Tentamentsskrivning i Statistisk slutledning MVE155/MSG200, 7.5 hp.

Tid: 27 augusti 2020, kl 14.00-18.00
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) A sample mean is found to be $\bar{x} = 3.14$. The random sample had size 100. The population distribution is normal, with the standard deviation being known to be equal 10.

(a) Suppose the random sample was taken without replacement, so that 10% of the population was covered by the sample. Find the standard error of $\bar{x}$. What does your answer say about the relationship between the point estimate 3.14 and the unknown population mean?

(b) Now assume that the sample is taken with replacement. Using an appropriate formula in the Lecture Notes, find the variance of the sample variance.

2. (5 points) In the dataset "Popular Kids" students in grades 4-6 were asked whether good grades, athletic ability, or popularity was most important to them. The survey involved 169 sibling pairs, with each pair consisting of one brother and one sister. According to the survey, of the girls, 60 ranked popularity most important, compared to 40 of the boys.

(a) Is popularity more important to girls than to boys? Write down the corresponding null and alternative hypotheses in terms of two population proportions p_b and p_g.

(b) Find in the Lecture Notes the most appropriate formula for the standard error of $\hat{p}_g - \hat{p}_b$ and explain it it your own words.

(c) How would you test the hypotheses from the item (a), provided you had access to the full survey data?

3. (5 marks) It is believed that in an election, the percentage of votes V a candidate gets, depends on the amount of money M they spend and their personal quality Q. The table below shows the data for 10 candidates collected from different elections. We wish to determine the relationship between votes, money and quality.

Votes	Money	Quality
70.4	10.8	4.3
41.8	8.1	3.5
7.2	10.7	1.3
57.4	2.8	7.8
48.3	6.2	4.6
19.6	4.5	3.0
72.1	6.8	5.7
40.8	2.1	6.8
55.5	7.9	4.3
50.8	3.1	6.8
37.7	7.6	3.4
60.9	4.6	6.4

(a) The least squares regression line based on the given data is found to be

$$V = -60.89 + 6.38 \cdot M + 14.05 \cdot Q.$$

Its adjusted R square is 77.4%. Explain how this formula can be used for prediction purposes. Do you see obvious limitations for the range of explanatory variables?

(b) A more advanced multiple regression model

$$V = -12.22 - 0.86 \cdot M + 4.86 \cdot Q + 1.56 \cdot M \cdot Q$$

produces a higher adjusted R square of 99.7%. Using an equivalent form

$$V = -12.22 + (1.56 \cdot Q - 0.86) \cdot M + 4.86 \cdot Q,$$

explain why the extra term $M \cdot Q$ can be called the *interaction* variable.

(c) What does the 99.7% say about the model in item (b)? Write down the corresponding design matrix.

4. (5 points) Three blends of a fertiliser are compared for four crops. The observed crop yields are summarised below

	Crop			
Fertilizer	Wheat	Corn	Soy	Rice
Blend X	123	128	166	151
	156	150	178	125
	112	174	187	117
	100	116	153	155
	168	109	195	158
Blend Y	135	175	140	167
	130	132	145	183
	176	120	159	142
	120	187	131	167
	155	184	126	168
Blend Z	156	186	185	175
	180	138	206	173
	147	178	188	154
	146	176	165	191
	193	190	188	169

SUMMARY	*Wheat*	*Corn*	*Soy*	*Rice*	*Total*
Blend X					
Count	5	5	5	5	20
Sum	659	677	879	706	2921
Average	131.8	135.4	175.8	141.2	146.05
Variance	844.2	707.8	278.7	354.2	782.366
Blend Y					
Count	5	5	5	5	20
Sum	716	798	701	827	3042
Average	143.2	159.6	140.2	165.4	152.1
Variance	498.7	978.3	165.7	217.3	511.042
Blend Z					
Count	5	5	5	5	20
Sum	822	868	932	862	3484
Average	164.4	173.6	186.4	172.4	174.2
Variance	443.3	428.8	212.3	175.8	330.695
Total					
Count	15	15	15	15	60
Sum	2197	2343	2512	2395	9447
Average	146.467	156.2	167.467	159.667	157.45
Variance	705.838	871.029	605.981	404.952	671.879

(a) Draw a plot to evaluate graphically if there is an interaction effect between blends and crops. What kind of interaction do you read from the graph?

(b) The p-value of the interaction effect is computed to be 0.046. Explain step by step the algorithm used for this computation.

(c) The yields of corn and rice are closest to each other. Describe a method for finding if this pairwise difference is significant.

5. (5 points) A patient named Boris took an HIV test which is an enzyme-linked immunosorbent assay, commonly known as an ELISA. ELISA test is known to have sensitivity of 78% and specificity of 99%. We want to test the two competing hypotheses below:

$$H_0: \text{ Patient does not have HIV}, \quad H_1: \text{ Patient has HIV}.$$

The possible decisions are:

$$d_0: \text{ Decide that the patient does not have HIV}, \quad d_1: \text{ Decide that the patient has HIV}.$$

(a) Describe two possible types of error in this setting. These two errors are associated with different costs for the society: c_0 and c_1. Which of the two errors do you think is more costly? Explain why.

(b) Boris' doctor prior the ELISA test believed that the odds of Boris having HIV are 40:60. How the doctor should update his believes given a positive result of the ELISA test?

(c) Find a restriction on the costs c_0 and c_1 which would result in rejecting H_0 given a negative result of the ELISA test.

6 (5 marks) Miscellaneous questions.

(a) Why do we need a tool like bootstrap for computing the standard error of the trimmed mean, while for the usual sample mean there is a ready formula?

