Cover Page

Department of Mathematical Sciences

Examination paper for ST2304 Statistical Modelling for Biologists and Biotechnologists

Examination date: 22nd May 2023
Examination time (from-to): 15:00 – 19:00

Permitted examination support material: C. Specified printed and hand-written support material is allowed. A specific basic calculator is allowed.

Academic contact during examination: Bob O'Hara

Phone: 91554416

Academic contact present at the exam location: NO

OTHER INFORMATION

Get an overview of the question set before you start answering the questions.

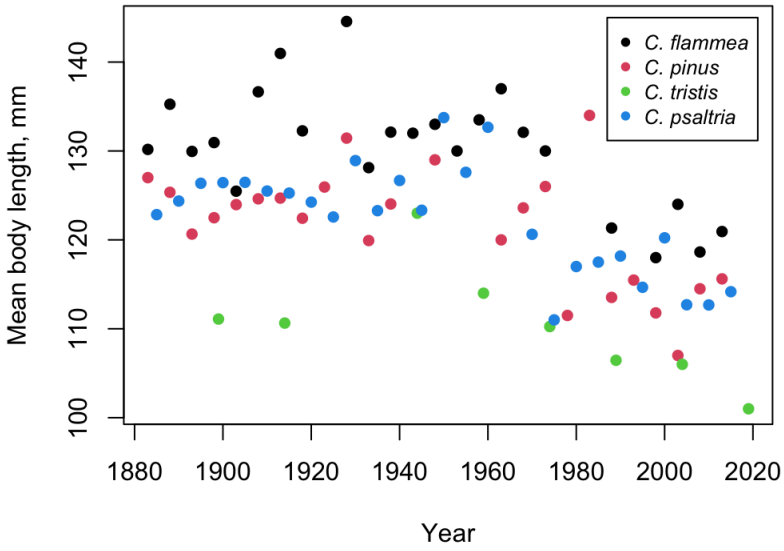
Read the questions carefully and make your own assumptions. If a question is unclear/vague, make your own assumptions and specify them in your answer. Only contact academic contact in case of errors or insufficiencies in the question set. Address an invigilator if you wish to contact the academic contact. Write down the question in advance.

Weighting: Maximum marks per question are given with the question

Notifications: If there is a need to send a message to the candidates during the exam (e.g. if there is an error in the question set), this will be done by sending a notification in Inspera. A dialogue box will appear. You can re-read the notification by clicking the bell icon in the top right-hand corner of the screen.

Withdrawing from the exam: If you become ill or wish to submit a blank test/withdraw from the exam for another reason, go to the menu in the top right-hand corner and click "Submit blank". This cannot be undone, even if the test is still open.

Access to your answers: After the exam, you can find your answers in the archive in Inspera. Be aware that it may take a working day until any hand-written material is available in the archive.



Biologists have become interesed in how the sizes of species have changed over time: are they becoming larger or smaller? And are different species changing in different ways?

There is a lot of data in museums that can be used to answer these question. We will look at a small part of a data full set of body sizes of birds in North America. The data plotted here is for four species of finch, and we will look at these.

1 Confidence Interval of Year Effect

An analysis with just year gave an estimate of the slope of -0.11 with a standard error or 0.020.

What is the 95% confidence interval for the slope?

(you can assume a normal distribution, and the critical value is 1.96)

Lower: -0.15 .

These are -0.11 +/- 1.96*0.020

Upper: -0.07

.

Maximum marks: 4

The data for this set of problems is from this paper:
Zheng et al. (2023) Increases in intraspecific body size variation are common among North American mammals and birds between 1880 and 2020. Nature Ecology & Evolution 7: 347–354 https://www.nature.com/articles/s41559-022-01967-w.
They were interested in the variation rather than the mean, but we haven't looked at those sorts of model.

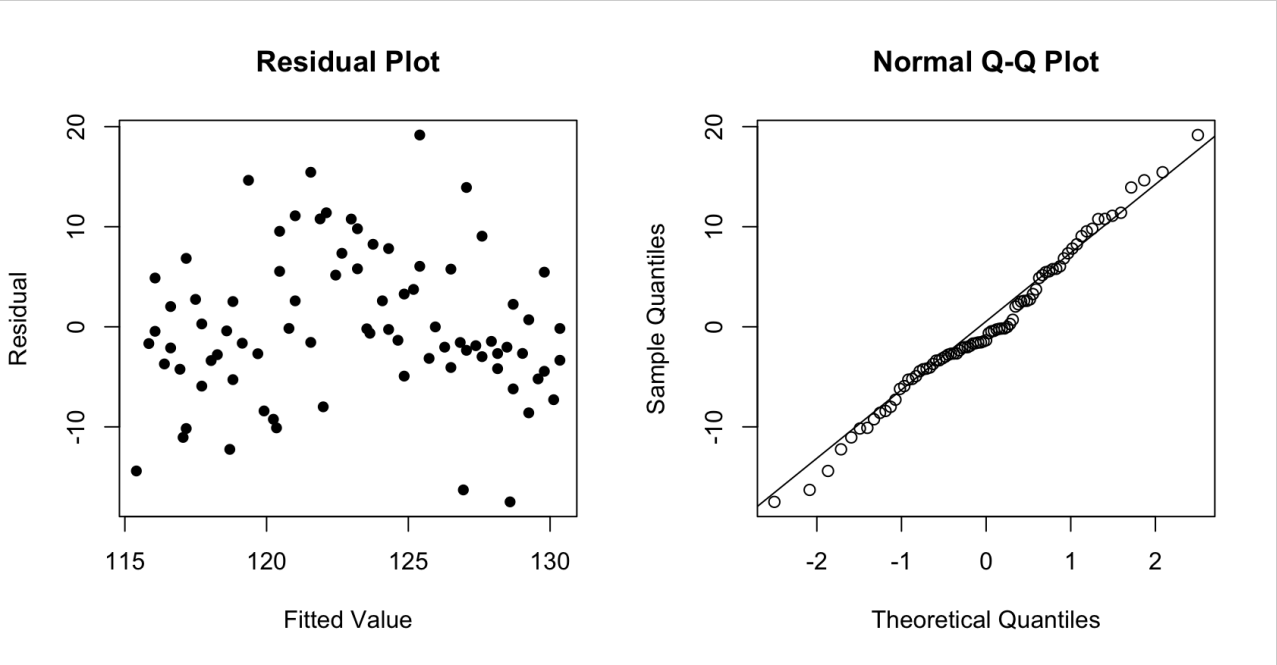
2 Body Size Change

How do you think body size is changing over time?
Select one alternative:

- Staying the same
- Decreasing
- Increasing
- Not sure: could be doing anything

Maximum marks: 1

3 Residual Checking



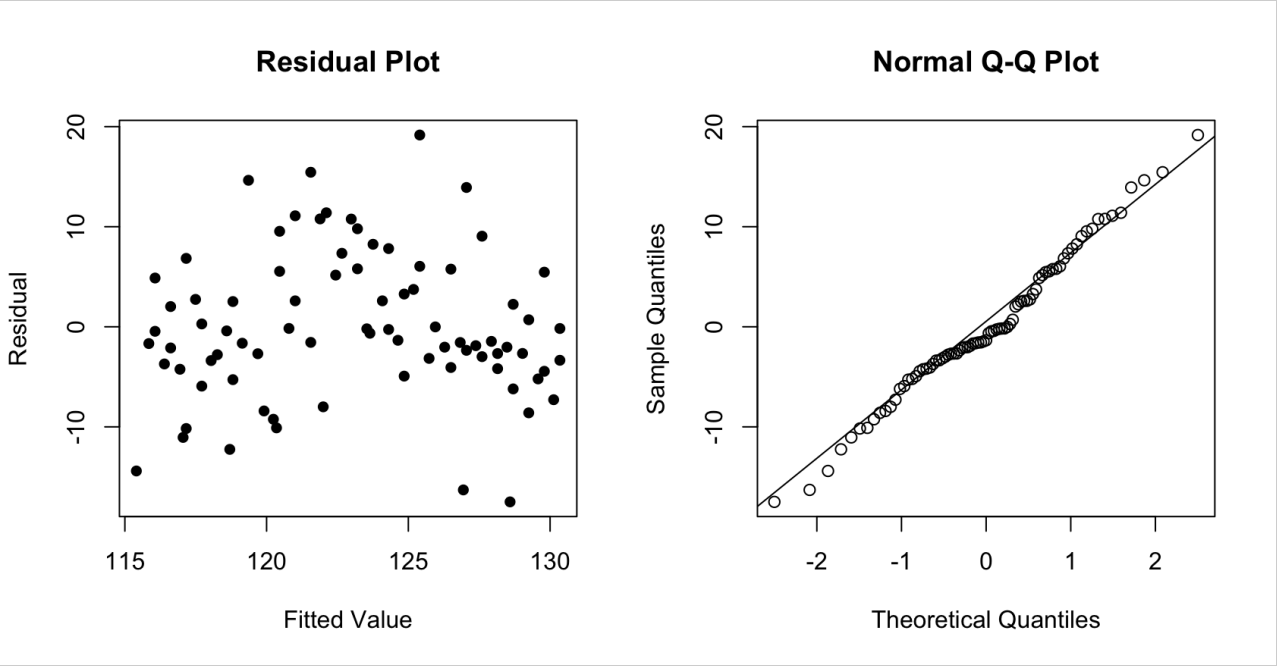
The simple model was checked to see if the assumptions of the model were reasonable by creating the residual plots you see here.

From studying this residual plot, which assumptions do you think are violated?
Select one or more alternatives:

- [ ] Normality
- [x] Linearity
- [ ] Outliers
- [ ] Heteroscedasticity
- [ ] It looks OK

Maximum marks: 2

4 What to do?



If there was one thing you would do to improve the model, what would it be?
Select one alternative:

- Add a quadratic term for Year
- Add an interaction
- Use a Box-Cox transformation
- Remove some outliers

Maximum marks: 2

```
Call:
lm(formula = mean ~ YearFrom1900 + Species, data = CardDat)

Residuals:
    Min      1Q  Median      3Q     Max
-9.1675 -3.5557 -0.5445  2.4257 15.9297

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               134.51956    1.25792 106.938  < 2e-16 ***
YearFrom1900               -0.09514    0.01442  -6.597 4.98e-09 ***
SpeciesCarduelis_pinus     -8.55261    1.51712  -5.637 2.80e-07 ***
SpeciesCarduelis_psaltria -18.24452    2.14958  -8.487 1.28e-12 ***
SpeciesCarduelis_tristis   -7.54649    1.48940  -5.067 2.76e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.197 on 76 degrees of freedom
Multiple R-squared:  0.6471,    Adjusted R-squared:  0.6285
F-statistic: 34.83 on 4 and 76 DF,  p-value: < 2.2e-16
```

Because there are 4 species, a model including all of them was fitted. This is the summary from that model.

Note: the year effect is YearFrom1900, i.e. year - 1900, so (for example) data from 1974 would be coded as 74.

5 Body Size Slope

What is the slope of the effect of Year?

-0.095

Maximum marks: 1

6 Body Size Prediction

For an individual of species *C. pinus*, what is the predicted size of an individual in 1970? (answer to one decimal place)

119.0 .

Maximum marks: 2

7 Where is the missing species?

There are 4 species, but only 3 are listed in the summary table. Why is *C. flammea* (the fourth species) not listed?

Fill in your answer here

Format

The missing species is the intercept level: the species effects listed are the contrasts to that level (i.e. the difference between the intercept for the species listed and the intercept for C. flammea.)

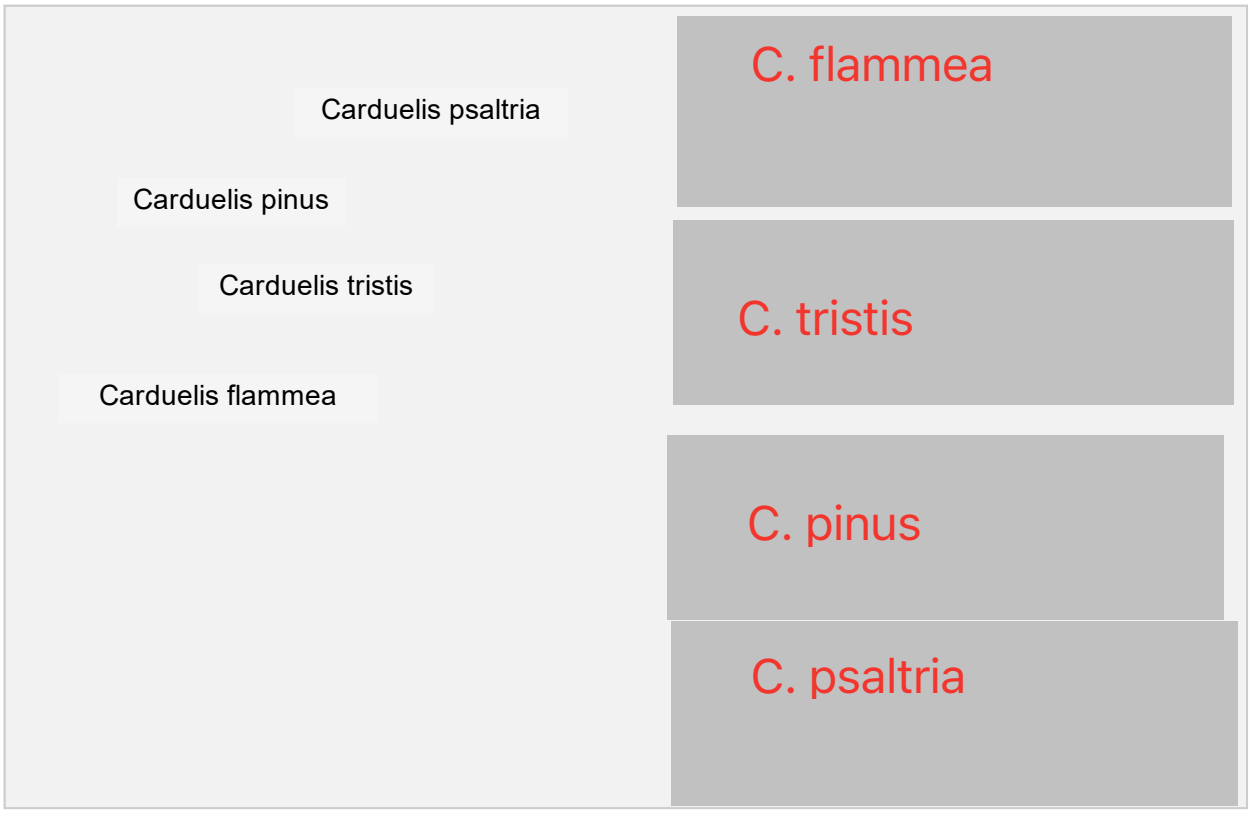
Words: 0

Maximum marks: 4

8 Order of Species

What is the order of the species' sizes?

Order the species from largest (drop area 1, at the top) to smallest (drop area 4, at the bottom)



Maximum marks: 4

9 How Certain is the Order?

Are there any pairs that you are not (fairly) certain that they are different? (i.e. chose the species pairs that you think the data suggests could be the same)
Select one or more alternatives:

- [ ] The order of C. pinus and C. psaltria might be wrong
- [ ] No, the order is pretty certain
- [ ] The order of C. psaltria and C. tristis might be wrong
- [X] The order of C. pinus and C. tristis might be wrong

Maximum marks: 2

10 How does the order affect the year effect?

How does the effect of year change the order of the species you gave in Question 8?

Fill in your answer here

Format

It doesn't. The year effect is the same for every species

Words: 0

Maximum marks: 2

11 Size By Species Interaction

```
> anova(card.main, card.int)
Analysis of Variance Table

Model 1: mean ~ Year + Species
Model 2: mean ~ Year * Species
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     76 2052.4
2     73 2050.2  3    2.2126 0.0263 0.9942
```

The change in size might be different for different species. This was tested with an ANOVA, which gave the output you can see here

Do you think the size change differs between species?

Select one alternative:

- ✗ No, the change seems to be the same in all species
- Yes, the change does seem to be different in different species
- The results are not clear: they may or may not be different.

Maximum marks: 2

12 What would you do next?

We have looked at 4 related species, but there are obviously many more. If you Wanted to look at all of the bird species with enough data (about 80 in the database being used), (a) what would you do, and (b) what would you be looking for?

Fill in your answer here

Format

This is open, so
(a) There is a bunch of stuff, so the possibilities are:
1. Fit the model with all species
2. Add Genus to the model
3. Test a Year: Species (or Year:Genus) interation with all species
4. Add traits, non-linear effects, or something similar

(b)
1. The intercepts will vary more (because some birds are bigger than others)
2. The interaction is more likely to be important, i.e. the variation in the slope between species will be bigger.
3. Any term not interacting with species will be more certain
4. Traits will probably have an effect: non-linear tersm too.

Words: 0

Maximum marks: 8

Some scientists have been testing whether giraffe have abilities as statisticians. Specifically, they tested whether giraffe can estimate the proportion of their preferred food (carrots) in containers holding carrots and zucchini.

The giraffe were shown containers with different proportions of carrots and zucchini, and then were offered a handful of food from each container, in a way that hid what was in the handful. if they were good statisticans, they would chose the handful from the container with more carrots.

Several slightly different experiments were done, to test different details of the process (some will be mentioned later, for now we can treat them as equivalent).

The predictors were

- **Subject**: which giraffe was tested (4 giraffe were used)
- **Side chosen**: whether the preferred food was in the Left or Right hand
- **Experiment**: which experiment was done (8 were done)
- **Trials**: each experiment was tried several times on each giraffe, so this is the order of the trial.

The response is whether the giraffe chose their preferred food: if they were able to estimate the proportion correctly, the probability of doing this should be above 0.5.

13 Intercept Estimates

A generalised linear model assuming a binomial response and a logit link function was used to test whether the data suggested giraffe could estimate proportions in the different experiments.

The simplest model (response~1) only has an intercept. For this model the estimate for the intercept is 1.17, with a confidence interval of 0.99 to 1.36. These are on the scale of the linear predictor.

What is the probability that a giraffe would chose the correct reward?

0.76 (answer to 2 decimal places)

What is the 95% confidence interval?

Lower: 0.73 .

Upper: 0.80 .

Maximum marks: 3

14 Interpreting the Intercept

Does this suggest that giraffes can chose the best food more often than by chance? Explain your answer.

Fill in your answer here

Format

Yes, it does suggest giraffe can chose their best food more often than by chance. The estimated probability is above 0.5 by some distance, and the confidence interval is a long way from 0.5, so this data are really really unlikely to have occured if the probability really waas 0.5.

Words: 0

Maximum marks: 4

The predictors were

- **Subject**: which giraffe was tested (4 giraffe were used)
- **Side chosen**: whether the preferred food was in the Left or Right hand
- **Experiment**: which experiment was done (8 were done)
- **Trials**: each experiment was tried several times on each giraffe, so this is the order of the trial.

A model with the main effects would have all of these, i.e. the formula would be

Response ~ Subject + Side + Experiment + Trials

The response is whether the giraffe chose their preferred food: if they were able to estimate the proportion correctly, the probability of doing this should be above 0.5.

15 Interaction Test

```
> anova(NoInteraction, Interaction, test="LRT")
Analysis of Deviance Table

Model 1: response ~ Experiment + Trial + Side + Animal
Model 2: response ~ Experiment * Trial + Side + Animal
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       587     615.09
2       580     611.32  7   3.7723   0.8056
```

The authors started from the model with the main effects by looking to see if there was an interaction between Experiment and Trial. This is the ANOVA test they used.

Which of these statements is an accurate summary of what we can conclude from this test?

Select one alternative:

- The data is too extreme to be likely if the null hypothesis was true
- X The data is not extreme, and so could have been obtained if the null hypothesis was true
- The null hypothesis is probably true
- The alternative hypothesis is probably true

Maximum marks: 1

16 Interpreting an Interaction

If there was an interaction, how would you interpret it?

Remember: Experiment is which experiment was done, and Trial was the order of the trials, so later trials have a higher number. A Trial effect would be a learning effect: a positive value would mean that the giraffe were learning.

Fill in your answer here

Format

The interaction would say that the rate of learning depended on the trial: in some trials they learned quicker than in others.

Words: 0

Maximum marks: 4

17 Model Choice

The Experiment should be in the analysis, but what about the other terms? Each model with Experiment and all combinatiobns of the main effect were fited to the data. These are the model comparison statistics

	df	Likelihood	AIC	BIC
None	8	-309.5	634.9	670.1
Side	9	-309.0	635.9	675.5
Animal	11	-308.2	638.4	686.8
Trial	9	-309.4	636.9	676.5
Animal + Side	12	-307.6	639.2	692.0
Trial + Side	10	-308.9	637.9	681.9
Trial + Animal	12	-308.1	640.3	693.1
Trial + Animal + Side	13	-307.5	641.1	698.3

Which is the best statistic to use to compare the models?

- df
- Likelihood
- AIC
- X BIC

(BIC is best, but AIC is not totally wrong, but we are not using this for prediction)

Which is the optimal model?

- Trial + Animal
- Trial + Side
- Animal
- Trial + Animal + Side
- Side
- X None
- Animal + Side

Maximum marks: 2

```
Call:
glm(formula = response ~ Experiment, family = "binomial", data = xdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1460   0.4590   0.5725   0.6912   1.1073

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.1671     0.2897   0.577 0.564156
Experiment1   1.5675     0.4266   3.675 0.000238 ***
Experiment2   1.4154     0.4053   3.492 0.000479 ***
Experiment3   1.5585     0.4178   3.730 0.000191 ***
Experiment4   0.4083     0.3359   1.215 0.224195
Experiment5   2.0302     0.7996   2.539 0.011119 *
Experiment7   1.1429     0.3983   2.870 0.004109 **
Experiment8   1.6006     0.5012   3.194 0.001404 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 656.65  on 599  degrees of freedom
Residual deviance: 618.92  on 592  degrees of freedom
AIC: 634.92

Number of Fisher Scoring iterations: 4
```

After looking at different models, we will look at the model with only Experiment. The parameter estimates for the model are shown here.

The experimenters carried out 8 different experiments:

Experiments 1 - 3 look at whether giraffes can estimate proportions (in different ways)
Experiments 4 - 5 tested whether they understood barriers (i.e. whether all of the food in a box could be given to the giraffe)
Experiments 6 - 8 tested alternative explanations for the results:
- *experiment 6*: if the giraffes were using smell instead
- *experiment 7*: if the giraffe were using visual cues from the experimenters instead
- *experiment 8*: test if the amount of visible food was the cue being used

You will not have to understand the details of how the experiments were done.

For all of the experiments except experiment 6, a giraffe would select their preferred food more often than by chance (i.e. have a 50% probability of selecting it) if they were using their statistical abilities. Experiment 6 is different: in it they would select their food more often if they were smelling it, and select it by chance if they were using their statistical skills.

18 Why a Change in Intercept

In the model with no effects (i.e. response ~ 1), the intercept was 1.17. Why is it different in this model?

Fill in your answer here

Format

The intecept has changed because before it was at the overall mean probability (on the logit scale!), but now it is the mean for the reference level, which (it turns out) is a long way from the average.

Words: 0

Maximum marks: 3

19 Intercept For Smell

The intercept level is for the experiment where smell was tested. If the giraffe had been using their statistical abilities rather than their sense of smell in the other experiments, they should have chosen their preferred food at random in this experiment (i.e. they would choose it about 50% of the time).

What is the estimated probability that the giraffe chose their preferred food in this experiment?

0.54 (answer to 2 decimal places).

Maximum marks: 1

20 Interpreting the Smell Intercept

Is there any evidence that this intercept differs from 0.5? Explain your reasoning.

Fill in your answer here

Format

No, there is between “none” and “almost no” evidence. The estimate is just above 0.5, but the confidence interval includes 0.5. Thus, it is not unlikely to have got data as extreme as this.

Words: 0

Maximum marks: 4

21 Difference Between Experiment 1 and 6

Experiment 1 is the first test of whether a giraffe is able to estimate proportions. What is the estimated difference between Experiment 1 and Experiment 6, on the linear scale (i.e. without transforming the estimate)?

1.57 (answer to 2 decimal places)

What is the probability that a giraffe in experiment 1 would chose their preferred food?

0.85 (answer to 2 decimal places)

Maximum marks: 2

22 What does Experiment 1 say?

Does Experiment 1 provide evidence that giraffe were able to chose their preferred food in Experiment 1 at a higher rate than by chance, or could this be random?

Fill in your answer here

Format

Yes it does provide evidence. It would provide evidence if the predicted mean (and its confidence interval) was above 0. Here we have a contrast that is positive (this includes the confidence interval), so is above the reference level. The intercept is above 0 (if not significantly so), so the experiment 1 effect must be too.

Words: 0

Maximum marks: 6

23 Can Giraffe Do Statistics?

Do you think giraffe are able to do statistics? i.e. are they able to estimate which box of food had a higher proportion of their preferred food?

We have not examined the experiments other than Experiments 1 and 6 in any detail, but do they change these conclusions? (you can assume the experiments tested what they claim, but there may be other experiments that could be tested)

Fill in your answer here

Format | B I U x_2 x^2 | I_x | | | Ω | Σ |

Yes, it looks like giraffe can do statistics: they are able to chose food from a "good" box more often than by chance (by some way).

The other experiments have similar estimates than experiment 1, which suggests that they do not change these conclusions (unless they were testing something really weird).

Words: 0

Maximum marks: 6