MSBD 5001

Topic 1: Mathematical Backgrounds

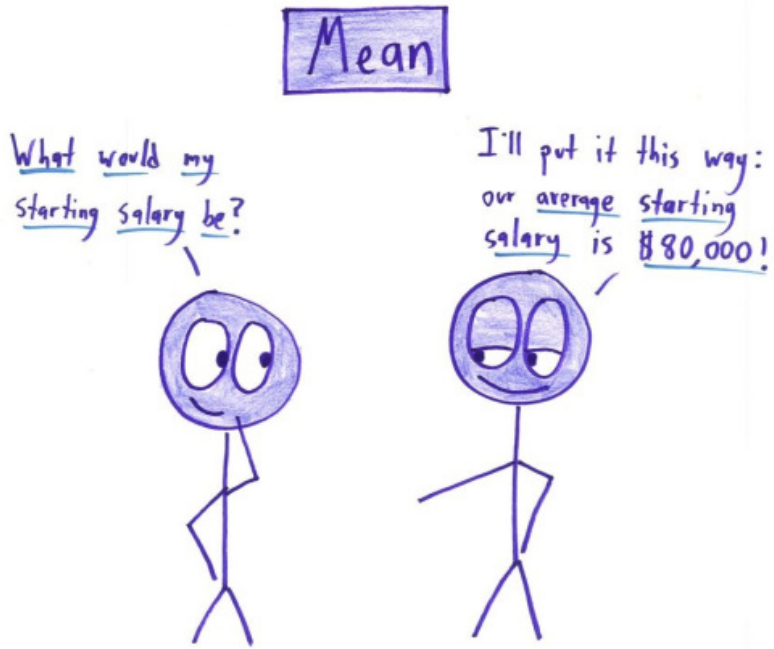
Prof. Kai Chen & Dr. Cecia Chan

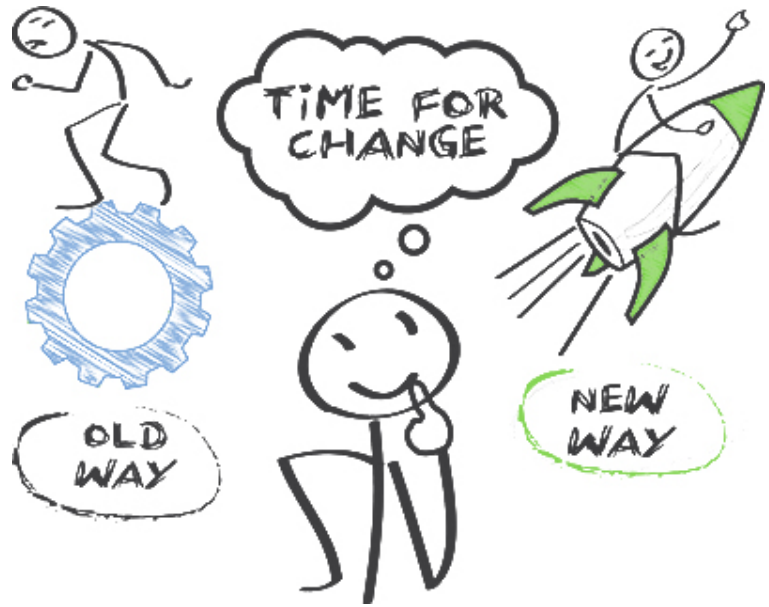
Department of Computer Science & Engineering
The Hong Kong University of Science and Technology
Hong Kong SAR, China

The slides are based on Dr. Desmond Tsoi's CSIT5800 slides

Statistical Analysis

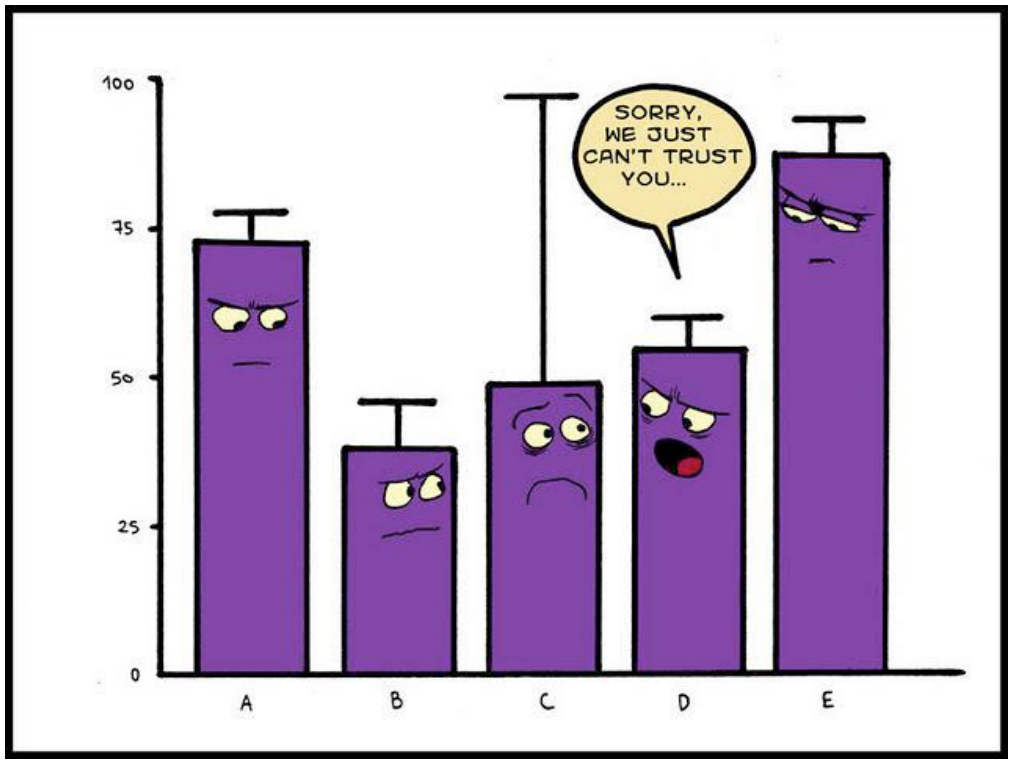
- Statistical analysis is the science of collecting, exploring and presenting large amounts of data (a.k.a. dataset) to discover underlying patterns and trends.
- It is used extensively in science, from physics to the social sciences.
- The following are the major tasks in statistical analysis:
  - Describing and summarizing the data
  - Identifying the relationship between variables
  - Forecasting the outcomes





Part I

Describing and Summarizing Data



Data Sets

- Data sets can be thought of as a bunch of number or a list of things.
  - Examples:
    - ★ Suppose we ask twenty students their weights and then record them as:

      ```
      122 146  65 162 148 155 136 151 151 153
      201 156 235 157 160 171 178 197 142 131
      ```

      This is a data set of 20 observations.
      Note: Number of items in a sample is called sample size, denoted as n
    - ★ Suppose we ask the students their hair color and get the responses:

      ```
      Red   Blond Blond Brown Brown Red   Blond Blond Brown Black
      Blond Red   Red   Brown Black Brown Red   Black Brown Blond
      ```

- Data come in two types:
  - Discrete (Example: Hair color data set)
  - Continuous (Example: Weight data set)

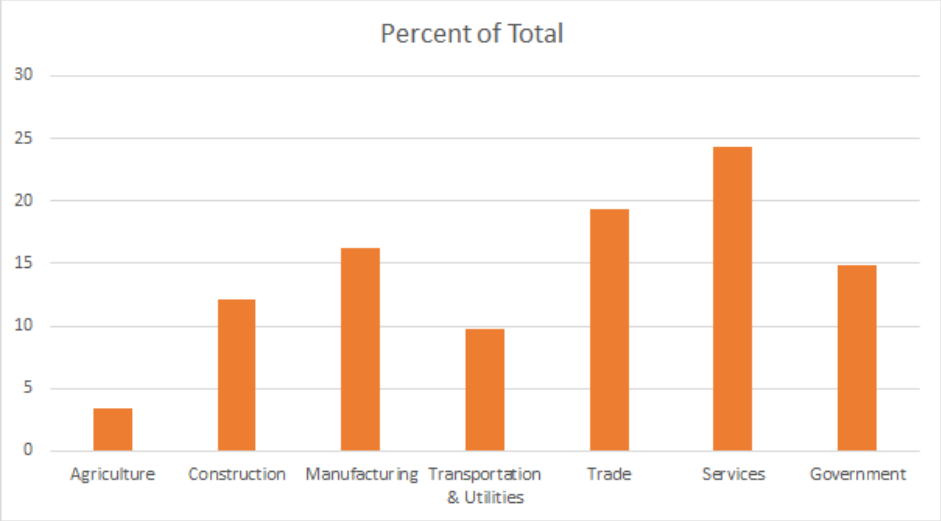
Describing and Summarizing Data

There are many ways to describe and summarize our data. We discuss a few below.

