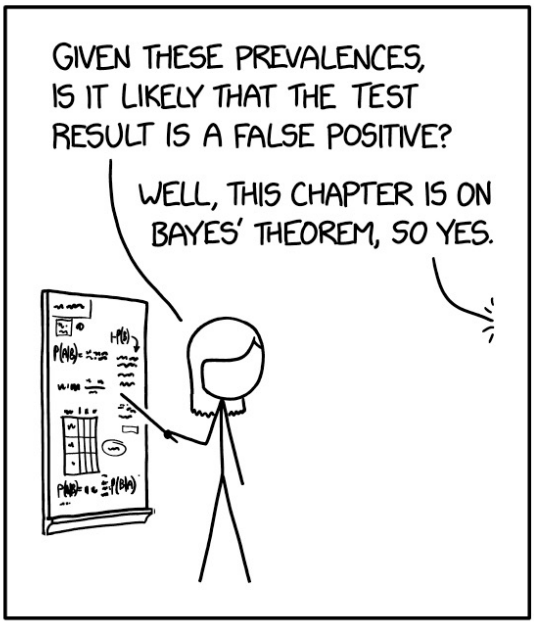
Stat 88: Probability & Math. Stat. in Data Science



https://xkcd.com/2575/

Lecture 9: 2/15/2022

CDF, waiting times

Sections 4.1,4.2, 4.3

Agenda

- Warm up with a cdf problem
- 4.2 Waiting times
- 4.3 Exponential approximations
- 4.4 The Poisson distribution

$0 \le f(x) \le 1$, $\sum_{x} f(x) = 1$

x: all possible values X takes

Story so far

- $X, f(x) = P(X = x)$, where X is a random variable and $f(x)$ is its pmf.
- We have talked about the binomial and the hypergeometric distributions
- You can also consider the *discrete uniform* where X is a random variable all of whose possible outcomes are equally likely, so if X has n possible outcomes, each of them has pmf $1/n$.

- Note that these are just special distributions, and there can be many situations in which we just define the random variable, and it is none of these situations. For example, suppose we have a box with 4 mangoes and 3 apples, and you draw out one fruit at a time, without replacement. Let X be the number of draws until you draw your first mango, including that last draw. Write down the pmf $f(x)$ of X (we will go over this after the next slide).

- At the end of lecture on Thursday, we also defined the ***cumulative distribution function*** (cdf) $\boldsymbol{F(x)}$ which totals up all the probability mass for all values up to and including x.

$$f(x) = \begin{cases} \frac{4}{7}, & x = 1 \\ \frac{4}{6}\cdot\frac{3}{7}, & x = 2 \end{cases}$$

$$X = \begin{cases} 1 & \text{w.p. } \frac{4}{7} \\ 2 & \text{w.p. } \frac{4}{6}\cdot\frac{3}{7} \\ 3 & \text{w.p. } \frac{3}{7}\cdot\frac{2}{6}\cdot\frac{4}{5} \\ 4 & \text{w.p. } \frac{3}{7}\cdot\frac{2}{6}\cdot\frac{1}{5}\cdot\frac{4}{4} \end{cases}$$

$P(X=2) = P(\text{first } \boxed{A}, \text{then } \boxed{M})$

$= P(\boxed{A}_1 \,\&\, \boxed{M}_2)$

$= P(\boxed{A}_1)\, P(\boxed{M}_2 \mid \boxed{A}_1)$

x	1	
$f(x)$		

$P(X=3) = P(A_1 A_2 M_3)$

$= P(A_1)\cdot P(A_2 \mid A_1) P(M_3 \mid A_1 \cap A_2)$

$= \frac{3}{7}\cdot\frac{2}{6}\cdot\frac{4}{5}$

Warm up: apples and mangoes

- Suppose we have a box with 4 mangoes 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. Write down the pmf $\boldsymbol{f(x)}$ of X. Why is it neither binomial nor hypergeometric?

