Variance

Rachel 31.01.23

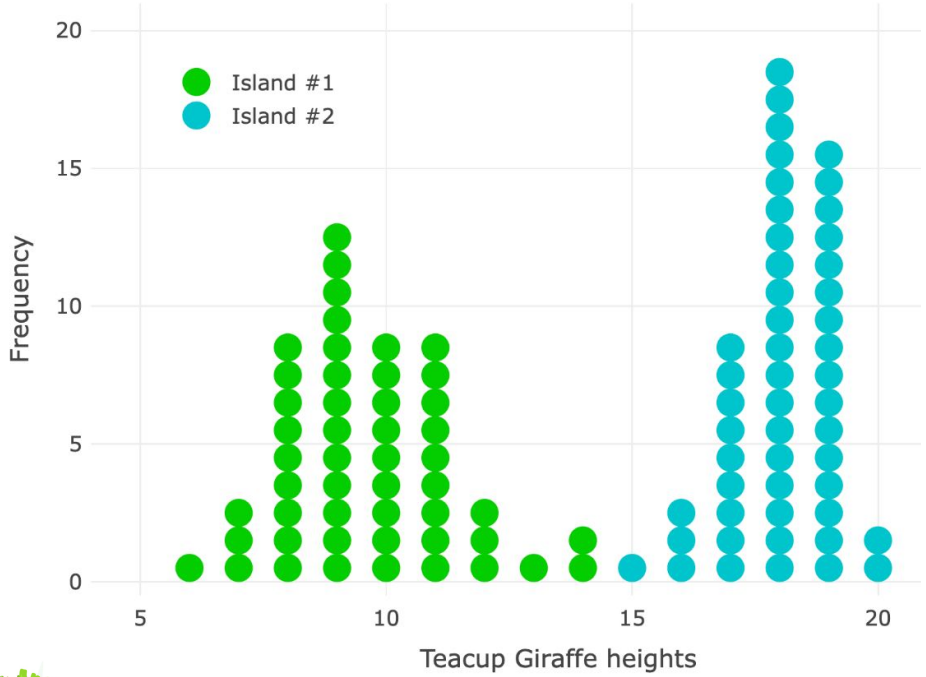
Learn more about the tiny giraffes @ tinystats.github.io

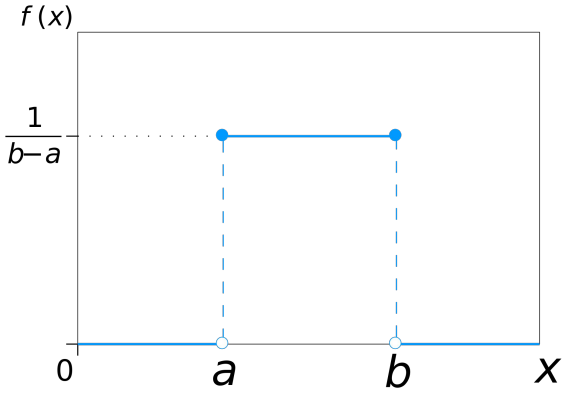
Recap

A distribution

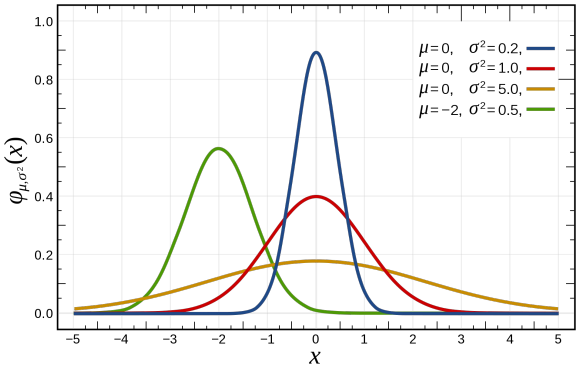
- Shows the values the variable (e.g., height) takes on in your data set
- How often each value occurs
- The **shape**, center, and **amount of variability** in the data

