STATISTICAL REPORT

MONTHLY RAINFALL IN CEDUNA (AUSTRALIA) 1980-2023

DONE BY:
Fernando Pastor Peralta - 897113
Juanjo José Muñoz Lahoz - 902677
Marcos San Julián Fuertes - 899849

GROUP 7

EINA - COMPUTER ENGINEERING - COURSE 2023-24

INDEX:

1. Introduction.

In this work we will carry out a statistical report on the monthly rainfall data in mm collected at the Ceduna region meteorological station in Australia; we have data from January 1980 to December 2023. We will also carry out a simulation following a system of independent components provided by the tutor.For more details, see the job description.

2. Rain with all available data.

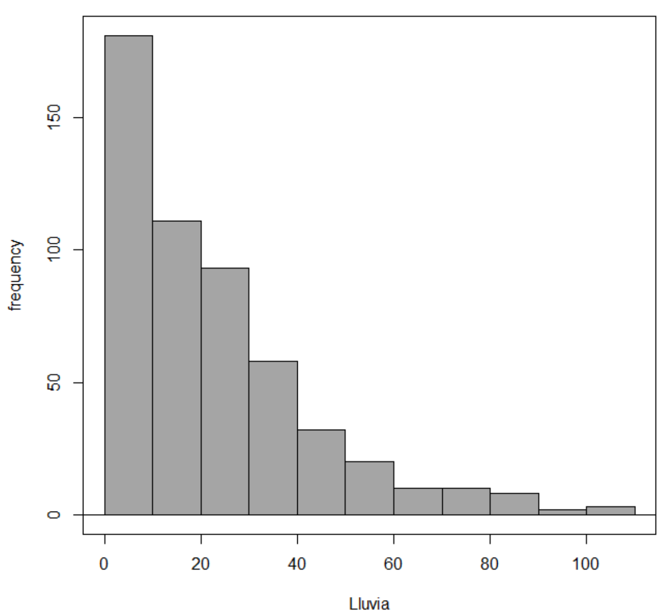
Next, we will begin exercise 1 and analyze the entire set of data we have and provide some general conclusions obtained from it.

2.1. Measures of location, dispersion and shape.

```
    mean       sd skewness 0%   25%  50%  75%  100%
22.73845 20.73703 1.432393  0 6.95 17.7 31.8 108.4
```

Data obtained:
- Media: 22.738.
- Standard deviation 20.737.
- Large positive skewness, meaning the data tends to accumulate to the left of the mean.
- More detailed quantile distribution in 2.3.

2.2. Histogram.

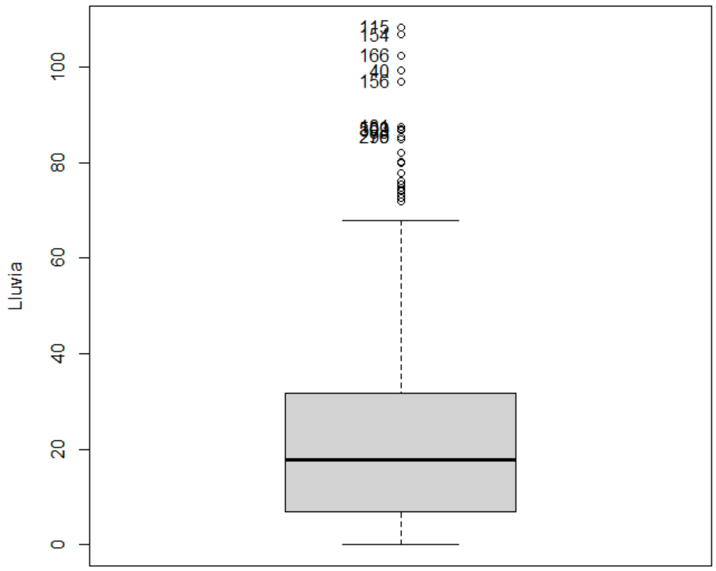


This is a histogram representing rainfall data for the Ceduna region of Australia. The data is distributed by month over 43 years, from 1980 to 2023.

The data on the y-axis, "frequency," represents the number of times the data on the x-axis, "rainfall," is repeated. The x-axis represents the volume of rainfall in the area throughout all the months of our dataset. It is noteworthy that the data drops exponentially.

We can observe that the vast majority of days have low rainfall and the days with the highest precipitation are not very abundant.

2.3. Box plot and outlier detection.



As can be seen in the histogram, the box plot shows that the data obtained exhibits positive skewness.

In addition, a significant number of outliers are visible that do not fall within the box plot. All of this indicates that the data does not clearly follow a normal distribution.

