Lecture Notes & Examples 5.1

Section 5.1 - Randomness, Probability & Simulation (pp. 287-299)

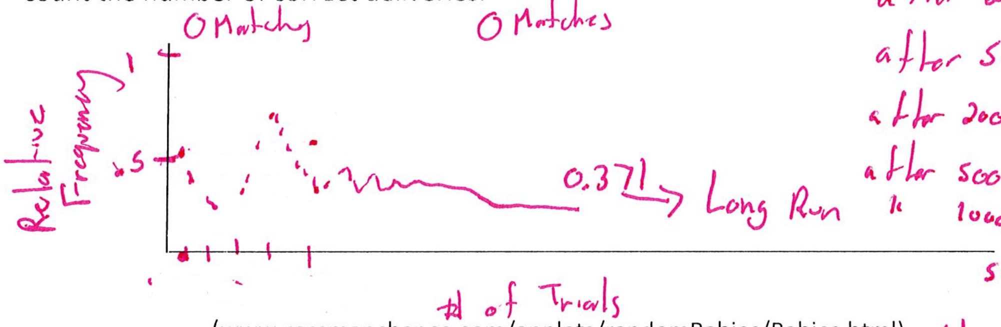
1. The Idea of Probability

- Random samples and randomized experiments
- Avoid bias by allowing chance to decide what individuals get selected.
- Chance behavior is *unpredictable in the* ***short run*** *but has a regular and predictable pattern in the* ***long run***.
- This is the basis for **probability**.

* Long Run Behavior *

Application - Random Babies

Suppose a stork randomly delivers four babies to four different houses. We are going to ***simulate*** this situation and count the number of correct deliveries.



after 25 trials rel freq was .280
after 50 trials rel freq .360
after 200 trials rel freq .260
after 500 trials rel freq .398
" 1000 trials rel freq .393
5000 trials rel freq .371
after 20000 trials rel freq .375

(www.rossmanchance.com/applets/randomBabies/Babies.html)

Plot the proportion of proportion of trials where there were 0 matches. What are we observing when we run the simulation a large number of times? "Convergence"

About 37.5% of the time none of the babies would end up in correct house.

Law of Large Numbers - The fact that the proportion of trials that had no babies delivered to the right house converges to 0.375 is guaranteed by the **law of large numbers**. This result says that if we observe more and more repetitions of a chance process, the proportion of times that a specific outcome occurs approaches a single value. The single value is called **probability**.

What implications does this have on sampling design and experimental design?

Should use large sample sizes and perform many replications (trials)

Definition: The **probability** of any outcome of a chance process is a number between 0 and 1 that describes the proportion of times the outcome would occur in a very long series of repetitions.

$0 \leq P(x) \leq 1$

0 - event never occurs
1 - event occurs every time (100% of time)

Example. How much should a company charge for an extended warranty for a specific type of cell phone? Suppose that 5% of these cell phones under warranty will be returned, and the cost to replace the phones is $150. If the company knew which phones would go bad, it could charge $150 for these phones and $0 for the rest. However, since the company cannot know which phones will be returned but knows that about 1 in every 20 will be returned. How much should they charge for the extended warranty?

They should charge at least $\frac{\$150}{20} = \$7.50$ for extended warranty

Other examples:

Life insurance - use probability to determine how much to charge
(see example page 2 re)

Example: The probability of getting a sum of 7 when rolling two dice is 1/6. Interpret this value.

If 2 dice were rolled many, many times, the proportion of rolls that resulted in the sum being 7 would be ≈ 1/6.

Example: Athletes are often tested for use of performance-enhancing drugs. Drug tests aren't perfect, they sometimes say that an athlete took a banned substance when that isn't the case (a "false positive"). Other times, the test concludes that the athlete is "clean" when he or she actually took a banned substance (a "false negative"). For one commonly used drug test, the probability of a false negative is 0.03. Interpret this probability.

