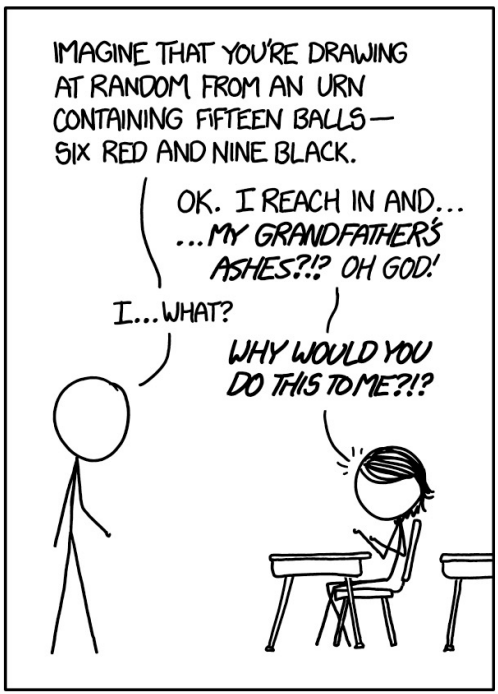
Stat 88: Probability and Statistics in Data Science



https://xkcd.com/1374/

Lecture 6: 2/3/2022

2.4, Random variables, distributions, and a special distribution

Sections 2.4, 3.1, 3.2, 3.3

Shobhana M. Stoyanov

Agenda

- 2.4: Use and interpretation of Bayes' rule
  - Disease, prevalence, base rate, base rate fallacy

- 3.1, 3.2, 3.3, 3.4
- Binary outcomes: success and failure
- Random variables
- The binomial distribution

2.4: Use and interpretation of Bayes' rule

"+" denotes pos test
− denotes neg test
D: event of having disease

- 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. $P(D) = 0.1\%$
- *False positive rate*: fraction of positive results among people who don't have the disease $5\% = P$
- *Positive result*: test is positive

$P(D \mid +)$

95% of the time it gives a correct negative result

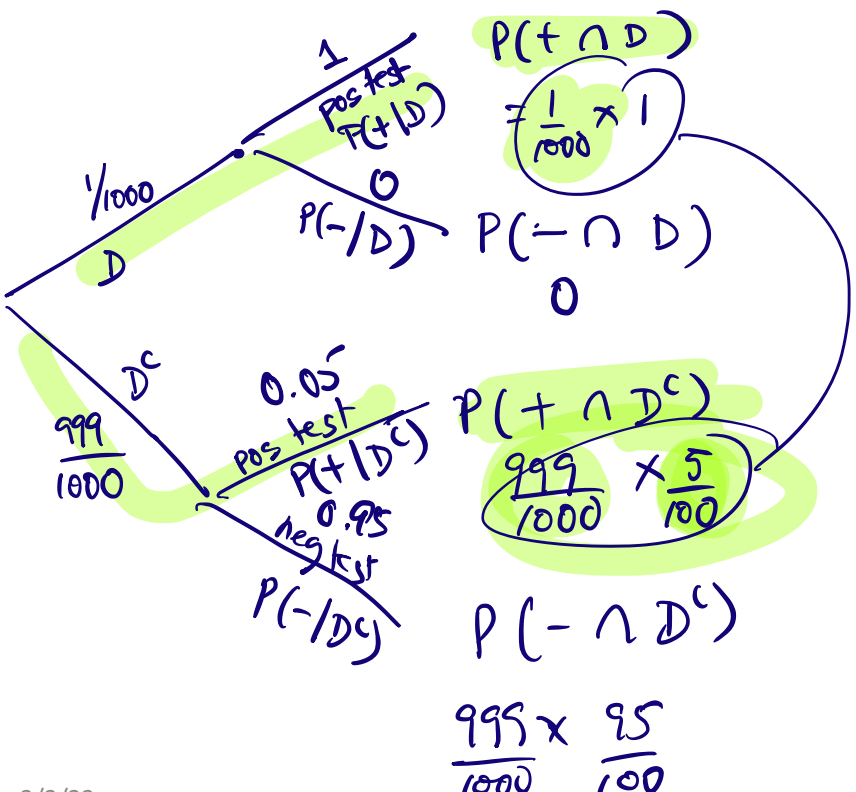
- What is your guess – without any computations?

