Consolidation of R programming skills

Distributions, functions, and variance

Rachel Warnock

19.01.2026

Recap: Data types


NOMINAL
unordered descriptions
I'm a TURTLE!
i'm a snail!
I'm a butterfly!
ORDINAL
ordered descriptions
I am unhappy.
I am O.K.
I am AWESOME!!!
BINARY
ONLY 2 MUTUALLY exclusive outcomes
I am EXTINCT!
HA.
@allison_horst

Part 1
Distributions and functions

Learn more about the tiny giraffes @ tinystats.github.io

Teacup giraffes

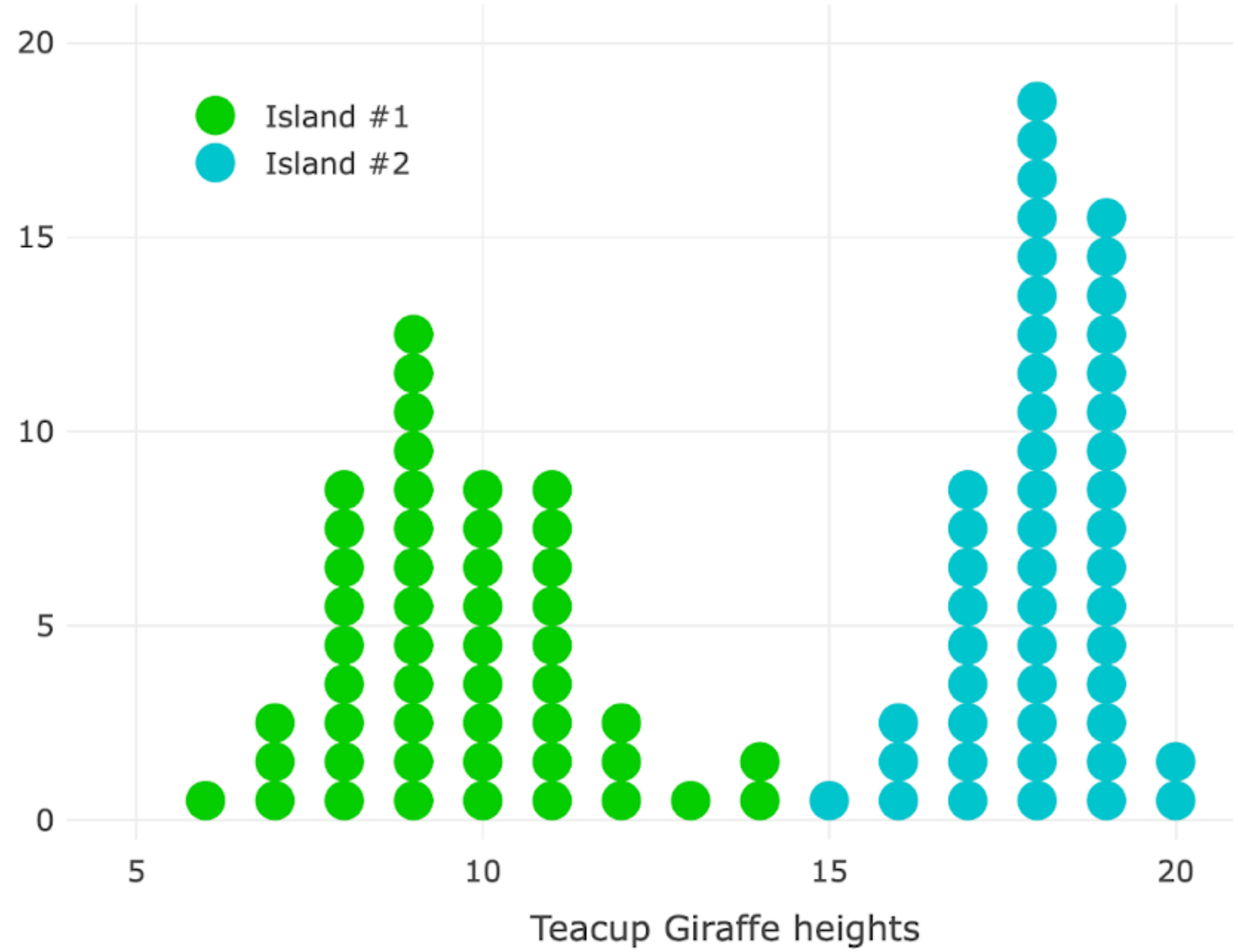
Imagine we've collected data for two populations that live on two different islands, like the tiny giraffes

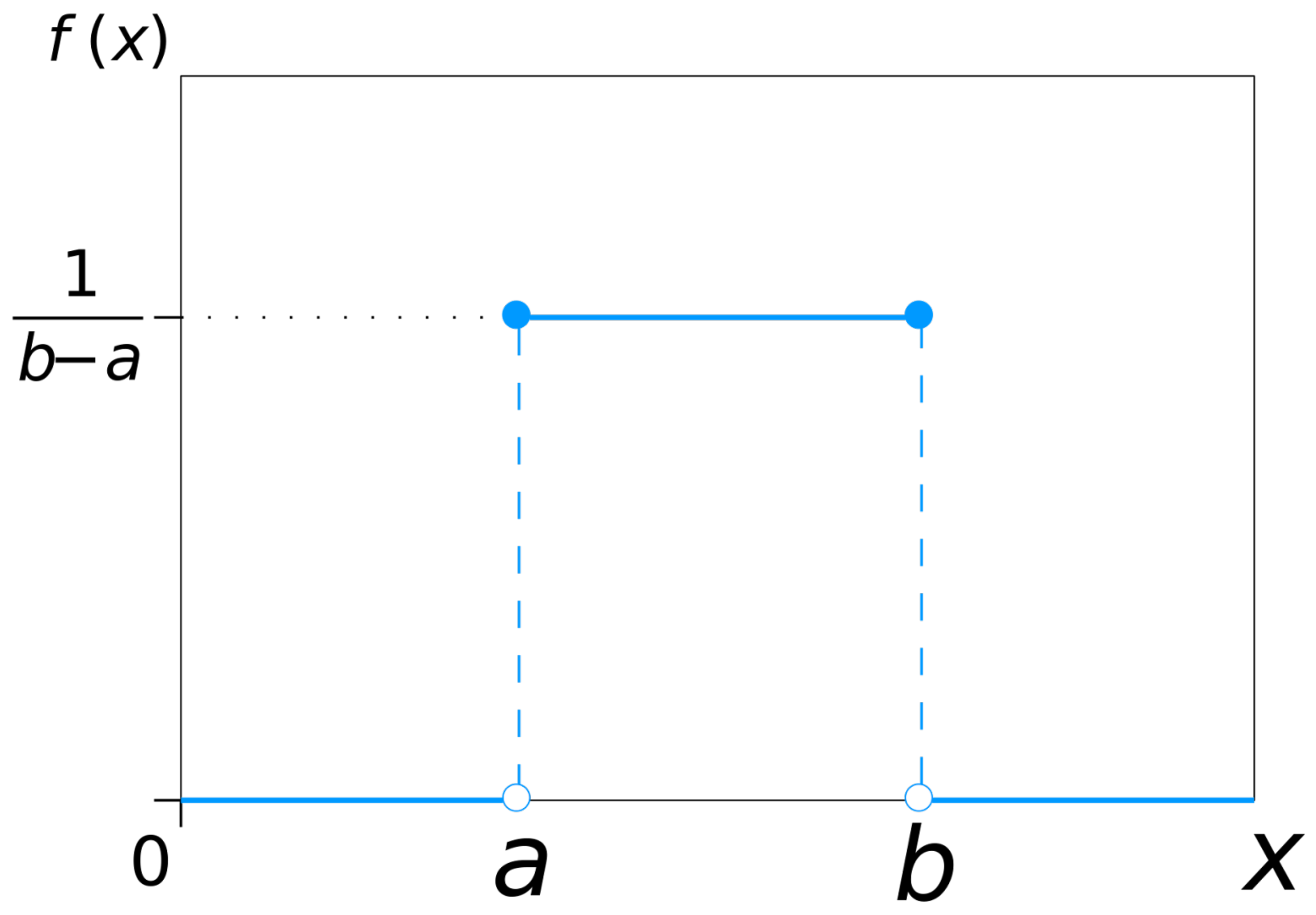
A distribution

Shows the range of values of your variable (e.g., height)

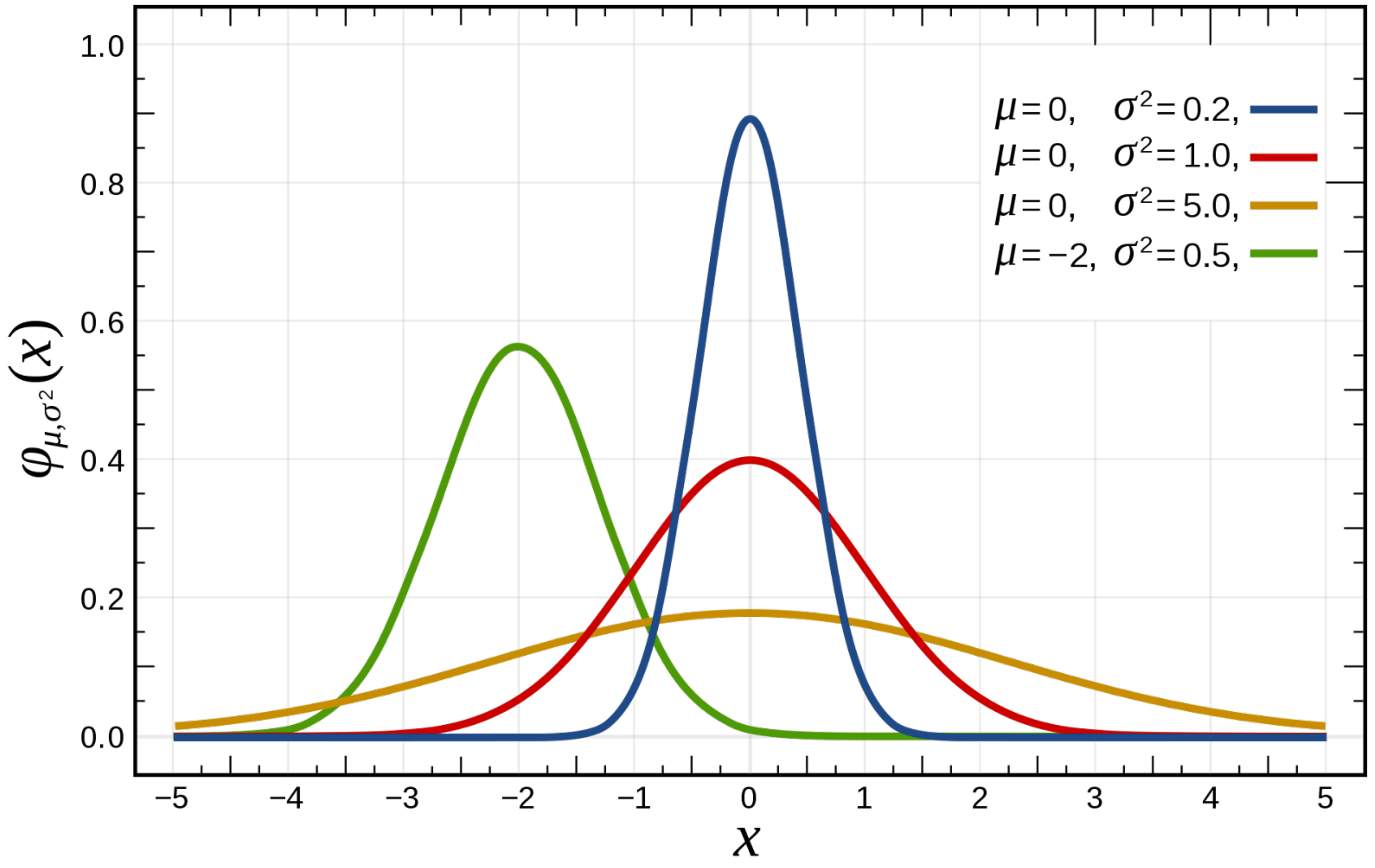
It captures: how often each value occurs

