Tentamentsskrivning i Statistisk slutledning MVE155/MSG200, 7.5 hp.

Tid: tisdagen den 17 mars, 2015 kl 14.00-18.00
Examinator och jour: Serik Sagitov, tel. 772-5351, mob. 0736 907 613, rum H3026 i MV-huset.
Hjälpmedel: Chalmersgodkänd räknare, **egen** formelsamling (fyra A4 sidor) samt utdelade tabeller.
CTH: för "3" fordras 12 poäng, för "4" - 18 poäng, för "5" - 24 poäng.
GU: för "G" fordras 12 poäng, för "VG" - 20 poäng.
Inclusive eventuella bonuspoäng.

Partial answers and solutions are also welcome. Good luck!

1. (5 points) Consider the problem of comparison of two simple hypotheses H_0: $p = p_0$, H_1: $p = p_1$ with $p_1 > p_0$ using the large-sample test for the proportion.

a. Let Y have a binomial distribution with parameters (n, p). The power function of the one-sided test is given by

$$\mathrm{Pw}(p_1) = \mathrm{P}(\tfrac{Y-np_0}{\sqrt{np_0q_0}} \geq z_\alpha \,|\, p = p_1).$$

Explain in detail all parts of this formula.

b. Suppose we want to plan for the sample size n to control the sizes of two types of error at levels α and β. Derive the following formula for the optimal sample size

$$\sqrt{n} = \frac{z_\alpha\sqrt{p_0q_0} + z_\beta\sqrt{p_1q_1}}{|p_1 - p_0|}.$$

Hint: under the alternative hypothesis, $\frac{Y-np_1}{\sqrt{np_1q_1}}$ is approximately normally distributed with parameters (0,1).

c. What happens to the planned sample size if the alternatives are very close to each other? What happens if we decrease the levels α and β?

2. (5 points) A sports statistician studied the relation between the time (Y in seconds) for a particular competitive swimming event and the swimmer's age (X in years) for 20 swimmers with age ranging from 8 to 18. She employed quadratic regression model and obtained the following result

$$\hat{Y} = 147 - 11.11X + 0.2730X^2.$$

The standard error for the curvature effect coefficient was estimated as $s_{b_2} = 0.1157$.

a. Plot the estimated regression function. Would it be reasonable to use this regression function when the swimmer's age is 40?

b. Construct a 99 percent confidence interval for the curvature effect coefficient. Interpret your interval estimate.

c. Test whether or not the curvature effect can be dropped from the quadratic regression model, controlling the α risk at 0.01. State the alternatives, the decision rule, the value of the test statistic, and the conclusion. What is the P-value of the test?

3. (5 points) In the Bayesian estimation framework we search for an optimal action

$$a = \{\text{assign value } a \text{ to unknown parameter } \theta\}.$$

The optimal choice depends on the particular form of the loss function $l(\theta, a)$. Bayes action minimizes the posterior risk

$$R(a|x) = \int l(\theta, a)h(\theta|x)d\theta \quad \text{or} \quad R(a|x) = \sum_{\theta} l(\theta, a)h(\theta|x).$$

a. Explain the meaning of the posterior risk function. What does $h(\theta|x)$ stand for? How is $h(\theta|x)$ computed?

b. The zero-one loss function is defined by $l(\theta, a) = 1_{\{\theta \neq a\}}$. Compute the posterior risk using the discrete distribution formula. Why is it called the probability of misclassification?

c. What Bayesian estimator corresponds to the optimal action with the zero-one loss function? Compare this estimator to the maximum likelihood estimator.

4. (5 points) See the picture to the right.
From this observation we would like to estimate the amount of work required to clean a street from chewing gums.

a. Describe a Poisson distribution model suitable for this particular observation. Summarize the data in a convenient way.

b. Write down the likelihood function for this particular observation.
Find the maximum likelihood estimate.

c. Without performing the required statistical test describe how to check whether the Poisson model fits to the data.

d. Estimate the proportion of tiles free from chewing gums using the fitted Poisson model.

5. (5 points) Miscellaneous questions.

a. Describe a situation when a stratified sampling is more effective than a simple random sampling for estimating the population mean. Which characteristics of the strata will influence your sample allocation choice?

b. Given a dataset how do you compute kurtosis? What is the purpose of this summary statistic? Why is it important to compute the coefficient of skewness for a proper interpretation of the kurtosis value?

c. What is the difference between the parametric and non-parametric bootstrap methods?

d. Suppose we are interested in the average height for a population of size 2,000,000. To what extend can a sample of 200 individuals be representative for the whole population?

