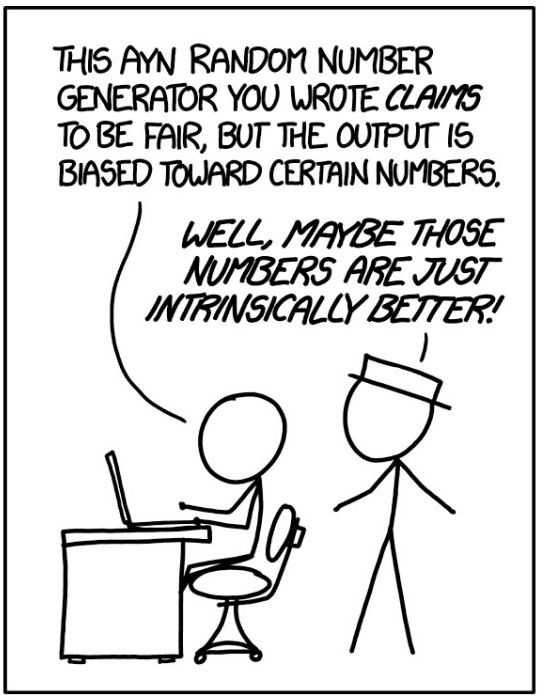
# Stat 88: Probability & Math. Stat in Data Science




https://xkcd.com/1277

Lecture 7: 2/8/2022

Random variables & their distributions + 2 special distributions

Sections 3.2, 3.3, 3.4

# Agenda

- 3.2, 3.3, 3.4
- Random variables
- The binomial distribution
- The hypergeometric distribution

# Recall:

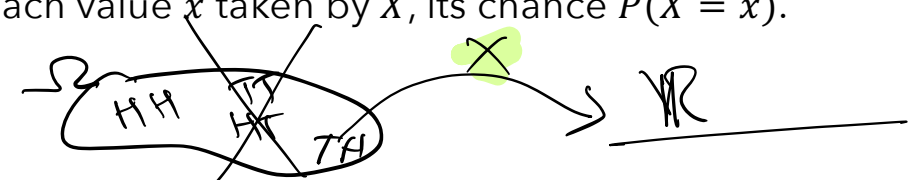
- When we have two kinds of tickets in a box and we draw tickets at random from this box, each draw is called a ***trial***
- We call the two kinds (binary) outcomes ***Success***, and ***Failure***

- Might be with replacement (like a coin toss) or without replacement (taking a simple random sample of residents and checking number of people who watched Ryan Cochran-Siegle win a silver yesterday.)

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

- 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 $P(X = x)$.



# 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 $X$.

Distribution ds Dsn of $X$ :

| $x$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $f(x)$ | 1/8 | 3/8 | 3/8 | 1/8 |

$f(x) = P(X = x)$

Exercise: how would $f(x)$ change if prob of correct answer $= 1/3$

Dsn of $Y$: $Y$ = # wrong

| $y$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $f(y)$ | 1/8 | 3/8 | 3/8 | 1/8 |

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

$$Y = 3 - X$$

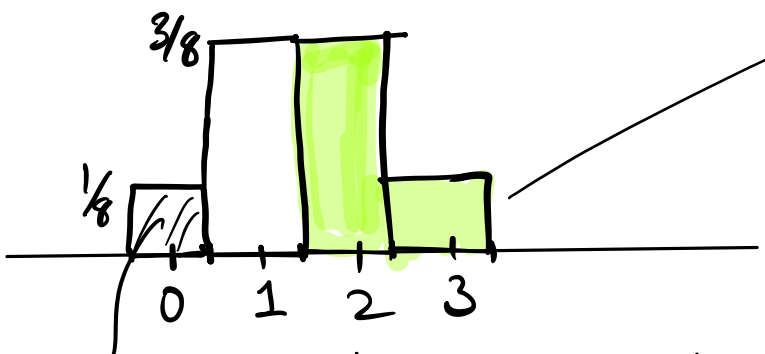
$$\Downarrow$$

$$X + Y = 3$$

$X, Y$ have the same prob dsn, but $X \neq Y$

# Probability histograms

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




$P(X>1) = P(X=2) + P(X=3) = \frac{3}{8} + \frac{1}{8} = \frac{4}{8} = \frac{1}{2}$

$\text{area} = w \cdot h = 1 \cdot \frac{1}{8} = \frac{1}{8} = P(X=0)$

$P(X>1) = 1 - P(X \leq 1)$

$= 1 - [P(X=0) + P(X=1)]$

$= 1 - [\frac{1}{8} + \frac{3}{8}] = \frac{1}{2}$

# 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.

Note P(S) does not have to be the same as P(F)

- 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.

