Introduction to Statistical Analysis

Stat Bootcamp
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

Manage Variation Concurrently and Proactively

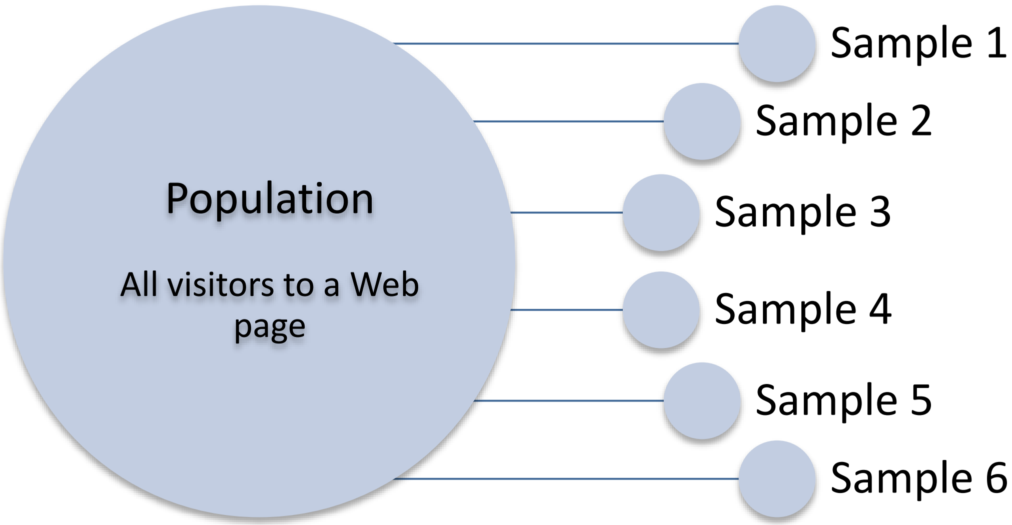
- Total variation in the attribute = systematic variation + random Variation
  - Variation in conversion rates = variation among regions + variation within a region
- Random variation: Due to unknown sources or inherent to the process – variation within a region (error).
- Systematic variation: Due to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

Variation due to Sampling from Population 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

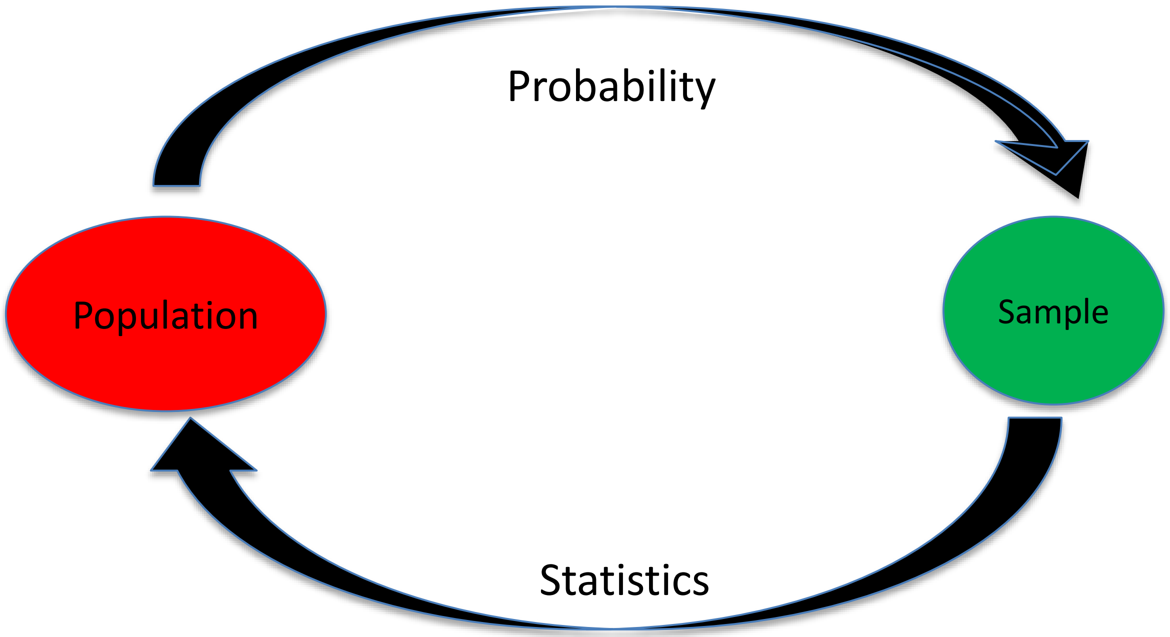


Types of Statistics

- 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 conversion 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 conversion rate?



Probability

Population

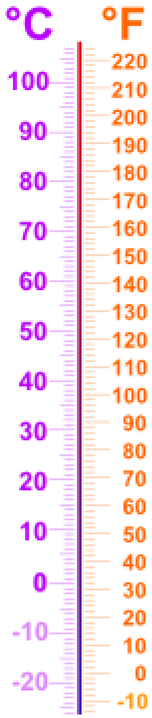
Sample

Statistics

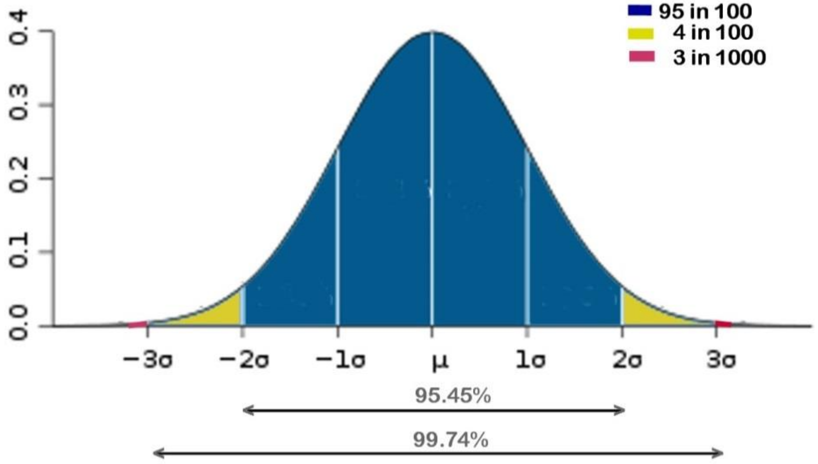
Given that average conversion rate rate is 18.5% in the sample of 100 websites.
Is this sample coming from the population of ordinary websites?

