NON-PARAMETRIC STATISTICS - WORKSHOP 1

Andrés Ospina Patiño
Santiago Cartagena Agudelo
Pablo Buitrago Jaramillo
Sofía Vega López
Juan Andrés Giraldo Aristizábal
Ingeniería Matemática - Universidad EAFIT

September 8, 2022

Exercise 1: Analysis of the ECDF of the temperatures from the last 35 years in Canada

The data used is the average daily temperatures in Canada for the past 35 years. The empirical cumulative distributions (ECDFs) for each year are presented in Figure 9, where the yellow-most curve represents the ECDF of the data recorded for the first year, and the blue-most is the last year.

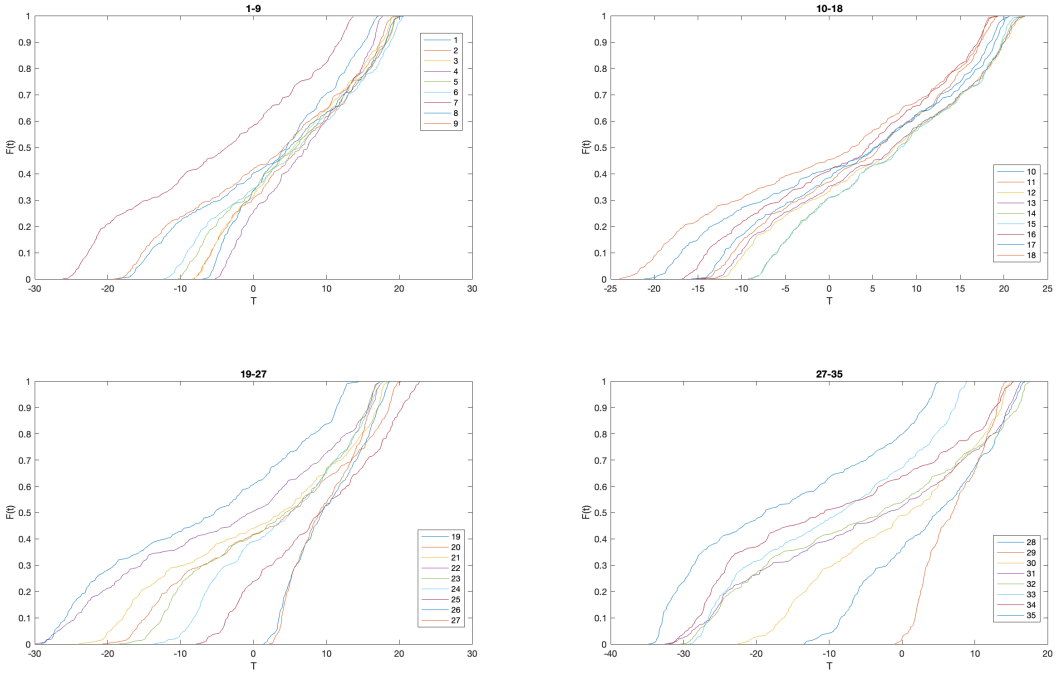

Figure 1: Empirical distributions by years.

Based on this plots, it can be observed a climate effect change because of the distribution displacement over the years. As we are seeing Empirical Cumulative Distribution Functions (ECDF), we get information about $P(T \leq t)$ where T is a random variable that represents the temperature and t an arbitrary temperature value; so changes in the ecdf over the years implies changes in the probability that the temperature is less or equal to a certain value, in the same time periods. As a consequence, we can identify distributions that are completely above or below others, so they are comparable. Given two comparable ecdfs, the distribution that is below represents a year that was probabilistically hotter than the one above.

Table that indicates the hottest from the consecutives years by mean and by ecdf

Year 1	Year 2	Mean year 1	Mean year 2	Hottest year by Mean	Hottest year by ecdf
1	2	4,68986301	6,149863014	2	0
2	3	6,14986301	5,509589041	2	0
3	4	5,50958904	6,811506849	4	0
4	5	6,81150685	5,232328767	4	0
5	6	5,23232877	5,263013699	6	0
6	7	5,2630137	-5,059178082	6	7
7	8	-5,0591781	3,1	8	7
8	9	3,1	2,24630137	8	0
9	10	2,24630137	4,080821918	10	9
10	11	4,08082192	4,119452055	11	0
11	12	4,11945205	6,130684932	12	11
12	13	6,13068493	5,81260274	12	0
13	14	5,81260274	7,26630137	14	0
14	15	7,26630137	7,312876712	15	0
15	16	7,31287671	2,461643836	15	16
16	17	2,46164384	2,472876712	17	0
17	18	2,47287671	-0,151232877	17	18
18	19	-0,1512329	-7,08630137	18	19
19	20	-7,0863014	2,745479452	20	19
20	21	2,74547945	0,682465753	20	21
21	22	0,68246575	-3,408767123	21	22
22	23	-3,4087671	2,29890411	23	0
23	24	2,29890411	3,986575342	24	0
24	25	3,98657534	8,748219178	25	24
25	26	8,74821918	9,959178082	26	0
26	27	9,95917808	9,623835616	26	0
27	28	9,62383562	3,835068493	27	28
28	29	3,83506849	7,002739726	29	0
29	30	7,00273973	-0,851780822	29	0
30	31	-0,8517808	-4,790684932	30	0
31	32	-4,7906849	-5,023561644	31	0
32	33	-5,0235616	-9,65260274	32	0
33	34	-9,6526027	-9,24	34	0
34	35	-9,24	-16,51835616	34	35

The first two columns in the table indicates the years to be compared, the third and fourth column indicates the mean temperature for the selected year 1 and the selected year 2 respectively. The fifth column indicates which year from the two being compared is the hottest based on their mean temperature. The last column indicates the hottest year from the sample based on their empirical cumulative distribution functions, but as mentioned before, only ecdfs that do not intersect are

comparable, a situation that does not hold for every comparison, in consequence, if they are comparable the number of the year with the ecdf that is below the other will be marked in the table, if they are not comparable a zero will be marked in the table (as occurred several times).

Exercise 2: Plug-In estimator of the expected value

The plug-in principle is a technique used in probability theory and statistics to estimate a parameter of a probability distribution (e.g., the expected value, the variance, a quantile) that cannot be exactly computed. In general, the plug-in principle says that a feature of a given distribution can be approximated by the same feature of the empirical distribution of a sample of observations drawn from the given distribution (Van der Vaart, 2000).
The feature of the empirical distribution is called a plug-in estimate of the feature of the given distribution. For example, a quantile of a given distribution can be approximated by the analogous quantile of the empirical distribution of a sample of draws from the given distribution. The following is a formal definition of plug-in estimate.

A statistical functional $T(F)$ is any function of F. The plug-in estimator of $\theta = T(F)$ is defined by

$$\widehat{\theta}_n = T\left(\widehat{F}_n\right)$$

A functional of the form $\int a(x)dF(x)$ is called a linear functional. The plug-in estimator for linear functional $T(F) = \int a(x)dF(x)$ is:

$$T\left(\widehat{F}_n\right) = \int a(x)d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} a\left(X_i\right)$$

It is important to note that $T\left(F_n\right)$ converges to $T(F)$ as the sample size n increases.
In the practice, the limited area is calculated in the first quadrant for each empirical curve of the temperature data, this allows to extract the plug-in estimator of the mean, since the difference between the area with positive values and the area with negative values is extracted from the empirical curve, which allows estimating the mean of the series. The plug-in estimation of the mean for each year is presented in Figure 2, where a comparison with the natural estimator of the mean is presented.

