Introduction to Statistical Analysis

Stat Bootcamp Autumn 2014
Session 1

Sema Barlas

What is Statistics?

- Statistics is a tool helping us to make intelligent decisions in the presence of uncertainty and variation.
- It is a tool to turn uncertainty into calculated risk.
- Decision: Conversion rates on a web page across regions are:
  - %13.8 %18.3 %32.2 %32.5
  - Probability of observing %32.5 conversion for a website at an ordinary region (OR) is 0.02 – so, this region is extra-ordinary
  - P(%18.3|OR) = 0.6 – so, this region is ordinary

What is Statistics?

- Communication – tell more with less
- Data on millions of website
- Summarizing data – descriptive statistics
  - Mean
  - Standard deviation
  - Distribution
  - Modality
  - Skewedness
  - Kurtosis

Variation

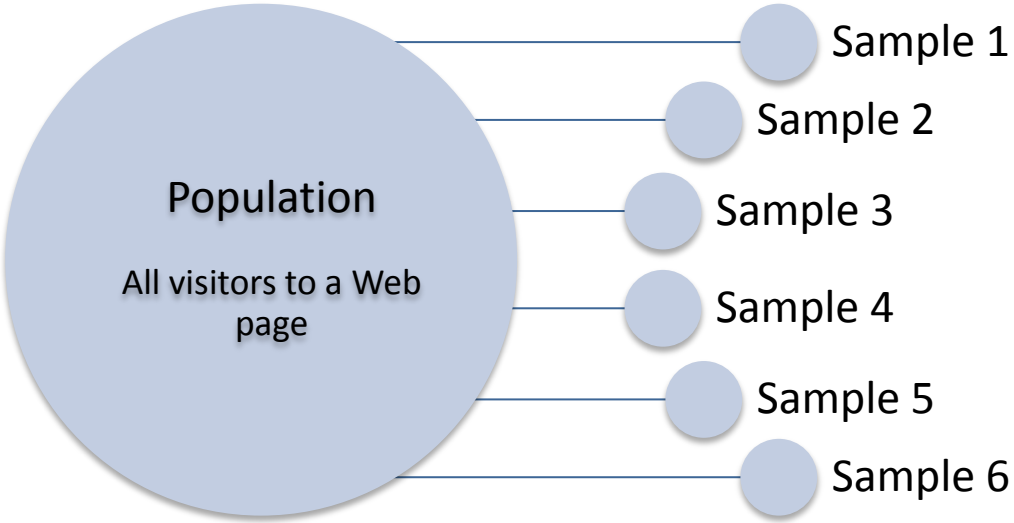
- Total variation = systematic variation + random Variation
  - Variation in conversion rates = variation among regions + variation within a region
- Random variation: Due to unknown sources or inherent to the process – variation within a region (error).
- Systematic variation: Due to known or knowable sources – variation across regions (explained variance).

Population, Sample, and Sampling Variance

Sampling Strategy

- Random – Samples 1 to 6 will be somewhat different
  - Sample is representative of the population depending on the size of sampling variance
- Stratified – Variation due to an important variable can be controlled for
  - Sample is unlikely to be representative of the population
- Probability – Sampling variation can be reduced
- Convenience - Sample is unlikely to be representative of the population

