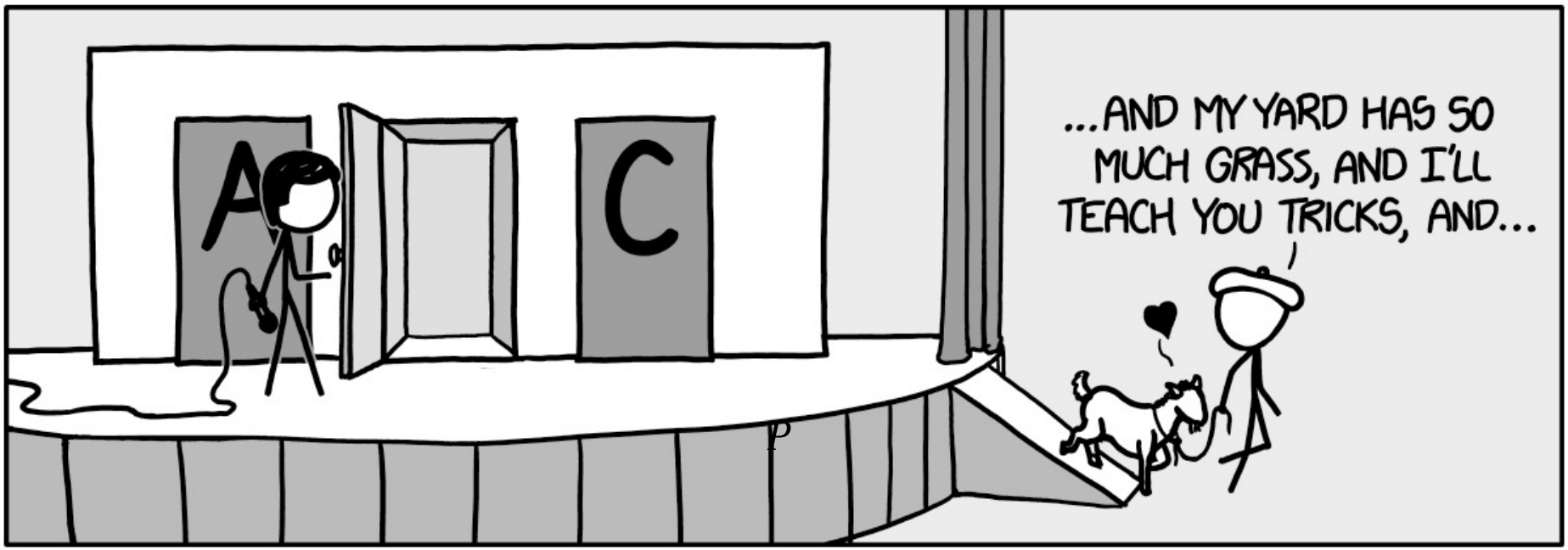
Stat 88: Probability and Statistics in Data Science



https://xkcd.com/1282/

Lecture 5: 2/1/2022

Symmetry in Sampling, Bayes' Rule, Random variables

Sections 2.2, 2.3, 2.4, 3.2

Shobhana M. Stoyanov

Agenda

- § 2.2: Symmetries in sampling & counting
- § 2.3: Bayes' rule
- § 2.4: Use and interpretation of Bayes' rule
- § 3.2: Random variables: intro

Review: Product rule and counting

- Recall the product rule for counting: if there are sequences constructed in n stages, with k_i options at each stage, then the total number of sequences is $k_1 \times k_2 \times \cdots \times k_n$

- Count the number of outcomes for each stage and multiply them. (Recall the tree diagrams, and how we count outcomes.)

- Deal 5 cards from a deck. Number of possible sequences?

- Number of outcomes from rolling three 6-sided dice?

- 10 students, choose 2 for committee (to be the president and secretary respectively). Number of possible committees?

Example

- The English language has 26 letters. 5 letters are chosen **with** replacement. What is the chance that the *middle* three letters are all *different*, and the *first* and *last* are the *same* as each other, and also the *same* as one of the three middle letters.

