MAT 105 - Summary of Lecture 1
Tuesday, January 5, 2016

> *It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge. The most important questions of life are indeed, for the most part, really only problems of probability.* (Laplace, 1812)

Probability is the language of statistics and the foundations of statistical inference are based on theoretical probability. Why does one study probability and statistics? Short answer: because it is useful and because it is fun.

History: As early as the 17th century, people used probability (just fancy counting at this point) to find odds in gambling. In fact, famous problems such as the Problem of Points were resolved during this time. At the same time, statistics was being used to keep track of mortality rates, economic and political data. Actuarial science developed, with insurance companies protecting investments against risk. Subsequent centuries saw development in these fields, with mathematicians using probabilistic models to solve such problems; in addition, physics and in particular quantum mechanics became intimately linked to probability ideas. For example, it was discovered that states of atoms are described by probability distributions, and not ruled by determinism, as previously believed. By 1940's, probability became a science in itself, whose theoretical foundations have been formulated and in which many of the initial unsolved paradoxes have been resolved.

Some of today's applications: We use probability and statistics every day. It is used in population genetics and medicine; data mining and machine learning; image processing; wireless communication system design; web search engine design; insurance risk; stock market and financial engineering; elections, census and public policy; psychology and social science research; weather (we would like to believe so); polymer chemistry and physics. It is still used to compute odds of winning games, odds in sporting events and in gambling. Some people even study it just for fun!

The best way to learn probability is via examples. Here are some of my favorite examples. The full solution for some of them will be developed throughout the semester.

1. This problem is courtesy of M. Kozdron of University of Regina: Each year, the Canadian coffee and donut chain Tim Hortons brings back its "Roll Up The Rim To Win" contest. The basic idea is that each coffee cup serves as a game piece and by looking under the rim of your cup, you can determine whether or not you won a prize. The official rules state that *there is a 1-in-9 chance of winning a prize.*

(a) Interpret this last phrase. In other words, what does Tim Hortons mean when they claim there is a 1-in-9 chance of winning a prize?

(b) Suppose that you buy 8 cups of coffee and do not win a prize with any of those cups. Are you guaranteed to win a prize when you buy your 9th cup of coffee?

(c) Suppose that you buy 18 cups of coffee. How many prizes will you win? How many prizes do you expect to win?

Solution:

(a) Here are a few ideas as to what the rules say:

- On average, 1 out of 9 cups of coffee will win.
- The more you play the game, the closer you will be to averaging 1/9 winning cups.
- 1/9 of all the cups have winning pieces.
- each cup has a 1/9 chance of being a winner.

(b) No, we are not guaranteed to win on the 9th cup, in fact, the probability that the 9th cup is a winner is 1/9.

(c) If we buy 18 cups of coffee, we *expect* to have about 2 winners. However, that is not guaranteed! We can only say with certainty that there will be somewhere between 0 and 18 winners.

(d) If we buy 36 cups, we find the following probabilities:

- $P(\text{no winning cups}) = \left(\frac{8}{9}\right)^{36}$, since each cup has 8/9 chance of not having a prize.
- $P(36 \text{ winning cups}) = \left(\frac{1}{9}\right)^{36}$, since each cup independently has 1/9 chance of having a prize.
- $P(1 \text{ winning cup}) = 36\left(\frac{1}{9}\right)\left(\frac{8}{9}\right)^{35}$, since each cup independently has 1/9 chance of having a prize and 8/9 chance of not having a prize. The 36 comes from the fact that the winning cup could be the 1st cup, or the 2nd cup, etc., with 36 possible choices as to which cup has the prize.

2. This problem was posed by Chevalier de Méré and was solved by Blaise Pascal and Pierre Fermat. Find the probability of rolling at least one six when a die is rolled 4 times. Also find the probability that a double six comes up at least once, when a pair of dice is rolled 24 times. Which probability is greater?

