Authors’ Attempts:
Chapter 6, Section 6.7 Exercises

**We have each had a go at producing suitable figures based on the ‘Stretch your understanding’ data sets we provide in bite\_force.csv and** **core\_stem.csv. See below for our attempts and what the other thought of them.**

# Bite Force

In the interest of challenging ourselves (and because we have too much time on our hands), we had a go at addressing both of the questions set out in the chapter for the naked mole-rat (hereafter NMR) data set (Hite et al. 2019).

***Question 1: How does the relationship between body mass and bite force in dominant and subordinate NMRs compare to the relationship seen across the other rodents in the data set?***

Rosalind’s attempt(*with comments by Graeme*)

Straight away the question tells me that I need to subset this data (I loaded in the full data set as ‘**bite**’). I’m only interested in: i) the dominant NMR, ii) the subordinate NMR, and iii) the other rodent species. So, once I had loaded in the full data set, I created subsets for each of these. To avoid plotting the same data point multiple times (and to avoid including the NMRs in the general trend for other rodents), I made sure that my NMRs were not included in the ‘other’ rodents subset:

**rodent <- subset(bite, group == 'rodentia' & scientific\_name !="Heterocephalus glaber")**

After checking the minimum and maximum values for both variables across all three subsets (using the **summary** function), I then plotted the data for the ‘other’ rodents. There looked to be a positive correlation, but I decided to fit a linear line of best fit between the x-axis range of the rodent data (using **summary**, **lm**,and **segments**) to be sure. The resulting line confirmed my suspicion of a moderate positive correlation, and (since I’d gone to the trouble of figuring it out) I decided to add the equation of the line to my plot. Next, I plotted the NMR data points from my subsets, making them a different point shape to the other rodent data (circles instead of squares) and choosing point shapes where I could differentiate the NMR castes using outline colour. I chose to keep them both filled with orange in order to match the other rodents (as they are also a species of rodent). Finally, I had a go at labelling the NMR data points by using the subsets’ bite\_force~body\_mass coordinates in the **text** command. Though I tried a range of positions (**pos**), the proximity of these points to each other and some of the other rodent’s data meant this didn’t turn out as neat as I’d hoped but was still OK.



Figure R.6.1.1: Body mass and bite force in dominant (black outline, n=1) and subordinate (purple outline, n=1) caste naked mole-rats (*Heterocephalus glaber*) (circles), compared to a range of other rodent species (squares, n=16). There is a moderate positive relationship between body mass and bite force across rodents, but subordinate naked mole-rats appear to exhibit a slightly higher bite force than would be predicted from their body size. Data adapted from Hite et al. (2019).

Overall, my Figure R.6.1.1 tells me that, firstly, there is a moderate positive relationship between body mass and bite force in rodents. Secondly, while dominant caste NMRs seem to fit fairly well with the relationship for rodents generally (almost positioned perfectly on the line of best fit), the subordinate caste NMRs seem to exhibit a slightly higher bite force than would be predicted from their body size. While the equation I added to the graph might be a bit of an unnecessary flourish, it would be useful if you wanted readers to be able to use your graph to calculate a predicted bite force (or body mass) of any hypothetical rodent of a given body mass (or with a known bite force). If I hadn’t added the line of best fit’s equation, I might have added grid lines behind the data points to make them easier to read off the graph, but here I thought they would make things just a bit too messy.

***Graeme’s comments:*** *I think Rosalind’s attempt is very effective. I do think the equation is useful, but I guess another alternative might have been to place it in the figure caption, and that might have allowed addition of the grid lines. I am not completely sold on the orange colour from an aesthetic standpoint, but that’s a trivial issue. More importantly, although the x axis runs from 0 to 700g, so much of the interest is really down the bottom end below 100g. I wonder if this argues for logging the x-axis.*

Graeme’s attempt(*with comments by Rosalind*)

When I fiddled a bit with Rosalind’s code to explore the issues I pointed out above, this is what I got:



Figure G.6.1.1: Body mass and bite force in dominant (black circle, n=1) and subordinate (purple circle, n=1) caste naked mole-rats (*Heterocephalus glaber*) (circles), compared to a range of other rodent species (green squares, n=16). Notice the x-axis is logarithmic. The green line is the linear regression line of best fit to the 16 other rodents: bite force =15.68 + (mass \* 0.0064). Data adapted from Hite et al. (2019).

