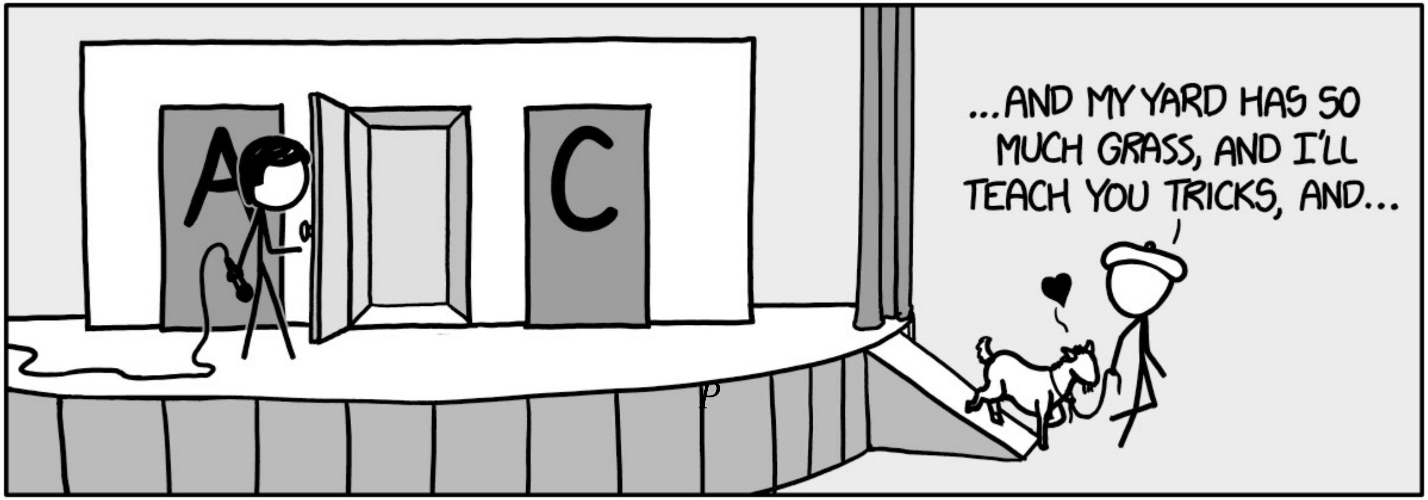
Stat 88: Probability and Statistics in Data Science



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?

$$\underbrace{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48} = \frac{52!}{47!}$$

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

$$6 \cdot 6 \cdot 6 = 6^3$$

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

$\underline{10} \cdot \underline{9}$

What if only <u>who</u> is chosen matters, not which <u>position</u> they are chosen for. $\frac{10 \cdot 9}{2} = 45 = \frac{10!}{8!} \cdot \frac{1}{2!}$

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.

$$\frac{3}{26} \cdot \frac{26}{26} \cdot \frac{25}{26} \cdot \frac{24}{26} \cdot \frac{1}{26} = \frac{\quad}{26^5}$$

Probabilities of dealing cards:

- 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? $\frac{26}{52}$ (no info about 1st card)

- $P(5^{th}$ card from top is red$)$ $= \frac{26}{52} = \frac{1}{2}$

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

$$P(R_{21})\,P(R_{35} \mid R_{21}) = \frac{26}{52} \cdot \frac{25}{51}$$

- $P(7^{th}$ card is a queen$)$ $= \frac{4}{52}$

- $P(\overset{A}{B_{52}} \mid \overset{B}{R_{21}R_{35}}) = \frac{26}{50}$

$$P(A \mid B) = \frac{P(A \cap B \cap C)}{P(B)}$$

Counting permutations & combinations

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

  $\underbrace{n \cdot (n-1)(n-2)(n-3)}_{4 \text{ spots}}$

- 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: $n \cdot (n-1) \cdot \ldots \cdot (n-(k-1))$ by $k!$

  $$\frac{n!}{(n-k)!} \cdot \frac{1}{k!}$$

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

- Prob. of "Full house" = P(triple & pair)

  3♡ 3♢ 3♣ 2♢ 2♡

  $$\frac{\binom{13}{1}\binom{4}{3}\binom{12}{1}\binom{4}{2}}{\binom{52}{5}}$$

  it matters which is the triple & which

26 choices

pick 3 $\frac{26 \cdot 25 \cdot 24}{6}$

abc = acb =

bac = bca =

cab = cba

is pair

Section 2.3: Bayes' Rule:

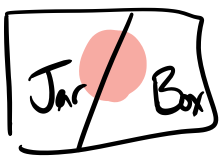
- I have two containers: a jar and a box. Each container has five balls: The jar has three red balls and two green balls, and the box has one red and four green balls.
- Say I pick one of the containers at random, and then pick a ball at random. What is the chance that I picked the box, if I ended with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

$P(\text{jar}) = P(\text{box}) = \frac{1}{2}$ (priors)

$P(\text{jar} \mid \text{red ball})$ = posterior prob.

