Central Limit Theorem

11 Monday 2019

PLSC 309

Roadmap

1. Law of large numbers (sample converges to population as sample gets larger)
2. What happens to distributions of sums?
    a. Bernoulli sums
    b. Random walk
    c. Exponential sums
    d. Poisson sums
3. CLT

Review: sampling distribution

Respondent	Age
1	20
2	18
3	22
4	23
5	17
6	21
7	20

Sample	Mean
1	20.14

Review: sampling distribution

Respondent	Age
8	22
9	19
10	24
11	16
12	20
13	18
14	21

Sample	Mean
1	20.14
2	20

Review: sampling distribution

Respondent	Age
15	18
16	22
17	19
18	21
19	19
20	23
21	20

Sample	Mean
1	20.14
2	20
3	20.28

Review: sampling distribution

Respondent	Age
22	23
23	16
24	18
25	22
26	23
27	17
28	19

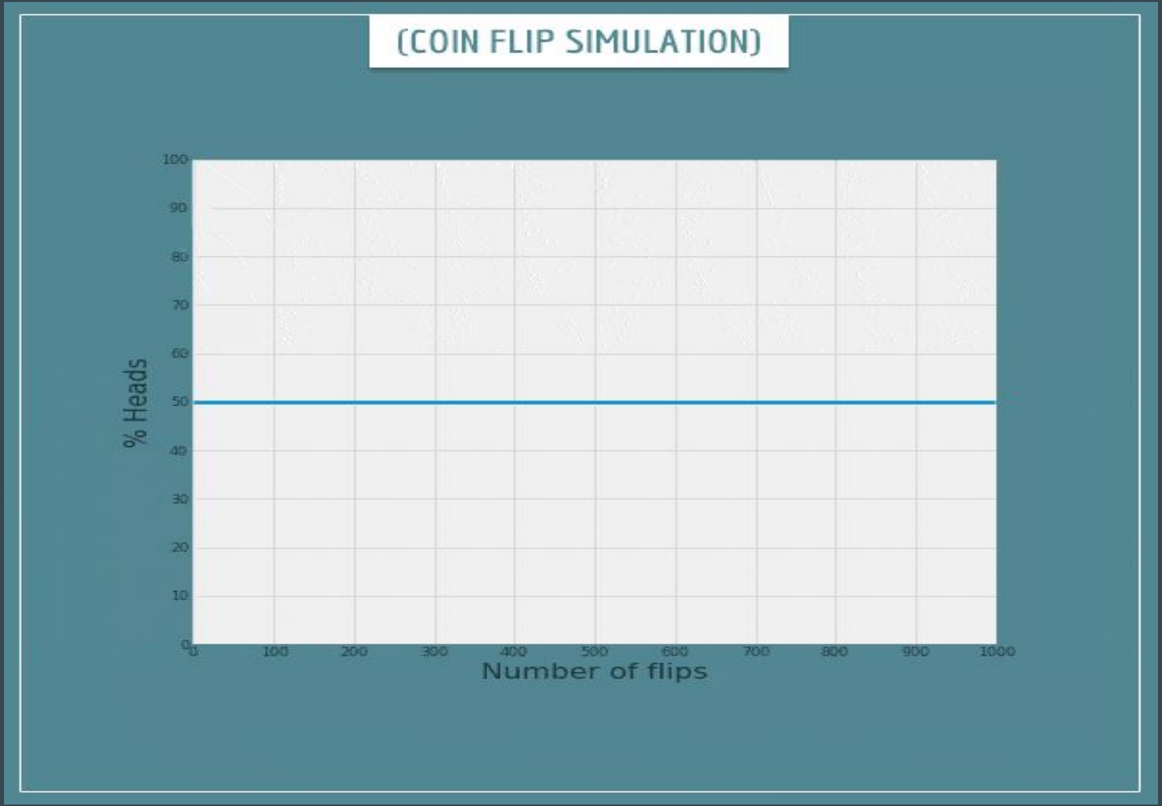
Sample	Mean
1	20.14
2	20
3	20.28
4	19.71

Review: sampling distribution

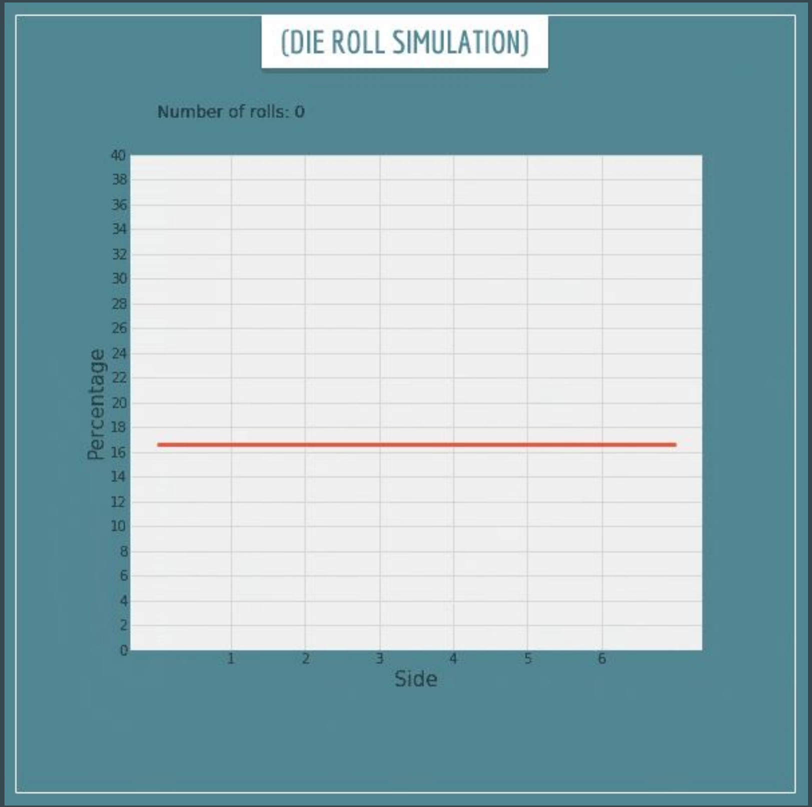
Respondent	Age
29	18
30	21
31	20
32	23
33	19
34	21
35	22

Sample	Mean
1	20.14
2	20
3	20.28
4	19.71
5	20.57

Law of Large Numbers



Law of Large Numbers



Law of Large Numbers

As N increases, mean of X approaches $E(X)$

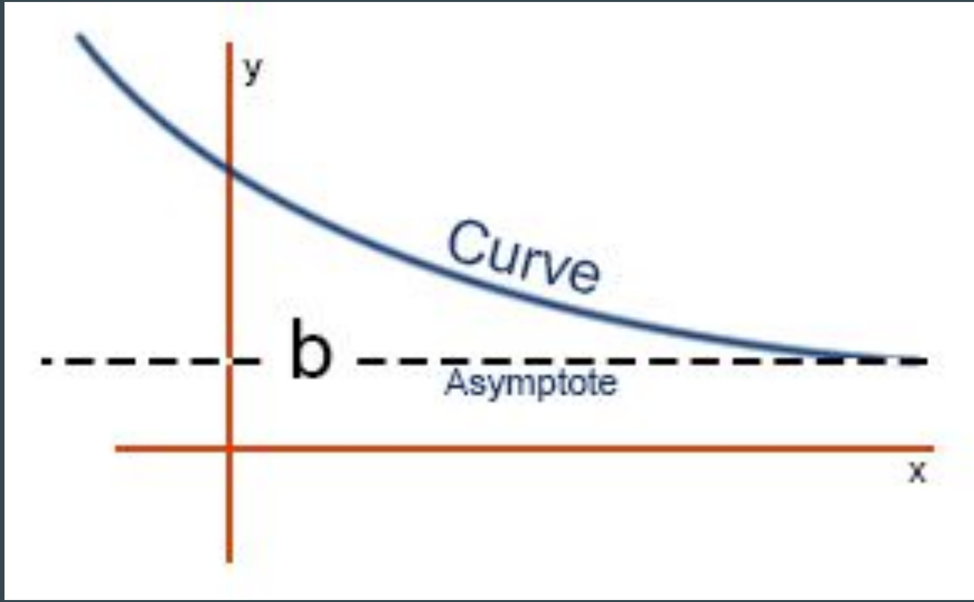
Law of Large Numbers

$$\sum(X_i\text{=outcome}) \,/\, N = P(X\text{=outcome})$$

$$\text{As } N \rightarrow \infty$$

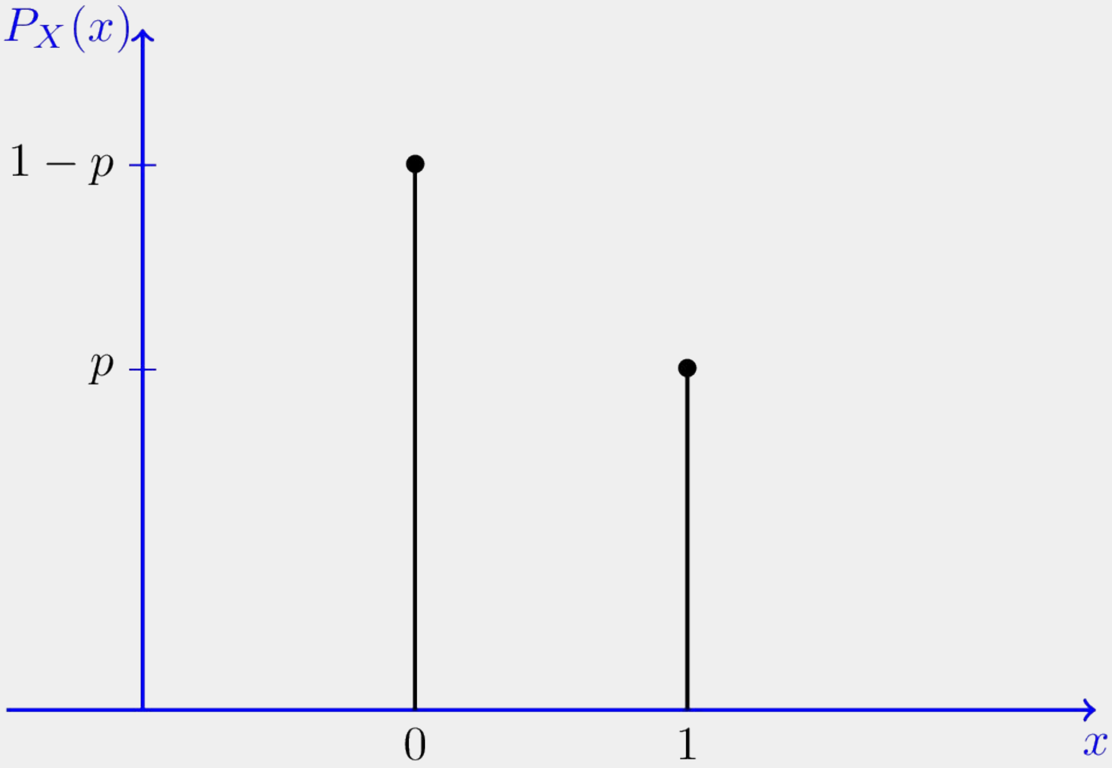
What are asymptotics?