However, we will also perform the relevant normality tests to verify whether these observations are true.

2.4. Study of normality.

```
> normalityTest(~Lluvia, test="shapiro.test", data=Dataset)

        Shapiro-Wilk normality test

data:  Lluvia
W = 0.8702, p-value < 2.2e-16


> normalityTest(~Lluvia, test="lillie.test", data=Dataset)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Lluvia
D = 0.13643, p-value < 2.2e-16


> normalityTest(~Lluvia, test="ad.test", data=Dataset)

        Anderson-Darling normality test

data:  Lluvia
A = 17.438, p-value < 2.2e-16
```

It can be clearly observed thatthe p-valuesIn all cases they are less (by a lot) than 0.05, which confirms that the data ultimately do not follow a normal distribution.

2.5 Conclusions.

The entire dataset forms a positively skewed graph, as shown in the analysis of measures of location, dispersion, and shape, and as can also be seen graphically in the histogram and box plot. Furthermore, this is confirmed after conducting the normality study, where we have definitively concluded that our distribution does not follow a Gaussian distribution.

The positive skewness of our data leads us to conclude that most of the rainfall in the region is of a small volume most of the time.

3. Rain for months.

We will analyze the data by month and provide some general conclusions drawn from it.

3.1. Measures of location, dispersion and shape.

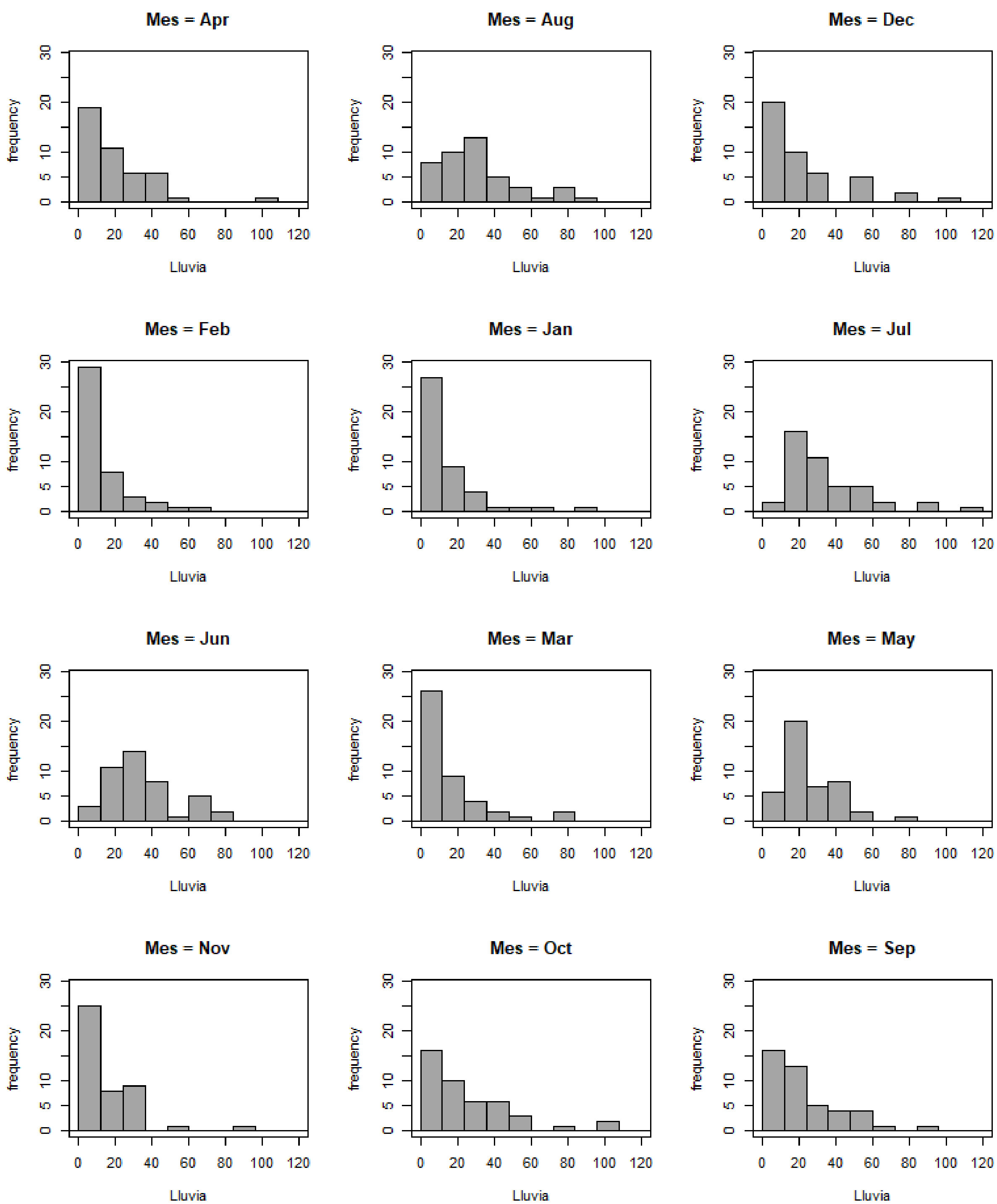
	mean	sd	skewness	0%	25%	50%	75%	100%
Apr	19.63864	19.17107	1.8840927	0.0	6.20	14.0	31.80	99.2
Aug	30.90000	21.30673	0.9320637	1.0	16.55	29.2	39.40	85.4
Dec	21.87273	23.77828	1.5425716	0.4	4.00	13.5	28.70	97.0
Feb	12.12727	14.91446	1.8726134	0.0	1.80	6.7	19.05	64.6
Jan	14.72273	18.35015	2.4592658	0.0	3.55	9.5	19.90	87.6
Jul	35.52273	22.00982	1.4187731	7.6	21.00	28.8	45.45	108.4
Jun	33.81591	18.37354	0.6332659	2.8	21.60	30.1	42.20	73.4
Mar	14.81364	18.58789	2.0338946	0.0	2.50	7.8	19.60	80.2
May	25.59773	14.91611	1.3704962	5.6	15.65	20.0	33.80	80.0
Nov	16.11364	15.99578	2.3268739	0.0	6.55	9.7	24.20	87.0
Oct	24.76818	25.19660	1.7077258	0.2	5.85	17.6	37.25	106.8
Sep	22.96818	19.21288	1.3546393	1.2	8.25	17.8	32.15	86.8