Solution: Note that the probability of a die landing on 6 is 1/6 and the probability of a double

6 for a pair of dice is 1/36. Let E be shorthand for rolling at least one 6 in 4 rolls. Then if E^c denotes the complement of E,

$$P(E) = 1 - P(E^c) = 1 - \left(\frac{5}{6}\right)^4 = .5177.$$

Let F be shorthand for rolling at least one double 6 in 24 rolls of two dice. Then

$$P(F) = 1 - P(F^c) = 1 - \left(\frac{35}{36}\right)^{24} = .4914.$$

It is more likely that a 6 six comes up in 4 single rolls, than a double 6 in 24 rolls of pairs of dice. In fact, history says that Chevalier de Mere first used to bet as in (a) and then switched to the bet in (b) and did not quite know why he was now losing money:

According to the reasoning of Chevalier de Mere, two 6's in two rolls are 1/6 as likely as one 6 in one roll. (Which is correct.) To compensate, de Mere thought, the two dice should be rolled 6 times. And to achieve the probability of one 6 in four rolls, the number of the rolls should be increased four fold - to 24. Thus reasoned Chevalier de Mere who expected a couple of 6's to turn up in 24 double rolls with the frequency of a 6 in 4 single rolls. However, he lost consistently.

3. The problem of points was considered by Fermat and Pascal and solved in 1654 by Pascal. We considered an easy version of the problem. Suppose two teams, Team A and Team B play a game and the team that first wins 5 rounds wins the game. Also suppose that each team puts down the same amount of money for the game and the winner takes all. Now due to unforseen circumstances, the game is stopped before any team can win 5 rounds. The score at this point is 3:1 in favor of Team A. How should the money be divided? Which would be the fairest way to split the money?

   Solution: Some possibilities:

   (a) The bookie keeps the money.

   (b) Each team takes back their money.

   (c) Teams flip a coin for the total amount.

   (d) Team A takes all the money, since it is ahead when the game is stopped.

   (e) Team A takes 3 parts of the pot and Team B takes one part, based on the current score.

   (f) Team A takes 13/16 of the pot and Team B takes 3/16, based on the probability of each

to win the game, if they continue to play. Note that we computed

$$\begin{aligned} P(B \text{ wins}) &= P(BBBB) + P(ABBBB) + P(BABBB) + P(BBABB) + P(BBBAB) \\ &= \left(\frac{1}{2}\right)^4 + 4\left(\frac{1}{2}\right)^5 = \frac{3}{16}, \end{aligned}$$

where $P(BABBB)$ denotes the chance that the next rounds are won by Team B, A, B, B, B, in this order and we assumed that in subsequent rounds, each team has equal winning chance (1/2). Then we conclude $P(A \text{ wins}) = 1 - P(B \text{ wins}) = 13/16$.

For example, if we assume that each team contributes \$48 for the game, then Team A receives \$78 and Team B receives \$18.

4. **The birthday problem.** Suppose there ar n people in the room. Would you bet that at least two share a birthday (common day and month, assuming 366 days in a year)?

   We decided that it depends on n. If $n < 2$, then clearly the probability of a shared birthday is 0, so we would not. If $n > 366$, then at least 2 *must* share a birthday, so we will win such a bet with certainty. How about for groups of sizes in between 2 and 366? Clearly, the closer to 0, the less likely it is to win the bet. One question of interest is *for which n is there a chance larger than 50% to win such a bet?*

   - if $n = 0$, $P\{win\} = 0$
   - if $n = 16$, $P\{win\} \approx .28$
   - if $n = 22$, $P\{win\} \approx .49$
   - if $n = 23$, $P\{win\} \approx .51$
   - if $n = 30$, $P\{win\} \approx .7$
   - if $n = 50$, $P\{win\} \approx .97$

   Here is how one might compute these probabilities at this point (we do it for 16 only, the other values less than 366 work the same). Let E denote the event that there is at least a pair sharing a birthday, for shorthand. It will be easier to compute the probability of the opposite event, which we call the *complement* E^c; that is, we find the probability of no two sharing a birthday. So, suppose the first person on my list of 16 is born on some day of the year, it does not matter which one, with probability $\frac{366}{366}$. Then the second person cannot share this day, hence it has a different birthday with chance $\frac{365}{366}$. The third person, cannot share the first two days that were picked, so they do not share a birthday with any of the first two, with probability $\frac{364}{366}$. We continue reasoning this way, and the 16th on the list has a different birthday than any of the others with probability $\frac{351}{366}$. Then the probability of this complementary event is

$$P(E^c) = \frac{366}{366} \times \frac{365}{366} \times \frac{364}{366} \times \cdots \times \frac{351}{366} \approx .72.$$

   Therefore, $P(E) = .28$. This argument uses a few of the ideas and principles we will study later in the semester, such as multiplication rule in counting, conditioning, complementary events, but I hope intuitively it makes sense as to why we arrive at this probability.

