Distributions and functions

Rachel 24.01.23

Teacup giraffes

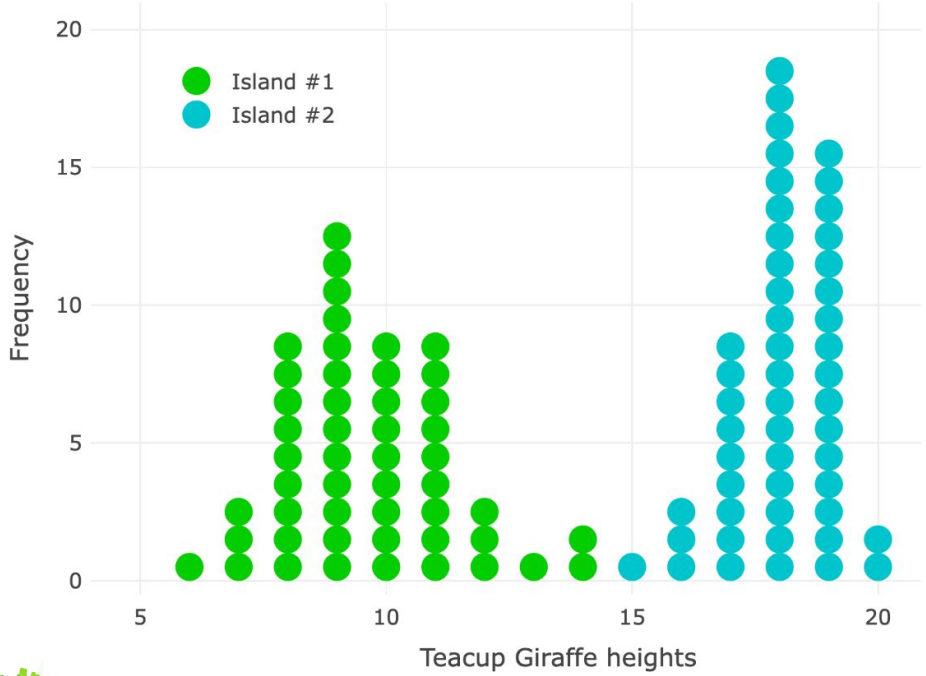
Imagine we've collected data for two populations of tiny giraffes that live on two different islands.

What is a distribution?

A distribution

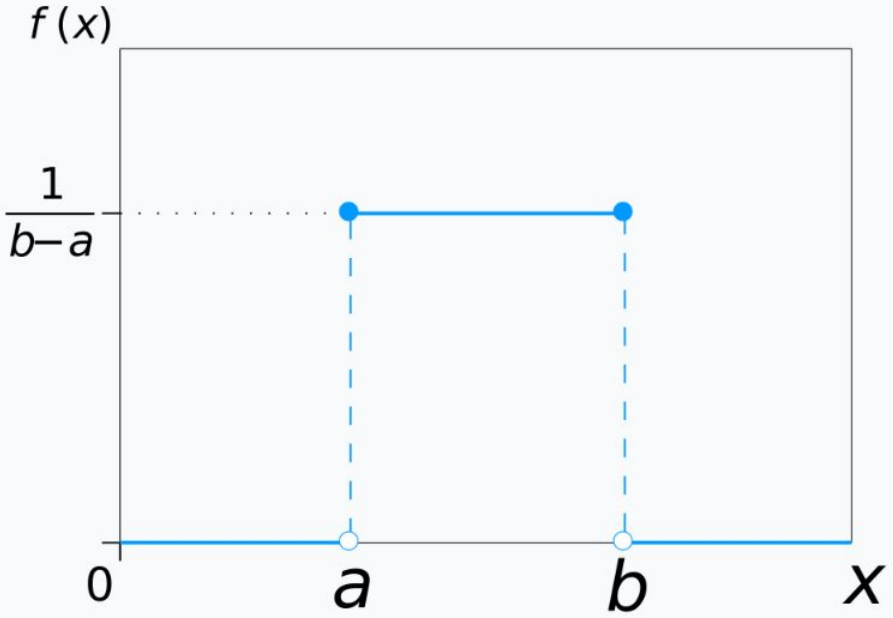
- Shows the values the variable (e.g., height) takes on in your data set
- How often each value occurs
- The **shape**, center, and **amount of variability** in the data

The first step of any analysis is often to **visualise the data**.

