Lecture Notes EMSE 4765:
Data Analysis - Statistics Review

Chapter 16: Exploratory Data Analysis: Numerical Summaries

Version: 1/19/2021

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Text Book: A Modern Introduction to Probability and Statistics, Understanding Why and How

By: F.M. Dekking. C. Kraaikamp, H.P.Lopuhaä and L.E. Meester

16 Exploratory Data Analysis: Numerical Summaries

16.0 Introduction . . .

- The classical way to describe important features of a dataset is **to give several numerical summaries** for:
  1) **The center of a dataset**
  2) **The amount of variability among the elements of a dataset,**
  3) **The quantiles for a dataset.**

- To distinguish them **from corresponding notions for probability distributions of random variables**, one adds the word **"sample"** or **"empirical"** ; For example:
  1) **The sample mean of a dataset**
  2) **The sample variance of a dataset,**
  3) **The empirical quantiles for a dataset.**

- **The boxplot**, a classical **graphical display** of some of these numerical summaries.

16 Exploratory Data Analysis: Numerical Summaries
16.1 The center of a data set . . .

- The best-known method to identify the center of a dataset $(x_1, \ldots, x_n)$ is to compute **the sample mean** :

$$\overline{x}_n = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

One often drops the subscripts n and simply writes $\overline{x}$ (read "x-bar").

Example: Wick Temperature Data

The following dataset consists of hourly temperatures in degrees Fahrenheit (rounded to the nearest integer), recorded at Wick in northern Scotland from 5 p.m. December 31, 1960, to 3 a.m. January 1, 1961.

43	43	41	41	41	42	43	58	58	41	41

Source: V. Barnett and T. Lewis. *Outliers in statistical data.* Third edition, 1994. © John Wiley & Sons Limited. Reproduced with permission.

16 Exploratory Data Analysis: Numerical Summaries

16.1 The center of a data set . . .

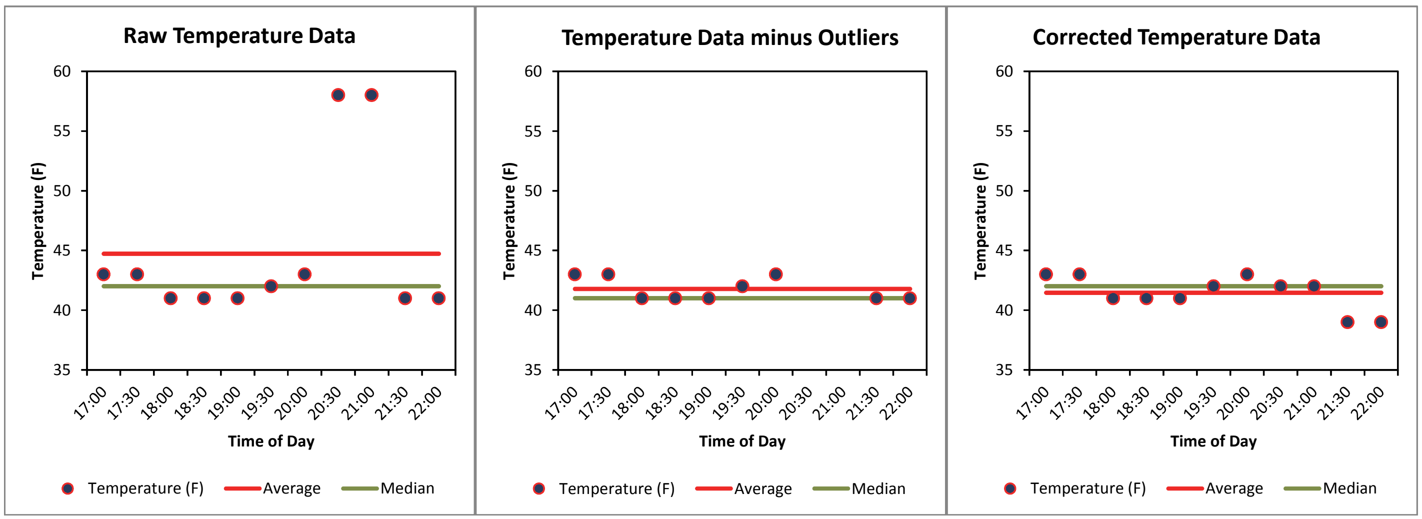
- The sample median $Med(x_1, x_2, \ldots, x_n)$ or Med_n is **the middle element of the data set when n is odd** and **the average of the two middle ones when n is even.**

