Conditional Probability and Bayes's Theorem
Descriptive Statistics

3.6 - 3.8, 2.1 , 2.2 & 2.5

Cathy Poliak, Ph.D.
cathy@math.uh.edu

Department of Mathematics
University of Houston

Lecture 3 - Online

Outline

Assigning Probability

- Suppose there are N total possible outcomes, the probability assigned to each is $\frac{1}{N}$.
- Consider an event A, with $N(A)$ denoting the number of outcomes contained in A. Then,

$$P(A) = \frac{N(A)}{N}$$

Basic Probability Rules

1. $0 \leq P(E) \leq 1$ for each event E.

2. $P(\Omega) = 1$

3. If $E_1, E_2, \ldots$ is a finite or infinite sequence of events such that $E_i \cap E_j = \varnothing$ for $i \neq j$, then $P(\bigcup_i E_i) = \sum_i P(E_i)$. If $E_i \cap E_j = \varnothing$ for all $i \neq j$ we say that the events $E_1, E_2, \ldots$ are **pairwise disjoint**.

★ Mutually exclusive

Other Probability Rules

4. **Complement Rule**: $P(E \sim F) = P(E \cap^{\sim} F) = P(E) - P(E \cap F)$.
   In particular, $P(^{\sim}E) = 1 - P(E)$.

   $= P(\Omega) - P(E)$

5. $P(\varnothing) = 0$

6. **Addition Rule**: $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.

7. If $E_1 \subseteq E_2 \subseteq \ldots$ is an infinite sequence, then
   $P(\bigcup_i E_i) = lim_{i \to \infty} P(E_i)$.

8. IF $E_1 \supseteq E_2 \supseteq \ldots$ is an infinite sequence, then
   $P(\bigcap_i E_i) = lim_{i \to \infty} P(E_i)$.

Example of Probability Rules

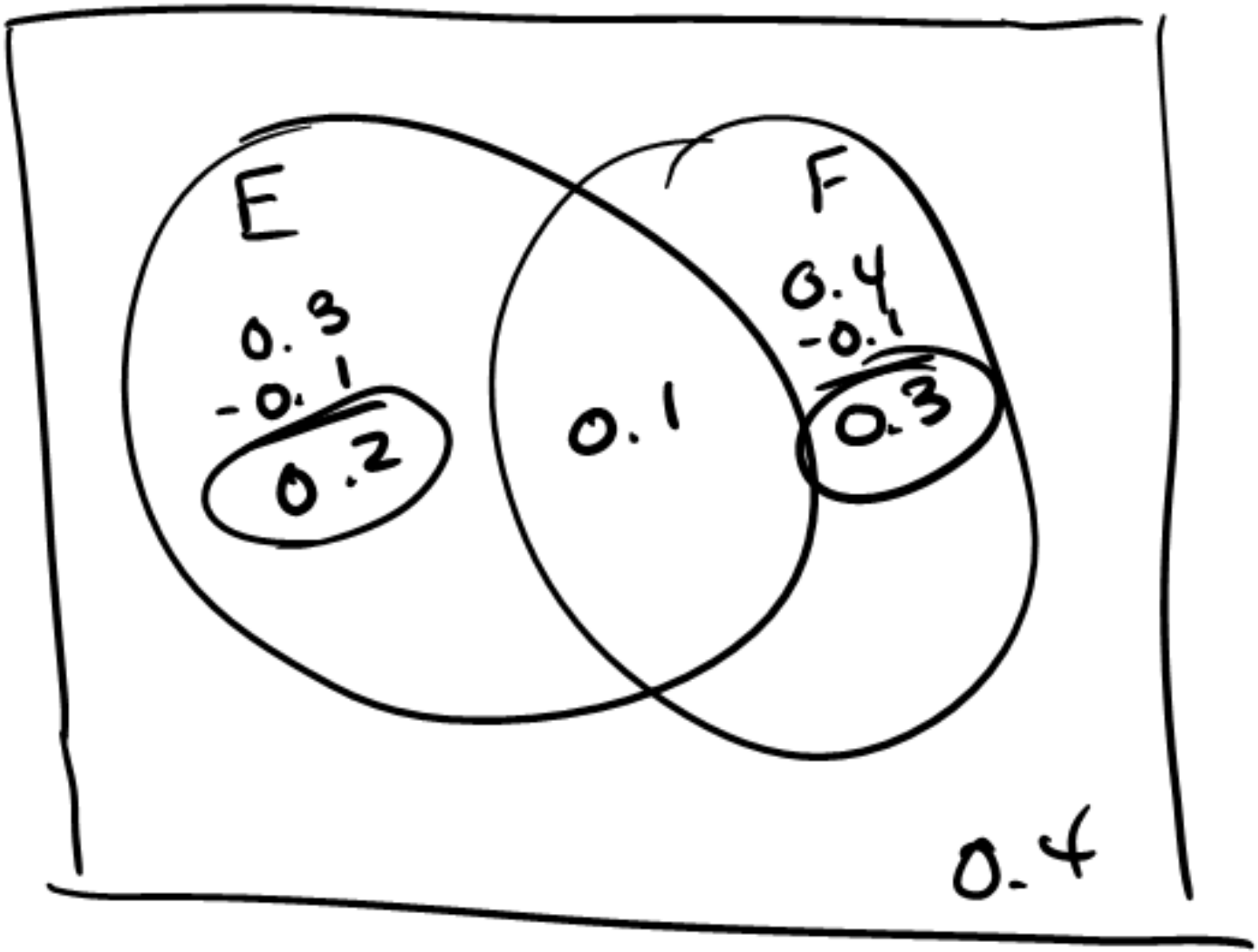
Suppose $P(E) = 0.3$, $P(F) = 0.4$ and $P(E \cap F) = 0.1$. Determine the following probabilities.

1. $P(E \cup F) = P(E) + P(F) - P(E \cap F)$
$= 0.3 + 0.4 - 0.1$
$= 0.6$

2. $P(\sim F) = 1 - P(F) = 1 - 0.4 = 0.6$

3. $P(E \sim F) = P(E \cap \sim F) = P(E) - P(E \cap F)$
$= 0.3 - 0.1$
$= 0.2$

Using the Venn Diagram



General Multiplication Rule

For any two events E and F

$$P(E \cap F) = P(E) \times P(F|E)$$

$$= P(E) * P(F \text{ , given } E)$$

or

$$P(E \cap F) = P(F) \times P(E|F)$$

Where $P(F|E)$ is the probability of F given that the event E has occurred. Similarly $P(E|F)$ is the probability of E given that F has occurred. These types of probabilities are called **conditional probability**. An easy way to determine this calculation is through a tree diagram.

Example General Multiplication Rule

A person must select one of three boxes, each filled with toy cars. The probability of box A being selected is 0.19, of box B being selected is 0.18, and of box C being selected is 0.63. The probability of finding a red car in box A is 0.2, in box B is 0.4, and in box C is 0.9. We are selecting one of the toy cars.

