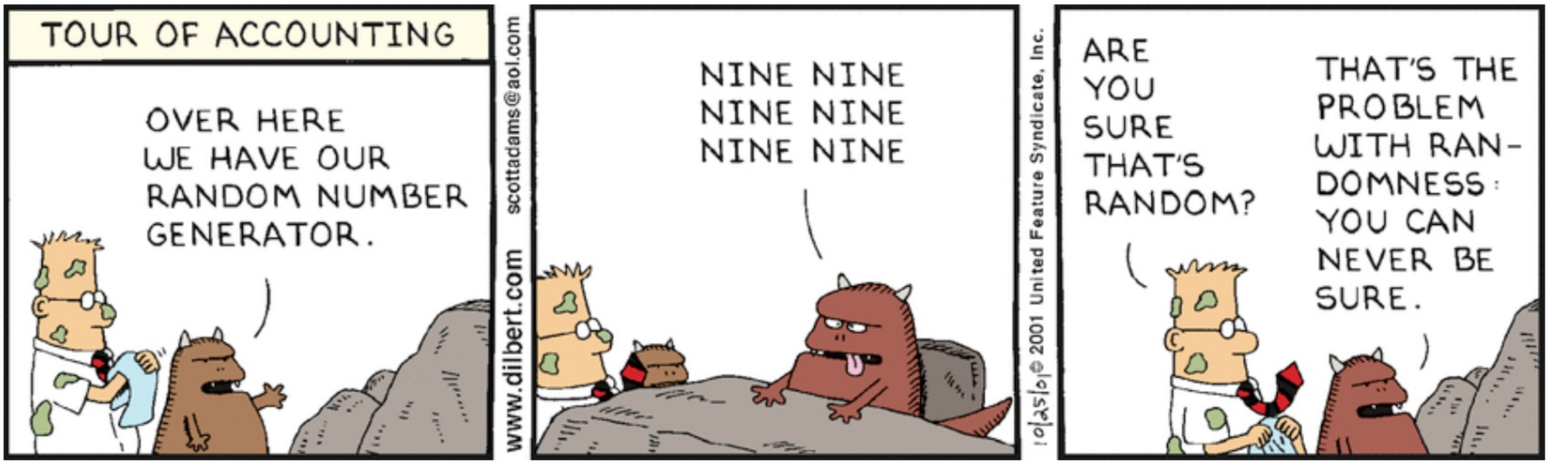
Stat 88: Probability & Math. Stat in Data Science



Lecture 8: 2/10/2022

Examples of computations2/9/22, CDF, waiting times

Sections 3.5, 4.1, 4.2

Shobhana M. Stoyanov

Agenda

- 3.5: Examples
- 4.1 The cumulative distribution function
- 4.2 Waiting times

Recap

- Last time, talked about classifying and counting.
- n **independent** trials each of which can result in one of two outcomes.
- We call these outcomes **S**uccess or **F**ailure, and can represent the random experiment by drawing n tickets with replacement from a box with tickets marked 0 or 1, where the proportion of tickets marked 1 is equal to the probability of a success in a trial.
- If X is the number of successes in n trials, then X is the ***sum of draws*** from such a box as described above.
- We say that $\boldsymbol{X \sim Bin(n, p)}$ and $\boldsymbol{P(X = k) = \binom{n}{k} \times p^k \times (1-p)^{n-k}, k = 0, 1, \dots n}$
- We might also draw ***without*** replacement, in which case, we say that X has the *hypergeometric*$(\boldsymbol{N, G, n})$ distribution, and

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

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?

$N = 1000$

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

Let X = # of women selected, if we assume
1. No gender bias
2. Employee picked at random
3. with replacement

$\sim$ "distributed as"

$X \sim Bin(10, \frac{1}{2})$

$$P(X=0) = \binom{10}{0}\left(\frac{1}{2}\right)^{10} = \frac{1}{1024} \approx 0.00097$$

Are we really sampling with replacement? No

$N = 1000$ (total # of employees)

$G = 500$

$N - G = 500$

$n = 10$

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

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

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?

For any non empty A, B cannot be BOTH Mut. exc & Indep

Advisors and their students

A: all students choose AM

B: " " " PM

- 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.
- What is the random variable?

P(a student choosing AM) $= \frac{1}{4}$

$P(A) = \left(\frac{1}{4}\right)^{10}$ $P(B) = \left(\frac{3}{4}\right)^{10}$

P(all meetings in AM OR all meetings in PM)

$= P(A \cup B) = \left(\frac{1}{4}\right)^{10} + \left(\frac{3}{4}\right)^{10}$

P(at least 1 meeting in AM & 1 meeting in PM)

$$P(\underbrace{\text{at least 1 meeting in AM}}_{B^c} \;\&\; \underbrace{\text{1 meeting in PM}}_{A^c})$$

$$P(A^c \cap B^c) = P\left((A \cup B)^c\right) = 1 - P(A \cup B) = 1 - \left(\frac{1}{4}\right)^{10} - \left(\frac{3}{4}\right)^{10}$$

De Morgan's Law

$\underbrace{P[(A \cup B)^c]}$ = Prob of at least 1 AM meeting & at least 1 PM meeting in any specific month

$$\boxed{1 - \left(\frac{1}{4}\right)^{10} - \left(\frac{3}{4}\right)^{10} = p}$$

Let X = # of times she has meeting in both the AM & PM in the academic year

$X \sim Bin(9, p)$

$$P(X=8) = \binom{9}{8} p^8 (1-p)^{9-8}$$

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?

$$20 \xrightarrow{10} T_1 \qquad 20 \xrightarrow{10} C_1$$

T_1: # of men in trt group of Expt 1

C_1: # of men in control " " "

$$P(T_1 = 10) = \frac{\binom{20}{10}\binom{80}{40}}{\binom{100}{50}}$$

$$T_1 \sim HG(\overset{N}{100}, \overset{G}{20}, \overset{n}{50})$$

$N = 100$
$n = 50$
$G = 20$

Exercise: (1) Make sure you could compute this if you needed to.

