Exam 1 - Statistics X_400004 26.10.18 Solutions

Exercise 1 [20 points]

Let $X_1, \ldots, X_n$ be an i.i.d. sample from the shifted exponential distribution with p.d.f.

$$p_\theta(x) = \begin{cases} e^{\theta - x}, & x \geq \theta, \\ 0, & x < \theta, \end{cases}$$

where $\theta \in \mathbb{R}$ is an unknown parameter.

Hint 1: To answer the questions below, you can use without derivation the following facts: $X_i = Y_i + \theta$ for all i, where the Y_i's $\overset{\text{i.i.d.}}{\sim}$ Exponential(1); in particular, $X_{(1)} = Y_{(1)} + \theta$; and $Y_{(1)} \sim$ Exponential(n). See Appendix 1.

(a) [6 points] Give the method-of-moments estimator $\widehat{\theta}_{\text{MM}}$ and the maximum-likelihood estimator $\widehat{\theta}_{\text{ML}}$ of the parameter θ.

By *Hint 1* and Appendix 1,

$$\mathbb{E}X_1 = \mathbb{E}Y_1 + \theta = \mathbb{E}\text{Exponential}(1) + \theta = 1 + \theta.$$

We have $\theta = \mathbb{E}X_1 - 1$, and the moment estimator is given by

$$\widehat{\theta}_{\text{MM}} = \overline{X} - 1.$$

Now we find the maximum-likelihood estimator for θ. The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} p_\theta(X_i) = \prod_{i=1}^{n} e^{\theta - X_i} = e^{n\theta - \sum_{i=1}^{n} X_i},$$

which is increasing in θ. Since all $X_i \geq \theta$, the feasible values of θ are

$$\theta \leq X_{(1)}.$$

Since the likelihood $L(\theta)$ is increasing, it is maximized by the largest feasible θ, and hence

$$\widehat{\theta}_{\text{ML}} := \text{argmax}_{\theta \leq X_{(1)}} L(\theta) = X_{(1)}.$$

(b) [5 points] Transform the estimators $\widehat{\theta}_{\text{MM}}$ and $\widehat{\theta}_{\text{ML}}$ that you found in part *(a)* into unbiased estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ of the parameter θ. (If $\widehat{\theta}_{\text{MM}}$ is already unbiased, just put $\widehat{\theta}_1 = \widehat{\theta}_{\text{MM}}$. If $\widehat{\theta}_{\text{ML}}$ is already unbiased, just put $\widehat{\theta}_2 = \widehat{\theta}_{\text{ML}}$.)

We have

$$\mathbb{E}\widehat{\theta}_{\text{MM}} = \mathbb{E}\overline{X} - 1 = \mathbb{E}X_1 - 1 \overset{(a)}{=} (1 + \theta) - 1 = \theta.$$

That is, $\widehat{\theta}_{\text{MM}}$ is an unbiased estimator for θ and we put

$$\widehat{\theta}_1 = \overline{X} - 1.$$

By *Hint 1* and Appendix 1,

$$\mathbb{E}\widehat{\theta}_{\text{ML}} = \mathbb{E}X_{(1)} = \mathbb{E}Y_{(1)} + \theta = \mathbb{E}\text{Exponential}(n) + \theta = \frac{1}{n} + \theta.$$

That is, $\widehat{\theta}_{\text{ML}}$ is biased. Note that

$$\mathbb{E}\left(\widehat{\theta}_{\text{ML}} - \frac{1}{n}\right) = \theta,$$

and hence,

$$\widehat{\theta}_2 := \widehat{\theta}_{\mathrm{ML}} - \frac{1}{n} = X_{(1)} - \frac{1}{n}$$

is an unbiased estimator for θ.

(c) [7 points] Which of the unbiased estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ that you found in part *(b)* is better?

Hint 2: If you have no answer to part *(b)*, compare $\widehat{\theta}_1 = \overline{X} - 1$ and $\widehat{\theta}_2 = X_{(1)} - \frac{1}{n}$ under the assumption that these are unbiased estimators for θ.