+ The shape, centre, and amount of variability in the data

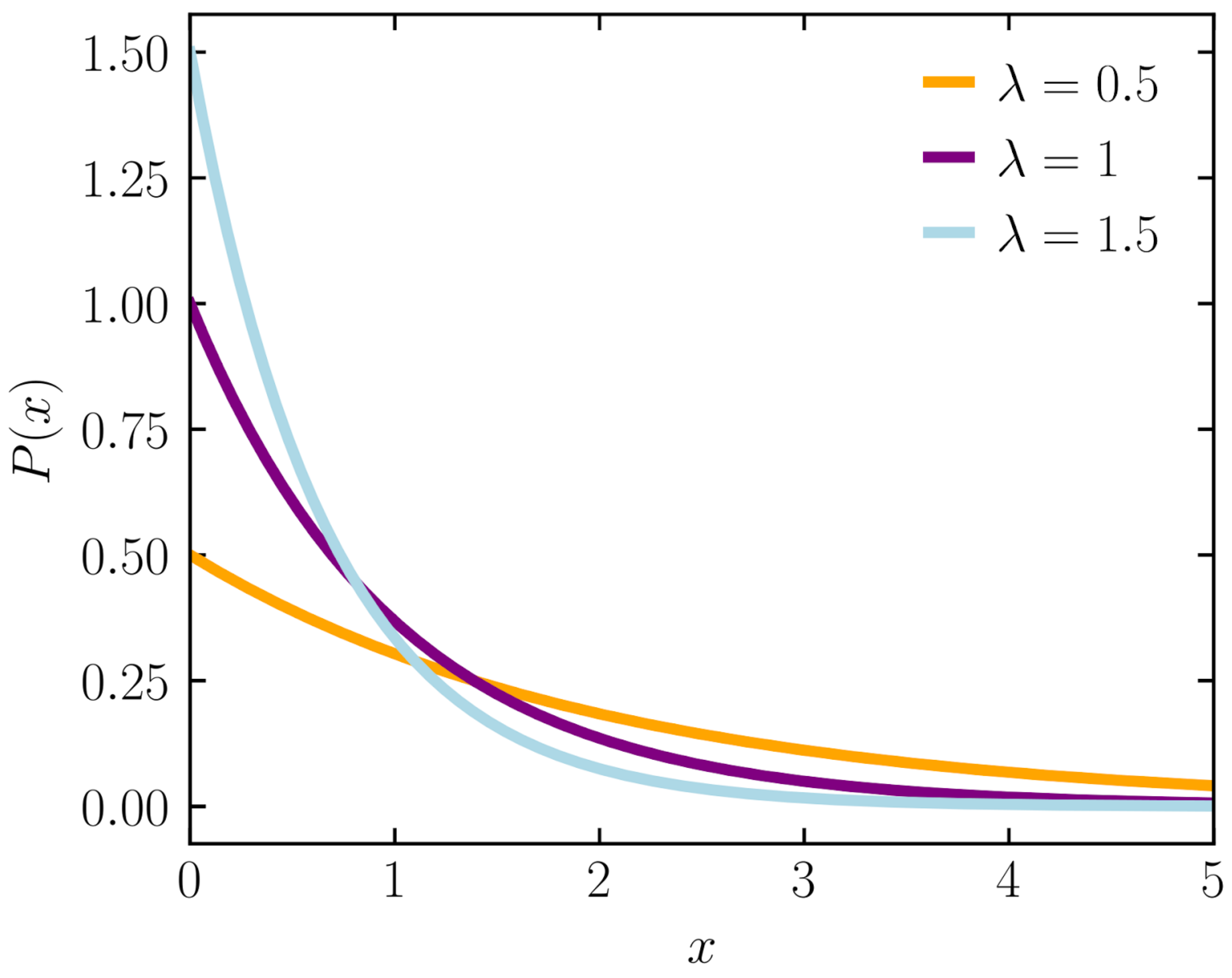




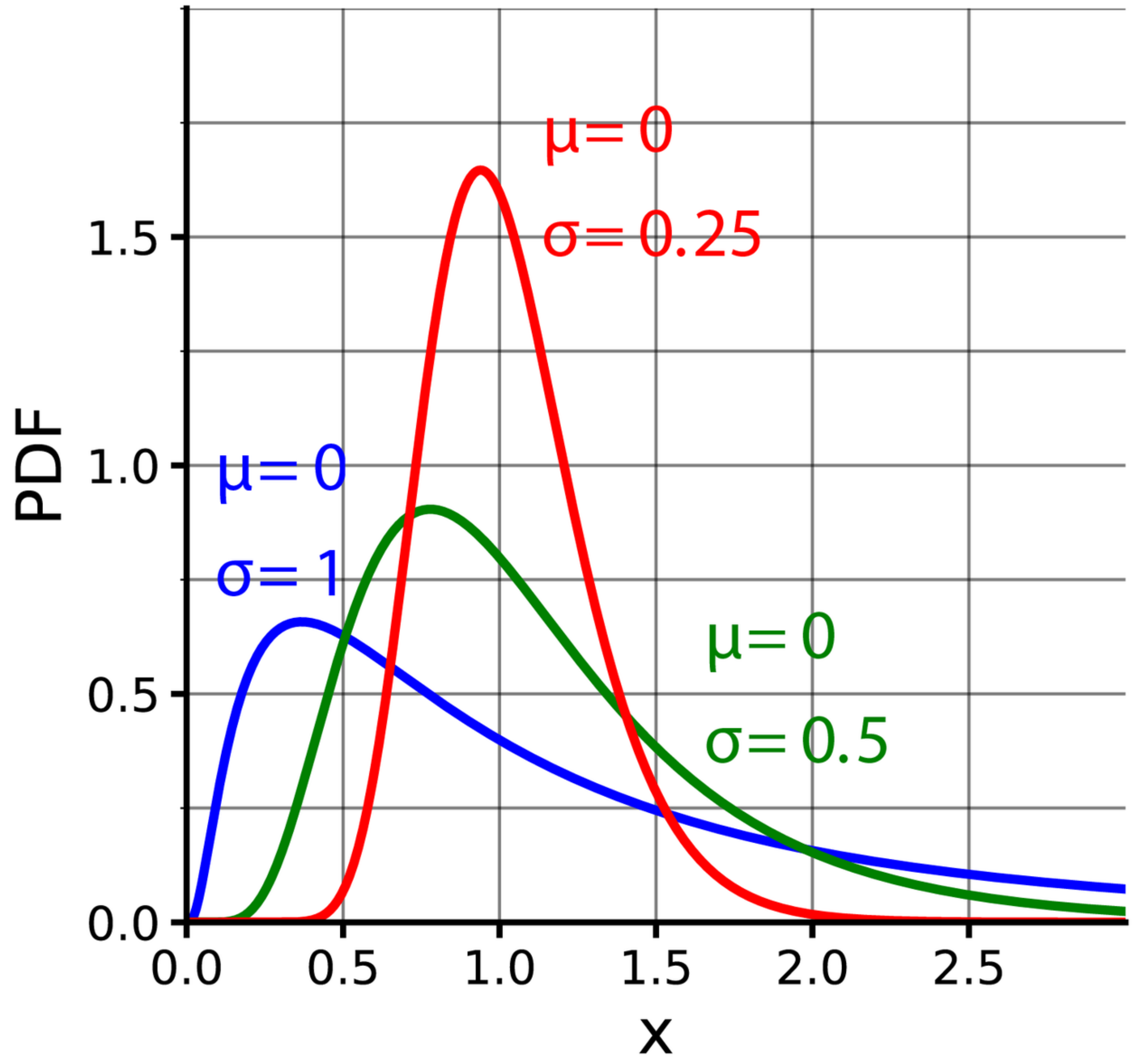
Uniform



Normal



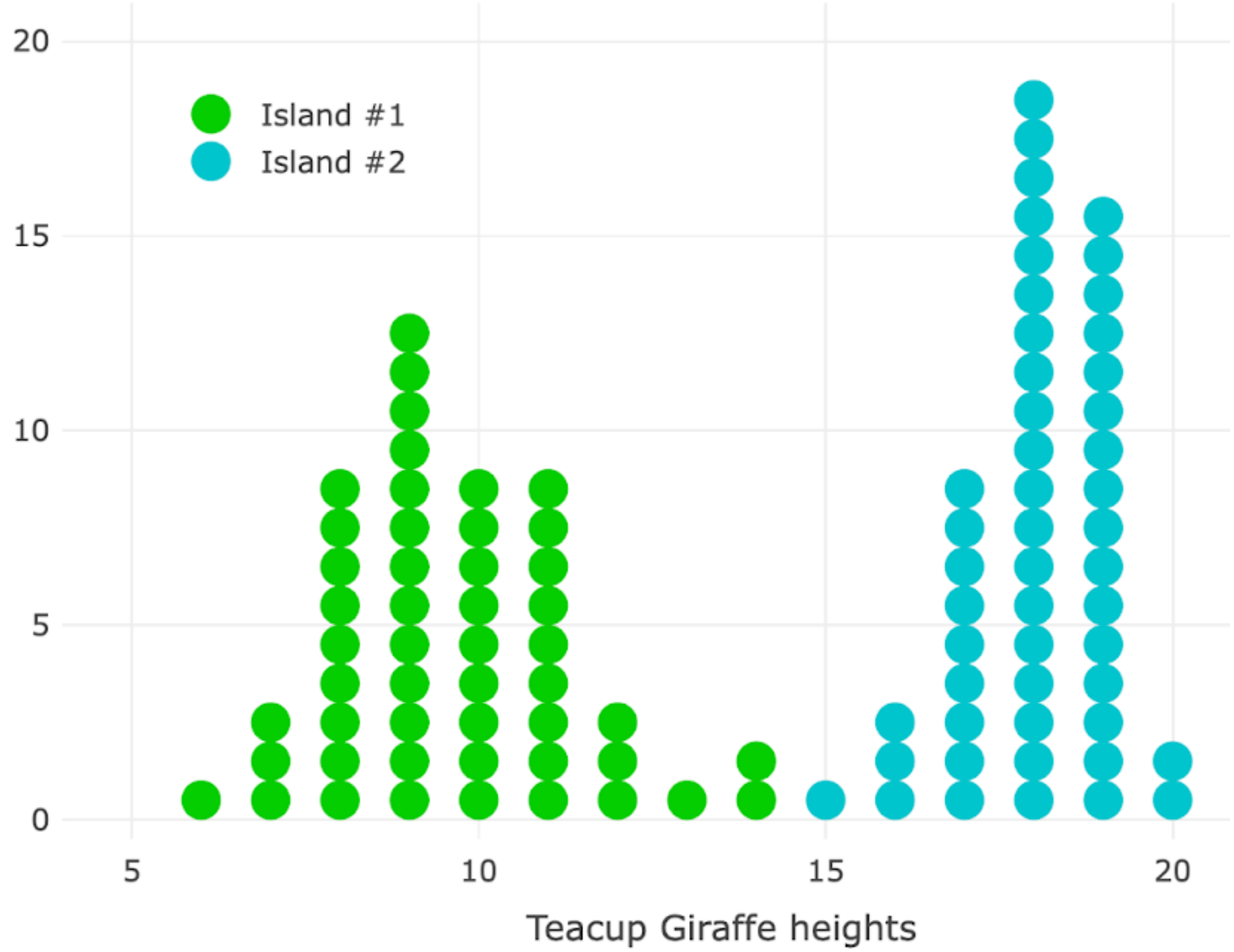
Exponential



Lognormal


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

What distribution provides a good description of our giraffe data?



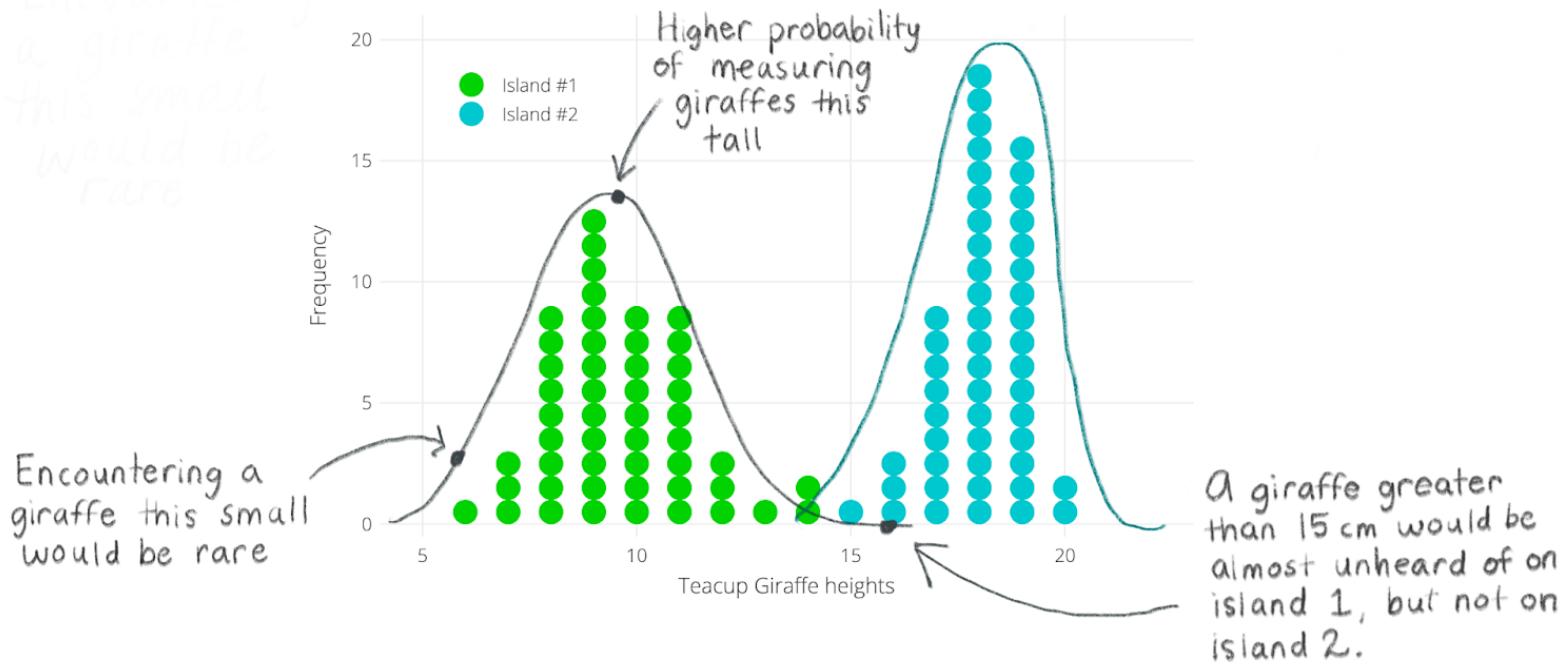
The normal distribution

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 a **Gaussian distribution** or a **bell curve**.

