Math 140
Introductory Statistics

Professor Silvia Fernández

Lecture 8

Based on the book *Statistics in Action*
by A. Watkins, R. Scheaffer, and G. Cobb.

5.1 Models of Random Behavior

- Outcome: Result or answer obtained from a chance process.
- Event: Collection of outcomes.
- Probability: Number between 0 and 1 (0% and 100%). It tells how likely it is for an outcome or event to happen.
  - $P = 0$ The event cannot happen.
  - $P = 1$ The event is certain to happen.

5.1 Models of Random Behavior

- $P(A)$ = the probability that event *A* happens
- $P(not\ A) = 1 - P(A)$ = the probability that event *A* doesn't happen.
- The event *not A* is called the **complement** of event *A.*

Where do Probabilities come from?

- *Observed data* (long-run relative frequencies).

  For example, observation of thousands of births has shown that about 51% of newborns are boys. You can use these data to say that the probability of the next newborn being a boy is about 0.51.
- *Symmetry* (equally likely outcomes).

  If you flip a fair coin, there is nothing about the two sides of the coin to suggest that one side is more likely than the other to land facing up. Relying on symmetry, it is reasonable to think that heads and tails are equally likely. So the probability of heads is 0.5.
- *Subjective estimates*.

  What's the probability that you'll get an A in this statistics class? That's a reasonable, everyday kind of question, and the use of probability is meaningful, but you can't gather data or list equally likely outcomes. However, you *can* make a subjective judgment.

Equally likely outcomes.

- If we have a list of all possible outcomes and **all of them are equally likely** then

$$P\text{(specific outcome)} = \frac{1}{\text{total number of equally likely outcomes}}$$

$$P\text{(event)} = \frac{\text{number of outcomes in event}}{\text{total number of equally likely outcomes}}$$

Examples

- Flipping a coin.
- P (heads) = 1/2
- P (tails) = ½

- Rolling a fair die.
- P (2) = 1/6
- P (5) = 1/6
- P (odd) = P (1 or 2 or 3) = 3/6
- P (even less than 5) = P (2 or 4) = 2/6 =1/3.

Tap vs. Bottled Water: The problem

- Jack and Jill, just won a contract to determine if people can tell tap water (T) from bottled water (B).
- They will give each person in their sample both kinds of water, in random order, and ask which is the tap water.
- Assuming that the tasters can't identify tap water, what is the probability that two tasters will guess correctly and choose *T*?

Tap vs. Bottled Water: Who is right?

Jack: There are three possible outcomes: Neither person chooses *T*, one chooses *T*, or both choose *T*. These three outcomes are equally likely, so each outcome has probability ⅓. In particular, the probability that both choose *T* is ⅓ .

Jill: Jack, did you break your crown already? I say there are *four* equally likely outcomes: The first taster chooses *T* and the second also chooses *T* (*TT*); the first chooses *T* and the second chooses *B* (*TB*); the first chooses *B* and the second chooses *T* (*BT*); or both choose *B* (*BB*). Because these four outcomes are equally likely, each has probability ¼ . In particular, the probability that both choose *T* is ¼ , not ⅓.

Tap vs. Bottled Water: Simulation.

- Jack and Jill use two flips of a coin to simulate the taste-test experiment with two tasters who can't identify tap water.
- Two tails represented neither person choosing the tap water, one heads and one tails represented one person choosing the tap water and the other choosing the bottled water, and two heads represented both people choosing the tap water.

Number Who Choose *T*	Frequency	Relative Frequency
0	782	0.26
1	1493	0.50
2	725	0.24
Total	**3000**	**1.00**

Display 5.2 Results of 3000 simulations for two tasters when $P(T) = 0.5$.

- So it seems that Jill is right.

Law of Large Numbers

- In a random sampling, the larger the sample, the closer the proportion of successes in the sample tends to be the proportion in the population.
- Example, simulation of flipping a coin.

Number of Flips	10	100	1000	10000	100000
Heads	2	45	525	4990	50246
Tails	8	55	475	5010	49754

Sample Space

- A Sample Space for a chance process is a complete list of disjoint outcomes.
- Complete means that no possible outcomes are left off the list.
- Disjoint (or mutually exclusive) means that no two outcomes can occur at once.

