Norwegian University of Science and Technology
Department of Mathematical Sciences

-

-

Problem 1 Dog mimics

In a recent experiment, researchers looked at whether dogs prefer people who mimic them. They had 30 dogs, and presented them with two researchers: one mimicked the dog, and the other did not. They then looked at which researcher the dog preferred. Their thesis was that dogs would prefer the human who mimicked them, so the proportion of trials where the dog preferred the mimicker should be larger than 50%. We want to estimate the actual proportion of trials where the dog preferred the mimicker from the experimental data. We can assume that the number of dogs preferring the mimicker follows a binomial distribution.

a) What are the parameters of the binomial distribution?

ANSWER: total = 2 marks: 1 each for N (number of trials) and p (probability of success)

b) What assumptions do we need to make when using this distribution?

ANSWER: total = 2 marks: 1 each for constant p, i.e. same probability of success for each trial and independence

c) How reasonable are these assumptions for this experiment?

ANSWER: total = 4 marks: 2 for independence probably OK, and 2 for constant p is a problem as the researchers might have different likabilities (or other aspects might vary) so probability of success is not the same for each trial

In the experiment, 16 dogs out of 30 preferred the mimicker.

d) What is the estimate for the preference?

ANSWER: total = 2 marks for 16/30 = 0.533

Below we plot the likelihood curve (Figure 1).

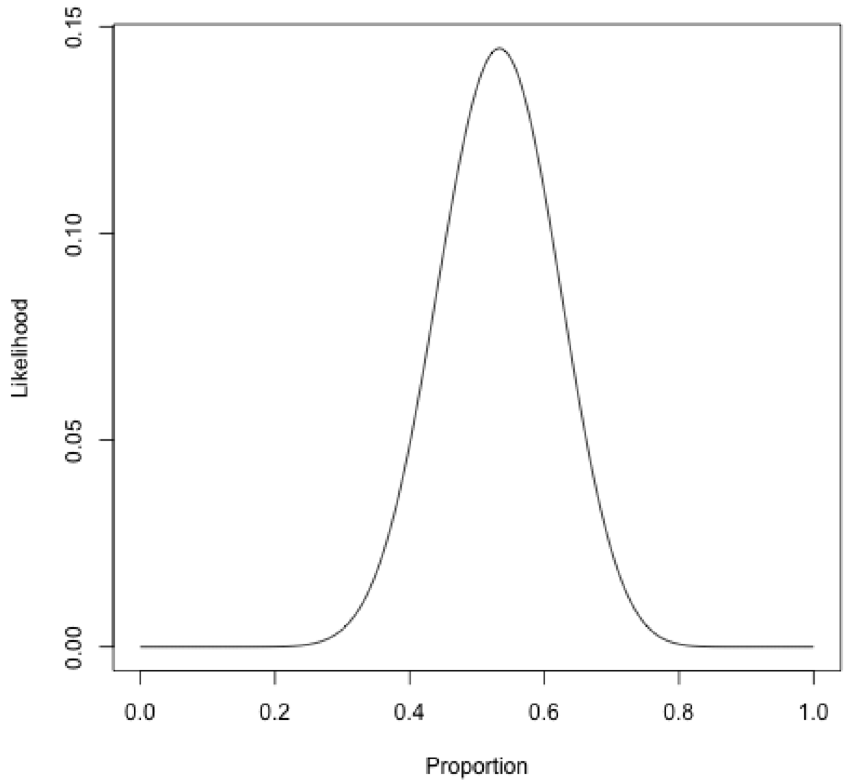

Figure 1: Plot of the likelihood for dog mimic data.

e) What is plotted on the y-axis against p, $Pr(n = 16 \ |N = 30 \ , p)$ or $Pr(p|n = 16 \ , N = 30 \ )$?

ANSWER: total = 1 mark $Pr(n = 16|N = 30, p)$

f) We want a confidence interval for the probability. Give one way we could we calculate an exact confidence interval for our estimate of the proportion of success?

ANSWER: total = 2 marks for either simulate the data and calculate the quantiles, OR calculate the quantiles from qbinom

We can calculate an approximate confidence interval by assuming a normal distribution. The standard error for a binomial distribution is $\sqrt{p(1-p)/N}$.

g) What is the standard error for this data?

ANSWER: total = 1 mark sqrt(0.53(1-0.53)/30) =0.09*

h) What is the approximate 95% confidence interval?

*ANSWER: total = 2 marks (0.53 - 1.96*0.09, 0.53 + 1.96*0.09) = (0.35, 0.70)*

i) Do the data show any evidence for dogs preferring mimics?

ANSWER: total = 2 marks. No, 0.5 is well within CI.

Problem 2 Circadian rhythms

It is well known that light has a strong effect on human circadian rhythm, e.g. sleeping patterns. In 1998 Campbell and Murphy claimed that the human circadian clock can be reset not just by light to the eyes, but by light applied to the back of the knee. In 2002 Wright and Czeisler decided to test this theory, they weren't completely convinced by the first study.

They constructed an experiment and measured daily cycles of melatonin (a hormone that regulates sleep-wake cycles) in 22 people. Each of these people were randomly assigned to one of three treatments; three hours of light to the eyes only, three hours of light to the knees only, and no light.

This produced data on:

- Melatonin shift in hours (**shift**), (negative = production later, positive = earlier production).
- Treatment group; control, eyes or knees (**treatment**).

10 rows of the data are shown below in Figure 2.

```
treatment shift
1        control       0.53
2        control       0.36
3        control       0.20
10        knee       0.31
11        knee       0.03
12        knee       -0.29
20        eyes       -1.52
21        eyes       -2.04
22        eyes       -2.83
```

Figure 2: Top 10 rows of circadian rhythm data.

a) Write a biological question you could answer with this data.

ANSWER total = 2 marks: for something related to: 'Does exposure to light on the knees produce a change in human circadian rhythms?' Other questions are acceptable but melatonin/circadian rhythm must be the response.

b) Would you choose an lm() or a glm() to address this question, why?

ANSWER total = 2 marks: 1 mark for either: choosing a linear model (lm) or glm() AND 1 mark for the reason: lm() because the residuals will be normally distributed and the response is continuous OR a glm() with family=Gaussian(link=identity) as this will give the same as the lm().

The figures below show some plots relating to the model fit of a linear model of this data.

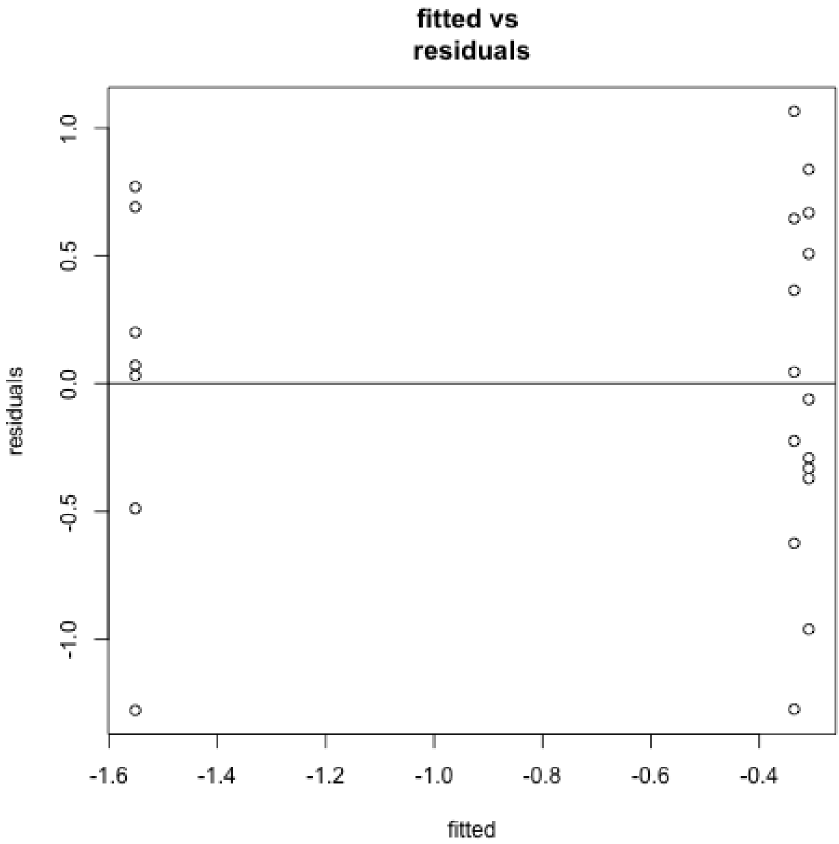

Figure 3: Plot of fitted versus residuals for model1.

c) Look at Figure 3, which assumption does this plot test? and how well does this model meet this assumption?

ANSWER total = 4 marks: 1 mark for residuals vs fitted tests equal variance and linearity. 2 marks for interpretting (some of this can be subjective when stating if good or bad but pattern should be as stated here). E.g. There seems to be equal variance across the fitted values, it is not heteroscedastic. Looks like two groups but that is because control and knees have very similar fitted values. There does not seem to be any evidence of non-linearity. 1 mark for statement of how well it is; good, ok, not bad, not great - all acceptable. Choice of 'does not meet assumptions' is not easily justified here.

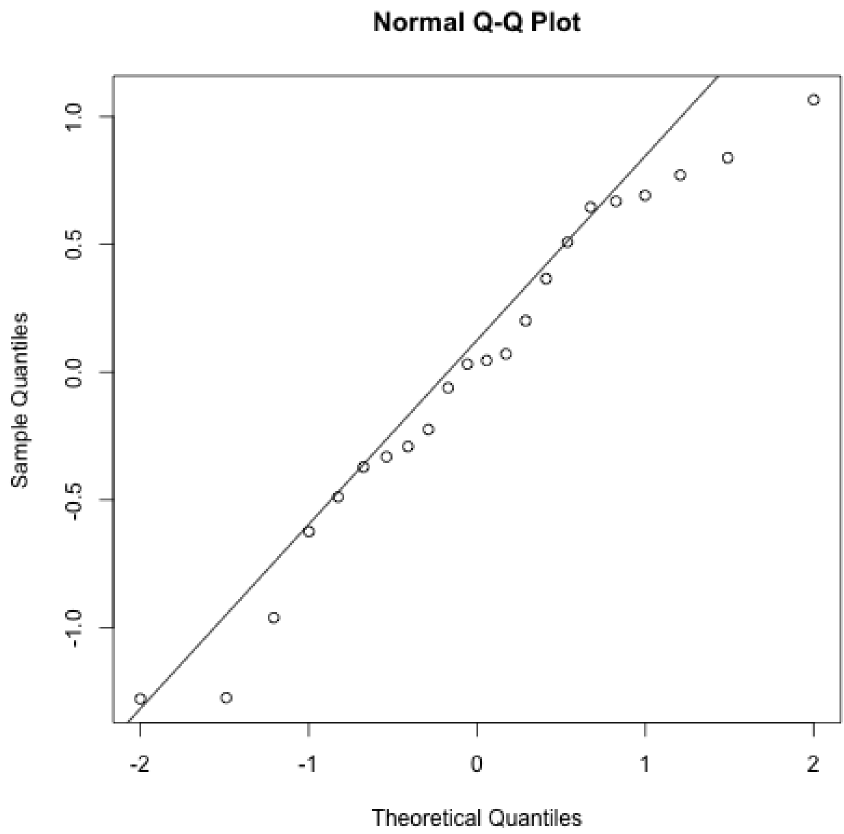

Figure 4: Normal QQ plot for model1.

d) Look at Figure 4, which assumption does this plot test? and how well does this model meet this assumption?

ANSWER total = 4 marks: 1 mark for normal QQ plot tests normality of residuals. 2 marks for: The residuals fit along the theoretical line in the normal QQ plot, but there is deviation at high theoretical values. 1 mark for statement of whether assumption is met; e.g. the normality assumption is met or the normality assumption is not quite met because of the deviation at high theoretical values. Justification for assumptions not being met at all is hard here.

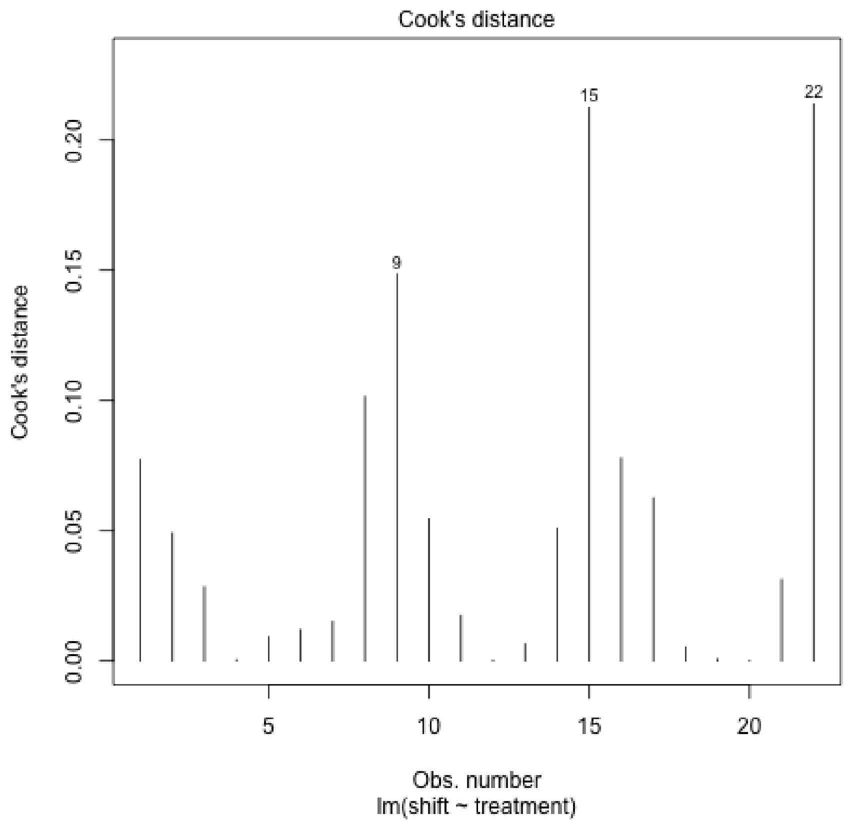

Figure 5: Cook's distance plot for model1.

e) Look at Figure 5, which assumption does this plot test? and how well does this model meet this assumption?

ANSWER total = 4 marks: 1 mark for Cook's distance tests outliers. 2 marks for: There are 3 outliers marked on the Cook's distance these points might need investigating, especially if they are the same points that deviate from normality. 1 mark for statement of whether assumption is met; both yes and no acceptable here but should be justified.

Suggestions to remove outliers without checking why they are outliers is not correct and leads to no mark when assessing if the assumption is met.

```
Call:
lm(formula = shift ~ treatment, data = LightData)

Residuals:
     Min       1Q   Median       3Q      Max
-1.27857 -0.36125  0.03857  0.61147  1.06571

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.30875    0.24888  -1.241  0.22988
treatmenteyes -1.24268    0.36433  -3.411  0.00293 **
treatmentknee -0.02696    0.36433  -0.074  0.94178
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7039 on 19 degrees of freedom
Multiple R-squared:  0.4342,        Adjusted R-squared:  0.3746
F-statistic: 7.289 on 2 and 19 DF,  p-value: 0.004472

                   2.5 %     97.5 %
(Intercept)   -0.8296694  0.2121694
treatmenteyes -2.0052265 -0.4801306
treatmentknee -0.7895122  0.7355836
```

Figure 6: Summary and confint output of model0

f) What do the coefficients from model0 (Figure 6) represent mathematically?

ANSWER total = 3 marks: 1 mark for the (Intercept) is the mean of the contol group (the contrast here) and is -0.3. 1 mark for treatmenteyes = the difference in mean from the control group to the group with light applied to eyes. The difference is -1.24. 1 mark for treatmentknees = the difference in mean from the control group to the group with light applied to knees. The difference is -0.03. Student's should be able to work out what model has been run from the 'call' section of the summary output.

g) What statistical conclusions can you draw from these results?

ANSWER total = 10 marks: 3 marks for first justified statement: e.g. The estimate of the mean of the control group is -0.3, however the confidence intervals span 0. This means when we consider uncertainty, 0 is included as a plausible value for this mean. Therefore statistically, the control treatment seems to have little or no effect on circadian rhythms. 3 marks for second justified statement: e.g. There is a negative effect of light applied to the eyes (negative meaning delay to melatonin production) compared to the control group. The reduction is between -2 and -0.48, therefore even with uncertainty it seems that a negative effect would be unlikely if the null was true. 3 marks for third justified statement: e.g. There is a negative estimate of the effect of light to the knees of -0.03. However, when we include the uncertainty in this estimate the 95% confidence interval spans from -0.79 to 0.74, this is a very wide band with 0 within in it. When we include uncertainty we cannot distinguish a direction of this effect, 0 remains a plausible option for the difference between the control and the knees treatment group. Each of the above needs three parts; the direction of relationship, the actual number (effect size) and confidence intervals. 1 mark for interpreting that approx 43% of the variation has been explained. This is good but not everything. Can be interpreted as either good or not enough, both answers are fine.

Problem 3 Clutch size

A decrease in number of eggs a bird produces (it's clutch size) the later in the year it breeds, has been observed across many bird species. The data below can be used to test whether this happens in great tits (kjøttmeis).

Researchers collected data on when the first egg appeared in great tit nests and then counted the total number of eggs that were laid. The data comes from a single year (1990).

This produced data on:

- The date the first egg appeared, in days since the 1st April 1990 (**LayDate**) - The total number of eggs counted in the nest (**ClutchSize**)

The first 10 rows of the data are shown in Figure 7.

```
LayDate ClutchSize
6043        25        12
6044        18         9
6045        28         6
6046        24         8
6047        16         9
6048        18         9
6049        19         9
6050         4        12
6051        27         8
6052        30         7
```

Figure 7: Top 10 rows of clutch size data.

To answer the question *Does the time eggs are laid influence the total clutch size?* researchers fit the model in Figure reffig:Model1:

```
model1 <- glm(ClutchSize ~ LayDate, data = BirdData, family = poisson(link = "log"))
```

Figure 8: R code for model1.

a) Why did researchers choose the model in (Figure 8) for this data?

ANSWER total = 2 marks: 1 mark for recognising that we have count data as the response so the residuals will be poisson distributed. 1 mark for stating that as it is poisson we need a glm, lm cannot handle non-normal error.|

b) Why has a log link function been used here?

ANSWER total = 2 marks: 1 mark for it is the default (canonical) link for a poisson family. 1 mark for it represents counting effort, by being exponential.

Before going too far with an analysis, the researchers decide to test whether there is any evidence of a quadratic effect of lay date on clutch size. As a specific hypothesis this is: Does lay date have a quadratic association with clutch size in great tits?*.

They run the following R code in Figure reffig:ANOVA:

```
model2 <- glm(ClutchSize ~ LayDate + I(LayDate^2), data = BirdData, family = poisson(l

anova(model1, model2, test="LRT")
```

Figure 9: R code.

```
Analysis of Deviance Table

Model 1: ClutchSize ~ LayDate
Model 2: ClutchSize ~ LayDate + I(LayDate^2)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       166     62.384
2       165     62.374  1 0.010234   0.9194
```

Figure 10: R code output of Figure 9.

c) Look at the output of the test in Figure reffig:ANOVA, is there any evidence of a quadratic effect of lay date? Explain your answer.

ANSWER total = 4 marks: 1 mark for no. 1 mark for the probability value (Pr(>Chi)) is 0.92. 1 mark for: this means that there is a 92% probability we would see the chi square value we got or higher if the null is true. 1 mark for: this means we cannot reject the null hypothesis so we say that the test is not statistically significant.

d) We have several variables (mean temperature, precipitation, and insect abundance) which might also affect clutch size. How would we select the variables that best explain the data?

ANSWER total = 4 marks: 1 mark for conduct exploratory model selection. 1 mark for construct several different models including all combinations of the explanatory variables. (lay date, mean temperature, precipitation, and insect abundance). 1 mark compare different models using the AIC or BIC (depending on the penalty you want to give to extra parameters). 1 mark for the code you would use e.g. AIC(), BIC(), bestglm().

```
deviance(model1)/df.residual(model1)

[1] 0.3758064
```

Figure 11: R code calculating deviance ratio.

e) Look at the code in Figure 11, what assumption of model1 is this code testing? and is this assumption met?

ANSWER total = 4 marks: 1 mark for testing the dispersion/variance assumption, that is should be controlled by the mean so the deviance ratio = 1. The variance should = the mean. 1 mark for: the model is not over-dispersed as the deviance ratio is < 1.2. 1 mark for: deviance ratio is not exactly 1 so the variance assumption is not perfectly met. 1 mark for: as the deviance ratio is less than 1, we could have under-dispersion, too little variation for what is expected under a Poisson.

```
coef(model1)

(Intercept)     LayDate
2.35205677 -0.01051424

confint(model1)

                  2.5 %       97.5 %
(Intercept)  2.22715164  2.475657352
LayDate     -0.01730312 -0.003797819
```

Figure 12: Coefficients and confidence intervals for model1.

As temperatures warm, lay dates are getting earlier. This year, 2019, the first great tit laid an egg 5 days before the start of April.

f) Predict the clutch size for a bird with a lay date of -5, using the coefficients in Figure 12.

ANSWER total = 2 marks: 2 marks for prediction: use the equation Y = exp(a + bX), where Y = predicted clutch size on original scale, a = intercept of GLM, b = slope of GLM, and X = lay date. exp(2.35 + (-0.01×-5)) = 11 eggs. Only 1 mark is given if the back transformation step, exp(), is forgotten.

g) The mean clutch size in 1990 was 9 eggs. What implications could the prediction for 2019 have for the great tit population, relative to 1990? Give a maximum of 3 implications.

ANSWER total = 4 marks: 1 mark for a summary of the finding: e.g. Over 29 years, the clutch size (based on the current model, data, and prediction) has increased for this population by 2 eggs. 1 mark for statements relating to a increase in the population if there is an increase in eggs as birds lay earlier. 1 mark for considering the rate of *increase, 2 eggs in 29 years. 1 mark for considering increased competition and density dependence. 2 marks for including limits to these predictions e.g. that physiological limits to egg production might being reached.*

h) Give two problems with generating predictions such as those in f (that predict outside of the original dataset)?

ANSWER total = 4 marks: 2 marks for explained statement: the relationship identified from the data may not hold. For instance, physiological limits to increasing clutch size might be reached so the relationship could become non-linear (even on the link scale). This could also be caused by temperature limits being reached, and any number of other causes. As long as a recognition that the relationship could change under new conditions is included and is sensible, exact reason for relationship change does not matter. 2 marks for uncertainty in the prediction was not included. This should always be included in predictions from models.