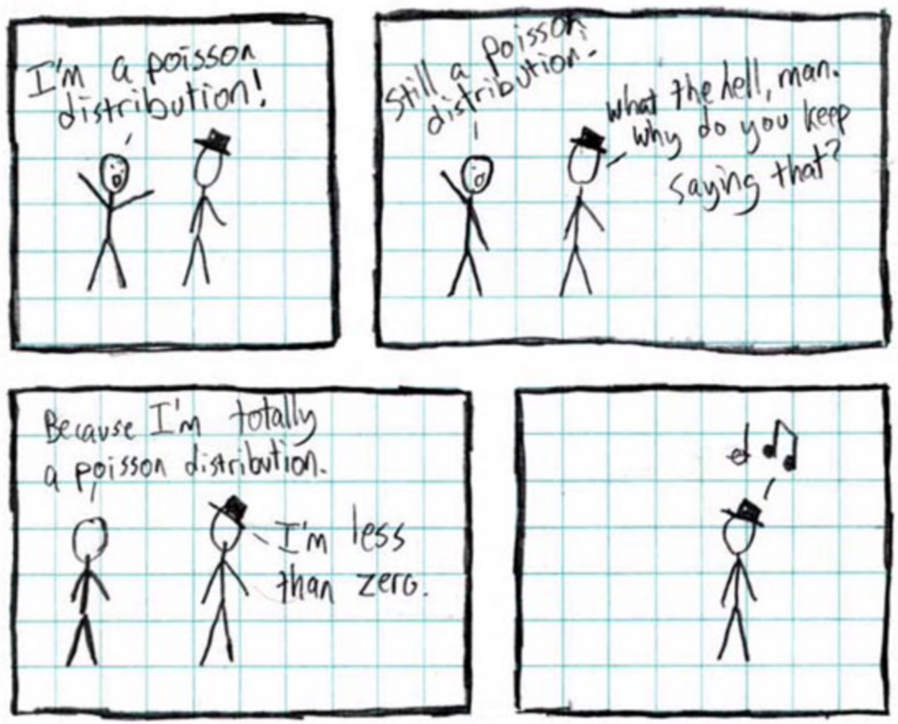
# Stat 88: Probability & Math. Stat. in Data Science




https://xkcd.com/12/

Lecture 10: 2/17/2022

Waiting times, exponential approximations, the Poisson distribution

Sections 4.2, 4.3, 4.4

# Agenda

- Go over solutions to the exercises from last lecture
- Finish up 4.2 (Waiting times)
- 4.3 Exponential approximations
- 4.4 The Poisson distribution

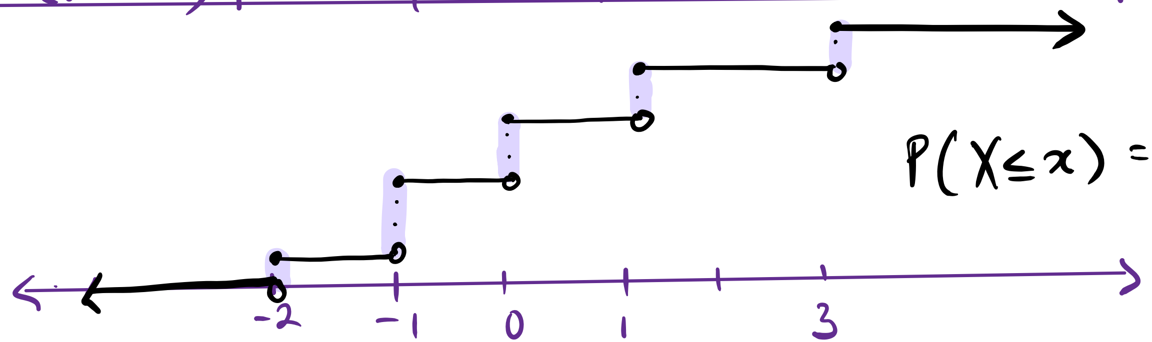
# $F(x) \longrightarrow f(x)$

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

$$
\begin{aligned}
f(x) &= P(X = x) \\
&= P(X \leq x) - P(X \leq x - 1) \\
&= F(x) - F(x - 1)
\end{aligned}
$$

