Tentamentsskrivning i Statistisk slutledning, TMS036/MSN560, 5p.

Tid: Måndagen den 23 maj 2005, kl 14.00-18.00.
Examinator och jour: Serik Sagitov, tel. 772-5351, 073 - 690 76 13, rum MC 1421.
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: carrying out a test make sure

- to state the hypotheses tested,
- state the statistical test you choose,
- explain your choice of the test by referring to the conditions assumed by the test.

---

1. (5 marks) Suppose your prior distribution for θ, the proportion of Californians who support the death penalty, is beta with mean 0.6 and standard deviation 0.3. (Reminder: Beta(a, b) distribution has density $f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}$, mean $\mu = \frac{a}{a+b}$, and variance $\sigma^2 = \frac{\mu(1-\mu)}{a+b+1}$.)

a. Determine the parameters of your prior beta distribution. Sketch the prior density curve. Judging from the curve shape what are our expectations about the value θ prior the data collection?

b. A random sample of 1000 Californians is taken, and 65% support the death penalty. What are your posterior mean and variance for θ? Sketch the posterior density curve and comment on the change of our opinion on the θ value after the sampling.

2. (5 marks) A study on the tensile strength (draghålfasthet) of aluminium rods is conducted. Forty identical rods are randomly divided into four groups, each of size 10. Each group is subjected to a different heat treatment, and the tensile strength, in thousands of pounds per square inch, of each rod is determined. The following data result:

Treatment	**1**	**2**	**3**	**4**	Combined data
	18.9	18.3	21.3	15.9	18.9 18.3 21.3 15.9
	20.0	19.2	21.5	16.0	20.0 19.2 21.5 16.0
	20.5	17.8	19.9	17.2	20.5 17.8 19.9 17.2
	20.6	18.4	20.2	17.5	20.6 18.4 20.2 17.5
	19.3	18.8	21.9	17.9	19.3 18.8 21.9 17.9
	19.5	18.6	21.8	16.8	19.5 18.6 21.8 16.8
	21.0	19.9	23.0	17.7	21.0 19.9 23.0 17.7
	22.1	17.5	22.5	18.1	22.1 17.5 22.5 18.1
	20.8	16.9	21.7	17.4	20.8 16.9 21.7 17.4
	20.7	18.0	21.9	19.0	20.7 18.0 21.9 19.0
mean	20.34	18.34	21.57	17.35	19.40
variance	0.88	0.74	0.88	0.89	3.58
skewness	0.16	0.14	-0.49	-0.08	0.05
kurtosis	2.51	2.59	2.58	2.46	1.98

Table 1: *The tensile strength of 40 aluminium rods*

a. Figure 5 depicts a kernel density estimate for the combined data consisting of all 40 observations. Explain how it is drawn. The bandwidth chosen by Matlab is 1.13. What will happen with the curve if the bandwith is increased? What will happen with the curve if the bandwith is decreased?

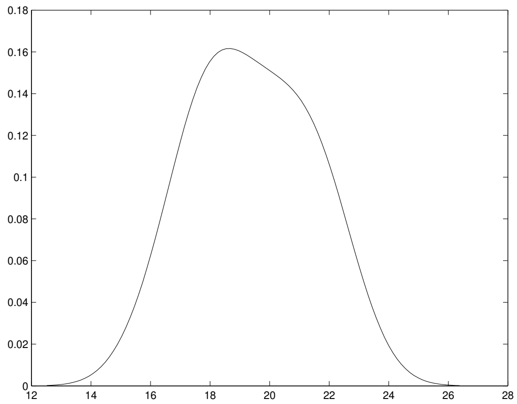

Figure 1: *A kernel density estimate for the combined data*

b. Explain which features of the curve on the Figure 5 are reflected in the values of skewness = 0.05 and kurtosis = 1.98.

c. The kurtosis of the curve on the Figure 5 is actually larger than 1.98 but still smaller than 3. The variance of the curve on the Figure 5 is also larger than the sample variance 3.58. Copy the curve of Figure 5 and try to sketch on top of it a normal probability curve with the same mean and variance.

