**EXAM:** Statistical inference (MVE155/MSG200)
Tuesday, March 14, 2023, at 14:00-18:00
**Examiner:** Aila Särkkä, phone 031 772 3542
**Allowed material:** Chalmers allowed calculator and your own summary (four A4 sidor) 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. a) Derive the maximum likelihood estimator for the population proportion and show that it is an unbiased and consistent estimator for the population proportion.

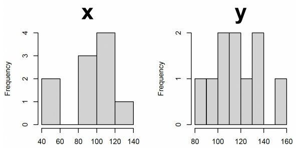
   b) What are the similarities and differences between random sampling and simple random sampling?

   c) If stratified sampling with $k$ strata is used, the sample mean becomes $\bar{X}_s = \sum_{i=1}^{k} w_i \bar{X}_i$, where $w_i$ is the stratum fraction and $\bar{X}_i$ the sample mean in stratum $i$, $i = 1, ..., k$. Compute the variance of $\bar{X}_s$ if the $n$ observations in the sample are allocated to the $k$ strata proportionally to the strata sizes. (6p)

2. A group of nurses wants to study if the number of maternity care visits (before birth) has a positive effect on the birth weight of the child. The data (see the table and histograms below) consist of 10 mothers who had had 5 or fewer maternity care visits before birth and 10 mothers who had had 6 or more. Let $X$ (5 or less) and $Y$ (6 or more) be the birth weights of the babies in these two populations.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $x$ | 49 | 108 | 110 | 82 | 93 | 134 | 114 | 96 | 52 | 101 |
| $y$ | 133 | 108 | 93 | 119 | 106 | 131 | 87 | 153 | 116 | 129 |

($\bar{x} = 93.9$, $s_x = 26.75$, $\bar{y} = 117.5$, $s_y = 19.92$)



   a) Formulate null and alternative hypotheses. Motivate your choice of the alternative hypotesis.

   b) Perform a parametric test corresponding to the hypotheses in a).

   c) Perform a non-parametric test. Reformulate the hypotheses in this case.

   d) Discuss and compare the results in b) and c). Why do they lead to the same/different conclusions?

   e) Which assumptions have you made in b) and c)? Do they seem to be valid? (8p)

3. Assume that we have a sample of size 20 from a normal distribution with mean $\mu$ and variance $\sigma^2 = 4$. We would like to test the null hypothesis $H_0 : \mu = 10$ against $H_1 : \mu \neq 10$ at the significance level 0.05. Assume that the true expected value is 9. What is the probability that the test cannot reveal this (that the true mean is 9 instead of 10)? (4p)

4. We want to study how the choice of detergent and water temperature affect the dirt removal of laundry by using two types of detergents (deter) and three different temperatures (temp) by using the analysis of variance. The values of the response variable, the amount of dirt removed, are given in the following table:

| | Cold | Warm | Hot |
|---|---|---|---|
| Detergent 1 | 4 | 7 | 10 |
| | 5 | 8 | 11 |
| | 5 | 9 | 12 |
| | 6 | 12 | 19 |
| | 5 | 3 | 15 |
| Detergent 2 | 4 | 12 | 10 |
| | 4 | 12 | 12 |
| | 6 | 13 | 13 |
| | 6 | 15 | 13 |
| | 5 | 13 | 12 |

An incomplete ANOVA table is given below:

| | Df | Sum Sq | Mean Sq | F value |
|---|---|---|---|---|
| temp | | 312.47 | | |
| deter | | 12.03 | | |
| temp:deter | | 60.47 | | |
| error | | 114.00 | | |

a) Give the model for the observations that corresponds to the (incomplete) ANOVA table. Give also the hypotheses of interest.

b) Complete the ANOVA table and interpret the results.

c) Give a formula to compute one of the p-values.

d) Give an estimate for the error/noise variance.

e) Discuss the choice of the model. Which assumptions are needed? (7p)

5. We have a sample $(x_1, x_2, ..., x_n)$ drawn from the Poisson distribution Pois$(\lambda)$ and want to estimate $\lambda$ by using Bayesian inference.

a) Show that gamma distribution is a conjugate prior for the Poisson distribution. The density function of $Y \sim$ Gam$(\alpha, \beta)$ is $f(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}$, $y > 0$, and the probability mass function for $X \sim$ Pois$(\lambda)$ is $P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$, $x = 0, 1, ...$

b) What are the advantages of choosing a conjugate prior (if possible)?