- Often by symmetry we can assume that the outcomes on a sample space are equally likely. But to verify this we need to collect data and see if indeed each of the outcomes occurs the same number of times (approximately).

Examples

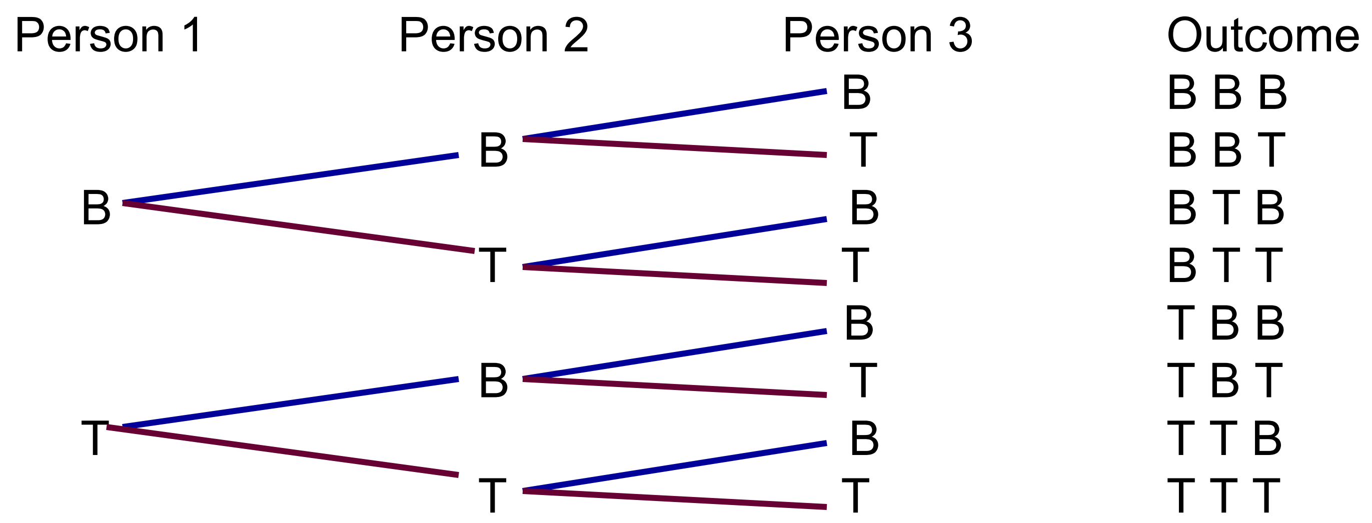
- Rolling a fair die.
  - Sample Space: {1,2,3,4,5,6}
  - P (4)= 1/6
  - P (number is even)= 3/6 = 1/2
- Selecting a card from a poker deck.
  - Sample Space: {A♥,2♥,3♥,…,Q♥,K♥,
    A♦,2♦,3♦,…,Q♦,K♦,
    A♣,2♣,3♣,…,Q♣,K♣,
    A♠,2♠,3♠,…,Q♠,K♠}

Examples

- Selecting a card from a poker deck.
  - Sample Space: {A♥,2♥,3♥,…,Q♥,K♥,
    A♦,2♦,3♦,…,Q♦,K♦,
    A♣,2♣,3♣,…,Q♣,K♣,
    A♠,2♠,3♠,…,Q♠,K♠}
  - P (3♥) = 1/52
  - P (Ace) = 4/52 = 1/13
  - P (♦) = 13/52 = 1/4

A random process is repeated several times

- To list the total list of outcomes when a random process is made up of many repetitions of another random process we can make a tree diagram.
- Example. Jack and Jill give samples of tap water (T) or bottled water (B) at random to three persons so that they taste it and see if they recognize tap water or not.



Fundamental Counting Principle

- For a two-stage process, with n_1 possible outcomes for stage 1 and n_2 possible outcomes for stage 2, the number of possible outcomes for the two stages together is $n_1 n_2$
- More generally, if there are k stages, with n_i possible outcomes for stage i, then the number of possible outcomes for all k stages taken together is $n_1 n_2 n_3 \ldots n_k$.