(b) The sample standard deviation systematically underestimates the population standard deviation. Explain in detail why.

(c) Explain using a graph the following statement: "The wage distribution is right-skewed: the majority of people earn less than the average wage."

NUMERICAL ANSWERS

1a. Using the formula

$$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}(1 - \frac{n-1}{N-1})},$$

where

$$n = 100, \quad , \sigma = 10, \quad \frac{n-1}{N-1} \approx 0.1,$$

we get

$$\sigma_{\bar{X}} = \sqrt{0.9} = 0.95.$$

The point estimate 3.14 is an observation of a random variable which has the mean equal to the unknown population mean μ and whose standard deviation is 0.95.

1b. Using

$$\mathrm{Var}(S^2) = \frac{\sigma^4}{n}\left(\mathrm{E}(\frac{X-\mu}{\sigma})^4 - \frac{n-3}{n-1}\right)$$

we find the variance of the sample variance as

$$\mathrm{Var}(S^2) = \frac{10000}{100}\left(3 - \frac{97}{99}\right) = 202,$$

where we used the value 3 of the kurtosis

$$\mathrm{E}(\frac{X-\mu}{\sigma})^4 = 3$$

due to the normality assumption.

2a. The sample proportion for girls is

$$\hat{p}_g = \frac{60}{169} = 0.355$$

and for boys it

$$\hat{p}_b = \frac{40}{169} = 0.237.$$

Is popularity more important to girls than to boys? The relevant hypotheses are

$$H_0 : p_g = p_b, \quad H_1 : p_g > p_b.$$

2b. The formula for comparing two population proportions from paired samples:

$$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{\pi}_{10} + \hat{\pi}_{01} - (\hat{\pi}_{10} - \hat{\pi}_{01})^2}{n-1}},$$

where $n = 169$ is the number of paired observations. Here $\hat{\pi}_{01}$ and $\hat{\pi}_{10}$ are the proportions of sibling pairs giving mismatching answers (no, yes) and (yes, no) respectively, to the question on popularity.

2c. Compute a one-sided confidence interval using the standard error according to the previous formula, and then reject the null hypothesis if the interval does not cover zero. Otherwise, apply a one-sided version of the McNemar test.

3a. This formula can be used for prediction purposes in the following way. Plugging the values M and Q available for a new candidate into the formula, one gets a prediction for V. Do you see obvious limitations for the range of explanatory variables? YES: if both M and Q are sufficiently small, we get negative V.

3b. The slope for M in

$$V = -12.22 + (1.56 \cdot Q - 0.86) \cdot M + 4.86 \cdot Q,$$

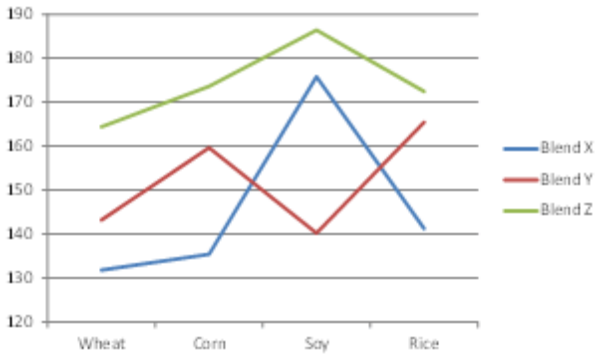
depends on Q, therefore interaction.

3c. 99.7% of the variation in the V variable is explained by the three explanatory variables in the multiple regression model with interaction. The 0.3% is due to noise.

The corresponding design matrix.

$$\begin{array}{cccc} 1 & 10.8 & 4.3 & 46.44 \\ 1 & 8.1 & 3.5 & 28.35 \\ 1 & 10.7 & 1.3 & 13.91 \\ 1 & 2.8 & 7.8 & 21.84 \\ 1 & 6.2 & 4.6 & 28.52 \\ 1 & 4.5 & 3.0 & 13.50 \\ 1 & 6.8 & 5.7 & 38.76 \\ 1 & 2.1 & 6.8 & 14.28 \\ 1 & 7.9 & 4.3 & 33.97 \\ 1 & 3.1 & 6.8 & 21.08 \\ 1 & 7.6 & 3.4 & 25.84 \\ 1 & 4.6 & 6.4 & 29.44 \end{array}$$

4a.



4b. Two-way ANOVA test for interaction.

4c. Simultaneous confidence interval.

5b. Solution

$$\begin{aligned} h_0 &= \mathrm{P}(H_0|\text{ELISA gives a negative result}) = \frac{0.6 \cdot 0.99}{0.6 \cdot 0.99 + 0.4 \cdot 0.22} = 0.87, \\ h_1 &= \mathrm{P}(H_1|\text{ELISA gives a negative result}) = 0.13 \end{aligned}$$

5c. Solution

$$6.7 = \frac{h_0}{h_1} < \frac{c_1}{c_0}$$

6a. After trimming the sample, we lose the nice IID property of the observations.

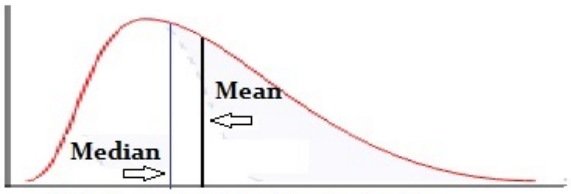
6b. Because

$$(\mathrm{E}S)^2 < \mathrm{E}(S^2) = \sigma^2$$

so that

$$\mathrm{E}S < \sigma.$$

6c.



Right skewed distribution: Mean is to the right