Tentamentsskrivning i Statistisk slutledning MVE155/MSG200, 7.5 hp.

Tid: 14 mars 2017, 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) 1600 British citizens were surveyed on the Prime Minister's job performance. Each citizen rated the Prime Minister as "A" = approve or "D" = disapprove. Then, after 6 months, each citizen re-rates the Prime Minister. The following two tables summarize the data.

	A	D
1st Survey	944	656
2nd Survey	880	720

	2nd A	2nd D
1st A	794	150
1st D	86	570

(a) Explain the relationship between these two tables.

(b) State a relevant pair of hypotheses: first in words, then in a parametric form. Test the null hypothesis at 1% significance level. Justify your choice of the test.

(c) Estimate the odds ratio measuring association between the first survey approval rate and the second survey approval rate. Explain the meaning of this odds ratio in terms of the conditional odds.

2. (5 points) Comment on the following reasoning of a student. Demonstrate your deeper understanding of the inference concepts used by the student.

> "I applied three tests to these data. The sign test has the largest p-value, one-sample t-test is the medium, signed-rank is the smallest. Therefore, the signed-rank test has more power."

3. (5 points) Consider the following sample of size 6

$$x_1 = 1.42, \quad x_2 = 0.58, \quad x_3 = -0.36, \quad x_4 = 3.76, \quad x_5 = 2.36, \quad x_6 = -1.76.$$

(a) Draw a normal probability plot for this data.

(b) Using the normal probability plot estimate the population mean and standard deviation. Compare these estimates with the sample mean and sample standard deviation.

(c) How the normal probability plot is used in connection to the two-sample t-test?

4. (5 points) Suppose that in a one-way layout there are 10 treatments and seven observations under each treatment.

(a) What is the ratio of the length of a simultaneous confidence interval for the difference of two means formed by Tukey's method to that of one formed by the Bonferroni method?

(b) How do both of these compare in length to an interval that does not take account of multiple comparisons?

5. (5 points) Assume the geometric distribution model $p_k = (1-p)^k p$, $k = 0, 1, 2, \ldots$ for the following discrete data

$$x_1 = 3, \quad x_2 = 0, \quad x_3 = 6, \quad x_4 = 2, \quad x_5 = 3.$$

(a) Show that beta-distribution is a conjugate prior. Clearly state the updating rules for the pseudocounts.

(b) Check that the variance of the posterior beta-distribution is smaller than the variance of the prior beta-distribution.

(c) Find a posterior mean estimate for the given data set.

6. (5 points) For the sample of size 9

$$\begin{aligned} &y_1 = 1.7, \quad y_2 = 1.9, \quad y_3 = 6.1, \\ &y_4 = 13.6, \quad y_5 = 19.8, \quad y_6 = 25.2, \\ &y_7 = 13.4, \quad y_8 = 20.9, \quad y_9 = 25.1, \end{aligned}$$

consider a multiple regression model

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2),$$

involving dummy variables through the following special design matrix

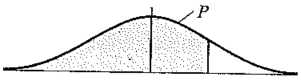
$$\mathbb{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

(a) In which cases the assumption of normality with the same variance can be justified by the central limit theorem argument?

(b) This multiple regression setting is equivalent to the one-way ANOVA model with three levels for the main factor. Express the corresponding three population means μ_1, μ_2, μ_3 in terms of the parameters of the multiple regression model.

(c) Estimate σ^2 using the ANOVA approach.