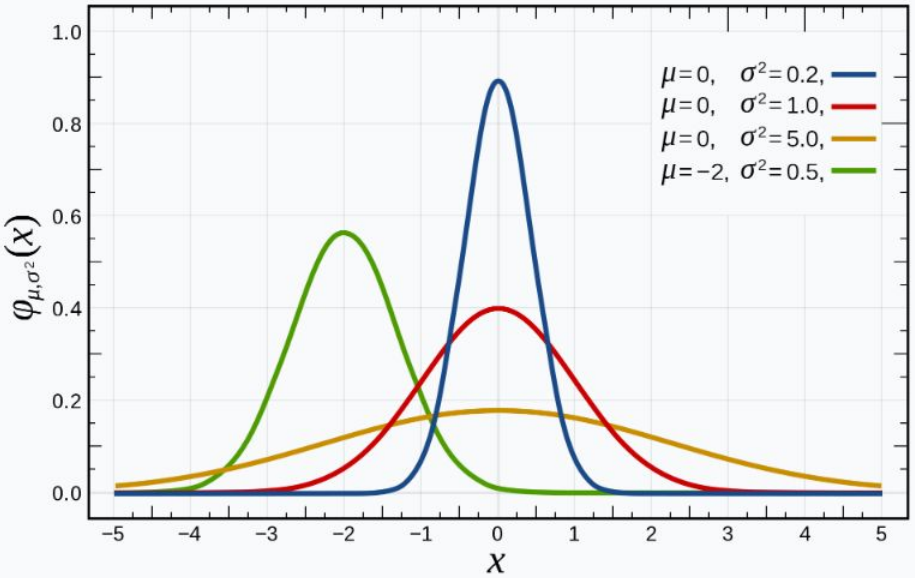


Can you name any distributions?

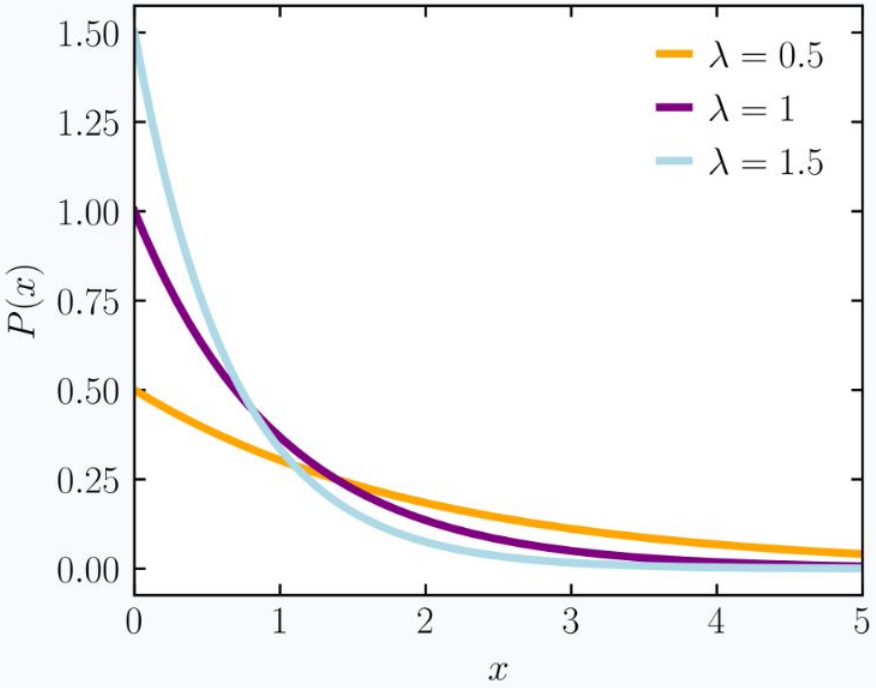
Can you name this distribution?



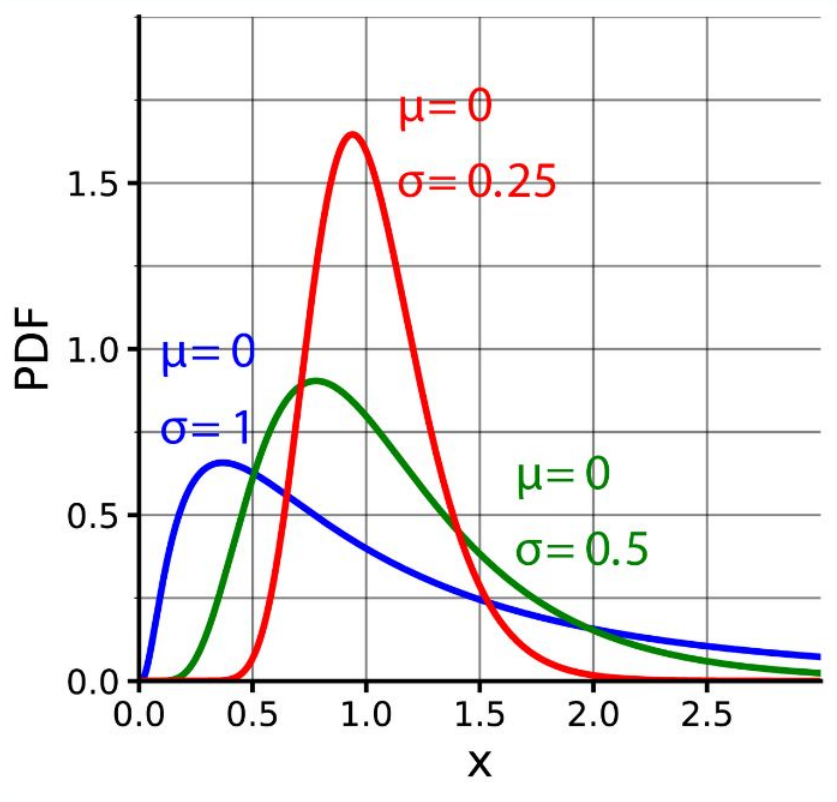
Can you name this distribution?

