2022 01 03 VB-STA5 Exam in Statistics

Monday 3rd of January.

The exam set consists of 3 main exercises with 9 sub exercises in total.

Each sub exercise is weighted equally when grading the hand-ins.

1. North America Rodents.

Dataset *data/surveys.csv* contains information about rodents sightings in North America from 1977 to 2002. Dataset *data/species.csv* contains information about species acronyms and their Genus.

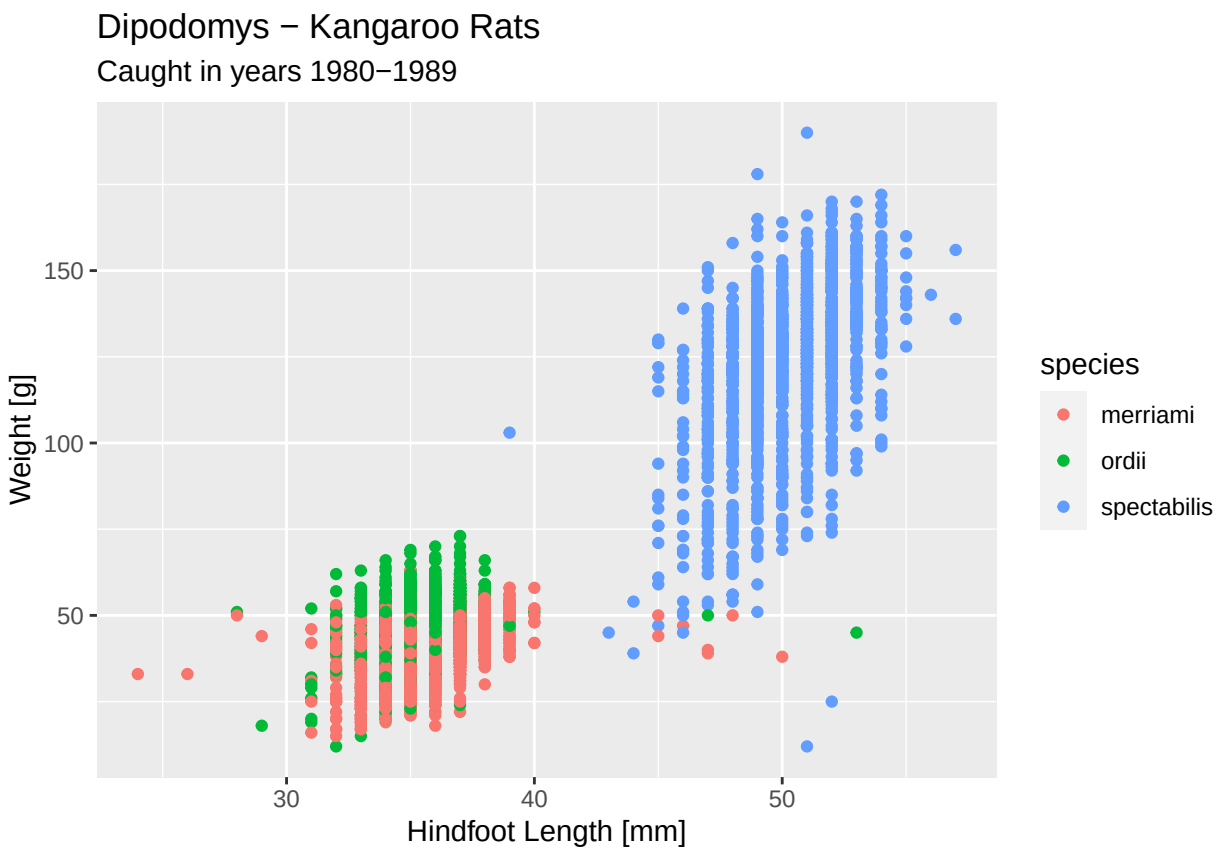
a) Join the two datasets.

b) 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.

The example shows 5 species with the shortest mean hindfoot length.

species	Mean Weight [g]	Mean Hindfoot Length [mm]
taylori	8.600000	13.00000
montanus	10.250000	15.37500
flavus	7.953093	15.58056
megalotis	10.585838	16.44257
fulvescens	13.479452	17.52055

c) Recreate the plot.



d) Describe the plot.

e) 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.

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.

Nationality	n
German	91
Hungarian	78
Multinational	51

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.

3. University salaries.

Dataset *data/salaries.csv* contains data about yearly salaries of random 52 academic workers at one of the U.S. Universities. Your 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. He wants to know, if his salary is appropriate, higher than expected, or lower than expected.

Variables in the dataset:

- *sx* - sex
- *rk* - position at University
- *yr* - years working at this University
- *dg* - degree
- *yd* - years since recieving a doctorate

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

b) Are the conditions for a valid linear fit fulfilled in this case?

c) Predict your friend's salary using the model. Help him make a decision - should he complain about being underpaid?