Tentamentsskrivning i Statistisk slutledning, TM, 5p.

Tid: Tisdagen den 28 maj 2002 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.
För "G" fordras 12 poäng, för "VG" - 17 poäng, för "MVG" - 22 poäng.

Important: carrying out a test make sure
1. to state the hypotheses tested,
2. state the statistical test you choose,
3. explain your choice of the test by referring to the conditions assumed by the test.

1. (6 marks) An experimental station wishes to test whether a growth hormone will increase the yield of wheat above the average value of 100 units per plot produced under currently standard conditions. Twelve plots treated with the hormone give the yields:

140, 103, 73, 171, 137, 91, 81, 157, 146, 69, 121, 134.

a. What would be your conclusion about the effectivness of the growth hormone? Hint: choose an appropriate test and find the P-value of the test.

b. What assumptions about the data do you make while answering to part (a) ?

c. Compute a 95% confidence interval for the average wheat yield per plot treated with the hormone.

d. Draw a normal probability plot and explain its use.

2. (7 marks) A research project studied how the amount of carbon fiber and sand additionds affect various characteristics of the molding process (mold = gjuta, forma). In the next table we give data on casting hardness and on wet-mold strength.

a. An ANOVA for wet-mold strength gives SS-sand = 705, SS-fiber = 1278, SSE = 843, and SST = 3105. Test for the presence of any effects using $\alpha = 0.05$.

b. Explain the equality $\text{SST} = 17 \cdot s^2$, where $s = 13.51$ is the sample standard deviation for wet-mold strength given at the bottom of the last column of the table.

c. Carry out an ANOVA on the casting hardness observations using $\alpha = 0.05$. Hint: for your convenience rearrange the columns 2-4 to the usual two-way layout format so that you have 3x3=9 cells with two observations per cell. To speed up calculations make use of columns 5-6 and various s-values given at the bottom of the table.

d. Plot sample mean hardness against sand percentage for different levels of carbon fiber. Is the plot consistent with your analysis in part (c) ?

e. What assumptions do you make for your analysis in parts (a) and (c) ?

No	Sand Addition (%)	Carbon Fiber Addition (%)	Casting Hardness	Cell Means	Level Means	Wet-Mold Strength
1	0	0	61.0			34.0
2	0	0	63.0	62.0		16.0
3	15	0	67.0			36.0
4	15	0	69.0	68.0		19.0
5	30	0	65.0			28.0
6	30	0	74.0	69.5	66.5	17.0
7	0	0.25	69.0			49.0
8	0	0.25	69.0	69.0		48.0
9	15	0.25	69.0			43.0
10	15	0.25	74.0	71.5		29.0
11	30	0.25	74.0			31.0
12	30	0.25	72.0	73.0	71.17	24.0
13	0	0.50	67.0			55.0
14	0	0.50	69.0	68.0		60.0
15	15	0.50	69.0			45.0
16	15	0.50	74.0	71.5		43.0
17	30	0.50	74.0			22.0
18	30	0.50	74.0	74.0	71.17	48.0
			$\bar{X}$=69.61	$\bar{X}$=69.61	$\bar{X}$=69.61	$\bar{X}$=39.94
			s=4.03	s=3.56	s=2.69	s=13.51

3. (7 points) Gamma distribution, Gamma(α, λ), has pdf

$$f(x) = \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x^{\alpha-1} e^{-\lambda x}, \ x \geq 0,$$

mean value $\mu=\frac{\alpha}{\lambda}$, and variance $\sigma^2=\frac{\alpha}{\lambda^2}$. In particular, gamma distribution was used as a model for times (measured in hours) to swim 1 km by 20 bowhead whales:

0.382, 0.205, 0.215, 0.364, 0.232, 0.296, 0.463, 0.352, 0.407, 0.284,
1.191, 0.350, 0.302, 0.327, 0.215, 0.369, 2.500, 1.563, 4.348, 0.134.

a. Why α is called a shape parameter, while λ is called a scale parameter? Draw a histogram for the whale data above. Can you say from the shape of the histogram what option fits best for the data: α is close to one, α is smaller than one, or α is larger than one?

b. What is the relationship between Gamma(α, λ) and exponential distribution in the case when α is a natural number? How this relationship explains close normal approximation of the gamma distribution when α is large?

c. Consider the whale data above. Fit the parameters of the gamma distribution by the method of moments. (Help: the sum of observed times is $t_1 + \ldots t_{20} = 14.50$, and the sum of squares $t_1^2 + \ldots t_{20}^2 = 30.63$.)

d. Describe the major computational steps for finding the maximum likelihood estimates $\hat{\lambda}$ and $\hat{\alpha}$. (Hint: at one of the steps the results of part (c) can be used.)

e. Explain in detail how would you estimate the standard error of the MLE $\hat{\alpha}$ using parametric bootstrap.