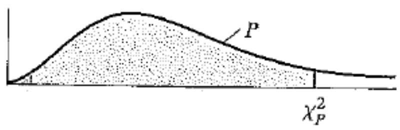
TABLE 2 Cumulative Normal Distribution—Values of P Corresponding to z_p for the Normal Curve



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



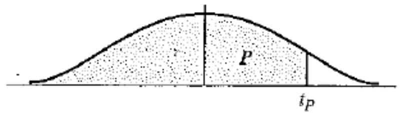
df	$\chi^2_{.005}$	$\chi^2_{.01}$	$\chi^2_{.025}$	$\chi^2_{.05}$	$\chi^2_{.10}$	$\chi^2_{.90}$	$\chi^2_{.95}$	$\chi^2_{.975}$	$\chi^2_{.99}$	$\chi^2_{.995}$
1	.000039	.00016	.00098	.0039	.0158	2.71	3.84	5.02	6.63	7.88
2	.0100	.0201	.0506	.1026	.2107	4.61	5.99	7.38	9.21	10.60
3	.0717	.115	.216	.352	.584	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,

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

where v = 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

TABLE 6 Percentiles of the Studentized Range: $q_{.95}$ (Continued)

v \ t	2	3	4	5	6	7	8	9	10
1	17.97	26.98	32.82	37.08	40.41	43.12	45.40	47.36	49.07
2	6.08	8.33	9.80	10.88	11.74	12.44	13.03	13.54	13.99
3	4.50	5.91	6.82	7.50	8.04	8.48	8.85	9.18	9.46
4	3.93	5.04	5.76	6.29	6.71	7.05	7.35	7.60	7.83
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99
6	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49
7	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92
9	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74
10	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60
11	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39
13	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20
16	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15
17	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11
18	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07
19	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04
20	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01
24	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92
30	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73
60	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65
120	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56
∞	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47

NUMERICAL ANSWERS

1a. This is an example of a pair-matched sample. The left table can be obtained from the margins of the right table, so that the full data is given by the right table.

	A	D
1st Survey	944	656
2nd Survey	880	720

	2nd A	2nd D	
1st A	794	150	944
1st D	86	570	656
	880	720	1600

1b. H_0: no change in the approval of the PM's performance between two surveys, against H_1: the approval rate has changed. Let π_{ij} be the joint probabilities for the cross-classification of 1600 pairs of answers. Then in the parametric form we an write H_0: $\pi_{12} = \pi_{21}$ and H_1: $\pi_{12} \neq \pi_{21}$.

Since the categorical data is paired, we apply the McNemar test. Observed test statistic $X^2 = \frac{(150-86)^2}{150+86} = 17.35$. The square root of this is larger than 4, which implies that we should reject the null hypothesis at 1% significance level. The aproval rate went down during the 6 month period.

1c. To estimate the odds ratio

$$\Delta = \frac{\text{odds}(A|\text{1st survey})}{\text{odds}(A|\text{ 2nd survey})} = \frac{P(A1)P(D2)}{P(D1)P(A2)},$$

we use the left table

$$\hat{\Delta} = \frac{944 \times 720}{656 \times 880} = 1.18.$$

The odds of approval in the first survey were 1.18 higher than the odds of approval in the second survey.

2. The student is confused about the meanings of the p-value on one hand, and the power of the test on the other hand. Suppose the null hypothesis of interest is no difference for two paired samples. To compare the powers of the three tests, one could simulate many samples from a distribution for the differences with a non-zero median. The test rejecting the null hypothesis most often (assuming the same significance level) will have the highest power.

3a. Plotting the ordered data [-1.76;-0.36; 0.58; 1.42; 2.36; 3.76] on y-axis against normal distribution quantiles [-1.38; -0.675; -0.21; 0.21; 0.675; 1.38] on the x-axis you will get a straight line $y = 1 + 2x$.

3b. The sample standard error $s = 1.96$ is pretty close to the estimated value 2 obtained in 3a.