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

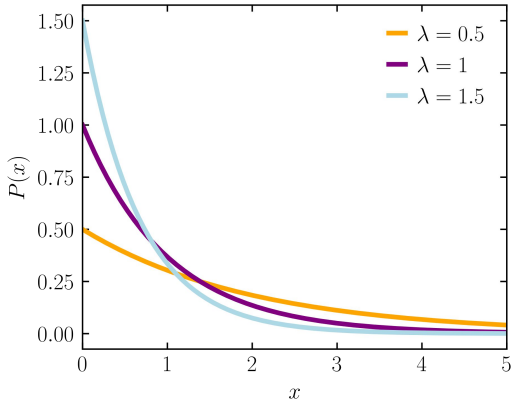




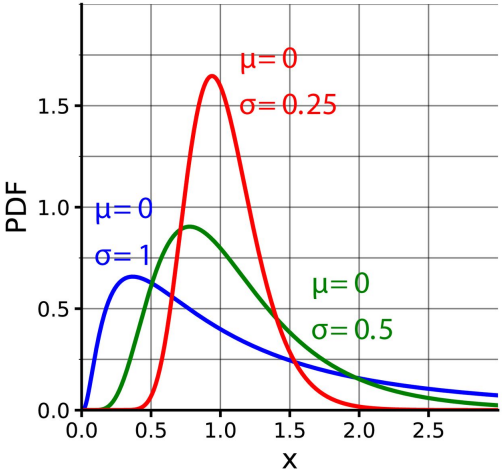
Uniform



Normal



Exponential

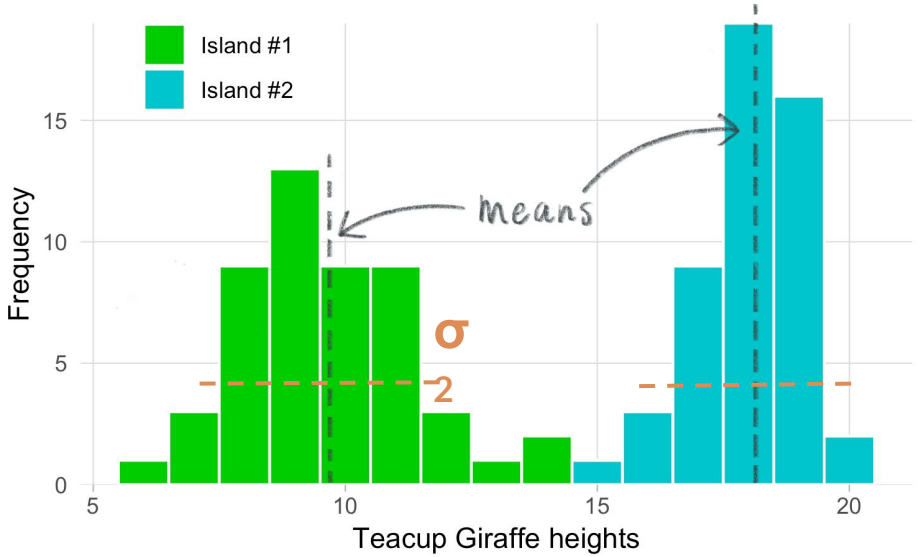


Lognormal

Parameters of the normal distribution

μ - the mean or expectation

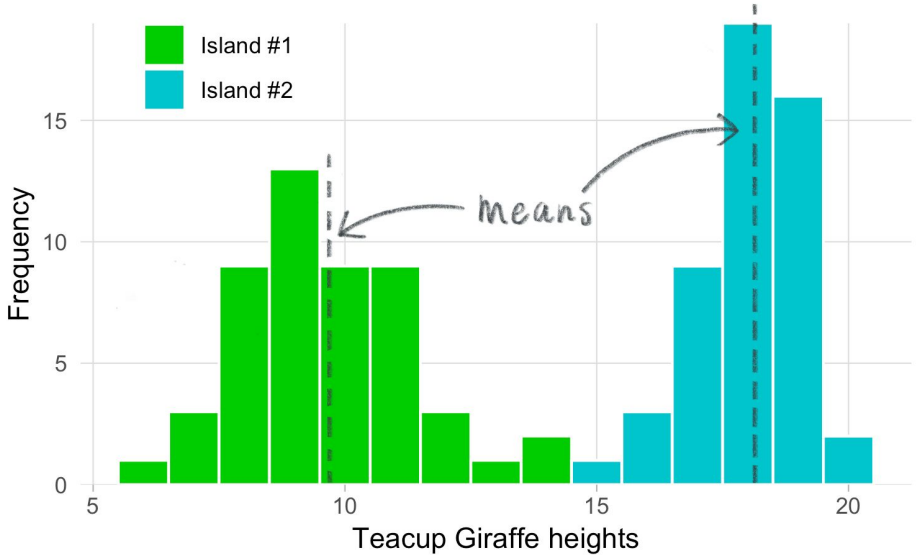
σ - the standard deviation
(or σ^2 - the variance)



Homework

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

