1) After running a sample of size $n = 50$ a researcher found the following values:

0.000	0.001	0.042	0.177	0.365	0.091	0.092	0.487	0.527	0.454
0.233	0.831	0.932	0.568	0.556	0.051	0.767	0.019	0.252	0.298
0.876	0.532	0.920	0.515	0.810	0.188	0.886	0.572	0.077	0.815
0.985	0.118	0.894	0.784	0.101	0.253	0.020	0.378	0.679	0.681
0.753	0.006	0.624	0.126	0.618	0.771	0.187	0.497	0.510	0.316

The same values ordered in increasing order are:

0.000	0.001	0.006	0.019	0.020	0.042	0.051	0.077	0.091	0.092
0.101	0.118	0.126	0.177	0.187	0.188	0.233	0.252	0.253	0.298
0.316	0.365	0.378	0.454	0.487	0.497	0.510	0.515	0.527	0.532
0.556	0.568	0.572	0.618	0.624	0.679	0.681	0.753	0.767	0.771
0.784	0.810	0.815	0.831	0.876	0.886	0.894	0.920	0.932	0.985

a) Knowing that $\sum_{i=1}^{50} x_i = 22.237$ and $\sum_{i=1}^{50} x_i^2 = 14.660$, compute the sample average $\bar{x}$ and sample variance s^2.

$$\bar{x} = \frac{1}{50}\sum_{i=1}^{50} x_i = \frac{22.237}{50} = 0.445 \qquad s^2 = \frac{1}{49}\sum_{i=1}^{50} x_i^2 - \frac{50}{49}\bar{x}^2 = 0.097$$

b) Compute the median $\tilde{x}$ and the fourth spread f_s.

$$\tilde{x} = \frac{0.497 + 0.487}{2} = 0.492 \qquad f_s = 0.753 - 0.126 = 0.627$$

Our researcher wants to draw an histogram of the data.

c) How may classesw is will he chose?

The best number of class is $\sqrt{50} \simeq 7$. He will chose $k = 7$ classes.

d) Write the intervals the define these classes.

We have $\min(x_i) = 0.00$ and $\max(x_i) = 0.985 \simeq 1$ so that we can take every class of size $\Delta = \frac{1}{7}$. The i-th class will have boundaries $(i-1)\Delta$ and $i\Delta$ so that we have:

1	2	3	4	5	6	7
$[0, 0.142]$	$[0.142, 0.285]$	$[0.285, 0.428]$	$[0.428, 0.571]$	$[0.571, 0.714]$	$[0.714, 0.857]$	$[0.857, 1]$

where the first line indicates the class and the second its interval.

e) Compute the value of the histogram for each class.

We have

class	1	2	3	4	5	6	7
frequency	13	6	4	9	5	7	6
relative frequency	0.26	0.12	0.08	0.18	0.1	0.14	0.12
histogram	1.82	0.84	0.56	1.26	0.7	0.98	0.84

where the first line indicates the class, the second the frequency of that class, the third the relative frequency and the fourth the histogram value.

Assume that the numbers come from a population characterized by a uniform distribution between 0 and θ.

f) Give a non biased estimate of θ.

You can use as an estimator $\hat{\theta}_2 = \frac{n+1}{n}\max(x_i) = \frac{51}{50}0.985 = 1.004$ or $\hat{\theta}_1 = 2\bar{x} = 0.890$

g) Derive a 95% CI for θ based on the above sample. (**Hint**: use the estimator $\hat{\theta}_1 = 2\bar{x}$.)

A 95% CI for the population average is given by:

$$\left[\bar{x} - \frac{1.96s}{\sqrt{50}}, \bar{x} + \frac{1.96s}{\sqrt{50}}\right] = [0.359, 0.531]$$

so that a 95% CI for θ is

$$\left[2\bar{x} - \frac{2 \cdot 1.96s}{\sqrt{50}}, 2\bar{x} + \frac{2 \cdot 1.96s}{\sqrt{50}}\right] = [0.718, 1.062]$$

h) **Bonus** Use the estimator $\hat{\theta}_2 = \frac{n+1}{n}\max(x_i)$ to derive a 95% upper confidence bound for θ.

Observe that if X_i is uniform between 0 and θ then $Y_i = X_i/\theta$ has uniform distribution between 0 and 1. From this we have that $Z = \max(X_i)/\theta$ has c.d.f. $P(Z < z) = z^n$ if $0 < z < 1$ and 0 otherwise. Finally

$$T = \frac{1}{\theta}\frac{n+1}{n}\max(X_i)$$

has c.d.f.

$$f(t) = P(T < t) = \left(\frac{n}{n+1}t\right)^n$$

for $0 < t < \frac{n+1}{n}$ and 0 otherwise. We now observe that $P(T > \bar{t}) = 0.95$ implies that $\bar{t} = \frac{n+1}{n}\sqrt[n]{0.05}$. This means that

$$P\left(\frac{1}{\theta}\frac{n+1}{n}\max(X_i) > \frac{n+1}{n}\sqrt[n]{0.05}\right) = 0.95$$

so that

$$P\left(\theta < \frac{\max(X_i)}{\sqrt[n]{0.05}}\right) = 0.95$$

so that an upper confidence bound at 95% is given by:

$$\theta < \frac{\max(X_i)}{\sqrt[n]{0.05}}$$

In particular in the present situation we get $\theta < 1.33$ at 95% confidence level. On the other hand you can say that $\theta > \max(X_i)$ with 100% confidence. So that the interval

$$\left[\max(X_i), \frac{\max(X_i)}{\sqrt[n]{0.05}}\right]$$

is a 95% CI for θ. With a little analysis you can see that the size of this interval is order $\frac{1}{n}$ in contrast with the one obtained in g) that is of order $\frac{1}{\sqrt{n}}$.