3. (5 marks) Turn back ot the Table 1 and consider the null hypothesis of equality between the four treatment means of tensile strength.

a. Test the null hypothesis applying an anova test. Show clearly how all the sums of squares are computed using ALL the means and variances given in the Table 1.

b. What are the assumptions of the anova model you used? Do they appear fulfilled?

c. The Bonferroni method suggests the following formula for computing simultaneuos 95% CIs for six pairwise differences between four treatment means

$$(\bar{X}_{i.} - \bar{X}_{j.}) \pm t_{36}(0.0083) \cdot 0.4472 \cdot s_p.$$

Using the simultaneuos 95% CIs check which of the pairs of treatments have significantly different means.

4. (5 marks) A recent study claims that an increasing proportion of engeneering firms are purchasing liability insurance. This claim is based on a survey of 753 engeneering firms. The status of each firm is recorded for the current and for the previous year. The data on which the claim is based are shown in the next table.

	This year insured	This year uninsured
Last year insured	650	5
Last year uninsured	28	70

a. Fill in the following table. Which of the two tables is more informative? Explain.

	Insured	Uninsured
Last year	?	?
This year	?	?

b. State the claim of the study as an alternative hypothesis in terms of appropriate population proportions. What is a relevant null hypothesis?

c. Do the data support the claim? Explain, based on the P-value of an appropriate test. Show how to use the normal probability table for finding the P-value.

5. (5 marks) Consider the random variable X with density given by

$$f(x) = (1+\theta)x^{\theta},\ 0 < x < 1,\ \theta > -1.$$

Five independent observations of this variable produced the numbers 0.5, 0.3, 0.1, 0.1, 0.2.

a. Find the method of moments estimate for θ.

b. Write down the likelihood function $L(\theta)$ for the data given above. Sketch the likelihood curve. Find the maximum likelihood estimate for θ and place it on the likelihood graph.

c. Find a sufficient statistic T for θ. Check that the MLE is a function of T. Why this fact makes the MLE a more sensible estimate of θ than the MME, which is a function of a non-sufficient statistic (which one?).

6. (5 marks) Information about ocean weather can ve extracted from radar returns with the aid of a special algorithm. A study is conducted to estimate the difference in wind speed as measured on the ground and via the Seasat satellite. To do so, wind speeds are measured using the two methods simultaneously at 12 specified times.

Time	Ground windspeed	Satellite windspeed	Difference
1	4.46	4.08	0.38
2	3.99	3.94	0.05
3	3.73	5.00	-1.27
4	3.29	5.20	-1.91
5	4.82	3.92	0.90
6	6.71	6.21	0.50
7	4.61	5.95	-1.34
8	3.87	3.07	0.80
9	3.17	4.76	-1.59
10	4.42	3.25	1.17
11	3.76	4.89	-1.13
12	3.30	4.80	-1.50
mean	4.18	4.59	-0.41
standard deviation	0.964	0.973	1.140

a. Draw a scatterplot showing the relationship between the two measurements. Does the plot strongly indicate positive dependence between the two measurements? Explain. Compute the sample correlation coefficient using the three standard deviations given in the table.

b. Find a 95% confidence interval on the mean difference in measurements taken by these methods. Based on this interval, is there reason to believe that, on the average, the satellite measurements differ from those taken on the ground?

c. Find a 95% prediction interval on a new difference X_{new} for a pair of measurements to be taken next. Use the formula

$$\operatorname{Var}(X_{\text{new}} - \bar{X}) = \operatorname{Var}(X_{\text{new}}) + \operatorname{Var}(\bar{X}).$$

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. $a = 1$, $b = 2/3$. The prior distribution curve shows that before the measurement we were rather uncertain about the parameter θ value. We were expecting it to be closer to one than to zero.