Census

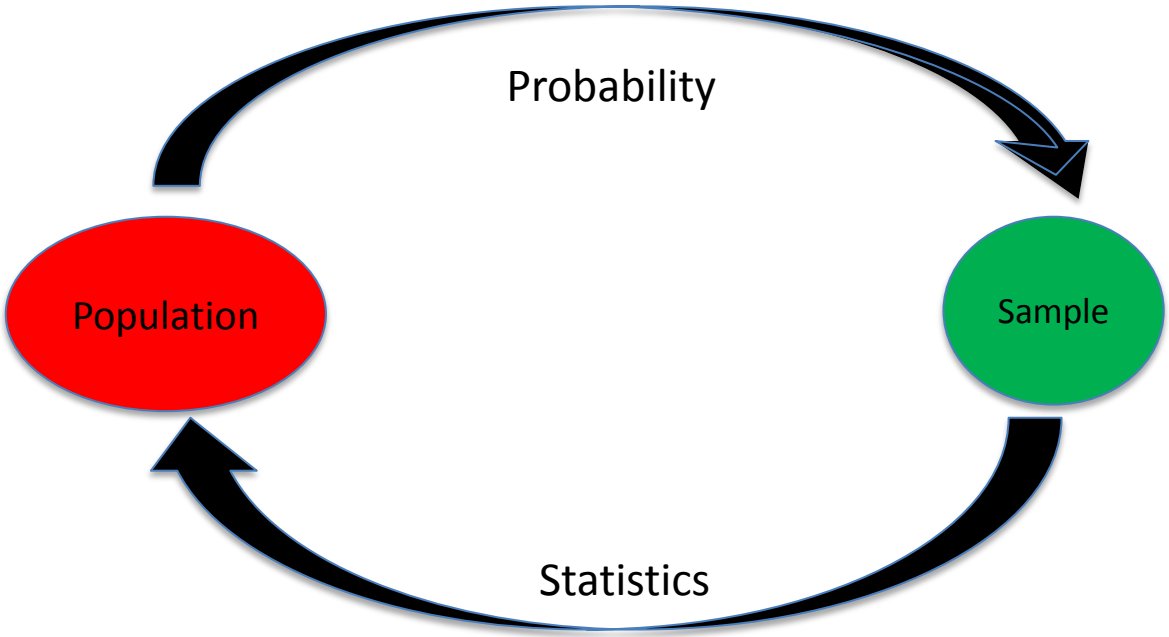


Types of Statistics

- Descriptive statistics
  - Understanding the sample- mean, median, mode, standard deviation, variance
- Probability
- Inferential statistics
  - Understanding the population on the basis of information from the sample – z test, t test, ANOVA, Regression

Does the Distribution in the Population Looks Like a Known Distribution?

Given that average hit rate is 15% in the population of ordinary websites.
How many of such sites in a sample of 100 can we expect to have %32 or more suspicious hits?

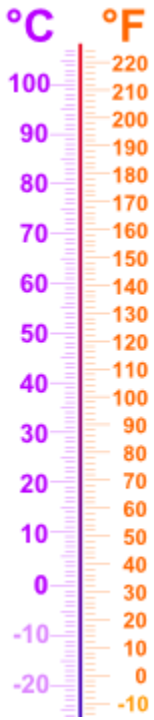


Given that average hit rate is %18.5 in the sample of 100 websites.
Is this sample coming from the population of ordinary websites?

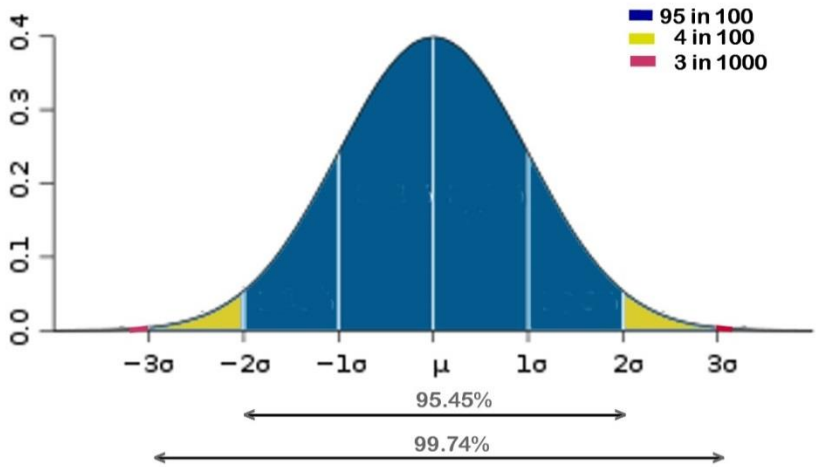
Variables

- Characteristics of some event, object, or person that take on different values (has variability)
  - Dependent variable – sales ($)
  - Independent variable – marketing expenditures

- Discrete
  - Nominal - race
  - Ordinal - latter grade
- Continuous
  - Interval - Fahrenheit
  - Ratio – weight in kilograms



Distributions of Variables



Normal distribution

Central tendency: Mean, median mode

Dispersion: Standard deviation, variance

Shape - Symmetric (skewness = 0)