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 - letter grade
- Continuous
  - Interval - Fahrenheit
  - Ratio – weight in kilograms



Distributions of Variables



Normal distribution

Central tendency: Mean, median, mode

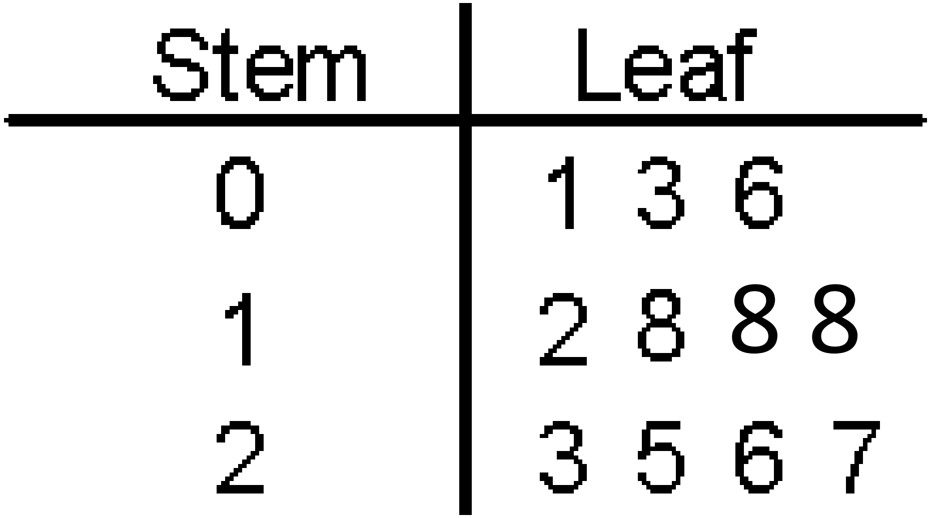
Dispersion: Standard deviation, variance

Shape - Symmetric (skewness = 0)

Uni-modal

Mezokurtic (Kurtosis = 0)

Central Tendency Measures

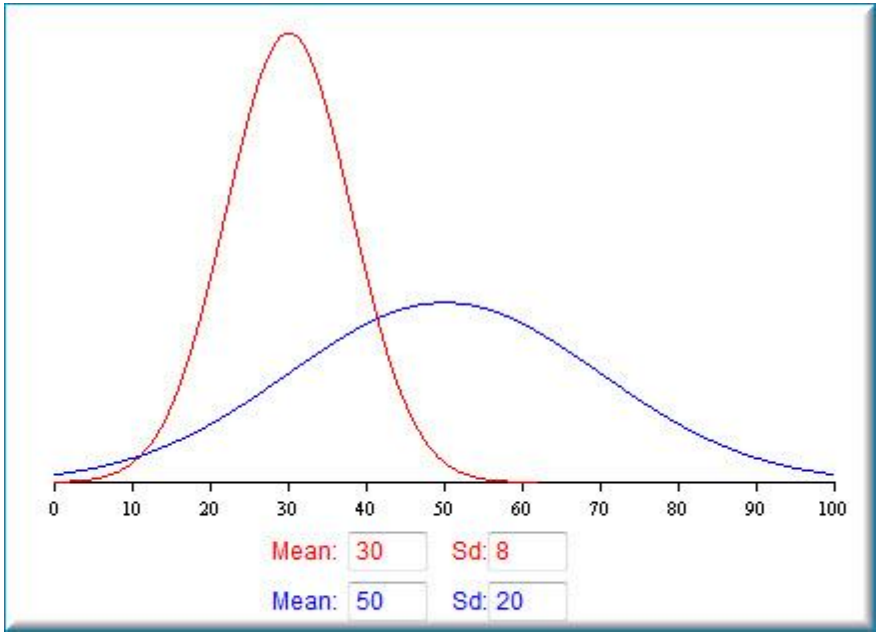


Stem	Leaf
0	1 3 6
1	2 8 8 8
2	3 5 6 7

N=11

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



Mean: 30 Sd: 8

Mean: 50 Sd: 20

- Variance
  - *$var = sum((x- \mu)^2 / (N-1)))$*
  - *87.29 in the example before*