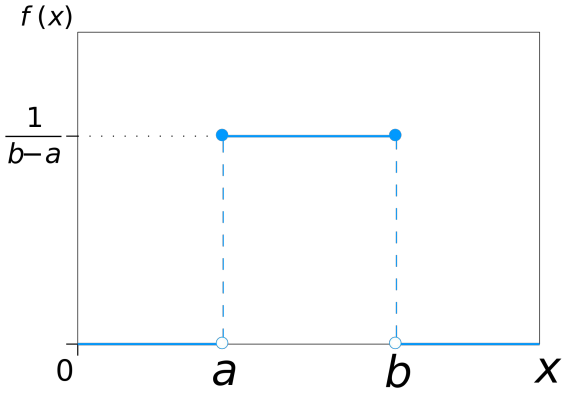


Can you name this distribution?

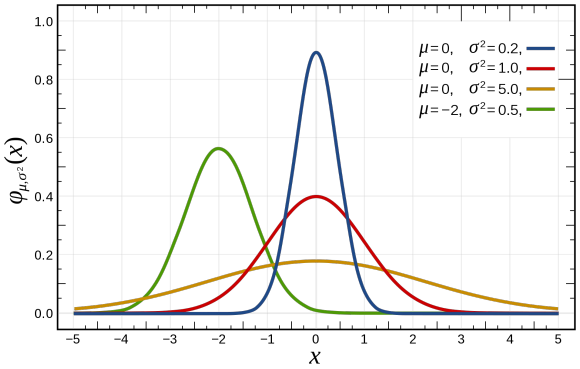


Can you name this distribution?

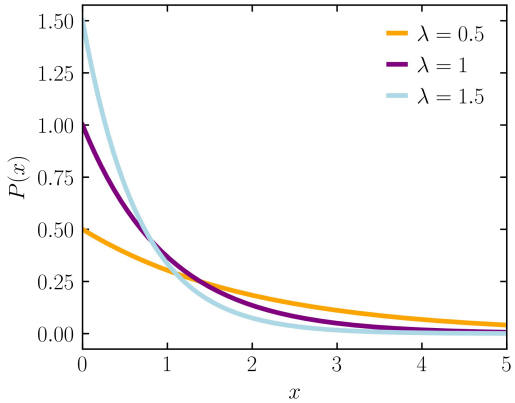




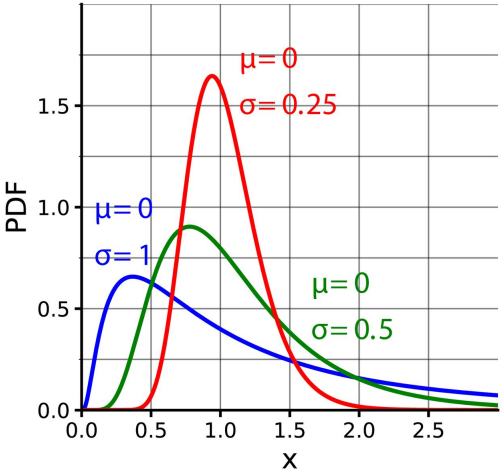
Uniform



Normal

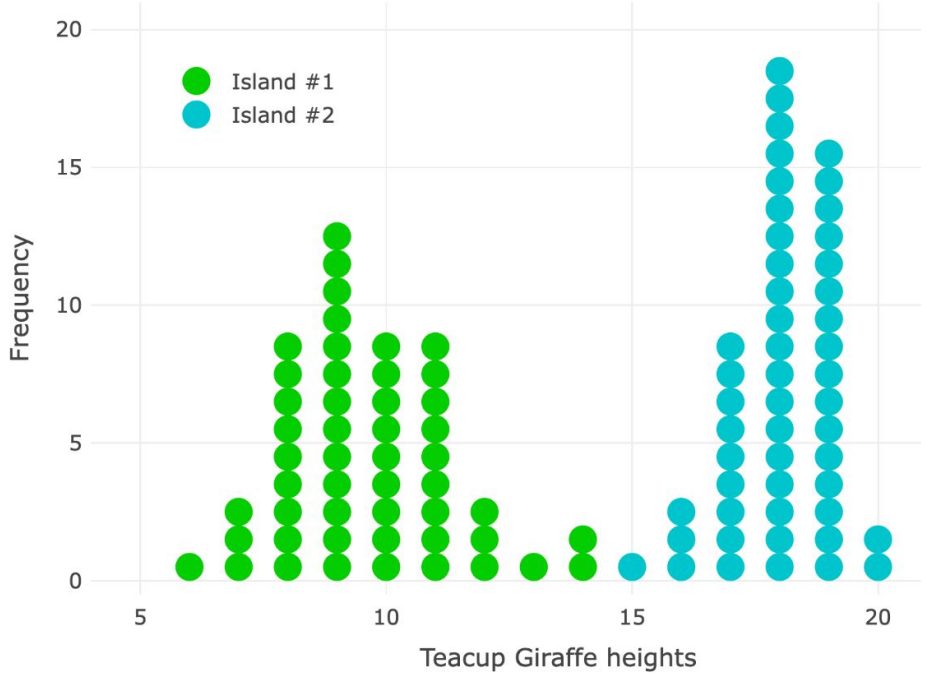


Exponential



Lognormal

What distribution provides a good description of our giraffe data?

