Tentamentsskrivning i Statistisk slutledning MVE155/MSG200, 7.5 hp.

Tid: 19 mars 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 sample $(x_1, \ldots, x_n)$ was taken from a stratified population consisting of k strata of relative sizes $(w_1, \ldots, w_k)$. The n observations were allocated among different strata as follows

$$n = n_1 + \ldots + n_k.$$

For the given allocation $(n_1, \ldots, n_k)$, consider the pooled sample mean

$$\bar{x}_{\mathrm{p}} = \frac{n_1\bar{x}_1 + \ldots + n_k\bar{x}_k}{n},$$

where $\bar{x}_j$ is the mean for the subsample taken from the stratum j. We assume that the k subsamples are mutually independent iid-samples taken from the respective strata.

(a) Show that $\bar{x}_{\mathrm{p}}$ is a biased estimate of the population mean μ, with the bias size

$$\mathrm{Bias} = \mathrm{E}(\bar{X}_{\mathrm{p}} - \mu) = \sum_{j=1}^{k} (\tfrac{n_j}{n} - w_j)\mu_j,$$

where $(\mu_1, \ldots, \mu_k)$ are the strata means.

(b) Assume that all strata have the same variance $\sigma_j^2 = \sigma^2$. Verify the following formula for the mean square error

$$\mathrm{MSE} = \mathrm{E}[(\bar{X}_{\mathrm{p}} - \mu)^2] = (\mathrm{Bias})^2 + \tfrac{\sigma^2}{n}.$$

(c) If $(n_1, \ldots, n_k)$ is a random allocation of n observation, then $\bar{x}_{\mathrm{p}}$ becomes a sample mean $\bar{x}$ for an iid-sample. Explain why despite (a), the estimate $\bar{x}$ is unbiased.

2. (5 points) An iid-sample from the normal distribution $\mathrm{N}(\mu, \sigma^2)$

124.9, 113.3, 114.5, 121.2, 123.7, 127.7, 128.2, 124.0, 124.6, 124.9,
124.9, 125.1, 125.5, 130.2, 125.9, 126.8, 128.3, 122.9, 128.5, 105.3,

produces the following summary statistics

$$\sum x_i = 2470.4, \qquad \sum x_i^2 = 305829.0.$$

(a) Find method of moments estimates $\tilde{\mu}$ and variance $\tilde{\sigma}^2$ for the population mean μ and variance σ^2.

(b) Are the point estimates $\tilde{\mu}$ and $\tilde{\sigma}^2$ unbiased? If not, compute the bias in terms of μ, σ^2, and sample size n.

(c) Explain why the method of moments produces consistent estimates for a broad set of parametric distributions.

3. (5 marks) An iid-sample $(x_1, \dots, x_n)$ was taken from a continuous population distribution with median m. Consider its ordered version $(x_{(1)}, \dots, x_{(n)})$.

(a) For a given k, show that

$$\mathrm{P}(X_{(k)} < m) = \mathrm{P}(Y \geq k),$$

where Y has distribution $\mathrm{Bin}(n, \frac{1}{2})$.

(b) Let $n = 10$. The binomial distribution table then gives

$$\mathrm{P}(Y \leq 8) = 0.989, \quad \mathrm{P}(Y \leq 9) = 0.999.$$

The interval $(x_{(2)}, x_{(9)})$ can be treated as a confidence interval for the median. Using (a), find the exact confidence level of this interval estimate for the median.

(c) In the large sample case, we get an approximate 95%confidence interval for the median of the form $(x_{(k)}, x_{(n-k+1)})$, where k is found as

$$k \approx \tfrac{n}{2} - 0.98\sqrt{n} + 0.5.$$

Explain this formula.

4. (5 points) To study the effect of cigarette smoking on platelet aggregation, Levine (1973) drew blood samples from 11 individuals before and after they smoked a cigarette and measured the extend to which the blood platelets aggregated. Platelets are involved in the formation of blod clots, and it is known that smokers suffer more often from disorders involving blood clots than do nonsmokers. The data are shown in the following table, which gives the maximum percentage of all the platelets that aggregated after being exposed to a stimulus.