6. (5 points) Three different varieties of tomato (Harvester, Pusa Early Dwarf, and Ife No. 1) and four different plant densities (10, 20, 30, and 40 thousands plants per hectare) are being considered for planting in a particular region. To see whether either variety or plant density affects yield, each combination of variety and plant density is used in three different plots, resulting in the following data on yields:

Variety	Density 10,000	Density 20,000	Density 30,000	Density 40,000	mean
H	10.5, 9.2, 7.9	12.8, 11.2, 13.3	12.1, 12.6, 14.0	10.8, 9.1, 12.5	11.33
Ife	8.1, 8.6, 10.1	12.7, 13.7, 11.5	14.4, 15.4, 13.7	11.3, 12.5, 14.5	12.21
P	16.1, 15.3, 17.5	16.6, 19.2, 18.5	20.8, 18.0, 21.0	18.4, 18.9, 17.2	18.13
mean	11.48	14.39	15.78	13.91	13.89

a. Fill in the ANOVA table for the missing numbers

Source of variation	SS	df	MS	F
Varieties				
Density				
Interaction	8.03			
Errors	38.04			

b. Clearly state the three pairs of hypotheses of interest. Test them using the normal theory approach.

c. Estimate the noise size σ.

Statistical tables:

1. Normal distribution table (attached)
2. Chi-square distribution table (attached)
3. t-distribution table (attached)
4. 95% percentiles of the F_{n_1,n_2} distribution (extrapolate for the missing values):

	$n_1 = 1$	$n_1 = 2$	$n_1 = 3$	$n_1 = 6$	$n_1 = 10$	$n_1 = 12$	$n_1 = 15$	$n_1 = 20$	$n_1 = 24$
$n_2 = 1$	161.4	199.5	215.7	234.0	241.9	243.9	245.9	248.0	249.1
$n_2 = 2$	18.51	19.00	19.16	19.33	19.40	19.41	19.43	19.45	19.45
$n_2 = 3$	10.13	9.55	9.28	8.94	8.79	8.74	8.70	8.66	8.64
$n_2 = 6$	5.99	5.14	4.76	4.28	4.06	4.00	3.94	3.87	3.84
$n_2 = 10$	4.96	4.10	3.71	3.22	2.98	2.91	2.85	2.77	2.74
$n_2 = 12$	4.75	3.89	3.49	3.00	2.69	2.62	2.53	2.54	2.51
$n_2 = 15$	4.54	3.68	3.29	2.79	2.54	2.48	2.40	2.33	2.29
$n_2 = 20$	4.35	3.49	3.10	2.60	2.35	2.28	2.20	2.12	2.08
$n_2 = 24$	4.26	3.40	3.01	2.51	2.25	2.18	2.11	2.03	1.98

NUMERICAL ANSWERS

1a. The null distribution of Y is approximately normally distributed with parameters (np_0, np_0q_0), where $q_0 = 1 - p_0$. At the significance level α, the rejection region for the one-sided alternative is $\frac{Y-np_0}{\sqrt{np_0q_0}} \geq z_\alpha$. The power function is the probability of rejecting the null hypothesis given the alternative one is true

$$\mathrm{Pw}(p_1) = \mathrm{P}(\tfrac{Y-np_0}{\sqrt{np_0q_0}} \geq z_\alpha \,|\, p = p_1).$$

1b. To compute the required sample size observe first that

$$\beta = \mathrm{P}(\tfrac{Y-np_0}{\sqrt{np_0q_0}} < z_\alpha \,|\, p = p_1) = \mathrm{P}(\tfrac{Y-np_1}{\sqrt{np_1q_1}} < \tfrac{z_\alpha\sqrt{p_0q_0}+\sqrt{n}(p_0-p_1)}{\sqrt{p_1q_1}} \,|\, p = p_1).$$

Now, since under the alternative hypothesis Y is approximately normally distributed with parameters (np_1, np_1q_1), we get

$$\beta \approx \Phi(\tfrac{z_\alpha\sqrt{p_0q_0}+\sqrt{n}(p_0-p_1)}{\sqrt{p_1q_1}}).$$

This leads to the equation

$$\tfrac{z_\alpha\sqrt{p_0q_0}+\sqrt{n}(p_0-p_1)}{\sqrt{p_1q_1}} = -z_\beta,$$

which brings the desired formula for the optimal sample size

$$\sqrt{n} = \frac{z_\alpha\sqrt{p_0q_0} + z_\beta\sqrt{p_1q_1}}{|p_1 - p_0|}.$$

1c. If the alternatives are very close to each other, the denominator goes to zero and the sample size becomes very large. This is very intuitive as it becomes more difficult to distinguish between two close parameter values.

If we decrease the levels α and β, the values z_α and z_β from the normal distribution table become larger and the sample size will be larger as well. Clearly, if you want have more control over both types of errors, you have to pay by collecting more data.

2a. The underlying parabola makes unrealistic prediction that $\hat{y}_{40} = 139$ sec compared to $\hat{y}_{10} = 63$ sec and $\hat{y}_{20} = 34$ sec. One should be careful to extend the range of explanatory variable from that used in the data.

2b. Using $t_{17}(0.005) = 2.898$ we get the exact confidence interval (under the assumption of normality and homoscedasticity for the noise component)

$$0.2730 \pm 2.898 \cdot 0.1157 = (-0.0623, 0.6083).$$

2c. Since the confidence interval from 2b covers zero, we reject the null hypothesis $H_0 : \beta_2 = 0$ at the 1% significance level. The observed t-test statistic $\frac{0.2730}{0.1157} = 2.36 \in (2.110, 2.567)$, and according to the t_{17}-distribution table says that the two-sided P-value is between 2% and 5%.

