Probability and Statistics

Exercise sheet 11

MC 11.1. Let $n \in \mathbb{N}$ and let $X_1, \dots, X_n$ be i.i.d. Bernoulli(p) random variables for some $p \in (0, 1)$. (Exactly one answer is correct in each question.)

(i) What is the maximum likelihood estimator of p?

(a) $T_{ML} = \overline{X}_n$.
(b) $T_{ML} = 1/\overline{X}_n$.
(c) $T_{ML} = \max\{X_i : i \in \{1, \dots, n\}\}$.
(d) $T_{ML} = X_n$.

(ii) Is the maximum likelihood estimator unbiased?

(a) Yes.
(b) No.

(iii) What is the distribution of the random variable $n \times T_{ML}$?

(a) $n \times T_{ML} \sim \mathrm{Binom}(n, p)$.
(b) $n \times T_{ML} \sim \mathrm{Geom}(p)$.
(c) $n \times T_{ML} \sim \mathrm{Binom}(n, p/n)$.
(d) $n \times T_{ML} \sim \mathrm{Bernoulli}(np)$.

MC 11.2. Let $X_1, \dots, X_n$, for $n \in \mathbb{N}$ be i.i.d. random variables with $X_i \sim \mathcal{U}([-\theta, \theta])$ for some unknown $\theta \in (0, \infty)$. What is the maximum likelihood estimator for θ? (Exactly one answer is correct.)

(a) $T_{ML} = \frac{2}{n} \sum_{i=1}^n X_i \mathbf{1}_{\{X_i > 0\}}$.
(b) $T_{ML} = \frac{1}{n} \sum_{i=1}^n 2X_i$.
(c) $T_{ML} = \max\{|X_i| : i \in \{1, \dots, n\}\}$.
(d) The maximum likelihood estimator does not exist.

MC 11.3. Let X be a random variable with well-defined positive variance. Find $a \in \mathbb{R}$ and $b > 0$ such that for $Z \coloneqq (X - a)/b$, it holds that $\mathbb{E}[Z] = 0$ and $\mathrm{Var}[Z] = 1$. (Exactly one answer is correct.)

(a) $a = \mathbb{E}[X]$, $b = \mathrm{Var}[X]$.
(b) $a = \mathbb{E}[X]$, $b = \sqrt{\mathrm{Var}[X]}$.
(c) $a = \mathbb{E}[X^2]$, $b = \mathrm{Var}[X]$.
(d) $a = (\mathbb{E}[X])^2$, $b = \mathrm{Var}[X]$.

Exercise 11.4. Let $X_1, \dots, X_n$ be independent random variables with $X_i \sim \mathcal{N}(\theta\alpha_i, 1)$, where $\alpha_i \neq 0$ are known parameters, but $\theta \in \mathbb{R}$ is unknown. Find the maximum likelihood estimator for θ.

Exercise 11.5. In a sawmill, the sawn timber of a certain sorting class undergoes quality control. Each production day, a sample of ten boards is taken, and each board is tested for its stiffness. Based on experience, it can be assumed that the stiffness of a board is normally distributed with a known standard deviation $\sigma = 1430$ MPa[1].

(a) Derive the formula for the 95% confidence interval for μ after 15 production days. You can assume that the samples are mutually independent.

(b) Compute the realized confidence interval from (a) for a sample mean of $\overline{x} = 11{,}000$ MPa (after 15 production days).

(c) How many samples would be required to ensure that the width of the confidence interval is less than 200 MPa?

Exercise 11.6. Let $\Theta = [1/2, 1]$. We consider the model family $(\mathbb{P}_\theta)_{\theta\in\Theta}$, where $X_1, \dots, X_n$ are independent and identically distributed under $\mathbb{P}_\theta$ with $X_1 \sim \text{Geom}(\theta)$. The maximum likelihood estimator for θ is given by (check!)

$$T_{ML} = \frac{n}{\sum_{i=1}^n X_i}.$$

Find an approximate 95% confidence interval for θ.

Hint: For all $\theta \in [1/2, 1]$, we have $\frac{\sqrt{1-\theta}}{\theta} \leq \sqrt{2}$. You can use this to simplify the expression. Note that enlarging the interval only increases the probability of coverage. However, for θ close to 1, the interval might be unnecessarily wide.

Hint 2: Recall that we know that

$$\mathbb{E}_\theta[X_1] = \frac{1}{\theta} \quad \text{and} \quad \text{Var}_\theta[X_1] = \frac{1-\theta}{\theta^2}.$$

Remark: In practice, if the variance is unknown, we typically replace it with an estimate. Either we use the generic estimator

$$\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2,$$

or, since in this case $\text{Var}[X_1] = \frac{1-\theta}{\theta^2}$ and we have an estimator for θ, we can also use

$$\tilde{\sigma}_n^2 = \frac{1-T_{ML}}{T_{ML}^2}.$$

As both estimators are consistent in this case, we still have by Slutsky's theorem that

$$\frac{\sum_{i=1}^n X_i - \frac{n}{\theta}}{\sqrt{n\hat{\sigma}_n^2}} \approx \mathcal{N}(0,1) \quad \text{and} \quad \frac{\sum_{i=1}^n X_i - \frac{n}{\theta}}{\sqrt{n\tilde{\sigma}_n^2}} \approx \mathcal{N}(0,1).$$

Exercise 11.7. To estimate the number N of trout in a lake, the following procedure is used (capture-recapture method): In a first step, 500 trout are caught, marked, and released back into the lake. In a second step, 200 trout are caught again, and the number X of marked trout is recorded.

[1]The pascal is a derived SI unit of pressure and mechanical stress. It is named after Blaise Pascal and defined as follows: 1 Pa = 1 kg×m^{-1}×s^{-2} = 1 N×m^{-2}. One pascal is the pressure exerted by a force of one newton on an area of one square meter. 1 MPa = 10^6 Pa = 10 bar.

(a) X is modeled as a binomial random variable, $X \sim \text{Binom}(n, \theta)$, where θ denotes the probability that a fish caught in the second step is marked. What is the value of n? What is the value of the parameter θ if the total number of trout in the lake is $N = 2000$ or $N = 5000$?

(b) The actual observed value for X is 40. Give a reasonable estimate for the parameter θ, and derive from it an estimate for the total number N of trout in the lake.

(c) Determine an approximate 95% confidence interval for θ, and from that, an approximate 95% confidence interval for N.

Quantile table for the standard normal distribution

0.5	0.75	0.9	0.95	0.975	0.99	0.995	0.999
0	0.6745	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902

For instance, $\Phi^{-1}(0.9) = 1.2816$, where Φ is the distribution function of $\mathcal{N}(0,1)$.

Table of standard normal distribution

	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

For instance, $\mathbb{P}[Z \leq 1.96] = 0.975$.