Math 130 – Jeff Stratton Name _____Solutions________

Lab – Simulation and Probability

Goal: The goal of this lab is two-fold. The first goal is to gain experience using a computer to do a simulation study and to look at a sample design for a survey. The second goal is to gain experience with different probability problems and methods for solving them.

Some rules that may be useful:

This worksheet reviews the four probability calculation rules of Chapter 7. The four rules are:

Complement Rule

$P(A) = 1 - P(A^C)$

General Additive Rule

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

If A and B are mutually exclusive, $P(A \text{ or } B) = P(A) + P(B)$

General Multiplication Rule

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

If A and B are independent, $P(A \text{ and } B) = P(A)P(B)$. This also gives us a way to test if two events are independent.

Simulation

Goal: To gain experience with a simulation, and to review and discuss the sample design of a survey.

Part 1 – The Birthday Paradox

We'll start by using simulation to study a famous statistical example known as the "Birthday Paradox." Suppose you are in a room with 24 other people. What are the chances that at least two people in the room have the same birthday? By birthday here I mean just the same day of the year, and am not worrying about the actual year of birth.

1. What would you guess this probability to be? No need for any calculation or anything, just take a stab at it.

The probability is difficult to calculate mathematically, but we can use simulation to investigate this. We'll begin with the following information:

- There are 365 days in a year. (We will ignore leap years).

- There are 25 people in the room. (You and 24 others).

- Each person could have any day of the year as his or her birthday, so each birthday is equally likely.

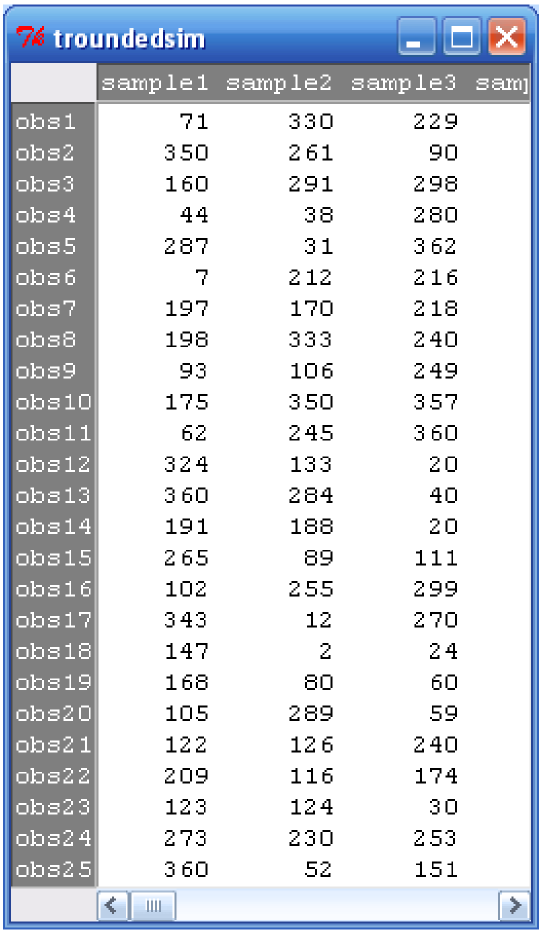
A trial in this case consists of looking at a class of 25 people, generating a birthday for each of them, and seeing if there are any in common.

Start your simulation with 1 room of 25 people:

Click on: **Distributions → Continuous Distributions → Uniform Distribution → Sample from uniform distribution**

- Set the minimum to 1 and the maximum to 365.
- We'll simulate 30 classrooms. Set the number of samples (rows) as 30 and set the number of observations (columns) to 25.
- Uncheck the "Sample Means" box.
- Click OK
- Go back to the R commander window and click the "View data set" button

To see if there is overlap, we can go through the row of data and look for any days that are the same. This is still tedious, though. We can get R to do this work for us.



	sample1	sample2	sample3	sam
obs1	71	330	229	
obs2	350	261	90	
obs3	160	291	298	
obs4	44	38	280	
obs5	287	31	362	
obs6	7	212	216	
obs7	197	170	218	
obs8	198	333	240	
obs9	93	106	249	
obs10	175	350	357	
obs11	62	245	360	
obs12	324	133	20	
obs13	360	284	40	
obs14	191	188	20	
obs15	265	89	111	
obs16	102	255	299	
obs17	343	12	270	
obs18	147	2	24	
obs19	168	80	60	
obs20	105	289	59	
obs21	122	126	240	
obs22	209	116	174	
obs23	123	124	30	
obs24	273	230	253	
obs25	360	52	151	

First, we will have R round off the generated data. Paste the following code into the script window.

roundedsim <- round(*datafilename*,0)

Put the name of your generated data file in where I've put ***datafilename***. Click on "Submit." We've created a new dataset **roundedsim** with the rounded values.

