(Re)Introducing some statistical terms

feat. teacup giraffes

Rachel Warnock with help from

tinystats.github.io/teacups-giraffes-and-statistics

21.02.2022

FAU

General concepts

- Intro to the Normal Distribution
- Mean, median, mode
- Spread of the data
- Standard Error

tinystats.github.io/teacups-giraffes-and-statistics

Types of Data

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

ORDINAL

ORDERED DESCRIPTIONS

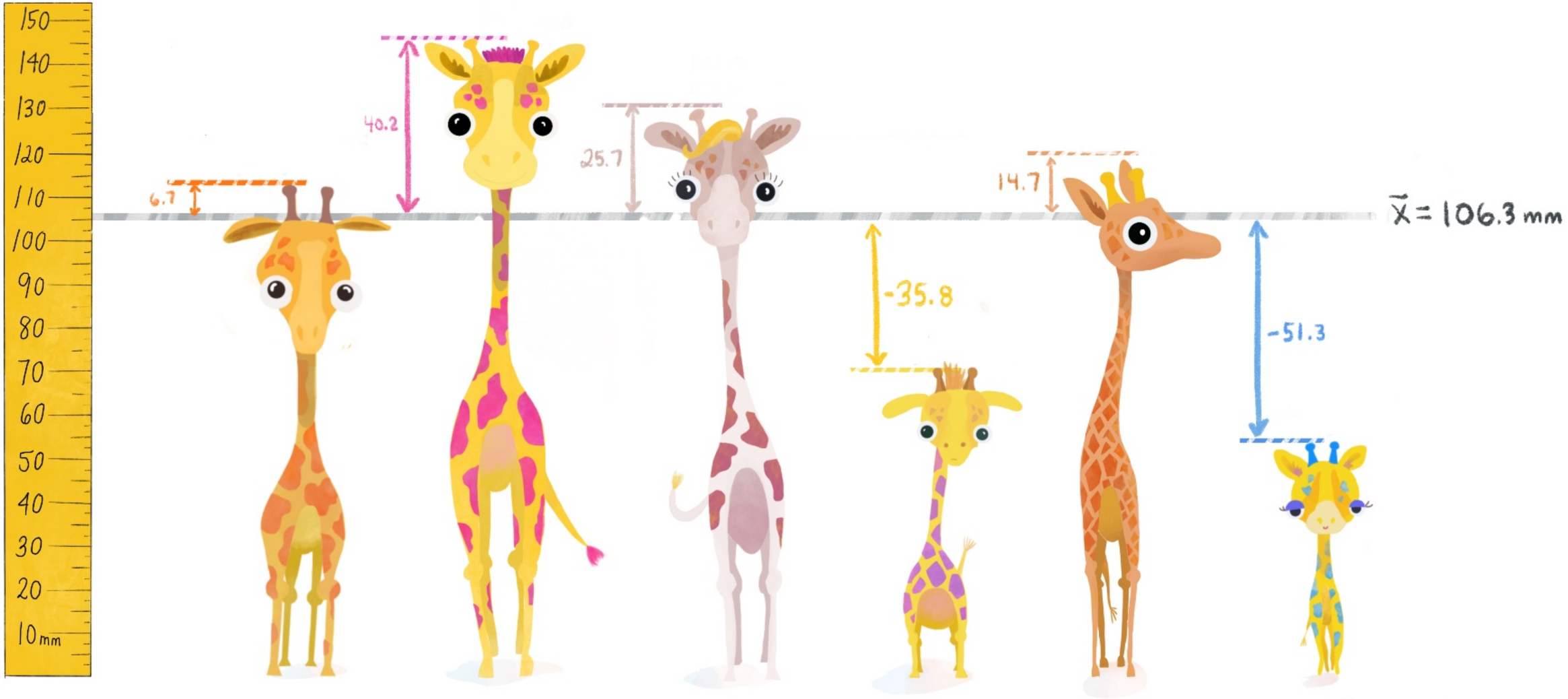
BINARY

ONLY 2 MUTUALLY EXCLUSIVE OUTCOMES

@allison_horst

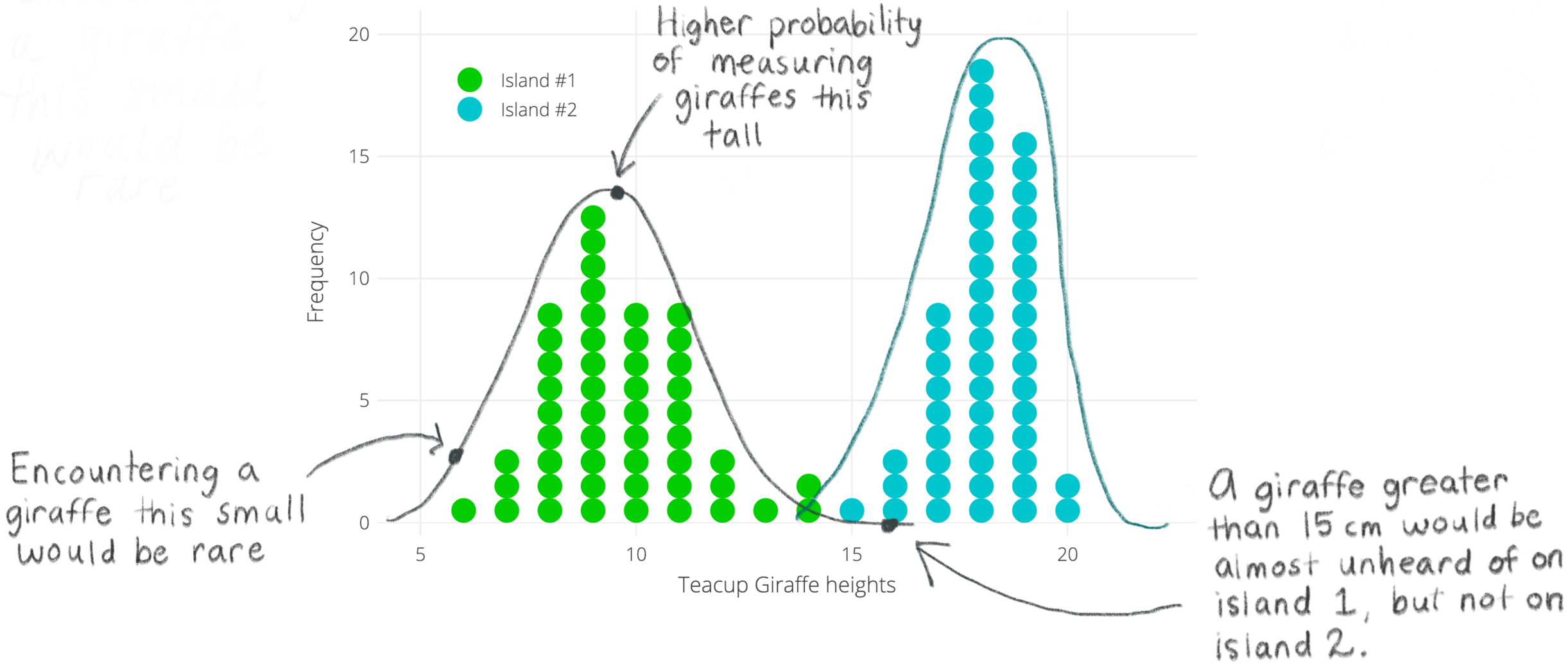
What types of data are used in the tests in your paper?

IMAGINE WE'VE VISITED TWO ISLANDS OF TINY GIRAFFES



teacups-giraffes-and-statistics

teacups-giraffes-and-statistics



the normal distribution

teacups-giraffes-and-statistics

Sample vs population

A **sample**, e.g. giraffes from each island (or different coloured birds), is a subset of a **population**.

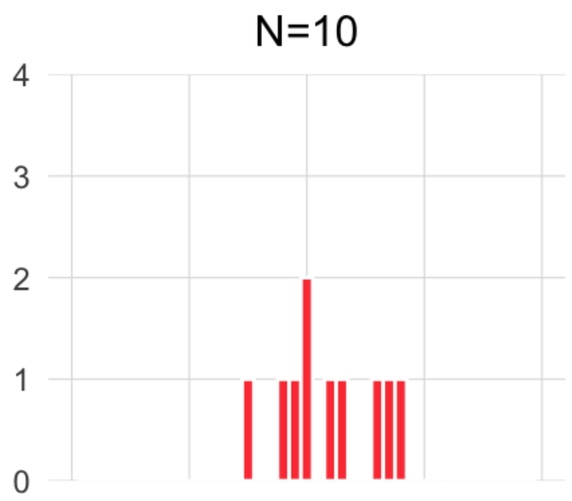
The population is defined as all available observations. In in this case, all existing giraffes on one of the islands or all existing coloured birds.