Example from last lecture:

- Consider X = number of heads 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
$F(x) = P(X \leq x)$	1/8	4/8	7/8	1=8/8

$F(-250) = 0$

$F(3.000001) = 1$

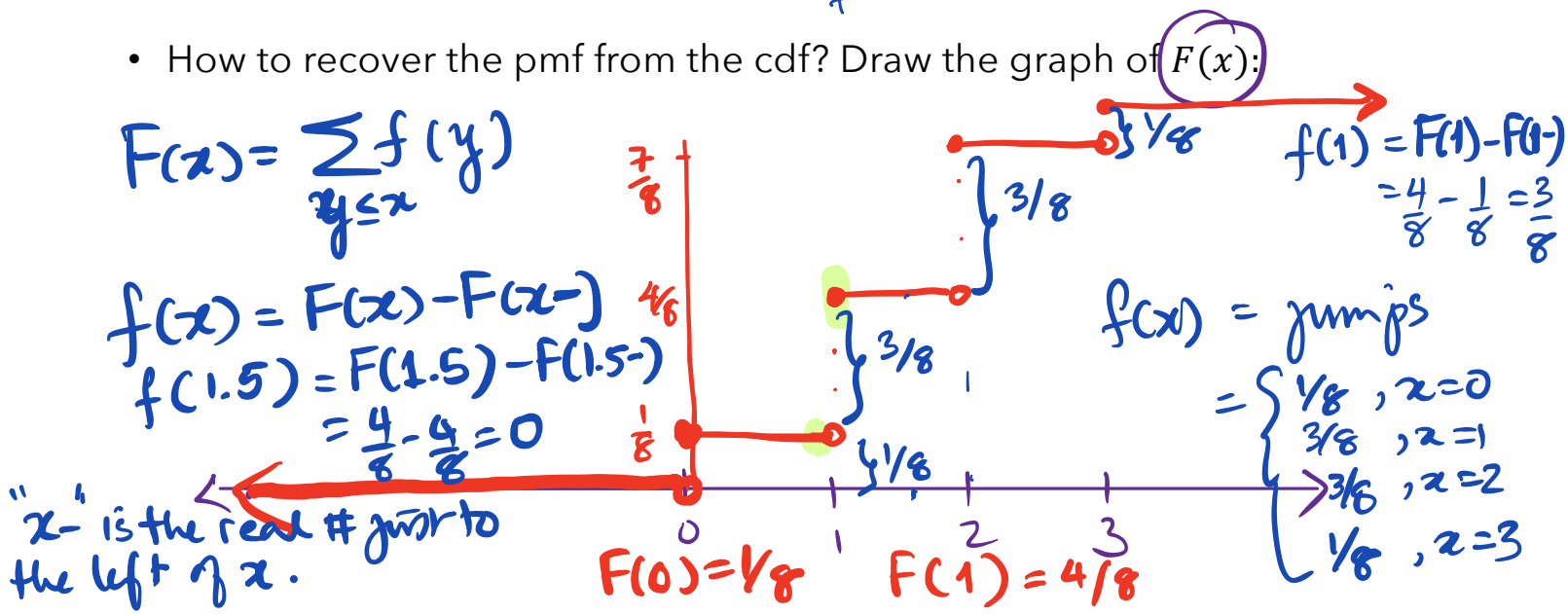
$F(\text{Elon Musk's net worth}) = 1$

$F(4) = 1$

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

$F(x) - F(x-1) = f(x)$

- 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?

Domain $F = (-\infty, \infty) = \mathbb{R}$.

Range $F = [0,1]$

F is a step function (jump discontinuities)

Right continuous. Increasing

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

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

$$f(x) = P(X = x)$$
$$= P(X \leq x) - P(X \leq x - 1)$$
$$= F(x) - F(x - 1)$$

- 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

Exercise

Back to apples and mangoes:

- Suppose we have a box with 4 mangoes and 3 apples, and you draw out one fruit at a time, without replacement. Let X be the number of draws until you draw your first mango, including that last draw. You wrote down the pmf $\boldsymbol{f(x)}$ of X, and now write down the cdf $\boldsymbol{F(x)}$ for $\boldsymbol{X}$.

Exe

4.2: Waiting times

(18 R, 18B, 2G)

$P(R) = 18/38$ $P(\text{not } R) = 20/38$

- 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)

Success: Red
Failure = not Red

$$P(FFFS) = \left(\frac{20}{36}\right)^3\left(\frac{18}{36}\right)$$

- 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 $\boldsymbol{P(S) = p}$ 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 $f(x)$?

$T_1 = 1, 2, 3, 4, 5, \ldots\ldots$ $P(S) = p$, $P(F) = 1-p$

$$f(k) = P(T_1 = k) = P(\underbrace{FFF\ldots F}_{k-1}S) = (1-p)^{k-1}p$$

Geometric distribution

Sum of an infinite geometric series: $b + br + br^2 + br^3 + \cdots = \frac{b}{1-r}$

- Say that T_1 has the **geometric distribution**, denoted $\boldsymbol{T_1 \sim Geom(p)}$ on {1, 2, 3,…}, when we have k-1 failures, and then first success is on k^{th} trial.