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

Specificity $95\% = P(- \mid D^c)$, $P(+ \mid D^c) = 5\%$

Tree diagram for disease and positive test

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



$$P(D|+) = \frac{P(D \cap +)}{P(+)} = \frac{1/1000}{\frac{1}{1000} + \frac{5 \times 999}{100 \times 1000}} \approx 0.0196 \approx 2\%$$

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?

If base rate increases from 0.1% to 10%,
Posterior prob ~ 69%

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)

$P(A \cap B) \neq P(A)P(B)$

$P(A \cap B) = P(B|A)P(A)$

Back to counting outcomes of tosses

- Toss a coin 8 times, how many possible outcomes? $2^8 = 256$
- What is the chance of ***all*** heads? $\frac{1}{256}$ $P(\text{not all H}) = 1 - \frac{1}{256} = \frac{255}{256}$
- If each of the students in this class present today flip a coin 8 times, what is the chance that at least 1 person gets all heads?

$P(\text{2 people don't get all heads}) = \left(\frac{255}{256}\right)^2$

$P(\text{n people don't get all heads}) = \left(\frac{255}{256}\right)^n$

$P(\text{at least 1 of n people gets all H}) = 1 - \left(\frac{255}{256}\right)^n$

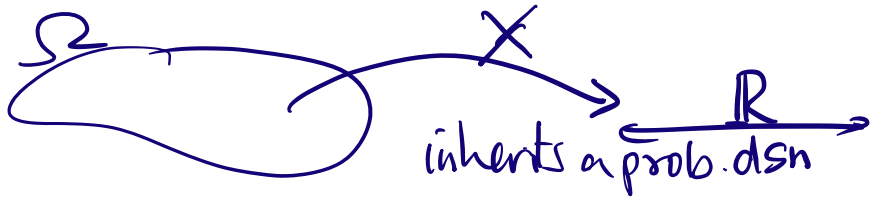
Section 3.1: Vocabulary

- 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) of outcomes ***Success***, and ***Failure***

- Might be with replacement (like a coin toss) or without replacement (drawing voters from a city and checking number of mask mandate supporters)

- Read about Paul the octopus and Mani the parakeet and their soccer predictions
- Note that Paul made 8 correct 2010 WC predictions. What is the chance of 8 correct if picking completely at random? (like tossing a coin and getting all heads)

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.

$$X = \begin{cases} 0 \text{ w.p } \frac{1}{4} \\ 1 \text{ w.p } \frac{1}{2} \\ 2 \text{ w.p } \frac{1}{4} \end{cases}$$

$\Omega = \{HH, HT, TH, TT\}$

$X =$ # of H in 2 tosses

$X(HH) = 2$

$X(HT) = X(TH) = 1$

$X(TT) = 0$

Random variables

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

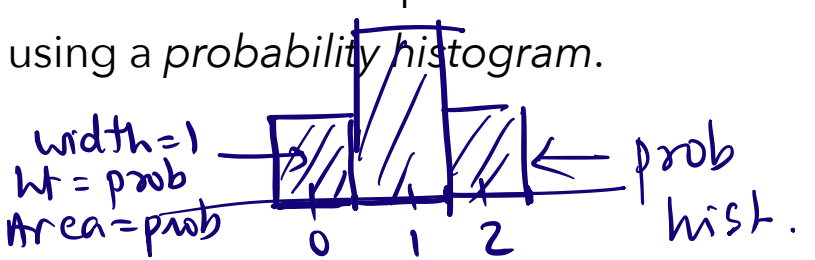
- It is a way to ***map*** the function space Ω to real numbers
- For example: Let X represent the number of heads in 3 tosses.