Characteristics of the normal distribution:

- a single peak
- the mass of the distribution is at its centre
- **symmetry** about the centre line



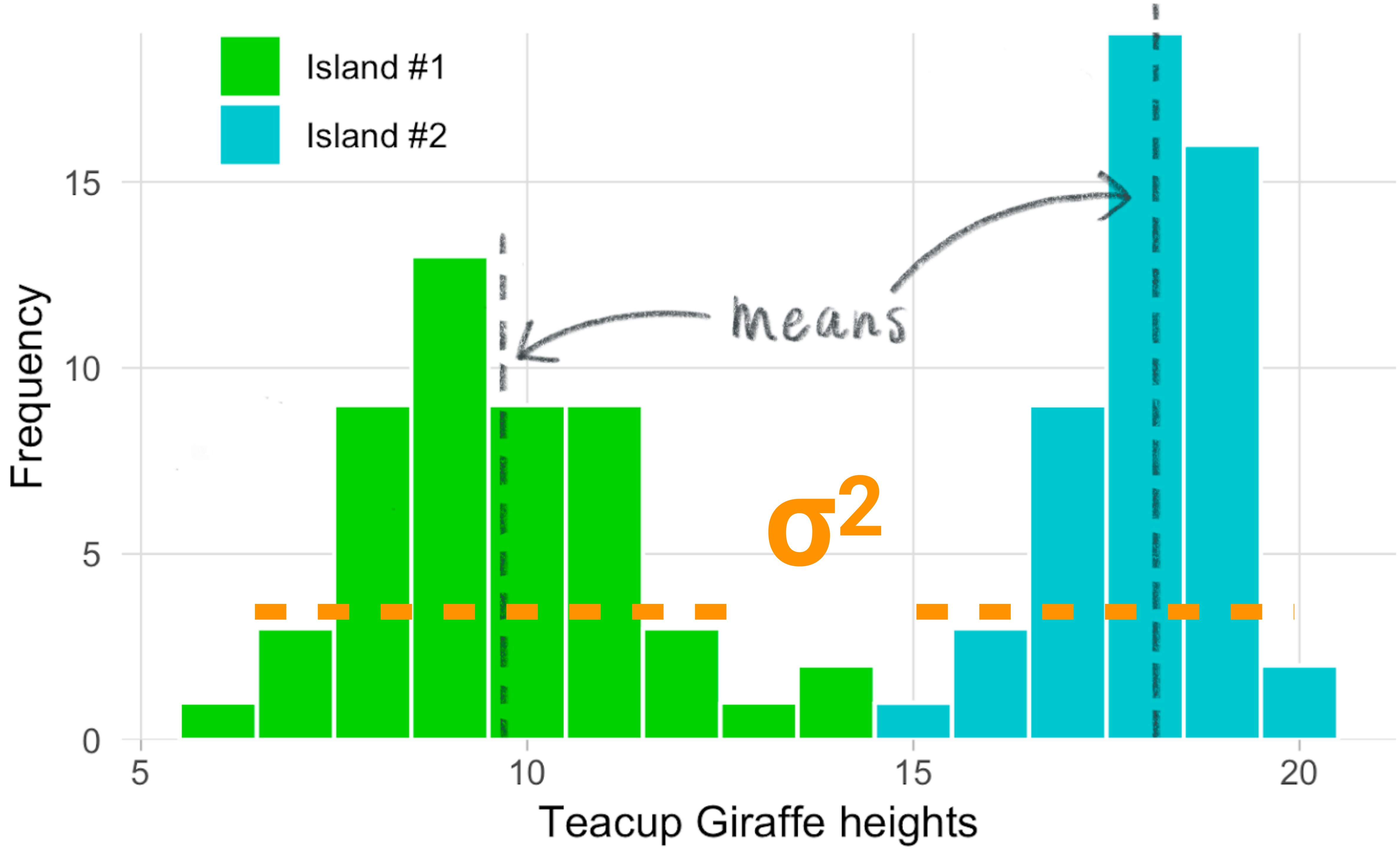
Island #1
Island #2
Frequency
Teacup Giraffe heights
Higher probability of measuring giraffes this tall
Encountering a giraffe this small would be rare
a giraffe greater than 15 cm would be almost unheard of on island 1, but not on island 2.

Parameters of the normal distribution

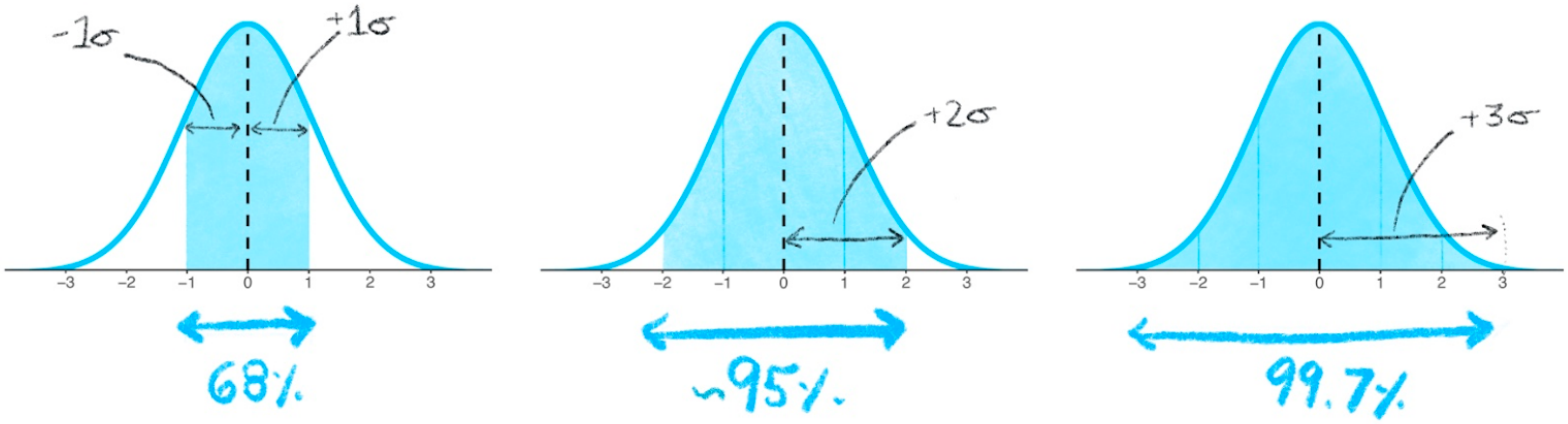
μ - the mean or expectation

σ - the standard deviation

or **σ²** - the variance



The standard deviation measures the spread of the data



Continuous distributions in R

Are associated with 4 standard functions, beginning d, p, q, and r:

`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

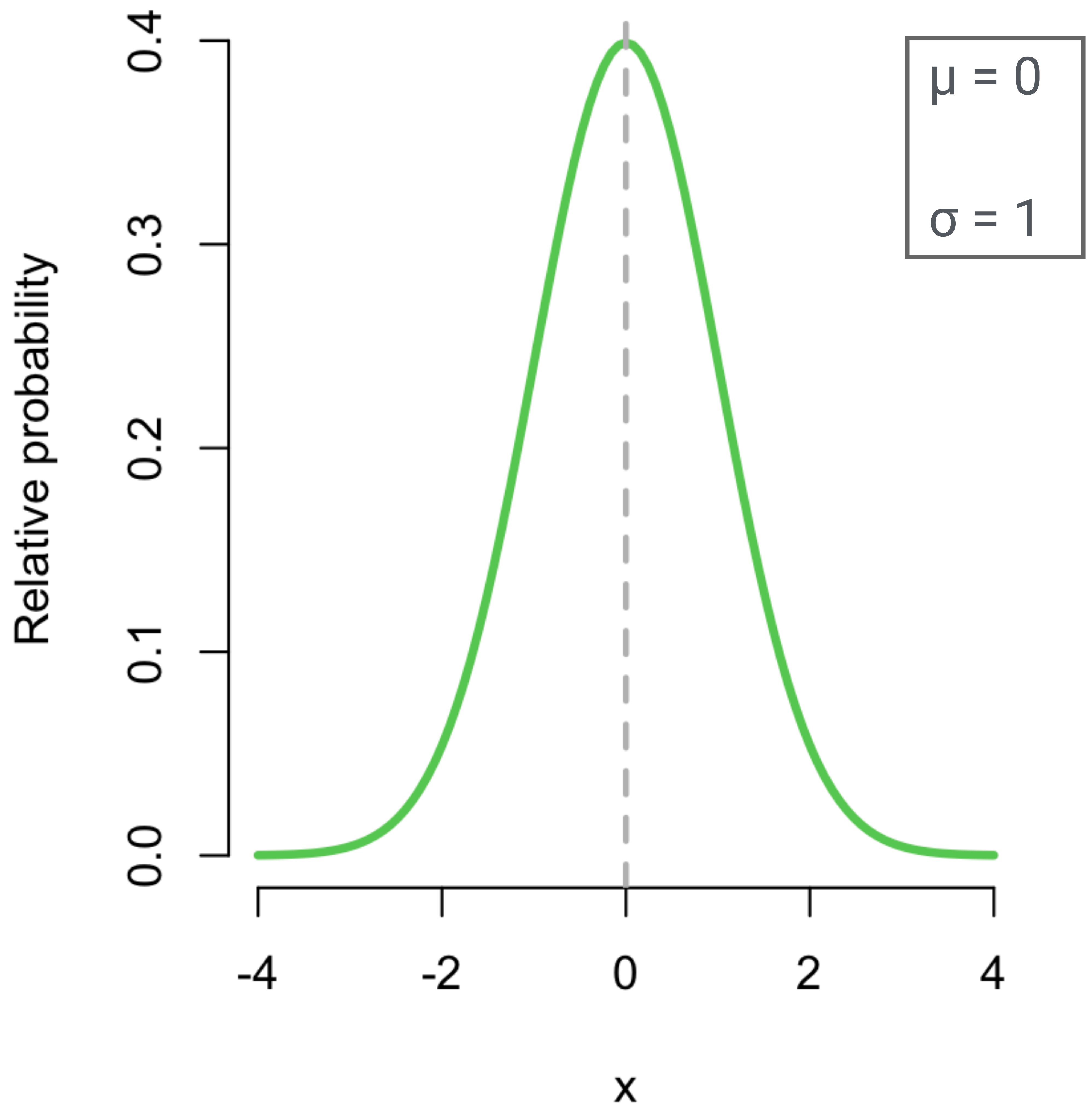
Let’s see these in action!

`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

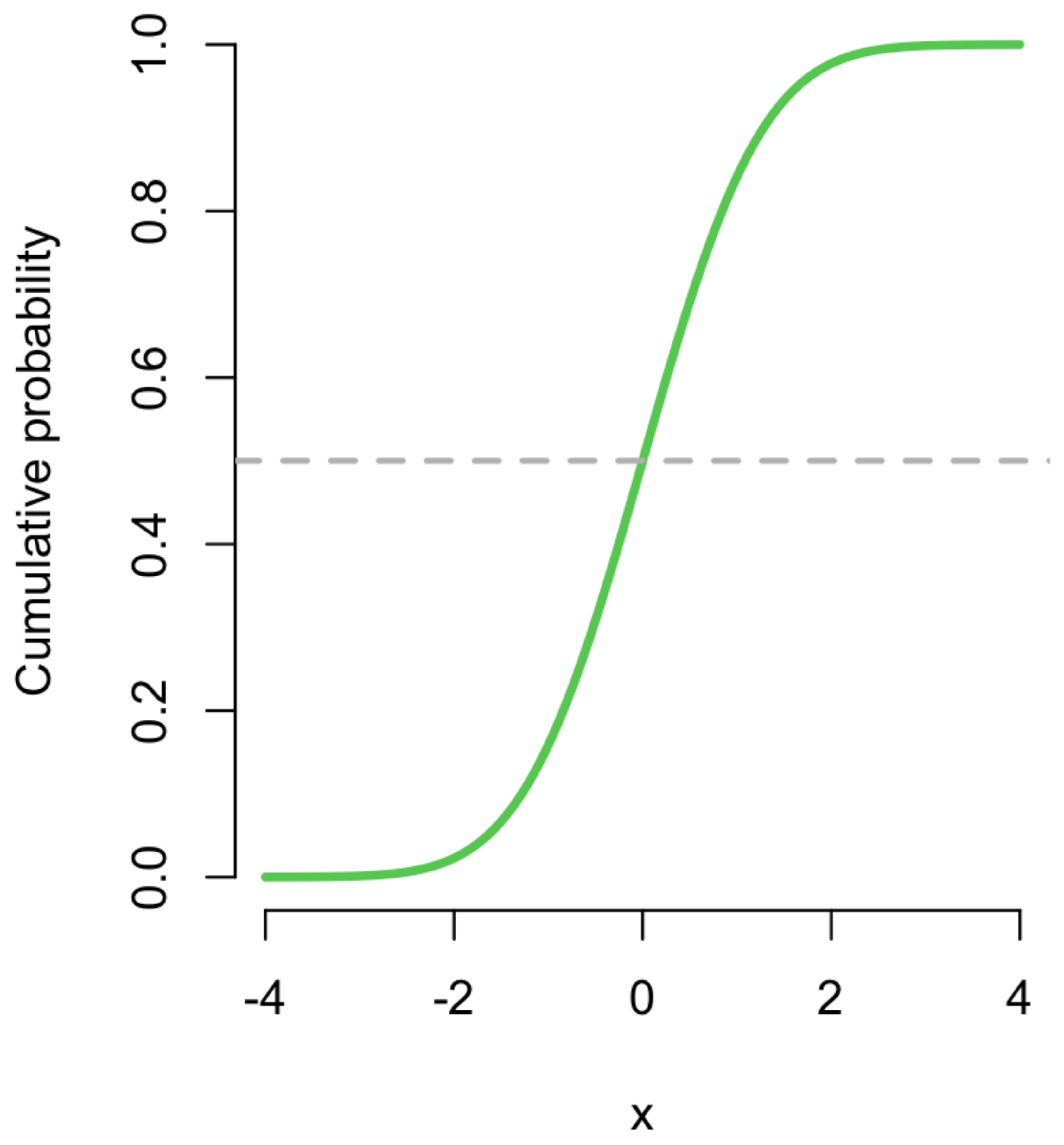


*Actually the probability of being a precise value can very small, so it's sometimes more useful to think of the probability of being within a range, e.g. the probability of being between 1 and 2