$P(A) = 0.19 \quad P(B) = 0.18 \quad P(C) = 0.63$

$P(R|A) = 0.2 \quad P(R|B) = 0.4 \quad P(R|C) = 0.9$

1. What is the probability that the toy car is red and in box A?

$P(R \cap A) = P(A) * P(R|A) = 0.19(0.2) = 0.038$

2. What is the probability that the toy car is red and in box B?

$P(R \cap B) = P(B) * P(R|B) = 0.18(0.4) = 0.072$

3. What is the probability that the toy car is red and in box C?

$P(R \cap C) = P(C) * P(R|C) = 0.63(0.9) = 0.567$

Law of Total Probability

$P(A_1)+P(A_2)+\cdots+P(A_k)=1$

Let $A_1, A_2, \ldots A_k$ be mutually exclusive and exhaustive events. Then for any other event B,

$P(B\cap A_1)+P(B\cap A_2)+\cdots+P(B\cap A_k)$

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_k)P(A_k)$$

$$= \sum_{i=1}^{k} P(B|A_i)P(A_i)$$

From the previous example, what it the probability that toy car is red?

$$P(R) = P(R\cap A) + P(R\cap B) + P(R\cap C)$$

$$= 0.038 + 0.072 + 0.567$$

$$= 0.677$$

Example of General Multiplication Rule

Suppose we draw two cards from a deck of 52 fair playing cards, what is the probability of getting an ace on the first draw and a king on the second draw?

- Without replacement.

$$P(A^{1st} \cap K^{2nd}) = P(A^{1st}) * P(K^{2nd} | A^{1st})$$
$$= \frac{4}{52} * \frac{4}{51} = 0.006$$

- With replacement. "independent events"

$$P(A^{1st}) * P(K^{2nd} | A^{1st})$$
$$\frac{4}{52} * \frac{4}{52} = 0.0059$$

Picture of Cards

Two events A and B are <u>independent</u> if the probability of one event does not change given the other event.

* $P(A, \text{given } B) = P(A \mid B) = P(A)$

Implies:

General rule: $P(A \cap B) = P(A) * P(B/A)$

If independent, $P(B \mid A) = P(B)$

thus $\boxed{P(A \cap B) = P(A) * P(B)}$ *

Ford Mustangs

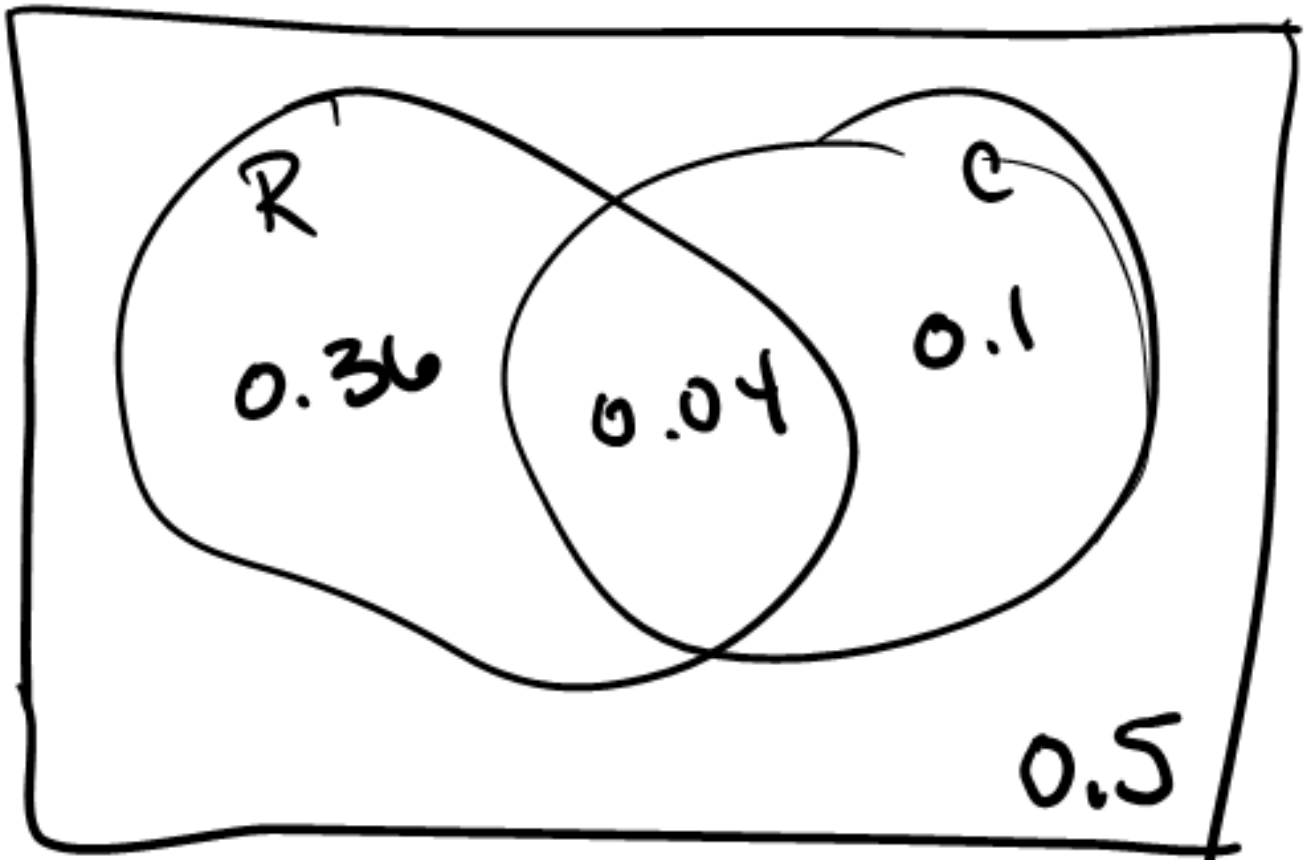
At a Ford dealership, if you select a Ford Mustang at random, the probability it is red is P(R) = 0.40, the probability it is a red convertible $P(R \cap C) = 0.04$, and the probability that it is red or a convertible $P(R \cup C) = 0.50$.

1. What is the probability that a randomly selected Ford Mustang is a convertible?

2. What is the probability that a randomly selected Ford Mustang is not red?

3. What is the probability of getting a convertible out of the red Mustangs?

$P(R) = 0.4$ $P(R \cap C) = 0.04$ $P(R \cup C) = 0.5$



$P(C) = 0.14$

$P(\sim R) = 1 - 0.4 = 0.6$

$P(C|R) = \frac{0.04}{0.4}$

$= 0.1$

Conditional Probability

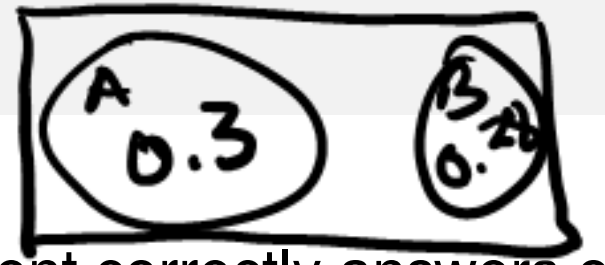
Let A and B be events with $P(B) > 0$. The **conditional probability** of A, given B is:

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

General rule for multiplication: For any two events E and F, $P(E \cap F) = P(E) \times P(F|E)$ or $P(E \cap F) = P(F) \times P(E|F)$.

$$P(R|C) = \frac{P(R \cap C)}{P(C)} = \frac{0.04}{0.14} = 0.2857$$

Example



The probability that a student correctly answers on the first try (the event A) is P(A) = 0.3. If the student answers incorrectly on the first try, the student is allowed a second try to correctly answer the question (the event B). The probability that the student answers correctly on the second try given that he answered incorrectly on the first try is 0.4. Find the probability that the student correctly answers the question on the first or second try.