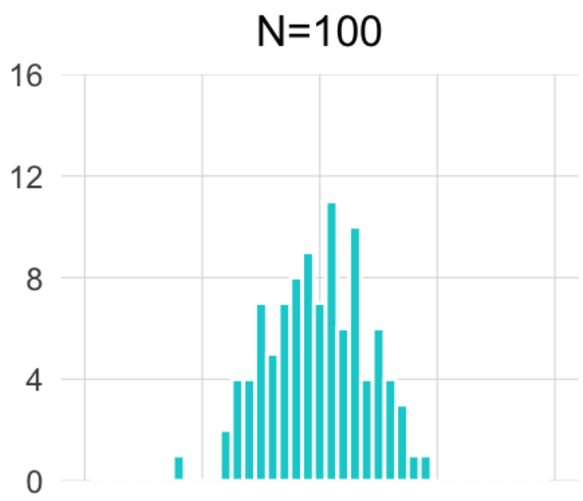
Different symbols are used to represent the mean for each of these.

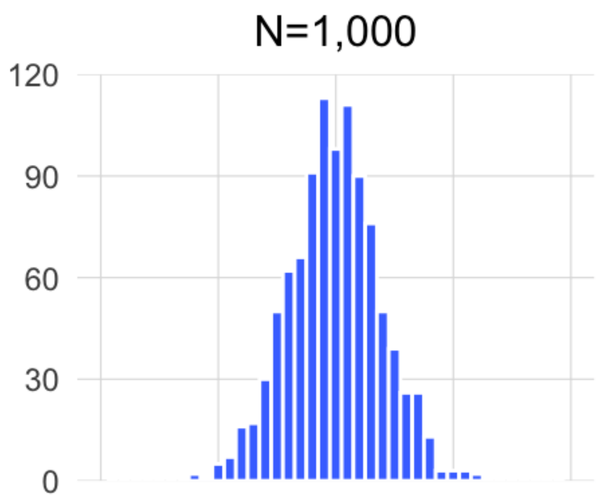
Sample vs. population

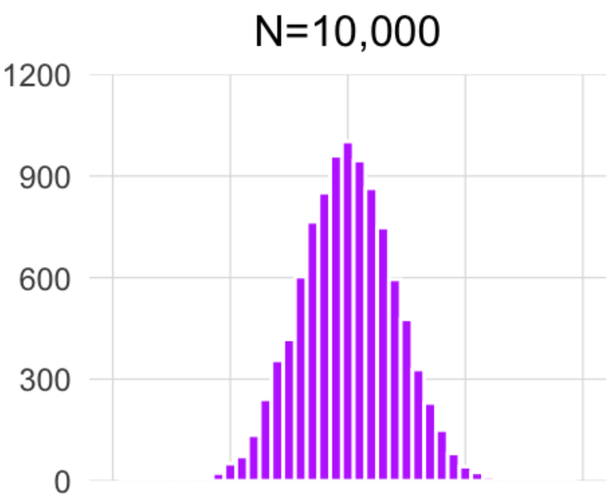
$\bar{X}$	Sample mean	Yes calculated from the raw data
μ	True population mean	Almost never known for sure
$\hat{\mu}$	Estimate of the population mean	Yes identical to the sample mean

Sample Size









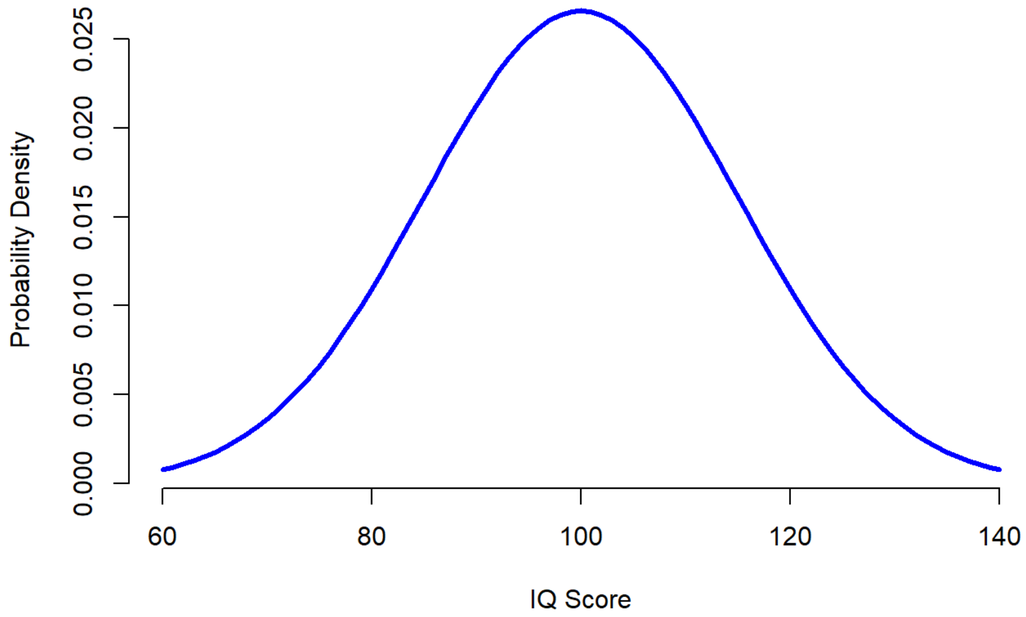
the larger our sample is, the more of the population it will include

Sampling

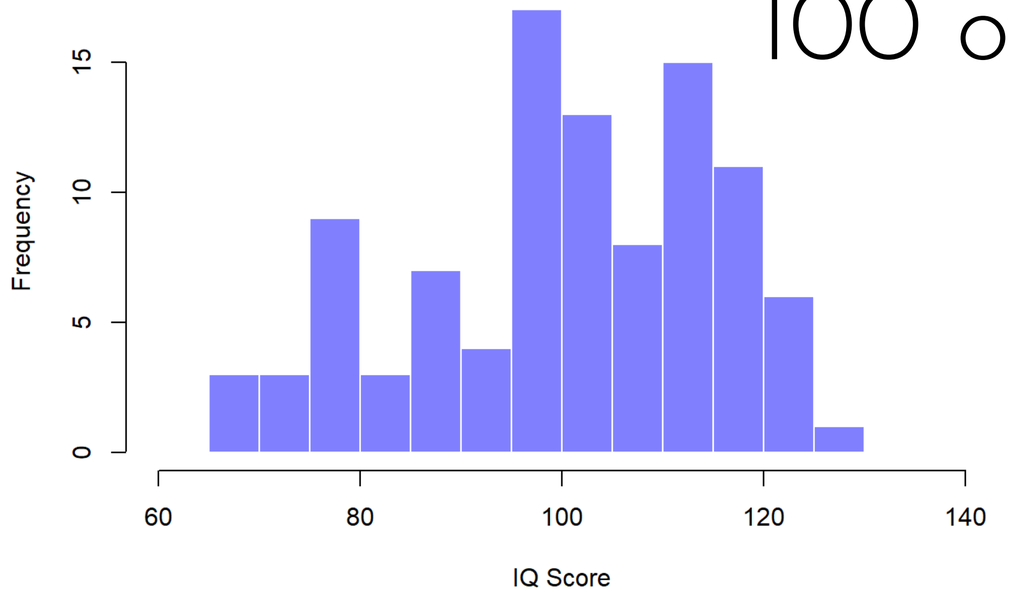
A procedure in which every member of the population has the same chance of being selected is called a **simple random sample**.

In reality, most samples are not random. Many samples are biased.

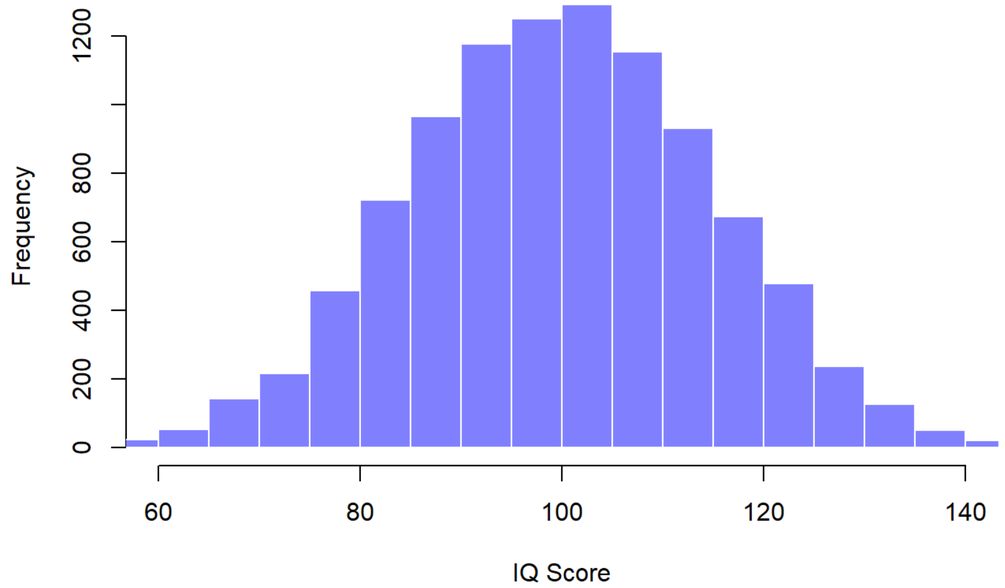
Population parameters



100 obs.