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

Test Scores

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

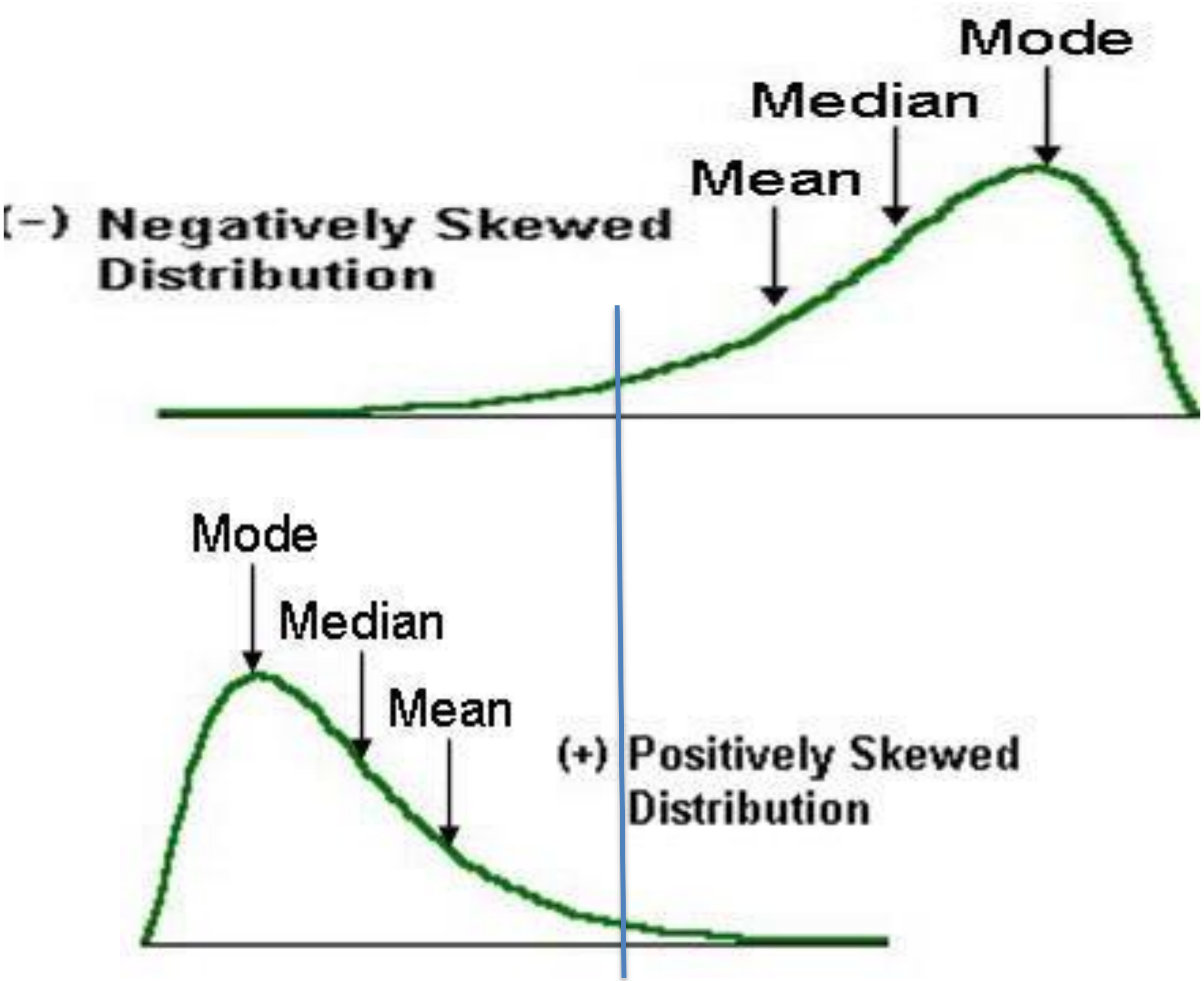
Mean = 70
Median = 75
Mod = 80

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

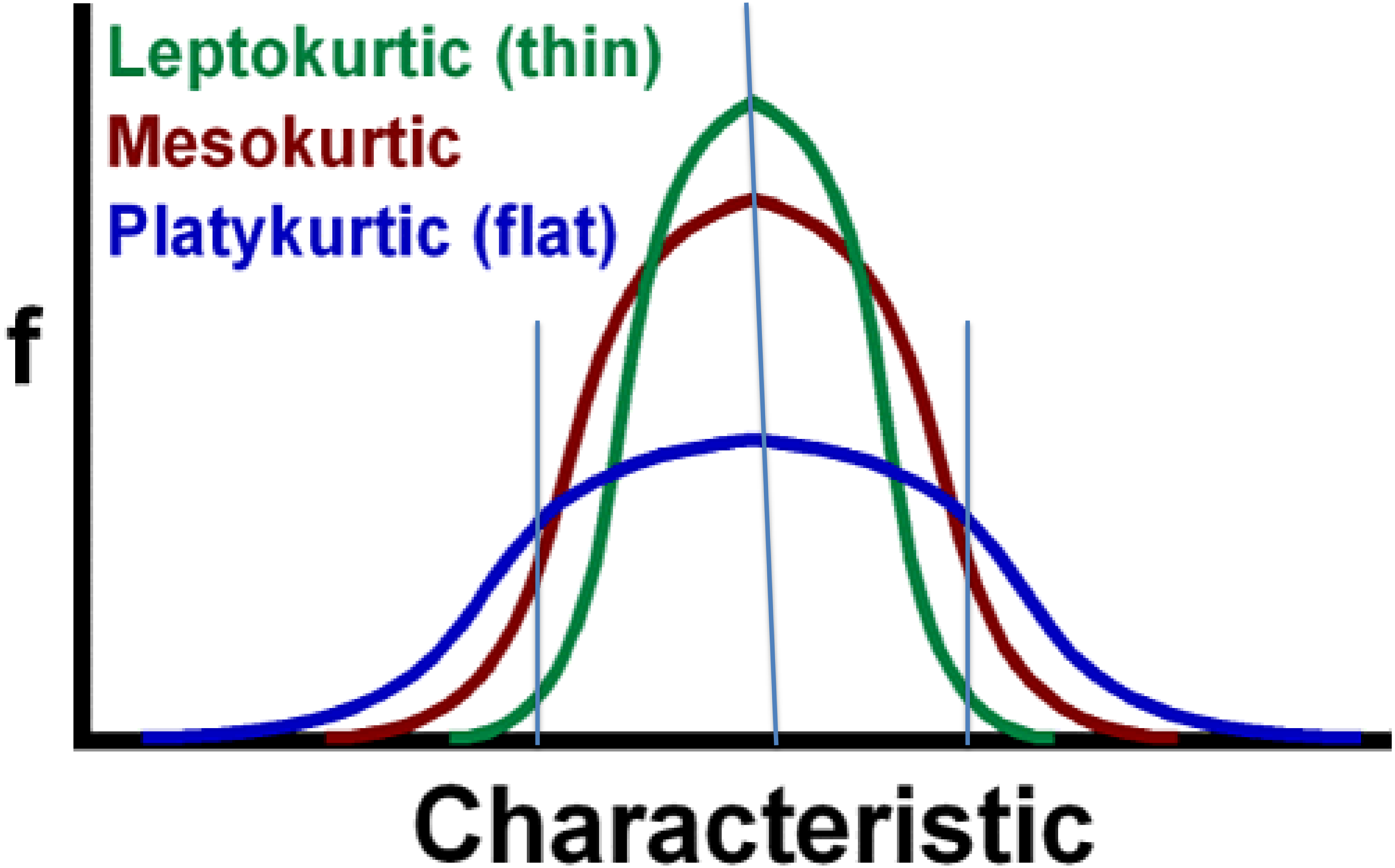