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

w.p. = with probability

- 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*.

$$X = \begin{cases} 0 & \text{w.p. } 1/4 \\ 1 & \text{w.p. } 1/2 \\ 2 & \text{w.p. } 1/4 \end{cases}$$



Random variables, distribution table & histogram

- For example: Let X represent the **number of heads in 3 tosses**.
- We can write down the ***distribution*** of X, which consists of the possible values of X and the probabilities of X taking these values & make a histogram:

Outcome ω	$x = X(\omega)$	$f(x) = P(X = x)$
HHH	3	1/8
HHT	2	
HTH	2	3/8
THH	2	
TTH	1	
THT	1	3/8
HTT	1	
TTT	0	1/8

$\Omega = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$

$$f(x) = \begin{cases} \frac{1}{8}, & x = 0, 3 \\ \frac{3}{8}, & x = 1, 2 \\ 0, & \text{otherwise} \end{cases}$$

- The function describing the distribution is called the ***probability mass function*** $f(x)$, where $\boldsymbol{f(x) = P(X = x)}$

$f(x) = P(X = x)$, x is a real #

$P(X \geq 1) = \frac{7}{8}$

$= 1 - P(X < 1)$

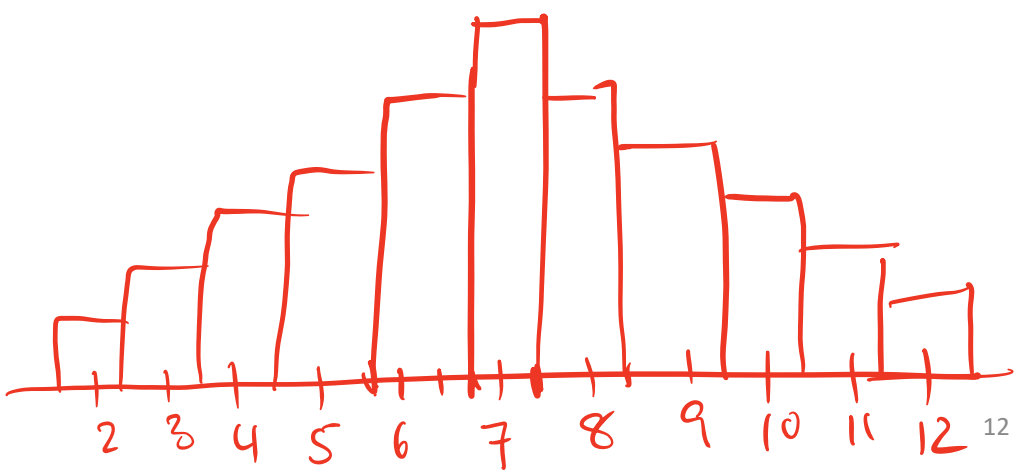
$= 1 - P(X=0) = 1 - \frac{1}{8} = \frac{7}{8}$

Another example

- Let X be the **sum of spots** when a pair of dice is rolled.
- Write down the probability distribution table of X :

x	2	3	4	5	6	7	8	9	10	11	12
$f(x)$	$1/36$	$2/36$	$3/36$	$4/36$	$5/36$	$6/36$	$5/36$	$4/36$	$3/36$	$2/36$	$1/36$

- Probability histogram:



Random Variables

X, Y, Z etc to denote 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.

$X: x = 0 \quad 1 \quad 2$

$f_X(x) = \frac{1}{4} \quad \frac{1}{2} \quad \frac{1}{4}$

$Y: y = 0 \quad 1 \quad 2$

$f_Y(y) = \frac{1}{4} \quad \frac{1}{2} \quad \frac{1}{4}$

but $Y = 2 - X$

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 (with1 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 $\boldsymbol{X \sim Bin(n, p)}$ if X takes values $0, 1, \ldots, n$ and

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

Characteristics of the binomial distribution

- There are n trials, where n is FIXED beforehand.
- The chance (p) of a success stays the SAME from trial to trial
- Each trial results in either success (S) or failure (F)
- The trials are INDEPENDENT of each other.
- $X \sim Bin(n, p)$, possible values of X: $0, 1, 2, \ldots, n$
- Use python to compute numerical values of probabilities (read section in text, in 3.3)

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.