MAT 105 - Summary of Lecture 2
Thursday, January 7, 2016

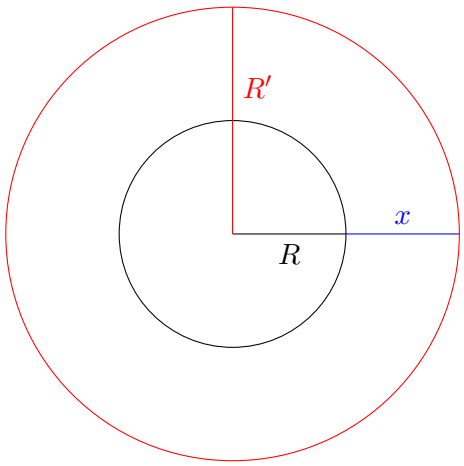
Chance favors the prepared mind. (Louis Pasteur, 1854)

While we may be tempted to use probability as a subjective measure of our belief, we must keep in mind that our intuition does not always lead to a correct solution. However, we can use rigourous arguments to verify or disprove these solutions. The following three problems stresses this point.

1. **Rope around the Earth.** Let's assume the Earth is perfectly round and we have a rope snugly placed around its equator. Expand this rope by 20 meters and stretch it so it forms a circle around the equator (think of it as suspended rope at some distance above the equator). What can be placed under the rope? An amoeba, an ant, a small child, a tall human standing up or nothing at all?

   Our intuition tells us that those 20 meters should be distributed to a large radius, so only something very small could fit under the rope. But our intuition is wrong!

   This problem has a very simple solution. It might be useful to draw a picture:



   Let R denote the radius of the Earth. The length of the original rope is the $2\pi R$. To this, we add 20 meters and the new rope has length $2\pi R + 20$. Since the new rope is placed around the circumference of a "suspended" circle of radius R', we can express the length of the new rope as

$$2\pi R' = 2\pi R + 20 \quad \Rightarrow \quad R' = \frac{2\pi R + 20}{2\pi} = R + \frac{10}{\pi}.$$

   The "walking" space between the equator and the suspended rope is $x = R' - R = \dfrac{10}{\pi}$, so as long as the creature is not taller than 3.18 meters, it can fit under the rope.

2. **Monty Hall Problem.** A car and two goats are hidden behind 3 doors. The point of the game show is to win the car. A contestant is asked to choose a door, after which the game show host opens one of the doors, always showing a goat. At this point, the contestant is asked if he wants to change his mind and switch doors. Should the player switch? Will that increase his odds to win the car?

   Yes, the player should always switch. Here is a heuristic argument. At the beginning, he picks a door with 1/3 chance of hiding the car. The remaining doors have a 2/3 chance of hiding the car. After the game show host shows a goat, the door that was not picked still holds 2/3 chance of hiding the car, so the contestant should switch doors and increase his chances two-fold. More precisely, we have the following 3 cases, each with equal chance of occurring

	Door1	Door2	Door3		Door1	Door2	Door3	Shown Door	Result
Case 1	C	G	G	pick Door 1 $\longrightarrow$	X	G	G	2 or 3	G
Case 2	G	C	G		X	C	G	3	C
Case 3	G	G	C		X	G	C	2	C

   That means, if the contestant switches, in 2 out of 3 cases he will win a car.

   If there are more than 3 doors with many goats and a car, as long as the game show host uncovers doors hiding goats, the contestant should always switch, as his odds will increase. Use the applet from Moodle to convince yourself that this is still the case. The settings in the app should be intuitive.

3. **A medical test problem.** Suppose that 1% of the population has some terrible disease. There is a medical test available, which detects this disease in a sick person 99% of the time. The test also gives false positives 10% of the time, that is, the test is positive for 10% of the people tested who do not have the disease. You are tested for this disease and the test is positive. What are your chances of actually having the disease?

