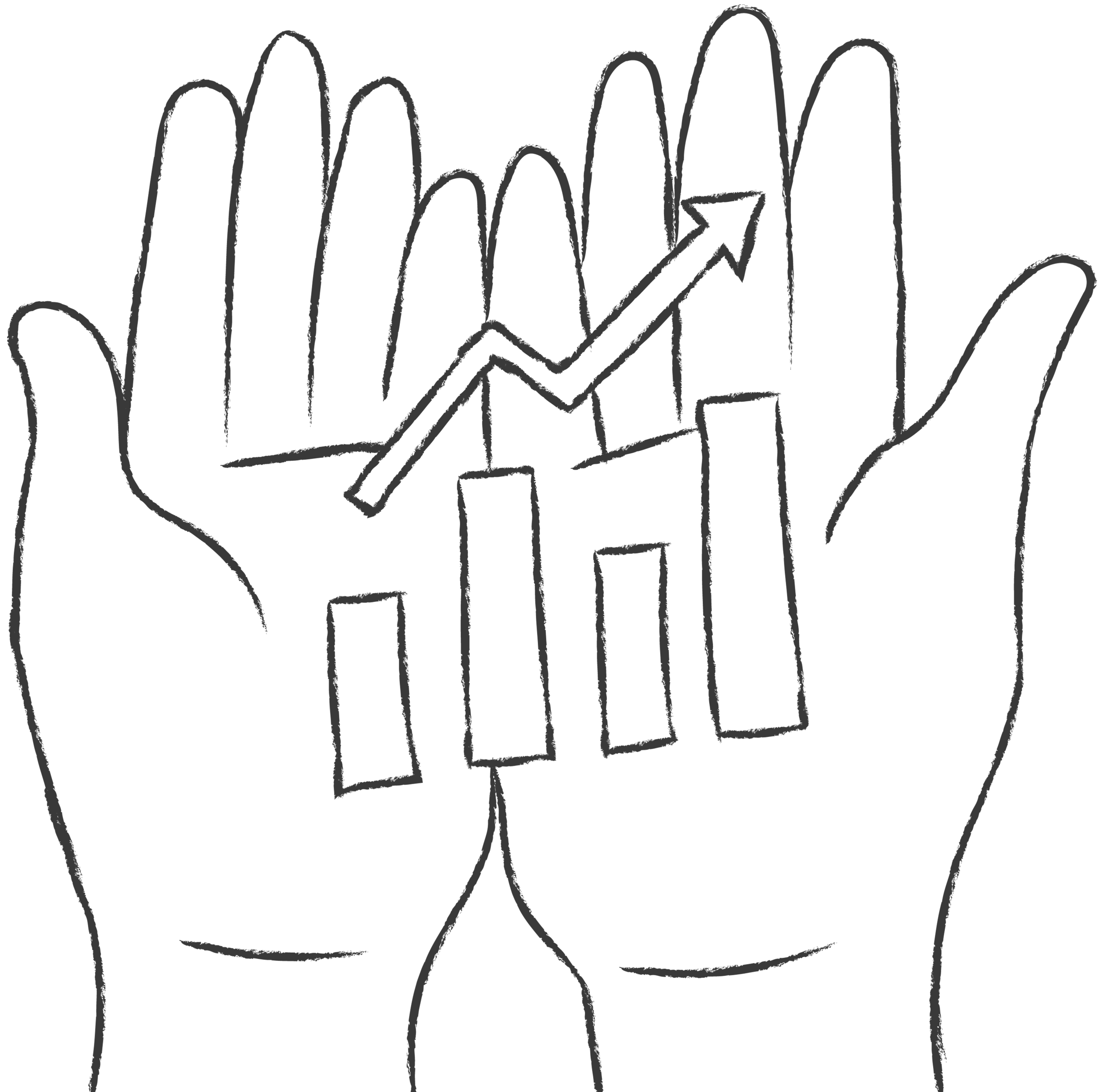
Hypothesis testing

Introduction

Rachel Warnock

13.02.2026

Course introduction



Today

Introduction to hypothesis testing

Introduction to reproducibility

Introduction to basic tests

We can take a break whenever you like!

About this course

The course focuses on **statistical hypothesis testing** and **reproducibility** in science

You will learn:

- how to develop and test hypotheses
- perform basic statistical tests
- apply this knowledge to reproduce (and potentially improve) published results in your assigned groups

Course evaluation

The goal is to reproduce the statistical results of a published scientific article

On **Fri 08 May 26 (14:00-17:00)** each group will present their findings

Use this **Google Slides Template** and **guidelines** to prepare your presentation

Within your group, make as much progress as you can today and on **Mon 13 April 26**

Will I have to use R?

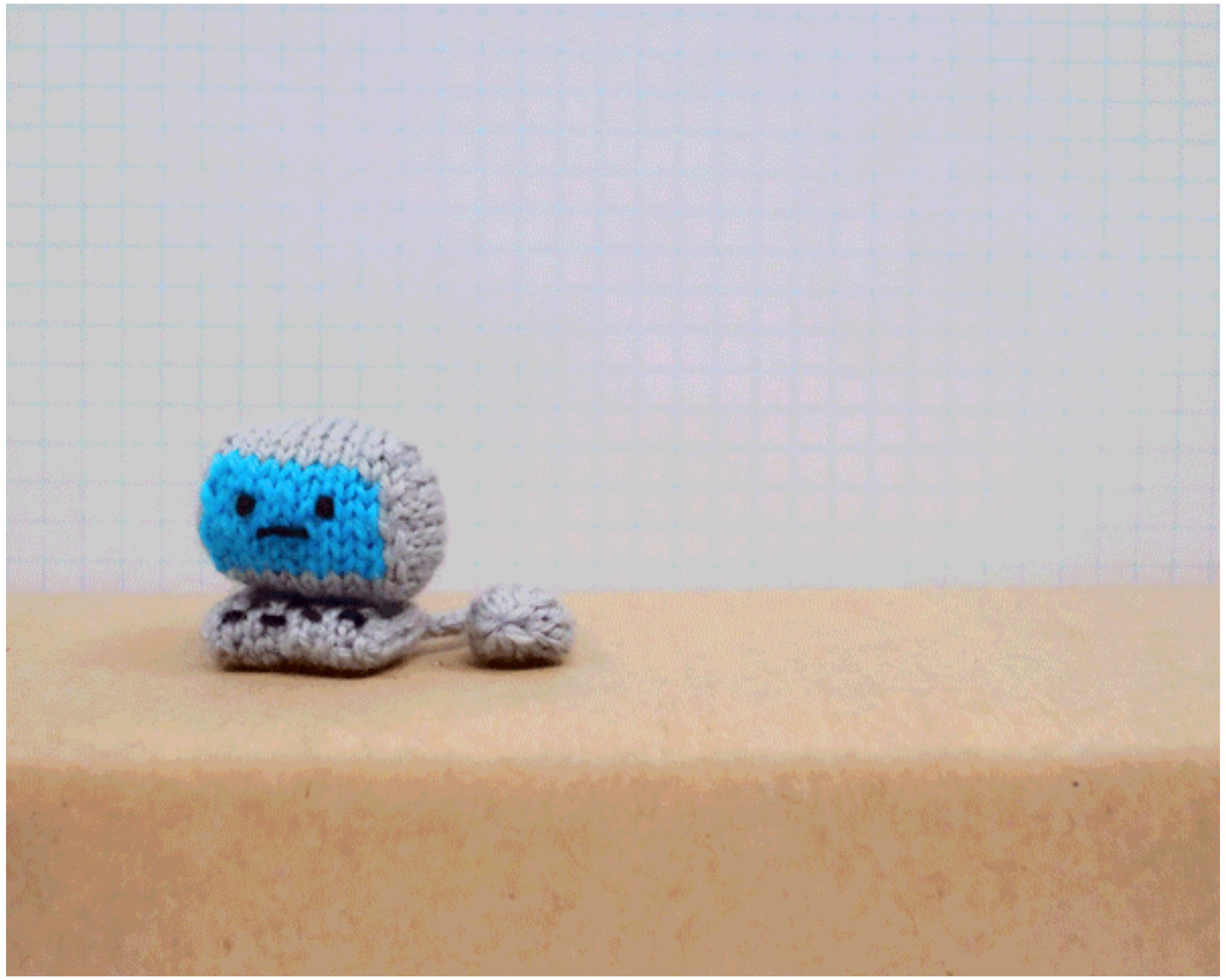
Not necessarily

The focus is on the concepts behind hypothesis testing and reproducibility, not programming, but both things are (probably) easier if you use R

It's up to each group how you divide the tasks


FULLY EXPECTING to HATE THIS CLASS!
STATS
CODE
@allison_horst

Here to help!



Discussion

Why does paleobiology involve so much statistics?

- Why do we need statistics?
- What is your experience using statistics so far?
- When might statistics be useful or important?

Statistics is the science of learning from data - *"the discipline that concerns the collection, organisation, analysis, interpretation, and presentation of data"*

Wikipedia
The Free Encyclopedia

Why study statistics?

Humans are biased

Living systems are complicated

It is deeply intertwined with **research design**

Where there's data there's statistics! → It makes the literature more accessible

Having some knowledge of statistics gives you a superpower 🦸

Reading group

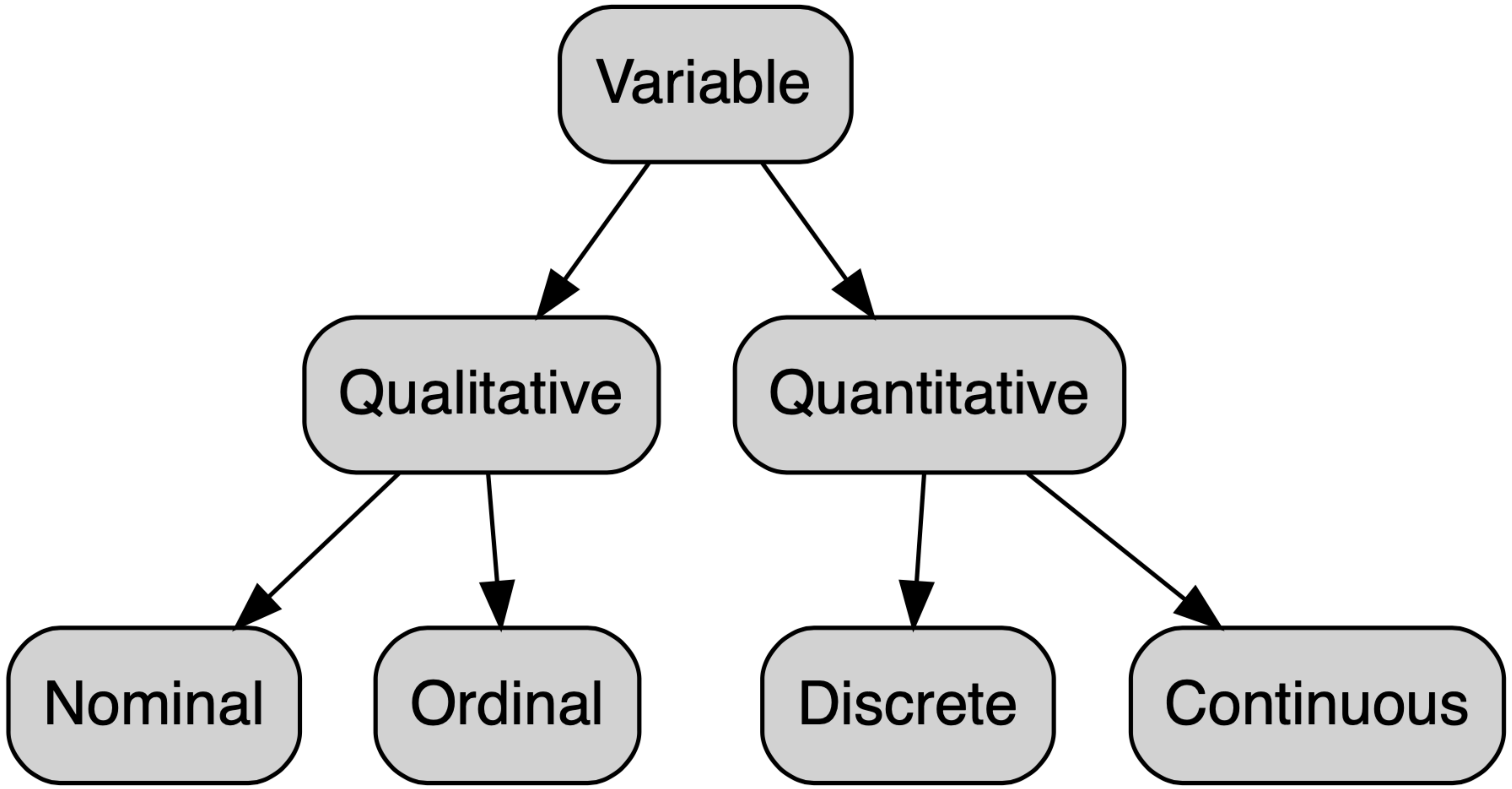
Group exercise

In your groups, discuss the following:

- What is your paper about?
- What was the general aim?
- What were the study's findings?