- Basically, note that the graph has a jump (or discontinuity) at each possible value $x$ of $X$, and the magnitude of the jump gives the mass function value at that $x$.
- 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 |
| $P(W \leq w)$ | 0.1 | 0.4 | 0.65 | 0.85 | 1 |




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

# 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 $X$.

$X$ = # of draws until you draw your first mango.

| $x$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $f(x)$ | $4/7 = \frac{20}{35}$ | $\frac{3}{7} \cdot \frac{4}{6} = \frac{10}{35}$ | $\frac{3}{7} \cdot \frac{2}{6} \cdot \frac{4}{5} = \frac{4}{35}$ | $\frac{3}{7} \cdot \frac{2}{6} \cdot \frac{1}{5} \cdot \frac{4}{4} = \frac{1}{35}$ |
| $F(x)$ | $\frac{20}{35}$ | $\frac{30}{35}$ | $\frac{34}{35}$ | $\frac{35}{35} = 1$ |

$$F(x) = \sum_{y \le x} f(y)$$

$$F(1) = P(X \le 1) = \sum_{y \le 1} f(y) = f(1)$$

$$F(2) = \sum_{y \le 2} f(y) = P(X \le 2) = f(1) + f(2)$$

# 4.2: Recall waiting times

- 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***.
- We 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 the kth trial

last one is a S

- $f(k) = P(T_1 = k) = P(\underbrace{FFF \ldots}_{\text{k-1 Failures}} FS) = (1-p)^{k-1}p$

$q = 1 - p$

$q = P(F)$

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

$\rightarrow P(T_1 > k) = P(1^{st}\ k \text{ trials are all } F) = q^k$

$= P(T_1 = k+1 \text{ or } T_1 = k+2 \text{ or } T_1 = k+3 \text{ or} \ldots)$

$= q^k \cdot p + q^{k+1} \cdot p + \ldots$

# Waiting time until $r^{th}$ success

$p = P(S) = \frac{1}{8}$, $q = P(F) = \frac{7}{8}$

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

$= P(\underbrace{FFFFFFFFFFS}_{10\ F,\ 1S}) = \left(\frac{7}{8}\right)^{10}\left(\frac{1}{8}\right)$

- 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 rolls.)

$T_4$

$= P(\underbrace{\_\_\_\_\_\_\_\_\_\_\_\_\_\_}_{3S,\ 11F}S)$ $\frac{1}{8} = 4^{th}\ S$

$\binom{14}{3}\left(\frac{1}{8}\right)^3\left(\frac{7}{8}\right)^{11}\left(\frac{1}{8}\right)$

$P(T_4 = 6)$ FFSSSS, SFSFSSS

$\underbrace{\_\_\_\_\_}$

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

$\rightarrow \underbrace{P(T_4 > 15)}_{\text{right tail prob.}} = P(\text{in 15 rolls we have at most 3 S})$

$= F(3)$, F is cdf of binomial$(15, \frac{1}{8})$

- Notice that the ***right-tail*** probability of $T_4$ is a left hand (cdf) of the Binomial distribution for (15,1/8), and where $k = 3$.

$k \geq r$

- In general, $P(T_r = k) = \binom{k-1}{r-1} p^r q^{k-r}$

→ first k-1 trials have r-1 S & k-r F

- And $P(T_r > k) =$

$P(T_4 = 6)$

$P(FFSSSS) = q^2 p^4$

$P(FSFSSS) = q^2 p^4$

$\binom{5}{3} q^2 p^4$

$P(T_r > k) = P(\text{in k tries, at most}$

r-1 success)

$= F(r-1)$, $F$ = cdf of Bin( , p)

$P(T_r > k)$ = P(We have at most r-1 successes in first k trials)

Let $X \sim Bin(k, p)$, $p = P(S)$

$P(X \leq r-1)$ = P(at most r-1 successes in k trials)