My graph invites a rather different interpretation. I agree that, across the rodents, bite force increases in species of larger characteristic mass. However, the logged data emphasizes that this relationship is not linear. The best-fit straight-line equation tends to overestimate bite force in smaller-bodied species and underestimate it in larger ones. When we plot the two naked mole-rat data points, dominants seem a little heavier than subordinates, but subordinates seem to have a slightly higher bite-force—but I am not sure we see much evidence to suggest that they deviate strongly in bite force from what might be expected of a rodent of that size.

The only R tricks of note in my code are that I used “log = ‘x’” in my plot commands to log the x-axis and that I created the line by producing a bunch of x values equally spaced between 5 and 700, then used the equation Rosalind obtained to calculate associated y values for each, and then plotted them just the same way that we suggested in the book for time series. The last thing to note about my code is that it is mostly Rosalind’s—I only steal from the best.

***Rosalind’s comments:*** *Overall, I think Graeme’s Figure G.6.1.1 probably gives us a better sense of what is going on here biologically—this is a great example of when logging axes can really help visualize a trend that spans a broad range of values. See Figure 6.9d in section 6.5 and the corresponding code in the ‘R script for chapter 6’ for another example of how to use a logarithmic axis effectively.*

***Question 2: How does the relationship between body mass and bite force in dominant and subordinate NMRs compare to the relationship seen across all of the mammalian groups in the data set?***

Rosalind’s attempt(*with comments by Graeme*)

As the chapter suggests, this question is a bit more tricky because we will have to log our data in order to show it all on one plot. To understand why, we can look at summaries of the data set:

summary(bite$bite\_force)

 Min. 1st Qu. Median Mean 3rd Qu. Max.

 1.00 12.44 56.00 239.69 312.50 1768.00

summary(bite$body\_mass)

 Min. 1st Qu. Median Mean 3rd Qu. Max.

 4 34 555 17116 10350 294600

The full set includes measurements from mammals ranging from mice and bats to lions and bears, so—understandably—the values for bite force and body mass vary massively. Log transformations can be used to make highly skewed distributions (such as the ones we have here) less skewed, helping to make patterns in data more visible. Rather than carry out a log transformation every time a variable was used in the creation of this plot, I decided to create two new variables in the full **bite** dataset by getting R to produce the natural logarithms of all the bite\_force and body\_mass observations:

**bite$bite\_log<-log(bite$bite\_force)**

**bite$mass\_log<-log(bite$body\_mass)**

Now that the full data set contained the log variables of interest, I subsetted the data again, separating out the NMRs and rodents as I did with the first question. This time, though, I thought it might be interesting to plot the other mammalian groups separately too, so I also created subsets for opossums, carnivores, bats, and primates. Then I plotted the new ‘bite\_log’ and ‘mass\_log’ variables for all of the mammal subsets except the NMRs, choosing different colours and point shapes for each subset and making sure I kept axes limits the same. Given that the relationship between log body mass and log bite force appeared clearly positive, I added a linear line of best fit. Originally, I plotted linear lines of best fit for each subset, which looked rather messy. But then I re-read the question: we are interested in the *overall* relationship across *all of the mammalian groups*. So even though the relationship seems positive across each subset, we only need one line of best fit to show the relationship across all groups:

lm(bite$bite\_log~bite$mass\_log)

Call:

lm(formula = bite$bite\_log ~ bite$mass\_log)

Coefficients:

 (Intercept) bite$mass\_log

 0.2868 0.5787

summary(bite$mass\_log)

 Min. 1st Qu. Median Mean 3rd Qu. Max.