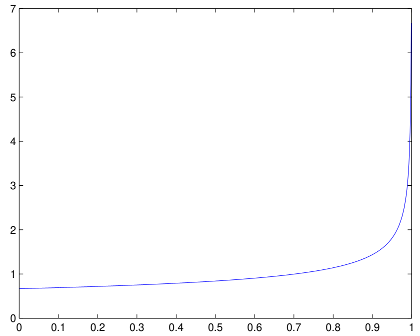

Figure 2: Beta$(1,\frac{2}{3})$ curve

1b. $a = 651$, $b = 350.7$. The posterior distribution curve is narrowly centered around the sample proportion 0.65. The prior distribution has very little contribution to the posterior distribution in comparison with the contribution of the data.

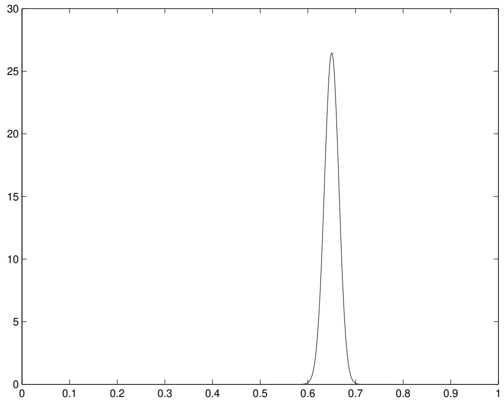

Figure 3: Beta(651,350.7) curve

2a. The dot-line has a larger bandwidth and the dash-lins has smaller bandwidth.

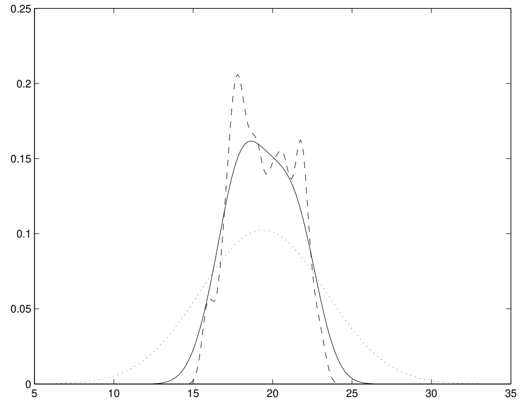

Figure 4: *Solution 2a*

2b. The skewness 0.05 witnesses an asymmetry of the curve around the mean. Indeed, the curve is skewed to the right so that the right tail is heavier. The kurtosis value 1.98 is smaller than the normal value 3. The latter implies that the curve has light tails: its peak is less sharp and the sides are closer to zero, if compared with the normal curve of the same mean and variance (see next figure).

2c.

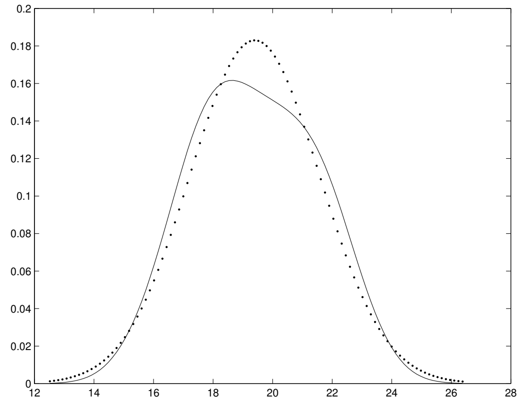

Figure 5: *Solution 2c*

3a. The sums of squares: between samples, within samples and total:
$\mathrm{SS_A} = 10((20.34 - 19.40)^2 + (18.34 - 19.40)^2 + (21.57 - 19.40)^2 + (17.35 - 19.40)^2) = 109.2$
$\mathrm{SS_E} = 9(0.88 + 0.74 + 0.88 + 0.89) = 30.5$
$\mathrm{SS_T} = 3.58 \cdot 39 = 139.7 = 109.2 + 30.5$

Source	SS	df	MS	F	P
Treatment	109.2	3	36.4	42.9	< 0.01
Error	30.5	36	0.85		
Total	139.7	39			

The result is highly significant and we reject the null hypothesis.

3b. The normality assumption is supported by the four skewness and kurtosis values, with the former being close to zero and the latter close to 3. On the other hand, the four sample variances are close to each other making realistic the assumption of equal variances.