= 1,600/5
= 320

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

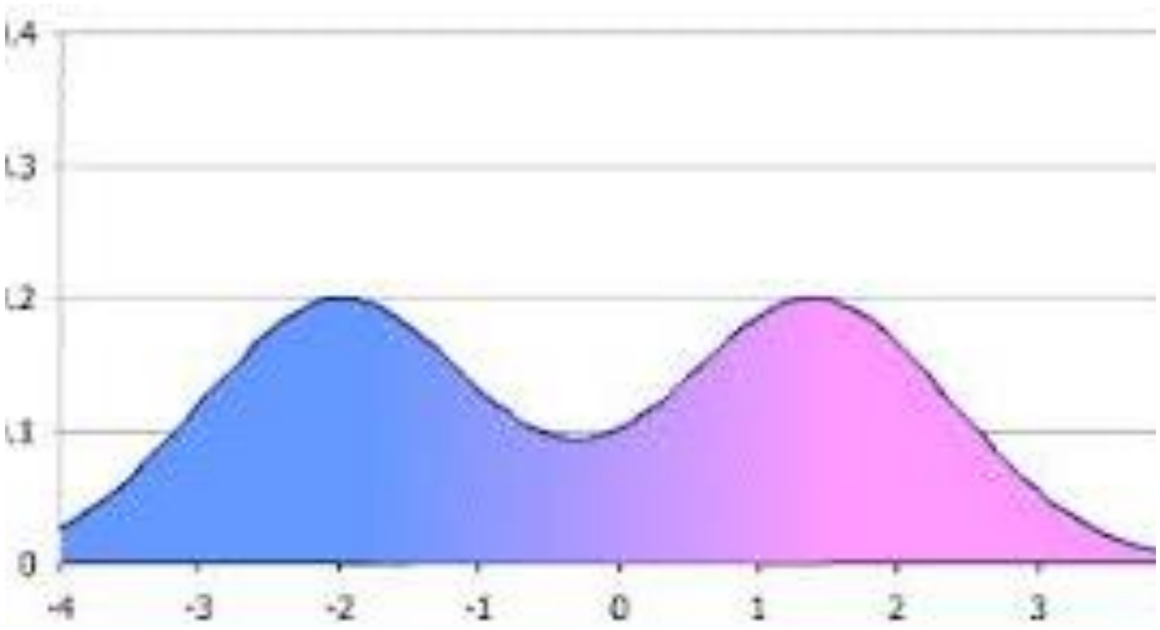
Skewness

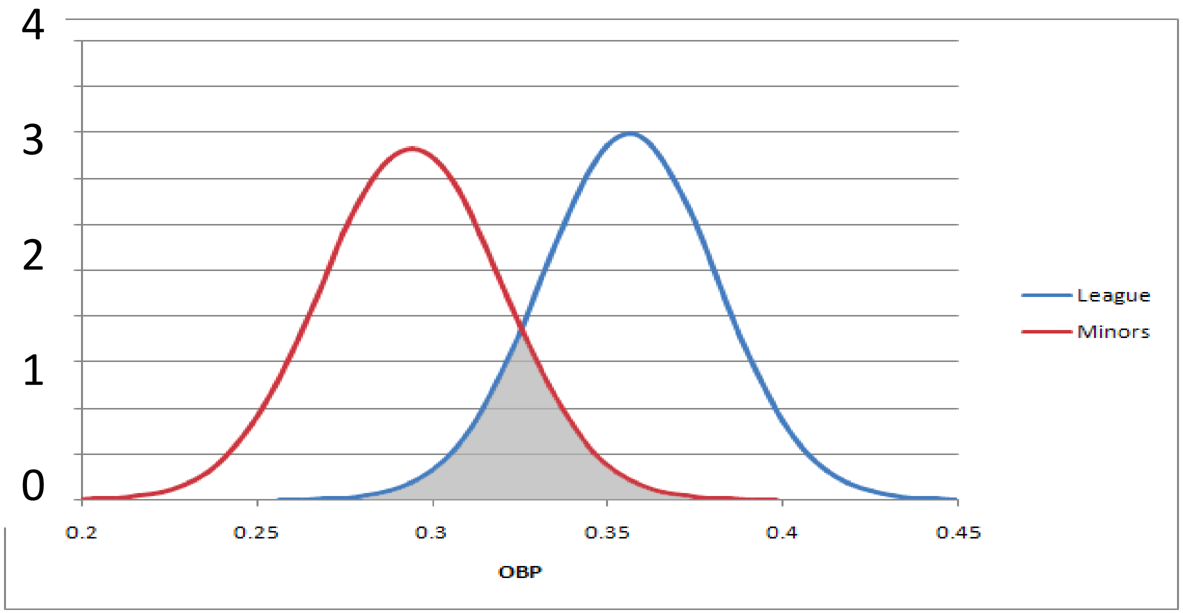

Mode
Median
Mean
(-) Negatively Skewed Distribution
Mode
Median
Mean
(+) Positively Skewed Distribution