Now if you click on the "Dataset" window in R Commander you can change the data file to your new data file. Click on "View data set" to see your rounded data values.

It's still going to be tedious to look through this by hand to see if we have any matches. We can get R to summarize this for us. However, we need to manipulate the data a little bit. The problem is that R thinks in terms of variables as column vectors. The simulated data is terms of a row for each sample rather than a column. We need to transpose our data and save it as a data frame.

Paste the following code into the script window.
troundedsim <- as.data.frame(t(roundedsim))

Now, you have a data set of 30 classrooms (as columns). We want a frequency summary of the birth days. To do that, we need to convert each sample to a factor.

Go to **Data → Manage variables in active data set → Convert numeric variables to factors**

Select all 30 columns and click OK. You'll get a bunch of annoying pop-up windows. You have to click on OK for each column.

Now, go to **Statistics → Summaries → Frequency distributions** and select all 30 samples. For each sample, you'll get something like this:

```
> .Table <- table(Newdata2$sample1)

> .Table  # counts for sample1

  7  44  62  71  93 102 105 122 123 147 160 168 175 191 197 198 209 265 273 287 
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1 
324 343 350 360 
  1   1   1   2 
```

```
> round(100*.Table/sum(.Table), 2)  # percentages for sample1

  7  44  62  71  93 102 105 122 123 147 160 168 175 191 197 198 209 265 273 287
  4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4
324 343 350 360
  4   4   4   8
```

There's a listing for both frequencies and percentages. Check if any day was tallied multiple times. In this example, two students had a birthday on the 360th day. Repeat this for all 30 days you simulated, and calculate the percentage of classes where at least one day occurred multiple times.

2. What percentage of your simulated classes had at least two people with the same birthday? How does this compare to your response to Q1? Are you surprised?

   Answers will vary here.

3. How close is your experimental probability to the theoretical probability, which is 0.592?

   Again, answers will vary here.

Note: The true probability

Note that "at least" means greater than or equal to 1. We can also calculate this as 1 minus the probability that no people match out of the 25.

Think of a sequence of 25 students. Student 1 can have any of the 365 days for their birthday. To avoid a match, though, Student 2 has the choice of only 364 days out of 365. Student 3 can choose from 363/365, and so on.

$$1-\left(\frac{364}{365}\cdot\frac{363}{365}\cdot\frac{362}{365}\cdots\frac{340}{365}\right)$$

Part 2 – Additional Discussion on Simulations

For this part of the lab, discuss why each of the following simulations fails to model the real situation properly (Exercises 11.11, 11.12 on page 300).

4. Use a random integer from 0 through 9 to represent the number of heads that appear when 9 coins are tossed.

   This strategy treats all sequences of heads and tails as equally likely, but a sequence with, say, 5 heads is more likely than one with 9 heads because there are many more ways to get a sequence with 5 heads.

5. A basketball player takes a foul shot. Look at a random digit, using and odd digit to represent a good shot and an even digit to represent a miss.

   This strategy assumes that the basketball player has a 50-50 chance of making the foul shot, and it may be different from that. Most players shoot better than that.

6. Use five random digits from 1 through 13 to represent the denominations of the cards in a poker hand.

   This strategy only deals with the card values, but doesn't account for the different suits in a card deck.

7. Use random numbers 2 through 12 to represent the sum of the faces when two dice are rolled.

   This strategy has the same problem as in question 4. In this situation, the outcomes of 2 to 12 are not equally likely. You are more likely to get a sum of 7 than any other outcome.

8. Use a random integer from 0 through 5 to represent the number of boys in a family of 5 children.

   This strategy has the same problem as in question 4. In this situation, the outcomes of, say, 2 boys are more probable than 0 boys. They are not equally likely.