Discussion D8 (p. 296)

- Suppose you flip a fair coin five times.
  - a. How many possible outcomes are there?
  - b. What is the probability you get five heads?
  - c. What is the probability you get four heads and one tail?

5.3 Addition Rule and Disjoint Events

- "OR" in mathematics means one, the other, or **both**.
- Two events A and B are called **disjoint** (mutually exclusive) if they have no outcomes in common.
- If A and B are disjoint then

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

- Similarly if A, B, and C are mutually exclusive then

$$P(A \text{ or } B) = P(A) + P(B) + P(C)$$

Discussion *A* or *B*

Type of Fishing	**Number (Thousands)**
All freshwater fishing	31,041
Saltwater fishing	8,885
Total	**35,578**

- Are the categories in the table of Display 6.8 complete? Are they disjoint?
- What is the probability that a randomly selected person who fishes does their fishing in freshwater or in saltwater? How many people fish in both freshwater and saltwater?

Discussion *A* or *B*

Type of Fishing	**Number (Thousands)**
All freshwater fishing	31,041
Saltwater fishing	8,885
Total	**35,578**

- Are the categories in the table of Display 6.8 complete? Are they disjoint?
- Complete: **YES**, any person that fishes does so in either fresh water or salt water (maybe both)
- Disjoint: **NO**, the events "Saltwater" and "Freshwater" have outcomes in common.

Discussion *A* or *B*

Type of Fishing	**Number (Thousands)**
All freshwater fishing	31,041
Saltwater fishing	8,885
Total	**35,578**

- What is the probability that a randomly selected person who fishes does their fishing in freshwater or in saltwater? How many people fish in both freshwater and saltwater?
- P(“Fresh” or “Salt”) = 1

  However,

  P(“Fresh”) = $\frac{31041}{35578}$

  P(“Salt”) = $\frac{8885}{35578}$

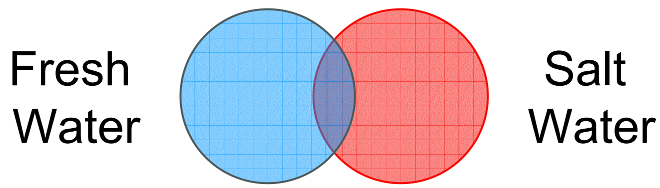
and then,

P(“Fresh”) + P(“Salt”) = $\frac{39926}{35578} > 1$

Discussion *A* or *B*

Type of Fishing	Number (Thousands)
All freshwater fishing	31,041
Saltwater fishing	8,885
Total	**35,578**

- What is the probability that a randomly selected person who fishes does their fishing in freshwater or in saltwater? How many people fish in both freshwater and saltwater?



("Only Salt") + #("Fresh") = 35578
("Only Salt") + 31041 = 35578
("Only Salt") = 35578 – 31041 = **4537**

Similarly
("Only Fresh") = 35578 – 8885
("Only Fresh") = **26693**.

("Only Salt" or "Only Fresh") = **31230**

Then,
#("Fresh" and "Salt") = 35578 – 31230
#("Fresh" and "Salt") = **4348**

A Property of Disjoint Events

- If A and B are disjoint then

$$P(A \text{ and } B) = 0$$

- Example. Suppose two dice are rolled. A is the event of getting a sum of 12, B is the event of getting two odd numbers.

  What is $P(A \text{ and } B)$?
- The only way to get a sum of 12 is (6,6) and both numbers are even. So A and B are disjoint and then $P(A \text{ and } B) = 0$.

Discussion D13 (p. 318)

- Suppose you select a person at random from your school. Which of these pairs of events must be disjoint?
  - a. the person has ridden a roller coaster; the person has ridden a Ferris wheel
  - b. owns a classical music CD; owns a jazz CD
  - c. is a senior; is a junior
  - d. has brown hair; has brown eyes
  - e. is left-handed; is right-handed
  - f. has shoulder-length hair; is a male

General Addition Rule

- For any two events A and B,

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

- In particular if A and B are **disjoint** then $P(A \text{ and } B) = 0$ and then,

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

Example: The Addition Rule for Events that are not Disjoint (p. 320)

- Use the Addition Rule to find the probability that if you roll two dice, you get doubles or a sum of 8.