Since $\widehat{\theta}_1$ is unbiased, we have

$$\mathrm{MSE}(\widehat{\theta}_1) = \mathbb{V}\widehat{\theta}_1 = \mathbb{V}(\overline{X} - 1) = \mathbb{V}\overline{X} = \frac{\mathbb{V}X_1}{n},$$

and then, by *Hint 1* and Appendix 1,

$$\mathrm{MSE}(\widehat{\theta}_1) = \frac{\mathbb{V}(Y_1 + \theta)}{n} = \frac{\mathbb{V}Y_1}{n} = \frac{\mathbb{V}\mathrm{Exponential}(1)}{n} = \frac{1}{n}.$$

Since $\widehat{\theta}_2$ is unbiased as well,

$$\mathrm{MSE}(\widehat{\theta}_2) = \mathbb{V}\widehat{\theta}_2 = \mathbb{V}\left(X_{(1)} - \frac{1}{n}\right) = \mathbb{V}X_{(1)},$$

and then, by *Hint 1* and Appendix 1,

$$\mathrm{MSE}(\widehat{\theta}_2) = \mathbb{V}(Y_{(1)} + \theta) = \mathbb{V}Y_{(1)} = \mathbb{V}\mathrm{Exponential}(n) = \frac{1}{n^2}.$$

For $n \geq 2$, we have $\mathrm{MSE}(\widehat{\theta}_2) < \mathrm{MSE}(\widehat{\theta}_1)$, and hence $\widehat{\theta}_2$ is a better estimator for θ than $\widehat{\theta}_1$. (Note that, for $n = 1$, the estimators coincide, $\widehat{\theta}_1 = \widehat{\theta}_2 = X_1 - 1$.)

(d) [2 points] Assume now that the statistical model is parametrized differently: $X_1, \ldots, X_n$ are i.i.d. with p.d.f.

$$p_a(x) = \begin{cases} e^{\ln a - x}, & x \geq \ln a, \\ 0, & x < \ln a, \end{cases}$$

where $a > 0$ is an unknown parameter. Construct an estimator for a based on one of the estimators you constructed in part *(a)*.

The old parameter $\theta \in \mathbb{R}$ and the new parameter $a > 0$ are in the following correspondence:

$$\ln a = \theta \quad \Leftrightarrow \quad a = e^{\theta}.$$

Then

$$\widehat{a}_{\mathrm{MM}} := e^{\widehat{\theta}_{\mathrm{MM}}} = e^{\overline{X} - 1}.$$

Also, since $a = e^{\theta}$ is a 1-to-1 correspondence (a is increasing in θ),

$$\widehat{a}_{\mathrm{ML}} = e^{\widehat{\theta}_{\mathrm{ML}}} = e^{X_{(1)} - 1/n}.$$

Exercise 2 [7 points]
Let $X_1, \ldots, X_n$ be an i.i.d. sample from the distribution with p.d.f.

$$p_\theta(x) = \begin{cases} 2\theta x e^{-\theta x^2}, & x > 0, \\ 0, & x \leq 0, \end{cases}$$

where $\theta > 0$ is an unknown parameter. Give the Bayes estimator for θ if the prior belief about θ is the gamma distribution with parameters $\alpha = 3$ and $\lambda = 1$. See Appendix 1.

Hint 3: The posterior belief about θ is a common distribution. See Appendix 1.

By Appendix 1, the prior density is

$$\pi(\theta) \propto \theta^{\alpha-1} e^{-\lambda\theta}.$$

The likelihood is

$$L(\theta) = \prod_{i=1}^{n} p_\theta(X_i) = \prod_{i=1}^{n} (2\theta X_i e^{-\theta X_i^2}) = 2^n \theta^n \left(\prod_{i=1}^{n} X_i \right) e^{-\theta(\sum_{i=1}^{n} X_i^2)} \propto \theta^n e^{-\theta n \overline{X^2}}.$$

The posterior density is proportional to the prior density $\times$ the likelihood,

$$p_{\overline{\theta}|X_1,\ldots,X_n}(\theta) \propto \pi(\theta) \times L(\theta) \propto (\theta^{\alpha-1} e^{-\lambda\theta}) \times (\theta^n e^{-\theta n \overline{X^2}}) = \theta^{\overbrace{(\alpha+n)}^{=:\,\alpha_{\text{new}}}-1} e^{-\overbrace{(\lambda + n\overline{X^2})}^{=:\,\lambda_{\text{new}}}\theta}.$$

