Tentamentsskrivning i Statistisk slutledning MVE155/MSG200/MSA840/MVE060, 7.5 hp.

Tid: Fredagen den 13 mars, 2009 kl 08.30-12.30
Examinator och jour: Serik Sagitov, tel. 772-5351, mob. 0736 907 613, rum H3026 i MV-huset.
Hjälpmedel: Chalmersgodkänd räknare, **egen** formelsamling (4 sidor på 2 blad A4) samt utdelade tabeller.
CTH: för "3" fordras 12 poäng, för "4" - 18 poäng, för "5" - 24 poäng.
GU: för "G" fordras 12 poäng, för "VG" - 20 poäng.

Important! For each problem

- describe and justify statistical models you apply,
- state clearly hypotheses you test,
- discuss different relevant approaches you have learned in the course.

---

1. (5 points) You are interested in finding the proportion p of pipe smokers in a population with 40% of males and 60% of females. Knowing that among pipe smokers only 10% are women, how would you allocate 1000 observations across gender to optimize your estimate of p?

2. (5 points) A researcher has the hypothesis that first-born babies take longer to be born than second-born babies. In order to test his hypothesis, he asks 20 of his colleagues who have at least two children, how many days their first child was born before or after the calculated date, and how many days their second child was born before or after the calculated day. Suppose the results look like this (negative numbers mean before the date):

Parent	1	2	3	4	5	6	7	8	9	10
First child X	0	5	-1	3	4	4	0	-7	5	12
Second child Y	0	0	-1	4	-6	2	10	0	-4	1

Parent	11	12	13	14	15	16	17	18	19	20
First child X	13	2	0	6	7	8	4	11	0	3
Second child Y	2	8	2	3	-1	10	6	4	5	0

a. Analyze the data. What are your conclusions?

b. Using summary statistics: $\bar{X} = 3.95$, $s_X = 4.82$, $\bar{Y} = 2.25$, $s_y = 4.18$, and $\overline{XY} = 9.50$ find a relevant measure of dependence between these two variables. Do they seem strongly dependent? Illustrate with a scatterplot.

3. (5 points) The data below represent a comparison of two media for culturing Mycobacterium tuberculosis. Fifty specimens were plated up on both media and the following results were obtained:

	Medium A: Growth	Medium A: No Growth
Medium B: Growth	20	12
Medium B: No Growth	2	16

a. Describe an appropriate statistical model and state clearly in terms of the key parameters of the model the null and alternative hypotheses which are relevant in this context.

b. Analyze the results. What are the conclusions?

4. (5 points) Elin is working on a reasearch project which will produce large amount of data. In a forthcoming paper she will present the results of 10 hypothesis tests, based on the data, and she is planning to use 5% significance level in each test.

a. If all the 10 null hypotheses are true, and if you assume that all 10 tests are independent, what is the probability that Elin will reject one ore more of the null hypotheses?

b. Instead of using 5% significance level, Elin decides to report ten P-values for the ten tests she performed. The P-values sorted by size are

$$0.0001, 0.0002, 0.001, 0.002, 0.003, 0.006, 0.02, 0.03, 0.05, 0.05.$$

How many hypotheses should she reject at the *combined* significance level of 5%? Explain why.

5. (5 points) Annika has measured her weight using low precision floor scales. The scales showed 72 kg. It was dissapointing, since her expectations about her weight were something around 68 kg plus minus $\sigma = 1$ kg. Assuming that the scales produces no systematic error and has a random error of size $\sigma' = 2$ kg find a 90% credibility interval for Annikas current weight.

6. (5 points) Given a sample 18.3, 16.7, 11.8, 21.0, 13.8, 15.6

a. Estimate the upper quartile of the population distribution. What assumptions do you make concerning that data?

b. How would you find the standard error of your estimate under the assumption of normality for the population distribution.

c. Answer the same question as in b) without the normality assumption.

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

1. The population consists of two strata with $W_1 = 0.4$ and $W_2 = 0.6$. If p_1 is the proportion of pipe smokers among males and p_2 is the proportion of pipe smokers among females, then $p = 0.4p_1 + 0.6p_2$. The optimal allocation would follow the ratio $W_1\sigma_1 : W_2\sigma_2$, where $\sigma_i^2 = p_i q_i$.

Since both p_1 and p_2 are relatively small, we approximate

$$\frac{W_1\sigma_1}{W_2\sigma_2} \approx \frac{0.4\sqrt{p_1}}{0.6\sqrt{p_2}} = 0.66\sqrt{\frac{p_1}{p_2}}.$$

If N is the total population size, the number of female pipe smokers is $0.6Np_2 = 0.1Np$, and the number of male pipe smokers is $0.4Np_1 = 0.9Np$. Thus $p_2 = p/6$, $p_1 = 9p/4$, and

$$0.66\sqrt{\frac{p_1}{p_2}} = 0.66\sqrt{\frac{6 \cdot 9}{4}} = 2.45.$$

Thus the optimal allocation of 1000 observations is $1000 \cdot \frac{2.45}{2.45+1} = 710$ males and 290 females.

2b. From a sample covariance

$$S_{xy} = \frac{n}{n-1}(\overline{XY} - \bar{X}\bar{Y}) = 0.645$$

we find a sample correlation $\rho = \frac{0.645}{4.82 \cdot 4.18} = 0.03$. It indicates that there is no much dependence between the first and second child dates. Which is also seen from the scatter plot on Figure 1.