Uni-modal

Mezokurtic – Kurtosis = 0

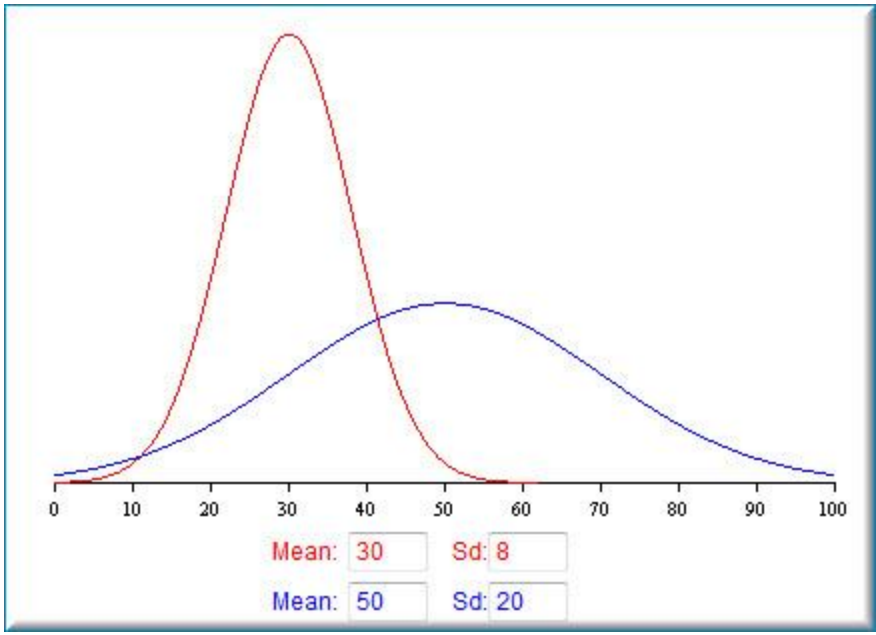
Central Tendency Measures

Stem	Leaf
0	1 3 6
1	2 8 8 8
2	3 5 6 7

Mean: 16.09 - Ratio, interval, and ordinal level variables
Median: 18 – the observation at the middle – Ratio, Interval, and ordinal variables
Mode: 18 – the most frequent observation – Ratio, interval, ordinal, and nominal variables

Dispersion Measures

Ratio, Interval, and ordinal variables



- Variance
  - *var = sum((x- μ)² / (N-1)))*
  - *87.29 in the example before*

- Standard Deviation
  - average distance from mean
  - Std = sqrt(var) = 9.34

Test Scores

Score	Score - Mean	Mean Square
90	20	400
80	10	100
80	10	100
70	0	0
60	-10	100
40	-30	900
420	0	1,600

Mean = 70
Median = 75
Mod = 80

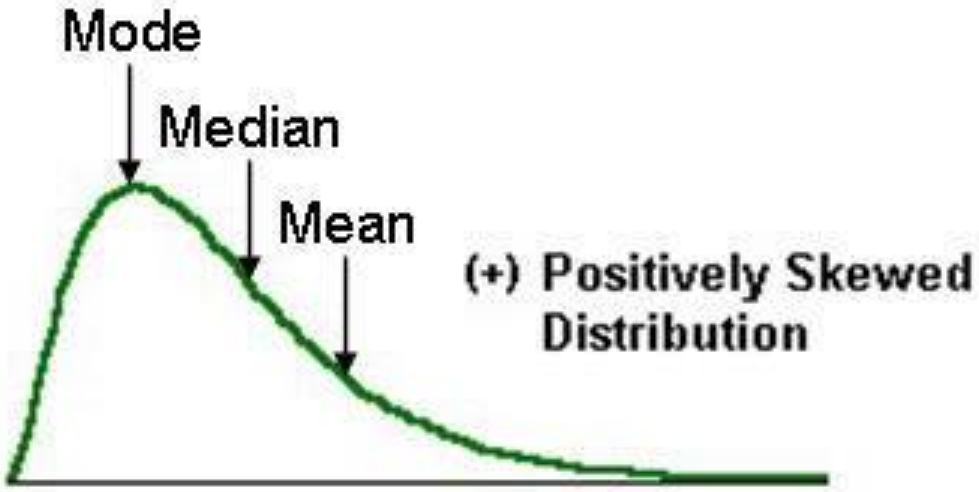
$$Variance = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