   While counterintuitive, there is only a 9.09% chance that you have the disease and that is because the odds of having the disease are very small to start with. Suppose 10,000 are tested for the disease. Then about 100 have the disease (1% of 10,000), out of which 99 test positive (99% 0f 100). The remaining 9900 people (99% of 10,000) do not have the disease, but 990 of them test positive (10% of 9900). So, out of 10,000 people, about $99 + 990 = 1089$ test positive, but only 99 of them are sick. So the fraction of the sick people to those who tested positive is $\frac{99}{1089} = \frac{1}{11} = .0909$, which give the probability that you have the disease when you tested positive.

   An important question we should ask ourselves is *"Is this test a good test?"*. It clearly detects most of the sick people, but it also gives a lot of false positives. We will consider these sort of questions when discussing types of error in hypothesis testing.

The last two problems will be derived in detail, once we develop the necessary theory behind their solutions.

1 Basic definitions

The main question we'll try to answer in this course is how can we draw good, reliable and useful conclusions from incomplete information? There are a few steps we take to answer this question.

1. Collect data (design experiments to verify hypothesis)
2. Organize and represent data (the focus of next week's lectures)
3. Analyze data (we use probability and statistics ideas)
4. Draw conclusions (we use common sense and statistical inference)
5. Measure confidence in the conclusion (we use confidence intervals derived from the Central Limit Theorem)

A **population** is the collection of all measurements or observations of interest. A **sample** is a subset, or a part of the population. A **random sample**, which is thought of as a representative sample, is determined by chance, in the sense that every member of the population has an equal chance of being picked in the sample.

Example: I want to study the height of people in the US. The population is the collection of all heights of all people in the US. A sample could be the heights all 10-20 year olds, the heights of all students in our class, the heights of all NBA players – these are all samples that are not random. A sample representative of the population might be achieved by randomly selecting Social Security Numbers and assessing the heights of these people.

One question to keep in mind is "how large should the sample be?" This depends on what we want to measure: the range of heights, the average height, the spread of all heights etc, as well the level of confidence and margin of error we would like for our measurements.

2 Data collection / experiments

We will touch upon the first question "how do we gather data?" throughout the semester. Of course, it depends on whether we want to prove a claim, or we just collect it without a specific purpose in mind and analyze it in order to draw conclusions about it. Probability and statistics can answer both questions, at least partially. To get us thinking about this topic, let's consider the following example.

Example: In 1778, Captain Cook "discovered" Hawaii, where he introduced several goats, which multiplied to several thousand over the years. Among other things, they enjoy eating the *silver sword plant*, which was at one point (1920) on the endangered species list. Steps have been taken since then to preserve this plant, but one question remains "Are goats to blame for the disappearance of silver sword plants from Maui?". Suppose we want to design an experiment to verify the claim that indeed the goats deserve part of the blame. How would we devise the experiment and what kind of data would we look for?

Here are a few ideas:

- Gather data on the number of plants before the goats were introduced and after. This is not only hard data to find, but might not be very relevant, as the goats were introduced a long time ago and many things have changed since then: other plants might have been introduced, climate might have changed, level of pollution increased etc.
- Design an experiment that shows preference of goats for this plant. However, the influence of goats on the plant population might not be directly related to which plant they prefer. Maybe they eat whatever is green without discrimination. Maybe they eat this plant less often, but it takes longer for it to grow back. Knowing this kind of information, would not help our analysis.
- Create a controlled environment: on Lot 1 we keep track of the plant development in the absence of goats, on Lot 2 we consider plant numbers in a space roamed by goats.
- Study relationships and correlations between the number of goats and the number of plants. While this can provide some insight, it is not the best approach. We should consider it in the absence of a controlled environment study.

The way the experiment was conducted was indeed by creating a controlled environment in the national park on Haleakala: one lot was fenced, so goats could not enter it and the second lot was left unfenced. Based on the data recorded, it was shown that goats reduce the number of silver sword plants by 1/4 or more in the unfenced lot, with a high level of confidence. Therefore, it was concluded that goats are partly to blame for the disappearance of this plant from Maui.