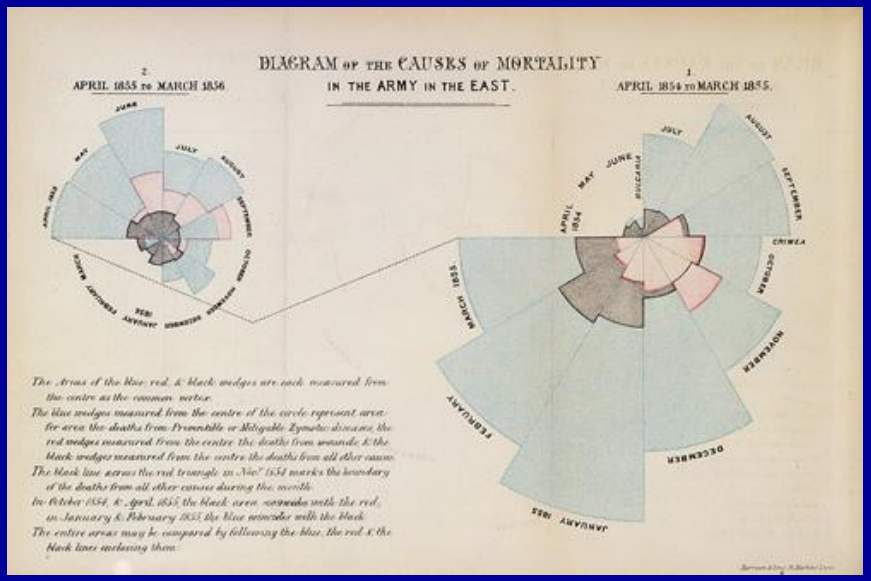
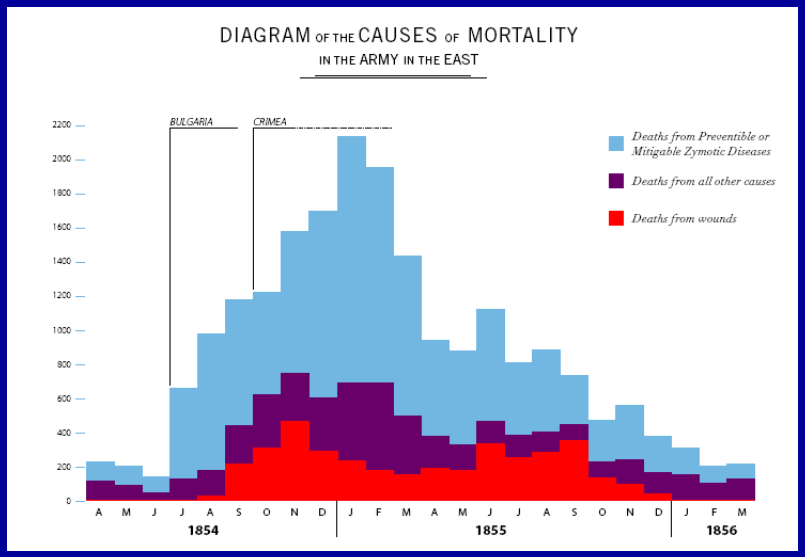
Stats 95

- Statistical analysis without compelling presentation is annoying at best and catastrophic at worst.
- From raw numbers to meaningful pictures

Stats 95

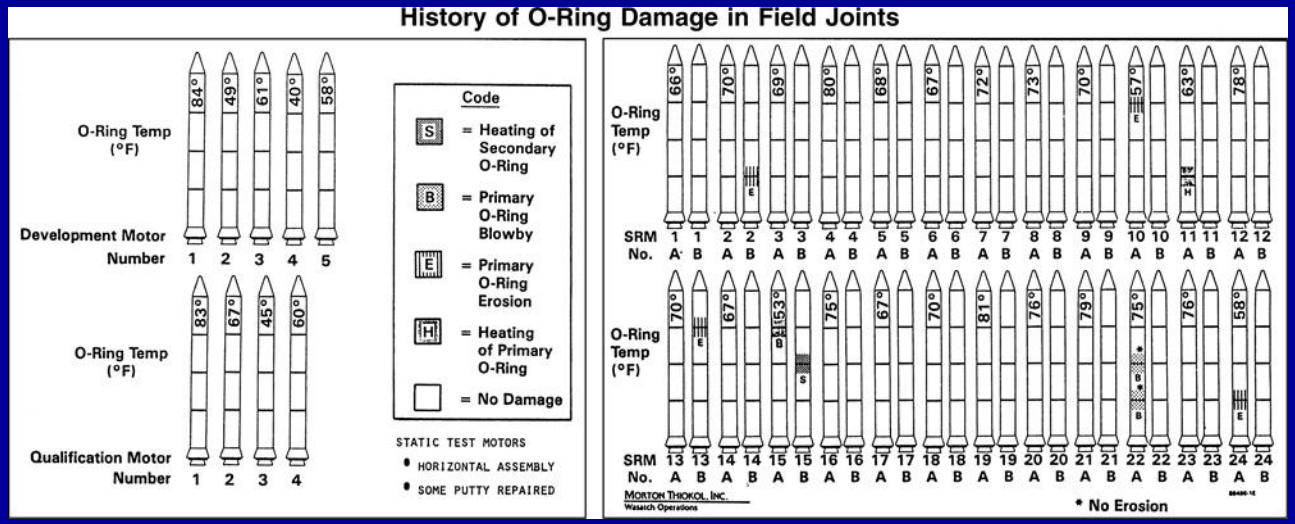
- Why Stats?

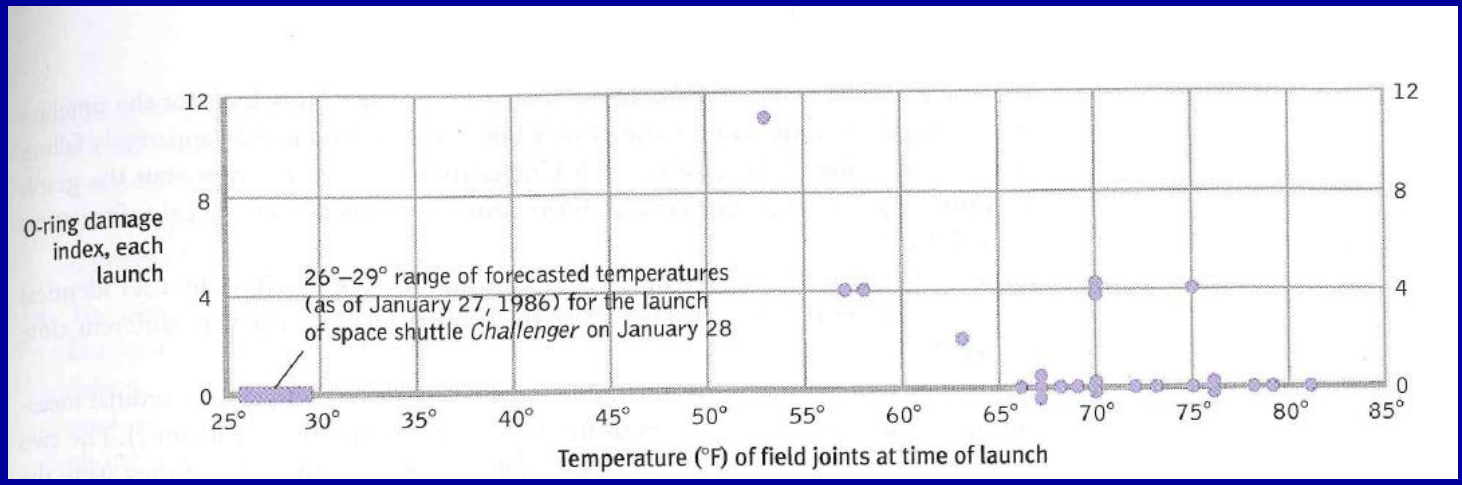




- 200 countries over 200 years
  http://www.youtube.com/watch?v=jbkSRLYSojo

BETTER GRAPHS COULD HAVE SAVED LIVES





Frequency Table

The first rule of statistics:
USE COMMON SENSE!

90% of the information is contained in the graph.

Frequency Distributions: Where It All Starts

- Number of graduate students of U.S. professors in chemistry departments who went on to have jobs at top 50 Chemistry departments (minimum of 3)
- 3 3 3 4 5 9 3 3 5 6 3 4 8 6 3 3 3 4 4 4 7 6 3 5 5 7 13 3 3 3 3 3 4 4 4 5 6 7 6 7 8 8 3 3 3 5 3 3 5 3 5 3 3

Frequency Table: It All Starts Here! **Not only does a Frequency table organize data in way a way that makes it intelligible, the link between the observation and its *statistical probability* starts here.**

TABLE 2-3. EXPANSION OF A FREQUENCY TABLE

This frequency table is an expansion of Table 2-2, which depicts the numbers of students placed on the top 50 chemistry faculties by each top-producing graduate advisor. It now includes percentages and cumulative percentages, which are often more descriptive than the actual counts.

FORMER STUDENTS NOW IN TOP JOBS	FREQUENCY	PERCENT	CUMULATIVE PERCENT
13	1	1.85	100.00
12	0	0.00	98.15
11	0	0.00	98.15
10	0	0.00	98.15
9	1	1.85	98.15
8	3	5.56	96.30
7	4	7.41	90.74
6	5	9.26	83.33
5	9	16.67	74.07
4	8	14.81	57.41
3	23	42.59	42.59

Data from Kuck et al. (2007).

Frequency Table: Steps

- Find Max and Min. scores
- Determine Range (Max-Min+1)
- Determine #of intervals
  - More art than science, judgment call
- Decide bottom number
- From Highest to lowest, count *number of scores* that belongs in each "bin" / interval.

Frequency Graphs

Descriptive Stats:

Bar Graphs Vs. Histograms

Always Draw A Graph!

• Bar Graphs

- **If Variable is Categorical *(labels)***
  - **Bars do NOT TOUCH**

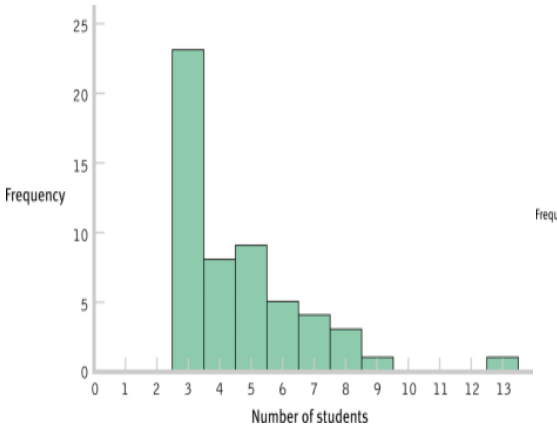
Bar Graphs can be horizontal

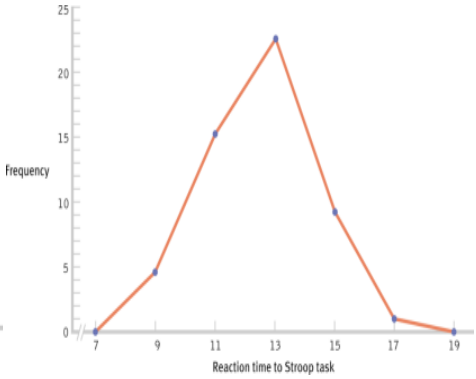
#	Answer		Response	%
1	Distance		10	12%
2	Time		18	21%
3	Inconvenience		13	15%
4	Unaware when/where		58	68%
5	Technical challenge		3	4%
6	Others (specify)		13	15%

Lei Lei * and Brian Hilton
2013

• Histogram

- **If Variable is Continuous *(numbers)***
  - **Bars TOUCH: Discrete *(no decimals)***
  - **Frequency Polygon: Scale/Continuous *(decimals)***

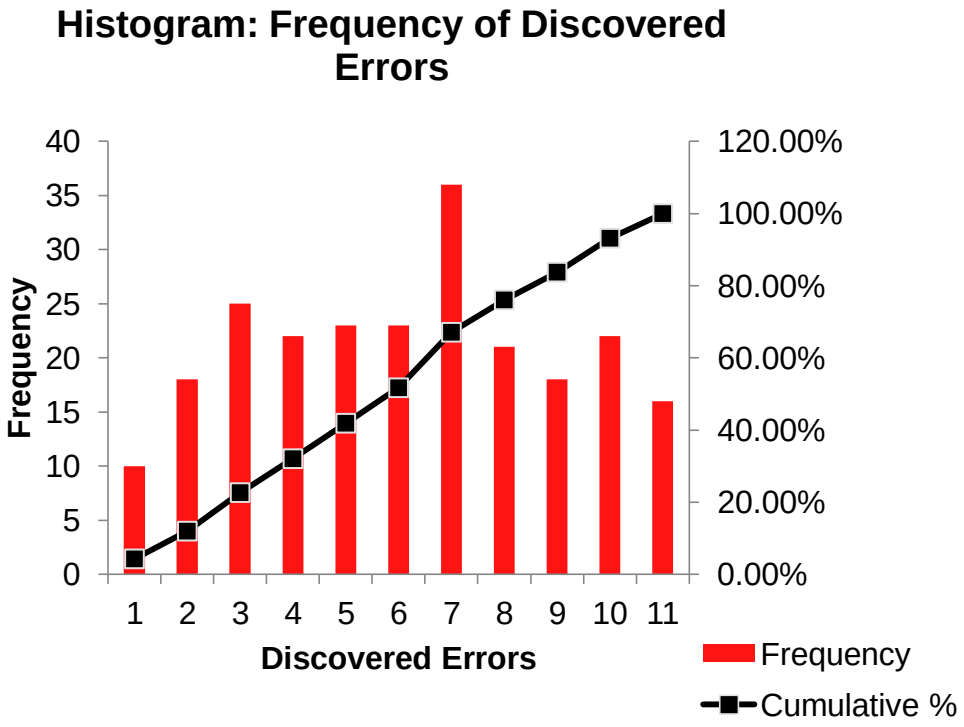




Plot Frequency Graph of Table

- *Example Using Fictitious Data: # of Discovered Errors*

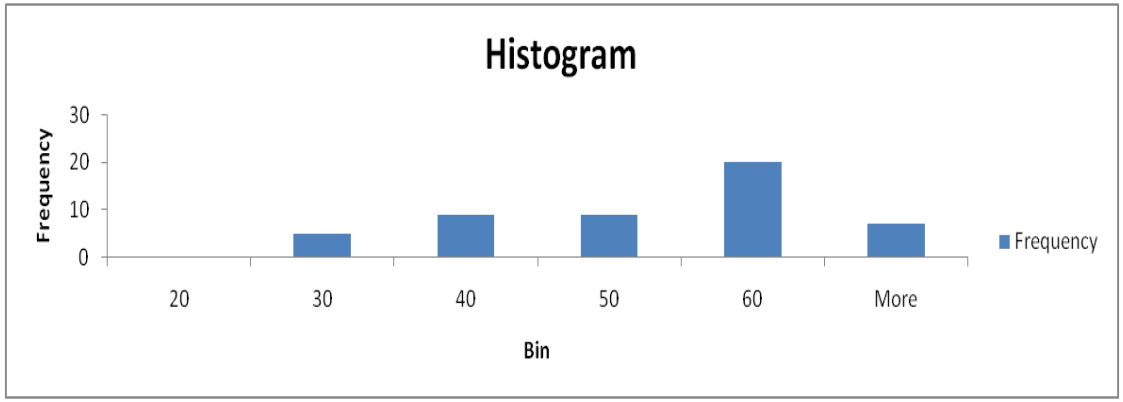
Errors Discovered	*Frequency*	*%*	*Cumulative %*
1	10	4.27%	4.27%
2	18	7.69%	11.97%
3	25	10.68%	22.65%
4	22	9.40%	32.05%
5	23	9.83%	41.88%
6	23	9.83%	51.71%
7	36	15.38%	67.09%
8	21	8.97%	76.07%
9	18	7.69%	83.76%
10	22	9.40%	93.16%
11	16	6.84%	100.00%