By Appendix 1, the posterior belief about θ is $\overline{\theta}|X_1, \ldots, X_n \sim \text{Gamma}(\alpha_{\text{new}}, \lambda_{\text{new}})$, and the Bayes estimator for θ is

$$\widehat{\theta}_{\text{B}} = \mathbb{E}[\overline{\theta}|X_1, \ldots, X_n] = \mathbb{E}\text{Gamma}(\alpha_{\text{new}}, \lambda_{\text{new}}) = \frac{\alpha_{\text{new}}}{\lambda_{\text{new}}} = \frac{\alpha + n}{\lambda + n\overline{X^2}} = \frac{3+n}{1+n\overline{X^2}}.$$

Exercise 3 [18 points]
A travel agency suspects that not all Amsterdammers have heard about the West Frisian Islands (Waddeneilenden). If more than 20% of the Amsterdam population does not know about the islands, the agency considers it worthwhile to invest into an advertising campaign to attract more visitors from Amsterdam. The agency surveyed 100 random Amsterdammers and it turned out 30 of them have not heard of the West Frisian Islands.

Use a suitable (approximate) test to help the agency decide whether to proceed with advertising. Allow for at most a 2.5%-chance of type-1 error. See Appendix 1 and 2.

(a) [2 points] Specify the statistical model and the null and alternative hypotheses.

Let p be the proportion of the Amsterdam population that does not know about the islands, and let $X_i = 0(1)$ if person i in the samlpe does (not) know about the islands. The statistical model is

$$X_1, \ldots, X_{100} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p),$$

in particular

$$X := \sum_{i=1}^{n} X_i \sim \text{Binomial(100,p)}.$$

The agency should proceed with advertising if there is evidence of $p > 0.2$. In testing, we look for evidence against H_0 and in favor of H_1. Hence, we will test

$$H_0\colon p \leq 0.2 \quad \text{VS} \quad H_1 : p > 0.2$$

(b) [5 points] Determine (approximately) the p-value.

We observe $X = \sum_{i=1}^{100} X_i = 30$. As or more unusual under $H_0\colon p \geq 0.2$ would be to observe $X \geq 30$. Hence,

$$p\text{-value} = \max_{p \leq 0.2} \mathbb{P}\left(X \geq 30\right) = \max_{p \leq 0.2} \mathbb{P}\left(\text{Binomial}(100, p) \geq 30\right) = \mathbb{P}\left(\text{Binomial}(100, 0.2) \geq 30\right).$$

Since $100 * 0.2 * 0.8 = 16 \geq 5$, the normal approximation

$$\text{Binomial}(100, 0.2) \approx N(100 * 0.2, 100 * 0.2 * 0.8) = N(20, 4^2)$$

is accurate. Applied with the continuity correction, this normal approximation gives

$$\begin{aligned} p\text{-value} &= \mathbb{P}\left(\text{Binomial}(100, 0.2) \geq 30\right) \approx \mathbb{P}(N(20, 4^2) \geq 29.5) \\ &= \mathbb{P}\left(\frac{N(20, 4^2) - 20}{4} \geq \frac{29.5 - 20}{4}\right) = \mathbb{P}(N(0,1) \geq 2.375) \approx \mathbb{P}(N(0,1) \geq 2.38) \\ &= 1 - \Phi(2.38) \overset{\text{Appendix 2}}{=} 0.0087. \end{aligned}$$

(c) [4 points] Determine (approximately) the critical region.

We will use a critical region of the form (i.e. reject H_0 if)

$$X = \sum_{i=1}^{n} X_i \geq c,$$

where c is integer. We want the test size (maximum chance of type-1 error)

$$\begin{aligned} \alpha &:= \max_{p \leq 0.2} \mathbb{P}(X \geq c) = \max_{p \leq 0.2} \mathbb{P}\left(\text{Binomial}(100, p) \geq c\right) \\ &= \mathbb{P}\left(\text{Binomial}(100, 0.2) \geq c\right) \leq 0.025. \end{aligned} \tag{1}$$