So, the exercise asks to calculate the area between the ecdf and the line $y = 1$ in the first quadrant, then substract to the resulting quantity the area between the ecdf an the line $y = 0$ in the second quadrant. Remeber the cdf is bounded between 0 and 1 in all its domain, this difference is calculated as the following integral:

$$\int_0^{\infty} [1 - F(t)]dt - \int_{-\infty}^{0} F(t)dt$$

$$= \int_0^\infty [F(x)\Big|_t^\infty]dt - \int_{-\infty}^0 [F(x)\Big|_{-\infty}^t]dt$$

$$= \int_0^\infty \int_t^\infty f(x)dxdt - \int_{-infty}^0 \int_{-infty}^t f(x)dxdt$$

$$= \int_0^\infty \int_0^x f(x)dtdx + \int_{-infty}^0 \int_x^0 -f(x)dtdx$$

$$= \int_0^\infty xf(x)]dx + \int_{-\infty}^0 xf(x)dx$$

$$= \int_{-\infty}^\infty xf(x)dx$$

Which is the definition of the expected value of the random variable X, *E(X)*

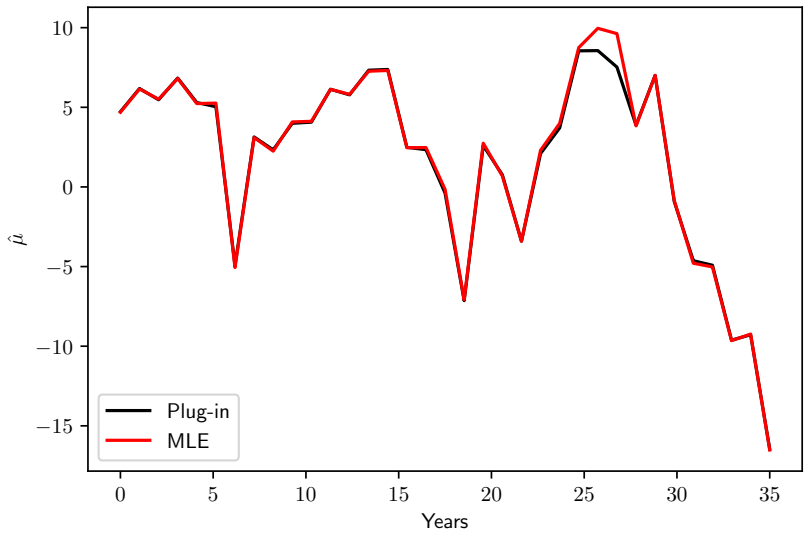

Figure 2: Mean estimation of the series.

From Figure 2, an almost identical estimation can be seen between the two estimators, however it shows a slight difference between the mean estimates, this is because the only case in which both estimates are identical is when the sample size is infinite.

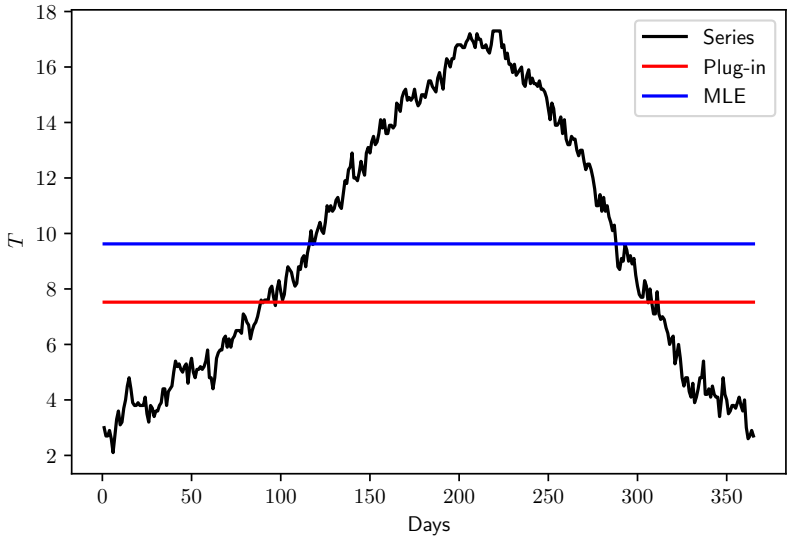

Figure 3: Comparison between the natural estimator and the plug-in estimator.

Exercise 3: Confidence intervals of the ECDF of the years with the least and the greater mean temperature

Calculate and plot the confidence bands for the empirical continuous distribution function (ECDF) of the coldest and hottest year in average with a confidence of 95 % . Are there any sectors that are not enclosed in the bands?
Let n be the size of the sample and $1 - \alpha$ the desired confidence for the bands. This method is based on the Dvoretzky–Kiefer–Wolfowitz inequality (see Wasserman (2006)). In order to calculate the confidence bands for the ECDF, we first define ϵ_n by the following formula:

$$\epsilon_n = \sqrt{\frac{1}{2n} \ln\left(\frac{2}{\alpha}\right)}$$

Let $\hat{F}_n(x)$ be the ECDF. Then, for each x in the ECDF we define the lower ($L(\cdot)$) and upper ($U(\cdot)$) bound by:

$$L(x) = \max\{\hat{F}_n(x) - \epsilon_n, 0\}$$
$$U(x) = \min\{\hat{F}_n(x) + \epsilon_n, 1\}$$

The results obtained by using the temperatures of the coldest and hottest year are seen in Figure 4.

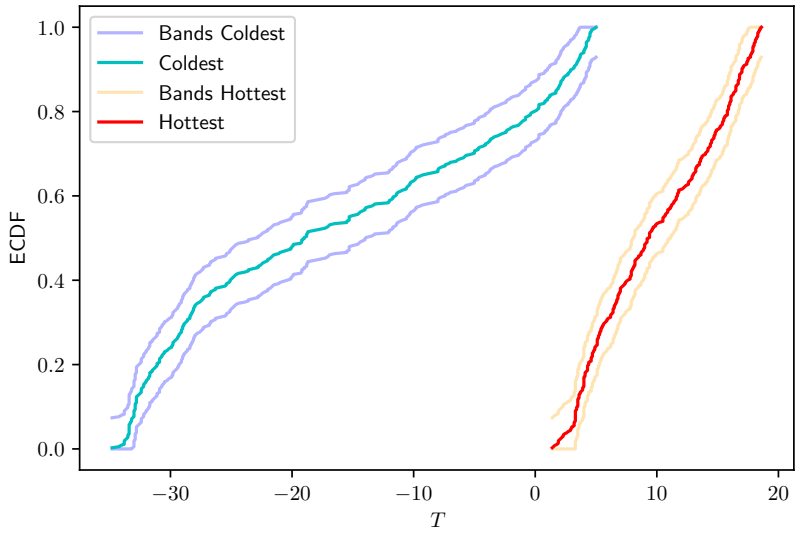

Figure 4: Bands for the coldest and hottest year.

It can be seen that the upper and lower bands fully enclose the ECDF. Nevertheless, there exists two points where the function and the bands meets. This happens in the lower band at the beginning and the upper band at the final point.
This phenomenon is due to the full certainty at those points in a sense that the lower bound, at the start, has to be the same point as it cannot go lower than 0. A similar reasoning can explain the upper bound and the final point.

Exercise 4: The Glivenko-Cantelli theorem simulation

Visualize the Glivenko Cantelli Theorem for a Weibull distribution with scale parameter, the grade that the work group expects to obtain in this course (5.0), and shape parameter, the average age of the group (20) over the square of the expected mark.

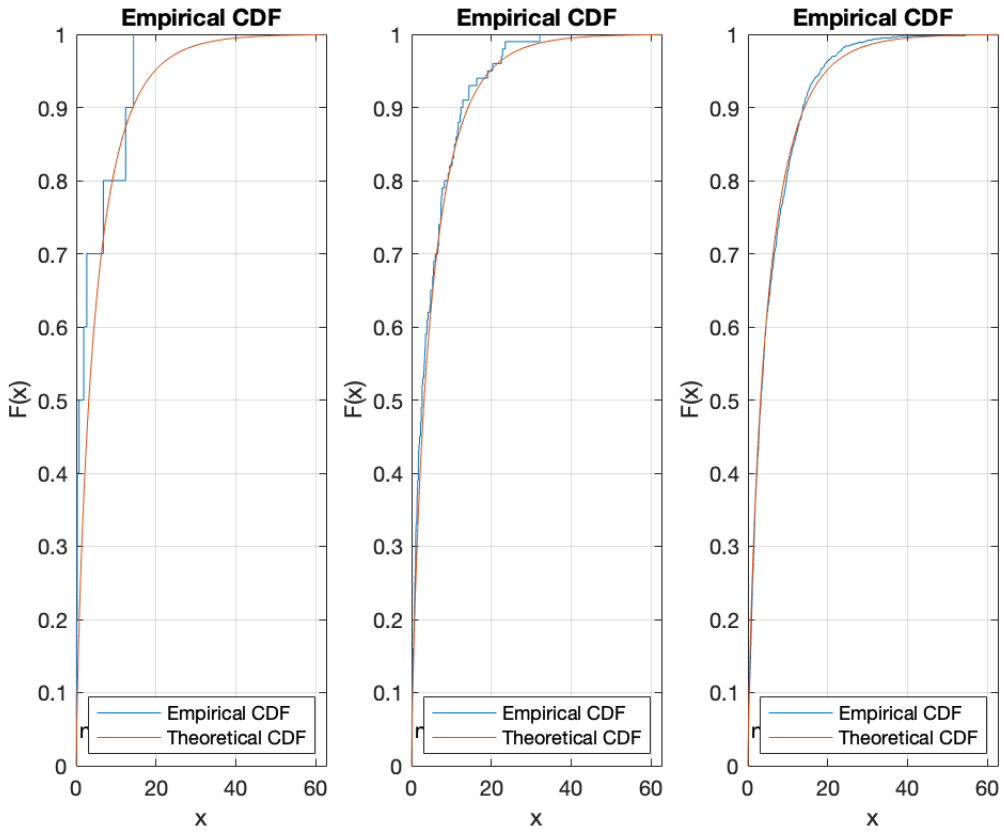

Figure 5: n=10, n=100, n=1000 respectively

Glivenko–Cantelli Theorem:

$$Sup_x|\hat{F}_n(x) - F(x)| \xrightarrow{\text{a.s}} 0$$

As can be seen in 5, as the number of observations grows, the difference between the empiric and the theoretical distribution is less as n is larger, since with $n = 1000$ the cumulative distribution functions are very similar. So, with n tending to infinity is almost sure that the supreme of the difference between the empiric and the theoretical cumulative distribution function is 0, what means that no difference should be noted on the plot with a large enough n, because the supreme is getting smaller and smaller just as you can see in 6.