$P(A) = 0.3 \quad P(\sim A) = 0.7$

a) 0.7

$P(B \mid \sim A) = 0.4$

b) 0.12

If independent $P(B \mid A) = P(B)$ Not mutually exclusive

c) 0.28

$P(B \cap \sim A) = P(\sim A) * P(B \mid \sim A)$

$= 0.7\,(0.4) = 0.28 = P(B$

d) 0.58

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ ↓ 0

Two Frequently Asked Questions

1. When do I add and when do I multiply?
   - Add when finding the chance of events A **or** B or both happening.

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

   - Multiply when finding the chance that both events A **and** B happen.

$$P(A \text{ and } B) = P(A \cap B) = P(A) \times P(B, \text{ given } A) = P(A)P(B|A)$$

Two Frequently Asked Questions

2. What's the difference between disjoint (mutually exclusive) and independent?
   - Two events are disjoint if the occurrence of one prevents the other from happening.
   $$P(A \cap B) = 0$$
   - Two events are independent if the occurrence of one does not change the *probability* of the other.
   $$P(A|B) = P(A)$$

If two events are mutually exclusive $P(A \cap B) = 0$, then they are most definetly dependent.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0}{P(B)} = 0 \neq P(A)$$

Example

Thirty percent of the students at a local high school face a disciplinary action of some kind before they graduate. Of those "felony" students, 45% go on to college. Of the ones who do not face disciplinary action 60% go on to college.

1. Show if events {faced disciplinary action} and {went to college} are independent or not.

A = Students that face disciplinary action

B = Student that go to college

$P(A) = 0.3 \quad P(\sim A) = 0.7$

$P(B|A) = 0.45 \quad P(B|\sim A) = 0.6$

$P(A \cap B) = 0.3(0.45) = 0.135$

$P(\sim A \cap B) = 0.7(0.6) = 0.42$

$P(B) = 0.135 + 0.42 = 0.555$

Prove if Independent:

① $P(B) = P(B|A) = P(B|\sim A)$

Not independent

② $P(B) * P(A) \stackrel{?}{=} P(B \cap A)$

$0.555 * 0.3 \neq 0.135$

Not independent

Dogs and Cats

The probability of owning a dog is 0.6, the probability of owning a cat is 0.4. The probability of owning a dog and a cat is 0.24.

1. What is the probability that out of cat owners, they also own a dog?

2. What is the probability that out of dog owners, they also own a cat?

3. Are "owning a dog" and "owning a cat" independent events?

Buyers of Computers

Approximately 5 months after the introduction of the iMac, Apple reported that 32% of iMac buyers were frist-time computer buyers. At the same time, approximately 5% of all computer sales were of iMacs. Of buyers who did not purchase an iMac, approximately 40% were first-time computer buyers. Let A = the event bought an iMac and B = the event of first-time computer buyer

1. What is the probability of a person buying an iMac, $P(A)$?

$P(A) = 0.05$

2. What is the probability that a person is a first-time computer buyer, given they bought an iMac, $P(B|A)$?

$P(B|A) = 0.32$

3. What is the probability that a person bought an iMac *and* is a first-time computer buyer, $P(A \cap B)$?

$0.05(0.32) = 0.016$

$P(A) = 0.05 \quad P(\sim A) = 0.95$

$P(B|A) = 0.32 \quad P(B|\sim A) = 0.4$

What is the probability that the buyer is a first-time buyer? P(B)

Law of Total Prob

$$P(B) = P(B \cap A) + P(B \cap \sim A)$$

$$= P(A)P(B|A) + P(\sim A)P(B|\sim A)$$

$$= 0.05(0.32) + 0.95(0.4)$$

$$= 0.016 + 0.38$$

$$= 0.396$$

1

Do we have independence between first-time buyers and iMac buyers?

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

$$0.3\ (0.396) \overset{?}{=} 0.016$$

$$0.1188 \neq 0.016$$

Not have Independence

4. What is the probability of a person buying an iMac, given they are first-time buyers?

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.016}{0.396}$$

$$= 0.0404$$

Law of Total Probability

Given: $P(B|A)$ want to find $P(A|B)$

Bayes' Rule

- The probability of a person buying an iMac, given they are first-time buyers is an example of using **Bayes' rule**.
- Given a prior (initial) probability then from sources we obtain additional information about the events.
- From these events we revise the probabilities and get a posterior probability.
- This is an application of the General Multiplication Rule.
- It might be easier to use either the tree diagram to calculate this probability.

Bayes' Rule

Let B and $A_1, A_2, \ldots, A_k$ be pairwise disjoint events such that each $P(A_i) > 0$ and $\Omega = A_1 \cup A_2 \cup \ldots \cup A_k$ and assume $P(B) > 0$. Then for each i,

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_1)}{\sum_{i=1}^{k} P(B|A_i)P(A_i)}$$

$$= \frac{P(A_i \cap B)}{P(B)}$$

Example

A rare disease exists in which only 1 in 500 are affected. A test for the disease exists but of course it is not infallible. A correct positive result (patient actually has the disease) occurs 95% of the time while a false positive result (patient does not have the disease) occurs 1% of the time. If a randomly selected individual is tested and the result is positive, what is the probability that the individual has the disease?

A = the person is affected

B = the test is positive

$P(A|B) = 0.16$

Aircraft Disappearance

Seventy percent of the light aircraft disappear while in flight in a certain country are subsequently discovered. Of the aircraft that are discovered, 60% have an emergency locator, whereas 90% of the aircraft not discovered do not have such a locator. Suppose a light aircraft has disappeared.