$P(\underbrace{T_r}_{} > k) = P(\underbrace{X}_{} \leq r-1)$, $X \sim Bin(k, p)$

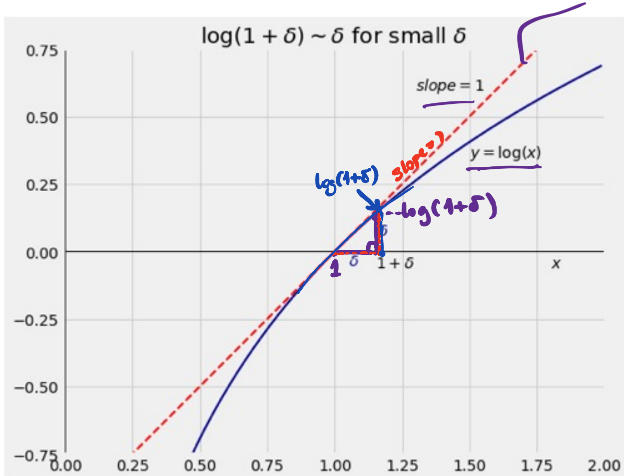
$T_r$ is called a NEGATIVE BINOMIAL r.v. parameters p, r.

Geometric r.v. is a special case with r=1

$\log x = \ln(x) = \log_e x$

# 4.3 Exponential Approximations

$\log(1+\delta) \approx \delta$ for small $\delta$




Slope of the tangent line to $y = f(x)$ is $\frac{df}{dx}$ at $x$

Slope of tangent line to $y = \log x$ at $x = 1$

$= \left.\frac{d(\log x)}{dx}\right|_{x=1} = \left.\frac{1}{x}\right|_{x=1} = 1$

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

Taylor's theorem: Let $x-a$ be close to 0

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots$$

go to 0

$$f(x) \approx f(a) + f'(a)(x-a) \Rightarrow \log(1+x) \approx \log 1 + x$$

Let $f(x) = \log(1+x)$, let $a=0$ $\quad f'(a)$ at $a=0$

$f'(x) = \frac{1}{1+x}$ , $f'(0) = \frac{1}{1+0} = 1$

# How to use this approximation

$\log(1+\delta) \approx \delta$

$\log(1-\delta) \approx -\delta$

$\log x = \log\left[\left(1-\frac{3}{100}\right)^{100}\right]$

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

$\log x = 100 \log\left(1-\frac{3}{100}\right)$ , $\delta = +\frac{3}{100}$ $\log\left(1-\frac{3}{100}\right) \approx -\frac{3}{100}$

$\approx 100 \times \left(-\frac{3}{100}\right) = -3$

$\log x = -3 \Rightarrow x = e^{-3}$

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

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

$\log x \approx 5\cancel{000} \cdot \left(\frac{-2}{1\cancel{000}}\right) = -10$

$x \approx e^{-10}$

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

$x = (1-p)^n \Rightarrow \log x = n \underbrace{\log(1-p)}_{\approx -p} \approx -np$

"This implies"

$x \approx e^{-np}$

In this class (lecture) $\log x = \ln x$ unless stated o/w.

# 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 : $p = 1/1{,}000{,}000 = \cancel{10-6}.$ $10^{-6}$

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

Can think of this as $100{,}000 = 10^5$ independent trials & each trial either has a success or not (success = Typo)

$$P(\text{no typo}) = 1-p$$

$$P(\text{no typo at all in chapter}) = (1-p)^n \approx e^{-np} = e^{-1/10}$$

$$-np = -10^5 \cdot 10^{-6} = -\frac{1}{10}$$

# Bootstraps and probabilities

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

$P(\text{picking Ali}) = \frac{1}{n}$

- 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.)

$P(\text{not picking Ali}) = \frac{n-1}{n} = 1-\frac{1}{n}$

$P(\text{Ali is chosen at least once})$

$= 1 - P(\text{Ali is never chosen in } n \text{ draws})$

$= 1 - \left(1-\frac{1}{n}\right)^{n}$ ⟸ # of draws.