Kurtosis


Leptokurtic (thin)
Mesokurtic
Platykurtic (flat)
f
Characteristic

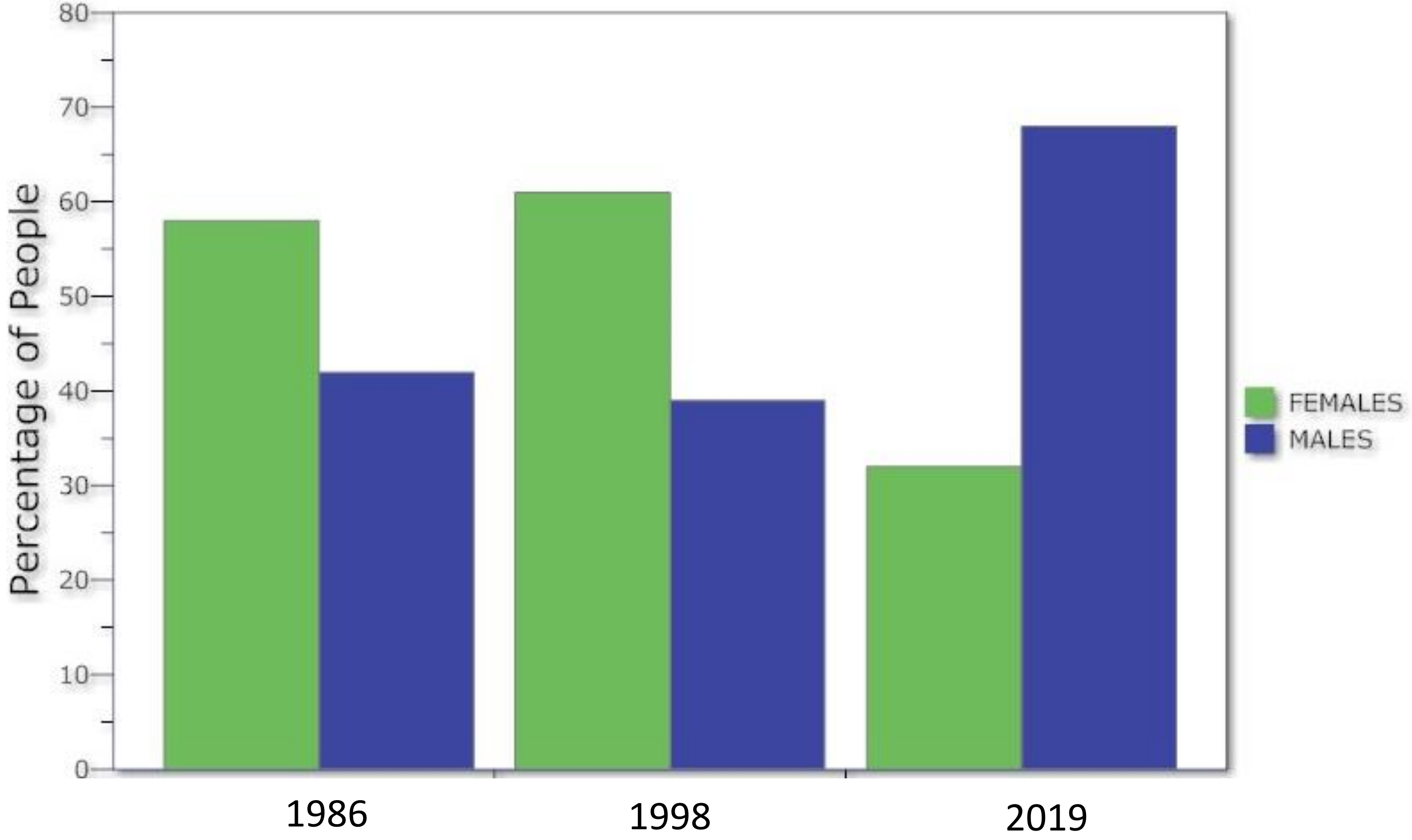
Bimodal Distribution



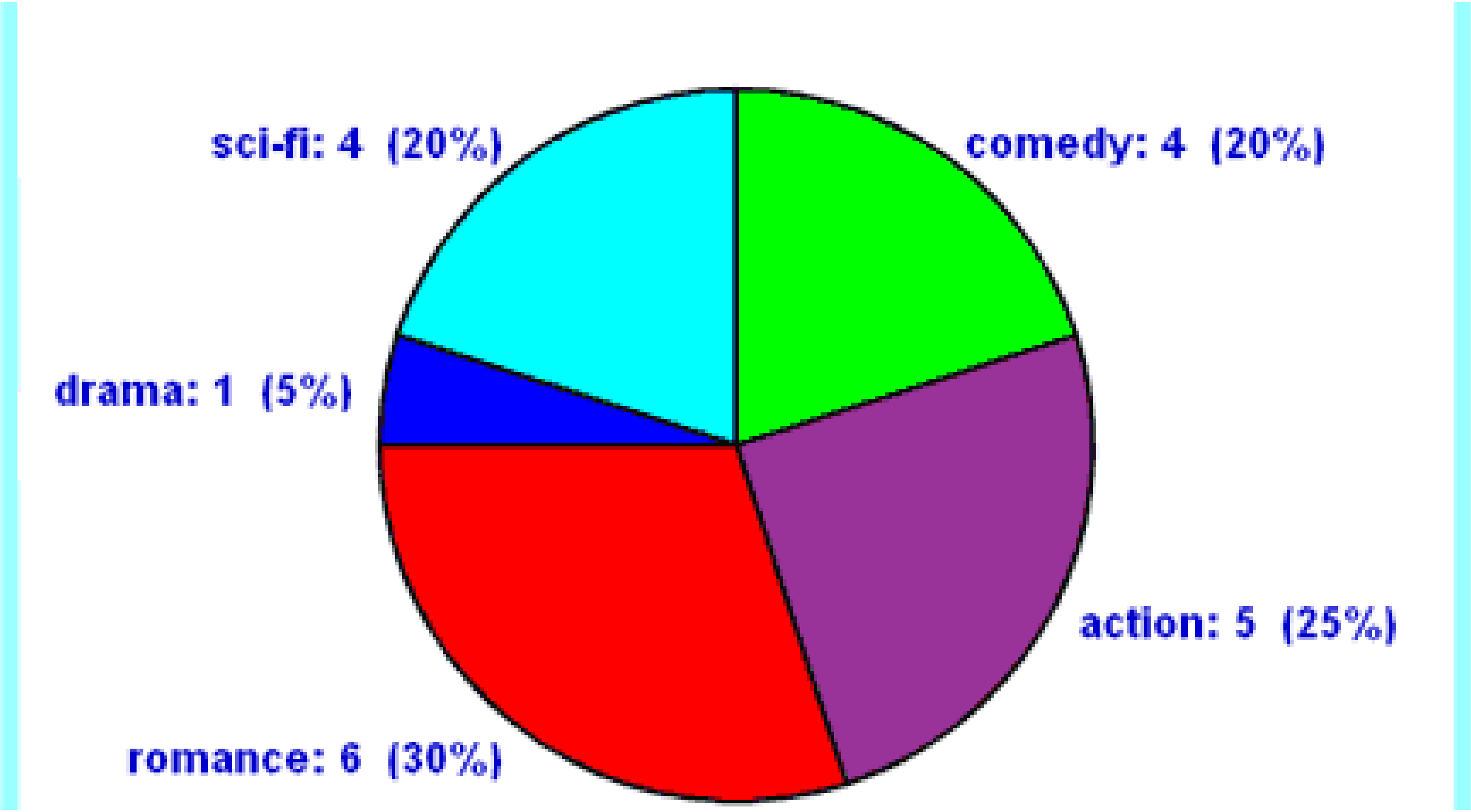


Nominal Variables – Bar Chart

Future Wealth Holder’s Gender Shift



Nominal Variables – Pie Chart



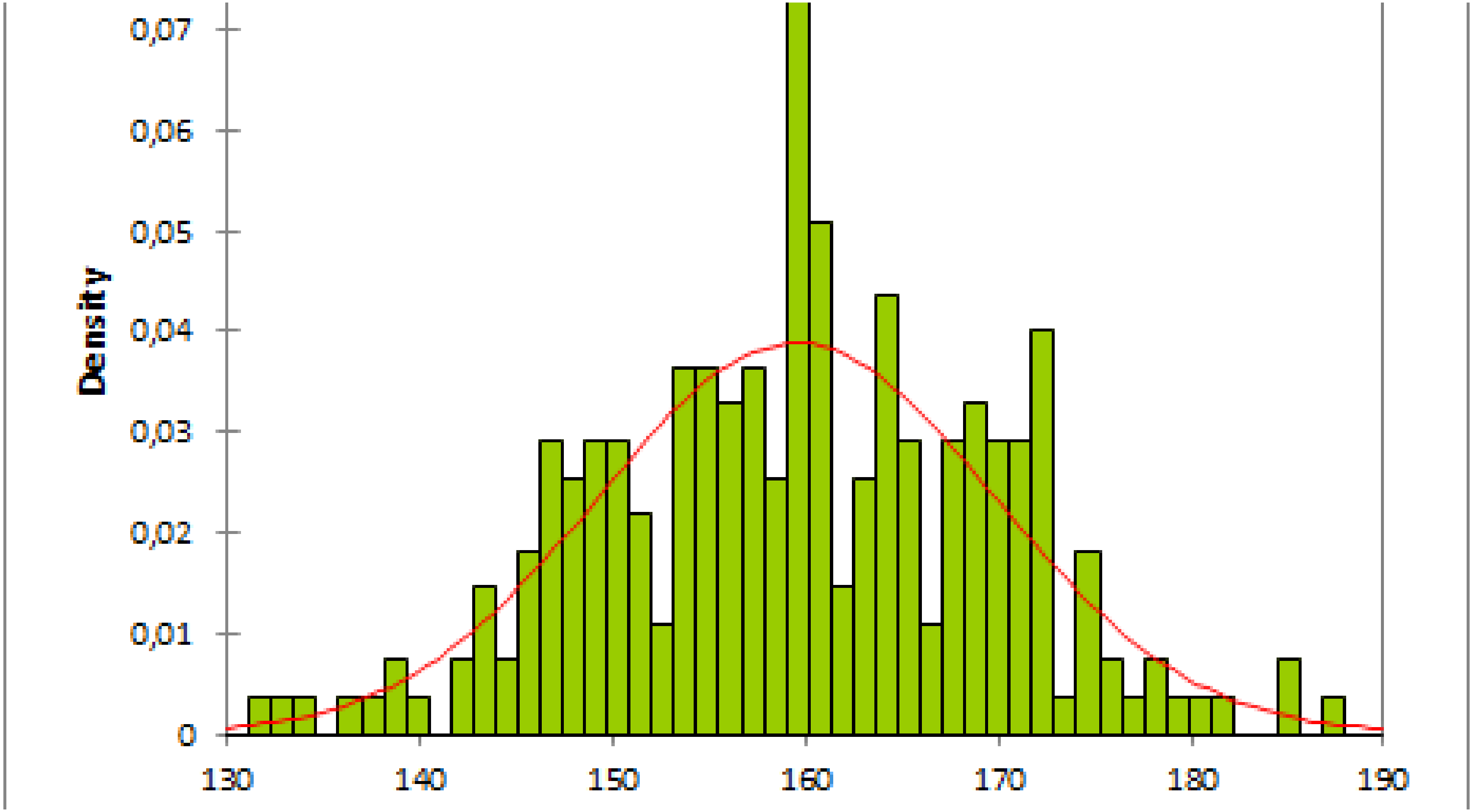
Frequency Tables

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	$(x-70)^2$	$(z-142)^2$
90	182	400	1600
80	162	100	400
80	162	100	400
70	142	0	0
60	122	100	400
40	82	900	3600
Mean	Mean	std	std
70	142	17.90	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

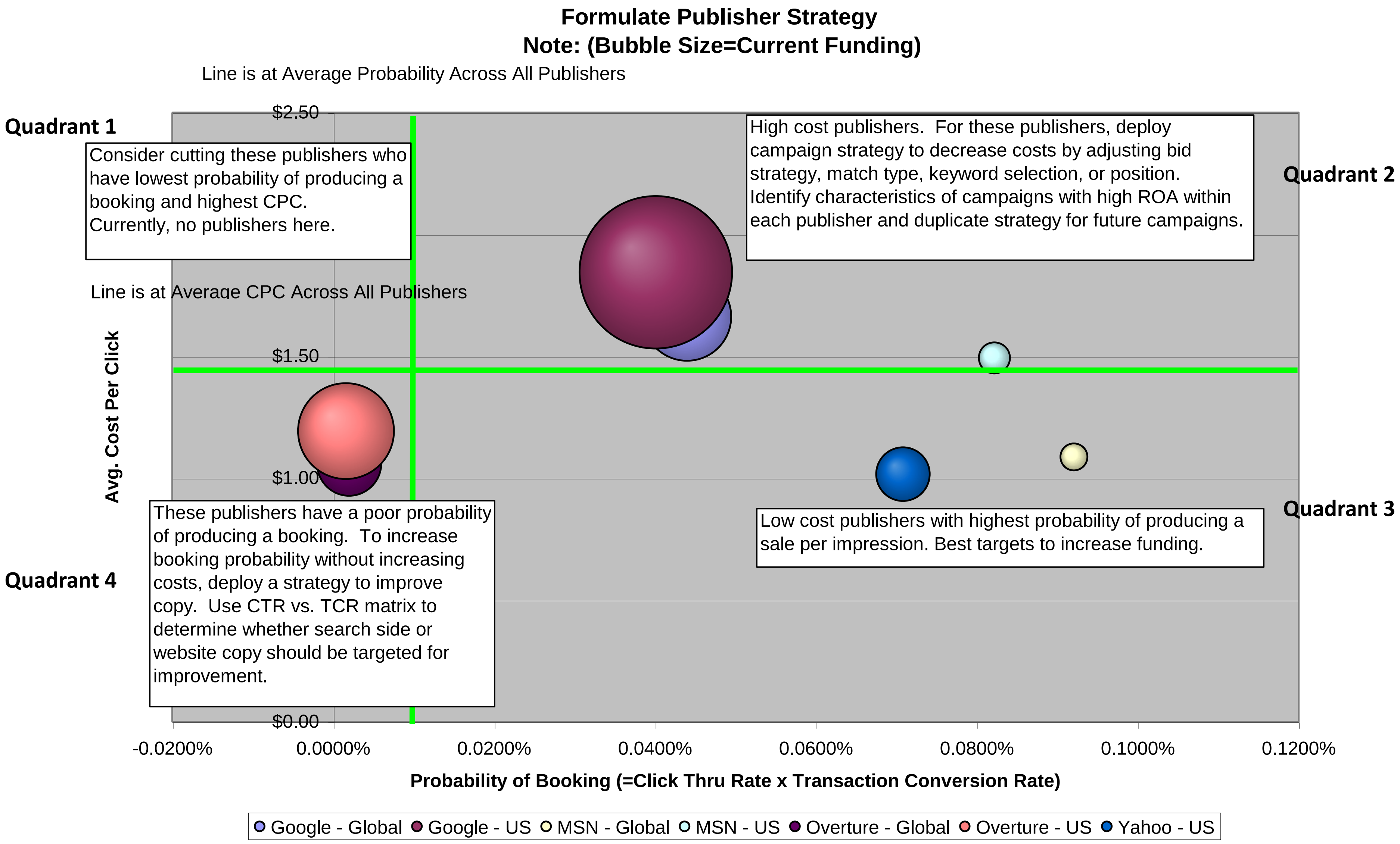
x	y	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})(y - \bar{y})$
23	11	1.4	0.8	1.12
20	9	-1.6	-1.2	1.92
14	4	-7.6	-6.2	47.12
27	15	5.4	4.8	25.92
22	10	0.4	-0.2	-0.08
20	11	-1.6	0.8	-1.28
26	11	4.4	0.8	3.52
16	7	-5.6	-3.2	17.92
25	13	3.4	2.8	9.52
23	11	1.4	0.8	1.12

Mean: 21.6 10.2 Sum: 106.8

Std: 4.20 3.05 Cov: 11.86

Corr: 0.98

Optimize Publisher Strategy—Results



Formulate Publisher Strategy
Note: (Bubble Size=Current Funding)

Line is at Average Probability Across All Publishers

Line is at Average CPC Across All Publishers

Consider cutting these publishers who have lowest probability of producing a booking and highest CPC. Currently, no publishers here.

High cost publishers. For these publishers, deploy campaign strategy to decrease costs by adjusting bid strategy, match type, keyword selection, or position. Identify characteristics of campaigns with high ROA within each publisher and duplicate strategy for future campaigns.

These publishers have a poor probability of producing a booking. To increase booking probability without increasing costs, deploy a strategy to improve copy. Use CTR vs. TCR matrix to determine whether search side or website copy should be targeted for improvement.