 1.386 3.526 6.319 6.434 9.240 12.593

x0<-1.386

x1<-12.593

a<-0.2868

b<-0.5787

segments(x0, a+b\*x0, x1, a+b\*x1, col = "black", lty = 1, lwd=2)

Before I then added the NMR data points, I thought I’d try out a different way of labelling the two castes. Because the labelling using text alone was messy in my previous attempt, I decided to try using the **segments** command to draw lines from where the data points would be positioned to areas on the plot where there would be more space for text labels (see chapters 7 and 8 for more ways to add lines, symbols, and shapes to figures). Remembering that text is positioned by an (x,y) coordinate structure and lines drawn with segments are positioned by a (x0, y0, x1, y1) coordinate structure, I used the known positions of the NMR subsets’ ‘bite\_log’ and ‘mass\_log’ variables in the code:

**segments(nmr\_dom$mass\_log, nmr\_dom$bite\_log,**

 **nmr\_dom$mass\_log+1, nmr\_dom$bite\_log+3.25,**

 **col = "black", lty = 2, lwd=1)**

**text(nmr\_dom$mass\_log+1, nmr\_dom$bite\_log+3.75, "Dominant NMR",col="black")**

**segments(nmr\_sub$mass\_log, nmr\_sub$bite\_log,**

 **nmr\_sub$mass\_log-1.5, nmr\_sub$bite\_log+2.5,**

 **col = "darkmagenta", lty = 2, lwd=1)**

**text(nmr\_sub$mass\_log-1.5, nmr\_sub$bite\_log+3, "Subordinate NMR",col="darkmagenta")**

Then, I just had to add on the NMR data points as I did in the first example. To finish, I added a legend to explain the point colours and shapes used for the different mammalian groups.



Figure R.6.1.2: Body mass and bite force in dominant (black outline, n=1) and subordinate (purple outline, n=1) caste naked mole-rats, NMR (*Heterocephalus glaber*) (orange circles), compared to 77 other mammalian species. There is a positive relationship between body mass and bite force across mammals, but naked mole-rats of both castes (though particularly the subordinate caste) appear to exhibit slightly higher bite force than would be predicted from their body size. Data adapted from Hite et al. (2019).

Figure R.6.1.2 tells me that there is a positive relationship between body mass and bite force across all the mammal groups presented. Compared to the general trend across all mammal groups, both dominant and subordinate NMRs appear to exhibit slightly higher bite force than would be predicted from their body size (though this is still more so the case for the subordinate NMRs).

***Graeme’s comments:*** *Rosalind’s attempt is a work of art. All I could think of to perhaps improve it was to change how she logged the data. She used the log function in R, which produces natural logs (logarithm to the base e ≈ 2.718). If we use base 10 instead, then I think the graph is easier to digest.*

Graeme’s attempt(*with comments by Rosalind*)

So, substituting the *log10* function into Rosalind’s code and doing a little tidying, I got this:



Figure G.6.1.2: Body mass and bite force in dominant (black outline, n=1) and subordinate (purple outline, n=1) caste naked mole-rats, NMR (*Heterocephalus glaber*) (orange circles), compared to 77 other mammalian species. On both axes, logarithms to the base 10 rather than raw data are plotted. There is a positive relationship between body mass and bite force across mammals, but naked mole-rats of both castes (though particularly the subordinate caste) appear to exhibit slightly higher bite force than would be predicted from their body size. Data adapted from Hite et al. (2019).

This is easier to interpret because if a value is 10*a* then log10 of that value is just *a*. So log10(10) = 1, log10(100) = 2, log10(1000) = 3. So even though we are using logged values, we can estimate values on the graph by eye more easily with logs done this way, I think.

***Rosalind’s comments:*** *I completely agree: using* ***log10*** *rather than the* ***log*** *function is a far better choice for easy interpretation. If I’m honest, it slipped my mind that the* ***log*** *function in R does not compute logs to the base 10! See Bigger Picture 4.1 from chapter 4 for further discussion on dealing with broad spans of values and the difference between* ***log*** *and* ***log10****—as I should have done before attempting this…*

# Women in STEM

Rosalind’s attempt(*with comments by Graeme*)

As stated in the chapter, we are interested in the numbers of men and women making up the core STEM workforce in the UK (WISE Campaign 2019). Because there is percentage (continuous) data provided (though we could have asked R to calculate percentages for us) and we are interested in changes over a given span of time—specifically 2009 to 2019—we can produce a time series.

