Tentamentsskrivning i Statistisk slutledning, TMS036/MSN560, 5p.

Tid: Tisdagen den 25 maj, 2004 kl 14.15-18.15.
Examinator och jour: Serik Sagitov, tel. 772-5351, rum MC 1421.
Hjälpmedel: miniräknare, egen formelsamling (4 sidor på 2 blad A4) samt utdelade tabeller.
CTH: för "3" fordras 12 poäng, för "4" - 17 poäng, för "5" - 22 poäng.
GU: för "G" fordras 12 poäng, för "VG" - 19 poäng.

Important: carrying out a test make sure

- to state the hypotheses tested,
- state the statistical test you choose,
- explain your choice of the test by referring to the conditions assumed by the test.

---

1. (5 marks) Two hundred female crabs were measured twice for their shell width: before and after molting (shell renewal). Next table contains summary statistics of the measurements.

	Sample mean	Sample standard deviation
Premolt size	129 mm	11 mm
Postmolt size	144 mm	10 mm

Sample correlation coefficient $r = 0.98$

a. Find a 95% confidence interval for the average increase in the shell size. What is the exact meaning of the obtained confidence interval?
Hint: the variance of the difference between two dependent variables X and Y with correllation coefficient ρ equals $\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 - 2\sigma_X\sigma_Y\rho$.

b. Describe a simple regression model for linear dependence of the postmolt shell width on the premolt width. Estimate the model parameters including the error variance.

c. Give a point estimate and a 95% prediction interval for the postmolt size of a female crab given its premolt size 140 mm.

2. (5 marks) The following results were obtained from a stratified random sample without replacement

Stratum 1:	$N_1 = 100$	$n_1 = 50$	$\bar{X}_1 = 10$	$s_1 = 50$
Stratum 2:	$N_2 = 50$	$n_2 = 50$	$\bar{X}_2 = 20$	$s_2 = 30$
Stratum 3:	$N_3 = 300$	$n_3 = 50$	$\bar{X}_3 = 30$	$s_3 = 25$

a. Give an unbiased estimate of the mean for the whole population.

b. Compute a 95% confidence interval for the population mean.

c. How would you allocate further 60 observations in a second round of the population mean estimation? Explain.

3. (5 marks) A particular county employs three assessors who determine the value of residential property in the county. To see whether these assesors differ systematically in their assessments, 5 houses are selected, and each assessor is asked to determine the market value of each house. With factor A denoting assesors and factor B denoting houses, it was found that $\mathrm{SS_A} = 11.7$,

$SS_B = 113.5$, and $SS_E = 25.6$.

a. State parametrically and test the hypothesis that there are no systematic differences among assesors.

b. Explain why a randomized block experiment with only 5 houses was used rather than a one-way ANOVA experiment involving a total of 15 different houses with each assessor asked to assess 5 different houses (a different group of 5 for each assessor).

c. Judging from the F_B ratio, do you think that blocking on houses was effective in this investigation? Explain

4. (5 points) Your own analysis of a certain stock value suggests that the stock will go up with probability 2/5 and go down with probability 3/5. Three possible actions - to buy, sell, or do nothing yield the following gains

	buy	sell	do nothing
If the stock's value goes up	+\$1000	-\$1000	\$0
If the stock's value goes down	-\$1000	+\$1000	\$0

a. What are your prior risks (expected losses) of buying, selling, and doing nothing?

b. Your investment advisor tells you that he believes that the stock is going up. On the basis of your experience, you belive that his prognosis is correct about 2/3 of the time. With the advisor's opinion in hand, compute the posterior probability of the stock going up.

c. Which action would you choose after comparing the posterior risks?

