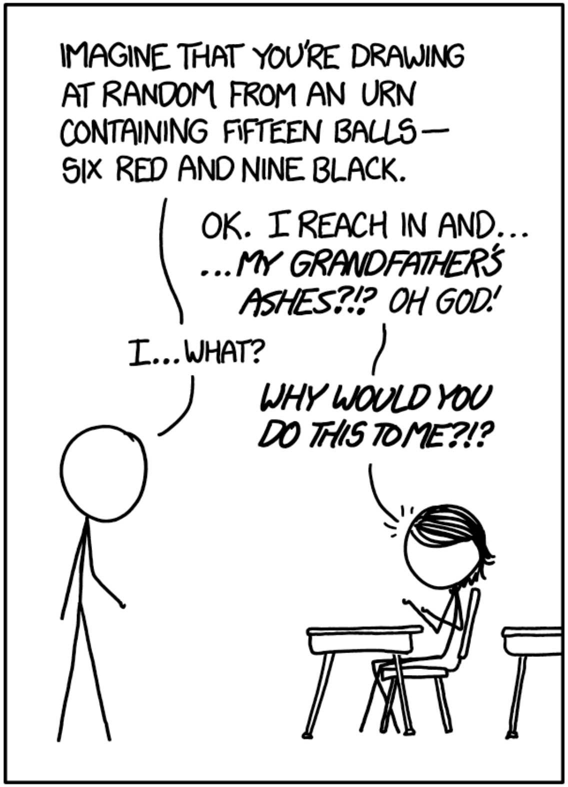
Stat 88: Probability & Math. Stat in Data Science



https://xkcd.com/1374/

Lecture 8: 2/5/2021

The hypergeometric distribution, examples, CDF

Sections 3.4, 3.5, 4.1

Agenda

- Finish up 3.3, Exercise 3.6.3
- 3.4: Sampling without replacement and the hypergeometric distribution
- 3.5: Examples of random variables
- 4.1 The cumulative distribution function

Identifying binomial random variables

Which of the following are binomial random variables?

A • Number of heads in 12 tosses of a fair coin. $Bin(12, \frac{1}{2})$

B • Number of tosses until we see two heads. Not bin b/c n not fixed

C • Number of queens in a five card hand not bin b/c not indep (p changes from draw to draw)

D • Number of Democrats in a simple random sample of 500 adult voters drawn from the SF Bay Area.

① Not bin b/c SRS (w/o repl.)

② Can model as bin b/c sample size much much smaller than population size.

Recall : $\boxed{X \sim Bin(n, p)}$, $P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$

$\binom{n}{k} = \frac{n!}{(n-k)!\,k!}$

$(n-k)!\,k!$

Exercise 3.6.3

1 Trial = 1 spin of the wheel.

$n = 10$, $p = P(S) = P(\text{Red}) = \frac{18}{38}$

- 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 successes in n trials.

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

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

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

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

$P(10^{th}\ \text{bet} = S)$

$$\underbrace{\_\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ \_}_{\text{First 9 bets have 2 S, 7 F}}\ \underline{S} = \binom{9}{2}\left(\frac{18}{38}\right)^2\left(\frac{20}{38}\right)^7\cdot\left(\frac{18}{38}\right)$$

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

$$P(X \le 4)$$

Counting permutations & combinations

- Recall # of ways to rearrange n things, taking them 1 at a time is n!
- If we have only $k \le n$ spots to fill, then $n \cdot (n-1) \cdot \ldots \cdot (n-(k-1))$
- # of perm. of n things taken k at a time.
- Example: How many 3 letter words from P A T I O

$n = 5$

# of permutations of n things taken k at a time

$60 = \underline{5} \cdot \underline{4} \cdot \underline{3} = \frac{5!}{2!}$ ← w/o replacement

w/repl. $\underline{5} \cdot \underline{5} \cdot \underline{5}$

PAT PIT POT APT AIT AOT TOP OPT TAP PTA TPA ATP

- If we don't care about order, then we are counting subsets, and this number is denoted by $\binom{n}{k}$, which we get by dividing: $n \cdot (n-1) \cdot \ldots \cdot (n-(k-1))$ by $k!$

$\left(\frac{n!}{(n-k)!}\right)\frac{1}{k!} = \frac{5 \cdot 4 \cdot 3}{3!} = \binom{5}{3}$

# of arrangements of 3 letters $\underline{3} \cdot \underline{2} \cdot \underline{1}$

- Example: How many 3 letter subsets from P A T I O

$\frac{60}{3!} = 10$

$\binom{5}{3} = \left(\frac{5!}{2!}\right)\frac{1}{3!} = 10$

- Note: $\binom{n}{n} = 1$, $\binom{n}{0} = 1$

Sampling binary outcomes without replacement

$n = 52$ 4 Aces, 48 non-Aces.

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

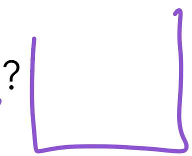
$$\frac{\binom{4}{2} \cdot \binom{48}{3}}{\binom{52}{5}}$$

\# of ways to get 2 aces

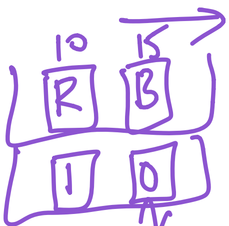
\# of ways to get the rest of 3 cards

\# of $\#(\Omega) = \binom{52}{5}$

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



25 balls
10 red
15 blue



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

X = # of R balls in sample of 5

$P(RRBBB)$

$$\left(\frac{10}{25} \; \frac{9}{24} \; \frac{15}{23} \; \frac{14}{22} \; \frac{13}{21}\right) \times \text{\# of sequences w/ 2R, 3B}$$

$P(R_1)P(R_2 | R_1)P(B_3 | R_1 R_2) \cdot \cdots$

Hypergeometric Random Variables

G N-G

1 0 → n

- **Two** kinds of tickets in box, but draws are ***without*** replacement (as opposed to the binomial setting, where the draws are independent). This situation is more common, in which we sample from a population ***without replacement,***

Total # of tickets in box N

# of tickets we are interested in : G

- What information will we need?

N, G, 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 N, G, n

# of successes in box

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

← # of F. in box

, g = 0, 1, 2, ... n

In example 1

N=52, G=4, N-G=48

Ex 3 N=25, G=10

n=5

X ~ HG(N, G, 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?

$N=1000$, $G=500$, $N-G=500$

$n=10$

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

$$X \sim Bin\left(n=10, p=\tfrac{1}{2}\right)$$

$$P(X=0) = \underbrace{\binom{10}{0}}_{=1} \underbrace{p^0}_{=1} (1-p)^{10} = (1-p)^{10} = \left(\tfrac{1}{2}\right)^{10}$$

$$= \frac{1}{1024} \approx 0.00097$$

Are we really sampling with replacement? No.

$N = 1000, \quad G = 500, \quad N - G = 500$

$X \sim HG(1000, 500, n = 10)$

$$P(X=0) = \frac{\binom{500}{0}\binom{500}{10}}{\binom{1000}{10}}$$