Tentamentsskrivning i Statistisk slutledning MVE155/MSG200, 7.5 hp.

Tid: 29 augusti 2019, 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) A state teachers association studied the educational qualifications of its membership, which consists of 50 000 teachers. The analyst used a sample design in which independent simple random samples of teachers are selected from among the member teachers in three school levels.

Stratum	Level of school	Number of members	Sample size
1	Elementary	25 000	100
2	High school	20 000	100
3	College	5 000	100
	Total	50 000	300

(a) What are possible advantages and disadvantages of this sampling design compared to a single random sample of size $n = 300$ taken from the whole population?

(b) The following data gives the sample means and variances for the reported numbers of years of education:

Level of school	Mean	Variance
Elementary	14.8	6.4
High school	17.3	2.7
College	19.3	3.6

Find a 99% confidence interval I_μ of the population mean. What are your assumptions?

(c) Without any prior knowledge on the variation within the strata, how would you allocate 300 observations among the three strata in a more effective way? Explain your choice.

2. (5 points) Turn to the data on the previous problem and treat the three strata as three separate populations. Denote by μ_1, μ_2, μ_3 the corresponding population means.

(a) Build a 95% confidence interval for the difference $\delta = \mu_1 - \mu_3$ based on a t-distribution. What is wrong with just checking if this interval does not contain zero, to reject the $H_0 : \mu_1 = \mu_2 = \mu_3$?

(b) Test the null hypothesis of no difference $H_0 : \mu_1 = \mu_2 = \mu_3$. What are your assumptions?

(c) A new study is planned to compare female and male means. There will be six independent samples of size 50 taken: so that for each of the level of school, 50 female members and 50 male members will be sampled at random. Describe a suitable parametric test.

3. (5 marks) Think of a chi-square distribution χ^2_k with k degrees of freedom. It is connected to the standard normal distribution as follows: if $Z_1, \ldots, Z_k$ are N(0,1) and independent, then

$$Z_1^2 + \ldots + Z_k^2 \sim \chi^2_k. \tag{1}$$

Denote by μ and σ^2 the mean and variance of χ_k^2-distribution.

(a) Using (1) compute μ.

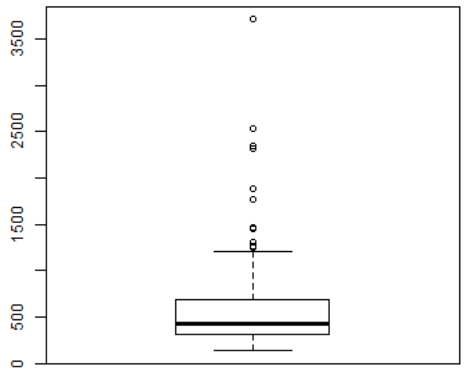
(b) Let $Z \sim \mathrm{N}(0,1)$. Using the fact that the curtosis of a normal distribution equals 3, show that

$$\mathrm{Var}(Z^2) = 2.$$

(c) Using (1) and (b) compute σ^2.

(d) The chi-square distribution has the same density as a gamma distribution with the shape parameter $\frac{k}{2}$ and the scale parameter $\frac{1}{2}$. Using this fact verify your answers for (a) and (c).

4. (5 points) Draw a copy of the boxplot on your answer paper.



(a) Compute the range and the inter-quartile range for the underlying sample.

(b) Is the distribution skewed to the right or to the left? Which is larger: the sample mean or the median? Explain.

(c) Roughly sketch the corresponding normal probability plot.

(d) Suggest a data transformation which will make the empirical distribution to look more like a normal distribution.

5. (5 points) Consider a cross-classification for a pair of categorical factors A and B. If factors A and B have three levels each, then the population distribution of a single cross classification outcome has the form

	b_1	b_2	b_3	Total
a_1	π_{11}	π_{12}	π_{13}	$\pi_{1\cdot}$
a_2	π_{21}	π_{22}	π_{23}	$\pi_{2\cdot}$
a_3	π_{31}	π_{I2}	π_{33}	$\pi_{3\cdot}$
Total	$\pi_{\cdot 1}$	$\pi_{\cdot 2}$	$\pi_{\cdot 3}$	1

Here

$$\pi_{ij} = \mathrm{P}(A = a_i, B = b_j)$$

are the joint the probabilities, and

$$\pi_{i\cdot} = \mathrm{P}(A = a_i), \qquad \pi_{\cdot j} = \mathrm{P}(B = b_j)$$

are the marginal probabilities.

(a) Clearly state the hypothesis of independence between factors A and B in terms of the above given parametric model.

