Week 1 lab solutions

Kasia Banas

11/01/2022

Setting things up

In this self-guided lab, you will practice exploring data in R using visualisations, tables and descriptive statistics.

The data we will be looking at come from a randomised controlled trial testing whether the use of a licorice gargle before intubation for elective thoracic surgery reduces sore throat after the surgery. You can find more information about the dataset and a codebook under licorice_gargle here: https://cran.r-project.org/web/packages/medicaldata/medicaldata.pdf

Let's start by installing the medicaldata package and loading it into R

```
# install.packages("medicaldata")
# The install.packages command is commented out, as we only require the installation once.
library(medicaldata)
```

We'll be using functions from the tidyverse library, so let's load it in too.

```
library(tidyverse)
```

Now let's tell R that we'll be using the licorice_gargle data and let's have a quick look at the dataset. Note: I have hidden the output here, as it's taking up a lot of space.

```
data("licorice_gargle")
str(licorice_gargle)
head(licorice_gargle)
```

You'll notice that all variables are recorded as numeric, even though some of them are categorical and should be recorded as factors. Let's fix this now, so we have a dataset ready for analysis.

```
licorice_gargle_clean <- licorice_gargle %>%
  mutate(preOp_gender = factor(preOp_gender, levels = c(0, 1),
                               labels = c("Male", "Female")),
         preOp_smoking = factor(preOp_smoking, levels = 1:3,
                                labels = c("Current", "Past", "Never")),
         treat = factor(treat, levels = c(0, 1),
                        labels = c("sugar", "licorice")),
         intraOp_surgerySize = factor(intraOp_surgerySize, levels = 1:3,
                                      labels = c("Small", "Medium", "Large")),
         pacu30min_cough = factor(pacu30min_cough, levels = 0:3,
                                  labels = c("No cough", "Mild", "Moderate", "Severe")))
```

Please note: Pain was measured using an 11-point Likert scale, which simply means that patients were asked to rate their pain on a scale between 0 and 10. We will analyse data from Likert scales as numeric.

Exercise 1

Task

In describing an RCT, we often want to know the demographic and baseline characteristics of people in each group. This is to check whether the groups are similar (which is what we would expect following a random assignment procedure) Please check the baseline characteristics (gender, age, smoking status) of people in both groups. Do they look similar? HINT: For categorical variables, try using the table function. For numerical variables, first check the distribution (is it normal?), and then try the group_by %>% summarise pattern, calculating the appropriate statistics to describe the centre and spread.

Solution

To see the gender and smoking status distribution in the two groups, we use the table function. You'll need to specify the two variables you're interested in.

```
# The first table has treatment group and gender as its variables:
table(licorice_gargle_clean$treat, licorice_gargle_clean$preOp_gender)
```

```
##
##            Male Female
##   sugar      73     44
##   licorice   69     49
```