9. Simulate a baseball player's performance at bat by letting 0 = an out, 1 = a single, 2 = a double, 3 = a triple, and 4 = home run.

   Again, this strategy treats all the events as equally likely, and they are not. An out is more likely than the other outcomes.

Part 3 – A Sample Design

In this part, we'll investigate the sample design of a survey. What follows here is a description of the National Survey of College Graduates, a survey conducted by the National Science Foundation. Please read through it and answer the questions below.

Source: http://www.nsf.gov/statistics/srvygrads/#design

10. What is the target population for this survey?

The target population of the 2003 survey consisted of all individuals:

- Under age 76 as of the survey reference date (i.e. born after September 30, 1927)
- Who received a bachelor's degree or higher prior to April 1, 2000
- Who were living in a housing unit or noninstitutionalized group quarters as of April 1, 2000, and
- Who resided in the 50 states, the District of Columbia, Puerto Rico, or the other outlying U.S. territories as of April 2000 and the survey reference week of October 1, 2003.

11. Where did NSF get the list of people to sample from? This master list of all the people available is the *sampling frame*.

The list came from several places. The 2000 Decennial Census long form respondents, the 1999 NSCG, and the 2001 NSRCG were used.

12. How many people were sampled? Were there any problems/changes that had to be made?

The initial sample size was set at 201,220. Problems with the data included

- blank names and addresses were removed
- interviewing due to cost reasons
- all cases with imputed educational attainment data on the Census were removed

The final 2003 survey sample size was 170,797.

13. What kind of sampling was done for this survey?

The NSCG sample was stratified by four sampling variables: demographic group, highest degree type, occupation, and sex.

14. How was the data collected? Why do you think the NSF used a combination of different methods?

Several methods were used: mail-in questionnaires, phone interviews, and personal interviews.

1. Overview (2003 survey cycle)

a. Purpose

The National Survey of College Graduates (NSCG) is a longitudinal survey, designed to provide data on the number and characteristics of experienced individuals with education and/or employment in science and engineering (S&E), or S&E-related fields in the United States. The 2003 NSCG provides a once in a decade opportunity to study the education and career paths of the nation's college-educated individuals and various characteristics of college-educated individuals in the workforce such as salaries, whether the college-educated population was working in their highest degree field of study, specific occupations, and a gender breakdown of the workforce.

The results of this survey are vital for educational planners within the federal government and in academia. Employers in all sectors (education, industry, and government) use the survey results to understand trends in employment opportunities and salaries in various degree fields and to evaluate the effectiveness of equal opportunity efforts. This survey is designed to complement the other surveys of scientists and engineers conducted by SRS in order to provide a comprehensive picture of the number and characteristics of individuals with education and/or employment in S&E or S&E-related fields in the United States. This combined system is known as SESTAT (Scientists and Engineers Statistical Data System).

Data from the 2003 NSCG are used to update the 1993 NSCG findings.

b. Respondents

2003 survey respondents were individuals living in the U.S. during the reference week of October 1st, 2003, holding a bachelor's or higher degree in any field, and under age 76.

c. Key variables

- Academic employment (positions, rank and tenure)
- Age
- Citizenship status
- Country of birth
- Country of citizenship
- Disability status
- Educational history (for each degree held: field, level, when received)
- Employment status (unemployed, employed full time, or employed part time)
- Geographic place of employment
- Immigrant module (year of entry, type of entry visa, reason(s) for coming to U.S., etc.)
- Labor force status
- Marital status
- Number of children
- Occupation (current or past job)
- Primary work activity (e.g., teaching, basic research, etc.)
- Publication and patent activities
- Race/ethnicity
- Salary
- Satisfaction and importance of various aspects of job
- School enrollment status
- Sector of employment (academia, industry, government)
- Sex
- Work-related training

2. Survey Design

a. Target population and sample frame

The target population of the 2003 survey consisted of all individuals:

- Under age 76 as of the survey reference date (i.e. born after September 30, 1927)
- Who received a bachelor's degree or higher prior to April 1, 2000
- Who were living in a housing unit or noninstitutionalized group quarters as of April 1, 2000, and
- Who resided in the 50 states, the District of Columbia, Puerto Rico, or the other outlying U.S. territories as of April 2000 and the survey reference week of October 1, 2003.

