Tentamentsskrivning i Statistisk slutledning MVE155/MSG200, 7.5 hp.

Tid: 9 juni 2020, kl 8.30-12.30
Examinator och jour: Serik Sagitov, available via Zoom
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.

This is a home-written exam.
The examination must be conducted individually, that is, COOPERATION IS NOT ALLOWED. Otherwise, all aids are allowed. Please mention what help have you used for each of your answers (book, lecture notes, Google, and so on). Checks on plagiarism will be carried out.

---

1. (5 points) Miscellaneous questions.

(a) Suppose a computer program has generated 100 random numbers $(x_1, \dots, x_n)$ according to the Beta(a, b) distribution with $a = b = 0.01$. How the histogram plotting these 100 numbers is expected to look like? Does it remind a certain discrete distribution? Explain the shape referring to the density formula.

(b) Provided you have resources to perform 1000 tests for the corona virus, how would you estimate the number of infected people in Sweden?

(c) An experimental drug is to be evaluated on laboratory rats. In 20 pairs of litter mates, one animal is given the drug and the other animal is given a placebo. A physiological measure of benefit is made after some time has passed. It was found that in 15 pairs the animal receiving the drug benefited more than its litter mate. Does the drug have an effect? Compute the p-value of your statistical test.

2. (6 points) A random sample $(x_1, \dots, x_n)$ is drawn from a normal population distribution N$(\mu, 1)$. Consider two alternative composite hypotheses

$$H_0 : \mu \leq 0, \qquad H_1 : \mu > 0.$$

(a) Rewrite H_0 and H_1 as a pair of nested hypotheses H_0 and H.

(b) Demonstrate that the likelihood function satisfies

$$L(\mu) \propto \exp\{-\tfrac{n}{2}(\mu - \bar{x})^2\},$$

where $\bar{x} = \frac{x_1 + \dots + x_n}{n}$ and $\propto$ means "proportional to".

(c) Show that the (generalised) likelihood ratio has the form

$$\Lambda = \begin{cases} 1 & \text{if } \bar{x} < 0, \\ e^{-\frac{n}{2}\bar{x}^2} & \text{if } \bar{x} \geq 0 \end{cases}$$

(d) Explain why the rejection region of the likelihood ratio test can be expressed as

$$\mathcal{R} = \{\bar{x} > c_\alpha\},$$

and show for $\alpha = 0.05$ and $n = 25$, that

$$\mathcal{R} = \{\bar{x} > 0.33\},$$

using the rule

$$\alpha = \max_{\mu \in H_0} \mathrm{P}(\bar{X} > c_\alpha | \mu).$$

3. (5 marks) Data on the hardening of eight boat propellor shafts follow, where x_1 is the number of hours the shaft was in the hardening process, x_2 is the process temperature in degrees Celsius, and y is a measure of hardness of the shaft

i	1	2	3	4	5	6	7	8
x_{1i}	11.0	13.0	15.0	14.5	12.0	12.5	13.5	14.0
x_{2i}	122	160	180	152	112	144	131	137
y_i	382	463	476	460	374	422	444	439

Results for three linear regression models are shown below

Statistic	Model 1 (x_1)	Model 2 (x_1, x_2)	Model 3 (x_2)
b_1	24.48	13.48	–
s_{b_1}	5.62	6.08	–
b_2	–	0.92	1.53
s_{b_2}	–	0.37	0.33

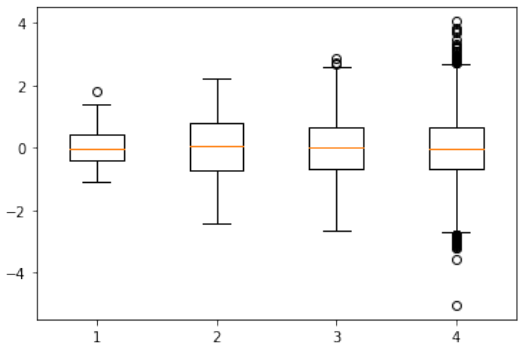
(a) Construct a scatter plot of x_2 against x_1. Is collinearity present in these data? Explain.