5. (5 marks) Suppose we have a sample with 5 observations: $(1, 4, 4, 4, 4)$ and we are interested in investigating the properties of the trimmed mean $\bar{X}_{0.4}$ and the sample median $\hat{M}$.

a. There are six possible ordered samples resulting from nonparametric bootstrapping:

$$(1,1,1,1,1),\ (1,1,1,1,4),\ (1,1,1,4,4),\ (1,1,4,4,4),\ (1,4,4,4,4),\ (4,4,4,4,4).$$

Compute probabilities for each of the six possibilities.

b. Assign the values of $\bar{X}_{0.4}$ to each of these outcomes and find the bootstrap distibution of the trimmed mean $\bar{X}_{0.4}$.

c. Repeat b) for the sample median and explain the differences in the bootstrap distibutions for $\bar{X}_{0.4}$ and $\hat{M}$.

6. (5 marks) A research project has been focused on the existence of any relationship between the date of patient admission for treatment of alcoholism and patient's birthday. The investigators established four different admission categories:
1. within 7 days following patient's birthday,
2. between 8 and 30 days, inclusive, from the birthday,
3. between 31 and 90 days, inclusive, from the birthday,
4. more than 90 days from the birthday.

A sample of 200 patients gave observed frequencies of 11, 24, 69, and 96 for categories 1, 2, 3, and 4 respectively.

a. State the null hypothesis of no relationship in terms of four probabilities of patient's birthday falling in one of the four categories. Without analysing the data try to give some reasons why this null hypothesis might fail.

b. Write down the likelihood of the observed counts (11, 24, 69, 96) as a function of four population proportions.

c. Test the null hypothesis using a significance level of 0.01. What are your conclusions concerning the relationship between the date of patient admission for treatment of alcoholism and patient's birthday?

Statistical tables supplied:
1. Normal distribution table
2. Chi-square distribution table
3. t-distribution table
4. F-distribution table

Partial answers and solutions are also welcome. Good luck!

ANSWERS

1a. If X = the premolt shell width, Y = the postmolt shell width, and $D = Y - X$, then $\bar{D} = 144 - 129 = 15$ mm is an unbiased point estimate of the average increase in the shell size $\mu = \mathrm{E}(D)$. The standard error of $\bar{D}$ equals

$$\sqrt{\frac{11^2 + 10^2 - 2 \cdot 11 \cdot 10 \cdot 0.98}{200}} = 0.16.$$

It implies $15 \pm 1.96 \cdot 0.16 = 15 \pm 0.32$ as a 95% confidence interval for μ. The interval 15 ± 0.32 covers the true value of μ with 95% confidence in the following sense. If we collect 100 independent samples of the same size and produce 100 similar confidence intervals, then it is expected that 95 of them will cover the true value of μ.

1b. $Y = \beta_0 + \beta_1 X + \epsilon$, where $\epsilon \in \mathrm{N}(0, \sigma^2)$. Least square estimates $b_1 = 0.98 \cdot \frac{10}{11} = 0.89$, $b_0 = 144 - 0.89 \cdot 129 = 29.1$.

Since $\mathrm{SSE} = (1 - r^2)\mathrm{SST}$ we obtain $s^2 = (1 - 0.98^2)\frac{199}{198}10^2 = 3.98$, an unbiased estimate of σ^2.

1c. Given $X = 140$ the expected value of Y is estimated as $29.1 + 0.89 \cdot 140 = 153.7$. The corresponding 95% prediction interval is $153.7 \pm 1.96 \cdot \sqrt{3.98(1 + \frac{1}{200} + \frac{1}{199}(\frac{140-129}{11})^2)} = 153.7 \pm 3.9$.

2a. Stratified sample mean $\bar{X}_s = \frac{100}{450}10 + \frac{50}{450}20 + \frac{300}{450}30 = 24.44$.

2b. Three standard errors $s_{\bar{X}_i} = \frac{s_i}{\sqrt{n_i}}\sqrt{1 - \frac{n_i}{N_i}}$: $s_{\bar{X}_1} = 5$, $s_{\bar{X}_2} = 0$, $s_{\bar{X}_3} = 3.23$. Standard error of $\bar{X}_s$: $s_{\bar{X}_s} = \sqrt{(\frac{5}{4.5})^2 + (\frac{3.23}{1.5})^2} = 2.42$. Approximate 95% CI for the population mean $24.44 \pm 1.96 \cdot 2.42 = 24.44 \pm 4.75$.