A = discovered B = emergency locator

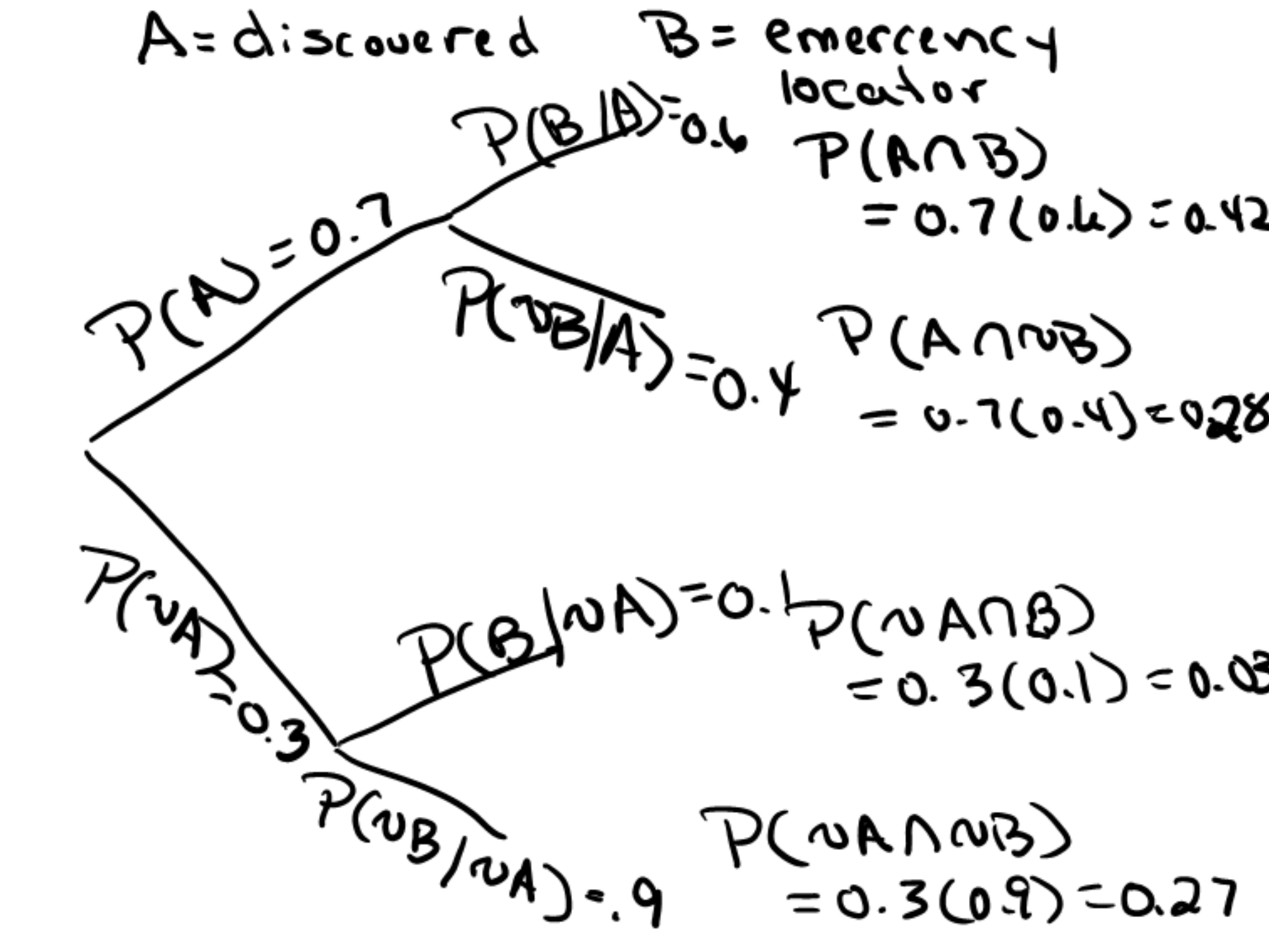
$P(A) = 0.7$ $P(B|A) = 0.6$ $P(\sim B|\sim A) = 0.9$

a) If it has an emergency locator, what is the probability that it will not be discovered?

$$P(\sim A|B) = \frac{P(\sim A \cap B)}{P(B)} = \frac{0.03}{0.03+0.42} = 0.0667$$

b) If it does not have an emergency locator, what is the probability that it will be discovered?

$$P(A|\sim B) = \frac{P(A \cap \sim B)}{P(\sim B)} = \frac{0.28}{0.28+0.27} = 0.5091$$


A = discovered
B = emergency locator
P(A) = 0.7
P(B|A) = 0.6
P(∼B|A) = 0.4
P(∼A) = 0.3
P(B|∼A) = 0.1
P(∼B|∼A) = .9
P(A∩B) = 0.7(0.6) = 0.42
P(A∩∼B) = 0.7(0.4) = 0.28
P(∼A∩B) = 0.3(0.1) = 0.03
P(∼A∩∼B) = 0.3(0.9) = 0.27

Example Two-Way Table

A clothing store targets young customers (ages 18 through 22) wishes to determine whether the size of the purchases related to the method payment. Suppose a customer is picked at random. The following is 300 customers the amount of the purchase and method payment.

	Cash	Credit	Layaway	Total
Under $40	60	30	10	100
$40 or more	40	100	60	200
Total	100	130	70	300

Example

1. What is the probability that the customer paid with a credit card?

$$P(\text{credit}) = \frac{130}{300} = 0.4333$$

2. What is the probability that the customer purchased under \$40?

$$P(\text{under } 40) = \frac{200}{300} = 0.6667$$

3. What is the probability that the customer paid with credit card given that the purchase was under \$40?

$$P(\text{credit} \mid \text{under } 40) = \frac{100}{200} = 0.5$$

4. What is the probability that the customer paid with credit card and that the purchase was under \$40?

$$P(\text{Credit} \cap \text{under } 40) = \frac{100}{300} = 0.3333$$

5. Are type of payment and amount of purchase independent?

No because $P(\text{Credit}) \neq P(\text{Credit} \mid \text{under } 40)$

Distributions

- When observing a data set, one of the first things we want to know is how each variable is *distributed*.
- The **distribution** of a variable tells us what values it takes and how often it takes these values based on the individuals.
- The distribution of a variable can be shown through tables, graphs, and numerical summaries.

Describing distributions

- An initial view of the distribution and the characteristics can be shown through the graphs. This is called **data visualization**.
- Then we use numerical descriptions to get a better understanding of the distributions characteristics.

Distributions for Categorical (Factor) Variables

- Lists the categories and gives either the count or the percent of cases that fall in each category.
- One way is a **frequency table** that displays the different categories then the count or percent of cases that fall in each category.
- Then we look at the graphs (bar or pie) to determine the distribution of a categorical variable.

A Data Set: Course Grades From Previous Semesters

https://www.math.uh.edu/~cathy/Math3339/data/grades.txt

Student	Score	Grade	Tests	Quiz	HW	Opt-out	Session
1	100.707	A	99.233	87.308	101.270	yes	Sp16
2	81.310	B	75	98.231	64.444	yes	Sp16
3	8.194	F	14.667	12.769	3.175	no	Sp16
4	90.449	A	91.533	77.231	82.222	yes	Sp16
5	68.461	D	65.783	81.769	68.571	no	Sp16
6	103.955	A	103.32	97.923	101.905	yes	Sp16
7	92.889	A	95.6	85.923	75.556	no	Sp16
8	84.805	B	83.2	79.385	75.238	yes	Sp16
9	91.640	A	89.967	91.231	85.079	yes	Sp16
10	22.316	F	17.433	40.615	44.444	no	Sp16
11	98.363	A	94.167	99.231	101.587	yes	Sp16
12	49.250	F	43.917	73.077	78.095	no	Sp16
13	16.967	F	15.5	20.077	29.841	no	Sp16
14	50.747	F	45.533	67.385	57.460	no	Sp16
15	43.184	F	72.983	47.462	38.413	no	Sp16
16	100.845	A	98.667	96.231	100.317	yes	Sp16
17	84.195	B	77.5	87.154	95.556	yes	Sp16
18	84.400	B	78.733	78.615	82.540	yes	Sp16
19	67.170	D	74.3	68.538	72.063	no	Fal15
20	87.413	B	92	82.077	77.778	yes	Fal15
21	67.899	D	71.8	71.077	84.127	no	Fal15
22	74.676	C	70.083	83.308	73.016	no	Fal15
23	40.054	F	44.133	21.308	33.333	no	Fal15
24	101.014	A	101.08	98.923	95.873	no	Fal15
25	11.972	F	17.1	10.385	3.810	no	Fal15
26	79.831	B	86.233	71.923	46.667	no	Fal15
27	83.301	B	94.6	69.692	60.317	no	Fal15
28	72.299	C	64.967	67.615	99.394	no	Sum16
29	83.821	B	77.2	80.923	83.030	yes	Sum16
30	90.703	A	83.617	87.923	80.000	no	Sum16