10000 obs.

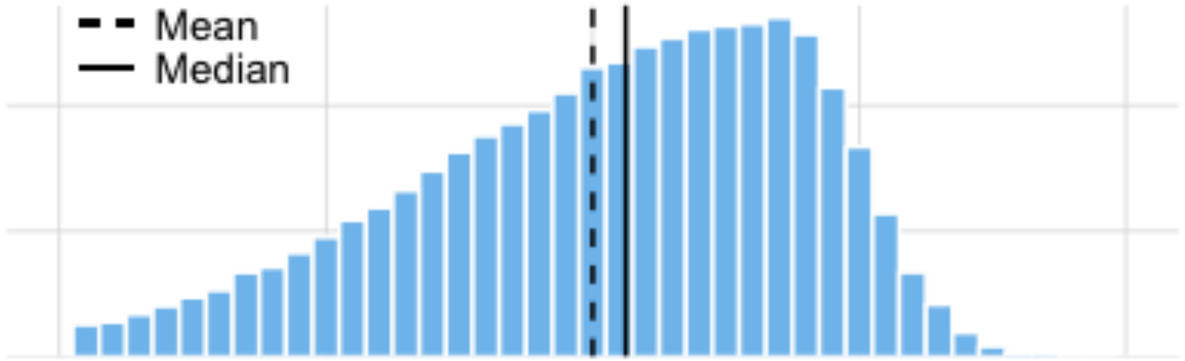


Learning statistics with R

Measures of centrality

Mean – the average and the measure of centrality.

Median – the value in the middle of the data set.

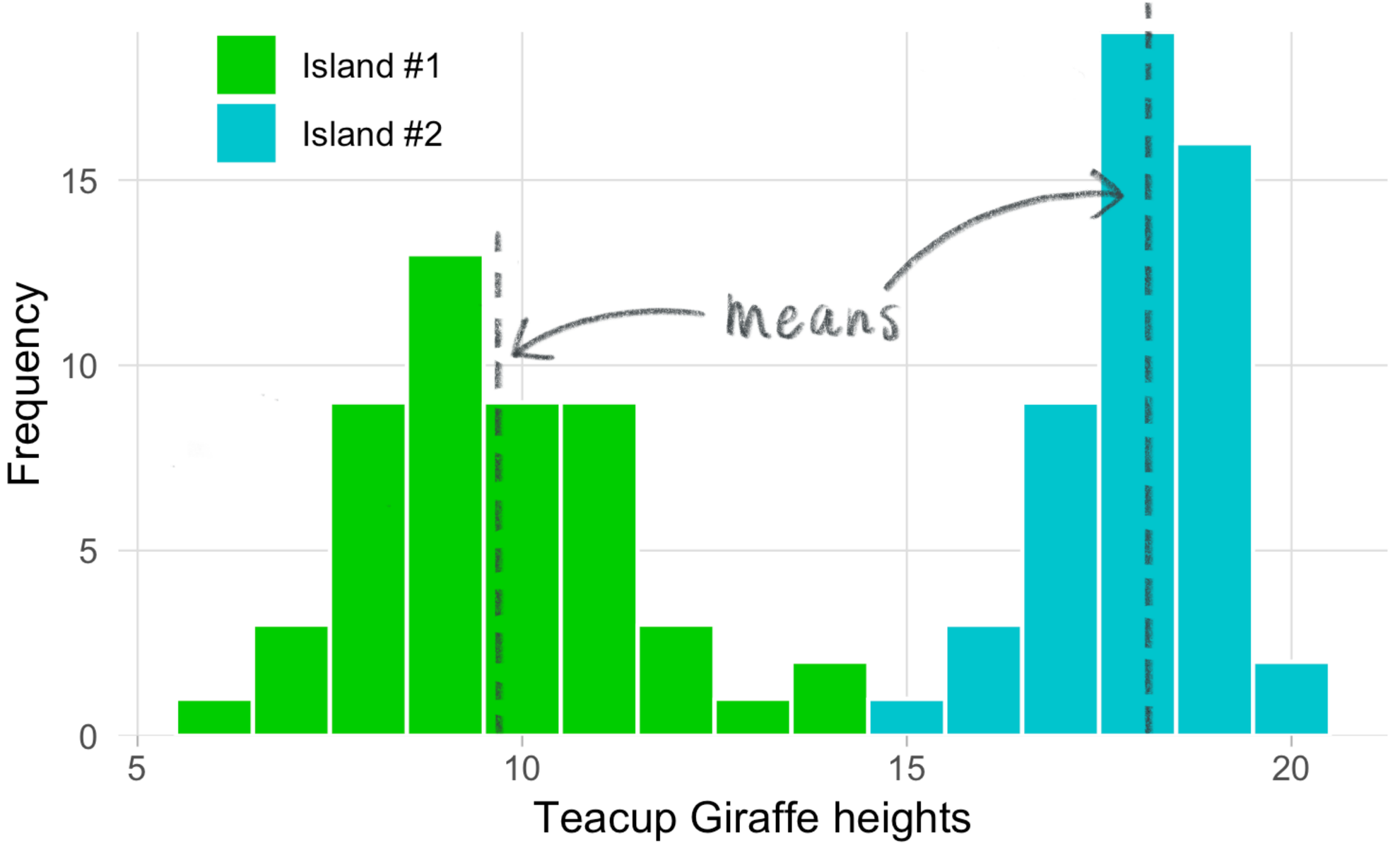


Measures of centrality

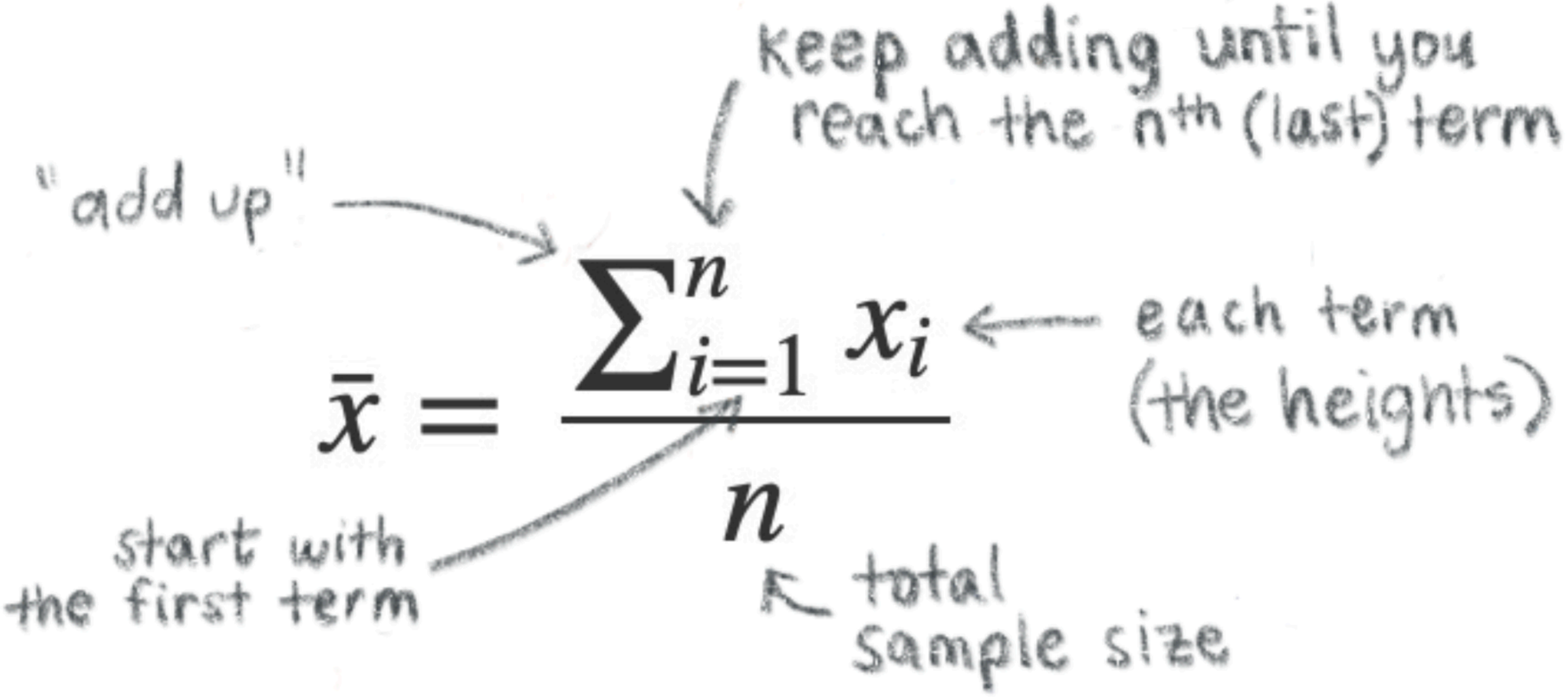
Mean – the average and the measure of centrality.

