Assignment 2: Data exploration

Biol 3550

by Levi Newediuk (based on work from Sean Johnson-Bice)

2023-09-11

Preamble

For this assignment we are using data from the paper:

Verbeke et al. (2023) The impact of plant diversity and vegetation composition on bumblebee colony fitness. *Oikos* e09790 doi: 10.1111/oik.09790

The authors placed colonies of lab-reared bumblebees at sites in the field to test how different variables associated with those sites affected the weight the colonies gained. They used sites in two different habitat types, grassland and heathland.

Your script visually explores changes in bumblebee colony weight gain according to the number of worker bees, plant species richness (the diversity of food options for the bees), and habitat type. We will be using barplots and scatterplots to visualize the variables and the relationships between them. We will also be calculating basic summary statistics like we did in Assignment 1.

To complete the assignment, you will be both copying chunks of code from this document and writing your own code into the template R script supplied for you.

Code chunks are provided in blocks that look like this:

```
code chunk
```

Your tasks in the assignment are indicated by font colours:

- Already-prepared code for you to copy or retype to your R script is indicated in red.
- Code you need to write or adjust yourself is indicated in orange.
- Outputs you need to copy to the word document to complete your assignment are indicated in blue.

You must submit **both your word document and completed R script to earn full marks on the assignment**.

Setting up your workspace

First, let's load the R packages we'll need for today's assignment. Copy the code below into your R script:

```
library(tidyverse)
library(ggpubr)
```

Next, let's load the bee data file you converted to a .csv file. Remember you can do this several different ways including:

- adding the csv file to your project and loading it directly from there
- setting your working directory to wherever you have stored the .csv file at on your computer.

If you've opted to set your working directory, change the path below to one that works and paste it into your R script:

```
setwd("User/levi/Documents/R-projects/Biol3350/")
```

Next, let's load our data:

```
bee_data <- read.csv('data/bee_data.csv', header = T)
```

Remember, you might not have a "data" folder in your project. You might not be using folders at all, or you might be using a different folder name. You'll need to adjust your path so R can find your data!

Next, let's rename some columns in our data frame. Rename the two variables 'species.richness' and 'number.of.bees'. You can use the names with underscores provided below or any other naming conventions that we learned in the first assignment, but remember you will need to refer to the exact variable names you chose when performing operations on the data frame (**hint**: you can always use the `head()` or `str()` functions to look at the variable names in your data frame.

```
bee_data <- rename(bee_data, 'species_richness' = `species.richness`)
bee_data <- rename(bee_data, 'n_bees' = `number.of.bees`)
```

Part A

Test the hypothesis that bumblebee colonies with more worker bees gain more weight.

This part of your assignment will be used to answer questions 1–4 in your word document by creating a series of plots.

Question 1: Which variable is the independent variable and which is the dependent variable? List the data type for each.

Complete this question in your word document. You do not need to include anything in your R script to answer this question.

Question 2: Make histograms to look at the distributions of each variable (with figure captions). Do they look normal?

Let's first make a histogram of the colony weight gain variable. Since the colony weights are listed twice (once for worker bees, once for queen bees) we first need to filter the data to include **only** the worker bee caste.

```
worker_bee_data <- bee_data %>%
  filter(caste == "worker")
```

Next, we need to use the **ggpubr** package to plot a histogram. We'll be using the `gghistogram()` function (remember last assignment we used `ggparplot()` function).

```
gghistogram(data = worker_bee_data,
            x = "weight",
            bins = 15,
            fill = "darkgray",
            title = "Histogram of colony weight gain")
```

You should recognize some of the arguments in this function from last week's assignment. Note that we only have an x variable and not a y variable here — we are only looking at the distribution of a single variable. We also see a new *"bins"* argument, which refers to the way R splits up the variable. In general, the more bins in your histogram, the fewer observations will land in each bin, and the "finer" your distribution will look. We also see a *"fill"* argument, which sets the colour of the bars in our histogram. Finally, we added a *"title"*. We won't normally add titles to figures, but we wanted to show you how to play around with the arguments.

Now let's change some of the attributes of this histogram and export the plot to a file using the code you learned in assignment 1. On your own, export a new histogram with a "lightblue" fill, exported in .png format, with a width of 800 and a height of 400. Paste this new histogram into your word document to answer the first part of question 2.