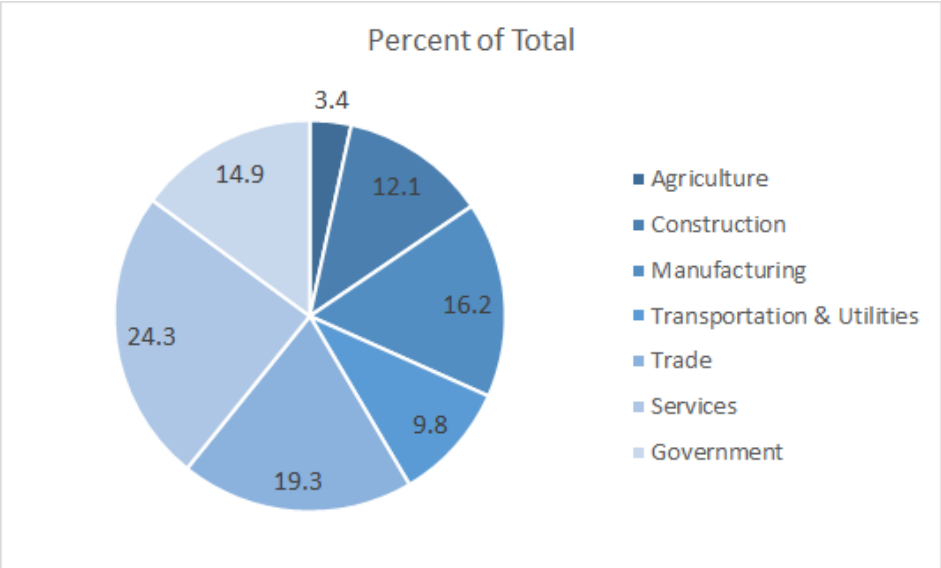
1. Table

Industry Group	**Number of disabling injuries (in 1,000s)**	**Percent of Total**
Agriculture	130	3.4
Construction	470	12.1
Manufacturing	630	16.2
Transportation & Utilities	300	9.8
Trade	380	19.3
Services	750	24.3
Government	580	14.9

2. Barchart



3. Pie chart



4. Stem-and-leaf plot

- Assume we have the data of maximum ozone reading (in parts per billion(ppb)) taken on 80 summer days in a large city.

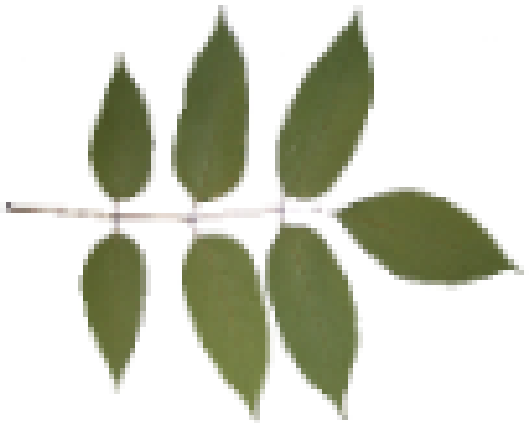
60	61	61	64	64	64	64	66	66	68
68	68	69	71	71	71	71	71	71	72
72	73	75	75	80	80	80	80	80	80
82	82	83	85	86	86	87	87	87	89
91	92	94	94	98	99	99	100	101	103
103	103	108	111	113	113	114	118	119	119
122	122	124	124	124	125	125	131	133	134
136	141	142	143	146	150	152	155	169	169

- A stem-and-leaf plot can be constructed using the first digit of the two-digit numbers and the first two digits of the three-digit numbers as the stem number and the remaining digits as the leaf number.

Stem	Leaf
6	0 1 1 4 4 4 4 6 6 8 8 8 9
7	1 1 1 1 1 1 2 2 3 5 5
8	0 0 0 0 0 0 2 2 3 5 6 6 7 7 9
9	1 2 4 4 8 9 9
10	0 1 3 3 3 8
11	1 3 3 4 8 9 9
12	2 2 4 4 4 5 5
13	1 3 4 6
14	1 2 3 6
15	0 2 5
16	9 9

Advantage of using Stem-and-leaf Plot

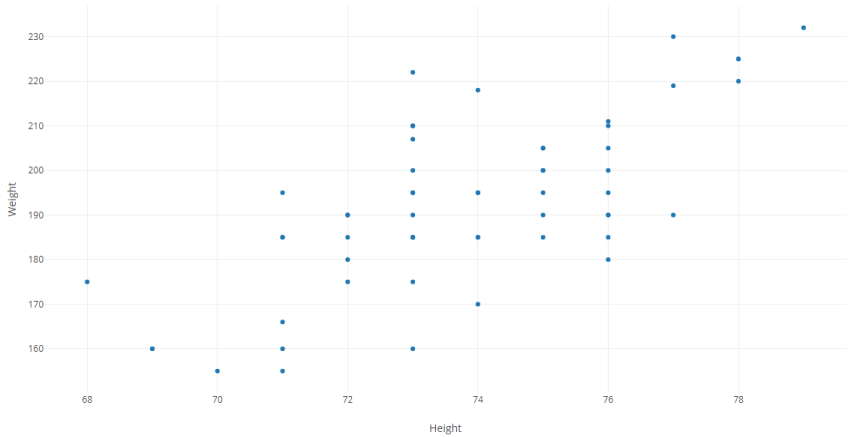
- The plot can be constructed quickly using pencil and paper.
- The values of each individual data point can be recovered from the plot.
- The data is arranged compactly since the stem is not repeated in multiple data points.



- Scatterplot
  - It uses Cartesian coordinates to display values for two variables for a set of data.
  - The following shows the height and weight of 57 baseball players.

(height, weight)

(74,218)	(75,185)	(77,219)	(73,185)	(69,160)	(73,222)	(78,225)
(76,205)	(77,230)	(78,225)	(76,190)	(72,180)	(73,185)	(73,200)
(74,195)	(75,195)	(72,185)	(75,190)	(76,200)	(76,180)	(72,175)
(76,195)	(68,175)	(73,185)	(69,160)	(76,211)	(77,190)	(74,195)
(75,200)	(73,207)	(79,232)	(72,190)	(75,200)	(78,220)	(73,195)
(75,205)	(74,195)	(71,185)	(73,210)	(76,210)	(73,195)	(75,205)
(73,175)	(73,190)	(74,185)	(72,190)	(73,210)	(71,195)	(71,166)
(71,185)	(73,160)	(74,170)	(76,185)	(71,155)	(76,190)	(71,160)
(70,155)						



Things to think about when looking at scatterplots.

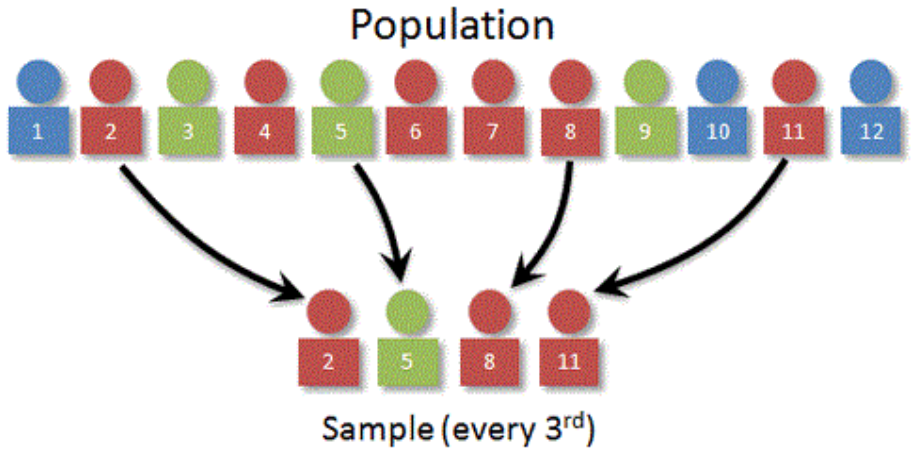
- Form: Does it have a shape?
- Direction: Does the data have a direction?
- Strength: Are the points close together or scattered?

Describing and Summarizing Data

- Apart from describing data graphically, data can also be described using numerical numbers.
- The following are common numerical descriptive measures of data.
  - Describing central tendency
    - ★ Mean (Arithmetic mean)
    - ★ Min
    - ★ Max
    - ★ Median
    - ★ Mode
  - Describing variability
    - ★ Range
    - ★ pth percentile
    - ★ Interquartile range (IQR)
    - ★ Variance
    - ★ Standard deviation
- Before elaborating all the numerical description measures above, we will first define a few basic concepts of statistics. They are population, sample, and sampling error.

Population, Sample and Sample Error

- Population: The collection of all individuals or items under consideration in a statistical study.
- Sample: Part of the population from which information is collected.
- Sampling error: Reflects the fact that the result we get from our sample is not going to be exactly equal to the result we would have got if we had been able to measure the entire population.



A sampling method is a procedure for selecting sample elements from a population.

Example

- A school takes a poll to find out what students to eat at lunch.
- 70 students are randomly chosen to answer the poll questions.
- What are the population and the sample of this study?

Answer

- Population: All students at the school.
- Sample: The 70 students polled.

