Authors’ Attempts:
Chapter 4, Section 4.6 Exercises

 **We have each had a go at producing suitable figures based on the ‘Stretch your understanding’ data sets we provide in ‘rat\_lick\_data.csv’ and ‘firefly\_spermatophore\_mass.csv’. See below for our attempts and what the other thought of them.**

# Rats licking

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

In the way this data is introduced, there is a heavy hint that the data might have more than one mode, and sure enough the simplest histogram shows just that.



Figure G.4.1.1: Sheng et al. 2016 used a rodent brief-access taste aversion (BATA) methodology to assess the bitterness of quinine. For 720 rats, they recorded how many licks they took from sipper tubes filled with a 3% solution of quinine hydrochloride dihydrate during an 8-second long trial.

The fact that the data has more than one mode seems like one of the dominant features of the data that you are likely going to want to discuss. In such situations, I think a histogram is going to be a much more effective means of displaying the data than a boxplot, so the thing to do is to tidy our basic histogram up a bit.



Figure G.4.1.2: Sheng et al. 2016 used a rodent brief-access taste aversion (BATA) methodology to assess the bitterness of quinine. For 720 rats, they recorded how many licks they took from sipper tubes filled with a 3% solution of quinine hydrochloride dihydrate during an 8-second-long trial.

There is nothing in my figure than we haven’t encountered in chapters. I think the added horizonal lines really help you to discuss the heights of the bars over on the right side. Just for comparison, let’s put together a boxplot.



Figure G.4.1.3: Sheng et al. 2016 used a rodent brief-access taste aversion (BATA) methodology to assess the bitterness of quinine. For 720 rats, they recorded how many licks they took from sipper tubes filled with a 3% solution of quinine hydrochloride dihydrate during an 8-second-long trial. This data is presented as a Tukey boxplot.

As we feared, this boxplot gives no hint of arguably the most interesting aspect of the data, the bimodality, so we would definitely prefer the histogram for this data set. But before leaving the boxplot, I will adjust the borders the way Rosalind showed me in chapter 4 so that there isn’t a big gap between the bottom of the figure and its legend.



Figure G.4.1.4: Sheng et al. 2016 used a rodent brief-access taste aversion (BATA) methodology to assess the bitterness of quinine. For 720 rats, they recorded how many licks they took from sipper tubes filled with a 3% solution of quinine hydrochloride dihydrate during an 8-second-long trial. This data is presented as a Tukey boxplot.

Anyway, let’s see if a violin plot gets us out of this pickle (see Scientific Approach 4.2 for a brief description of violin plots).



Figure G.1.5: Sheng et al. 2016 used a rodent brief-access taste aversion (BATA) methodology to assess the bitterness of quinine. For 720 rats, they recorded how many licks they took from sipper tubes filled with a 3% solution of quinine hydrochloride dihydrate during an 8-second-long trial. This data is presented here as a Tukey Violin plot.

The violin plot is really easy to produce—I barely changed my code from the boxplot. All the information from the boxplot is reproduced in the centreline of the violin plot, but we have the added information from the blue coloured wings that give the violin plot its name—this is a probability distribution that is a guess based on the sample of the underlying distribution from which the sample is drawn. The wider the wings the more likely values are to be to fall in that region of values. What we see here is a bulge at low values and then another bulge peaking in the high forties—so the violin graph does capture the bimodality that we saw in the histogram. For this data, the violin plot does seem a much more appropriate representation of the data than a boxplot—but is it better than a histogram? The simple truth is that most folk are just not used to looking at violin plots. If your report is aimed at a non-scientific audience, then I think the histogram will be your best bet, but the violin plot could be the better choice for addressing a more technical audience because it gives you the added ‘boxplot’ information from the midline of the violin. In fact, working through this example is really converting me to the violin plot: it gives you all the information of the boxplot and some extra information as well. That extra information won’t always be all that valuable, but sometimes (like in this example) it will be really handy. Further, even when the extra information isn’t all that useful, the cost to this (in terms of either distracting chart clutter or computational difficulty of production) doesn’t seem very high.