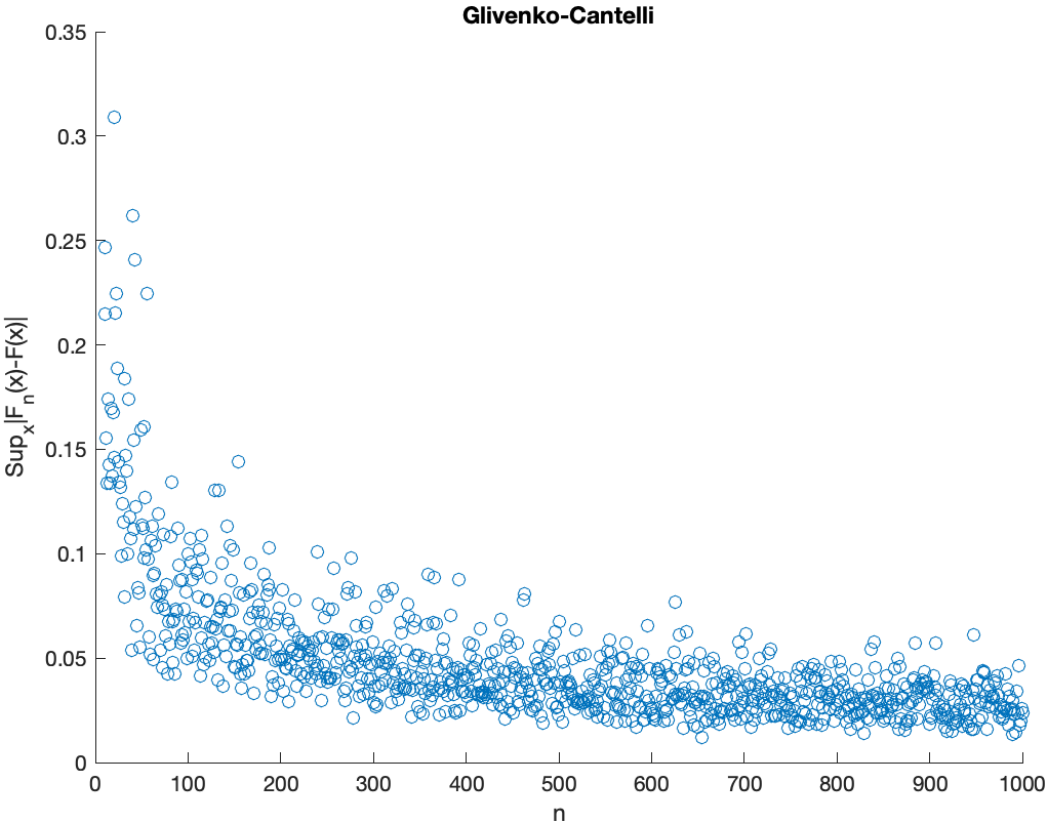

Figure 6: Glivenko-Cantelli theoreme

Exercise 5: Jensen inequality for non-convex functions

Given a function g, if it is concave then $E[g(X)] \leq g(E[X])$
Proof:

By definition g is concave if for any point x_0 the graph of g lies completely below its tangent at point x_0, i.e.

$$g(x) \leq g(x_0) + b(x - x_0) \quad \forall x$$

where b is the slope of the tangent line. Then let us consider $x = X$ and $x_0 = E[X]$, getting

$$g(X) \leq g(E[X]) + b(X - E[X])$$

Finally, if the expecting value is toked in both sides of the inequality and properties are applied, we have the next

$$E[g(X)] \leq E[g(E[X])] + E[b(X - E[X])]$$

$$E[g(X)] \leq g(E[X]) + b(E[X] - E[X])$$

Taking into account that $E[g(E[X])] = g(E[X])$, since $g(E[X])$ is a constant

$$E[g(X)] \leq g(E[X]) + b(0)$$
$$E[g(X)] \leq g(E[X])$$

Example:

The Jensen inequality is commonly used in financial time series, such as calculating the rate of return. For example, the average rate of return can be calculated over an interval such as monthly over one year. It could be calculated using the arithmetic mean, but this value would be optimistic (a larger value) and would exclude the reinvestment of returns and would assume any losses would be topped up at the start of each period. Instead, the geometric mean must be used to give the real average rate of return over the interval (a smaller value), correctly taking into account losses with reinvesting and compounding gains. The arithmetic average return is always larger than the geometric average return. (see Brownlee (2019))

Exercise 6: L-statistics v.s. M-statistics

L-statistics:
Take the form $\mu^* = \sum_i w_i x_{(i)}$, where $x_{(i)}$ is the ith ordered value of a sample of size n and w_i are weights. (Breakdown: fraction of data which can be given arbitrary values without making the estimator, arbitrarily too large or too small.)
Examples:

- *Mean:* If w_i takes the same weight for all value, then the mean is obtained. Breakdown = $\frac{1}{n}\%$
- *Median:* If $n = 2m + 1$, then $w_m = 1$ and $w_i = 0\ \forall i \neq m$, that is the $\frac{n-1}{2}$-th ordered value is the median. If $n = 2m$, then $w_i = 0.5$ for $i = m,\ m + 1$, 0 otherwise. Breackdown = 50%
- *Minimum:* $w_1 = 1$, $w_i = 0\ \forall i \neq 1$, since the $x_{(1)}$ is the first ordered value of the sample, what means that it is the smallest. Breakdown = $\frac{1}{n}\%$
- *Maximum:* $w_n = 1$, $w_i = 0\ \forall i \neq n$, since the $x_{(n)}$ is the last ordered value of the sample, what means that it is the greatest. Breakdown = $\frac{1}{n}\%$

M-statistics:

They are the ones who solves the optimization problem $\rho(x, \theta)$, relative to a statistic θ.
Examples:
Let $(X_1, ..., X_n)$ be a set of independent, identically distributed random variables, with distribution F.

- *Mean:* If
$$\rho(x, \theta) = \frac{(x - \theta)^2}{2}$$
then, it is observed that this is minimized when θ is the mean of the X_s

- *Median:* If
$$\rho(x, \theta) = |x - \theta|$$
then, it is observed that this is minimized when θ is the median of the X_s
- *Maximum likelihood estimator:* To density function defined as $f'(x, \theta)$, seeks to minimize $f(x, T) = -log(f; (x, T))$
- *Non-linear least squares.*

Assuming sorted data, L-statistics involving only a few points can be calculated with far fewer mathematical operations than efficient estimates (see Mosteller (1946)), also are often much more robust than maximally efficient conventional methods. However, with little data they are not very efficient.
On the other, M-statistics are more robust under small data, but have a high computational cost given the optimization.