Describing and Summarizing Data

- Suppose we have a sample of size n, denoted as $x_1, x_2, \ldots, x_i, \ldots, x_n$, the followings are the formal definitions of all the descriptive measures mentioned earlier.
  - Mean of a sample: $\overline{x} = \sum_{i=1}^{n} x_i$
  - Minimum of a sample: Minimum of $\{x_1, x_2, \ldots, x_i, \ldots, x_n\}$
  - Maximum of a sample: Maximum of $\{x_1, x_2, \ldots, x_i, \ldots, x_n\}$
  - Median of a sample: Middle ordered of $\{x_1, x_2, \ldots, x_i, \ldots, x_n\}$
  - Mode of a sample: The value that appears most often in $\{x_1, x_2, \ldots, x_i, \ldots, x_n\}$
  - Range of a sample: Minimum to Maximum
  - pth percentile of a sample: The value so that roughly $p\%$ of the sample are smaller and $(100 - p)\%$ of the sample are larger.
  - Interquartile range (IQR) of a sample: Third quartile - First quartile
    - ★ First quartile: Median of the first half of the data
    - ★ Third quartile: Median of the second half of the data
  - Variance of a sample: $\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$
  - Standard deviation of a sample: $\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2}$

Example

- Suppose we ask twenty students their weights and record them as:

  65 122 131 136 142 146 148 151 151 153
  155 156 157 160 162 171 178 197 201 235

  - Mean of a sample:
  - Minimum of a sample:
  - Maximum of a sample:
  - Median of a sample:
  - Mode of a sample:
  - Range of a sample:
  - 60th percentile of a sample:
  - Interquartile range (IQR) of a sample:
    - ★ Third quartile:
    - ★ First quartile:
  - Variance of a sample:
  - Standard deviation of a sample:

Example

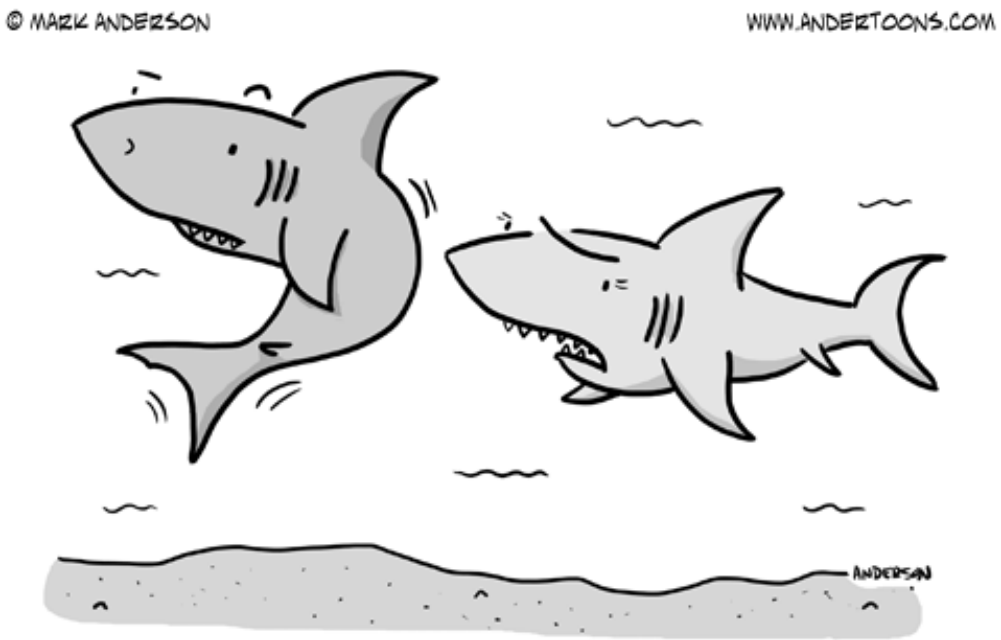
- Suppose we ask twenty students their weights and record them as:

  65 122 131 136 142 146 148 151 151 153
  155 156 157 160 162 171 178 197 201 235

  - Mean of a sample: 155.85
  - Minimum of a sample: 65
  - Maximum of a sample: 235
  - Median of a sample: 155
  - Mode of a sample: 151
  - Range of a sample: 65 to 235
  - 60th percentile of a sample: 156.39
  - Interquartile range (IQR) of a sample: 177 - 146 = 25
    - ★ Third quartile: 171
    - ★ First quartile: 146
  - Variance of a sample: $\frac{(65-155.85)^2+(122-155.85)^2+\ldots+(201-155.85)^2+(235-155.85)^2}{(20-1)} = 1136.34473$
  - Standard deviation of a sample: 33.7097

Part II

Identifying Relationship between Variables

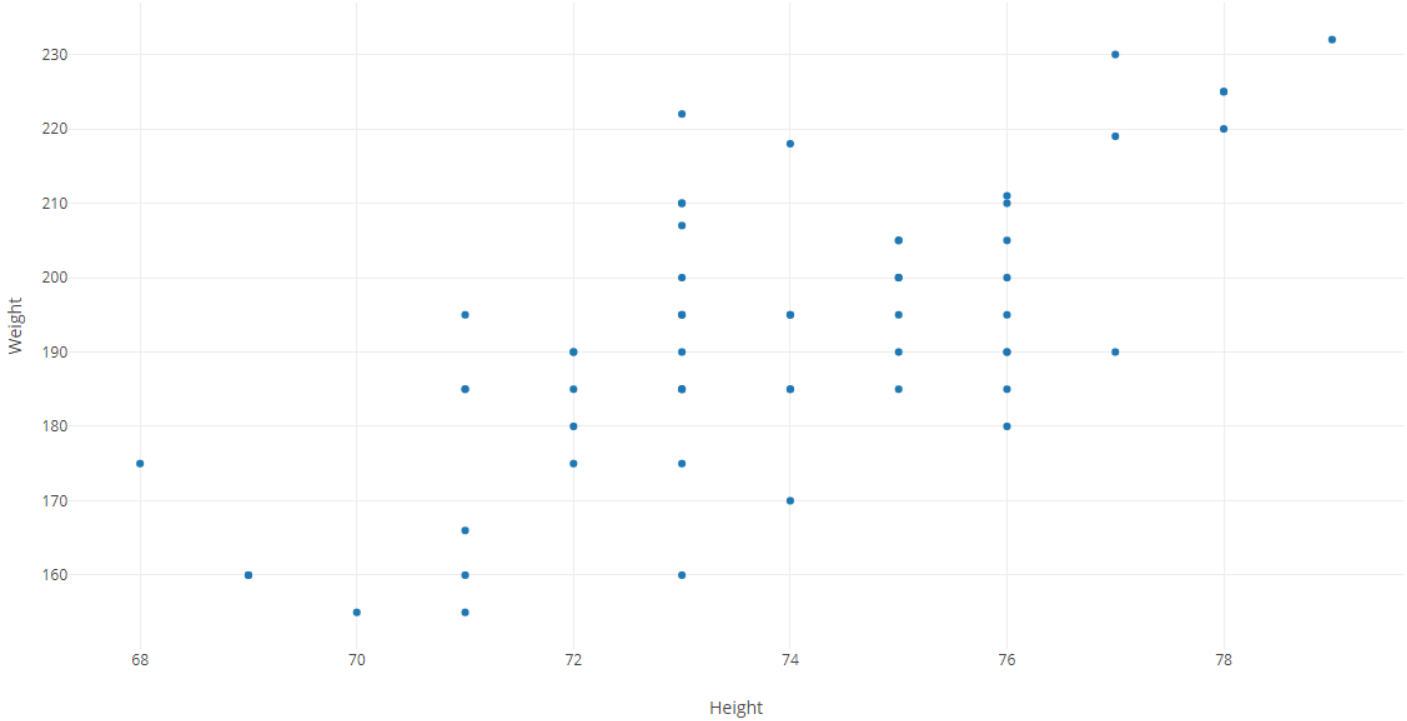


Relationship between Variables

- We often collect data from several different variables on a subject.
- A simple example is a form, such as an application form, which are collected from a group of people. Each item on the form corresponds to a variable.
  - Example: Suppose the form is that students are filling out at an university. Items might include the GPA, major, weight, height, gender, etc.
- We may describe each variable separately using the descriptive statistics, but often we also want to investigate the relationship between the variables, e.g., weight and height of students, denoted as Y and X, respectively.

Relationship between Variables

- Plot data with Y (weight) on the vertical axis and X (height) on the horizontal axis of a scatterplot.



Observations

- On the basis of the plot, a linear model is certainly worthy of a first try.

Relationship between Variables

- Linear model

$$Y = a + bX + e$$

where a is the y-intercept, b is the slope, and e is the random error (i.e., if there were no error Y would be a deterministic linear function of X).

- Note: X is called independent variable and Y is called dependent variable.