***Rosalind’s comments:*** *Overall, I think all of Graeme’s final versions of figures are clear and effective. Comparing the aesthetics of our histograms, one difference is that I neatened up both my axes with* ***xaxs="i"*** *and* ***yaxs="i"****—but this can really just be a personal choice as to what is more visually-appealing. Comparing the information communicated, I chose to include the detail about the trials’ duration in my accompanying caption rather than in my x-axis label, but the emphasis you choose to put on such a methodological detail in a figure might depend on the importance of that facet to your study or as a comparison to previous studies. Across all Graeme’s figures, I liked the way he ordered the information in his captions—starting big picture with the purpose of the study before homing in on the specifics and explaining the ‘Tukey’ method for his boxplots and violin plot.*

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

As Graeme pointed out, the chapter hints that this data is unimodal. We are also asked to look specifically at the *distribution* of the licks rats took, rather than a *summary* of the distribution (i.e. descriptive statistics). For these reasons, a histogram is my weapon of choice.

When plotting my histogram, R automatically selected easy-to-digest bin spacing that made the bimodal distribution of the data clear, so I didn’t need to adjust this (using **breaks** to specify bins) myself. However, I did neaten up the default axes by setting rounded axes ranges and using **xaxs="i"** and **yaxs="i"** to force my histogram cells to sit directly beside and on the axes. I also added some grid lines to improve interpretation.



Figure R.4.1.1: The number of licks rats took from sipper tubes filled with 0.03mM concentration quinine hydrochloride dihydrate over a duration of 8 seconds (n=720). Data from a study utilizing rodent brief-access taste aversion (BATA) experiments to assess the bitterness of quinine (Sheng et al., 2016).

This histogram looks pretty clear and gives viewers a good sense of the distribution of licks—rats tended towards either licking very few times or licking ~50 times during trials. As Graeme demonstrated well, using a boxplot for data like this obscures the interesting finding of the bimodal distribution, but his violin plot managed to capture the bimodality while also showing some useful descriptive statistics. I next had a go at adding a few descriptive statistics to my histogram to see how this compared. To do this, I looked at a summary of the data (with the **summary** function) and then added lines (see section 7.3.4), text (see section 7.3.5), and arrows (see section 8.6) to my plot.



Figure R.4.1.2: The number of licks rats took from sipper tubes filled with 0.03mM concentration quinine hydrochloride dihydrate over a duration of 8 seconds (n=720). Measures of range and central tendency are included on the plot. Data from a study utilizing rodent brief-access taste aversion (BATA) experiments to assess the bitterness of quinine (Sheng et al., 2016).

I am not convinced that this histogram with added information is a better alternative to Graeme’s violin plot. The histogram might be simpler to interpret and more familiar to most readers, but adding the descriptive statistics as extra lines and text looks a bit cluttered and somewhat undermines the simplicity of the histogram. If you want readers to be able to identify key descriptive statistics while getting a sense of the distribution, then a violin plot might be your best bet. If all that you want to convey is the distribution alone, then a histogram can be very effective, particularly where your data is not unimodal.

***Graeme’s comments:*** *There is no question Rosalind’s way of neatening the axes is a good idea. I also think the added summary statistics do add real value to the histogram without over-cluttering it. Whether you would go to this added trouble would, of course, depend on the points you wanted to make with your figure.*

# Fruitfly spermatophores

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

Again, it seems to make sense to use a rough histogram to look at the data as a starting point.



Figure G.4.2.1: Masses (in milligrams) of the spermatophores dissected from 35 *Photinus ignitus* fireflies. These values were inferred by Whitlock and Schluter (2014) from a figure in a paper exploring female preference for male courtship ﬂashes (Cratsley and Lewis, 2003).

It looks like there is one really big value and some small ones. I ran the summary function over the data set, and this gave:

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

0.03700 0.05550 0.06600 0.06877 0.07800 0.17200

It’s a relief that the small values don’t seem unrealistically small, they are positive, and the smallest is about 50% of the median value. The biggest value is really big, almost three times the size of the median values, but this too does not seem utterly implausible. Also, I don’t think it is likely to be a typing error, since 0.0172 would be a lot smaller than any of the other values with the decimal in the wrong place. Having got to grips with the data, the way to go now is to polish up this histogram.