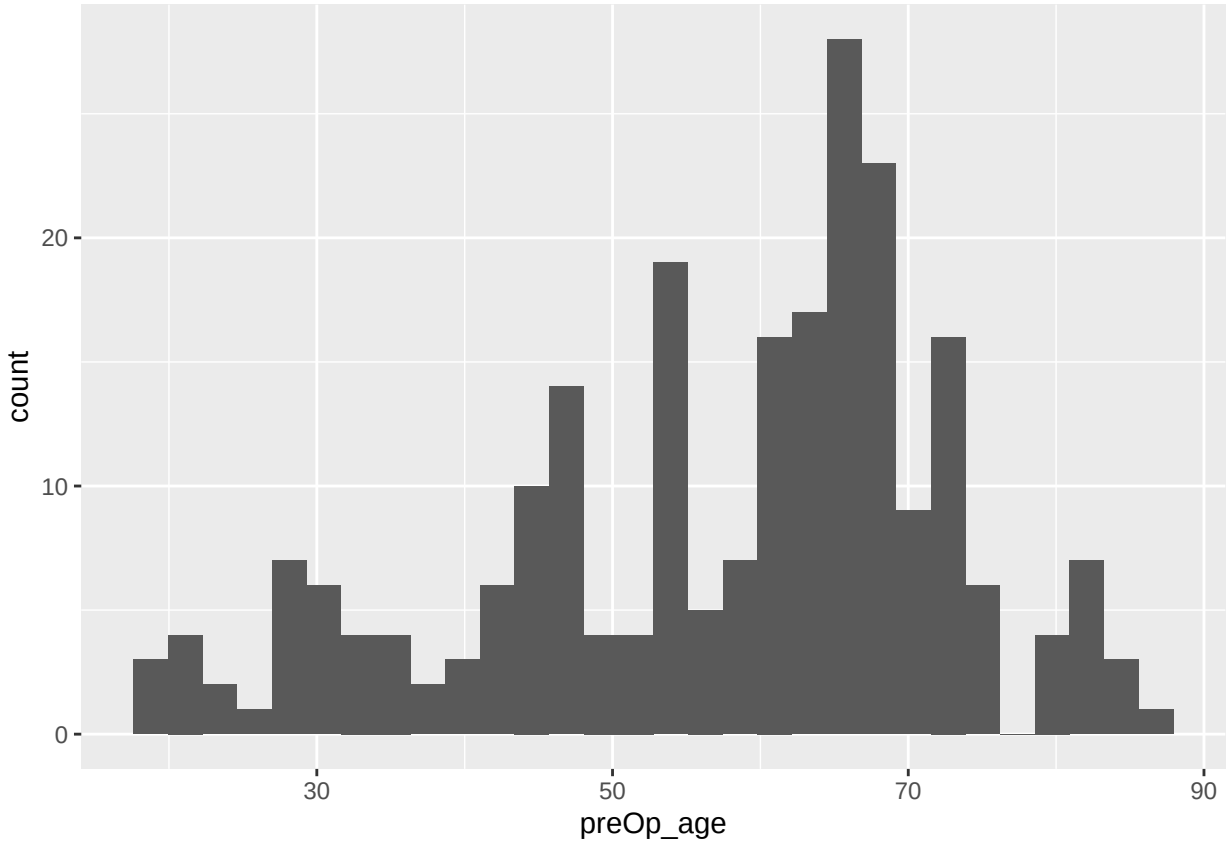
```
# and similar for the distribution of smoking status:
table(licorice_gargle_clean$treat, licorice_gargle_clean$preOp_smoking)
```

```
##
##            Current Past Never
##   sugar         45   36    36
##   licorice      45   36    37
```

Age is a numeric variable, so we'll be using descriptive statistics to look at differences between the groups. But, we need to check the distribution first, so we know which descriptive statistics will be appropriate. The mean and standard deviation can be used if the distribution is normal, and the median and interquartile range would be more appropriate for a skewed or otherwise non-normal distribution. In order to check the distribution visually, we can create a histogram:

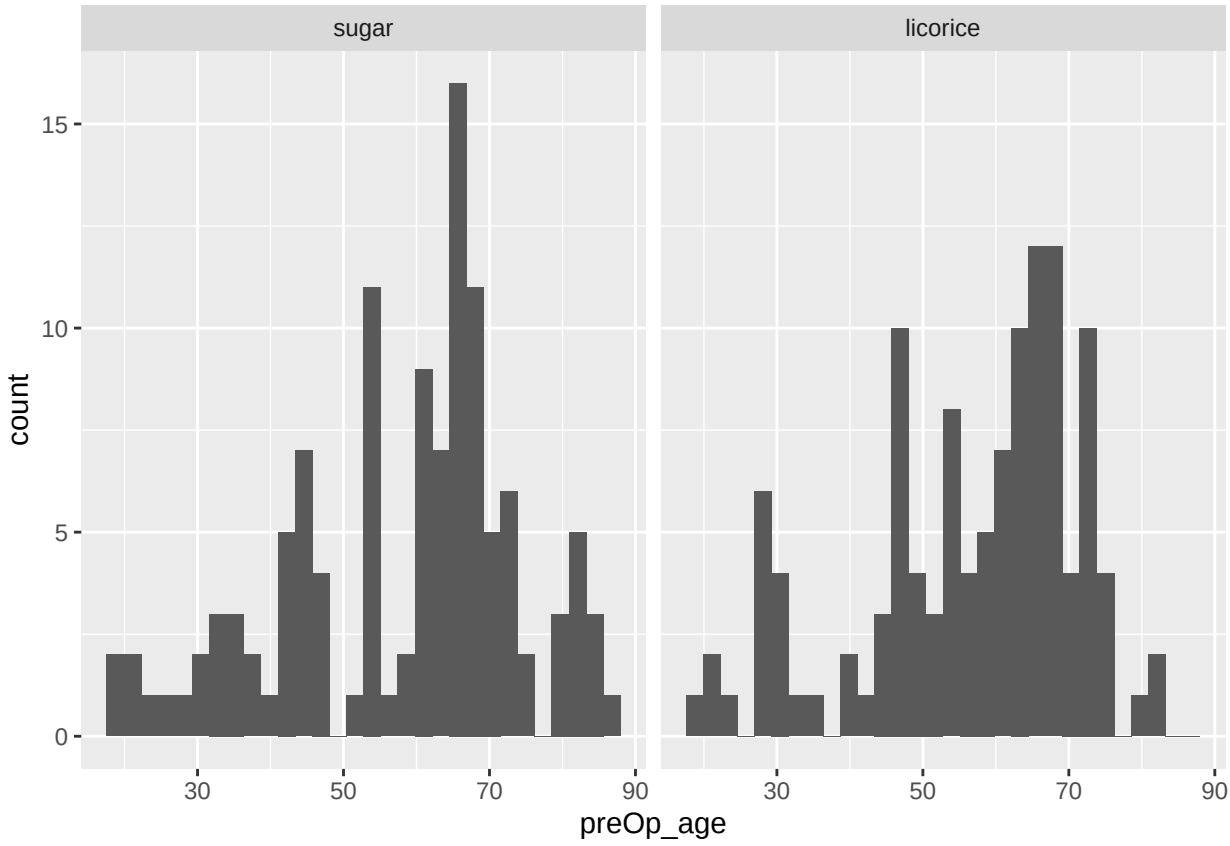
```
# We can start by plotting the entire sample
licorice_gargle_clean %>%
  ggplot(aes(x = preOp_age)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# And now we can plot the two treatment groups separately by using the facet_wrap function:
licorice_gargle_clean %>%
  ggplot(aes(x = preOp_age)) +
  geom_histogram() +
  facet_wrap(~treat)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



NOTE: When you fun this code, R gives you a warning: 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'. This means that it used the default value of the number of bins, which is 30. If you would like to change the number of bins, you can do that in the function call (see lecture videos and R scripts on how to do this).

It looks like the distribution is skewed to the left, indicating that there were more people in the older age range. Given the skew, it seems more sppropriate to use the median and IQR, rather than mean and standard deviation.