Before smoking	After smoking
25	27
25	29
27	37
44	56
30	46
67	82
53	57
53	80
52	61
60	59
28	43

(a) Suppose that the difference (After smoking – Before smoking) is normally distributed with an unknown mean μ and known standard deviation $\sigma = 10$. Assuming a normal prior $\mathrm{N}(\mu_0, \sigma_0^2)$ for the mean difference μ with $\mu_0 = 5$, and $\sigma_0 = 100$, compute the posterior distribution for μ.

(b) How would you justify the choice of a large value for σ_0?

(c) How can one find quantiles of the normal distribution using the attached t-distribution table? Illustrate by finding the 0.9995-quantile of the standard normal distribution.

(d) Compute a 95% credibility interval for μ. Would you reject the null hypothesis of no difference?

5. (5 points) Three different varieties of tomato and four different plant densities (10, 20, 30, and 40 thousand plats per hectar) are being considered for planting in a particular region. To see whether either variety or plant density affects yield, each combination of variety and plant density is used in three different plots, resulting in the data on the yield averaged over three replications

	10	20	30	40	Mean
Variety 1	9.20	12.43	12.90	10.80	11.33
Variety 2	8.93	12.63	15.1	12.77	12.21
Variety 3	16.30	18.10	19.93	18.17	18.13
Mean	11.48	14.39	15.78	13.91	13.89

(a) Using the table draw a graph with three profiles and make your preliminary conclusions. What does the graph say about the interaction effect?

(b) The following values of sum of squares are given: 327.60 for tomato varieties, 8.03 for interaction, and 460.36 for the total sum of squares. Apply three relevant F-tests and present your findings.

(c) Explain how would you verify the key assumption of normality for the F-tests using all 36 yield values.

6 (5 marks) A study was conducted to determine a woman's risk of transmitting HIV to her unborn child. A sample of 114 HIV-infected women who gave birth to two children found that HIV infection occurred in 19 of the 114 older siblings and in 20 of the 114 younger siblings.

	Older sibling	Younger sibling
HIV	19	20
no HIV	95	94
Total	114	114

(a) Denote by p_1 the probability of HIV infection for older siblings, and by p_2 the probability of HIV infection younger siblings. Find an unbiased estimate of the difference $p_1 - p_2$. Why is it unbiased?

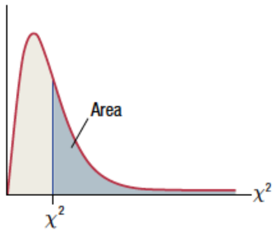
(b) The following table presents the data in a different format

	Younger sibling HIV	Younger sibling no HIV
Older sibling HIV	2	17
Older sibling no HIV	18	77

Explain in which way this table is more informative than the first one. Connect to the underlying joint and marginal distributions.

(c) Apply an appropriate test to verify whether the probability of HIV infection is the same for older siblings as it is for younger siblings.