This data represents the data for each month of the year throughout all the years in which data has been obtained.

The highest average of all months is July, so that is the month in which it usually rains the most, although its standard deviation of data is among the highest, so the data obtained can vary considerably between different years.

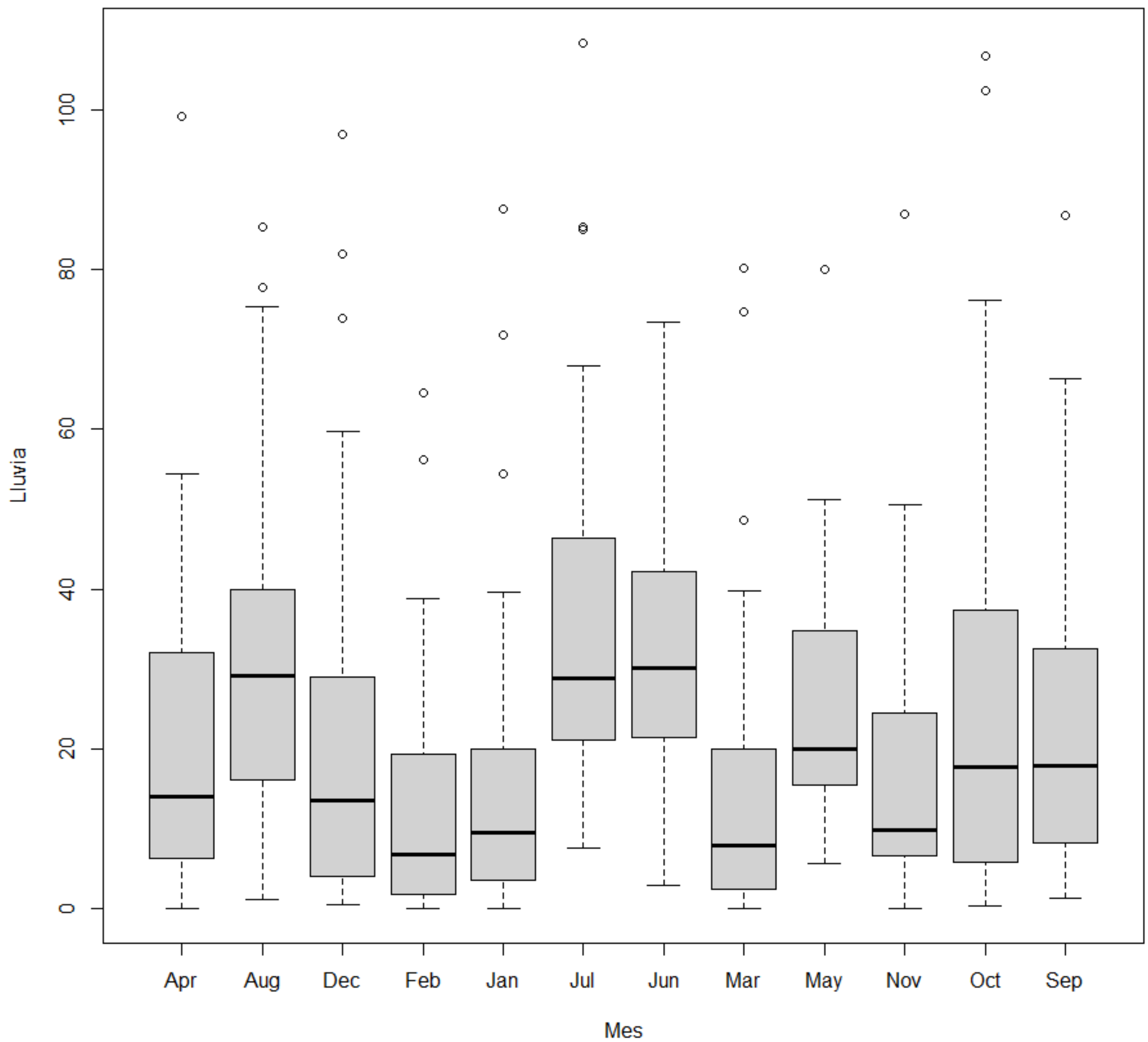
February is the month with the least rainfall on average, and it also has one of the lowest standard deviations. Finally, we can see that most months exhibit positive skewness, with the exception of June and August.