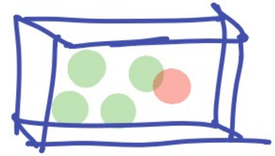
$P(\underset{\text{ball}}{\text{red}} \mid \overset{J}{\text{jar}}) = \frac{3}{5}$, $P(\underset{\text{ball}}{\text{red}} \mid \text{box}) = \frac{1}{5}$ } likelihoods

$$P(\overset{B}{\text{box}} \mid \overset{R}{\text{red}}\text{ ball}) = \frac{P(B \cap R)}{P(R)} = \frac{P(R \mid B)P(B)}{P(R)} = \frac{(\frac{1}{5})(\frac{1}{2})}{P(R)}$$



$$P(R) = P(R \cap B) + P(R \cap J)$$
$$= P(R \mid B)P(B) + P(R \mid J)P(J)$$

Jars and boxes



$$P(R) = \left(\frac{1}{5}\right)\left(\frac{1}{2}\right) + \left(\frac{3}{5}\right)\left(\frac{1}{2}\right)$$

$$= \frac{1+3}{10} = \frac{4}{10} = \frac{2}{5}$$

$$P(B \mid R) = \frac{\left(\frac{1}{\cancel{5}}\right)\left(\frac{1}{2}\right)}{\left(\frac{2}{\cancel{5}}\right)} = \frac{1}{4}$$

Prior and Posterior probabilities

- The ***prior*** probability of drawing the box = 0.5 (before we knew anything about the balls drawn)

- The ***posterior*** probability of drawing the box = 0.25 (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) =

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

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

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B \cap A) + P(B \cap A^c)}$$

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

- 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) $P(A \mid M_1)$

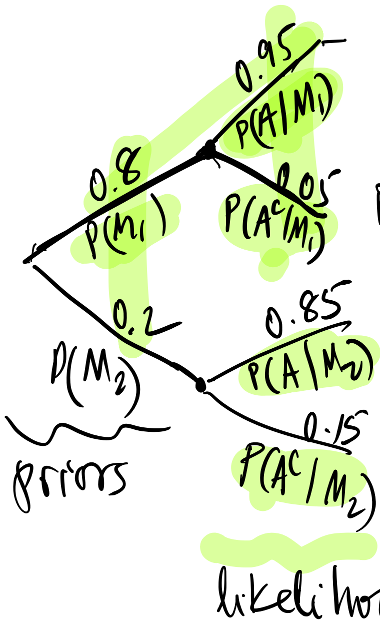
b) Find the chance that the widget is produced by Machine I, given that it is acceptable. (posterior)

M_1, M_2, A: widget is acceptable

Machine I, Machine II,

$P(A \mid M_1) = 0.95$, $P(A \mid M_2) = 0.85$

$$P(M_1 \mid A) = \frac{P(A \cap M_1)}{P(A)}$$



$$P(A \cap M_1) = P(A \mid M_1)\,P(M_1) = (0.8)(0.95)$$

$$P(A^c \cap M_1) = P(A^c \mid M_1)\,P(M_1)$$

$$P(A \cap M_2) = P(A \mid M_2)\,P(M_2)$$
$$= (0.2)(0.85)$$

$$P(A^c \cap M_2) = P(A^c \mid M_2)\,P(M_2)$$

$$P(M_1 \mid A) = \frac{P(A \cap M_1)}{P(A)} = \frac{(0.8)(0.95)}{(0.8)(0.95) + (0.4)(0.85)}$$

posterior

Example: Binge drinking & Alcohol related accidents

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

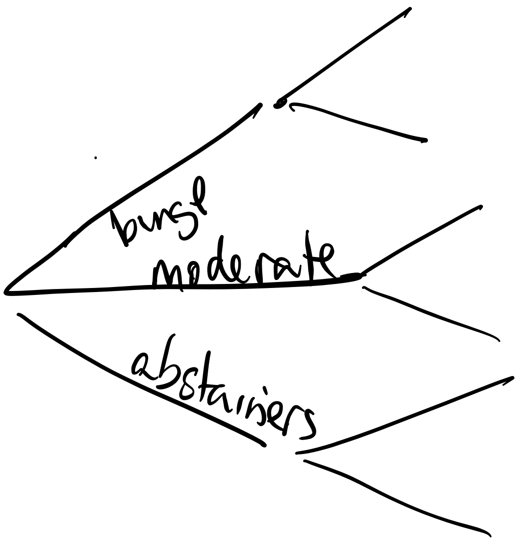
For men, binge drinking is defined as having 5 or more drinks in a row and for women as having 4 or more drinks in a row.
(The difference is because of the average difference in weight.)

According to a study by the Harvard School of Public Health (H.Wechsler, G. W. 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. (*priors*)
- 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. (*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?