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. For $I = 10$ treatments and $J = 7$ observations under each treatment, we have Bonfrerroni's formula of the half-width

$$t_{I(J-1)}(\frac{\alpha}{I(I-1)})s_p\sqrt{\frac{2}{J}} = 0.53 s_p t_{60}(\frac{\alpha}{90})$$

and Tukey's formula

$$q_{I,I(J-1)}(\alpha)s_p\sqrt{\frac{1}{J}} = 0.38 s_p q_{10,60}(\alpha).$$

With $\alpha = 5\%$, their ratio becomes

$$\frac{\text{Tukey}}{\text{Bonferroni}} = \frac{0.38 \times 4.65}{0.53 \times t_{60}(0.00055)} \approx \frac{0.38 \times 4.65}{0.53 \times 3.3} \approx 1,$$

where $t_{60}(0.00055)$ is approximated by $z(0.00055) = 3.3$ using the normal distribution table.

4b. The half-width of the interval that does not take account of multiple comparisons is

$$t_{I(J-1)}(\frac{\alpha}{2})s_p\sqrt{\frac{2}{J}} = 0.53 s_p \times 2.00$$

so that

$$\frac{\text{Tukey}}{\text{single pair}} = \frac{0.38 \times 4.65}{0.53 \times 2.00} = 1.67.$$

Without taking account of multiple comparisons the CI is much narrower producing an excess of false positive results.

5a. A prior beta-distribution

$$g(p) \propto p^{a-1}(1-p)^{b-1}$$

and a geometric likelihood function

$$f(k|p) = p(1-p)^k$$

give a posterior beta-distribution

$$h(p|k) \propto g(p)f(k|p) \propto p^a(1-p)^{b+k-1}.$$

The udating rule for the parameters of the beta distributions is

$$a' = a+1, \quad b' = b+k.$$

For several observations $k_1, \ldots, k_n$, the updating rule becomes

$$a' = a+n, \quad b' = b+k_1+\ldots+k_n.$$

5b. For the given data the updating rule is

$$a' = a+5, \quad b' = b+14.$$

Since we are not given parameters for the prior we will use the non-informative Beta(1,1) distribution. The mean of the prior beta-distribution is $\mu = \frac{a}{a+b} = \frac{1}{2}$, and variance is

$$\sigma^2 = \frac{\mu(1-\mu)}{a+b+1} = \frac{1}{12} = 0.083.$$

The mean of the posterior beta-distribution is $\mu' = \frac{6}{21}$, and the variance is much smaller

$$(\sigma')^2 = \frac{(6/21)(15/21)}{22} = 0.009.$$

5c. A posterior mean estimate for p is $\hat{p}_{\text{PME}} = \mu' = \frac{6}{21} = 0.29$.

6a. The normality assumption can be justified in the case when the noise value is the sum of many independent and relatively small factors. Equal variance is realistic if the external factors are more or less the same across the three different experiments.

6b. From the given design matrix we obtain

$$\begin{aligned}\mu_1 =& EY_1 = EY_2 = EY_3 = \beta_0,\\ \mu_2 =& EY_4 = EY_5 = EY_6 = \beta_0 + \beta_1,\\ \mu_3 =& EY_1 = EY_2 = EY_3 = \beta_0 + \beta_2.\end{aligned}$$

6c. We estimate σ^2 using the formula $s_p^2 = MS_E = \frac{SS_E}{df_E}$. From the data

$$\begin{aligned}&y_1 = 1.7, \quad y_2 = 1.9, \quad y_3 = 6.1, \quad \hat{\mu}_1 = 3.23;\\ &y_4 = 13.6, \quad y_5 = 19.8, \quad y_6 = 25.2, \quad \hat{\mu}_2 = 19.53;\\ &y_7 = 13.4, \quad y_8 = 20.9, \quad y_9 = 25.1, \quad \hat{\mu}_2 = 19.3;\end{aligned}$$

we find

$$SS_E = \sum_{j=1}^{3}(y_j - \hat{\mu}_1)^2 + \sum_{j=4}^{6}(y_j - \hat{\mu}_2)^2 + \sum_{j=7}^{9}(y_j - \hat{\mu}_3)^2 = 153.4.$$

Since $df_E = 3 \cdot (3-1) = 6$, we conclude $s_p^2 = \frac{153.4}{6} = 25.6$.