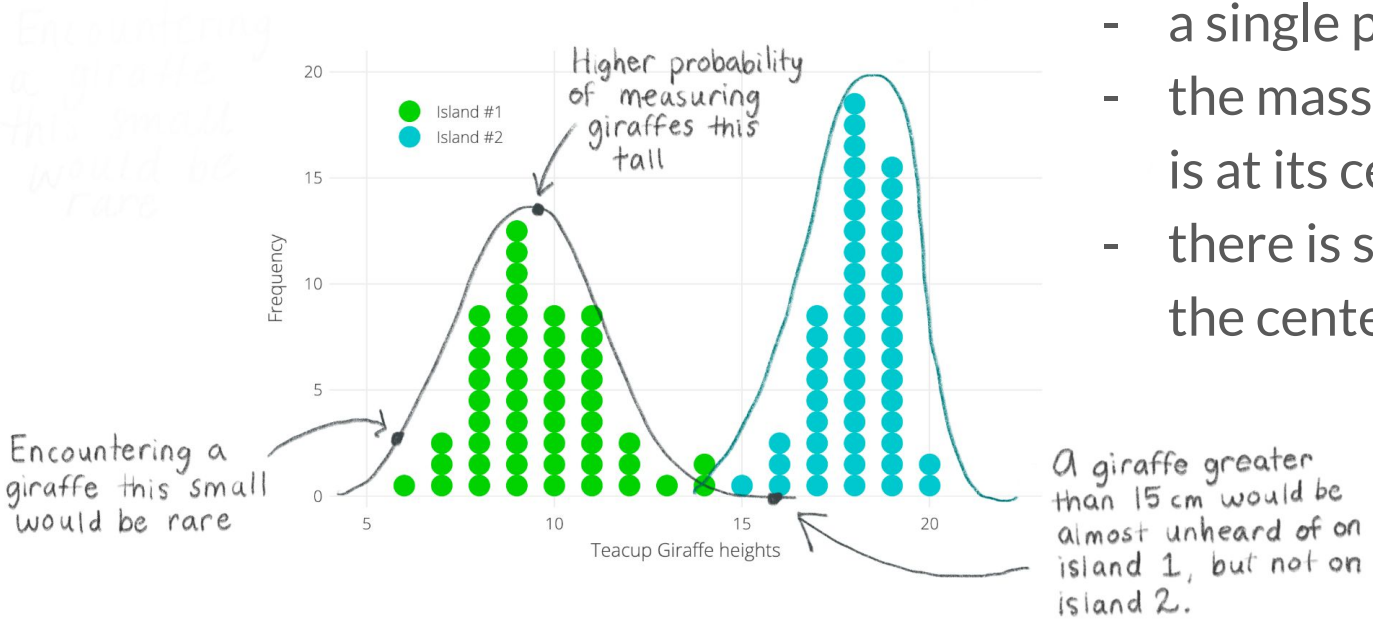


The normal distribution

Distributions of data can take on many shapes but there are some general shapes that occur frequently in nature.

The normal distribution is one of the most well-known. Also known as the **Gaussian distribution** or a **bell curve**.

Characteristics of the normal distribution

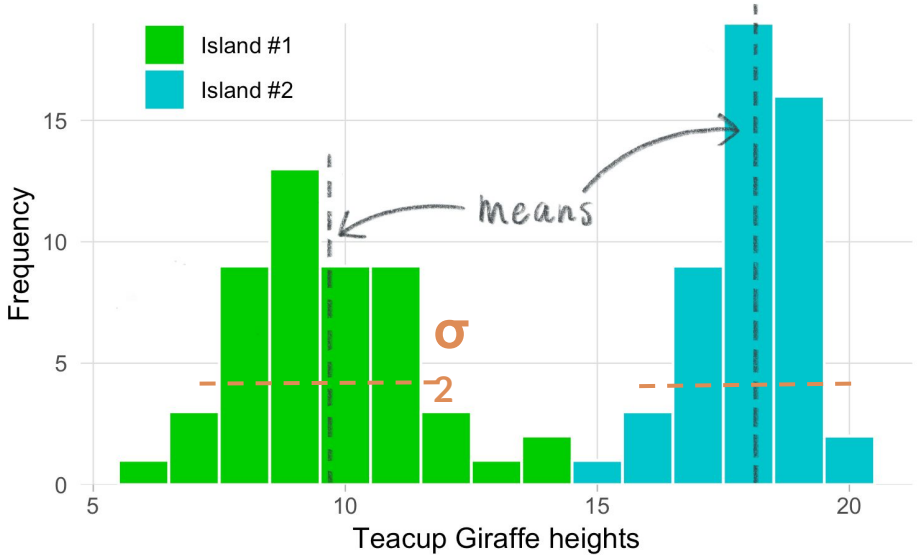


- a single peak
- the mass of the distribution is at its center
- there is symmetry about the centerline