(b) Present a realistic example of such two categorical factors. Try to make your example different from those mentioned in the lecture notes.

(c) Aiming at the chi-square test of independence, the following data was collected

	b_1	b_2	b_3
a_1	6	3	9
a_2	7	2	8
a_3	13	6	15

Explain how these nine counts were produced. In particular, how many independent samples were collected?

(d) Apply the chi-square test of independence to the data in (c). What can be said about the p-value of the test?

6 (5 marks) An experiment was conducted to study the effects of two types of promotional expenditures on sales of a certain product sold in supermarkets. Sixteen localities were selected for the test. Different combinations of media advertising expenditures (x_1) and point-of-sale expenditures (x_2) were specified for the study, and the localities were assigned at random to each of these combinations (x_{1i}, x_{2i}). Then the dollar sales (y_i) were recorded for $i = 1, \ldots, 16$.

(a) The sample correlation coefficient computed from the 16 pairs (x_{1i}, x_{2i}) is zero. Is it a good or bad feature of the experimental design? Explain.

(b) The following table presents a part of the computer output for multiple regression

Variable	Ref. coeff.	Std. dev.	T stat.
Constant	2.13438	.61036	3.50
x1	3.02925	.12028	25.18
x2	.70575	.12028	5.87

Perform three utility tests. Clearly state your conclusions.

(c) The following table presents another part of the computer output for multiple regression

Source	Sum of squares
Regression	193.4888
Residual	3.7616
Total	197.2503

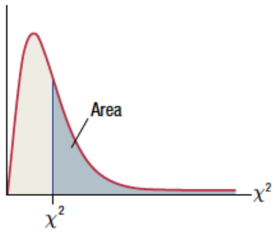
Compute the adjusted coefficient of determination. What does it say about the underlying model?

(d) For a supermarket with media advertising expenditures being $x_1 = 2$ and point-of-sale expenditures being $x_2 = 4$, what would be your prediction for the dollar sales of the product under the study?

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
18	.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
	.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
19	.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
	.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
20	.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
	.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
21	.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
	.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
22	.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
	.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
23	.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
	.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
24	.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
	.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
25	.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
	.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
26	.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
	.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
27	.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
	.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. An advantage of this stratified sampling compared to a single sample is in the possibility of comparing qualification levels of the teachers from different school categories. A possible disadvantage is a larger standard error of the estimate of the overall mean.

1b. For the stratified sample mean with $w_1 = 0.5, w_2 = 0.4, w_3 = 0.1$, and $n_1 = n_2 = n_3 = 100$, we find the stratified sample mean

$$\bar{x}_{\mathrm{s}} = 0.5 \cdot 14.8 + 0.4 \cdot 17.3 + 0.1 \cdot 19.3 = 16.25,$$

and its standard error

$$s_{\bar{x}_{\mathrm{s}}} = \sqrt{\tfrac{(0.5)^2 6.4}{100} + \tfrac{(0.4)^2 2.7}{100} + \tfrac{(0.1)^2 3.6}{100}} = 0.144.$$

The 99% confidence interval for the overall mean becomes

$$I_\mu = 16.25 \pm 2.58 \cdot 0.144 = 16.25 \pm 0.37 = (15.88, 16.62).$$

1c. Using proportional allocation allocation, we put 150 observations on the Elementary level, 120 observations on the High school level, and 30 observations on the College level. This would make smaller standard error compared to the simple random sample.

2a. Two sample 95% confidence interval

$$I_\delta = 14.8 - 19.3 \pm 1.96\sqrt{\tfrac{3.6}{100} + \tfrac{6.4}{100}} = -4.50 \pm 0.62.$$

Since this interval does not cover zero, it indicates that the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ should be rejected. However, this conclusion is based on the comparison of the two extreme strata means and might be false due to missing of the multiple comparison effect.

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

$$\tfrac{14.8+17.3+19.3}{3} = 17.13,$$
$$SS_A = 100 \cdot ((14.8 - 17.13)^2 + (17.3 - 17.13)^2 + (19.3 - 17.13)^2) = 1016.7,$$
$$SS_E = 6.4 \cdot 99 + 2.7 \cdot 99 + 3.6 \cdot 99 = 1257.3.$$

Putting these into the anova table

Source of variation	SS	df	MS	F
Main factor	1016.7	2	508.3	121.0
Error	1257.3	297	4.2	
Total	2274.0	299		

and checking the F-distribution table with 2 degrees of freedom in the numerator, we see that the p-value of the F-test is much less than 0.1%. We conclude that the difference between the three means is statistically significant.