3a. For a given action a, the posterior risk function

$$R(a|x) = \sum_\theta l(\theta, a)h(\theta|x) = E(l(\Theta, a)|x).$$

is the expected loss when the unknown parameter θ is treated as a random variable Θ with the posterior distribution:

$$P(\Theta = \theta|x) = h(\theta|x).$$

3b. For the 0-1 loss function in the discrete distribution case,

$$R(a|x) = \sum_{\theta \neq a} h(\theta|x) = 1 - h(a|x) = P(\Theta \neq a|x)$$

is the probability of misclassification, that is the posterior probability that the chosen action a is different from the true value of the parameter θ.

3c. The corresponding Bayesian estimator minimizing the risk function $R(a|x) = 1 - h(a|x)$ maximizes $h(a|x)$, the posterior probability. It is denoted $\hat{\theta}_{\text{MAP}}$ and called the maximum a posteriori probability estimate. In the case the prior distribution is non-informative, so that the posterior distribution is proportional to the likelihood function, we have $\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{MLE}}$.

4a. The numbers of chewing gums for different tiles are summarized in the form of observed counts

Number of gums per tile	0	1	2	3	4	≥ 5
Counts	11	8	2	0	1	0

with the total number of tiles $n = 22$. The Poisson model assumes that the number of gums $X_1, \ldots, X_n$ are independent random variable with the common one-parameter distribution

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, \ldots$$

4b. The likelihood function

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{e^{-n\lambda} \lambda^{n\bar{x}}}{x_1! \cdots x_n!} = \frac{e^{-22\lambda} \lambda^{16}}{2!2!4!}.$$

The log likelihood

$$l(\lambda) = const - 22\lambda + 16 \log \lambda.$$

The equation $l'(\lambda) = 0$ gives

$$0 = -22 + \frac{16}{\lambda}.$$

The MLE becomes $\hat{\lambda} = \frac{16}{22} = 0.73$, which is $\bar{x}$.

4c. Use the Pearson chi-square test of goodness of fit. Combine the cells for 2 and more gums. Compute the expected counts by

$$E_0 = n \cdot e^{-\hat{\lambda}}, \quad E_1 = n \cdot \hat{\lambda} e^{-\hat{\lambda}}, \quad E_2 = n - E_0 - E_1.$$

Then find the test statistic $X^2 = \sum \frac{(O_k - E_k)^2}{E_k}$ and use the chi-square distribution with df $= 3 - 1 - 1 = 1$ table to see if the result is significant. For example, if $\sqrt{X^2} > 1.96$, we reject the Poisson model hypothesis at $\alpha = 5\%$.

4d. Using the Poisson model we estimate p_0 by $\hat{p}_0 = e^{-\hat{\lambda}} = 0.48$, which is close to the sample proportion $\frac{11}{22} = 0.50$.

5a. When the population under investigation has a clear structure it is more effective to use stratified sampling for estimating the overall population mean. In accordance with the optimal allocation formula:

$$n_i = n \frac{W_i \sigma_i}{\bar{\sigma}},$$

the allocation of observations should follow the next two key rules: put more observations in the larger strata, and put more observations in the strata with higher variation.

5b. The sample kurtosis is computed from a sample $(X_1, \ldots, X_n)$ as

$$b_2 = \frac{1}{ns^4} \sum_{i=1}^{n} (X_i - \bar{X})^4,$$

where $\bar{X} = \frac{X_1 + \ldots + X_n}{n}$ is the sample mean and $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ is the sample variance. If the corresponding coefficient of skewness is close to zero and $b_2 \approx 3$, then we get an indication that the shape of the population distribution curve is close to the normal distribution.

5c. Bootstrap is a resampling technique used to study the sampling distribution of a parameter estimator. In the parametric bootstrap resampling is done from the given parametric distribution with the unknown parameters replaced by their estimates obtained from the underlying sample. In the non-parametric bootstrap resampling is performed with replacement directly from the the underlying sample.

5d. The standard error for the sample mean is $s_{\bar{X}} = \frac{s}{\sqrt{200}}$. Roughly: the range of heights $160 - 200$ covers 95% of the population distribution. Then $s \approx 10$ cm and $s_{\bar{X}}$ is something like 0.7 cm. Thus a random sample of size 200 may give a decent estimate of the population mean height.

6a.

Source of variation	SS	df	MS	F
Varieties	328.24	2	164.12	103.55
Density	86.68	3	28.89	18.23
Interaction	8.03	6	1.34	0.84
Errors	38.04	24	1.59	

6b. Using the critical values

$$F_{2,24} = 3.40, \qquad F_{3,24} = 3.01, \qquad F_{6,24} = 2.51,$$

we reject both null hypotheses on the main factors and do not reject the null hypothesis on interaction.

6c. $s = \sqrt{1.59} = 1.26$.