If a drug test is conducted many times, we can expect a false negative 3% of the time.

Example: In the popular Texas hold'em variety of poker, players make their best five-card poker hand by combining the two cards they are dealt with three of five cards available to all players. You read in a book on poker that if you hold a pair (two cards of the same rank) in your hand, the probability of getting four of a kind is 88/1000

a) Explain what this probability means.

If we look at many, many hands of poker in which you hold a pair, the proportion of times in which you make four of a kind is ≈ 88/1000.

b) Why doesn't this probability say that if you play 1000 such hands, exactly 88 will be four of a kind?

It does not mean that exactly 88 out of 1000 such hands would give 4 of a kind: It just means over the long run the probability converges to 88/1000 (or 8.8%). Exact # of 4 of a kinds will vary sample to sample

CHECK YOUR UNDERSTANDING

1. According to the "Book of Odds," the probability that a randomly selected U.S. adult usually eats breakfast is 0.61.

(a) Explain what probability 0.61 means in this setting.

This means that if you asked a large sample of U.S. adults if they usually eat breakfast, ≈ 61 % would say yes.

(b) Why doesn't this probability say that if 100 U.S. adults are chosen at random, exactly 61 of them usually eat breakfast?

Exact # who eat breakfast vary sample to sample.

2. Probability is a measure of how likely an outcome is to occur. Match one of the probabilities that follow with each statement. Be prepared to defend your answer. 0 0.01 0.3 0.6 0.99 1

(a) This outcome is impossible. It can never occur.

Probability is 0. If outcome can never occur, it occurs in 0 % of cases.

b) This outcome is certain. It will occur on every trial.

Probability is 1, which means it occurs in 100 % of cases.

(c) This outcome is very unlikely, but it will occur once in a while in a long sequence of trials.

Probability is .01. Outcome is rare, but will occur in 1% of cases.

(d) This outcome will occur more often than not. (more than 50 %) so 0.6 or 0.99

Best answer is 0.6. This means outcome will occur in 60 % of cases

Wording leads us to believe event occurs often, but not nearly every time, so .99 is not best answer.

2. Myths about Randomness - The idea of probability seems straightforward. It answers the question "What would happen if we did this many times?" In fact, both the behavior of random phenomena and the idea of probability are a bit subtle. We meet chance behavior constantly, and psychologists tell us we deal with it poorly. Unfortunately, our intuition about randomness tries to tell us that random phenomena should also be predictable in the short run. When they aren't, we look for some explanation other than chance variation.

1. The Myth of Short-Term Regularity

Myth: Random phenomena is predictable in the short run

Truth: Patterns are very common in the short run

Example – Runs in Coin Tossing
What looks Random
Toss a coin six times and record heads (H) or tails (T) on each toss. Which of the following outcomes is more probable?

HTHTTH TTTHHH

Almost everyone says that HTHTTH is more probable, because TTTHHH does not "look random." In fact, both are equally likely. That heads and tails are equally probable says only that about half of a very long sequence of tosses will be heads. It doesn't say that heads and tails must come close to alternating in the short run. The coin has no memory. It doesn't know what past outcomes were, and it can't try to create a balanced sequence.

The outcome TTTHHH in tossing six coins looks unusual because of the runs of 3 straight tails and 3 straight heads. Runs seem "not random" to our intuition but are quite common.

Example: Roll a die 12 times and record the result of each roll. Which of the following outcomes is more probable?

123456654321 154524336126

These outcomes are equally likely, even though the 1st set of rolls has a more noticeable pattern.
(So people might say the 1st set is less likely, but this is incorrect. Both are equally likely!)

Application - Suppose that a basketball announcer suggests that a certain player is *streaky*. That is, the announcer believes that if the player makes a shot, then he is more likely to make his next shot. As evidence, he points to a recent game where the player took 30 shots and had a streak of 7 in a row. Is this evidence of streakiness or could it have occurred by chance? Assuming the player makes 50% of his shots and the results of a shot do not depend on previous shots, how likely is it for the player to have a streak of 7 or more made shots in a row?

