Problem 1 Tree Distributions

In ecology there are theories that abundances (i.e. numbers) of species in a community should follow either a Poisson-lognormaldistribution or a negative binomial distribution (the exact forms of these are not important for this question).

Trees of different species on a central American island were counted. Table 1 summarises the distribution of abundances, i.e. how many species were counted once, how many counted twice etc. Some abundances have been pooled together (e.g. 6-10 is how many species were counted between 6 and 10 times), to avoid small counts. Both a Poisson-lognormal distribution and a negative binomial distribution were fitted to this data, to see which describes the data better.

Abundance Class	1	2	3	4	5	6-10	11-20	21-50	51-100	100+
Counts	19	13	9	5	8	19	25	49	34	44
Poisson log-normal	13.4	11.3	9.7	8.4	7.4	27.5	31.9	42.0	26.6	46.8
Negative Binomial	9.8	7.1	5.8	5.0	4.4	17.2	23.4	41.4	37.6	73.4

Table 1: Observed counts of abundance classes oftropical trees, and expected counts assuming a Poisson log-normal or negative binomial distribution.

a) Write down H_0 and H_1, for the test of whether the data follow a Poisson lognormal distribution. Then carry out the test. What do you conclude?

A negative binomial distribution was also fitted to the data. The goodness of fit statistic was 31.9, with 10 degrees of freedom.

b) Based on these statistics, test whether negative binomial distribution fits the data, and state your conclusion.

c) Which distribution do you think is a better fit to the data, and why?

Problem 2 Bird Palatability

A British zoologist in the last century suspected that birds that were more colourful, i.e. easier to see were not tastyHe collected 38 birds and gave them a score based on **visibility**, i.e. how easy they are to see (a higher score meant easier to see), and used taste tests to estimate **palalability**: a higher score meant they were better tasting. The data are plotted in Figure 1.

The analysis consisted of fitting a simple linear regression to the data. The regression model is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

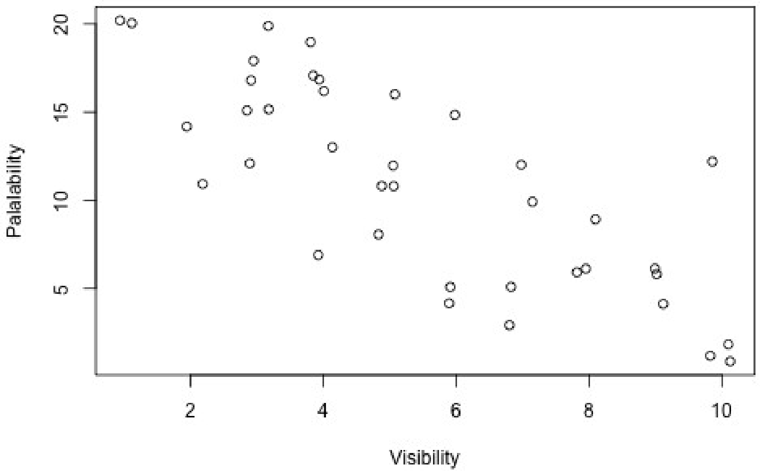

Figure 1: Relationship between Palatability and visibility of 38 species of bird.

a) Write down the likelihood for this model.

b) What is the fitted line (i.e. what are the values of the intercept & slope?)?

You might find some of these helpful:

$$\sum_{i=1}^{38} x_i = 210$$

$$\sum_{i=1}^{38} x_i^2 = 1424$$

$$\sum_{i=1}^{38} y_i = 416$$

$$\sum_{i=1}^{38} y_i^2 = 5778$$

$$\sum_{i=1}^{38} x_i y_i = 1853$$

The residual variance, σ^2, is estimated as $\hat{\sigma}^2 = 13.0$

c) Calculate the sampling variance of β_1

d) Calculate a 95% confidence interval for β_1

e) What is the R^2 for this data? What does R^2 tell you about the relationship between visibility and palalability?

Hint: calculate r first.

f) What would you conclude about the relationship between visibility and palalability? If you were offered a colourful bird to eat, how do you think it would taste?

Problem 3 Two Way ANOVA with Replication

We have looked at a two-way ANOVA with a treatment and block, but this design can be extended by having multiple observations for each treatment/block combination.

Assume we have a treatment, $i = 1, ..., T$ and blocks $j = 1, .., B$ For each combination of i and j we have $k = 1, ..., n$ observations. Our model for this data is

$$y_{ijk} = \mu_j + \beta_i + \gamma_{ij} + \varepsilon_{ijk}$$

which splits the data into a treatment effect, μ_j, a block effect, β_j, an interaction effect, γ_{ij}, and a residual error, ε_{ijk}. The γ_{ij} allows the block effect to be different in different treatments.

a) Show that the total sum of squares ($SS_T = \sum_{i=1}^{T} \sum_{j=1}^{B} \sum_{k=1}^{n} (y_{ijk} - \bar{y}_{...})^2$) can be written as $SS_T = SS_{T R} + SS_B + SS_{T B} + SS_E$

where

$$SS_{TR} = \sum_{i=1}^{T}\sum_{j=1}^{B}\sum_{k=1}^{n}(\bar{y}_{i..} - \bar{y}_{...})^2$$

$$SS_B = \sum_{i=1}^{T}\sum_{j=1}^{B}\sum_{k=1}^{n}(\bar{y}_{.j.} - \bar{y}_{...})^2$$

$$SS_{TB} = \sum_{i=1}^{T}\sum_{j=1}^{B}\sum_{k=1}^{n}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$$

$$SS_E = \sum_{i=1}^{T}\sum_{j=1}^{B}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij.})^2$$

And, for example, $\bar{y}_{i..} = \frac{1}{nB}\sum_{j=1}^{B}\sum_{k=1}^{n} y_{ijk}$.