Median – the value in the middle of the data set.

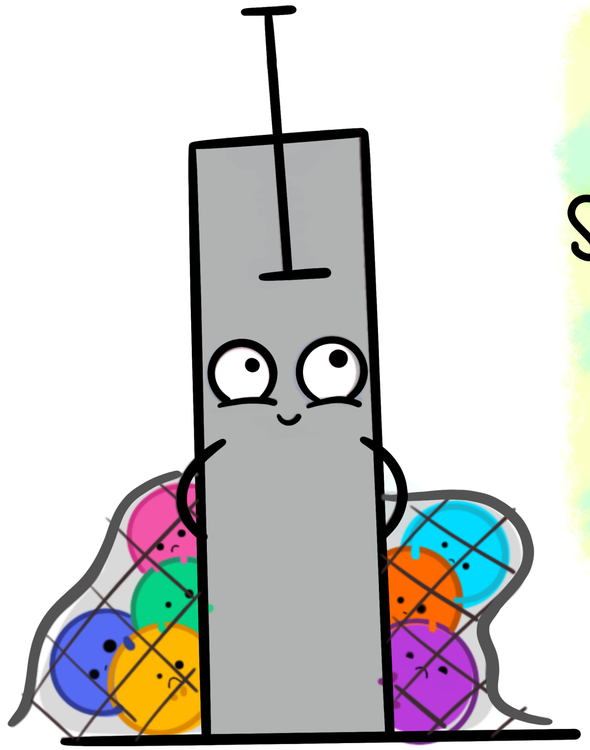
Mode – the value (e.g. height of giraffes, bird colour) that occurs most frequently.



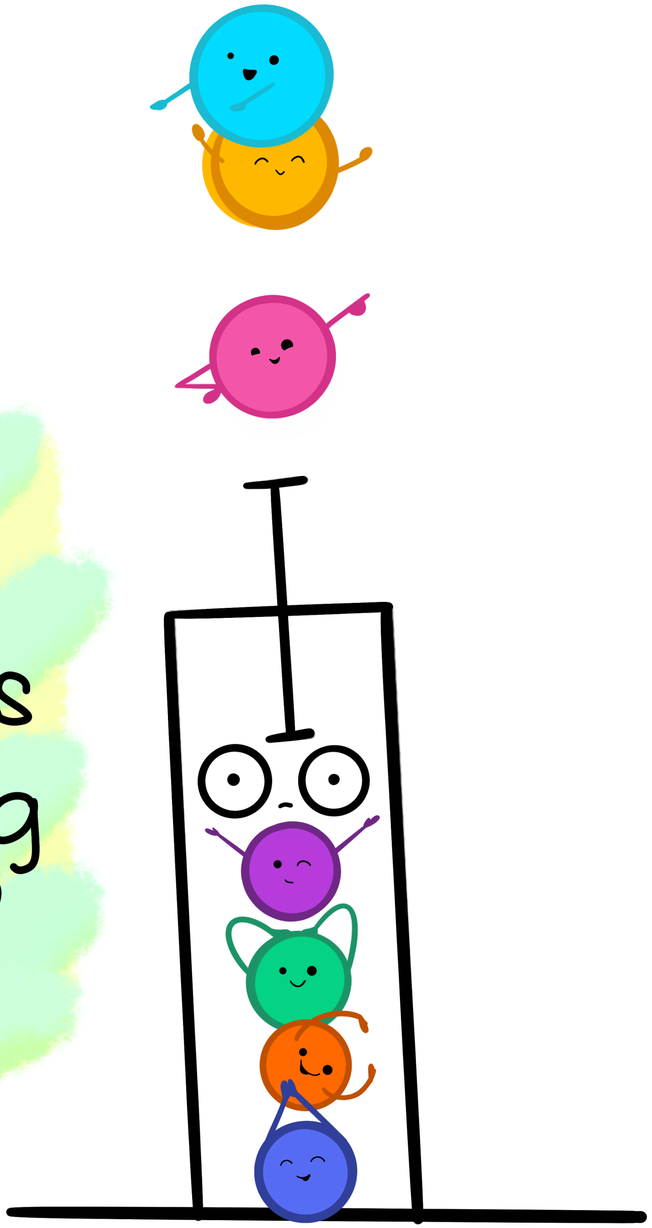
teacups-giraffes-and-statistics



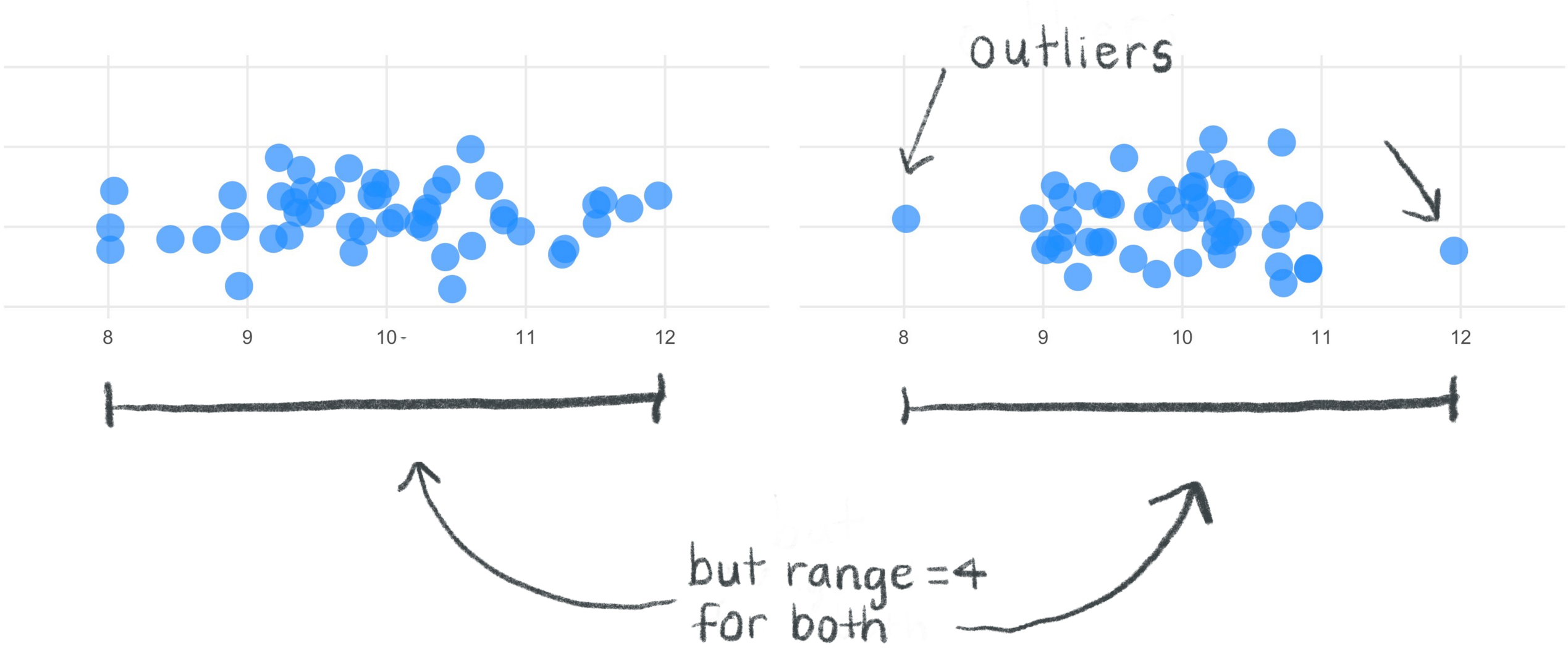
teacups-giraffes-and-statistics



are your summary statistics hiding something interesting?

