ST1201/ST6201 20022h Exam Solutions

Bob O'Hara

2022-12-09

Problem 1

(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?

H_0: The distribution follows a Poisson lognormal distribution

H_1: The distribution does not follow a Poisson lognormal distribution. i.e. the counts are too far away from the expected values

Now we need to calculate $(O_i - E_I)^2/E_i$. O_i is the Counts row, E_i is the Poisson log-normal row:

	1	2	3	4	5	6-10	11-20	21-50	51-100	100+
O_i	19.0	13.0	9.0	5.0	8.0	19.0	25.0	49.0	34.0	44.0
E_i	13.4	11.3	9.7	8.4	7.4	27.5	31.9	42.0	26.6	46.8
O_i - E_i	5.6	1.7	-0.7	-3.4	0.6	-8.5	-6.9	7.0	7.4	-2.8
(O_i - E_i)^2	31.7	2.8	0.5	11.7	0.3	72.2	48.0	49.4	55.1	7.7
(O_i - E_i)^2/E_i	2.4	0.2	0.1	1.4	0.0	2.6	1.5	1.2	2.1	0.2

Then we need to add these up:

$2.37 + 0.24 + 0.05 + 1.39 + 0.04 + 2.63 + 1.5 + 1.18 + 2.07 + 0.16 = 11.64$

(Note: there may be some rounding error, which will not be penalised. I got 11.59 by hand, and 11.64 without rounding first)

We need to test this. The statistic is 11.64. The degrees of freedom was not stated. In the next section it was given as 10, but it should have been 8 (as there are 2 parameters estimated). I thus marked it correct if 8, 9, or 10 was used. The critical value for a test at 5% is 15.51, so the statistic is less than this (in fact the p-value is 0.17). We can thus declare the test non-significant, and conclude that there is no evidence that the distribution does not follow a Poisson-lognormal distribution.

(b) does a negative binomial distribution fit the data?

(The goodness of fit statistic was 31.93, with 10 degrees of freedom)

The critical value is the same, so this suggests that the negative binomial distribution is not a good fit to the data. If anyone is interested, the p-value is 4×10^{-4}.

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

The Goodness of fit statistics are different: the Poisson lognormal has a lower statistic, which suggests it fits better. The difference is quite large, so this is probably not random noise.

Note: the data are real, from a "classic" data set from Barro Colorado Island in Panama. There was a small industry in ecology in fitting different distributions to this data: thankfully most people seem to have got bored by doing this now.

Problem 2

(a) Write down the likelihood for this model.

$$L(\theta|y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y_i-\beta_0-\beta_1 x_i)^2}{\sigma^2}} = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\beta_0-\beta_1 x_i)^2}{\sigma^2}}$$

or

$$L(\theta|y) = -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

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

First the slope. This is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}.$$

We can calculate $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 210/38 = 5.53$, and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 416/38 = 10.95$. So

$$\hat{\beta}_1 = \frac{1853 - 38 \times 5.53 \times 10.95}{1424 - 38 \times 5.53^2} = \frac{-445.95}{263.47} = -1.69.$$

Then

$$\beta_0 = \bar{y} - \beta_1\bar{x} = 10.95 + 1.69 \times 5.53 = 20.30.$$

(c) Calculate the sampling variance of β_1

We are given $\hat{\sigma}^2$, so

$$Var(\beta_1) = \hat{\sigma}^2 / \sum (x_i - \bar{x})^2 = 13.0/263.47 = 0.053.$$

(d) Calculate 95% confidence intervals for β_1

Assuming a t distribution with $n - 2 = 36$ degrees of freedom we get

$$\hat{\beta}_1 \pm 2.03 \times \sqrt{\hat{\sigma}^2 / \sum (x_i - \bar{x})^2},$$

which is

$$-1.69 \pm 2.03 \times \sqrt{0.053} = -1.69 \pm 0.47 = (-2.16, -1.22).$$

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

First, follow the hint and calculate r.

$$r = -445.95/\sqrt{263.4 \times 1223.90} = -0.78$$

so $R^2 = 0.62$, or 62%.

This tells us that most of the variation in palatability is explained by visibility. The relationship is strong, but not super strong. (judging the strength is a bit subjective, which means some leeway in the marking).

(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?

There is a fairly strong negative relationship between visibility and palalability: the more colourful a bird is, the less palatable it is. So if I was offered a colourful bird to eat, I do not think it would taste very good.

Note: the data is real. This is the paper:

Cott, H.B. (1947), The Edibility of Birds: Illustrated by Five Years' Experiments and Observations (1941–1946) on the Food Preferences of the Hornet, Cat and Man;and considered with Special Reference to the Theories of Adaptive Coloration. Proceedings of the Zoological Society of London, 116: 371-524. https://doi.org/10.1111/j.1096-3642.1947.tb00131.x

The data is actually for taste tests with hornets, but he found that humans gave similar rankings. apparently cats do too.