- As a function gets infinitely large it approaches infinity



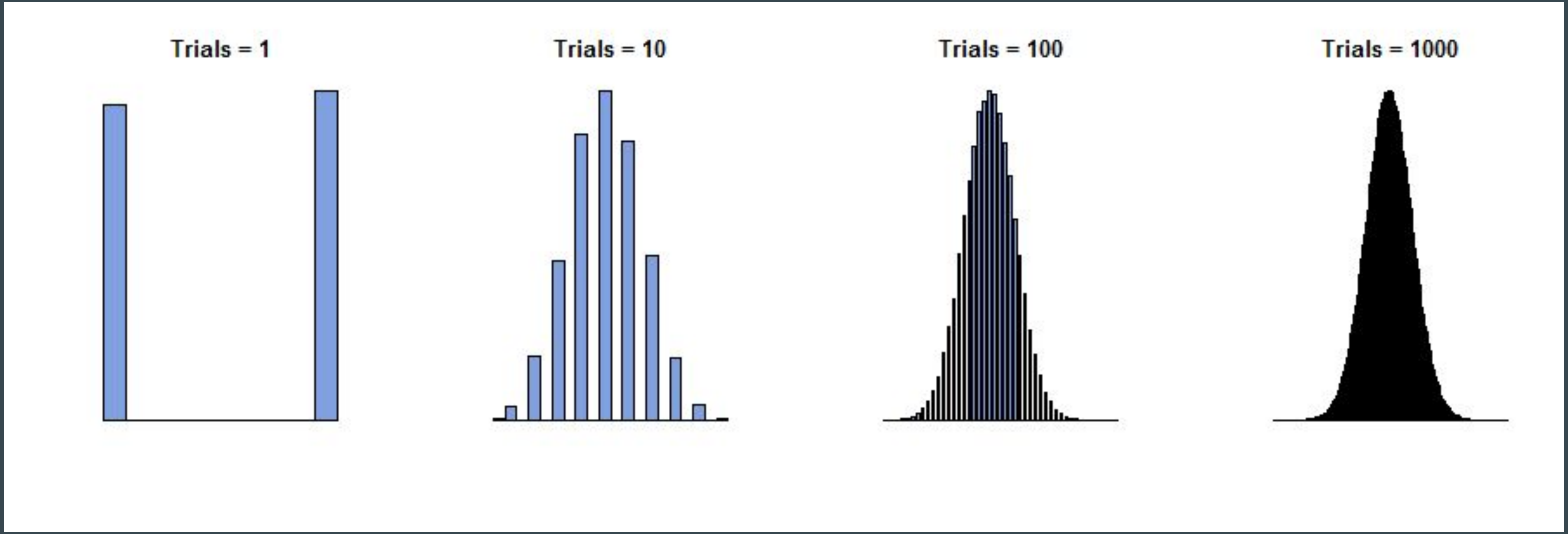
Bernoulli Distribution

$$X \sim Bernoulli(p)$$



$$\begin{cases} q = 1 - p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases}$$

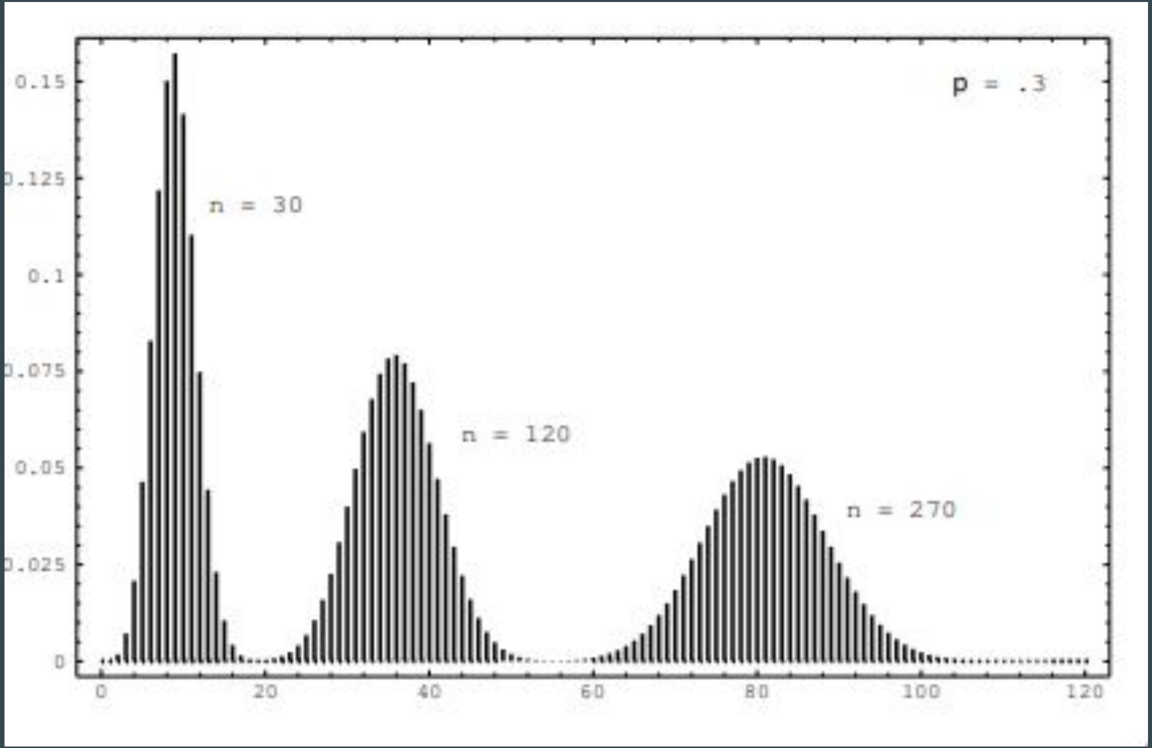
Series of bernoulli trials



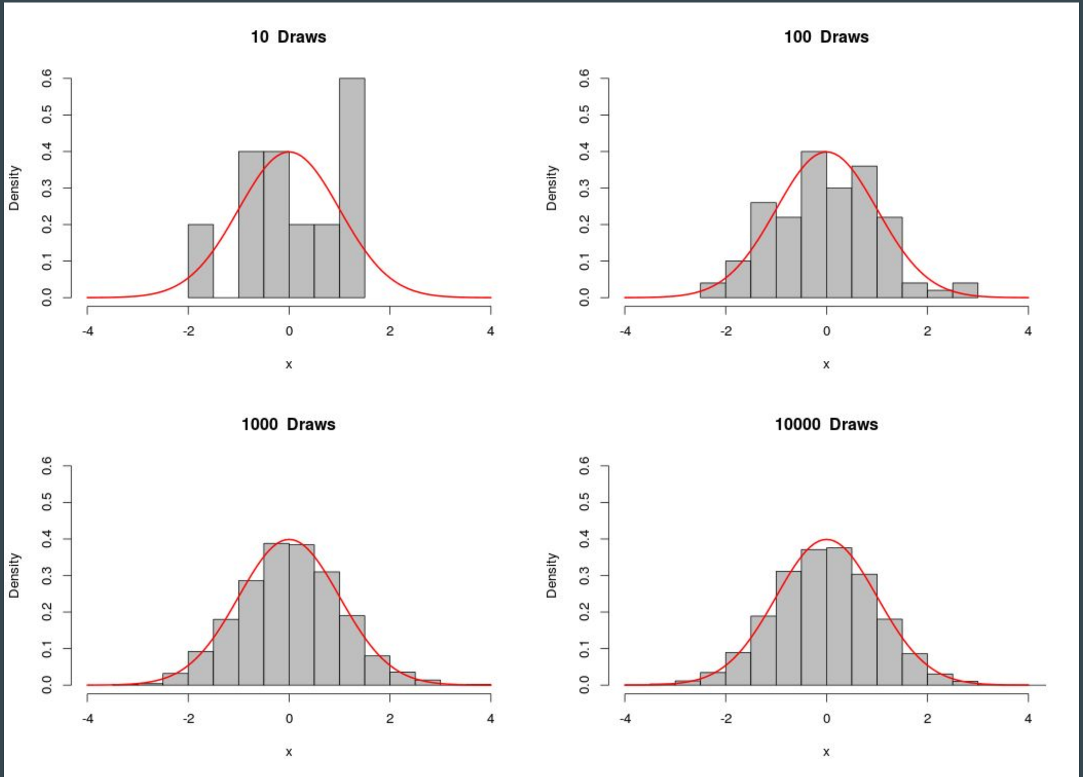
Binomial distribution is a sum of Bernoullis

- If X is a series of Bernoulli trials, each X_i is a 0 or 1
  - $X = \{0, 1, 0, 1, 1, 0, 1, 0, 0\}$
- A binomial trial is predicting *k* successes out of n trials
- So a binomial expected value is just the sum of a Bernoulli variable
  - $k = \sum(X)$
  - $k = np = E(X)$