Exercise 7: j-th statistic distribution

Deduce the distribution and density of the j-th ordered statistic. Explain with detail what would be a simple procedure to simulate the j-th ordered statistic. Simulate 1000 observations of some ordered statistic of a sample of size n that comes from a Weibull distribution. Draw in a same graph the ECDF and theoretical distribution.
Let's first deduce the distribution and density of the j-th ($X_{[j]}$) ordered statistic. Let $X_1, \ldots, X_n$ be independent random variables that come from a distribution $F(\cdot)$. Hence, the probability that $X_i \leq t$ is given by:
$$\mathrm{P}(X_i \leq t) = F(t)$$
Let Z_t be a random variable that represents the number of variables whose value are less than t. Hence:
$$Z_t \in \{0, 1, \ldots, n\}$$
Hence:
$$\begin{aligned}
\mathrm{P}(Z_t = 0) &= \mathrm{P}(X_1 > t \wedge \ldots \wedge X_n > t) \\
&= \mathrm{P}(X_1 > t) \ldots \mathrm{P}(X_n > t) \\
&= (1 - F(t)) \ldots (1 - F(t)) \\
&= (1 - F(t))^n \\
\mathrm{P}(Z_t = 1) &= \binom{n}{1} \mathrm{P}(X_1 \leq t \wedge X_2 > t \wedge \ldots \wedge X_n > t) \\
&= \binom{n}{1} F(t)(1 - F(t)) \ldots (1 - F(t)) \\
&= \binom{n}{1} F(t)(1 - F(t))^{n-1} \\
&\vdots
\end{aligned}$$

$$\mathrm{P}\left(Z_t = j\right) = \binom{n}{j} F(t)^j (1 - F(t))^{n-j}$$

Hence, the distribution j-th ordered statistic is given by:

$$\mathrm{P}\left(X_{[j]} \leq t\right) = \mathrm{P}\left(Z_t \geq j\right) = \sum_{i=j}^{n} \binom{n}{i} F(t)^i (1 - F(t))^{n-i}$$

In these manner, we obtained the distribution for the j-th ordered statistic. Then, to obtain the density ($f_{[j]}(\cdot)$) we just differentiate hence:

$$\begin{aligned}
f_{[j]}(t) &= \frac{d}{dt}\mathrm{P}\left(X_{[j]} \leq t\right) \\
&= \sum_{i=j}^{n} \binom{n}{i} \left[iF(t)^{i-1}(1 - F(t))^{n-i} f(t) - (n-i)F(t)^i (1 - F(t))^{n-i-1} f(t)\right] \\
&= f(t) \sum_{i=j}^{n} \binom{n}{i} F(t)^{i-1}(1 - F(t))^{n-i-1} \left[i(1 - F(t)) - (n-i)F(t)\right] \\
&= f(t) \sum_{i=j}^{n} \binom{n}{i} F(t)^{i-1}(1 - F(t))^{n-i-1} (i - nF(t)) \\
&= f(t) \left[\sum_{i=j}^{n} i\binom{n}{i} F(t)^{i-1}(1 - F(t))^{n-i-1} - \sum_{i=j}^{n} n\binom{n}{i} F(t)^i (1 - F(t))^{n-i-1}\right] \\
&= nf(t) \left[\sum_{i=j}^{n} \binom{n-1}{i-1} F(t)^{i-1}(1 - F(t))^{n-i-1} - \sum_{i=j}^{n} \binom{n}{i} F(t)^i (1 - F(t))^{n-i-1}\right] \\
&= n\binom{n-1}{j-1} F(t)^{j-1}(1 - F(t))^{n-j} f(t)
\end{aligned}$$

Furthermore, simulating a j-th statistic can be done easily. In first place, generate samples of size n and obtain the j-th statistic of that sample. These generates one data of the j-th ordered statistic. Repeat the previous process until the desired number of data is obtained.
The result of the simulation of the 5-th ordered statistic for a Weibull distribution can be found in Figure 7.

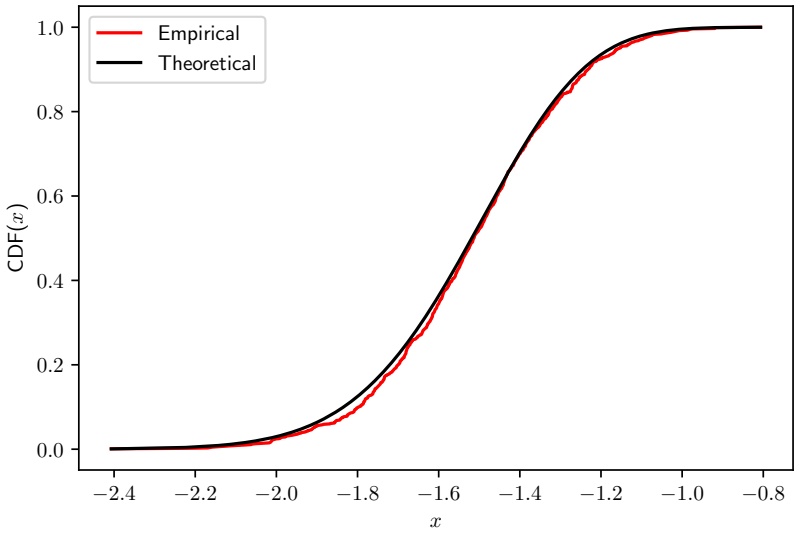

Figure 7: Comparison between ECDF and theoretic distribution.

Exercise 8: Chebyshev inequality bound comparision using the exponential distribution

Suppose X is an exponentially random variable of parameter β. Calculate:

$$P(|X - \mu| > k\sigma)$$

for $k > 1$.
Recall Chebyshev's inequality: let Y be a random variable and let $\mathbb{E}[Y] = \mu$ and $\text{Var}[Y] = \sigma^2$, then

$$P(|Y - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

Let

$$\begin{aligned}
P(|X - \mu| > k\sigma) &= P(-k\sigma > X - \mu > k\sigma) \\
&= P(-k\sigma + \mu > X > k\sigma + \mu) \\
&= P(X < \mu - k\sigma) + P(X > k\sigma + \mu) \\
&= P\left(X < \frac{1-k}{\beta}\right) + 1 - P\left(X < \frac{k+1}{\beta}\right) \qquad \text{but } 1 - k < 0 \\
&= 1 - \left(1 - e^{-\frac{k+1}{\beta^2}}\right) \\
&= P(|X - \mu| > k\sigma) \\
&\leq \frac{\sigma^2}{(k\sigma)^2} \\
&\leq \frac{1}{k^2}
\end{aligned}$$

Clearly $e^{-\frac{k+2}{\beta^2}}$ is always lesser than 1 and because $k > 1$ then $\frac{1}{k^2}$ is also lesser than one. Thus, $P(|X - \mu| > k\sigma)$ is bounded by Chebyshev's inequality.

Exercise 9: Chebyshev inequality in the Poisson distribution

Prove that if $X \sim \text{Poisson}(\lambda)$, then

$$P(X \geq 2\lambda) \leq \frac{1}{\lambda}.$$

Proof: Let $X \sim \text{Poisson}(\lambda)$. Recall Chebyshev's inequality: let Y be a random variable and let $\mathbb{E}[Y] = \mu$ and $\text{Var}[Y] = \sigma^2$, then

$$P(|Y - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

Clearly, $\mathbb{E}[X] = \text{Var}[X] = \lambda$. If we set $t = \lambda$, then

$$\begin{aligned}
P(|X - \lambda| \geq \lambda) &\leq \frac{1}{\lambda} \\
P[(X - \lambda \geq \lambda) \cup (X - \lambda \leq -\lambda)] &\leq \frac{1}{\lambda} \\
P[(X \geq 2\lambda) \cup (X \leq 0)] &\leq \frac{1}{\lambda} \\
P(X \geq 2\lambda) + P(X \leq 0) &\leq \frac{1}{\lambda} \\
P(X \geq 2\lambda) &\leq \frac{1}{\lambda}
\end{aligned}$$

Exercise 10: Quadratic mean convergence implies probability convergence

The sequence $\{X_n\}$ of random variables is said to converge in probability to X if

$$\forall \epsilon > 0, \ \lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$$

The sequence $\{X_n\}$ of random variables is said to converge in mean-square to X if

$$\lim_{n\to\infty} \mathbb{E}\left[(X_n - X)^2\right] = 0$$

The convergence in mean-square implies convergence in probability. Proof: Let $\{X_n\}$ be a sequence of random variables that converges in mean-square to X. Recall Markov's inequality: Let Y be a non-negative random variable and suppose $\mathbb{E}[Y]$ exists. Then for any $t > 0$,

$$P(Y > t) \leq \frac{\mathbb{E}[Y]}{t}.$$

Take $Y = |X_n - X|$ and $t = \epsilon$, then

$$\begin{aligned}
P(|X_n - X| > \epsilon) &= P\left[(X_n - X)^2 > \epsilon^2\right] \\
&\leq \frac{\mathbb{E}\left[(X_n - X)^2\right]}{\epsilon^2}
\end{aligned}$$

Since $\{X_n\}$ converges in mean-square to X, $\mathbb{E}\left[(X_n - X)^2\right] \to 0$ as $n \to \infty$, which directly implies that $P(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$.