Figure G.4.2.2: Masses (in milligrams) of the spermatophores dissected from 35 *Photinus ignitus* fireflies. These values were inferred by Whitlock and Schluter (2014) from a figure in a paper exploring female preference for male courtship ﬂashes (Cratsley and Lewis, 2003).

Now I admit this looks a lot like the histogram I used for the first example, and I indeed modified the code I used for that one because I can’t really see anything I would do differently. Now I think the thing to do is to have the courage of my convictions and go for a violin plot rather than a boxplot (again, see Scientific Approach 4.2 for a brief description of violin plots).



Figure G.4.2.3: Tukey violin plot of masses (in milligrams) of the spermatophores dissected from 35 *Photinus ignitus* fireflies. These values were inferred by Whitlock and Schluter (2014) from a figure in a paper exploring female preference for male courtship ﬂashes (Cratsley and Lewis, 2003).

Ok, I am a wee bit disappointed in this. All the things I liked about the violin plot in the last example are on show here, but I am not crazy about how the high-mass outlier is handled. It looks a bit odd, but maybe I just have to get used to this. Let’s see a boxplot for comparison.



Figure G.4.2.4: Tukey boxplot of masses (in milligrams) of the spermatophores dissected from 35 *Photinus ignitus* fireflies. These values were inferred by Whitlock and Schluter (2014) from a figure in a paper exploring female preference for male courtship ﬂashes (Cratsley and Lewis, 2003).

I like this better. I think the single value that seems unusually high is an important part of this sample, and the boxplot shows you where that value lies a bit more clearly than the violin plot does. So I think I am maybe reining in my newfound enthusiasm for violin plots a little—they definitely deal with multi-modality better than boxplots, but maybe box plots deal with outliers more clearly. So, which of these is best for your data is likely to depend on the features in your sample that you think are most important.

***Rosalind’s comments:*** *I totally agree with Graeme’s conclusion—the violin plot made it less clear where the outlier lay, and its visual look also seemed to give more weight to the outlier than the simple empty circle point of the boxplot did. There’s nothing I would really change about these figures, except perhaps using my* ***xaxs = "i"*** *trick on the histogram.*

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

Like Graeme, I started by looking at this data as a histogram. The masses appeared to be unimodal (not multi-modal like the rat lick data) and fairly normally distributed (think bell-shaped curve) with the exception of one outlier. Because there is no multi-modal distribution, but there is an important outlier, a boxplot that summarizes the values while highlighting the unusual data point seems to me like the most helpful way to present this information.

I couldn’t think of a way to meaningfully improve on Graeme’s boxplot, but I did have a go at adding grid lines behind the boxplot, rotating the y-axis tick mark labels (using **las**, and then **oma** and **mtext** to reposition the y-axis label, see sections 7.5.1, 7.2.1, and 7.3.5, respectively), and choosing a nicer colour 😊:



Figure R.4.2.1: The masses (in milligrams) of spermatophores dissected from *Photinus ignitus* fireflies (n=35) in a study by Cratsley and Lewis (2003) exploring female preference for male courtship ﬂashes. Values inferred by Whitlock and Schluter (2014).

Whether the grid lines are included or not, the outlier in the data is really apparent in our boxplots—and it is clear that most fireflies had spermatophores of mass close to 0.07mg.

***Graeme’s comments:*** *All of Rosalind’s little tweaks definitely improve ease of reading. How she imagines that hers is a nicer colour is beyond me—you have to pray that Rosalind never knits you a scarf for your birthday!*

# References:

CRATSLEY, C. K. & LEWIS, S. M. 2003. Female preference for male courtship flashes in *Photinus ignitus* fireflies. *Behavioral Ecology,* 14**,** 135-140.

SHENG, Y., SOTO, J., ORLU GUL, M., CORTINA-BORJA, M., TULEU, C. & STANDING, J. F. 2016. New generalized poisson mixture model for bimodal count data with drug effect: An application to rodent brief-access taste aversion experiments. *CPT: pharmacometrics & systems pharmacology,* 5**,** 427-436.

WHITLOCK, M. C. & SCHLUTER, D. 2014. The Analysis of Biological Data. 2 ed.: W. H. Freeman.