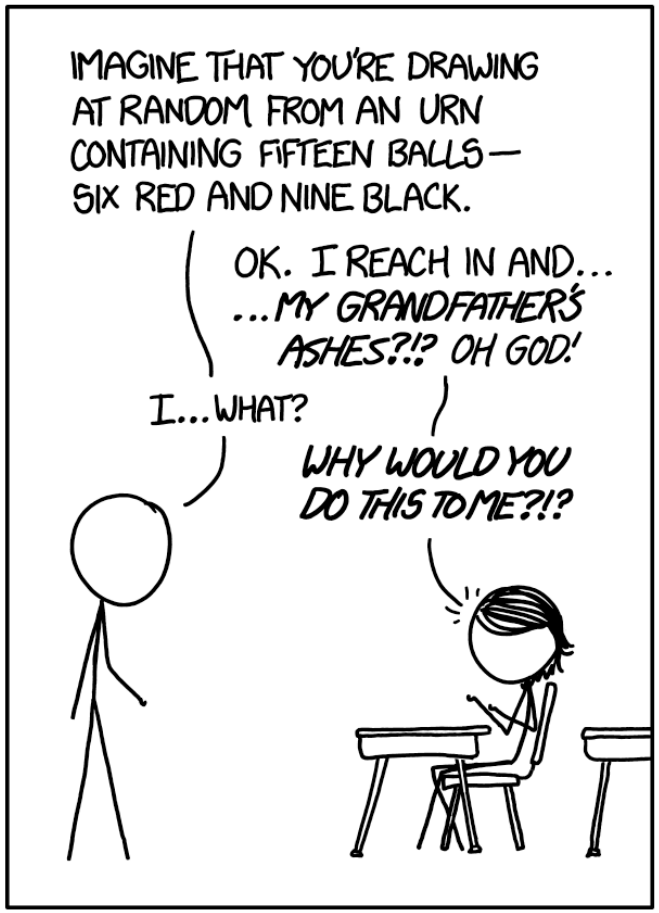
Stat 88: Probability and Statistics in Data Science



https://xkcd.com/1374/

Lecture 6: 1/29/2024

Finishing up chapter 2

Sections 2.2, 2.3, 2.4, 2.5

Shobhana M. Stoyanov

Warm up: Please answer at poll~~e~~v.com/shobhana

POLLEV

Remember: use the product rule of counting, and when you are counting, make dashes: _ · _ · _ · _ · _ (put the number of choices for the first stage or draw, times the number of choices for the second etc.)

$26 \cdot 26 \cdot 26 \cdot 26 \cdot 26 = 26^5$

The English language has 26 letters. 5 letters are chosen **with** replacement.

1. How many possible sequences of 5 letters are there?

2. How many sequences of letters are there which have 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

A. $26 \cdot 25 \cdot 72$

B. $26 \cdot 25 \cdot 24 \cdot 23 \cdot 22$

~~D~~. $5!/3!2!$

b b c d b

$3 \cdot 26 \cdot 25 \cdot 24 \cdot 1$

$3 \cdot 26 \cdot 25 \cdot 24 \cdot 1$

$26 \cdot 25 \cdot 72$

Agenda

- Review warm up problems
- Examples: Symmetry of sampling
- Use and interpretation of Bayes' rule
    - Disease, prevalence, base rate, base rate fallacy
- Examples: Independence

Phataphat = QUICK

Warm up: Poll Everywhere (pollev.com/shobhana)

1. What is the probability that the top card in a standard 52 card deck is a queen, *and* the bottom card is a queen? (write multiplication rule)

$$P(AB) = P(A)P(B|A) = \frac{4}{52} \cdot \frac{3}{51}$$

2. What is the probability that the top card in a standard 52 card deck is a queen *or* the bottom card is a queen? (share pollev results)

$$P(A \cup B) = P(A) + P(B) - P(AB) \overset{??}{=} \frac{4}{52} + \frac{4}{52} - \frac{4}{52} \cdot \frac{3}{51}$$

3. 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 the host, 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. What should you do?

Example: Symmetries in cards

$52 \cdot 51$

- Deal 2 cards from top of the deck.
  - How many possible sequences of 2 cards? $52 \cdot 51$
  - What is the chance that the second card is red? $\frac{26}{52}$

- P(5th card is red) $= \frac{26}{52} = \frac{1}{2}$
- $P(R_{21} \cap R_{35})$ is the prob that 21st card and 35th cards are red.