Frequency Tables

Oup-out	Percent
Yes	40%
No	60%

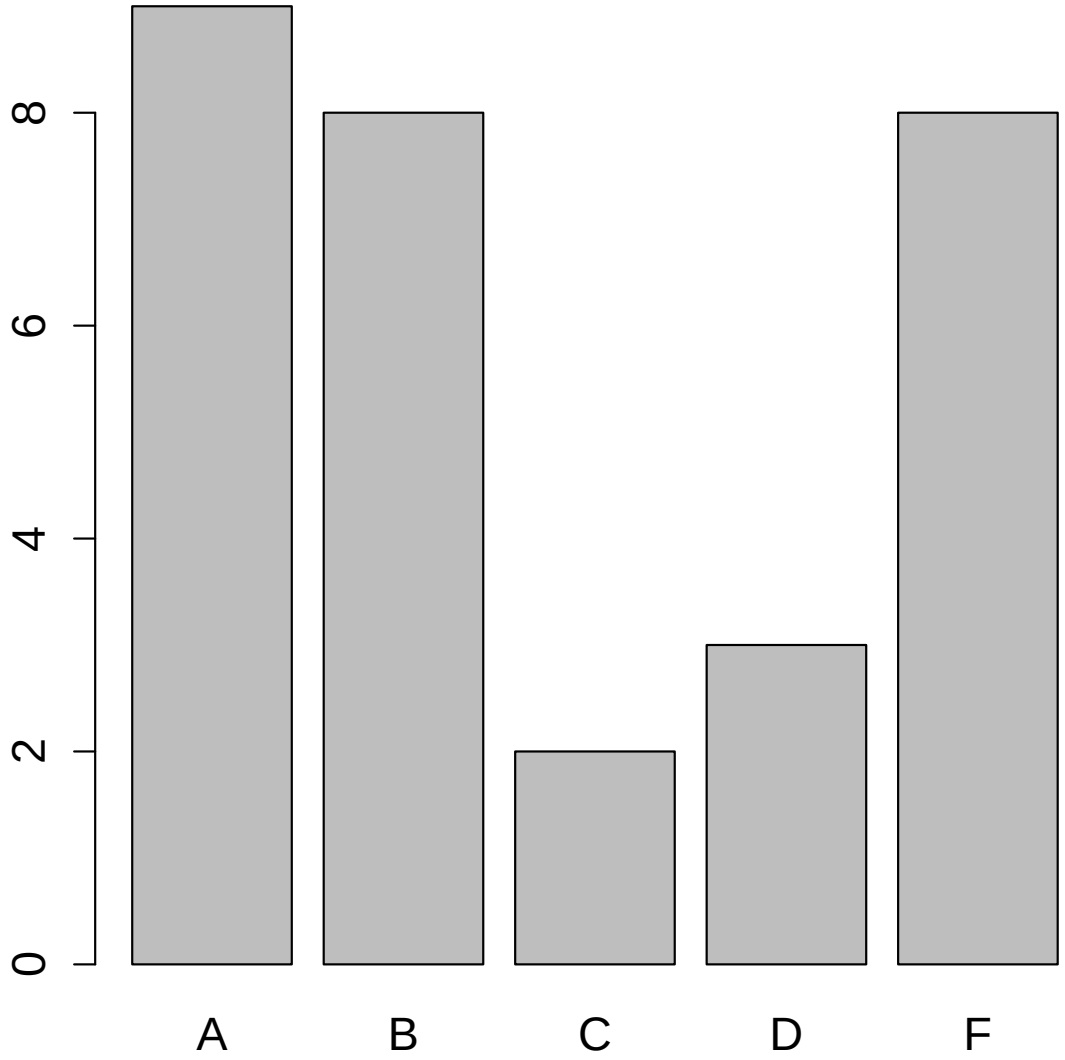
Relative Frequencies

Grade	Percent
A	30%
B	26.67%
C	6.67%
D	10%
F	26.67%

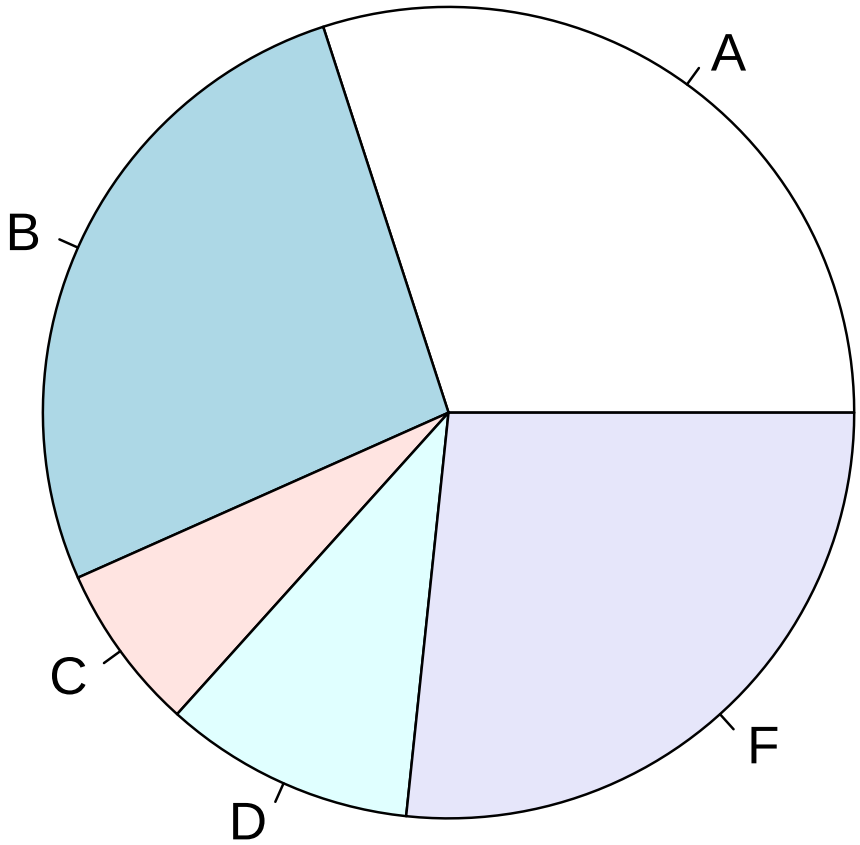
Describing Data By Graphs

- Graphs are an easy and quick way to describe the data.
- Types of graphs that we use depends on the type of data that we have.
- Graphs for **categorical** (factor) variables.
  - **Bar graphs**: Each individual bar represents a category and the height of each of the bars are either represented by the count or percent.
  - **Pie charts**: Helps us see what part of the whole each group forms.
- Graphs for **quantitative** variables.
  - Dot plot
  - Stem plot
  - Histogram
  - Box plot
  - Cumulative Frequency Plot (Ogive)