$= 1 - \frac{1}{e}$

$x = \left(1-\frac{1}{n}\right)^{n} \approx e^{-1}$

# The Poisson Distribution

prob small

- 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

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

$X \sim \text{Poisson}(\mu)$

- The parameter of the distribution is $\mu$

$$P(X=0) = e^{-\mu} \cdot \frac{\mu^0}{0!} = e^{-\mu}$$

$$P(X=1) = e^{-\mu} \cdot \frac{\mu^1}{1!} = e^{-\mu} \cdot \mu$$

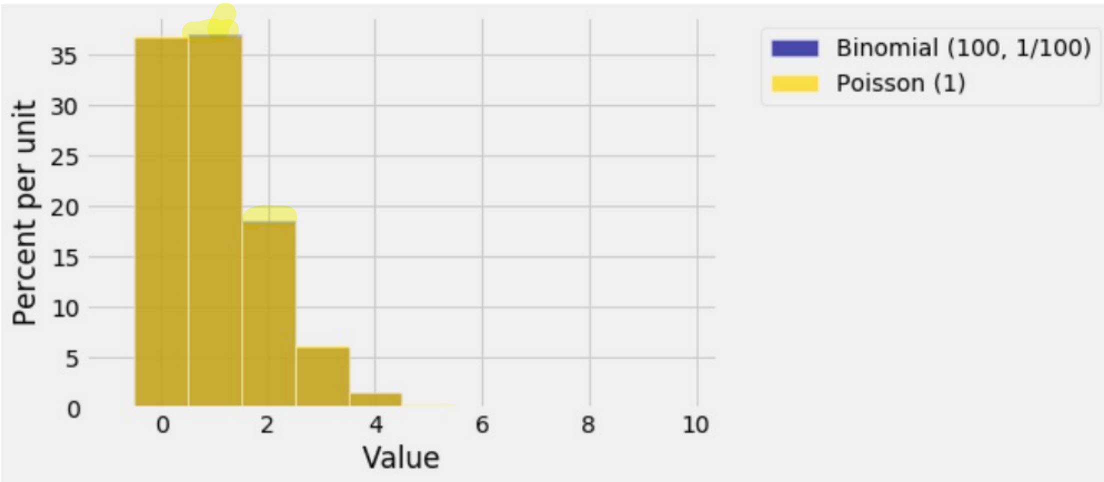
$$P(X=2) = e^{-\mu} \cdot \frac{\mu^2}{2}$$

# Relationship between Poisson and Binomial distributions

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

$X \sim \text{Poisson}(\mu)$, $P(X=k) = e^{-\mu} \frac{\mu^k}{k!}$

$k = 0, 1, 2, 3, \ldots$



# 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($\mu$) distribution, and
- $Y$ has the Poisson($\lambda$)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

# 5.1: Expected Value of a random variable

- The ***Expectation*** or ***Expected Value*** of a random variable $X$ is defined to be the sum of all the products $(\boldsymbol{x} \times \boldsymbol{f}(\boldsymbol{x}))$ over all possible values $x$ of the random variable $X$: that is, the expectation of a random variable is a ***weighted average*** of all the possible values that the rv can take, weighted by the probability of each value.

