Assignment 2: statistical testing

Statistics 2023 – 2024

Deadline: Friday 8 December, 23:59h

- Make sure your solutions are your own. Collaborating on the assignments (including the use of generative tools such as ChatGPT) is not allowed.
- The solutions should be in English.
- For exercises containing a calculation, giving only the solution is not sufficient: show how you obtained the solution.
- The maximum number of points is indicated for each exercise. You can lose points for writing irrelevant or incorrect information. Your grade will be computed as $points/9*10$.
- Hand in a PDF version of your solutions via Brightspace, created using your favorite text editor. Scans of hand-written solutions are not allowed and will not be graded.
- Pay attention: if you don't get a digital submission receipt (on your screen after submission and via e-mail), your solutions were not submitted correctly.

Good luck!

(2) 1. Eight different species of penguins live in Antarctica. Scientists have collected a data set with different measurements of many penguins of each species. The following variables are present in the data set:

- ID (unique to each penguin)
- Species (8 options)
- Sex (male or female)
- Body mass (grams)
- Flipper length (mm)
- Age group (chick, juvenile or adult)
- Beak length (mm)
- Partner ID (if a penguin has a known monogamous partner, this variable contains the ID of the partner)

For each of the following research question, give 1) if applicable, the explanatory variable and its measurement scale; 2) the response variable and its measurement scale; 3) the appropriate statistical test (if relevant, mention paired/unpaired and for one or two groups); and 4) the hypotheses. Keep your answers concise without additional explanations.

($\frac{1}{2}$) (a) Is the distribution of penguins over the different age groups the same for all eight species?

($\frac{1}{2}$) (b) Are the beaks of penguins that are in a monogamous relationship the same length?

($\frac{1}{2}$) (c) Are Macaroni penguins heavier than Chinstrap penguins, on average?

($\frac{1}{2}$) (d) Are beak length and flipper length correlated?

(4) 2. Below you see the outcome of a linear regression analysis that predicts song popularity (score on a scale from 0-100) from several song characteristics: duration (in minutes), tempo (in beats per minute), instruments (the number of instruments used) and pop (whether or not the genre of the song is pop). The data set contains 8000 observations.

Variable	Mean	St.dev.
duration	4.6	1.2
tempo	129	40
instruments	4.9	2.6
pop	0.16	0.37
popularity	49.3	21.3

Variable	Estimate	Std. Error
(intercept)	7.32	1.121
duration	1.63	0.174
tempo	0.25	0.005
instruments	0.41	0.081
pop	1.45	0.564

($1\frac{1}{2}$) (a) Give (formally) the prediction equation and interpret the estimated parameters. Either use the given variable names, or clearly state what your notation means.

(1) (b) The explanatory variables have very different units of measurement, so the estimated regression coefficients cannot be readily compared. How can we compare them? Do this comparison and determine which explanatory variable has the largest effect on predicted song popularity.

(1) (c) Using the given regression model and a confidence interval, determine whether the association between popularity and instruments is significant at $\alpha = 0.05$.

($\frac{1}{2}$) (d) Knowing that for this model $SSE = 2776112$, what is the expected absolute difference between the predicted and actual song popularity for a random song in the data set?

(3) 3. Researchers are interested in studying whether different generations have different preferences for transportation method to a holiday destination within Europe. They send out a survey and obtain the results in the table below.

	Car	Train	Plane	Total
Generation Z	55	138	107	300
Millennials	103	109	138	350
Generation X	191	24	55	270
Baby boomers	27	4	49	80

(1) (a) Use a chi-squared test to determine whether there is an association between generation and preferred transportation method with $\alpha = 0.05$. For the first cell (Gen Z preferring car), show your calculation. You don't have to show the calculations for the other cells and may use software to perform them, but make sure to describe all steps of the test.

(1) (b) What are the odds that someone who prefers travelling by plane is a millennial?

(1) (c) Calculate the odds ratio comparing generation Z and baby boomers in their preference for travelling by car (vs not by car). What is the interpretation of this odds ratio?