`pnorm()` - cumulative distribution function

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

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



$\mu = 0$

$\sigma = 1$

`qnorm()` - quantile function

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

Inverse of `pnorm()`

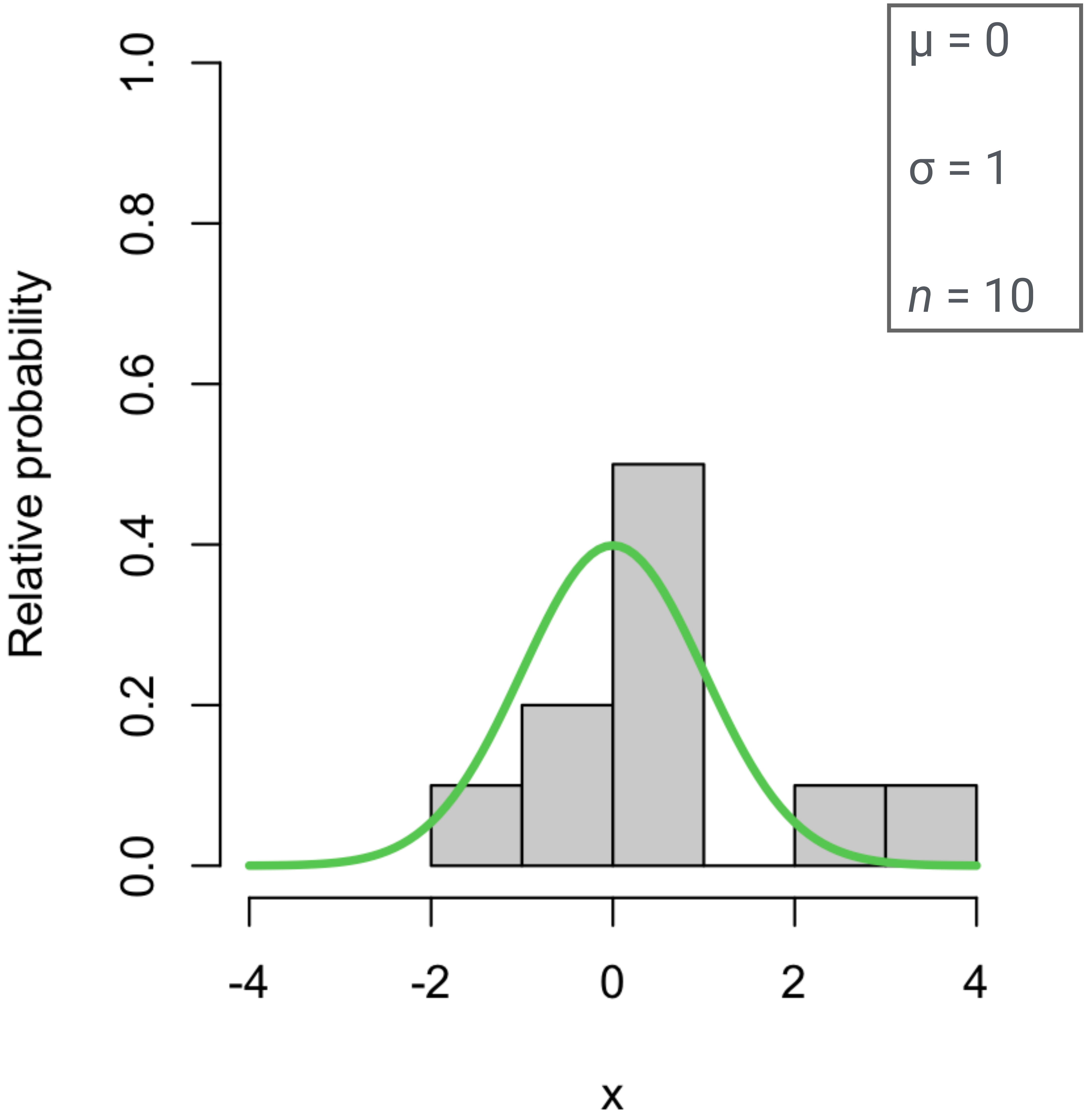
Shown right: the 10th, 25th, 50th, 75th, 90th percentiles

```
> qnorm(0.1, mean = 0, sd = 1)
[1] -1.281552
> qnorm(0.25, mean = 0, sd = 1)
[1] -0.6744898
> qnorm(0.5, mean = 0, sd = 1)
[1] 0
> qnorm(0.75, mean = 0, sd = 1)
[1] 0.6744898
> qnorm(0.90, mean = 0, sd = 1)
[1] 1.281552
```

`rnorm()` - generates pseudo random numbers

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

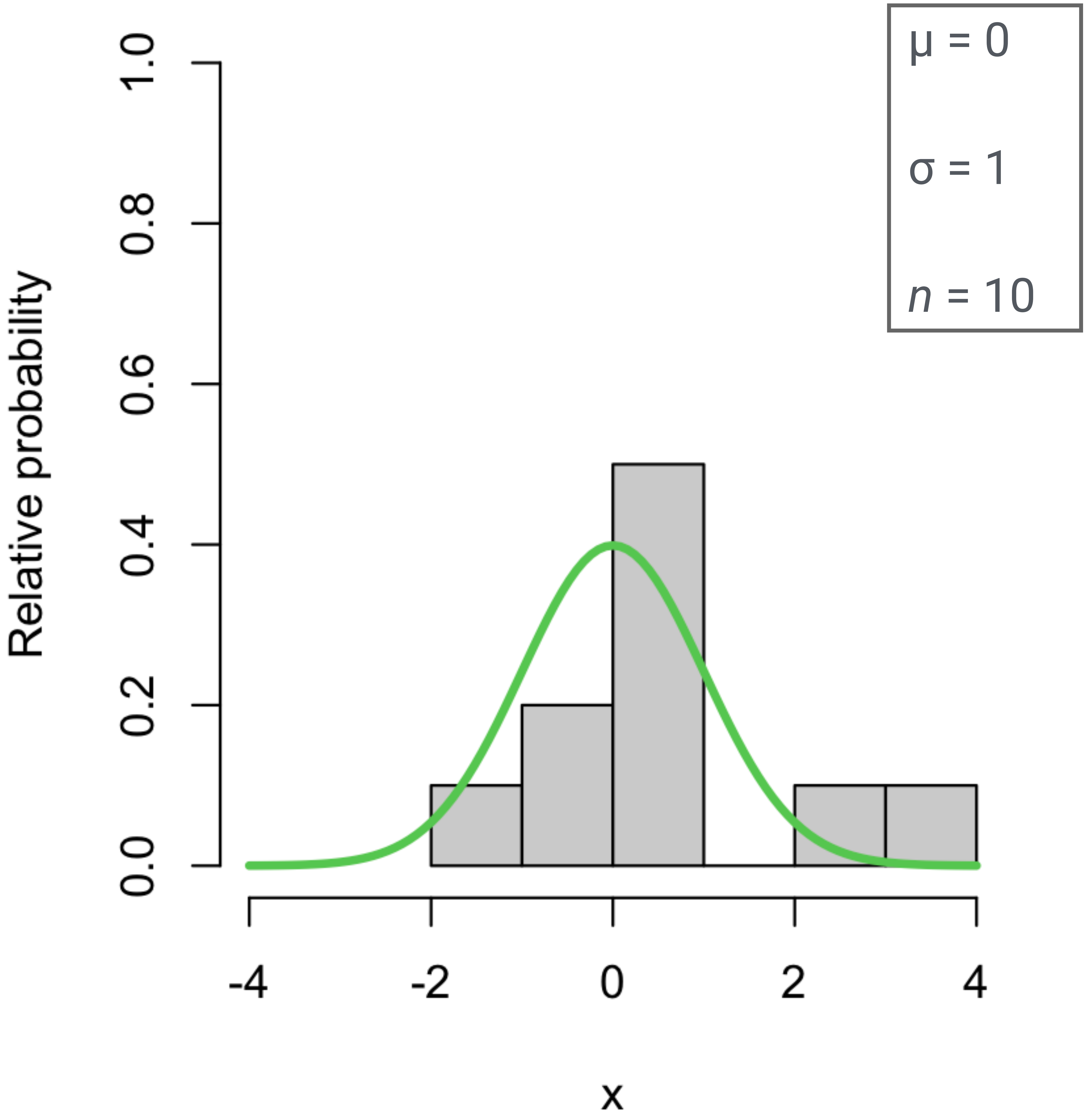
This plot shows random draws (*n*) from a normal distribution as a histogram

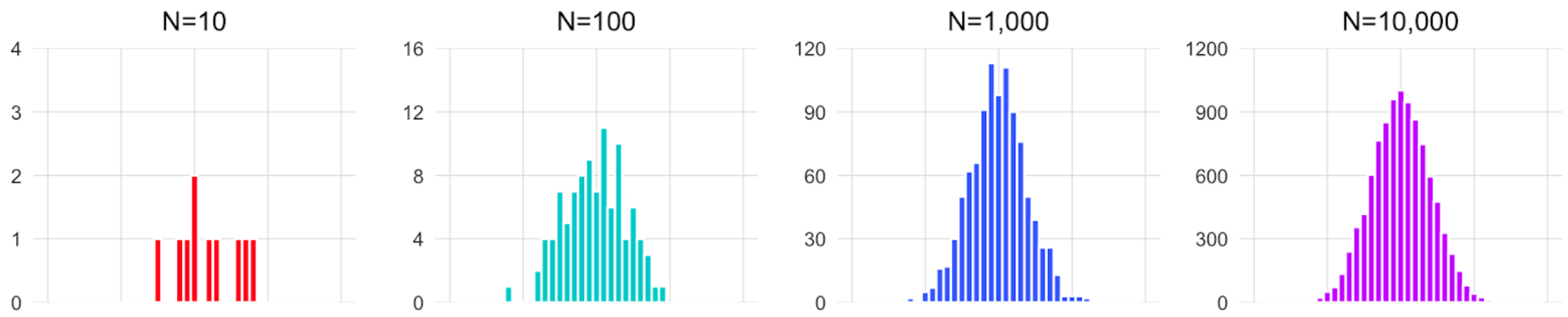


Why is it pseudo random?

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

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



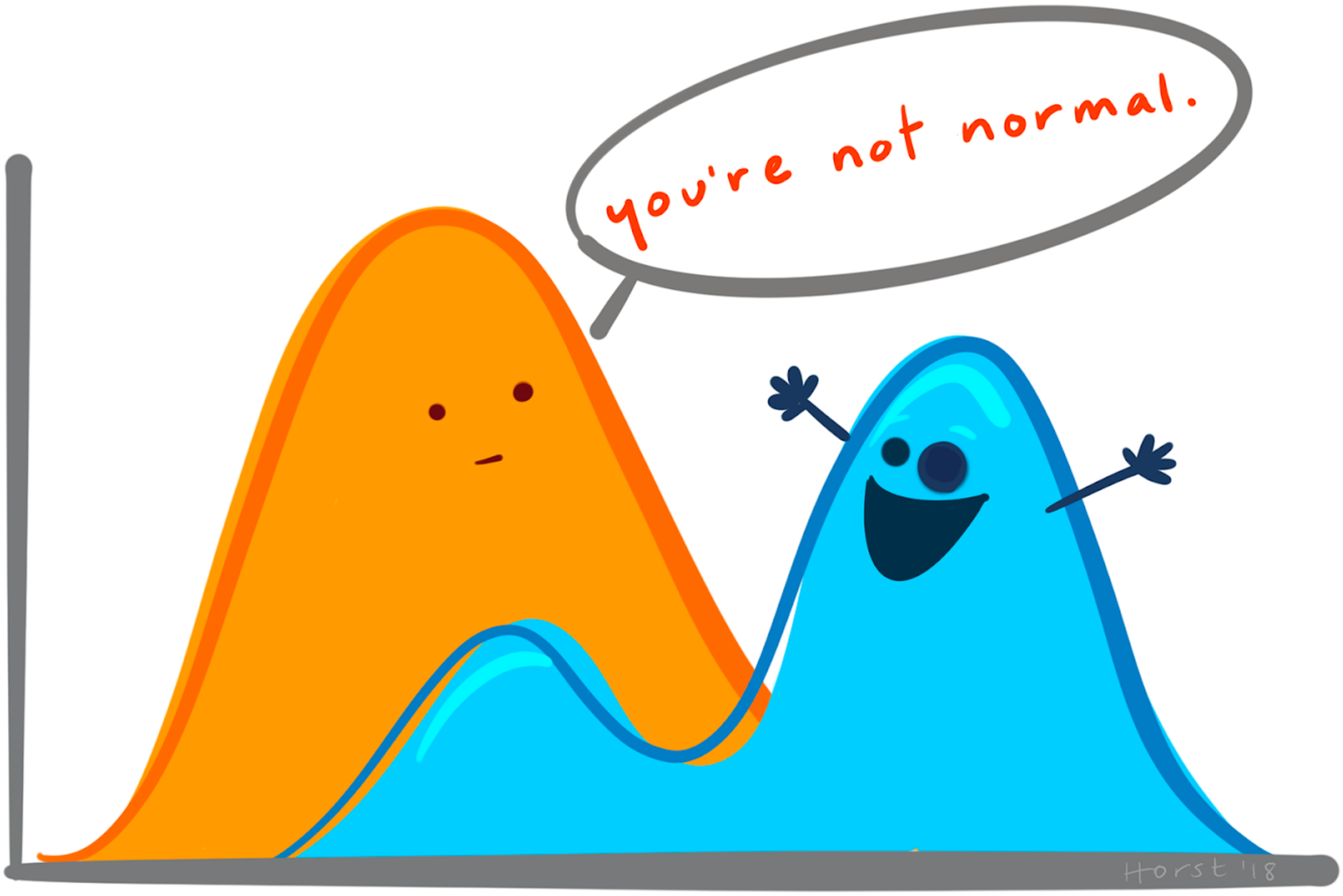

N=10
N=100
N=1,000
N=10,000
4
3
2
1
0
16
12
8
4
0
120
90
60
30
0
1200
900
600
300
0

Exercise

Explore the properties and plot an alternative distribution to the normal

Hint: Plot the normal distribution

Hint: see earlier examples! (link to slide)