Solution (4 Step Process):

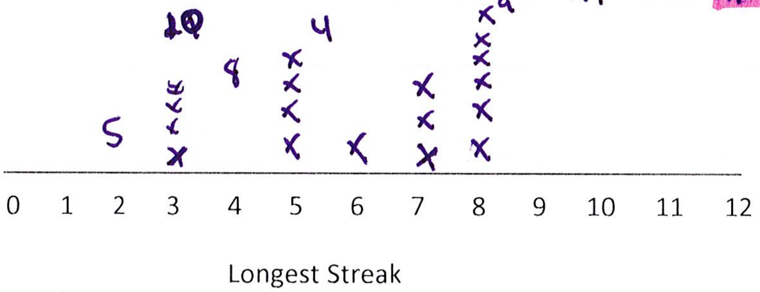
1. **State** - How likely is it for the player to have a streak of 7 or more made shots in a row?

2. **Plan** - Use the random number generator on the calculator to generate 30 random 0s and 1s. 0 = misses shot, 1 = makes shot. Record the outcome of each trial. Record the longest streaks on a dot plot.

3. **Do** - Have each student do this 2 times. Rand Int (0, 1) or Rand Int (0,1, 30)

Trial 1 000100 111 00001 0001 011 011 00 1100

Trial 2 11100 10000 10110 00111 000 11 11111



4. **Conclude** $\frac{12}{50}$ = 24% of the time is likely to make 7 or more shots in a row. Since it is somewhat likely to have a streak of 7 or more just by chance, we do not have convincing evidence that this player is streaky

2. The myth of the "law of averages"

The Myth of the "law of averages": is that when a long streak of a certain outcome occurs, the future outcomes must make up for an imbalance in the short term.

When the shooter in the dice game, craps, rolls several winners in a row, some gamblers think she has a "hot hand" and bet that she will keep on winning. Others say that "the law of averages" means that she must now lose so that wins and losses will balance out. Believers in the law of averages think that if you toss a coin six times and get TTTTTT, the next toss must be more likely to give a head. It's true that in the long run heads will appear half the time.

✱ Don't confuse the law of large numbers, which describes the big idea of probability, with the "law of averages" described here. ✱

What is a myth is that future outcomes must make up for an imbalance like six straight tails. Coins and dice have no memories. A coin doesn't know that the first six outcomes were tails, and it can't try to get a head on the next toss to even things out. Of course, things do even out in the long run. That's the law of large numbers in action. After 10,000 tosses, the results of the first six tosses don't matter. They are overwhelmed by the results of the next 9994 tosses.

Myth: If I flip Tails 6 times in a row, I am more likely to get a Heads the 7th time.

Truth: If I flip Tails 6 times in a row, my chance of getting a Heads the 7th time is still 50%

Example: To pass the time during a long drive, you and a friend are keeping track of the makes and models of cars that pass by in the other direction. At one point, you realize that among the last 20 cars, there hasn't been a single Ford. (Currently, about 16% of cars sold in America are Fords). Your friend says, "The law of averages says that the next car is almost certain to be a Ford." Explain to your friend what he doesn't understand about probability.

Assuming that the brand of each car is independant of other cars, the probability that the next car is a Ford does not change, regardless of what brands preceeded it. Only in the long run can we be sure that the # of Fords will approach whatever the expected probability is.

3. Simulations - The application that we just conducted is called a **simulation**. In fact, the vaunted Hyena Problem on day 1 was a simulation. To perform a simulation, we are going to use the venerable 4 Step Process.

Simulation: The imitation of chance behavior, based on a model that accurately reflects the situation.
It is used to estimate probabilities when they are difficult to calculate theoretically.

Performing a Simulation

1. **State** - What is the question of interest about some chance process?

