Authors’ Attempts:
Chapter 5, Section 5.7 Exercises

 **We have each had a go at producing suitable figures based on the ‘Stretch your understanding’ data sets we provide in haemoglobin\_and\_elevation.csv, nova\_scotia\_birds.csv, and** **caffeine.csv. See below for our attempts and what the other thought of them.**

# Haemoglobin and Elevation

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

Without further ado, here is my attempt:



Figure G5.1: Tukey Boxplots of measured haemoglobin concentration (g/dl) of residents of four regions around the world. Above each plot are the sample size and the average altitude above sea level of that region.

I have arranged the four regions in order of altitude, and provided sample sizes and altitudes as text in the figure (I got the sample sizes by using the summary function). I decided to include altitudes in the figure rather than the caption since it is so fundamental to the scientific question we are invited to explore. I also thought sample sizes should be highlighted because they are very uneven—and that likely explains the high number of outliers for the USA. I decided to start the y-axis from a non-zero value because I don’t think zero haemoglobin concentration is a particularly natural reference point—nobody has no haemoglobin.

Turning to the scientific question, it does look as though people from the Andes have a higher haemoglobin concentration. But we should be slow to explain this simply in terms of them living at altitude since the Ethiopians and Tibetans also live at high altitude but appear to be pretty similar in haemoglobin concentration to sea-level dwelling folk from the US.

Let’s just look at the equivalent violin plot (see Scientific Approach 4.2 for a brief description of violin plots and the ‘Authors’ Attempts for chapter 4’ for more violin plot examples).

 Figure G5.2: Violin plots of measured haemoglobin concentration (g/dl) of residents of four regions around the world. Above each plot are the sample size and the average altitude above sea level of that region.

I guess this might be preferable: it reassures us that in each region the data is unimodal. I don’t especially miss seeing the outliers that we saw with the boxplot, but I respect your right to view that differently. [On reflection, I think I should have capitalized the first letter of the y-axis title in both of these. I mention this because I know Rosalind is not petty enough to bring up such a minor detail. 😉]

***Rosalind’s comments:*** *I think both of these figures are easy to interpret and full of useful information. The figure captions include details important to the interpretation of the figures, but including the source of the data here too would have made them ideal. The violin plot certainly looks neater, given that outliers are incorporated into the shapes, but personally I still prefer the data summaries that the boxplots offer—I find it easier to compare boxes of equal width than to try to incorporate the varying widths of the violin plot shapes into my comparisons across samples. Violin plots would definitely be worth considering if any of the samples consisted of multi-modal data, though, as important patterns in the distribution of the data could be obscured by a boxplot.*

*In terms of aesthetics, I took slightly different approaches to the presentation of the elevations and sample sizes in my attempt, but presenting all the information for each sample immediately above it does make it very easy to associate the information with the relevant sample. I also prefer my y-axes to run right to the very edges of the plot, but this is just a personal preference.*

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

I opted to stick with a boxplot for this data. As we are primarily interested in comparing data summaries rather than distributions, and there are more than three samples, it does not seem that histograms would be much help to us here. Further, although I see the value of showing the shape of each sample’s data distribution alongside summary statistics (as demonstrated well by Graeme’s Figure G5.2), I also decided against trying a violin plot for this data. This is simply because, to me, a boxplot is more intuitive; and I find it easier to compare the median lines, quartiles, and confidence limits across samples when they do not vary substantially in shape (like violin plots can). Also, because all of the samples here consist of unimodal rather than multimodal data, I don’t think the choice to use a boxplot obscures any important patterns.

So, here is my boxplot:



Figure R5.1: Tukey boxplots showing the haemoglobin concentration (g/dl) of residents of four populations around the world, with the sample size from each population given above the respective box. The residences varied in altitude above sea level, as detailed by the legend. Data adapted from a study by Beall et al. (2002).

Like Graeme, I thought that including sample sizes in the plot itself was important because they differ so greatly, so I have also positioned these above the boxes. But I have varied the height of the sample sizes up the y-axis depending on the maximum value of each sample (see Additional Guide 5.1 Bar Charts for Quantitative Multiple-samples Data for further examples of this). I also included the elevation of each locality, as this is key to our question of whether high-altitude living selects for greater oxygen-carrying capabilities. However, I used colour and a legend to communicate the different elevations—a large part of this decision was driven by a desire to highlight the fact that the Andes and Tibet have similar elevation, but their residents seem to differ significantly in haemoglobin concentration. To me, colour draws attention to this more effectively than text alone.

