Tentamentsskrivning i Statistisk slutledning MVE155/MSG200, 7.5 hp.

Tid: 17 mars 2020, kl 14.00-18.00
Examinator och jour: Serik Sagitov, tel. 031-772-5351
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.

This is a home-written exam.

The examination must be conducted individually, that is, cooperation is not allowed.

You have 4 hours to complete the exam. Solutions are written on paper, or digitally on a digital writing pad if you have access to it. Never write more than one task on each sheet. After 4 hours, you have 30 minutes to scan / photograph your solutions and organize and submit your solutions according to one of the following ways in order of priority:

1. A single pdf file where the pages are arranged in the order of the questions.
2. Image files (jpg or png) or pdf files where each file contains solution to just one question and named according to ”Question 1”, or ”Question 1 page 1”, Question 1 page 2 ”etc if there are multiple pages to a question.

---

1. (5 points) Miscellaneous questions.

(a) The Oxford English Dictionary is widely regarded as the accepted authority on the English language. It is an unsurpassed guide to the meaning, history, and pronunciation of 600,000 words, past and present, from across the English-speaking world. Propose a statistical inference algorithm for estimating your English vocabulary size using the online version of the Oxford English Dictionary.

(b) Explain the difference between the Kruskal-Wallis and the Fridman tests referring to the dataset

Placebo	Treatment 1	Treatment 2
174	263	105
224	213	103
260	231	145
225	291	108

(c) Given that a beta posterior distribution is skewed to the right, which of the two estimates for the population proportion would be larger: the MAP estimate or the PME estimate? Explain by drawing a skewed posterior curve.

2. (5 points) A population consists of three subpopulations whose relative sizes are $w_1 = 0.5$, $w_2 = 0.25$, $w_3 = 0.25$. The variable of interest X, characterising a generic element of the population, is normally distributed over each of the subpopulations. The normal distribution of subpopulation i has mean μ_i (possibly different for different $i = 1, 2, 3$) and standard deviation σ which is the same across subpopulations.

(a) In the current setting, what is the optimal allocation of 100 sample observations among three strata for estimation of the population mean μ? Explain.

(b) Three independent samples each of size 10 were drawn from the three strata. The data produced three sample means $\bar{x}_1 = -0.3$, $\bar{x}_2 = 0.8$, $\bar{x}_3 = -0.5$ and the pooled sample standard deviation $s_p = 1.1$. Do the data reveal a significant difference among three strata means μ_1, μ_2, μ_3?

(c) Suppose we know that $\mu_1 = 0, \mu_2 = 1, \mu_3 = 2$, and $\sigma = 1$. What is the variance $\mathrm{Var}(X)$ for the whole population in this case?

(d) Given $\mu_1 = 0, \mu_2 = 1, \mu_3 = 2$, and $\sigma = 1$, draw a sketch depicting three subpopulation distributions. On top of these three curves draw the overall population distribution.

3. (5 marks) A computer program has simulated two independent values (x_1, x_2) from a normal distribution $\mathrm{N}(0, \sigma)$ with zero mean and unknown to you standard deviation σ.

(a) Show that

$$\hat{\sigma}^2 = \frac{x_1^2 + x_2^2}{2}$$

is an unbiased estimate of the variance σ^2. Would this estimate be biased if the random variables (X_1, X_2) are positively correlated?

(b) The scaled estimate

$$\frac{2\hat{\sigma}^2}{\sigma^2}$$

has a particular sampling distribution, what is it? Using the table for this distribution, construct a formula for an exact 95% confidence interval for σ^2.

(c) Show that

$$\tilde{\sigma}^2 = \frac{(x_1 - x_2)^2}{2}$$

is also is an unbiased estimate of the variance σ^2. What its relation to the sample variance s^2? Would this estimate be biased if the random variables (X_1, X_2) are positively correlated?

(d) Which of these two unbiased point estimates of σ^2 would you prefer? Explain why.

