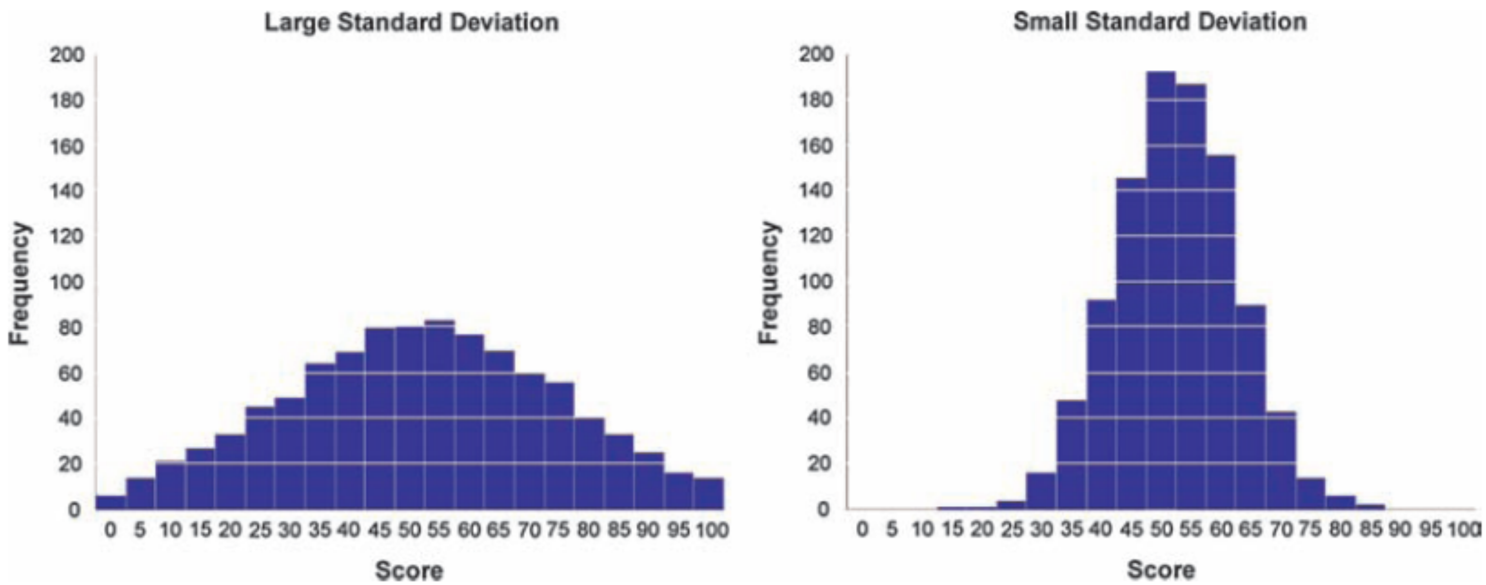


FIGURE 2.6 Two distributions with the same mean, but large and small standard deviations

- Visualizing data doesn’t have to be constrained by the limitations of Excel, PowerPoint, SPSS, or any other ‘tool’

- *www.**youtube**.com/watch?v=jbkSRLYSojo*

Next:

- PROBABILITY