3.2. Histograms.



In the histograms of the different months we can observe that low rainfall predominates in most months, as well as positive skewness, with some exceptions such as August and June as previously mentioned that may approach a normal distribution.

3.3. Box plot and outlier detection.



These box plots show that all months exhibit high outliers and a tendency to accumulate many low rainfall values, with the exception of June, which shows no outliers. However, the summer months show a more homogeneous data distribution and a higher average rainfall, suggesting that this is the rainiest time of year. Finally, we can see that June is not only the only month without outliers, but its histogram also closely approximates a normal distribution, which we will verify below.

3.4. Study of normality.

3.4.1. Shapiro-Wilk test.

Abril	0.000016830705
Agosto	0.00501295
Diciem	0.000005340333
Febrero	0.000000954330
Enero	0.000000059789
Julio	0.00014213
Junio	0.02769845
Marzo	0.000000353199
Mayo	0.00042450
Noviem	0.000001417050
Octubre	0.000007958794
Septiem	0.00024544

In this case, most months do not follow a normal distribution, as we had already observed with the box plot data. We can also see that although August does not have a normal distribution, July appears to be close to it.

3.4.2. Test de Anderson-Darling.

Abril	0.00037767
Agosto	0.00813916
Diciem	0.0000003362471
Febrero	0.0000000889388
Enero	0.0000000014032
Julio	0.0000738150924
Junio	0.02921740
Marzo	0.0000000211375
Mayo	0.00104751
Noviem	0.0000144392175
Octubre	0.0000104844549
Sepiem	0.00038122

As in the Shapiro-Wilk test, no month has a normal distribution except June, which is quite close to the critical value of 0.05.

3.4.3. Kolmogorov-Smirnov test.

Abril	0.01147782
Agosto	0.02434657
Diciem	0.00061569
Febrero	0.0000035192
Enero	0.0000357057
Julio	0.00052977
Junio	0.33085955
Marzo	0.0000296884
Mayo	0.00142237
Noviem	0.00014451
Octubre	0.00422545
Septiem	0.00165958

As we have observed with the two previous tests, this test confirms that June does indeed approximate a normal distribution, while the rest of the data does not.

3.5. Conclusions.

As we have observed, the months with the highest average rainfall are the summer months of June, July, and August. Furthermore, these months also show the closest approximation to a normal distribution, especially June, which, according to normality tests, exhibits this distribution. Regarding the rest of the data, most show significant positive skewness and very low rainfall values. However, every single month has outliers except for June, which is unusual given the large number of years observed.

4. Rainfall per year.

We will analyze the data by year and provide some general conclusions drawn from it.