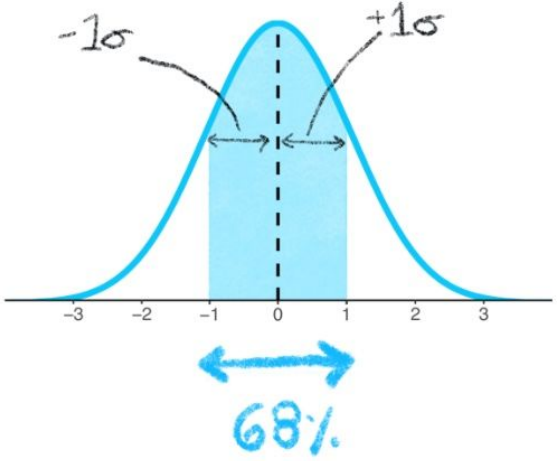
Parameters of the normal distribution

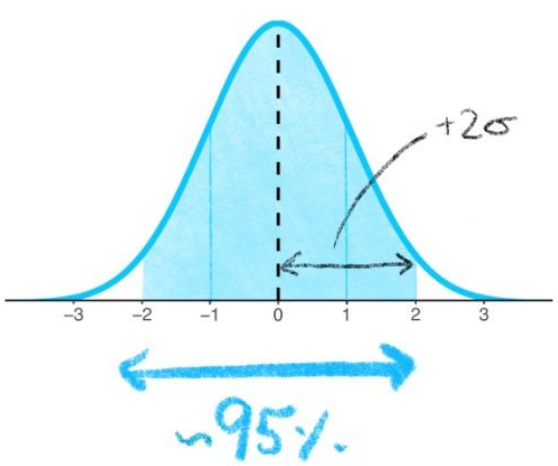
μ - the mean or expectation

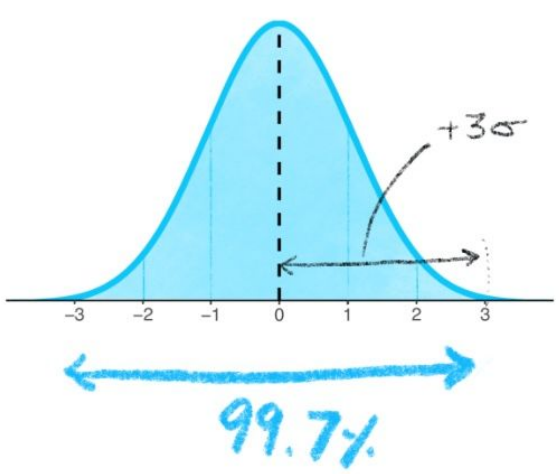
σ - the standard deviation
(or σ^2 - the variance)



Standard deviation measures the spread of the data







More about this next week.

Tools for exploring the normal distribution

The normal distribution

Compare two normal distributions

Continuous distributions in R

Are associated with 4 standard functions:

`dnorm(x, mean = 0, sd = 1)` - probability density function

`pnorm(q, mean = 0, sd = 1)` - cumulative distribution function (% of values < than q)

`qnorm(p, mean = 0, sd = 1)` - quantile function (inverse of cumulative distribution)

`rnorm(n, mean = 0, sd = 1)` - generates random numbers

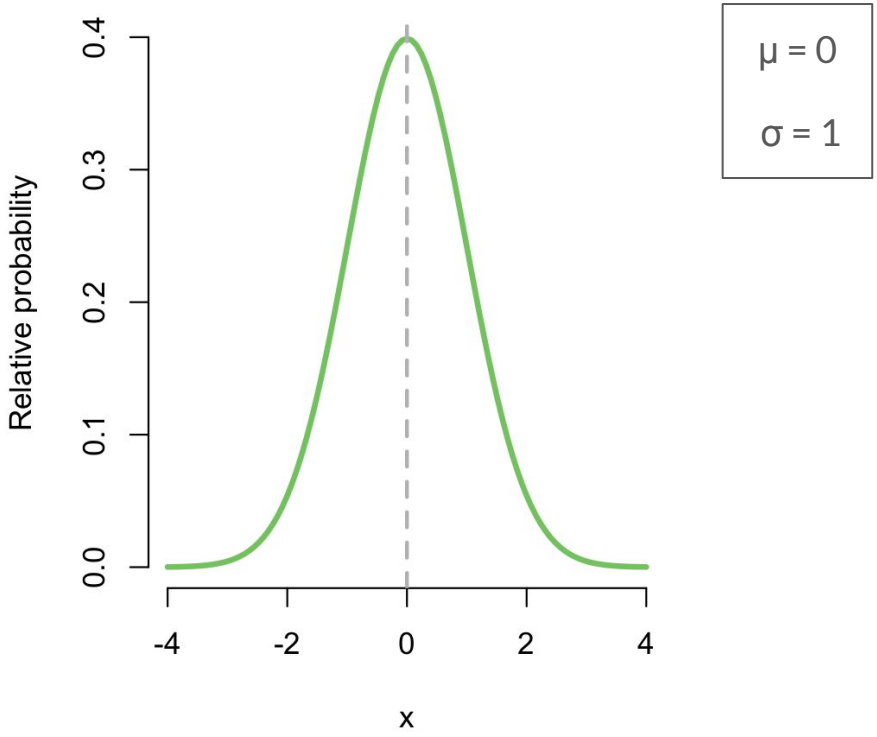
Let's see these in action!

`dnorm()` - probability density function

Describes the probability of a value at any point along the x-axis*.

This function can be used to draw the distribution.

The dashed line shows the mean.