4. (5 marks) Suppose we are interested in two normally distributed variables $X \in \mathrm{N}(\mu_x, 4^2)$ and $Y \in \mathrm{N}(\mu_y, 5^2)$ with known variances and unknown mean values.

a. To estimate the difference $\theta = \mu_x - \mu_y$ two independent samples were collected: one of size n for X values and another of size m for Y values. Compute the variance of the point estimate $\hat{\theta} = \bar{X} - \bar{Y}$.

b. What is the sampling distribution of $\hat{\theta}$ and what precisely does it describe? Is this estimate unbiased and consistent? Explain.

c. We want to spend 9 observations measuring $\hat{\theta}$. How should you allocate them to observations on X and observations on Y? Explain.

5. (5 marks) In the 1954 trials of the Salk polio vaccine, 401,974 subjects were observed. The results were as follows:

	Did not get polio	Got polio
Control (not vaccinated)	201,114	115
Vaccinated	200,712	33

a. Describe the design of this experiment.

b. The odds of contracting polio decreases by vaccination. Compute the odds ratio measuring this decrease.

c. How significant is the result of this experiment?

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. Test $H_0: \ \mu = 100$ against one-sided (because the yield is expected to increase) $H_1: \ \mu > 100$ using one-sample t-test.

Observed test statistic $T = \frac{118.58-100}{9.91} = 1.875$, where $\bar{X} = 118.58$ and $s_{\bar{X}} = \frac{34.34}{\sqrt{12}} = 9.91$. One-sided p-value from the t_{11}-distribution table lies between 2.5% and 5%.

Conclusion: the increase of the yield by the new growth hormone is significant at 5% level.

1b. Assumptions: observations are independent and taken form a normal population distribution.

1c. An exact 95% CI for μ is $(118.58 \pm 2.201 \cdot 9.91) = (118.58 \pm 21.81)$ or $(96.77, 140.39)$.

1d. Normal dprobability plot: plot the normal distribution quantiles
$\Phi_{-1}(0.042) = -1.73$, $\Phi_{-1}(0.125) = -1.15$, $\Phi_{-1}(0.208) = -0.81$,
$\Phi_{-1}(0.292) = -0.55$, $\Phi_{-1}(0.375) = -0.32$, $\Phi_{-1}(0.458) = -0.10$,
$\Phi_{-1}(0.542) = 0.10$, $\Phi_{-1}(0.625) = 0.32$, $\Phi_{-1}(0.708) = 0.55$,
$\Phi_{-1}(0.792) = 0.81$, $\Phi_{-1}(0.875) = 1.15$, $\Phi_{-1}(0.958) = 1.73$
against the ordered sample values
69, 73, 81, 91, 103, 121, 134, 137, 140, 146, 157, 171.