4.1. Measures of location, dispersion and shape.

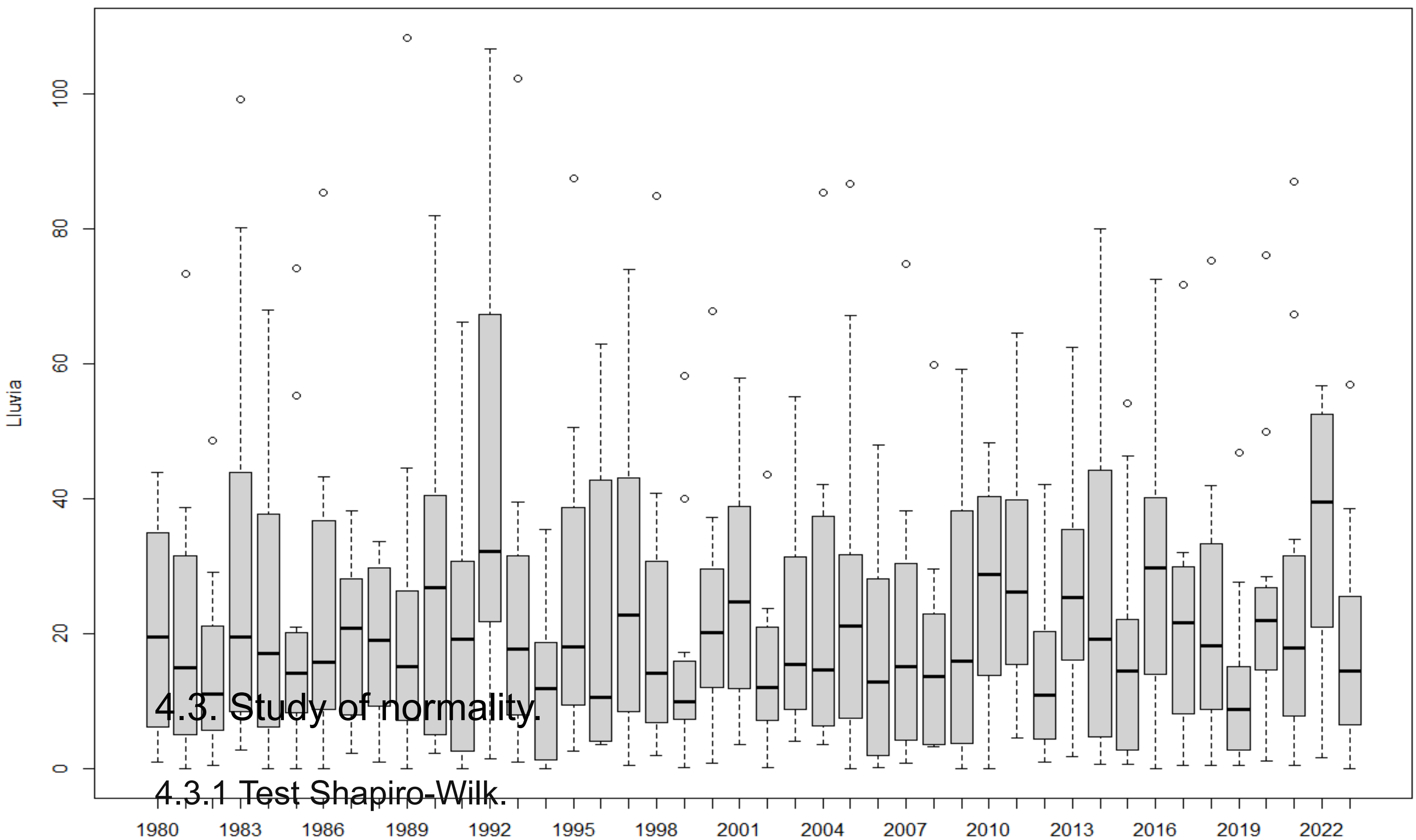
	mean	sd	skewness	0%	25%	50%	75%	100%
1980	20.66667	16.16468	0.32919316	1.0	6.35	19.5	32.150	44.0
1981	20.90000	21.05413	1.46528022	0.0	6.90	15.0	30.300	73.4
1982	15.46667	13.63267	1.36625621	0.4	5.85	11.1	20.900	48.6
1983	30.40000	31.93323	1.34977100	2.8	8.70	19.6	36.000	99.2
1984	23.00000	20.90463	1.03755247	0.0	6.25	17.1	36.100	68.0
1985	20.60000	22.07722	1.76958101	0.0	8.95	14.1	19.650	74.2
1986	23.50000	24.16075	1.65519039	0.0	8.90	15.8	36.200	85.4
1987	19.01667	12.02254	0.06691601	2.2	8.90	20.8	27.850	38.2
1988	18.75000	10.99888	-0.13717836	1.0	10.25	19.1	29.700	33.6
1989	23.38333	29.47091	2.51546438	0.0	7.15	15.2	25.800	108.4
1990	28.01667	24.27007	0.95106372	2.2	6.05	26.9	36.650	82.0
1991	19.85000	19.53070	1.13442560	0.0	3.10	19.2	29.500	66.2
1992	43.51667	33.97705	0.84344428	1.4	22.90	32.2	62.050	106.8
1993	25.41667	27.39057	2.21170996	1.0	10.70	17.7	30.550	102.4
1994	12.80000	11.84629	0.69667265	0.0	1.65	11.8	17.750	35.4
1995	25.88333	24.66112	1.57503549	2.6	10.80	18.0	38.650	87.6
1996	21.20000	21.73259	0.95711183	3.6	4.15	10.6	41.300	63.0
1997	27.98333	24.55279	0.82679015	0.4	9.60	22.8	39.050	74.0
1998	22.08333	23.30953	2.00139822	2.0	7.35	14.1	30.500	85.0
1999	16.05000	16.55429	1.94918206	0.2	7.45	9.9	15.400	58.2
2000	23.15000	17.47832	1.50627917	0.8	14.00	20.2	28.200	67.8
2001	26.75000	17.30415	0.49961552	3.6	12.35	24.7	34.850	58.0
2002	14.28333	11.90186	1.34577105	0.2	7.35	12.1	20.400	43.6
2003	20.68333	15.63178	1.01386788	4.0	9.15	15.4	30.600	55.2
2004	23.56667	24.07471	1.71056947	3.6	7.15	14.6	35.350	85.4
2005	25.98333	26.62582	1.41590686	0.0	8.00	21.1	31.550	86.8
2006	16.23333	15.45523	0.76242936	0.2	2.00	12.8	27.950	48.0
2007	19.83333	21.39581	1.68961288	0.8	4.90	15.2	28.700	74.8
2008	17.23333	16.32841	1.70764247	3.2	3.60	13.7	22.700	59.8
2009	20.55000	19.28525	0.79020872	0.0	3.75	15.9	37.550	59.2
2010	26.86667	15.40775	-0.37331604	0.0	15.25	28.8	39.450	48.4
2011	28.30000	17.20771	0.58033651	4.6	18.60	26.2	39.300	64.6
2012	13.53333	11.88837	1.28805616	1.0	5.10	10.9	20.250	42.2
2013	27.90000	18.39768	0.62353084	1.8	18.25	25.3	33.925	62.5
2014	26.73333	26.79842	0.88763600	0.6	6.55	19.2	38.350	80.0
2015	17.00000	17.63550	1.21821950	0.6	3.30	14.5	20.800	54.2
2016	30.31667	21.22065	0.57220733	0.0	15.00	29.7	39.800	72.6
2017	22.28333	19.10116	1.53386833	0.4	8.95	21.7	29.200	71.8
2018	23.86667	20.81928	1.40881593	0.4	10.45	18.2	31.100	75.4
2019	12.14167	13.50774	1.79266236	0.4	2.85	8.7	14.150	46.9
2020	25.40000	20.07169	1.64695371	1.2	14.75	21.9	26.150	76.2
2021	25.58333	26.57346	1.50274894	0.4	10.60	17.9	30.250	87.0
2022	35.53333	18.41192	-0.48623317	1.6	21.00	39.5	51.550	56.8
2023	18.28333	16.52116	1.28363412	0.0	6.55	14.4	24.550	57.0

