Tentamentsskrivning i Statistisk slutledning MVE155/MSG200, 7.5 hp.

Tid: 13 mars 2018, kl 14.00-18.00
Examinator och jour: Serik Sagitov, tel. 031-772-5351, rum H3026 i MV-huset.
Hjälpmedel: Chalmersgodkänd räknare, **egen** formelsamling (fyra A4 sidor).
CTH: för "3" fordras 12 poäng, för "4" - 18 poäng, för "5" - 24 poäng.
GU: för "G" fordras 12 poäng, för "VG" - 20 poäng.
Inclusive eventuella bonuspoäng.

Partial answers and solutions are also welcome. Good luck!

1. (5 points) The intensity if light reflected by an object is measured. Suppose there are two types of possible objects, A and B. If the the object is of type A, the measurement is normally distributed with mean 100 and standard deviation 25; if it is of type B, the measurement is normally distributed with mean 125 and standard deviation 25. A single measurement is taken with the value $X = 120$.

(a) What is the likelihood ratio? What does it say?

(b) If the prior probabilities of A and B are 0.75 and 0.25, what is the posterior probability that the item is of type B?

2. (5 points) What is the difference between

(a) the parametric bootstrap and the non-parametric bootstrap?

(b) a confidence interval, a credibility interval, and a prediction interval?

(c) the chi-square test of independence, chi-square test of homogeneity, and goodness of fit chi-square test?

(d) the significance level and the P-value of a test? In what sense the P-value is a random variable?

3. (5 points) Ten paired observations

x	0.34	1.38	-0.65	0.68	1.40	-0.88	-0.30	-1.18	0.50	-1.75
y	0.27	1.34	-0.53	0.35	1.28	-0.98	-0.72	-0.81	0.64	-1.59

have sample means, sample standard deviations, and a sample correlation coefficient given below

$$\bar{x} = -0.046, \quad \bar{y} = -0.075, \quad s_x = 1.076, \quad s_y = 0.996, \quad r = 0.98.$$

(a) Treating x as an explanatory variable and y as the response variable, write down a regression line fitted to the data above. Estimate the size of the noise.

(b) Treating y as an explanatory variable and x as the response variable, write down a regression line fitted to the data above. Estimate the size of the noise.

(c) Draw the regression lines from (a) and (b) on the same plot. Explain the relation between the lines.

(d) Explain the meaning of r^2 both in (a) and (b).

4. (5 points) Consider a single parameter distribution with the probability density function

$$f(x) = c(\theta)x^{\theta-1}e^{-x}, \quad x > 0.$$

Here θ is a positive parameter and $c(\theta)$ is a normalisation constant ensuring that $\int_0^\infty f(x)dx = 1$. It is known that the distribution mean and variance are both equal to θ. For this parametric model with a certain parameter value θ, Matlab generated the following five independent random numbers

$$0.6687, \quad 1.0037, \quad 0.7563, \quad 0.0745, \quad 0.5942,$$

which give three summary statistics

$$x_1 + \ldots + x_5 = 3.0974, \quad x_1^2 + \ldots + x_5^2 = 2.3852, \quad x_1 \cdots x_5 = 0.0225.$$

(a) Using the method of moments, find a point estimate of θ. Sketch the corresponding distribution curve.

(b) For the given data write down the likelihood function. What is a sufficient statistic here?

(c) Describe step by step: how would you compute the maximum likelihood estimate of θ using computer.

(d) How would you estimate the standard error of the maximum likelihood estimate?

5. (5 points) Twenty five independent pairs of observations (X_i, Y_i) are taken from a joint distribution with unknown means (μ_1, μ_2). It is known that the differences $X_i - Y_i$ are normally distributed with unknown variance σ^2. An exact 98% confidence interval for the difference $\mu_1 - \mu_2$ is computed to be 5 ± 5.86.

(a) Shall we reject the null hypothesis of equality $\mu_1 = \mu_2$ to the favour of $\mu_1 \neq \mu_2$ at two percent significance level?

(b) Find $\bar{X} - \bar{Y}$ and its standard error.

