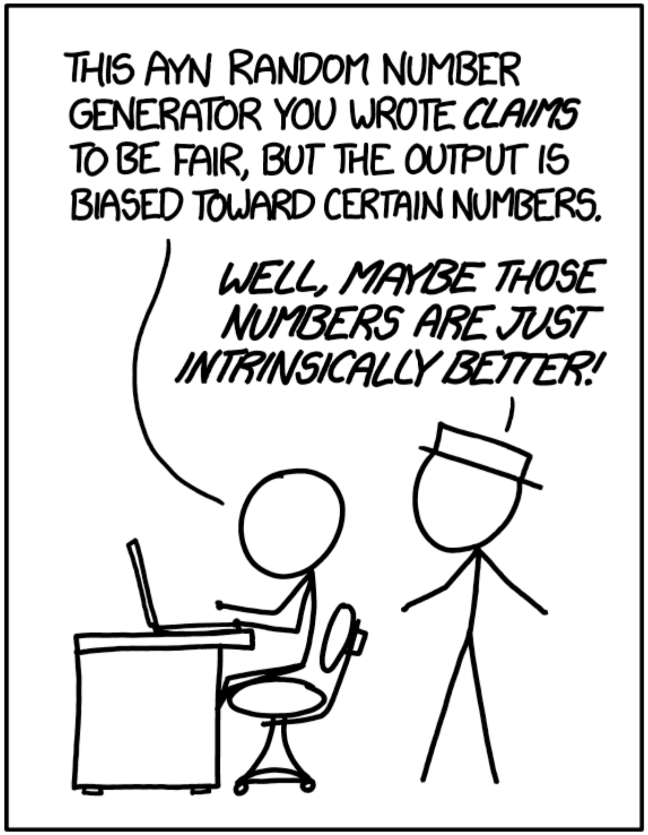
Stat 88: Probability & Math. Stat in Data Science



https://xkcd.com/1277

Lecture 7: 1/31/2024

Finish up base rate fallacy, Random variables & their distributions

Finish chapter 2, and 3.1, 3.2

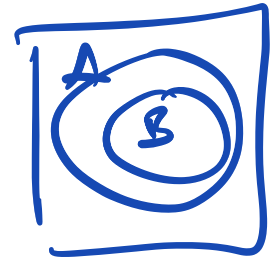
Shobhana Stoyanov

Warm up: Please answer at pollev.com/shobhana

For the questions below, we have that $0 < P(A), P(B) < 1$

1) True or false: If A and B are mutually exclusive, they must be independent.

$P(AB) = 0 \neq P(A) \cdot P(B)$



2) True or false: If $B \subset A$, then A and B must be independent.

$P(AB) \overset{?}{=} P(A)P(B)$ | $AB = B \; (B \subset A)$

$P(AB) = P(B) \neq P(A) \cdot P(B)$ bc $P(A) < 1$

3) True or false: If A and B are independent, then so are their complements.

Exercise

$P(A^c) \cdot P(B^c) = (1 - P(A)) \cdot (1 - P(B))$

$= P(A^c B^c)$

4) If A, B, C are mutually independent events with $P(A) = 0.3, P(B) = 0.8, P(C) = 0.4$, what is the probability that at least one of these events occurs (that is, find $P(A \cup B \cup C)$).

Exercise for Chocolate

3) Need to show $P(A^c \cap B^c) \overset{?}{=} P(A^c)P(B^c)$

Left hand side

$$P(A^c B^c) = P(A^c)\,P(B^c)$$

$$P(A^c \cap B^c) = P((A \cup B)^c) = 1 - P(A \cup B)$$

$$= 1 - [P(A) + P(B) - P(AB)]$$

$$P(A^c \cap B^c) = 1 - P(A) - P(B) + P(AB)$$

$$= 1 - P(A) - P(B) + P(A)P(B)$$

$$= 1 - x - y + xy$$

$$= (1-x) - y(1-x)$$

$$= (1-x)[1-y]$$

$$P(A^c \cap B^c) = (1-P(A))(1-P(B))$$

$$P(A^c \cap B^c) = P(A^c) \cdot P(B^c) \checkmark$$

A^c & B^c are independent.

Exercises Assume A & B are independent

(1) Are A & B^c independent?

(2) What about A^c & B ?

Agenda

- Base rate fallacy
- The Monty Hall Problem
- Random variables and their distributions
- The binomial distribution

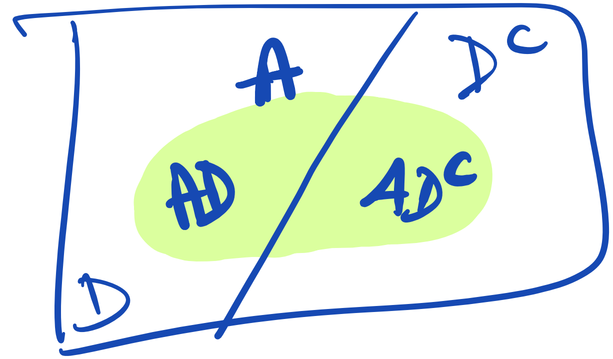
Recall Harvard study of physicians

- 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?" $P(D) = 0.1\%$
- *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
- Note that the test is assumed to be accurate in the sense that it will never give a false negative, so prob of positive test given disease is 100%.
- We saw that P(disease | pos test) ≈ 2% (≈ 1/51) $\longrightarrow P(D \mid +) = ??$

D +

$P(+ \mid D) = 100\%$

Base Rate Fallacy



- $P(D\,|\,\text{pos. test})$ or ***posterior*** probability =
- Recall that ***prior*** probability of disease = 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 $P(D\,|\,+)$ 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?

$$P(D|+) = \frac{P(D \,\&\, +)}{P(+)} = \frac{P(+|D)\cdot P(D)}{P(+\&D) + P(+\&D^c)}$$

$$= P(+|D)P(D)$$

$$P(+|D)P(D) + P(+|D^c)P(D^c)$$

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 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?

$\boxed{A}$ $\boxed{B}$ $\boxed{C}$

<u>Exercise</u> Show this using Bayes' thm

using A, B, C are events of car behind doors A, B, C

$P(A) = P(B) = P(C) = 1/3$

Let D be event that host opens door B to show you a goat. What is $P(A|D)$?

$$P(A|D) = \frac{P(AD)}{P(D)} = \frac{P(D|A)P(A)}{P(AD) + P(BD) + P(CD)}$$