EXAM: Statistical inference (MVE155/MSG200)
Tuesday, August 15, 2023, at 14:00-18:00

Examiner and jour: Aila Särkkä, phone 031 772 3542
Allowed material: Chalmers allowed calculator and your own summary (four A4 pages) of the course.
Passing limits: Chalmers students: 12p for '3', 18p for '4', and 24p for '5'; GU students: 12p for 'G' and 20p for 'VG'.

1. All three questions below (a-c) concern the Bayesian approach to estimate parameters:

   a) Give the idea behind the Bayesian approach.

   b) How do you choose the prior distribution?

   c) What is a conjugate prior? (5p)

2. We investigate the effect of a new drug on 50 different factors by computing 50 95% confidence intervals, one for the mean value of each factor (based on the sample size 40 in each case). Suppose that the drug has no effect on any of the 50 factors. What is the probability that at least two of these 50 confidence intervals show that there would be a significant effect? What assumptions do we have to make to rely on the result? (4p)

3. To investigate whether the proportion of red M&M candies in the plain and peanut variants of the candy is the same, a random sample of each variant was collected. In the sample of 56 plain candies, 12 were red, and in the sample of 32 peanut candies, 8 were red.

   a) Construct a 95% confidence interval for the difference in the proportions of red candies for the plain and peanut variants. Does there seem to be a difference in the proportions of the red candies for the plain and peanut variants? Explain.

   b) Perform a hypothesis test to determine whether there is a significant difference in the proportions of red candies for the two variants. Use the significance level 5% and compare to the result in a).

   c) Give the assumption we need to make in a) and b).

   d) When should we use a non-parametric test instead of the one in b)? Which non-parametric test would be suitable in this case? Describe the idea of that test. (8p)

4. An experiment was conducted to compare the glare characteristics of four types of rearview mirrors of cars. Ten drivers were randomly selected to participate in the experiment. Each driver was exposed to glare produced by a headlight located nine meters behind the rear window of the car. The driver then rated the glare produced by the rearview mirror on a scale 1

(low) to 10 (high). Each of the four mirrors was tested by each driver and the mirrors were assigned to a driver in a random order. An analysis of variance produced the following incomplete ANOVA table:

Source	Df	Sum Sq	Mean Sq	F value
Mirrors	?	12.32	?	?
Drivers	?	?	8.42	?
Error	?	?	?	
Total		158.32		

a) What kind of design (type of ANOVA) has been used?

b) Write down the hypotheses that are tested with the analysis of variance.

c) Fill in the ?'s in the ANOVA table.

d) Interpret the results in the ANOVA table. (Take advantage of the F-distribution table.)

e) Why is the order of the mirrors randomized for each driver?

f) Based on the results in d), what are the practical implications of this experiment for the manufactures of the rearview mirrors? (8p)

5. To pass a course one has to pass an exam and do a project. The teacher has graded the exams and the projects and computed (estimated) the correlation coefficient between the two scores. (The correlation coefficient between two stochastic variables X and Y with standard deviations σ_X and σ_Y is defined as the covariance between X and Y divided by $\sigma_X \sigma_Y$.)

a) What can the teacher use this correlation coefficient for?

b) What conclusions can the teacher draw if the correlation coefficient between the two scores is 0.78?

c) Say, that the correlation coefficient was 0.32 with a 95% confidence interval 0.32 ± 0.43. How can we interpret this result?

d) Which assumptions do we have to make to be able to compute the correlation coefficient and the confidence interval for it? (5p)

Good luck!

Solutions

1. a) The parameter of interest is treated as a random variable generated from some prior distribution $g(\theta)$. Given θ, data have the distribution or likelihood $f(x|\theta)$. The parameter is estimated by finding the posterior distribution $h(\theta|x) \propto f(x|\theta)g(\theta)$ (or $h(\theta|x) = f(x|\theta)g(\theta)/\phi(x)$, where $\phi(x) = \int f(x|\theta)g(\theta)\,d\theta$ or $\phi(x) = \sum \mathrm{P}(X = x|\theta)g(\theta)$ depending on whether X is continuous or discrete). E.g. the mean or median of the corresponding posterior distribution can be used as a point estimate for the parameter.

   b) The prior distribution is chosen by the user/researcher. If we do not have any prior information on the parameter, we can choose an uninformative, uniform prior. If we have some prior information, we can take it into account when choosing the prior. The prior distribution should be chosen before the data are collected.

   c) Let the data be generated from a parametric model having the likelihood $f(x|\theta)$ and let us have a parametric family of prior distributions $\mathcal{G}$. Then, $\mathcal{G}$ is called a family of conjugated priors for the likelihood function $f(x|\theta)$ if for any prior $g(\theta) \in \mathcal{G}$, the posterior $h(\theta|x)$ also belongs to $\mathcal{G}$.