you're not normal.
Horst '18

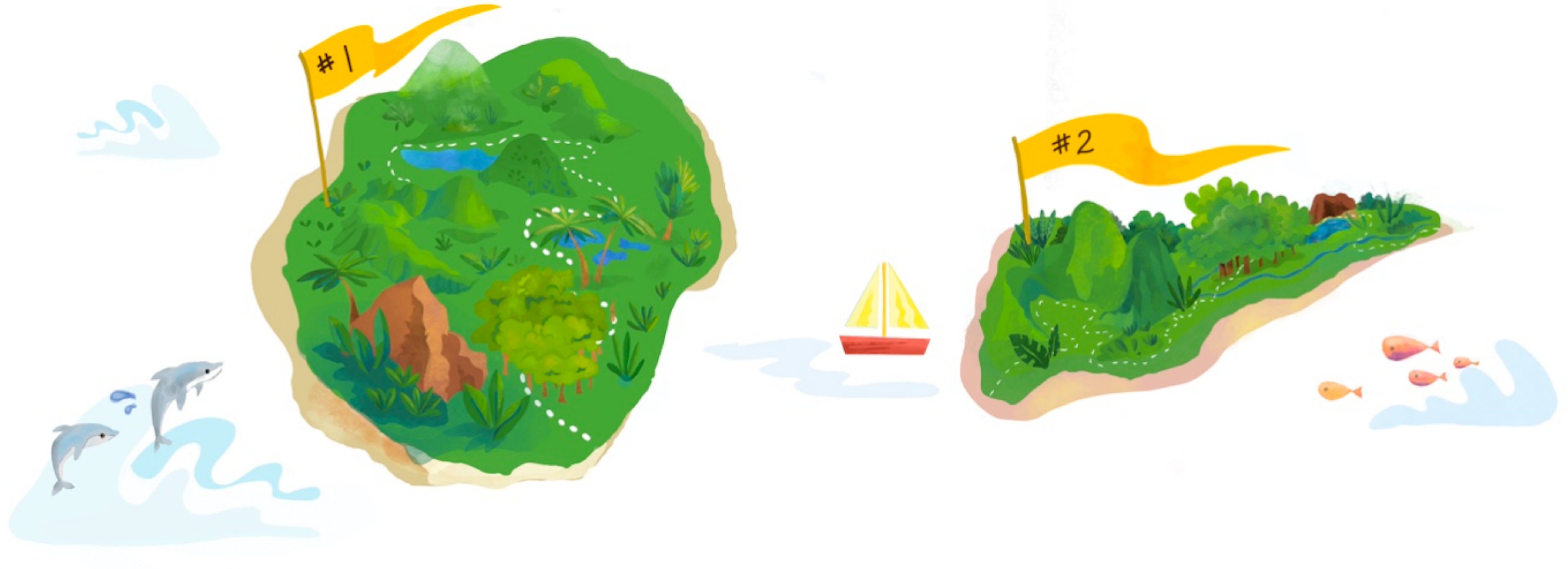
Part 2
Standard deviation and variance

Learn more about the tiny giraffes @ tinystats.github.io

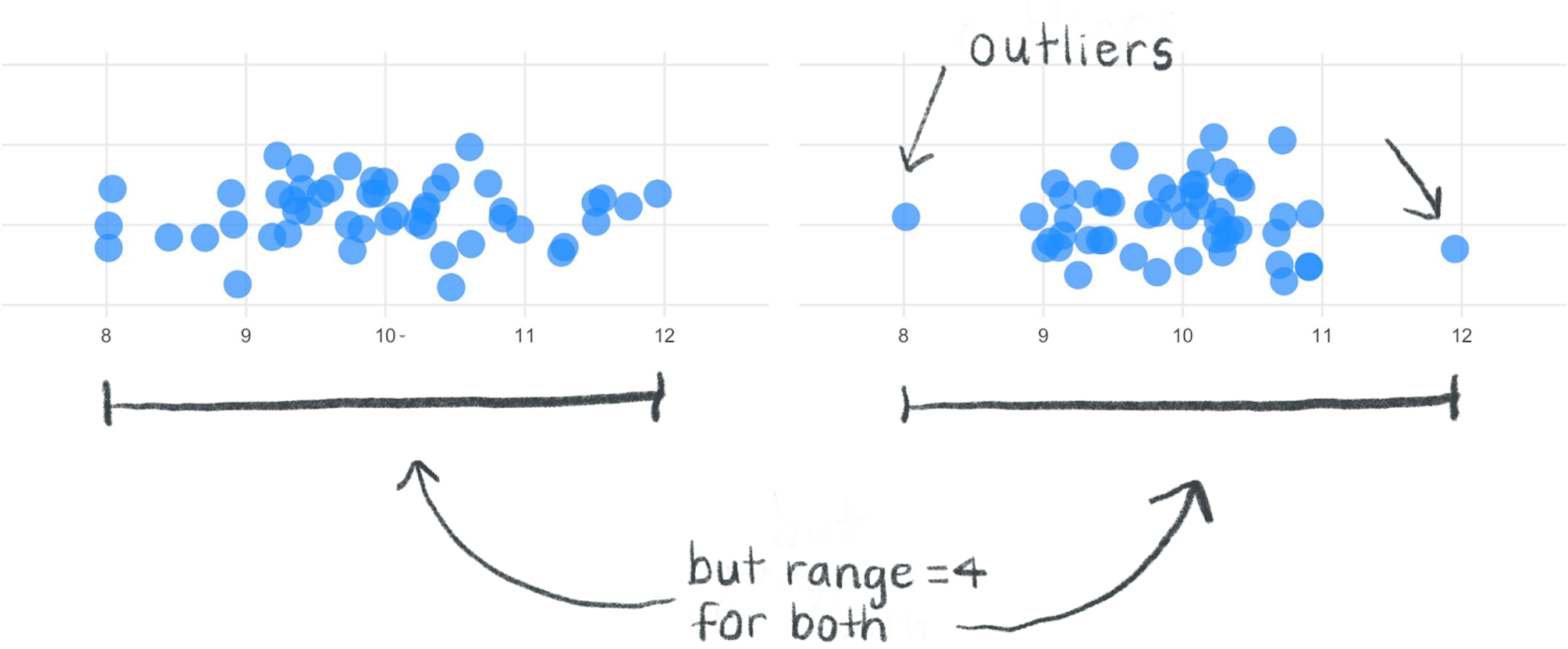
Teacup giraffes

Reminder

Imagine we've collected data for two populations that live on two different islands, like the tiny giraffes



Spread of the data: the range



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

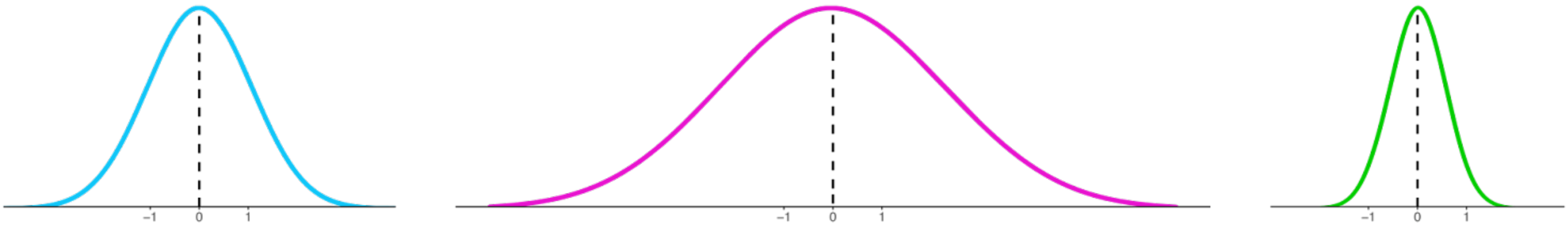
If we want to avoid undue influence of the outliers, the range is not good measure


are your
summary statistics
hiding something
interesting?
@allison_horst

Spread of the data: the standard deviation

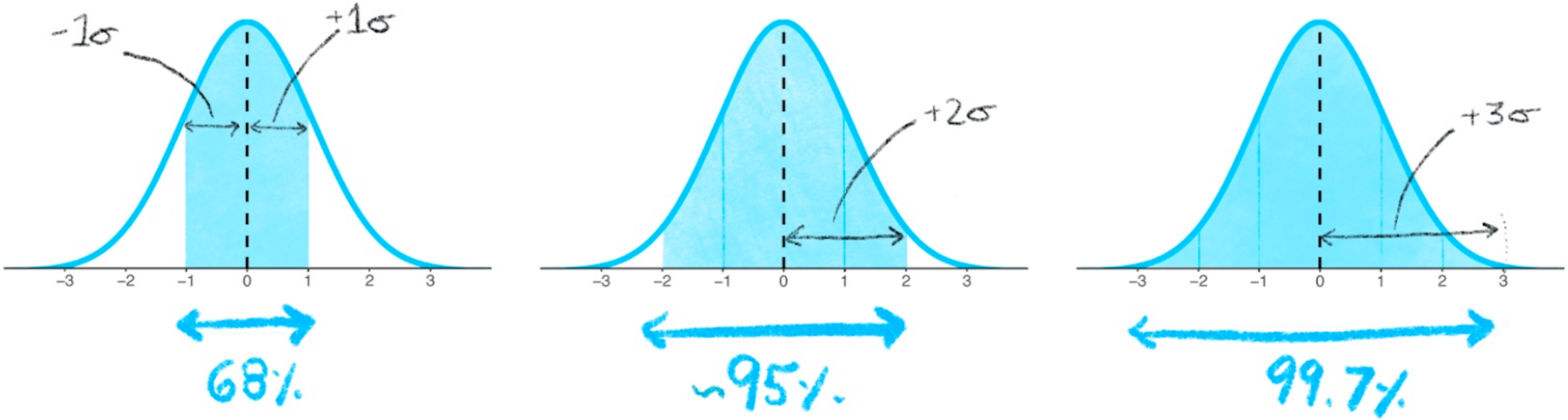
The **standard deviation** (**σ**) and **variance** (**σ**^{2}) account for outliers

It is a measure of how many your data **scatter around the mean**



To grasp the mechanics of common statistical tests, it is useful to have a good understanding of the s.d.

Standard deviation

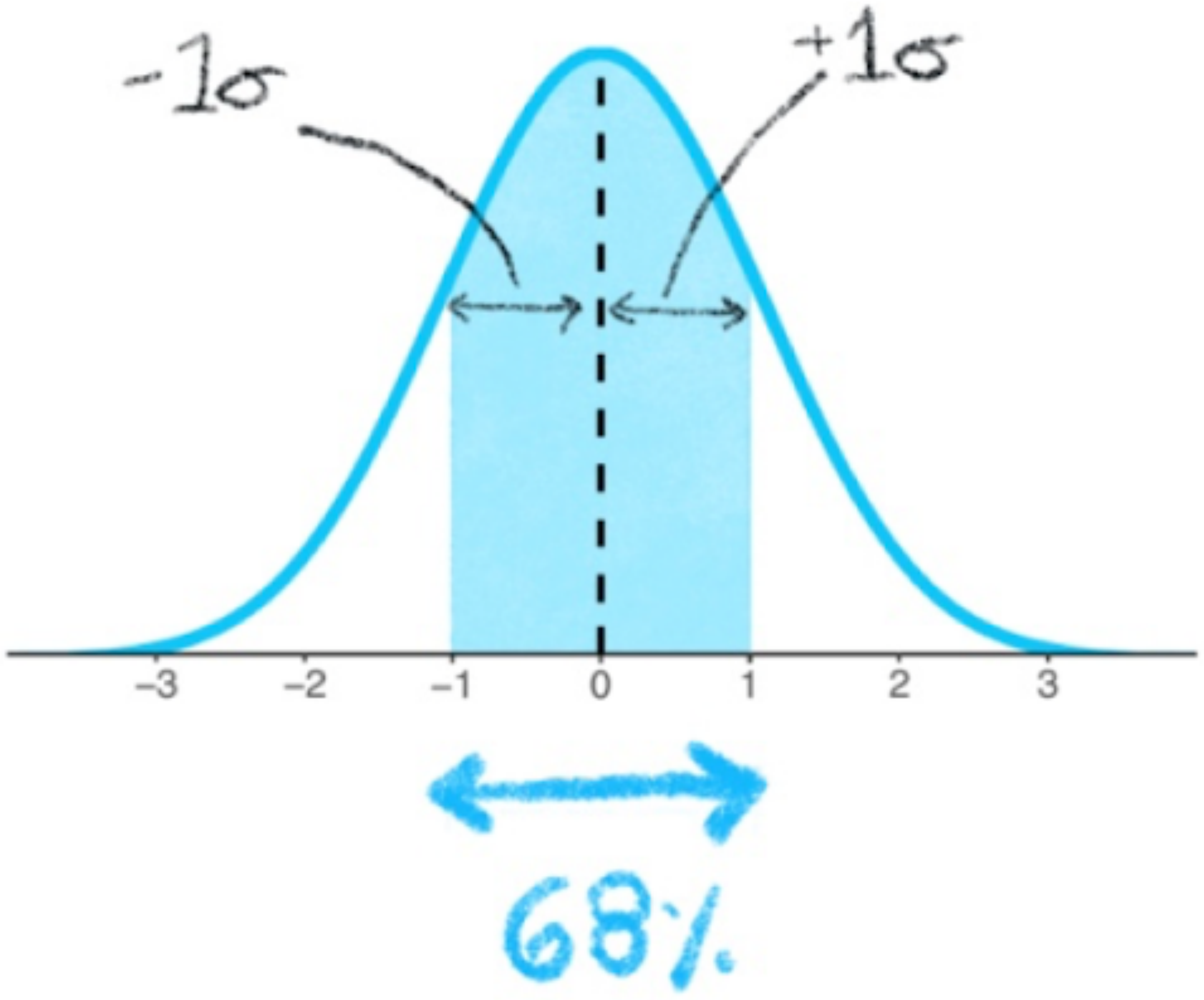


The 68–95–99.7 rule - a property of the normal distribution

Standard deviation

A measure of the amount of **variation** or **dispersion** in a set of normally distributed values

How do we calculate this?