Handily, the percentages of women and men making up the workforce are already given to us as separate variables, so there was no need to subset the original data set here. So, I first plotted the women\_percent and men\_percent data separately against the year, choosing different colours and point symbols for the two samples. I then ordered the original data set by year (as in the chapter, this data was already provided in order, but it’s a good habit to confirm this!) and joined the data points for each sample using **lines**. It is clear from looking at the time series that there was evidently a discrepancy between the % of men and women making up the STEM workforce. However, the percentage of women did seem to be increasing. Given that WISE are particularly interested in the inclusion of women in STEM, I thought it might also be interesting to show the total counts of men and women (in millions) for each year to see if the number as well as the percentage of women working in STEM was increasing over time. To do this, I used the **text** function to place labels of values from the variables women\_millions and men\_millions alongside their corresponding time series data points. Because of how the data is distributed, this worked out a lot more neatly than it did for the naked mole-rat data! I also added asterisk symbols and text below my legend to explain these values within the figure.



Figure R.6.2: The percentage (and total count in millions) of men and women in core STEM workplace roles in the UK from 2009 to 2019 (adapted from WISE Campaign 2019). Core STEM includes science, engineering, and information and communications technology.

In the final Figure R.6.2, the trends in percentages for both men and women look pretty clear. I also like the inclusion of the total count data because—importantly for WISE—a milestone of 1 million women in core-STEM occupations was reached in 2019, and this would not be evident from the percentage data alone.

***Graeme’s comments:*** *My tiny beef with Rosalind’s attempt is that by including men in the graph, all the data on women ends up being squashed down the bottom of the graph a little—and we aren’t really interested in men in this context. Further, I would expect the percentage for women and men to add up to 100% or very near to this, so the information on men seems redundant.*

Graeme’s attempt(*with comments by Rosalind*)

I will continue my lazy cannibalizing of Rosalind’s code and replot the data for females only. I definitely support her decision to start the y-axis from zero in this case, since zero seems a very real reference point. However, in the interest of really focussing in on the data as much as possible to see trends, I am going to deviate from this. I also like including the data on total numbers, since we have the space. Notice that we need to give the total number separately for each year since the size of the whole workforce can change as well as the fraction of the workforce made up of women.



Figure G.6.2: The percentage (and total count in millions) of women in UK core STEM jobs from 2009 to 2019 (adapted from WISE Campaign 2019). Core STEM includes science, engineering, and information and communications technology.

My graph, I think, makes the general trend towards increase very clear. However, I accept that there is always the danger that a casual reader could miss how low the percentages still remain—there is admittedly no danger of this kind with Rosalind’s graph. We see the interesting trend that the total number of women has increased year on year, but we have the unusual case in the last two years that although the gross number of women employed continued to rise, it did so at a slower rate than the sector grew as a whole. So the percentage of women in the STEM workplace actually dropped a little.

***Rosalind’s comments:*** *Overall, I think Graeme’s attempt does a really good job of highlighting the women’s data: the general trend of increase over time is certainly more apparent than it was in mine. However, depending on the intended audience, a quick glance could suggest that the number of women in STEM has shot up in recent years and that nothing needs to be done to encourage any further increase. Fully absorbing the non-zero y-axis is key to accurate interpretation here. Perhaps one way to highlight the increase while keeping it in context would be to include an inset of a version of the time series with a y-axis starting at zero—see sections 3.7, 6.6 and 8.2 for further information on inset plots. In terms of the figure’s design, I only have a couple of minor issues. Firstly, the unrounded y-axis ranges look a little untidy to my eye (15.8 instead of 16 seems an oddly specific minimum value), and this consequently sits a major grid line somewhat awkwardly immediately above the x-axis. Also concerning the grid lines, they look to have been drawn on top of the data points and connecting lines, rather than being prepared on a base plot ahead of adding on the actual data—visually, this might emphasize the grid lines rather than the data itself. Finally, I think the figure caption would have been more informative if it had highlighted that the ‘total count in millions’ is shown in red text on the figure.*

# References

HITE, N. J., GERMAIN, C., CAIN, B. W., SHELDON, M., PERALA, S. S. N. & SARKO, D. K. 2019. The Better to Eat You With: Bite Force in the Naked Mole-Rat (*Heterocephalus glaber*) Is Stronger Than Predicted Based on Body Size. *Frontiers in Integrative Neuroscience,* 13.

WISE CAMPAIGN. 2019. *2019 Workforce Statistics – One million women in STEM in the UK* [Online]. Available: <https://www.wisecampaign.org.uk/statistics/2019-workforce-statistics-one-million-women-in-stem-in-the-uk/> [Accessed 28/09/2020 2020].