Prepare a brief summary 📝

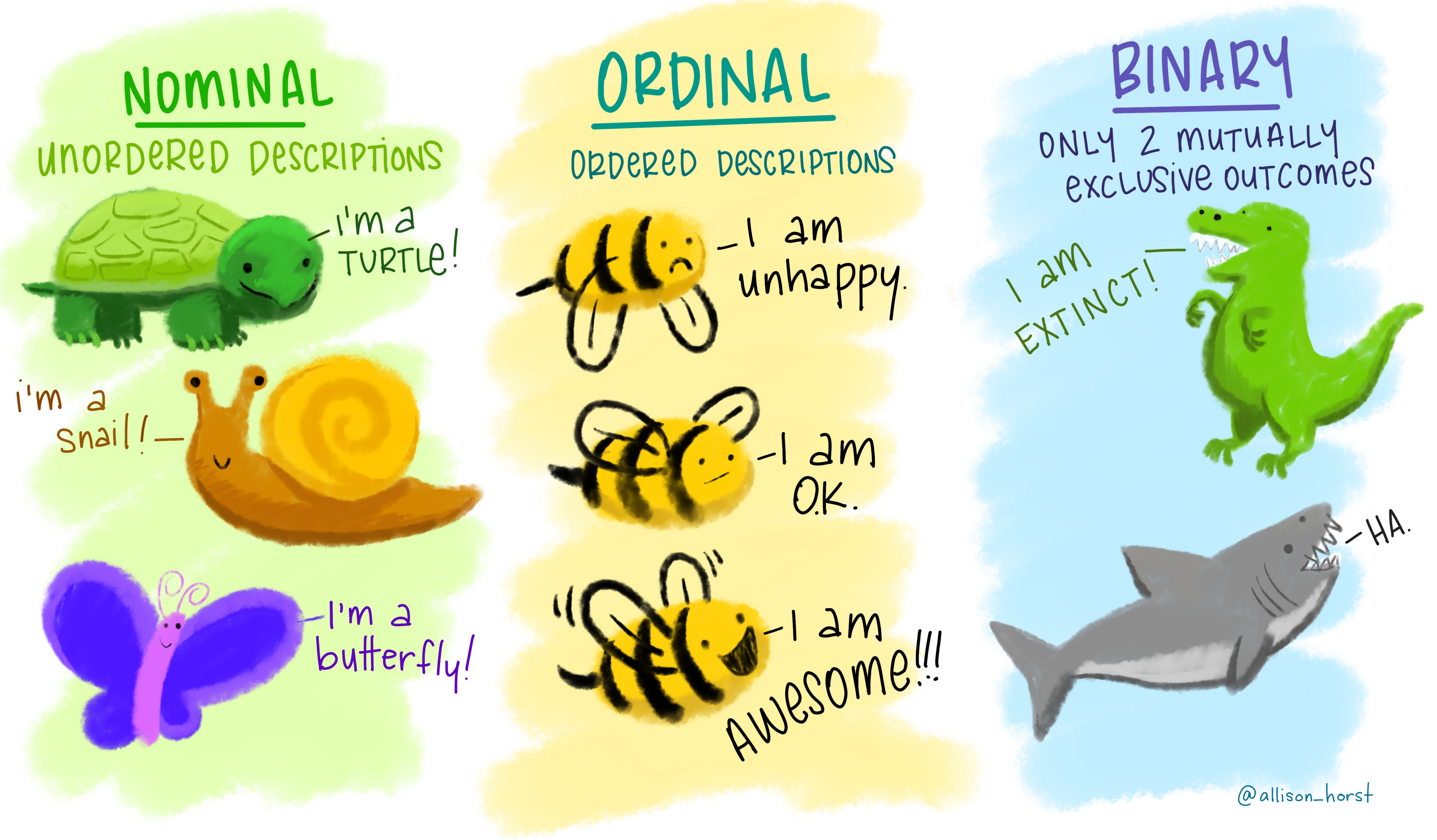
Quick break?



Recap
Types of variables

Stats and R: [Variable types and examples]


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

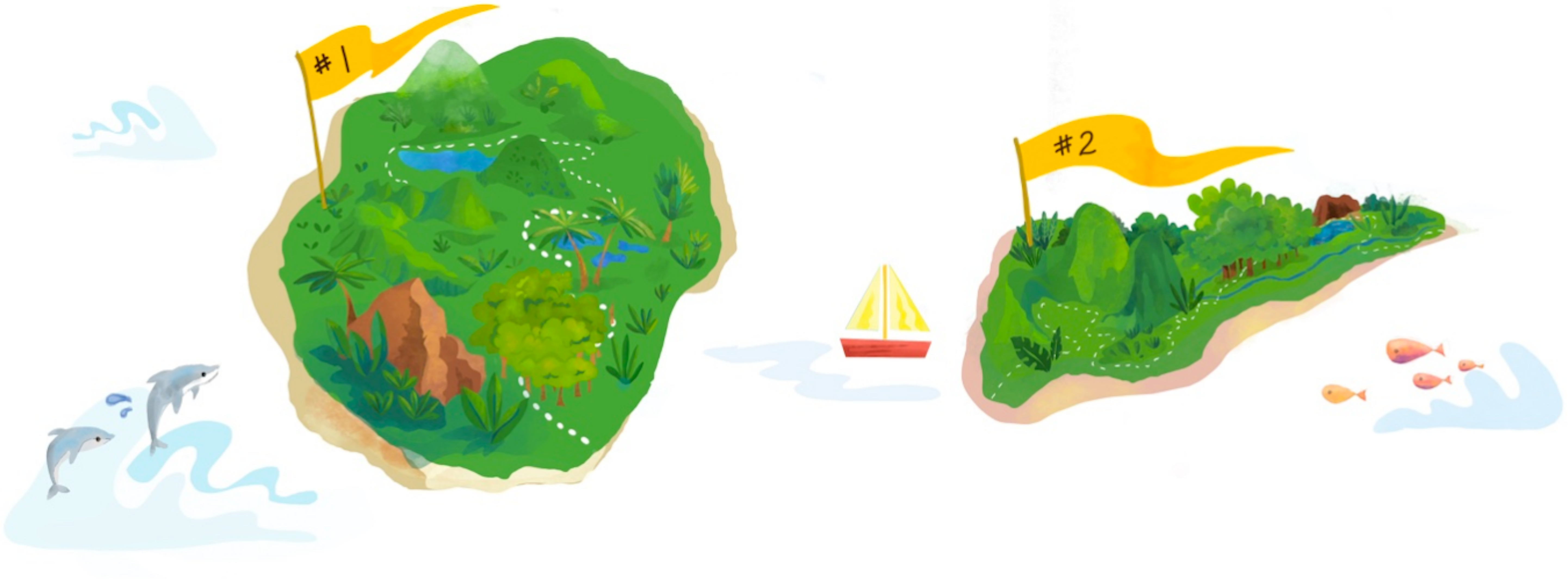

NOMINAL
unordered descriptions
I'm a TURTLE!
i'm a snail!
I'm a butterfly!
ORDINAL
ordered descriptions
I am unhappy.
I am O.K.
I am AWESOME!!!
BINARY
ONLY 2 MUTUALLY exclusive outcomes
I am EXTINCT!
HA.
@allison_horst