Sum of n Bernoulli trials



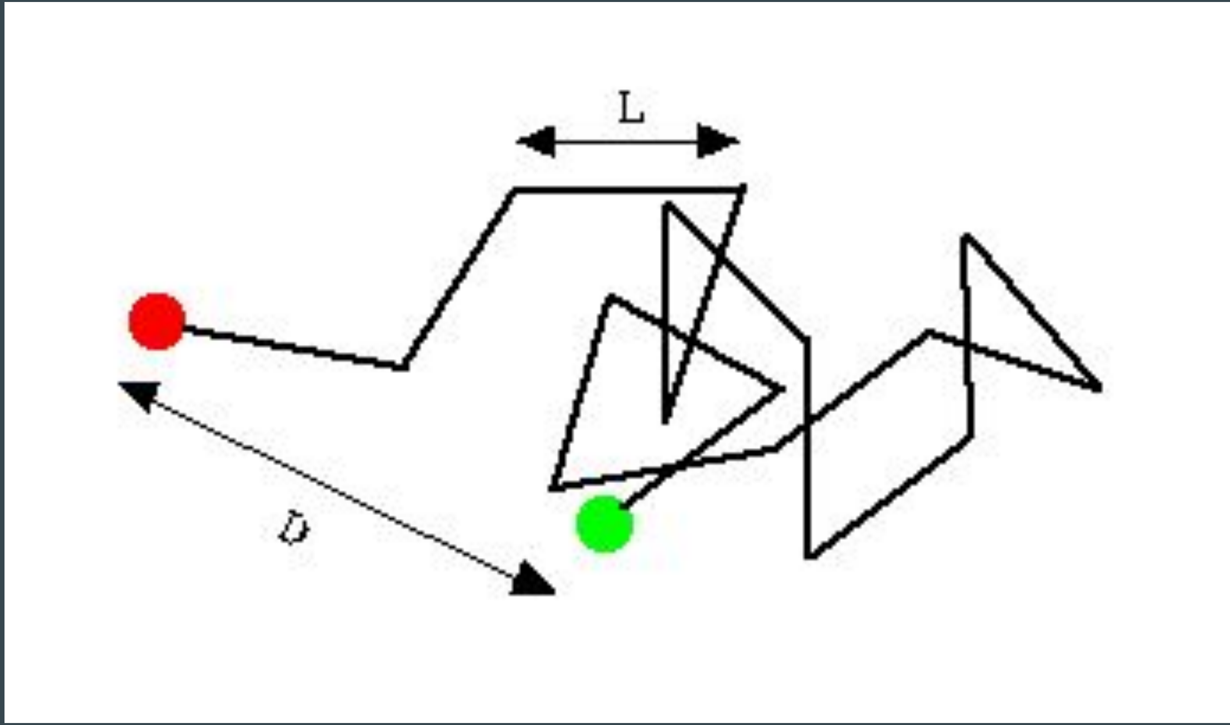
Normal approximation to binomial



Random walks

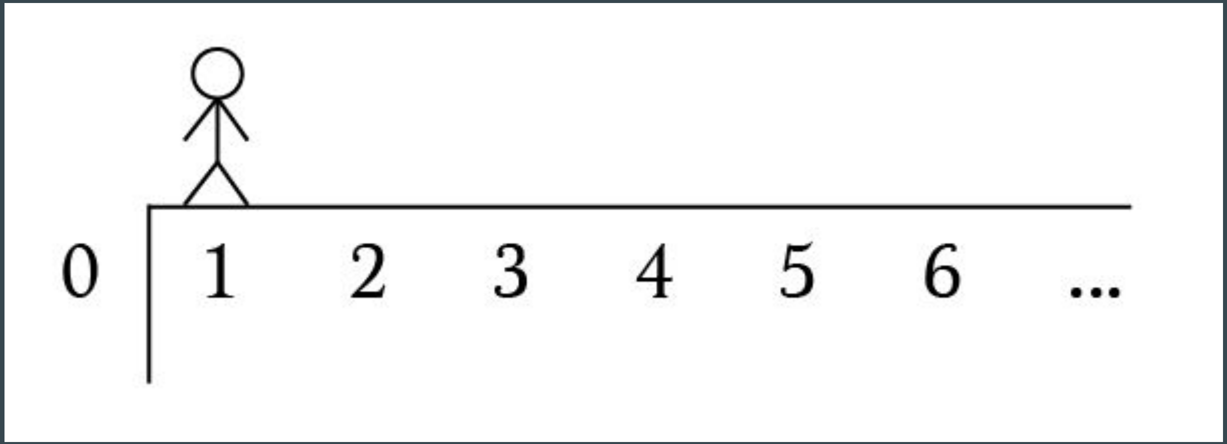
- For an N dimensional space
- Move some fixed distance b
- With direction a random variable p

Random walks

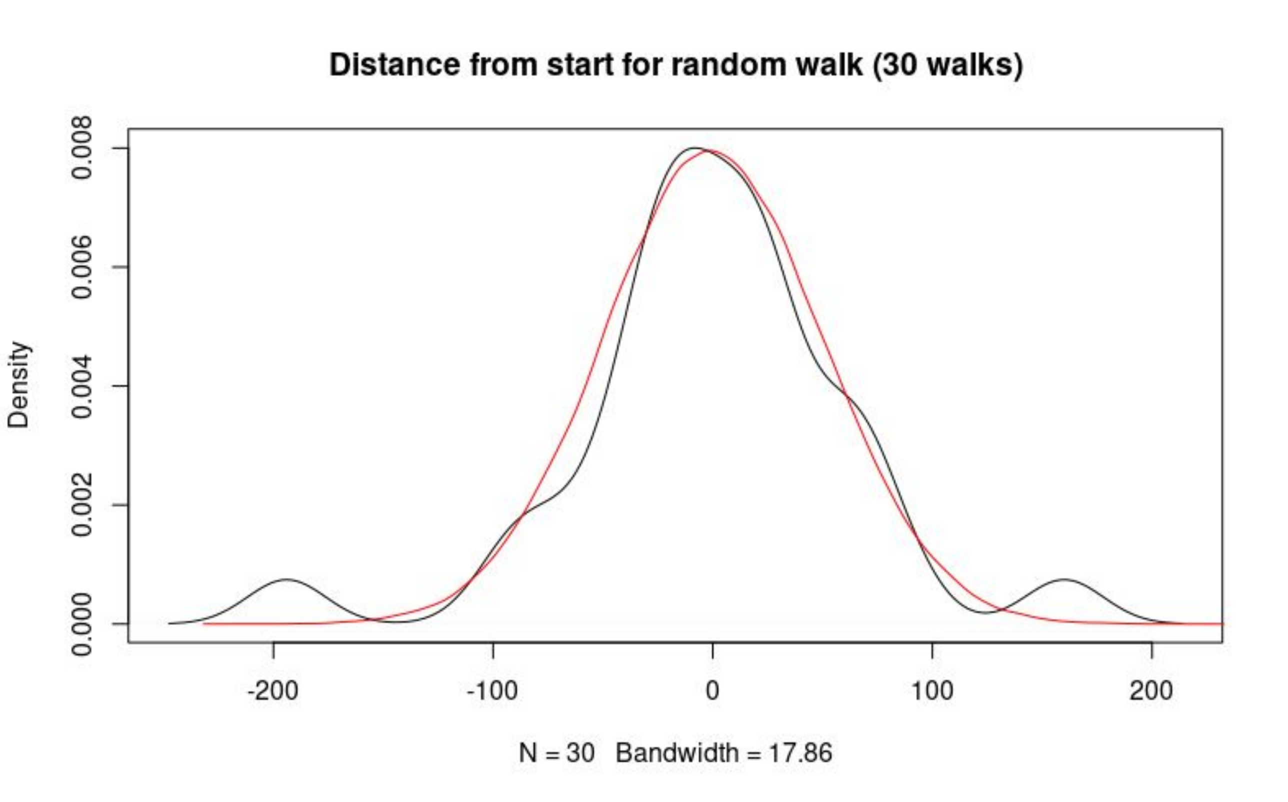


One dimensional random walk

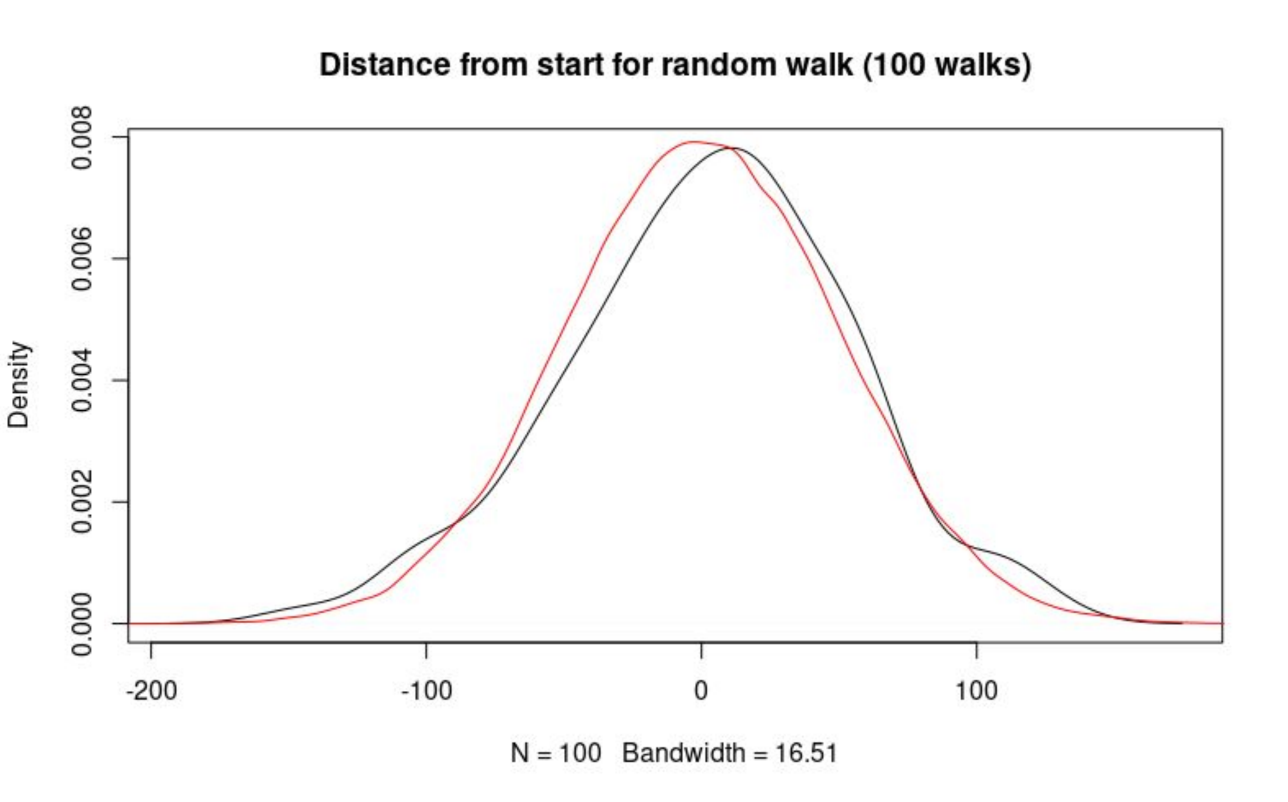
- Move one integer with probability p for either direction



