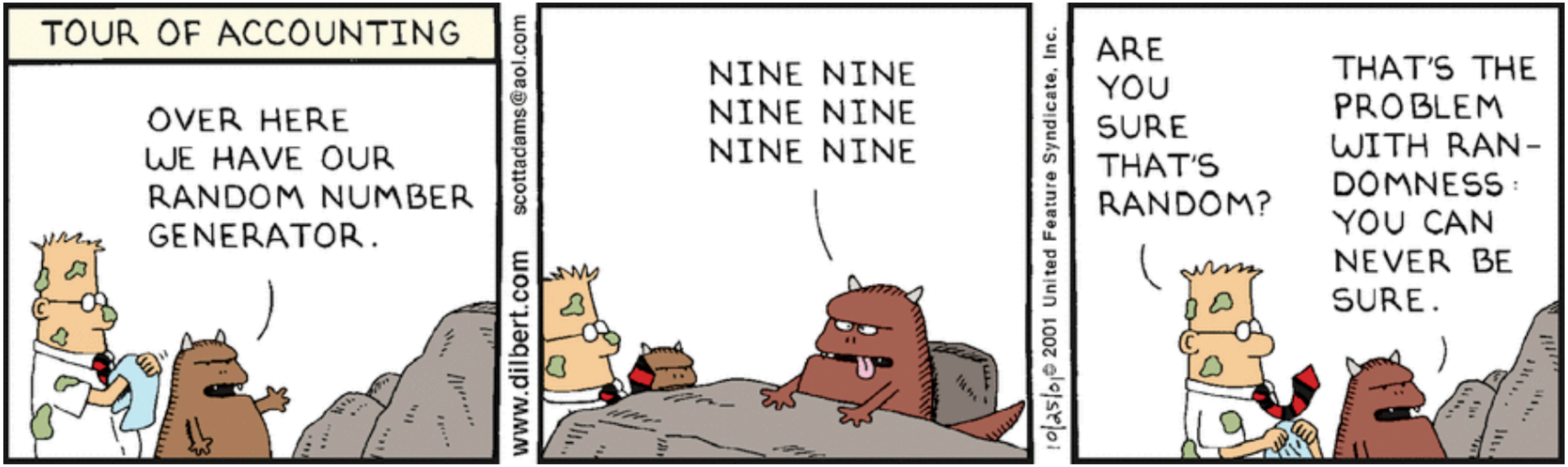
Stat 88: Probability & Math. Stat in Data Science


TOUR OF ACCOUNTING
OVER HERE WE HAVE OUR RANDOM NUMBER GENERATOR.
scottadams@aol.com
www.dilbert.com
NINE NINE NINE NINE NINE NINE
10/25/01 © 2001 United Feature Syndicate, Inc.
ARE YOU SURE THAT'S RANDOM?
THAT'S THE PROBLEM WITH RANDOMNESS: YOU CAN NEVER BE SURE.

Lecture 12: 2/12/2024

Examples, cdf, waiting times

3.5, 4.1, 4.2

Shobhana Stoyanov

Warm up: apples and mangoes

- Suppose we have a box with 4 mangoes, 3 oranges, and 3 apples and you draw out one fruit at a time, ***at random and without replacement***. Let X be the number of draws until you draw your **first mango**, including that last draw. Is it binomial, hypergeometric, or neither? Write down the pmf $\boldsymbol{f}(\boldsymbol{x})$ of X.

$f(x) = P(X=x)$ for all poss values of X

we know that $X = 1, 2, 3, \ldots 7$

B = draw is not M.

A: draw is mango

x	1	2	3	4	5	6	7
$f(x)$	$4/10$	$\frac{4}{9} \cdot \frac{6}{10}$	$\frac{6}{10} \cdot \frac{5}{9} \cdot \frac{4}{8}$	$\frac{6}{10} \cdot \frac{5}{9} \cdot \frac{4}{8} \frac{4}{7}$	$\frac{6}{10} \cdot \frac{5}{9} \cdot \frac{4}{8} \cdot \frac{3}{7} \cdot \frac{4}{6}$		

for X to be 2, 1st draw is $\underline{\text{not mango}}$ & $\underbrace{\text{2nd is mango}}_{A}$

$$P(X=2) = P(BA) = P(A|B)P(B) = \frac{4}{9} \cdot \frac{6}{10}$$

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.

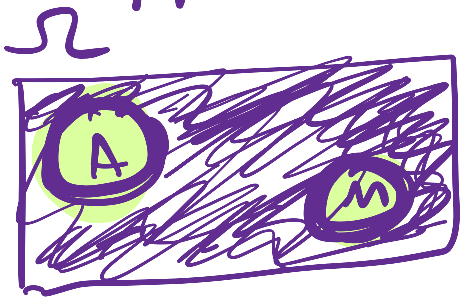
What is the chance that she has ***both*** morning and afternoon meetings in ***all*** of the months except one?