(b) For this example, describe how the regression results demonstrate the problems encountered in assessing the separate effects of the independent variables in a multiple regression model when collinearity is present.

(c) Describe in detail the underlying multiple regression model referring to the corresponding design matrix.

4. (5 points) Four boxplots were built from four samples generated by the standard normal distribution. The four sample sizes were

$$n_1 = 10, \quad n_2 = 100, \quad n_3 = 1000, \quad n_4 = 10000.$$

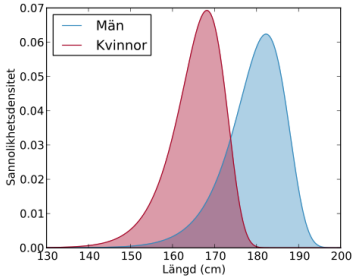


(a) Consider the boxplot number 2 having no outliers. What percentage of the data lies within each of the whiskers? Explain.

(b) Boxplots 3 and 4 have seemingly the same whisker sizes. The upper whisker reaches the hight 2.7. Explain this value referring to the normal distribution table.

(c) Comment on the numbers of outliers on the four different boxplots.

5. (5 points) The graph below describes two strata distributions of the heights in the Swedish population.



(a) For the normal density curve, express its highest value as a function of the standard deviation.

(b) Roughly estimate two standard deviations for female and male heights using the graph.

(c) Estimate the total population (women and men together) mean and variance using the graph.

6 (4 marks) In the analysis of data produced by a two-way ANOVA design, three F-distributions with degrees of freedom $(4, 70)$, $(6, 70)$, and $(24, 70)$ were used.

(a) Describe the following features of the corresponding ANOVA setting: the numbers of levels for two main factors, the numbers of measurements.

(b) What is the rejection region at $\alpha = 0.05$ for the null hypothesis of no interaction? Explain in detail.

(c) Carefully state the normality assumption of the underlying ANOVA model. How can it be verified given the full dataset?

NUMERICAL ANSWERS

1a. Similar to the Bernoulli distribution $\mathrm{Bin}(1, 0.5)$.

1b. Sample at random 1000 individuals and compute the sample proportion $\hat{p}$ of infected. Then multiply it with the population size.

1c. This is an example of $n = 20$ paired observations, with $x = 15$ positive results. Using $X \sim \mathrm{Bin}(n, p)$ model we test $H_0 : p = 0.5$ (no effect) against one-sided alternative (the drug is effective) $H_1 : p > 0.5$. The p-value of the test

$$\begin{aligned}\mathrm{P}(X \geq 15|p = 0.5) &= \mathrm{P}(\tfrac{X-10}{\sqrt{5}} \geq \tfrac{15-10}{\sqrt{5}}|p = 0.5) = \mathrm{P}(\tfrac{X-10}{\sqrt{5}} > \tfrac{14-10}{\sqrt{5}}|p = 0.5)\\ &\approx 1 - \Phi(\tfrac{14.5-10}{\sqrt{5}}) = 1 - \Phi(2.01) = 0.22\end{aligned}$$

says that the drug effect is significant at 2.5% level.