Relationship between Variables

1. Eyeball Fit – Pick two points on the plot so that the line passing through them gives a "fairly" good fit.
   - To estimate the slope, take two points, say (X_1, Y_1) and (X_2, Y_2), then

     $$\hat{b} = \frac{Y_2 - Y_1}{X_2 - X_1}$$

     For the student data, we chose the points (69, 160) and (78, 225). Hence the estimate of slope is

     $$\hat{b} = \frac{225 - 160}{78 - 69} = \frac{65}{9} = 7.2$$

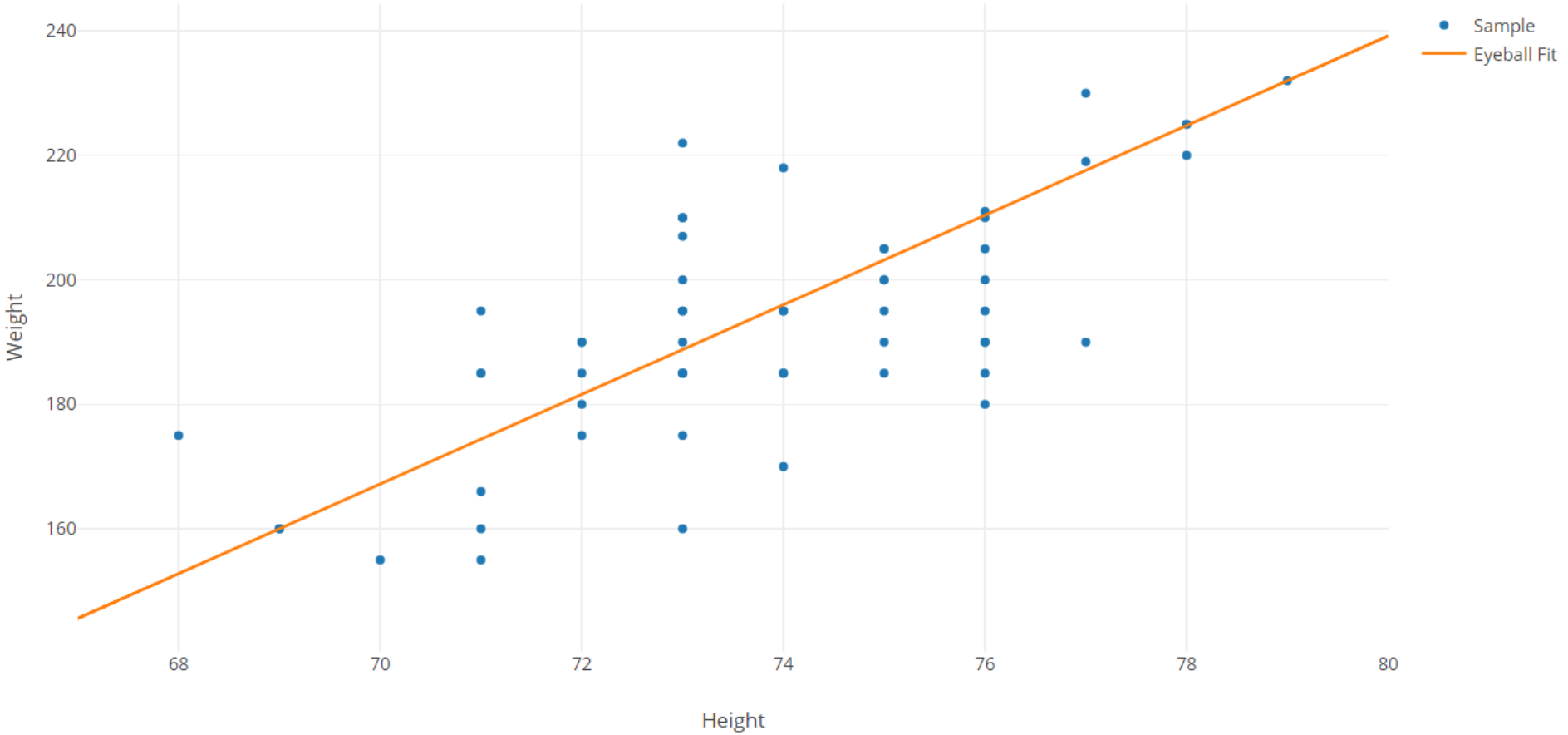
   - To estimate the y-intercept, simply take one of the points, say, (X_1, Y_1), then estimate the intercept by solving the linear equation for a, i.e., $\hat{a} = Y_1 - \hat{b}X_1$. For the student data, we chose the the point (69, 160),

     $$\hat{a} = 160 - 7.2(69) = -336.8$$

   - Thus, the predicted equation is

     $$Y = -336.8 + 7.2X$$

Eyeball Fit


Sample
Eyeball Fit
240
220
200
180
160
Weight
68
70
72
74
76
78
80
Height

Relationship between Variables

2. Least Square Fit

- Fit a line $Y = a + bX$ such that it minimizes the error S

$$S(a,b) = \sum_{i=1}^{n}(a + bx_i - y_i)^2$$

$$\frac{\partial S}{\partial a} = 2\sum_{i=1}^{n}(a + bx_i - y_i) = na + b\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i = 0$$

$$\frac{\partial S}{\partial b} = 2\sum_{i=1}^{n}(a + bx_i - y_i)x_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i = 0$$

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$a = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

Relationship between Variables

- Least Square Fit (Cont'd)
  - Alternatively, a and b can be found using the following:

$$b = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{X})^2}$$

$$a = \overline{Y} - b\overline{X}$$

where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

- ★ $\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})$ is covariance between x and y.
- ★ $\sum_{i=1}^{n}(x_i - \overline{X})^2$ is variance of x.

Can you prove the above?

- ★ Hint: $n\overline{X} = \sum_{i=1}^{n} x_i$ and $\overline{Y} = \sum_{i=1}^{n} y_i$.

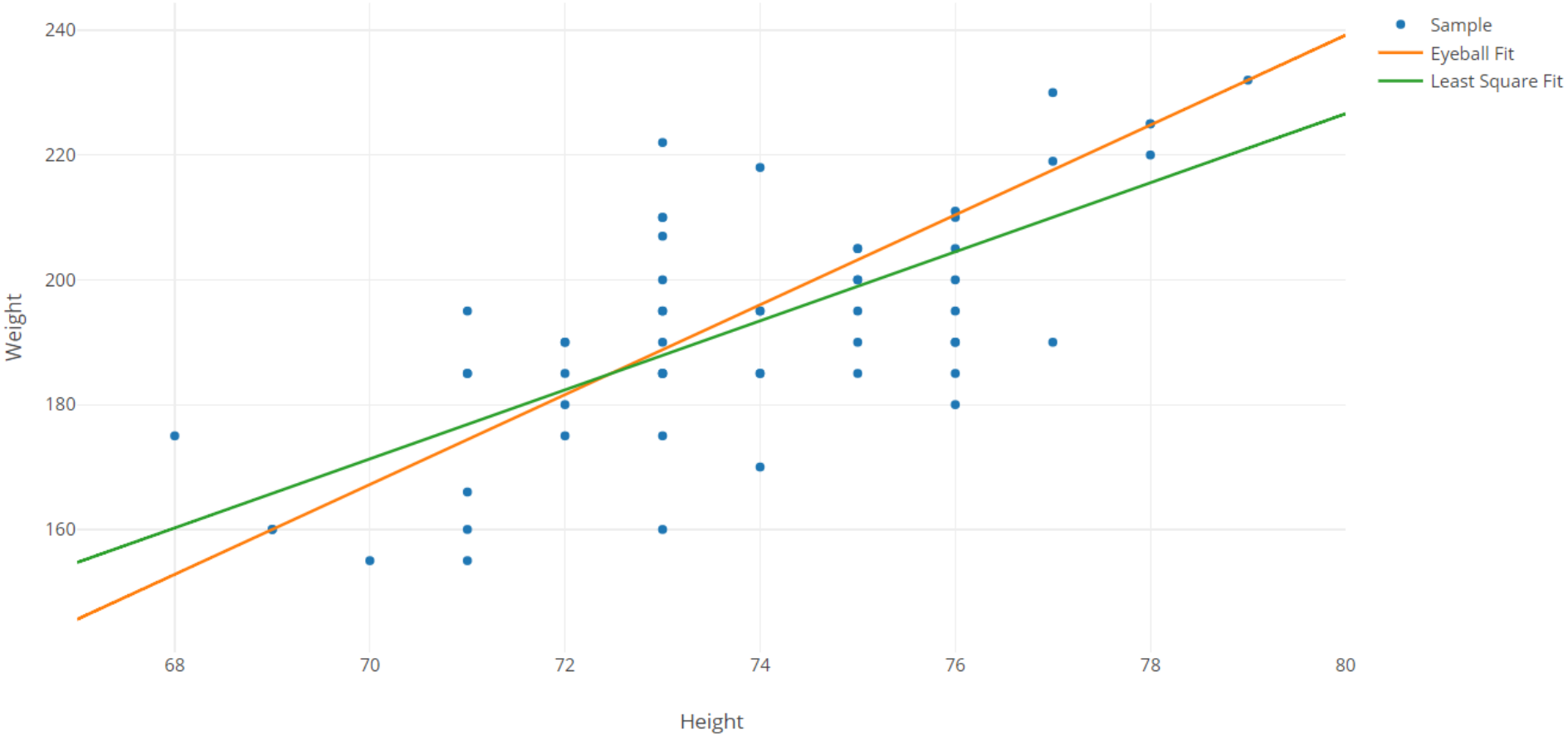
  - For the student data, we got

$$b = 5.530918$$

$$a = -215.861$$

$$Y = -215.861 + 5.530918X$$

Least Square Fit



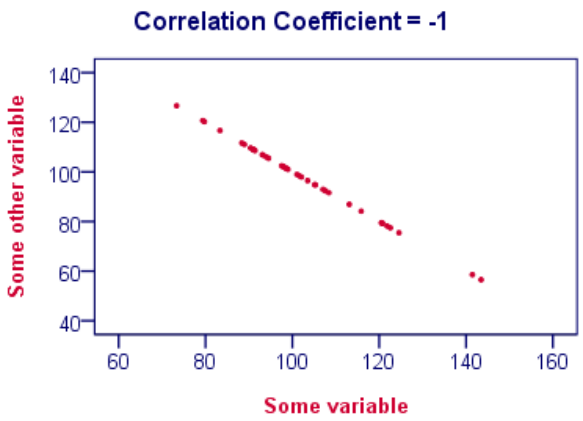
Correlation Coefficient

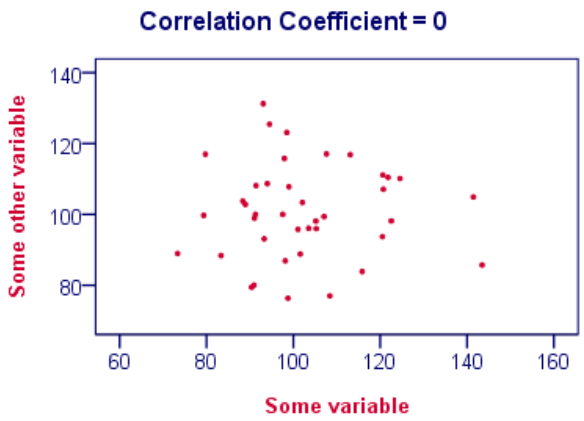
- Correlation coefficient, denoted as r, measures the degree to which two variables' movements are associated.