(c) Give an unbiased estimate of σ^2.

(d) What is the P-value of the test based on the confidence interval 5 ± 5.86?

6. (5 points) The following data gives the amount of time (in minutes) it took a certain person to drive to work, Monday through Friday, along four different routes.

Route	Mon	Tue	Wed	Thu	Fri	Mean
1	22	26	25	25	31	25.8
2	25	27	28	26	29	27
3	26	29	33	30	33	30.2
4	26	28	27	30	30	28.2
Mean	24.75	27.5	28.25	27.75	30.75	27.8

(a) How would you justify the experimental design chosen for this study?

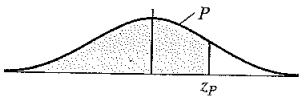
(b) Fill in the missing values into the ANOVA table for the data above. Explain.

Source	SS	df	MS	F
-	52.8	-	-	-
-	73.2	-	-	-
Error	27.2	-	-	
Total	-	-		

(c) State the most relevant H_0 and H_1. Which statistical table would you need for testing? Explain.

(d) Compute the residual value is obtained from the data value 22. How would you verify the normality assumption using the normal probability plot?

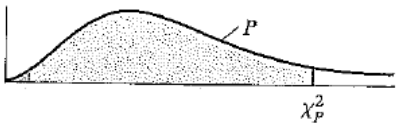
TABLE 2 Cumulative Normal Distribution—Values of P Corresponding to z_p for the Normal Curve



z is the standard normal variable. The value of P for $-z_p$ equals 1 minus the value of P for $+z_p$; for example, the P for -1.62 equals $1-.9474=.0526$.

z_p	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

TABLE 3 Percentiles of the χ^2 Distribution—Values of χ^2_P Corresponding to P



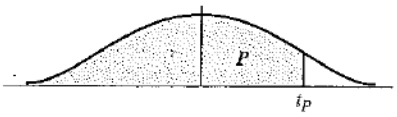
df	$\chi^2_{.005}$	$\chi^2_{.01}$	$\chi^2_{.025}$	$\chi^2_{.05}$	$\chi^2_{.10}$	$\chi^2_{.90}$	$\chi^2_{.95}$	$\chi^2_{.975}$	$\chi^2_{.99}$	$\chi^2_{.995}$
1	.000039	.00016	.00098	.0039	.0158	2.71	3.84	5.02	6.63	7.88
2	.0100	.0201	.0506	.1026	.2107	4.61	5.99	7.38	9.21	10.60
3	.0717	.115	.216	.352	.584	6.25	7.81	9.35	11.34	12.84
4	.207	.297	.484	.711	1.064	7.78	9.49	11.14	13.28	14.86
5	.412	.554	.831	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	.676	.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
120	83.85	86.92	91.58	95.70	100.62	140.23	146.57	152.21	158.95	163.64

For large degrees of freedom,

$$\chi^2_P = \tfrac{1}{2}(z_P + \sqrt{2\upsilon - 1})^2 \text{ approximately,}$$

where υ = degrees of freedom and z_P is given in Table 2.

