1. (a) Name and describe three types of data collection methods.

   (b) i. What is the Delta method?

   ii. We consider an i.i.d. sample $X_1, \cdots, X_n$ from the Bernoulli distribution with parameter θ. What is the asymptotic distribution of the sample mean $\bar{X}$?

   iii. Use the Delta method to derive the asymptotic distribution of

$$\hat{o} = \frac{\bar{X}}{1 - \bar{X}}.$$

   (c) Recall the Jensen inequality. Explain how it relates to the concept of skewness in a distribution of data.

2. This problem is based on the following table

1.15	1.12	1.19	1.50	1.26	1.64	1.08	1.09	1.17	1.79
1.54	1.26	1.41	1.08	1.75	1.28	2.84	1.87	1.21	1.03
1.21	1.04	1.16	1.82	1.34	1.05	1.50	1.34	1.48	1.39

   (a) Give a 95% confidence interval for the mean.

   (b) Check the independence of the sequence with the Wald–Wolfowitz run test.

   (c) We suspect that this dataset is i.i.d. from the Pareto distribution with index 5/2, which is defined by the density function

$$f(x) = c \cdot x^{-5/2}, \qquad x \geq 1.$$

   i. Calculate the value of the constant c.

   ii. Calculate the cumulative distribution function.

   iii. Use an adequate binning and perform a χ^2 goodness of fit test to check if we may assume that the table follows this distribution.

   iv. We also perform a Kolmogorov–Smirnov test. Explain how this test is performed. In this case, it yields a statistic value of 0.176. How should we interpret it?

3. We consider an i.i.d. sample $X_1, \cdots, X_n$ following the exponential distribution with parameter λ.

   (a) Explain why $\lambda\bar{X} - 1$ is an asymptotic pivot.

   (b) Use the Neyman–Pearson lemma to construct the most powerful test of level α of $H_0 : \lambda = \lambda_0$ against $H_1 : \lambda = \lambda_1$, where $\lambda_1 > \lambda_0$ and n is large.

4. We consider an i.i.d. sample $X_1, \cdots, X_n$ with probability mass function

$$p(-1|\theta) = \theta, \quad p(1|\theta) = 1 - \theta$$

   for a parameter $\theta \in [0, 1]$.

   (a) Calculate the expectation of a random variable following this distribution and deduce an estimator of θ.

   (b) Find the maximum likelihood estimator of θ.

   (c) Use the Cramér–Rao bound to prove that the maximum likelihood estimator is a minimum variance unbiased estimator of θ.

Solutions

1. (a) See the lecture notes.

   (b) i. Suppose that θ and σ^2 are finite and that $\hat{\theta}_n$ is such that

   $$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

   then for all C^2 function g such that $g'(\theta) \neq 0$

   $$\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta)\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2 \cdot g'(\theta)^2).$$

   ii. We recall the central limit theorem in this case

   $$\sqrt{n}(\bar{X} - \theta) \xrightarrow{d} \mathcal{N}(0, \theta(1-\theta)).$$

   iii. We use the Delta method with the function

   $$g(\theta) = \frac{\theta}{1-\theta}, \quad g'(\theta) = \frac{1}{(1-\theta)^2}$$

   to get

   $$\sqrt{n}\left(\hat{o} - \frac{\theta}{1-\theta}\right) \to \mathcal{N}\left(0, \frac{\theta}{(1-\theta)^3}\right)$$

   (c) The Jensen inequality states that if f is a convex function

   $$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

   In particular, we can reduce left-skewness by applying a convex function and conversely we can reduce right-skewness by applying a concave function.

2. (a) In this case the dataset is sufficiently large ($n = 30$), thus we may use the confidence interval based on the normal distribution

$$\left[\bar{x} - \frac{1.96S}{\sqrt{30}},\ \bar{x} + \frac{1.96S}{\sqrt{30}}\right] = [1.254, 1.519]$$

(b) First of all we need to find the median (there can be slight differences depending on your convention for calculating the median)

$$m = 1.26$$

The dataset can then be transformed into a binary sequence that will be used for the run test

$$(000101000110101111000001101111).$$

the number of runs, zeros and ones are

$$R = 14, \quad n_0 = 15, \quad n_1 = 15.$$

if the sequence was i.i.d. we would have the following mean number of runs

$$\mu = \frac{2n_0n_1}{30} + 1 = 16.$$

and the variance

$$\sigma^2 = \frac{(\mu-1)(\mu-2)}{29} = 7.24.$$

We compare the statistic

$$\frac{R-\mu}{\sigma} = -0.74$$

to the threshold of the standard normal distribution. Since $-1.96 < -0.74 < 1.96$, we observe no significant difference with an independent sample.

(c) i. The constant c should be such that the density function integrates to 1

$$\frac{1}{c} = \int_1^\infty x^{-5/2}dx = \left[-\frac{3}{2}x^{-3/2}\right]_1^\infty = \frac{3}{2} \Leftrightarrow c = \frac{2}{3}.$$

ii.

$$F(x) = \int_1^x f(t)dt = \frac{2}{3} \cdot \int_1^x t^{-5/2}dt = \left[-t^{-3/2}\right]_1^x = 1 - x^{-3/2}.$$

iii. For simplicity we partition the range into three bins, each containing 10 entries

$$[1, 1.19), \quad [1.19, 1.48), \quad [1.48, \infty).$$

the expected frequencies of the bins are

$$e_1 = F(1.19) = 0.23, \quad e_2 = F(1.48) - e_1 = 0.21, \quad e_3 = 1 - F(1.48) = 0.56$$