As in previous instances, the data for this section shows the mean, standard deviation, skewness, and quartiles, this time for all the years included in our data. Let's briefly analyze the most interesting parts:

The years with the highest rainfall were 1992, 2022, 1983, and 2016, with averages of 43.517, 35.533, 30.400, and 30.317 mm, respectively. It is also worth noting that 1992 and 1983 are the years with the greatest data dispersion, as their standard deviations are 33.977 and 31.933 mm, respectively.

Regarding symmetry, the years with the greatest positive skewness were 1989 with 2.515 and 1993 with 2.211. On the other hand, negative skewness is very scarce and practically imperceptible; it is only noticeable in the years 1988, 2010, and 2022 with their negative values. However, since these values are so close to zero, we can consider them more symmetrical than asymmetrical.

4.2. Box plot and outlier detection.



4.3. Study of normality.

4.3.1 Test Shapiro-Wilk.

1980	0.12082324	1.000000	2002	0.13837552	1.000000
1981	0.05766582	1.000000	2003	0.13900060	1.000000
1982	0.05519844	1.000000	2004	0.00979835	0.362539
1983	0.00946667	0.359733	2005	0.02585015	0.788509
1984	0.13179326	1.000000	2006	0.13804607	1.000000
1985	0.00429135	0.175946	2007	0.01767605	0.583310
1986	0.02397589	0.767229	2008	0.01210644	0.423725
1987	0.66702600	1.000000	2009	0.11991235	1.000000

Those p-values (left column) that are greater than 0.05 will pass the Shapiro-Wilk test.