2. For each factor, with probability 95% we have found an interval that covers the "normal" value and therefore, with probability 5% probability the interval does not cover the "normal" value. Let X be the number of intervals that do not cover the "normal" value. Then, $X \sim Bin(50, 0.05)$ and

$$\begin{aligned} P(X \geq 2) &= 1 - P(X \leq 1) = 1 - P(X = 1) - P(X = 0) \\ &= 1 - \binom{50}{0} 0.95^{50} - \binom{50}{1} 0.05 \cdot 0.95^{49} = 0.72. \end{aligned}$$

   We have to assume that the tests based on the confidence intervals are independent which is not the case if they are computed based on data from e.g. the same people in each case.

3. a) Let p_1 and p_2 be the proportions and n_1 and n_2 the number of red candies in the plain and peanut samples, respectively. A 95% confidence interval for the difference can be computed as

$$\hat{p}_1 - \hat{p}_2 \pm z_{0.025}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1 - 1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2 - 1}}.$$

   Here, $\hat{p}_1 = 12/56 = 0.214$, $\hat{p} = 8/32 = 0.25$, $n_1 = 56$, and $n_2 = 32$ resulting in the confidence interval -0.036 ± 0.187, i.e. $(-0.223, 0.151)$. Since the confidence interval contains 0, we cannot say that the proportion of red candies would be significantly different among the two variants.

b) To test the null hypothesis $H_0 : p_1 = p_2$ against $H_1 : p_1 \neq p_2$, we can use the test statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1-1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2-1}}}$$

which is approximatively $N(0, 1)$-distributed if the sample sizes are large enough. The test statistic obtains the value $z = -0.036/0.095 = -0.38$. Using the significance level 5%, we cannot reject the null hypothesis since $|z| < z_{0.025} = 1.96$. Therefore, there is not enough evidence to indicate that the proportions of red candies would be different for the two variants and we get the same result as in a).

c) We have to assume that the two samples are independent and that the sample sizes are large enough (at least 25) so that the test statistic is approximately normally distributed.

d) If the samples were small, we should use a non-parametric test and Fisher's exact test would be an appropriate non-parametric test to test the same pair of hypotheses as in b). In Fisher's test, we summarize the data as a 2×2 table of counts (the two samples as columns and the number of red and non-red candies as rows). The test statistic would be the number of red candies for the plain variant which has hypogeometric distribution $Hg(n_1 + n_2, n_1, p)$, where p is the portion of red candies in the two samples.

4. a) This is a random block design, where drivers are the blocks. The main interest is to compare the rearview mirrors, not the drivers.

b) Let μ_i^M, $i = 1, 2, 3, 4$, be the mean glare ratings for the ith rearview mirror. The main task is to test $H_0 : \mu_1^M = \mu_2^M = \mu_3^M = \mu_4^M$ against H_1 : all the means are not the same.

c) The numbers of degrees of freedom (Df) are $4 - 1 = 3$ for Mirrors, $10 - 1 = 9$ for Drivers, and $3 \times 9 = 27$ for Error. Mean Sq for Mirrors is $12.32/3 = 4.11$ (Sum of Sq/Df) and Sum of Sq for Drivers $8.42 \times 9 = 75.78$ (mean Sq $\times$ Df). Then, Sum Sq for Error is $158.32 - 12.32 - 75.78 = 70.22$ (Total Sum Sq-Mirrors Sum Sq-Drivers Sum Sq) and Mean Sq for Error $70.22/27 = 2.60$. The two F values are computed by dividing Mirrors Mean Sq and Drivers Mean Sq by Error Mean Sq. Finally, we obtain the complete ANOVA table

Source	Df	Sum Sq	Mean Sq	F value
Mirrors	3	12.32	4.11	1.58
Drivers	9	75.78	8.42	3.24
Error	27	70.22	2.60	
Total		158.32		

d) We compare the F value 1.58 to $F_{3,27}(0.05) = 2.96$ and since $1.58 < 2.96$ we cannot reject the null hypothesis that there is no difference between the mirrors at the significance level 5%. We can also compare the F value 3.24 to $F_{9,27}(0.05) = 2.25$ and since $3.24 > 2.25$ there seems to be a significant difference between the drivers.

e) The order is randomized since, for example, fatigue can effect the result. If all the drivers got the mirrors in the same order and we found differences between the mirrors it would be difficult to say whether the difference would be due to the type of mirror or due to the driver being tired.

f) It does not matter which type of rearview mirror is used since they all give similar results. The manufacturer can, for example, use the cheapest one.

5. a) To measure whether there is a linear relationship between the two scores.

b) It seems that there is a positive linear relationship between the two grades meaning that if one of them increases, also the other one increases. However, without knowing the variance of the estimated correlation coefficient, we cannot be sure that the observed correlation is significant.

c) The confidence interval covers zero and therefore, it seems that there is no significant linear relationship between the two grades.

d) We do not have to make any assumptions of the underlying stochastic variables to estimate the correlation coefficient but to estimate a confidence interval, we have to make some assumptions on the distribution of them, typically that they are normally distributed.