Comparing our code, Graeme used a really neat line of code to re-order the four localities by the elevations given in the chapter, whereas I went with the slightly more laborious route of subsetting each locality and inputting each into my boxplot separately. However, this then let me get R to print the sample size for each residence onto my plot (using the **length** function)—as seen in chapter 5—and thereby reduce my chance of human error when typing in values. I also decided to position the Andes ahead of Tibet in the descending elevation order, even though they were both 4000m, partly because this made sense in terms of the alphabet but also, handily, because this locality has residents with the higher haemoglobin concentrations of the two and—to my mind—starting with the highest values makes the plot easier to interpret.

In terms of the scientific question, we draw the same conclusions from this plot as from either of Graeme’s: a high-altitude lifestyle alone does not explain greater oxygen-carrying capabilities in residents from the Andes because residents from Tibet and Ethiopia live at high altitude but have similar haemoglobin concentrations to residents from the USA at sea level.

***Graeme’s comments:*** *There isn’t really much to choose between our boxplots—but in a number of small ways, Rosalind’s comes out on top. I agree that her ordering of the four groups makes interpretation a little easier. I also agree that her use of colour emphasizes similarities and differences in altitude a little better than my text. Although it is pretty obvious, you need to read my caption to find reference to elevation—whereas this word appears in her legend, as does confirmation that 0m is equivalent to sea level. These additions perhaps make absorbing the major points of the figure more obvious at first glance. Most minor of all, having the text for each region immediately above the highest value (rather than in a line across the top as in my case) is probably a bit easier on the eye.*

# Nova Scotian Birds

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

Really this is an exercise in data manipulation. The first thing I did was check that there was a value recorded for each area and each year (and there was). Then I aggregated the dataframe to give me a total value across the five years for each region. Finally, I reordered the regions by these totals and produced the bar chart below. I think a bar chart is probably the better option compared to a table, since we are looking for a trend rather than for exact values. However, I think this is a marginal call, and a table would likely be just as good. Anyway, here is my final bar chart:



Figure G.5.3. Aggregated data on number of bird species observed in different areas of Nova Scotia from the National Audubon Society Christmas Bird Count from the 5 years (2014-2018) (Audubon, 2020). Notice that the value given is an aggregate across 5 years, if the same species was seen in a given area in each of the five years, then it contributes five to that area’s aggregative count.

Now I feel that this figure provides the clearest way to address the question as set, but I am a little bit uneasy about aggregating data across years, since we seem to be throwing away information we have about year-to-year variation. I think we could have gone for a grouped bar chart in the way we describe in chapter 3, section 3.4, but my guess is that this would be a lot messier for a situation where we haven’t actually been asked about year-to-year variation. Actually, I gave this a go and it looked like this:



Figure G.5.4. Annual data on number of bird species observed in different areas of Nova Scotia from the National Audubon Society Christmas Bird Count from 2014-2018 (Audubon, 2020).

On reflection, I guess my second attempt is better. It shows pretty clearly that the trends are consistent across all five years—Wolfville wins every time. I have kept the order of areas alphabetical this time, since there wasn’t any other obvious way to order them. I was never tempted to use a boxplot or histogram for a situation where I only have five values for each area.

***Rosalind’s comments:*** *It is important to emphasize that, while we advise against using bar charts for quantitative data in chapter 5, here we only have five recorded values for each area. This sits within our ‘seven or fewer different values’ limit which suggests that using a figure type best suited for qualitative data types is a better call. I agree with Graeme that bar charts are probably the best call for this data set, and the first bar chart Figure G.5.3 presenting aggregated data does clearly show us that the total counts of species do tend to vary noticeably between the areas. It is great that the aggregated approach is also clearly explained in the legend. However, Figure G.5.4 shows the same trend, but with added detail on year-to-year variation. Although it shows more than the question specified, it gives us a fuller picture of what is going on biologically. The trends in Figure G.5.3 were not due to extreme outlying species counts in different areas; they result from longer-term trends in the different areas seen consistently across most years. I cannot think of many ways to improve on the approach Graeme took with this data so, for the sake of exploration, I instead tried some alternative ways to present the data.*

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

I think that Graeme’s grouped bar chart (Figure G.5.4) was a great way, and perhaps the best way, to present this data clearly. However, in the interest of demonstrating a couple of alternatives, I will try some different presentation tools.

