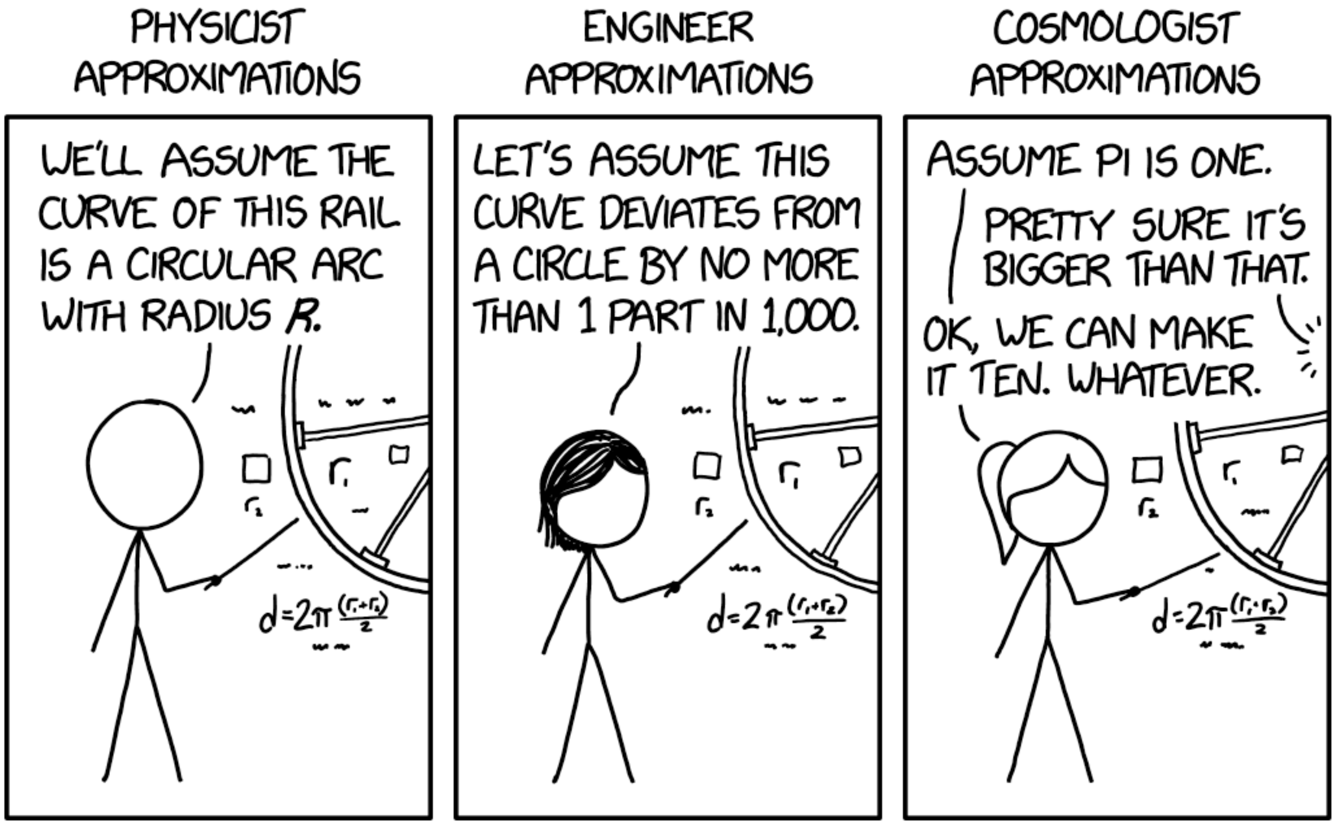
# Stat 88: Probability & Math. Stat in Data Science