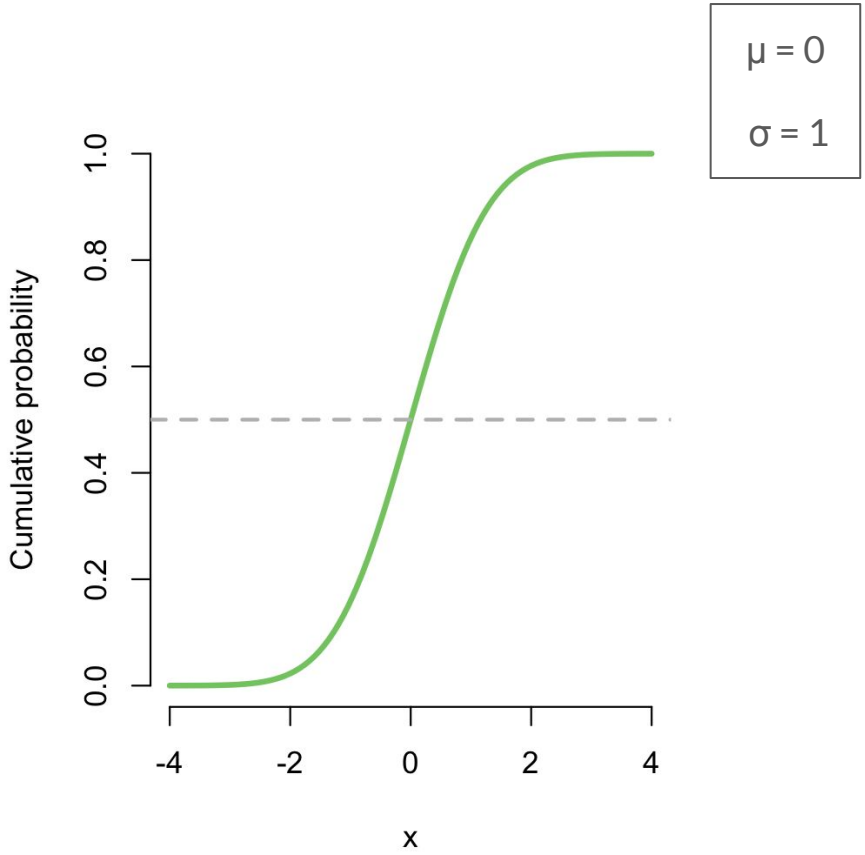


*Actually the probability of being a precise value can very small, so it's sometimes more useful to think of the probability of being within a range, e.g. the probability of being between 1 and 2.

`pnorm()` - cumulative distribution function

The probability that a value will be less than or equal to *x*.

The dashed line shows the cumulative probability = 0.5 at 0, the mean of this distribution. i.e., 50% of values ≤ 0.



`qnorm()` - quantile function (inverse of qnorm)

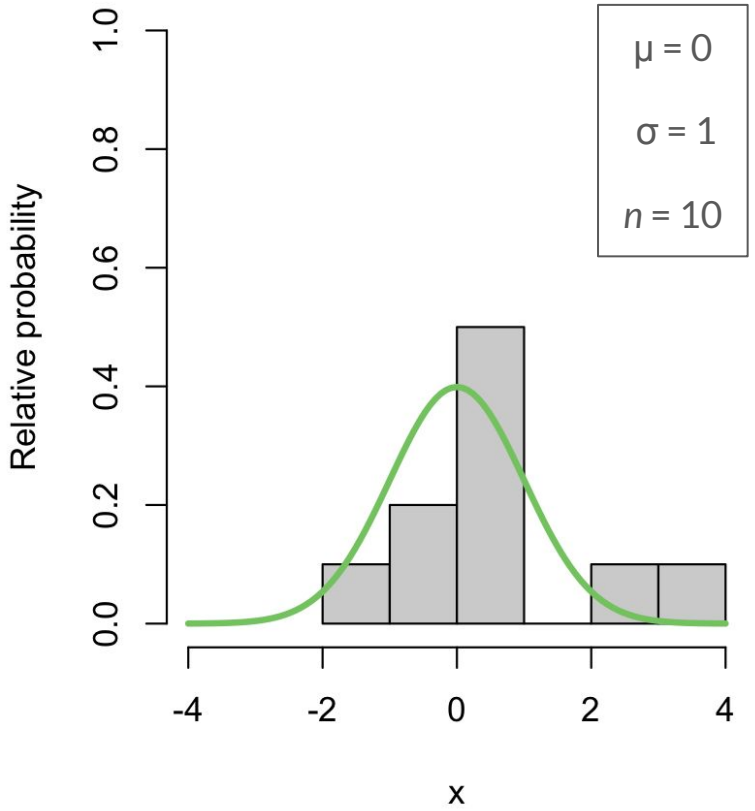
It gives you the value of *x* at a given quantile, i.e., at a given cumulative probability.

`rnorm()` - generates "pseudo" random numbers

The are lots of reasons we might want to generate random numbers.