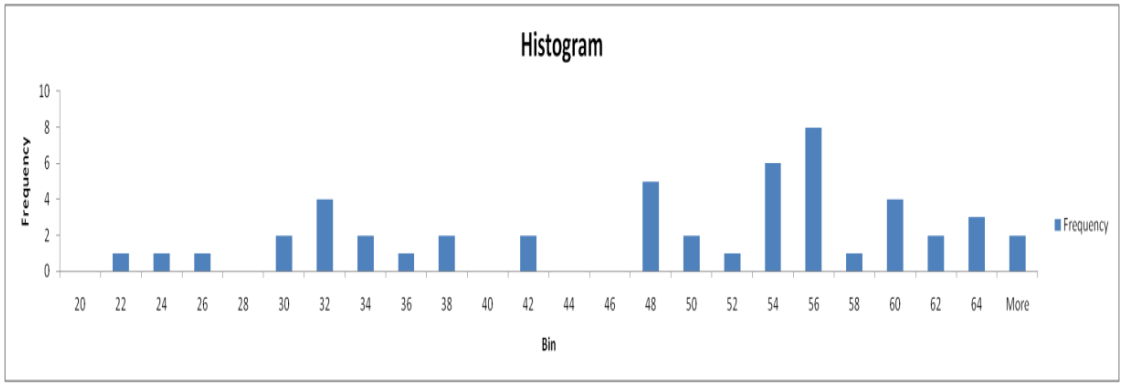


Descriptive Stats: Histograms Vs. Bar Graphs

Bin size can help you tell the story

- Find Max, Min and Range (Max-Min+1)
- Make Bins
  - More art than science, judgment call



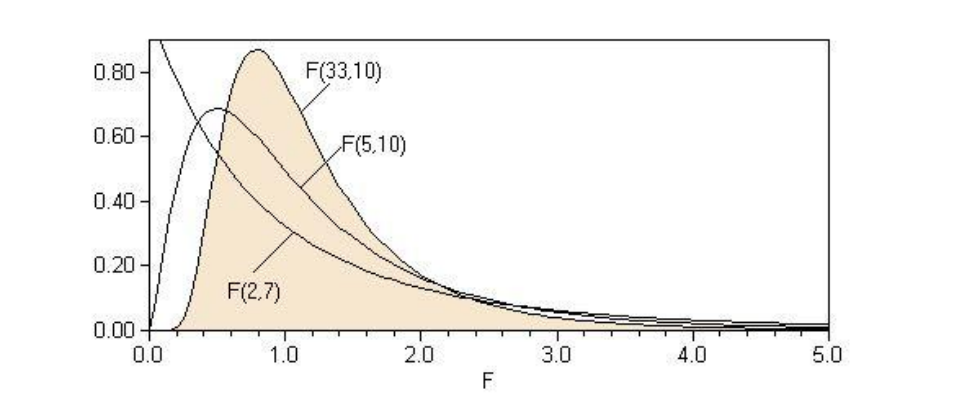


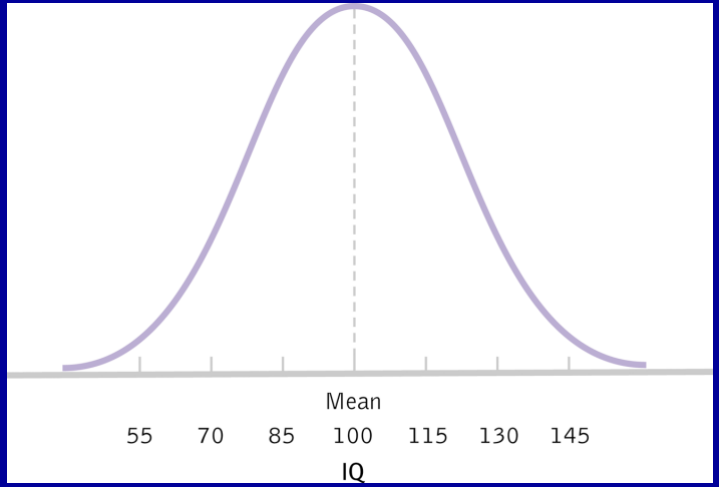
Histogram: Steps

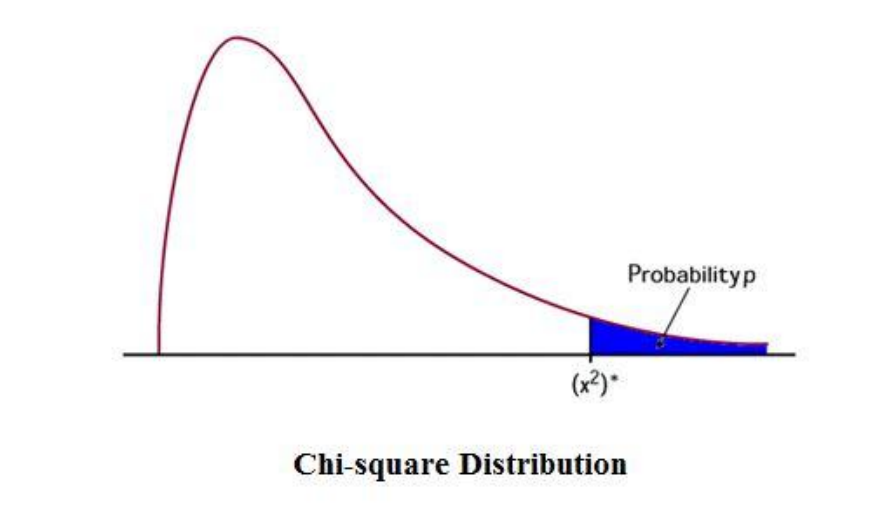
- Define X-axis
- Define the range of X-axis variable.
- Define range of frequency on the Y-axis .
- CHOOSE the bin size (wisely).

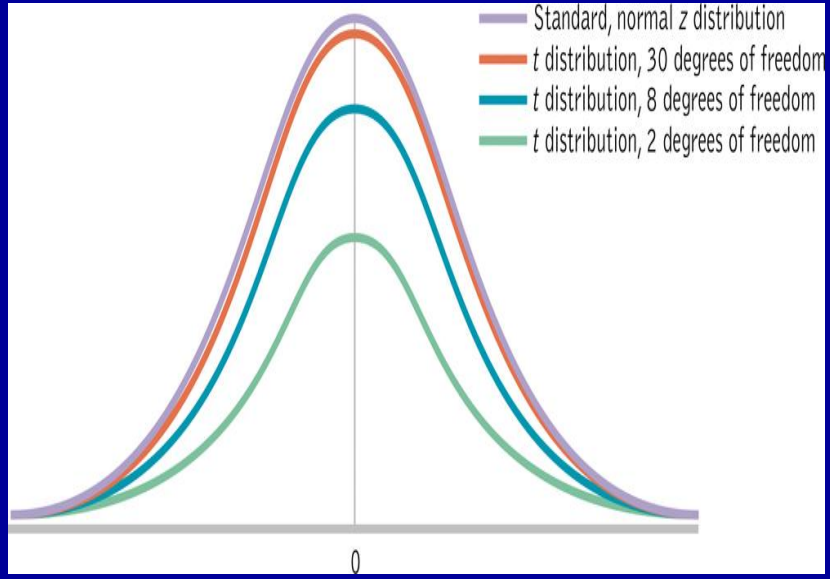
Frequency Distributions
Central Tendencies
Variability

Distributions

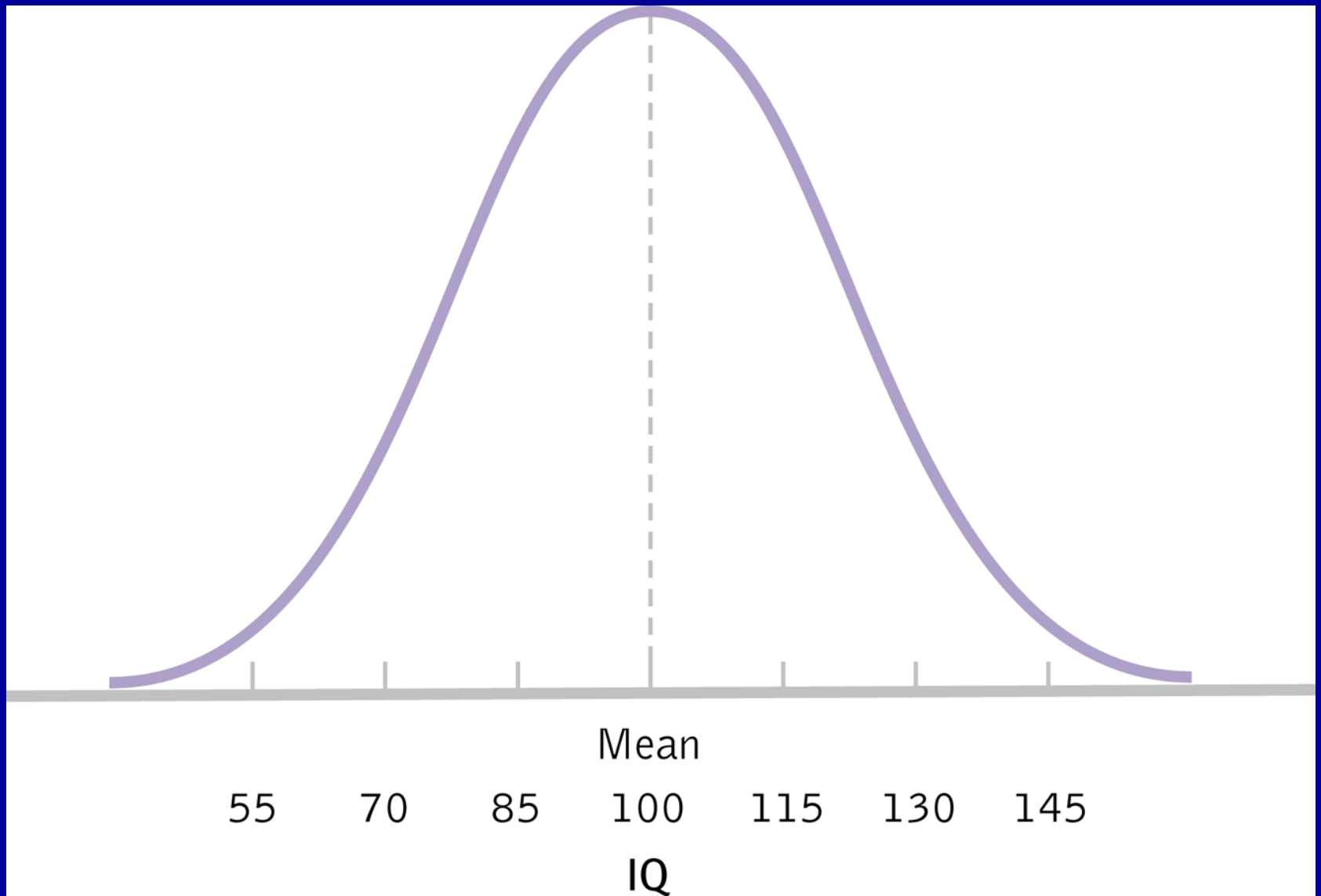




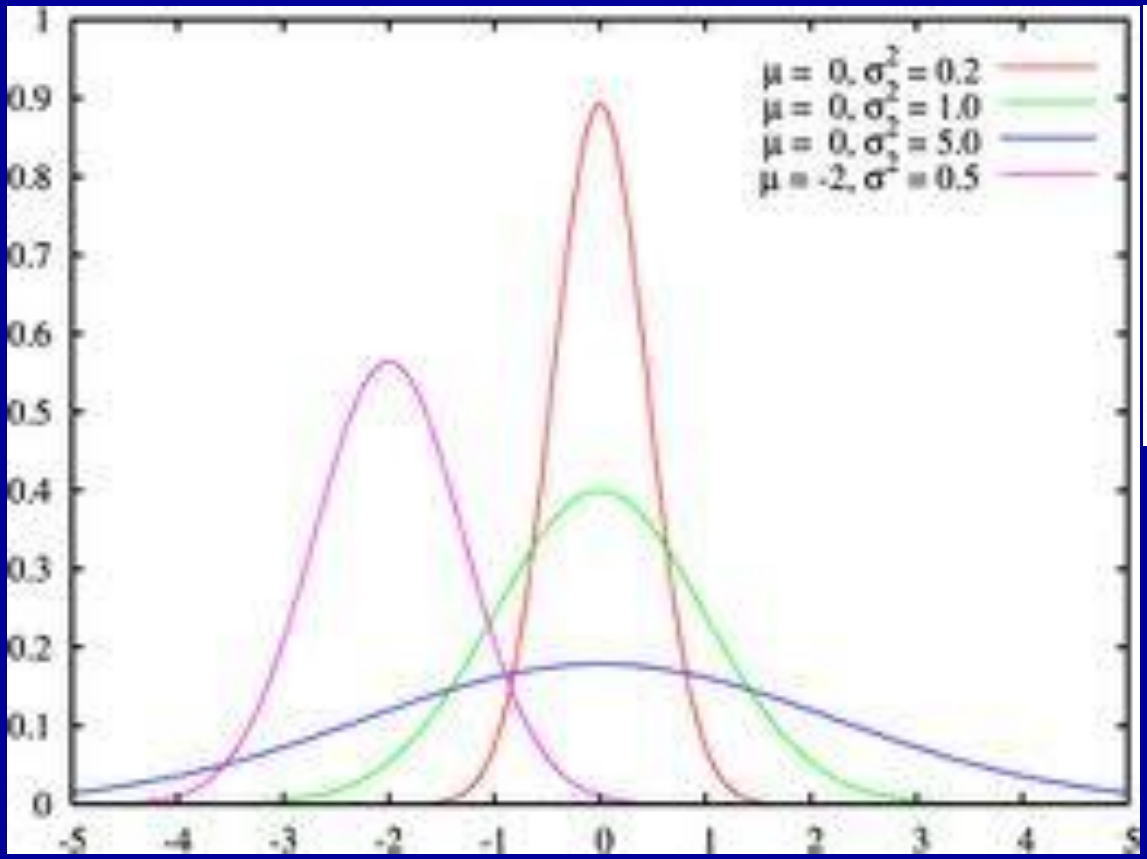




“God Loves The Normal Distribution”