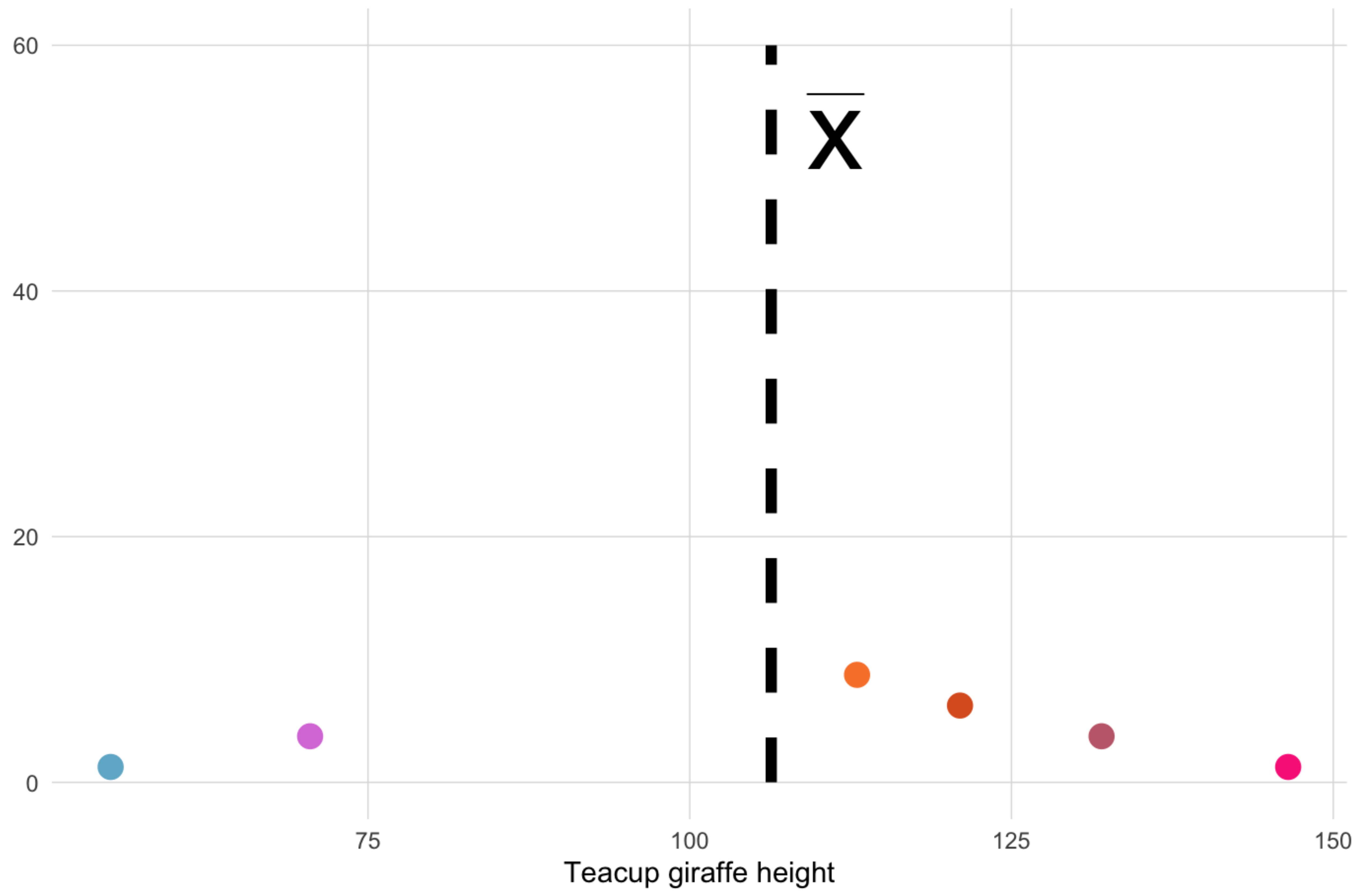
Calculating the variance and standard deviation

1. Calculate the **sample mean** ($\bar{x}$)

2. Square the deviations from the mean (this ensures the values are all positive)

→ How much do our data points deviate from the mean on average?

See [Variance and Standard Deviation]



Calculating the variance

$$\sum_{i=1}^{N}(x_i - \mu)^2$$

3. Calculate **the sum of squared deviations**

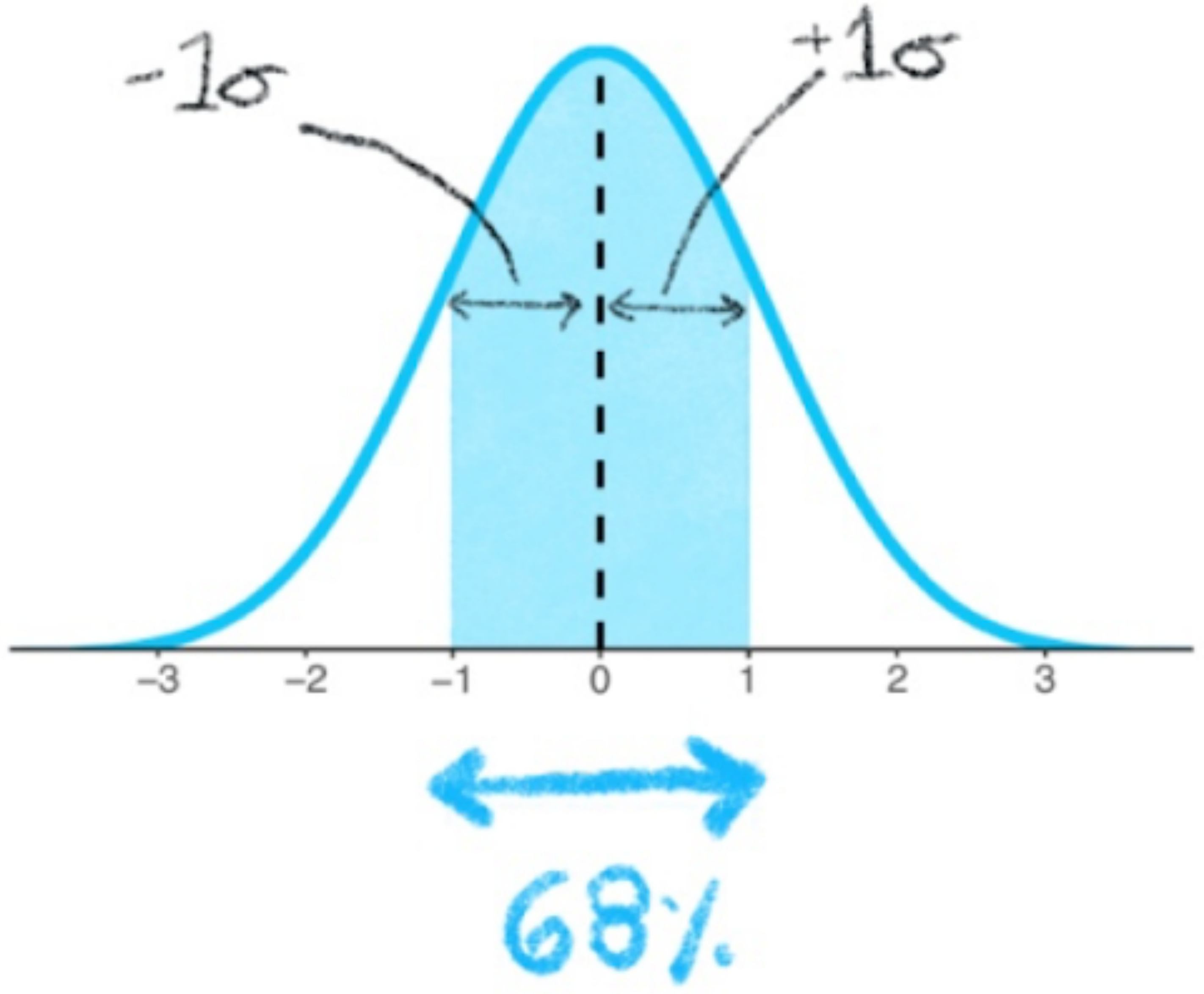
4. Calculate the **average squared differences from the mean** (i.e., the average of step 3)

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Calculating the standard deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

5. Variance is not easily interpretable → so we "unsquare" the variance to return to the data's original units (e.g., cm, ml, etc.)