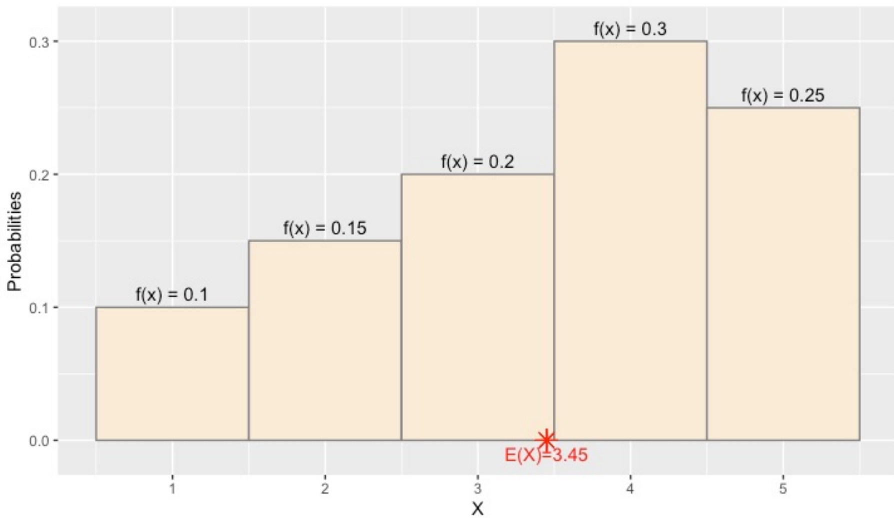
- $\boldsymbol{E}(\boldsymbol{X}) = \sum \boldsymbol{x} \cdot \boldsymbol{P}(\boldsymbol{X} = \boldsymbol{x}) = \sum \boldsymbol{x} \cdot \boldsymbol{f}(\boldsymbol{x})$

- For example, toss a coin 3 times, let $X$ = # of heads. Write down $f(x)$, and then use the formula to compute the expectation of $X$.

- Note: If $X$ takes finitely many values, no problem. If it takes infinitely (countable, but not finite) many values (such as Poisson, or Geometric), then we have to be more careful when we take the sums.

# Example (from text)

| x | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| f(x) | 0.1 | 0.15 | 0.2 | 0.3 | 0.25 |



# Notes about E(X)

- Same units as X
- Not necessarily an attainable value (for example, in In 2020, there was an average of 1.93 children under 18 per family in the United States)
- Expectation is a long-run average value of X
- Center of mass or center of gravity (balancing point) for the distribution
- Expectation of a constant is that constant $E(c) = c$
- Linearity: $E(aX + b) = aE(X) + b$
- Example: for the X defined in the previous slide, find E(2X-1)

# Special examples

- Bernoulli (Indicators)

- Uniform

- Poisson

# 5.2: Functions of random variables

- $X$~unif{-1,0,1}, $Y = X^2$, find $E(Y)$

- In general, if $Y = g(X), E(Y) = E\big(g(X)\big) = \sum g(x) \cdot f(x) = \sum g(x) \cdot P(X = x)$

- Example: $W = \min(X, 0.5)\,.\,Find\ E(W)$ (write out the values of $W$, and probs)

- Note that $Y = g(X)$ is also a random variable, just defined via $X$.

# Multiple random variables on the same outcome space

- **Joint distributions**: Recall rolling a pair of dice. Draw a table of outcomes and their probabilities. What event does each cell represent?
- Sum of probabilities across *all* the cells is 1, since this is all the probabilities across all possible outcomes

- Now, suppose we draw two tickets without replacement from a box that has 5 tickets marked 1, 2, 2, 3, 3. Let $X_1$ and $X_2$ represent the values of the tickets drawn on the first and second draws respectively.
- Create a table of all possible outcomes for the pair $(X_1, X_2)$ (which is also a random variable),and write down the probabilities using the multiplication rule.

# Joint distributions

- Draw two tickets without replacement from a box that has 5 tickets marked 1, 2, 2, 3, 3. Let $X_1$ and $X_2$ represent the values of the tickets drawn on the first and second draws respectively.
- Create a table of all possible outcomes for the pair $(X_1, X_2)$ (which is also a random variable), and write down the probabilities using the multiplication rule.

| | $X_2 = 1$ | $X_2 = 2$ | $X_2 = 3$ |
|---|---|---|---|
| $X_1 = 1$ | | | |
| $X_1 = 2$ | | | |
| $X_1 = 3$ | | | |

# Marginal distributions

- What is $\mathrm{P}(X_1 = 1)$? Write down the pmf for $X_1$ and $X_2$

- Are they independent?

- Use the table to compute $P(X_1 + X_2 = 5)$

- Use the table to compute $E\big(g(X_1, X_2)\big)$, where $g(X_1, X_2) = |\ X_1 - X_2\ |$ (the expected distance between the two draws)

- If $S = X_1 + X\_2$, find $E(S)$.