Normal Distributions: Different Shapes, Same Formula



$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Normal (Gaussian)

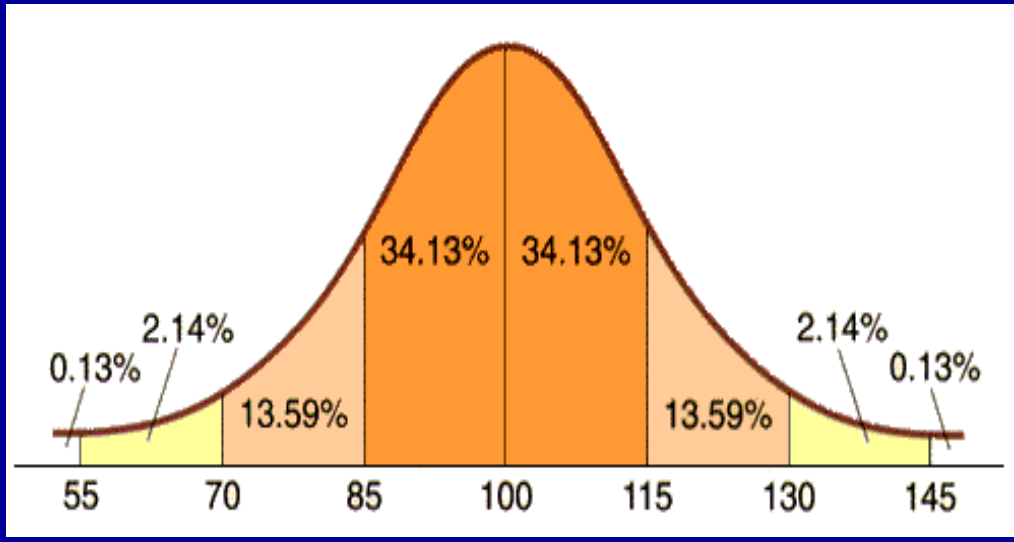
μ is the mean

σ is the standard deviation

The Normal Distribution

Described by:

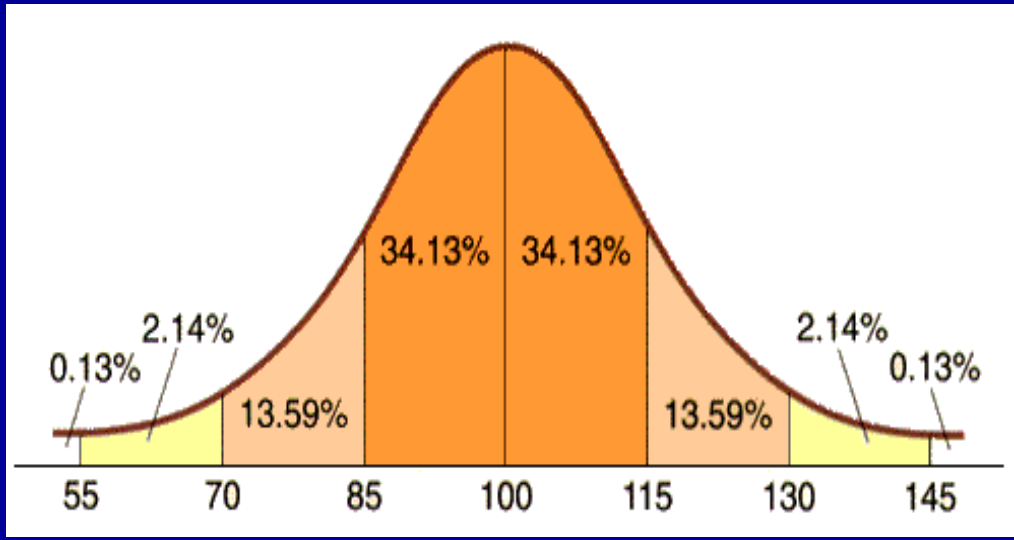
- Shape
- Central Tendency
- Variability



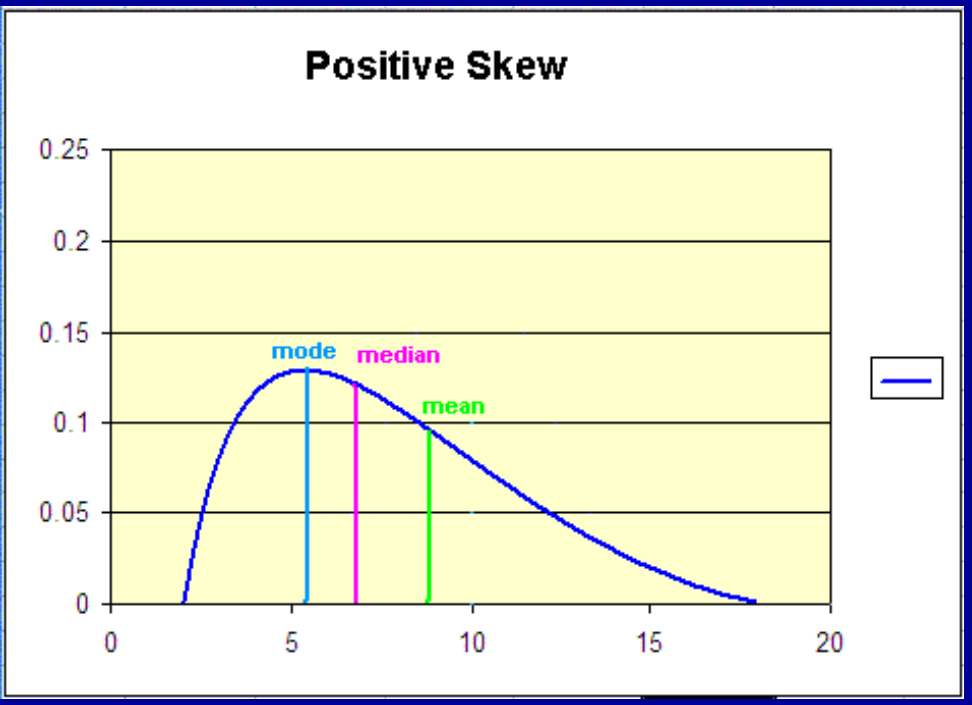
2.2 The Science of Observation

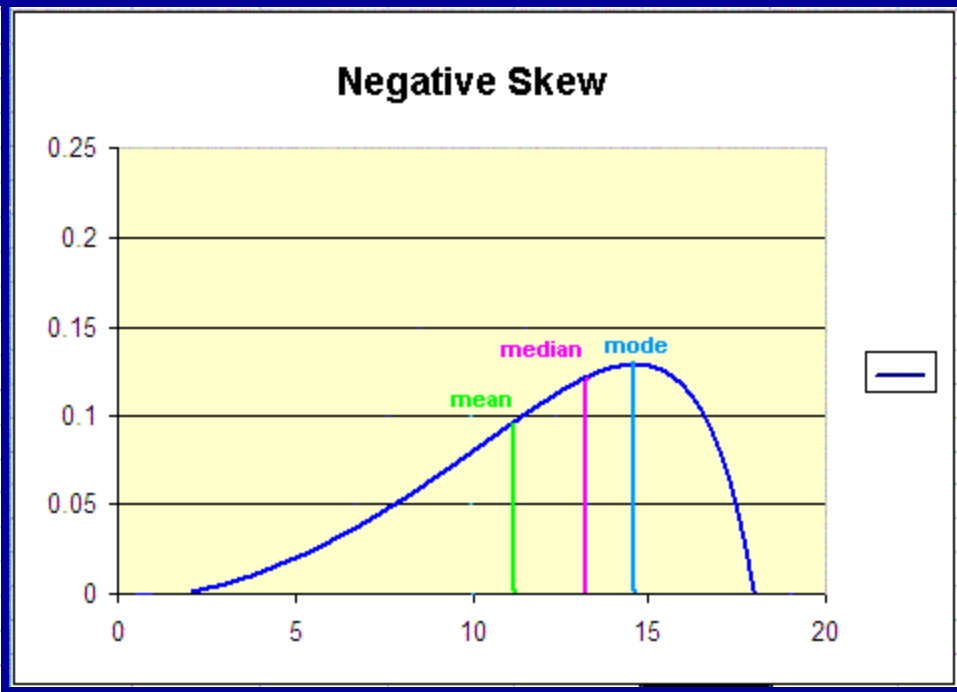
Normal distribution

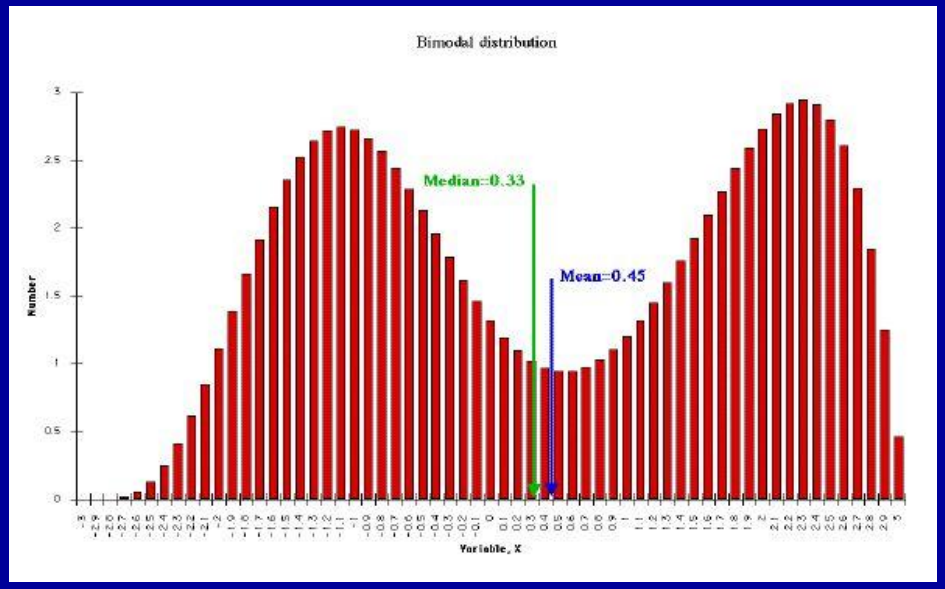
- Shape
  - •Symmetrical
  - •Positively skewed
  - •Negatively skewed
  - •Bi modal



- Tails go on to infinity
- Central tendencies overlap exactly
- Most biological measures (height, IQ), random events (coin tosses, dice) and measurement error falls in a Normal Distribution.

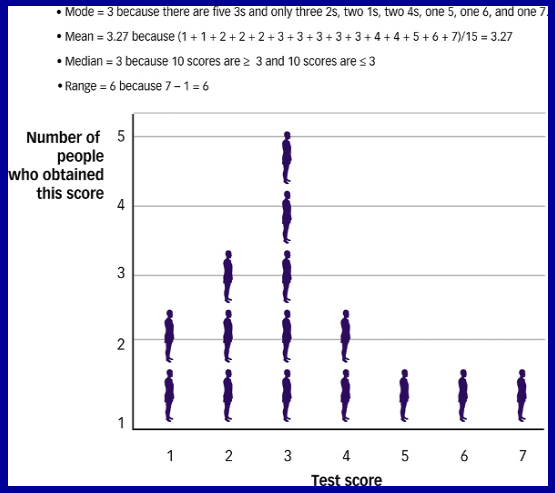

Positive Skew
0.25
0.2
0.15
0.1
0.05
0
mode
median
mean
0
5
10
15
20


Negative Skew
0.25
0.2
0.15
0.1
0.05
0
median
mode
mean
0
5
10
15
20

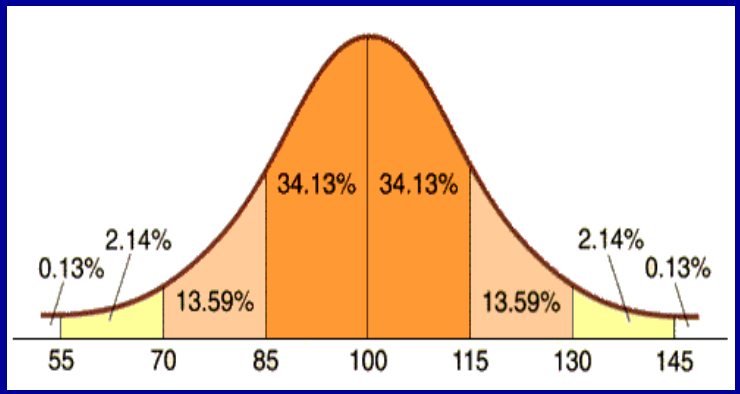

Bimodal distribution
Median=0.33
Mean=0.45
Number
Variable, X

Central Tendencies

- Central tendency
  - mode—most frequent
  - mean—average
  - median— 50th percentile



Normal Distribution: All three central tendencies **overlap**



From Grad Student Example: What is the Median?