2c. A suitable parametric model would be a two-way anova model with two main factors being
factor A = level of school with 3 levels,
factor B = gender with 2 levels.

3a. Since

$$\mathrm{E}(Z^2) = \mathrm{Var}(Z) = 1,$$

we get

$$\mu = \mathrm{E}(Z_1^2 + \ldots + Z_k^2) = k.$$

3b. Kurtosis of the standard normal distribution is

$$3 = \mathrm{E}((Z)^4),$$

and therefore,

$$\mathrm{Var}(Z^2) = \mathrm{E}((Z)^4) - (\mathrm{E}(Z^2))^2 = 2.$$

3c. Due to independence,

$$\mathrm{Var}(Z_1^2 + \ldots + Z_k^2) = k\mathrm{Var}(Z^2) = 2k.$$

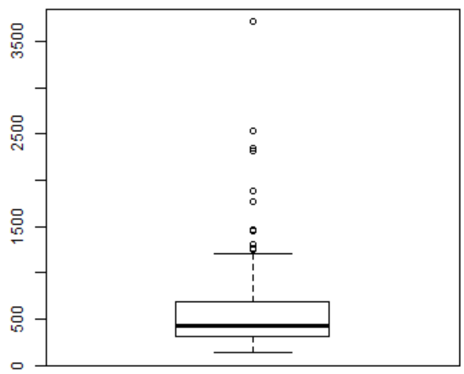
3d. Gamma distribution $\mathrm{Gam}(\alpha, \lambda)$ has the mean and variance

$$\mu = \tfrac{\alpha}{\lambda}, \quad \sigma^2 = \tfrac{\alpha}{\lambda^2}.$$

With $\alpha = \frac{k}{2}$ and $\mu = \frac{1}{2}$, we obtain the same values as above

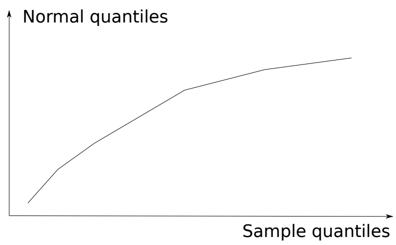
$$\mu = \tfrac{k/2}{1/2} = k, \quad \sigma^2 = \tfrac{k/2}{1/4} = 2k.$$

4a. The range is close to 3500 and the inter-quartile range is like 500.



4b. Skewed the right. The sample mean is larger than the sample median as the sample values larger than the median contribute to the arithmetic mean such that the mean will be larger than the median.

4c. For a right skewed distribution the normal probability plot would have the following typical profile reflecting the fact that the largest sample values are larger than would be predicted by the normal distribution.



4d. A logarithmic transformation

$$\text{new data} = \log(\text{original data})$$

might do the job.

5a. Hypothesis of independence

$$\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}, \quad i = 1, 2, 3, \quad j = 1, 2, 3.$$

5c. The nine numbers

	b_1	b_2	b_3
a_1	6	3	9
a_2	7	2	8
a_3	13	6	15

are the counts obtained after cross-classification of a single sample.

5d. The observed and expected counts

	b_1	b_2	b_3
a_1	6 (6.8)	3 (2.9)	9 (8.3)
a_2	7 (6.4)	2 (2.7)	8 (7.9)
a_3	13 (12.8)	6 (5.4)	15 (15.8)

Here the problems is with the expected counts being smaller than recommended 5. Combining the first two rows we get a new table with observed and expected counts

	b_1	b_2	b_3
$a_1 + a_2$	13 (13.2)	5 (5.6)	17 (16.2)
a_3	13 (12.8)	6 (5.4)	15 (15.8)

The corresponding chi-square test statistic is as small as 0.2, and according to the χ^2_2-distribution table the p-value is close to 90%. Thus we do not reject the null hypothesis of independence.

6a. A problematic case in a multiple regression setting

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon,$$

is when two explanatory variables are linearly dependent (collinearity). In our case the X_1 and X_2 are uncorrelated (orthogonality) which is a very good feature for the multiple regression design of experiment.

6b. Using t_{13} as the null distribution for all three test statistics for three utility tests we reject the following three null hypotheses

$$H_0 : \beta_0 = 0, \quad H_0 : \beta_1 = 0, \quad H_0 : \beta_2 = 0.$$

6c. The adjusted coefficient of determination

$$R^2_a = 1 - \tfrac{15\cdot 3.76}{13\cdot 197.25} = 0.98$$

gives a high score (98 out of 100) on how well the two explanatory variables explain the observed variation in the response variable.

6d. The predicted mean response is

$$2.13 + 3.03 \cdot 2 + 0.71 \cdot 4 = 11.3.$$