First Solution.

- A = getting doubles
- B = getting a sum of 8

- $P(A) = P(\{1,1\}, \{2,2\}, \{3,3\}, \{4,4\}, \{5,5\}, \{6,6\}) = 6/36 = 1/6$
- $P(B) = P(\{2,6\}, \{3,5\}, \{4,4\}, \{5,3\}, \{6,2\}) = 5/36$
- $P(A \text{ and } B) = P(\{4,4\}) = 1/36$

- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
  $= 6/36 + 5/36 - 1/36 = 10/36$

Example: The Addition Rule for Events that are not Disjoint (p. 320)

- Use the Addition Rule to find the probability that if you roll two dice, you get doubles or a sum of 8.
- Second Solution.

Doubles?

Sum of 8?

	Yes	No	Total
Yes	**1**	**4**	5
No	**5**	26	31
Total	6	30	36

$$P = \frac{1+5+4}{36} = \frac{10}{36}$$

Example: Computing $P(A \text{ and } B)$ (p. 320)

- In a local school, 80% of the students carry a backpack, *B*, or a wallet, *W*. Also, 40% carry a backpack and 50% carry a wallet. If a student is selected at random, find the probability that the student carries both a backpack and a wallet.

	B	No B	Total
W	10%	40%	50%
No W	30%	20%	50%
Total	40%	60%	100%

5.4 Conditional Probability

- The Titanic sank in 1912 without enough lifeboats for the passengers and crew. Almost 1500 people died, most of them men. Was that because a man was less likely than a woman to survive? Or did more men die simply because men outnumbered women by more than 3 to 1 on the Titanic?

Gender

Survived?

	Male	**Female**	**Total**
Yes	367	344	711
No	1364	126	1490
Total	1731	470	2201

Definition of Conditional Probability

- Conditional Probability refers to the probability of a particular event where additional information is known.
- We write it the following way

  $P(S|T)$ = "probability of S given that T is known to have happened"
- For short we refer to $P(S|T)$ as the **probability of S given T**.

Example: The Titanic and Conditional Probability (p. 326)

- Suppose you pick a person at random from the list of people aboard the Titanic. Let S be the event that this person survived, and let F be the event that the person was female.

  Find $P(S|F)$

Gender

Survived?

	Male	**Female**	**Total**
Yes	367	344	711
No	1364	126	1490
Total	1731	470	2201

- To find $P(S|F)$ restrict your sample space to only the 470 females (the outcomes for which we know the condition F is true).
- Then among these the favorable outcomes are 344. Thus

$$P(S \mid F) = \frac{344}{472} \approx 0.732$$

- So 73.2 % of the females survived even though only 32.3 % of the people survived.

Discussion P23. (p. 334)

- For the Titanic data, let S be the event a person survived and F be the event a person was female. Find and interpret these probabilities.
  - a. $P(F)$ and $P(F|S)$
  - b. $P(\text{not } F)$, $P(\text{not } F|S)$, and $P(S|\text{not } F)$

Gender

Survived?		Male	Female	Total
	Yes	367	344	711
	No	1364	126	1490
	Total	1731	470	2201

- $P(F)$ = Probability of a female passenger

$$P(F) = \frac{470}{2201} \approx 0.213$$

- $P(F|S)$ = Probability that a survivor is female.

$$P(F \mid S) = \frac{344}{711} \approx 0.483$$

- $P(\text{not } F)$ = Probability of a male passenger

$$P(\text{not } F) = \frac{1731}{2201} \approx 0.786$$

Discussion P23. (p. 334)

- For the Titanic data, let S be the event a person survived and F be the event a person was female. Find and interpret these probabilities.
  - a. $P(F)$ and $P(F|S)$
  - b. $P(\text{not } F)$, $P(\text{not } F|S)$, and $P(S|\text{not } F)$

Gender

		Male	Female	Total
Survived?	**Yes**	367	344	711
	No	1364	126	1490
	Total	1731	470	2201

