Part A Statistics TT 2004 **Exercises** Dr. Gesine Reinert

Note: Exercises 1-8 refer to Lectures 1-4; Exercises 9-16 refer to Lectures 5-8. In particular Question 16 refers to Lecture 8.

1. If gene frequencies are in equilibrium, the genotypes AA, Aa and aa occur in a population with frequencies $(1-\theta)^2$, $2\theta(1-\theta)$, and θ^2, according to the so-called *Hardy-Weinberg equilibrium*. In a sample from the chinese population of Hong Kong in 1937, blood types occurred with the following frequencies, where M and N are erythrocyte antigens.

	M	MN	N	Total
Frequency	342	500	187	1029

Test the hypothesis that the multinomial distribution is as specified by the Hardy-Weinberg equilibrium, with unknown parameter θ, against the alternative that the multinomial distribution does not have probabilities of that specified form. Use a 5% significance level. (Hint: If X_2, X_2 and X_3 are the counts in the three cells, show that the maximum likelihood estimator is $(2X_3 + X_2)/(2n)$.)

2. Parallel surveys of 506 women in 1993 and 295 women in 1994, who were in technical and scientific professions, asked whether various family-oriented benefits existed in their workplaces. One of the benefits of interest was family leave. In 1993, 63% of the women said yes, they did have this benefit, 19% said no, and 18% said that they were not sure. In 1994 the distribution of reponse was 83% yes, 8% no, and 9% not sure. Is there a significant difference between years in the response offered by the professional women? Use an appropriate χ^2-test, taking into account that the observed percentages are only estimates of the true population percentages.

3. In 30 settings of gill nets in a lake, these counts were obtained.

Number of fish	0	1	2	3	4
Frequency	12	10	6	0	2

Assess the question whether a Poisson distribution describes the distribution of fish, using an appropriate significance test.

4. Show that if $Y_1, \ldots, Y_k$ are independent Poisson variables, where Y_i has the Poisson distribution with parameter np_i, for $i = 1, \ldots, k$, with $\sum_{i=1}^k p_i = 1$, then their conditional joint distribution given that $\sum_{j=1}^k Y_j = n$ is multinomial with parameters $(n; p_1, \ldots, p_k)$.

   Suppose that gill nets in a lake are set until 30 fish are caught. The number of settings needed turns out to be 30. The counts obtained are as in Exercise 3. Assess the question whether a Poisson distribution describes the distribution of fish in a catch, using an appropriate significance test.

5. 185 people are randomly surveyed about the number of hours each week they spend watching television. The following table presents the data, using interval notation.

hours spent	number of people
$[0, 5)$	6
$[5, 10)$	4
$[10, 15)$	12
$[15, 20)$	15
$[20, 25)$	28
$[25, 30)$	38
$[30, 35)$	34
$[35, 40)$	22
$[40, 45)$	15
$[45, 50)$	6
more than 50	5

(a) Use a χ^2-goodness of fit test to assess the null hypothesis that the unbinned data come from a normal distribution with mean 30 and variance 100.
(b) Use a χ^2-goodness of fit test to assess the null hypothesis that the unbinned data come from some normal distribution, assuming that the sample mean was 30, and the sample variance was 100. How do the degrees of freedom change compared to (a)?

6. Show that in the case of a population with just two categories, with p being the probability of the first category, Pearson's χ^2-test for testing $p = p_0$ is equal to Z^2, where Z is the usual z-statistic that we would use for this purpose.

7. (a) What, if any, is the effect on Pearson's χ^2-test of increasing the sample size? More specifically, suppose that the number of observations is multiplied by a factor of $c > 1$, and that the category frequencies are controlled to bear the same proportion to one another as in the smaller sample size data set. What change occurs to Pearson's χ^2-statistic? Say something about the power of the test using larger as opposed to smaller sample size.
(b) If the population parameters must be estimated, does that increase or decrease the probability that the null hypothesis of a goodness-of-fit test will be rejected?

8. Prove the following fact, which was used in lecture: If n balls are thrown into m cells, independently of each other, with probability p_i for landing in cell i, $i = 1, \ldots, m$, and if x_i is the ball count in cell i, for $i = 1, \ldots, m$, then the maximum-likelihood estimate for p_i is
$\hat{p_i} = \frac{x_i}{n}$, for $i = 1, \ldots, m$.
Note: You should use Lagrange multipliers for this question. These are covered in *multivariable calculus*. If you do not take multivariable calculus, then give a heuristic of why the above should be true.