Exercise 11: ECDF converges in probability to the theoretic distribution

Show that the ECDF converges in probability to the theoretical continuous distribution function.
Proof: Let X_i, for $i = 1, \ldots, n$, be an independent data sample. Then, the empirical distribution function is defined as

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$$

where

$$I(X_i \leq x) = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{otherwise} \end{cases}$$

Hence, to see that it converges in probability lets see if the MSE tends to 0 when the number of samples is bigger. Let's calculate the expectancy of the estimator.

$$\begin{aligned} \mathbb{E}\left[\hat{F}_n(x)\right] &= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)\right] \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[I(X_i \leq x)\right] \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathrm{P}\left(X_i \leq x\right) \\ &= \frac{1}{n} \sum_{i=1}^{n} F(x) \\ &= F(x) \end{aligned}$$

In this manner, it is a non-biased estimator for the theoretical distribution. Hence, the MSE is calculated by the variance. Then:

$$\begin{aligned} \text{MSE} &= \text{Var}\left[\hat{F}_n(x)\right] + \text{Bias}\left(\hat{F}_n(x), F(x)\right) \\ &= \text{Var}\left[\frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)\right] \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}\left[I(X_i \leq x)\right] \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \left[\mathrm{P}\left(X_i \leq x\right)\left(1 - \mathrm{P}\left(X_i \leq x\right)\right)\right] \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \left[F(x)(1 - F(x))\right] \\ &= \frac{F(x)(1 - F(x))}{n} \end{aligned}$$

Therefore, when $n \to \infty$ the MSE tends to 0. Therefore,

$$\hat{F}_n(x) \xrightarrow{P} F(x)$$

Exercise 12: Confidence interval for the maximum temperature in the coldest year

Consider the daily temperatures of the hottest year in average. Calculate a confidence interval for the maximum temperature. Calculate the bias of $T_{[n]}$ and the variance.
To calculate the interval for the maximum temperature, we applied the normal bootstrap method. The normal bootstrap method consists of, in a sample $X_1, \ldots, X_n$ and a confidence of $1 - \alpha$:

1. First, extract k samples of the size n with repetition and calculate the desired statistic to find it's confidence intervals. Let S_i be the sample of each statistic calculated.

2. Calculate:
$$v_{\text{boot}} = \frac{1}{k} \sum_{i=1}^{k} \left(S_i - \frac{1}{n} \sum_{i=1}^{k} S_i \right)^2$$

3. Then, the interval of confidence for the statistic is, with S the statistic calculated in the original sample:
$$(S - z_{\alpha/2}\sqrt{v_{\text{boot}}}, S + z_{\alpha/2}\sqrt{v_{\text{boot}}})$$
where $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of a standard normal distribution.

The result for the maximum temperatures interval is, with $\alpha = 0.05$ and $k = 1000$:

$$T_{[n]} \in (18.486, 18.714)$$

On the other hand, to calculate the variance and bias the Jackknife method was used. The Jackknife method consists is explained in the following exercise. Hence, we find that the bias and the variance and we obtain:

$$b_{\text{jack}} = -0.10$$
$$v_{\text{jack}} = 7.49 \cdot 10^{-8}$$

Exercise 13: Bootstrap and Jacknife on the uniform distribution

In this section, a sample of a uniform distribution in the interval [0,1] is generated. Then a bootstrap method to calculate the variance and a Jackknife method to calculate the bias of this sample are implemented.
The bootstrap and the jackknife are non-parametric methods for computing standard errors and confidence intervals Rodgers (1999). A short review of the methodology will be presented, based on Wasserman (2006).

The Jackknife

Jackknife is a simple method for approximating the bias and variance of an estimator. Let $T_n = T(X_1, \ldots, X_n)$ be an estimator of some quantity θ and let $\text{bias}(T_n) = \mathbb{E}(T_n) - \theta$ denote the bias. Let $T_{(-i)}$ denote the statistic with the i^{th} observation removed. The jackknife bias estimate is defined by

$$b_{\text{jack}} = (n-1)\left(\bar{T}_n - T_n\right)$$

where $\bar{T}_n = n^{-1} \sum_i T_{(-i)}$. The bias-corrected estimator is $T_{\text{jack}} = T_n - b_{\text{jack}}$

The Bootstrap

Bootstrap is a method for estimating the variance and the distribution of a statistic $T_n = g(X_1, \ldots, X_n)$. Let $\mathbb{V}_F(T_n)$ denote the variance of T_n. We have added the subscript F to emphasize that the variance is a function of F. If we knew F we could, at least in principle, compute the variance. For example, if $T_n = n^{-1}\sum_{i=1}^{n} X_i$ then

$$\mathbb{V}_F(T_n) = \frac{\sigma^2}{n} = \frac{\int x^2 dF(x) - \left(\int x dF(x)\right)^2}{n}$$

which is clearly a function of F.
With the bootstrap, we estimate $\mathbb{V}_F(T_n)$ with $\mathbb{V}_{\widehat{F}_n}(T_n)$. In other words, we use a plug-in estimator of the variance. since, $\mathbb{V}_{\widehat{F}_n}(T_n)$ may be difficult to compute, we approximate it with a simulation estimate denoted by v_{boot}

Its important to know that The jackknife is a linear approximation of the bootstrap.

In the simulation, the Variance Bootstrap obtained from the uniform sample of the Order statistic is $v_{\text{boot}} = 3.5815\text{x}10^{-8}$, this indicates that almost all of the Order data values are nearly identical. Moreover, the theoretical bias and the Jackknife bias generated similar results, $b_{\text{jack}} = -1.6166\text{x}10^{-5}$ and $b_{\text{theoretical}} = -9.9990\text{x}10^{-5}$, which shows that the bias of the estimator is unbiased, hence the sampling distribution has a mean that is equal to the parameter being estimated.
It should be noted that the jackknife method is less computationally expensive, but the bootstrap has some statistical advantages, such as its simplicity to derive estimates of standard errors and confidence intervals for complex estimators of complex parameters of the distribution.

Exercise 14: Parametric Bootstrap vs Non-Parametric Bootstrap

Show the differences between the parametric and non-parametric bootstrap. Investigate robust versions of the bootstrap method and show examples about their performance.
Proof: Following the ideas from Wasserman (2006) and Kulperger (2019), Bootstrap is performed over a sample $X_1, ..., X_n$. In the case of the nonparametric Bootstrap, we generate new samples (resampling) based on the ECDF $\hat{F}_n$ of the sample; the generation of samples from the ECDF is equivalent to draw samples $X_1^j, ..., X_n^j$ from the original data **with replacement** for $j = 1, ..., N$, where N is the number of Bootstrap resamples. On the other hand, the parametric Bootstrap takes into consideration that the data comes from an specific distribution F_θ that, clearly, depends on an unknown parameter θ; instead of drawing from $\hat{F}_n$, we draw from $F_{\hat{\theta}}$, where $\hat{\theta}$ is an estimator of θ based on the sample. This method is just as accurates as the nonparametric, but under certain scenarios could not behave properly. An excellent example of the parametric and nonparametric Bootstrap is presented in point 11 of page 40 from Wasserman (2006).

Exercise 15: Mahalanobis Distance

Explain a nonparametric and robust way of calculating the Mahalanobis distance. Apply the technique to trim the bivariate sample of temperatures corresponding to the mean, hottest and coldest years. Compare with the cut obtained in the usual way. Then randomly choose 30 observations

and add a noise that comes from a normal with mean 10 and variance 0.25 and perform the exercise again.
Creating a robust non-parametric method of the Mahalanobis distance we made some changes based on the definition of the normal one, specifically in the covariance matrix $\sum$:

- Given $Cov(X,Y) = Ppearson(X,Y)S_X S_Y$, we change the correlation $Ppearson$ for its robust equivalent $PSpearman$
- Instead of using the estimators of the standard deviation S_X and S_Y, we use their robust equivalents of median absolute deviation (MAD), calculated as the median of the vector defined by $\{\mid X_n - \tilde{X} \mid\}$ where $\tilde{X}$ is the median of the vector $\{X_n\}$.
- Since the resulting covariance matrix $\tilde{\sum}$ may not be well conditioned, it implies that its inverse $\tilde{\sum}^{-1}$ may not exist or give values that do not correspond to reality (complex values matrix), so we change the inverse matrix to its inverse of Moore-Penrose $\tilde{\sum}^{+}$, which if $\tilde{\sum}$ is well conditioned is equal to its inverse $\tilde{\sum}^{-1}$

So, making this changes, we define the robust version of the Mahalanobis distance, the next way:

$$d^2(X) = [X - \tilde{X}]^T \tilde{\sum}^{-1} [X - \tilde{X}]$$

where $\tilde{\sum}$ is calculated the following way:

$$\tilde{\sum}_{ij} = p_{ij} mad_i mad_j$$

where p es the Spearman covariance and mad_i is the median absolute deviation for the variable t_i of each X_n
If we consider the bivariate sample of temperatures corresponding to the years, on average, hottest and coldest, and filter them with the traditional Mahalanobis distance, we see in figure 7, that the data labeled as outliers are only given by the traditional version of the Mahalanobis distance; the distance with the robust fit does not offer outliers, since it does not capture only the linear dependency, but also any relationship that exists between the two variables.