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.
- What is the random variable?

Let A=event that all students choose afternoon

M= " " " morning.



$$P(\Omega \underset{\substack{\uparrow \\ \text{"set minus"}}}{\setminus} (A \cup M)) = P((A \cup M)^c)$$

$$= 1 - (P(A) + P(M))$$

$$= \underbrace{1 - (0.25)^{10} - (0.75)^{10}} = p$$

Prob that in any of the months advisor has meetings BOTH in AM & PM

Let X = # of months out of 9 in which she has meetings in both AM & PM.

We want $P(X=8)$, $X \sim Bin(9,p)$

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

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.
- Exercise for Choc

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?

Exercise for Choc.

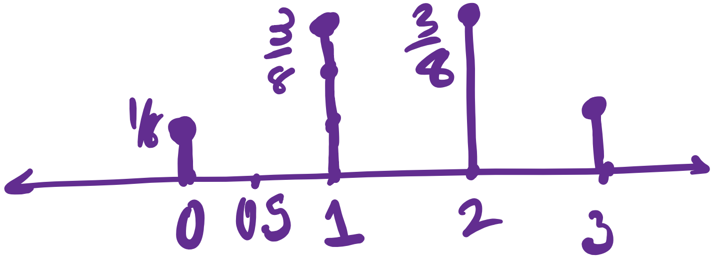
Hypergeometric but don’t know N

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

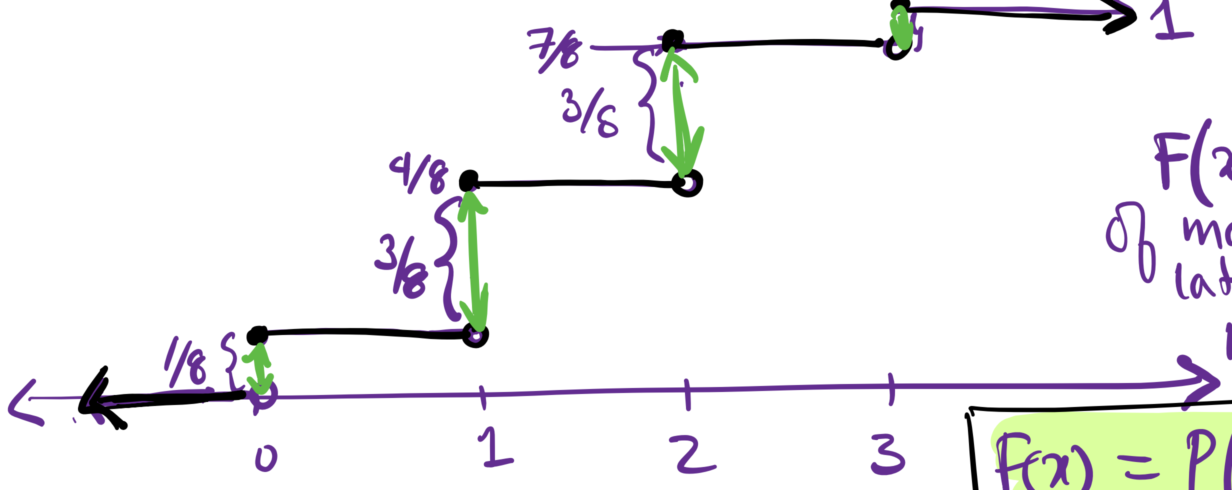
4.1: Back to random variables and their distributions