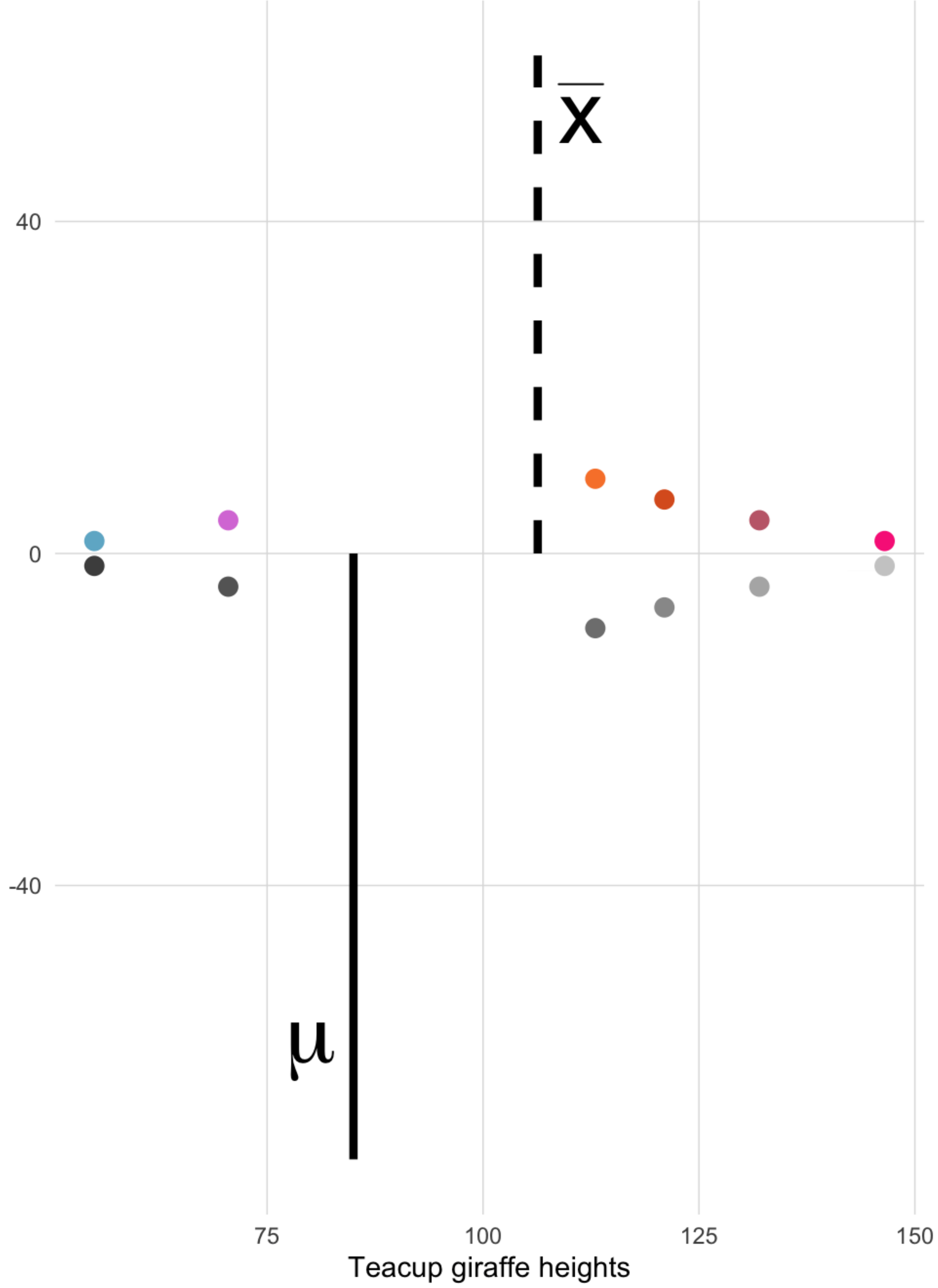
Population vs. sample

however

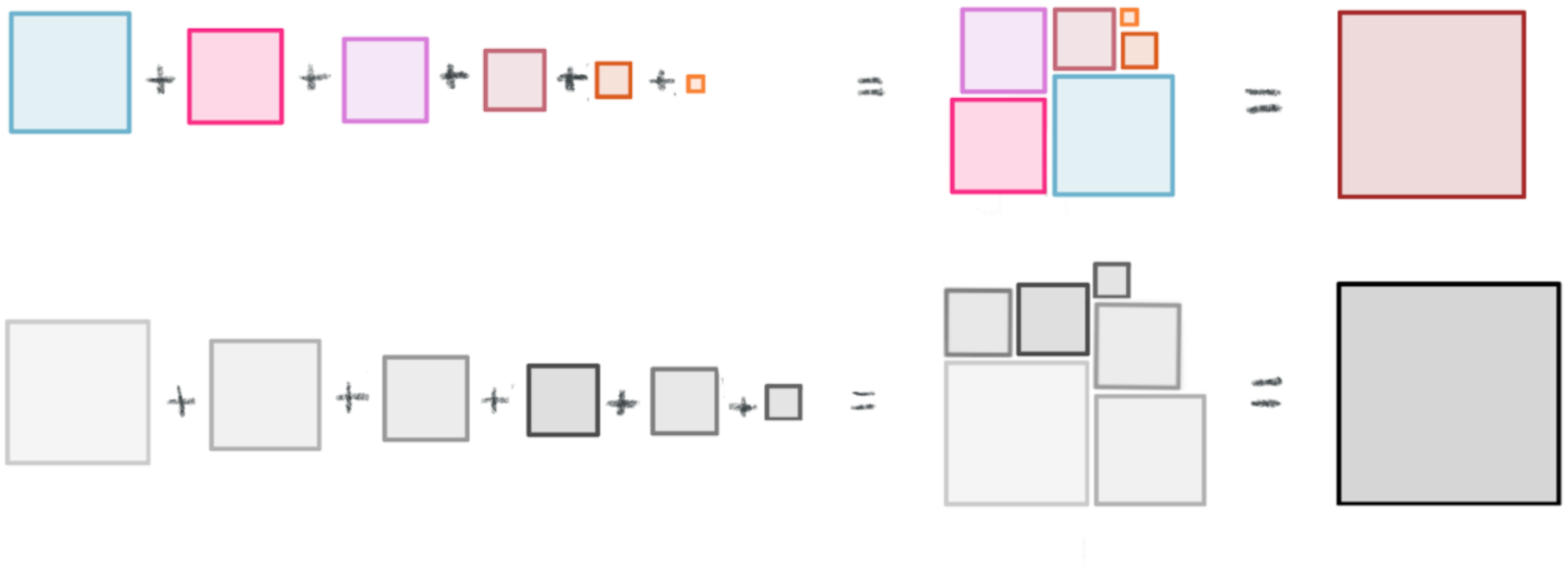
We only have the sample mean as our centre point

The true population mean μ is unknowable

The smaller the sample, the less likely the sample mean *x* will be close to the true mean μ



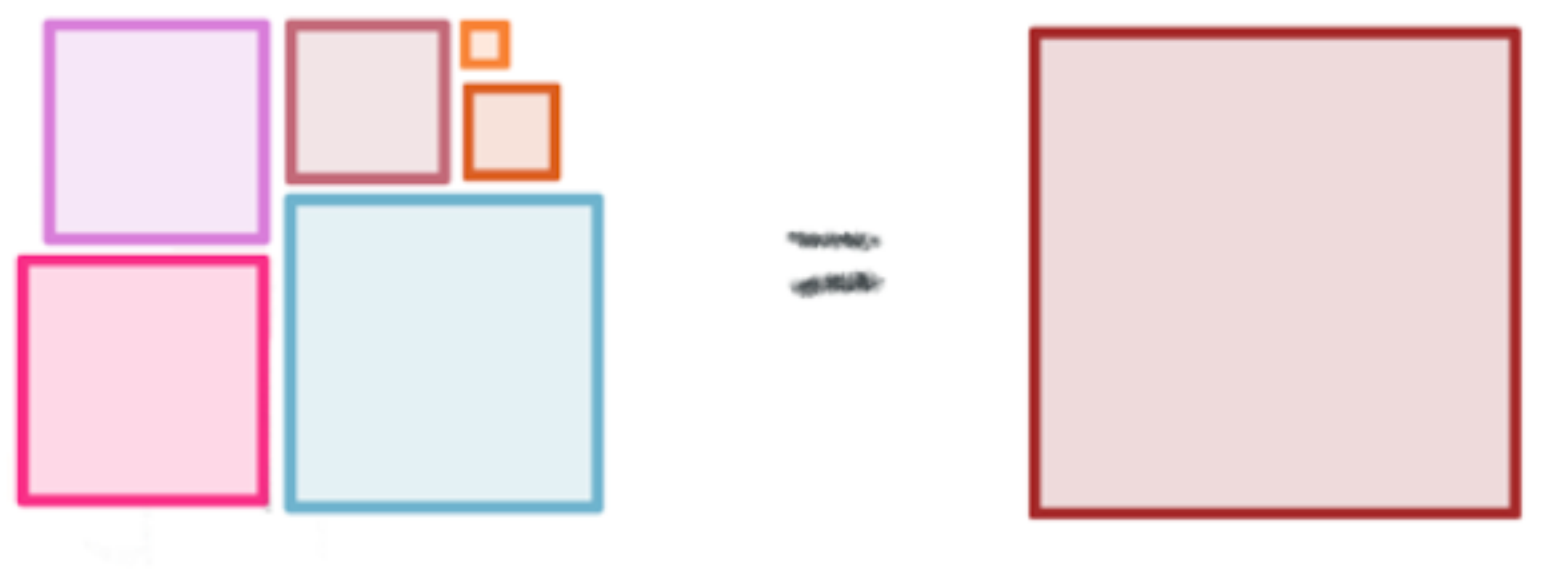
Population vs. sample

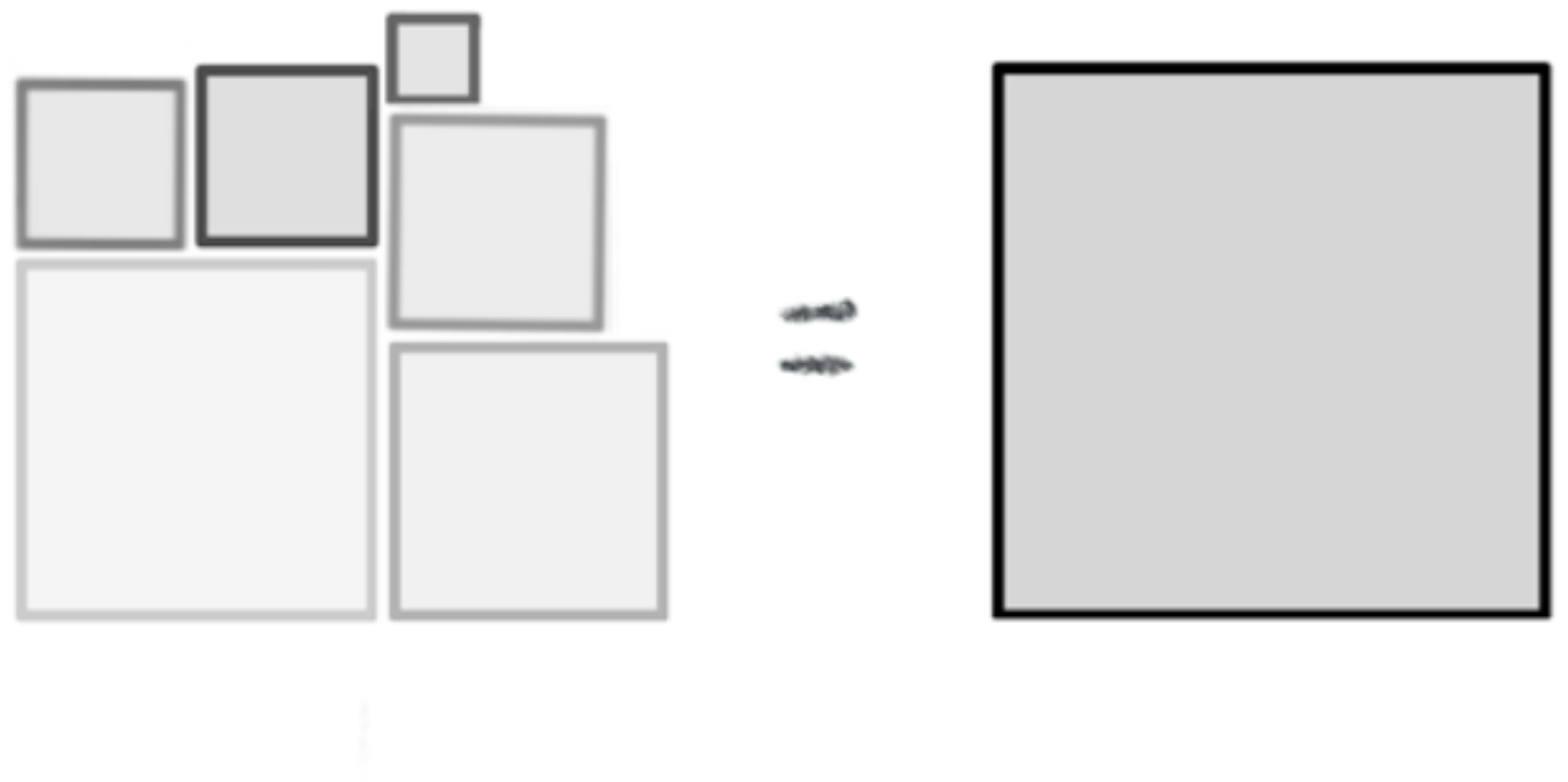


Population vs. sample

The sum of squares from μ will *always be greater* than the sum of squares from *x*

By definition, the location of *x* minimises the total distance of all the observations to the centre





Solution: *n* - 1

→ If we divide by *n* - 1, we ensure the overall variance and standard deviation is a little larger, correcting for this bias (**4.**)

This means if we calculate the sum of squares (and thus, the variance and standard deviation) using the sample mean, our estimate will (most likely) be biased downwards

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

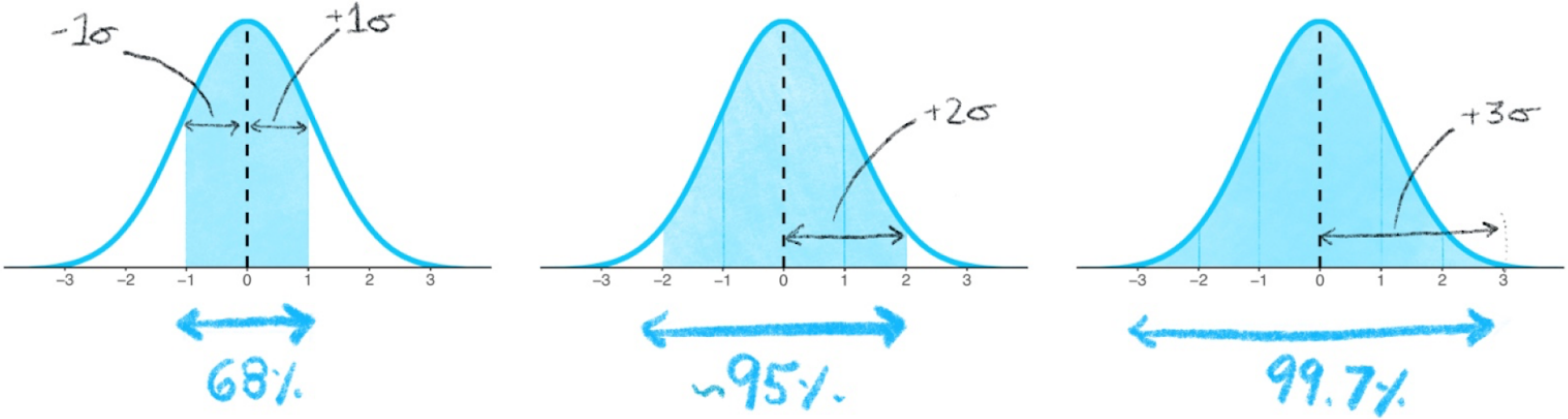
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}$$

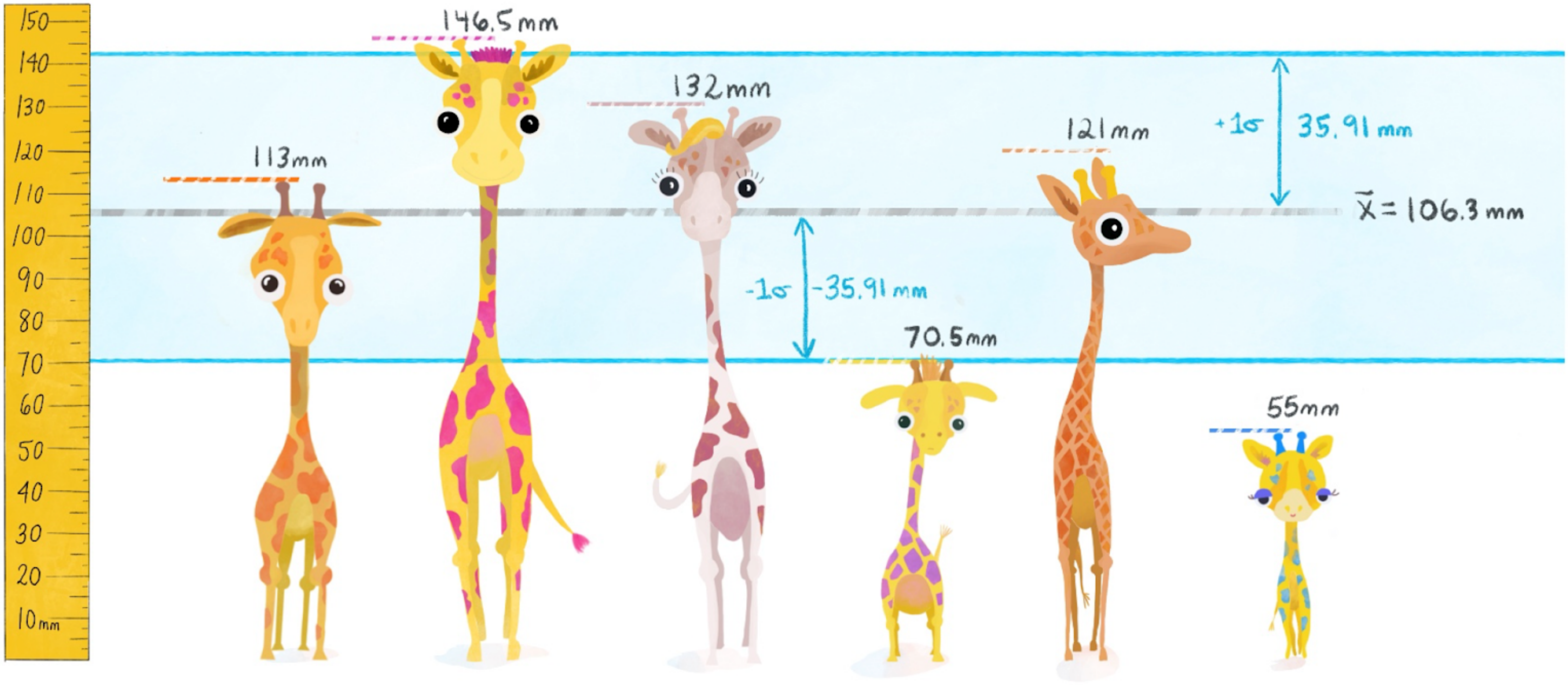
$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}}$$