= 1,600/5
= 320

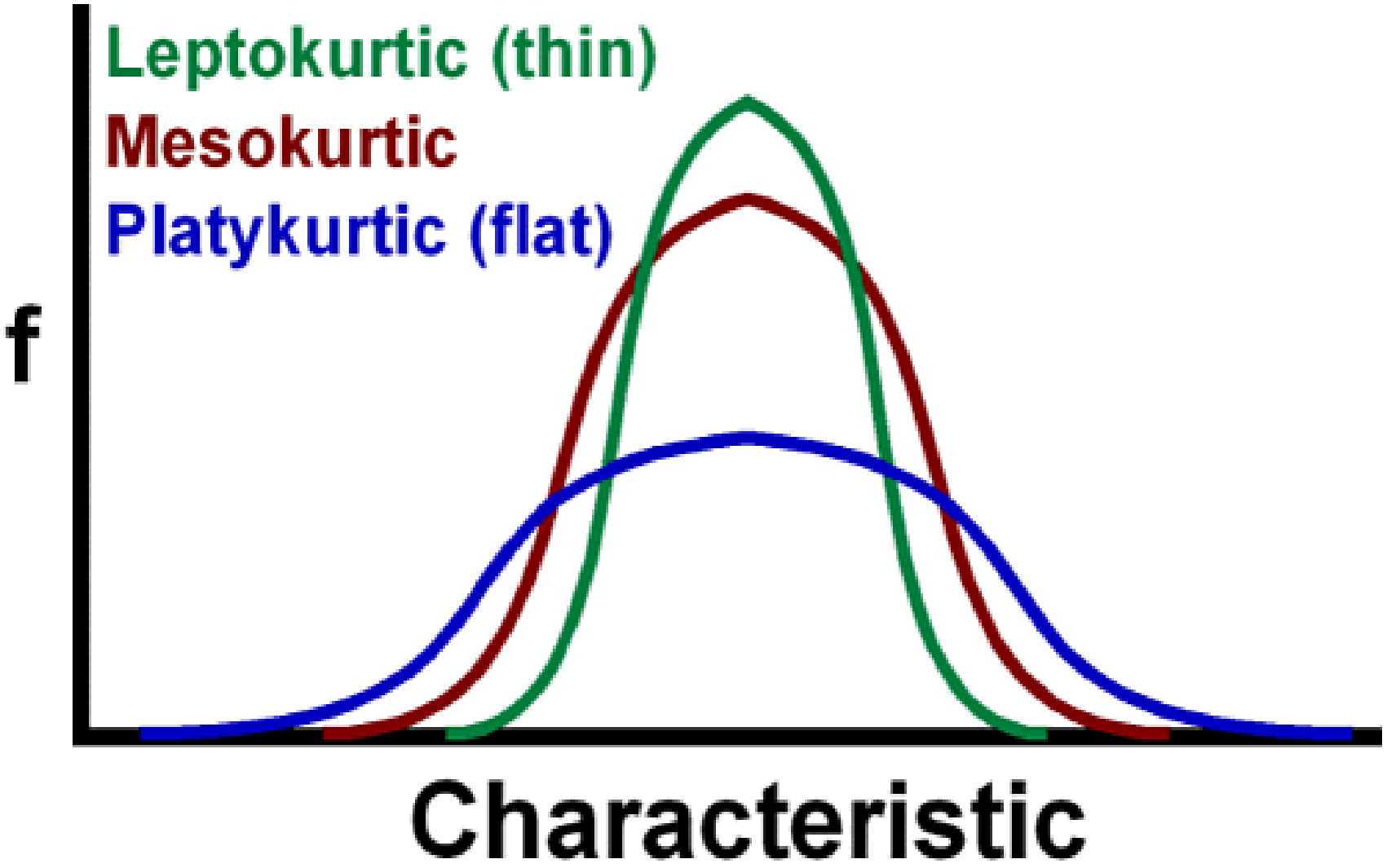
Std = $\sqrt{320}$ = 17.89
Range = highest – lowest = 90 – 40 = 50
IQR = Q3 – Q1 = 77.5 - 52.5 = 25