2c. The second stratum has been fully observed, therefore further 60 observations should be allocated between the first and the third starta. Since $\sigma_1 : \sigma_3 \approx 2 : 1$, we have $W_1\sigma_1 : W_3\sigma_3 \approx 2 : 3$. According to the optimal allocation formula new sample sizes should relate as $n_1 : n_3 = 2 : 3$ implying the optimal allocation $n_1 = 24$ and $n_3 = 36$.

3a. $H_0 : \mu_1 = \mu_2 = \mu_3$, where μ_i is the average house price set by i-th assessor. The ANOVA table

Source	SS	df	MS	F	P
A	11.7	2	5.85	1.83	> 0.10
B	113.5	4	28.38	8.87	< 0.01
Error	25.6	8	3.20		
Total	150.8	14			

implies that we can not reject the null hypothesis.

3b. With a one-way ANOVA there is a risk of a significant F-value due to differences between houses assigned to different assessors rather then due to systematic differences between the assessors.

3e. The F_{B} ratio is very significant. This means that the investigation included a wide range of house types, and we conclude that blocking on houses was effective.

4a. Prior probabilities $\mathrm{P}(Up) = 2/5$, $\mathrm{P}(Down) = 3/5$.

Action	buy	sell	do nothing
Prior risk	\$200	–\$200	\$0

4b. Advisor'r forecast is an example of uncertain measurement. If $A = \{\text{advisor predicts Up}\}$, then $\mathrm{P}(A|Up) = 2/3$ and $\mathrm{P}(A|Down) = 1/3$. Thus $\mathrm{P}(Up|A) = \frac{(2/3)\cdot(2/5)}{(2/3)\cdot(2/5)+(1/3)\cdot(3/5)} = 4/7$.

4c. Buy - because this is the action with the lowest posterior risk:

Action	buy	sell	do nothing
Posterior risk	$-\$142.9$	$+\$142.9$	$\$0$

5a-c. Possible values of $\bar{X}_{0.4}$ and $\hat{M}$

	Prob	$\bar{X}_{0.4}$	$\hat{M}$
(1,1,1,1,1)	0.00	1	1
(1,1,1,1,4)	0.01	1	1
(1,1,1,4,4)	0.05	2	1
(1,1,4,4,4)	0.20	3	4
(1,4,4,4,4)	0.41	4	4
(4,4,4,4,4)	0.33	4	4

Here the probabilities follow the Binomial distribution:

$$\mathrm{P}(\text{Number of ones} = k) = \binom{5}{k}(0.2)^k(0.8)^{5-k}$$

Bootstrap distributions

Values k	1	2	3	4
$\mathrm{P}(\bar{X}_{0.4} = k)$	0.01	0.05	0.20	0.74
$\mathrm{P}(\hat{M} = k)$	0.06	0.00	0.00	0.94

Both $\bar{X}_{0.4}$ and $\hat{M}$ are robust against outliers, with $\hat{M}$ beeing a more rigid measure of location, since it focuses on the middle observation neglecting the distribution of other observations.

6a. H_0: $p_1 = 7/365, p_2 = 23/365, p_3 = 60/365, p_4 = 275/365$ assuming 365 day in a year. No relationship means uniform distribution for the admission date over the year counted from patient's birthday. The null hypothesis might fail to be true because people are used to drink on their birthday.

6b. $L(p_1, p_2, p_3, p_4) = \binom{200}{11,24,69,96} p_1^{11} p_2^{24} p_3^{69} p_4^{96}$.

6c. Apply chi-square test for the simple null hypothesis stated above. Expected counts: $(7/365)\cdot 200 = 3.84, (23/365)\cdot 200 = 12.60, (60/365)\cdot 200 = 32.88, (275/365)\cdot 200 = 150.68$. The chi-square test statistic $X^2 = 83.19$ is highly significant according to the χ^2_3 distribution table. There exists a relationship between the admission date and patient's birthday: more patients are admitted sooner after their birthday than expected under H_0.