4. (5 points) The following (ordered) 16 numbers are computer generated from $\mathrm{N}(\mu, \sigma)$

0.11	1.60	1.61	1.72
2.28	3.12	3.15	3.53
3.70	4.15	4.25	4.74
5.33	5.49	6.39	6.59

(a) Compute the inter-quartile range for this sample. In what sense this measure of dispersion is robust against outliers?

(b) Estimate σ using the the inter-quartile range of the standard normal distribution.

(c) Sketch the normal probability plot using only lower quartiles, medians, and upper quartiles. Explain how you did it step by step.

5. (5 points) In a political poll survey two independently chosen at random groups of voters were asked whether they would vote for the political party $\mathcal{P}$. Each person answered either yes or no (which even includes don't know option). Group 2, which consisted of 2000 voters, was contacted in August 2019, while group 1, which also consisted of 2000 voters, was asked first in August 2018 and then again in August 2019. The purpose of the survey was to compare two population proportions:

- p_1 proportion of people supporting party $\mathcal{P}$ in August 2018,
- p_2 proportion of people supporting party $\mathcal{P}$ in August 2019.

The percentages of yes answers obtained by the survey were as follows

- group 1 in August 2018: 10%,
- group 1 in August 2019: 12%,
- group 2 in August 2019: 13%.

(a) What is you best point estimate of the population proportion p_2. Compute its standard error.

(b) Find a 95% confidence interval for the difference $p_2 - p_1$. Justify the choice of the formula you apply.

(c) Using the additional information that 200 out 2000 people in the group 1 have changed their answers between 2018 and 2019 (either from yes to no, or from no to yes), and disregarding the response of group 2, test $H_0 : p_1 = p_2$ against $H_1 : p_1 \neq p_2$.

6 (5 marks) The data below show rounded-to-integer values of x = frequency (MHz) and y = output power (W) for a certain laser configuration.

x	60	63	77	100	125	157	186	222
y	16	17	19	21	22	20	15	5

The Matlab 'regress' command yields the following information for a quadratic regression model:

b =
-1.5127
0.3919
-0.0016
bint =
-2.9440 -0.0814
0.3678 0.4160
-0.0017 -0.0015
r =
-0.1283 0.2980 0.0089 -0.3634 0.0158 0.1968 0.0593 -0.0870

where b stands for the estimated parameters β_0, β_1, β_2, bint gives three 95% confidents intervals, and r gives 8 residuals. The sum of squares of the residuals is 0.2874. The sample standard deviation of y is 5.3835.

(a) Does the quadratic model appear to be suitable for explaining observed variation in output power by relating it to frequency? Answer by applying a relevant parametric statistical test. What are your assumptions about the underlying statistical model? How do you verify the key assumption?

(b) Find the adjusted coefficient of determination. What does is say?

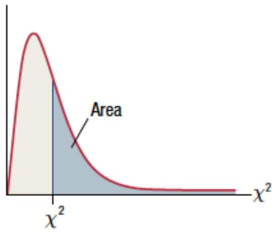
(c) The sum of the 8 residuals is 0. Prove in the simple linear regression setting that the sum of the residuals equals zero.

(d) Draw by hand the scatter plot for the data and then on top of the scatter plot draw the line predicted by the quadratic model.

Normal distribution table

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Chi-square distribution table



Area to the Right of the Critical Value of χ^2

df	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Critical values of t-distribution

df/ α =	.40	.25	.10	.05	.025	.01	.005	.001	.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.385	3.646
35	0.255	0.682	1.306	1.690	2.030	2.438	2.724	3.340	3.591
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	3.307	3.551
50	0.255	0.679	1.299	1.676	2.009	2.403	2.678	3.261	3.496
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	3.160	3.373
inf.	0.253	0.674	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Critical values of the F-distribution (continued)