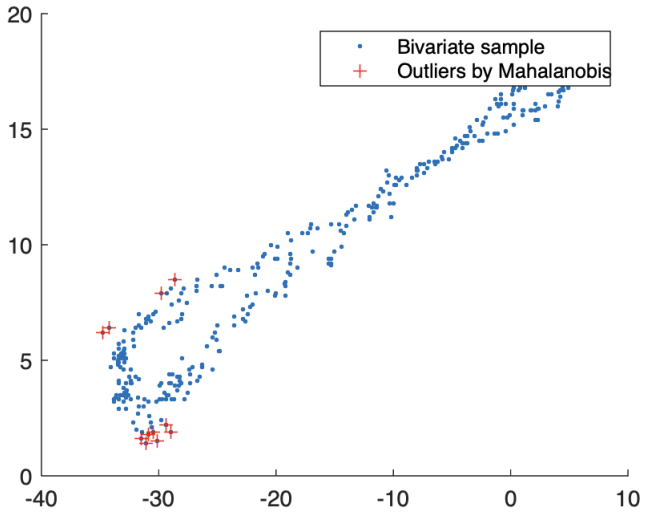

Figure 8: Scatter plot of the temperature for the hottest year and the coldest mean

However, if we contaminate the sample with a normal noise of parameters $\mu = 10$ and $\sigma^2 = 0.25$ in 30 observations, the result in figure 8, captures outliers in both ways.

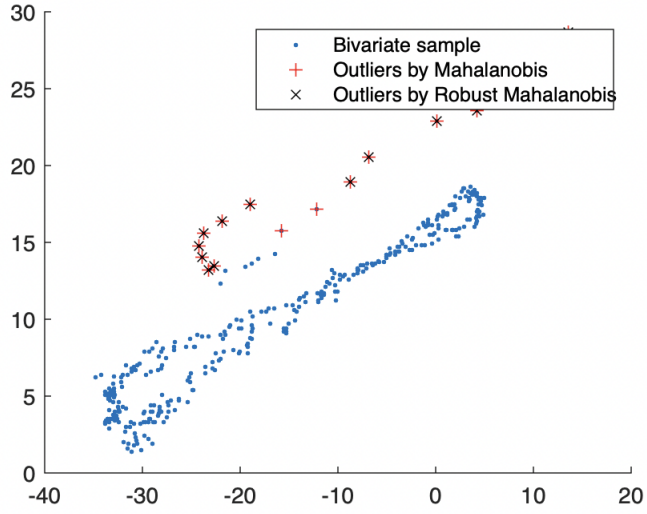

Figure 9: Contaminated sample

Exercise 16: Wasserman book exercises

Exercise 2. Page 10.

(Computer experiment). Compare the coverage and length of 1 and 2 by simulation. Take $p = 0.2$ and use $\alpha = 0.05$. Try various sample sizes n. How large must n be before the pointwise interval has accurate coverage? How do the lengths of the two intervals compare when this sample size is reached?

$$\hat{p}_n \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \tag{1}$$

$$\hat{p}_n \pm \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)} \tag{2}$$

where $\hat{p_n} = n^{-1}\sum_{i=1}^{n} X_i$ and $X_1, ..., X_n$ follows a Bernoulli(p) distribution. In 10 can be observed the amplitude of the confidence interval given by the two equations where can be seen that 1 decreases faster than 2. Now, it was generated for $n = 2000$ 1000 samples of this length and the confidence interval was calculated for each of them. Then, it was estimated the proportion of how many times the parameter was contained in the interval and it was obtained that 95.3% of the times, the parameter wherein the interval generated by 1 and 99.9% of the times in the intervals generated by 2, what makes sense since intervals of equation 1 are smaller.

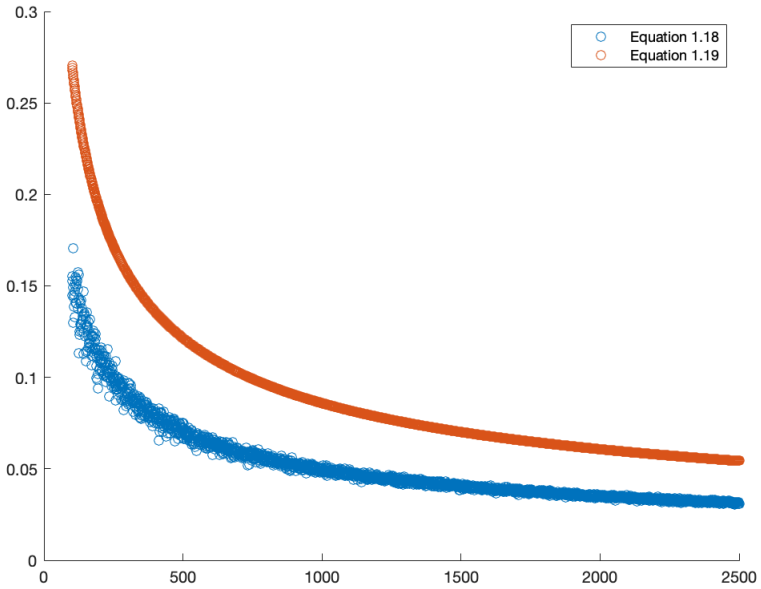

Figure 10: Amplitudes vs n

Exercise 3. Page 24.

(Computer experiment.) Generate 100 observations from a N(0,1) distribution. Compute a 95 percent confidence band for the CDF F. Repeat this 1000 times and see how often the confidence band contains the true distribution function. Repeat using data from a Cauchy distribution.

It was generated 1000 random variables of 100 observations each from a N(0, 1) distribution and cauchy distribution for the second case. Later, the times that the cumulative distribution function went out of the interval were counted to then obtain the proportion, finally this process was repeated 1000 times and the average of the proportions was obtained.

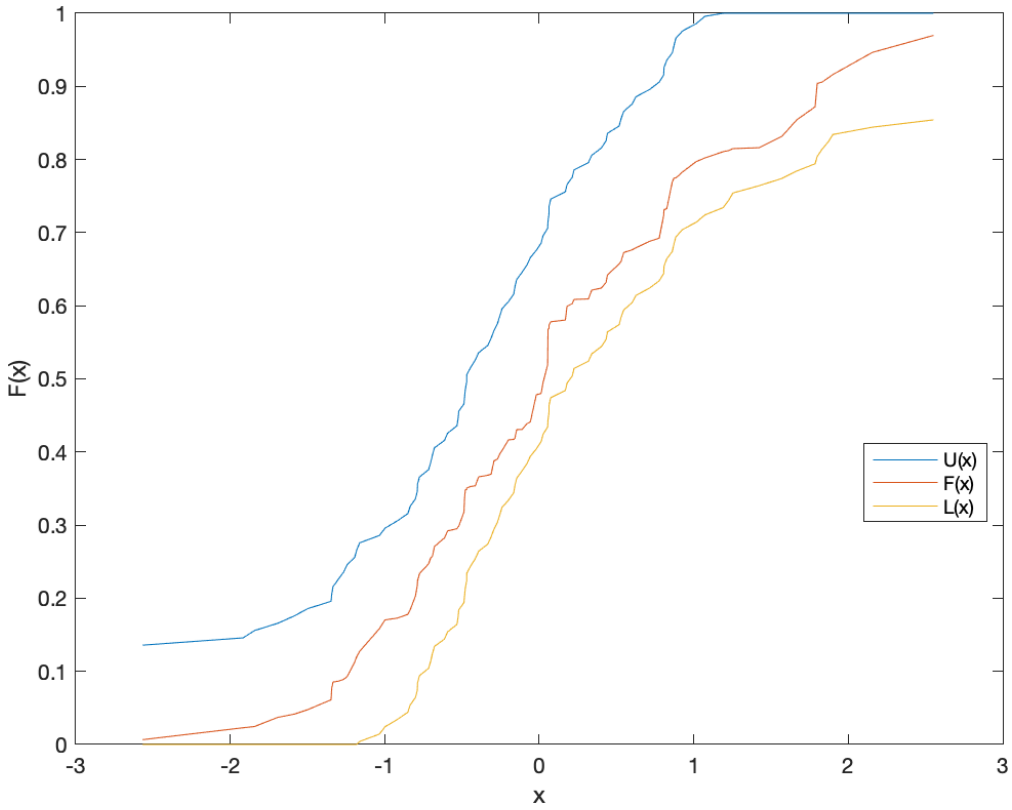

Figure 11: Graphic of the last simulation of normal values

97% of the time the confidence band contains the true distribution function.