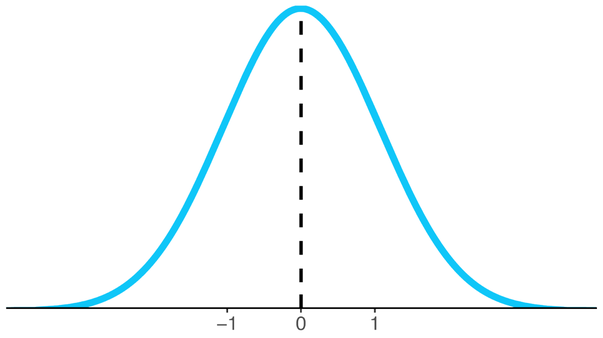


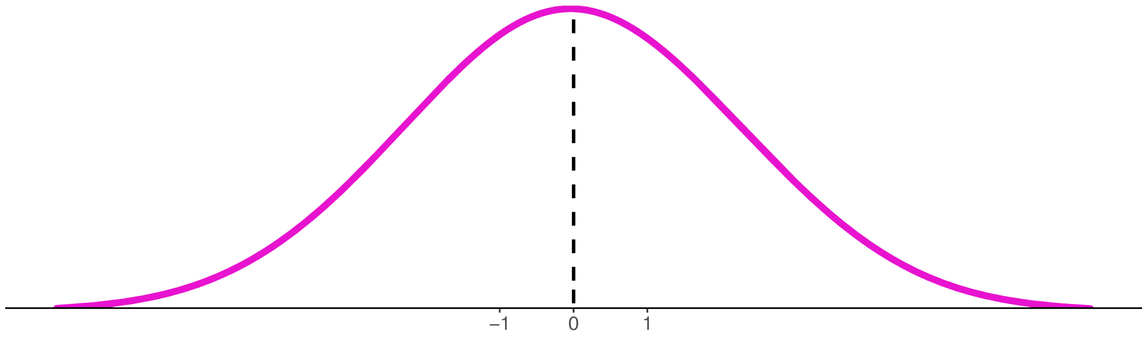
Spread of the data - the range

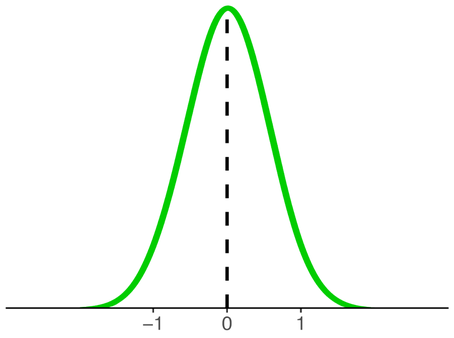


teacups-giraffes-and-statistics

Spread of the data - Variance







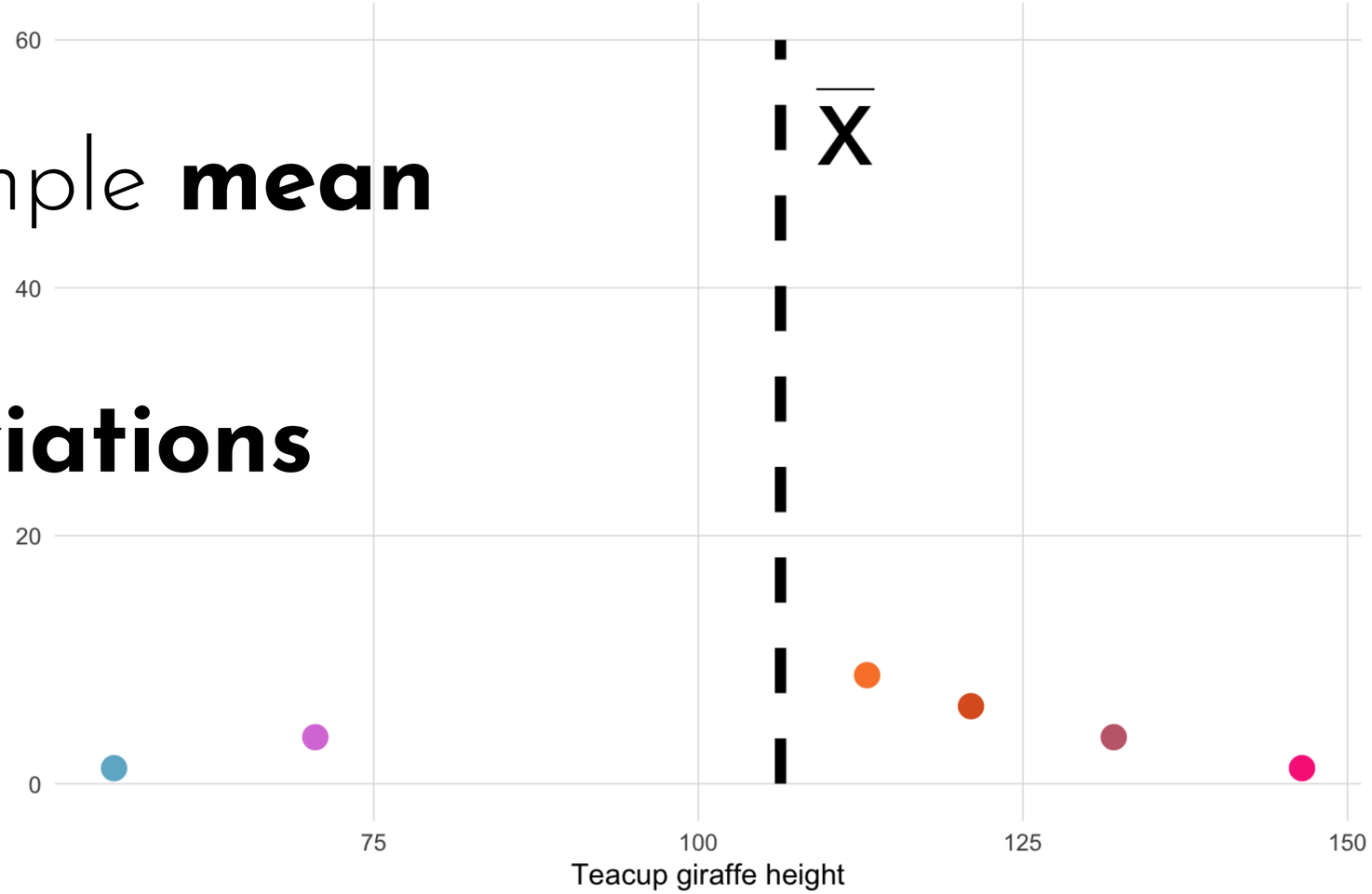
How do we calculate this?

teacups-giraffes-and-statistics

CALCULATING VARIANCE

1. Calculate the sample **mean**

2. **Square the deviations** from the mean



teacups-giraffes-and-statistics

CALCULATING VARIANCE

3. Calculate the sum of squares

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

CALCULATING VARIANCE

4. Calculate the average of the squared differences from the mean.

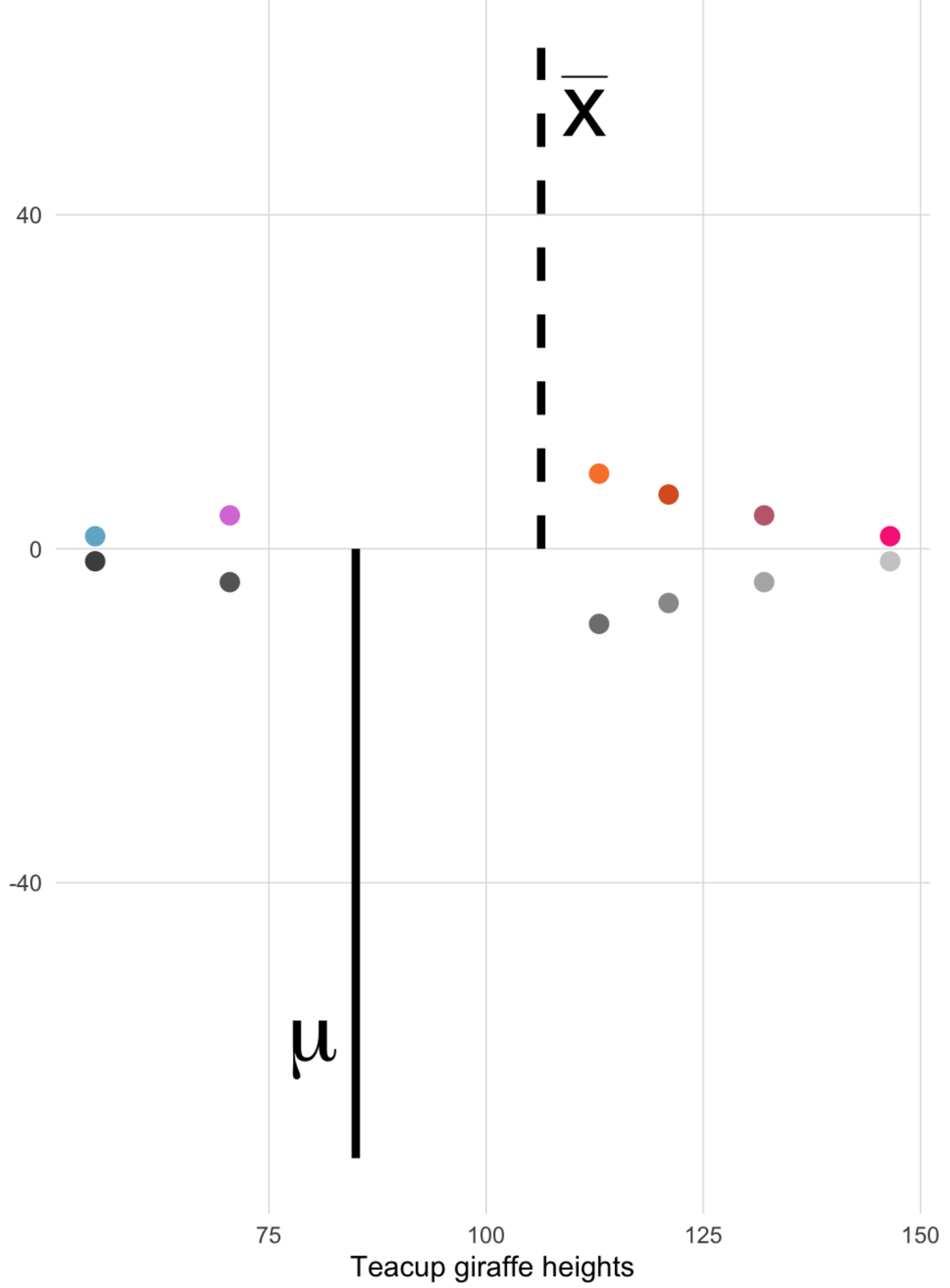
$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Standard deviation