3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 5 5 5 5 5 5 5 6 6 6 6 6 7 7 7 7 8 8 8 9 13

Sources of Variability

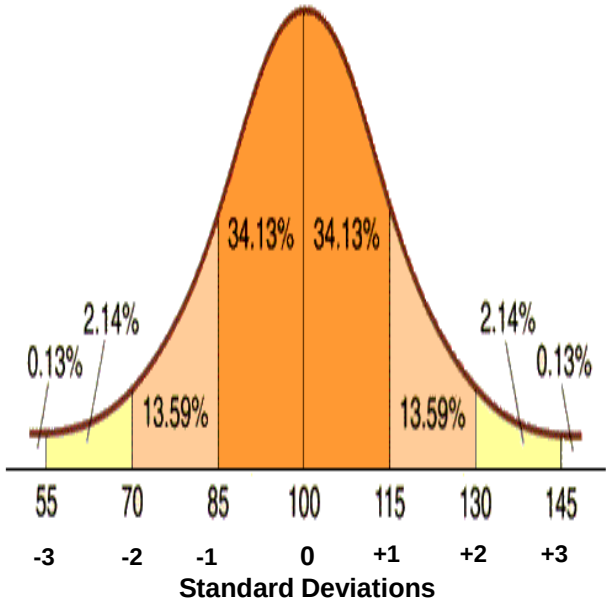
- Environmental and Natural forces
- Random variation
- Measurement error

- Measures of Variability:
  - Range (max-min)
  - Quartiles (divide cumulative percentiles in equal quarters)
  - Variance
  - Standard deviation (avg distance of all observations to the mean and the square root of Variance)

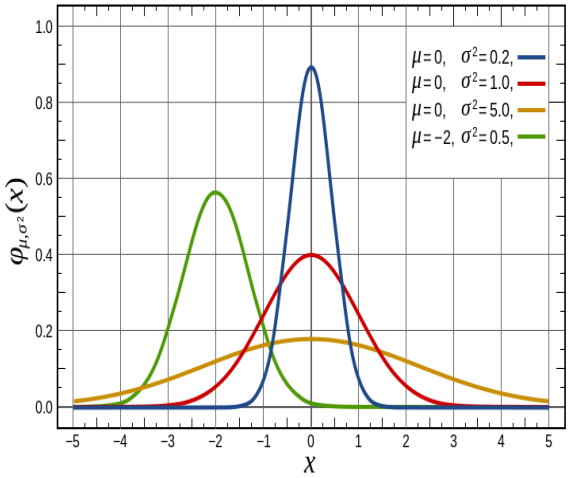
Descriptive Statistics: Standard Deviation

Variability

- **Standard Deviation = average distance of all measures from the mean**
- **As STDV decreases, Normal Distribution becomes narrower and taller**
- **In Normal Distribution, Mean = 0 STDV**
- **As distance from mean increases = probability decreases**
- **In normal distribution for a Population, STDV is called Z-score**



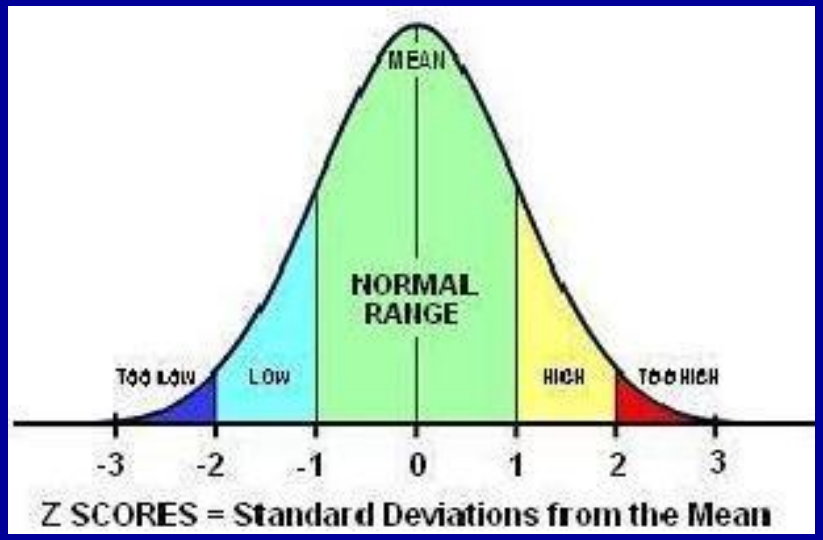
Normal Distribution



Normal Distributions with diff. Standard Deviations.
The Smaller the STDV The Narrower the Curve

2.2 The Science of Observation

- Variability
  - Range
- Standard deviation
  - The "average range", the average distance from a point to the mean
  - Measure of "Specialness"



Symbols for Standard Deviation

s ***(sample)*** σ **(population)**

Expressed in z-scores

Plotting Frequencies with Box Plots and Stem & Leaf

Stem-and-Leaf: Exam 1 & 3

Selection of ranges & bins like Histogram, but usually simpler.

Exam1 Stem-and-Leaf Plot

Frequency	Stem	&	Leaf
3.00	Extremes		(=<19)
1.00	2	.	1
1.00	2	.	3
1.00	2	.	4
2.00	2	.	67
4.00	2	.	8899
13.00	3	.	0000011111111
8.00	3	.	22333333
14.00	3	.	44444455555555
7.00	3	.	6666777
15.00	3	.	888899999999999
14.00	4	.	00000111111111
12.00	4	.	222222222333
10.00	4	.	4444555555
3.00	4	.	667
4.00	4	.	8889

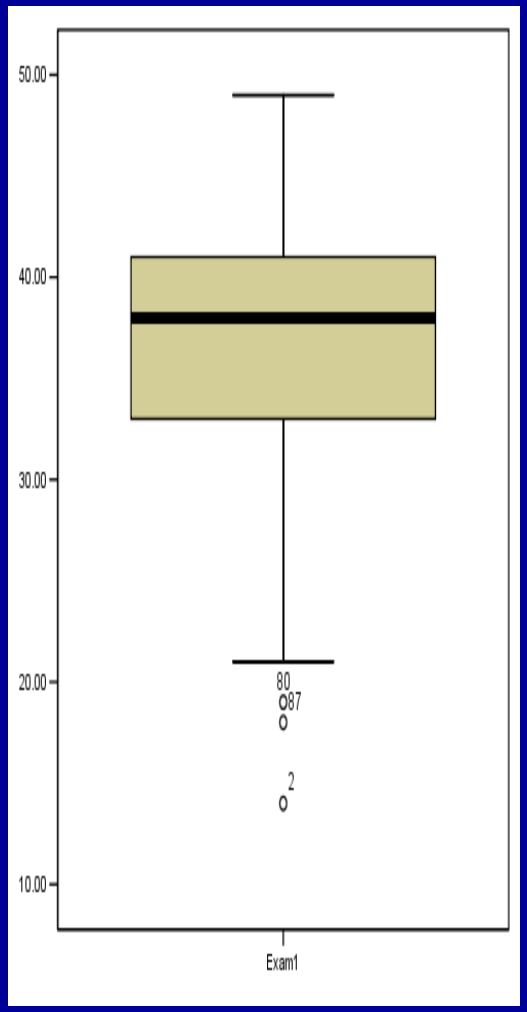
Exam3 Stem-and-Leaf Plot

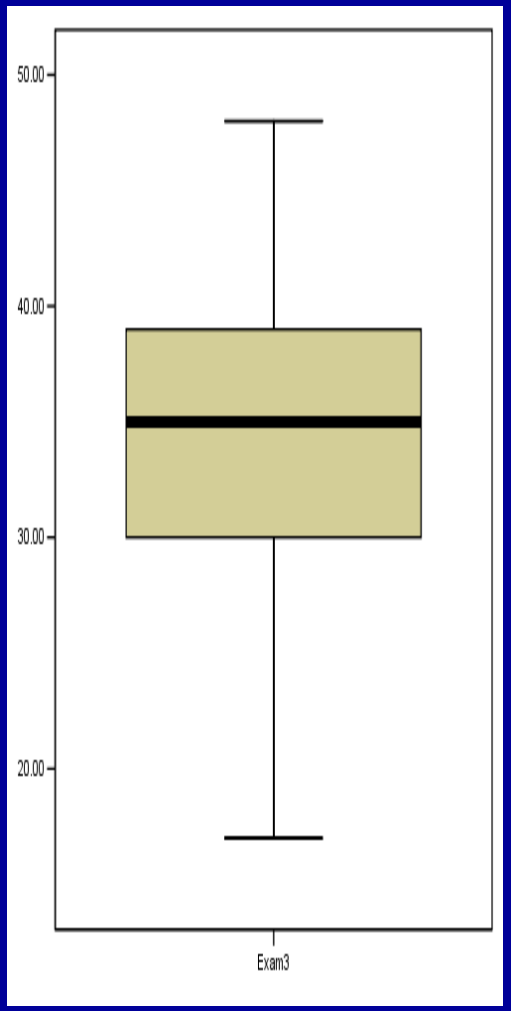
Frequency	Stem	&	Leaf
1.00	1	.	7
3.00	1	.	899
1.00	2	.	0
2.00	2	.	23
1.00	2	.	5
7.00	2	.	6667777
9.00	2	.	888889999
9.00	3	.	000000011
8.00	3	.	22233333
12.00	3	.	444444455555
12.00	3	.	666666677777
16.00	3	.	8888888899999999
14.00	4	.	00000000001111
4.00	4	.	3333
1.00	4	.	4
.00	4	.	
1.00	4	.	8

These plots represent the scores on an exam given to two different sections for the same course.

Box-Plot: Exam 1 & 3

- Box ‘Bottom’ = 1st quartile
- Box ‘Midline’ = 2 quartile
- Box ‘Top’ = 3rd Quartile
- Stems = Lowest & Highest scores, range.

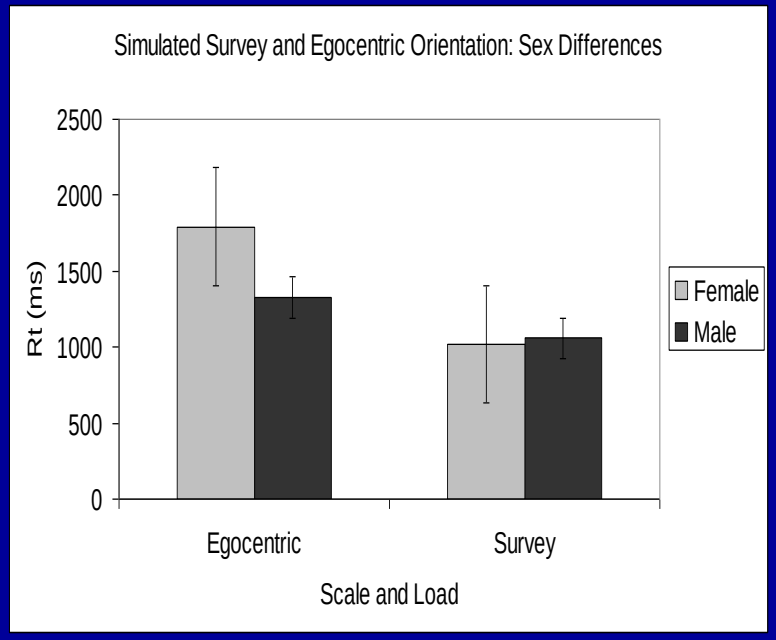


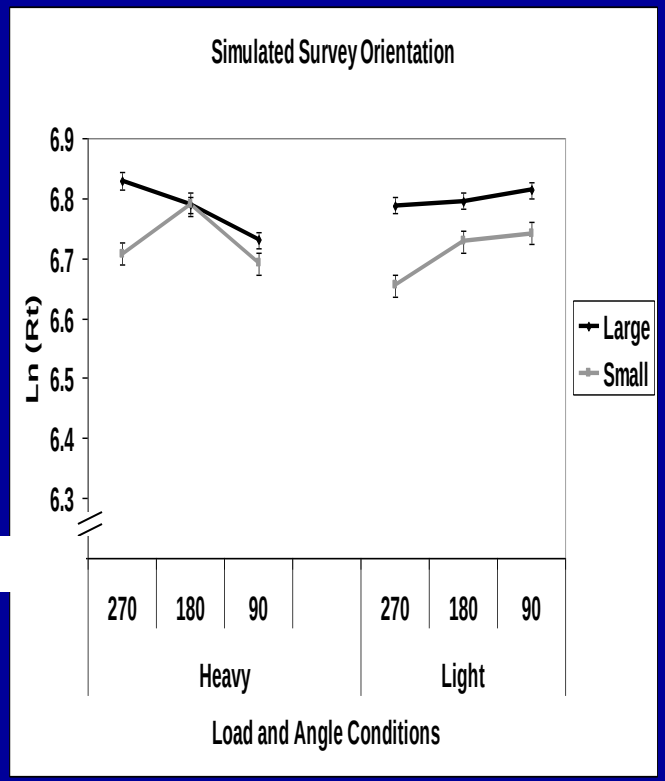


How to Lie with Graphs

Tricks in Describing Stats

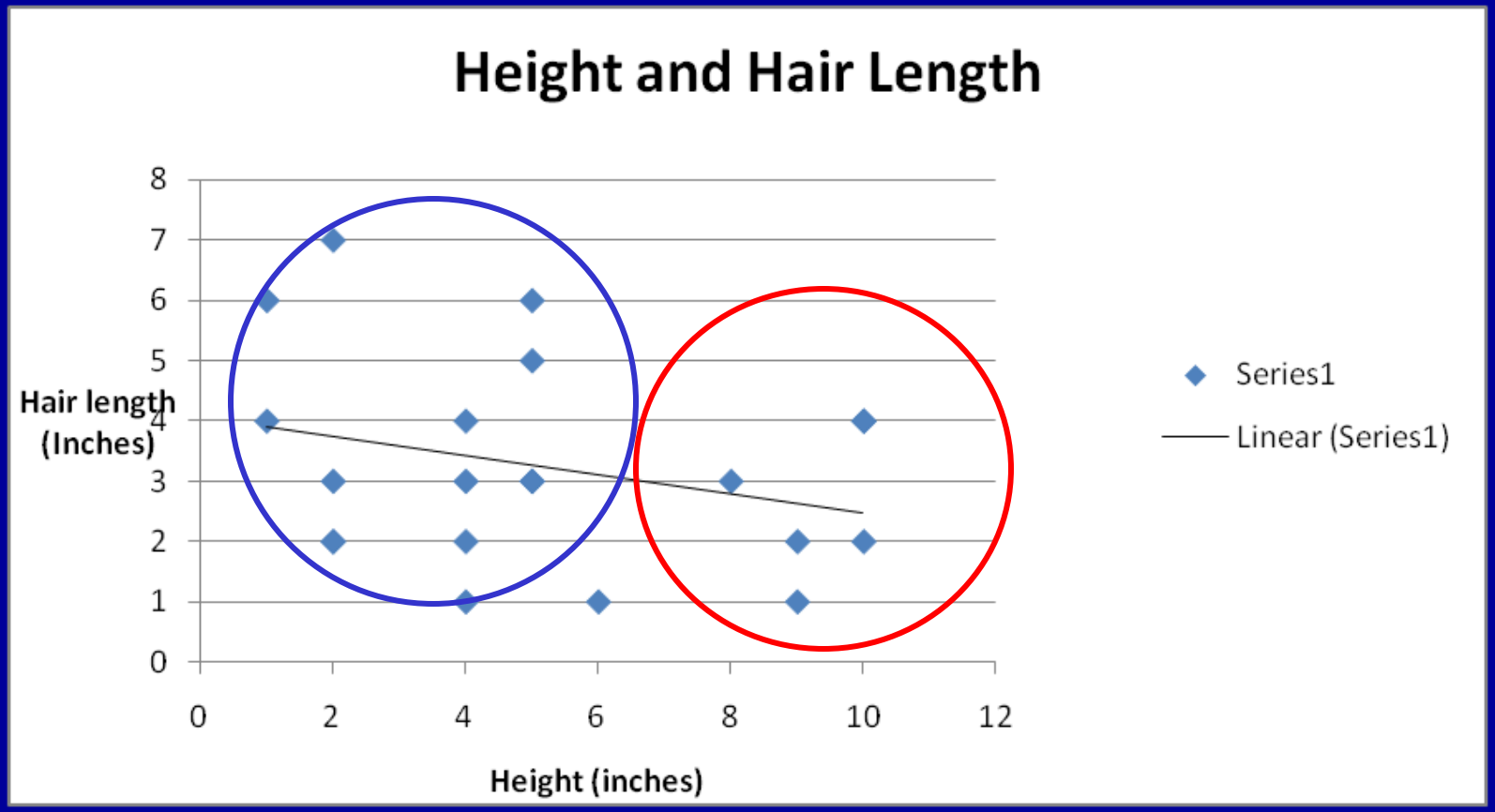
- If you omit Zero in your scale, must indicate with ellipsis