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

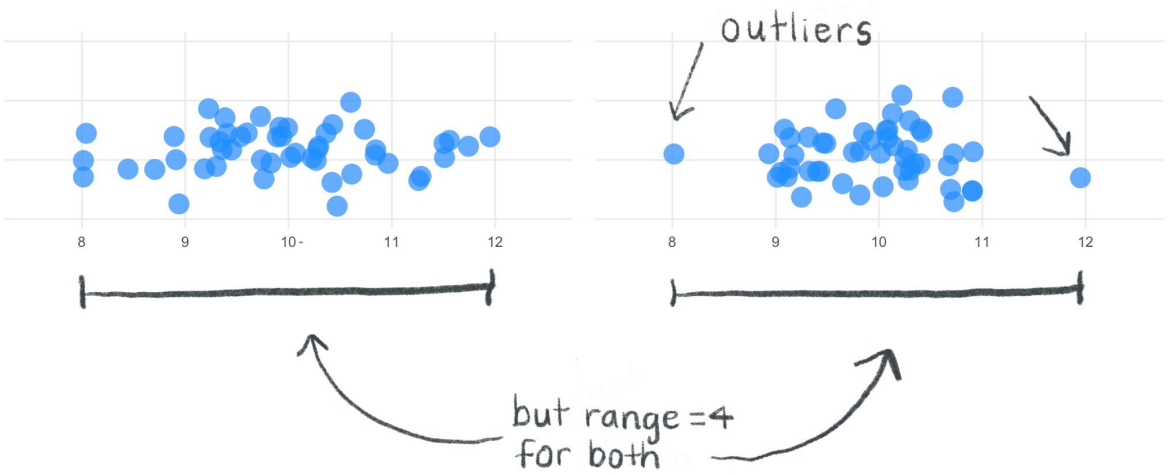
`rnorm(n, mean = 0, sd = 1)` - generates random numbers

Let's see these in action!

Teacup giraffes

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

Spread of the data - range

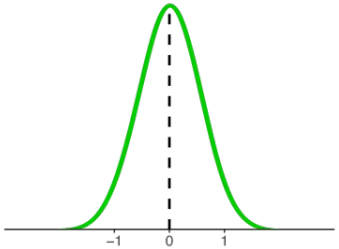


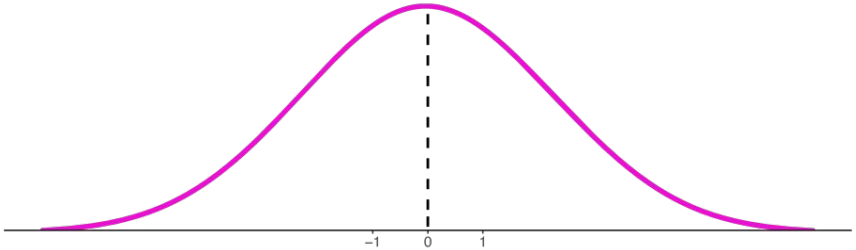
If we want to avoid undue influence of the outliers, the range is not good measure.

