

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.

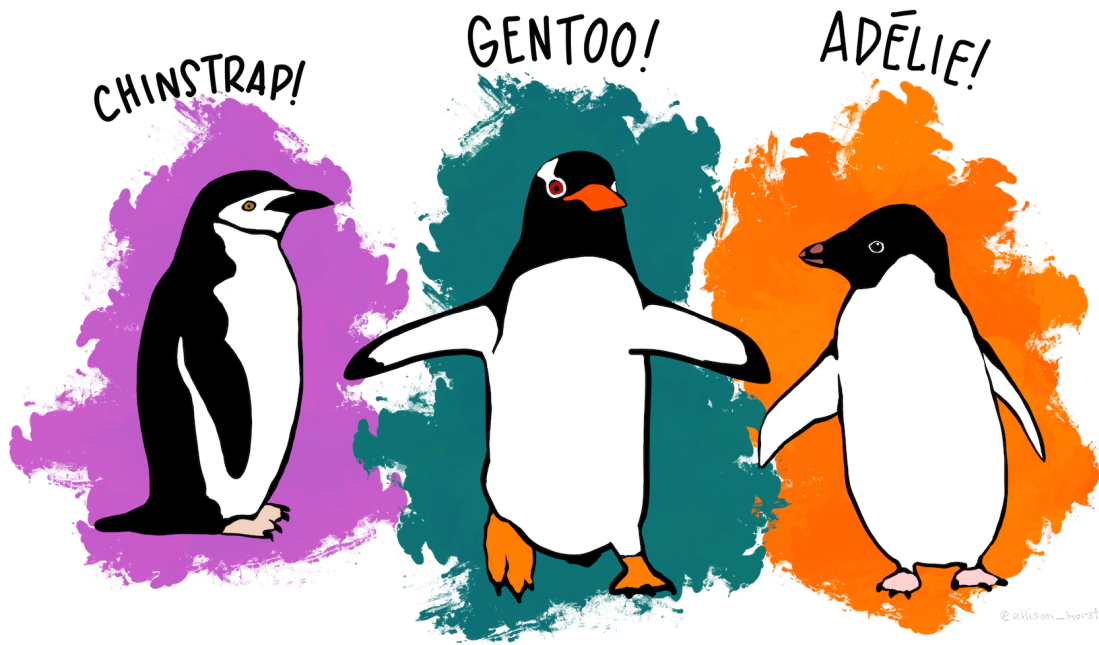


Figure 1.1: Meet the penguins

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 of the three species.

Null hypothesis: There is no difference in body 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 calculating 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) |
|-----------|----|--------------------|--------------------|
| Chinstrap | 34 | 3527.21 | 48.93 |
| Gentoo | 58 | 4679.74 | 36.97 |
| Adelie | 73 | 3368.84 | 31.53 |

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. Assuming normally distributed values, it would be very surprising if the true population means were more than two standard errors away from the sample means.

Standard errors are calculated from the standard deviations (SD) of the sample using the formula $SE = \frac{SD}{\sqrt{n}}$ where n is the sample size. This means that standard errors *do* get systematically smaller, the larger the sample. The larger the sample, the closer the sample mean is likely to be to the true population mean. Who knew?

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%. If you include error bars in a plot you can use any of these three errors, depending on the story you want to tell. Whichever, just **must** state in the figure caption which of them you have used. Failing to do this can seriously mislead the reader, since they can be of very different magnitudes.

The errors calculated here are **standard errors of the mean**. We use these because we want to get an idea, from our samples, of how plausible it is that the population means differ from each other. These population means could plausibly lie anywhere in the range that is our sample means plus or minus two of these standard errors.

- Does it look as though there is evidence from the data for a difference between Adelie and Chinstrap 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!)

1.2 Plot the data

After summarising the data, the next thing we nearly always do in deciding what the data is telling us is to plot the data. We have several choices of how to do so and each has its pros and cons. Let's run through a few of them.

1.2.1 Bar charts

1.3 Histograms

In histograms the range of a variable is split into bins of certain width, then the number of observations that fall within each bin is displayed.

They can be used to inspect a data set, even one with multiple categories, as with the penguin data. Unlike bar charts they do show the distribution of the dataset, including its central value, spread and symmetry, or lack thereof.

They do need care however in choice of the width of the bins. Make these too narrow and the histograms can look gappy, with too much scatter introduced by there not being many observations in each bin. Make the bins too wide and much of the detail of the distribution is lost. You need to find, approximately, the 'Goldilocks' width, one that is just right. Sometimes, though, you choose a binwidth that has meaning to you and the reader, such as widths of 1 m/s if you were doing a histogram of a set of wind speed measurements.