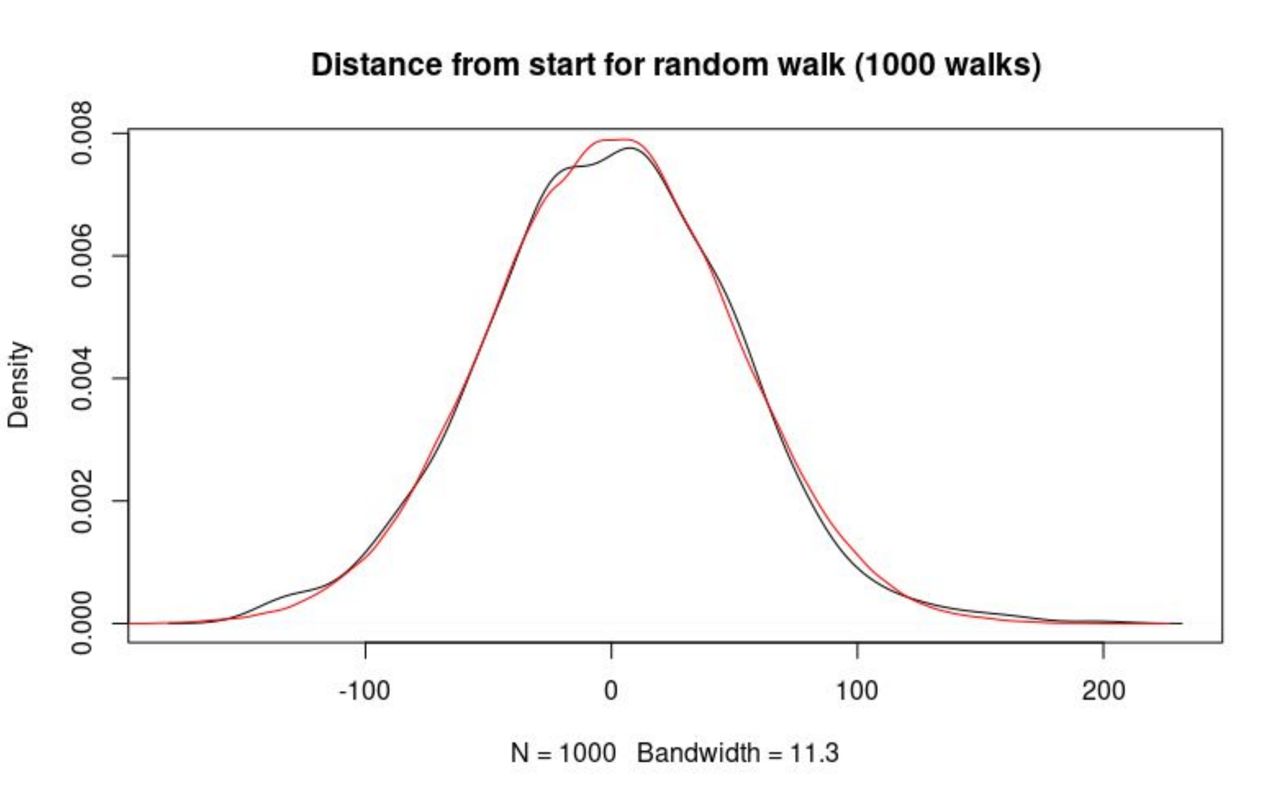
Distance from start for random walk



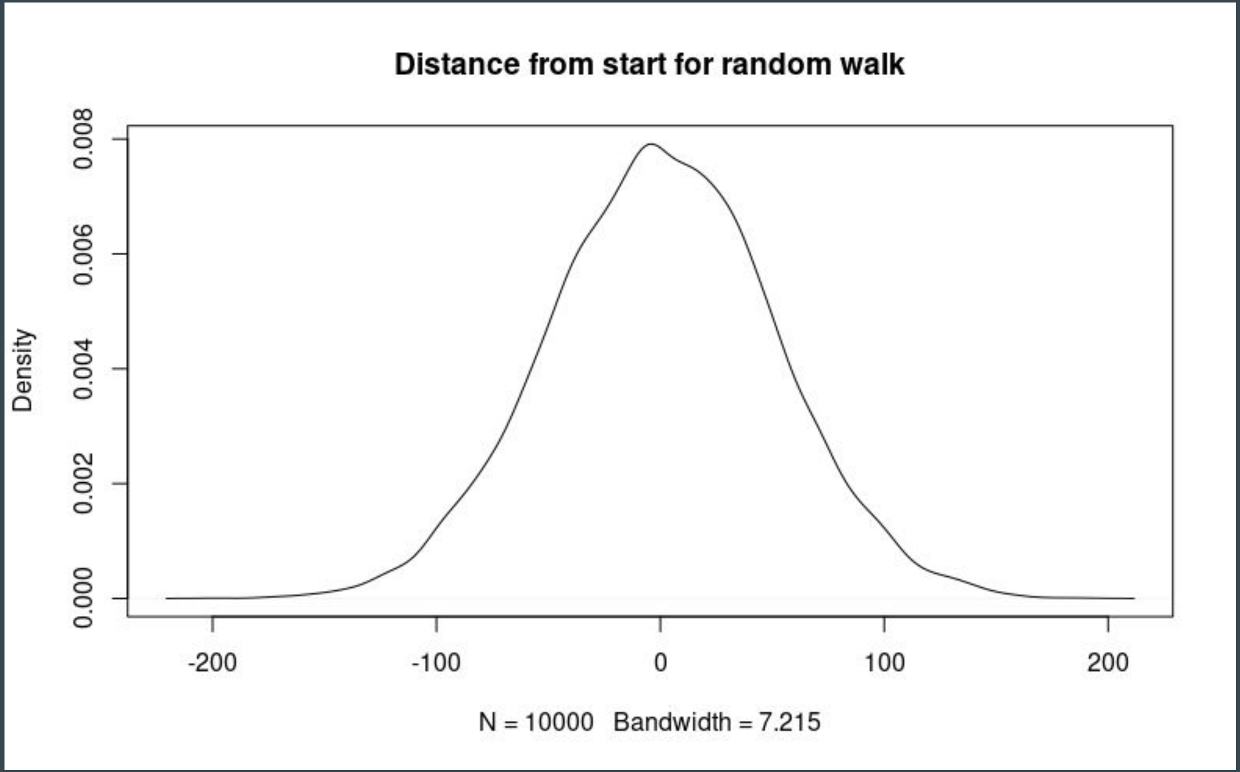
Distance from start for random walk



Distance from start for random walk



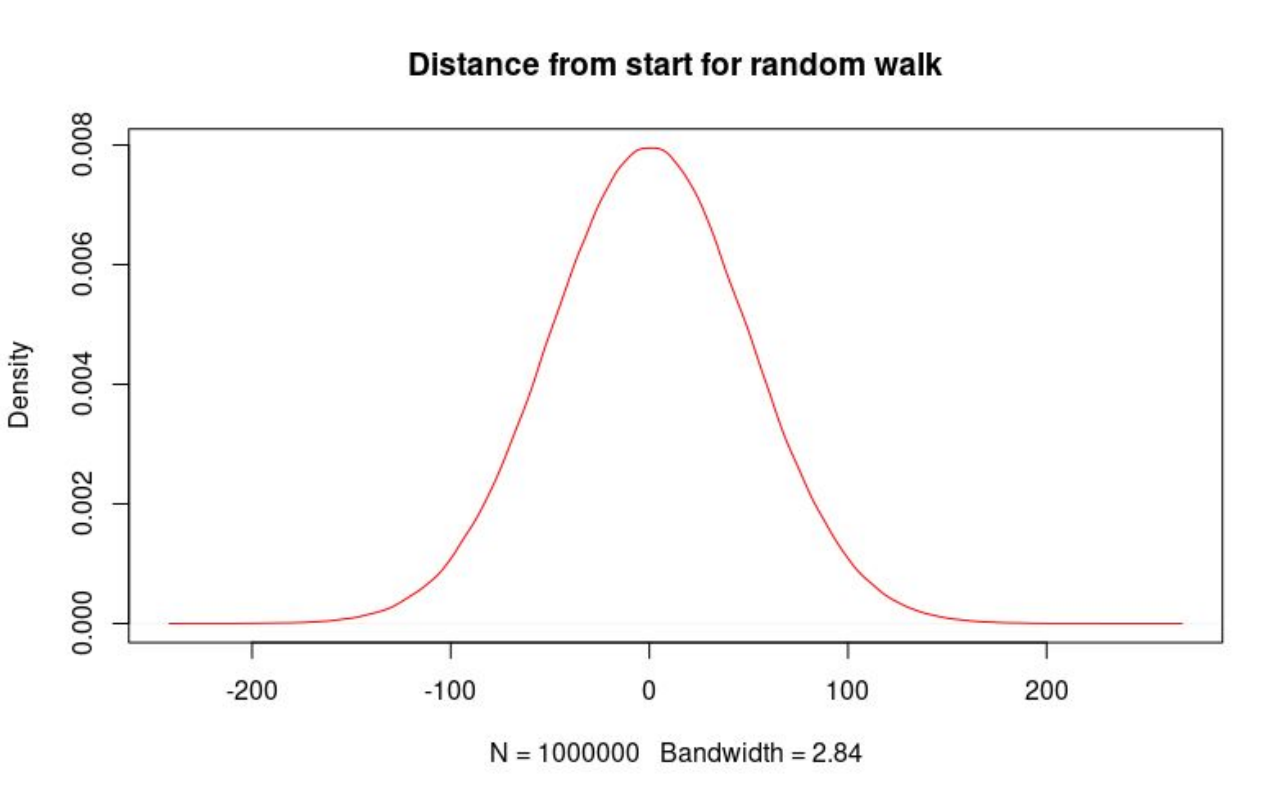
Why is this useful?