2a. We test $H_0 : \mu_1 = \mu_2$, no difference between the means, against $H_1 : \mu_1 > \mu_2$, first-born babies take longer to be born than second-born babies. Since the samples are paired we take the differences and use a one-sample t-test assuming that the differences are normally distributed.

The 20 differences

$$0, 5, 0, -1, 10, 2, -10, -7, 9, 11, 11, -6, -2, 3, 8, -2, -2, 7, -5, 3$$

have sample mean $\bar{D} = \bar{X} - \bar{Y} = 1.7$ and sample variance

$$s_X^2 + s_y^2 - 2S_{xy} = 39.4.$$

Thus the t-test statistic is $T = \frac{1.7}{\sqrt{39.4/20}} = 1.21$. According to the t-distribution table with df=19 the ONE-sided P-value is larger than 10%. We can not reject $H_0 : \mu_1 = \mu_2$ in favor of $H_1 : \mu_1 > \mu_2$.

3a. The data consists of 50 independent PAIRS of observations. Statistical model: n pairs of observations fall into 4 groups of outcomes with probabilities:

	Medium A: Growth	Medium A: No Growth
Medium B: Growth	π_{11}	π_{12}
Medium B: No Growth	π_{21}	π_{22}

We would like to test $H_0 : \pi_{11} + \pi_{12} = \pi_{11} + \pi_{21}$ against $H_1 : \pi_{11} + \pi_{12} \neq \pi_{11} + \pi_{21}$.

3b. The McNemar test statistics $X^2 = \frac{(12-2)^2}{12+2} = 7.14$ has approximate null distribution χ_1^2. Using the normal distribution table for the $\sqrt{7.14} = 2.67$ we obtain the P-value of the McNemar test to be $2 \cdot (1 - 0.9962) = 0.008$. We conclude that there is a significant difference between two media.

4a. Due to independence the probability to accept all 10 null hypotheses is $0.95^{10} = 0.6$. Therefore, with probability 0.4 Elin gets at least one false positive result.

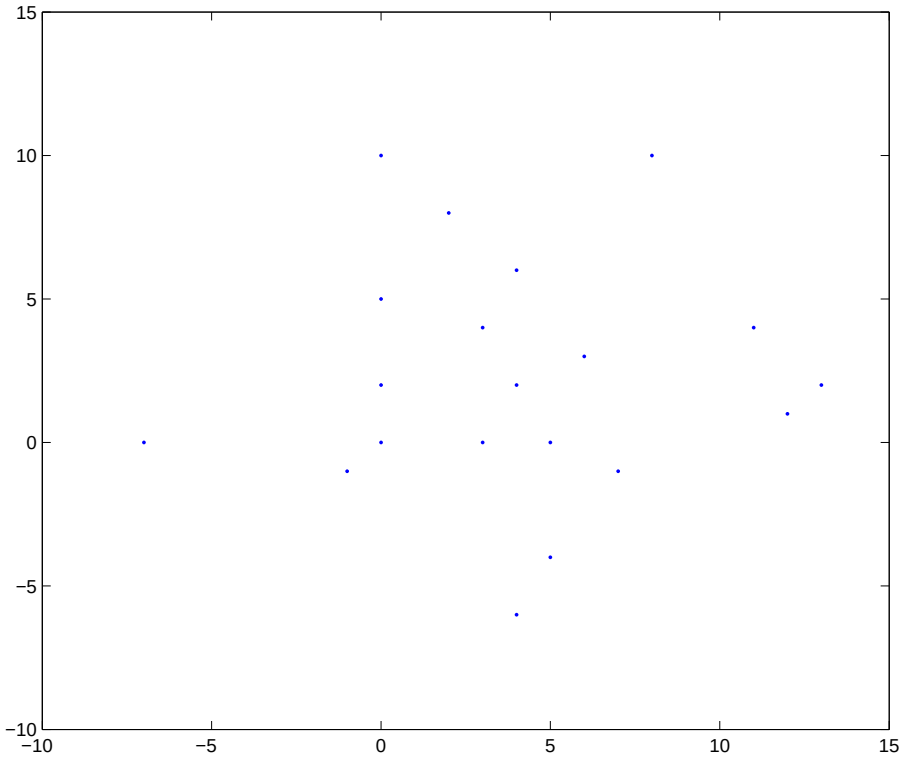

Figure 1: *Scatter plot: first child's birth (X) against the scond's (Y)*

4b. In view of the previous observation one should be aware of the multiple testing problem and adjust the significance level for a single test using the Bonferronni correction: $0.05/10 = 0.005$. It follows that with the combined 5% significance level we reject 5 out of 10 null hypotheses.

5. If we take Annika's expectations as a normal prior N(68,1) of her weight θ, then given the data N(72,4) we get the posterior distribution of θ to be

$$N(c \cdot 68 + (1-c) \cdot 72; c \cdot 1) = N(68.8; 0.8),\ c = \frac{4}{4+1} = 0.8.$$

Thus a 90% credibility interval for θ is $68.8 \pm 1.645 \cdot \sqrt{0.8} = 68.8 \pm 1.5$.

6a. The 3/4-quantile is estimated by $X_{(5)} = 16.7$ since for $k = 5$ and $n = 6$ we have $\frac{k-1/2}{n} = 0.75$. Here we assume that the sample is IID.

6b. Parametric bootstrap.

6c. Non-parametric bootstrap.