Recap
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

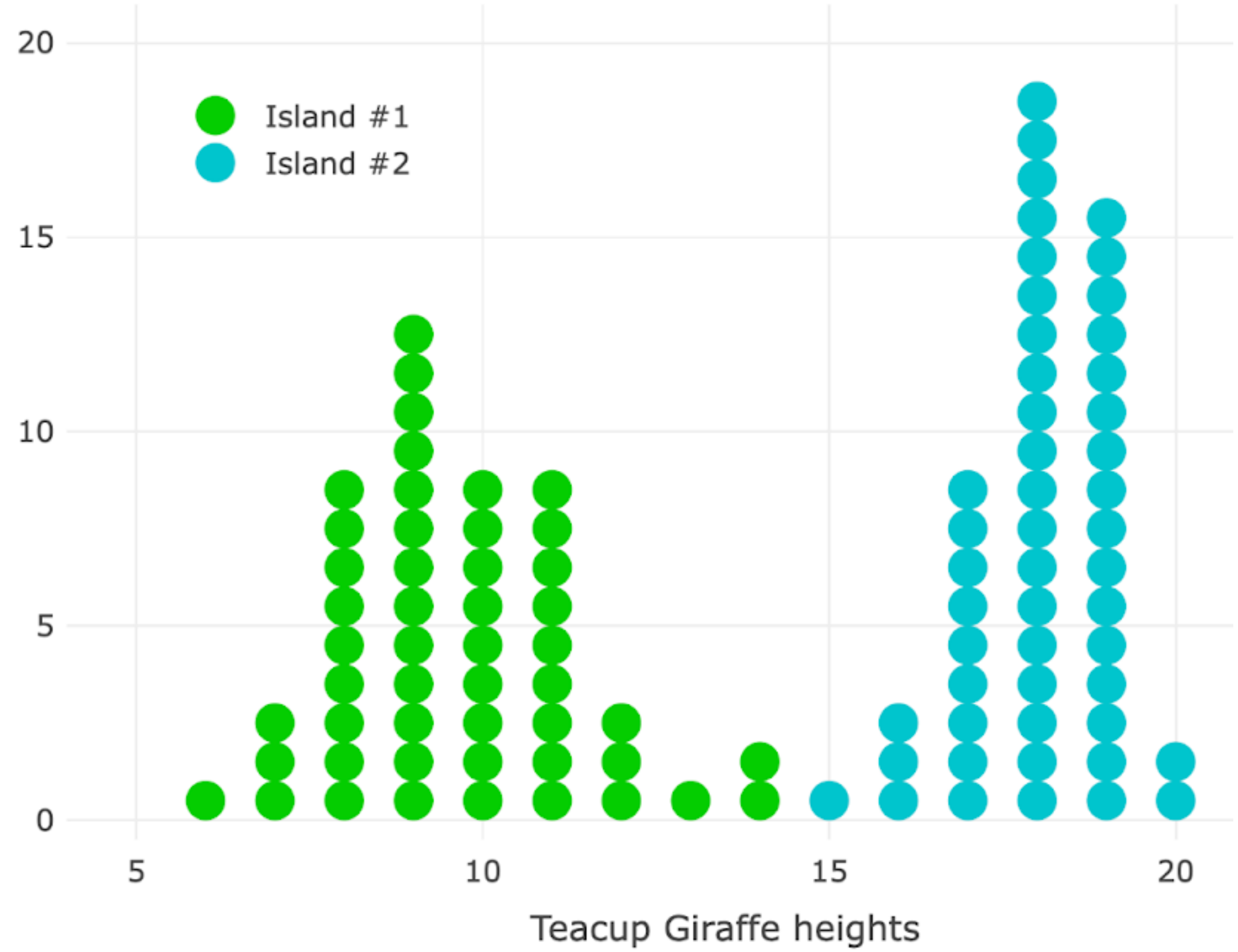


A distribution

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

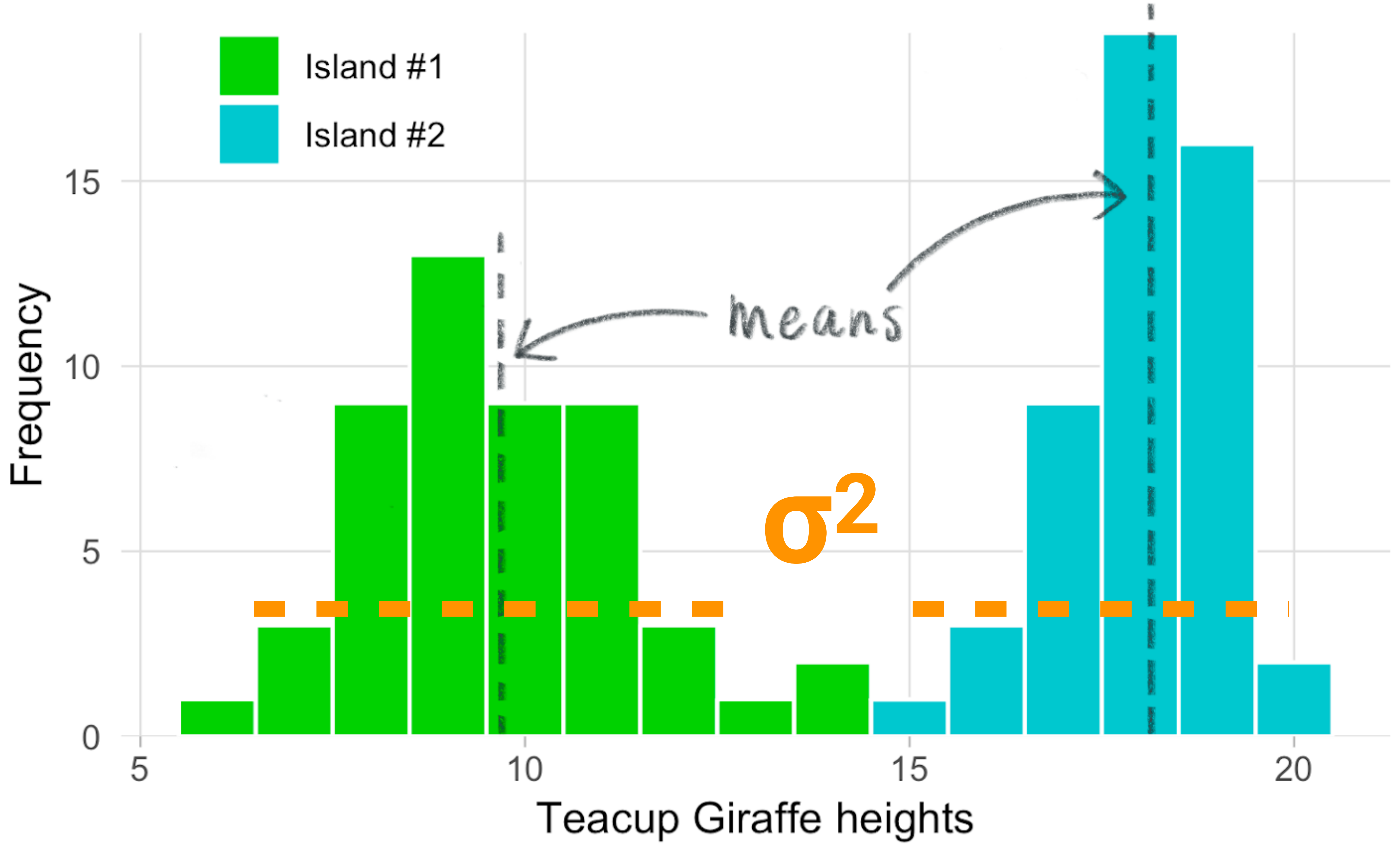


Parameters of the normal distribution

μ - the mean or expectation

σ - the standard deviation

or **σ2** - the variance



Continuous distributions in R

Are associated with 4 standard functions:

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

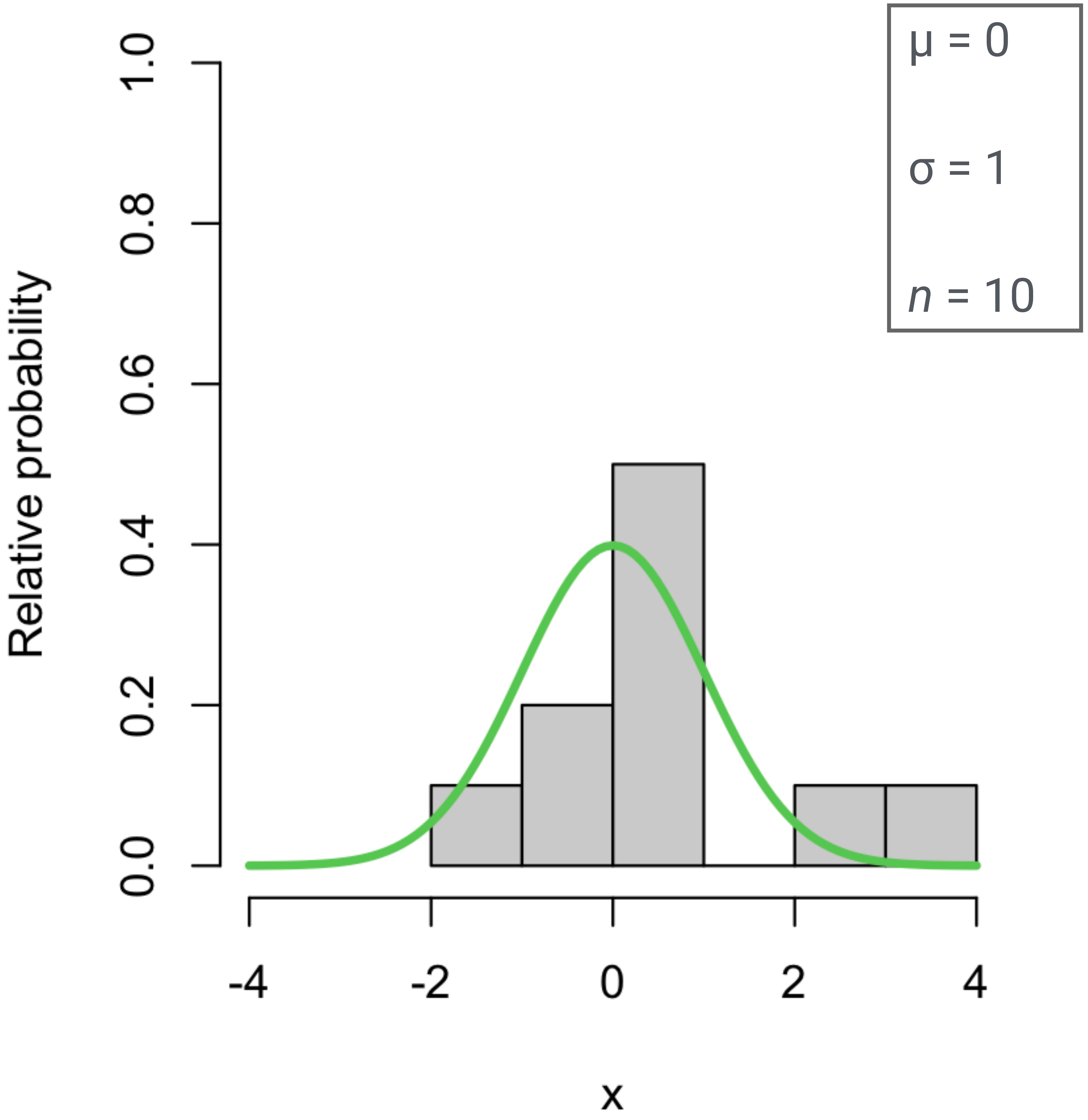
`pnorm(q, mean = 0, sd = 1)` - cumulative distribution function (% of values < than q)

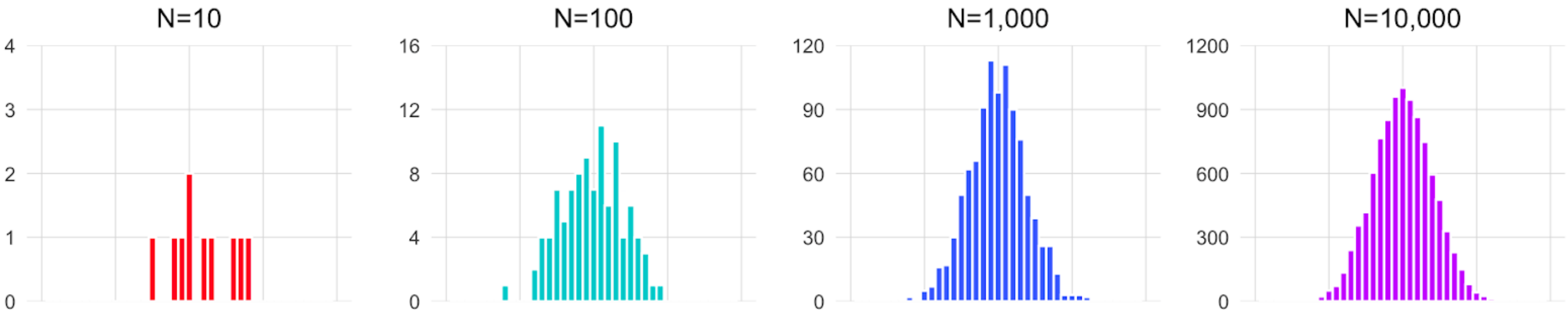
`qnorm(p, mean = 0, sd = 1)` - quantile function (inverse of cumulative distribution)