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 F-distribution for $\alpha = 1\%$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9
1	4052	4999·5	5403	5625	5764	5859	5928	5981	6022
2	98·50	99·00	99·17	99·25	99·30	99·33	99·36	99·37	99·39
3	34·12	30·82	29·46	28·71	28·24	27·91	27·67	27·49	27·35
4	21·20	18·00	16·69	15·98	15·52	15·21	14·98	14·80	14·66
5	16·26	13·27	12·06	11·39	10·97	10·67	10·46	10·29	.0·16
6	13·75	10·92	9·78	9·15	8·75	8·47	8·26	8·10	7·98
7	12·25	9·55	8·45	7·85	7·46	7·19	6·99	6·84	6·72
8	11·26	8·65	7·59	7·01	6·63	6·37	6·18	6·03	5·91
9	10·56	8·02	6·99	6·42	6·06	5·80	5·61	5·47	5·35
10	10·04	7·56	6·55	5·99	5·64	5·39	5·20	5·06	4·94
11	9·65	7·21	6·22	5·67	5·32	5·07	4·89	4·74	4·63
12	9·33	6·93	5·95	5·41	5·06	4·82	4·64	4·50	4·39
13	9·07	6·70	5·74	5·21	4·86	4·62	4·44	4·30	4·19
14	8·86	6·51	5·56	5·04	4·69	4·46	4·28	4·14	4·03
15	8·68	6·36	5·42	4·89	4·56	4·32	4·14	4·00	3·89
16	8·53	6·23	5·29	4·77	4·44	4·20	4·03	3·89	3·78
17	8·40	6·11	5·18	4·67	4·34	4·10	3·93	3·79	3·68
18	8·29	6·01	5·09	4·58	4·25	4·01	3·84	3·71	3·60
19	8·18	5·93	5·01	4·50	4·17	3·94	3·77	3·63	3·52
20	8·10	5·85	4·94	4·43	4·10	3·87	3·70	3·56	3·46
21	8·02	5·78	4·87	4·37	4·04	3·81	3·64	3·51	3·40
22	7·95	5·72	4·82	4·31	3·99	3·76	3·59	3·45	3·35
23	7·88	5·66	4·76	4·26	3·94	3·71	3·54	3·41	3·30
24	7·82	5·61	4·72	4·22	3·90	3·67	3·50	3·36	3·26
25	7·77	5·57	4·68	4·18	3·85	3·63	3·46	3·32	3·22
26	7·72	5·53	4·64	4·14	3·82	3·59	3·42	3·29	3·18
27	7·68	5·49	4·60	4·11	3·78	3·56	3·39	3·26	3·15
28	7·64	5·45	4·57	4·07	3·75	3·53	3·36	3·23	3·12
29	7·60	5·42	4·54	4·04	3·73	3·50	3·33	3·20	3·09
30	7·56	5·39	4·51	4·02	3·70	3·47	3·30	3·17	3·07
40	7·31	5·18	4·31	3·83	3·51	3·29	3·12	2·99	2·89
60	7·08	4·98	4·13	3·65	3·34	3·12	2·95	2·82	2·72
120	6·85	4·79	3·95	3·48	3·17	2·96	2·79	2·66	2·56
∞	6·63	4·61	3·78	3·32	3·02	2·80	2·64	2·51	2·41

NUMERICAL ANSWERS

1a. The statement follows from

$$\mu = w_1\mu_1 + \ldots + w_k\mu_k$$

and

$$\mathrm{E}(\bar{X}_\mathrm{p}) = \sum_{j=1}^{k} \tfrac{n_j}{n}\mathrm{E}(\bar{X}_j) = \sum_{j=1}^{k} \tfrac{n_j}{n}\mu_j.$$

1b. In view of

$$\mathrm{E}(Y^2) = \mathrm{Var}(Y) + (\mathrm{E}Y)^2,$$

we get

$$\mathrm{E}[(\bar{X}_\mathrm{p} - \mu)^2] = \mathrm{Var}(\bar{X}_\mathrm{p} - \mu) + (\mathrm{E}(\bar{X}_\mathrm{p} - \mu))^2.$$

Given $\sigma_j^2 = \sigma^2$, we have

$$\mathrm{Var}(\bar{X}_\mathrm{p} - \mu) = \mathrm{Var}(\bar{X}_\mathrm{p}) = \frac{n_1^2\mathrm{Var}(\bar{X}_1) + \ldots + n_k^2\mathrm{Var}(\bar{X}_k)}{n^2} = \frac{n_1\sigma^2 + \ldots + n_k\sigma^2}{n^2} = \tfrac{\sigma^2}{n},$$

so that

$$\mathrm{MSE} = \mathrm{E}[(\bar{X}_\mathrm{p} - \mu)^2] = (\mathrm{Bias})^2 + \tfrac{\sigma^2}{n}.$$

1c. With a random allocation, the sample sizes n_j are random with $\mathrm{E}(n_j) = nw_j$. The bias formula in 1 (a) is conditional on the allocation $(n_1, \ldots, n_k)$. The averaging over possible random outcomes $(n_1, \ldots, n_k)$ removes the bias.

