Exam of *Statistics for Data Science*

It is forbidden to consult any material during the test, with the exception of a scientific calculator. Duration of written exam is 1.5h.

Exercise 1 (6 points). Let $X, Y \sim Ber(0.5)$ be i.i.d. random variables. Define the random variables:

$$U = X + Y \qquad\qquad V = |X - Y|$$

a Determine the joint and marginal distributions of U and V.

b Find out whether U and V are dependent or independent.

c Determine the covariance $Cov(U, V)$ and the correlation coefficient $Cor(U, V)$.

Solution. (a) see solution of Ex. 9.6 (a) at page 437 of [1].
(b) Dependent, because e.g.,

$$P(U = 0, V = 0) = \frac{1}{4} \neq P(U = 0)P(V = 0) = \frac{1}{4}\frac{1}{2} = \frac{1}{8}$$

(c) We have

$$E[U] = 0\frac{1}{4} + 1\frac{1}{2} + 2\frac{1}{4} = 1 \qquad\qquad E[V] = 0\frac{1}{2} + 1\frac{1}{2} = \frac{1}{2}$$

$$E[UV] = \sum_a a \cdot P(UV = a) = 0P(UV = 0) + 1P(U = 1, V = 1) + 2P(U = 1, V = 2) = \frac{1}{2}.$$

Therefore

$$Cov(U, V) = E[UV] - E[U]E[V] = \frac{1}{2} - \frac{1}{2} = 0$$

and then $Cor(U, V) = 0$ as well.

Exercise 2 (6 points). Suppose that $x_1, \ldots, x_n$ is a dataset, which is a realization of a random sample from a Rayleigh distribution, which is a continuous distribution with probability density function:

$$f_\theta = \frac{x}{\theta^2} e^{-\frac{x^2}{\theta^2}} \quad \text{for } x \geq 0.$$

In this case what is the maximum likelihood estimate of θ?

Solution. The likelihood is

$$L(\theta) = \prod_{i=1}^{n} \frac{x_i}{\theta^2} e^{-\frac{x_i^2}{\theta^2}}$$

and then the log-likelihood is

$$\ell(\theta) = \sum_{i=1}^{n} (log x_i - 2 log \theta - \frac{1}{2\theta^2} x_i^2)$$

The (log-)likelihood has maximum when the derivative (w.r.t. θ) is zero, i.e., when

$$\frac{d\ell(\theta)}{d\theta} = -\frac{2n}{\theta} + \frac{1}{\theta^3} \sum_{i=1}^{n} x_i^2 = 0$$

which occurs for

$$\theta = \sqrt{\frac{1}{2n} \sum_{i=1}^{n} x_i^2}$$

Exercise 3 (6 points). Consider a linear regression model *without intercept*:

$$Y_i = \beta x_i + U_i \quad \text{for } i = 1, \dots, n$$

where $U_1, \dots, U_n$ are independent random variables with $E[U_i] = 0$ and $Var(U_i) = 2$. Consider the following three estimators for the parameter β:

$\hat{\beta}_1 = \frac{1}{n}\sum_{i=1}^n Y_i/x_i$

$\hat{\beta}_2 = (\sum_{i=1}^n Y_i)/(\sum_{i=1}^n x_i)$

$\hat{\beta}_3 = (\sum_{i=1}^n x_i Y_i)/(\sum_{i=1}^n x_i^2)$

Show that all three estimators are unbiased for β. Compute their variance and discuss their efficiency.

Solution. We observe $E[Y_i] = \beta x_i + E[U_i] = \beta x_i$ and $Var(Y_i) = Var(U_i) = 2$. We calculate:

- $E[\hat{\beta}_1] = \frac{1}{n}\sum_{i=1}^n E[Y_i]/x_i = \frac{1}{n}\sum_{i=1}^n \beta x_i/x_i = \beta$
- $E[\hat{\beta}_2] = (\sum_{i=1}^n E[Y_i])/(\sum_{i=1}^n x_i) = (\sum_{i=1}^n \beta x_i)/(\sum_{i=1}^n x_i) = \beta$
- $E[\hat{\beta}_3] = (\sum_{i=1}^n x_i E[Y_i])/(\sum_{i=1}^n x_i^2) = (\sum_{i=1}^n \beta x_i^2)/(\sum_{i=1}^n x_i^2) = \beta$

and:

- $Var(\hat{\beta}_1) = \frac{1}{n^2}\sum_{i=1}^n Var(Y_i)/x_i^2 = \frac{2}{n}\frac{1}{n}\sum_{i=1}^n \frac{1}{x_i^2}$
- $Var(\hat{\beta}_2) = (\sum_{i=1}^n Var(Y_i))/(\sum_{i=1}^n x_i)^2 = 2n/(\sum_{i=1}^n x_i)^2 = \frac{2}{n}\frac{1}{(\frac{1}{n}\sum_{i=1}^n x_i)^2}$
- $Var(\hat{\beta}_3) = (\sum_{i=1}^n x_i^2 Var(Y_i))/(\sum_{i=1}^n x_i^2)^2 = \frac{2}{\sum_{i=1}^n x_i^2} = \frac{2}{n}\frac{1}{\frac{1}{n}\sum_{i=1}^n x_i^2}$

Since $1/x^2$ is a convex function, by the Jensen's inequality

$$\frac{1}{n}\sum_{i=1}^n \frac{1}{x_i^2} \geq \frac{1}{(\frac{1}{n}\sum_{i=1}^n x_i)^2}$$

and then $Var(\hat{\beta}_1) \geq Var(\hat{\beta}_2)$. Since x^2 is also convex, by the Jensen's inequality:

$$\frac{1}{n}\sum_{i=1}^n x_i^2 \geq (\frac{1}{n}\sum_{i=1}^n x_i)^2$$

and then $Var(\hat{\beta}_2) \geq Var(\hat{\beta}_3)$.