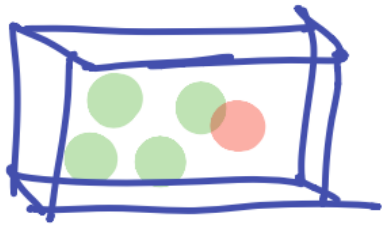
$$P(\underbrace{21\text{st card is red}}_{R_{21}} \;\&\; \underbrace{35^{th} \text{ is } R}_{R_{35}}) = \underbrace{P(R_{21})}_{\frac{1}{2}} \cdot P(R_{35} \mid R_{21}) \quad \underbrace{\frac{25}{51}}$$

- P(7th card is a queen) $\frac{4}{52}$
- $P(B_{52} \mid R_{21} R_{35}) = \frac{26}{50}$

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?

Jar

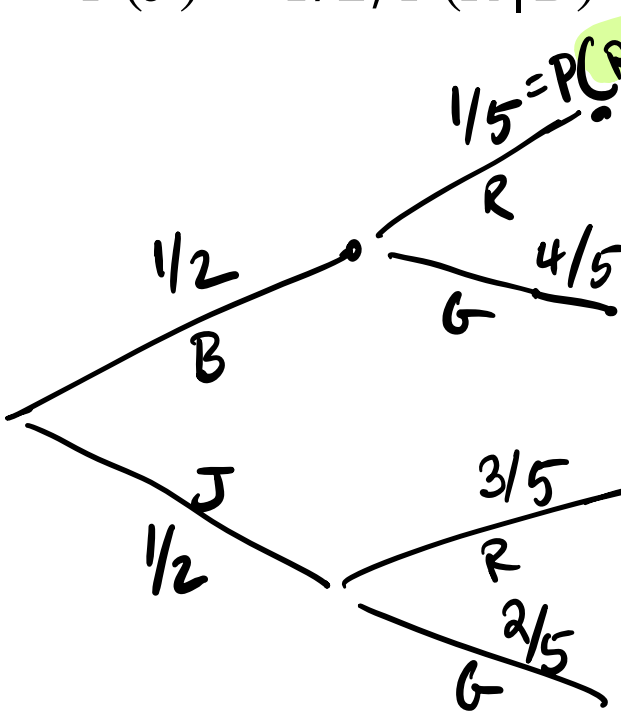


Box

- Let B be the event of picking the box, J the event of picking the jar
- Let R, G be the events of drawing a red ball and a green ball, respectively.

Jars and boxes

$P(B) = P(J) = 1/2;\ P(R\,|\,B) = ?, P(R\,|\,J) = ?$



$\frac{1}{2} \cdot \frac{1}{5} = P(B)\,P(R\,|\,B) = P(B \& R)$

$P(BG) = P(G\,|\,B)P(B) = \frac{4}{5} \cdot \frac{1}{2}$

$P(RJ) = P(R\,|\,J)P(J) = \frac{3}{5} \cdot \frac{1}{2}$

$P(GJ) = \frac{2}{5} \cdot \frac{1}{2}$

$$P(B\,|\,R) = \frac{P(BR)}{P(R)} = \frac{(1/2)(1/5)}{P(RB) + P(RJ)} = \frac{1/2 \cdot 1/5}{1/2 \cdot 1/5 + 1/2 \cdot 3/5}$$

Division rule

$$P(B\,|\,R) = \frac{P(R\,|\,B)P(B)}{P(RB) + P(RJ)} = \frac{1}{5} / 4/5 = \frac{1}{4}$$

P(BR) + P(RJ)

Prior and Posterior probabilities

- The ***prior*** probability of drawing the box = $\frac{1}{2}$ (before we knew anything about the balls drawn)

- The ***posterior*** probability of drawing the box = $\frac{1}{4}$ (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) = $P(B|A) = \frac{P(AB)}{P(A)}$

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(AB)}{P(AB)+P(A^cB)}$$

- 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.)

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

Please Read

- What is your guess – without any computations? You may assume the test is accurate in the sense that if a person does *not* have the disease, the test will be correct. Make a guess at pollev.com/shobhana

Tree diagram for disease and positive test

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

In 1000 people 50 FP
1 disease pos.

$$P(D \mid +) = \frac{1}{51}$$

Base Rate Fallacy

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

- P(+ test) = P(+ & disease) + P(+ & 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 $P(D \mid +)$ 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 the posterior probability if we change the prior 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...?