		Degrees of freedom in the numerator								
Degrees of freedom in the denominator	*p*	1	2	3	4	5	6	7	8	9
	.100	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00
	.050	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
18	.025	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93
	.010	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
	.001	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56
	.100	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98
	.050	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
19	.025	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88
	.010	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
	.001	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39
	.100	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96
	.050	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
20	.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84
	.010	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
	.001	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24
	.100	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95
	.050	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
21	.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80
	.010	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
	.001	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11
	.100	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93
	.050	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
22	.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76
	.010	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
	.001	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99
	.100	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92
	.050	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
23	.025	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73
	.010	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
	.001	14.20	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89
	.100	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91
	.050	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
24	.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70
	.010	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
	.001	14.03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80
	.100	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89
	.050	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
25	.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68
	.010	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
	.001	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71
	.100	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88
	.050	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
26	.025	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65
	.010	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
	.001	13.74	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64
	.100	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87
	.050	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
27	.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63
	.010	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
	.001	13.61	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.57

NUMERICAL ANSWERS

1a. Choose at random $n = 1000$ English words from the dictionary and compute the proportion $\hat{p}$ of the words you knew. A good point estimate of your English vocabulary size would be

$$\hat{N} = \hat{p} \cdot 600000.$$

1b. The assumption for the Kruskal-Wallis is three independent samples and you rank the pooled sample

Placebo	Treatment 1	Treatment 2
5	11	2
7	6	1
10	9	4
8	12	3

and the main idea is to check if the three rank averages are significantly different.

For the Fridman test the three columns are not independent since they involve four subjects responding to the three treatments. Here the ranking is done separately by each subject

Placebo	Treatment 1	Treatment 2
2	3	1
3	2	1
3	2	1
2	3	1

and the Fridman idea is (again) to see if the rank means for treatments are significantly different.

1c. Typically, a distribution skewed to the right has a mean larger than its mode. In this case, PME > MAP.

2a. Given the three standard deviations are equal, the optimal allocation is the same as the proportional

$$n_1 = 50, n_2 = 25, n_3 = 25.$$

2b. We apply the one-way anova test. Using the the samples means we obtain

$$\bar{x}_{..} = (10 \cdot (-0.3) + 10 \cdot (0.8) + 10 \cdot (-0.3))/30 = 0,$$
$$SS_A = 10 \cdot ((0.3)^2 + (0.8)^2 + (0.3)^2) = 9.8,$$
$$MS_A = 9.8/2 = 4.9.$$

Moreover, we know that $MS_E = s_p^2 = 1.21$. Thus the observed value of the F-test statistics is

$$F = \tfrac{MS_A}{MS_E} = \tfrac{4.9}{1.21} = 4.05.$$

Turning to the table for $F_{2,27}$ we find that the p-value of the test is between 2.5% and 5%. We conclude that the difference between the three population means is significant at 5% significance level.

2c. Using

$$\mu = w_1\mu_1 + w_2\mu_2 + w_3\mu_3 = 0.75,$$
$$\sigma^2 = \overline{\sigma^2} + \sum_{j=1}^{3} w_j(\mu_j - \mu)^2,$$

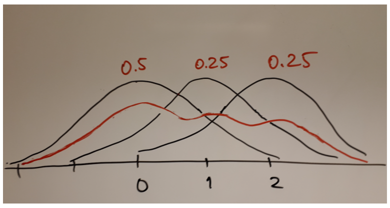
where

$$\overline{\sigma^2} = w_1\sigma_1^2 + w_2\sigma_2^2 + w_3\sigma_3^2 = 1,$$

we find

$$\sigma^2 = 1.69.$$

2d.



3a. Since $\mu = 0$, we have

$$\mathrm{E}(X_i^2) = \mathrm{Var}(X_1) = \sigma^2, \quad i = 1, 2,$$

and

$$\mathrm{E}(\tfrac{X_1^2+X_2^2}{2}) = \sigma^2.$$

This is true even if (X_1, X_2) are dependent. The estimate $\hat{\sigma}^2$ is unbiased.