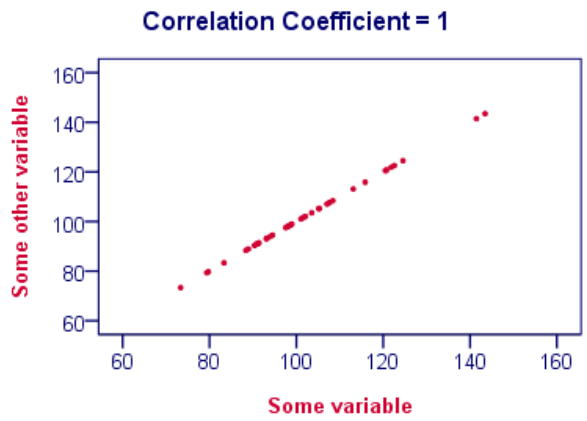
$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{X})^2 \sum_{i=1}^{n}(y_i - \overline{Y})^2}}$$

  where $-1 \leq r \leq 1$.
  - $r = 1$ means a perfect positive relationship, i.e., every positive increase of 1 in one variable, there is a positive increase of 1 in the other.
  - $r = -1$ means a perfect negative relationship, i.e., every positive increase of 1 in one variable, there is a negative decrease of 1 in the other.
  - r close to zero indicate little or no linear relationship, i.e., for every increase, there is not a positive or negative increase.

Correlation Coefficient







- For the student data, $r = 0.704583$.

Part III

Forecasting Outcomes

Experiment, Sample Space, Event

- An experiment is an action where the result is uncertain.
- A sample space is all the possible outcomes of an experiment, denoted as S.
  Examples:
  1. Flip a coin: $S = \{H, T\}$
  2. Roll a six sided die: $S = \{1, 2, 3, 4, 5, 6\}$
  3. Roll a pair of six-sided dice: $S = \{(1, 1), (1, 2), (1, 3), \ldots, (6, 6)\}$. That is, S consists of 36 pairs of integers.
- A event is a subset of S, denoted by A, B, C, etc.
  Examples:
  1. Flip a coin: $A = \{H\}$
  2. Roll a six-sided die: $B = \{1, 2\}$
  3. Roll a pair of six-sided dice: $A =$ sum of up-faces 7 or 11

Probabilities

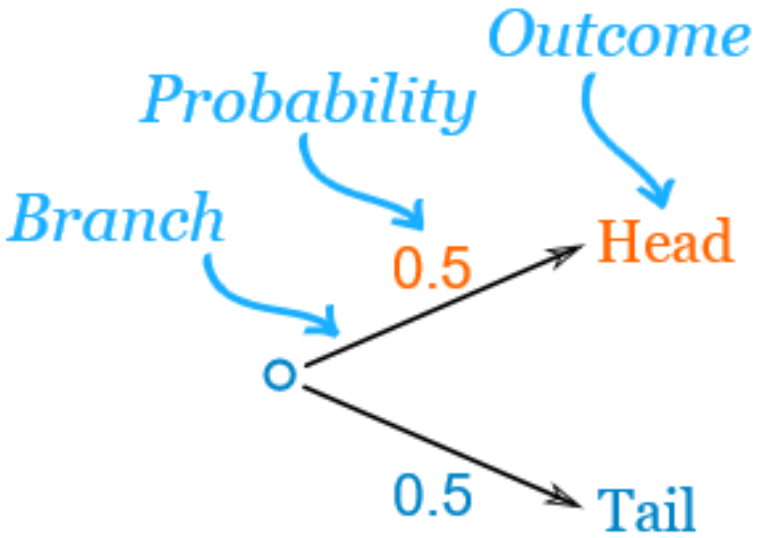
- Probability is the measure of how likely an event is to occur out of the number of possible outcomes.
- In other words, it is a ratio where we compare how many times an outcome can occur compared to all possible outcomes, i.e.,

$$\texttt{Probability} = \frac{\texttt{The number of wanted outcomes}}{\texttt{The number of possible outcomes}}$$

- The followings are some facts about probability.
  1. The probability of an event A is a number between 0 and 1.
  2. The probability of the sample space is 1.
  3. If two events cannot occur at the same time, the probability that one or the other occurs is the sum of the probabilities of the individual events.
- We denote the probability of event A by $P(A)$.

Determination of Probabilities: Tree Diagrams

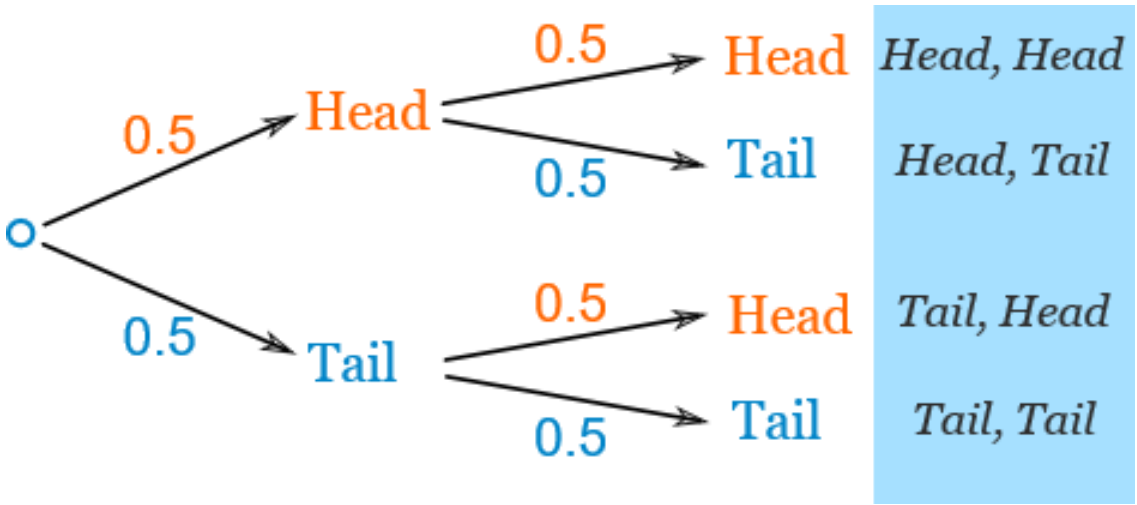
- Calculating probabilities can be hard. Sometimes we add them, sometimes we multiply them, and often it is hard to figure out what to do.
- To remedy this, we often construct a tree diagram.
- Here is a tree diagram for the toss of a coin:



  - The probability of each branch is written on the branch.
  - The outcome is written at the end of the branch.

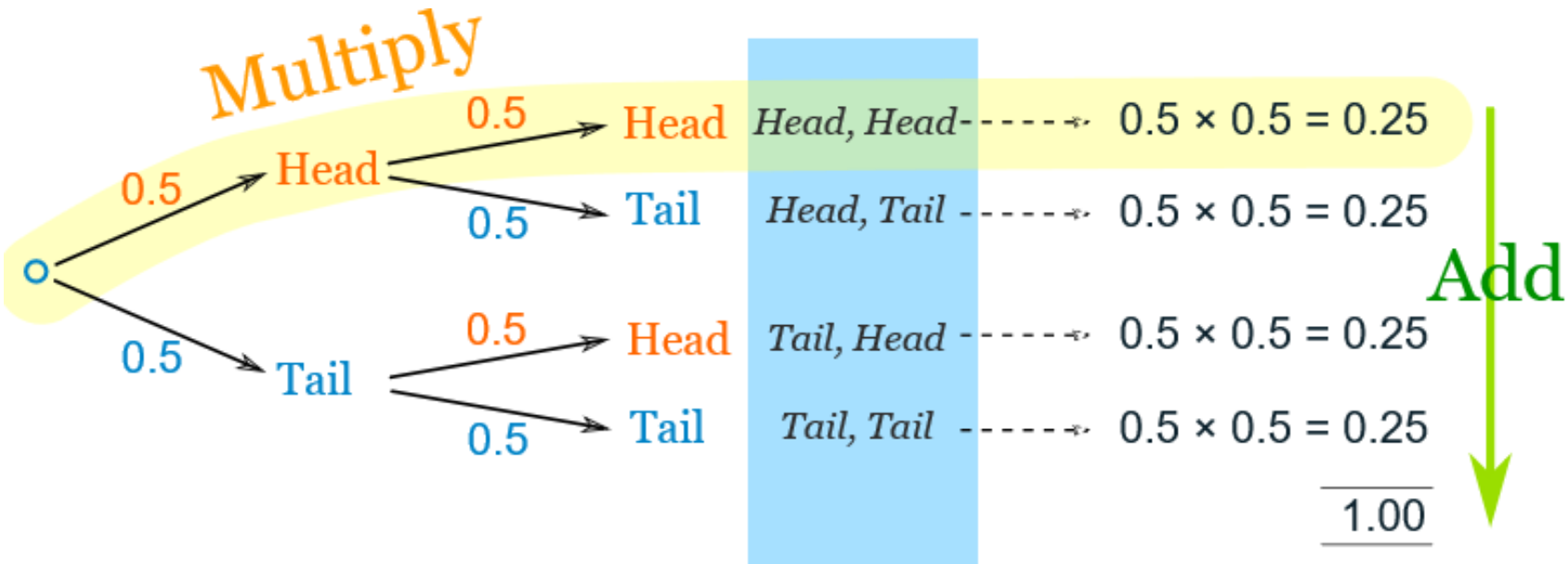
Determination of Probabilities: Tree Diagrams

- We can extend the tree diagram to two tosses of a coin:



Determination of Probabilites: Tree Diagrams

- How to calculate the overall probabilities?
  - Multiply probabilities along the branches.
  - Add probabilities down columns.