2a. We have two sample moments

$$\bar{x} = \tfrac{2470.4}{20} = 123.52, \qquad \overline{x^2} = \tfrac{305829}{20} = 15291.45.$$

Method of moments estimates for the mean μ is

$$\tilde{\mu} = \bar{x} = 123.52,$$

and for the variance $\sigma^2 = \mathrm{E}(X^2) - \mu^2$ is

$$\tilde{\sigma}^2 = 15291.5 - (123.5)^2 = 34.26.$$

2b. Estimate $\tilde{\mu}$ is unbiased, while $\tilde{\sigma}^2$ is biased. We have

$$\mathrm{E}(\tilde{\sigma}^2) = \mathrm{E}(\overline{X^2}) - \mathrm{E}(\bar{X}^2) = \mathrm{E}(X^2) - \mathrm{Var}(\bar{X}) - (\mathrm{E}(\bar{X}))^2 = \sigma^2 - \tfrac{\sigma^2}{n} = \tfrac{n-1}{n}\sigma^2.$$

So the bias is equal to

$$\mathrm{E}(\tilde{\sigma}^2) - \sigma^2 = -\tfrac{\sigma^2}{n}.$$

2c. By the law of large numbers,

$$\bar{x} \to \mathrm{E}(X), \qquad \overline{x^2} \to \mathrm{E}(X^2), \qquad n \to \infty,$$

and therefore, the method of moments based on a continuous function

$$\theta = g(\mathrm{E}(X), \mathrm{E}(X^2)),$$

produces a consistent estimate

$$\tilde{\theta} = g(\bar{x}, \overline{x^2}).$$

3a. An iid-sample $(x_1, \ldots, x_n)$ was taken from a continuous population distribution with median m. Consider its ordered version $(x_{(1)}, \ldots, x_{(n)})$. For a given k, the event $\{X_{(k)} < m\}$ means that at least k sample values fell below the median m. If Y stands for the number of sample values below m, then we get

$$\{X_{(k)} < m\} = \{Y \geq k\}.$$

It remains to show that Y has distribution $\text{Bin}(n, \frac{1}{2})$. This follows from the fact that each observation X_i is smaller than m with probability $\frac{1}{2}$, and Y is the number of successes in such n Bernoulli trials.

3b. Let $n = 10$. The binomial distribution table then gives

$$\text{P}(Y \leq 8) = 0.989, \quad \text{P}(Y \leq 9) = 0.999.$$

The interval $(x_{(2)}, x_{(9)})$ can be treated as a confidence interval for the median. Using 3a, we get

$$\text{P}(X_{(2)} \geq m) = \text{P}(Y < 2) = \text{P}(Y > 8) = 1 - 0.989 = 0.011,$$

and

$$\text{P}(X_{(9)} > m) = \text{P}(X_{(9)} \geq m) = \text{P}(Y < 9) = \text{P}(Y \leq 8) = 0.989.$$

Therefore,

$$\text{P}(X_{(2)} < m < X_{(9)}) = \text{P}(m < X_{(9)}) - \text{P}(X_{(2)} \geq m) = 0.989 - 0.011 = 0.978,$$

giving the exact confidence level of this interval estimate for the median to be 97.8%.

3c. We have due to the normal approximation with a continuity correction

$$0.025 \approx \text{P}(Y < k) = \text{P}(Y \leq k - 1) \approx \Phi\left(\frac{k - \frac{1}{2} - \frac{n}{2}}{\frac{\sqrt{n}}{2}}\right),$$

which leads to the equation

$$-1.96 = \frac{k - \frac{1}{2} - \frac{n}{2}}{\frac{\sqrt{n}}{2}}.$$

Thus k kan be found as

$$k \approx \tfrac{n}{2} - 0.98\sqrt{n} + 0.5.$$

4a. We assume that differences $D_i \sim \text{N}(\mu, (10)^2)$, $i = 1, \ldots, 11$, where $\mu \sim \text{N}(5, (100)^2)$. The observed differences

Before smoking	After smoking	d_i
25	27	2
25	29	4
27	37	10
44	56	12
30	46	16
67	82	15
53	57	4
53	80	27
52	61	9
60	59	-1
28	43	15

give

$$\bar{d} = \tfrac{113}{11} = 10.27.$$

Using the normal-normal conjugate prior formula we find the posterior distribution to be normal $\text{N}(\gamma_n \mu_0 + (1 - \gamma_n)\bar{d}; \gamma_n \sigma_0^2)$ with

$$\gamma_n = \frac{\sigma^2}{\sigma^2 + n\sigma_0^2} = 0.00091.$$

The normal posterior has the mean close to 10.27 and the variance 9.08.