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

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

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



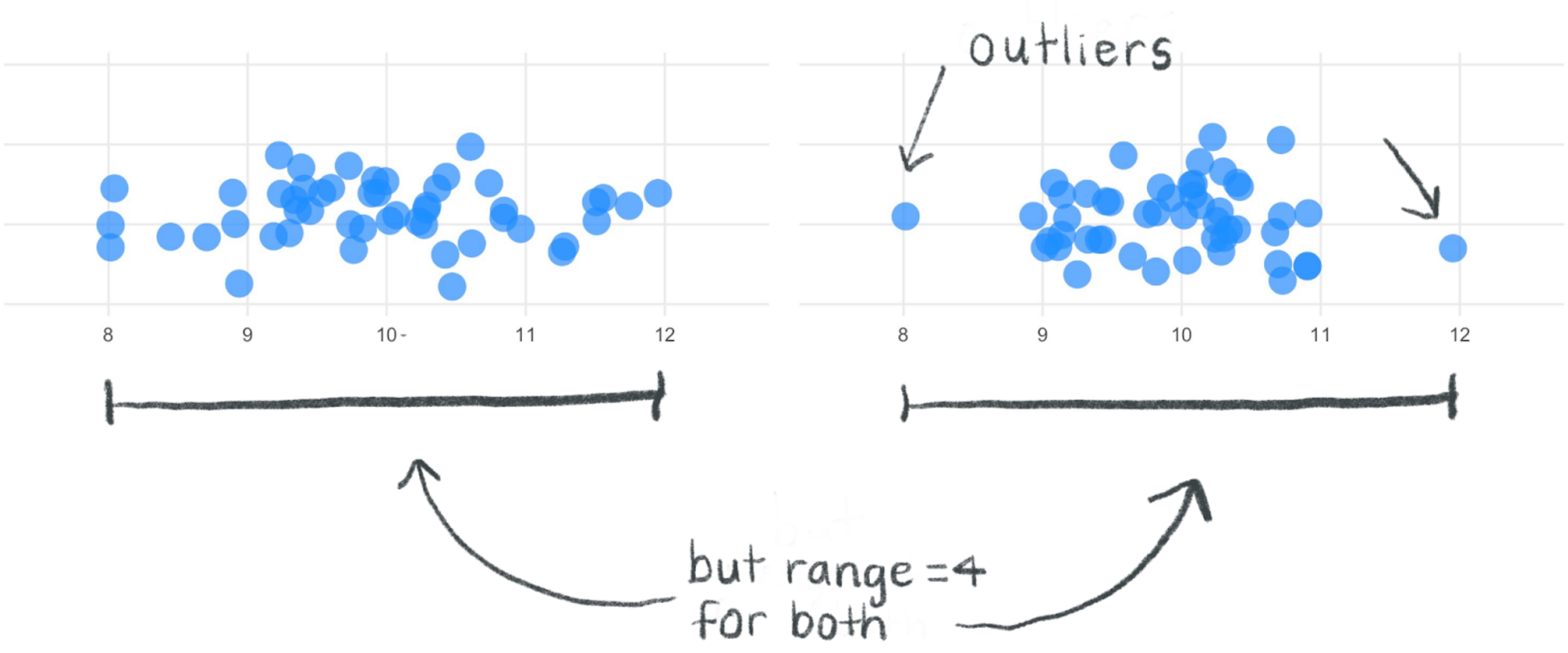

N=10
4
3
2
1
0
N=100
16
12
8
4
0
N=1,000
120
90
60
30
0
N=10,000
1200
900
600
300
0

Recap
Standard deviation and variance

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


are your
summary statistics
hiding something
interesting?
@allison_horst

Spread of the data - range



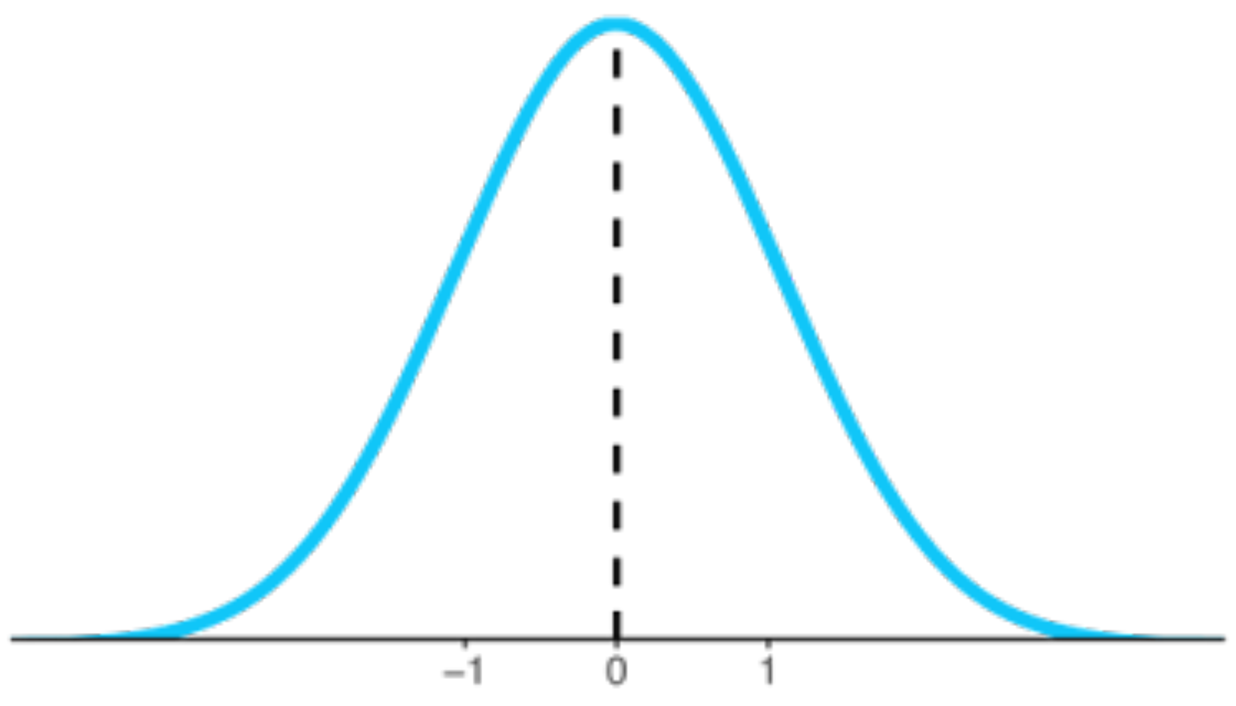
The first step of any analysis is often to **visualise the data**

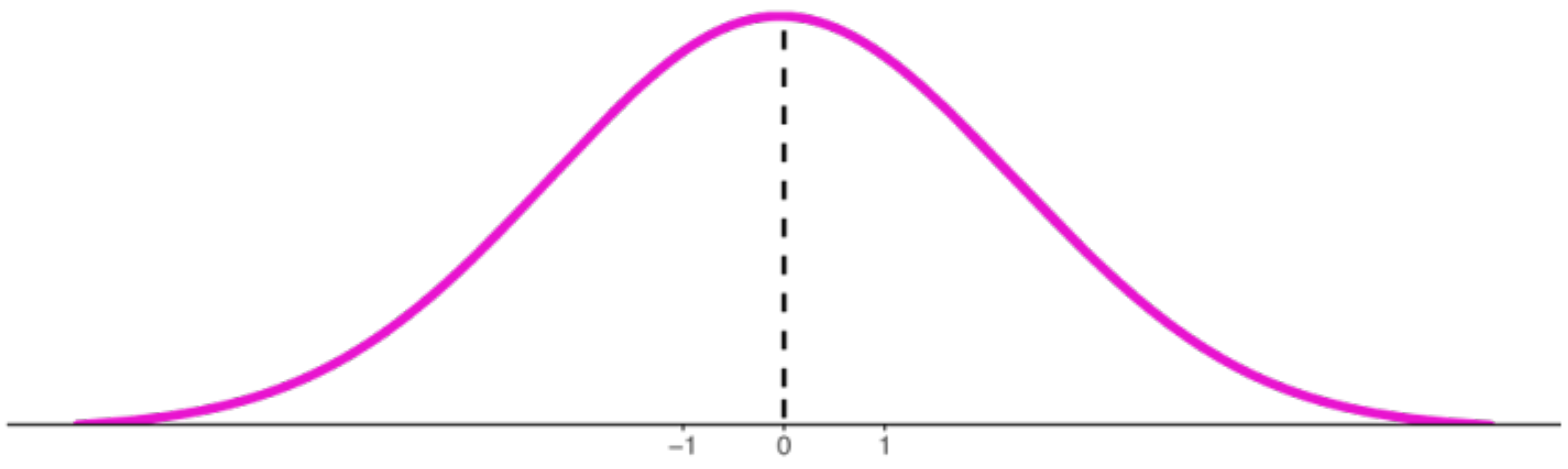
If we want to avoid undue influence of the outliers, the range is not good measure.

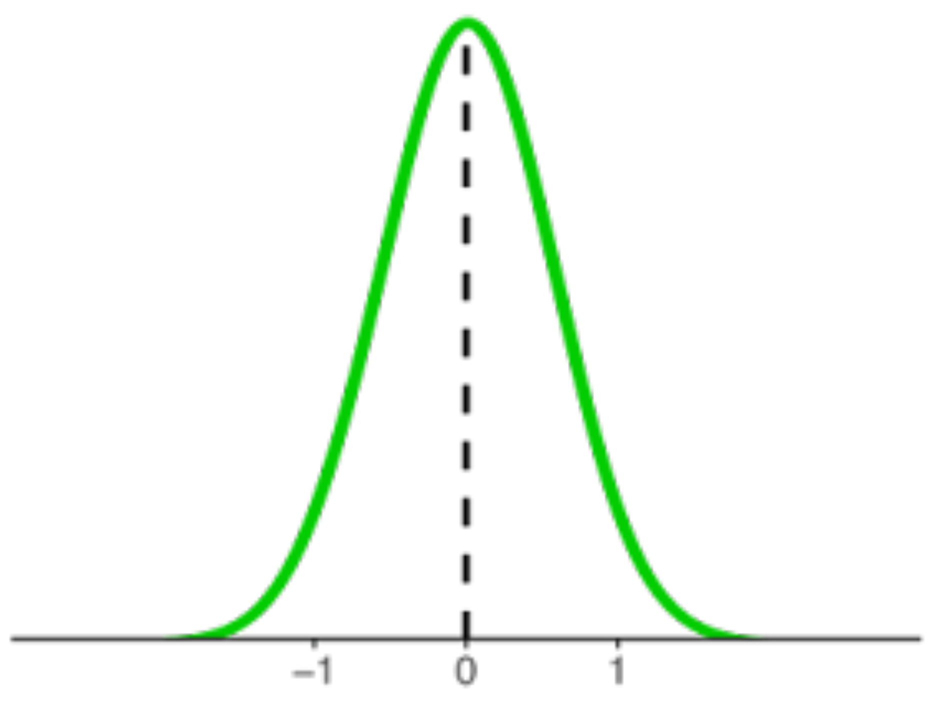
Spread of the data

The **standard deviation** (**σ**) and **variance** (**σ²**) account for outliers

We need a good understanding of this to grasp the mechanics of common statistical tests



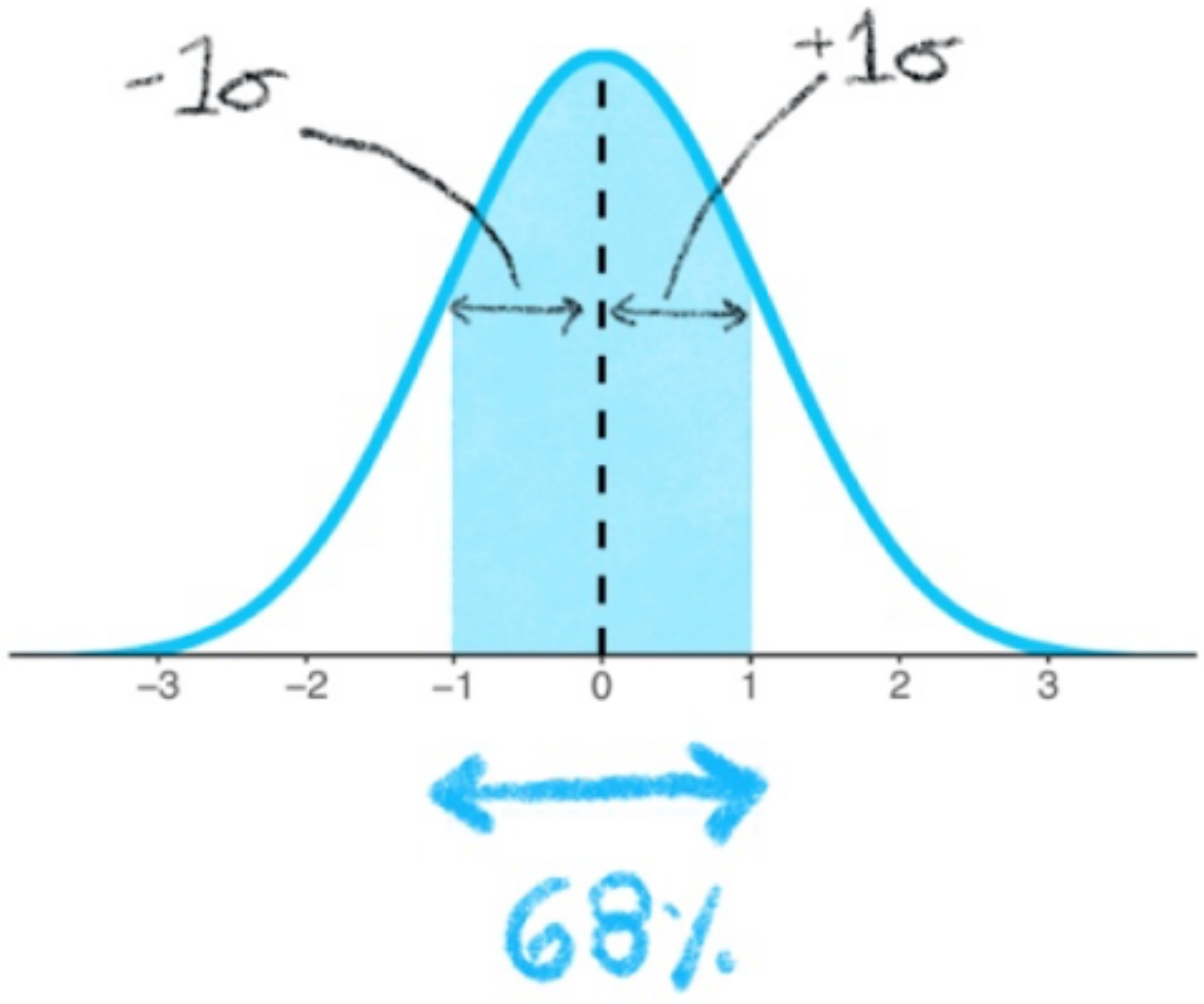




Standard deviation

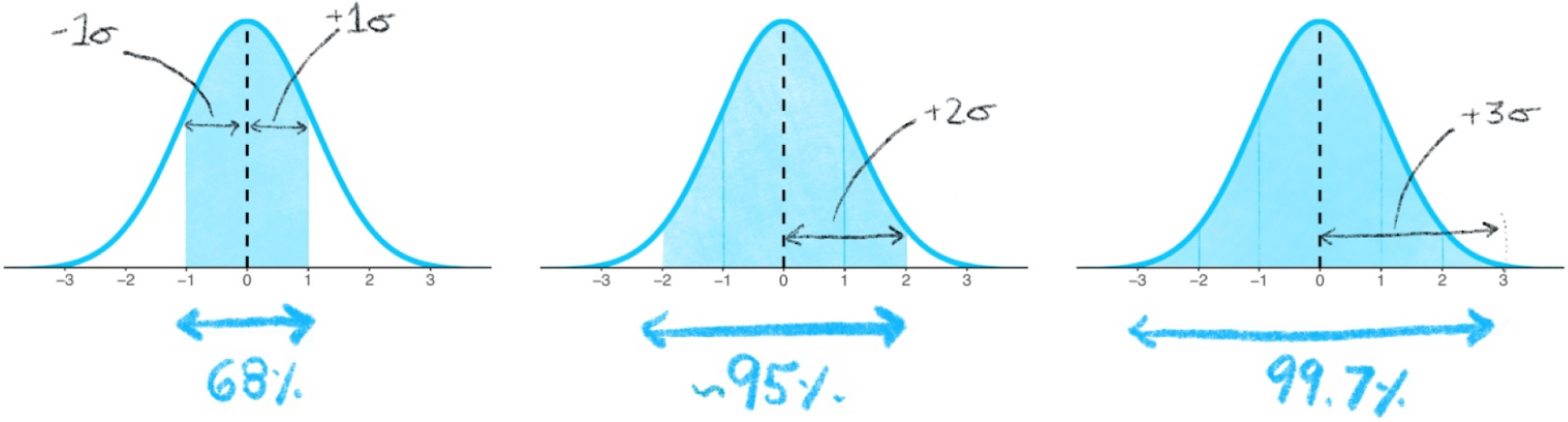
A measure of the amount of **variation** or **dispersion** in a set of normal distributed values

How do we calculate this?
→ See [Variance and Standard Deviation]



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

Interpreting the standard deviation



Introduction to hypothesis testing

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

Objectives

- Research hypotheses vs. statistical hypotheses
- Null vs. alternative hypotheses
- Significance and *p*-values

Hypotheses testing is an integral part of the scientific method

A research question

A clearly formulated question about a phenomenon that can be answered using data

Identifies a relationship, difference, or pattern we want to understand and defines the scope of our investigation

It’s about real world data and phenomena, not parameters

Helps guide what data need to be collected

A research hypotheses

Your proposed answer to the research question

A testable prediction about the relationship between real world variables, stated in conceptual terms

Translates a research question into a falsifiable claim about a relationship between variables

It is grounded in theory, and prior evidence. It can be wrong!