2. **Plan** - Describe how to use a *chance device* to imitate one repetition of the process. Explain clearly how to identify the outcomes of the chance process and what variable to measure.

3. **Do** - Perform *many* repetitions of the simulation. (At least 30.)

4. **Conclude** - Use the results of the simulation to answer the question of interest.

↑ in context

Examples of 4-Step Process

Refer to the table on p. 290

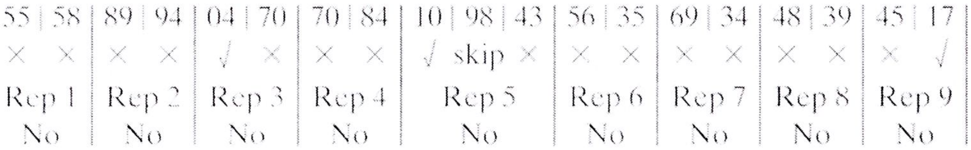
Example – Golden Ticket Parking Lottery
Simulations with a Table of Random Digits

At a local high school, 95 students have permission to park on campus. Each month, the student council holds a "golden ticket parking lottery" at a school assembly. The two lucky winners are given reserved parking spots next to the school's main entrance. Last month, the winning tickets were drawn by a student council member from the AP Statistics class. When both golden tickets went to members of that same class, some people thought the lottery had been rigged. There are 28 students in the AP Statistics class, all of whom are eligible to park on campus. Design and carry out a simulation to decide whether it's plausible that the lottery was carried out fairly.

STATE: What's the probability that a fair lottery would result in two winners from the AP Statistics class?

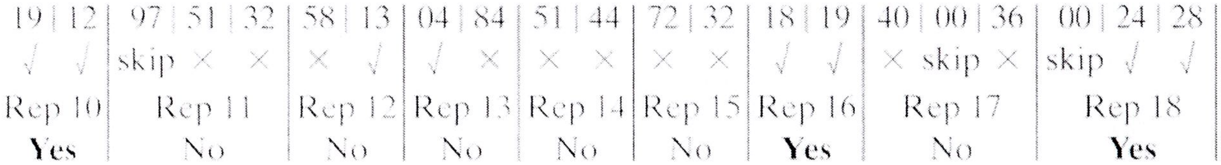
PLAN: We'll use table D to simulate choosing the golden ticket lottery winners. Since there are 95 eligible students in the lottery, we'll label the students in the AP Statistics class from 01 to 28, and the remaining students from 29 to 95. numbers from 96 to 00 will be skipped. Moving left to right across the row, we'll look at pairs of digits until we come across two different labels from 01 to 95. the two students with these labels will win the prime parking spaces. We will record whether both winners are members of the AP Statistics class (Yes or no)

DO: let's perform many repetitions of our simulation. We'll use table D starting at line 139. the digits from that row are shown below. We have drawn vertical bars to separate the pairs of digits. Underneath each pair, we have marked a √ if the chosen student is in the AP Statistics class, × if the student is not in the class, and "skip" if the number isn't between 01 and 95 or if that student was already selected. (note that if the consecutive "70" labels had been in the same repetition, we would have skipped the second one.)



55	58	89	94	04	70	70	84	10	98	43	56	35	69	34	48	39	45	17
×	×	×	×	√	×	×	×	√	skip	×	×	×	×	×	×	×	×	√
Rep 1		Rep 2		Rep 3		Rep 4		Rep 5			Rep 6		Rep 7		Rep 8		Rep 9	
No		No		No		No		No			No		No		No		No	

In the first 9 repetitions, the two winners have never both been from the AP Statistics class. But 9 isn't many repetitions of the simulation. Continuing where we left off,



19	12	97	51	32	58	13	04	84	51	44	72	32	18	19	40	00	36	00	24	28
√	√	skip	×	×	×	√	√	×	×	×	×	×	√	√	×	skip	×	skip	√	√
Rep 10		Rep 11			Rep 12		Rep 13		Rep 14		Rep 15		Rep 16		Rep 17			Rep 18		
Yes		No			No		No		No		No		**Yes**		No			**Yes**		

So after 18 repetitions, there have been 3 times when both winners were in the AP Statistics class. If we keep going for 32 more repetitions (to bring our total to 50), we find 30 more "no" and 2 more "Yes" results. all totaled, that's 5 "Yes" and 45 "no" results.