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



F(x) = amount of mass accumulated upto & including x

$$\boxed{F(x) = P(X \le x)}$$

$F(-27) = 0$

$F(2 \text{ billion } \& 700{,}000) = 1$

$F(+27) = 1$

$F(\text{JB's wealth}) = 1$

F(x) is a step function

Domain $F(x) = \mathbb{R}$

Range $F(x) = [0,1]$

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

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

look at the jumps

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

Exercise 4.5.2

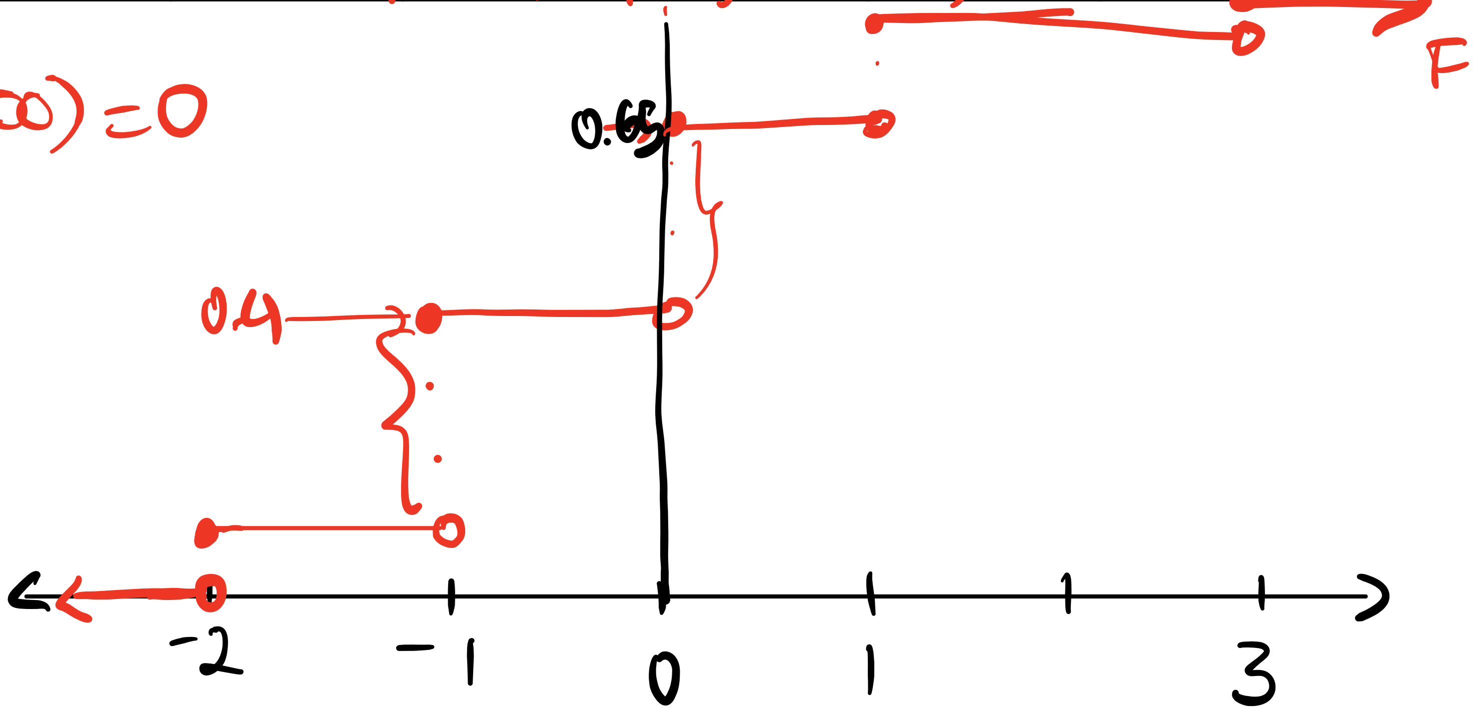
- A random variable W has the distribution shown in the table below. Sketch a graph of the cdf of W.

w	-2	-1	0	1	3
f(w)	0.1	0.3	0.25	0.2	0.15
F(w)	F(-2)=0.1	F(-1)=0.4	F(0)=0.65	F(1)=0.85	F(3)=1

$F(4)=1$

$F(400)=1$

$F(-200)=0$



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)

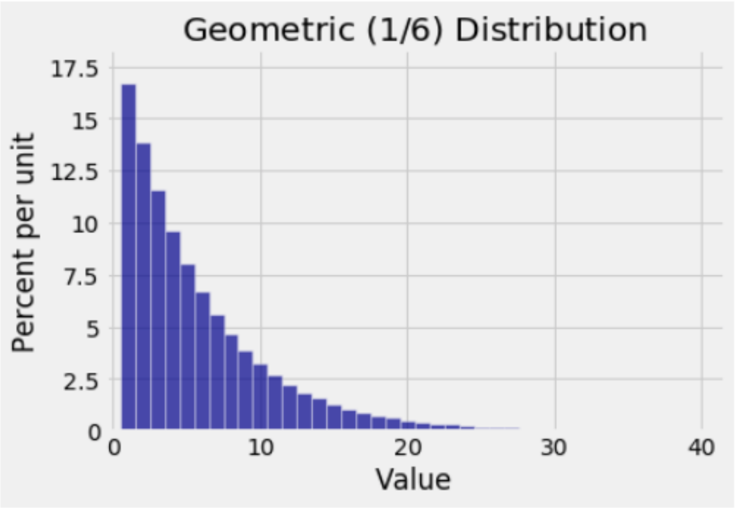
- Say we have a sequence of ***independent*** trials (roulette spins, coin tosses, die rolls etc) each of which has outcomes of success or failure, and $\boldsymbol{P(S)} = \boldsymbol{p}$ on each trial.
- Let T_1 be the number of trials up to and including the first success. Then T_1 is the ***waiting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

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

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

- What is the chance that we need **more** than 15 rolls to roll an eight 4 times?