Summary of steps used to calculate the standard deviation

1. Calculate the **sample mean**
2. Square the **deviations from the mean**
3. Calculate the **sum of squares**
4. Calculate **average squared differences** (apply the *n*-1 correction)
5. Unsquare to get the standard deviation

Interpreting the standard deviation




150
140
130
120
110
100
90
80
70
60
50
40
30
20
10mm
146.5mm
132mm
113mm
121mm
+1σ
35.91 mm
X̄ = 106.3 mm
-1σ
-35.91 mm
70.5mm
55mm

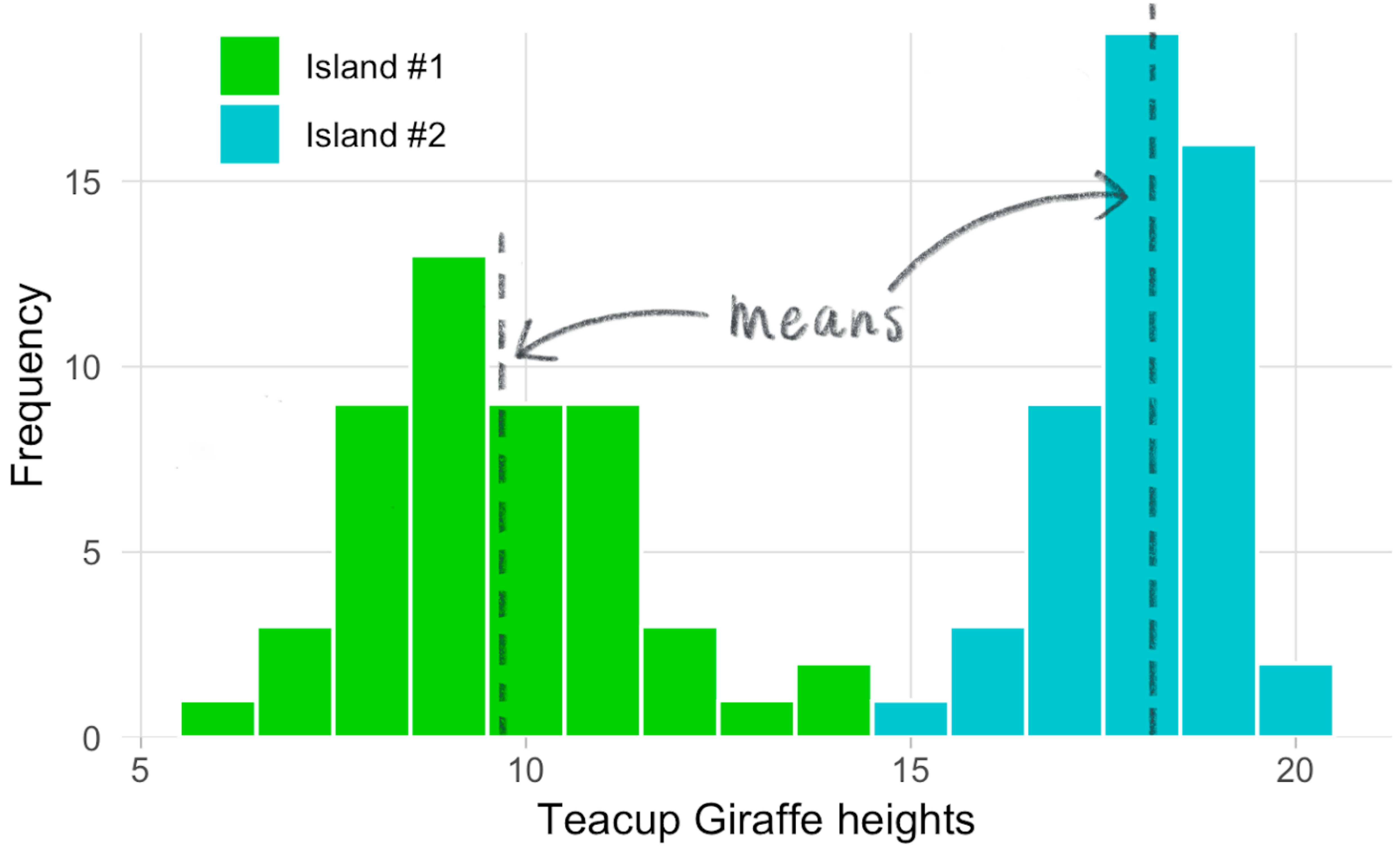
Functions in R

demo

```
my_function <- function(){
   print("hello world")
}
```

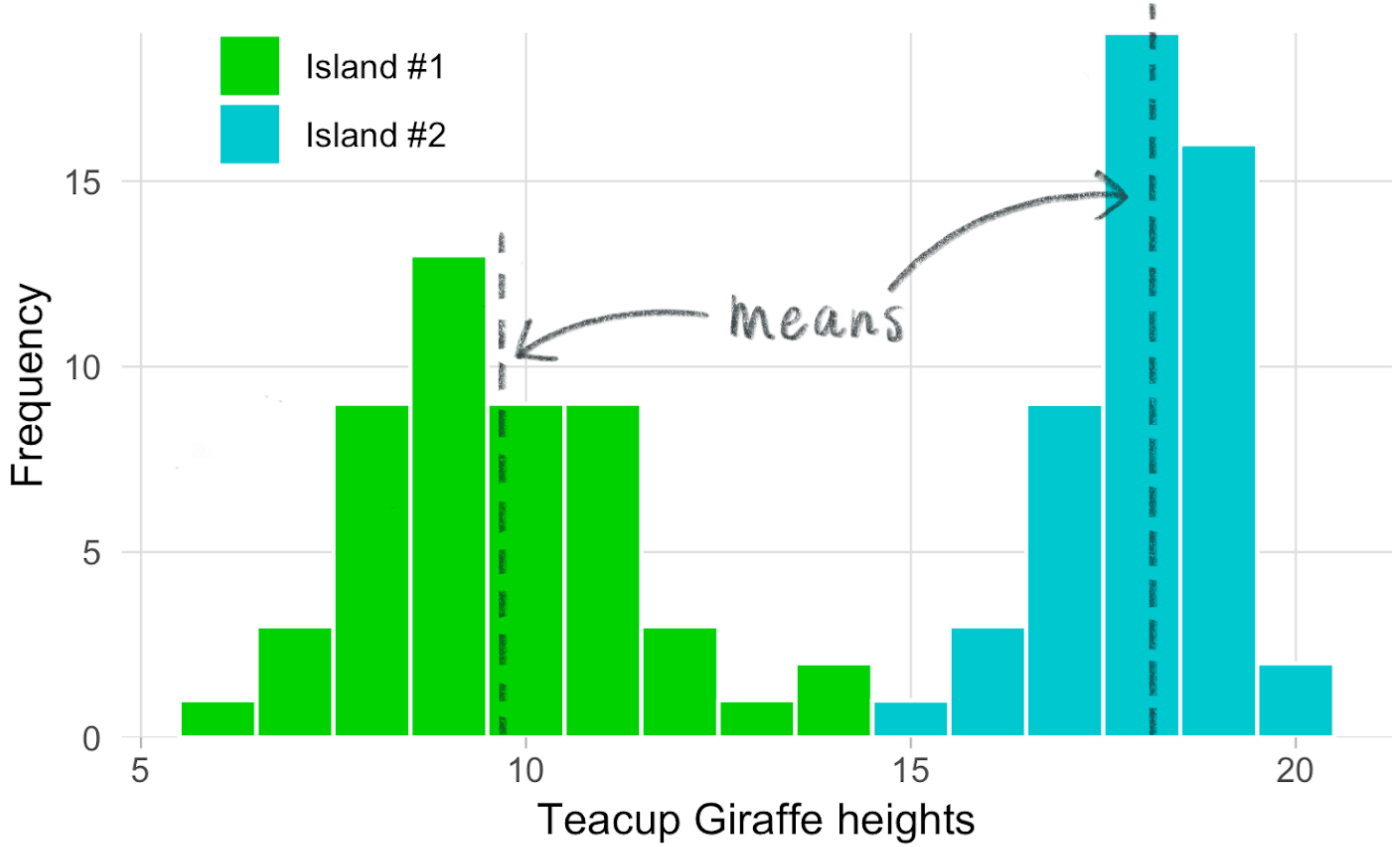
Exercise

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



Exercise

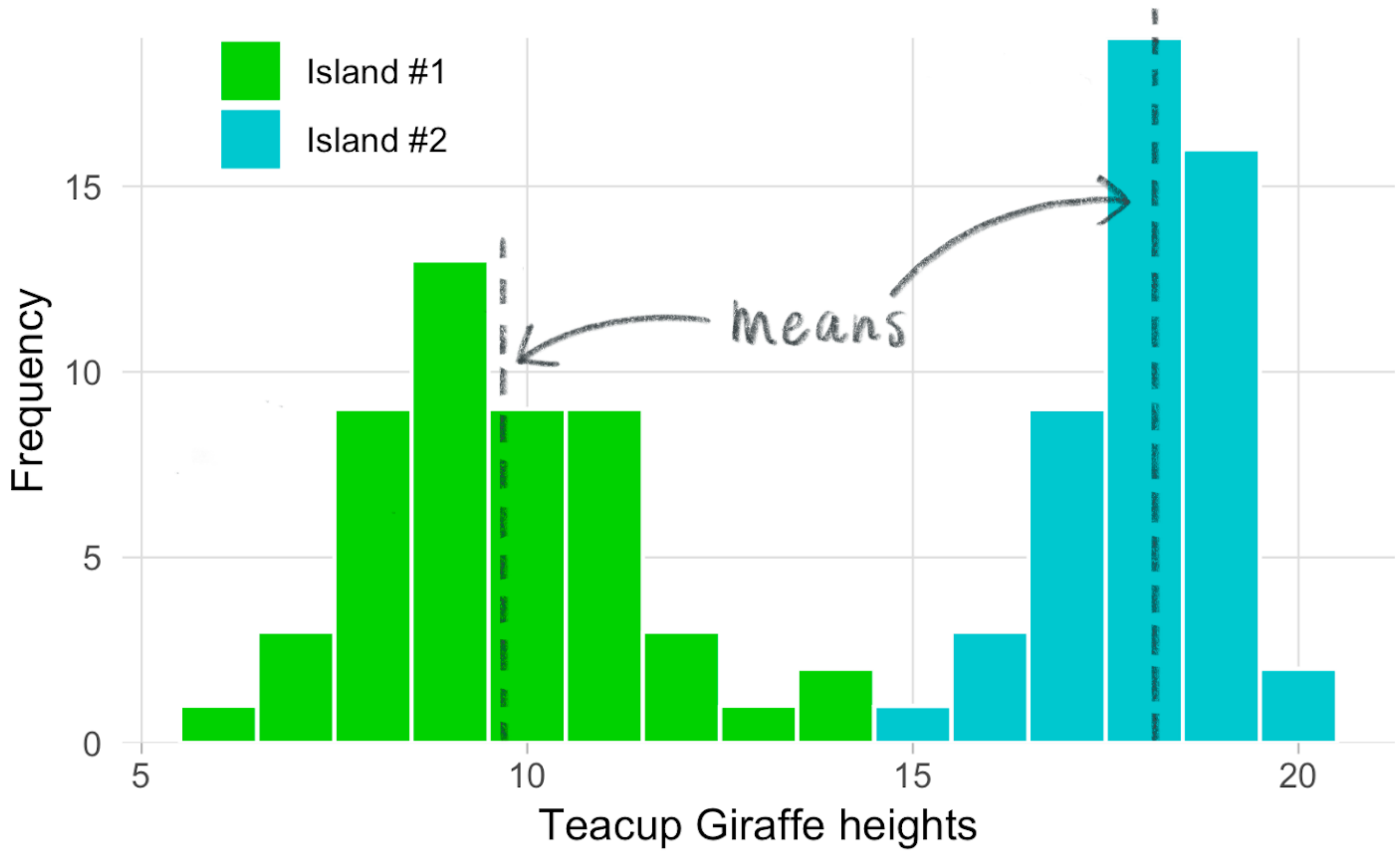
Write a **function** that calculates the **standard deviation** for a **vector**



Homework

Can you write a function that calculates **mode** or **median**?

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



Resources

Tools for exploring the normal distribution

Compare two normal distributions

Plot the normal distribution

Learn more about s.d. and variance

Variance and Standard Deviation: Why divide by n-1? Zed Statistics

Standard deviation (simply explained) DATAtab