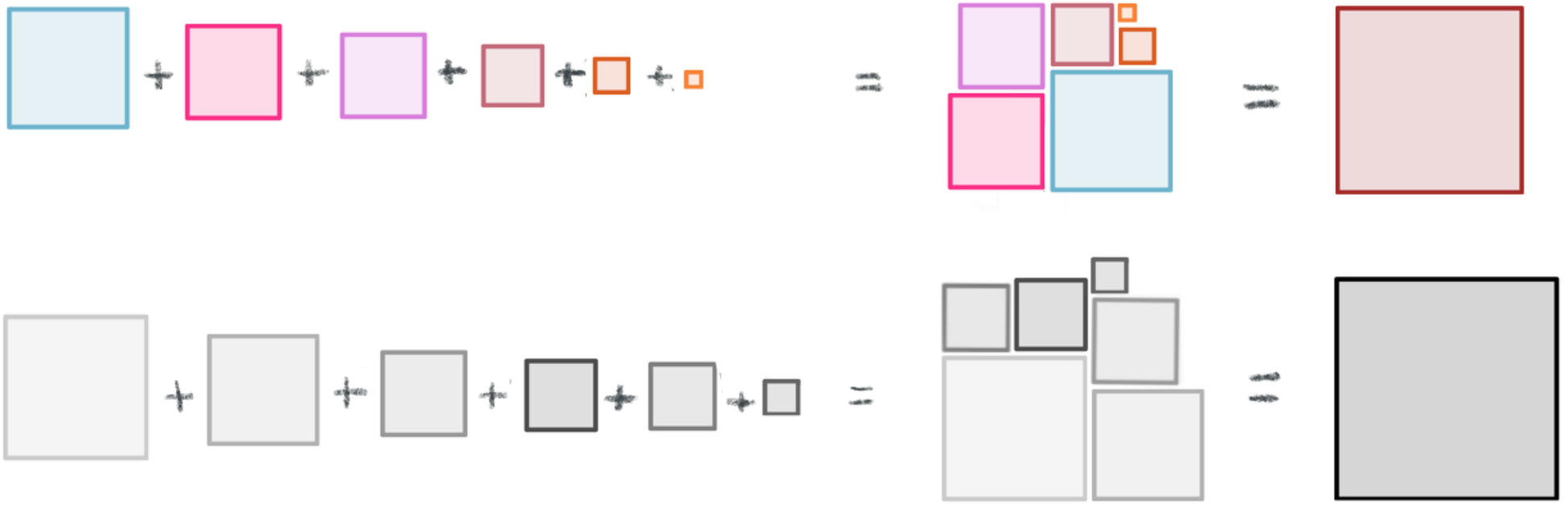
Variance is not easily interpretable → "unsquare" the variance to return to the data's original units.

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

Population versus sample



Population versus sample



teacups-giraffes-and-statistics

Solution : N - 1

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

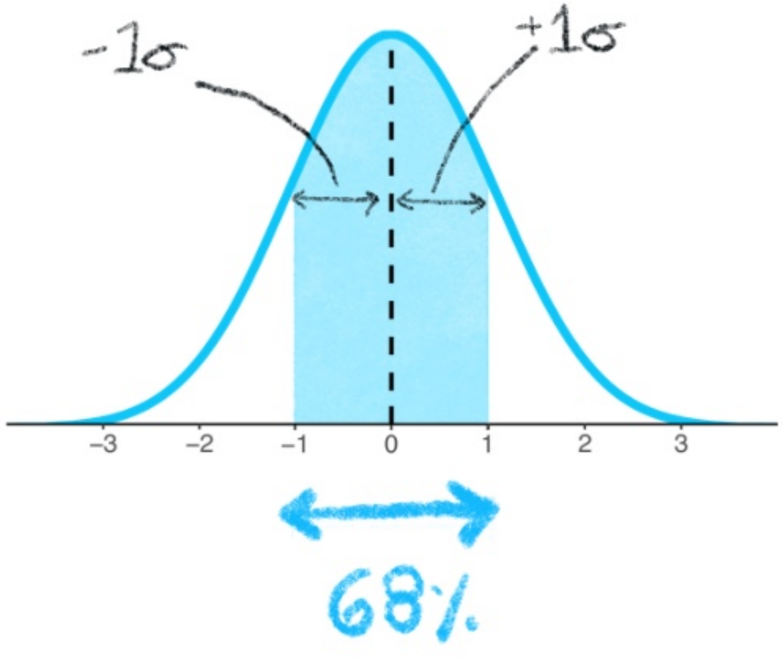
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \qquad s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

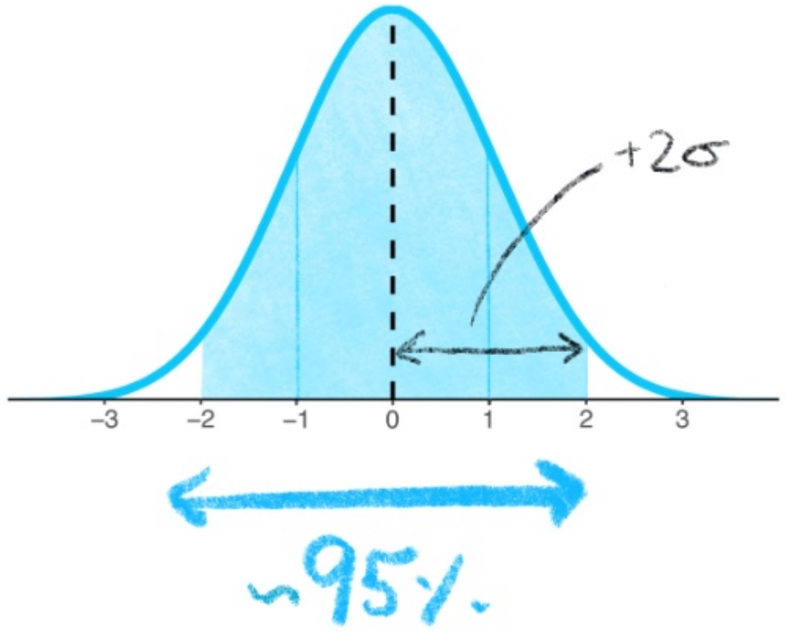
Also note the change in symbols

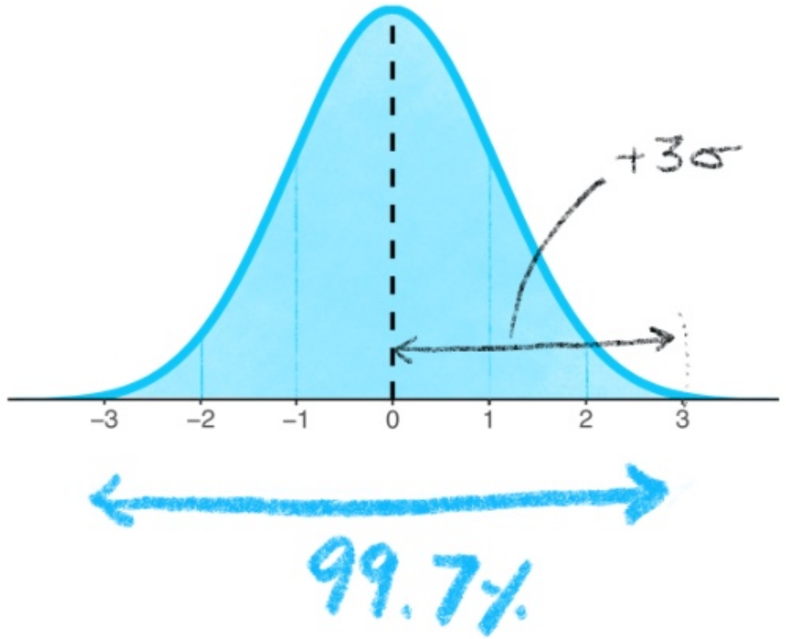
Steps summary

1. Calculate the sample **mean**
2. Square the **deviations from the mean**
3. Calculate the **sum of squares**
4. Calculate **average squared differences**
5. Apply the N-1 correction
6. Compare your results

