2024 01 08 VB-STA5 Exam in statistics

Monday 8th of January, 14:00 - 18:00.

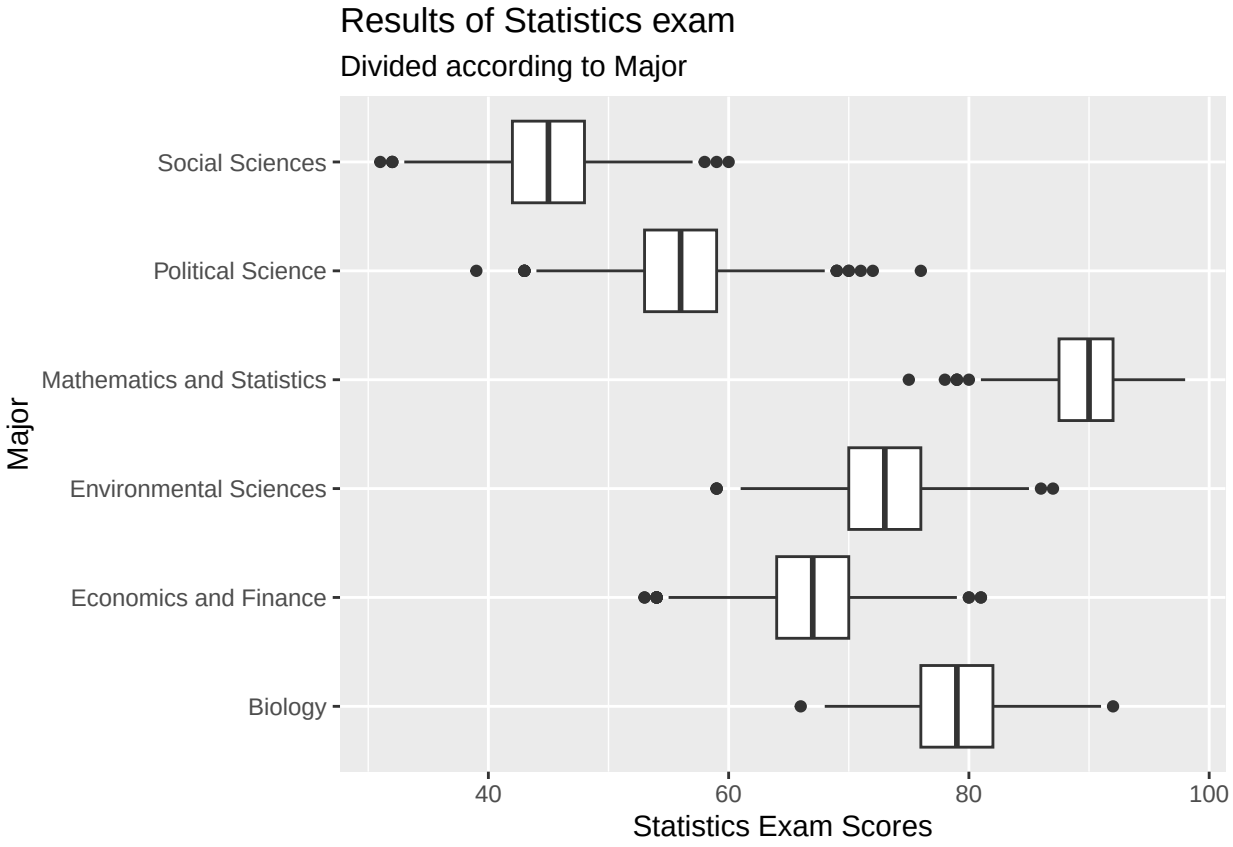
The exam set consists of 3 main exercises with 9 sub exercises in total.

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

1. Math and Statistics exam performance

Dataset *data/students_exam_performance.csv* contains information about students that participated both in Mathematics and Statistics class.

a) Recreate the plot:



b) Describe the plot.

c) Check whether there is a significant difference between a Mathematics Exam score for *Economics and Finance* major students with minor in *Mathematics and Statistics*, and *Economics and Finance* major students with other minors. Form hypothesis, check for conditions, and conduct a statistical test.

2.Childs seatbelt - car seat legislation

The dataset *data/CarSeatLegislation.csv* contains information about car accidents with children passengers.

a) Present the number and proportion of 'No Injury' accidents divided according to the implemented restraint (protection). *The example shows similar table, but for 'Incapacitating' injury*

Restraint	n	Proportion
Car Seat	1136	0.1139190
Lap and Shoulder Belt	1088	0.1091055
Lap-Only Belt	1103	0.1106097
No Restraint	6645	0.6663658

b) Is there a correlation in between type of injury and implemented restraint? Form hypothesis, check for conditions, and conduct a statistical test.

Injury	No Restraint	Car Seat	Lap and Shoulder Belt	Lap-Only Belt
Fatal	6201	1241	978	772
Incapacitating	6645	1136	1088	1103
Non-Incapacitating	468	1610	1233	1190
Possible Injury	1881	1111	772	683
No Injury	1769	1532	974	871

3.Wild blueberries yield prediction

Three datasets about wild blueberry farming are provided:

- *data/blueberries_insects.csv* contains information about pollinating insects presence
- *data/blueberries_weather.csv* contains information about weather (temperature and rain)
- *data/blueberries_yield.csv* contains information about size of the fruit, seeds, and final yield.

a) Join all three datasets.

b) Which variables have statistically significant influence on the blueberry yield? Create multiple regression model and tune it.

c) What should be satisfied for model (3b) to be valid. Check if the model you created is valid?

d) Construct a 95% confidence interval for the multiplication parameter of the '*seeds*' variable.