Examples

Question: *Does temperature influence species richness across elevation gradients?*

Hypothesis: *Species richness decreases with increasing elevation.*

Question: *Is sedimentation rate associated with fossil preservation quality?*

Hypothesis: *Higher sedimentation rates lead to better fossil preservation.*

Reading group

Group exercise

In your groups, discuss the following:

- Can you identify a clear research question?
- What is the the research hypothesis?

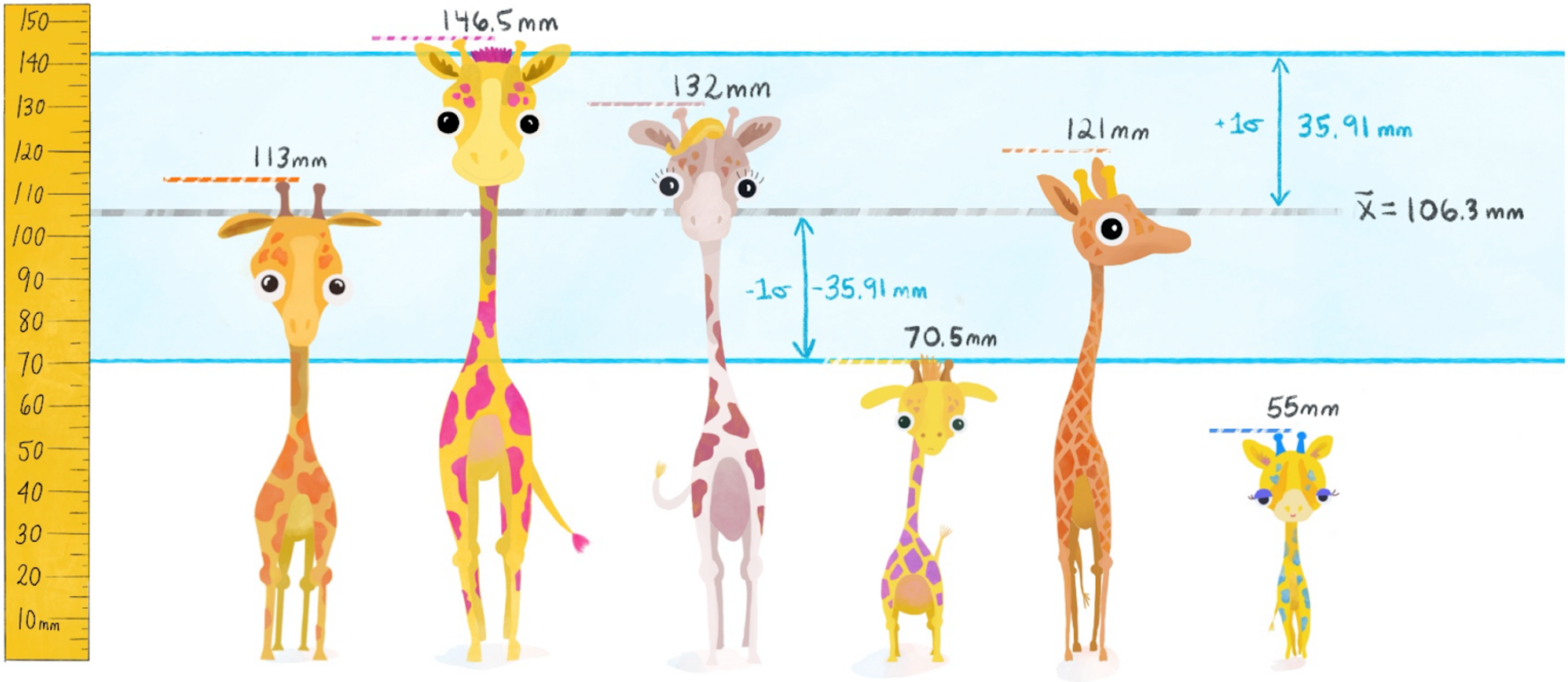
Make a note 📝

Research question: a clearly formulated question about a phenomenon that can be answered using data

Research hypotheses: a testable statement that provides a specific, directional expectation about the answer to the research question

Statistical hypothesis: formalises our prediction in terms of parameters

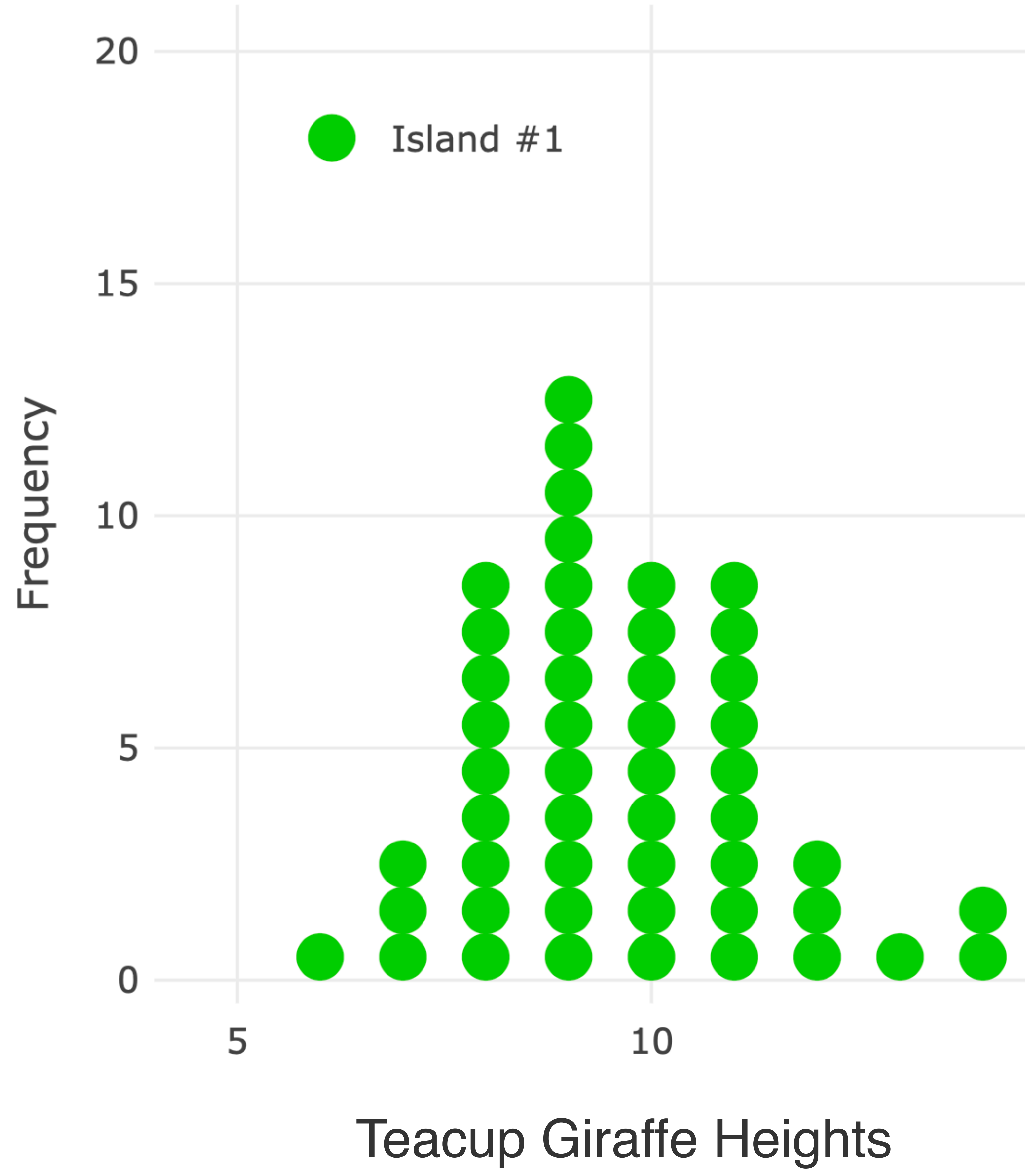
⚠️ A statistical hypothesis (or test) is a test of the statistical hypothesis, not the research hypothesis!



Imagine we have a hypothesis about the average height of giraffes from one of the Islands

Example hypothesis

My hypothesis: **the average height of a giraffe from Island 1 is 10 cm**



My hypothesis: **the average height of a giraffe from Island 1 is 10 cm**

Say, I collected **20 giraffes** and the **average is 9.7 cm**

How much doubt does this cast on my hypothesis?

What if, I collect **20 giraffes** and the **average is 8.9 cm**

*How much doubt does **this** cast on my hypothesis?*

How likely is it, that this sample is **randomly less** that the true mean?

Statistical hypothesis testing

The **null hypothesis** H_0 is $\mu = 10$ cm

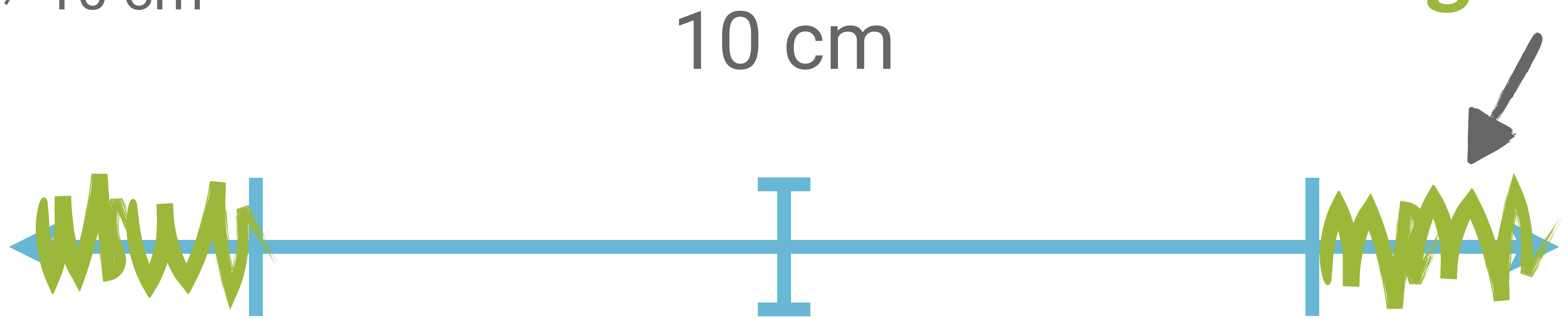
The **alternative hypothesis** H_1 is $\mu \neq 10$ cm

Is our sample mean ($\bar{x}$) far enough away from 10 cm for us to **reject** the null hypothesis?

H_0 is $\mu = 10$ cm

H_1 is $\mu \neq 10$ cm

Rejection regions



Where does the sample mean ($\bar{x}$) lie?

Hypotheses establish critical values beyond which, we can say, ‘we **reject** H_0’

Another example

My hypothesis: **the average age of first semesters students is 21**

Based on a sample of **5 students**, the **average age is 24**

How much doubt does this cast of my hypothesis?

Based on a different sample of **200 students**, the **average age is 24**

*How much doubt does **this** cast of my hypothesis?*

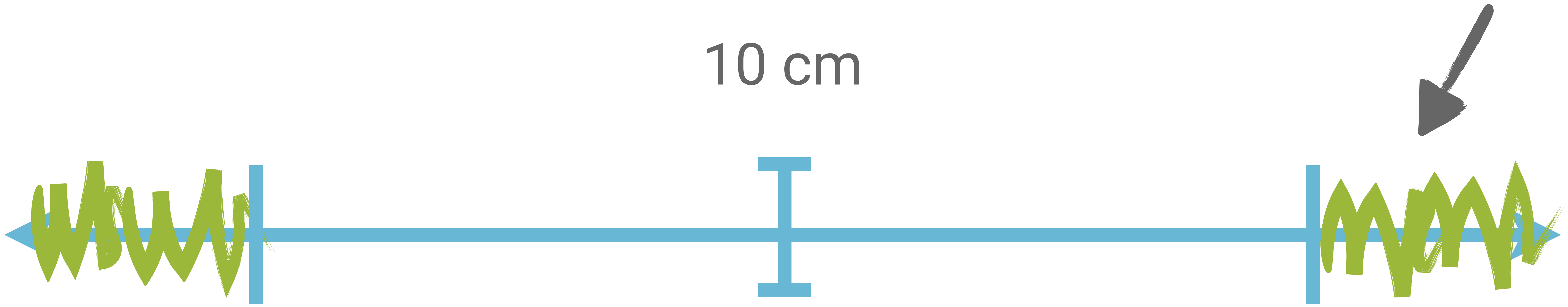
If H_0 is true, how extreme is our sample?