- $P(RBBB) = \left(\frac{1}{12}\right)\left(\frac{11}{12}\right)^3$
- How many such sequences are there? 4
- What is the probability of all such sequences (1 R, 3B)?

$$4 \times \left(\frac{1}{12}\right)\left(\frac{11}{12}\right)^3$$

# 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.
- $P(RRBB) = \left(\frac{1}{12}\right)^2\left(\frac{11}{12}\right)^{12}$ $\binom{4}{2}$
- There are 6 such sequences (how?), so if we let $X$ = # of red balls in 4 draws with replacement, we have that

$$P(X = 2) = \binom{4}{2} \times p^2 \times (1-p)^2$$

where p = P(red) = 1/12

- We say that $X$ has the **Binomial distribution with parameters n and p**, and write it as ***X~Bin(n, p)*** if X takes values 0, 1, …, n and

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

pmf

$$P(X \le M) = \sum_{k=0}^{M} \binom{n}{k} p^k (1-p)^{n-k},$$

CDF

$P(X \leq 2) = P(X=0) + P(X=1) + P(X=2)$

$= \binom{n}{0} p^0 (1-p)^{n-0} + \dots$

## Characteristics of the binomial distribution

- There are $n$ trials, where $n$ is FIXED beforehand.
- The chance 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 $X$: $0, 1, 2, \dots, n$
- Can use python to compute probabilities, on tests can leave a simplified algebraic expression

$F(x) = P(X \leq x)$

# Identifying binomial random variables

Which of the following are binomial random variables?

- Number of heads in 12 tosses of a fair coin. $n=12, p=\frac{1}{2}$
- Number of tosses until we see two heads. Not binomial. number of tosses not fixed
- Number of queens in a five card hand not indep
- Number of Democrats in a simple random sample of 500 adult voters drawn from the SF Bay Area.

No. w/o repl.

# Exercise 3.6.3


- Yi likes to bet on "red" at roulette. Each time she bets, her chance of winning is 18/38, independently of all other times. Suppose she bets repeatedly on red. Find the chance that:

$X$ = # of wins in 10 spins

$X \sim Bin(10, \frac{18}{38})$

a) she wins four of the first 10 bets

$$P(X=4) = \binom{10}{4}\left(\frac{18}{38}\right)^4\left(\frac{20}{38}\right)^6$$

b) she wins at most four of the first 10 bets

$$P(X \le 4) = \sum_{k=0}^{4} P(X=k) = \sum_{k=0}^{4} \binom{10}{k}\left(\frac{18}{38}\right)^k\left(\frac{20}{38}\right)^{10-k}$$

c) the third time she wins is on the 10th bet (spin)

first 9 spins $\binom{9}{2}\left(\frac{18}{38}\right)^2\left(\frac{20}{38}\right)^7$ · $\left(\frac{18}{38}\right)$ ← last spin

W — — — — — — — W

2 W, 7 Losses

d) she needs more than 10 bets to win five times

same as (b) $P(X \le 4)$

# Sampling binary outcomes without replacement

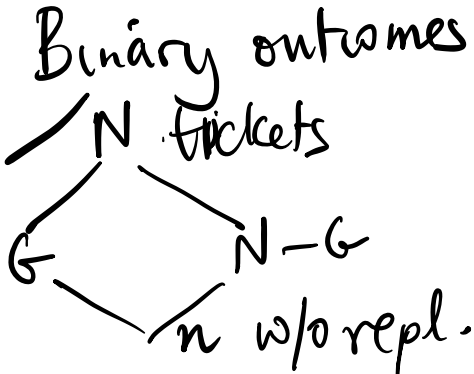
- Deck of cards, deal 5, chance of 2 aces in hand? What about chance of 3 hearts in a hand of 5?