D = distance from center

P(D < -100) = 2.92%

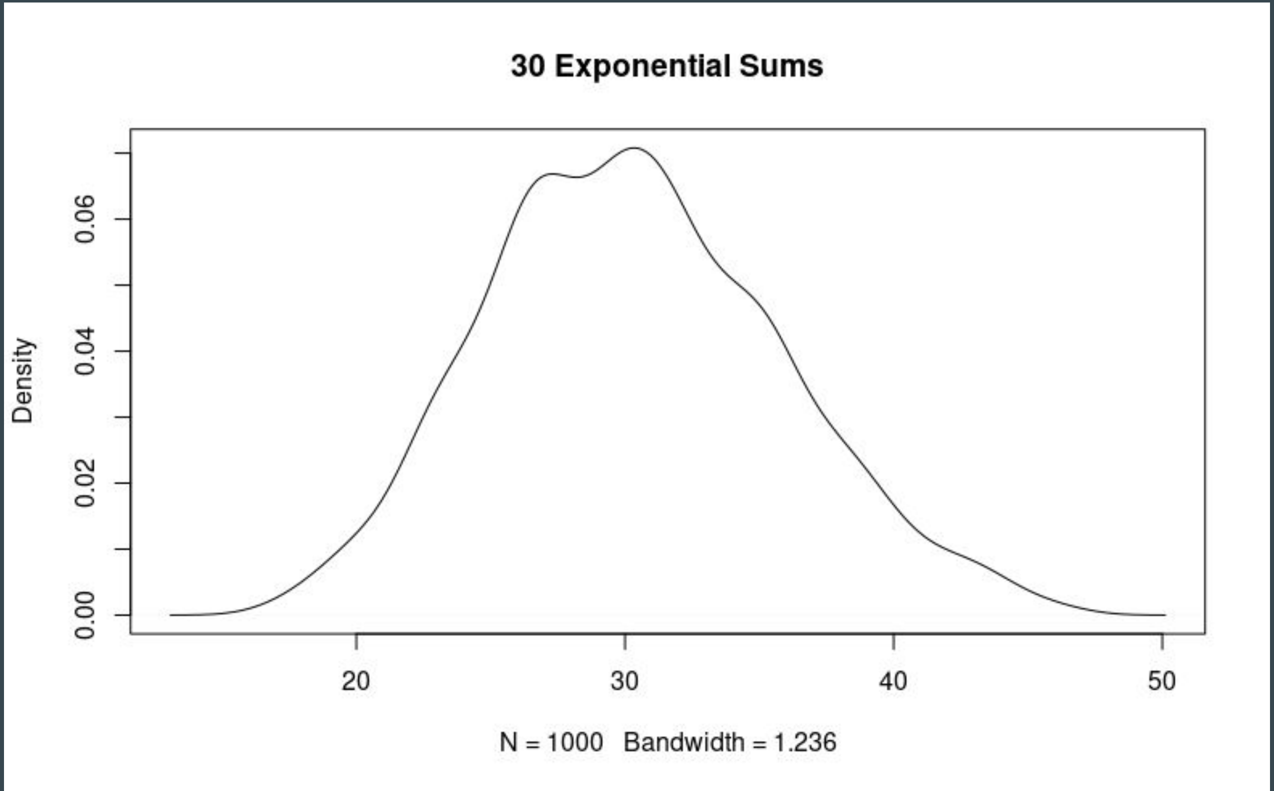
Why is this useful?



