Workshop 2: Displaying data using graphs and tables

The purpose of this exercise is to tour the table and graphics capabilities of R, and to explore the best methods for displaying patterns in data.

Data set 1: Body mass of late Quaternary mammals

These data were published as a data paper in Ecology and deposited in the Ecological Archives (F. A. Smith, S. K. Lyons, S. K. M. Ernest, K. E. Jones, D. M. Kaufman, T. Dayan, P. A. Marquet, J. H. Brown, and J. P. Haskell. 2003. Body mass of late Quaternary mammals. *Ecology* 84: 3403.) See the metadata for a description.
Most of the variables are categorical, with multiple named categories. "Continent" includes mammals on islands ("Insular" category) whereas "Oceanic" refers to marine mammals. Body mass is the sole numeric variable. The "status" variable indicates whether species is currently present in the wild (extant), extinct as of late Pleistocene (extinct), extinct within the last 300 years (historical), or an introduced species (introduction).

Read and examine the data

The original data were saved in mammals.csv Download the file to your computer and open in a spreadsheet program (e.g., Excel, Calc) to have a look at it.
Start R and use `read.csv` to read the contents of the file to a data frame. You will need to modify the default `na.strings='NA'` argument to `na.strings=''` because in this data file `'NA'` is used to symbolize North America in the continent column rather than missing data (don't do this in your own data). Remember to set your working directory if you are not working with an RStudio project.
Carry out the following inspections of the data

1. Use the `head` function to view the first few lines of the data frame on the screen. You'll see that every row represents the data for a different mammal species.

2. For the two most interesting character variables, tabulate the frequency of cases in each group (remember that the category `'NA'` in continent stands for North America, not missing data). Use the function `table` in base R.

   If you want to use `dplyr` you can use the `group_by` and `summarise` functions. Remember to load `dplyr` first using the function `library`.

3. You'll notice in the frequency table for the variable "continent" that there's a typo in the data. One case is shown as having the continent `'Af'` rather than `'AF'`. Fix this using R (you may want to refer back to the last workshop to do this to change ecomorph names).

4. Create a new variable in the mammal data frame: the log (base 10) of body mass. You can do this by using the `$` symbol (see below code):

```
mammals$log_body_mass <- ____________
```

   and a `dplyr` option:

```
mammals <- mutate(mammals, log_body_mass = ____________)
```

Visualizing frequency distributions

To visualize the data we will be creating plots with base R (or `ggplot` for those who are interested). If you want to use `ggplot`, make sure you install `ggplot2` using the `install.packages('ggplot2')` function and you load it using the `library` function. `ggplot2` is a package based on layers of code. Every time you want to see something new added to your plot, you must add a new layer with each layer being separated by the `+` symbol.

1. Which continent has the greatest number of mammal species? Which has the least? Use a simple bar graph to find out. Rotate the labels on the y-axis to horizontal. **Hint:** use the `barplot` function (or `geom_bar()` if using `ggplot`).

2. Redo the bar graph using the `cex.names` option to control the size of the category names (this might be needed to ensure that there is room for them all.

3. Redo the bar graph in color. Add a label to the y-axis.

4. Redo the bar graph to increase the limit of the y-axis to 1500 species. (The result might not be immediately evident in the axis labeling because by default R applies internal rules to make graphs "pretty". Try increasing the limit to 1600 or 1700 and see what happens.)

5. The plot categories are listed in alphabetical order by default, which is arbitrary and makes the visual display less efficient than other possibilities. Redo the bar graph with the continents appearing in order of decreasing numbers of species. **Hint:** use the `sort` function.

6. Generate a histogram of the body masses of mammal species. How informative is that?!. **Hint:** use the `hist` function (or `geom_histogram()` if using `ggplot`).

7. Generate a histogram of log body mass. Is this more informative? Morphological data commonly need a log-transformation when analyzing.

8. Redo the previous histogram but use the breaks option to force a bin width of 2 units (i.e., generate breaks between 0 and 10 by 2 units). How much detail is lost? (note: if you used `log` rather than `log10` to create your variable of log body mass you will need to use breaks between 0 and 20). Redo the previous histogram but vary the bin width. Try a bin width of of 1; then try 0.5; and then 0.1. Which bin width is superior?

9. Redo the histogram, but display probability density instead of frequency. **Hint:** use the `freq=FALSE` option (or `geom_density()` if using `ggplot`).

10. How does the frequency distribution of log body mass depart from a normal distribution? Answer by visual examination of the histogram you just created. Now answer by examining a normal quantile plot instead. Which display is more informative? **Hint:** use the `qqnorm` function.

11. Redo the normal quantile plot but use the options `pch=16, cex=0.5` to use a smaller plotting symbol.

12. If time permits, redraw the histogram of log body mass and superimpose a normal density curve to assess help detect deviations from normality.