Note: you may assume any cross-product terms (such as AB in $(A - B)^2$) are zero

We have a long-term data set on yields of barley, which was designed to test the effects of fertilisers. The data are mean yields from each decade. There are four treatments:

Control No fertiliser

Fertilised Artificial fertiliser used

Manure manure (animal feriliser) used

Stopped fertiliser had been applied, but was later stopped

The second factor is whether the decade is before or after 1971 (as there were changes to the experiment in that year).

	df	SS	MS	F value
Time	1	13.1	13.1	56.6
Treatment	3	93.9	a	b
Time:Treatment	3	c	6.6	28.6
Residuals	64	14.8	0.23	

Table 2: Analysis of Variance for Barley Yield Data. df: Degrees of Freedom, SS: Sums of Squares, MS: Mean Squares

Based on this model, we can write down the ANOVA table in Table 2.

b) Fill in the missing values (a, b and c) in the ANOVA

c) Test if the 'Time:Treatment' effect is significant. Explain which statistics you use.

The Time:Treatment effect is the interaction: here it estimates how the treatments changed from before to after 1971. Table 3 shows the contrasts of each treatment to the control, i.e. testing if the change in treatment effect was different from the change in the control.

	Estimate	Std. Error	t value
TimeAfter:TreatmentFertilised	1.02	0.34	3.00
TimeAfter:TreatmentManure	2.61	0.34	7.70
TimeAfter:TreatmentStopped	-0.19	0.34	-0.57

Table 3: Contrasts for interaction terms for Barley Yield Data.

d) Which of the tests in Table 3 are different from 0? Test this at 5%, after correcting for multiple tests.

Problem 4 Hobbit Hairs

You have been hired by a new boss, Sharkey, who wants to shave the feet of the hobbits he has taken control of. He needs to know how many razor blades to make, for which he needs to know how hairy the hobbits' feet are. You have collected data from 100 hobbits, which are plotted in Figure 2

The sample mean is 77.0, the sample median is 26 and the sample standard deviation is 109.6.

a) Calculate the 95% confidence interval for the mean number of hairs, assuming the data come from a normal distribution.

If the mean number of hairs per hobbit is less than 55, Sharkey will have enough razor blades to shave all of the hobbit feet (if he does not have enough, he will have to build a factory to make more).

b) Test if your data is likely if the actual mean number is 55 hairs per hobbit.

Sharkey points out that the data do not look normally distributed, and demands you use a Wilcoxon signed rank test. this is based on the ranks of $|x_i - \mu_0|$. The test statistic is W, the sum of the ranks of the data for which $x_i > \mu_0$.

Table 4 give the data for 10 of the 100 hobbits, along with statistics relevant for a Wilcoxon signed rank test.

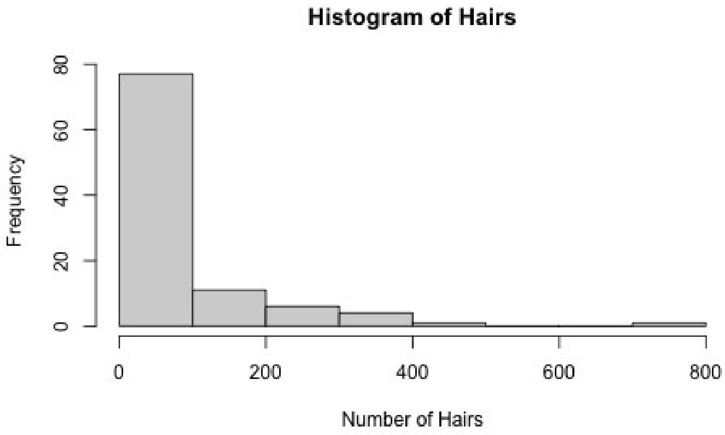

Figure 2: Histogram of number of hairs on hobbit feet.

c) Use these 10 data points to test if the median is less than 55. State H_0 and H_1, and the distribution of W under the null hypothesis. You can assume a large sample size.

Number of Hairs, h	204	401	26	125	21	24	23	78	28	16
h-55	149	346	-29	70	-34	-31	-32	23	-27	-39
Rank of |h-55|	9	10	3	8	6	4	5	1	2	7

Table 4: Sample of data on number of hairs on hobbits' feet.

d) Which test do you prefer for this problem: the t-test or the Wilcoxon signed rank test (carried out on the full data, rather than the subset above)? Why do you prefer this test?