Dirk Metzler SS 2026

STATISTICS FOR EES — EXERCISE SHEET 8

1. `moreqqplots.pdf` shows normal-qqplots for data sampled according to 6 different distributions. Draw the density polygons to show for each of the 6 distributions how it differs from a normal distribution with the same mean and the same variance.

2. A breeder crossed 200 pairs of plants of some crop. The yield of the father plant, the mother plant and the offspring (F1) is given in table yield.csv (in some measuring unit). For the next generation, the breeder crosses plants from the F1 generation, but selects for this only plants of a yield of more than 20.

   (a) Visualize the distributions of the yield in the parent population, in the offspring population before selection and in the offspring population after selection.

   (b) Add lines indicating the mean values to your plot.

   (c) Predict the average yield of the plants of the F2 generation and show it in your plot.

3. *Daphnia* can develop structures, so-called head crests, neckteeth, or helmets, when they sense chemical cues of certain predators. Assume that for a certain *Daphnia* species the probability that an individual develops neckteeth is modeled by a logistic-regression GLM (that is, of type binomial with logit-link) with a linear predictor

$$\eta_i = -10.1 + 4.2 \cdot c_i,$$

   where c_i is the concentration [in some unit] of the chemical cue in the water. Answer the following questions for this model.

   (a) Calculate the probability that a daphnia develops neckteeth if the cue concentration is 2.3.

   (b) Visualize how the probability of a daphnia to develop neckteeth depends on the cue concentration.

   (c) You carry out an experiment with 10 daphias under exact same conditions with a cue concentration of 2.8. Calculate the probabilty that excatly 6 of the daphnias will develop neckteeth.

   (d) You carry out many experiments, each with 10 daphnias in a bottle of water with a cue concentration of 2.6. Calculate the standard deviation describing how the numbers of daphnias that develop neckteeth vary between the bottles. (Still assume that all bottles follow the GLM specified above.)

   (e) For the same experiment as in the question before, calculate the standard deviation of the fractions of daphnia developing neckteeth.

   (f) Assume that the model parameter values -10.2 and 4.2 are not yet known. To estimate these parameter values you carry out a similar experiment as assumed in the previous two questions with 10 daphnias per bottle but with different cue concentrations in the bottles. Let `c[i]` be the cue concentration in bottle i and `k[i]` the number of daphnias in bottle i that developed neckteeth. Give the R command to estimate the parameters of the model described above.

(g) Give an R command that fits a variant of the model allowing that the some unknown, hard-to-control conditions (water quality, temperature,...) may vary between the bottles and my add variation between the bottles in the probability that daphnias develop neckteeth.

4. One morning at certain spot in a forest, bird songs were recorded in three 5-Minutes intervals starting at 6 a.m., 7 a.m. and 8 a.m. For a certain bird species the following tables shows the numbers of male bird calls and female answer calls during these time intervals:

```
> dat
  time calls answers
1    6     3       2
2    7     7       5
3    8     9       3
```

(a) With the following R command a Poisson GLM with log link has been fitted to predict how the numbers of male bird calls in a five-Minutes interval depended on the time of the day.

```
> poismod <- glm(calls~time, dat, family=poisson)
> poismod$coef
(Intercept)        time
 -1.6817850   0.4926147
```

i. Give, according to this model, the formula to calculate the expected number of calls in a five-Minutes intervall for any time (expressed in hours) in the morning of this day.

ii. What was according to this formula the expected value for the number of calls in the five-Minutes interval starting at 7 a.m.?

iii. Given this expected value, what was the probability to observe 7 calls in this time interval?

iv. Calculate the likelihood of the model (with the fitted parameters).

v. Calculate the likelihood of the saturated model with the same data.

vi. Calculate the residual deviance of the fitted model.

vii. Calculate the deviance residuals of the fitted model.

(b) With the following R command a logistic regression model (that is, a binomial GLM with logit link) has been fitted to predict how the fraction of male calls that were answered by female calls depended on the hour.

```
> binmod <- glm(cbind(answers,calls-answers)~time, dat,
+               family=binomial)
> binmod$coef
(Intercept)        time
  6.9361009  -0.9303041
```

i. Give the formula according to this model to calculate for any time (expressed in hours) in the morning of this day the predicted probability of a male call to be answerd.

ii. What was according to this formula the probabilty of a male call at 7 a.m. to be answered?

iii. Given this probability and given that there were 7 male calls at 7 a.m. (and neglecting that some may have been a few Minutes after 7 a.m.), what is the probability that exactly 5 of the calls were answered?

iv. Calculate the likelihood of the model (with the fitted parameters).

v. Calculate the likelihood of the saturated model with the same data.

vi. Calculate the residual deviance of the fitted model.

vii. Calculate the deviance residuals of the fitted model.

5. You fitted a GLM of type Poisson to a dataset with the standard log scaling. To check whether certain model assumptions are fulfilled, you can use a qqnorm plot, but instead of the residuals that you would use in normal linear model, you use residuals that are…

   (a) differences between the predicted Poisson expected value and the observed value.

   (b) for each observation the contribution to the residual deviance, which is based on the log likelihood-ratio compared to a saturated model.

   (c) differences between the linear predictor (that is, the logarithm of the predicted expected value) and logarithm of the observed value.

   (d) the overdispersion rate for each line.

   (e) the square root of each predicted expected value because this is the standard deviation of the corresponding Poisson distribution.

   One and only one answer is correct.

6. Bats of three different species A, B, and C were sampled from 100 different caves and tested for virus infections. With the R command
   `read.csv("https://evol.bio.lmu.de/_statgen/StatEES/bats.csv")`
   the (hypothetical) data can be loaded into R. For each cave the data table contains one line, containing the species that was found there (note that the three species strictly avoid each other and are never found in the same cave), the temperature in °C and the humidity that were measured in the cave, the total number n of sampled bats and the number v of bats for which a virus infection was detected.

   (a) Fit a GLM to model how the infection rate depends on the species, the temperature in the cave and the humidity in the cave.

   (b) Compare a model with pairwise interaction terms between the three variables (species, temperature and humidity) and tests whether it fits the data significantly better than the model without interaction terms.

   (c) Visualize the model predictions in a way that shows how, according to the two models, the bat infection probability depends on the three variables.