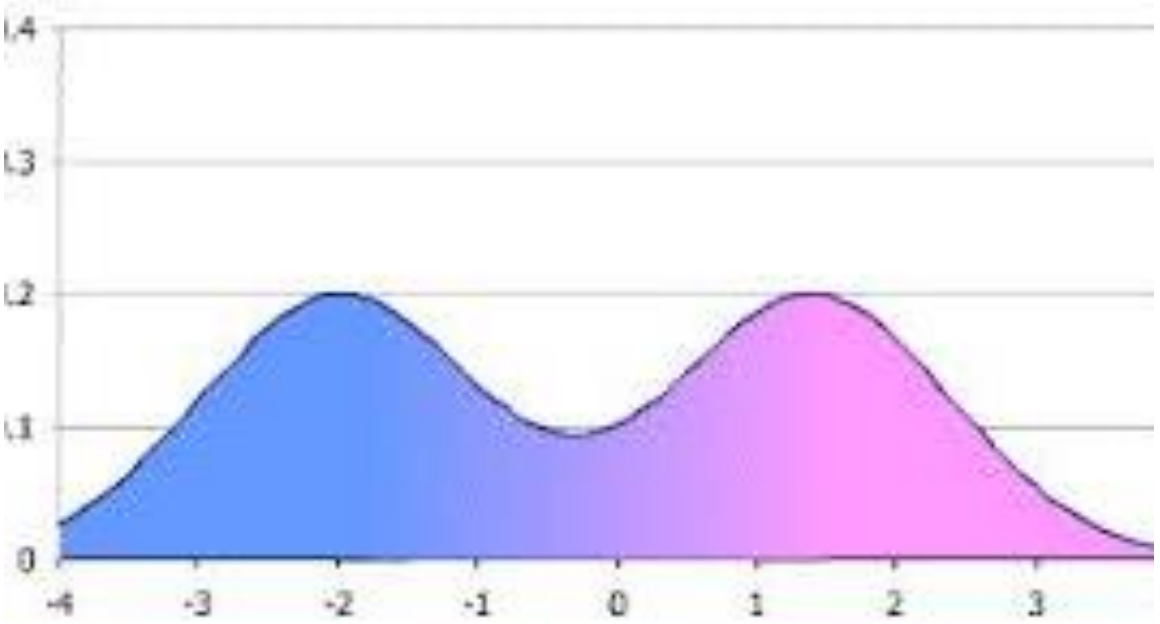
Skewness


Mode
Median
Mean
(-) Negatively Skewed
Distribution

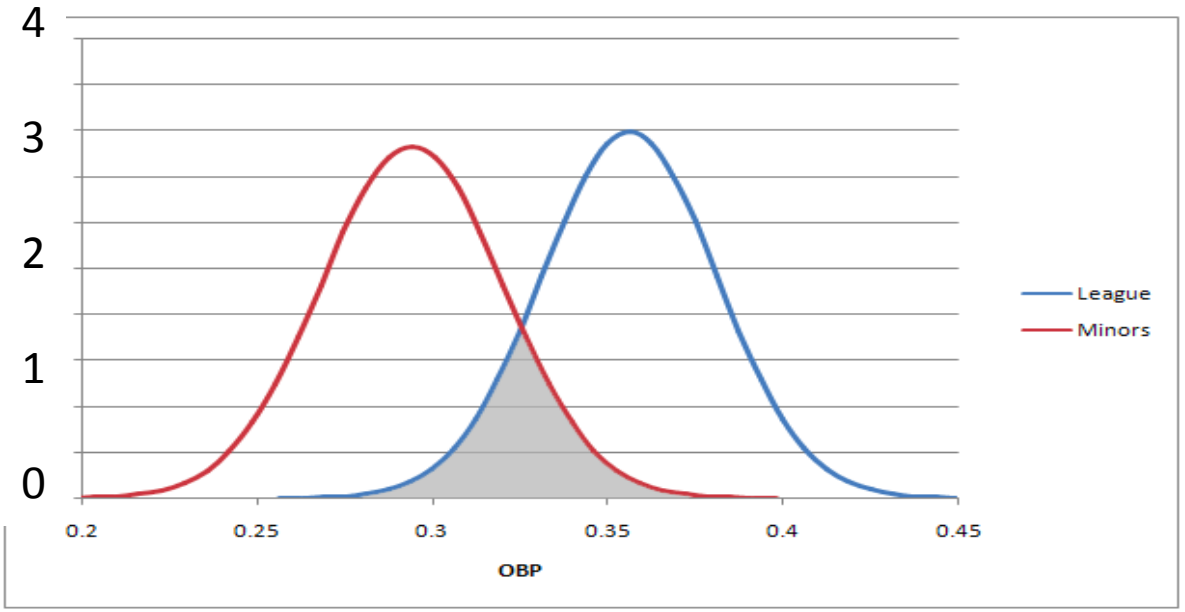

Mode
Median
Mean
(+) Positively Skewed
Distribution

Kurtosis


Leptokurtic (thin)
Mesokurtic
Platykurtic (flat)
f
Characteristic

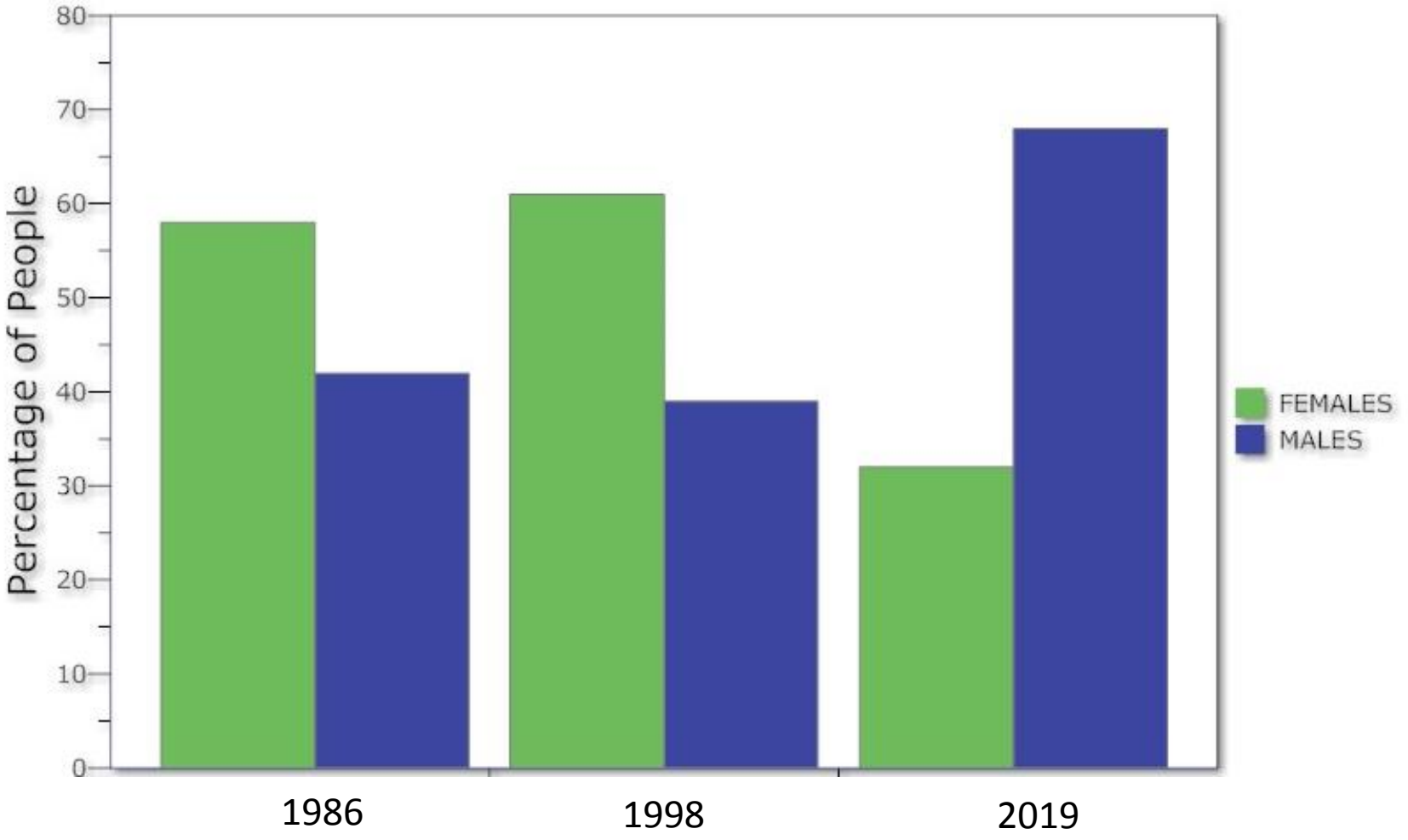


Bimodal Distribution

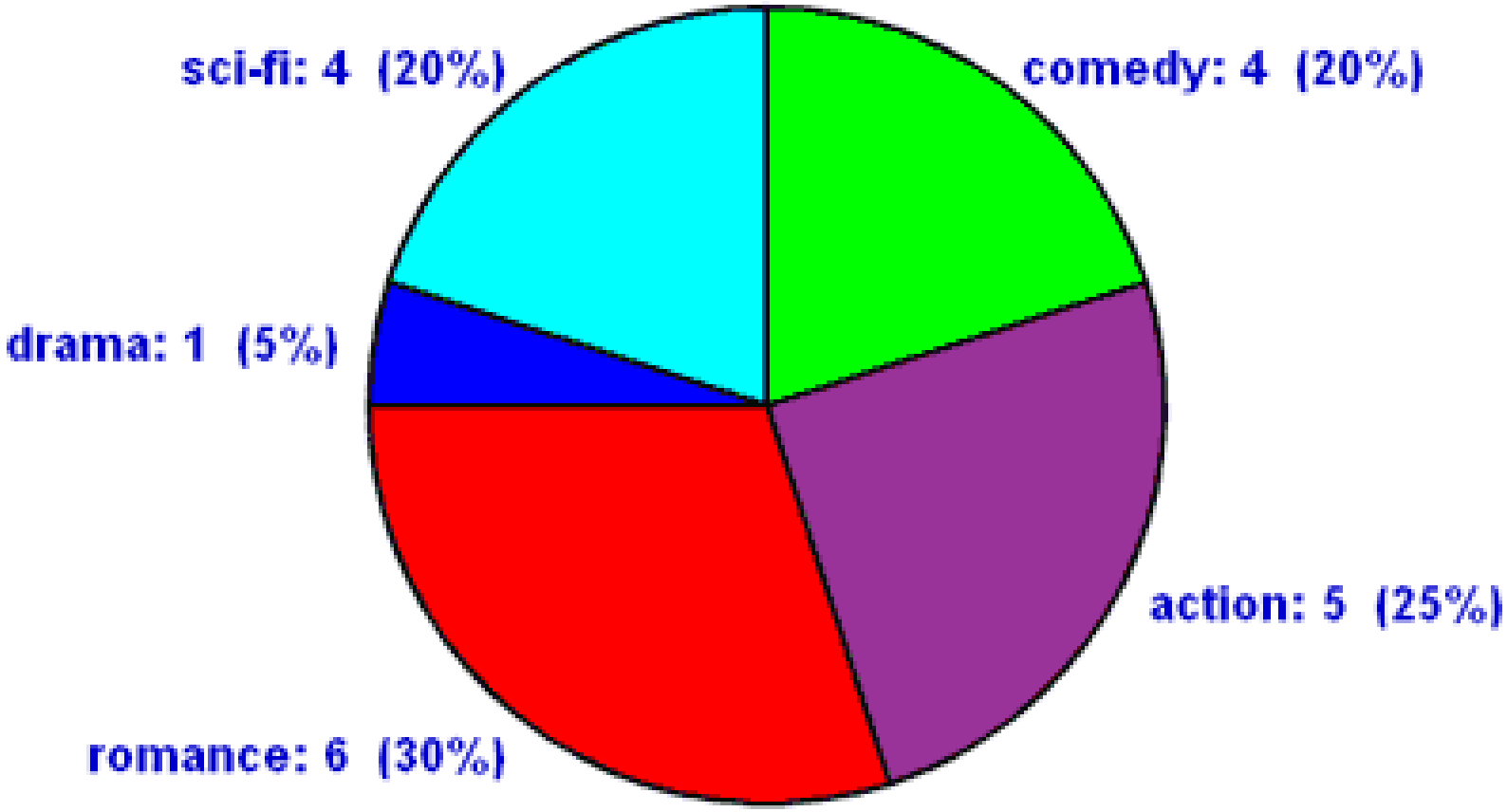


Nominal Variables – Bar Chart

Future Wealth Holder's Gender Shift