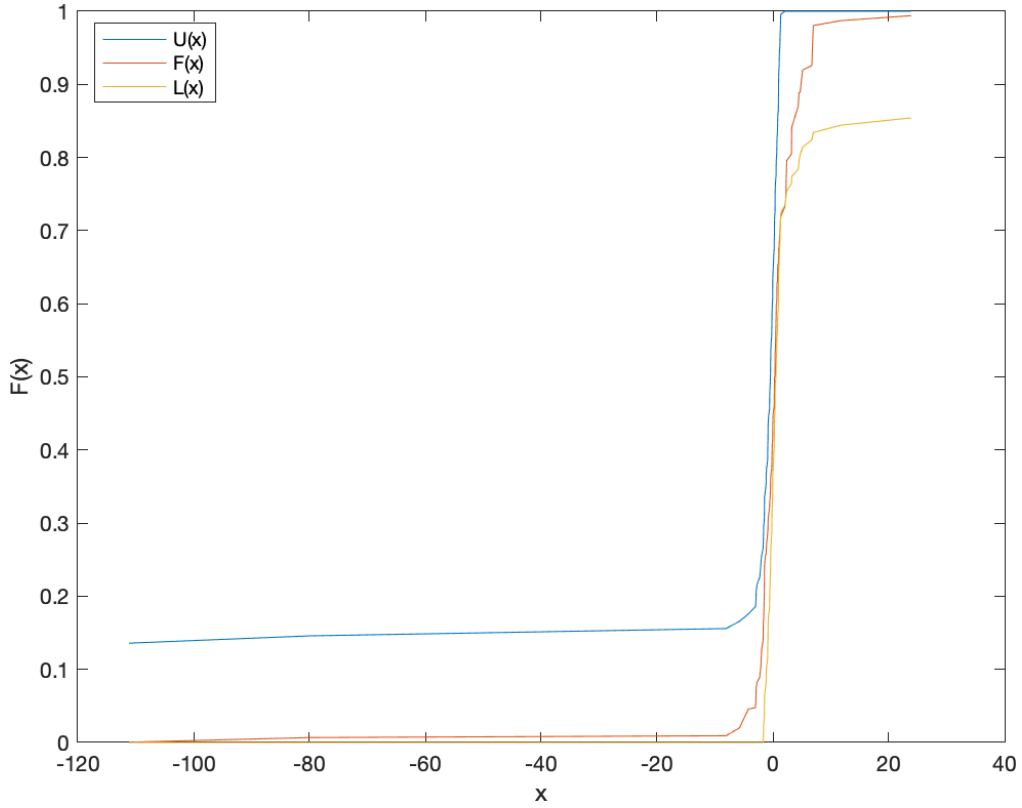

Figure 12: Graphic of the last simulation of cauchy values

96% of the time the confidence band contains the true distribution function.

Exercise 7. Page 24.

Let $X_1, ..., X_n \sim Bernoulli(p)$ and let $Y_1, ..., Y_m \sim Bernoulli(q)$. Find the plug-in estimator and estimated standard error for p. Find an approximate 90 percent confidence interval for p. Find the plug-in estimator and estimated standard error for p q. Find an approximate 90 percent confidence interval for p q.

Standard error plug-in estimator for p:

The plug-in estimator for a linear functional $T(F) = \int a(x)dF(x)$ is:

$$T(\hat{F}_n) = \int a(x)d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} a(X_i)$$

Now, taking into account that we are working wit Bernoulli(p) variables it is known that $E(x) = p$, so it will be used to estimated the parameter taking into account the above:

$$E(x) = \int x d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n$$

Now, is well know that to approximate the standard error of a plug-in estimator the influence function is used, which bring us to: $L(x) = x - p$ and $\hat{L}(x) = x - \overline{(X)}$.

Confidence interval for p

Using the Nonparametric Delta Method, we have:

$$\hat{se} = \sqrt{\frac{\hat{\Gamma}}{n}} \ \ and \ \ \hat{\Gamma^2} = \frac{1}{n}\sum_{i=1}^{n} \hat{L}^2(X_i)$$

Now, if we take into account what was obtained previously

$$\hat{se} = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X_n})^2}{n}}$$

For a 90 percent confidence interval we have an $\alpha = 0.1$, which brings us to an interval given by $\hat{p} \pm z_{\alpha/2}\hat{se}$ with $z_{\alpha/2} = P(Z \geq z) = \alpha/2 = -1.6449$ since Z is a variable that follows a standard normal. Finally

$$\bar{X} \pm (-1.6449)\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X_n})^2}{n^2}}$$

Estimator for p-q:

Since $Y_i \sim Bernoulli(q)$ from the above procedure it is known that $E(Y) = q$ and $\hat{q} = \bar{Y}$. Then, taking into account the linearity of the estimators:

$$p \hat{-} q = \int a(x)d(\hat{F_n}(x) - \hat{G_n}(y)) = \int x d\hat{F_n}(x) - \int y d\hat{G_n}(y) = \bar{X} - \bar{Y}$$

Confidence interval for p-q

Since $\hat{p}$ is independent of $\hat{q}$, then $V(\hat{p} - \hat{q}) = V(\hat{p}) - V(\hat{q})$, thus

$$\hat{se} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X_n})^2}{n^2} + \frac{\sum_{i=1}^{m}(Y_i - \bar{Y_n})^2}{m^2}}$$

Finally

$$(\bar{X} - \bar{Y}) \pm (-1.6449)\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X_n})^2}{n^2} + \frac{\sum_{i=1}^{m}(Y_i - \bar{Y_n})^2}{m^2}}$$

Exercise 2. Page 39.

The following data was provided in the exercise:
Given Z = (X,Y) where X = LSAT and Y = GPA. The correlation between X and Y is defined as follows:

$$\rho = T(F) = \frac{E[(X - \mu x)(Y - \mu y)]}{\sigma x \sigma y}$$

576	635	558	578	666	580	555	661
651	605	653	575	545	572	594	

Table 1: LSAT

3.39	3.30	2.81	3.03	3.44	3.07	3.00	3.43
3.36	3.13	3.12	2.74	2.76	2.88	3.96	

Table 2: GPA

Where F(x,y) is bivariate. The **plug-in** estimator for the correlation is:

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}}$$

Which is the **sample correlation**. The correlation of the given data is $\rho = 0.5449$. The Standar Error calculated using the influence function is (i) 0.1658, using Jackknife (ii) 0.1732 and using Bootstrap (iii) 0.1944. The SE calculated with the three methods do not differ a lot, and are low values for this plug-in estimator. The confidence interval for the estimator, using Bootstrap, is [0.1620, 0.8595].

Exercise 6. Page 40.

(Computer experiment.) Conduct a simulation to compare the four bootstrap confidence interval methods. Let n = 50 and let $T(F) = \int (x-\mu)^3 dF(x)/\sigma^3$ be the skewness. Draw $Y_1, ..., Y_n \sim N(0,1)$ and set $Xi = e^{Y_i}, i = 1, ..., n$. Construct the four types of bootstrap 95 percent intervals for T(F) from the data $X_1, ..., X_n$. Repeat this whole thing many times and estimate the true coverage of the four intervals.
The objective of this experiment is to compare the four different confidence intervals exposed for the Bootstrap method. The Bootstrap is performed over a sample of lognormally distributed data, taking 50 standard normal observations and making the exponentiation with the natural base. The statistic to be evaluated is the skewness of the data, which is well known to be positive for lognormal data. The results are presented in Table 3.

Method	**95% C.I.**
Normal	(1.23, 3.41)
Pivotal	(1.07, 3.22)
Studentized	(1.63, 3.57)
Percentile	(1.05, 3.20)

Table 3

Exercise 7. Page 40.

Let $X_1, ..., X_n \sim t_3$ where $n = 25$. Let $\theta = T(F) = \frac{(q.75 - q.25)}{1.34}$ where q_p denotes the p^{th} quantile. Do a simulation to compare the coverage and length of the following confidence intervals for θ ::

- Normal interval with standard error from the jackknife.
- Normal interval with standard error from the bootstrap.
- Bootstrap percentile interval.

Remark: The jackknife does not give a consistent estimator of the vari- ance of a quantile.
Let

$$X_1, \ldots, X_n \sim t_3$$

where $n = 25$ and t_3 is a student's t-distribution with $\nu = 3$. Let $\theta = T(F) = (q.75 - q.25)/1.34$. A simulation was implemented to compare the coverage and length of the following confidence intervals for θ: Normal interval with standard error from the bootstrap : [1.5486, 0.1564], bootstrap percentile interval:[0.3823, 1.3026] and a normal interval with standard error from the jackknife: [0.8141, 0.8905].
This results shows that the jackknife does not give a consistent estimator of the variance of a quantile, showing a very short coverage for the estimator, however, both bootstrap methods show a coverage and length very similar to each other, with the difference that the bootstrap percentile interval has a smaller lower bound, this is because it is a quantile method, With a statistics that have a small lower bound.

Exercise 10. Page 40.

(Computer experiment) Let $X_1, ..., X_n \sim \text{Normal}(\mu, 1)$.Let $\theta = e^{\mu}$ and let $\hat{\theta} = e^{\bar{X}}$ be the MLE. Create a data set (using μ=5) consisting of n=100 observations.