```
licorice_gargle_clean %>%
  group_by(treat) %>%
  summarise(median_age = median(preOp_age, na.rm = TRUE),
            IQR_age = IQR(preOp_age, na.rm = TRUE))
```

```
## # A tibble: 2 x 3
##   treat    median_age IQR_age
##   <fct>         <dbl>   <dbl>
## 1 sugar          63        23
## 2 licorice       60.5      20
```

Note that we have to specify that we'd like any missing values to be removed from the calculation (na.rm = TRUE). If we didn't do this, R would return missing values as a result, as we do have missing data for age.

Exercise 2

Task

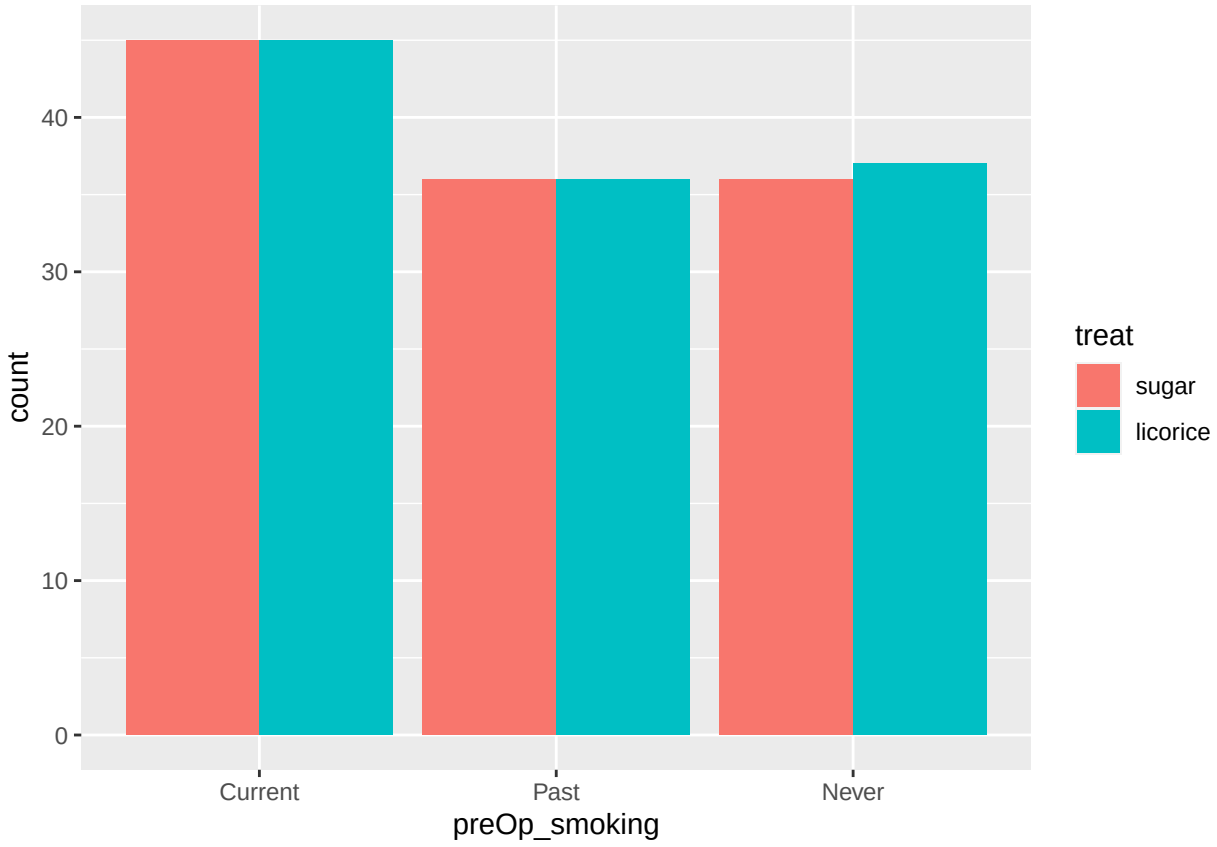
