Tentamentsskrivning i Statistisk slutledning, TM, 5p.

Tid: tisdagen den 29 maj 2007 kl 14.00-18.00.
Examinator och jour: Serik Sagitov, tel. 772-5351, mob. 0736-907613, rum MV:H3026.
Hjälpmedel: Chalmersgodkänd räknare, egen formelsamling (4 sidor på 2 blad A4) samt utdelade tabeller.
För "G" fordras 12 poäng, för "VG" - 18 poäng, för "MVG" - 24 poäng.

1. (5 points) Two types of coating are being compared for use as a rust preventive. Fifty pieces of pipe, each of the same type and size, are used in the equipment. Half of each pipe is coated with a 0.5-mil layer of compound A; the other half receives a 5-mil layer of compound B. Each pipe is then subjected to 1000 hours of salt fog. At the end of the experiment an impartial judge compares the two compounds for effectiveness in preventing rust. The data gathered are shown in the next table.

	A effective	A not effective
B effective	35	5
B not effective	8	2

Is there a difference in the proportion of pipes that are deemed effective for the two compounds? Explain using an appropriate test.

2. (5 points) The volume, height and diameter at 4.5 ft above ground level were measured for a sample of 31 black cherry trees. The data were collected to provide a basis for determining an easy way of estimating the volume of a tree.

a. Suggest a simple formula relating volume to height and diameter (hint: use common sense and geometric considerations). Such a formula implies a multiple regression model explaining the logarithm of the volume in terms of logarithms the height and diameter. What values seem to be appropriate for the two slope parameters in this model?

b. Statistical analysis of real tree data produced three 95 % CIs

$$(0.6984; 1.53590), (-8.2699; -4.9933), (1.8290; 2.1363).$$

for the two slope parameters and the intercept parameter for the multiple regression model mentioned above. These three CIs (which are given in an arbitrary order) confirm the slope coefficients' values implied by the model a. What are the three point estimates and their standard errors?

c. It was found that the determination coefficient equals $R^2 = 0.9777$. Compute the adjusted R_a^2 and explain the reason for such an adjustment.

3. (6 points) A population consisting of three strata of equal sizes is characterizing by a random variable X with stratum means μ_1, μ_2, μ_3 and equal

stratum variances $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_0^2$. An optimal stratified sample of size $n = 30$ was taken to estimate the overall mean μ. The three samples from each of the strata gave sample means $\bar{X}_1 = 15.1$, $\bar{X}_2 = 14.2$, $\bar{X}_3 = 16.2$ and sample standard deviations $s_1 = 3.1$, $s_2 = 3.2$, $s_3 = 2.7$.

a. How does the formula for the total variance $\sigma^2 = Var(X)$ look in this particular case?

b. Find the stratified sample mean and its standard error.

c. Test the null hypothesis of equal means $H_0 : \mu_1 = \mu_2 = \mu_3$. State clearly all your assumptions and how would you check them.

d. Which non-parametric test can be used here given all 30 sample values? Describe this test.

4.(5 points) A 20% trimmed mean for an iid sample of size 20 taken from a gamma distribution was found to be 30.6. How would you estimate the standard error of the trimmed mean using the sample values and computer?

5. (5 points) Explain the difference between the confidence interval and the credibility interval. Illustrate with a simple example based on the self-conjugate property of the normal distribution.

6. (4 points) On top of the density curve for the standard normal distribution draw another probability density curve with zero mean, variance one, zero skewness and curtosis > 3. Explain your drawing. Sketch the corresponding normal probability plot.

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!

Numerical ANSWERS

1. This is an example of data gathered using a matched pairs design. An appropriate test here is the McNemar test. The parametric form of the hypotheses of interest are

$$H_0 : \pi_{1.} = \pi_{.1},$$
$$H_1 : \pi_{1.} \neq \pi_{.1}.$$

The observed test statistic is $\frac{(5-3)^2}{5+8} = 0.69$. Since the null distribution is approximately χ_1^2 we can use the normal probability table to find an approximate two-sided P-value as $2 \times (1 - \Phi(\sqrt{0.69})) = 0.40$. Thus the observed difference between the two types of coating statistically is not significant, and we do not reject the null hypothesis.

2a. Viewing a tree as a cylinder or as a cone we can regard its volume V as a constant times the height H times the squared diameter D. Therefore, we expect a relation of the form

$$\log V = \beta_0 + \beta_H \log H + \beta_D \log D + \epsilon$$

with $\beta_H = 1$ and $\beta_D = 2$.