Applying the normal approximation from *(a)* again (with the continuity correction), we get

$$\alpha \approx \mathbb{P}\left(N(0,1) \geq \frac{c - 0.5 - 20}{4}\right).$$

Hence, $\alpha \approx 0.025$ if

$$\frac{c - 0.5 - 20}{4} \approx z_{0.025} \overset{\text{Appendix 2}}{=} 1.96 \quad \Leftrightarrow \quad c \approx 1.96 * 4 + 0.5 + 20 = 28.34.$$

Finally,

$$c = \lceil 28.34 \rceil = 29,$$

where we round up because in (1)

$$\text{Binomial}(100, 0.2) \geq 28.34 \quad \Leftrightarrow \quad \text{Binomial}(100, 0.2) = 29, 30, \ldots, 100.$$

(d) [2 points] Should the null hypotheses be rejected? Should the agency proceed with advertising?

Testing with the p-value from *(a)*: p-value $= 0.0087$ and the significance level $\alpha_0 = 0.025$. Since p-value $\leq \alpha_0$, we reject H_0.

Equivalently, testing with the critical region from *(b)*: we reject H_0 if $X \geq 29$. We observe $X = 30 \geq 29$, hence we reject H_0.

There is evidence of $p \geq 0.2$, the agency should proceed with advertising.

(e) [5 points] Determine (approximately) the power of the test if in reality 30% of Amsterdammers do not know about the islands.

Hint 4: If you have no answer to part *(c)*, use 29 as the border of the critical region.

Using the critical region we found in *(c)*, the power at $p = 0.3$ is

$$\beta(0.3) := \mathbb{P}_{p=0.3}(\text{reject } H_0) = \mathbb{P}_{p=0.3}(X \geq 29) = \mathbb{P}(\text{Binomial}(100, 0.3) \geq 29)$$

Since $100 * 0.3 * 0.7 = 21 \geq 5$, we apply the normal approximation

$$\text{Binomial}(100, 0.3) \approx N(100 * 0.3, 100 * 0.3 * 0.7) = N(30, 21)$$

with the continuity correction and get

$$\beta(0.3) \approx \mathbb{P}(N(30,21) \geq 28.5) = \mathbb{P}\left(\frac{N(30,21) - 30}{\sqrt{21}} \geq \frac{28.5 - 30}{\sqrt{21}}\right) \approx \mathbb{P}(N(0,1) \geq -0.33)$$
$$= \Phi(0.33) \overset{\text{Appendix 2}}{=} 0.6293.$$

Appendix 1: Some relevant distributions

Exponential(λ) distribution, $\lambda > 0$

The p.d.f. is given by $\begin{cases} \lambda e^{-\lambda x}, & x > 0, \\ 0, & x \leq 0. \end{cases}$

The expectation is $\frac{1}{\lambda}$, the variance is $\frac{1}{\lambda^2}$.

Gamma(α,λ) distribution, $\alpha > 0$, $\lambda > 0$

The p.d.f. is proportional to $\begin{cases} \theta^{\alpha-1} e^{-\lambda\theta}, & \theta > 0, \\ 0, & \theta \leq 0, \end{cases}$.

The expectation is $\frac{\alpha}{\lambda}$, the variance is $\frac{\alpha}{\lambda^2}$.

Binomial(n,p) distribution, $n = 1, 2, \ldots$, $p \in [0, 1]$

The p.m.f. is given by $\binom{n}{k} p^n (1-p)^{n-k}$, $k = 0, 1, \ldots, n$.

The expectation is np, the variance is $np(1-p)$.

Appendix 2: Table Standard Normal Distribution

	0	1	2	3	4	5	6	7	8	9
0.0	0.5	0.504	0.508	0.512	0.516	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	0.591	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	0.648	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.67	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.695	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.719	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	0.758	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.791	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	0.834	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	0.877	0.879	0.881	0.883
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.898	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	0.937	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	0.975	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	0.983	0.9834	0.9838	0.9842	0.9846	0.985	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.989
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	0.992	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.994	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	0.996	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.997	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	0.998	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	0.999	0.999
3.1	0.999	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
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1	1	1	1	1	1	1	1	1	1

Table 1: Distribution function of the standard normal distribution on the interval $[0, 3.9]$. The value in the table is $\Phi(z)$ for $z = a + b/100$ where a indicates the row and b indicates the column.