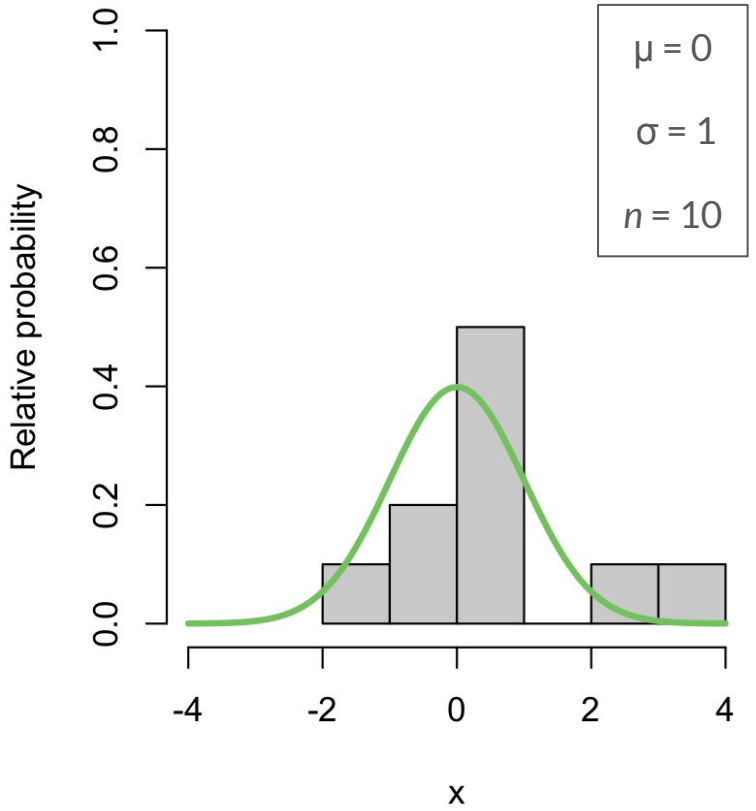
This plot shows random draws (*n*) from a normal distribution as a histogram.

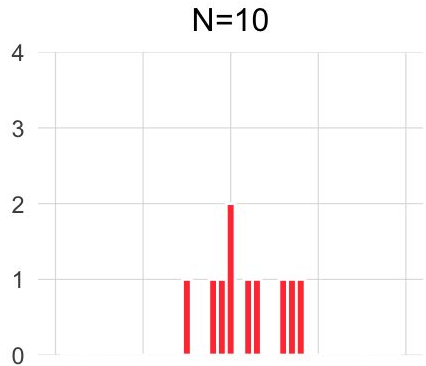
Why is it pseudo random?

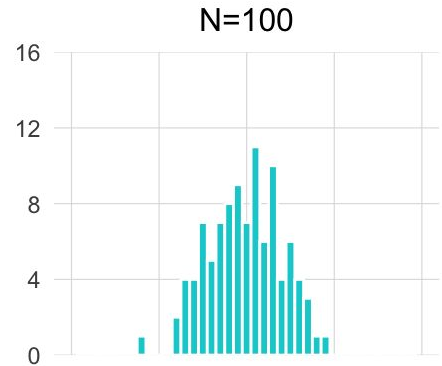


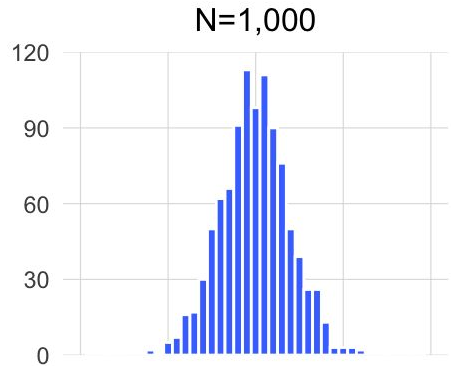
What do you predict will happen if we increase the number of random draws?

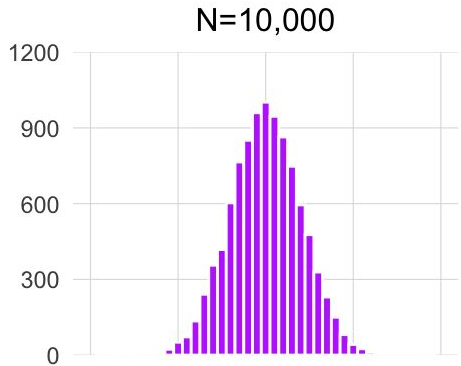
e.g., $n = 100$, $n = 1000$ etc.