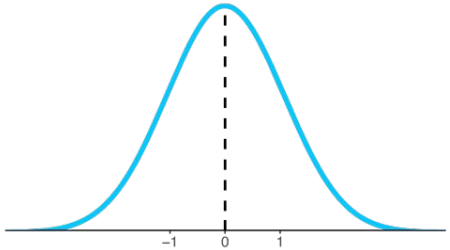
Spread of the data - variance and standard deviation

The variance and s.d. can account for outliers.

We need a good understanding of s.d. to grasp the mechanics of commonplace statistical tests.



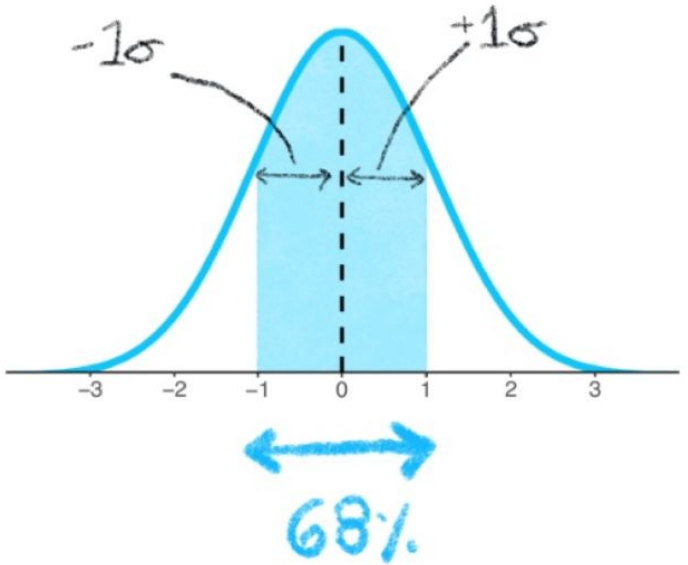




Standard deviation

A measure of the amount of variation or dispersion of a set of values.

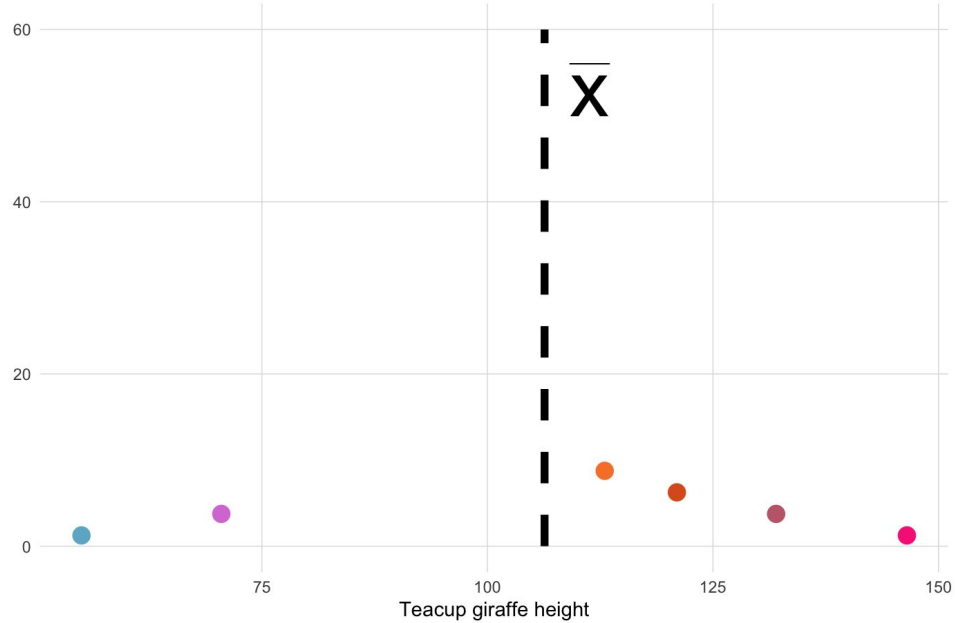
How do we calculate this?