The values marked in yellow indicate the years in which the test has not been passed, and those that have not been underlined are those that are greater than 0.05 and therefore have passed the test.

4.3.2. Test Anderson-Darling.

1980	0.16534357	1.000000	2002	0.27361073	1.000000
1981	0.27014521	1.000000	2003	0.31687697	1.000000
1982	0.07180980	1.000000	2004	0.18494316	1.000000
1983	0.01005511	0.412259	2005	0.04693147	1.000000
1984	0.06592859	1.000000	2006	0.24331257	1.000000
1985	0.00094234	0.041463	2007	0.11411220	1.000000
1986	0.07706579	1.000000	2008	0.23544037	1.000000
1987	0.94770078	1.000000	2009	0.18840685	1.000000
1988	0.42704868	1.000000	2010	0.91325404	1.000000
1989	0.01182211	0.472884	2011	0.85363982	1.000000
1990	0.42235495	1.000000	2012	0.46228299	1.000000
1991	0.37074700	1.000000	2013	0.60732244	1.000000
1992	0.04499732	1.000000	2014	0.24768430	1.000000
1993	0.11548412	1.000000	2015	0.23004222	1.000000
1994	0.61787014	1.000000	2016	0.69038212	1.000000
1995	0.06451772	1.000000	2017	0.10448936	1.000000
1996	0.00186003	0.079981	2018	0.32291152	1.000000
1997	0.47131908	1.000000	2019	0.11614979	1.000000
1998	0.14008629	1.000000	2020	0.01325946	0.517119
1999	0.00284302	0.119407	2021	0.09068960	1.000000
2000	0.43929046	1.000000	2022	0.43132831	1.000000
2001	0.56513699	1.000000	2023	0.69803550	1.000000

Again, the values in the left column are the p-values, which tell us whether the year passes the test or not, using the same criteria as before.

We can observe that the values that failed the test are the same as in the previous test. However, there are years that previously failed the test but now pass (1995 and 2017).

4.3.3. Kolmogorov-Smirnov test.

As in previous tests, we will follow the same guidelines.

The years that have not passed the tests have been the same as in previous cases and 1992 has been added to the years not passed. On the other hand, the years 1986, 1993, 1998, 2004, 2007, 2008, 2015, 2019 and 2021 have passed

1980	0.1788490	1.000000	2002	0.2227005	1.000000
1981	0.1309824	1.000000	2003	0.1739469	1.000000
1982	0.0709895	1.000000	2004	0.0192810	0.674834
1983	0.0064627	0.258508	2005	0.0307371	1.000000
1984	0.1345436	1.000000	2006	0.1740284	1.000000
1985	0.0022535	0.094646	2007	0.0394310	1.000000
1986	0.0438604	1.000000	2008	0.0323682	1.000000
1987	0.7688647	1.000000	2009	0.1229812	1.000000
1988	0.5734565	1.000000	2010	0.7793019	1.000000
1989	0.0009151	0.040264	2011	0.7018485	1.000000
1990	0.2906220	1.000000	2012	0.2145783	1.000000
1991	0.1444452	1.000000	2013	0.5045469	1.000000
1992	0.1217110	1.000000	2014	0.1021181	1.000000
1993	0.0084439	0.329310	2015	0.0375097	1.000000
1994	0.3299418	1.000000	2016	0.5412210	1.000000
1995	0.0501748	1.000000	2017	0.1098439	1.000000
1996	0.0023978	0.098309	2018	0.1917782	1.000000
1997	0.2003331	1.000000	2019	0.0170257	0.629950
1998	0.0090622	0.344363	2020	0.0199242	0.677423
1999	0.0010205	0.043883	2021	0.0173887	0.629950
2000	0.1478994	1.000000	2022	0.2592887	1.000000
2001	0.5106411	1.000000	2023	0.2065628	1.000000

the test when in previous cases they had not passed it.

