Paper

Andrei Nistor

#1. North America Rodents. ##Dataset data/surveys.csv -rodents sightings in North America from 1977 to 2002 ##Dataset data/species.csv -species acronyms and their Genus

## a) Join the two datasets

surveys <- readr::read\_csv('data/surveys.csv')

## New names:  
## \* `` -> ...1

## Rows: 30738 Columns: 10

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (2): species\_id, sex  
## dbl (8): ...1, record\_id, month, day, year, plot\_id, hindfoot\_length, weight

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

spec(surveys)

## cols(  
## ...1 = col\_double(),  
## record\_id = col\_double(),  
## month = col\_double(),  
## day = col\_double(),  
## year = col\_double(),  
## plot\_id = col\_double(),  
## species\_id = col\_character(),  
## sex = col\_character(),  
## hindfoot\_length = col\_double(),  
## weight = col\_double()  
## )

species <- readr::read\_csv('data/species.csv')

## Rows: 54 Columns: 4

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (4): species\_id, genus, species, taxa

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

spec(species)

## cols(  
## species\_id = col\_character(),  
## genus = col\_character(),  
## species = col\_character(),  
## taxa = col\_character()  
## )

rodents <- surveys %>% left\_join(species, by = c('species\_id' = 'species\_id'))  
spec(rodents)

## cols(  
## ...1 = col\_double(),  
## record\_id = col\_double(),  
## month = col\_double(),  
## day = col\_double(),  
## year = col\_double(),  
## plot\_id = col\_double(),  
## species\_id = col\_character(),  
## sex = col\_character(),  
## hindfoot\_length = col\_double(),  
## weight = col\_double()  
## )

1. Present the 5 rodent species having the highest mean weight in a table showing species, mean weight and mean hindfoot length as in the example below.

mean\_weight <-mean(surveys$`weight`)   
 mean\_length <-mean(surveys$`hindfoot\_length`)

rodents %>%   
 group\_by(species) %>%   
 summarize("Mean Weight [g]" = mean(weight),"Mean Hindfoot Length [mm]"=mean(hindfoot\_length)) %>%   
 top\_n(5, `Mean Weight [g]`) %>%  
 arrange(desc(`Mean Weight [g]`)) %>%  
 knitr::kable()

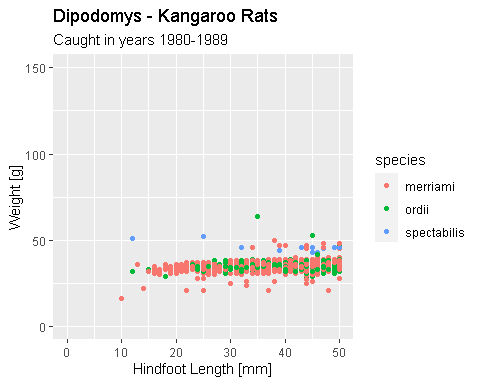
| species | Mean Weight [g] | Mean Hindfoot Length [mm] |
| --- | --- | --- |
| albigula | 158.72371 | 32.25048 |
| spectabilis | 120.21668 | 49.99260 |
| hispidus | 64.84906 | 28.05031 |
| fulviventer | 59.12500 | 26.70000 |
| ochrognathus | 55.37500 | 25.60000 |

## c) Recreate the plot.

It seems like the plot is a scatter type.

bigrodents<- filter(rodents,genus=='Dipodomys')   
 ggplot( data = bigrodents) +  
 geom\_point(mapping=aes( x = weight, y = hindfoot\_length, color = species)) +  
 labs(title = 'Dipodomys − Kangaroo Rats',  
 subtitle ='Caught in years 1980−1989',  
 x = 'Hindfoot Length [mm]',  
 y = 'Weight [g]')+  
 xlim(0,50) +   
 ylim(0,150)