Calculating the variance and standard deviation

1. Calculate the **sample mean (*x*)**

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



Calculating the variance

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

3. Calculate **the sum of squares**

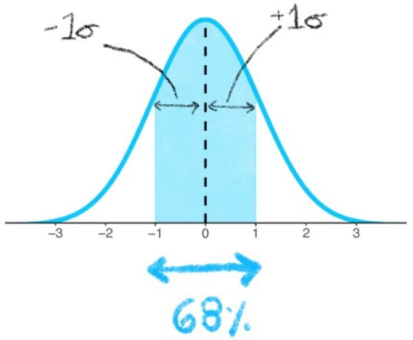
4. Calculate the average **squared differences from the mean** (i.e., the average of 3)

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

Calculating the standard deviation

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

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

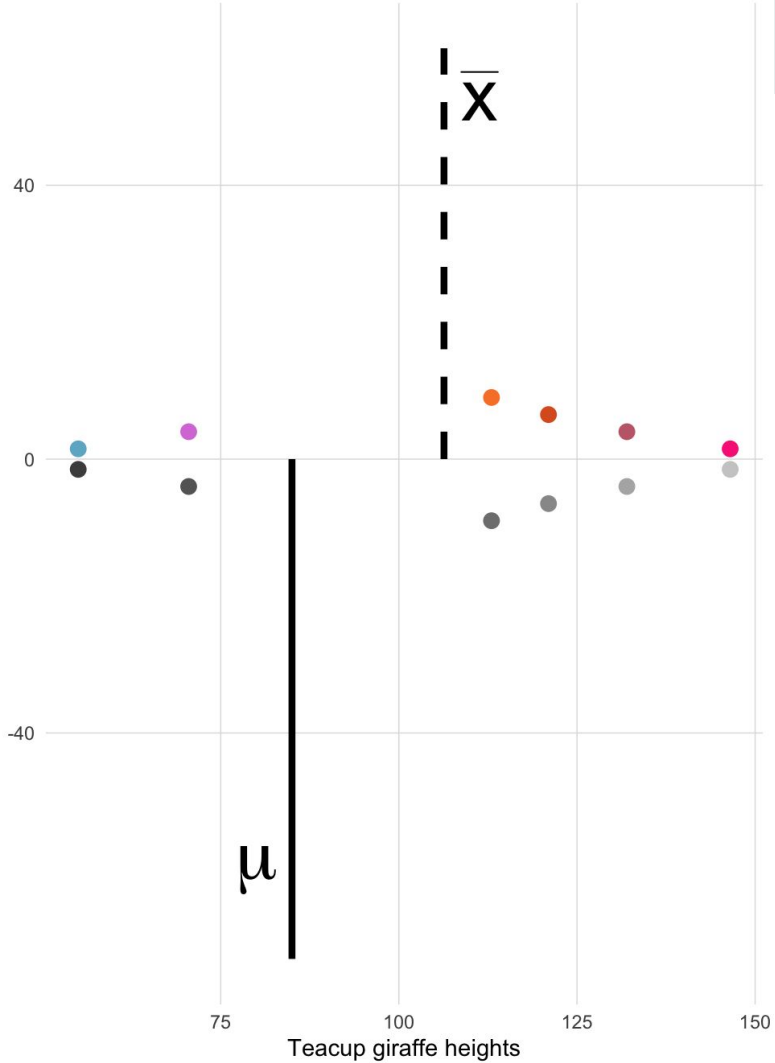


Population vs. sample