- Use the delta method, the parametric and the non-parametric bootstrap to get the se and 95 percent confidence interval for θ. Compare your answers.
- Plot a histogram of the bootstrap replications for the parametric and nonparametric bootstraps. These are estimates of the distribution and $\hat{\theta}$. The delta method also gives an approximation to this distribution, namely, Normal$(\hat{\theta}, \hat{se}^2)$. Compare these to the true sampling distribution of $\hat{\theta}$. Which approximation–parametric bootstrap, bootstrap or delta method–is closer to the true distribution?

The statistic is $\theta = e^{\mu} = e^5 = 148.41$. Approximating the delta method using Jackknife and with the sample of size 100 the SE is 17.0675, and with the non-parametric Bootstrap the SE is 16.8382. The confidence intervals at 95% confidence are respectively [128.5428, 184.8488] and [128.4849, 185.3347]

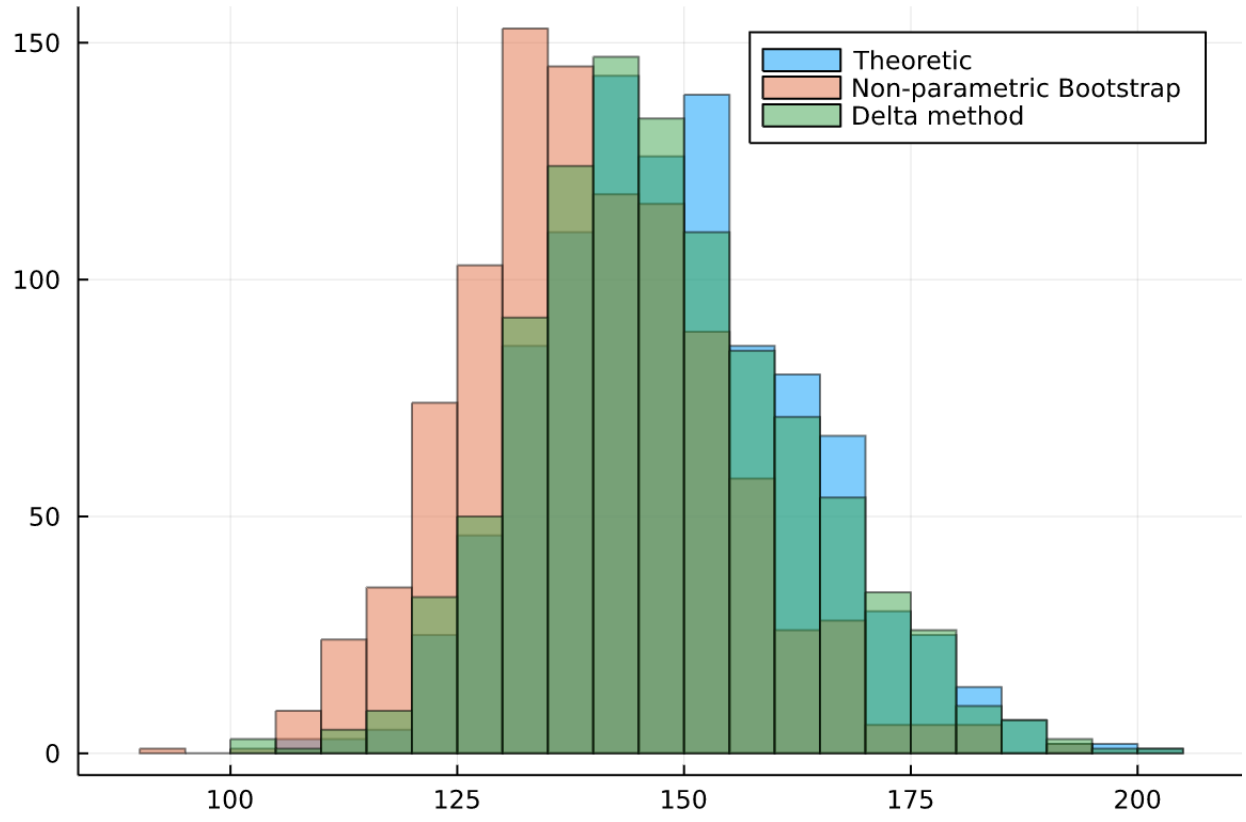

Figure 13: Histograms

As can be seen in the plot, both methods approximate the theoretical distribution similarly, but the Delta method is the closest one to the real distribution, as its peak does not have a significant bias, in contrast to the other ones.

Exercise 11. Page 41.

Let $X_1, ..., X_n \sim \text{Uniform}(0, 1)$, the maximum likelihood estimator (MLE) for $\theta = 1$ is $\hat{\theta} = X_{\max} = \max\{X_1, ..., X_n\}$. The distribution of $\hat{\theta}$ is a direct consequence of the proof made in Subsection **??** and it is given by

$$F(t) = t^n.$$

In Figure 14 shows the theoretical density of $X_{\max}$ and the approximation in histograms by parametric and nonparametric Bootstrap.

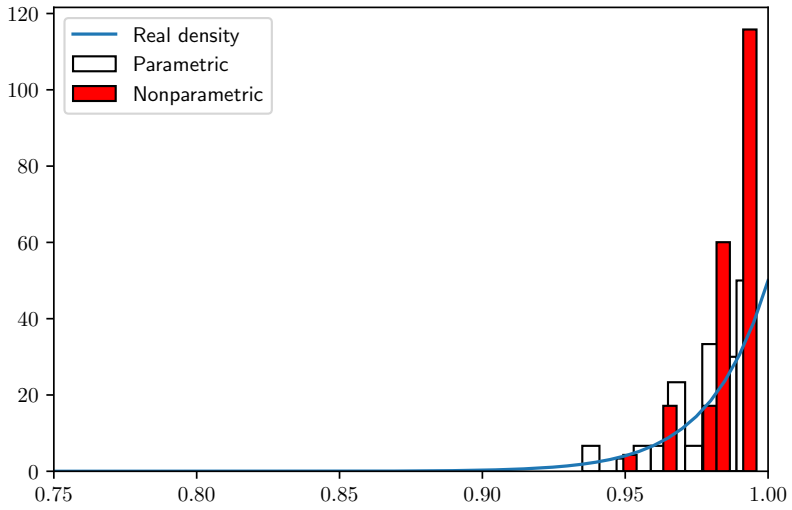

Figure 14: Density of $X_{\max}$, parametric and nonparametric histograms of $X_{\max}$.

Let $\hat{\theta}^*$ be the estimation of the parameter θ for each Bootstrap replication. For a parametric Bootstrap, one could draw samples from $F_{\hat{\theta}}$, instead of drawing samples from the empirial distribuion $\hat{F}_n$. Therefore, we would draw a N (number of Bootstrap iterations) new samples $X_1^j, ..., X_n^j \sim \text{Uniform}(0, X_{\max})$, for $j = 1, ..., N$. Clearly, $P(\hat{\theta}^* = \hat{\theta}) = 0$, since it is the probability of a point on a continuous random variable, which is obvious that it has a probability measure 0.
On the other hand, for a nonparametric Bootstrap, the procedure would be to draw new samples from the empirical distribution, i.e. simply draw samples with replacement from the initial sample $X_1, ..., X_n$ and calculate the new estimation. In this case, $P(\hat{\theta}^* = \hat{\theta}) \approx 1 - (1 - \frac{1}{n})^n$, since the event $\hat{\theta}^* = \hat{\theta}$ could be interpreted as the probability of $\hat{\theta}^*$ appearing at least once within the j-th resample for the nonparametric Bootstrap.
Now, in order to calculate the true probabilty of the event $\hat{\theta}^* = \hat{\theta}$, take the limit as $n \to \infty$. It is well known that $(1 - \frac{1}{n})^n \to e^{-1}$ as $n \to \infty$. Then, $P(\hat{\theta}^* = \hat{\theta}) = 1 - e^{-1} \approx 0.632$.

References

Brownlee, Jason. 2019. A Gentle Introduction to Jensen's Inequality.

Kulperger, R.J. 2019. Parametric and Nonparametric Bootstrap. *Handouts for Introduction to the Theory of Statistics*.

Mosteller, Frederick. 1946. On Some Useful "Inefficient" Statistics. *Handouts for Introduction to the Theory of Statistics*.

Rodgers, Joseph Lee. 1999. The Bootstrap, the Jackknife, and the Randomization Test: A Sampling Taxonomy. *Multivariate Behavioral Research*, **34**(4), 441–456.

Van der Vaart, Aad W. 2000. Asymptotic Statistics. **3**.

Wasserman, Larry. 2006. All of Nonparametric Statistics.