$\sum_{k=1}^{\infty} f(k) = 1$??

- $f(k) = P(T_1 = k) = P(FFF \ldots FS) = (1-p)^{k-1} p$ ($k-1$ failures, k^{th} trial is a success)

$\sum_{k=1}^{\infty} f(k) = \sum_{k=1}^{\infty} (1-p)^{k-1} p = p \sum_{k=1}^{\infty} (1-p)^{k-1} = p\left[\frac{1}{1-(1-p)}\right] = 1$

- $F(k) = P(T_1 \le k) = 1 - P(T_1 > k) = 1 - q^k$

$P(T_1 > k) = q^k$ (b/c if first S is after k trials, the first k trials were F) (Let $q = 1-p$)

- 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?

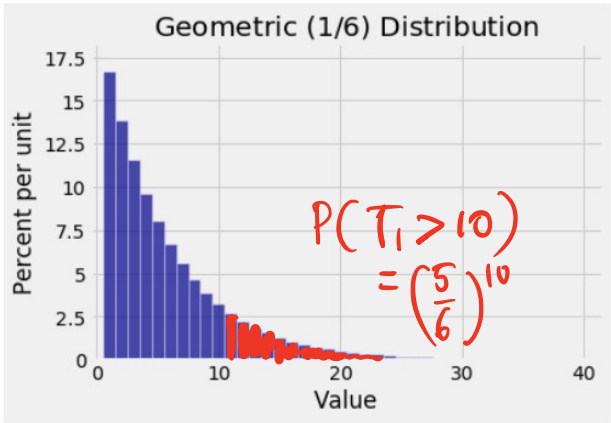
$F(k) = 1 - q^k$, $q = P(F) = 1-p$

Roll a die until ***first*** ace (1 spot):

$P(T_1 = 1) = \frac{1}{6} = 1$

$P(T_1 = 2) = \frac{5}{6} \cdot \frac{1}{6} = \left(\frac{5}{6}\right)^{2-1} \cdot 1$

$P(T_1 = 10) = \left(\frac{5}{6}\right)^9 \cdot \frac{1}{6}$, $P(T_1 > 10) = \left(\frac{5}{6}\right)^{10}$



$P(T_1 > 10) = \left(\frac{5}{6}\right)^{10}$

Waiting time until r^{th} success

$P(8) = \frac{1}{8}$ $P(\text{not } 8) = \frac{7}{8}$

- 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?

$= P(FFFFFFFFFFS) = \left(\frac{7}{8}\right)^{10}\left(\frac{1}{8}\right)$

10 F, 1

- 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)

$= P(\_\_\_\_\_S)$ $1/8$

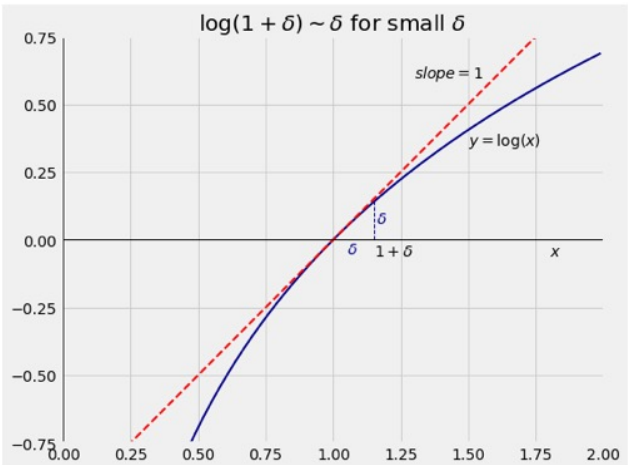
14 rolls, 11 F, 3 S

15^{th} roll & 4^{th} Success.

- What is the chance that we need **more** than 15 rolls to roll an eight 4 times?
- Notice that the ***right-tail*** probability of T_4 is a left hand (cdf) of the Binomial distribution for (15,1/8), and where k=3.