- Results:
  - The probability of "Head, Head" is $0.5 \times 0.5 = 0.25$.
  - All probabilities add to 1.0.
  - The probability of getting at least one Head from two tosses is $0.25 + 0.25 + 0.25 = 0.75$.

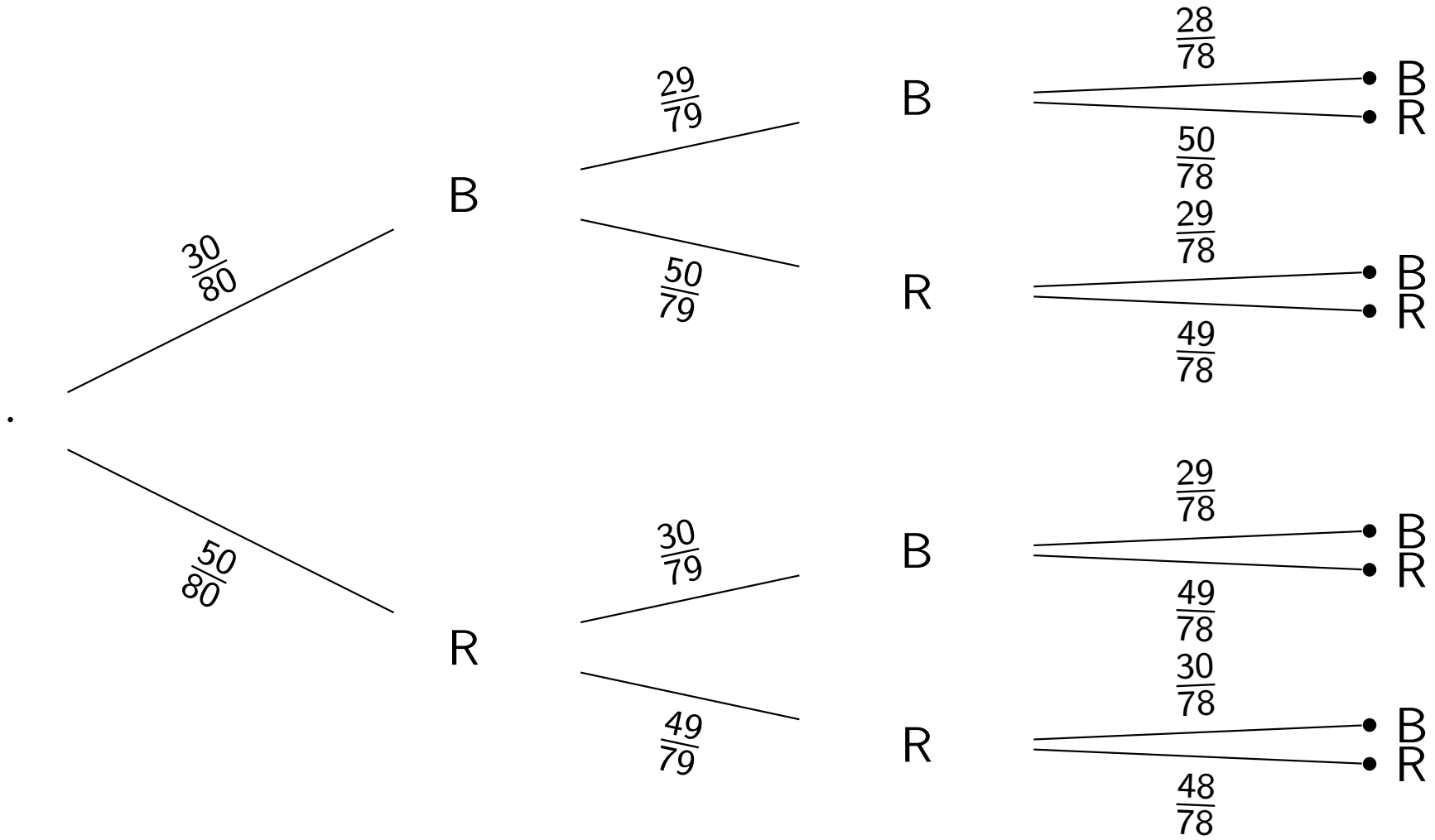
Example

- Problem:
  - Suppose we have an urn with 30 blue balls and 50 red balls in it and that these balls are identical except for color.
  - Suppose further the balls are well mixed and that we draw 3 balls, without replacement.
  - Determine the probability that the balls are all of the same color.

Example

- Problem:
  - Suppose we have an urn with 30 blue balls and 50 red balls in it and that these balls are identical except for color.
  - Suppose further the balls are well mixed and that we draw 3 balls, without replacement.
  - Determine the probability that the balls are all of the same color.
- Solution:
  - Step 1: Trace a branch up for blue, putting the probability of the first ball being blue, 30/80, on it and a "B" and the end. Likewise, trace a branch down for red with 50/80 on it and an "R" at the end.
  - Step 2: The second ball is either blue or red, i.e., at the "B", draw one branch up for second ball blue with the probability of 29/79 on it, and end it with a "B". Next draw one branch down for second ball red with the probability 50/79 and end it with an "R".
  - Step 3: The ball can be blue or red so there will be two branches at the end of the four second step branches.

Determination of Probabilities: Tree Diagrams



Independence

- Referring to the final tree diagram of the last example, the probabilities on the branches are called conditional probabilities.
- Example
  - Let B_2 denote the event that the second ball is blue and
  - Let B_1 denote the event that the first ball is blue

  Then the probability on the first step upward branch is the probability that B_2 occurs given that B_1 has occurred, i.e., $29/79$. This is called conditional probability of B_2 given B_1 and we will denote it by $P(B_2|B_1)$. The bar is pronounced "given".
- In general for two events A and B, if $P(B|A) = P(B)$, i.e., knowledge of A did not change the prediction of B, then we say that A and B are independent events.

Independence

Question

What is $P(B_2)$? Look at all the end nodes for which the second ball is blue in the final tree diagram.

Independence

Question

What is $P(B_2)$? Look at all the end nodes for which the second ball is blue in the final tree diagram.

- Answer:

$$\frac{30}{80} \times \frac{29}{79} + \frac{50}{80} \times \frac{30}{79} = \frac{30}{80}$$

Observation

$P(B_2|B_1) = 29/79$ and $P(B_2) = 30/80$, which means B_1 and B_2 are not independent events. In fact, they are dependent events.

Conditional Probability

- Let A and B be arbitrary events and we want to determine $P(B|A)$.
- Formula of conditional probability

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

Assume we repeat the experiment many many times.

How to compute $P(B|A)$?

Count how many times that A occurs and count those times that B has occurred.

- If A and B are independent events, we get

$$P(B) = \frac{P(A \text{ and } B)}{P(A)}$$

$$P(A \text{ and } B) = P(A)P(B)$$

Example

- Problem:
  - A jet airplane has 3 engines which function independently of one another. The probability that an engine fails in flight is 0.0001. Furthermore, the plane can fly if at least one engine is functioning. Determine the probability the airplane has a successful flight.

Example

- Problem:
  - A jet airplane has 3 engines which function independently of one another. The probability that an engine fails in flight is 0.0001. Furthermore, the plane can fly if at least one engine is functioning. Determine the probability the airplane has a successful flight.
- Solution:
  - The event we want to consider is $A =$ at least one engine operates throughout the flight.
  - Consider the complement of A, A^c which is the event all the three engines fail.
  - Let B_1 be the event that engine one fails, B_2 be the event that engine two fails, B_3 be the event that engine three fails. Hence, A^c is the event B_1 and B_2 and B_3 occurs. Thus

$$P(A) = 1 - P(A^c) = 1 - P(B_1 \text{ and } B_2 \text{ and } B_3)$$

- Solution (Cont'd):
  - As the engines function independently of one another, hence B_1, B_2, and B_3 are independent events. So,

$$P(B_1 \text{ and } B_2 \text{ and } B_3) = P(B_1)P(B_2)P(B_3)$$

  - Therefore,
    $P(B_1 \text{ and } B_2 \text{ and } B_3) = 0.0001 \times 0.0001 \times 0.0001 = 0.0000000001$
  - Hence, $P(A) = 0.999999999999$.