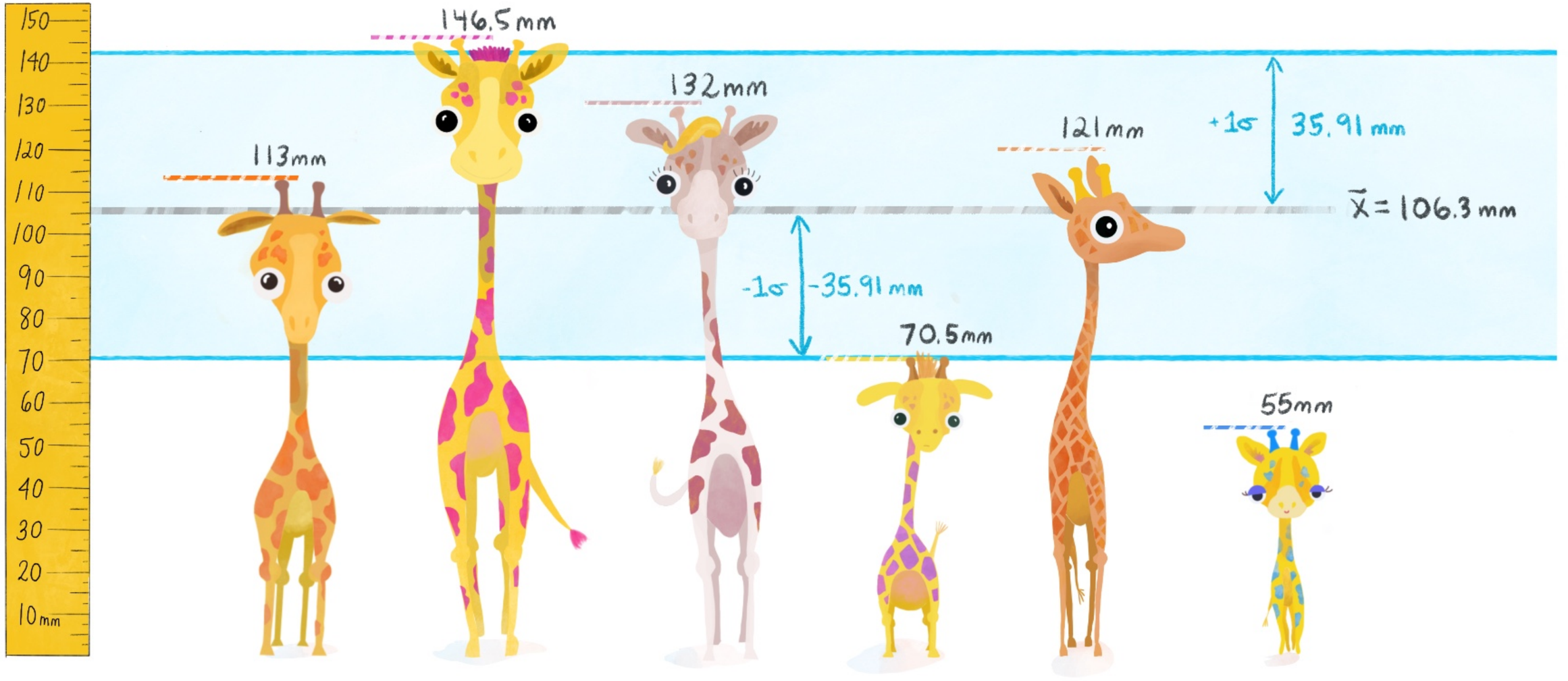
Interpreting the standard deviation





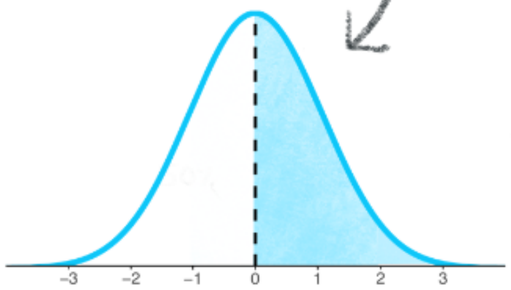


teacups-giraffes-and-statistics



teacups-giraffes-and-statistics

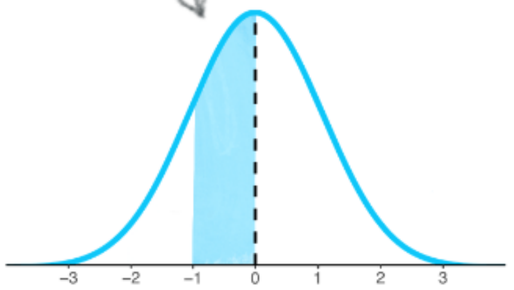
All giraffes above the mean



50%

+

Giraffes less than the mean up to -1σ

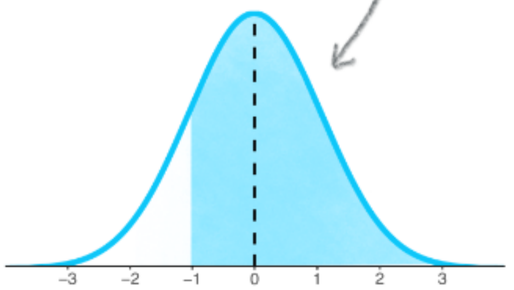


34%

Half the 68% that we memorized from the empirical rule

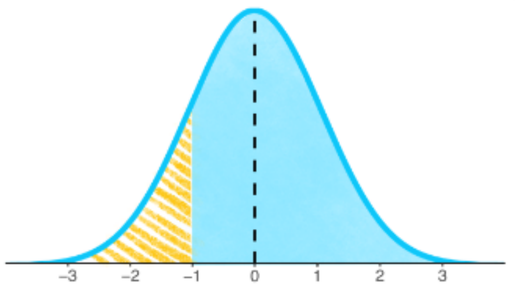
=

% of giraffes taller than our giraffe of interest



84%

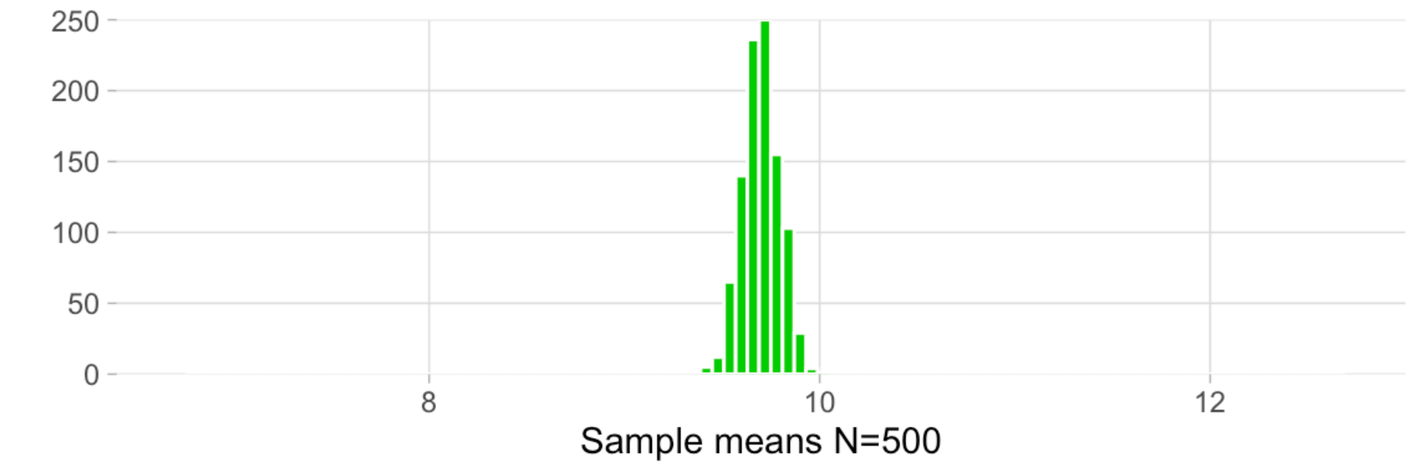
% of giraffes shorter than our giraffe of interest.