3b. The distribution in question is the chi-square distribution with two degrees of freedom. From the table we find the 95% confidence interval to be

$$I_{\sigma^2} = (0.27\hat{\sigma}^2, 39.22\hat{\sigma}^2).$$

3c. It is easy to show that

$$\tilde{\sigma}^2 = s^2,$$

which implies that $\tilde{\sigma}^2$ is also an unbiased estimate of σ^2. It is not unbiased if (X_1, X_2) are positively correlated, $\tilde{\sigma}^2$ would systematically underestimate σ^2.

3d. The variance $\tilde{\sigma}^2$ is twice as large as $\hat{\sigma}^2$, therefore $\hat{\sigma}^2$ is a better estimate.

4a. From the sample values we find

$$\hat{x}_{0.25} = 2, \quad \hat{x}_{0.5} = 3.615, \quad \hat{x}_{0.75} = 5.035,$$

which gives IQR$= 3.04$. Robustness means: if we add an unusually large sample value s^2 would change dramatically, but not IQR.

4b.

$$\tfrac{3.04}{1.35} = 2.25.$$

4c. The simplified normal probability plot is the scatter plot of three points on the plane $(2, -0.675)$, $(3.615, 0)$, $(5.035, 0.675)$.

5a. Pooling together 4000 observations for August 20019 we get

$$\hat{p}_2 = 0.125, \quad s_{\hat{p}_2} = 0.00523.$$

5b. We apply the formula

$$I_{p_1-p_2} \approx \hat{p}_1 - \hat{p}_2 \pm 1.96\sqrt{\tfrac{\hat{p}_1(1-\hat{p}_1)}{n-1} + \tfrac{\hat{p}_2(1-\hat{p}_2)}{m-1}}.$$

which requires independence of two samples. Therefore we use only group 1 in 2018 and group 2 in 2019

$$I_{p_1-p_2} \approx 0.10 - 0.13 \pm 1.96\sqrt{\tfrac{0.1(1-0.1)}{1999} + \tfrac{0.13(1-0.13)}{1999}} = -0.03 \pm 0.02.$$

The interval does not cover zero and we conclude that the difference between p_1 and p_2 is significant at 5% level.

5c. The matched pairs design results in the following observed counts

	2019 yes	2019 no	Total
2018 yes	120	80	200
2018 no	120	1680	1800
Total	240	1760	2000

The McNemar's test statistic is

$$\frac{(120-80)^2}{120+80} = 8.$$

Taking the square root of 8 and using the normal distribution table we find the p-value

$$2(1 - 0.9977) = 0.005$$

to be 0.5%. We reject $H_0 : p_1 = p_2$.

6a. We apply a multiple regression setting

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon,$$

where ϵ is a normally distributed $\mathrm{N}(0, \sigma)$ homoscedastic noise. Using the model utility test for

$$H_0 : \beta_2 = 0$$

we observe that the corresponding confidence interval

$$I_{\beta_2} = (-0.0017, -0.0015)$$

does not cover 0 and we reject the null hypothesis at 5% significance level. We conclude that the quadratic model does appear to be suitable. To check the assumption concerning the noise one may draw a normal probability plot for the residuals.

6b. The adjusted coefficient of determination

$$R_a^2 = 1 - \frac{s^2}{s_y^2} = 1 - \frac{0.2874/5}{(5.3835)^2} = 0.998$$

says that 99.8% of the variation in the respons variable is explained by the quadratic model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2.$$

6c. Because

$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n}(y_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y}) = s_y \sum_{i=1}^{n} \frac{\hat{y}_i - \bar{y}}{s_y} = s_y r \sum_{i=1}^{n} \frac{\hat{x}_i - \bar{x}}{s_x} = 0.$$

6d. The scatter plot of the data clearly indicates non-linear relationship - seemingly quadratic.