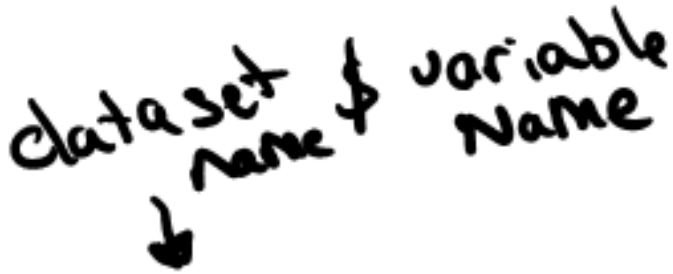
Bar Graph of Letter Grades



Pie Chart of Letter Grades



R code



- First create a table: counts = table(grades$Grade)
- For bar graph: barplot(counts)
- For pie chart: pie(counts)

Describing distributions of Quantitative (Numeric) Variables

- The **distribution** of a variable tells us what values it takes and how often it takes these values.
- There are four main characteristics to describe a distribution:
  1. Shape
  2. Center
  3. Spread
  4. Outliers

Describing a distribution

- Shape
  - 
    A distribution is **symmetric** if the right and left sides of the graph are approximately mirror images of each other.
  - 
    A distribution is **skewed to the right** if the right side (higher values) of the graph extends much farther out than the left side.
  - 
    A distribution is **skewed to the left** if the left side (lower values) of the graph extends much farther out than the right side.
  - 
    A distribution is **uniform** if the graph is at the same height (frequency) from lowest to highest value of the variable.

- Center - the values with roughly half the observations taking smaller values and half taking larger values.

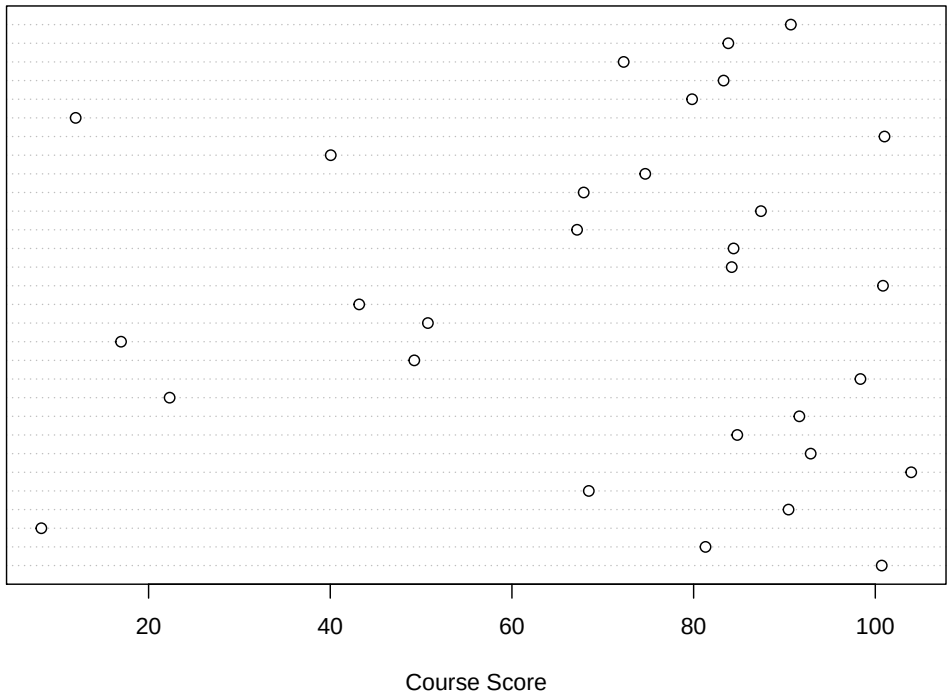
- Spread -from the graphs we describe the spread of a distribution by giving *smallest and largest values*.

- Outliers - individual values that falls outside the overall pattern.

Dot plots

A **dot plot** is made by putting dots above the values listed on a number line. When a value occurs more than once, there is a dot for each occurrence, and these dots are stacked vertically. Rcode:

```
dotchart(grades$Score, xlab = "Course Score")
```



Steps for Creating A Stem - and - leaf plot

1. Select one or more leading digits for the stem values. The trailing digits become the leaves.
2. List possible stem values in a vertical column
3. Record the leaf for every observation beside the corresponding stem value.
4. Indicted the units for stems and leaves someplace in the display.

Rcode: stem(**dataset name$variable name**)

Example Of Stem-and-Leaf Plot

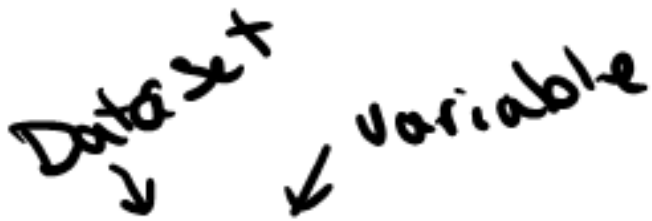
The following is the number of wins of the 2015 baseball season for each pitcher in the MBL. Put together a stem-and-leaf display.

86.4 78.6 69.6 63.6 70.4 61.9 56.5 53.8 66.7 66.7
~~26.1~~ 52.2 56.3 45.5 47.8 63.2 70.8 55.0 56.5 52.6
52.4 58.8 57.9 50.0 65.0 53.3 52.2 57.9 ~~40.9~~ 44.0
45.8 47.4 51.9 ~~37.5~~ ~~27.3~~ ~~42.1~~ 60.9 ~~28.6~~

Stem = tens , leaves = ones

2	6 7 9
3	8
4	1 2
5	
6	
7	
8	

Stem-and-leaf Plot in R studio




```
> stem(Era$Wins)

The decimal point is 1 digit(s) to the right of the |

2 | 679
3 | 8
4 | 1246678
5 | 022223345677889
6 | 1234577
7 | 0019
8 | 6
```

Stem-and-leaf Plot of ERA

```
> stem(Era$ERA)

The decimal point is at the |

1 | 78
2 | 1
2 | 567889
3 | 00023344
3 | 67777889
4 | 0001111233
4 | 579
```

1.7, 1.8

2-1

skewed left

center ≈ 3.3

spread: 4.9 − 1.7 = 3.2

The Information from a Stem-and-Leaf and Dot Plots

A stem-and-leaf and dot plots convey information about the following aspects of the variable:

- Identification of a typical or representative value
- Extent of spread about the typical value
- Presence of any gaps in the data
- Extent of symmetry in the distribution of values
- Number and location of "peaks"
- Presence of any outlying values

Stem-and-Leaf Plot of Grades

```
> stem(grades$Score,scale=0.5)

The decimal point is 1 digit(s) to the right of the |

0  | 827
2  | 2
4  | 0391
6  | 78825
8  | 0134445701238
10 | 1114
```

1. What is the "shape" of this distribution?

   a) skewed left b) skewed right
   c) symmetric d) uniform

2. What is the approximate center of this distribution?

   a) 50 b) 82 c) 8.5 d) 4

Frequency and Relative Frequency

- The **frequency** for any particular counting variable x is the number of times that value occurs in the data set.
- The **relative frequency** of a value is the fraction or proportion of times the value occurs:

$$\text{relative frequency of a value} = \frac{\text{number of times the value occurs}}{\text{number of observations in the data set}}$$

- Constructing a histogram for measurement data entails subdividing the measurement axis into a suitable number of **class intervals** or **classes**, such that each observation is contained in exactly one class.

