Stat 88: Probability & Math. Stat in Data Science

https://xkcd.com/2328

Lecture 10: 2/7/2024

The binomial and hypergeometric distributions

3.3, 3.4

Shobhana Stoyanov

Agenda & warm-up

- Warm up
- The binomial distribution
- The hypergeometric distribution

Warm up

- A quiz has 3 multiple choice questions. Each question has 2 possible answers, one of which is correct. A student answers all the questions by guessing at random. Let X be the number of questions the student gets right, and Y the number that the student gets wrong. What is the distribution of the student's score on the exam, if each correct answer is worth 1 point? Note that this value is X.

- Write down an expression for Y in terms of X, and the distribution of Y. Do X and Y have the same distribution?

Probability histograms

- Draw the probability histogram for X and mark the area for $P(X > 1)$. What is the value of this area?

Recall:

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

- Might be with replacement (like a coin toss) or without replacement (taking a simple random sample of Cal students and checking the number of people who are planning on going to cheer on the rugby team against BYU on Saturday.)

- Random variables (usually denoted by X, Y etc) are numbers that ***map*** the outcome space Ω to real numbers, so they inherit a probability distribution.

- The probability distribution of a random variable X, is a description of the values taken by X, and the probabilities that X takes these values.

- The ***probability mass function*** of X, denoted by $f(x)$, is a function that gives, for each value x taken by X, its chance $P(X = x)$.

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 $X \sim Bin(n, p)$ 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)

Sampling binary outcomes without replacement

- Deck of cards, deal 5, chance of 2 aces in hand? What about chance of 3 hearts in a hand of 5?

- 25 balls, 10 red, 15 blue, pick 5 w/o repl. Chance of 2 red balls?

Hypergeometric Random Variables

- Two kinds of tickets in box, but draws are ***without*** replacement (as opposed to the binomial setting, where the draws are independent).

- What information will we need?

- In this setting of drawing tickets without replacement, let X be the sample sum of tickets drawn from a box with tickets marked 0 and 1. Say that X has the ***hypergeometric*** distribution with parameters _____

$$P(X = g) = \frac{\binom{G}{g}\binom{N-G}{n-g}}{\binom{N}{n}}$$

Example

- A large supermarket chain in Florida occasionally selects employees to receive management training. A group of women there claimed that female employees were passed over for this training in favor of their male colleagues. The company denied this claim. (A similar complaint of gender bias was made about promotions and pay for the 1.6 million women who work or who have worked for Wal-Mart. The Supreme Court heard the case in 2011 and ruled in favor of Wal-Mart.)
- Suppose that the large employee pool of the Florida chain (more than a 1000 people) that can be tapped for management training is half male and half female. Since this program began, none of the 10 employees chosen have been female. What would be the probability of 0 out of 10 selections being female, if there truly was no gender bias?

- Method 1: pretend we are sampling with replacement, use Binomial dsn.

Are we really sampling with replacement?

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.