Examination Statistics, Bachelor Computer Sciences $\triangle$ 1
Prof. Dr. Falkenberg
24.02.2025

1. The data set fuel_efficiency_raw.csv contains data on 84 cars including their fuel efficiency in miles per gallon measured on a track.

   (a) Import the dataset into a tibble. In your solution, include the R commands and the first 10 lines of the tibble.

   (b) Explain, why the dataset is not tidy and make the dataset tidy. In your solution, include the R commands and the first 10 lines of the tidy data set.

   (c) Add variables containing the consumption in l per 100 km in the city and the highway (1 mile = 1.609 km, 1 gallon = 3.785 l).
   In your solution, include the R commands and the first 10 lines of the changed data set.

   (d) Determine for the different car types the cars with highest City.Mpg.
   In your solution, include the R commands and their output.

   (e) Determine for each country the minimum, maxium, mean, median, standard deviation and interquartil range of Highway.Mpg and number of cars for that country.
   In your solution, include the R commands and their output.

   (f) Draw a a side-by-side boxplot of Highway.Mpg depending on the car type. Discuss the differences visible in the diagram.
   In your solution, include the R commands and the resulting diagram.

   (g) Draw a scatterplot of Highway.Mpg against the Weight including the regression line.
   In your solution, include the R commands and the resulting diagram.

   (h) Interpret the diagram and the parameters of the regression line. Which percentage of the variation of Highway.Mpg can be explained by the Weight via a linear relationship (coefficient of determination) ?
   In your solution, include the R commands and their output.

2. Squash matches are played until one player has won 3 games.

   (a) Player A has a 40% probability of winning each individual game in a match. What is the probability that A wins the match.

   (b) Let X be the length of a match, i.e. number of games played in the match. Given A's 40% in chance, what is the density, the expected value and the variance of X.

   (c) Another player B takes part in a tournament with 8 players. The tournament uses a three round knockout system, i.e. players get eliminated as soon as they lose a match. The probabilities of B winning each individual game are p1 = 0.4 for the first round, p2=0.3 for the second round and p3=0.3 for the third round. What is the probability that B wins the whole tournament.

   In your solution, include your calculations as well as all R statements and their outputs.

3. The length of two kinds of wooden parts A and B are normally distributed, with means mu.A = 2 inches and mu.B = 4 inches, and standard deviations sigma.A = 0.009 inch, sigma.B = 0.04 inch. An A-part and a B-part are randomly selected and are joined end zo end into a single assembly. The assembly can be used further if the total length is between 5.92 and 6.08 inches. If the length is between 6.08 and 6.12 inches, it can still be fiexed manually. In all other cases, the assembly can no longer be used.

   (a) Determine the percentage of assemblies, which
   - fits
   - must be adjusted by hand
   - can not be used

   (b) Fixing an assembly costs 5 Euro. For every assembly that can not be used the loss is 12 Euro. The random variable X denote the additional costs per assembly. Determine the density of X, E(X) and Var(X).

   (c) If 100 A and B parts are randomly selected, what is (approximately) the probability that the total lost money is bigger than 100 Euro during the production process?

   In your solution, include your calculations as well as all R statements and their outputs.

4. Mr. Müller is running in the next election for mayor. According to the latest poll, 40% of voters wanted to vote for him. Mr. Müller believes that support for him has grown significantly in recent weeks and commissions a new survey. In this new survey of 822 voters in total 400 answered that they would vote for Mr. Müller.

   (a) Find a 95% lower confidence bound for the true but unknown proportion voters of Mr. Müller.

   (b) Give an interpretation of the lower confidence bound and of the confidence level.

   (c) You want to have a margin of error of ≤ 0.01 for the lower 95% confidence bound. What is the minimal amount of people that need to participate in the survey for that confidence level and margin of error? Use a normal approximation of the lower confidence bound.

   (d) Mr. Müller believes that the true proportion is at most 45%. Conduct an appropriate statistical test on a 1 percent level to check whether his believes are supported by the survey.
   Write down the null hypothesis. What is the p-value? What is your decision? Give the acceptance range of the null hypothesis.

   (e) Find the probability of type 2 error if the true proportion is 55%.

   In your solution, include your calculations as well as all R statements and their outputs.

5. In 2004, the state of North Carolina released a large data set containing birth records to the public. This data set has been of interest to medical researchers who are studying the relation between habits and practices of expectant mothers and the birthweight of their children. The csv-file weight_habit.csv is a random sample of 1000 cases from this data set containing the variables

   - weight = Weight of the baby at birth in pounds
   - habit = Status of the mother as a nonsmoker or a smoker.

   (a) Import the dataset into a tibble.

   (b) Create a side-by-side boxplots of the weight for each habit and interpret the boxplots.

   (c) Assume weight is a normally distributed random variable. Use an appopriate statistical to check whether the variances of the weight differs between smokers and nonsmokers at a 10% level. What is your decision?

   (d) You want to check if the birthweight of babies from smokers is lower. Use an appropriate statistical test at 5% level. What is your decision?

   In your solution, include your calculations as well as all R statements and their outputs.