Problem 3

(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 $SS_T = SS_{TR} + SS_B + SS_{TB} + SS_E$

Take a deep breath! The main trick is to write $(A-B)^2$ as $(A-C+C-B)^2 = ((A-C)+(C-B))^2$ (we did this a couple of times in the course). Then it's just doing it in the right places, so you get the right sums of squares, and simplifying.

$$
\begin{aligned}
SS_T &= \sum_{i=1}^{T}\sum_{j=1}^{B}\sum_{k=1}^{n}(y_{ijk} - \bar{y})^2 \\
&= \sum_{i=1}^{T}\sum_{j=1}^{B}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{i..} + \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{.j.} - \bar{y}_{ij.} + \bar{y}_{ij.} - \bar{y}_{...} + \bar{y}_{...} - \bar{y}_{...})^2 \\
&= \sum_{i=1}^{T}\sum_{j=1}^{B}\sum_{k=1}^{n}([y_{ijk} - \bar{y}_{ij.}] + [\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + +\bar{y}_{...}] + [\bar{y}_{.j.} - \bar{y}_{...}] + [\bar{y}_{i..} - \bar{y}_{...}])^2 \\
&= \sum_{i=1}^{T}\sum_{j=1}^{B}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij.})^2 + (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + +\bar{y}_{...})^2 + (\bar{y}_{.j.} - \bar{y}_{...})^2 + (\bar{y}_{i..} - \bar{y}_{...})^2 + CP \\
&= \sum_{i=1}^{T}\sum_{j=1}^{B}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij.})^2 + \sum_{i=1}^{T}\sum_{j=1}^{B}\sum_{k=1}^{n}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 + \\
&\sum_{i=1}^{T}\sum_{j=1}^{B}\sum_{k=1}^{n}(\bar{y}_{.j.} - \bar{y}_{...})^2 + \sum_{i=1}^{T}\sum_{j=1}^{B}\sum_{k=1}^{n}(\bar{y}_{i..} - \bar{y}_{...})^2 \\
&= SS_{TR} + SS_B + SS_{TB} + SS_E
\end{aligned}
$$

(yes, this is a difficult question: I wanted to stretch you)

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

$$a = 93.9/3 = 31.3$$

$$b = a/0.23 = 135.2$$

$$c = 6.6 \times 3 = 19.8$$

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

From the (full) ANOVA table above, 'Time:Treatment' has an F ratio of 28.6, with 3 and 64 degrees of freedom. The critical p-value at 5% is 2.75, so the actual value is much higher. (the p-value is about 10^{-11}).

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

From the table, and assuming a t distribution with 64 degrees of freedom, the p-values are 0.0039, 0, 0.5696 (a normal distribution would give very similar values). We can use a few corrections to find the critical p-value:

- Bonferroni: 0.05/3 = 0.0167
- Sidak: 1-(1-0.05)^(1/3) = 0.0170
- Benjamani-Hochberg: 0.0167, 2*0.0167=0.033, 0.05

In all cases the first 2 contrasts (TimeAfter:TreatmentFertilised and TimeAfter:TreatmentManure) are significant: TimeAfter:TreatmentStopped is not.

Note: the data are from a field trial at Rothamsted in the UK, which has been running since 1852 and is still going (www.era.rothamsted.ac.uk/Hoosfield). Rothamsted is where R.A. Fisher did a lot of his work modernising statistics in the 1920s. His motivation was to help design and analyse field trials like this one.

Problem 4

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

The test assumes the mean follows a t distribution: $\mu \pm t_{99,0.975}\sigma/\sqrt{n} = 77.0 \pm 1.98 \times 109.6/\sqrt{100} = 77.0 \pm 21.7 = (55.3, 98.7)$.

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

the test statistic is $t = \frac{55-\mu}{\sigma/\sqrt{n}} = \frac{55-77.0}{109.6/\sqrt{100}} = -2.01$

The critical value for the t test (with 99 degrees of freedom) is -1.98, so this is (just) significant at 5%. The actual p-value is 0.047. Note that as stated this is a two-tailed test.

(c) Use these 10 data points to test if the median is less than 55. State H0 and H1, and the distribution of W under the null hypothesis. You can assume a large sample size.

H_0: $\mu_D = 55$ (i.e. the median)

H_1: $\mu_D < 55$

Under the null hypothesis (and a large sample size), $W \sim N(\frac{1}{4}n(n+1), \sqrt{\frac{1}{24}n(n+1)(2n+1)})$

Now the calculation. This is the table, with the ranks for the data above the median in red:

HS	204	401	26	125	21	24	23	78	28	16
Diff	149	346	-29	70	-34	-31	-32	23	-27	-39
Rank	9	10	3	8	6	4	5	1	2	7

We can calculate the test statistic for either above (W_+) or below (W_-) the median. The statistics are:

$W_+ = 9 + 10 + 8 + 1 = 28$

$W_- = 3 + 6 + 4 + 5 + 2 + 7 = 27$

Under the null hypothesis, $W \sim (27\frac{1}{4}, \sqrt{96\frac{1}{4}})$

Using W_-, the test statistic is $\frac{27-27.5}{\sqrt{96\frac{1}{4}}} = 0.05$

The critical value is 1.64 (because this is a one-tailed test), so it is massively non-significant.

(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?

The correct answer is probably "neither of the above"! Either of these is appropriate as the right answer:

- the data are horribly non-normal, so the Wilcoxon signed rank test might be more appropriate. Except that it assumes that the distributions are symmetrical.
- BUT the mean is a more appropriate, because Starkey wants to shave a total number of hairs (55 times the number of hobbits). So a t-test might be more appropriate.

Note: the data are real, but are of water heights in Shipley in the UK, rounded to the nearest integer. But that would have been boring, so I changed them to hobbit hairs. Sharkey appears at the end of *The Lord of the Rings*. I won't give the ending away.