N=10
4
3
2
1
0


N=100
16
12
8
4
0

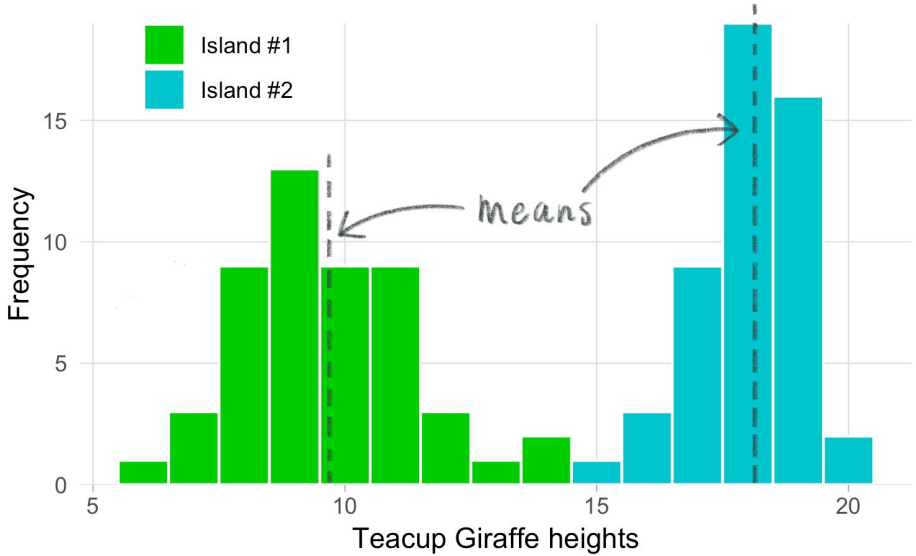

N=1,000
120
90
60
30
0

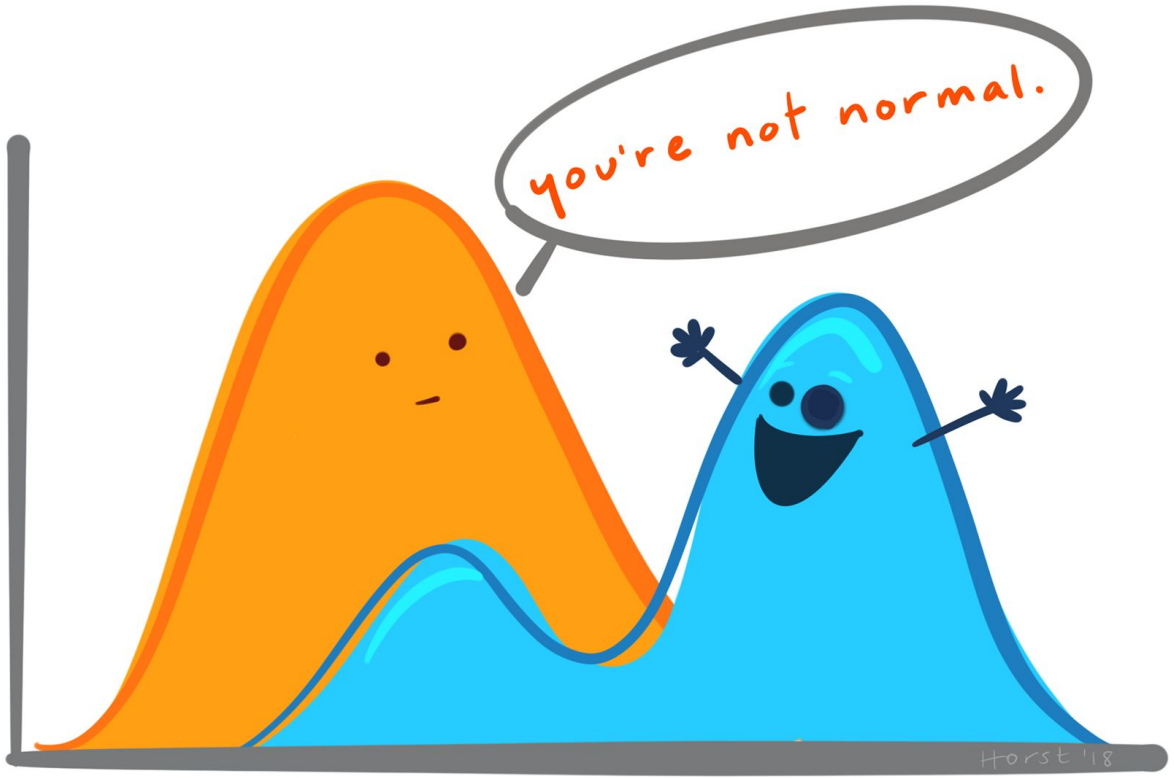

N=10,000
1200
900
600
300
0

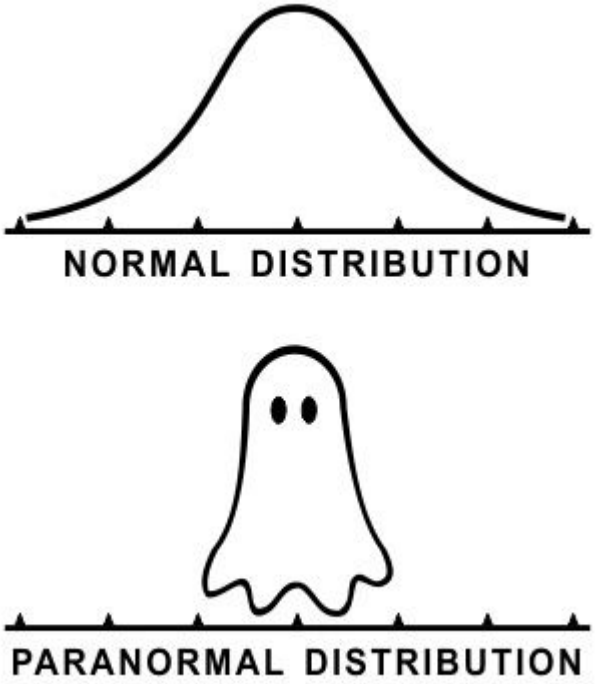
Homework

Write a **function** that calculates the mean for a **vector**.

Next week, we'll talk more about the standard deviation / variance.




you're not normal.
Horst '18


NORMAL DISTRIBUTION
PARANORMAL DISTRIBUTION

CONTINUOUS

measured data, can have ∞ values within possible range.

I AM 3.1" TALL

I WEIGH 34.16 grams

DISCRETE

OBSERVATIONS CAN ONLY EXIST AT LIMITED VALUES, OFTEN COUNTS.

I HAVE 8 LEGS *and* 4 SPOTS!

@allison_horst


are your
summary statistics
hiding something
interesting?
@allison_horst