Row	Time of Day	A Temperature (F)	B Ordered Column A	C Collumn A Minus utliers	D Ordered Column C	E Correctled Column A	F Ordered Column E
1	17:00	43	41	43	41	43	39
2	17:30	43	41	43	41	43	39
3	18:00	41	41	41	41	41	41
4	18:30	41	41	41	41	41	41
5	19:00	41	41	41	41	41	41
6	19:30	42	42	42	42	42	42
7	20:00	43	43	43	43	43	42
8	20:30	58	43		43	42	42
9	21:00	58	43		43	42	43
10	21:30	41	58	41		39	43
11	22:00	41	58	41		39	43
Sample Mean		44.73		41.78		41.45	
Sample Median		42		41		42	
Sample St. Dev		6.62		0.97		1.44	

16 Exploratory Data Analysis: Numerical Summaries

16.1 The center of a data set . . .

- The sample mean $\overline{x}$ is the natural analogue for $E[X]$ where X is a random variable pdf $f(x)$. **However, the sample mean $\overline{x}$ is very sensitive to outliers, whereas the sample median Med_n is not.**

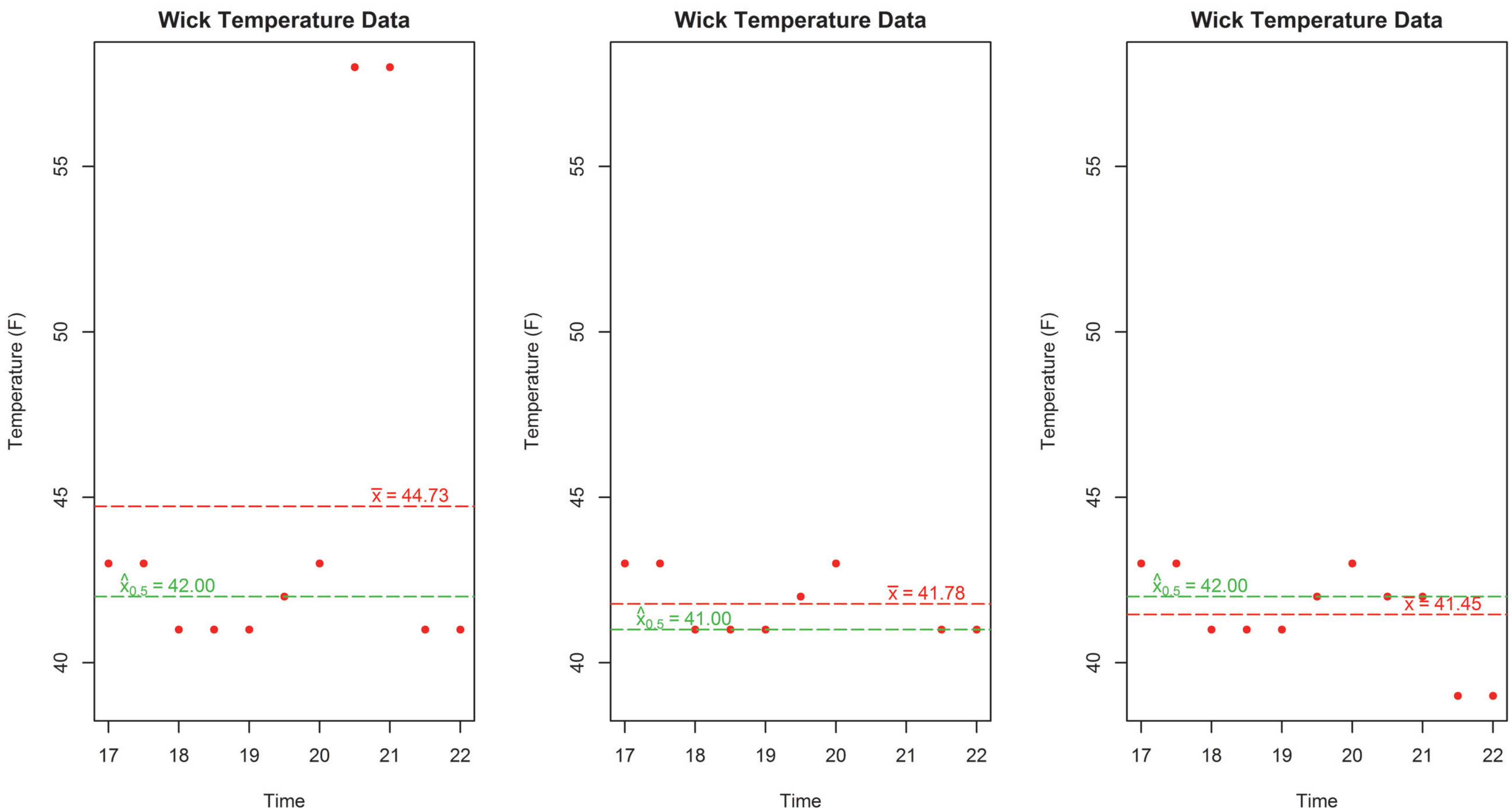


- **By no means should one leave out measurements that deviate a lot from the bulk of the data! One should be aware of outliers and only correct them if one concludes an error occurred in the data recorded process.**

16 Exploratory Data Analysis: Numerical Summaries

16.1 The center of a data set . . .

Same analysis in file "Wick_Temperature.R"



- To quantify the amount of variability among the elements of a dataset $(x_1, x_2, \ldots, x_n)$, one often uses the sample variance defined by :

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$$

- Up to a scaling factor this is equal to **the average squared deviation from $\overline{x}_n$.** At first sight, it seems more natural to define the sample variance by :