c) What are the advantages of having a large sample size? (5p)

**Good luck!**

**Solutions**

1. a) We have a random sample from $\text{Bin}(1, p)$, where $p$ is the population proportion. A maximum likelihood estimator can be found by maximizing the log-likelihood function

$$l(p) = \ln(p) \sum x_i + (n - \sum x_i)\ln(1 - p)$$

giving the maximum likelihood estimator $\bar{X}$, which is the sample proportion. We have that

$$\mathbb{E}(\bar{X}) = \frac{1}{n} \sum \mathbb{E}(X_i) = \frac{1}{n} \cdot np = p$$

and

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \sum \text{Var}(X_i) = \frac{1}{n^2} \cdot np(1 - p) = \frac{p(1 - p)}{n}.$$

The estimator is unbiased since $\mathbb{E}(\bar{X}) = p$ and since the variance goes to zero as $n \to \infty$, it is also consistent.

b) A random sample is taken with replacement and a simple random sample without replacement. In both cases, the observations are identically distributed but only in the random sample case, the observations are independent. If the sample size is small compared to the population size, these two methods are almost the same.

c) In proportional allocation, the number of observations in stratum $i$ is $n_i = nw_i$. Then

$$\begin{aligned} \text{Var}(\bar{X}_s) &= \text{Var}(\sum_{i=1}^{k} w_i \bar{X}_i) = \sum_{i=1}^{k} \text{Var}(w_i \bar{X}_i) \\ &= \sum_{i=1}^{k} w_i^2 \text{Var}(\bar{X}_i) = \sum_{i=1}^{k} w_i^2 \frac{\sigma_i^2}{n_i}, \end{aligned}$$

where $\sigma_i^2$, $i = 1, ..., k$, is the (unknown) variance in stratum $i$. Taking into account that $n_i = nw_i$, we obtain

$$\text{Var}(\bar{X}_s) = \frac{1}{n} \sum_{i=1}^{k} w_i \sigma_i^2 = \frac{\overline{\sigma^2}}{n}.$$

2. a) Let $\mu_x$ and $\mu_y$ be the population means of the two groups. We test

$$H_0 : \mu_x = \mu_y \quad \text{against} \quad H_1 : \mu_x < \mu_y$$

The nurses suspect that the number of visits would affect the birth weight positively, i.e. mothers who have had more visits have heavier babies.

b) The two samples are independent and we use a two-sample t-test with the test statistic

$$t = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}},$$

where

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$$

and $T$ (the stochastic version of $t$) is $t_{n_x+n_y-2}$-distributed under $H_0$. Here, $s_p = 23.58$ and $t = -2.24$. Since the critical value $t_{18}(0.05)$ is $-1.73$, the null hypothesis is rejected at the significance level 0.05. It seems that the number of visits affects the birth weight of the baby positively.

c) Rank sum test to test the hypotheses

$$H_0 : F_x = F_y \quad \text{against} \quad H_1 : F_x \neq F_y$$

where $F_x$ and $F_y$ are the distribution functions of $X$ and $Y$. The rank sum $r_x$ in the $x$-group (observed value of $R_1$) is 81 and the test statistic

$$Z = \frac{R_1 - \mathbb{E}(R_1))}{\sqrt{\text{Var}(R_1)}} \approx N(0, 1)$$

since both sample sizes are at least 10. Here, $\mathbb{E}(R_1) = \frac{1}{2} \cdot n_x(n_x + n_y + 1)$ is 105, $\text{Var}(R_i) = \frac{1}{12} \cdot n_x n_y(n_x + n_y + 1)$ 175, and the test statistic -1.814. The value of the test statistic should be compared to the critical value $z(0.975) = -1.96$. The null hypothesis cannot be rejected at the significance level 0.05 so that there is not enough evidence that the two distributions would be different.

d) The two-sample t-test rejects the null hypothesis but the rank sum test does not. One possible explanation is that in b) the alternative hypothesis is one-sided while in c) it is double-sided making it more difficult to reject $H_0$. (In b), we would have rejected $H_0$ even with a double-sided $H_1$.) If all the assumptions for a parametric test hold, it is always better to use a parametric test since it uses more information on the data than non-parametric tests and makes it easier to reject the null hypothesis.

e) Both tests assume that the two samples are independent. In b), one also has to assume that $X \sim N(\mu_x, \sigma)$ och $Y \sim N(\mu_y, \sigma)$. According to the histograms, the two samples are quite close to being normal but the sample variances are quite different, $s_x^2 = 715$ and $s_y^2 = 397$. To be able to use the normal approximation in the rank sum test, the number of observations in each group should be at least ten which they are.