Produce a bar chart that shows the distribution of smoking status in each group. HINT: Use a grouped bar chart.

Solution

We'll start by creating a simple bar chart, and then we will customise it a little.

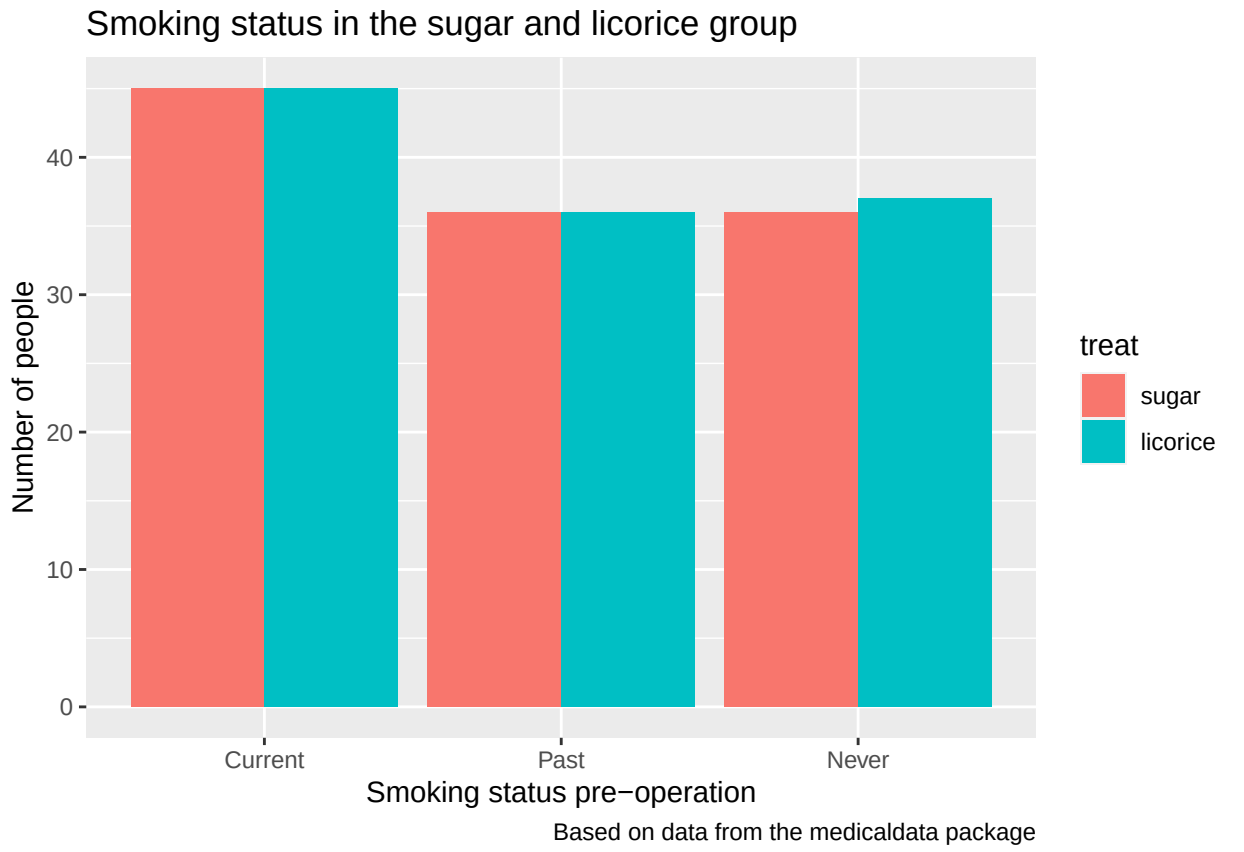
```
barchart1 <- licorice_gargle_clean %>%
  ggplot() +
  geom_bar(aes(x = preOp_smoking, fill = treat), position = "dodge")
# The position = "dodge" argument will ensure that the bars are not stacked
# on top of each other.
# While a stacked bar chart would also be an option here,
# a grouped one is easier to read, as all bars start at the 0.