- In general, $P(T_r = k) =$
- And $P(T_r > k) =$

4.3 Exponential Approximations



Very useful approximation: $\log(1+\delta) \approx \delta$, for δ close to 0

How to use this approximation

- Approximate the value of $x = \left(1 - \frac{3}{100}\right)^{100}$

- $x = \left(1 - \frac{2}{1000}\right)^{5000}$

- $x = (1 - p)^n$, for large n and small p

Example

- A book chapter $n = 100{,}000$ words and the chance that a word in the chapter has a typo (independently of all other words) is very small : $p = 1/1{,}000{,}000 = 10^{-6}$.

Give an approximation of the chance the chapter ***doesn't*** have a typo. (Note that a typo is a *rare event*)

Bootstraps and probabilities

- Bootstrap sample: sample of size n drawn with replacement from original sample of n individuals

- Suppose one particular individual in the original sample is called Ali. What is the probability that Ali is chosen ***at least once*** in the bootstrap sample? (Use the complement.)

The Poisson Distribution

- Used to model rare events. X is the number of times a rare event occurs, $X = 0, 1, 2, \ldots$

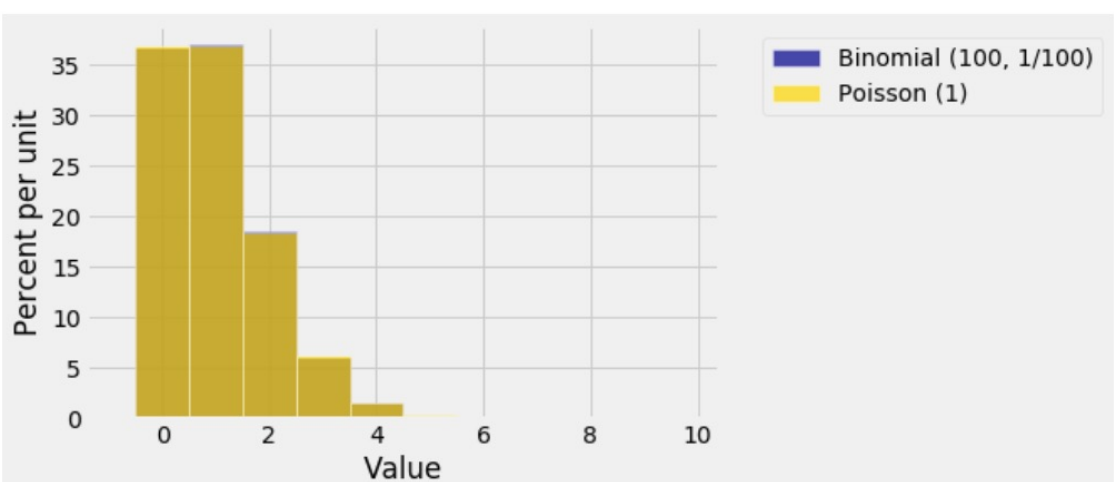
- We say that a random variable X has the ***Poisson*** distribution if

$$\boldsymbol{P(X = k) = e^{-\mu} \frac{\mu^k}{k!}}$$

- The parameter of the distribution is μ

Relationship between Poisson and Binomial distributions

- ***The Law of Small Numbers***: when n is large and p is small, the binomial (n,p) distribution is ***well approximated*** by the Poisson(μ) distribution where $\mu=np$.



Exercise 4.5.7

A book has 20 chapters. In each chapter the number of misprints has the Poisson distribution with parameter 2, independently of the misprints in other chapters.

a) Find the chance that Chapter 1 has more than two misprints.

b) Find the chance that the book has no misprints.

c) Find the chance that two of the chapters have three misprints each.

Sums of independent Poisson random variables

- If X and Y are random variables such that
- X and Y are independent,
- X has the Poisson(μ) distribution, and
- Y has the Poisson(λ)distribution,
- then the sum $S=X+Y$ has the Poisson ($\mu+\lambda$) distribution.

Exercise 4.5.8

In the first hour that a bank opens, the customers who enter are of **three** kinds: those who only require teller service, those who only want to use the ATM, and those who only require special services (neither the tellers nor the ATM). Assume that the numbers of customers of the three kinds are independent of each other, and also that:

- the number that only require teller service has the Poisson (6) distribution,

- the number that only want to use the ATM has the Poisson (2) distribution, and

- the number that only require special services has the Poisson (1) distribution.

Suppose you observe the bank in the first hour that it opens. In each part below, find the chance of the event described.

a) 12 customers enter the bank

b) more than 12 customers enter the bank

c) customers do enter but none requires special services