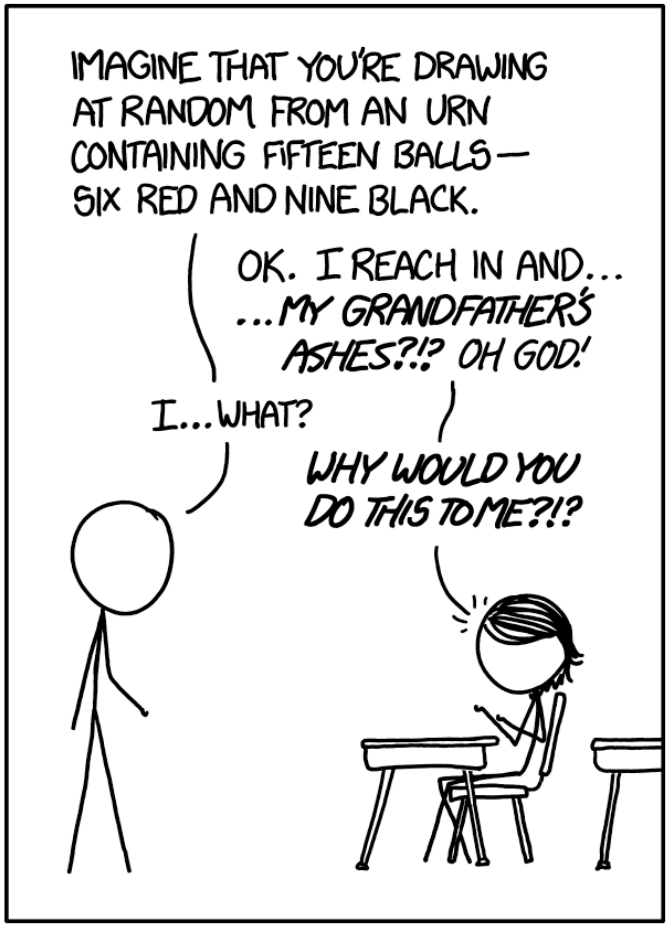
Stat 88: Probability & Math. Stat in Data Science



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?

- 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.

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:

a) she wins four of the first 10 bets

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

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

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

Counting permutations & combinations

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

- 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: $n \cdot (n-1) \cdot \ldots \cdot \big(n-(k-1)\big)$ by $k!$

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

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

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). This situation is more common, in which we sample from a population ***without replacement,***

- 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?

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

- 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?

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 ~ Bin(3,1/2)
- We can also define a new function 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.
- $F(x) = P(X \leq x) = \sum_{k \leq x} \binom{3}{k} p^k (1-p)^{n-k}$.
- Write table for F, draw the graph.