2a. Two composite nested hypotheses

$$H_0 : \mu \leq 0, \quad H : -\infty < \mu < \infty.$$

2b. The likelihood function

$$L(\mu) = \prod_{i=1}^{n} \tfrac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2}} \propto \exp\{-\sum_{i=1}^{n} \tfrac{(x_i-\mu)^2}{2}\} \propto \exp\{\tfrac{2n\bar{x}\mu - n\mu^2}{2}\} \propto \exp\{-\tfrac{n}{2}(\mu - \bar{x})^2\}.$$

2c. The maximum likelihood estimate under H is $\hat{\mu} = \bar{x}$. The maximum likelihood estimate under H_0 is $\hat{\mu}_0 = \bar{x}$ if $\bar{x} < 0$, and $\hat{\mu}_0 = 0$ if $\bar{x} \geq 0$. Therefore, the (generalised) likelihood ratio equals

$$\Lambda = \frac{L(\hat{\mu}_0)}{L(\hat{\mu})} = \frac{\exp\{-\frac{n}{2}(\hat{\mu}_0 - \bar{x})^2\}}{\exp\{-\frac{n}{2}(\hat{\mu} - \bar{x})^2\}} = \begin{cases} 1 & \text{if } \bar{x} < 0, \\ e^{-\frac{n}{2}\bar{x}^2} & \text{if } \bar{x} \geq 0 \end{cases}$$

2d. The likelihood ratio test rejects H_0 for small values of Λ or equivalently, for large positive values of $\bar{x}$. The 5% rejection region

$$\mathcal{R} = \{\bar{x} > \tfrac{1.645}{5}\} = \{\bar{x} > 0.33\}$$

is found from the exact null distribution $\bar{X} \sim \mathrm{N}(0, \frac{1}{5})$.

3a. A scatterplot reveals collinearity.

3b. Three models: their t-values and two-sided p-values

Model 1 (x_1)	Estimates	t-value	p-value
b_1	24.48	4.36	0.005
s_{b_1}	5.62	–	

Model 2 (x_1, x_2)	Estimates	t-value	p-value
b_1	13.48	2.22	0.077
s_{b_1}	6.08	–	
b_2	0.92	2.49	0.055
s_{b_2}	0.37	–	

Model 3 (x_2)	Estimates	t-value	p-value
b_2	1.53	4.67	0.003
s_{b_2}	0.33	–	

show that the multiple regression model utility tests have p-values which are much larger than those of two simple linear regression models. This can be explained by the collinearity effect.

4a. 25%

4b. For a large sample taken from the standard normal distribution the upper end of the whisker is predited to be

$$0.5 \cdot \mathrm{IQR} + 1.5 \cdot \mathrm{IQR} = 2 \cdot \mathrm{IQR} = 2 \cdot 1.35 = 2.7.$$

5a. The maximum of the bell curve

$$f(x) = \tfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

is achieved at $x = \mu$ and equals $\frac{1}{\sqrt{2\pi}\sigma}$.

5b. Solving two equations

$$\tfrac{1}{\sqrt{2\pi}\sigma_1} = 0.07, \qquad \tfrac{1}{\sqrt{2\pi}\sigma_2} = 0.06,$$

we find $\hat{\sigma}_1 = 5.70$ for females, and $\hat{\sigma}_2 = 6.65$ for males.

5c. Assuming equal strata proportions, we get

$$\bar{x}_s = \tfrac{168+182}{2} = 175 \text{ cm}.$$

The total variance is computed using the formula

$$\sigma^2 = \overline{\sigma^2} + \sum_{j=1}^{k} w_j(\mu_j - \mu)^2 = \tfrac{(5.70)^2+(6.65)^2}{2} + \tfrac{(168-175)^2+(182-175)^2}{2} = 87.36.$$

6a. Number of levels for the factor 1 is $I = 4 + 1 = 5$, the number of levels for the factor 2 is $J = 6 + 1 = 7$. The number of observations per cell is $n = \frac{70}{5 \cdot 7} + 1 = 3$. So that the total number of the measurements is $3 \cdot 35 = 105$.

6b. Compute the test statistic $F_{\mathrm{AB}} = \frac{MS_{\mathrm{AB}}}{MS_{\mathrm{E}}}$ and reject the null hypothesis of no interaction if $F_{\mathrm{AB}} > 1.67$, where the critical value 1.67 is obtained from the F-distribution with the degrees of freedom (24,70).