CONCLUDE: In our simulation of a fair lottery, both winners came from the AP Statistics class in 10% of the repetitions. So about 1 in every 10 times the student council holds the golden ticket lottery, this will happen just by chance. It seems plausible that the lottery was conducted fairly.

AP EXAM TIP On the AP exam, you may be asked to describe how you will perform a simulation using rows of random digits. If so, provide a clear enough description of your simulation process for the reader to get the same results you did from only your written explanation.

In the previous example, we could have saved a little time by using randInt(1, 95)repeatedly instead of a table of random digits (so we wouldn't have to worry about numbers 96 to 00).We'll take this alternate approach in the next example.

Example – NASCAR Cards and Cereal Boxes

Simulations with technology

In an attempt to increase sales, a breakfast cereal company decides to offer a NASCAR promotion. Each box of cereal will contain a collectible card featuring one of these NASCAR drivers: Jeff Gordon, Dale Earnhardt, Jr., Tony Stewart, Danica Patrick, or Jimmie Johnson. The company says that each of the 5 cards is equally likely to appear in any box of cereal. A NASCAR fan decides to keep buying boxes of the cereal until she has all 5 drivers' cards. She is surprised when it takes her 23 boxes to get the full set of cards. Should she be surprised? Design and carry out a simulation to help answer this question.

STATE: What is the probability that it will take 23 or more boxes to get a full set of 5 NASCAR collectible cards?

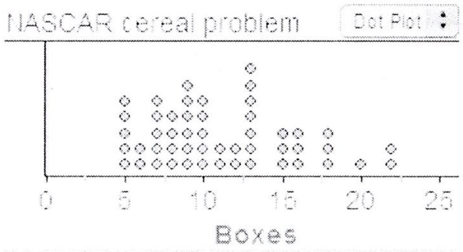
PLAN: We need five numbers to represent the five possible cards. let's let 1 = Jeff Gordon, 2 = Dale earnhardt, Jr., 3 = tony Stewart, 4 = Danica Patrick, and 5 = Jimmie Johnson. We'll use randInt(1,5) to simulate buying one box of cereal and looking at which card is inside. In the golden ticket lottery example, we ignored repeated numbers from 01 to 95 within a given repetition. That's because the chance process involved sampling students without replacement. In the NASCAR example, we allowed repeated numbers from 1 to 5 in a given repetition. That's because the chance process of pretending to buy boxes of cereal and looking inside could have resulted in the same driver's card appearing in more than one box. Since we want a full set of cards, we'll keep pressing enter until we get all five of the labels from 1 to 5. We'll record the number of boxes that we had to open.

DO: It's time to perform many repetitions of the simulation. Here are our first few results:

352152354 9 boxes   5125141412224453 16 boxes   5552412153 10 boxes

435351115315452 15 boxes   3322124334223332334225 22 boxes

The Fathom dotplot shows the number of boxes we had to buy in 50 repetitions of the simulation.



CONCLUDE: We never had to buy more than 22 boxes to get the full set of NASCAR drivers' cards in 50 repetitions of our simulation. So our estimate of the probability that it takes 23 or more boxes to get a full set is roughly 0. The NASCAR fan should be surprised about how many boxes she had to buy.

Example: On her drive to work every day, Sara passes through an intersection with a traffic light. The light has a probability 1/3 of being green when she gets to the intersection. Explain how you would use each chance device to simulate whether the light is green or not green on a given day.

A) A six-sided die.