Visualizing associations between variables

1. Use a box plot to compare the distribution of body sizes (log scale most revealing) of mammals having different extinction status. Are extinct mammals similar to, larger than, or smaller than, extant mammals? (You may need to use the `cex.axis` option to shrink the labels so that they all fit on the graph). **Hint:** use the `boxplot` function (or `geom_boxplot()` if using `ggplot`).

2. Examine the previous box plot. How do the shapes of the body size distributions compare between extinct and extant mammals?

3. Redo the previous box plot but make box width proportional to the square root of sample size. Add a title to the plot.

4. Use the `tapply` command to calculate the median log body mass of each extinction-status group of mammals. Check that these are consistent with the box plot results.

5. Calculate the mean of log body mass of each extinction-status mammal group. Why is the mean log size of extant mammals larger than the median, but the mean log size for extinct mammals smaller than the median?

6. Create a two-way frequency table (contingency table) describing the frequencies of mammal species in different extinction status groups on different continents. Which continent has seen the most extinctions? Which continent has the greatest number of extinctions relative to the number of extant species? **Hint:** use the `table` function, but this time with two arguments, `table(____, ____)`.

7. Draw a mosaic plot illustrating the relative frequencies of mammal species in different extinction status groups on different continents. Try switching the order of the variables in the mosaicplot command to change the explanatory and response variable. Which continent has experienced the greatest number of extinctions relative to total numbers of species? (A mosaic plot is perhaps not ideal for these data because the frequencies are so small for some categories, such as "introduction". In this case R also has difficulties squeezing in the labels. Perhaps this is a case in which a table is superior to a graph. You could also try a stacked barplot.) **Hint:** use the `mosaicplot` and `barplot` functions.

Data set 2: Fly sex and longevity

The data are from L. Partridge and M. Farquhar (1981), Sexual activity and the lifespan of male fruitflies, *Nature* 294: 580-581. The experiment placed male fruit flies with varying numbers of previously-mated or virgin females to investigate how mating activity affects male lifespan. The data are in the file fruitflies.csv.

Visualizing associations between variables

Download the file to your computer and open in a spreadsheet program to have a look at it. Read the data file into a new data frame. Our goal here is to find a plot type that most clearly and efficiently visualizes the differences among treatment groups.

1. Use the `head` command to view the first few lines of the data frame on the screen, and familiarize yourself with the variable names.

2. Use a box plot to examine the distribution of longevities in the treatment groups. Add a label to the y-axis. Do the treatment groups differ in longevities? Describe the pattern of differences between treatments.

3. Use a strip chart (`stripchart` function) to examine the distribution of longevities in the treatment groups. Try the jitter method to reduce overlap between points. Adjust the treatment label sizes so that they all fit on the graph. **Hint:** try using the `method='jitter'` option (or the `geom_jitter()` if using `ggplot`). Compare with the box plot results. Which is more revealing?

4. The variable "thorax" stands for thorax length, which was used as a measure of body size. The measurement was included in case body size also affected longevity. Produce a scatter plot of thorax length and longevity. Make "longevity" the response variable (i.e., plot it on the vertical axis). Is there a relationship? **Hint:** use the `plot` function (or the `geom_point()` if using `ggplot`).

5. Use the `lowess` smoother to draw a smooth curve through the scatter plot of longevity on thorax length. **Hint:** use the `lines` function (or the `geom_smooth()` if using `ggplot`).

6. Redraw the scatter plot but this time use different symbols or colors for the different treatment groups. Add a legend to identify the symbols. Describe the pattern of differences between treatments. **Hint:** use the `legend` function.

Multipanel plots

The above graphs of the fly data indicate that it is a challenge to visualize relationships between longevity and size for different treatment groups all in a single frame. Alternatively, use the `layout` command to explore graphs having multiple panels, one for each group, all on the same scale. Might the results be more revealing about differences among groups than superimposing all points onto a single frame?

```
layout(matrix(1:3, ncol=3, nrow=1))
add.panel <- function(vals, colour) {
  plot(vals, xaxt='n', yaxt='n', las=1, pch=16, col=colour)
}
add.panel(vals=1:10, colour='red')
add.panel(vals=10:1, colour='blue')
add.panel(vals=c(rep(2,5), rep(3,5)), colour='green4')
```

```
ggplot(data = fruitflies, aes(x = ____, y = _____)) + geom___() +
facet_wrap(~treatment)
```

1. Plot a frequency distribution of male longevity for all treatment groups separately. In this plot, how easy is it to visualize differences among treatments in the distributions?

2. Repeat the previous command but stack the panels one above the other. Consider on how this affects your ability to compare the distributions among treatments compared with side-by-side plots.

3. Create a panel of scatter plots showing the relationship between male longevity and male size (as measured by thorax length) separately for each treatment group. Compare this with the previous exercise in which all points were placed on the same scatter plot with different symbols. Which is more revealing

4. Try making a multi-panel graph using an `lapply` statement. In other words, write a function that makes a single panel, and then use `lapply` to call that function multiple times.