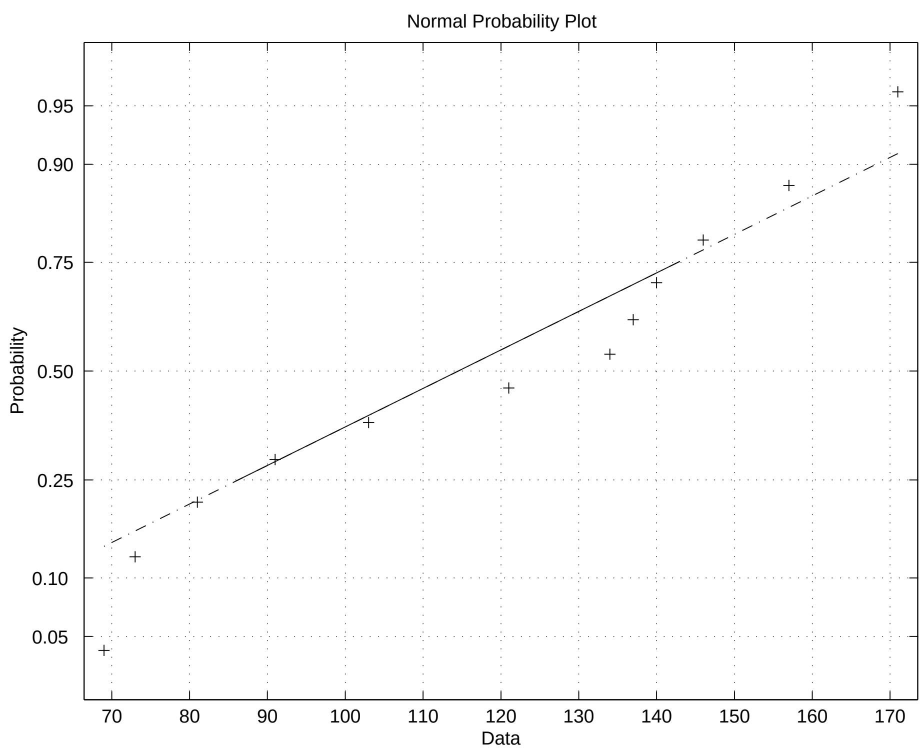

Figure 1: *Normal probability plot*

Tails are lighter than of the normal distribution. Normal distribution model poorly fits to the data.

2a. The ANOVA table for the wet-mold strength

Source	SS	df	MS	F	P
Fiber	1278	2	639	6.83	0.016
Sand	705	2	353	3.77	0.065
Interaction	279	4	70	0.75	0.59
Error	843	9	94		
Total	3105	17			

2b. Follows from the definition of the sample variance.

2c. The ANOVA table for the casting hardness

Source	SS	df	MS	F	P
Fiber	87	2	43.6	5.33	0.03
Sand	107	2	53.4	6.54	0.02
Interaction	8.9	4	2.2	0.27	0.89
Error	73.5	9	8.17		
Total	276.3	17			

Here, in particular, $276.3 = 17 \cdot 4.03^2$ and $87 = 2 \cdot 2.69^2 \cdot 3 \cdot 2$. The SS for sand and error are calculated from the next table.

	Fiber 0	Fiber 0.25	Fiber 0.50	Mean
Sand 0	61 63 (62.0)	69 69 (69.0)	67 69 (68.0)	66.33
Sand 15	67 69 (68.0)	69 74 (71.5)	69 74 (71.5)	70.33
Sand 30	65 74 (69.5)	74 72 (73.0)	74 74 (74.0)	72.17
Mean	66.5	71.17	71.17	69.61

The SS for interaction is calculated by substracting from SST the SS for fiber, sand and error.

2d. On the casting hardness averages graph the crossing lines indicate a possible interaction between two factors. Such an interaction effect is not significant due to the analysis in part (c).

3a. The histogram for the whales data indicates a sharp increase of the pdf near $x = 0$. This feature is expected with the shape parameter α smaller than 1.

3b. If $\alpha = k$ is a natural number, then $X \in \text{Gamma}(\alpha, \lambda)$ can be viewed as a sum of k exponentially distributed iid random values. For large k the distribution of such a sum is well approximated by a normal distribution according to the CLT.

3c. MME: solve two equations $\frac{\tilde{\alpha}}{\tilde{\lambda}} = \frac{14.5}{20}$ and $\frac{\tilde{\alpha}}{\tilde{\lambda}^2} = \frac{30.63}{20} - (\frac{14.5}{20})^2$ to find $\tilde{\lambda} = 0.72$ and $\tilde{\alpha} = 0.52$.

4a. Due to independence $\text{Var}(\bar{X} - \bar{Y}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) = \frac{16}{n} + \frac{25}{m}$.

4b. In accordance with the properties of the normal distribution: $\hat{\theta} \in \text{N}(\mu_x - \mu_y, \frac{16}{n} + \frac{25}{m})$. The sampling distribution describes the variation in values of $\hat{\theta}$ calculated for independent samples of the same size from the same population.

The estimate $\hat{\theta}$ is unbiased since $\text{E}(\hat{\theta}) = \mu_x - \mu_y$, and it is consistent since its variance tends to zero as the sample sizes tend to infinity.

4c. To minimize the variance $\frac{16}{n} + \frac{25}{9-n}$ take the derivative and solve the equation $\frac{16}{n^2} = \frac{25}{(9-n)^2}$. It gives the optimal allocation $n = 4$ and $m = 5$.