$$\widetilde{s}\,_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 = \frac{n-1}{n} \times s_n^2$$

In Chapter 19, it is explained why one prefers s_n^2 over $\widetilde{s}\,_n^2$.

16 Exploratory Data Analysis: Numerical Summaries

16.2 The variability of a data set . . .

- Because s_n^2 is in different units from the elements of the dataset, one often evaluates instead **the sample standard deviation :**

$$s_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}$$

		A	B	C	D	E	F
Row	**Time of Day**	**Temperature (F)**	**Ordered Column A**	**Column A minus Outliers**	**Ordered Column C**	**Corrected Column A**	**Ordered Column E**
1	17:00	43	41	43	41	43	39
2	17:30	43	41	43	41	43	39
3	18:00	41	41	41	41	41	41
4	18:30	41	41	41	41	41	41
5	19:00	41	41	41	41	41	41
6	19:30	42	42	42	42	42	42
7	20:00	43	43	43	43	43	42
8	20:30	58	43		43	42	42
9	21:00	58	43		43	42	43
10	21:30	41	58	41		39	43
11	22:00	41	58	41		39	43
Sample Mean		**44.73**		**41.78**		**41.45**	
Sample Median		**42**		**41**		**42**	
Sample St. Dev.		$\boxed{\textbf{6.62}}$		$\boxed{\textbf{0.97}}$		$\boxed{\textbf{1.44}}$	

16.2 The variability of a data set . . .

- Just as the sample mean, **the sample standard deviation is very sensitive to outliers.**

- **A more robust measure of variability** is **the median of absolute deviations or MAD;** First evaluate Med_n, second evaluate

$$|x_i - Med_n|,\ i = 1, \ldots, n$$

Third, set :

$$MAD(x_1, \ldots, x_n) = Med(|x_1 - Med_n|,\ |x_2 - Med_n|,\ \ldots, |x_2 - Med_n|)$$

- **Just as the sample median, the MAD is hardly affected by outliers.**

- Just as the sample median, the MAD is hardly affected by outliers.