Bayes Theorem

- In many situations, you will know one conditional distribution $P(x|y)$ and $P(x)$ but you are really interested in the other conditional distribution $P(y|x)$.

$$P(y|x) = \frac{P(x,y)}{P(x)} = \frac{P(x|y)P(y)}{P(x)}$$

- Let $A_1, A_2, \ldots, A_n$ be a set of mutually exclusive events that together form the sample space S. Let B be any event from the same sample space, such that $P(B) > 0$. Then

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

Example

- For a magazine, the probability that the reader is male given that the reader is at 35 years old is 0.3. The probability that a reader is male, given that the reader is under 35, is 0.65. If 75% of the reader are under 35, what is the probability that a randomly chosen reader is
  (a) Male
  (b) Female
  (c) Under 35 and it is given the reader is a female

Example

- For a magazine, the probability that the reader is male given that the reader is at 35 years old is 0.3. The probability that a reader is male, given that the reader is under 35, is 0.65. If 75% of the reader are under 35, what is the probability that a randomly chosen reader is
  (a) Male
  (b) Female
  (c) Under 35 and it is given the reader is a female
- Solution:
  (a) Let A_1 be the event of the reader being at least 35 years old, A_2 the event of the reader being under 35 years old, M be the event of the reader is being a male, and F be the event of the reader is being a female.
  $P(A_2) = 0.75$, $P(A_1) = 1 - 0.75 = 0.25$, $P(M|A_1) = 0.3$,
  $P(F|A_1) = 0.7$, $P(M|A_2) = 0.65$, $P(F|A_2) = 0.35$
  $P(M) = P(A_1, M) + P(A_2, M) = P(A_1)P(M|A_1) + P(A_2)P(B|A_2) = 0.25 \times 0.3 + 0.75 \times 0.65 = 0.5625$.
  (b) $P(F) = 1 - P(M) = 1 - 0.5625 = 0.4375$.
  (c) $P(A_2|F) = \frac{P(F|A_2)P(A_2)}{P(F|A_1)P(A_1)+P(F|A_2)P(A_2)} = \frac{0.35 \times 0.75}{0.7 \times 0.25 + 0.35 \times 0.75} = 0.6$

Parameters – Mean, Expected Value, or Expectation

- The mean, expected value, or expectation of a random variable X is written as $E(X)$ or μ.
- If we observe n random values of X, i.e., $x_1, x_2, \ldots, x_n$, then the mean of n values will be approximately equal to $E(X)$ for large n defined as follows:

$$E(X) = \sum_{i=1}^{n} xf(x) = \sum_{i=1}^{n} x_i p(X = x_i)$$

where $f(x)$ is the probability function of X.

Parameters – Mean, Expected Value, or Expectation

- Referring to the spinner example, the sample mean $(\overline{x}) = 183 \ / \ 100 = 1.83$, which can be calculated in one of the following ways:

$$\overline{x} = \frac{(1 \times 43) + (2 \times 31) + (3 \times 26)}{100}$$

or

$$\overline{x} = \left(1 \times \frac{43}{100}\right) + \left(2 \times \frac{31}{100}\right) + \left(3 \times \frac{26}{100}\right)$$

- From the last line, $\overline{x}$ is estimating

$$\mu = (1 \times p(1)) + (2 \times p(2)) + (3 \times p(3))$$

where μ is called the mean (or the parameter) of the probability model.

Parameters – Variance

- Variance is another parameter of probability model.
- The variance of a random variable X is written as $Var(X)$ or σ^2.
- It is a measure of how spread out it is.
- Are the values of X clustered tightly around their mean?
- The variance measures how far the values of X are from their mean, on average.
- Variance of X is

$$\begin{aligned} Var(X) &= E((X-\mu)^2) = E(X^2) - (E(X))^2 \\ &= \sum_{i=1}^{n} (x_i - \mu)^2 \times p(x_i) \end{aligned}$$

Variance

- Variance of the spinner sample is calculated by

$$\sigma^2 = \left((1-1.83)^2 \times \frac{43}{100}\right) + \left((2-1.83)^2 \times \frac{31}{100}\right) + \left((3-1.83)^2 \times \frac{26}{100}\right) = 0.6611$$

Binomial Probability Model

- A binomial model is characterized by trials (called Bernoulli trials) which either end in success or failure.
- Suppose we have n Bernoulli trials and p is the probability of success on a trial. Then this is a binomial model if
  - The Bernoulli trials are independent of one another.
  - The probability of success, p, remains the same from trial to trial.
- The binomial random variable, X, is the number of successes in the n trials.
  - Over the n trials, there could be one success, two successes, etc., up to n successes.
  - So the range of X is the set $\{0, 1, 2, \ldots, n\}$.
  - The probability of observing x success out of n trials is given by

$$P(X = x) = C_x^n p^x (1 - p)^{(n-x)}$$

    where $x = 0, 1, \ldots, n$.
- If the probabilities of X are distributed in this way, we write

$$X \sim bin(n, p)$$

Example

- Suppose we want the probability of getting 7 heads in ten flips of an unfair coin for which the probability of getting a head is $2/3$ and the probability of a tail is $1/3$.
- In other words, X is $bin(10, 2/3)$ and we want to compute $P(X = 7)$.
  - One possible way of obtaining 7 heads is if we observe the pattern HHHHHHHTTT and the probability of obtaining this pattern is

$$P(HHHHHHHTTT) = \frac{2}{3} \times \frac{2}{3} \times \frac{2}{3} \times \frac{2}{3} \times \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3}$$
$$= \left(\frac{2}{3}\right)^7 \left(\frac{1}{3}\right)^3 = \frac{128}{59049} = 0.00216769$$

  - There are C_7^{10} of the patterns contain 7 heads.
  - So, $P(X = 7)$ can be computed by

$$P(X = 7) = \text{No. of patterns} \times \text{Probability of pattern}$$

$$P(X = 7) = C_7^{10} \times \frac{128}{59049}$$
$$= 120 \times \frac{128}{59049} = \frac{5120}{19683} = 0.2601229$$

Mean and Standard Deviation of Binomial Probability Model

- Suppose we have an unfair coin for which the probability of getting a head is $2/3$, and the probability of a tail is $1/3$.
- Consider tossing the coin five times in a row and counting the number of times we observe a head.
- We denote this number as $X =$ No. of heads in 5 coin tosses, where $0 \leq X \leq 5$.
- Consider the example of the Binomial distribution below

x	0	1	2	3	4	5
P(X=x)	0.004	0.041	0.165	0.329	0.329	0.132

The mean value of the distribution can be calculated as

$$\begin{aligned}\mu =& 0 \times 0.004 + 1 \times 0.041 + 2 \times 0.165 + 3 \times 0.329 + \\ & 4 \times 0.329 + 5 \times 0.132 = 3.333\end{aligned}$$

Mean and Standard Deviation of Binomial Probability Model

- In general, there is a formula for the mean of a binomial distribution, μ. There is also a formula for the standard deviation, σ.

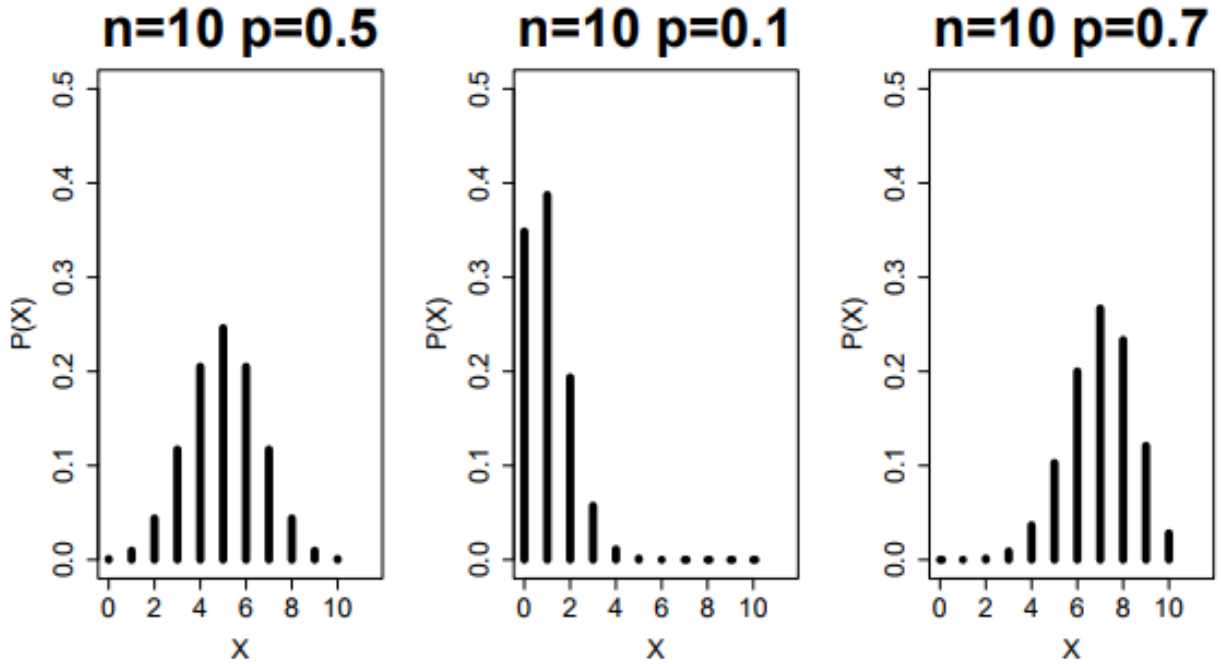
$$\mu = np$$
$$\sigma = \sqrt{np(1-p)}$$

- In the example above, X is $bin(5, 2/3)$ and so the mean and standard deviation are given by

$$\mu = np = 5 \times (2/3) = 3.3333$$
$$\sigma = \sqrt{np(1-p)} = 5 \times (2/3) \times (1/3) = 1.111$$

Shape of Binomial Distribution

- Different values of n and p lead to different distributions with different shapes.