https://xkcd.com/2205/

## Lecture 14: 2/16/2024

## Waiting times, Exponential approximations, Introducing the Poisson

4.2, 4.3, 4.4

Shobhana Stoyanov

# Geometric distribution

$T_1$ = # of trials up to & including the 1st trial which is a success

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

$(k-1)$ F, last trial = S

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

$= p(1-p)^0 + p(1-p)^{2-1} + p(1-p)^{3-1}$

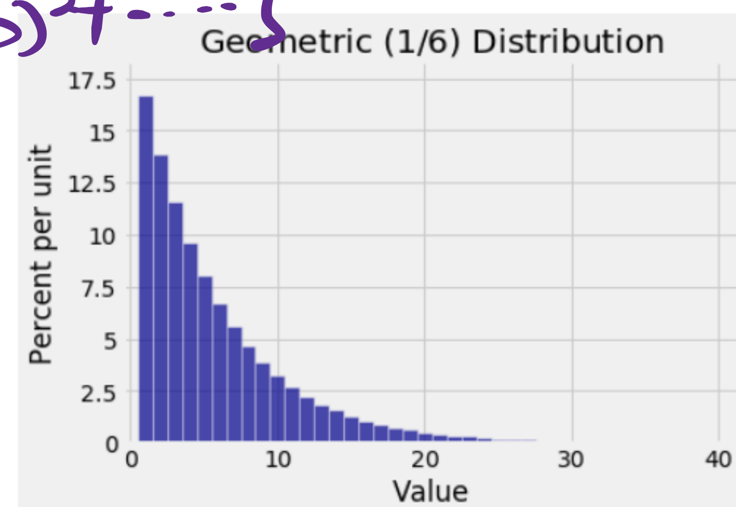
$k=1 \qquad = p + p(1-p) + p(1-p)^2 + \ldots$

- $f(k) = P(T_1 = k) = (1-p)^{k-1} p$

- 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?

$p \sum_{k=1}^{\infty} (1-p)^{k-1} = p\left[1 + (1-p) + (1-p)^2 + (1-p)^3 + \ldots\right]$

$= p\left[\frac{1}{1-(1-p)}\right] = 1$



# Geometric distribution

$P(T_1 > n) = q^n \quad (q = 1-p)$

$P(T_1 \leq n) = 1 - q^n$

- $f(k) = P(T_1 = k) = (1-p)^{k-1} p$
- $F(n) = P(T_1 \leq n) = \sum_{k=1}^{n} (1-p)^{k-1} p = p + p(1-p) + p(1-p)^2 + \cdots p(1-p)^n$
- $P(T_1 > n) = \sum_{k=n+1}^{\infty} f(k)$