9. Find the function $\alpha + \beta x$ (representing a straight line) that best fit these points in the sense of least squares: (0,2), (1,1), (4,3), (5.2).

10. Express each of the following models in matrix notation $\underline{Y} = \mathbf{X}\underline{\theta} + \underline{\epsilon}$, identifying the matrix $\mathbf{X}$ and the vector $\underline{\theta}$ in each case.
(a) $Y_i = \gamma x_i^2 + \epsilon_i$, for $i = 1, \ldots, n$
(b) $Y_i = \alpha x_i + \beta z_i + \epsilon_i$, for $i = 1, \ldots, n$
(c) $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$, for $i = 1, \ldots, n$.

11. For each of the assumed regression functions in Exercise 10, write down the normal equations. Then use the method of least squares to obtain estimators of each of these assumed regression functions.

12. The following data give temperatures corresponding to chirp frequencies of the striped ground cricket.

x (chirps per sec)	20	16	20	18	17	16	15	17	15	16
y (degrees F)	89	72	83	84	81	75	70	82	69	83

Find the least squares line and give 90% confidence limits for the slope.

What is the corresponding R^2?

For $x_0 = 19$ chirps per seconds, predict the temperature and give a 90% prediction interval for the temperature.

For $x_0 = 0$ chirps per seconds, predict the temperature and give a 90% prediction interval for the temperature. Do you find this prediction reasonable? Explain briefly.

13. The observations (x_i, Y_i) satisfy the equation

$$Y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i, \quad i = 1, \ldots, n,$$

where α and β are unknown constants, $\bar{x} = n^{-1} \sum x_i$ and the ϵ_i's are independent normal random variables with mean zero and variance σ^2. The x_i's are not all equal. Derive the covariance matrix of the least squares estimators of α and β. Consider estimating the value of Y when $x = 0$. What estimate would you use, and what is the appropriate variance?

Now suppose that Y_i depends on two explanatory variables x_i and z_i according to the model

$$Y_i = \alpha + \beta(x_i - \bar{x}) + \gamma z_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where the vectors $(x_1, \ldots, x_n)$, $(z_1, \ldots, z_n)$ and $(1, \ldots 1)$ are linearly independent. Show that the variance of the least squares estimator of β is

$$\frac{\sigma^2 \sum (z_i - \bar{z})^2}{\sum (x_i - \bar{x})^2 \sum (z_i - \bar{z})^2 - \left(\sum (x_i - \bar{x})(z_i - \bar{z}) \right)^2}.$$

14. Data giving Y, water absorption of wheat flour, and corresponding values of these explanatory variables: x_1 = flour protein percentage, and x_2 = starch damage, were analyzed using a computer software package with results printed out as follows.

Explan. var.	coefficient	t	P
constant	11.36	2.93	.0074
x_1	2.2265	5.94	.0000
x_1x_2	-.04303	2.43	0.231
x_2	.80962	4.13	.0004

Overall F: 272.76 ($P = .0000$)
MSE: 1.0011 (d.f.= 24)
R^2: .9715

(a) What was the model fitted?
(b) Which of the coefficients in the fitted regression model are significantly different from 0, at the 5% level?
(c) Calculate an F-test for the whole regression.
(d) From what is given above, find the number of data triplets, the number of degrees of freedom for regression, and the regression sum of squares.

15. The following data were reported in a National Institute of Health study. The first component of each pair is the midpoint of an age group classification, and the second is the percentage of people in that age group who have ever used marijuana. First use a linear function to model the dependence of lifetime prevalence of marijuana use on age group. Check appropriate residual graphs, then try another function. Are there still ptoential problems with your revised model?
$(18, 33), (19.5, 42), (21.5, 53), (23.5, 60), (25.5, 64), (27.5, 68), (29.5, 70), (31.5, 71)$.

16. Given are divorce rates in several countries, grouped by continent. Test at level 5% for a difference in divorce rate by continent.

Oceania:	1.1	2.5	1.0	2.7
Africa:	7.0	1.4	0.6	0.7
Asia:	1.3	1.6	1.2	0.8
Europe:	0.8	2.1	3.4	1.9