STEPHAN PEISCHL

STATISTIK FUER BIOLOGIE FS 2023

Solutions

Solutions to the exercises will be posted here each week.

Back to main page. Back to exercises.

Week 1

Exercise 1

a) No. Collectors did not choose randomly, but preferred rarer types over more common ones.

b) It is a sample of convenience.

c) There is bias in every year's sample because rare types are over-represerented. Further this bias might change across years as the frequencies of the two morphs have changed over time.

Exercise 2

a) Discrete.

b) Technically it is a discrete variable (because fractions can be enumerated / are restricted to finitely many values in a finite sample) but it makes sense to treat it as a continuous variable if the sample is very large.

c) Discrete. There are no half crimes.

d) Continuous. Body mass is continuous and hence the log as well.

Exercise 3

a) E(xplanatory): Altitude (categorical: high vs low) R(esponse): Growth rate S(tudy type): Observational

b) E: Treatment (standard vs. tasploglutide) R: Rate of insulin release S: Experimental

c) E: Health status (schizophrenia vs. healthy) R: Frequency of illegal drug use S: Observational

d) E:Number of legs R: Survival propability S: Experimental

e) E: Treatment (advanced communication therapy vs. social visits without formal therapy) R: Communication ability S: Experimental

Exercise 4

The main problem is a strong bias in the sample. To see this, consider the sample of planes that were used in this study. Only planes that were *not* hit in critical areas were available to estimate the distribution of bullet holes. Planes that were hit in a critical area, i.e., one that leads to a crash, were not available because they did not return to base. With this knowledge, it becomes clear that it would have been better to reinforce the areas were no, or very little, bullet holes were found, namely the cockpit and engine.

Exercise 5

a) The population parameter being estimated is all the small mammals of Kruger National Park.

b) No, the sample is not likely to be random. In a random sample, every individual has the same chance of being selected. But some small mammals might be easier to trap than others (for example, trapping only at night might miss all the mammals active only in the day time). In a random sample individuals are selected independently. Multiple animals caught in the same trap might not be independent if they are related or live near one another (this is harder to judge).

c) The number of species in the sample might underestimate the number in the Park if sampling was not random (e.g., if daytime mammals were missed), or if rare species happened to avoid capture. Even if the sample is perfectly random, the estimator will underestimated the true number in most cases. You can sample fewer species just by chance, but not more species as there actually are. Thus - on average - you will underestimate the true number.

Week 2

Exercise 6

a) You should see a blank plot. This is what ggplot does, it simply creates a "canvas" to which you can add "layers".

b) How many rows are in mpg? How many columns?

```
str(mpg)
```

```
## tibble [234 x 11] (S3: tbl_df/tbl/data.frame)
##  $ manufacturer: chr [1:234] "audi" "audi" "audi" "audi" ...
##  $ model       : chr [1:234] "a4" "a4" "a4" "a4" ...
##  $ displ       : num [1:234] 1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int [1:234] 1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int [1:234] 4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr [1:234] "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr [1:234] "f" "f" "f" "f" ...
##  $ cty         : int [1:234] 18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int [1:234] 29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr [1:234] "p" "p" "p" "p" ...
##  $ class       : chr [1:234] "compact" "compact" "compact" "compact" ...
```

There are 234 observations of 11 variables. thus, there are 11 columns and 234 rows. we can confirm this by looking at the dimension of the underlying data matrix:

```
dim(mpg)
```

```
## [1] 234  11
```

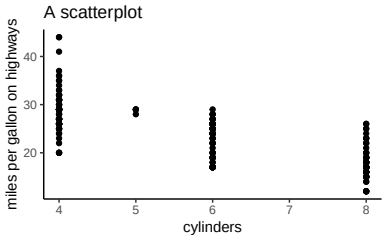
c)

```
?mpg
```

drv is a categorical vaiable indicating the drive type: f = front-wheel drive, r = rear wheel drive, 4 = 4wd

d)

```
ggplot(data  = mpg) +
  geom_point(mapping = aes(y = hwy,x = cyl)) +
  theme_classic() +
  labs(title = "A scatterplot", y = "miles per gallon on highways"
```



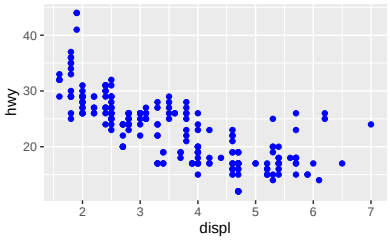
e)

```
ggplot(data  = mpg) +
  geom_point(mapping = aes(x = class,y = drv)) +
  theme_classic() +
  labs(title = "A useless plot", x = "class",y="drive type")
```

both drv and class are categorical variable and it does not make sense to visualize them this way. What would be a better way to show these data?

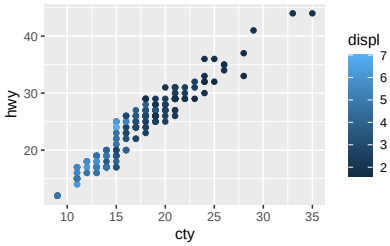
f) Color is used wihtin the aesthetics function, where we specifiy which varibales should be used. "blue" is however not a variable in our data frame. Thus it is ignored. The follwoing code creates the correct plot:

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy), color = "blue")
```



g) A continuous variable will lead ot a continuous color gradient rather than a discrete set of colors.

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = cty, y = hwy, color = displ))
```



h) The "+" symbol always has to be in the top row:

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy))
```