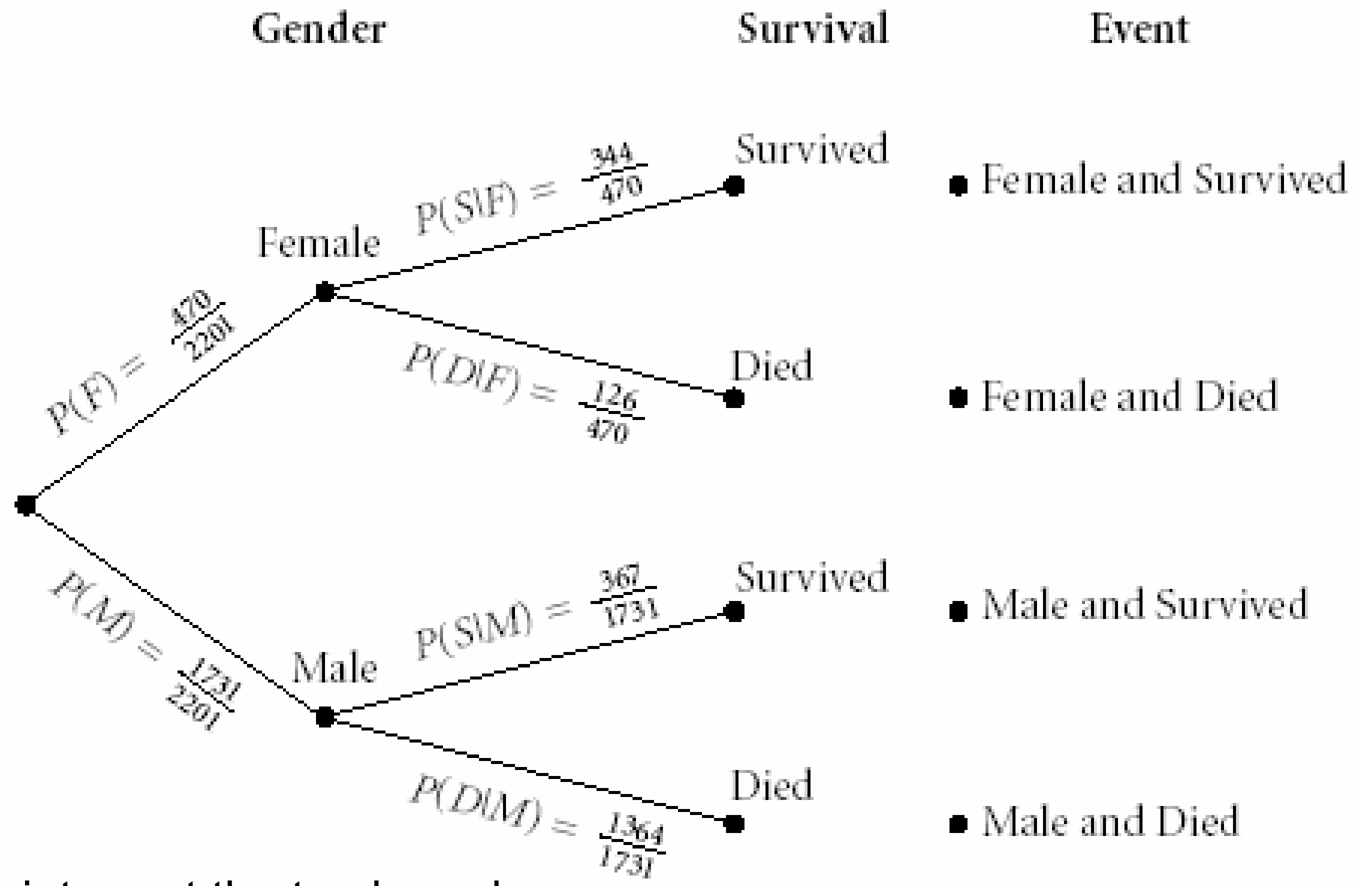
- $P(\text{not } F|S)$ = Probability that a survivor is male.

$$P(\text{not } F \mid S) = \frac{367}{711} \approx 0.516$$

- $P(S \mid \text{not } F)$ = Probability of surviving given that the person selected is male.

