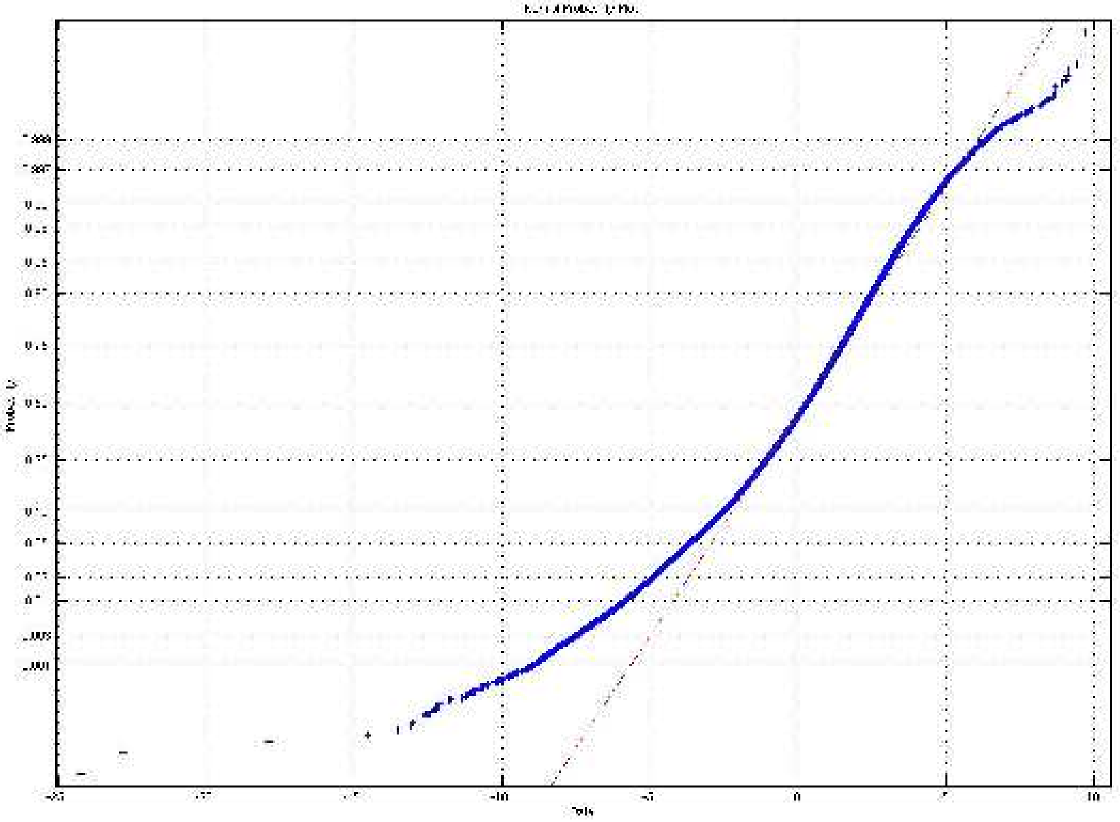
Tentamentsskrivning i Statistisk slutledning MVE155/MSG200, 7.5 hp.

Tid: Fredagen den 14 mars, 2008 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.

1. (5 points) The picture below diplays certain data in the form of normal probability plot.



The x-axis represents the data and the y-axis gives normal ditsribution’s quantiles.

a. Is this a case of heavy tails or light tails distribution? Explain.

b. What do you think is the sign of the coefficient of skewness? Why? What can you say about the value of the kurtosis coefficient?

c. Draw a sketch of the corresponding histogram together with a normal distribution curve having the **same** mean and variance as the data.

2. (5 points) Next table gives the numbers of fumbles (missar) made by 110 football teams in Division IA (American football) during 55 games:

2	1	2	2	3	1	3	4	3	4	5
5	2	1	3	2	5	2	4	1	2	2
1	0	4	2	4	1	2	0	2	0	3
0	1	2	0	1	2	2	3	5	1	3
2	3	4	5	4	3	6	0	3	1	2
1	2	2	1	2	1	3	2	4	2	4
4	2	0	5	4	3	6	5	3	5	1
3	1	1	3	1	4	3	1	5	1	2
1	3	4	4	4	2	7	4	2	5	3
1	3	6	2	1	1	4	1	2	3	0

with the mean number of fumbles per team being 2.55.

a. Suggest a simple one-parameter statistical model for the dataset. Justify your choice by refferring to a relevant asymptotic result for the binomial distribution.

b. How well does your model fit the data? Use a relevant statistical test procedure. State clearly the null and alternative hypotheses.

3. (5 points) A certain fraction of antibiotics are bound to serum proteins. This phenomenon bears directly on the effectiveness of the medication, because the binding decreases the systemic uptake of the drug.

The following table lists the binding percentages in bovine serum measured for five widely prescribed antibiotics.