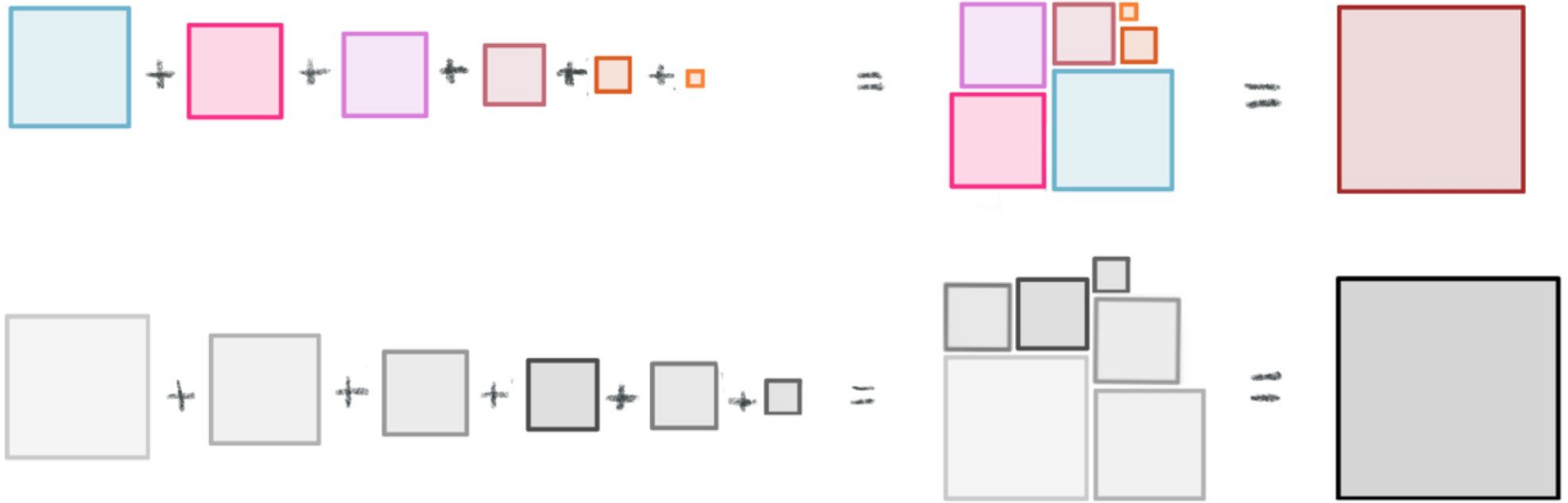
We only have the sample mean as our center point.

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.



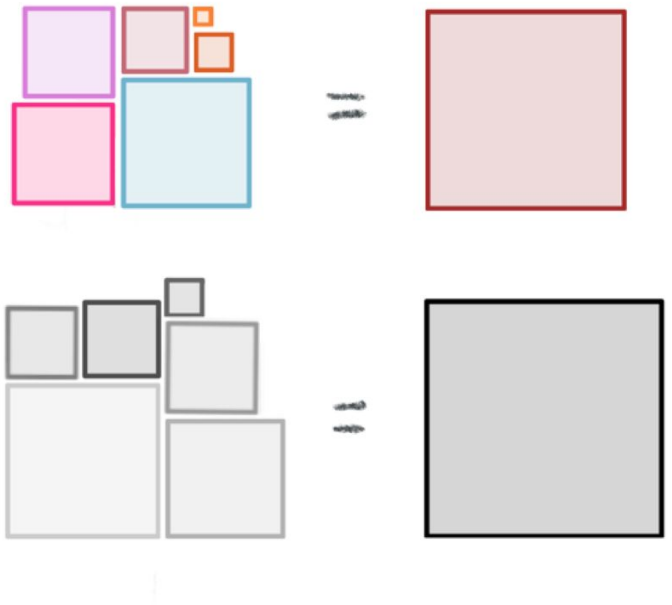
Population vs. sample



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* minimizes the total distance of all the observations to the center.



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.

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

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

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

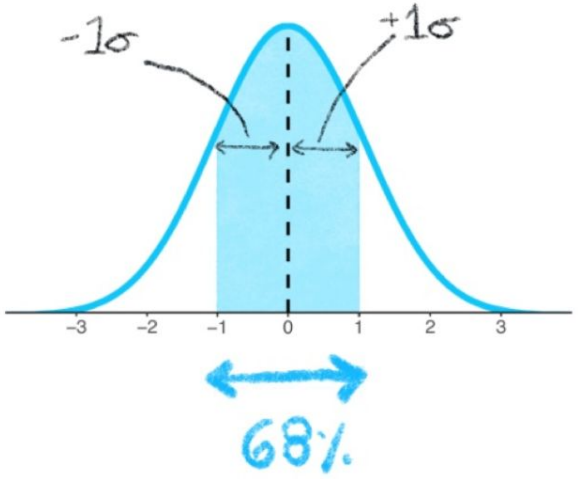
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}$$