Nominal Variables – Pie Chart



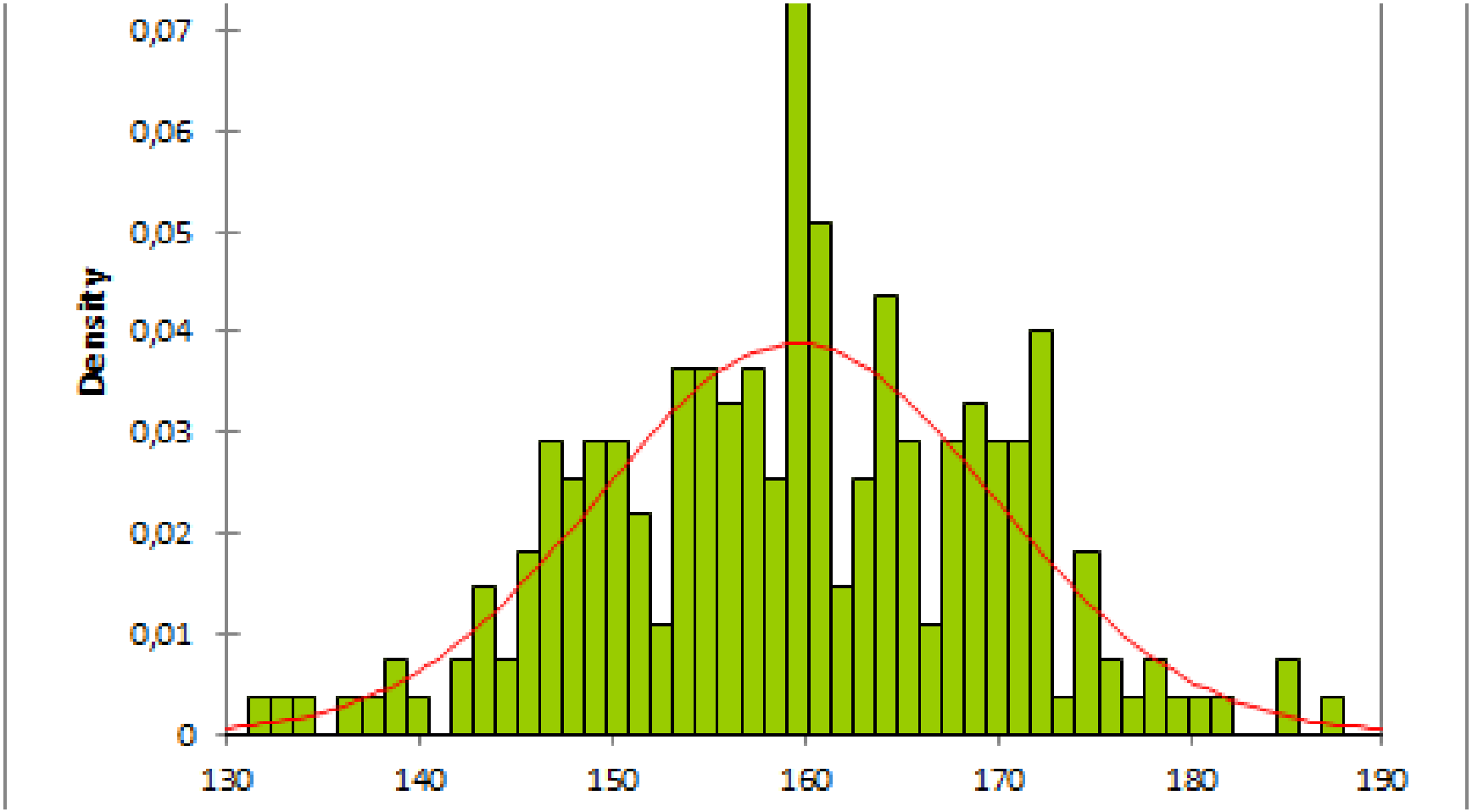
Frequency Tables

Ratings	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency
2	1	1	.006	.006
3	2	3	.011	.017
4	13	16	.074	.091
5	45	61	.256	.347
6	33	94	.187	.534
7	56	150	.318	.852
8	21	171	.119	.972
9	5	176	.028	100

N=176 Mean = 6.18 Mod = 7 median = 6 std = 1.33

Central Tendency Variability

Continuous Variables - Histogram



Linear Transformation of Variables

- $Z = a + b*X$
- $\bar{z} = a + b*\bar{x}$