3c. Since $s_p = \sqrt{\mathrm{MS_E}} = 0.92$ and the normal distribution table gives approximately $t_{36}(0.0083) \approx 2.40$ (the true value according to Matlab is 2.51), we get

$$(\bar{X}_{i.} - \bar{X}_{j.}) \pm 0.99$$

Therefore all observed pairwise differences (possibly except 2-4) are significant:

Pairs	1-2	1-3	1-4	2-3	2-4	3-4
Differences	2.00	-1.23	2.99	-3.23	0.99	4.22

4a. The first table is more informative than the next one, since one can not recover the former from the latter.

	Insured	Uninsured
Last year	655	98
This year	678	75

4b. In terms of the proportion parameters

	This year insured	This year uninsured	Total
Last year insured	π_{11}	π_{12}	$\pi_{1.}$
Last year uninsured	π_{21}	π_{22}	$\pi_{2.}$
Total	$\pi_{.1}$	$\pi_{.2}$	1

the relevant null hypothesis states that there is no difference between two years $H_0 : \pi_{1.} = \pi_{.1}$. The claim of the study is a one-sided alternative $H_1 : \pi_{1.} < \pi_{.1}$.

4c. The observed McNemar test statistics is $X^2 = \frac{(28-5)^2}{28+5} = 16.03$. Its null distribution is approximated by the χ^2_1 distribution. The two-sided P-value of the test can found from the normal distribution table first taking the square root of X^2, which is 4.00, and then computing $\mathrm{P} = 2(1 - \Phi_{-1}(4.00)) < 0.0004$. The difference between the numbers 28 and 5 in the data table is highly significant.

5a. This is a Beta$(\theta + 1, 1)$ distribution with $\mathrm{E}(X) = \frac{\theta+1}{\theta+2}$. Since $\bar{X} = 0.24$ we solve $\frac{\theta+1}{\theta+2} = 0.24$ to obtain the MME $\tilde{\theta} = -\frac{0.52}{0.76} = -0.684$.

5b. The likelihood function is computed using independence $L(\theta) = (1+\theta)^5(0.0003)^\theta$. Its curve is given in Figure 6. Control values: $L(0) = 1$, $L(1) = 0.01$, $L(-0.5) = 1.80$.

5c. The log-likelihood function $\log L(\theta) = 5\log(1+\theta)^5 - 8.11 \cdot \theta$ and its derivative $\frac{d}{d\theta} \log L(\theta) = \frac{5}{1+\theta} - 8.11$. Solve the equation $\frac{5}{1+\theta} = 8.11$ to find the MLE $\hat{\theta} = -0.384$. MLE and MME are negative but different. Both methods give rough estimates when the sample size is small.

5d. The joint distrubution for the data values $f(x_1, \ldots, x_n|\theta) = (1+\theta)^n(x_1 \ldots x_n)^\theta$. According to the factorization criterium $T = X_1 \ldots X_n$ is a sufficient statistic. It summarizes all information in the sample related to the value θ. The MLE is a function of T since $\hat{\theta}$ is chosen to maximize $L(\theta) = (1+\theta)^5(x_1 \ldots x_n)^\theta$. The MME is a function of a non-sufficient statistic $\bar{X}$ and therefore it misses some information concerning θ.

6a. The scatter plot on the Figure 7 indicates moderate positive correlation. Without the top-rightmost outlier the remaining points indicate independence.