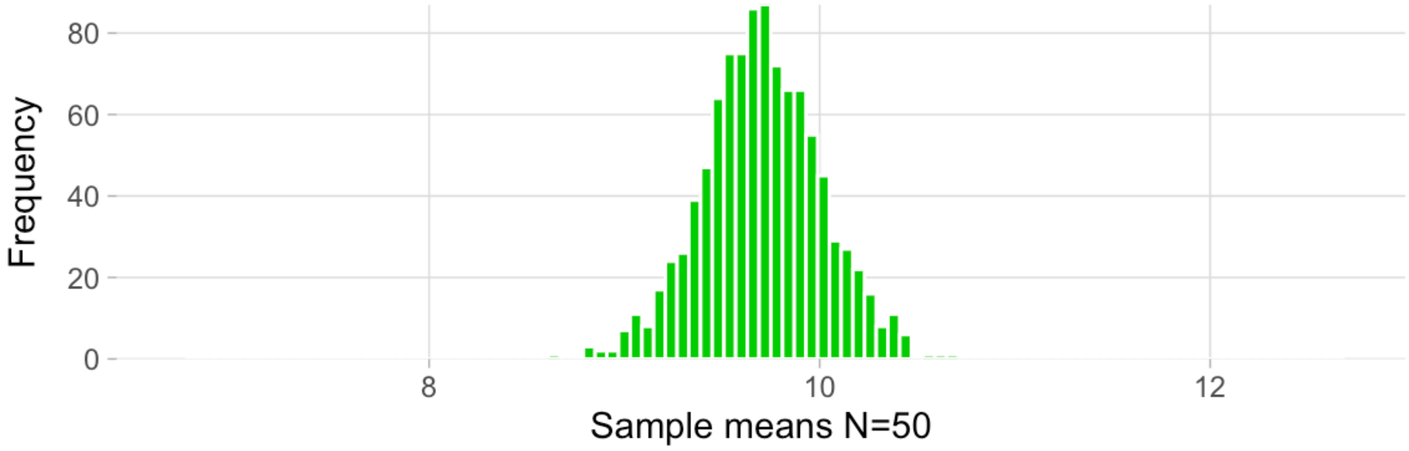


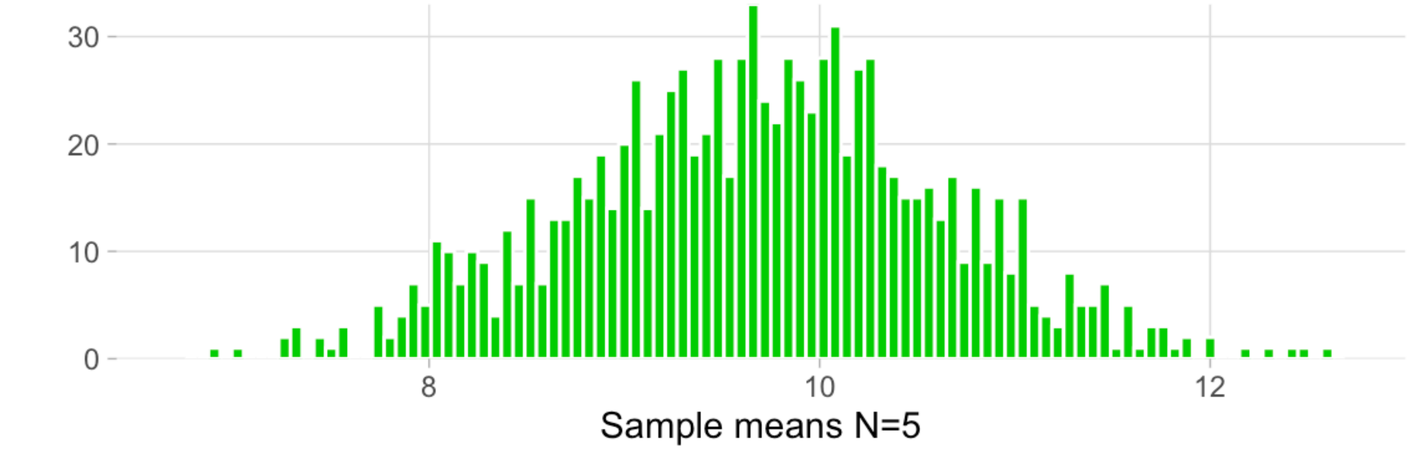
16% = 100% - 84%

Sample Size

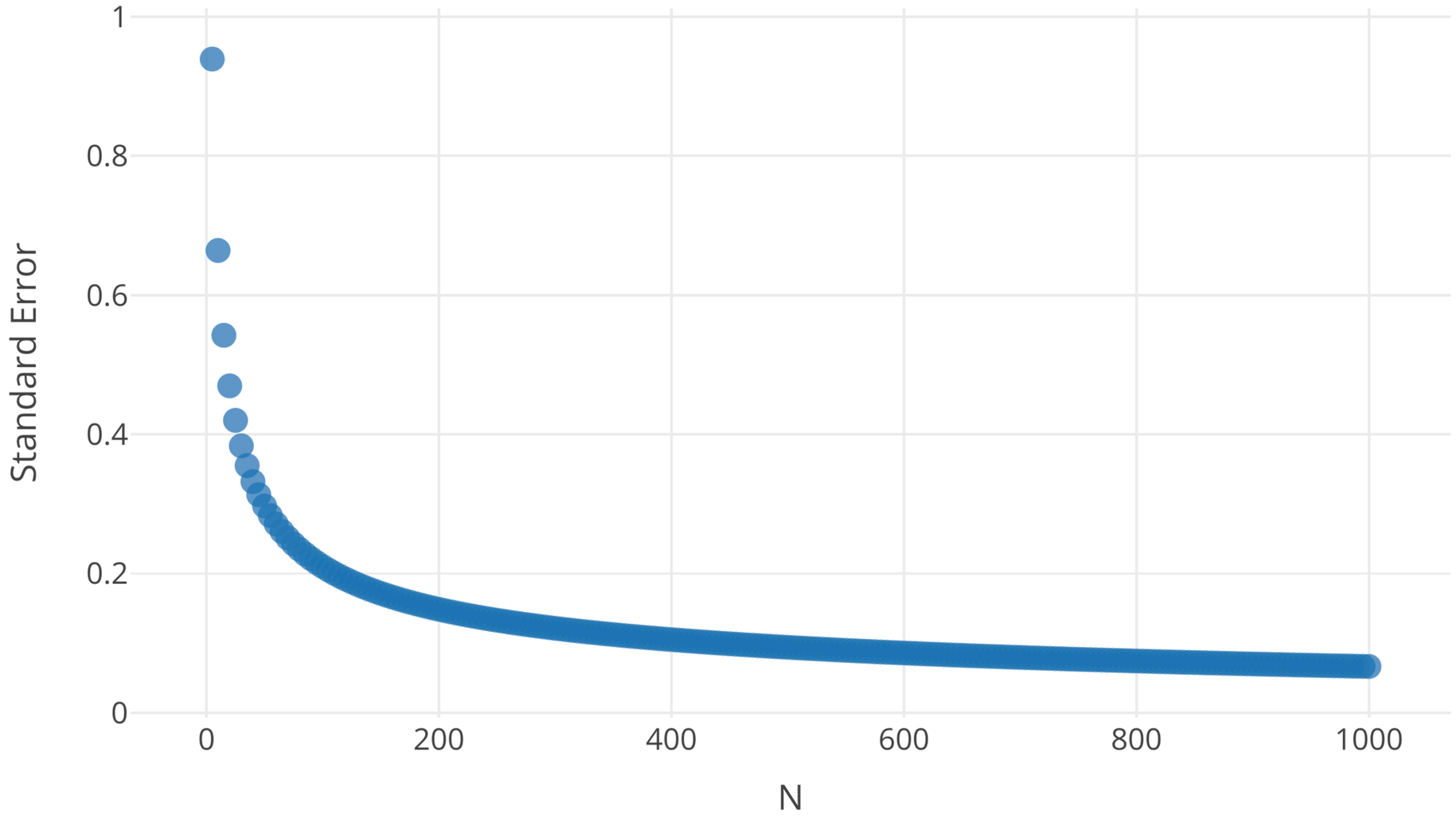
distributions of the mean based on random samples of different sizes







Standard error



Standard error

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}}$$

- a measure of the statistical accuracy of an estimate
- equal to the standard deviation of the theoretical distribution of a large population of equivalent estimates

Further reading

tinystats.github.io/teacups-giraffes-and-statistics

QUESTIONS?

teacups-giraffes-and-statistics