To start with, I decided to try a boxplot. Although we recommend that if you have discrete data with seven or fewer different values, you should consider treating it as qualitative (aka categorical) data instead (see chapters 2 and 3), some believe that a sample size of 5 is OK as an absolute minimum for boxplots; and that is what we have for each area in Nova Scotia. I agree with Graeme that the data we have on species counts is perhaps best presented in full, including year-to-year variation rather than just the total counts and/or summary statistics, so here I tried overlaying my boxplots with strip-charts showing all of the individual data values. See section 6.5 in chapter 6 for details on strip-charts (aka univariate scatterplots) and another example. However, with this approach I chose to only present four of the seven areas, as the plot became quite messy when I tried to fit in boxplots and strip-charts for all seven areas. Overall, the resulting figure looks a bit odd, as with the small sample sizes most boxes are missing one or both whiskers. But at least including the strip-charts alongside makes it easy to identify which years some outlier counts were recorded in, e.g. fewer species were recorded than usual in 2015 for Apple River, and in Tatamagouche more species than usual were recorded in 2015, but fewer species were recorded than usual in 2017. Overall, the distribution of species counts across years and the differences between areas match what we saw for these four areas in Graeme’s grouped bar chart, but the grouped bar chart was clearer, contained more information, and was a lot more intuitive to read. Boxplots are arguably a better choice if you really want summary statistics to compare (e.g. medians), but they are not a good call if you have very small sample sizes and/or lots of samples.



Figure R.5.2. Tukey boxplots overlaid by strip-charts showing the number of bird species observed in four different areas of Nova Scotia across a period of 5 years (2014-2018), from the National Audubon Society Christmas Bird Count (Audubon, 2020).

Next, I decided to try presenting all of the information in the data set as a table in Word, using the guidelines outlined in section 2.3.1 and some of the tools recommended in section 2.3.2 of chapter 2. I arranged the areas by descending total/mean counts to help give viewers a sense of the overall trend as they scan down the table (although this may not be immediately obvious as a trend because of the inclusion of all of the other columns of numerical data). I rounded mean species counts to the nearest whole number and used a footnote to keep my heading in the ‘Mean’ column brief while including a full explanation of this. I also presented the values for each year running left-to-right across the table (an intuitive way to present time variables), to demonstrate the year-to-year variation (though such variation is harder to discern from numbers alone than it is in a visualization like a bar chart). By including the years as column headings, I reduced redundancy, as I didn’t need to split the areas into separate rows for each year, and kept the table compact. This layout also allows for easy comparison of the total counts of species in each year, as well as the total and mean counts, between the different areas—this comparison *between areas* is what we are interested in. I used alternate shading here, as there are multiple columns (seven in total) containing numerical data and shading could help readers follow along the values associated for each area. I also used emboldening on the column headings and restricted cell borders to just the top and bottom of the table for structure.

Table R.5.1. The number of bird species observed in seven different areas of Nova Scotia during annual surveys running from 2014-2018, including the total number of species recorded over the 5-year period and the mean number per year. Data from the National Audubon Society Christmas Bird Count (Audubon, 2020).

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Year** |  |  |
| **Area** | **2014** | **2015** | **2016** | **2017** | **2018** | **Total** | **Mean1** |
| Wolfville | 77 | 73 | 70 | 75 | 79 | 374 | 75 |
| Broad Cove | 68 | 57 | 69 | 61 | 68 | 323 | 65 |
| Sheet Harbour | 51 | 48 | 56 | 50 | 56 | 261 | 52 |
| Apple River | 54 | 42 | 52 | 54 | 54 | 256 | 51 |
| Truro | 51 | 55 | 47 | 55 | 47 | 255 | 51 |
| Tatamagouche | 47 | 59 | 48 | 36 | 50 | 240 | 48 |
| Springville | 45 | 45 | 47 | 48 | 47 | 232 | 46 |
| 1*Rounded to nearest whole number* |

A table such as this would be easier to use if it was necessary for precise values to be clearly read, but I am still not convinced that it is as helpful as the grouped bar chart if what we are interested in are the general trends across areas and between years.