The 2003 NSCG serves as the baseline survey for future survey cycles in the current decade, much as the 1993 NSCG did. The 2003 survey included a sample of respondents to the 2000 Decennial Census long form who indicated they had a baccalaureate degree or higher in any field of study. However, those holding a Ph.D. earned in the U.S. in an S&E field will not be followed in the future NSCG survey cycles as these individuals are covered in a companion SESTAT survey—the Survey of Doctorate Recipients (SDR).

2003 NSCG also included two different samples of "old cohort" panel cases from earlier surveys. Cases in the "old cohort" sampling frame were either drawn from the 1999 NSCG respondents who originated from the 1993 NSCG (based on the 1990 Decennial Census long form) or panel respondents from the 1993-2001 National Survey of Recent College Graduates (NSRCG). A sample of respondents to the 1999 NSCG (1993 NSCG cases and 1993-1999 NSRCG cases) were included in the 2003 NSCG to allow methodological research to compare the quality and reliability of estimates based on continuation of the cases from the 1990s with those based on a new frame from the 2000 census. A sample of respondents from the 2001 NSRCG were included in the 2003 NSCG to cover new S&E degree recipients between April 15, 2000 and June 30, 2000 to provide complete coverage of the S&E degree holders in the 2003 SESTAT integrated database. The 2003 NSCG file does not include any of the "old cohort" cases and these cases are omitted from the discussions below **[1]**.

b. Sample design

The sample frame for the 2003 survey was drawn from the 2000 Decennial Census long form responses to determine initial eligibility status and stratification cells for the sampling frame. Cases were eligible if they were living in a housing unit or non-institutionalized group quarters; had received a bachelor's degree or higher; resided in the 50 states, the District of Columbia, Puerto Rico, or the other outlying U.S. territories; all as of April 2000; and were age 75 or less as of the 2003 survey reference date of October 1st (i.e., born after September 30, 1927).

The goals of the 2003 NSCG sample design were:

- Develop a design that was comparable to the sample design used in the other two SESTAT surveys;
- Address the sampling inefficiencies that existed in the 1990s NSCG design based on the 1990 Census long form responses to the extent possible;
- Use caution when developing sampling rates in an effort to reduce the sampling rate range across all strata; and
- Allocate the sample to provide complete coverage of science and engineering (S&E) and non-S&E degree holders.

The NSCG sample was stratified by four sampling variables:

- Demographic Group (8 levels) - a composite variable that captured disability status, ethnicity, race, citizenship status at birth, and likelihood of a U.S.-earned degree.
- Highest Degree Type (3 levels) - Bachelor's, Master's, Professional degree, or Doctoral degree. A small number of professional degree cases were grouped with the bachelor's degree cases for stratification purposes.
- Occupation (30 levels) - the other two SESTAT surveys (NSRCG and SDR) categorize the sample degree into one of the five specific S&E fields (Life Sciences, Mathematics/Computer Sciences, Physical Sciences, Social Sciences, and Engineering). However, because the 2000 Decennial Census long form did not collect degree field information, Decennial Census reported occupation was used as a proxy variable for degree field.
- Sex (2 levels) - male and female.

The initial sample size of the 2003 NSCG was set at 201,220. After the cases with blank names and addresses were removed, 197,834 cases remained in the sample. The final sample size was reduced to 177,320 for interviewing due to cost reasons. After data collection, analysis showed that a disproportionate number of cases requiring personal follow-up who had been imputed as college degree holders on the 2000 Census data actually did not have at least a bachelor's degree. To address concerns about coverage bias related to the imputed degree cases being identified as ineligible (i.e., no bachelor's degree), all cases with imputed educational attainment data on the Census were removed from the sample to increase the probability that all the necessary sampling criteria were satisfied. There were 6,523 of these "imputed degree" cases in the sample resulting in a final 2003 survey sample size of 170,797.

For the old cohort, the initial sample size was set at 40,073.

c. Data collection techniques

The Bureau of the Census conducted the NSCG for NSF. Initial data collection was done through the use of a self-administered mail survey using a prenotification letter, a first mailing, a reminder letter, and a second mailing. If the sample person did not respond to any of the various mailings, automated reminder phone calls were made using a software package called Phone Tree.