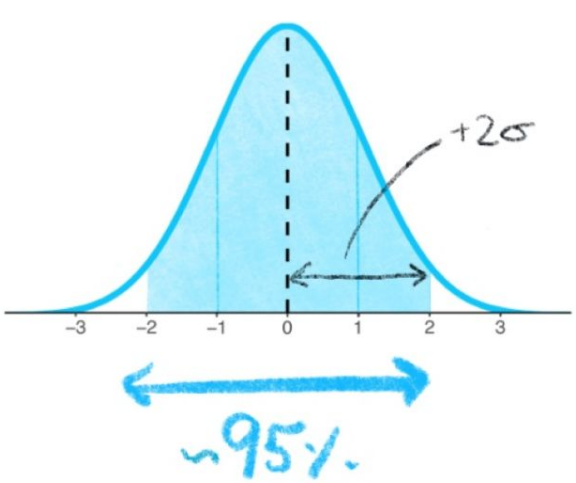
$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}}$$

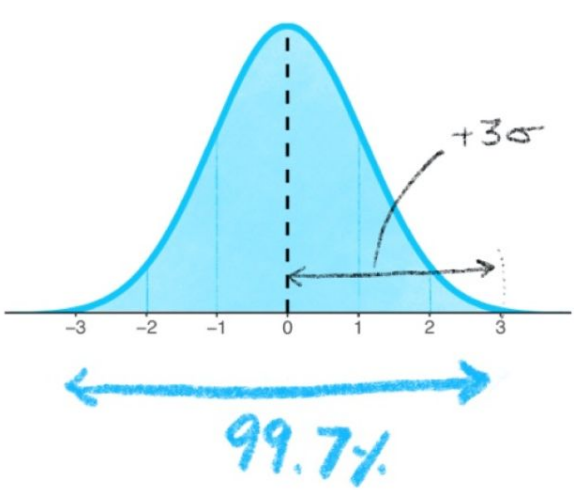
Summary of steps used to calculate the standard deviation