Frequency Table of Scores

Cumulative Relative Frequency

Score	Tally	Frequency	Relative Frequency	
0 - 20		3	$\frac{3}{30} = 0.1$	0.1
20 - 40		1	$\frac{1}{30} = 0.0333$	0.1333
40 - 60		4	$\frac{4}{30} = 0.1333$	0.2667
60 - 80		6	$\frac{6}{30} = 0.2$	0.4667
80 - 100		12	$\frac{12}{30} = 0.4$	0.8667
100 - 120		4	$\frac{4}{30} = 0.1337$	1.000

Drawing Histograms

1. Determine the frequency and relative frequency for each class.
2. Mark the class boundaries on a horizontal measurement axis.
3. Above each class interval, draw a rectangle whose height is the corresponding frequency (or relative frequency).

If the classes are of unequal class widths, the height of each rectangle is:

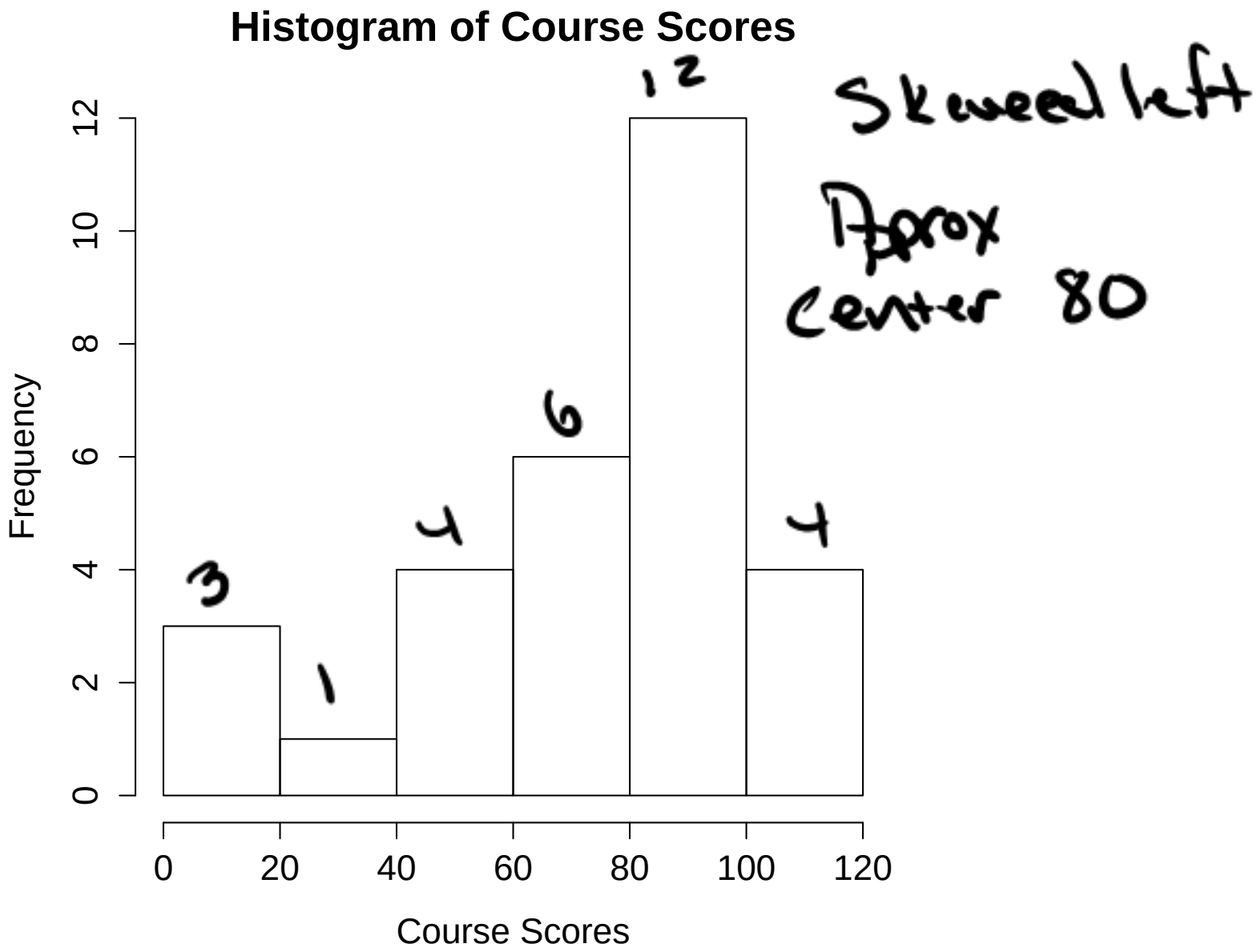
$$\text{rectangle height} = \frac{\text{relative frequency of the class}}{\textbf{class width}}$$

The resulting rectangle heights are usually called **densities**, and the vertical scale is the **density scale**.

Histograms

- Bar graph for quantitative variables.
- Values of the variable are grouped together.
- Bars are touching.
- The width of the bar represents an interval of values (range of numbers) for that variable.
- The height of the bar represents the number of cases within that range of values.