Statistical Control

- Using Multivariate Analysis



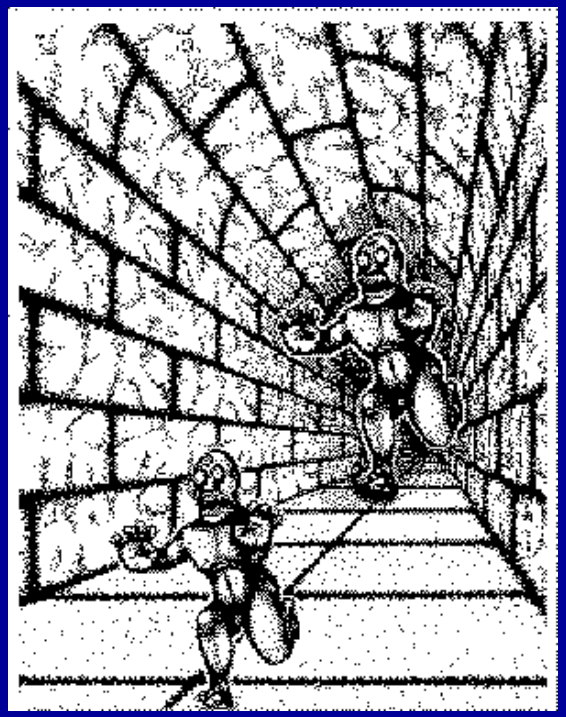
Tricks in Describing Statistics: Distorting Proportions

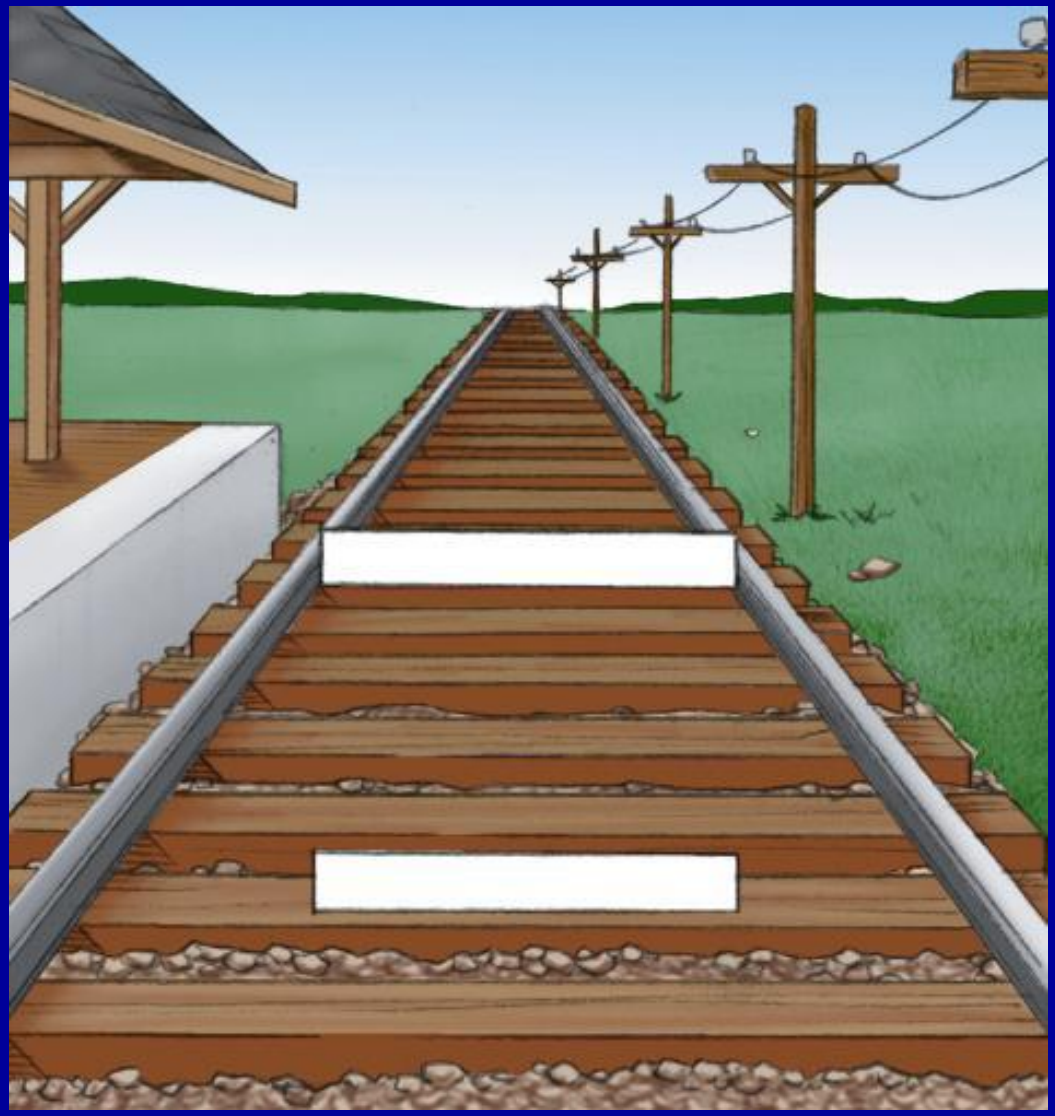
- What's wrong with this comparison?

- Value of Dollar in 1960's

- Value of Dollar in 2010's

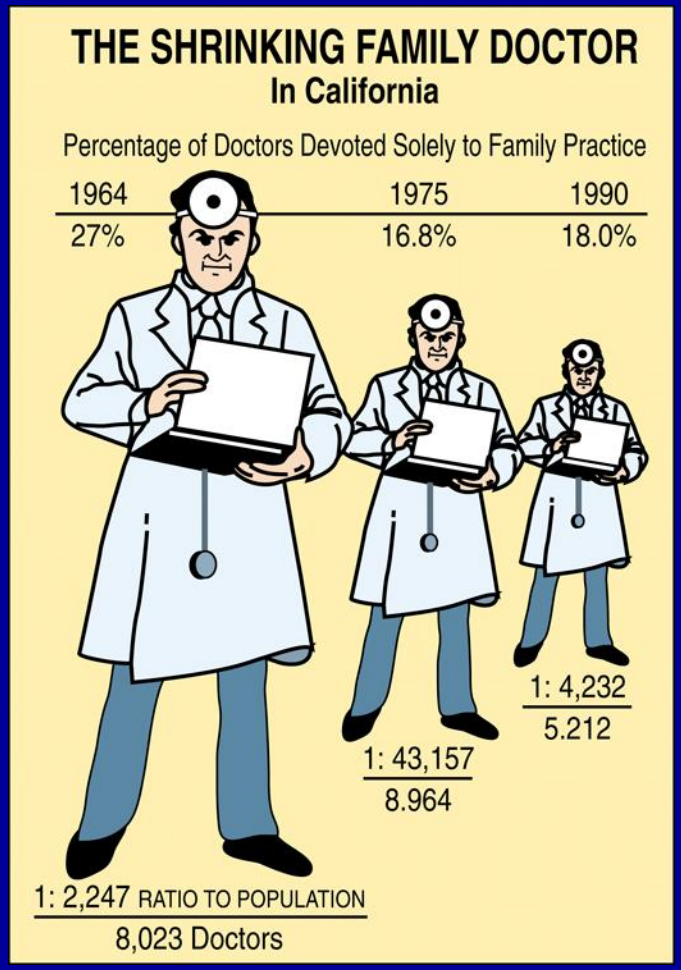
A simple visual size illusion

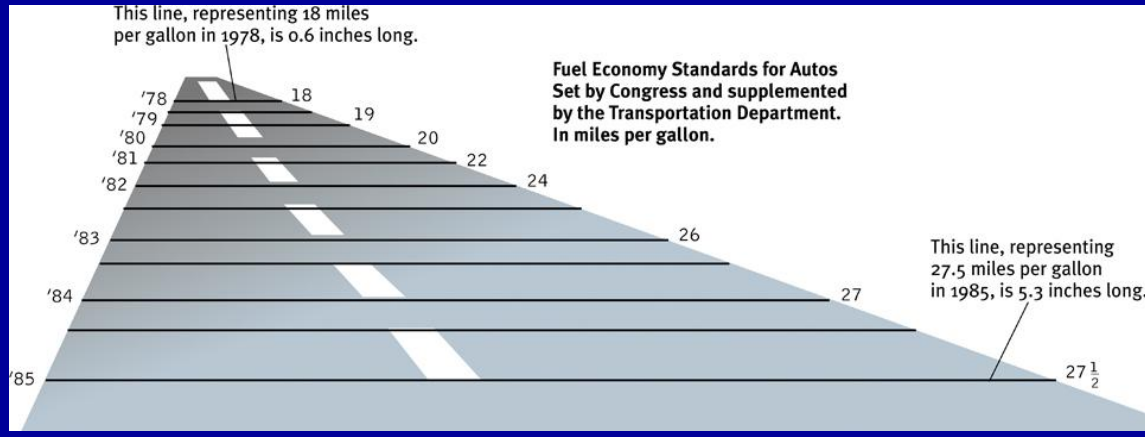




Ponzo Illusion: visual depth illusion

How To Lie with Graphs





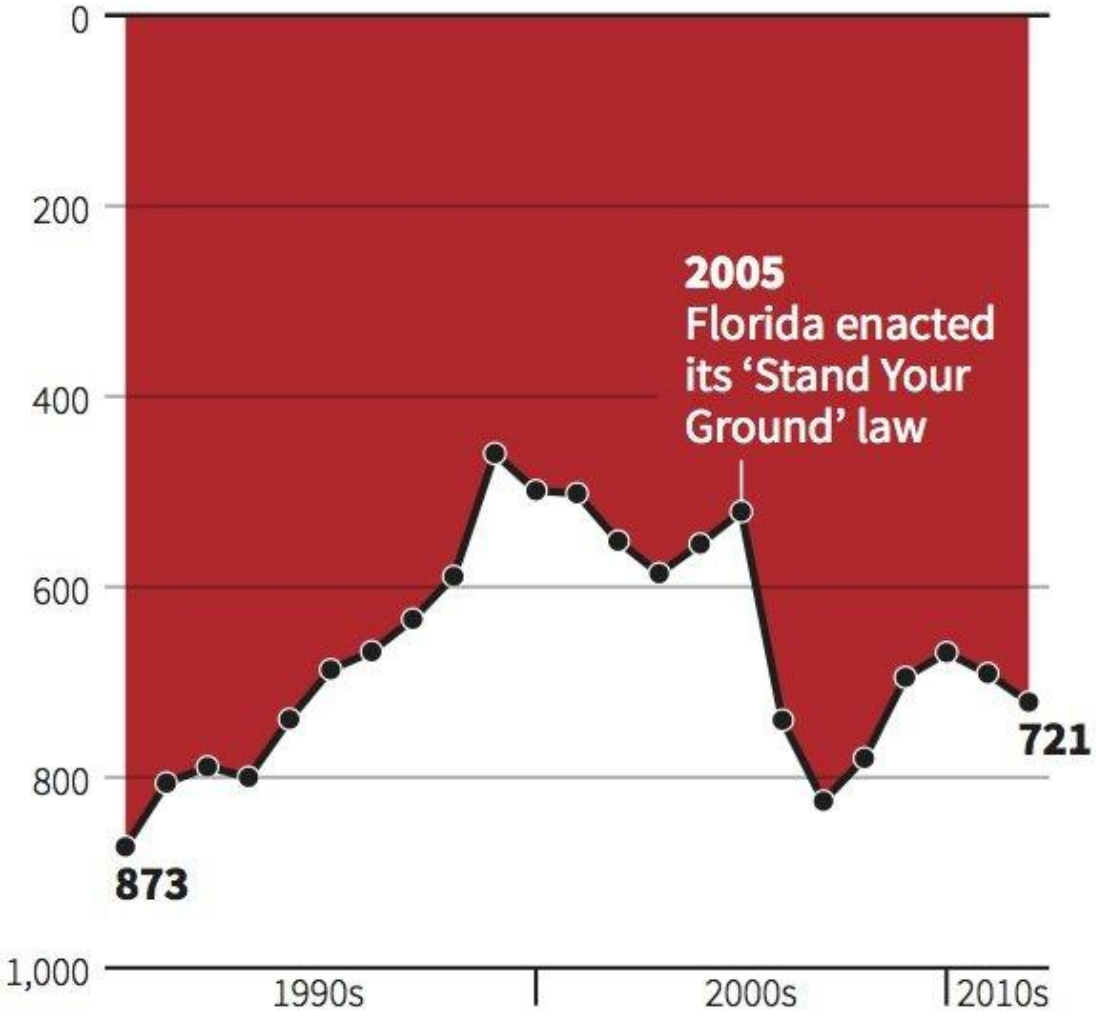
aphs

- Th
de


ITHACA TIMES
Visit us online at http://www.ithacatimes.com
Planning Board approves Widewaters development
IC students occupy Job Hall
Cayuga Vocal Ensemble ushers in the holidays with "Judas Maccabaeus"
Rising Signs
1999
Cornell University's Tuition
1965
Cornell University's Ranking
1989
1999
Why does college have to cost so much?
Vol. XXIII. No. 18 / Our Twenty Sixth Year / December 7, 2000 / FREE