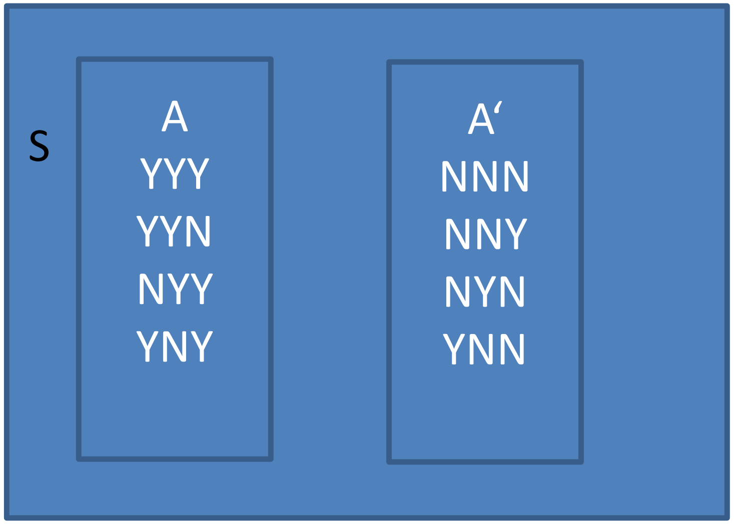
Low cost publishers with highest probability of producing a sale per impression. Best targets to increase funding.

Probability

- Experiment: Whether a person would click a web page A – Bernoulli trial?
- Sample space (S): Yes and No
- Event (success): p(x)
- Experiment: Whether 3 people you observe would click a web page A – binomial trial? X: # of people clicking.
- Sample space: YYY YYN YNY YNN NYY NYN NNY NNN
- Event: at least two people click
- $p(x \geq 2)$.
- Outcome for a single experiment: 2, Replication: 3
- Total number of outcomes in sample space: 2^3 = 8
- $p(x \geq 2) = p(x=3) + p(x=2) = \frac{1}{8} + \frac{3}{8} = \frac{4}{8} = 0.5$
- $\sum_{i=1}^{8} P(O_i) = 1$

Complement

- Complement of event A, A', is the set of all outcomes that are not in A.
- A: at least two clicks {YYY, YYN, NYY, YNY}
- A‘: {NNN, NNY, NYN, YNN}



S

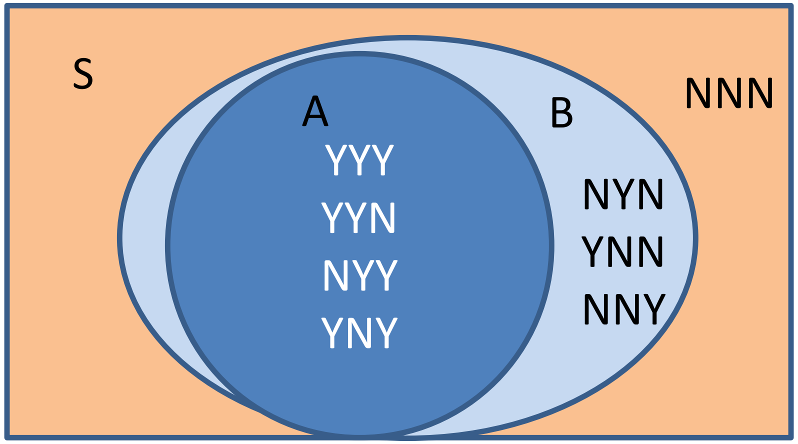
A
YYY
YYN
NYY
YNY

A‘
NNN
NNY
NYN
YNN

$0 \leq p(A) \leq 1$
$p(S) = 1$
$p(A) + p(A') = 1$
$p(A) = 1 - p(A')$
$p(A') = 1 - p(A)$

Union and Intersection

- Union: A or B – A U B (most women want rich or handsome man)
- Intersection: A and B – A ∩ B (most women want rich and handsome man)
- A: at least two clicks {YYY, YYN, NYY, YNY}
- B: at least one click {YYY, YYN, NYY, YNY, NYN, YNN, NNY}
- A ∩ B: {YYY, YYN, NYY, YNY}
- A U B : {YYY, YYN, NYY, YNY, NYN, YNN, NNY}
- Events are disjoint or independent if A ∩ B = Ø –> P(A ∩ B) = 0.



p(A U B) = p(A) + p(B) – p(A ∩ B)

Interpretation of Probability

Person #	1	2	3	4	5	6	7	8	9	10
A: Clicked?	N	Y	Y	Y	N	N	Y	Y	N	N

P	0	.5	.667	.75	.6	.5	.571	.625	.556	.5

$$P(A) = \frac{N(A)}{N} = \frac{5}{10} = .5$$

Permutation

- There are 3 (n) web pages on your web site and visitors can access from one page to all other pages. Visitors usually select 2 pages (k). How many ways are there to select the 2 pages?
- 12, 13, 21, 23, 31, 32 (ordered subsets)
- 3*2
- $P_{k,n} = \frac{n!}{(n-k)!} = \frac{3!}{(3-2)!} = \frac{3*2*1}{1} = 6$

Combination

- There are 3 (n) web pages on your web site and visitors can access from one page to all other pages. Visitors usually select 2 pages (k). Which 2 pages are selected?
- 21, 13, 23 (unordered subsets)
- $\binom{n}{k} = \frac{P_{k,n}}{k!} = \frac{n!}{k!(n-k)!} = \frac{3*2*1}{2*1*1} = 3$

Conditional Probability

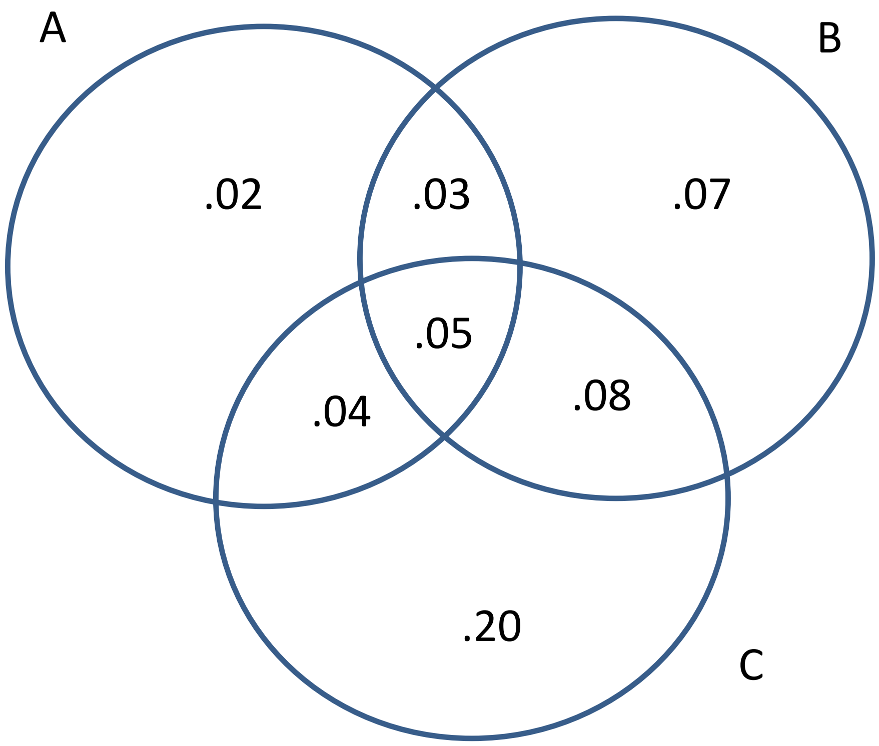
	B - faulty	B’ – not faulty
Line A	2	6
Line A’	1	9

