Tentamentsskrivning i Statistisk slutledning MVE155/MSG200, 7.5 hp.

Tid: Fredagen den 18 mars, 2011 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 do your best to
- 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) A large sample has the following summary statistics

$$\bar{X} = 10$$
$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = 4$$
$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3 = 3$$
$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4 = 50.$$

a. Sketch a possible population density curve and comment on various features of it.

b. How the corresponding normal probability plot should look like?

2. (5 points) The following table shows admission rates for the six most popular majors at the graduate school at the University of California at Berkeley.

Major	Men: number of applicants	Men: percentage admitted	Women: number of applicants	Women: percentage admitted
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	34
F	373	6	341	7

a. If the percentage admitted are compared, women do not seem to be unfavorably treated. But when the combined admission rates for all six majors are calculated, it is found that 44% of the men and only 30% of the women were admitted. How this paradox is resolved?

b. This is an example of an observational study. Suggest a controlled experiment testing relevant statistical hypotheses.

3. (5 points) A new drink is claimed to hardly ever cause a hangover. To test this wonder beverage, 30 people who were particularly susceptible to hangovers volunteered to first drink 2 ounces of pure alcohol and note the effect, and then a week later to drink 6 glasses of the new drink (containing an equivalent amount of alcohol).

If the result were as follows, could the new drink be said to be significantly less prone to cause hangovers than pure alcohol?

	Hangover after alcohol	No hangover after alcohol
Hangover after new drink	12	3
No hangover after new drink	13	2

4. (5 points) Three hundred patients suffering from chronic backache were divided at random into 3 equal-sized groups. The first group was given Treatment A, which relieved 48% of them. The second group was given Treatment B, which relieved 61%, and the third group was given Treatment C, which was successful in 71%.

Do these figures indicate a statistically significant difference between the efficacy of the 3 treatments?

5. (5 points) Tukey's simultaneous confidence intervals formula is based on the following observation. Suppose we have k independent samples, each of size n, taken from possibly different $\mathrm{N}(\mu_i, \sigma^2)$ distributions $i = 1, \ldots, k$. With $\bar{Y}_i$ standing for the sample means, the normalized maximum of pairwise differences

$$W_{k,n} = \frac{\sqrt{n}}{s_p} \max_{i,j} |\bar{Y}_i - \bar{Y}_j - (\mu_i - \mu_j)| \tag{1}$$

has distribution $\mathrm{SR}(k, k(n-1))$ called the Studentized range distribution.

a. Explain how the pooled sample variance s_p^2 is computed. What is the relation between the number of degrees of freedom $k(n-1)$ to the pooled sample variance formula?

b. The distribution $\mathrm{SR}(k, k(n-1))$ is free from the parameter σ^2. Explain this fact referring to the expression (1).

c. Let $n = 11$, $k = 4$, $s_p = 2.9$. Applying $P(W_{4,11} \leq 3.79) = 0.95$, write down the formula for the simultaneous 95% confidence intervals for the 6 pairwise mean differences. Given $\bar{Y}_i = 11.3$, $\bar{Y}_j = 14.6$, would you reject $H_0 : \mu_i = \mu_j$?

6. (5 points) You are studying a two-dimensional data set of size $n = 12$ whose scatter plot has an elliptic shape.

a. Write down a simple linear regression model for this data mentioning all key assumptions.

b. The sample covariance is computed to be -3.5 while two sample variances are 4.9 for the explanatory variable and 3.8 for the response variable. Estimate the size of the noise.

c. A closer inspection indicates that the elliptic form of the scatter plot is slightly distorted. You try then to test a quadratic relationship instead of the linear. Write down a model for this purpose.

d. The coefficient of multiple determination for the quadratic model is found be 67%. Compare it to the coefficient of determination for the linear model to decide which model is better.

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

1a. The provided summary statistics imply the next estimates for the population distribtuion:

- mean = 10
- standard deviation = 2
- coefficiient of skewness = 0.375
- kurtosis = 3.125.

The corresponding curve should look pretty much like a normal distribution curve with mean 10 and st. dev. 2, though slightly skewed to the right and with a bit heavier tails.

2a. This is another example of the Simpson paradox. The confounding factor here is the difficulty to enter the programmes. Men tend to apply for easy programs, while women more often apply for programs with low admission rates.

3. Matched pairs design. McNemara's test statistics is $\frac{(3-13)^2}{3+13} = 6.25$. The null distribution is approximated by the χ^2_1-distribution. Since the square root of 6.25 is 2.5, the standard normal distribution gives a (two-sided) P-value 1.2%. We reject the null hypothesis stating that two drinks are equally damaging in terms of hangovers. The data clearly demonstrates that the new drink is less damaging.

4. The data can be summarized by the table

Treatments	A	B	C	Totals
Patients relieved	48	61	71	180
Patients not relieved	52	39	29	120
Sample sizes	100	100	100	300

The chi-square test of homogeneity produces the statistics $X^2 = 11.08$ is large enough to reject the null hypothesis of no difference among three treatments. The approximate null distribution is the χ^2_2-distribution, whose table gives the P-value larger than 0.5%.

5c. The simultaneous confidence interval is

$$(\bar{Y}_i - \bar{Y}_j) \pm 3.31.$$

Thus the observed difference $14.6 - 11.3 = 3.3$ is not statistically significant.

6b. $s^2 = \frac{n-1}{n-2} 3.8(1 - r^2) = 1.42$. So the noise size is estimated as $s = 1.2$.

6d. Compare $R^2_a = 0.63$ for the linear model and $R^2_a = 0.60$ for the quadratic model.