We compare the expected number of entries per bin to the observations

$$\frac{(10 - 30 \cdot 0.23)^2}{30 \cdot 0.23 \cdot (1 - 0.23)} + \frac{(10 - 30 \cdot 0.21)^2}{30 \cdot 0.21 \cdot (1 - 0.21)} + \frac{(10 - 30 \cdot 0.56)^2}{30 \cdot 0.56 \cdot (1 - 0.56)}.$$

For the χ^2 with 2 degrees of freedom, this would yield a p-value of 0.042 which is below the threshold of 5%. According to this test we would reject the assumption that the data follow the Pareto distribution with index 5/2.

iv. In the Kolmogorov–Smirnov distribution, we compare the empirical cumulative distribution

$$F_n(x) = \sum_{i=1}^{n} \frac{\mathbf{1}(x_i \leq x)}{n},$$

to the assumed distribution, via the statistic

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$

which should follow the Kolmogorov–Smirnov distribution with n trials. In this case, we should compare the value 0.176 to the Kolmogorov–Smirnov distribution with 30 trials. This yields a p-value of 0.14, suggesting that we can't reject our assumption.

3. (a) This follows immediately from the central limit theorem, as the mean of the exponential distribution with rate parameter λ is $1/\lambda$ and its variance is $1/\lambda^2$.

(b) According to the Neyman–Pearson lemma, the test should have the following acceptance region

$$A = \left\{ \mathbf{x} : \frac{\mathcal{L}(\lambda_0|\mathbf{x})}{\mathcal{L}(\lambda_1|\mathbf{x})} > k_\alpha \right\}$$

where k_α is such that

$$\mathbb{P}[\mathbf{X} \in A^c|\lambda = \lambda_0] = \alpha.$$

Let us express the acceptance region in a more tractable way

$$\frac{\mathcal{L}(\theta_0|\mathbf{x})}{\mathcal{L}(\theta_1|\mathbf{x})} > k_\alpha \Leftrightarrow \left(\frac{\lambda_0}{\lambda_1}\right)^n e^{(\lambda_1-\lambda_0)n\bar{\mathbf{x}}} > k_\alpha \Leftrightarrow e^{(\lambda_1-\lambda_0)n\bar{\mathbf{x}}} > k_\alpha \left(\frac{\lambda_1}{\lambda_0}\right)^n$$

$$\Leftrightarrow (\lambda_1 - \lambda_0)n\bar{\mathbf{x}} > \log(k_\alpha) + n(\log(\lambda_1) - \log(\lambda_0)) \Leftrightarrow \bar{\mathbf{x}} > \frac{\log(k_\alpha) + n(\log(\lambda_1) - \log(\lambda_0))}{(\lambda_1 - \lambda_0)n}$$

$$\Leftrightarrow \sqrt{n}\left(\bar{\mathbf{x}} - \frac{1}{\lambda_0}\right) > \sqrt{n}\left(\frac{\log(k_\alpha) + n(\log(\lambda_1) - \log(\lambda_0))}{(\lambda_1 - \lambda_0)n} - \frac{1}{\lambda_0}\right)$$

in particular, using the central limit theorem, we can define the best acceptance region by taking k_α such that

$$1.96 = \sqrt{n}\left(\frac{\log(k_\alpha) + n(\log(\lambda_1) - \log(\lambda_0))}{(\lambda_1 - \lambda_0)n} - \frac{1}{\lambda_0}\right)$$

that is

$$k_\alpha = \exp\left(\left(\frac{1.96}{\sqrt{n}} + \frac{1}{\lambda_0}\right)(\lambda_1 - \lambda_0)n - n(\log(\lambda_1) - \log(\lambda_0))\right)$$

4. (a)
$$\mathbb{E}[X_i|\theta] = -\theta + (1-\theta) = 1-2\theta.$$
In particular we can define a estimator with the method of moments
$$\bar{\mathbf{X}} = 1 - 2\hat{\theta} \Leftrightarrow \hat{\theta} = \frac{1-\bar{\mathbf{X}}}{2}.$$

(b) The likelihood of the sample is given by
$$\mathcal{L}(\theta|\mathbf{X}) = \prod_{i=1}^{n} p(X_i|\theta) = \theta^k(1-\theta)^{n-k},$$
where k is the number of entries equal to -1. We want to maximise this function
$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ \theta^k(1-\theta)^{n-k} = \underset{\theta}{\operatorname{argmax}}\ k\log(\theta) + (n-k)\log(1-\theta).$$
The maximum likelihood estimator must cancel the first order derivative of the log-likelihood
$$0 = \frac{k}{\hat{\theta}} - \frac{n-k}{1-\hat{\theta}} \Leftrightarrow \hat{\theta} = \frac{k}{n}.$$
Finally, we check the sign of the second order derivative
$$-\frac{k}{\hat{\theta}^2} + \frac{n-k}{(1-\hat{\theta})^2} = -\frac{n^2}{k} + \frac{n^2}{n-k}.$$
It is larger than 0 if $n > k$.

(c) First of all we observe that k follows the Binomial distribution with success rate θ and n trials. In particular k/n is an unbiased estimator and its variance is $\theta(1-\theta)/n$. We should check if this variance coincides with the Cramér–Rao bound. First of all, we calculate the Fisher information
$$I(\theta) = n\mathbb{E}\left[\left(\frac{\partial\ell(X_1,\theta)}{\partial\theta}\right)^2\right] = n\mathbb{E}\left[\left(\frac{\mathbf{1}(X_1=-1)}{\theta} - \frac{\mathbf{1}[X_1=1]}{1-\theta}\right)^2\right] = \frac{n}{\theta} + \frac{n}{1-\theta}.$$
In particular we have
$$\mathbb{V}[k/n] = \frac{1}{I(\theta)},$$
thus we have a minimum variance unbiased estimator.