2) The time between two successive arrivals of a bus at a bus stop is assumed to be distributed exponentially with parameter λ. An actual observation of a sample of size 10 gives the following numbers (in hours):

$$0.01 \quad 2.13 \quad 0.11 \quad 0.24 \quad 2.29 \quad 1.37 \quad 3.92 \quad 0.97 \quad 0.38 \quad 0.32$$

Let X_i, $i = 1, \ldots, 10$ be a random sample for this problem.

a) Write the p.d.f. of the X_i.

$$f(x_i, \lambda) = \lambda e^{-\lambda x_i}$$

b) Write the joint p.d.f. of the random sample X_i, $i = 1, \ldots, 10$.

$$F(x_1, x_2, \ldots, x_10, \lambda) = \prod_{i=1}^{10} f(x_i) = \lambda^{10} \prod_{i=1}^{10} e^{-\lambda x_i} = \lambda^{10} e^{-\lambda \sum_{i=1}^{10} x_i} = \lambda^{10} e^{-10\lambda \bar{x}}$$

c) Find a MLE $\hat{\lambda}$ estimator for λ. You must find the maximum of $F(x_1, x_2, \ldots, x_10, \lambda)$ as a function of λ. Derivating we get:

$$10\lambda^9 e^{-10\lambda \bar{x}} - 10\bar{x}\lambda^{10} e^{-10\lambda \bar{x}} = 0$$

that implies

$$\lambda = \frac{1}{\bar{x}}$$

d) What value of λ you get for the above sample?

We get $\bar{x} = 1.174$ so that $\theta = 0.8517$.

e) Is $\hat{\lambda}$ unbiased? If not, can you make it unbiased?

The MLE estimator is thus:

$$\hat{\lambda} = \frac{1}{\bar{X}}$$

Observe that in general:

$$E(\hat{\lambda}) = E\left(\frac{1}{\bar{X}}\right) \neq \frac{1}{E(\bar{X})} = \lambda$$

so that $\hat{\lambda}$ is biased. We can evaluate the bias computing:

$$E\left(\frac{1}{\bar{X}}\right) = \int_0^\infty dx_1 \int_0^\infty dx_2 \cdots \int_0^\infty dx_{10} \frac{1}{\bar{x}} F(x_1, x_2, \ldots, x_{10}, \lambda) =$$
$$= 10\lambda^{10} \int_0^\infty dx_1 \int_0^\infty dx_2 \cdots \int_0^\infty dx_{10} \frac{1}{\tilde{x}} e^{-\lambda \sum_{i=1}^{10} x_i} =$$
$$= 10\lambda^{10} \int_0^\infty d\tilde{x} \int_0^{\tilde{x}} dx_1 \int_0^{\tilde{x}-x_1} dx_2 \cdots \int_0^{\tilde{x}-x_1-\cdots-x_9} dx_9 \frac{1}{\tilde{x}} e^{-\lambda\tilde{x}} =$$
$$= \frac{10}{9!}\lambda^{10} \int_0^\infty \tilde{x}^8 e^{-\lambda\tilde{x}} d\tilde{x} = \frac{10 \cdot 8!}{9!}\lambda = \frac{10}{9}\lambda$$

where we set $\tilde{x} = \sum_{i=1}^{10} x_i$. This implies that

$$\hat{\lambda}_1 = \frac{9}{10}\frac{1}{\bar{X}}$$

is an unbiased estimator.

3) Flipping a coin 10000 times you get the following result: 5100 Head and 4900 Tail.

a) Estimate the probability of observing a Head flipping the coin.

The probability is given by

$$P(\mathrm{H}) = \frac{5100}{10000} = 0.51$$

b) Give a 99% CI of the probability of observing a Head.

The variance is given by $s = \sqrt{0.51 \cdot 0.49} = 0.5$ so that the CI is

$$\left[0.51 - \frac{2.58 \cdot 0.5}{\sqrt{10000}}, 0.51 + \frac{2.58 \cdot 0.5}{\sqrt{10000}}\right] = [0.497, 0.523]$$

c) In your opinion, is the coin fair?

A 99% CI is compatible with a fair coin because $0.5 \in [0.497, 0.523]$. At this level we cannot say that the coin is not fair although it looks suspicious.

4) A researcher run a small sample of size $n = 100$ on a given population and obtain a sample average $\bar{x} = 1$ and a sample variance $s^2 = 2$. He wants a 95% CI on the average of the population with a precision of 0.001. Is the above sample large enough? If he has to run a new sample, which size will he choose?

A 95% CI with the given sample will give the interval $[1 - 1.96\sqrt{2}/10, 1 - 1.96\sqrt{2}/10] = [0.723, 1.277]$ with a precision of 0.277. The sample is not large enough. He can assume that the variance of the population is close to 2 so that he will get that he needs a sample of size size at least

$$N = \left(\frac{1.96\sqrt{2}}{0.001}\right)^2 = 7683200$$