Probabilities of dealing cards:

- Deal 2 cards from top of the deck.
    - How many possible sequences of 2 cards?
    - What is the chance that the second card is red?

- P(5^{th} card from top is red)

- 21^{st} card *and* 35^{th} cards are red = $P(R_{21} \cap R_{35})$ = (write it using conditional prob)

- P(7^{th} card is a queen)

- $P(B_{52} \mid R_{21}R_{35})$

Counting permutations & combinations

- Recall # of ways to rearrange n things, taking them 1 at a time is n!
- If we have only $k \leq n$ spots to fill, then $n \cdot (n-1) \cdot \ldots \cdot \big(n-(k-1)\big)$
- # of perm. of n things taken k at a time.

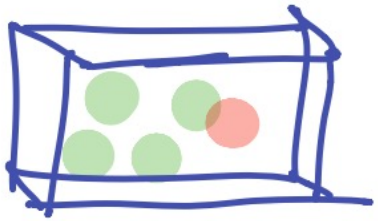
- If we don't care about order, then we are counting subsets, and this number is denoted by $\binom{n}{k}$, which we get by dividing: $n \cdot (n-1) \cdot \ldots \cdot \big(n-(k-1)\big)$ by $k!$

- Note: $\binom{n}{n} = 1, \binom{n}{0} = 1$

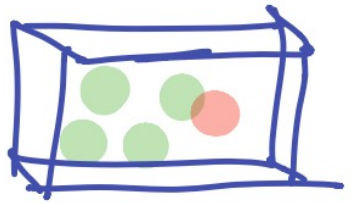
- Prob. of "Full house" =

Section 2.3: Bayes' Rule:

- I have two containers: a jar and a box. Each container has five balls: The jar has three red balls and two green balls, and the box has one red and four green balls.
- Say I pick one of the containers at random, and then pick a ball at random. What is the chance that I picked the box, if I ended with a red ball?



Jars and boxes



Prior and Posterior probabilities

- The ***prior*** probability of drawing the box = ___ (before we knew anything about the balls drawn)

- The ***posterior*** probability of drawing the box = ____ (this is after we *updated* our probability, *given* the information about which ball was drawn)

Computing Posterior Probabilities: Bayes' Rule

- We want the *posterior* probability. That is, the conditional prob for the first stage A, ***given*** the second stage B.
- Division rule (for conditional probability) =

- Using the multiplication rule on $P(A \cap B)$, we get:

- Rule first written down by Rev. Thomas Bayes in the 18th century. Helps us compute posterior probability, given prior prob. And **likelihoods** (which are conditional probabilities for the *second* stage given the first, which are generally easier to compute.)

Exercise 2.6.9

A factory has two widget-producing machines. Machine I produces 80% of the factory's widgets and Machine II produces the rest. Of the widgets produced by Machine I, 95% are of acceptable quality. Machine II is less reliable – only 85% of its widgets are acceptable.

Suppose you pick a widget at random from those produced at the factory.

a) Find the chance that the widget is acceptable, given that it is produced by Machine I. (likelihood)

b) Find the chance that the widget is produced by Machine I, given that it is acceptable. (posterior)

Example: Binge drinking & Alcohol related accidents

(This example is from the text *Intro Stats* by De Veaux, Velleman, and Bock)

For men, binge drinking is defined as having 5 or more drinks in a row and for women as having 4 or more drinks in a row.
(The difference is because of the average difference in weight.)

According to a study by the Harvard School of Public Health (H.Wechsler, G. W. Dow dall, A. Davenport, and W. Dejong, "*Binge Drinking on Campus: Results of a National Study*"):

- 44% of college students engage in binge drinking,37% drink moderately, and 19% abstain entirely. (*priors*)
- Another study, published in American journal of Health Behavior, finds that amon g binge drinkers aged 21 to 34, 17% have been involved in an alcohol-related automobile accident, while among nonbingers of the same age, only 9% have been involved in such accidents. (*likelihoods*)
- Given that a student has been in a car crash, what is the chance that they were a binge drinker? (*posterior*)

Example: Binge drinking & Alcohol related accidents

- Make a tree diagram. What are we given? What do we want to compute?