Observations

- In general, the probabilities of a binomial will increase until np and then decrease.
- The probability distribution will be symmetric if $p = 1/2$, skewed right if $p < 1/2$, and skewed left if $p > 1/2$.

Probability Models for Continuous Data

- So far, we consider discrete data and discrete probability distributions.
- In practice, many data that we collect from experiments consist of continuous measurements.
- So, we need to study probability models for continuous data.
- For continuous data, we do not have equally spaced discrete values so instead we use a curve or function that describes the probability density over the range of the distribution.
- The curve is chosen so that the area under the curve is equal to 1.
- If we observe a sample of data from such a distribution, we should see that the values occur in regions where the density is highest.

Expectation and Variance of Continuous Random Variables

- The expectation is defined differently for continuous and discrete random variables.
- Let X be a continuous random variable with probability density function $f_X(x)$.
- The expected value of X is

$$E(X) = \mu = \int_{-\infty}^{\infty} x f_X(x) dx$$

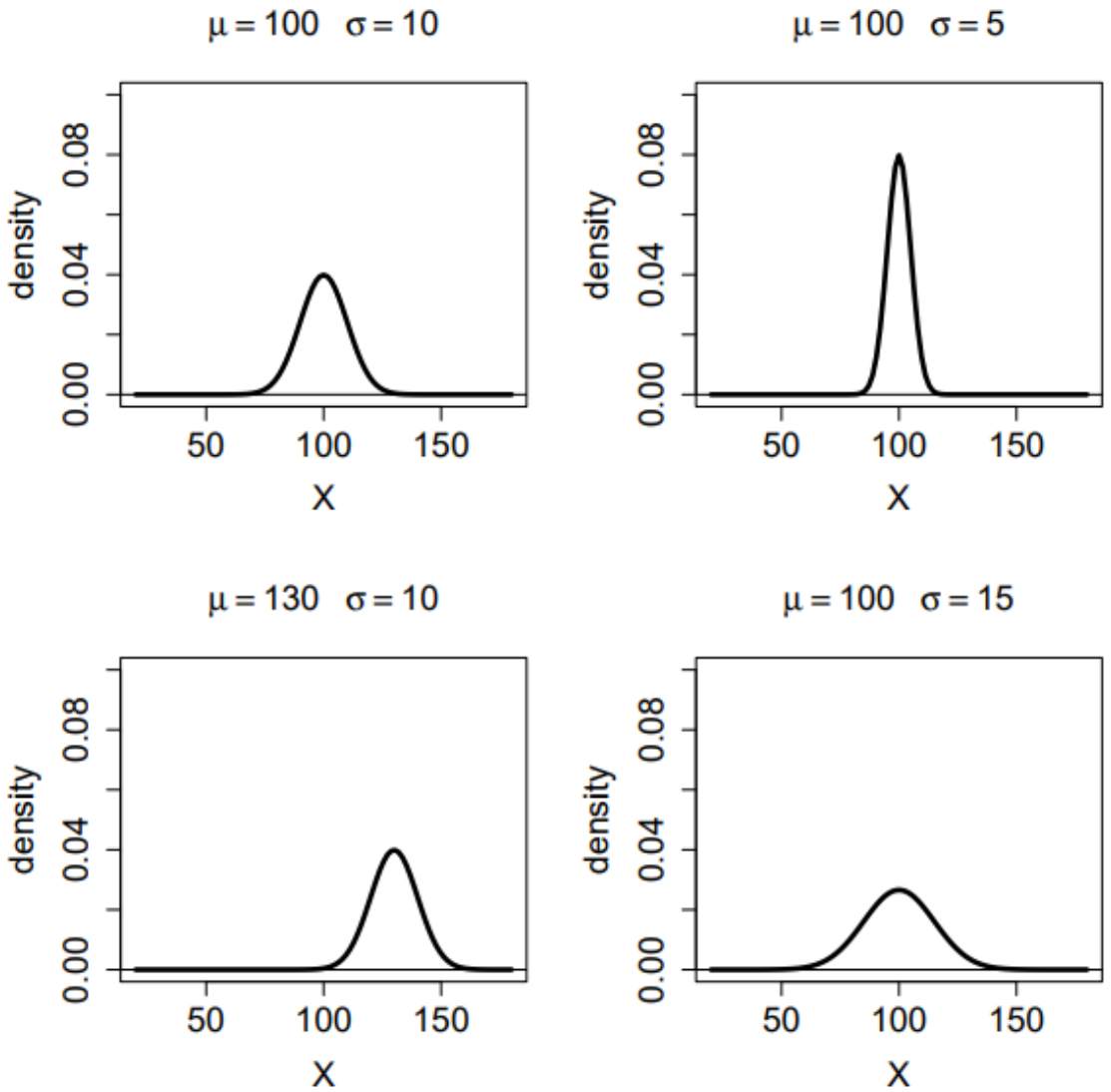
- Similarly, variance is also defined differently.

$$Var(X) = \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx$$

Normal Probability Model (a.k.a Gaussian or Gauss or Laplace-Gauss Distribution)

- There will be many possible probability density functions over a continuous range of values.
- The normal distribution describes a special class of such distributions that are symmetric and can be described by two parameters.
  - $\mu =$ The mean of the distribution.
  - $\sigma =$ The standard deviation of the distribution.
- Changing the values of μ and σ alters the positions and shapes of the distributions.

Normal Probability Model
(a.k.a Gaussian or Gauss or Laplace-Gauss Distribution)



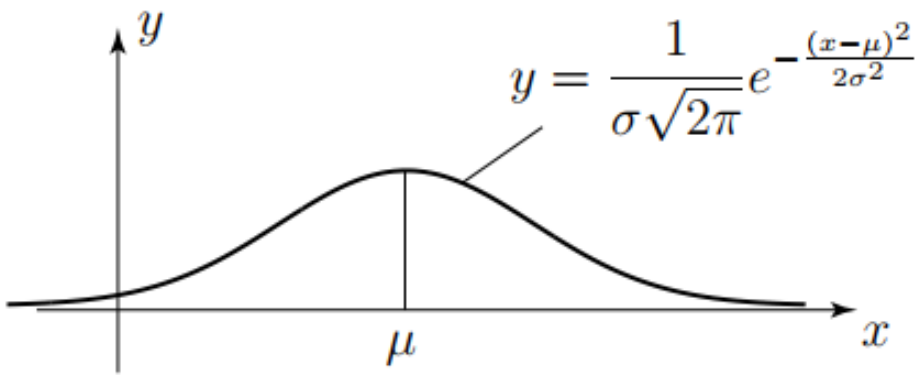
Normal Probability Model (a.k.a Gaussian or Gauss or Laplace-Gauss Distribution)

- If X is normally distributed with mean μ and standard deviation σ, we write

$$X \sim N(\mu, \sigma^2)$$

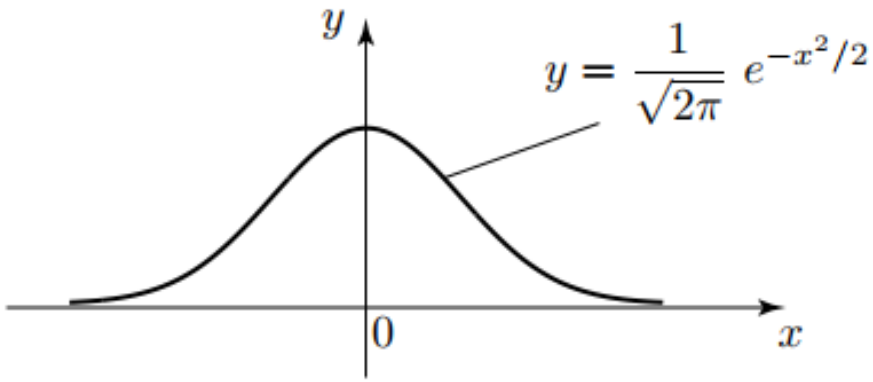
- The probability density function of normal distribution is given by

$$y = f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Standard Normal Distribution

- The standard normal distribution has a mean of zero and a variance of one.
- The following shows the graph of the standard normal distribution which has probability density function $y = f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$



- If the behavior of a continuous random variable X is described by the distribution $N(\mu, \sigma^2)$ then the behavior of the random variable $Z = \frac{x-\mu}{\sigma}$ is described by the standard normal distribution $N(0, 1)$.

We call Z the standardized normal variable.

Example

1. If the random variable X is described by the distribution $N(45, 0.000625)$ then what is the transformation to obtain the standardized normal variable?
   - Given $\mu = 45$, $\sigma^2 = 0.000625$ and so that $\sigma = 0.025$, hence $Z = (X - 45)/0.025$ is the required transformation.
2. When the random variable X takes value between 44.95 and 45.05, between which values does the random variable Z lie?
   - When $X = 45.05, Z = (45.05 - 45)/0.025 = 2$.
   - When $X = 44.95, Z = (44.95 - 45)/0.025 = -2$.
   - Hence Z lies between -2 and 2.

Other math pre-requisites

Please also have a brief review on the following topics on your own:

- Linear Algebra: Basic operations on a vector and a matrix; eigenvalues.
- Calculus: Derivatives and gradients.