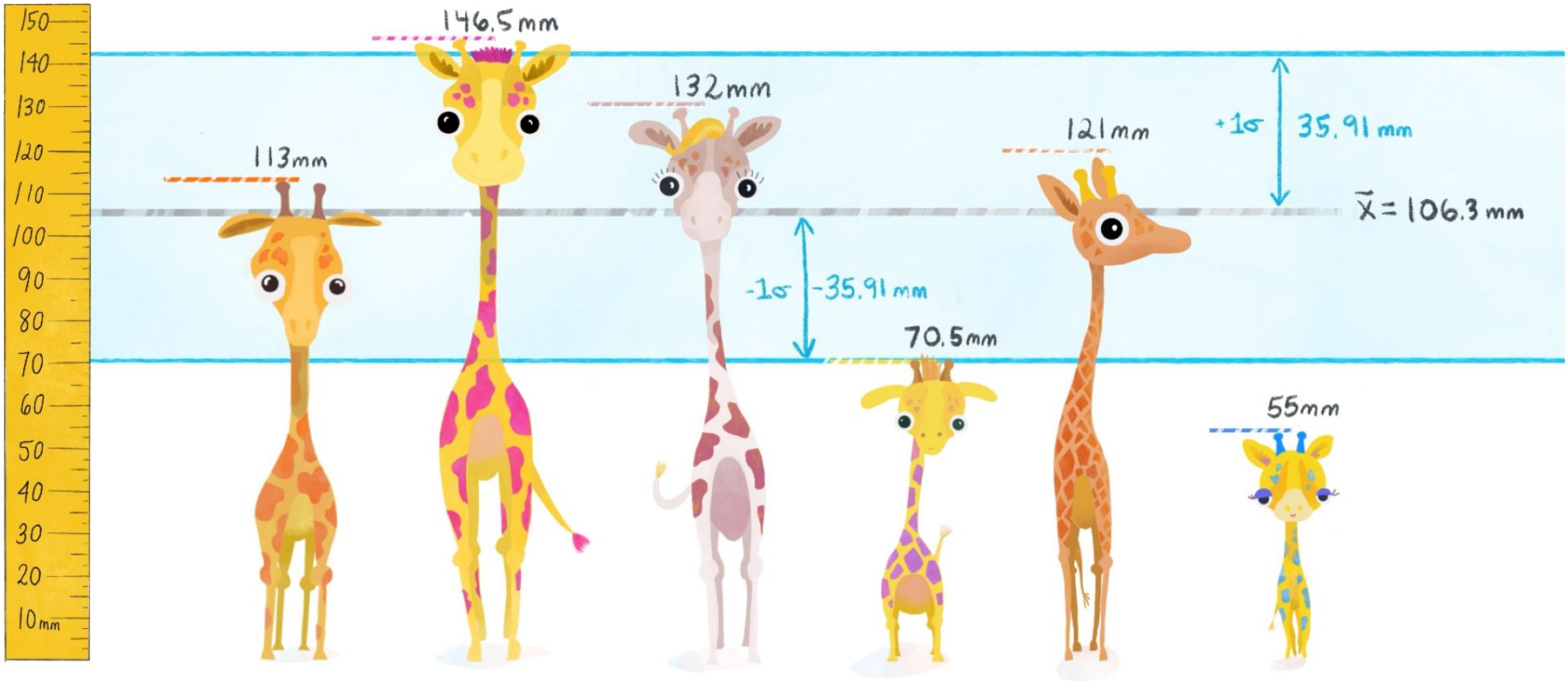
1. Calculate the **sample mean**
2. Square the **deviations from the mean**
3. Calculate the **sum of squares**
4. Calculate **average squared differences**
5. Apply the *n*-1 correction

Interpreting the standard deviation



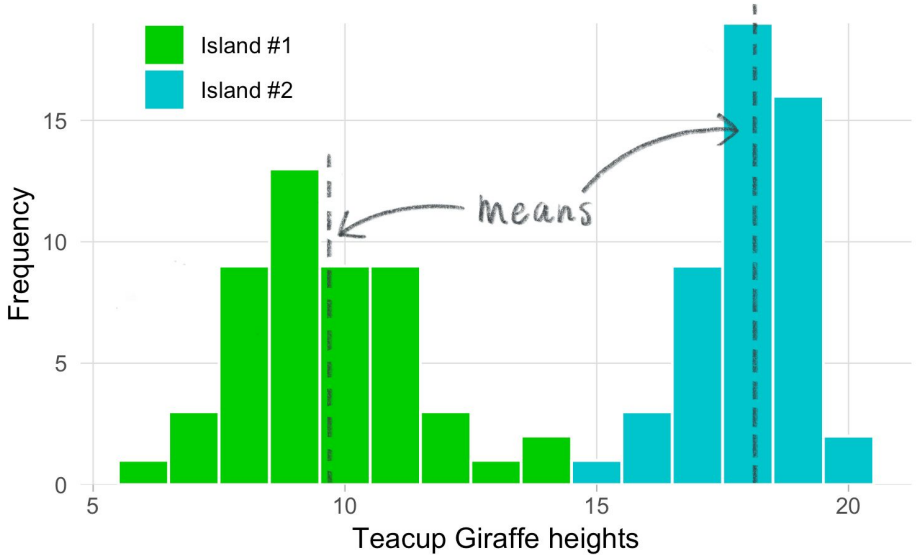





150
140
130
120
110
100
90
80
70
60
50
40
30
20
10 mm
146.5 mm
132 mm
121 mm
113 mm
70.5 mm
55 mm
+1σ
35.91 mm
$\bar{X}$ = 106.3 mm
-1σ
-35.91 mm

Homework

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



Always plot your data!


The only thing that spreads better than data is marmalade!