$$P(S \mid \text{not } F) = \frac{367}{1731} \approx 0.212$$

The Multiplication Rule for P (A and B) (p. 327)



We can interpret the top branch as:

$$P(F)\cdot P(S\mid F)=\frac{470}{2201}\cdot\frac{344}{470}=\frac{344}{2201}=P(F\text{ and }S)$$

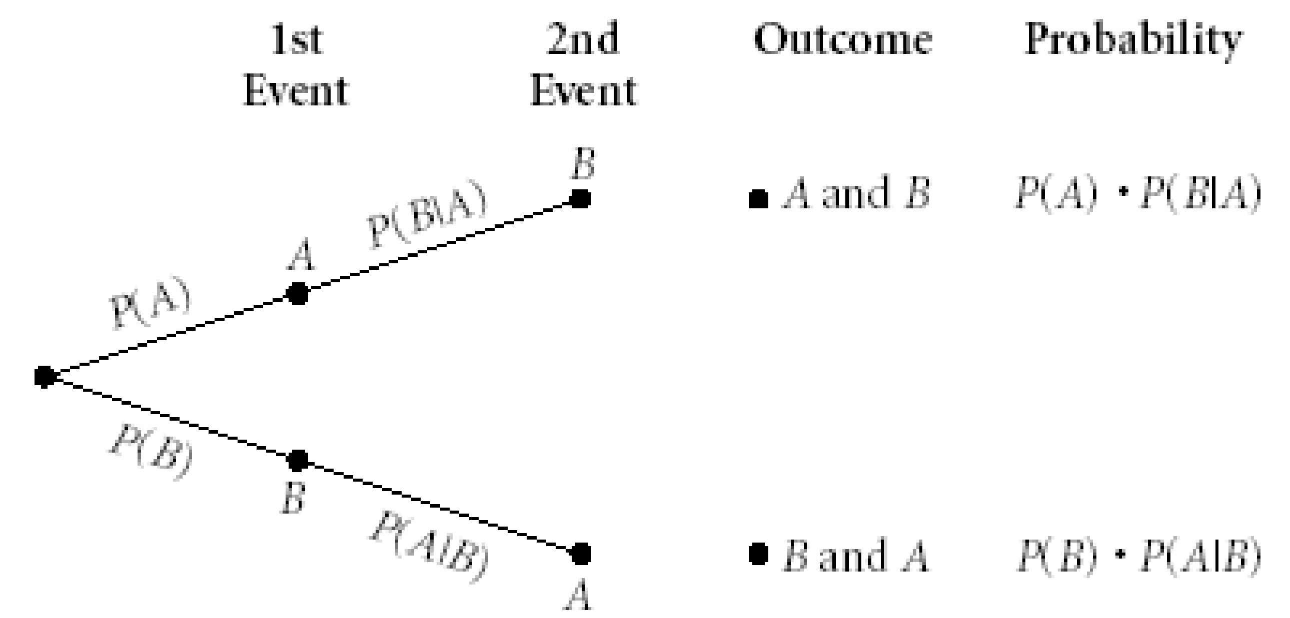
The Multiplication Rule

- The probability that event A and event B both happen is given by

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

  or alternatively

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



Definition of Conditional Probability

- For any two events A and B, $P(B)>0$,

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

- Example: Suppose you roll two dice. Use the definition of conditional probability to find the probability that you get a sum of 8 given that you rolled doubles.

$$P(\text{sum8} \mid \text{doubles}) = \frac{P(\text{sum8 and doubles})}{P(\text{doubles})}$$

$$= \frac{1/36}{6/36} = \frac{1}{6}$$

Important Uses of Conditional Probability

- To compare sampling with or without replacement.
- To study effectiveness of medical tests
- To study effectiveness of statistical testing

Conditional Probability and Medical Tests

- In medicine, ***screening tests*** give a quick indication of whether or not a person is likely to have a particular disease. Because screening tests are intended to be relatively quick and noninvasive, they are often not as accurate as other tests that take longer or are more invasive.
- A two-way table is often used to show the four possible outcomes of a screening test.