Gun deaths in Florida

Number of murders committed using firearms

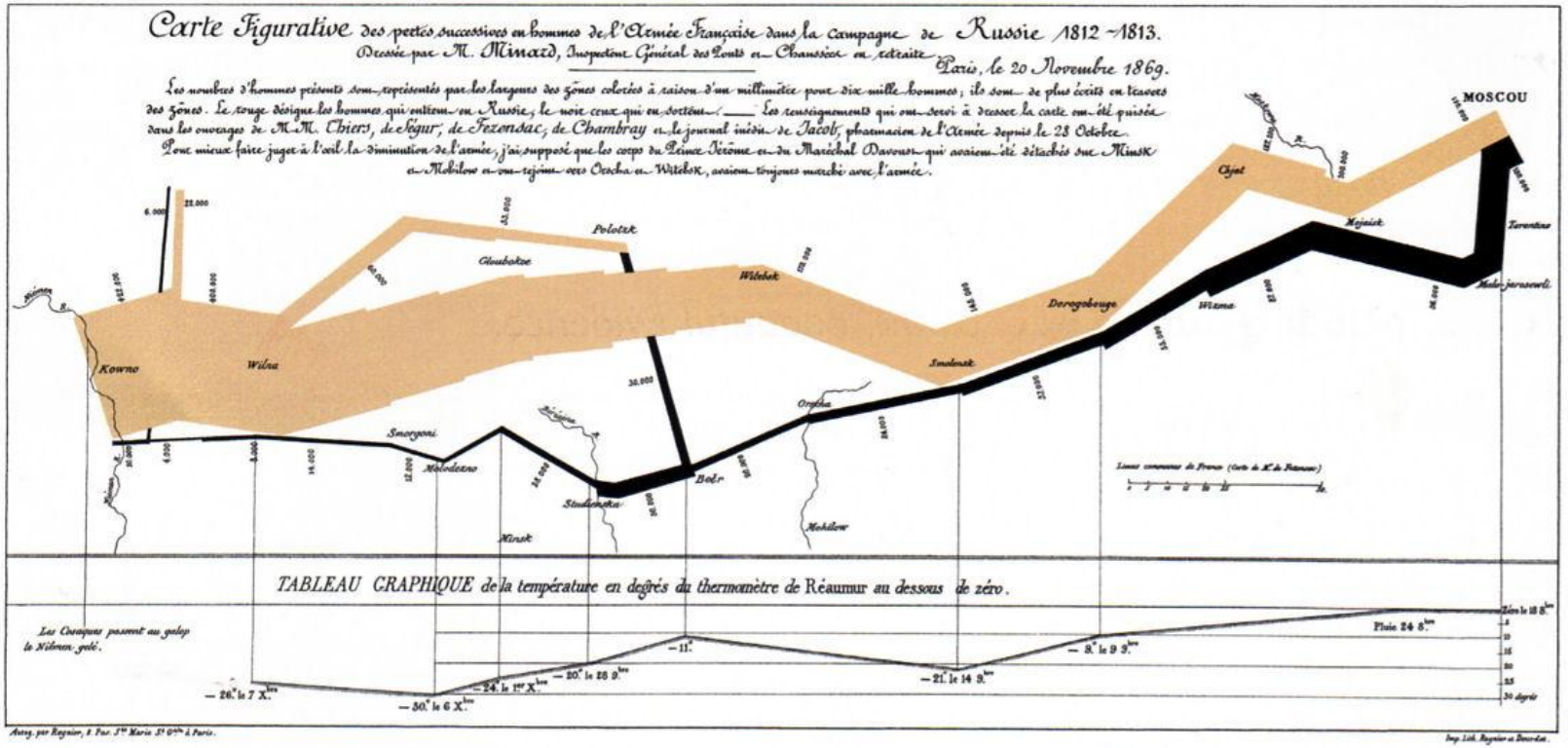


Source: Florida Department of Law Enforcement

C. Chan 16/02/2014

REUTERS

World's Best Graph



This map drawn by Charles Joseph Minard portrays the losses suffered by Napoleon's army in the Russian campaign of 1812. Beginning at the left on the Polish-Russian border near the Niemen, the thick band shows the size of the army (422,000 men) as it invaded Russia. The width of the band indicates the size of the army at each position. In September, the army reached Moscow with 100,000 men. The path of Napoleon's retreat from Moscow in the bitterly cold winter is depicted by the dark lower band, which is tied to temperature and time scales. The remains of the Grande Armée struggled out of Russia with 10,000 men. Minard's graphic tells a rich, coherent story with its multivariate data, far more enlightening than just a single number bouncing along over time. Six variables are plotted: the size of the army, its location on a two-dimensional surface, direction of the army's movement, and temperature on various dates during the retreat from Moscow. It may well be the best statistical graphic ever drawn. Napoleon's March poster $14 postpaid; English/French version $18 postpaid.

The End

- Back-up slides

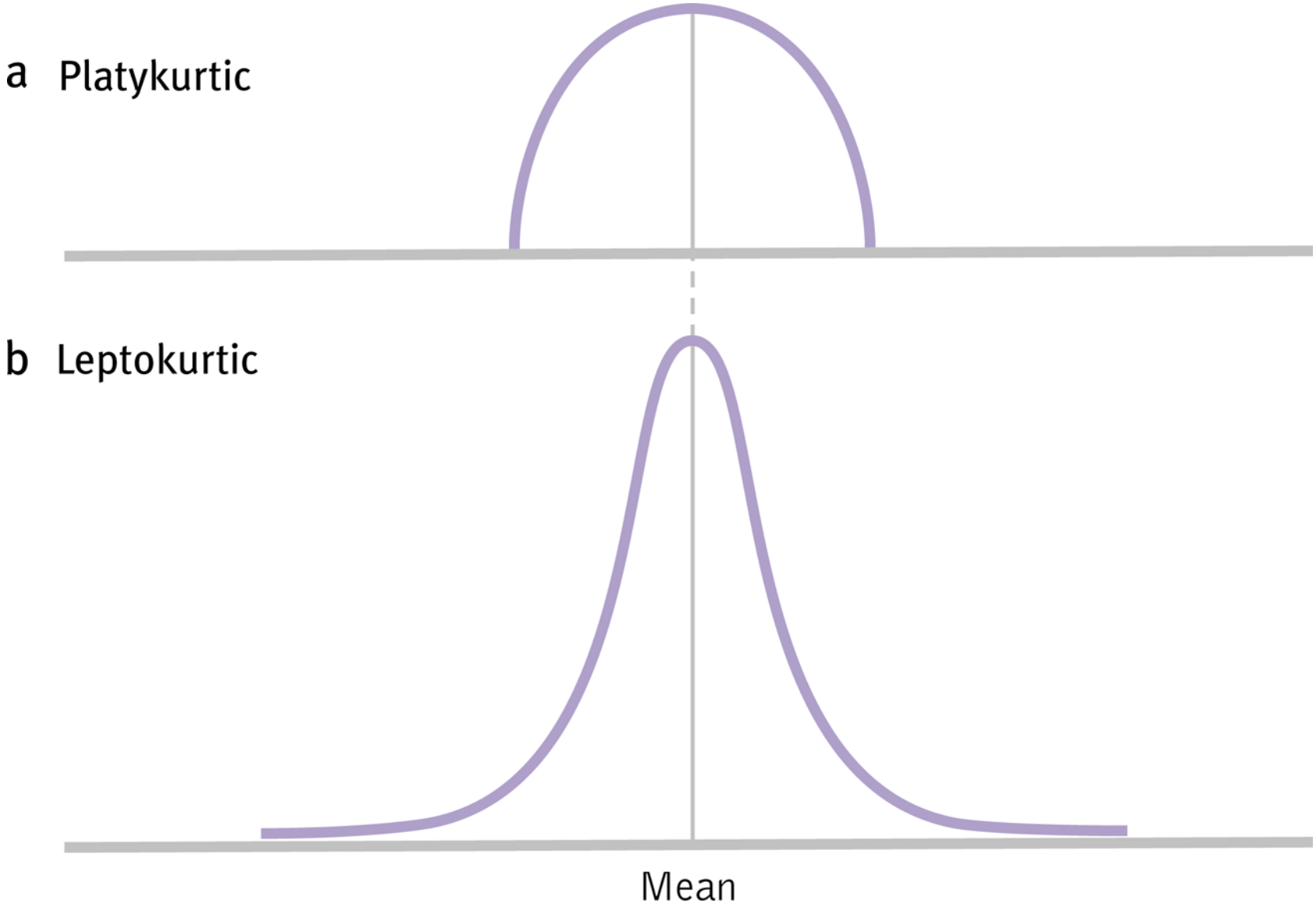


Figure 2.14 Kurtosis
Nolan and Heinzen: Statistics for the Behavioral Sciences, First Edition

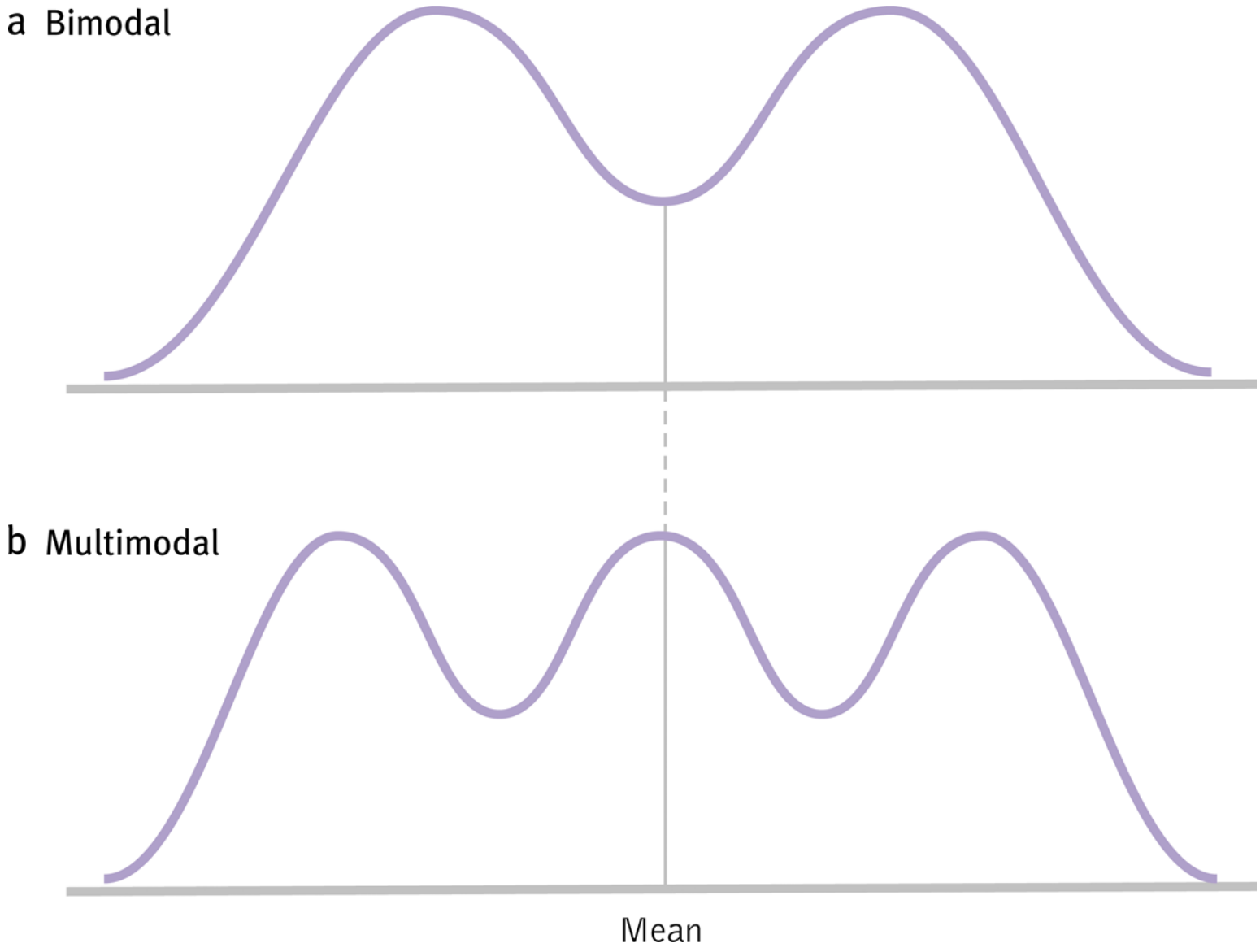
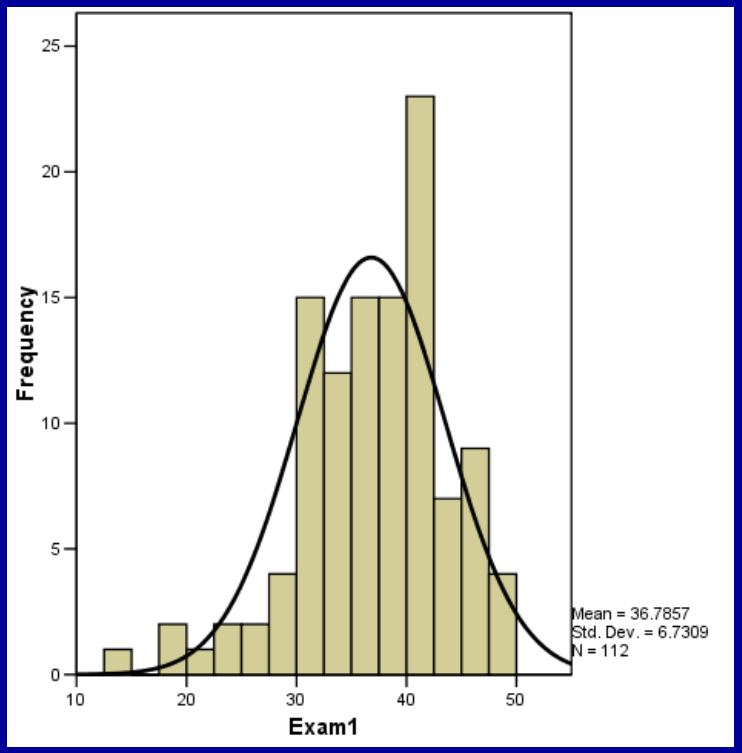
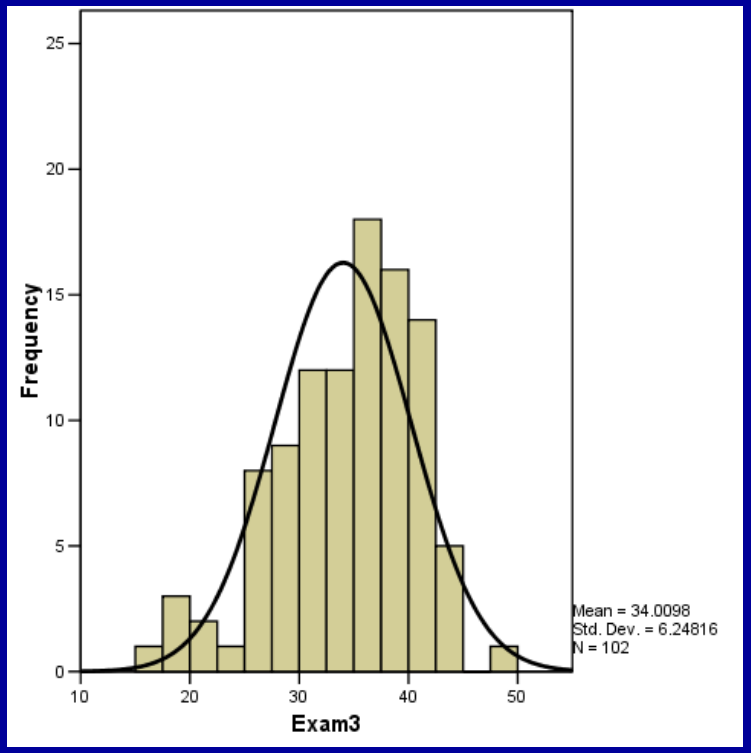


