BDSI R workshop – Learning from data

Robert Cope & Terry Neeman

2022

This document is an ‘R markdown’ document. You write normal text in the white spaces and R code in the grey ‘chunks’. To create a new chunk you can either type the chunk code manually or use the insert button in the middle of the top of this section of Rstudio.

## Data reorganisation

In this workshop we will look at some data from the following paper:

Mantella, N. M., & Youngentob, S. L. (2014). Prenatal alcohol exposure increases postnatal acceptability of nicotine odor and taste in adolescent rats. PloS one, 9(7), e102255.

In this study, rat dams were randomised to either alcohol exposure or an ad-lib alcohol-free diet. Their litters were then the test subjects. Each baby rat was offered water with a range of concentration of nicotine or sucrose. The sucrose was a control substance; no group differences were expected.

The research hypothesis is that rats exposed to alcohol prenatally will be more attracted to nicotine flavour compared with control rats. There are two response variables in this experiment: latency until first lick and average licks per period.

In the first part, we explore the response - average licks.

These data are organised in a quite reasonable way – they did four experiments and each gets its own sheet of the spreadsheet, and then within each sheet each row corresponds to an individual rat (measured at different stimulus concentrations). This is a little more complex than the csv style data that we have previously provided (and that statistics).

We will start by just looking at the third sheet. We need to load the tidyverse, janitor, and readxl packages.

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.1 ✔ stringr 1.4.1   
## ✔ readr 2.1.2 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(janitor)

##   
## Attaching package: 'janitor'  
##   
## The following objects are masked from 'package:stats':  
##   
## chisq.test, fisher.test

library(readxl)  
library(gtsummary)

Now we read in the data, specifying the sheet, and then check its structure.

nicotine\_avg <- read\_xlsx('Brief Access Lick Data.xlsx',  
 sheet = 'Nicotine Average Lick Data')  
glimpse(nicotine\_avg)

What do you notice about this?

The column names are going to be a problem here – we will start by using clean\_names from the janitor package to automatically help:

nicotine\_avg <- clean\_names(nicotine\_avg)  
glimpse(nicotine\_avg)

Your task as a group is to explore the data and develop an initial understanding of what is going on in this experiment – perhaps like if you were a reviewer or statistical consultant.

First, use what you learned last time to get a sense of the makeup of this dataset. How many rats in each treatment group? For each treatment group, provide summary statistics for weight and sex? Are the groups roughly balanced?

nicotine\_avg %>%  
 group\_by(sex,prenatal\_exposure) %>%  
 summarise(count = n(), mean\_wt = mean(weight), sd\_weight = sd(weight))

nicotine\_avg %>%  
 select(prenatal\_exposure, sex, weight) %>%  
tbl\_summary(by = prenatal\_exposure)

Average licks are recorded across 6 columns. In order to explore the patterns visually, we should restructure the data so that the response variable is in a single column.

Restructure data

nicotine\_long <- nicotine\_avg %>%  
 pivot\_longer( cols = starts\_with("average\_licks"),   
 names\_to = "Concentration",  
 values\_to = "Licks",  
 names\_prefix = "average\_licks\_")

Create a plot that allows you to compare average licks by Treatment over concentration. In this plot, connect the values with the same animal\_id. You may need to reset the levels of Concentration.

nicotine\_long <- mutate(nicotine\_long, Concentration = factor(Concentration,   
 levels = c("water",   
 "0\_1m\_m",  
 "0\_3m\_m",   
 "1\_0m\_m",  
 "3\_0m\_m",  
 "6\_0m\_m")))  
  
ggplot(nicotine\_long, aes(x=Concentration,y=Licks, col=prenatal\_exposure))+  
 geom\_point()+  
 geom\_line(aes(group = animal\_id))

Another way of looking for potential patterns is using geom\_smooth(). In order to do this, Concentration should be numeric (not a factor). Create a new variable Concentration2 that is a numeric field.

nicotine\_long$Concentration2 <- factor(nicotine\_long$Concentration,   
 labels = c("0","0.1","0.3","1.0","3.0","6.0")) %>%  
 as.character()%>%  
 as.numeric()  
  
ggplot(nicotine\_long, aes(x=Concentration2,y=Licks, col=prenatal\_exposure))+  
 geom\_point()+  
 geom\_smooth()

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'

We can also put concentration (x-axis) on the log scale:

ggplot(nicotine\_long, aes(x=Concentration2,y=Licks, col=prenatal\_exposure))+  
 geom\_point()+  
 scale\_x\_log10()+  
 geom\_smooth()

How would you interpret these data? What can you say about the association between pre-natal alcohol exposure and preference for nicotine-flavoured water?

## Exercise: Repeat for the latency endpoint.

nicotine\_lat <- read\_csv("Brief access lick data latency.csv") %>%  
 clean\_names()

glimpse(nicotine\_lat)

How many rats in each treatment group? For each treatment group, provide summary statistics for weight and sex? Are the groups roughly balanced?

nicotine\_lat %>%  
 group\_by(sex,prenatal\_exposure) %>%  
 summarise(count = n(), mean\_wt = mean(weight), sd\_weight = sd(weight))

## 3 male alcohol 10 72.5 5.08  
## 4 male free choice liquid diet 9 78 5.83

nicotine\_lat %>%  
 select(prenatal\_exposure, sex, weight) %>%  
tbl\_summary(by = prenatal\_exposure)

## Table printed with {flextable}, not {gt}. Learn why at  
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html  
## To suppress this message, include `message = FALSE` in the code chunk header.

| Characteristic | alcohol, N = 191 | free choice liquid diet, N = 191 |
| --- | --- | --- |
| sex |  |  |
| female | 9 (47%) | 10 (53%) |
| male | 10 (53%) | 9 (47%) |
| weight | 71.0 (67.0, 74.0) | 75.0 (70.0, 77.5) |
| 1n (%); Median (IQR) | | |

Restructure data and create a plot to explore average latency across a range of concentrations. Choose a different colour for each group.

nicotine\_long <- nicotine\_lat %>%  
 pivot\_longer( cols = starts\_with("average\_latency"),   
 names\_to = "Concentration",  
 values\_to = "Latency",  
 names\_prefix = "average\_latency\_to\_first\_lick\_sec\_") %>%  
 mutate(Concentration = factor(Concentration,   
 levels = c("water",   
 "0\_1m\_m",  
 "0\_3m\_m",   
 "1\_0m\_m",  
 "3\_0m\_m",  
 "6\_0m\_m")))  
  
ggplot(nicotine\_long, aes(x=Concentration,y=Latency, col=prenatal\_exposure))+  
 geom\_point()+  
 geom\_line(aes(group = animal\_id))

Also try exploring patterns using geom\_smooth(). Why might a pattern be more evidence with geom\_smooth. Try also making concentration numeric and log-transform the x-axis. Interpret these plots wrt the research question.

nicotine\_long$Concentration2 <- factor(nicotine\_long$Concentration,   
 labels = c("0","0.1","0.3","1.0","3.0","6.0")) %>%  
 as.character()%>%  
 as.numeric()  
  
ggplot(nicotine\_long, aes(x=Concentration2,y=Latency, col=prenatal\_exposure))+  
 geom\_point()+  
 geom\_smooth()

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'

ggplot(nicotine\_long, aes(x=Concentration2,y=Latency, col=prenatal\_exposure))+  
 geom\_point()+  
 scale\_x\_log10()+  
 geom\_smooth()

## Warning: Transformation introduced infinite values in continuous x-axis  
## Transformation introduced infinite values in continuous x-axis

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 38 rows containing non-finite values (stat\_smooth).