You will also need to create a histogram of worker bee counts using the worker bee data you already filtered. Look at the first histogram you made to help you make this one. You will need to use 15 bins with a "darkred" fill. Export this histogram in a .png format with a width of 800, height of 400, and a resolution of 100 ppi. Paste this histogram into your word document to answer the second part of question 2.

Question 3: Make a scatter plot testing your hypothesis. Make sure your dependent variable is on the y-axis and the independent variable is on the x-axis. Label your axes, including units. Add a trend line.

Let's make a new type of plot, a scatter plot, using the `ggscatter()` function.

```
weight_n_workers_scatter <- ggscatter(data = worker_bee_data,
          x="n_bees",
          y="weight",
          add="reg.line",
          xlab = "Number of worker bees",
          ylab="Colony weight gain (grams)")
```

Most of these arguments should look familiar other than "add='reg.line' ". Here, we are telling R to add a linear regression line to the plot. This line is evaluating the linear relationship between the number of worker bees and colony weight gain.

Export the scatter plot in any format of your choosing with a width of 600 and height of 500. Paste the scatterplot into your word document to answer question 3.

Question 4: Based on your graph, what do you conclude about the hypothesis? If you reject the hypothesis, suggest an alternative explanation.

Complete this question in your word document. You do not need to include anything in your R script to answer this question.

Part B

Test the hypothesis that bumblebee colonies placed in sites with greater plant species richness gain more weight. Evaluate grassland and heathland habitats separately.

Question 5: List the independent and dependent variables from the hypothesis, and the type of data for each.

Complete this question in your word document. You do not need to include anything in your R script to answer this question. (**Hint:** you should list your data types in a statistical sense, not in terms of R data types)

Question 6: What are the means and standard errors of colony weight gain for each of the variables, i.e., sites with 21-30 versus 31-40 species in each habitat category (grassland versus heathland)?

Recall that I showed you at least two different ways to calculate means and standard errors in Assignment 1. Here is one way we learned to calculate the mean colony weight gain and SE for the grassland habitat:

```
worker_bee_data %>%
  filter(habitat == 'grassland') %>%
  group_by(species_richness) %>%
  summarise(mean_grassland_weight = mean(weight),
            se_grassland_weight = sd(weight)/sqrt(length(species_richness)))
```

This method uses the `filter()` function to isolate only the grassland sites. Use the output of your code chunk above to answer the first part of question 6.

Write some code to calculate the mean colony weight gain and SE for the heathland habitat. Use the output to answer the second part of question 6.

Challenge question: This challenge question is worth 1 bonus point in your assignment. In the above code, we created two separate data frames for grassland and heathland habitats. Using the dplyr function 'group_by', create a single data frame that contains the mean & SE values for both ranges of species richness in both grassland and heathland.

Write code that performs the task described in the question. Paste the code you wrote into your word document.

Question 7: Test your hypothesis by making a bar graph. Be sure to label your axes, including units where necessary. Include error bars and a figure caption indicating what the error bars represent (use standard error).

We want to create a barplot that shows colony weights at sites with 21-30 versus 31-40 species. We are also going to need to distinguish grassland and heathland habitats so we can evaluate the hypothesis for each habitat. This can be done pretty simply using the `ggbarplot()` function. Refer to the help pages of ?ggbarplot to see how you can create a bar plot (mean and se bars) for multiple groups (habitats) with dodged bars, and a fill command based on treatment.

```
weight_richness_bar <- ggbarplot(bee_data,
          x="habitat",
          y="weight",
          fill="species_richness",
```

```
          add="mean_se",
          position=position_dodge(),
          ylab = "Colony weight gain (grams)",
          xlab="Species richness")
```

The code above should look familiar to you aside from a couple of new arguments. This time we are using *add="mean_se"* instead of *add='reg.line'*. This argument adds our error bars. We also included the argument *"position=position_dodge()"*, which tells R to leave some space between the bars.

Export this grouped barplot using any resolution and output format. Paste it in your word document to complete question 7.

Question 8: Based on your graph, what do you conclude about the hypothesis? Follow the statistical axiom, "If it looks different, it probably is!". Answer this question for both the grassland and heathland habitats separately. Propose a biological explanation for your conclusions. If you rejected the hypothesis for either or both of the habitats, propose an alternative explanation.

Complete this question in your word document. You do not need to include anything in your R script to answer this question.