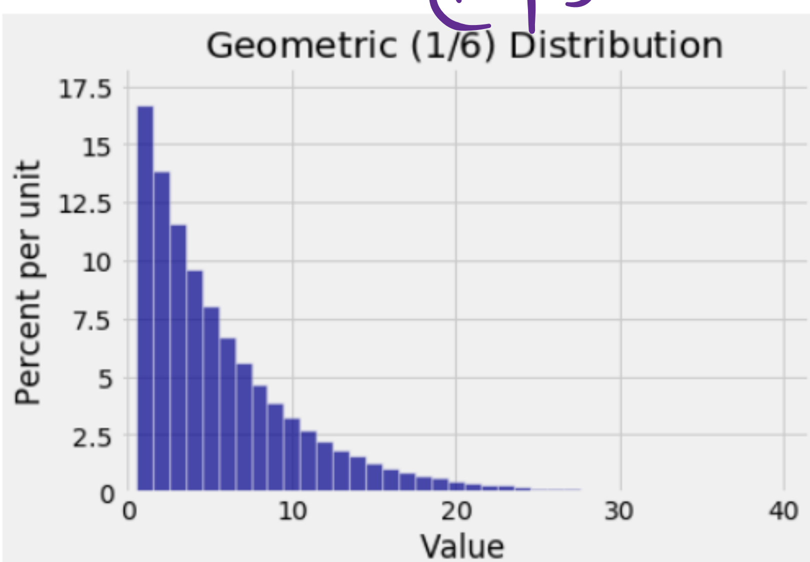
TAIL PROBABILITY

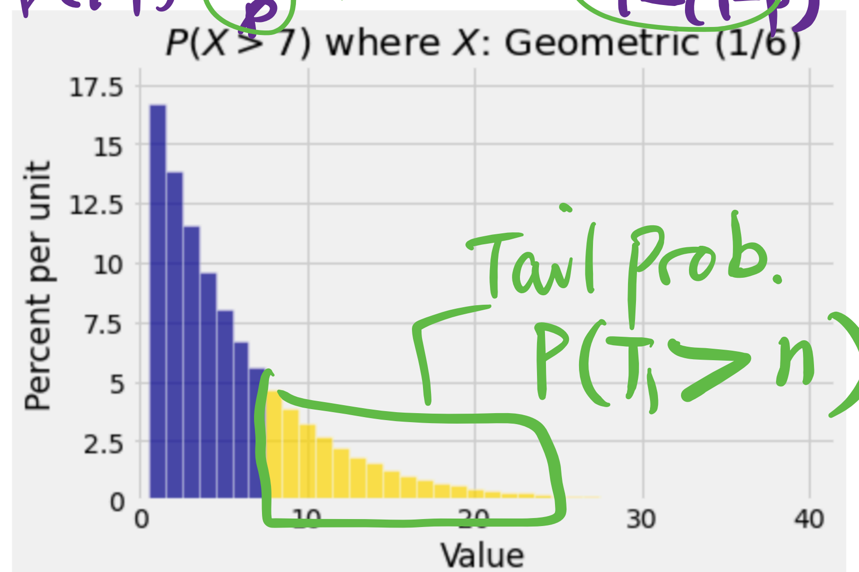
$= \sum_{k=n+1}^{\infty} (1-p)^{k-1} p = p(1-p)^n + p(1-p)^{n+1} + \cdots$

$= p(1-p)^n \left[ 1 + (1-p) + (1-p)^2 + \cdots \right]$, where $1 + (1-p) + (1-p)^2 + \cdots = \frac{1}{1-(1-p)}$

$= p(1-p)^n \cdot \frac{1}{p}$

$= (1-p)^n$



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

- Say we roll a 8 sided die. $P(S) = P(\boxed{8}) = \frac{1}{8}$
- What is the chance that the first time we roll an eight is on the 11$^{th}$ try?

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

- 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)

$$= P(\underbrace{\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_}_{14\ \text{rolls}}S) \quad \binom{14}{3}\left(\frac{7}{8}\right)^{11}\left(\frac{1}{8}\right)^{3}\frac{1}{8}$$

$$P(T_r = 15) = \binom{14}{3}\left(\frac{7}{8}\right)^{11}\left(\frac{1}{8}\right)^{4} \quad r=4$$

- 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.

Ex Suppose I draw from a std deck with repl until I draw 7 Queens. What is the prob that I need to draw 60 times

$$\binom{59}{6}\left(\frac{12}{13}\right)^{53}\left(\frac{1}{13}\right)^{7}$$

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

$P(S) = p$

# The Negative Binomial Distribution

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

Negative Binomial r.v

$P(T_r > k)$

TAIL PROB OF NB

For Ex $P(T_4 > 15)$

P( more than 15 rolls 8 sided die to get 4 S )

In order for $T_4 > 15$

In the first 15 rolls we see at most 3 S

If X is the # of S in 1st 15 rolls

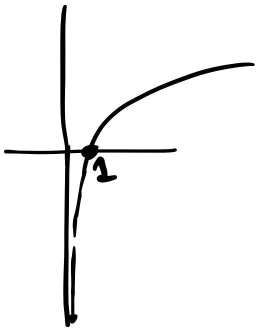
this is the prob that $X \sim Bin(15, p)$, $X \le 3 = F(3)$ for Bin. F

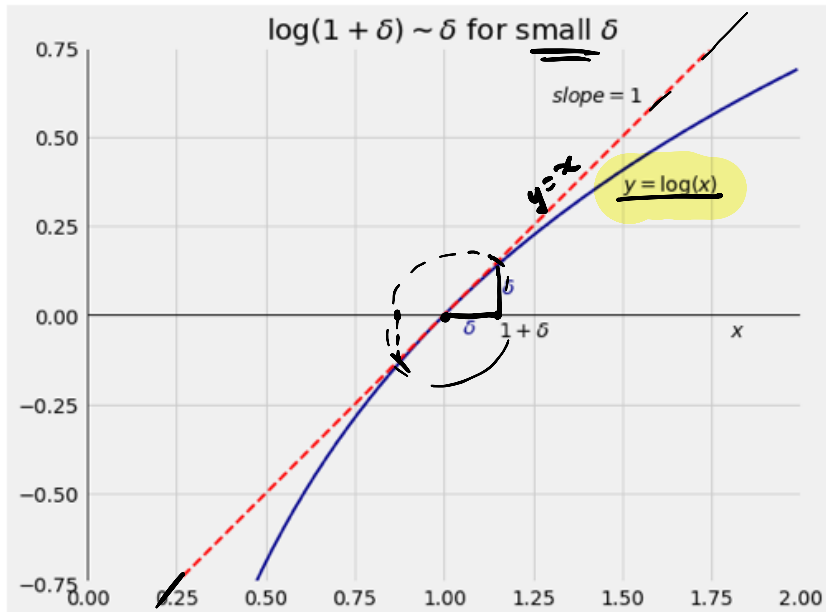
$T_r \sim NB(r, p)$

Geometric r.v $T_1 \sim NB(p, r=1)$

# 4.3 Exponential Approximations

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




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

Taylor's theorem is a linearization rule. If we have a near 0 (

$$f(x) = f(a) + f'(a)(x-a) + f''(a)\frac{(x-a)^2}{2} + \cdots + f^{(3)}(a)\frac{(x-a)^3}{3!} + \cdots$$