4b. The large variance for the prior reflects our lack of prior knowledge about the mean difference. The resulting distribution density curve is flat as an informative prior.

4c. Since the t-distribution with k degrees of freedom is asymptotically normal as $k \to \infty$, we find from the t-distribution table that

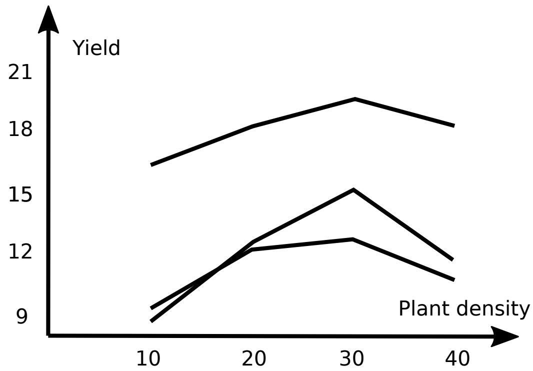
$$\Phi^{-1}(0.9995) = 3.291.$$

4d. A 95% credibility interval

$$J_\mu = 10.27 \pm 1.96\sqrt{9.08} = 10.27 \pm 5.91.$$

It is far from covering 0, so we can reject the null hypothesis of no difference.

5a. Variety 1 has higher yields over all densities. There is an indication of interaction as the lines are not parallel.



5b. The sum of squares for densities is computed as

$$3 \cdot 3 \cdot [(11.48 - 13.89)^2 + (14.39 - 13.89)^2 + (15.78 - 13.89)^2 + (13.91 - 13.89)^2] = 86.68.$$

This allows us to fill in the ANOVA table

Source of variation	SS	df	MS	F
Varieties	327.60	2	163.8	103.0
Density	86.68	3	28.9	18.2
Interaction	8.03	6	1.34	0.8
Errors	38.05	24	1.6	
Total	460.36	35		

Using the table for critical values for the F-distribution we find for $\alpha = 0.01$

$$F_{2,24} = 5.61, \quad F_{3,24} = 4.72, \quad F_{6,24} = 3.67.$$

Conclusions: 1) we reject the null hypothesis of no variety effect, 2) we reject the null hypothesis of no density effect, 1) we do not reject the null hypothesis of no interaction.

5c. We compute 36 residuals $y_{ijk} - \bar{y}_{ij\cdot}$ and check the normality assumption using a normal probability plot.

6a. We have a paired sample of size $n = 114$. The difference between two population proportions $p_1 - p_2$ is estimated by the difference of two sample proportions

$$\hat{p}_1 - \hat{p}_2 = \tfrac{19}{114} - \tfrac{20}{114} = -\tfrac{1}{114} = -0.0088.$$

This is an unbiased estimate because despite the dependence between $\hat{p}_1$ and $\hat{p}_2$ we have

$$\mathrm{E}(\hat{P}_1 - \hat{P}_2) = \mathrm{E}(\hat{P}_1) - \mathrm{E}(\hat{P}_2) = p_1 - p_2.$$

6b. The new table contains more information

	Younger sibling HIV	Younger sibling no HIV	Total
Older sibling HIV	2	17	19
Older sibling no HIV	18	77	95
Total	20	94	114

so that the previous table is recovered by computing the totals. The relationship between them is similar as the relationship between the underlying joint and marginal distributions.

6c. We apply the McNemar test for the matched pair design. The observed test statistic is $\frac{(18-17)^2}{18+17} = 0.03$. Using the table for χ^2_1-distribution we see that the p-value is much large than 10%. We do not reject the null hypothesis of no difference between the probabilities of HIV infection for older and younger siblings.