Nonrespondents to the mail questionnaire were followed up using computer-assisted telephone interviewing. If the paper questionnaire was not returned and telephone follow-up attempts failed, for certain selected cases, a personal visit (CAPI) was conducted.

Information in the 2003 survey was collected for the week of October 1, 2003. Data collection took place between October 2003 and August 2004.

Probability

Part 4 – Basic Probability

15. Describe the sample space for the following situations:

a.) Ask whether a student did or did not take a math class in each of the 2 previous years.

Let Y be the event that the student took a math class in one year.
S = {YY, YN, NY, NN}

b.) Record a student's grade at the end of the course.

S = {A, B, C, D, F}

c.) Ask how much time the student spent studying during the past 24 hours.

S = {any time between 0 and 24 hours}

d.) A basketball player shoots 3 free throws. You record the number of baskets she makes.

S = {0, 1, 2, 3}

e.) A basketball player shoots 3 free throws. You record the sequence of hits and misses.

Let H be the event that the player makes a single shot.
Let m be the event that the player misses the shot.
S = {HHH, HHM, HMH, MHH, HMM, MHM, MMH, MMM}

16. Car colors: Choose a new car or light truck at random and note its color. Here are the probabilities of the most popular colors for vehicles made in North America in 2001:

Color	Silver	White	Black	Blue	Red	Green
Probability	.210	.156	.112	.112	.099	.076

a.) What is the probability that the vehicle you choose has any color other than the 6 listed?

P(not listed) = 1 – P(listed)

= 1 – (0.210 + 0.156 + 0.112 + 0.112 + 0.099 + 0.076)

= 1 – 0.765

= 0.235

There is a 23.5% change that the vehicle chosen is one other than those listed.

b.) What is the probability that a randomly chosen vehicle is either silver or white?

Because the events are disjoint,
P(silver or white) = P(silver) + P(white) = 0.210 + 0.156 = 0.366

There's a 36.6% probability that a randomly chosen vehicle is silver or white.

EXAMPLE SCENARIO: As discussed in class, consider rolling two dice and recording the numbers on the faces showing. The sample space of such a random circumstance is given below.

Rolling Two Dice Sample Space

		Second Die					
		1	2	3	4	5	6
First Die	1	(1 , 1)	(1 , 2)	(1 , 3)	(1 , 4)	(1 , 5)	(1 , 6)
	2	(2 , 1)	(2 , 2)	(2 , 3)	(2 , 4)	(2 , 5)	(2 , 6)
	3	(3 , 1)	(3 , 2)	(3 , 3)	(3 , 4)	(3 , 5)	(3 , 6)
	4	(4 , 1)	(4 , 2)	(4 , 3)	(4 , 4)	(4 , 5)	(4 , 6)
	5	(5 , 1)	(5 , 2)	(5 , 3)	(5 , 4)	(5 , 5)	(5 , 6)
	6	(6 , 1)	(6 , 2)	(6 , 3)	(6 , 4)	(6 , 5)	(6 , 6)

Let A = the event that the sum of the two dice is equal to 7.
Let B = the event that one of the dice shows a 3.

17. Use the probability rules and the sample space to calculate the following probabilities:

a. Find the probability that the sum of the two dice is equal to 7. i.e., P(A)
There are a total of 36 outcomes, and 6 of them give a sum of 7.

$$P(A) = \frac{6}{36} = \frac{1}{6} = 0.1667$$

b. Find the probability that one of the dice shows a 3.
In class I told some of you to ignore the case where both show a 3. I'm going to include it in these solutions. The question wording is unclear.

$$P(\text{one dice is a 3}) = \frac{11}{36} = 0.3056$$

c. Find the probability that the sum of the two dice is not equal to 7. i.e., $P(A^C)$

$$P(A^C) = 1 - P(A) = \frac{30}{36} = \frac{5}{6} = 0.8333$$

d. Calculate the probability that the sum of the two dice is equal to 7 or that one of the dice shows a 3. i.e., P(A or B).

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = \frac{6}{36} + \frac{11}{} - \frac{2}{} = \frac{15}{} = 0.4167$$

Part 5 – Probability and the Lottery

In this section, we'll explore calculating probabilities using the popular Mega Millions Lottery game.