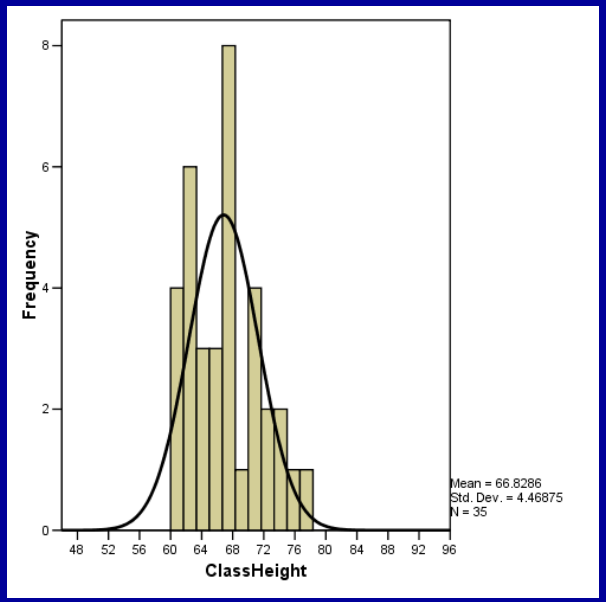
Figure 2.13 Distributions with More Than One Mode
Nolan and Heinzen: Statistics for the Behavioral Sciences, First Edition

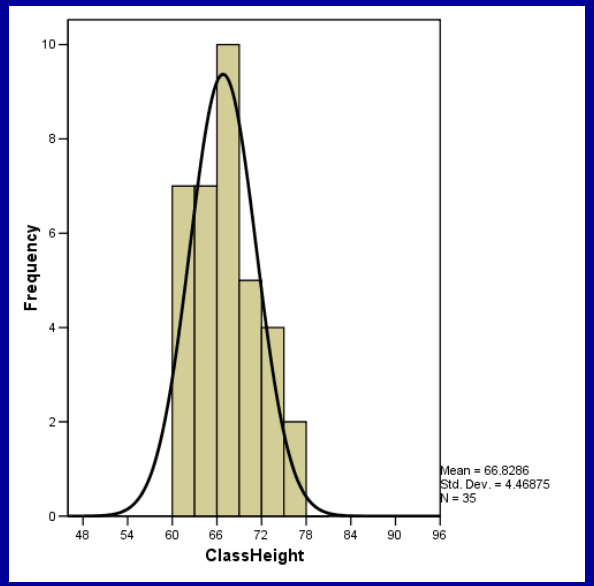
Histogram: Exam 1 & 3

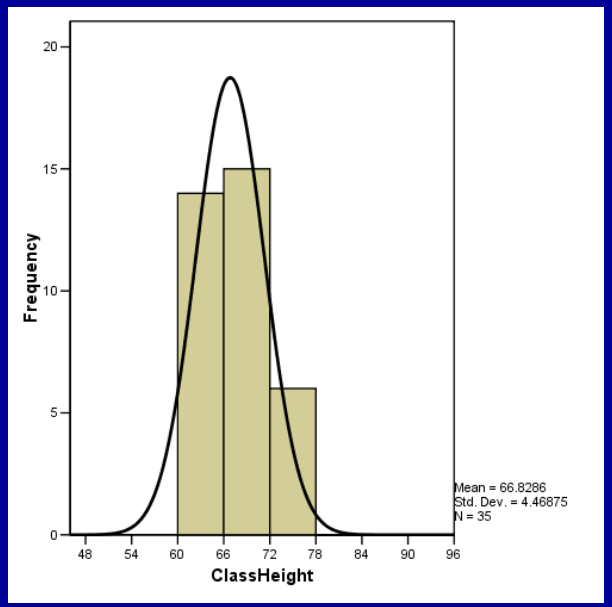




Histogram







- Bins. Both histograms show the same data, first has bins of 1.666, and second bins of 3, third bins of 6. Which conveys more information and which conveys information most clearly? Histogram changes, but normal curve stays more stable

2.2 The Science of Observation

- Variability
  - Range
  - Standard deviation
    - The "average range", the average of the distances of all points to the mean

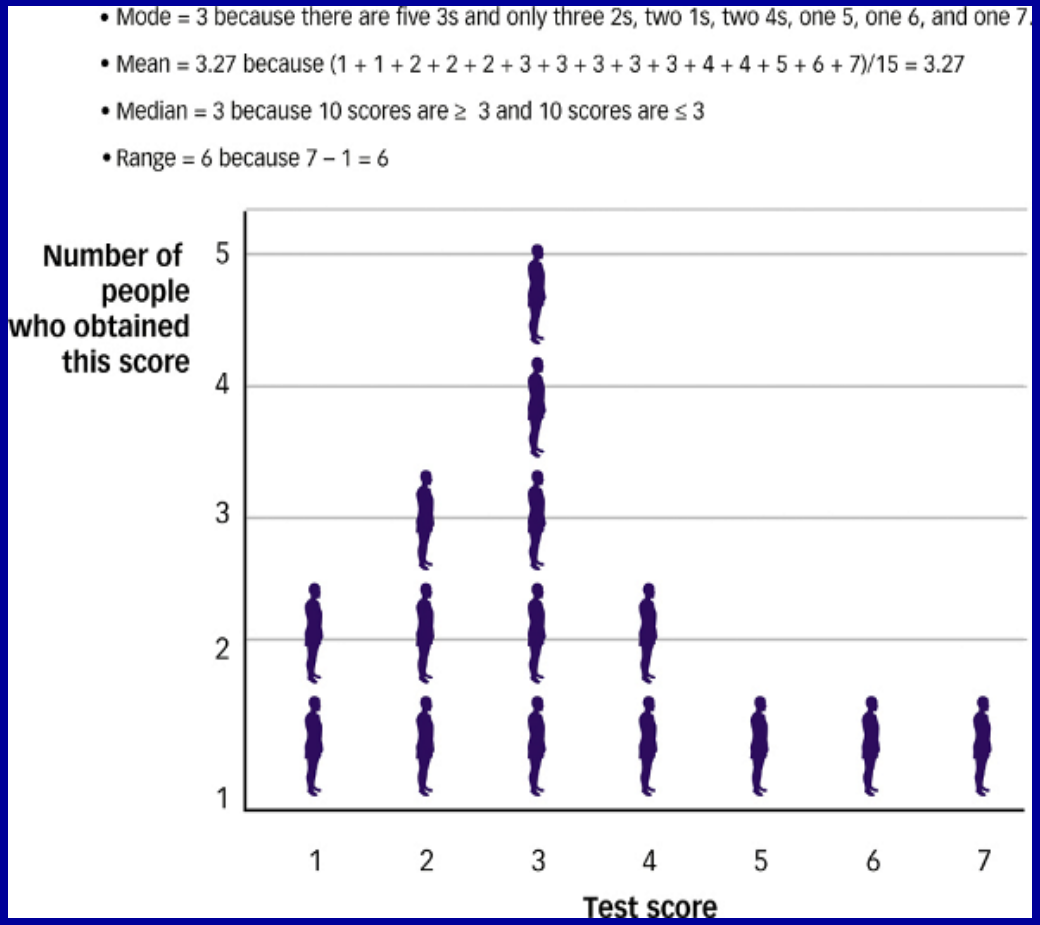


TABLE 2-2. FREQUENCY TABLES AND GRADUATE ADVISING

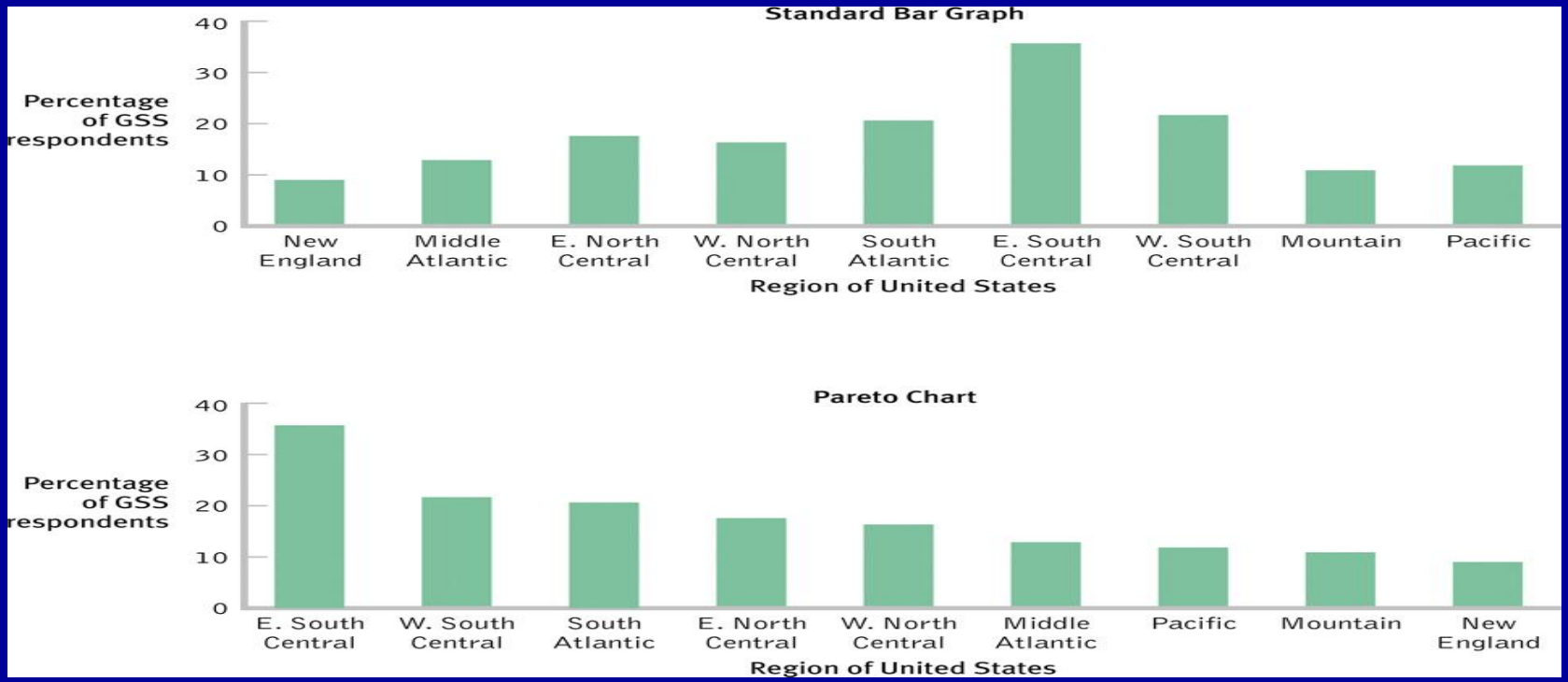
This frequency table depicts the numbers of students placed on the top 50 chemistry faculties by each top-producing graduate advisor. If you wanted a high-profile professorial job, which advisor would you want?

FORMER STUDENTS NOW IN TOP JOBS	FREQUENCY
13	1
12	0
11	0
10	0
9	1
8	3
7	4
6	5
5	9
4	8
3	23

Data from Kuck et al. (2007).