- $Std_z = b*(std_x)$

X	z=2+2X	$(x\text{-mean})^2$
90	182	1600
80	162	400
80	162	400
70	142	0
60	122	400
40	82	3600
Mean	Mean	std
70	142	35.77

Two Variables, X and Y

- Covariance between X and Y

$$Covariance = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x}) \ (y_i - \bar{y})$$

$\bar{x}$

$\bar{y}$

X is below and y is above the mean	Both X and Y are above mean
Both, x and y are below the mean	X is above and y is below the mean

- Correlation between X and Y

$$Correlation = \frac{Cov(x,y)}{std(x)std(y)}$$

Covariance is about the Direction of the Relationship

x	y	$(x - \bar{x})$	$(y - \bar{y})$	Covariance
23	11	1.4	0.8	1.12
20	9	-1.6	-1.2	1.92
14	4	-7.6	-6.2	47.12
27	15	5.4	4.8	25.92
22	10	0.4	-0.2	-0.08
20	11	-1.6	0.8	-1.28
26	11	4.4	0.8	3.52
16	7	-5.6	-3.2	17.92
25	13	3.4	2.8	9.52
23	11	1.4	0.8	1.12

Mean: 21.6 10.2

Std: 4.20 3.05

Sum: 106.8

Cov: 11.86

Corr: 0.98

Optimize Publisher Strategy—Results