$$P(\text{2 aces in hand of 5}) \qquad \frac{\binom{4}{2}\cdot\binom{48}{3}}{\binom{52}{5}} \qquad N = 52,\ G = 4,\ n = 5$$

$$P(\text{3 hearts in 5}) = \frac{\binom{13}{3}\binom{39}{2}}{\binom{52}{5}}$$

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

$$\frac{\binom{10}{2}\binom{15}{3}}{\binom{25}{5}}$$



$X \sim HG(N, G, n)$

Count the # of "good" tickets in sample of $n$. Call it $X$

## 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? Total # $N$; # of "Good" tkts $G$; # of sample $= n$

- 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?

# Problem solving techniques

- See if problem can be broken into smaller problems
- See which distribution applies to the situation
- Identify the parameters
- Use the addition and multiplication rules carefully

---

An advisor at a university provides guidance to 10 students. Each student has to meet with her once a month during the school year which consists of nine months.

Each month the advisor schedules one day of meetings. Each student has to sign up for one meeting that day. Students have the choice of meeting her in the morning or in the afternoon.

Assume that every month each student, independently of other students and other months, chooses to meet in the afternoon with probability 0.75.

What is the chance that she has both morning and afternoon meetings in all of the months except one?

# Advisors and their students

- Need to figure out a random variable. First fix **one** month, any month.
- Figure out the chance in that month, *all* the students choose the afternoon OR *all* the students choose the morning: this would mean that the meetings happen *only* in the morning OR *only* in the afternoon.
- We need the chance of the complement of this event.
- Where is the random variable?

# Randomized Controlled Experiments

Two randomized controlled experiments are being run independently of each other. In each experiment, a simple random sample of **half** the participants will be assigned to the treatment group and the other half to control. Expt 1 has 100 participants of whom 20 are men. Expt 2 has 90 participants of whom 30 are men.

What is the chance that the treatment and control groups in Experiment 1 contain the same number of men?

# Problems, continued

What is the chance that the treatment groups in the two experiments have the **same** number of men?

- Notice this is a bit tricky. There are many disjoint cases (each of the treatment groups has 1 man, or 2 men or 3 men etc. What is the max?
- We will have to split the chance into the chance of each of the cases and add them.
-

# Did the treatment have an effect?

- RCE with 100 participants, 60 in Treatment, 40 in Control
- T: 50 recover, out of 60 (83%), C: 30 recover out of 40 (75%)

- Suppose treatment had no effect, and these 80 just happened to recover. What is the chance they would have recovered no matter what and 50 were assigned to the treatment group by chance?

# Hypergeometric but don't know N

- A state has several million households, half of which have annual incomes over 50,000 dollars. In a simple random sample of 400 households taken from the state, what is the chance that more than 215 have incomes over 50,000 dollars?

How should we do this? $n = 400, k = 215, G = N/2, N = ???$

# 4.1: Back to random variables and their distributions

- $X, f(x) = P(X = x)$
- Consider $X$ = number of H in 3 tosses, then $X \sim Bin(3,1/2)$
- We can also define a new function $\boldsymbol{F}$, called the ***cumulative distribution function,*** that, for each real number x, tells us how much mass has been accumulated by the time X reaches x.

$$\boldsymbol{F(x) = P(X \leq x) = \sum_{k \leq x} \binom{3}{k} p^k (1-p)^{n-k}}$$

| $x$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $f(x) = P(X = x)$ | 1/8 | 3/8 | 3/8 | 1/8 |
| $F(x) = P(X \leq x)$ | | | | |

## $F(x) \longrightarrow f(x)$?

- How to recover the pmf from the cdf? Draw the graph of $F(x)$:

- What are the properties of $F(x)$? What is its domain? Range?

# Exercise 4.5.2

- A random variable $W$ has the distribution shown in the table below. Sketch a graph of the cdf of $W$.

| $w$ | -2 | -1 | 0 | 1 | 3 |
|---|---|---|---|---|---|
| $P(W = w)$ | 0.1 | 0.3 | 0.25 | 0.2 | 0.15 |