	Penicillin	Tetracycline	Streptomycin	Erythromycin	Chloramphenicol
	29.6	27.3	5.8	21.6	29.2
	24.3	32.6	6.2	17.4	32.8
	28.5	30.8	11.0	18.3	25.0
	32.0	34.8	8.3	19.0	24.2
means	28.6	31.4	7.8	19.1	27.8

The corresponding error sum of squares is $\sum_{j=1}^{5}\sum_{i=1}^{4}(X_{ij}-\bar{X}_{\cdot j})^2 = 135.83$.

a. According to this table Tetracycline and Streptomycin have the largest difference. Is this difference statistically significant? Explain why is it wrong to answer this question using a two-sample confidence interval treating Tetracycline and Streptomycin samples as two independent samples.

b. Apply Tukey's method to make all the pairwise comparisons with 95% confidence. Here comes an extract from the Studentized range distribution:

$$q_{4,15}(0.05) = 4.08, q_{4,16}(0.05) = 4.05, q_{5,15}(0.05) = 4.37, q_{5,16}(0.05) = 4.33.$$

4. (5 points) Suppose $H_0 : p_1 = p_2$ is being tested against $H_1 : p_1 \neq p_2$ on the basis of two independent sets of 100 Bernoulli trials. If the number of successes in the first set is 60 and the number of successes in the second set is 48, what P-value would be associated with the data?

5. (5 points) Suppose that 100 items are sampled from a manufacturing process and 3 are found to be defective. A beta prior is used for the unknown proportion θ of defective items. Consider

two cases: (1) $a = b = 1$, and (2) $a = 0.5$, $b = 5$. Sketch the two posterior distribitution curves and compare them. Explain the differences referring to the prior distributions.

(Reminder: the beta density is $f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}$, $0 < p < 1$ with mean $\mu = \frac{a}{a+b}$ and variance $\sigma^2 = \frac{\mu(1-\mu)}{a+b+1}$.)

6. (5 points) Bailey, Cox, and Springer (1978) discuss a method of measuring the concentrations of food dyes by high-pressure chromotography. Measurements of the chromotographic peak areas (y_i) corresponding to sulfanilic acid were taken for 21 known concentrations (x_i) of FD&C Yellow No.5. The quadratic regression model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

fits the data with the following coefficients

Coefficient	Estimate	Standard error
β_0	0.58	0.54
β_1	11.17	1.20
β_2	-1.90	5.53

a. Apply the model utility test to see if a quadratic term is really needed. State clearly the null hypothesis and specify the assumptions you make.

b. What exactly would the negative answer imply about a fair description of the relationship in question?

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. Since the extreme data values are obviously more extreme than they would be according the normal distribution we conclude that this is an example of heavy tail distribution.

1b. The left tail is heavier, thus the distribution is skewed to the left and this should be reflected in the negative skewness coefficent. Indeed, for the data the coeficcient of skewness is (-0.68). The coefficient of kurtosis is 4.91 which is larger than 3, this is what one should expect for heavy tailed data.

2a. A sutable model for the number of misses during a game is a Poisson distribution $\mathrm{Pois}(\lambda)$. It is an approximation for the Binomial distribution $\mathrm{Bin}(n, p)$ for large n and small p so that $\lambda = np$. Here n can be treated the number of "trials" during a single game which may result in a miss with probability p.

2b. We test
H_0 : the number of misses in a game follows a Poisson distribution with unspecified parameter λ
against
H_1 : the number of misses in a game does not follow a Poisson distribution.
We use the Pearson chi-square test based on the data summary

Number of fumbles	0	1	2	3	4	5	6	7
Observed counts	8	24	27	20	17	10	3	1

The MLE $\hat{\lambda} = 2.55$ leads to ceratain expected counts which after grouping for ≥ 6 gives a very small value for the observed test statistic $X^2 \approx 2$. The chi-square distribution table with df=5 shows that the p-value of the test is larger than 10%. Therefore we can not reject the null hypothesis and we conclude that a Poisson model gives a fair description of the data in hand.

3a. Suppose you have observed five independent random variables with the same distribution (null hypothesis assumption). If you pick the largest and the smallest values, then these two can not be treated as two independent random variables with the same mean. Therefore, you can not apply a two sample test to see if the difference between the two means is significant. This is a part of the multiple comparisons problem.

3b. Tukey's formula for simultaneous confidence interval

$$(\bar{Y}_{u.} - \bar{Y}_{v.}) \pm q_{I,I(J-1)}(\alpha) \cdot \frac{s_p}{\sqrt{J}} = (\bar{Y}_{u.} - \bar{Y}_{v.}) \pm 6.57.$$

Thus those treatments whose means differ less that 6.57 (for example Pen-Tet) are not significantly different.

4. The observed test statistic is 1.71 for the large sample test for the difference between two independent samples. It follows that the two-sided P-value is 0.09. Two high to reject the null hypothesis of equality between two population proportions against the two sided alternative.

5. The two posterior distributions are Beta(4,98) and Beta(3.5,102). They look similar with slightly different peaks at (0.039, 0.033) and almost equal standard deviations (0.019, 0017). The second mean is smaller because the second prior distribution predicts smaller values for the population proportion in question.

Both curves are bell shaped skewed to the right.

6a. The model utility test is applied to test $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$ under the assumption of the quadratic model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

where all ϵ_i are independent and normally distributed with mean zero and the same variance σ^2. The observerd test statistic (-0.34) is not significant according to the t_{18}-distribution.

6b. The negative answer implies that adding the quadratic term to the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

is irrelevant.