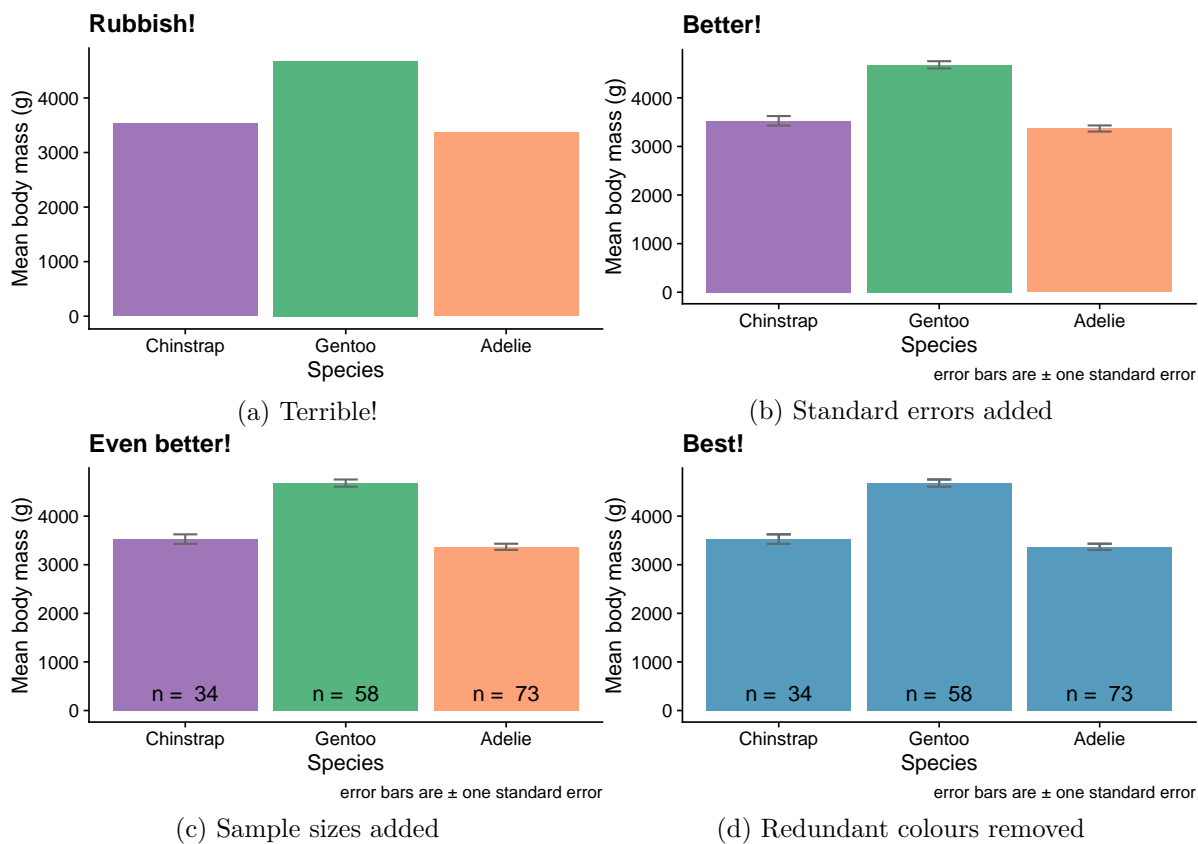


Figure 1.2: Bar charts, from worst to best

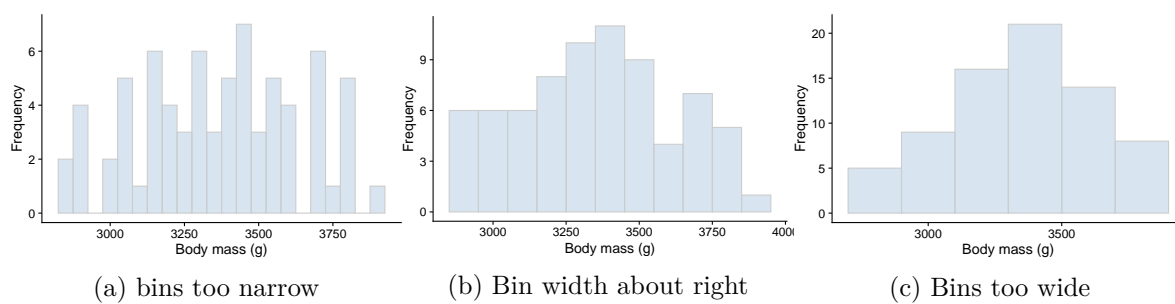


Figure 1.3: Histograms of different bin width