***Graeme’s comments:*** *I agree with every word that Rosalind wrote. I think we agree that for the biological question that was asked, the bar chart is definitely the best approach. If we wanted to discuss specific numbers of species as well as look at the broad trend, then that might have tipped the scales in favour of using the table. Rosalind drew the best boxplot she could, but it doesn’t really come close to the ease of reading that the bar chart gives. That is because boxplots are good for summarizing a lot of the key features of a (unimodal) distribution of data—but five points just don’t qualify as the sort of distribution that benefits from summarizing. You could imagine dropping the boxplots and just presenting the four strip-charts—but these contain no added information over the grouped bar charts and would be harder to interpret quickly. Overall, I was really lucky to get in first with the bar charts, and Rosalind was a really good sport to try different solutions.*

# Caffeine and Finger Tapping

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

Here we have ten values for each combination of person type (gamer or not) and caffeine dose (three different levels). This is not a huge number, but probably enough that a boxplot is a reasonable summary. However, I would be sure to emphasize the small sample sizes in the figure caption. Following the methodology described in the chapter, this is what I end up with:



Figure G5.5: Tukey boxplots of maximum finger tapping frequency in relation to caffeine ingestion and experience of gaming. 30 professional gamers (shown in grey) and 30 college students (shown in green) were each randomized into 3 treatment groups of ten. Treatment groups were given different doses of caffeine (0, 100, and, 200mg). Two hours after treatment administration, the maximum number of times the participant could tap their finger in 60 seconds was recorded. Thus each boxplot is a summary of 10 values (Draper and Smith, 1981).

It seems pretty clear that caffeine improved performance but that there was no obvious strong difference between the two types of participants.

***Rosalind’s comments:*** *Aesthetically, there isn’t much I’d change about Graeme’s grouped boxplot, except I might trim the y-axis a little to reduce the space at the bottom of the plot. However, the results are not what I got when I plotted the data. Looking at Graeme’s code, it looks like he has made the tiny (but fatal!) error of using = rather than == when subsetting. R is very picky about its ‘operators’ and == is needed to specify specific levels ‘exactly equal to’ within a variable. Using = alone means that both of the samples plotted here—Gamers and Non-gamers—both still include the entire data set.*

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

I liked the look of Graeme’s grouped boxplot so much that I decided to simply cannibalize his code for my attempt with this data set. Crucially, though, I used == rather than = when subsetting (see section 5.6.2 for another example of using **==** in chapter 5):

**gamers <- subset(tapping, participant == "gamer")**

**non <- subset(tapping, participant == "non-gamer")**

This produced two data sets: ‘gamers’ with only the data for the gamers and ‘non’ with only the data for the non-gamers.

I also noticed that the axes labels and the y-axis tick labels were given in the code for the boxplots for both samples, which can cause them to appear thicker (as they essentially get positioned twice by R)—so I removed these from the code for the second boxplot.



Figure R.5.2: Tukey boxplots showing the number of finger taps achieved by professional gamers (n=30) and college students (n=30) treated with different doses of caffeine (n=10 for each group-treatment combination) in 60 seconds. Data from Draper and Smith (1981).

When looking at the resulting boxplot, it seems clear that caffeine consumption improved finger-tapping performance across both groups, but also that gamers out-performed non-gamers across all caffeine treatments. It is still hard to say whether caffeine has more or less of an effect on finger-tapping by professional gamers compared to college students, as both exhibited an increase in ability, but perhaps the distance between treatment medians could cautiously be suggested to be slightly greater for the group of gamers.

***Graeme’s comments:*** *Feel my embarrassment! It doesn’t matter how pretty the design features of your graph are if you plot the wrong data! I should not have made this mistake. Firstly, I was lazy—after subsetting, I should have looked at my two subsets to see that they looked as expected. In this case, just looking to see how many data-points I had in each subset would have alerted me to the problem. I didn’t do that, and this mistake slipped through. Even then, once I had drawn my figure, it should have made me a little suspicious that the three boxes were absolutely identical for gamers and non-gamers. With these sample sizes, even if gaming-experience had no effect on average, you would still expect that general person-to-person variation in performance would have led to some difference between groups of ten. So, I was insufficiently critical as well as lazy, and now must wipe egg from my face. Rosalind’s plot is clear and polished, her caption is more succinct than mine, and—best of all—she actually plots the correct data!*

# References:

AUDUBON. 2020. *Christmas Bird Count* [Online]. Available: [https://netapp.audubon.org/CBCObservation/Historical/ResultsByCount.aspx#](https://netapp.audubon.org/CBCObservation/Historical/ResultsByCount.aspx) [Accessed 2020].

BEALL, C. M., DECKER, M. J., BRITTENHAM, G. M., KUSHNER, I., GEBREMEDHIN, A. & STROHL, K. P. 2002. An Ethiopian pattern of human adaptation to high-altitude hypoxia. *Proceedings of the National Academy of Sciences,* 99**,** 17215.

DRAPER, N. R. & SMITH, H. 1981. *Applied regression analysis,* New York, John Wiley & Sons.