Exercise 4 (6 points). Environmentalists have taken 16 samples from the wastewater of a chemical plant and measured the concentration of a certain carcinogenic substance. They found $\bar{x}_{16} = 2.24$ (ppm) and $s_{16}^2 = 1.12$, and want to use these data in a lawsuit against the plant. It may be assumed that the data are a realization of a normal random sample.

a Construct the 97.5% one-sided confidence interval that the environmentalists made to convince the judge that the concentration exceeds legal limits.

b The plant management uses the same data to construct a 97.5% one-sided confidence interval to show that concentrations are not too high. Construct this interval as well.

Hint. $t_{15,0.025} = 2.131$

Solution. See full solution of Ex. 24.6 (a,b) at page 470 of [1].

Exercise 5 (6 points). Write an R function to compute p-value in 2-sided t-test without using the pre-defined built-in `t.test()`.

Solution.

```
pvalue = function(data, m, n)  # data = vector of n values, m = actual mean
{
  xbar <- mean(data) # sample mean
  sbar <- sd(data) # sample variance
  t0 <- sqrt(n)*(xbar-m)/sbar # studentized mean
  v = pt(t0, n-1) # P(t <= t0)
  p <- min(v, 1-v) # lower tail
  return (2*p) # 2-sided
}
```

References

[1] F.M. Dekking, C. Kraaikamp, H.P. Lopuhaä, and L.E. Meester. *A Modern Introduction to Probability and Statistics*. Springer, 2005.

Table B.1. Right tail probabilities $1 - \Phi(a) = \mathrm{P}(Z \geq a)$ for an $N(0,1)$ distributed random variable Z.

a	0	1	2	3	4	5	6	7	8	9
0.0	5000	4960	4920	4880	4840	4801	4761	4721	4681	4641
0.1	4602	4562	4522	4483	4443	4404	4364	4325	4286	4247
0.2	4207	4168	4129	4090	4052	4013	3974	3936	3897	3859
0.3	3821	3783	3745	3707	3669	3632	3594	3557	3520	3483
0.4	3446	3409	3372	3336	3300	3264	3228	3192	3156	3121
0.5	3085	3050	3015	2981	2946	2912	2877	2843	2810	2776
0.6	2743	2709	2676	2643	2611	2578	2546	2514	2483	2451
0.7	2420	2389	2358	2327	2296	2266	2236	2206	2177	2148
0.8	2119	2090	2061	2033	2005	1977	1949	1922	1894	1867
0.9	1841	1814	1788	1762	1736	1711	1685	1660	1635	1611
1.0	1587	1562	1539	1515	1492	1469	1446	1423	1401	1379
1.1	1357	1335	1314	1292	1271	1251	1230	1210	1190	1170
1.2	1151	1131	1112	1093	1075	1056	1038	1020	1003	0985
1.3	0968	0951	0934	0918	0901	0885	0869	0853	0838	0823
1.4	0808	0793	0778	0764	0749	0735	0721	0708	0694	0681
1.5	0668	0655	0643	0630	0618	0606	0594	0582	0571	0559
1.6	0548	0537	0526	0516	0505	0495	0485	0475	0465	0455
1.7	0446	0436	0427	0418	0409	0401	0392	0384	0375	0367
1.8	0359	0351	0344	0336	0329	0322	0314	0307	0301	0294
1.9	0287	0281	0274	0268	0262	0256	0250	0244	0239	0233
2.0	0228	0222	0217	0212	0207	0202	0197	0192	0188	0183
2.1	0179	0174	0170	0166	0162	0158	0154	0150	0146	0143
2.2	0139	0136	0132	0129	0125	0122	0119	0116	0113	0110
2.3	0107	0104	0102	0099	0096	0094	0091	0089	0087	0084
2.4	0082	0080	0078	0075	0073	0071	0069	0068	0066	0064
2.5	0062	0060	0059	0057	0055	0054	0052	0051	0049	0048
2.6	0047	0045	0044	0043	0041	0040	0039	0038	0037	0036
2.7	0035	0034	0033	0032	0031	0030	0029	0028	0027	0026
2.8	0026	0025	0024	0023	0023	0022	0021	0021	0020	0019
2.9	0019	0018	0018	0017	0016	0016	0015	0015	0014	0014
3.0	0013	0013	0013	0012	0012	0011	0011	0011	0010	0010
3.1	0010	0009	0009	0009	0008	0008	0008	0008	0007	0007
3.2	0007	0007	0006	0006	0006	0006	0006	0005	0005	0005
3.3	0005	0005	0005	0004	0004	0004	0004	0004	0004	0003
3.4	0003	0003	0003	0003	0003	0003	0003	0003	0003	0002

Table B.2. Right critical values $t_{m,p}$ of the t-distribution with m degrees of freedom corresponding to right tail probability p: $\mathrm{P}(T_m \geq t_{m,p}) = p$. The last row in the table contains right critical values of the $N(0,1)$ distribution: $t_{\infty,p} = z_p$.

	Right tail probability p							
m	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
∞	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291