		Test Result		
		Positive	Negative	Total
Disease	Present	a	b	$a + b$
	Absent	c	d	$c + d$
	Total	$a + c$	$b + d$	$a + b + c + d$

Conditional Probability and Medical Tests

Test Result

		Positive	Negative	Total
Disease	Present	a	b	$a + b$
	Absent	c	d	$c + d$
	Total	$a + c$	$b + d$	$a + b + c + d$

- **False Positive Rate** = P(no disease | test positive) = $\frac{c}{a+c}$
- **False Negative Rate** = P(disease present | test negative) = $\frac{b}{b+d}$
- **Sensitivity** = P(test positive | disease present) = $\frac{a}{a+b}$
- **Specificity** = P(test negative | no disease) = $\frac{d}{c+d}$

Example of a Rare Disease on 10,000 patients

Test Result

Disease

	Positive	Negative	Total
Present	9	1	10
Absent	50	9,940	9,990
Total	59	9,941	10,000

- **False Pos. Rate** = P(no disease | test positive) = $\frac{50}{59} \approx 0.8474$
- **False Neg. Rate** = P(disease present | test negative) = $\frac{1}{9941} \approx 0.0001$
- **Sensitivity** = P(test positive | disease present) = $\frac{9}{10} = 0.90$
- **Specificity** = P(test negative | no disease) = $\frac{9940}{9990} \approx .9949$

Example P34 (p. 335)

- A laboratory technician is being tested on her ability to detect contaminated blood samples. Among 100 samples given to her, 20 are contaminated, each with about the same degree of contamination. Suppose the technician makes the correct decision 90% of the time **(regardless of contamination or not)**. Make a table showing what you would expect to happen. What is her false positive rate? What is her false negative rate? How would these rates change if she were given 100 samples with 50 contaminated?

Detection of Contamination (test)

Contaminated?		Positive	Negative	Total
	Yes			
	No			
	Total			100

Example P34 (p. 335)

Detection of Contamination (test)

Contaminated?

	Positive	Negative	Total
Yes	18	2	20
No	8	72	80
Total	26	74	100

- False Pos. Rate = 8/26 = 0.3076
- False Neg. Rate = 2/74 = 0.027
- Sensitivity = 18/20 = 0.90
- Specificity = 72/80 = 0.9

Conditional Probability and Statistical Inference

- ***Statistician:*** Suppose you draw 3 workers at random from the set of 10 hourly workers. This establishes random sampling as the model for the study.
- ***Lawyer:*** Okay.
- ***Statistician:*** It turns out that there are 120, possible samples of size 3, and only 6 of them give an average age of 58 or more.
- ***Lawyer:*** So the probability is 6/120 , or .05.
- ***Statistician:*** Right.
- ***Lawyer:*** There's only a 5% chance the company didn't discriminate and a 95% chance that they did.
- ***Statistician:*** No, that's not true.
- ***Lawyer:*** But you said . . .
- ***Statistician:*** I said that *if* the age-neutral model of random draws is correct, then there's only a 5% chance of getting an average age of 58 or more.
- ***Lawyer:*** So the chance the company is guilty must be 95%.
- ***Statistician:*** Slow down. If you start by assuming the model is true, you can compute the chances of various results. But you're trying to start from the results and compute the chance that the model is right or wrong. You can't do that.

$$P(\text{av. age} \geq 58 \mid \text{random draws}) = .05$$

$$P(\text{no discr.} \mid \text{av. age} = 58)\,??$$

5.5 Independent Events

- Events A and B are independent if and only if $P(A \mid B) = P(A)$.

  Equivalently, A and B are independent if and only if $P(B \mid A) = P(B)$.

- In other words, knowing that B happened does not affect the probability of A happening, and conversely knowing that A happened does not affect the probability of B.

Example: Water, Gender, and Independence (p.340)

Drinks Bottled Water ? / Identified Tap Water ?

	Yes	No	Total
Yes	24	6	30
No	36	34	70
Total	60	40	100

Gender / Identified Tap Water ?

	Yes	No	Total
Male	21	14	35
Female	39	26	65
Total	60	40	100

- Show that the events ***is a male*** and ***correctly identifies tap water*** are independent and that the events ***drinks bottled water*** and ***correctly identifies tap water*** are not independent.