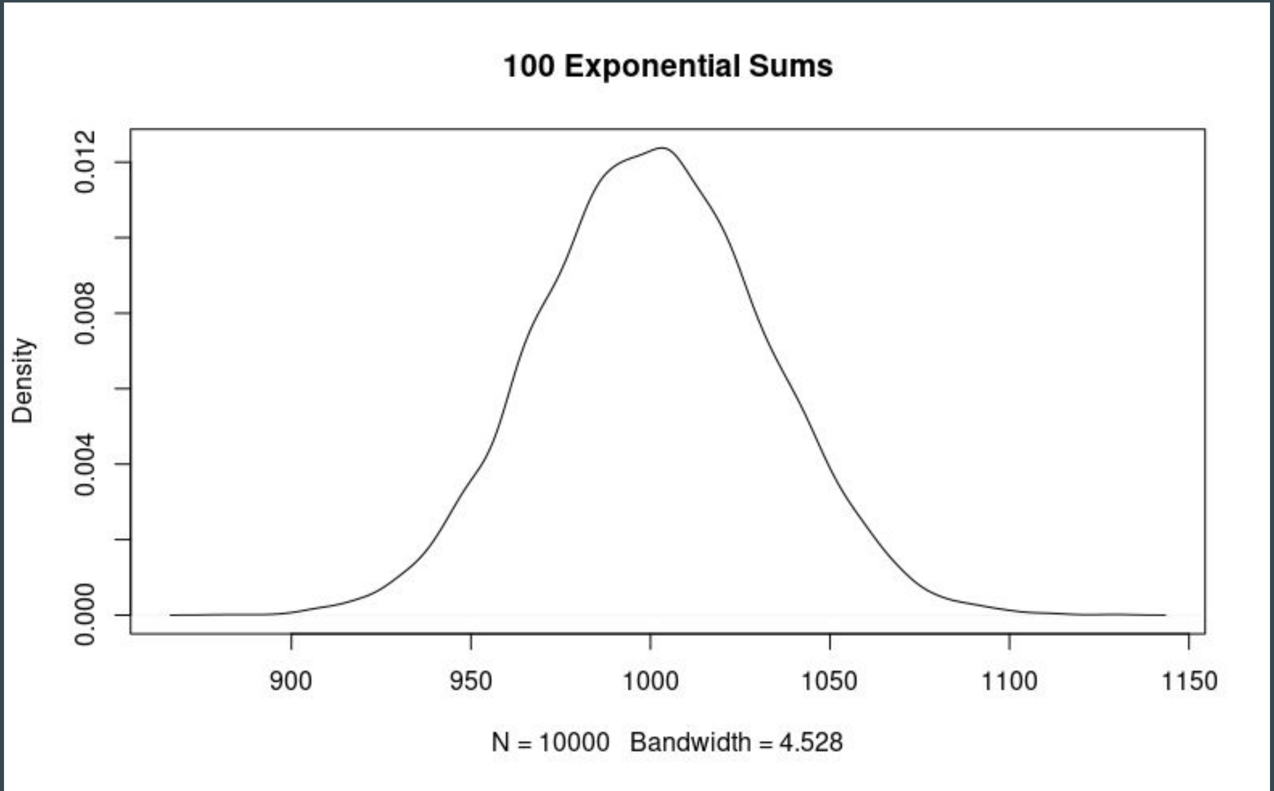
D = distance from center

P(D < -100) = 3%

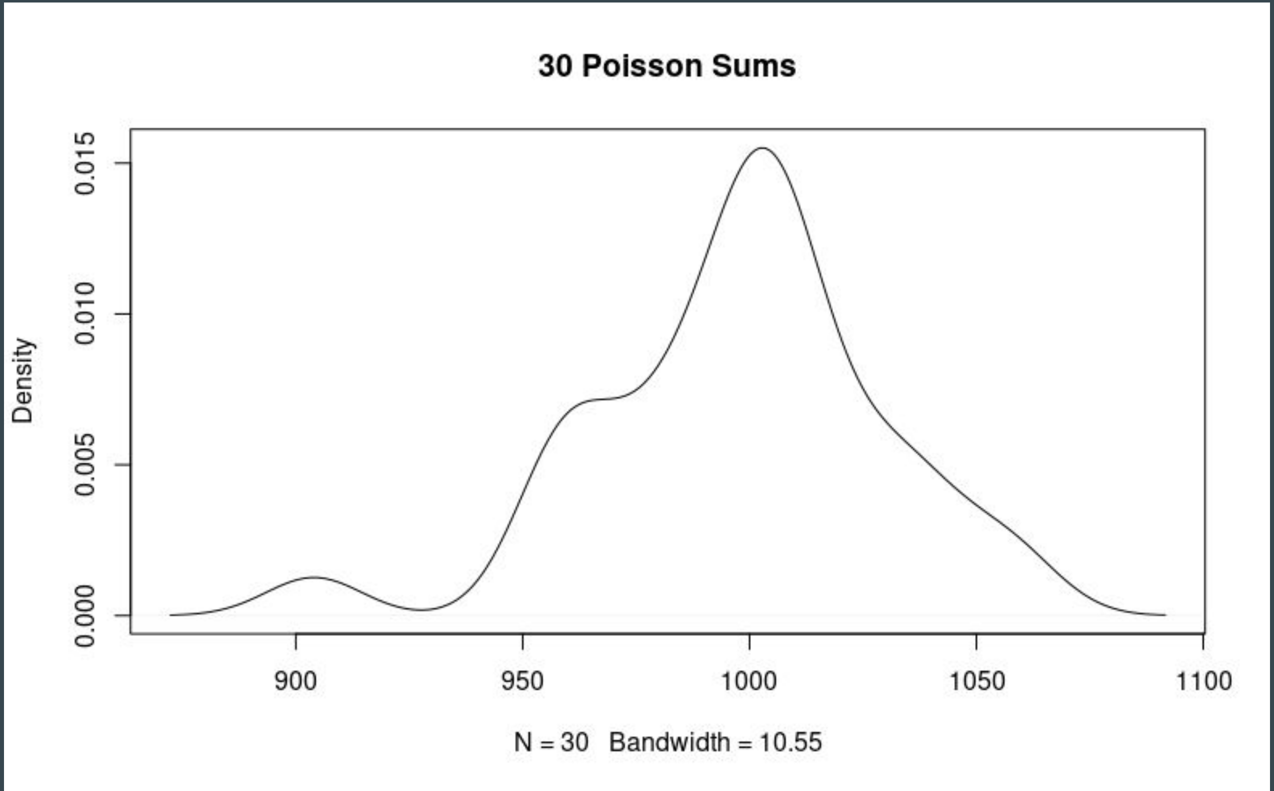
Exponential sums



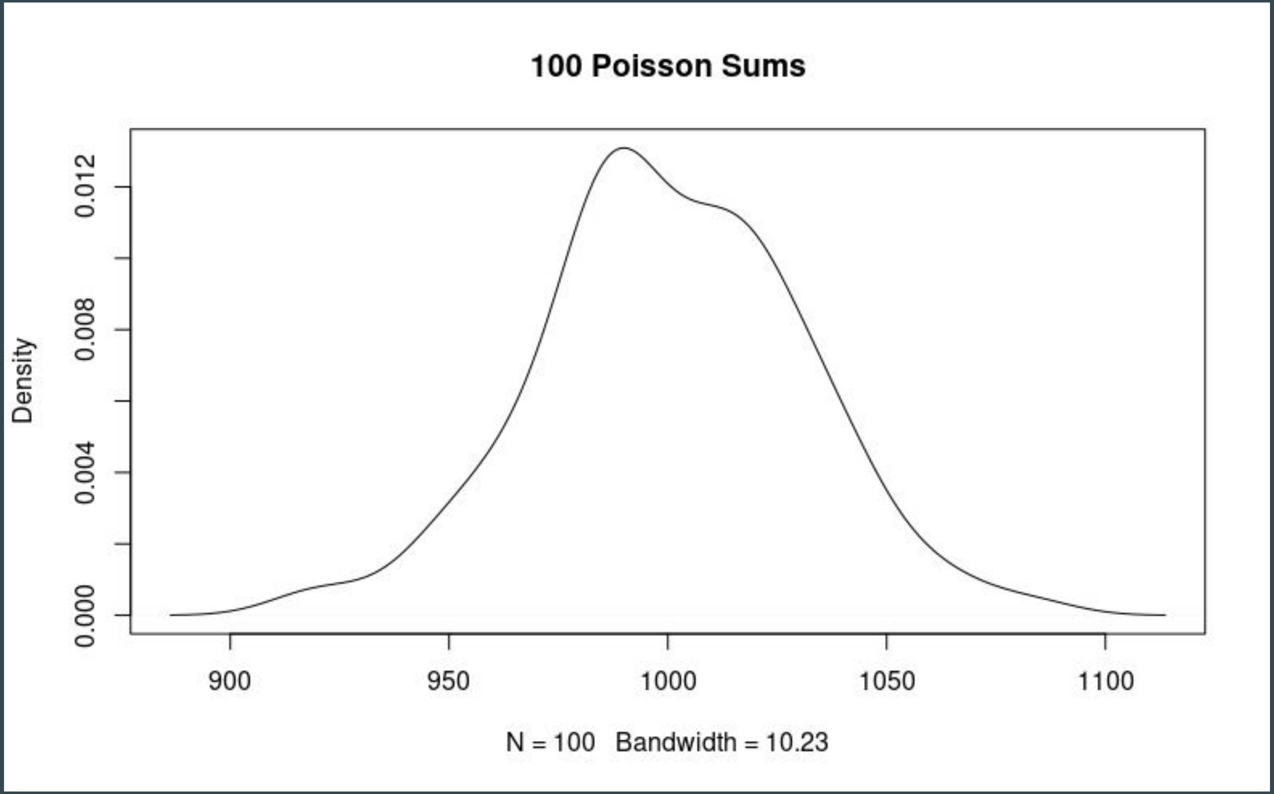
Exponential sums



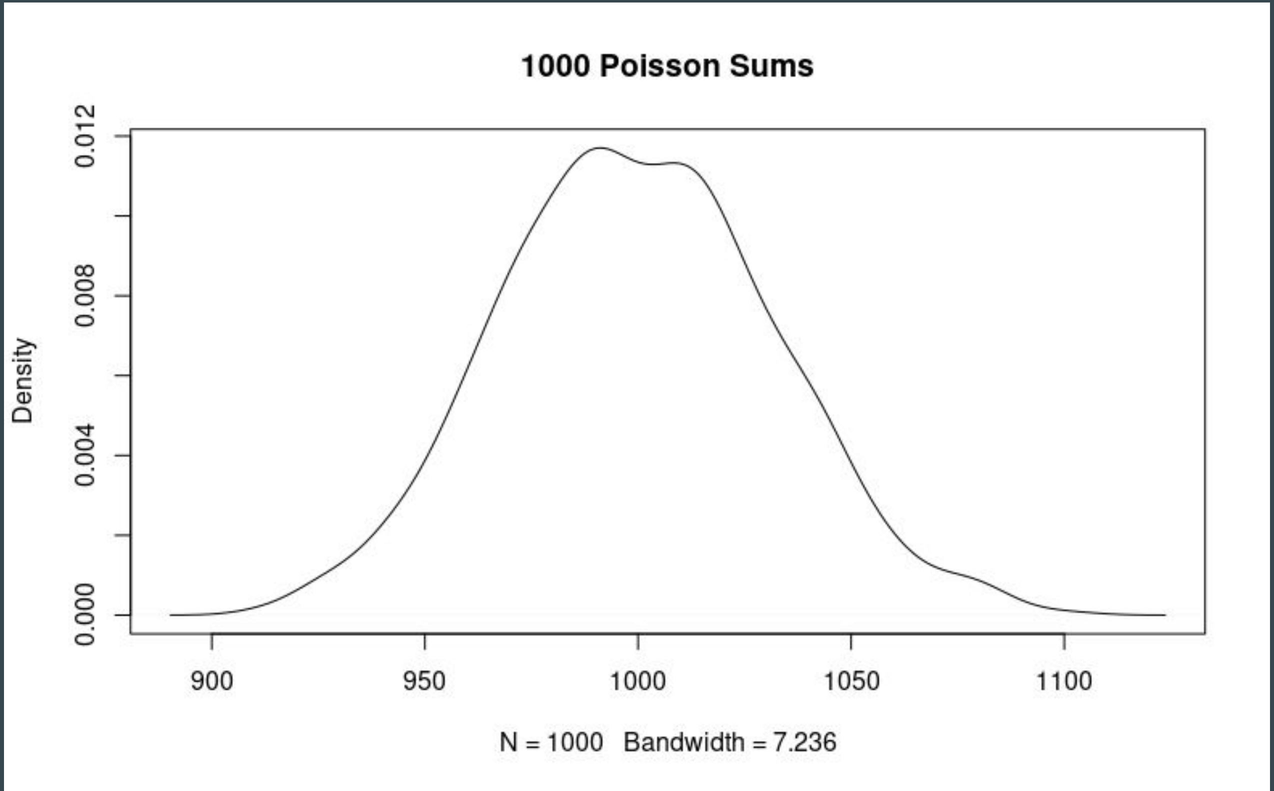
Poisson sums



Poisson sums



Poisson sums



Central Limit Theorem

For any i.i.d. variables, given a large enough sample size, any sums drawn from those variables will be normally distributed.

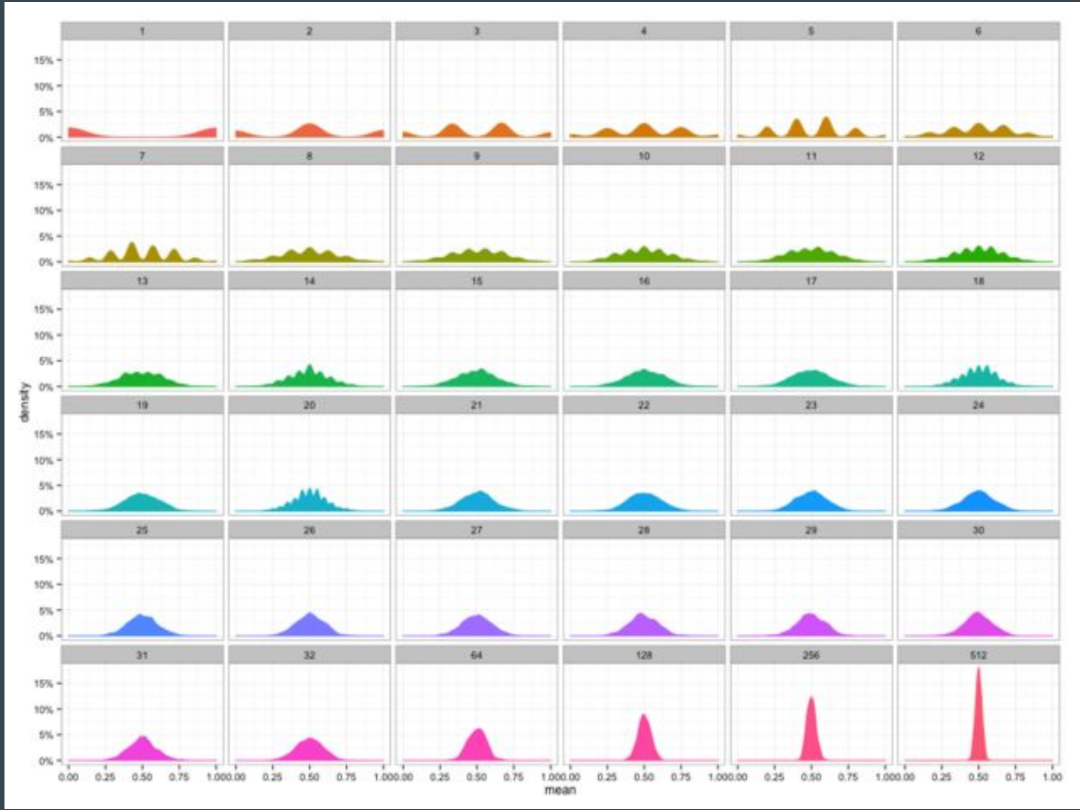
What is i.i.d.?

