CORN276-RMGIS

Michael Hunt

2026-02-06

Table of contents

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

1 Visualising data

When we have got our data safely tucked into a spreadsheet. Now we need to tease out of it the answers to our question(s) and to decide whether we have evidence enough to reject our null hypotheses, or not, in which case we will fail to reject them.

Let's take the example of the Palmer penguins dataset. This set contains measurements of bill depth, bill length, flipper length and body mass of males and females of three species of penguins: Adelie, Chinstrap and Gentoo observed on any one of three islands in the Palmer Archipelago, Antarctica.

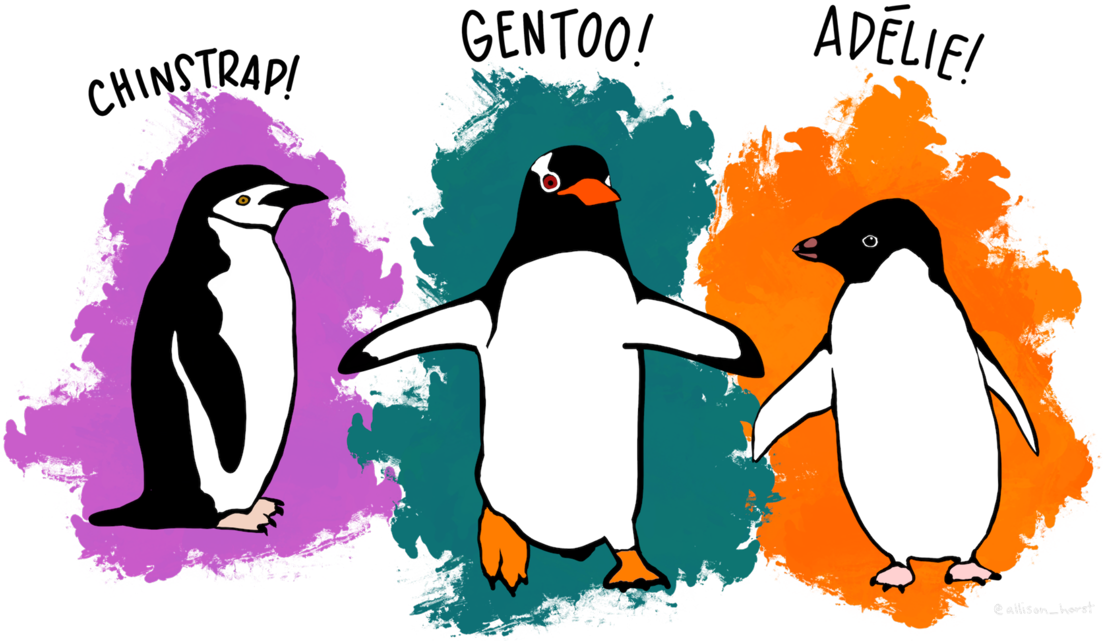

Figure 1.1: Meet the penguins

Let's consider only the females and ask the question:

Question: Is there any difference in body weight between the three species:

from which we can generate a hypothesis:

Hypothesis: There is a difference in body weight between females ofthe three species.

Null hypothesis: There is no difference in bosy weight between females of the three species.

and hence a prediction of what we will find if the hypothesis is true:

Prediction if the hypothesis is true: At least one species will have a different average body mass than at least one other species.

1.1 Summarise the data

The first thing we can do to investigate our hypotheses is to summarise the data. More often than not this means caclulating three things for each sample - the sample sizes, the mean values and the standard errors of those means.

Species	N	Mean body mass (g)	Standard error (g)
Adelie	73	3368.84	31.53
Chinstrap	34	3527.21	48.93
Gentoo	58	4679.74	36.97

ℹ Types of error bar

standard deviations: These tell us about the spread of values in a sample or a population. They do not systematically get bigger or smaller as the sample size increases. The standard deviation of a sample can be used as an estimate of the standard deviation of the population. We use standard deviations for *descriptive purposes*
standard errors of the mean These are used to indicate how precisely a sample mean estimates the true population mean. They are used for *inferential purposes*, whereby we try to infer from the sample mean the range of values in which the true population mean might be.
confidence intervals These are also inferential tools. They tell us the range of values within which the true mean might plausibly lie, at some level of confidence, usually 95%.

The errors calculated here are **standard errors of the mean**. They give us an indiciation as to how well the sample means estimate the population means. Assuming normally distributed values, it would be very surprising if the true population means were more than two of these standard errors away from the sample means.

On the flip side, the population means could plausibly lie anyhwere in the range that is our sample means plus or minus two of these standard errors.

That said:

- Does it look as though there is evidence form the data for a difference between Adelie and Chinstrop penguins?
- What about the Gentoos compared to either of the other two?
- Do we have evidence to reject the null hypothesis. (Clue: yes we do!)