(2) Pretend T_1 is binomial. Find the

prob.

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.
-

Treatment = Trt

T_1 = # of men in trt gp of Expt 1

T_2 = # " " " Expt 2

$$P(T_1 = T_2) = P(T_1 = T_2 = 0) + P(T_1 = T_2 = 1) + P(T_1 = T_2 = 2) + \dots$$

$$= \sum_{k=0}^{20} P(T_1 = T_2 = k)$$

$$= \sum_{k=0}^{20} P(T_1 = k \ \& \ T_2 = k) = \sum_{k=0}^{20} P(T_1 = k) P(T_2 = k)$$

Exercise! Finish this

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?

Q: What is the chance that 80 recovered no matter what & 50 of these 80 were assigned to the trt gp.?

$N=100, \; G=80, \; n=60$

X = # that recover in Trt

$$P(X=50) = \frac{\binom{80}{50}\binom{20}{10}}{\binom{100}{60}}$$

$P(X \geq 50)$

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? $n = 400, k = 215, G = N/2, N = ???$

Since $n << N$, pretend X is binomial

n is much muchmuch smaller than N

where X = # of households in sample w/ incomes over \$50K.

$X \sim Bin(n, p)$, $n = 400$, $p = 0.5$

$P(X > 215)$

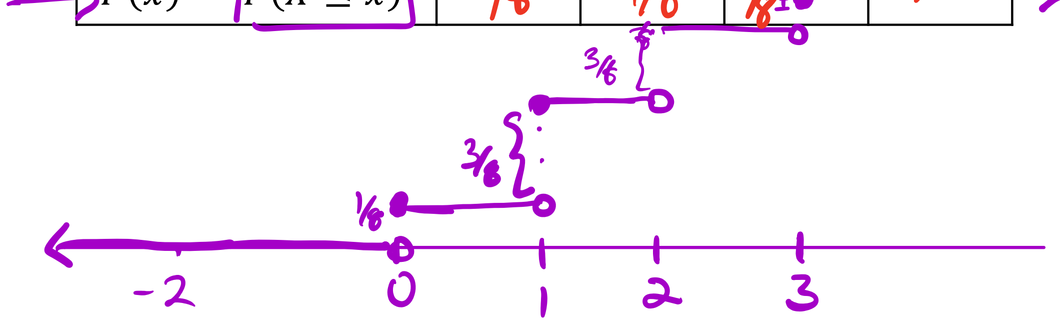
4.1: Back to random variables and their distributions

- $X, f(x) = P(X = x)$
- Consider X = number of H in 3 tosses, then $X \sim Bin(3, \frac{1}{2})$
- 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
$F(x) = \boxed{P(X \leq x)}$	1/8	4/8	7/8	8/8

← prob mass fn



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

- 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
$F(w)$					

4.2: Waiting times

- Say Ali keeps playing roulette, and betting on red each time. The waiting time of a red win is the number of spins until they see a red (so the number of spins until and including the time the ball lands on a red pocket).

What is the probability that Ali will wait for 4 spins before their first win? (That is, the first time the ball lands in red is the 4th spin or trial)

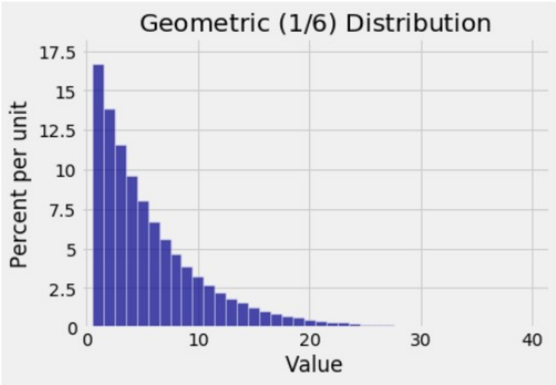
- Say we have a sequence of ***independent*** trials (roulette spins, coin tosses, die rolls etc) each of which has outcomes of success or failure, and $\boldsymbol{P(S)} = \boldsymbol{p}$ on each trial.
- Let T_1 be the number of trials up to and including the first success. Then T_1 is the ***waiting time until the first success***.
- What are the values T_1 takes? What is its pmf $f(x)$?

Geometric distribution

- Say T_1 has the **geometric distribution**, denoted $\boldsymbol{T_1 \sim Geom(p)}$ on {1, 2, 3,…}

- $f(k) = P(T_1 = k) =$

- Check that it sums to 1. What is the cdf for this distribution? Can you think of an easy way to write down the cdf?



Waiting time until r^{th} success

- Say we roll a 8 sided die.
- What is the chance that the ***first*** time we roll an eight is on the 11^{th} try?

- What is the chance that it takes us 15 times until the 4^{th} time we roll eight? (That is, the waiting time until the 4^{th} time we roll an eight is 15)

- What is the chance that we need **more** than 15 rolls to roll an eight 4 times? (Like the last part of roulette problem from Tuesday's lecture)