p(A) = $\frac{8}{18}$ = 0.44

p(A|B) = $\frac{2}{3}$ = $\frac{\frac{2}{18}}{\frac{3}{18}}$ = $\frac{P(A \cap B)}{P(B)}$

Reading Habits
A: Art, B: Books, C: Cinema

Read Regularly	A	B	C	A ∩ B	A ∩ C	B ∩ C	A ∩ B ∩ C
P	.14	.23	.37	.08	.09	.13	.05



$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{.08}{.23} = .348$$

$$\mathrm{P(A|B \cup C)} = \frac{P(A \cap (B \cup C))}{P(B \cup C)} = \frac{.04+.05+.03}{.47} = .225$$

P(A|reads at least one) = P(A|A U B U C)

$$= \frac{P(A \cap (A \cup B \cup C))}{(A \cup B \cup C)}$$

$$= \frac{P(A)}{(A \cup B \cup C)} = \frac{.14}{.49} = .286$$

$$\mathrm{P(A \cup B|C)} = \frac{P((A \cup B) \cap C)}{P(C)} = \frac{.04+.05+.08}{.37} = .459$$

Multiplication Rule for P(A∩B)

- $P(A \cap B) = P(A|B) * P(B)$

Player Brand	Market Share	Repair Rate
M	50%	25%
L	30%	20%
N	20%	10%

- Probability that a consumer bought Brand M that will need repair?
- Probability that customer has a player that will need repair?
- Given that player needs repair, what is the probability that it is brand M? Brand L? Brand N?

Independence

- If two events A and B are independent:
- $P(A|B) = P(A)$
- P(A∩B) = P(A) * P(B)

A	B	AB	O
.40	.11	.04	.45

- What is the probability that blood phenotypes of two randomly selected individuals match?

Binomial Probability Distribution

- The experiment consists of a sequence of n smaller experiments called trials, where n is fixed in advance of the experiment.
- Each trial can result in one of the same two possible outcomes, success (S) or Failure (F).
- The trials are independent, so that the outcome on any particular trial does not influence the outcome on any other trial.
- The probability of success p(S) is constant from trial to trial; we denote this probability by p.

- Examples: The number of heads when one flips a coin 10 times. Number of customers who pay with credit card among 10 customer who visit the store.

- $b(x; n, p)$
- $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$

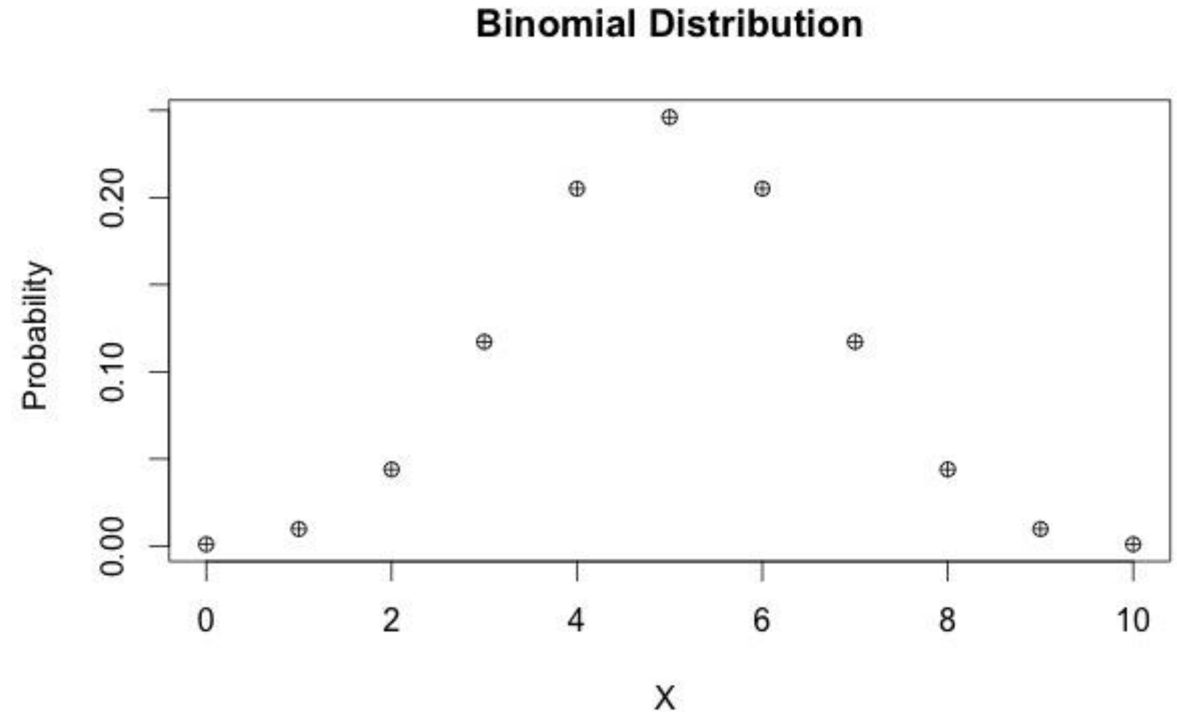
Example

- 20% of customers click your ad.
- Select random 5 people
- X: # of customers who click your ad.
- What is the probability that at most 3 customers click your ad?
- P(X=3) = b(3; 5, .20) = $\binom{5}{3}.20^3\ .80^2$= .0512
- P(X=2) = $\binom{5}{2}.20^2\ .80^3$= 0.2048
- P(X=1) = $\binom{5}{1}.20^1\ .80^4 = 0.4096$
- P(X=0) = $.80^5$=0.32768
- Answer: 0.99328

- Mean = E(X) = np = 5*.20 = 1
- Variance(X) = np(1-p)

Flipping a Fair Coin 10 Times

0	1	2	3	4	5	6	7	8	9	10
.0001	.001	.044	.117	.205	.246	.205	.117	.044	.001	.0001


Binomial Distribution
Probability
X