## Warning: Removed 4346 rows containing missing values (geom\_point).



1. Describe the plot. First of all,the plot illustrates the size of the rodents from the genus Dipodomys, specifically their weight and hindfoot length. The first thing I noticed when analyzing the dataset was that those rodents are mainly the ones that are bigger in size than rodents from other genus, with spectabilis being the heaviest reaching even above 150g, at almost 200g, and with a hindfoot length close to 60 mm. Furthermore, this result gives an insight into why they were named like that ,thanks to their spectacular sizes. Although, the ordii and the merriami are smaller sometimes they can reach a hindfoot length closer to the spectabilis even though they might be 3 times lighter in weight. For example, in the dataset we can see the longest hindfoot recorded was that of an ordii with 64 mm. Based on my assesment of the dataset, I can confidently say that the Dipodomys are mainly the bigger rodents , along with albigula from the Neotoma genus, which seems to be the heaviest rodent, but lacks the hindfoot length of the Dipodomys genus. Thus, from our dataset we can conclude that the Dipodomys are the rodents with the longest hindfoot length.
2. Kangaroo Rats (genus Dipodomys) are small rodents moving similarily to kangaroos using jumping.steps. Is the mean male hindfoot length different between ordii and merriami species? Comment on the results

Analyzing the dataset , I noticed that the biggest hind foot recorder was by a female ordii with 64mm, which is very impressive compared to her lighter build of 35g . It seems like the male merriami does not go above 40 mm in hind foot length, while the ordii seem to vary more , with the highest hind foot recorder by a male being 58 mm.

#2. Medical students smoking habits

A study was conducted on various Medical Universities within Germany and Hungary. The students were asked about their smoking habits. 2883 students took part, 44% of them were German, 36% were Hungarian and 20% other nationalities. The table below lists number of students per nationality declaring daily smoking habit.

##a) Are the proportions of nationalities within smoking students a true representation of proportions of the whole student body? Conduct a suitable test to check this hypothesis

-2883 students took part -44% of them were German, 36% were Hungarian and 20% other nationalities. -91 germans,78 hungarians, 51 multinational

smoking <- tribble(~'Nationality', ~'n',   
'German', 91,  
'Hungarian', 78,  
'Multinational', 51)  
print(smoking)

## # A tibble: 3 x 2  
## Nationality n  
## <chr> <dbl>  
## 1 German 91  
## 2 Hungarian 78  
## 3 Multinational 51

### I will start by checking if i can apply the central limit theorem:

-1st condition: if the data is independent.We have to assume the data is independent. -2st condition : check if sample size larger than 30. It is not so we cannot do CLT.

Making the hypothesis with the help of null hypothesis: H0: There is no relation between the percentage of smoking students and the percentage of nationalities within the student body. HA: : There is an accurate representation of the relation between the smoking students and the percentage of nationalities within the student body. This proves that smoking habits can be influenced by one’s nationality.

Now we have to choose an alpha significance value. I will choose alpha signivicant value= 0.05 We can now run the function for the Chi-squared test.

smoking %>% select(-1) %>%  
chisq.test(.)

##   
## Chi-squared test for given probabilities  
##   
## data: .  
## X-squared = 11.355, df = 2, p-value = 0.003423

Since the result states the p value is lower than the significant alpha value I picked, I can reject the null hypothesis in favor of the alternative. This means that the table is an accurate representation of the relation between the smoking students and the percentage of nationalities within the student body.

#3. University salaries -yearly salaries of random 52 academic workers at one of the U.S. Universities. -friend has been working there for the past 10 years. He achieved his doctorate 12 years ago, and is an associate professor in the Geology Department. -He earns $30.000 a year. Is it a fair salary or not?

salary <- readr::read\_csv('data/salaries.csv')

## Rows: 52 Columns: 6

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (3): sx, rk, dg  
## dbl (3): yr, yd, sl

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

spec(salary)

## cols(  
## sx = col\_character(),  
## rk = col\_character(),  
## yr = col\_double(),  
## dg = col\_character(),  
## yd = col\_double(),  
## sl = col\_double()  
## )

## a) Create a model to predict salaries at this University. Tune it, so that it is most statisticaly significant

I will create a linear model to predict ‘Average Salary’ using the variables in the dataset.

average\_salary <- read\_delim('data/salaries.csv', delim = ',')

## Rows: 52 Columns: 6

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (3): sx, rk, dg  
## dbl (3): yr, yd, sl

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

Since there is data about the salaries in the company and we know the amount of workers(52) I came up with the following formula to find out the average income.

workers<-52

average\_salary=total\_salaries÷workers We calculate the total budget of salaries

(total\_salaries <- sum(salary$sl))

## [1] 1237478

(average\_salary<- total\_salaries / workers)

## [1] 23797.65

The conclusion is that he earns more than average.