- Stands for Independently and Identically Distributed
  - Independent: the value of one observation does not affect the other
  - Identically distributed: all variables are generated by the same distribution
- Practically, we are way more concerned about *independence* than *identically distributed*

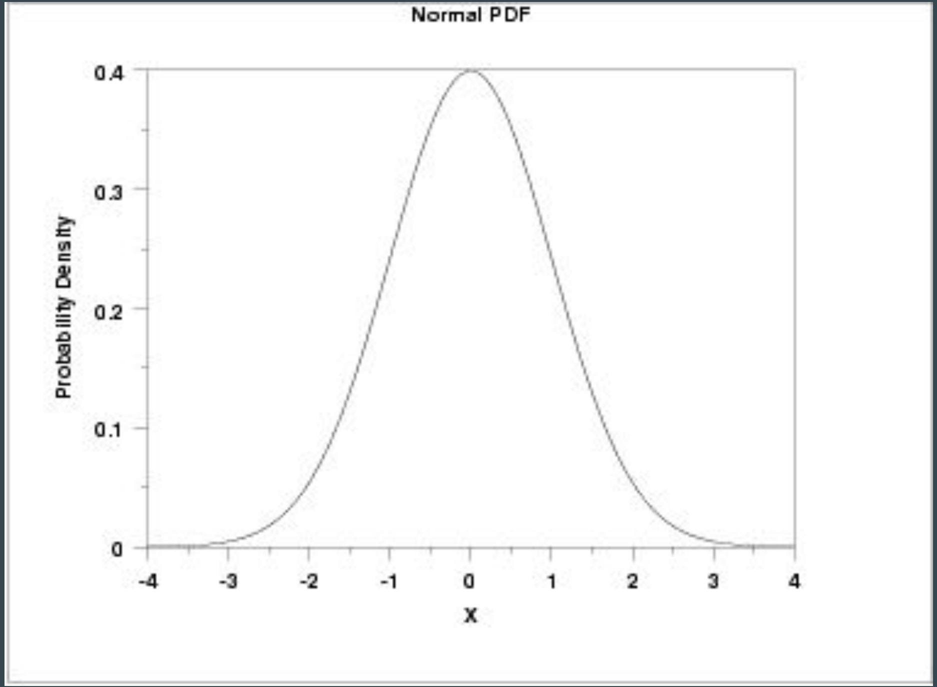
Central Limit Theorem

Given a large enough sample size, the sample means will be normally distributed.

Central Limit Theorem

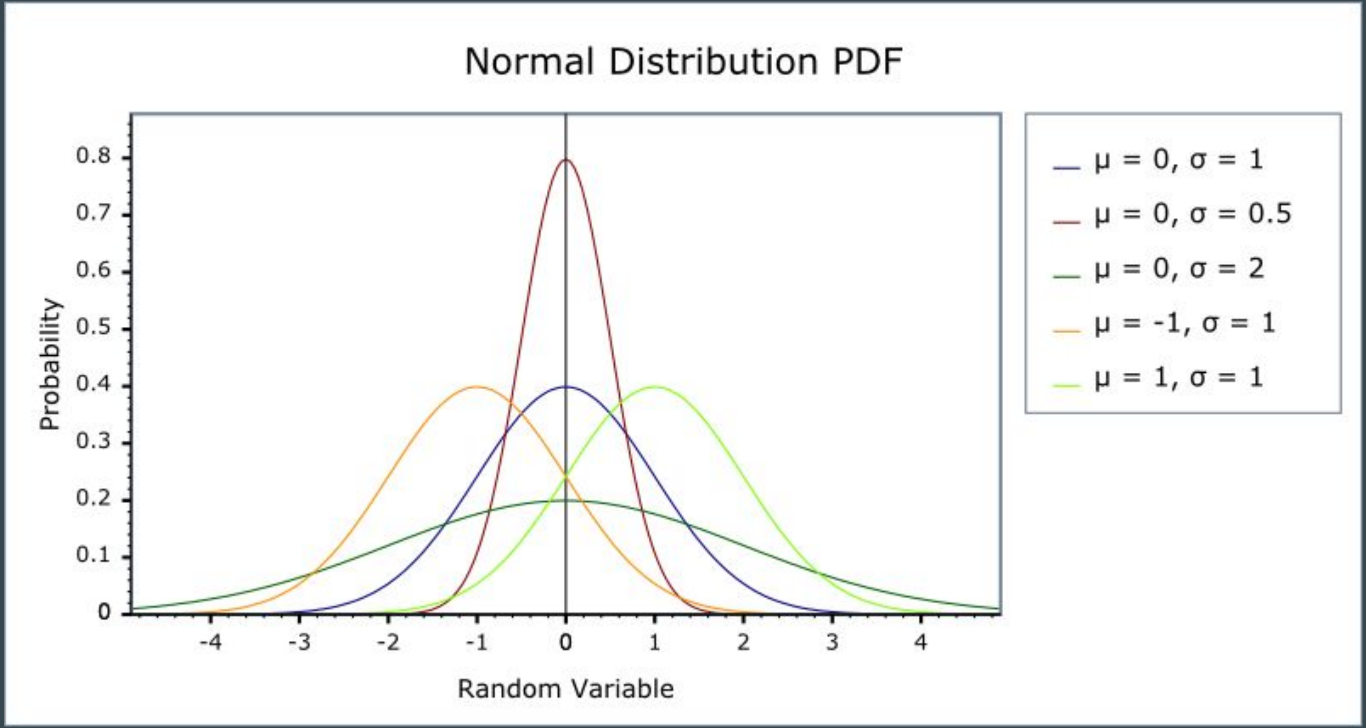


Normal distribution



$$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Changing normal parameters



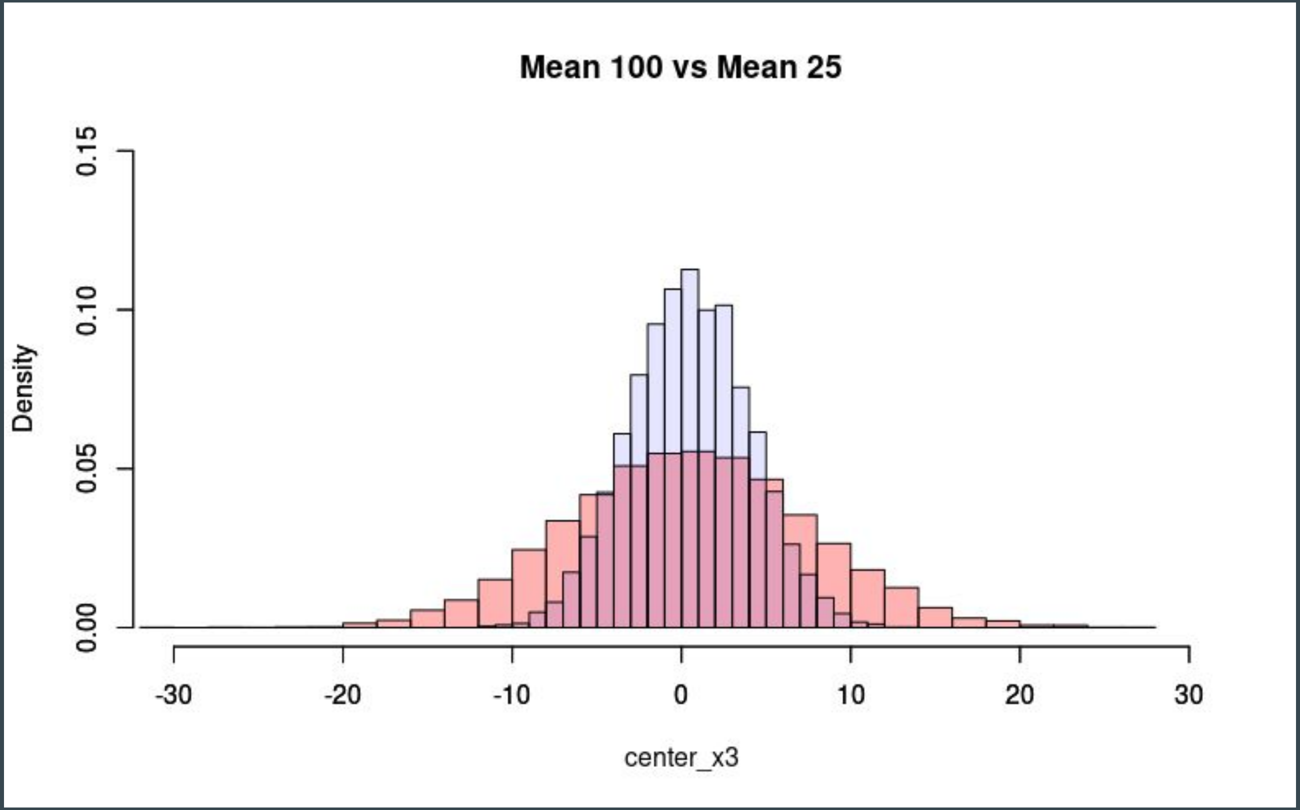
Standardizing distributions

- Take two sets of data from the same distribution
- Two parts of standardization
  - Centering data (lining up E(X))
  - Scaling data (line up the variance)
- Standardized variables have mean 1, variance 0

Centering variables

X - mean(X)

