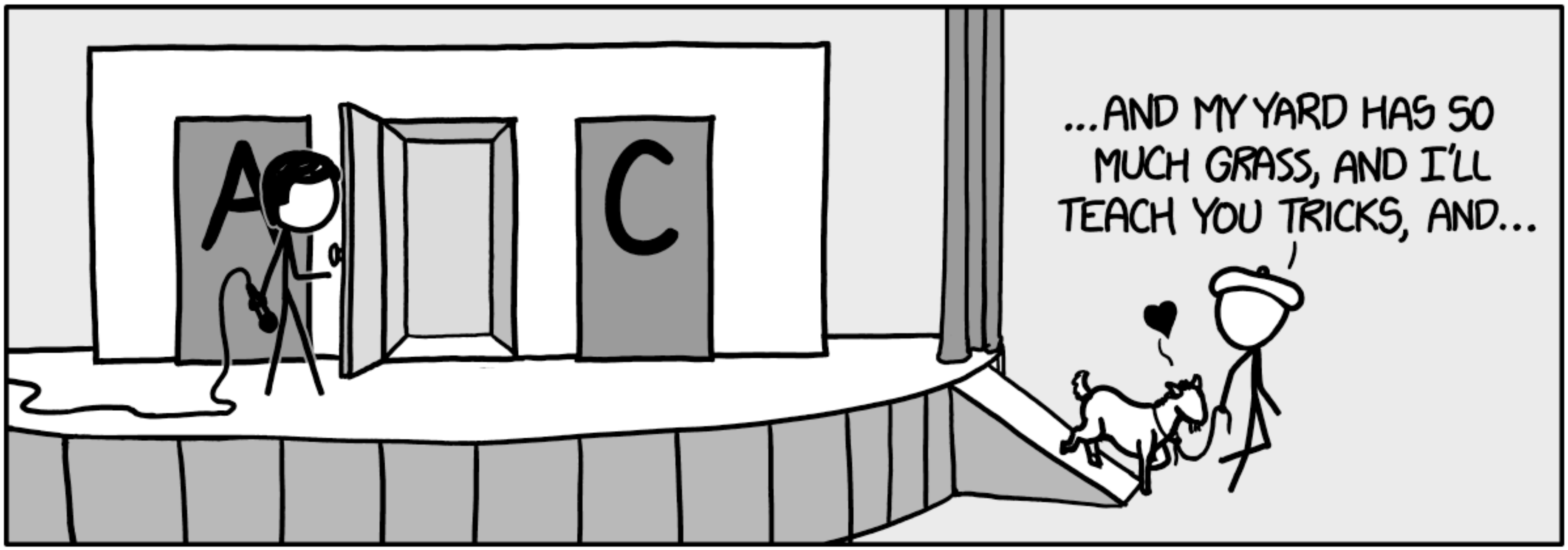
Stat 88: Probability and Mathematical Statistics in Data Science



https://xkcd.com/1282/

Lecture 6: 2/1/2021

Finish up chapter 2, Random variables & their distributions

Sections 2.4, 3.1, 3.2

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…?

Agenda

- Examples from last time, talk about use and interpretation of Bayes' rule
  - Binge drinking
  - Disease, prevalence, base rate, base rate fallacy

- 3.1, 3.2, 3.3
- Review counts, permutations and combinations, $\binom{n}{k}$
- Success and failure
- Random variables
- The binomial distribution

Examples

- $P(B_{52} \mid R_{21}R_{35}) = 26/50$ (2 cards been pulled, both red, 26 black cards left)

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. Dowdall, 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.
- Another study, published in American journal of Health Behavior, finds that among 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.
- Given that a student has been in a car crash, what is the chance that they were a binge drinker?

Tree diagram:

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?"

- What is your guess – without any computations?

- 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

Tree diagram for disease and positive test

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

Base Rate Fallacy

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

- P(+ test) = P(+ & disease) +P(+ & no disease)

- 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(no disease| neg)

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

Case of Sally Clarke and SIDS

- 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)

Back to counting outcomes of tosses

- Toss a coin 8 times, how many possible outcomes?
- What is the chance of all heads?
- If each of the students in this class present today flip a coin 5 times, what is the chance that at least 1 person gets all heads?

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 3.1: 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 (drawing voters from a city and checking number of Bernie supporters)

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

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 are 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. That is, we can talk about the particular values being equal and distributions being equal.

- 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.

3.3 The Binomial distribution

- Many situations can be modeled using the following set up:
  - We have a ***fixed*** number of ***independent*** trials, each of which has ***two*** possible outcomes. "success"(S) and "failure"(F)
  - The probability of success stays constant from trial to trial.

- Example: toss a coin 10 times, count the number of heads
  - Each toss is an independent trial
  - A success is a head.
  - P(success) = 0.5

- Need to specify number of trials (***n***), and P(success) (***p***)
  - Example: number of people who accept credit card offer from bank
  - Number of aces in 10 rolls of a die.

Binomial distribution: Example

- Consider a box with one red ball and eleven blue ones.

- One draw is made. What is the probability that the ball is red?
  - n=1, p=1/12
  - P(R) = 1/12

- Now 4 draws are made, ***with replacement***. What is the probability that ***exactly*** 1 draw is red (out of the 4)?
  - Notice that this is like a tossing a coin 4 times, with P(head) = 1/12.

- P(RBBB) =
- How many such sequences are there?
- What is the probability of all such sequences (1 R, 3B)?

Binomial distribution: Example

- What if we want to compute the probability of 2 red balls in 4 draws? We need the number of sequences of R and B that have 2 R and 2 B.

- P(RRBB) =

- There are 6 such sequences (how?), so if we let X = # of red balls in 4 draws with replacement, we have that

$$P(X = 2) = \binom{n}{k} \times p^2 \times (1-p)^2$$

where p = P(red)

- We say that X has the **Binomial distribution with parameters n and p**, and write it as ***X~Bin(n, p)*** if X takes values 0, 1, …, n and

$$P(X = k) = \binom{n}{k} \times p^k \times (1-p)^{n-k}$$

Identifying binomial random variables

Which of the following are binomial random variables?

- Number of heads in 12 tosses of a fair coin.
- Number of tosses until we see two heads.
- Number of queens in a five card hand
- Number of Democrats in a simple random sample of 500 adult voters drawn from the SF Bay Area.

- Please read about how to compute binomial probabilities using Python.