log ⟺ ln  log ⟺ $\log_e$

## How to use this approximation

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

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

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

$\log x = y$

$x = e^y$

$\log x = 100 \log_e\left(1 - \frac{3}{100}\right) \approx 100 * \left(-\frac{3}{100}\right) = -3$

$\log_e\left(1 - \frac{3}{100}\right) \approx -\frac{3}{100}$

$\log x \approx -3$, $x = e^{-3}$

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

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

exercise for chocolate.

# 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, $X = 0, 1, 2, \ldots$

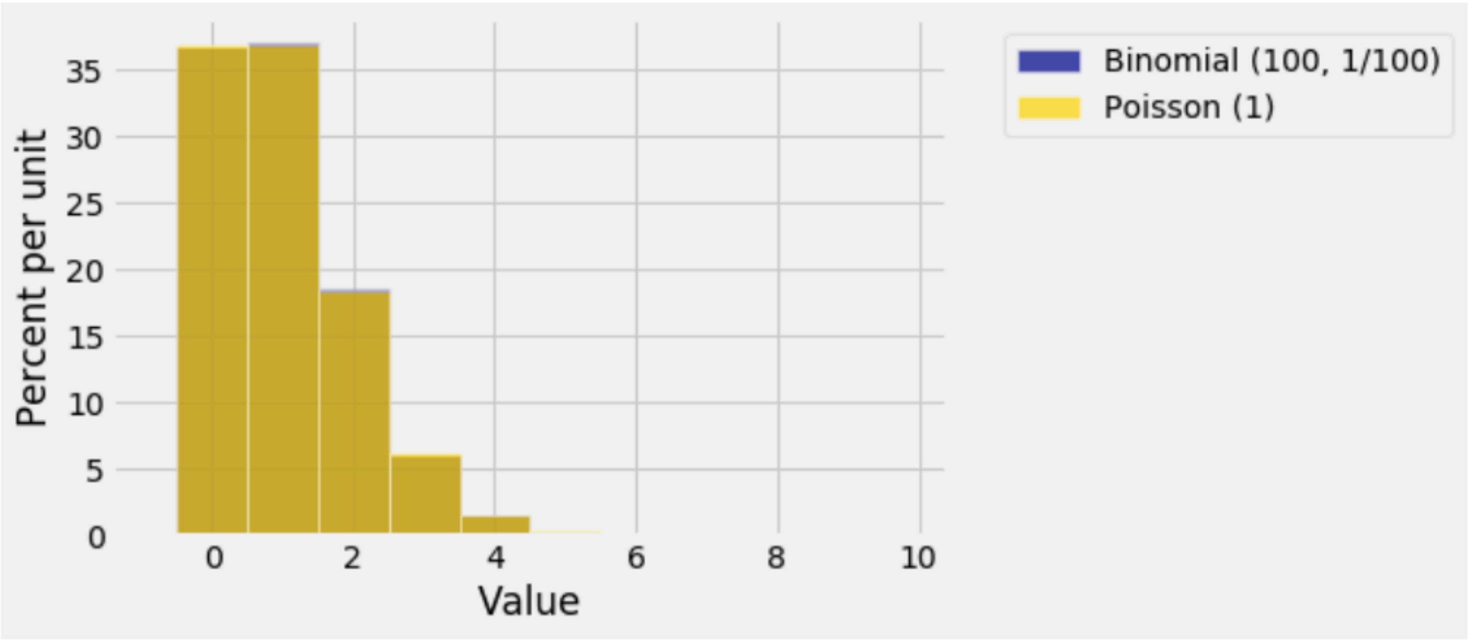
- We say that a random variable $X$ has the ***Poisson*** distribution if

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

- The parameter of the distribution is $\mu$

# 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($\mu$) 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($\mu$) 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