Centering variables



Scaling Variables

$$X / \mathrm{sqrt}(\mathrm{Var}(X))$$

Scaling formula / z-score

$$z = \frac{x - \mu}{\sigma}$$

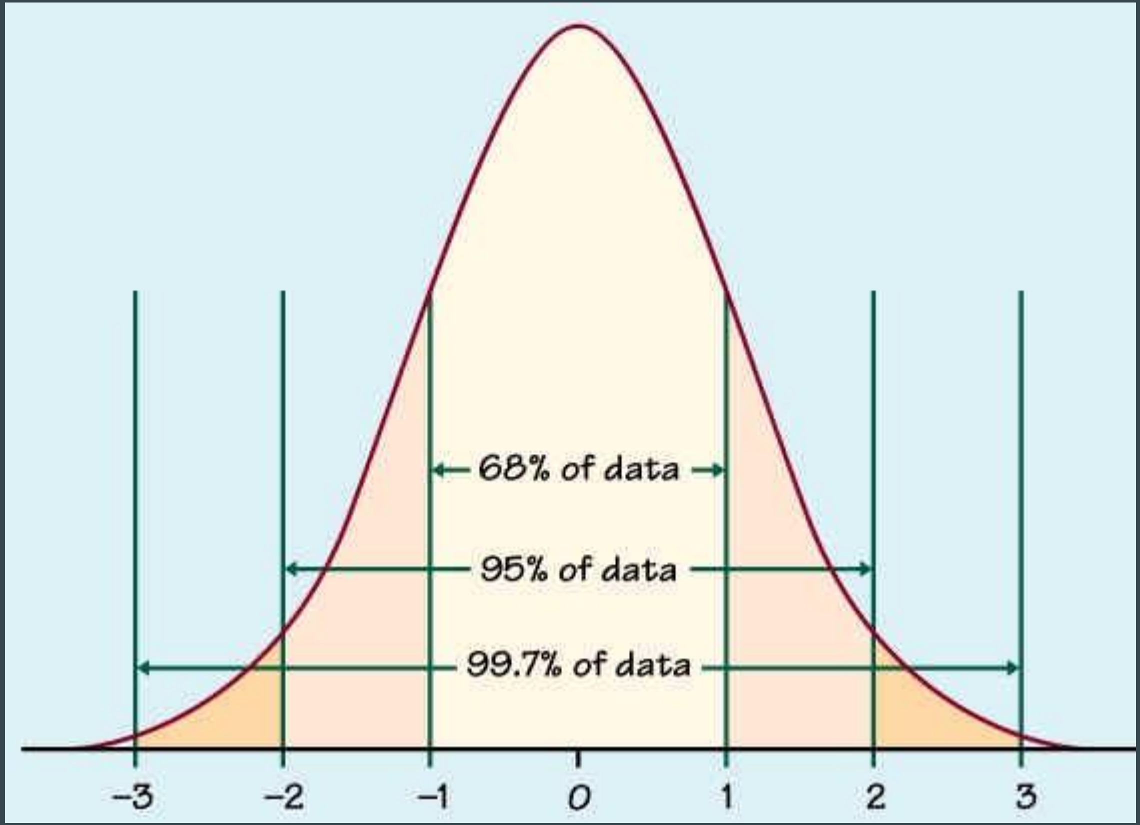
$\mu =$ Mean

$\sigma =$ Standard Deviation

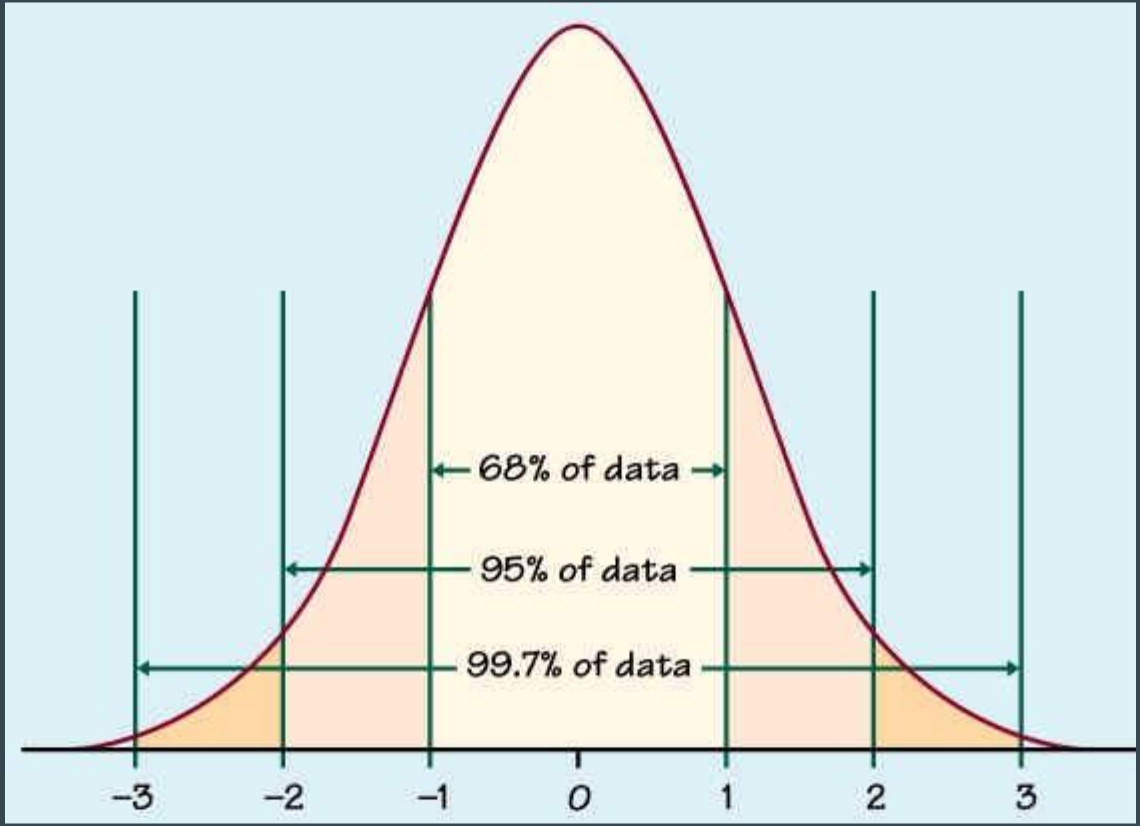
Use of z-scores

- Z-scores allow us to determine the probability of only a single observation
  - If you assume the normal distribution
- Mean of 0 and Variance of 1 provided simple interpretation

Interpretation of z-score

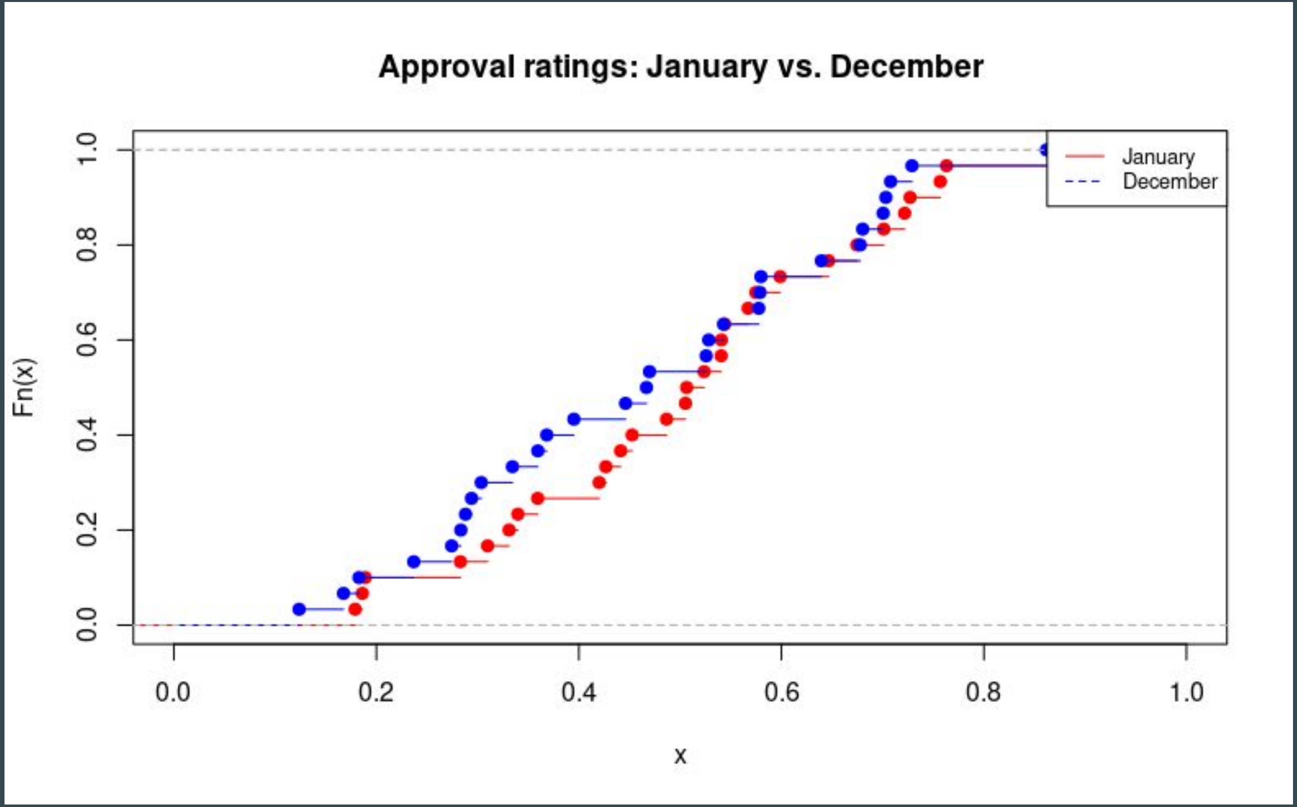


Interpretation of z-score



- $-1 < z < 1$
  - 68% chance
- $-2 < z < 2$
  - 95% chance
- $-3 < z < 3$
  - 99.7% chance

Let's take a step back



Use of the normal

- If we are interested in the mean of a sample with sufficient size all we need is…
  - Expected Value (mean)
  - Variance (standard deviation)
- …and we can assign it a probability
  - Answer questions whether something changed in the population or random chance
- We don’t need to know the distribution of our data, or anything else about the population

Z-scores in action

After opening a new school, a city wants to determine if the new school is performing well. The average school in the district has a graduation rate of 74%, with a standard deviation of 8%. The new school's graduation rate is 82%.

$$z = x - mean(x) / stdev(x)$$

$$z = (.82 - .74) / .08 = -1$$

Review

- Assuming i.i.d, any type of sums drawn from any type of distribution will be normally distributed, given a large enough sample
- Z-scores rescale distributions to have:
  - $E(X) = 0$
  - $Var(X) = 1$
- Z-scores allow us to determine probability *without* knowing what distribution it came from