# We've put the plot inside an object, so need to call the object's name now to view it:
barchart1
```




```
# We can make it nicer by adding custom labels and title
barchart1 +
  labs(title = "Smoking status in the sugar and licorice group",
```

```
       x = "Smoking status pre-operation",
       y = "Number of people",
       caption = "Based on data from the medicaldata package")
```




```
# You can experiment with the different arguments inside the labs function -
# see the function help page to see what other labels you could add.
```

Exercise 3

Task

Does it look like the licorice gargle reduced throat pain 30 minutes after the surgery? And what about cough? Use a graph and a table to answer each question.

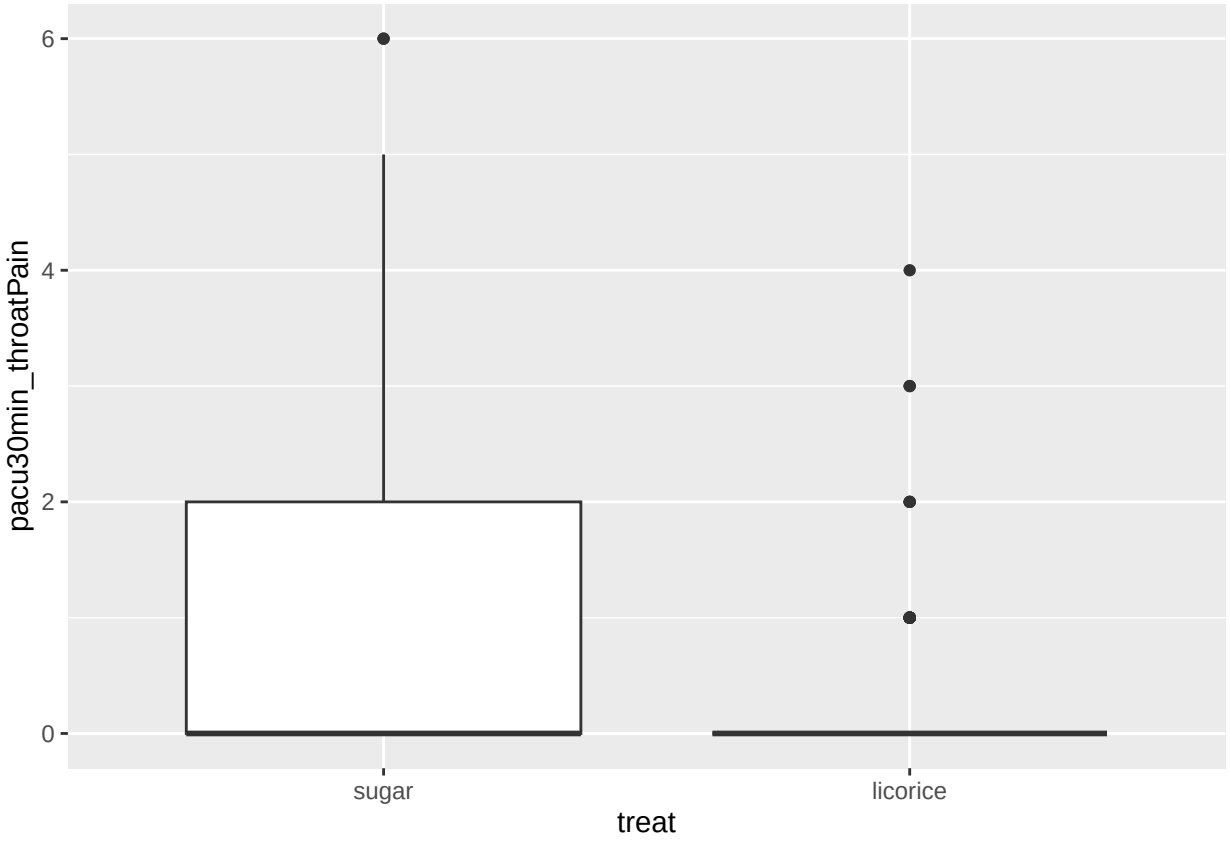
Solution

The first thing to do is decide which variable names correspond to the variables we're interested here. From the codebook, we can see that the two key variables are pacu30min_throatPain (throat pain at 30 minutes) and pacu30min_cough (cough severity after 30 minutes).

Let's look at the pacu30min_throatPain variable first. We can create a box plot to see the distribution for our two groups:

```
licorice_gargle_clean %>%
  ggplot(aes(x = treat, y = pacu30min_throatPain)) +
  geom_boxplot()
```

```
## Warning: Removed 2 rows containing non-finite values (‘stat_boxplot()‘).
```




```
# Remember the convention for vertical box plots -
# this means that the outcome variable is on the y axis,
# and the grouping variable on the x axis.
```

The box plot has highlighted very interesting things about the data. First of all, the median seems the same in both groups, and it is equal to 0.

In the licorice group, the interquartile range (IQR) is equal to 0 as well, indicating very little variability in the data. It seems that everyone gave the score of 0, apart from 4 patients, who are marked here as outliers. Remember from the videos that an outlier is a value that appears extreme relative to the rest of the data. In this case, with most patients reporting a score of 0, every value that's different from 0 appears extreme.

In the sugar group, the interquartile range only goes up from 0 (as you could not have a negative pain score), and the whisker goes up to 5. We have one outlier with a value of 6.

Overall, this is definitely not a normal distribution - it's not symmetrical around the median. If we looked at the median as an indicator of intervention effectiveness, we would conclude that there was no difference between the two groups on the pain score after 30 minutes.

The exercise is asking us to create a table as well, so let's create one that includes 4 descriptive statistics: median, interquartile range, mean and standard deviation. We can use this table to reflect on how the two sets of statistics provide a different insight into the data:

```
licorice_gargle_clean %>%
  group_by(treat) %>%
  summarise(mean_pain = mean(pacu30min_throatPain, na.rm = TRUE),
            sd_pain = sd(pacu30min_throatPain, na.rm = TRUE),
            median_pain = median(pacu30min_throatPain, na.rm = TRUE),
            iqr_pain = IQR(pacu30min_throatPain, na.rm = TRUE))
```

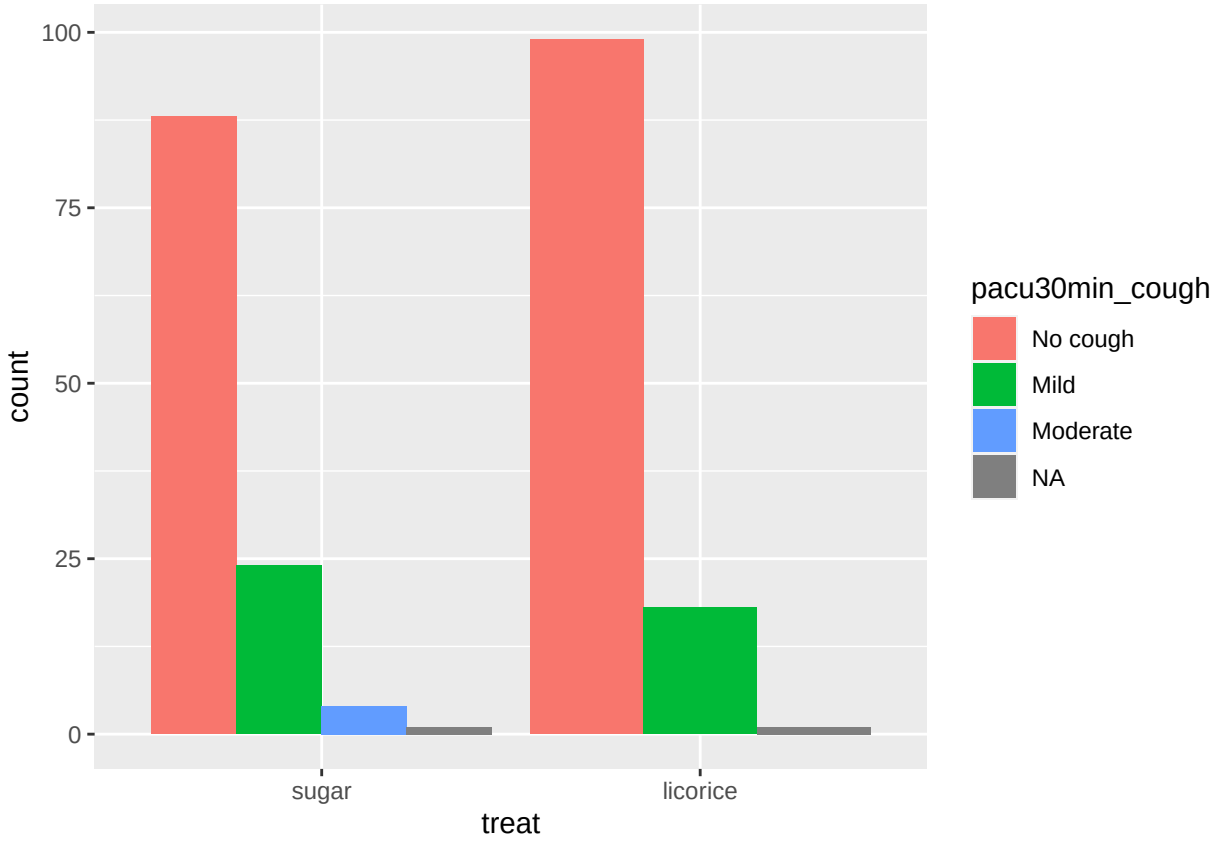
```
## # A tibble: 2 x 5
##   treat    mean_pain sd_pain median_pain iqr_pain
##   <fct>        <dbl>   <dbl>       <dbl>    <dbl>
## 1 sugar        1.03    1.55            0        2
## 2 licorice     0.274   0.678           0        0
```

The median and IQR columns confirm what we saw in the box plot - the median in both groups is equal to 0, and the IQR in the licorice group is also equal to 0. The means of the two groups are somewhat different (by 0.76 on an 11-point scale), and so is the standard deviation (SD is about twice as large in the sugar group).

If you only looked at the means of the two groups, you might conclude that they are different enough to warrant a conclusion that the licorice gargle was effective (and you will learn later about statistical tests that would let you test this conclusion formally). However, inspection of the median and IQR shows that most patients in both groups reported a pain score of 0, and it looks like the licorice gargle helped shift the scores above 0 further down, and resulted in even more people reporting a score of 0. So, while the median has not shifted, the range of the data has changed significantly, becoming much narrower in the licorice group. In other words, the effect of the intervention seems to lie not so much in shifting the centre of the distribution (i.e. the mean or median), but rather in reducing the spread. This is a good outcome - even if most people already reported no pain, we are still interested in reducing pain for those who did experience it.

Now, let's look at cough, which is a categorical variable. Starting with a graph - a grouped bar chart will be useful here:

```
licorice_gargle_clean %>%
  ggplot(aes(x = treat, fill = pacu30min_cough)) +
  geom_bar(position = "dodge")
```



Notice how in both groups, most patients did not have any cough. But, there does seem to be a little bit of a difference in distribution - in the licorice group no one had moderate cough, while in the sugar group a few people did. Looking at the mild cough category, almost 25 people in the sugar group had it, compared to around 20 in the licorice group (it's difficult to read these numbers off a graph, but we will see them more clearly in a table).

Now let's look at this distribution in a table:

```
table(licorice_gargle_clean$treat, licorice_gargle_clean$pacu30min_cough)
```

```
##
##            No cough Mild Moderate Severe
##   sugar          88   24        4      0
##   licorice       99   18        0      0
```

The findings are very much the same as in the graph (we're describing the same data), but we can see the numbers more accurately.

Exercise 4

Task

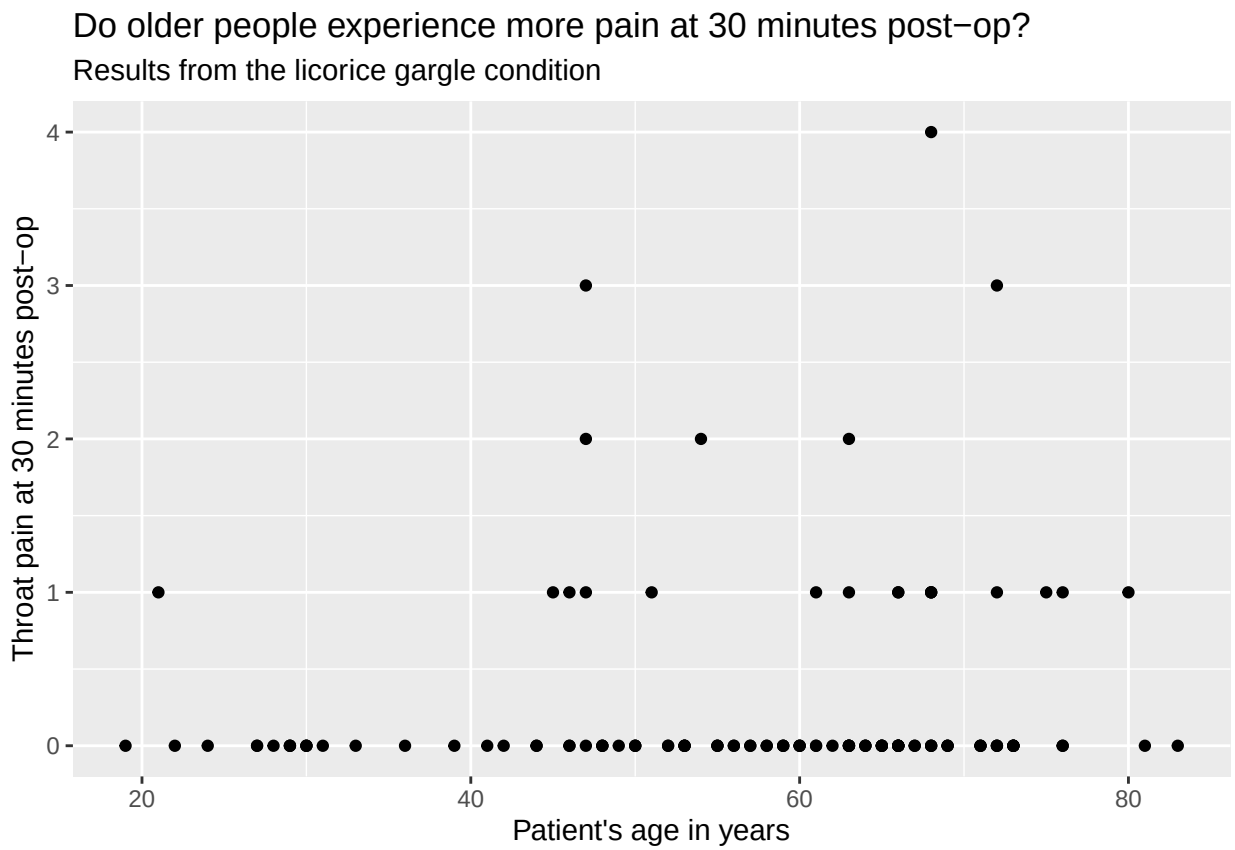
Is there any evidence that in the intervention (licorice gargle) group, older people experienced more post-operative throat pain? Create a scatterplot to illustrate the relationship between age and throat pain at 30 minutes after the surgery, and comment on your findings. Remember to filter the data, so you only look at the licorice condition.

Solution

The task suggests that we should filter our data first (to only include the licorice condition), and then create a scatterplot of post-operative pain at 30 minutes plotted against age. We can do these two things in one code chunk:

```
licorice_gargle_clean %>%
  filter(treat == "licorice") %>%
  ggplot() +
  geom_point(aes(x = preOp_age, y = pacu30min_throatPain)) +
  labs(title = "Do older people experience more pain at 30 minutes post-op?",
       subtitle = "Results from the licorice gargle condition",
       x = "Patient's age in years",
       y = "Throat pain at 30 minutes post-op")
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



Before we think about the relationship between age and pain, it's good to spend a few seconds reflecting on the distribution of the two variables. For pain, we have seen in Exercise 3 that most people reported no pain (0 on the scale), and this graph shows this too - most dots lie on the x-axis, which indicates the score of 0. For age, we see dots on the full range of the x-axis, with a few more on the right hand side of the graph - this indicates the slight skew that we saw in Exercise 1 - there are somewhat more older patients, compared to younger patients.

Now, in terms of a relationship, it does not seem like older patients reported more pain. Most pain scores were at 0 regardless of age. And if you imagine a straight line fitted to the data points, it would have to be

on the x-axis, i.e. to lie flat on the score of 0, with the non-zero data points located quite far away from the line. This indicates that there is hardly any relationship between the two variables.

Final note

I hope you have found this lab useful in demonstrating that you should always think about the distribution of your data first, and that this is often best done by creating a plot. By now, you should be comfortable creating histograms, bar charts, box plots and scatter plots - these are all extremely useful in data science. You have also practiced making simple tables using the table function and the group_by %>% summarise pattern.

As you may have noticed in the documentation of the medicaldata library, the data that we used in this lab come from a real clinical trial. If you're interested in reading more about it, this is the paper where it was reported: Ruetzler K, Fleck M, Nabecker S, Pinter K, Landskron G, Lassnigg A, You J, Sessler DI. A randomized, double-blind comparison of licorice versus sugar-water gargle for prevention of postoperative sore throat and postextubation coughing. Anesthesia & Analgesia. 2013 Sep 1;117(3):614-21.