4.4. Conclusions.

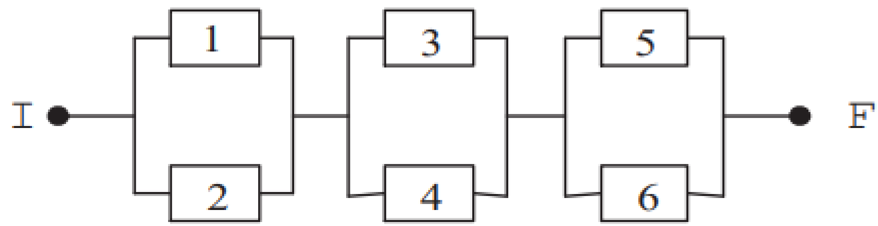
We have conducted a study of location, dispersion and shape that has helped us to know data on the mean, standard deviation, skewness and quartiles of the years, data that have been confirmed and seen more visually in the box plot.

We have also studied the normality of the data,Therefore, the years that will possibly follow a normal distribution will only be those that have passed the three tests, which are: 1980, 1981, 1982, 1984, 1987, 1988, 1990, 1991, 1994, 1997, 2000, 2001, 2002, 2003, 2006, 2009, 2010, 2011, 2012, 2013, 2014, 2016, 2018, 2022 and 2023.

5. Independent component systems.

In this part we begin exercise 2, in which we will solve a system of independent components obtaining different values.

5.1. System.



To solve this circuit, we can divide it into the three clearest groups: components 1 and 2, which are in parallel, as well as components 3 and 4, and components 5 and 6. Finally, we just need to connect these three blocks in series.

5.2 Approximations of the mean, variance, and standard deviation of the variableT.

First, we will calculate the parameters approximately using the program "R Commander". The following code is used for circuit simulation. Note that the operating time of each circuit component is a random variable that follows an exponential distribution with parameter 1/(2*i), where i is the component number in each case, and that the variable n in the code varies depending on the chosen sample size (see table):

```
rm(list=ls())
n<-500000
X1<-rexp(n,1/(2*1))
X2<-rexp(n,1/(2*2))
X3<-rexp(n,1/(2*3))
X4<-rexp(n,1/(2*4))
X5<-rexp(n,1/(2*5))
X6<-rexp(n,1/(2*6))
X12<-pmax(X1,X2)
X34<-pmax(X3,X4)
X56<-pmax(X5,X6)
X1234<-pmin(X12,X34)
T<-pmin(X1234,X56)
muT<-mean(T)
wasT<-was(T)
desT<-sqrt(varT)
where; where; where
```

N	E[T]	Was[T]	sqrt(Var[T])
10000	3.484504	6.11107	2.472058
50000	3.517013	6.286436	2.507277
100000	3.505085	6.331292	2.516206
500000	3.513901	6.315292	2.513024

5.3 Exact value of the mean, variance and standard deviation of the variable T.

For the exact calculation of the parameters, we will calculate the distribution function of the variable T using the distribution functions of the different circuit components. Subsequently, we will find the probability density function by taking the derivative of the distribution function and follow the definitions of E[T] and Var[T] to calculate the parameters. The program to be used will be SageMath due to its simplicity and our familiarity with it. The calculation procedure is as follows:

```
[15]: Par(x) = 1/(2*x)

[16]: X1 = Par(1)
      X2 = Par(2)
      X3 = Par(3)
      X4 = Par(4)
      X5 = Par(5)
      X6 = Par(6)

[17]: X12 = ( ( 1-e^(-Par(1)*x ) ) * ( 1-e^( -Par(2)*x ) ) )

[18]: X34 = ( ( 1-e^(-Par(3)*x ) ) * ( 1-e^( -Par(4)*x ) ) )

[19]: X56 = ( ( 1-e^(-Par(5)*x ) ) * ( 1-e^( -Par(6)*x ) ) )

[20]: X1234 = 1-( ( 1-X12 ) * ( 1-X34 ) )
```

```
[21]: DistT = 1-( ( 1-X1234 ) * ( 1-X56 ) )

[22]: DensT = diff(DistT,x)

[23]: EX = integrate(x*DensT,x,0,oo)
      EX.n()

[23]: 3.51190505179759

[24]: VarX1 = integrate(x*x*DensT,x,0,oo)

[25]: VarX2 = EX*EX

[26]: VarX = VarX1-VarX2
      VarX.n()

[26]: 6.30021542873418

[27]: DesTX = sqrt(VarX)
      DesTX.n()

[27]: 2.51002299366643
```

Therefore, the exact values of the parameters to be calculated are:

- E[X] = 3.51190505179759
- Where[x] =6.30021542873418
- sqrt(Var[X]) = 2.51002299366643

5.4. Conclusions.

In short, to calculate the function that defines the variable T, it is necessary to segment the circuit to be able to calculate the function in parts. Once the function is obtained, it is easy to calculate, both approximately and exactly, the values of the mean, variance, and standard deviation, applying their theoretical definitions.