School of Mathematics and Statistics

Te Kura Mātai Tatauranga

STAT 292 **Assignment 3: Due Thursday, 6 May 2021 at 11:59 PM**

Note: Your assignment can be typed or handwritten (and scanned). Be sure to submit your assignment as a PDF and follow the instructions specified on the course Blackboard page. Where calculations are performed in `R`, be sure to include relevant code and output with your answer.

1. (20 marks)

   In psychology, there are tests to classify people into one of many personality types. An experiment is run to find the extent of the influence of personality type on the subject's score in a certain test. A random sample of four personality types is taken, and within each type a random sample of eight subjects is taken. Each subject is given the test and the score is recorded, with data as follows:

Type	Test Score
T1	44, 46, 59, 48, 49, 60, 51, 39
T2	43, 45, 57, 51, 51, 49, 44, 61
T3	51, 34, 53, 45, 41, 46, 41, 43
T4	52, 47, 59, 56, 49, 61, 57, 54

   a. Explain why this is a random effects design, rather than a fixed effects design.
   b. Write up the results of the analysis, following the Assignment Guidelines given at the end of these questions. Your report should include:
      - Boxplots, Levene's test and the usual diagnostic graphs.
      - Comments on whether the assumptions seem satisfied.
      - An ANOVA table.
      - Estimated components of variance and the percentage of the total variance in the test scores due to personality. Also give the percentage of the total variance that is unexplained.
      - An interpretation—is personality type important in determining the score on the test? Give a reason.

      Note: you will not need to present ANOVA hypotheses, as an F test is not required.

2. (20 marks)

Certain plants take up toxic metals (e.g. zinc, cadmium, uranium) and accumulate them in their vacuoles as protection against chewing insects and infections. Suppose that four species of plant were tested, at low and neutral soil pH, and their uptake of zinc, measured in parts per million (ppm) of dry plant weight at the end of the trial, was recorded as follows.

Plant Name	pH 5.5 (acid)	pH 7 (neutral)
Alpine pennycress	6340, 4280, 5170	2880, 4330, 3050
Bladder campion	3690, 4750, 5100	2360, 1990, 2140
Lettuce	250, 470, 330	400, 310, 430
Martin red fescue	2850, 2380, 3130	1070, 960, 1300

a. Specify what kind of design this is and give the relevant model equation, including an interaction term.
b. Analyse the data in R using the model from part (a), but try both the untransformed response variable and the log-transformed response variable. Choose one of these models for the presentation of your results, explaining the reason for your choice.
c. Present the report in the usual way, using a 5% significance level. Whether or not you found a significant interaction, include an interaction graph in your report (using your choice of raw or logged data), and refer to it in the interpretation section, to help illustrate your results.

3. (20 marks)

Children in a school class are given a test of comprehension of English, marked out of 100. The children are from three different ethnic groups, which is thought to be an important factor. The question of interest to the teacher is whether there are sex differences after allowing for ethnicity. The data follow.

Ethnic group	Females	Males
E1	67, 66, 75, 76, 71, 70, 72	63, 72, 62, 61, 69, 64, 71, 68, 56
E2	69, 57, 55, 63, 65, 55	59, 47, 49
E3	30, 47	39, 33

a. Do a two-way ANOVA on the data, presenting your results with the usual headings. Please include the interaction term in the model, and include an interaction graph plus comments on what it shows in the interpretation section at the end of your report.

   Note: Please present the ANOVA on the raw data, whether or not you find the diagnostic graphs acceptable. You should say if you think the ANOVA is not valid, or if you are undecided. If you think a log transformation is needed, just state this—with your reason—but do not actually do it.

b. If a one-way ANOVA is done with factor Sex, the resulting ANOVA table is given below. Explain (relatively briefly) the discrepancy between the outcome of this test and the test for Sex in part (a).

Source	Df	Sum of Squares	Mean Square	F value	Pr $> F$
Sex	1	144	144.2	0.987	0.329
Residuals	27	3943	146.0		

4. (20 marks)

The following data are from Peake and Quinn (1993), "Temporal variation in species-area curves for invertebrates in clumps of an intertidal mussel", *Ecography* **16**, 269–277. The response variable of interest is the number of different species of macroinvertebrates found in mussel clumps. The explanatory variable is the area (in dm^2) of the mussel clumps. The data are given below.

Clump	Area	Species
1	516	3
2	469.06	7
3	462.25	6
4	938.6	8
5	1357.15	10
6	1773.66	9
7	1686.01	10
8	1786.29	11
9	3090.07	16
10	3980.12	9
11	4424.84	13
12	4451.68	14
13	4982.89	12
14	4450.86	14
15	5490.74	20
16	7476.21	22
17	7138.82	15
18	9149.94	20
19	10133.07	22
20	9287.69	21
21	13729.13	15
22	20300.77	24
23	24712.72	25
24	27144.03	25
25	26117.81	24

Investigate whether it is appropriate to model the relationship between the number of different species of macroinvertebrates found in mussel clumps and the area (in dm^2) of those mussel clumps using simple linear regression.

Ensure you try fitting two models: one with **Area** as the explanatory variable and one using **log(Area)** as the explanatory variable. For both models, include a scatterplot of the data along with the line of best fit from the simple regression model that you have estimated using `R`. Pick the most appropriate of those two models and include in your report the model equation, the null and alternative hypotheses that you test, the model assumptions, diagnostic graphs and comments on whether the analysis is valid, as well as statistical conclusions and an appropriate interpretation.

Assignment Guidelines

The following Assignment Guidelines are relevant for all the assignments in Parts 2 and 3 of the course.

When you do a statistical test of a particular hypothesis, it is assumed you will state the following, **if relevant**:

- Model equation.
- Assumptions about the data, and comments about whether diagnostic graphs support those assumptions.
- Null and alternative hypotheses.
- ANOVA Table (if relevant) and p-value.
- Statistical conclusion(s) (*e.g.*, "We reject $\mathcal{H}_0$ and conclude $\mathcal{H}_1$, that μ_1 and μ_2 differ at the 5% significance level").
- Interpretation of the statistical conclusions back to the original problem, using the original meaning of the response variable and any factors or covariates. For example, if comparing heights of two groups, "Female and male adults have different mean heights, with males being taller on average".