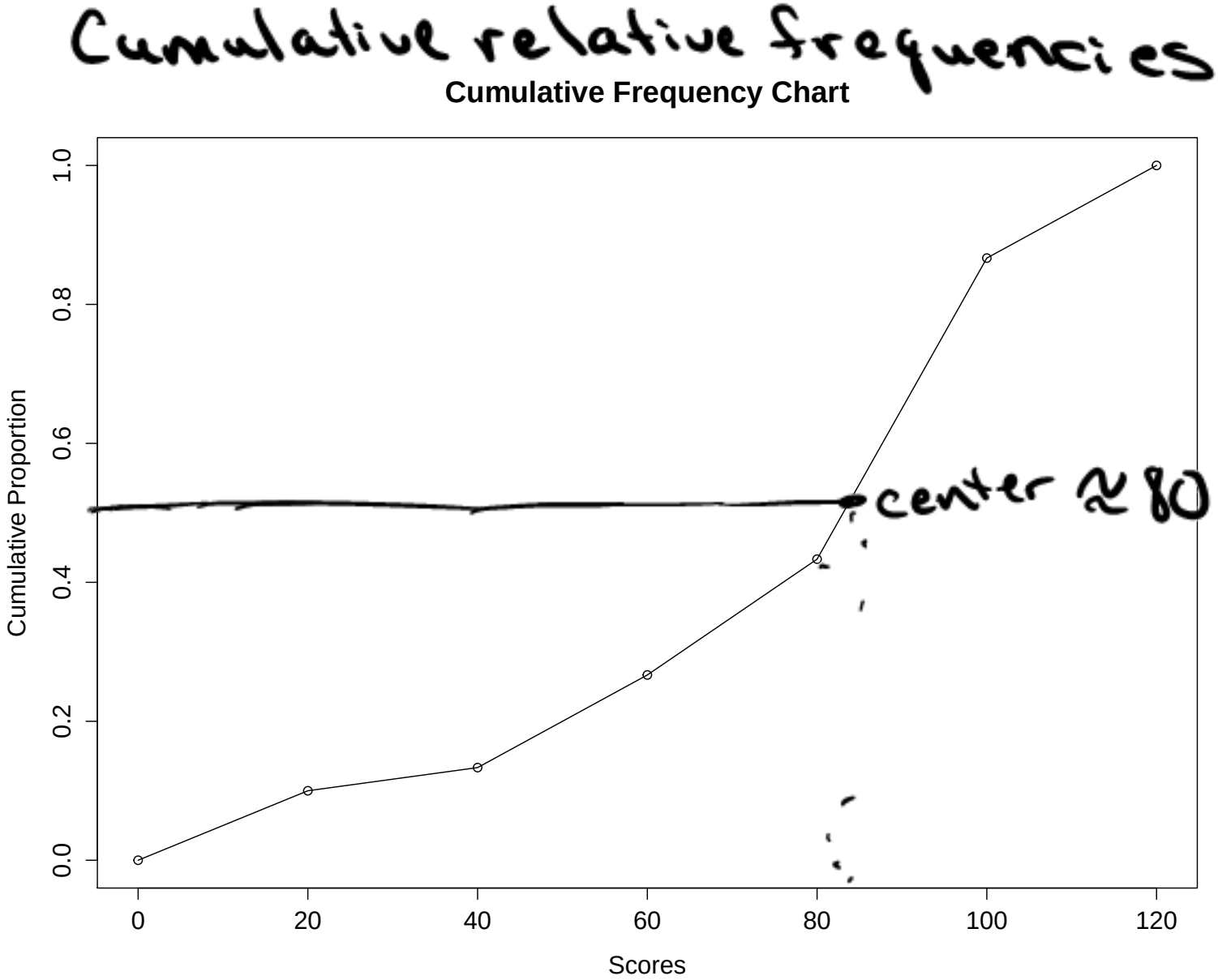
Histogram of Course Score



Cumulative Frequency Polygon

- Plot a point above each upper class boundary at a height equal to the cumulative frequency of the class.
- Connect the plotted points with line segments.
- A similar graph can be used with the cumulative relative frequency.

Cumulative Percent Polygon



Describing Quantitative Variables with Numbers

- Location - mean, median, quartile, percentiles and trimmed means
- Variability (Spread) - range, interquartile range, variance, or standard deviation, and coefficient of variation

Parameters and Statistics

- A **parameter** is a number that describes the **population**. A parameter is a fixed number, but in practice we usually do not know its value.
- A **statistic** is a number that describes a **sample**. The value of a statistic is known when we have taken a sample, but it can change from
- The purpose of sampling or experimentation is usually to use statistics to make statements about unknown parameters, this is called **statistical inference**.

Notation of Parameters and Statistics

Name	Statistic	Parameter
mean	$\bar{x}$	μ mu
standard deviation	s	σ sigma
correlation	r	ρ rho
regression coefficient	b	β beta
proportion	$\hat{p}$	p

Example

A carload lot of ball bearings has a mean diameter of **2.503** centimeters. This is within the specifications for acceptance of the lot by the purchaser. The inspector happens to inspect 100 bearings from the lot with a mean diameter of **2.515** centimeters. This is outside the specified limits, so the lot is mistakenly rejected. Is each of the bold numbers a parameter or a statistic?

Parameter = $\mu = 2.503$

Statistic = $\bar{x} = 2.515$

Presidential Approval Rating

On January 25, 2017 by Gallup.com, 46% of Americans approved of how Trump is doing as President. Gallup tracks daily the percentage of Americans who approve or disapprove of the job Donald Trump is doing as president. Daily results are based on telephone interviews with approximately 1,500 national adults; Margin of error is $\pm$ 3 percentage points.

Is this 46% a statistic or parameter?

$\tilde{p} = 0.46$ Statistic

Estimate the parameter with ME

$0.43 < p < 0.49$

The Mean

- Most common measure of center.
- Arithmetic average.
- To calculate the mean of a set of observations $x_1, x_2, \ldots, x_n$, add their values and divide by the number of observations ***n***.
- Denoted: $\bar{x}$ called x-bar if the data is from a sample, μ, called "mu" if the data is from the entire population.

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{n} x_i$$

- Where n is the size of the sample and N is the size of the population.

The Median

- The **median** M is the midpoint of a data set such that half of the observations are smaller and the other half are larger.
- Arrange all observations in order of size, from smallest to largest (with any repeated values included so that every sample appears in the ordered list).
- Then

$$M = \begin{cases} \text{The single middle value if } n \text{ is odd} = & \left(\frac{n+1}{2}\right)^{th} \text{ ordered value} \\ \text{The average of the two middle values if } n \text{ is even} = & \text{average of } \left(\frac{n}{2}\right)^{th} \text{ and } \left(\frac{n}{2}+1\right)^{th} \text{ ordered values} \end{cases}$$

Calculate the mean and median

The following is a stem-and-leaf plot of the course scores. Determine, the mean and median of the course scores.

```
The decimal point is 1 digit(s) to the right of the |

0  |  8
1  |  27
2  |  2
3  |
4  |  039
5  |  1
6  |  788
7  |  25
8  |  01344457
9  |  01238
10 |  1114
```

$$\text{mean} = \frac{2131}{30} = 71.0333$$

$$\text{median} = \frac{81 + 83}{2} = 82$$

Finding mean and median in R

```
scores=c(8,12,17,22,40,43,49,51,67,68,68,72,75,80,81,
83,84,84,84,85,87,90,91,92,93,98,101,101,101,104)
mean(scores)
[1] 71.03333
median(scores)
[1] 82
```

Example: Test Scores

The test scores of a class of 20 students have a mean of 71.6 and the test scores of another class of 14 students have a mean of 78.4. Find the mean of the combined group.

$n(\text{class1}) = 20 \quad \text{mean}(\text{class1}) = 71.6$

$n(\text{class2}) = 14 \quad \text{mean}(\text{class2}) = 78.4$

$$\text{mean}(\text{combined}) = \frac{\text{sum}}{n}$$

$$= \frac{20(71.6) + 14(78.4)}{34}$$

$$= 74.4$$

Mean vs. Median

- If the mean and the median are both numbers that describe the center of the values then why do we have different values?
- If the data has values that are **outliers** values that are beyond the range of the others, the mean is going toward these outliers.

Mean vs. Median

$\bar{x}$(score) = 71 and M(score) = 82 skewed left mean < median

- If the mean and the median are both numbers that describe the center of the values then why do we have different values?
- If the data has values that are **outliers** values that are beyond the range of the others, the mean is going toward these outliers.
- The median is resistant to extreme values (outliers) in the data set.
- The mean is NOT robust against extreme values.

Basketball Team

"Should we scare the opposition by announcing our mean height or lull them by announcing our median height?"