

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

- 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?

Tree diagram for disease and positive test

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

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?

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)

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 8 times, what is the chance that at least 1 person gets all heads?

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.

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.
- 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, 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	$x = X(\omega)$	$f(x) = P(X = x)$

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

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
$f(x)$											

- Probability histogram:

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.

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.
 - $P(\text{success}) = 0.5$
- Need to specify number of trials (n), and $P(\text{success})$ (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 $P(\text{head}) = 1/12$.
- $P(RBBB) =$
- How many such sequences are there?
- What is the probability of all such sequences (with 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.
- $P(\mathbf{RRBB}) =$
- 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 $p = P(\text{red})$

- We say that X has the **Binomial distribution with parameters n and p** , and write it as $\mathbf{X \sim Bin(n, p)}$ if X takes values $0, 1, \dots, n$ 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.
- $X \sim \text{Bin}(n, p)$, possible values of X : $0, 1, 2, \dots, n$
- 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.