TABLE 4 Percentiles of the t Distribution



df	$t_{.60}$	$t_{.70}$	$t_{.80}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$
1	.325	.727	1.376	3.078	6.314	12.706	31.821	63.657
2	.289	.617	1.061	1.886	2.920	4.303	6.965	9.925
3	.277	.584	.978	1.638	2.353	3.182	4.541	5.841
4	.271	.569	.941	1.533	2.132	2.776	3.747	4.604
5	.267	.559	.920	1.476	2.015	2.571	3.365	4.032
6	.265	.553	.906	1.440	1.943	2.447	3.143	3.707
7	.263	.549	.896	1.415	1.895	2.365	2.998	3.499
8	.262	.546	.889	1.397	1.860	2.306	2.896	3.355
9	.261	.543	.883	1.383	1.833	2.262	2.821	3.250
10	.260	.542	.879	1.372	1.812	2.228	2.764	3.169
11	.260	.540	.876	1.363	1.796	2.201	2.718	3.106
12	.259	.539	.873	1.356	1.782	2.179	2.681	3.055
13	.259	.538	.870	1.350	1.771	2.160	2.650	3.012
14	.258	.537	.868	1.345	1.761	2.145	2.624	2.977
15	.258	.536	.866	1.341	1.753	2.131	2.602	2.947
16	.258	.535	.865	1.337	1.746	2.120	2.583	2.921
17	.257	.534	.863	1.333	1.740	2.110	2.567	2.898
18	.257	.534	.862	1.330	1.734	2.101	2.552	2.878
19	.257	.533	.861	1.328	1.729	2.093	2.539	2.861
20	.257	.533	.860	1.325	1.725	2.086	2.528	2.845
21	.257	.532	.859	1.323	1.721	2.080	2.518	2.831
22	.256	.532	.858	1.321	1.717	2.074	2.508	2.819
23	.256	.532	.858	1.319	1.714	2.069	2.500	2.807
24	.256	.531	.857	1.318	1.711	2.064	2.492	2.797
25	.256	.531	.856	1.316	1.708	2.060	2.485	2.787
26	.256	.531	.856	1.315	1.706	2.056	2.479	2.779
27	.256	.531	.855	1.314	1.703	2.052	2.473	2.771
28	.256	.530	.855	1.313	1.701	2.048	2.467	2.763
29	.256	.530	.854	1.311	1.699	2.045	2.462	2.756
30	.256	.530	.854	1.310	1.697	2.042	2.457	2.750
40	.255	.529	.851	1.303	1.684	2.021	2.423	2.704
60	.254	.527	.848	1.296	1.671	2.000	2.390	2.660
120	.254	.526	.845	1.289	1.658	1.980	2.358	2.617
∞	.253	.524	.842	1.282	1.645	1.960	2.326	2.576

NUMERICAL ANSWERS

1a. The likelihood ratio is

$$\frac{f(120|A)}{f(120|B)} = \frac{e^{-(\frac{120-100}{25})^2/2}}{e^{-(\frac{120-125}{25})^2/2}} = e^{-0.3} = 0.74.$$

The ratio is less than one being in favour of the outcome B: the model B gives a higher likelihood than the model A.

1b. If the prior probabilities of A and B are 0.75 and 0.25, then the posterior probability that the item is of type B is

$$\mathrm{P}(B|120) = \frac{f(120|B)\cdot 0.25}{f(120|A)\cdot 0.75 + f(120|B)\cdot 0.25} = \frac{1}{\frac{f(120|A)}{f(120|B)}\cdot 3 + 1} = \frac{1}{3.22} = 0.31.$$

3a. Simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2).$$

Fitting a straight line using

$$y - \bar{y} = r \cdot \tfrac{s_y}{s_x}(x - \bar{x})$$

we get the predicted response

$$\hat{y} = -0.033 + 0.904 \cdot x.$$

Estimated σ^2 is

$$s^2 = \tfrac{n-1}{n-2} s_y^2 (1 - r^2) = 0.044.$$

3b. Simple linear regression model

$$X = \beta_0 + \beta_1 y + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2).$$

Fitting a straight line using

$$x - \bar{x} = r \cdot \tfrac{s_x}{s_y}(y - \bar{y})$$

we get the predicted response

$$\hat{x} = 0.033 + 1.055 \cdot y.$$

Estimated σ^2

$$s^2 = \tfrac{n-1}{n-2} s_x^2 (1 - r^2) = 0.052.$$

3c. First fitted line

$$y = -0.033 + 0.904 \cdot x$$

is different from the second

$$y = -0.031 + 0.948 \cdot x.$$

The first line minimises the vertical residuals while the second minimises the horizontal residuals.

3d. Both models have the same coefficient of determination $r^2 = 0.96$. Explain its meaning in both cases.

3c. The two-sample t-test assumes that two independent samples $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_m)$ are taken from two normal distributions with equal variance. To test this normality assumption one may use a normal probability plot for $n+m$ residuals $(X_1 - \bar{X}, \ldots, X_n - \bar{X}, Y_1 - \bar{Y}, \ldots, Y_m - \bar{Y})$.