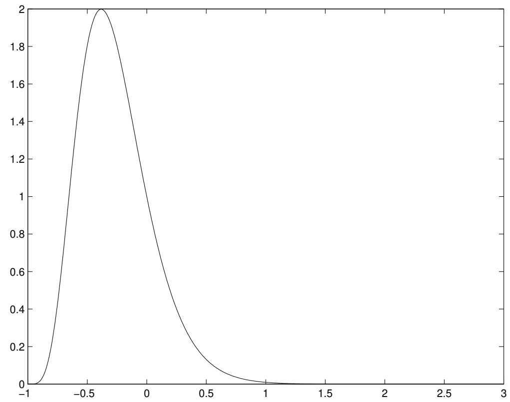

Figure 6: Likelihood curve

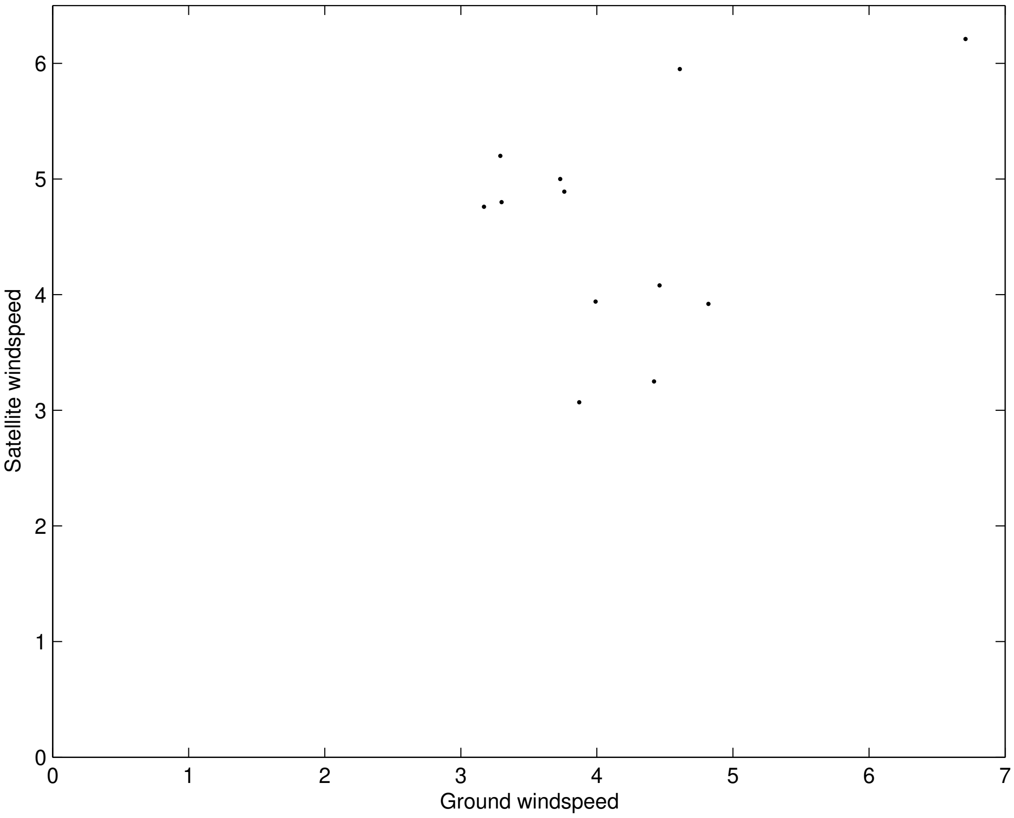

Figure 7: Scatter plot

To estimate ρ use

$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y - 2\sigma_X\sigma_Y\rho$$

which implies $\hat{\rho} = \frac{0.964^2+0.973^2-1.140^2}{2\cdot 0.964\cdot 0.973} = 0.307$.

6b. Assuming normally distributed difference the 95% CI for the average difference is $-0.41 \pm 2.201\cdot 1.14/\sqrt{12} = -0.41 \pm 0.72$. Since the CI covers zero, we conclude that on average the satellite mesurement does not differ on that taken on the ground.

6c. Assuming normally distributed difference we can compute a rough prediction interval in the form $\bar{X} \pm 1.96 s_{\text{new}}$, where $s^2_{\text{new}} = s^2 + s^2_{\bar{X}} = 1.14^2(1 + \frac{1}{12})$ and $s_{\text{new}} = 1.19$ giving a rough prediction interval for the new difference -0.41 ± 2.33. To be more exact we should replace 1.96 taken from the normal distribution table by 2.20 taken from the $t_{11}-$distribution. This results in a wider interval -0.41 ± 2.62.