2b. The given CIs have the form $b \pm 2.048 s_b$, where 2.048 is taken from the t_{28}-distribution table. We conclude that the three point estimates and their standard errors are

$$b_0 = -6.63,\ s_{b_0} = 0.80;$$
$$b_H = 1.12,\ s_{b_H} = 0.20;$$
$$b_D = 1.98,\ s_{b_D} = 0.07.$$

2c. The required adjustment is done by

$$R_a^2 = 1 - \frac{n-1}{n-p} \cdot \frac{\text{SSE}}{\text{SST}} = 1 - \frac{30}{28} \cdot (1 - R^2) = 0.9761.$$

It is meant to punish for unnecessary extra parameters in a multiple regression model.

3a. The total variance is

$$\begin{aligned} \sigma^2 &= \overline{\sigma^2} + \sum W_l (\mu_l - \mu)^2 \\ &= \sigma_0^2 + \frac{(\mu_1 - \mu)^2 + (\mu_2 - \mu)^2 + (\mu_3 - \mu)^2}{3}. \end{aligned}$$

3b. Stratified sample mean:

$$\bar{X}_s = W_1 \bar{X}_1 + W_2 \bar{X}_2 + W_3 \bar{X}_3 = 15.2$$

and its squared standard error

$$\begin{aligned} s^2_{\bar{X}_s} &= (W_1 s_{\bar{X}_1})^2 + (W_2 s_{\bar{X}_2})^2 + (W_3 s_{\bar{X}_3})^2 \\ &= \frac{1}{9}\left(\frac{s_1^2}{10} + \frac{s_2^2}{10} + \frac{s_3^2}{10}\right) = 0.302. \end{aligned}$$

Thus the standard error equals $s_{\bar{X}_s} = 0.55$.

3c. One-way ANOVA table

Source	df	SS	MS	F	P-value
Strata	2	20.1	10	1.11	> 0.10
Errors	27	244.3	9		

shows that the difference between the sample means is not significant. Here the sum of squares bewtween the samples and within the samples are computed as follows

$$\begin{aligned} &20.1 = 10((15.1 - 15.2)^2 + (14.2 - 15.2)^2 + (16.2 - 15.2)^2) \\ &244.3 = 9(3.1^2 + 3.2^2 + 2.7^2). \end{aligned}$$

3d. The Kruskal-Wallis test. The 30 values in the sample are ranked from 1 to 30 and the ranks are groupped in three groups according to the strata allocation. Then the average group ranks are compared with the midrank value 15.5 using a Kruskal-Wallis test statistic K based on the sum of the squared differences. In this case the null distribution of K is approximated by the χ^2_2-distribution and the null hypothesis of equality is rejected for large values of K.

4. Since gamma distribution is chosen as a statistical parametric model we can apply a parametric bootstrap procedure. Using the original sample we can estimate the scale λ and the shape α parameters using the maximum likelihood method. Denote these MLEs by $\hat{\alpha}, \hat{\lambda}$.

Substituting the true population distribution Gamma(α, λ) with Gamma$(\hat{\alpha}, \hat{\lambda})$ we can generate many independent samples of size $n = 20$. For each such sample it is easy to compute a 20% trimmed mean. The standard deviation for the set of these trimmed means gives an estimate for the stardard error in question.

5. A confidence interval is a random interval for an unknown constant parameter θ. The randomness comes from the random sampling procedure. Among 100 independent CIs on average $(1-\alpha)\cdot 100$ intervals would cover the true value of θ.

On the other hand, a credibility interval is a nonrandom interval covering $(1 - \alpha) \cdot 100$ procent of the posterior distribution of the unknown population parameter θ considered to be random. The randomness of θ reflects our uncertainty about its true value.

An IQ measurement example clearly illustrates the difference. The IQ value of a randomly chosen person can be modelled by the normal distribution $X \sim$

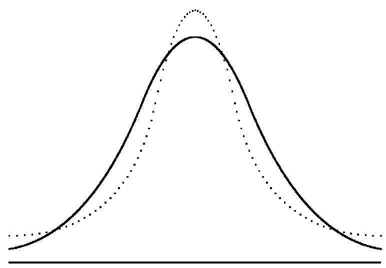
N(μ, 100). An appropriate prior distribution for the mean $\mu \sim$ N(100,225) is the population IQ distribution. If the observed IQ is $x = 130$, then the posterior distribution is $\theta \sim$ N(120.7, 69.2). Now we can compare the 95% CI

$$130 \pm 1.96 \cdot 10 = 130 \pm 19.6$$

for μ with its 95% CrI

$$120.7 \pm 1.96 \cdot \sqrt{69.2} = 120.7 \pm 16.3.$$

6. The dotted line depicts a heavy tailed distribution curve compared to the N(0,1)-curve. Both curves are symmetric around zero since both have zero mean and zero skewness. Heavy tails are seen from the ends of the curve being above the ends of the N(0,1)-curve. The center of the curve goes above the N(0,1)-curve since the variances are equal and the total area below both curves is one.



The next figure is a rough sketch of the corresponding probability plot.