How to play Mega Millions

Pick 5 different numbered balls between **1 and 56**, and a megaball number between **1 and 46** for each $1 play. The megaball number can be the same as one of the first five numbers on your play line.

Formulas we'll need

$$\text{Probability of Event A} = P(A) = \frac{\text{Number of ways to get event A}}{\text{Total number of possibilities}}$$

$$\text{Number of Combinations of } N \text{ items taken } n \text{ at a time} = \binom{N}{n} = \frac{N!}{n!\,(N-n)!}$$

The $N!$ does not mean we are excited about the number N. It is shorthand:

$$N! = N(N-1)(N-2)\cdots 2 \cdot 1$$

For example, $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$

Hitting the Jackpot

Find the probability of hitting the jackpot. Let A be the event that we hit the jackpot. What does this mean? This means that we matched all 5 numbers drawn, and also matched the megaball.

18. How many ways can we match the winning five numbers and the powerball? ____1____

Now, how many total possibilities are there? Let's think of the steps involved:

- Draw 5 white balls out of 56. One example would be the balls 1-15-34-11-46. Another would be 2-1-31-19-17. Obviously, there are way too many possibilities to write down. But, we can use the combinations formula to find the number of ways to pick 5 balls out of 56.

- Next, we draw a single megaball from the balls numbered 1 to 46. So, there are 46 possible choices of megaball.

We can use R to compute this. There's no function native to R that will compute the number of combinations, but we can write one. To do this, we'll open R without R commander.

- Open R
- In the "R Console" window, paste the following onto the line with the ">" symbol and press ENTER.
```
combinations <- function(N,n){
  outnum <- factorial(N)/(factorial(n)*factorial(N-n))
  outnum
  }
```
- Now, to compute the number of combinations of 5 balls out of 56, enter
```
> combinations(56,5)
```

19. Calculate the number of ways to get 5 balls out of 56. _3,819,816__

20. Take 46 × your answer to number 5 to get the total number of possibilities:
_175,711,536_

Note: You can use R like a calculator simply by typing expressions at the ">" prompt.
> 3*5

21. So, the probability of hitting the jackpot is: $\frac{1}{175{,}711{,}536} \approx 0$

Now you can see why lotteries are sometimes called a tax on people who don't know statistics. ☺

Hitting the $10,000 prize Finding this probability is a bit trickier. We already know the total number of possibilities, though, so we just need to worry about finding the number of ways to win the $10,000 prize. Let B be the event that we won the $10,000 prize.

What does this mean? It means we matched four of the 5 winning numbers, and the megaball.

How many ways can we match 4 numbers out of 5? We can use combinations again. To win, we need choose any 4 of the 5 winning numbers, or find $\binom{5}{4}$. Now, there are 51 numbers left over, and we are picking 1 of them, or $\binom{51}{1}$. So the total number of winning white ball combinations is $\binom{5}{4} \times \binom{51}{1}$. For the megaball, we need to pick the 1 winning ball. We multiply these together to get the total number of winning $10,000 prize tickets possible.

22. The total number of winning $10,000 prize tickets is ___255__________

23. The probability of winning the $10,000 prize is $\frac{255}{175{,}711{,}536} \approx \frac{1}{689{,}065}$

See if you can verify the probabilities for some of the other prizes in the table below:

Match	+	Megaball	Probability	Top Prize
5	+	1	1 in 175,711,536	JACKPOT
5			1 in 3,904,701	$250,000**
4	+	1	1 in 689,065	$10,000**
4			1 in 15,313	$150**
3	+	1	1 in 13,781	$150**
2	+	1	1 in 844	$10
3			1 in 306	$7
1	+	1	1 in 141	$3
0	+	1	1 in 75	$2

* Overall odds are 1 in 39.89.
* Prizes based on a $1 wager.
* The jackpot will be divided evenly among winners.

Part 6 – Conditional Probability and Independence

Conditional probability is the idea that the probability of an event can change if we are given additional information. If knowing that one event has happened changes the probability of another event, they are *dependent*.

The keyword phrase for conditional probability is "given that" or "if it is known that."

A good example of *dependent* events is drawing cards without replacement. Suppose you are interested in drawing two aces in a row. On the first draw, the probability of an ace is $\frac{4}{52}$. However, the probability of an ace on the second draw depends on what happened on the first draw. If the first card was an ace, then the probability of an ace on the second draw is $\frac{3}{51}$. Otherwise, the probability is $\frac{4}{51}$.