3. Compute the probability that $H_0$ is not rejected (i.e. type II error), when $\mu = 9$, i.e. compute

$$\begin{aligned} & P(-z_{0.025} < \frac{\bar{X} - 10}{\sigma/\sqrt{n}} < z_{0.025} | \mu = 9) \\ & = P(-z_{0.025} + \sqrt{n}/\sigma < \frac{\bar{X} - 9}{\sigma/\sqrt{n}} < z_{0.025} + \sqrt{n}/\sigma), \end{aligned}$$

where $\frac{\bar{X}-9}{\sigma/\sqrt{n}} = Z \sim N(0,1)$. The probability becomes

$$\begin{aligned} P(0.28 < Z < 4.20) &= \Phi(4.20) - \Phi(0.28) \\ &= 1 - 0.6103 = 0.3897 \end{aligned}$$

(where $\Phi(4.20) \approx 1$).

4. a) The model for the observations is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \sigma Z_{ijk},$$

where $\alpha$ is the effect of temperature, $\beta$ the effect of detergent, $\delta_{ij}$ the interaction effect, $Z \sim N(0,1)$, and $\sigma^2 > 0$ a constant noise variance. We test the hypotheses

$$\begin{aligned} H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0 \quad & \text{against} \quad H_1 : \text{at least one of the } \alpha\text{'s is not } 0 \\ H_0 : \beta_1 = \beta_2 = 0 \quad & \text{against} \quad H_1 : \text{at least one of the } \beta\text{'s is not } 0 \\ H_0 : \delta_{ij} = 0 \text{ for all } i \neq j \quad & \text{against} \quad H_1 : \text{at least one of the } \delta_{ij}\text{'s is not } 0 \end{aligned}$$

b)

| | Df | Sum of Sq | Mean Sq | F value |
|---|---|---|---|---|
| temp | 2 | 312.47 | 156.23 | 32.891 |
| deter | 1 | 12.03 | 12.03 | 2.533 |
| temp:deter | 2 | 60.47 | 30.23 | 6.365 |
| Error | 24 | 114.00 | 4.75 | |

Temperature seems to have a significant effect on removing dirt (compare to $F_{2,24}(0.05) = 3.40$) but not which detergent is used (compare to $F_{1,24}(0.05) = 4.26$). However, there is a significant interaction effect (compare to $F_{2,24}(0.05) = 3.40$).

c) p-value for the effect of temperature is computed as $P(F \geq 32.891 | H_0)$, where $F \sim F_{2,24}$.

d) The error variance can be estimated by $ss_e/df_e = ms_e = 4.75$.

e) This is a two-way ANOVA model with interaction term. The observations in each group should be iid normally distributed and the variances in the groups should be equal.

5. a) The likelihood function is

$$f(x_1, ..., x_n|\lambda) = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod x_i!},$$

the prior distribution

$$g(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

and the posterior distribution

$$h(\lambda|x_1, ..., x_n) \propto e^{-n\lambda}\lambda^{\sum x_i}\lambda^{\alpha-1}e^{-\beta\lambda} = \lambda^{\alpha+\sum x_i-1}e^{-\lambda(\beta+n)}$$

which corresponds to the distribution $\mathrm{Gam}(\alpha+\sum x_i, \beta+n)$. Taking into account the normalizing term

$$\phi(x_1, ..., x_n) = \int f(x_1, ..., x_n|\lambda)g(\lambda)\,d\lambda,$$

you obtain the correct multiplier $(\beta+n)^{\alpha+n\bar{x}}/\Gamma(\alpha+n\bar{x})$ for the posterior distribution. Since the posterior belongs to the same family of distributions as the likelihood, we have shown that gamma distribution is a conjugate prior for Poisson distribution.

b) The advantage of choosing a conjugate prior is that we can easily compute the posterior distribution without computing any complicated integrals. Therefore, e.g. posterior mean is readily available as a point estimator.

c) If the sample size is large, the estimators are more precise (variance of the posterior distribution narrower) and the prior distribution has less effect. Also, frequentistic (ML) and Bayesian approaches typically give similar results.