Intuitively, we are more likely to reject H_0 if:

- sample difference is greater
- number of observations is greater

We can *never* disprove the null hypothesis. It's always possible to get an extreme sample

There's always a possibility of incorrectly rejecting it



H_0 could still be true, even if the sample mean is here

Type I and II errors

A **type I error** occurs when we reject a null hypothesis that is *true*
Prob = α

A **type II error** occurs when we **accept** (don’t reject) a null hypothesis that is *false*
Prob = β (also known as the **power** of the test)

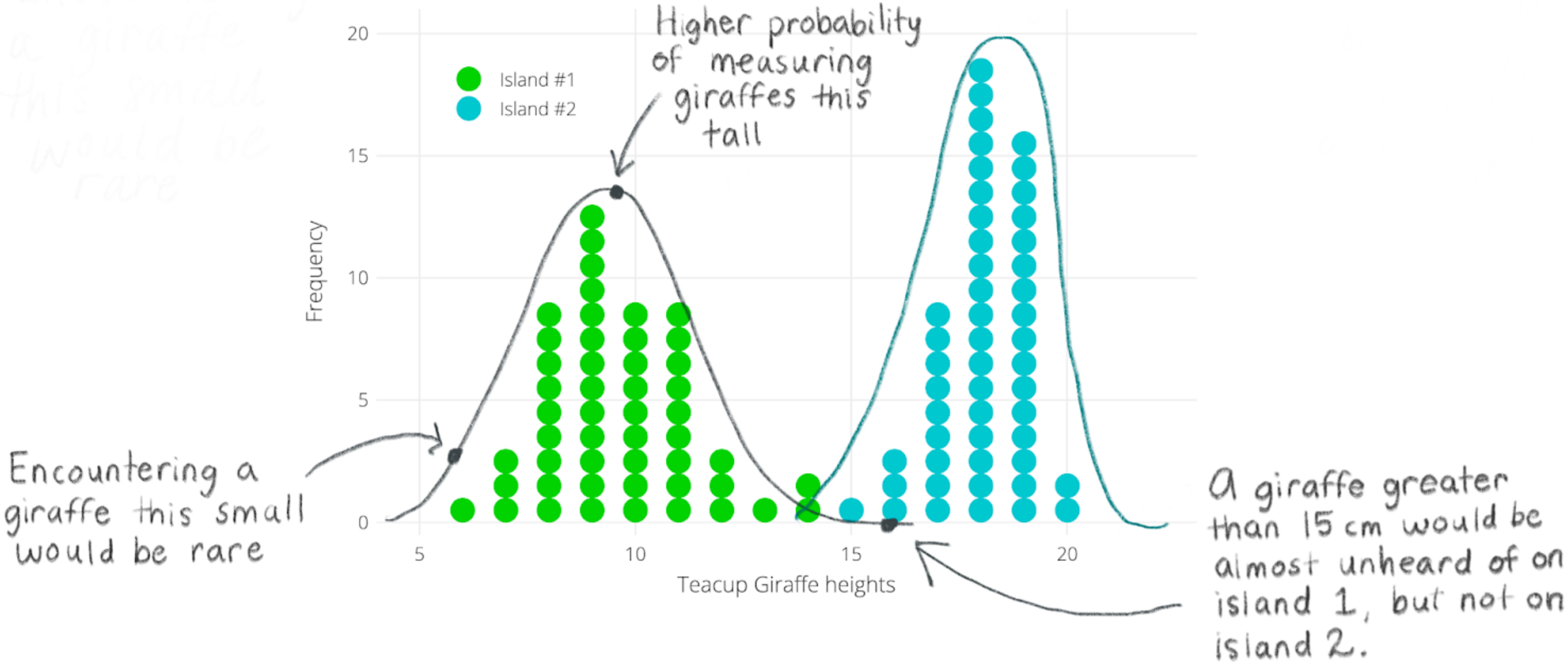
Type I and type II errors are *always* possible

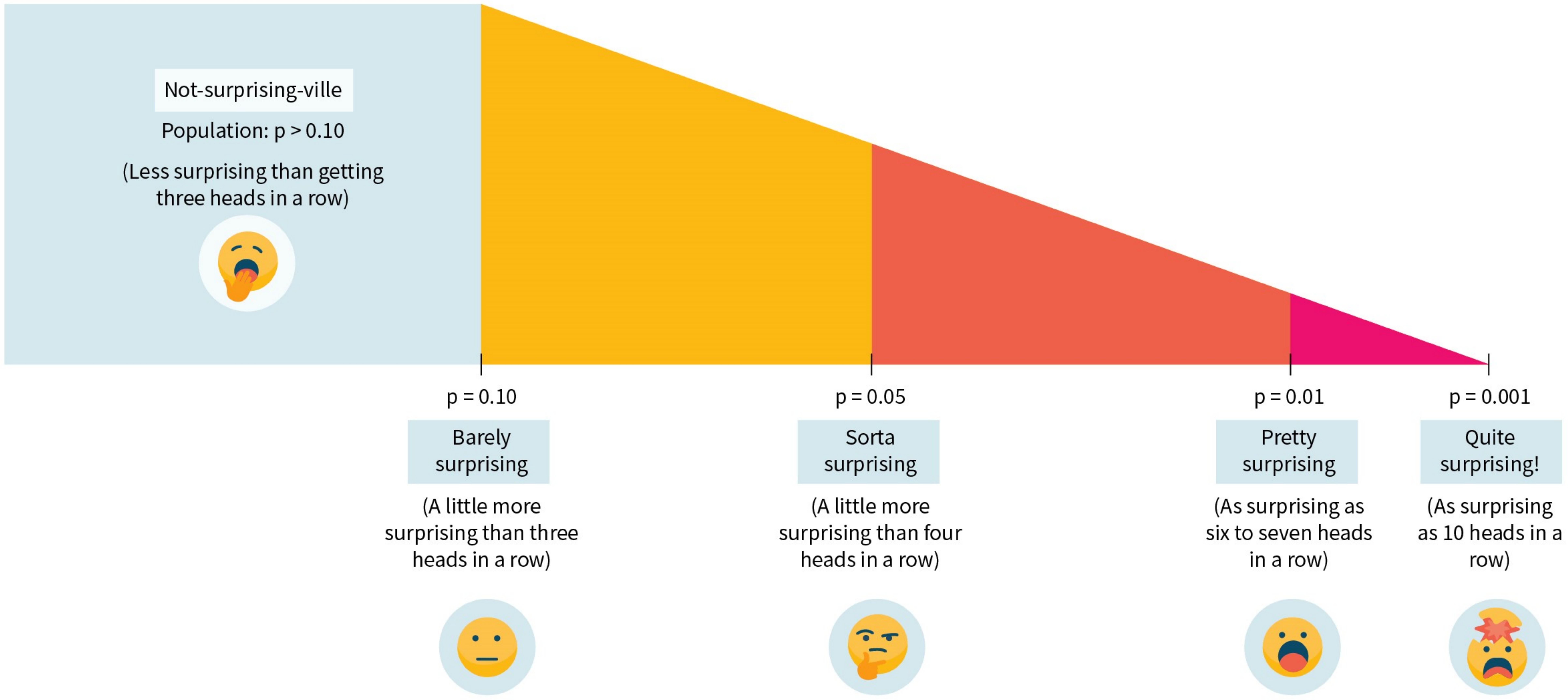
p-values

α is a predetermined threshold

The *p*-value is the probability of observing the results if the null hypothesis is true. This calculated from the data

If the *p*-value is **less than or equal to** alpha, the result is considered **statistically significant**


Higher probability of measuring giraffes this tall
Island #1
Island #2
Frequency
20
15
10
5
0
5
10
15
20
Teacup Giraffe heights
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.


Not-surprising-ville
Population: p > 0.10
(Less surprising than getting three heads in a row)
p = 0.10
Barely surprising
(A little more surprising than three heads in a row)
p = 0.05
Sorta surprising
(A little more surprising than four heads in a row)
p = 0.01
Pretty surprising
(As surprising as six to seven heads in a row)
p = 0.001
Quite surprising!
(As surprising as 10 heads in a row)

Reporting *p*-values

Option 1: you can state only that $p < \alpha$ for a significance level that you chose in advance, e.g., $p<0.05$.

But this means we're treating $p=0.051$ in a fundamentally different way to $p=0.049$

Option 2: just report the actual *p*-value and let the reader make up their own mind

If you get $p=.062$, it means that you have to be willing to tolerate a type I error rate of 6.2% to justify rejecting the null. If you find 6.2% intolerable, you can retain the null

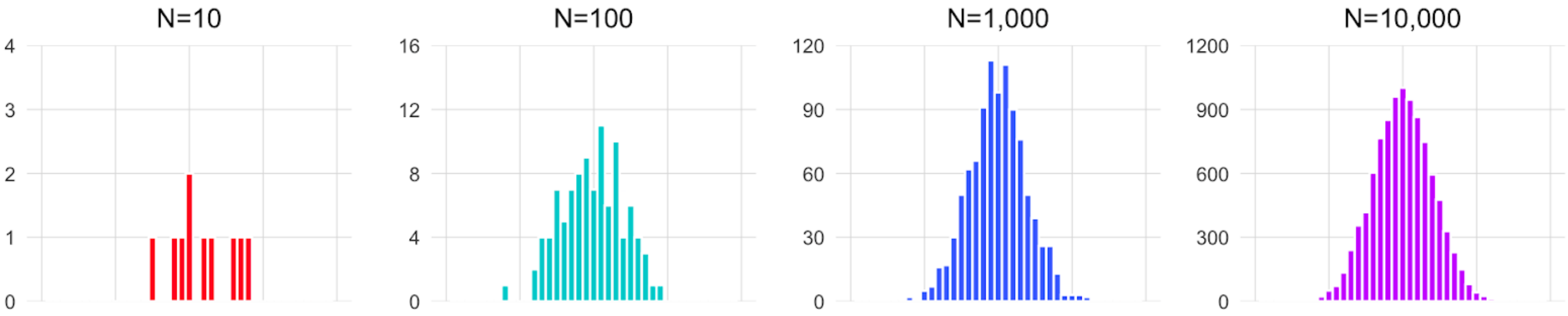
The (student's) *t*-distribution and *t*-test

t-distribution and *t*-test

Described by William Sealy Gosset in 1908 under a pseudonym ‘A student’

Overcomes the problem of ‘small sample statistics’

Assumes the underlying distribution is normal (e.g., height). Used when the true standard deviation is **unknown**