Assign #'s 1 and 2 to represent a green light and #'s 3, 4, 5, 6 to represent not green light Roll the die and record the result.

B) Table D of random digits

Assign #'s 0, 1, 2 to represent a green light. Assign #'s 3, 4, 5, 6, 7, 8 to represent a not green light. Choose a line in Table D, and look up the # listed in the row. If it is 9, ignore it.

Example:

Suppose I want to choose a simple random sample of size 6 from a group of 60 seniors and 30 juniors. To do this, I write each person's name on an equally sized piece of paper and mix them up in a large grocery bag. Just as I am about to select the first name, a thoughtful student suggests that I should stratify by class. I agree, and we decide it would be appropriate to select 4 seniors and 2 juniors. However, since I already mixed up the names, I don't want to have separate them all again. Instead, I will select names one at a time from the bag until I get 4 seniors and 2 juniors. This means, however, that I may need to select more than 6 names (e.g. I may get more than 2 juniors before I get the 4 seniors). Design and carry out a simulation to estimate the probability that you must draw 8 or more names to get 4 seniors and 2 juniors.

State: What is the probability that you must draw 8 or more names to get 4 seniors and 2 juniors?

Plan: We will use table D to simulate drawing 4 seniors + 2 juniors. Assign the seniors #'s 01-60, and the juniors 61-90. (Ignore #'s 00, and 91-99) Choose a row in table D and move left to right across a row. We will look at pairs of digits until we have 4 unique labels from 01 to 60 and 2 unique labels from 61-90 (this means ignore repeats) Then, we will count how many different labels from 01-90 we looked at. Plot results in Dot Plot.

DO: (2 Trials per student)

Conclude:

(We are not finishing this problem, below is an example of using table with line 101)

Example of 1 Trial with Line 101:

19	22	39	50	34	05	75	62
S	S	S	S	S	S	J	J
1	2	3	4				

It took 8 selections to get at least 4 seniors and 2 juniors

S S S 5 6 S 7 S 8 S J J

Line								
101	19223	95034	05756	28713	96409	12531	42544	82853

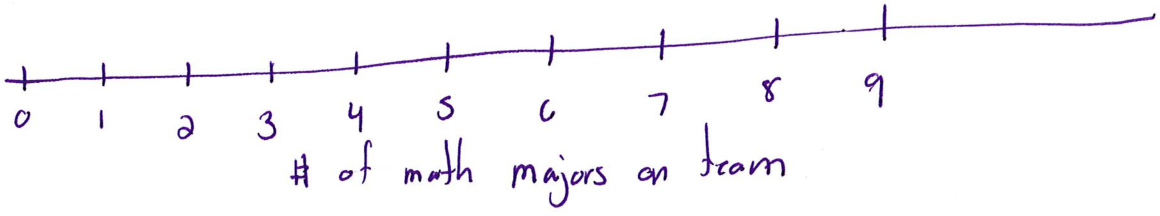
Application - At a department picnic, 18 students in the mathematics/statistics department at a university decide to play a softball game. Twelve of the 18 students are math majors and 6 are stats majors. To divide into two teams of 9, one of the professors put all the players' names into a hat and drew out 9 players to form one team, with the remaining 9 players forming the other team. The players were surprised when one team was made up entirely of math majors. Is it possible that the names were not adequately mixed in the hat, or could this have happened by chance? Design and carry out a simulation to help answer this question.

1) State: What is the probability that when randomly assigning 12 math majors + 6 Stats majors to 2 teams, there will be one team with all Math Majors?

② Plan: Using a random # generator, generate random integers 1-18, 1-12 for Math Major, 13-18 for Stats Major. 1st 9 #'s go to 1st team, remainder go to 2nd team. ignore repeats Count # of math majors on each team and record results. Plot results in a dot plot.

③ Do: (2 Trials per student)

Results: Trial 1: Team 1 Team 2



④ Conclude: