Categorical Data
Proposals for Exercises

Theoretical Problems

1. Derive the formula of the Wilson confidence interval.

2. Consider Model II: Show that the distribution of the counts is a product of multinomial distributions.

3. Verify the asymptotic distribution of the difference $\widehat{\theta}$ of two relative frequencies in Model II.

4. Verify the properties of the odds ratio formulated in the statements 1 to 3 in both models.

5. Prove the asymptotic normality of the log odds ratio in Model I. (Use the delta method.)

6. Consider Model I: Compute $\mathsf{P}(N_{11} = 1|N_{1+} = 3, N_{+1} = 5)$ and $\mathsf{P}(N_{11} = 3|N_{1+} = 3, N_{+1} = 5)$

7. Prove that under the assumption that X and Y are independent the conditional distribution of N_{11} given $N_{1+} = n_{1+}$ and $N_{+1} = n_{+1}$ is the hypergeometric distribution, i.e. verify formula (1).

8. Derive the conditional distribution of N_{11} given $N_{1+} = n_{1+}$ and $N_{+1} = n_{+1}$ for arbitrary ϑ, i.e. verify formula of the Theorem on slide ??

9. Consider Model II with $I = J = 2$ Show that application of the delta method yields

$$\sqrt{n}\left(\log\widehat{\vartheta} - \log\vartheta\right) \xrightarrow{\mathcal{D}} \mathsf{N}(0, \sigma^2)$$

with $\sigma^2 = p_{11}^{-1} + p_{12}^{-1} + p_{21}^{-1} + p_{22}^{-1}$

10. In Model I: Show that the likelihood ratio statistic has the form $\Lambda = 2\sum_{ij} n_{ij} \log\left(\frac{n\, n_{ij}}{n_{i+}n_{+j}}\right)$

11. Show that the logistic regression model is also suitable for the retrospective sampling. Let $\pi(x) = \mathsf{P}(Y = 1|X = x)$ be the probability of the disease of interest satisfying a logistic regression model with parameters α and β. Let Z be the variable indicating whether an individual is sampled: $q_1 = \mathsf{P}(Z = 1|Y = 1)$ and $q_0 = \mathsf{P}(Z = 1|Y = 0)$. Show that

$$\mathsf{P}(Y = 1|Z = 1, X = x) = \frac{\exp(\alpha^* + \beta)}{1 + \exp(\alpha^* + \beta)}$$

with $\alpha^* = \alpha + \log(q_1/q_0)$. That is $\mathsf{P}(Y = 1|Z = 1, X = x)$ has the same coefficient β, the intercept depends on the sampling probabilities.

12. For a given value π_0 one has to estimate the value x such that $\pi(x) = \pi_0$. Show that in the logit model $\operatorname{logit}(\pi(x)) = \alpha + \beta x$ an asymptotic confidence interval with coverage probability $(1 - \gamma)$ is given by the set of values x for which

$$\frac{|\widehat{\alpha} + \widehat{\beta}x - \operatorname{logit}(\pi_0)|}{\sqrt{\widehat{\sigma}^2(\widehat{\alpha}) + x^2\widehat{\sigma}^2(\widehat{\beta}) + 2x\widehat{\sigma}^2(\widehat{\alpha}, \widehat{\beta})}} \leq z_{1-\frac{\gamma}{2}}$$

Here $\widehat{\sigma}^2(\widehat{\alpha})$, $\widehat{\sigma}^2(\widehat{\beta})$ and $\widehat{\sigma}^2(\widehat{\alpha}, \widehat{\beta})$ are the estimated asymptotic variances of the parameter estimators; $z_{1-\frac{\gamma}{2}}$ is the $(1 - \frac{\gamma}{2})$-quantile of the standard normal distribution.

13. Consider the log likelihood equation in the logistic regression model with a single explanatory variable taking only the values 0 and 1. Show that the MLE for the parameter β is the empirical log odds ratio.

Practical Problems

1. In Costa Rica, the vampire bat Desmodus rotundus feeds on the blood of domestic cattle. If the bats respond to a hormonal signal, cows in estrous (in heat) may be bitten with a different probability than cows not in estrous. (The researcher could tell the difference by harnessing painted sponges to the undersides of bulls who would leave their mark during the night.)

	In estrous	Not in estrous	Total
Bitten by a bat	15	6	21
Not bitten by a bat	7	322	329
Total	22	328	350

2. (from Agresti) The following table shows data from the 2002 General Social Survey cross classifying a person's perceived happiness with their family income.

	Happiness		
Income	Not Too Happy	Pretty Happy	Very Happy
Above Average	21	159	110
Average	53	372	221
Below Average	94	249	83

Are Happiness and Income independent?

3. To investigate the the rate of sprouted seeds in two different kinds of water rainwater, muddy water, and tap water were used to water 100 seeds each. Then they were checked and noted how many of those seeds sprouted. The result is presented in the following table:

	Sprouted		
Water Type	Yes	No	Total
Rain Water	64	36	100
Muddy Water	74	26	100
Tap Water	60	40	100
Total	198	102	300

The question is whether the type of the water has an effect on the probability for a seed to sprout.

4. (From Agresti) The following table comes from one of the first studies of the link between lung cancer and smoking, by Richard Doll and A. Bradford Hill. In 20 hospitals in London, UK, patients admitted with lung cancer in the previous year were queried about their smoking behavior. For each patient admitted, researchers studied the smoking behavior of a noncancer control patient at the same hospital of the same sex and within the same 5-year grouping on age. A smoker was defined as a person who had smoked at least one cigarette a day for at least a year.

   (a) Identify the response variable and the explanatory variable.

   (b) Identify the type of study this was.

   (c) Can you use these data to compare smokers with nonsmokers in terms of the proportion who suffered lung cancer? Why or why not?

   (d) Summarize the association, and explain how to interpret it.

   (e) Apply a suitable test procedure.

	Lung Cancer	
Have Smoked	Not Cases	Control
Yes	688	650
No	21	59
Sum	709	709

5. A company filling grass seed bags wants to evaluate their filling machine. The following distribution is advertised on their bags, where T1-T5 are different kinds of grass seeds:

Kind	T1	T2	T3	T4	T5
Prop.	0.5	0.25	0.15	0.05	0.05

The company wants to check if the seed distribution in the bags fits the advertised distribution. They take a sample of size 1000 and find the following summarized data:

Kind	T1	T2	T3	T4	T5
Count	480	233	160	63	64

Does the seed distribution in the bags correspond correctly to the advertised distribution?

6. Consider the example ”Blood pressure” discussed in the slides.

   (a) We can fit a logistic regression model taking as explanatory variables the values

   $$\text{pressure} = x = (111.5, 121.5, 131.5, 141.5, 151.5, 161.5, 176.5, 191.5).$$

   (b) Fit the logistic regression and compare the fitted probabilities with the relative interval frequencies.

   (c) In the dataset `bpchol.r` also the cholesterol level is given. Fit a logistic regression that describes the effect of cholesterol on heart disease.

   (d) Fit a logistic regression model that simultaneously describes the effects of cholesterol and blood pressure on heart disease.

7. On February 28, 1986, the space shuttle Challenger took on the 25th flight in NASA’s space shuttle program. Less than 2 minutes into the flight, the spacecraft exploded, killing all on board. The space shuttle uses two booster rockets to help lift it into the orbit. Each booster rocket consists of several pieces whose joints are sealed with rubber O-rings, which are designed to prevent the release of hot gases produced during combustion. Each booster contains 3 primary O-rings (for a total of 6 for the orbiter). In the 23 previous flights the O-rings were examined for damage. One interesting question is the relationship of O-ring damage to temperature, particularly since it was forecast to be cold on the morning of January 28, 1986 (31^o F). There was a good deal discussion among the engineers the previous day as to whether the flight should go on as planned or not. Use the data given in `SpaceShuttle.r` to estimate failure probability depending on the temperature.

8. A random sample of 100 children aged 3-15 years from a village in Ghana is considered. The children were followed for a period of 8 months. At the beginning of the study, values of a particular antibody were assessed. Based on observations during the study period, the children were categorized into two groups: individuals with and without symptoms of malaria. Use a logistic regression model to estimate the probability of developing malaria depending on age and the log-transformed antibody level. Data in `malaria.r`.

9. Kyphosis (`kyphosis.r`) is a spinal deformity found in young children who have corrective spinal surgery. The incidence of spinal deformities following corrective spinal surgery is thought to be related to the *Age* (in months) at the time of surgery, *Start* (the starting vertebra for the surgery) and *Num* (the number of vertebrae involved in the surgery).

   (a) Plot the binary response for the incidence of Kyphosis versus the age of the child. Fit a simple logistic regression of incidence of Kyphosis on *Age*.

   (b) Fit a quadratic logistic regression model in *Age*. Compare both models.

   (c) Include now the input *Num*.

   (d) Regress on *Age, Num, AgeSq, NumSq, Age*Num.*

(e) Finally include the variable *Start*. Comment!

10. The data set `infection.r` records the occurrence of a human parasitic worm infection in residents of a rural community in China. The variables are: *Age*, the age of the resident, *Infection* (presence (1) or absence (0) of infection) and *Sex* (male (1) or female (2)) Estimate the probability of the occurrence of the infection depending on age. What is the influence of the sex?

11. In the R-package ElemStatLearn in the data set `SAheart` and also in the file `SAheart.r` you find data from a coronary risk-factor study baseline survey, carried out in three rural areas of the Western Cape, South Africa. The aim of the study was to establish the intensity of ischemic heart disease risk factors. The data represent white males between 15 and 64, and the response variable is the presence or absence of myocardial infarction at the time of the survey, There are 160 cases in our data set, and a sample of 302 controls.

    Find the significant risk factors!

12. Consider the dataset `womlab.r`, which is based on the dataframe `Womenlf` from the package carData about Canadian women's labour-force participation. As response we consider, whether a woman is working outside home (yes or no) and as explanatory variables the factor children in the household (present or absent) and the income of the husband (in 1000$).

    Fit a logistic regression model.