Conditional probability is one of the most difficult concepts in this course. Intuition doesn't work quite so well here. I find that the best way to compute conditional probability is to use the computing formula or the idea of a restricted sample space.

<u>Conditional Probability</u>

- Formula $P(A\,|\,B) = \dfrac{P(A \text{ and } B)}{P(B)}$

- Another approach is to think of the restricted sample space. Use the given information to eliminate some possibilities from the sample space, and recomputed the probability.

24. For the dice example earlier in the lab, calculate the probability that the sum of the two dice is equal to 7, given that one of the dice shows a 3. i.e., P(A | B).

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{\frac{2}{36}}{\frac{11}{36}} = \frac{2}{11} = 0.1818$$

The probability that the sum of the two dice is 7 given that one of the dice shows a 3 is 0.1818.

25. Calculate the probability that one of the dice shows a 3, given that the sum of the two dice is equal to 7. i.e., P(B | A).

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{\frac{2}{36}}{\frac{6}{36}} = \frac{2}{6} = \frac{1}{3} = 0.3333$$

The probability that one of the dice shows a 3 given that the sum of the two dice is 7 is 0.3333.

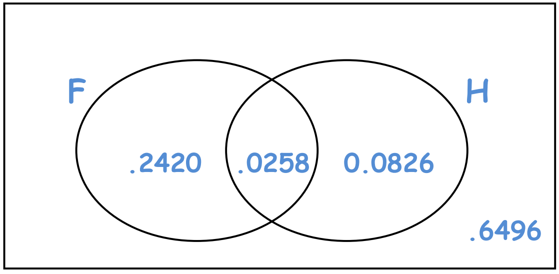
26. There were 1,744 students at Amherst as of October 15, 2009. We find that 467 of them were first-time degree-seeking students, 189 of them were Hispanic, and 45 were both first-time students and Hispanic.

Define the following events:

F: Student is a first-time degree-seeking student

H: Student is Hispanic.

a. Use a Venn diagram to represent this information.



If we randomly select one student, find the following probabilities:

b. P(F or H)

$$\boldsymbol{P(F \text{ or } H) = P(F) + P(H) - P(F \text{ and } H) = 0.2678 + 0.1084 - 0.0258 = 0.3504}$$

c. P(F^{C})

$$\boldsymbol{P(F^{C}) = 1 - P(F) = 1 - 0.2678 = 0.7322}$$

d. P(H^{C})

$$\boldsymbol{P(H^{C}) = 1 - P(H) = 1 - 0.1084 = 0.8916}$$

e. P(F and H)

$$\boldsymbol{P(F \text{ and } H) = 0.0258}$$

f. P(F | H)

$$\boldsymbol{P(F|H) = \frac{P(F \text{ and } H)}{P(H)} = \frac{0.0258}{0.1084} = 0.2380}$$

g. P(H | F)

$$\boldsymbol{P(H|F) = \frac{P(H \text{ and } F)}{P(F)} = \frac{0.0258}{0.2678} = 0.0963}$$

27. Data was collected from a sample of Connecticut residents and is displayed in the following two-way table:

		Income			
		<$25k	$25k – $70k	> 70k	Total
Age (years)	< 25	952	1,050	53	2,055
	25 – 45	456	2,055	1,570	4,081
	> 45	54	952	1,008	2,014
	Total	1,462	4,057	2,631	8,150

Select one person at random from this study. Calculate the probability:

a. P(Age < 25) $= \frac{2055}{8150} = 0.2521$

b. P(Age < 25 and Income > $70,000) $= \frac{53}{8150} = 0.0065$

c. P(Income < $70,000) $= \frac{1462+4057}{8150} = 0.6772$

d. P(Age 25 – 45 or Income < $25,000) $= \frac{4081}{8150} + \frac{1462}{8150} - \frac{456}{8150} = 0.6242$

e. P(Age < 25 given Income > $70,000)