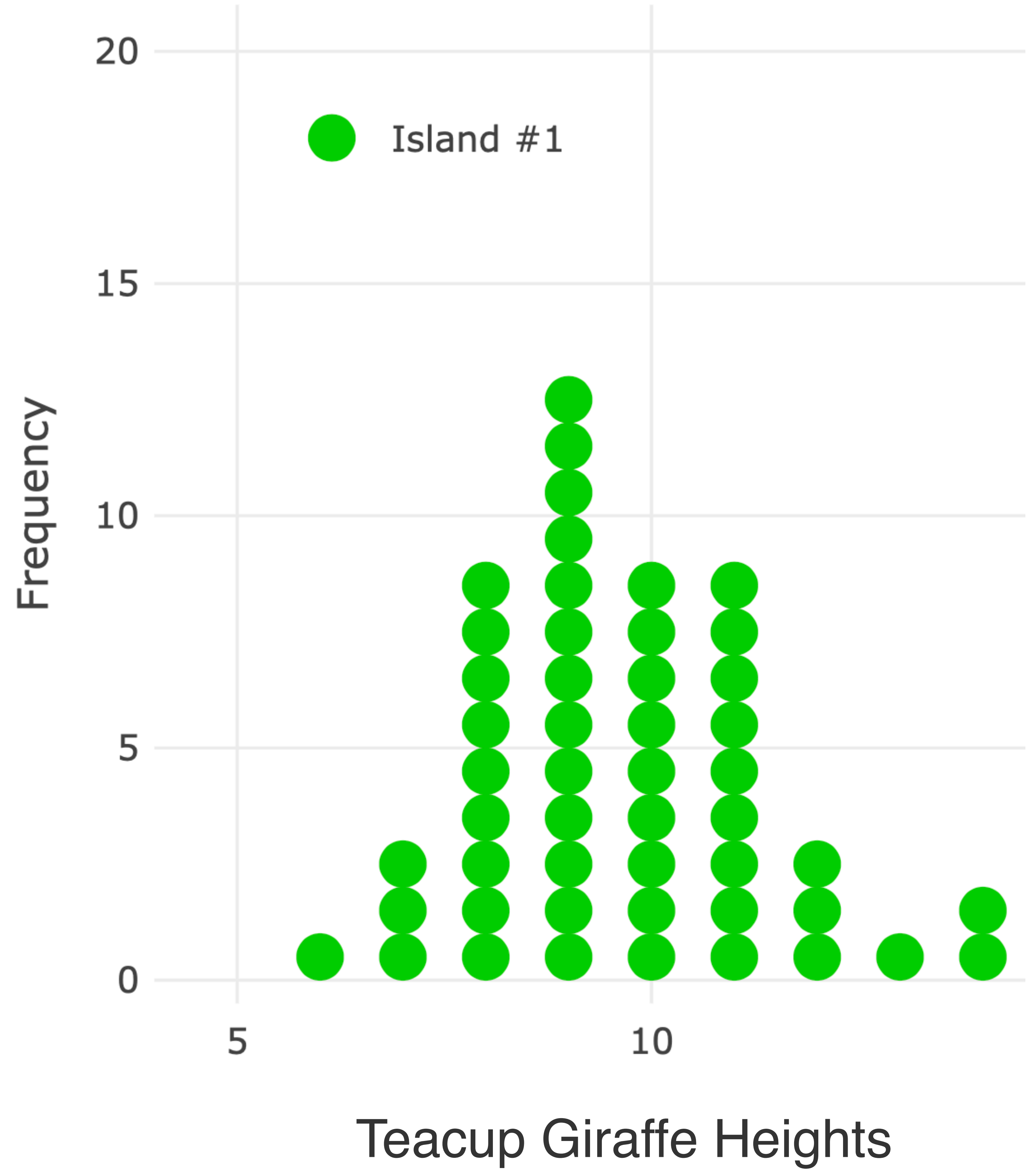

N=10
4
3
2
1
0
N=100
16
12
8
4
0
N=1,000
120
90
60
30
0
N=10,000
1200
900
600
300
0

My hypothesis: **the average height of a giraffe from Island 1 is 10 cm**

H_0 is $\mu = 10$ cm

H_1 is $\mu \neq 10$ cm

If H_0 is true, we would expect the sample mean ($\bar{x}$) minus the population mean (μ) to be zero →



$$t = \frac{\bar{x} - \mu}{s\sqrt{n}}$$

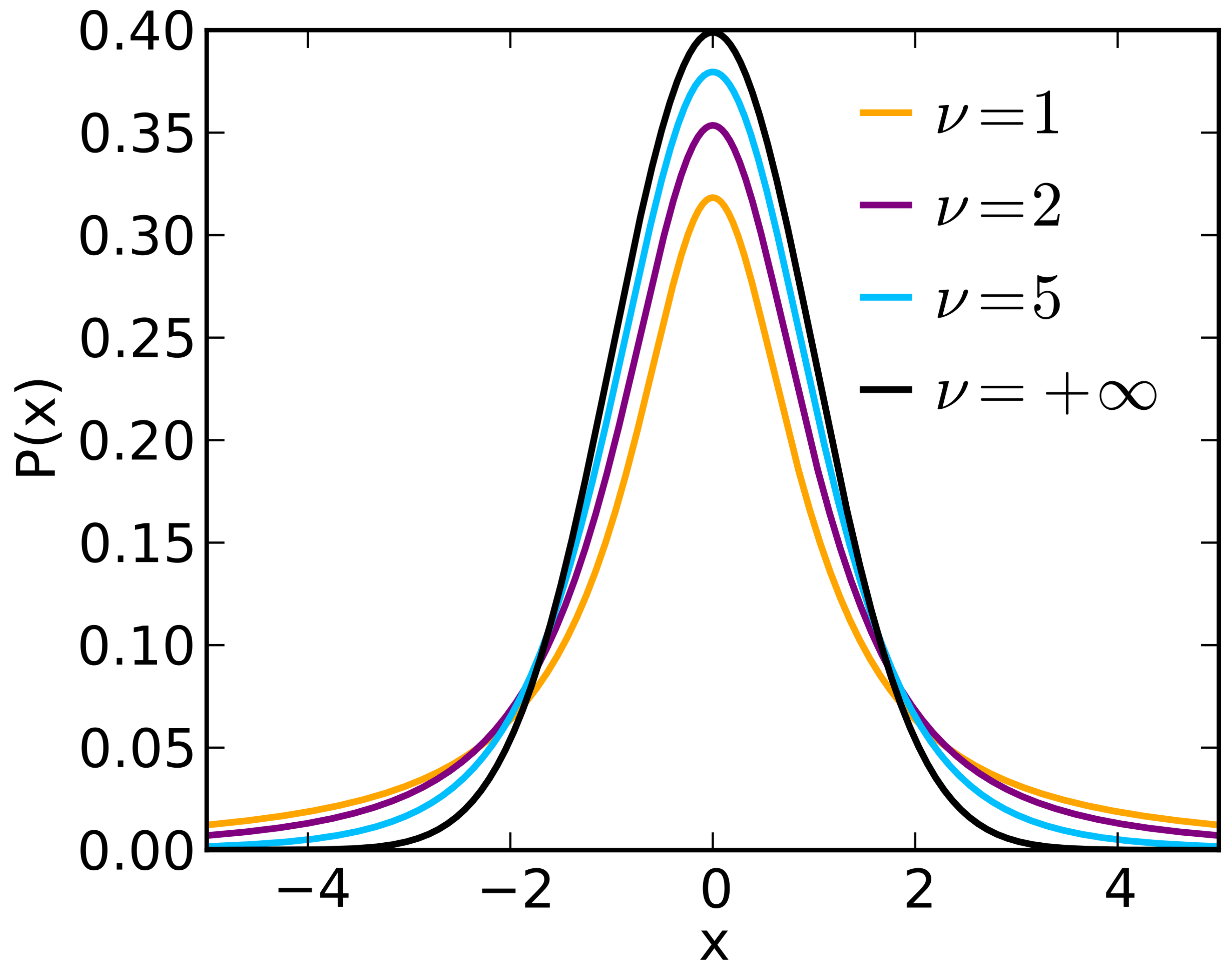
The *t*-statistic

Sample mean minus our reference value

$$t = \frac{\bar{x} - \mu}{s\sqrt{n}}$$

Standard deviation

If our hypothesis is true the mean of *t* will be zero



The parameters of the distribution are **degrees of freedom (*n* - 1)**, accounts for uncertainty in s

We can look up the critical regions in a table

Highlighted: the value of t at $\alpha = 0.05$

Significance level (α)

Degrees of freedom *(df)*	.2	.15	.1	.05	.025	.01	.005	.001
1	3.078	4.165	6.314	12.706	25.452	63.657	127.321	636.619
2	1.886	2.282	2.920	4.303	6.205	9.925	14.089	31.599
3	1.638	1.924	2.353	3.182	4.177	5.841	7.453	12.924
4	1.533	1.778	2.132	2.776	3.495	4.604	5.598	8.610
5	1.476	1.699	2.015	2.571	3.163	4.032	4.773	6.869
6	1.440	1.650	1.943	2.447	2.969	3.707	4.317	5.959
7	1.415	1.617	1.895	2.365	2.841	3.499	4.029	5.408
8	1.397	1.592	1.860	2.306	2.752	3.355	3.833	5.041
9	1.383	1.574	1.833	2.262	2.685	3.250	3.690	4.781
10	1.372	1.559	1.812	2.228	2.634	3.169	3.581	4.587
11	1.363	1.548	1.796	2.201	2.593	3.106	3.497	4.437
12	1.356	1.538	1.782	2.179	2.560	3.055	3.428	4.318
13	1.350	1.530	1.771	2.160	2.533	3.012	3.372	4.221
14	1.345	1.523	1.761	2.145	2.510	2.977	3.326	4.140
15	1.341	1.517	1.753	2.131	2.490	2.947	3.286	4.073
16	1.337	1.512	1.746	2.120	2.473	2.921	3.252	4.015
17	1.333	1.508	1.740	2.110	2.458	2.898	3.222	3.965
18	1.330	1.504	1.734	2.101	2.445	2.878	3.197	3.922
19	1.328	1.500	1.729	2.093	2.433	2.861	3.174	3.883
20	1.325	1.497	1.725	2.086	2.423	2.845	3.153	3.850
21	1.323	1.494	1.721	2.080	2.414	2.831	3.135	3.819
22	1.321	1.492	1.717	2.074	2.405	2.819	3.119	3.792
23	1.319	1.489	1.714	2.069	2.398	2.807	3.104	3.768
24	1.318	1.487	1.711	2.064	2.391	2.797	3.091	3.745
25	1.316	1.485	1.708	2.060	2.385	2.787	3.078	3.725
26	1.315	1.483	1.706	2.056	2.379	2.779	3.067	3.707
27	1.314	1.482	1.703	2.052	2.373	2.771	3.057	3.690
28	1.313	1.480	1.701	2.048	2.368	2.763	3.047	3.674
29	1.311	1.479	1.699	2.045	2.364	2.756	3.038	3.659
30	1.310	1.477	1.697	2.042	2.360	2.750	3.030	3.646
40	1.303	1.468	1.684	2.021	2.329	2.704	2.971	3.551
50	1.299	1.462	1.676	2.009	2.311	2.678	2.937	3.496
60	1.296	1.458	1.671	2.000	2.299	2.660	2.915	3.460
70	1.294	1.456	1.667	1.994	2.291	2.648	2.899	3.435
80	1.292	1.453	1.664	1.990	2.284	2.639	2.887	3.416
100	1.290	1.451	1.660	1.984	2.276	2.626	2.871	3.390
1000	1.282	1.441	1.646	1.962	2.245	2.581	2.813	3.300
Infinite	1.282	1.440	1.645	1.960	2.241	2.576	2.807	3.291