2.4: Use and interpretation of Bayes' rule

- Harvard study: 60 physicians, students, and house officers at the Harvard Medical school were asked the following question:
- "If a test to detect a disease whose ***prevalence*** is 1/1,000, has a false positive rate of 5 per cent, what is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person's symptoms or signs?"
- *Prevalence* aka *Base Rate* = fraction of population that has disease.
- *False positive rate*: fraction of positive results among people who don't have the disease
- *Positive result*: test is positive

- What is your guess – without any computations?

Tree diagram for disease and positive test

- $P(D|\text{pos. test})$ or *posterior* probability =
- Recall that prior probability = 0.001 = 0.1%

Base Rate Fallacy

- $P(D|\text{pos. test})$ or *posterior* probability =
- Recall that prior probability = 0.001 = 0.1%

- $P(+\text{ test}) = P(+ \text{ \& disease}) + P(+ \text{ \& no disease})$ (since either you have the disease or not, so we have a partition of the event "positive test")

- Base rate fallacy: Ignore the base rate and focus only on the likelihood. (Moral of this story: ignore the base rate at your own peril)

- Note: Want $P(D|+)$ but most people focus on the test giving correct results for negative tests 95% of the time, that is $P(\text{no disease}|\text{ neg})$

- What happens to the posterior probability if we change the prior probability?

Case of Sally Clarke and SIDS: Was this justice? Or quite the opposite?

- Around 2003, Sally Clark, in a famous murder trial had two children one year apart who both died mysteriously. Sally Clarke's defence was that the babies both died of Sudden Infant Death Syndrome (SIDS)
- A = event the first child dies of SIDS
- B = event the second child dies of SIDS.
- Assumption: $P(A) = P(B) = 1/8543$ (based on stats, unconditional probability)

Let's make a deal!: The Monty Hall Problem

There are 3 doors, A, B, C, behind one is a new car (a Ferrari, say), and behind the other two are goats.

Now suppose you are the contestant, and you choose door A. Then Monty Hall opens one of the other two doors, say B, to show you a goat!

He asks you if you want to switch to C or stick with your original choice A, you say...?

Section 3.1: Random variables: Vocabulary

- When we have two kinds of tickets in a box and we draw tickets at random from this box, each draw is called a ***trial***
- We call the two kinds (binary) outcomes ***Success***, and ***Failure***

- Might be with replacement (like a coin toss) or without replacement (taking a sample of students, and checking number that want classes to remain online)

- Read about Paul and Mani.
- Note that Paul made 8 correct predictions. What is the chance of 8 winners if picking completely at random? (Like 8 coin tosses)

3.2 Random Variables

- A real number – we don't know exactly *what* value it will take, but we know the possible values.

- The number of heads when a coin is tossed 3 times could be 0, 1, 2, or 3.
- The sum of spots when a pair of dice is rolled could be 2, 3, 4, 5, …, 12.

- These are both examples of *random variables*.
- *Variable* because the number takes different values
- *Random variable* because the outcomes are not certain.

Random variables

- Using random variables helps to write the event more clearly and concisely.

- It is a way to ***map*** the function space Ω to real numbers
- For example: Let X represent the number of heads in 3 tosses.

- We can write down the ***distribution*** of X, which consists of its possible values and their probabilities.

- The function describing the distribution is called the ***probability mass function***$(f(x))$
- Note that the probabilities must add up to 1.
- We can visualize it using a probability histogram.

Random Variables

- Note that even if two random variables have the same distribution, they are not necessarily equal. For example, let X be the number of heads in 2 tosses of a fair coin, and Y be the number of tails.
- That is, we can talk about the particular values being equal and distributions being equal – and these are not the same thing.

- Mark on table, and the probability histogram, the area $P(X > 0)$ where X is the number of heads in 3 tosses of a fair coin.