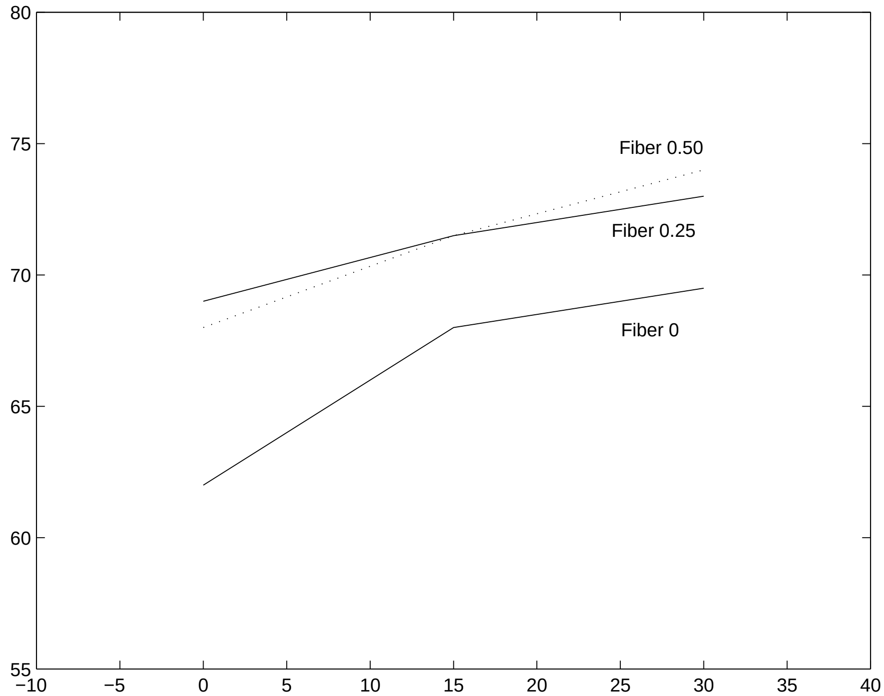

Figure 2: *The casting hardness averages*

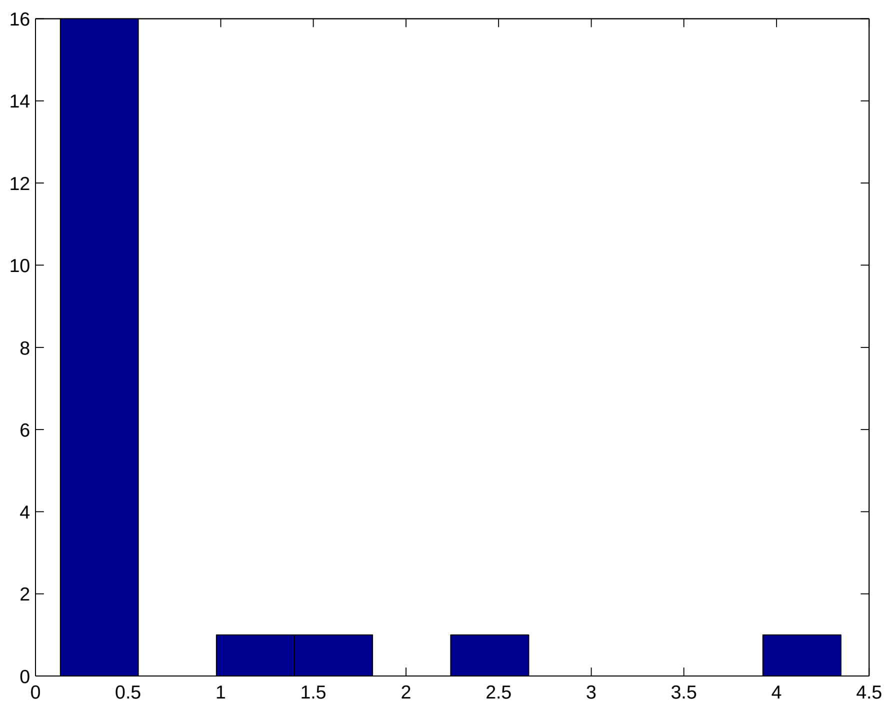

Figure 3: *Histogram for the whales data*

5a. It is a radomized controlled experiment. To avoid possible placebo effect the experiment should be double-blind. Read: http://www.detroitnews.com/history/polio/polio.htm

5b. The odds ratio is $\frac{33}{200745} : \frac{115}{201229} = 1 : 3.5$ so that the decrease of odds due to vaccination is 3.5 times.

5c. Test H_0: $p_1 = p_2$ against two-sided alternative. Here p_1 and p_2 are probabilities of contracting polio without and with vaccination. Apply the chi-squre test of homogeneity. Observed test statistic $X^2 = 45.25$. The P-value of the test $P = 1.7 \cdot 10^{-11}$ shows that the results are very significant.