https://numiqo.com/tutorial/t-distribution

Performing a *t*-test in R

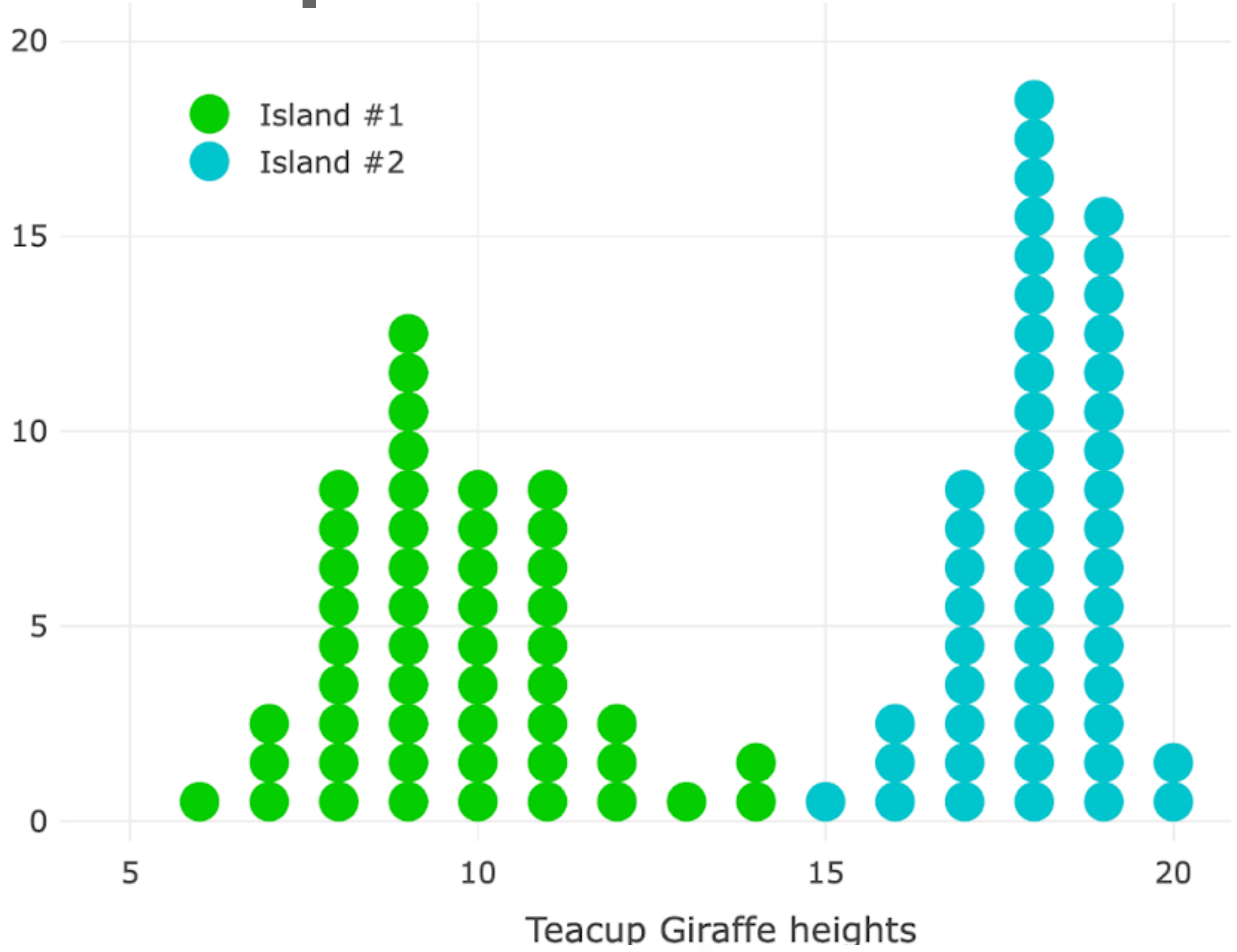
demo

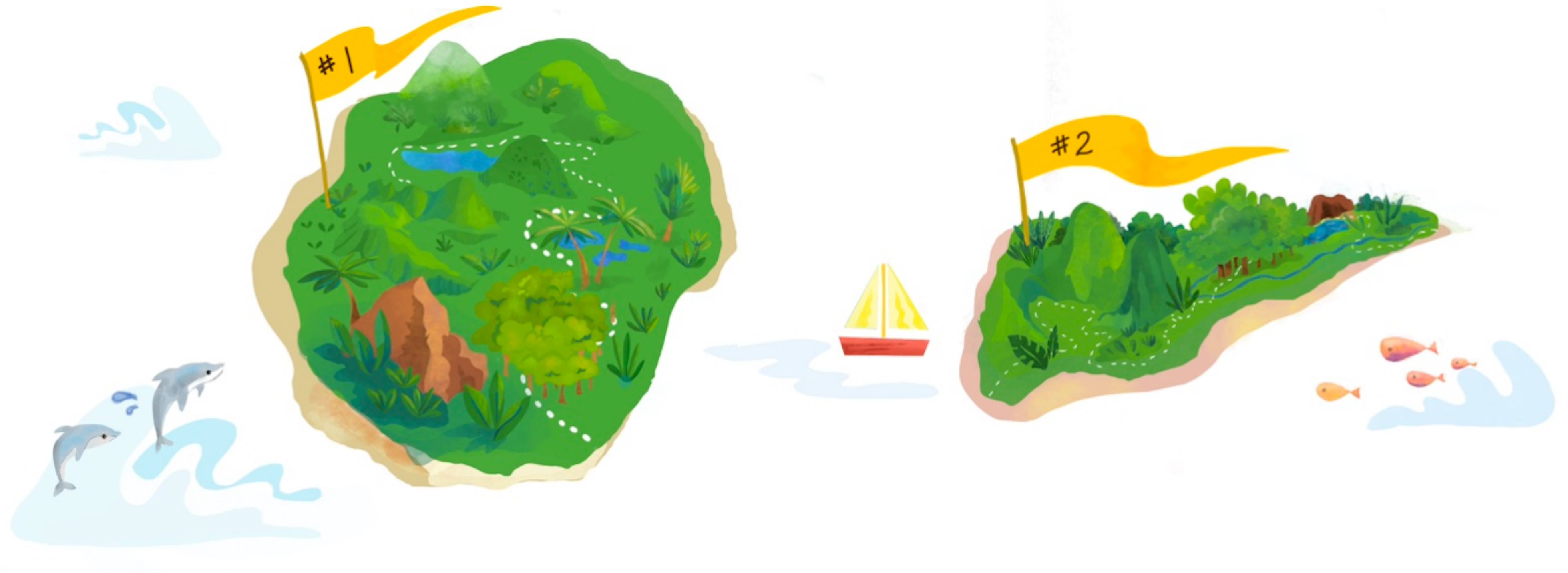
```
t.test(heights, mu = pop_mean)
```

Differences between samples

Independent samples *t*-test

Is there a significant difference between the heights of giraffes on Island 1 and Island 2?





Defining hypothesis

We have two sample means for Island 1 and Island 2

The **null hypothesis** H_0 is ???

The **alternative hypothesis** H_1 is ???

Defining hypothesis

We have two sample means for Island 1 and Island 2

The **null hypothesis** H_0 is there is no difference between the means

The **alternative hypothesis** H_1 is that the means are different

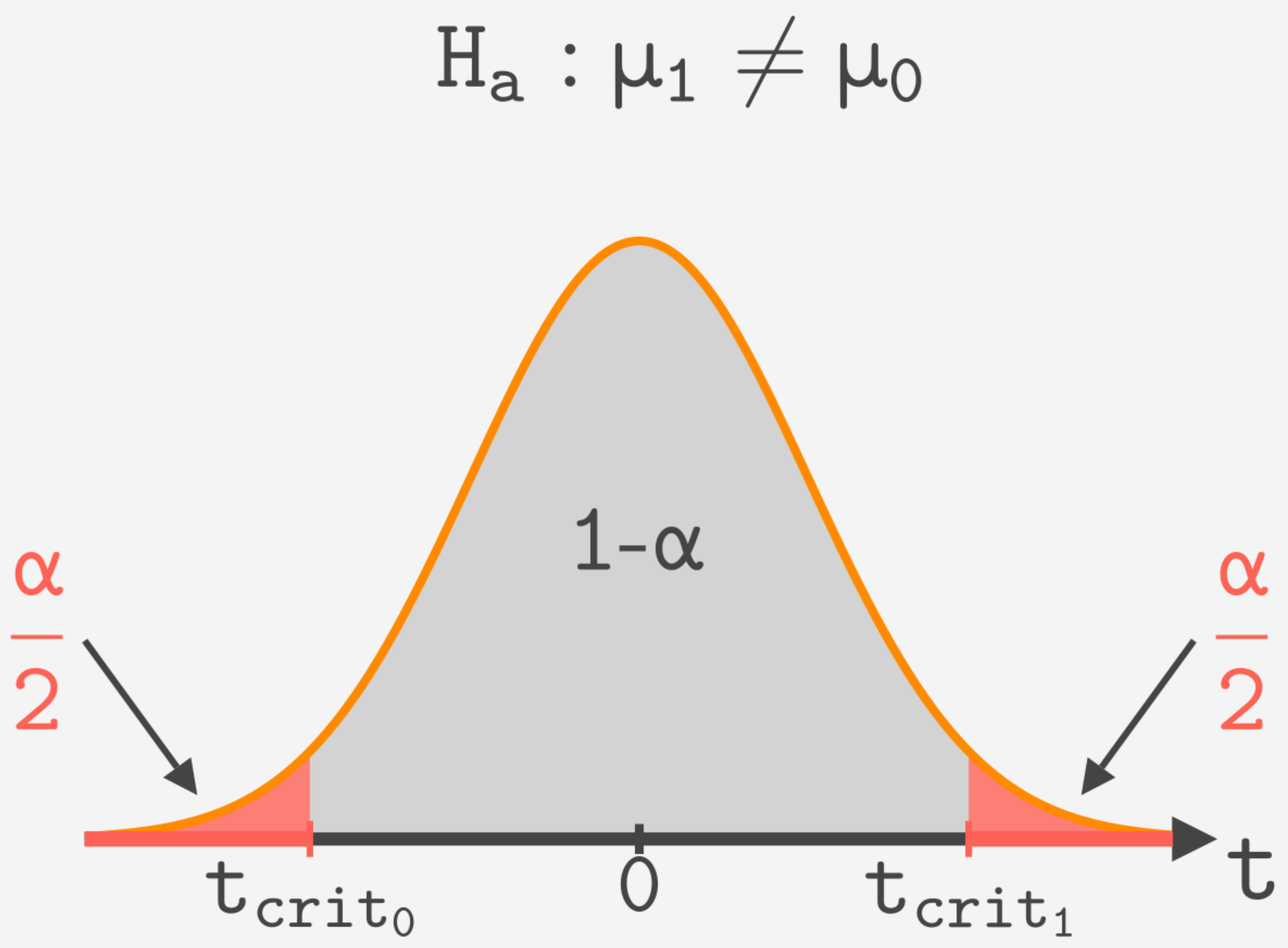
$$\rightarrow \quad \mathrm{t} = \frac{\text{Difference Between Sample Means}}{\text{Standard Error of the Difference}} \quad \rightarrow \quad t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

The **degrees of freedom are (n_1 - n_2 - 2)**, taking into account the sample size of each group

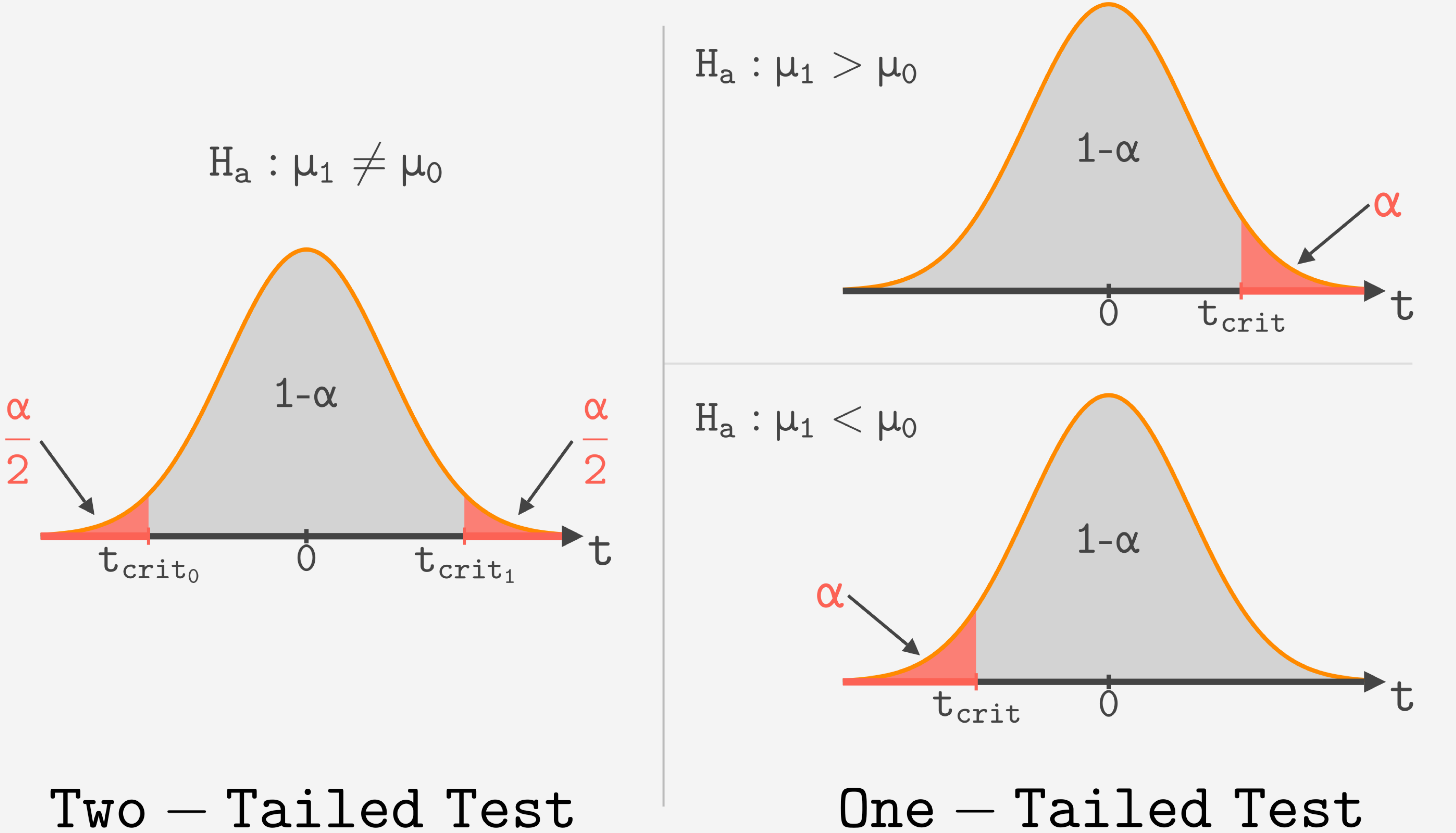
Performing an independent *t*-test in R

demo

```
t.test(group_a, group_b)
```



Two — Tailed Test



Ha : μ1 ≠ μ0
1-α
α/2
α/2
tcrit0
0
tcrit1
t
Two – Tailed Test
Ha : μ1 > μ0
1-α
α
0
tcrit
t
Ha : μ1 < μ0
1-α
α
tcrit
0
t
One – Tailed Test

Break?

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