4a. Since θ is the first population moment, its method of moment estimate is obtained as

$$\tilde{\theta} = \bar{x} = \tfrac{3.1}{5} = 0.62.$$

4b. This is a gamma distribution with the shape parameter $\alpha = \theta$ and the scale parameter $\lambda = 1$. The likelihood function is

$$L(\theta) = c^n(\theta)(x_1 \cdots x_n)^{\theta-1} e^{-x_1-\ldots-x_n} = c^n(\theta)(0.0225)^{\theta-1} e^{-3.0974},$$

where $c(\theta) = 1/\Gamma(\theta)$. There is a single sufficient statistic $t = x_1 \cdots x_n$, so that the MLE of θ is a function of t, all remaining information in the sample data being irrelevant.

4c. Compute log-likelihood

$$l(\theta) = \ln L(\theta) = -n \ln \Gamma(\theta) + (\theta - 1) \ln t - x_1 - \ldots - x_n,$$

and then its derivative

$$l'(\theta) = -5 \tfrac{\Gamma'(\theta)}{\Gamma(\theta)} + \ln 0.0225.$$

Putting this to zero we arrive at the equation

$$\Gamma'(\theta) = \Gamma(\theta) \tfrac{\ln 0.0225}{5} = -0.76 \cdot \Gamma(\theta).$$

Solve this equation numerically using computer with the starting value given by the method of moments estimate 0.62.

4d. Describe the corresponding parametric bootstrap algorithm.

5a. Since the CI covers zero, we do not reject the null hypothesis of equality $\mu_1 = \mu_2$ to the favour of $\mu_1 \neq \mu_2$ at two percent significance level.

5b. From the CI formula for two paired samples

$$5 \pm 5.86 = \bar{X} - \bar{Y} \pm t_{24}(0.01) \cdot s_{\bar{X}-\bar{Y}} = \bar{X} - \bar{Y} \pm 2.5 \cdot \tfrac{s}{\sqrt{n}} = \bar{X} - \bar{Y} \pm \tfrac{s}{2},$$

we find the point estimate $\bar{X} - \bar{Y} = 5$ and its standard error to be $s_{\bar{X}-\bar{Y}} = \frac{5.86}{2.5} = 2.34$.

5c. An unbiased estimate of σ^2 is given by

$$s^2 = (\sqrt{n} \cdot 2.34)^2 = (5 \cdot 2.34)^2 = 137.$$

5d. The observed value of the t-test statistic is $T = \frac{\bar{X}-\bar{Y}}{s_{\bar{X}-\bar{Y}}} = \frac{5}{2.34} = 2.14$. Since $t_{24}(0.025) = 2.064$, the two-sided P-value of the test is slightly smaller than 5%.

6a. The applied experimental design is randomised block design. The main interest in the comparison of different routes. The observations are blocked using five days of the week, which is reasonable as the traffic is expected to be different across the days of the week. If alternatively, all five observations for each route are taken on Mondays, then the comparison will not be valid for different traffic regimes.

6b. To fill in the missing values into the ANOVA table, check first that the mean times for the four routes give the sum of squares

$$\{(25.8 - 27.8)^2 + (27 - 27.8)^2 + (30.2 - 27.8)^2 + (28.2 - 27.8)^2\} \cdot 5 = 52.8.$$

Source	SS	df	MS	F
Route	52.8	3	17.6	7.75
Days	73.2	4	18.3	8.06
Error	27.2	12	2.27	
Total	153.2	19		

6c. One expects significant differences of traffic intensities across the week days. Therefore, the most interesting H_0 is that stating no difference among four routes versus H_1 of significant difference among the routes. The collected data indicates that the route 1 is the fastest. To test that this result is significant we could use the $F_{3,12}$-distribution table.

6d. The residual value corresponding to the data value 22 is

$$\hat{\epsilon}_{11} = 22 - 25.8 - 24.75 + 27.8 = -0.75.$$

To verify the normality assumption, compute on a similar way all 20 residuals and then draw the normal probability plot based on these 20 residuals.