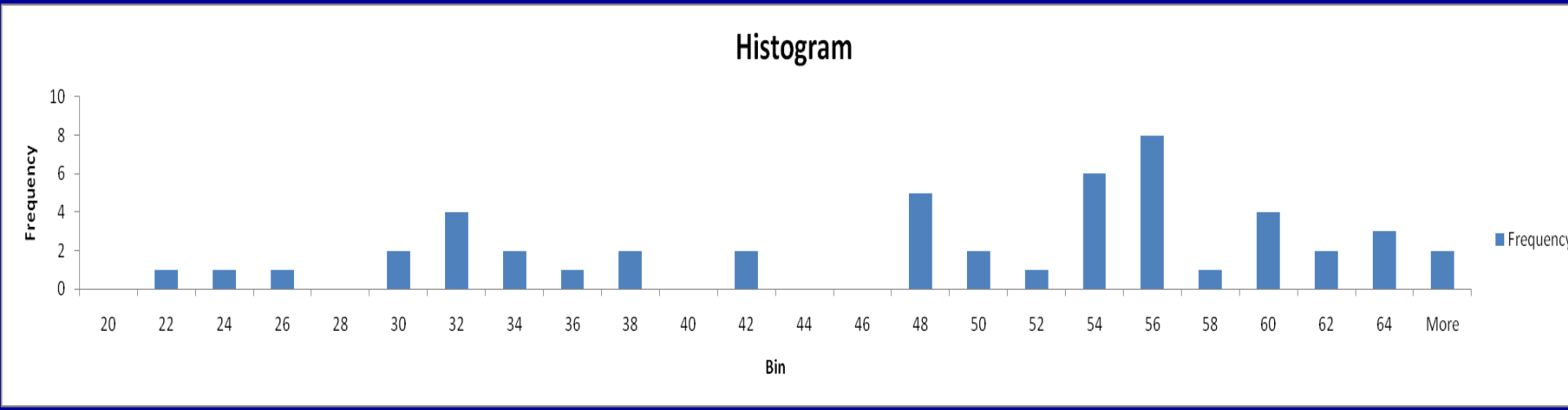
Table 2.2
Nolan and Heinzen: Statistics for the Behavioral Sciences, First Edition

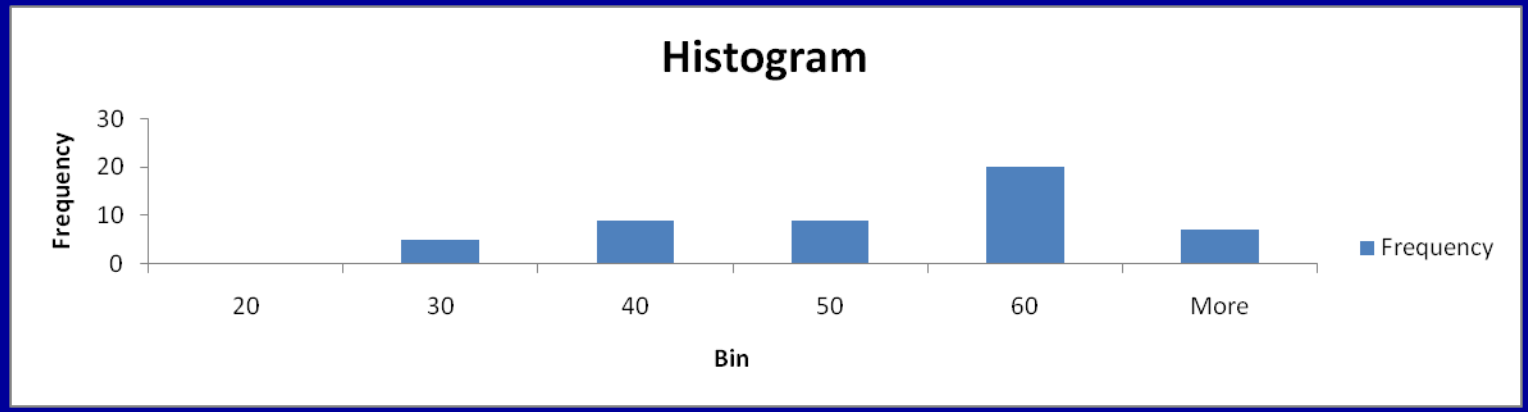
Special Histogram: Pareto Chart



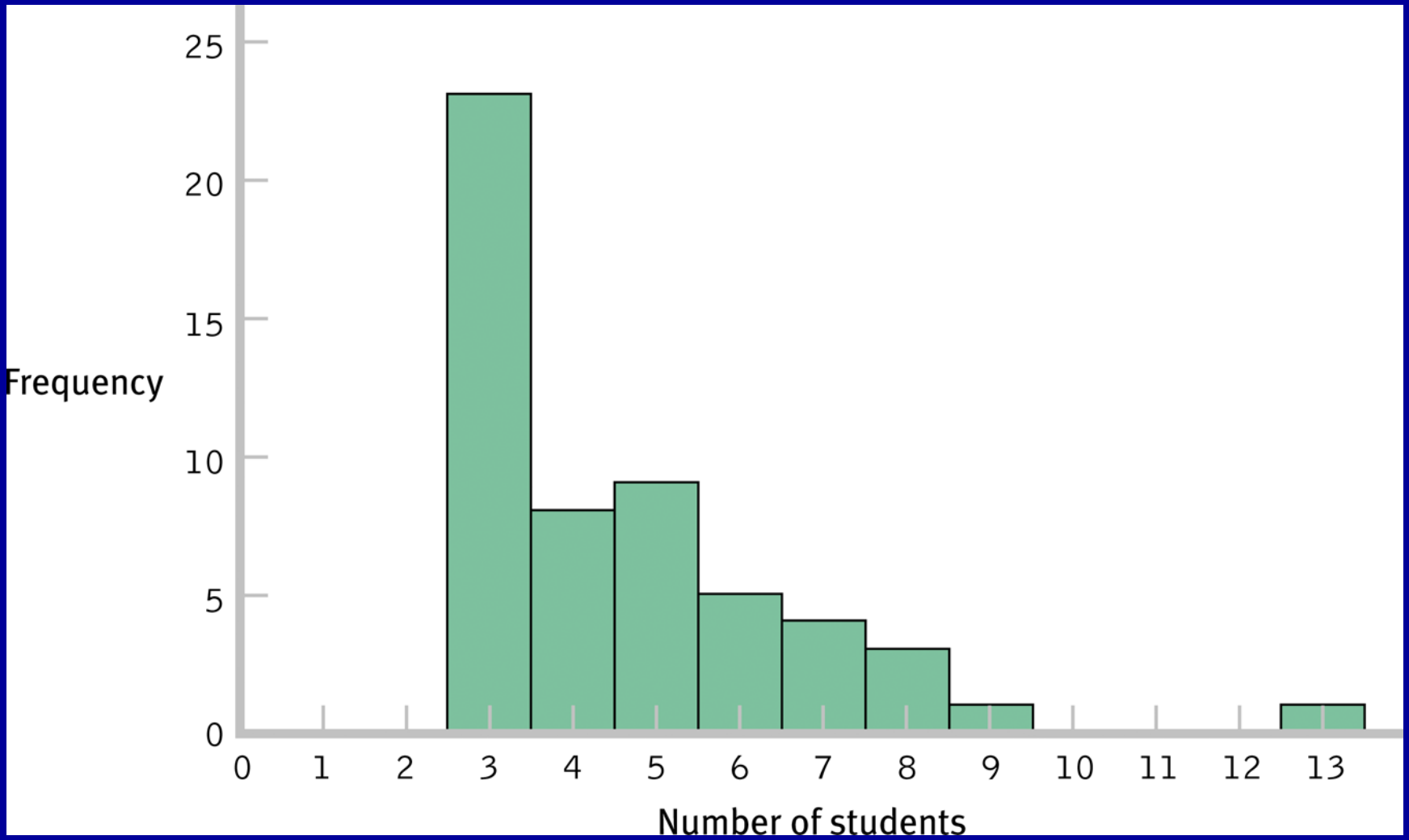
- Histograms show percent who said that avowed homosexuals were “not allowed” to speak.

Age Discrimination Case



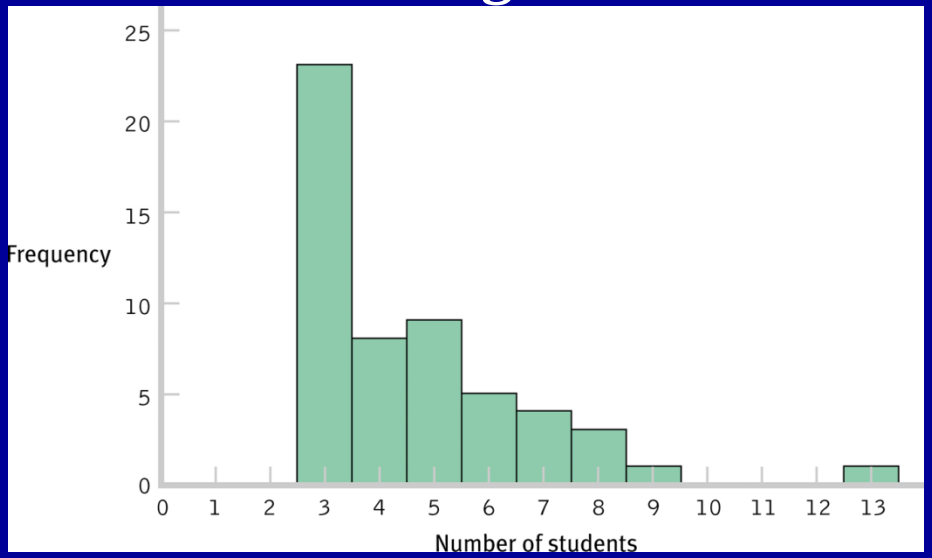


Histogram: Showing a Frequency Table

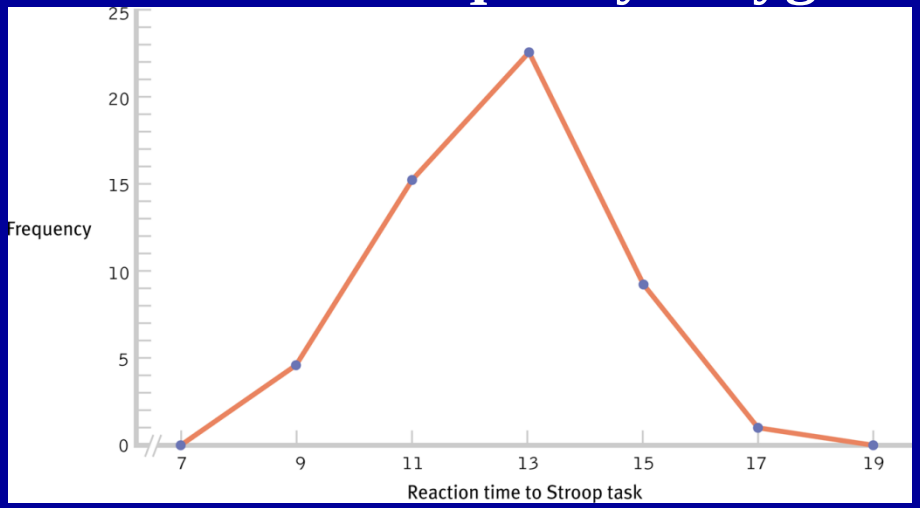


Graphing Frequency

Discrete: Histogram



Continuous: Frequency Polygon



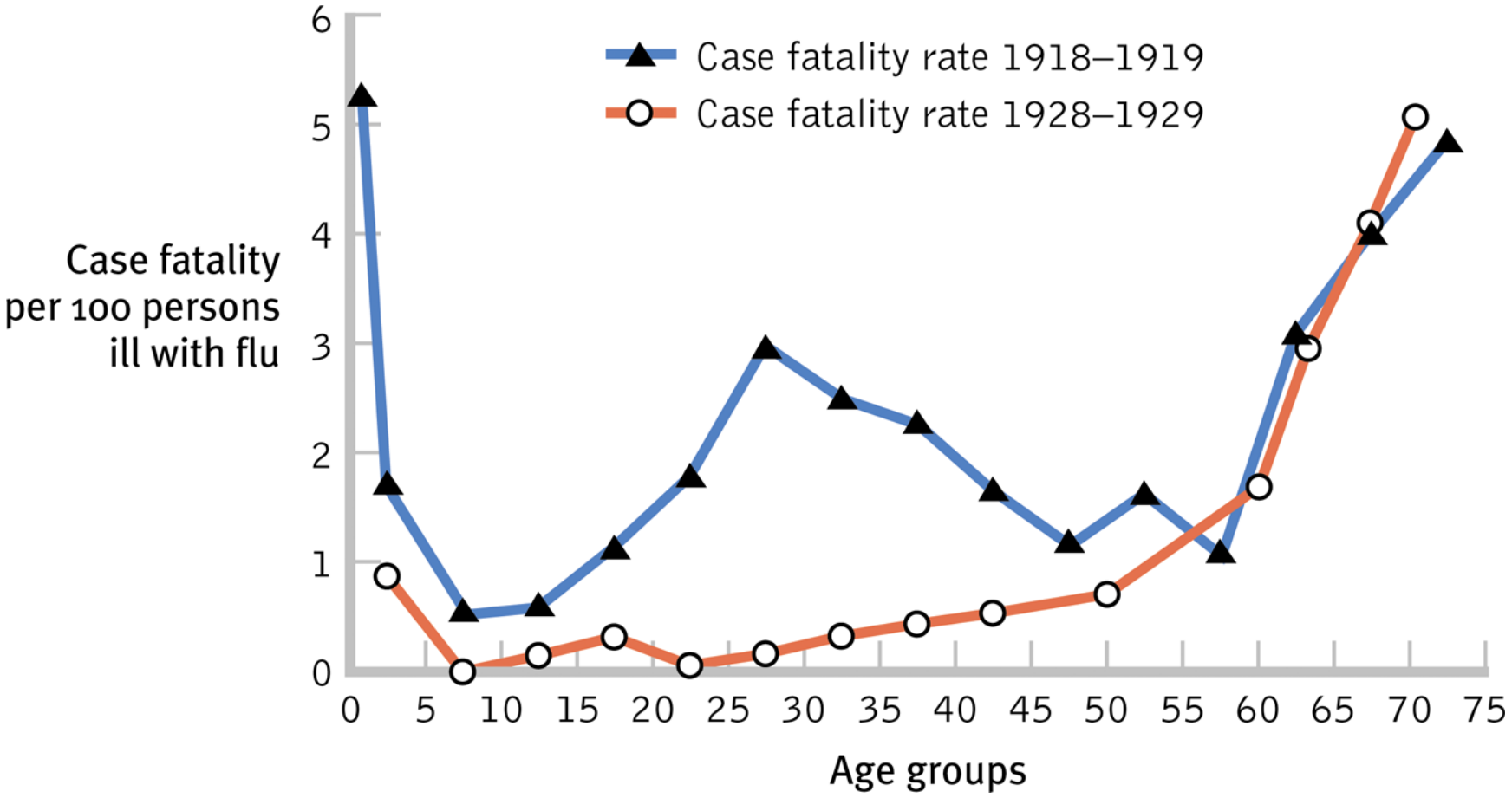


Figure 2.16 Skewed Distributions
Nolan and Heinzen: Statistics for the Behavioral Sciences, First Edition