Let A = Event age is <25
Let I = Event that income is > 70k

$$P(A|I) = \frac{P(A \text{ and } I)}{P(I)} + \frac{0.0065}{0.3228} = 0.0201$$

f. P(Age 25 – 45 given Income < $25,000)

$$P(\text{Age 25-45}|\text{Income} < 25\text{k}) = \frac{\text{P(Age 25-45 and Income} < 25\text{k)}}{\text{P(Income} < 25\text{k)}} = \frac{\frac{456}{8150}}{\frac{1462}{8150}} = 0.3119$$

g. P(Income > $70,000 given Age > 45)

$$P(\text{Income} > 70\text{k}|\text{Age} > 45) = \frac{\text{P(Income} > 70\text{k and Age} > 45)}{\text{P(Age} > 45)} = \frac{\frac{1008}{8150}}{\frac{2014}{8150}} = 0.5005$$

Part 7: Tree Diagrams

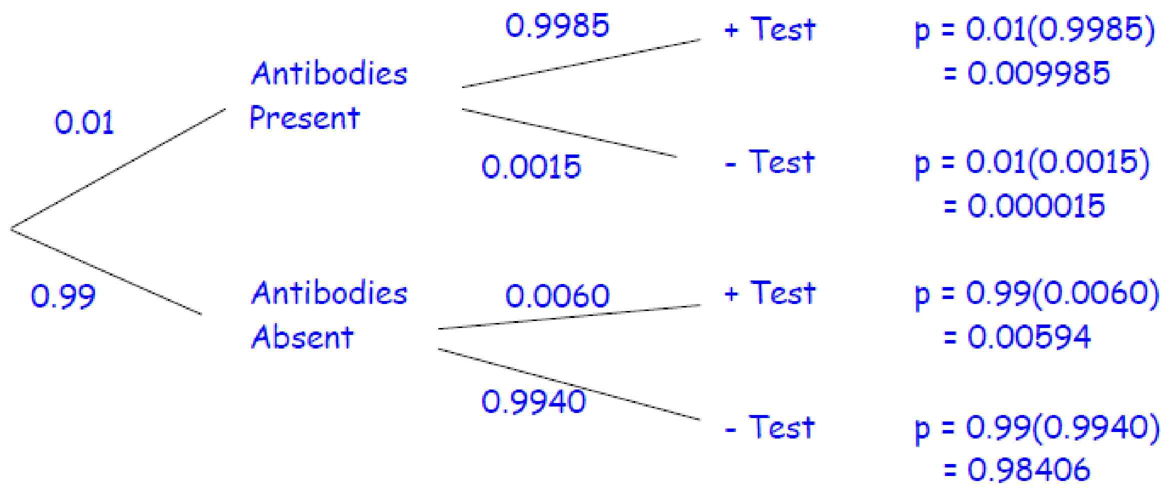
Testing for HIV: Enzyme immunoassay (EIA) tests are used to screen blood specimens for the presence of antibodies to HIV, the virus that causes AIDS. Antibodies indicate the presence of the virus. The test is quite accurate but not always correct. Here are approximate probabilities of positive and negative EIA outcomes when the blood tested does and does not actually contain antibodies to HIV:

Test Result:

	+	-
Antibodies Present	0.9985	0.0015
Antibodies absent	0.0060	0.9940

Suppose that 1% of a large population carries antibodies to HIV in their blood.

28. Draw a tree diagram for selecting person from the population and testing his or her blood.


0.01
Antibodies Present
0.9985
+ Test
p = 0.01(0.9985) = 0.009985
0.0015
- Test
p = 0.01(0.0015) = 0.000015
0.99
Antibodies Absent
0.0060
+ Test
p = 0.99(0.0060) = 0.00594
0.9940
- Test
p = 0.99(0.9940) = 0.98406

29. What is the probability that the EIA test is positive for a randomly chosen person from this population?

We can have a positive test in two ways:

Antibodies present – Test +
Antibodies not present – Test –

We can add up the probabilities: 0.009985 + 0.00594 = 0.015925

30. What is the probability that a person has the antibody, given that the EIA test is positive?

$$P(\text{Antibody} \mid \text{Test } +) = \frac{P(\text{Antibody and Test } +)}{P(\text{Test } +)} = \frac{0.009985}{0.015925} = 0.6270$$

Note: This illustrates a fact that is important when considering proposals for widespread testing for HIV, illegal drugs, or agents of biological warfare: if the condition being tested is uncommon in the population, many positives will be false positives.