		A	B	C	D	E	F
Row	Time of Day	|Temperature (F) - Sample Median|	Ordered Column A	|Temperature (F) - Sample Median|	Ordered Column C	|Temperature (F) - Sample Median|	Ordered Column E
1	17:00	1	0	2	0	1	0
2	17:00	1	1	2	0	1	0
3	17:00	1	1	0	0	1	0
4	17:00	1	1	0	0	1	1
5	17:00	1	1	0	0	1	1
6	17:00	0	1	1	1	0	1
7	17:00	1	1	2	2	1	1
8	17:00	16	1		2	0	1
9	17:00	16	1		2	0	1
10	17:00	1	16	0		3	3
11	17:00	1	16	0		3	3
	MAD	1		0		1	

- **The sample median** divides the dataset in two more or less equal parts: about **half of the elements are less than the median** and about **half of the elements are greater than the median**.

- **The pth empirical quantile is denoted by $q_n(p)$: It divides the dataset in two parts** in such a way that **a proportion p is less than a certain number** and **a proportion $1 - p$ is greater than this number.**

- **The order statistics** consist of the same elements as in the original dataset $x_1, x_2, \ldots, x_n$, but in ascending order. Denote by $x_{(k)}$ the kth element in the ordered list. **Then**

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n-1)} \leq x_{(n)}.$$

are called the order statistics of $x_1, x_2, \ldots, x_n$.

- One often sets $x_{(0)} \equiv 0$ and $x_{(n+1)} \equiv \infty$ for positive RV's.

Example: Wick Temperature Data

	A	B	C	D	E	F
Time of Day	Temperature (F)	Ordered Collumn A	Collumn A minus Outliers	Ordered Collumn C	Corrected Collumn A	Ordered Collumn E
17:00	43	41	43	41	43	38
17:30	43	41	43	41	43	39
18:00	41	41	41	41	41	41
18:30	41	41	41	41	41	41
19:00	41	41	41	41	41	41
19:30	42	42	42	42	42	42
20:00	43	43	43	43	43	42
20:30	58	43		43	42	42
21:00	58	43		43	42	43
21:30	41	58	41		39	43
22:00	41	58	41		38	43
Average		44.73		41.78		41.36
Median		42		41		42

Order Statistics Wick Temperature Data

- Note that by putting the elements in order, it is possible that successive order statistics are the same, for instance, $x_{(1)} = x_{(2)} = \ldots = x_{(5)} = 41$.

- **To compute empirical quantiles** one **linearly interpolates** between order statistics of the dataset. Let $0 \leq p \leq 1$, and suppose for a data set of size n,

$$\frac{k}{n} \leq p \leq \frac{k+1}{n} \Leftrightarrow k \leq np \leq k+1$$

Then one evaluates the $q_n(p)$ as follows from $x_{(k)}$ and $x_{(k+1)}$:

$$q_n(p) = \alpha[x_{(k+1)} - x_{(k)}] + x_{(k)}, \text{ where } \alpha = np - k$$

Exercise: Compute $q_{11}(0.55)$ for the Wick temperature data.

Answer: We have $n = 11, p = 0.55$ and $np = 6.05 \Rightarrow 6 \leq np < 7$ and thus $k = 6$. We have $\alpha = 6.05 - 6 = 0.05$. With $x_{(6)} = 42$ and $x_{(7)} = 43$ it follows that

$$q_{11}(0.55) = 0.05 \cdot (43 - 42) + 42 = 42.05$$

Analysis in file "Empirical_Quantile_Example.R"

Example data set for positive continuous random variable X :

$$(x_{(1)}, x_{(2)}, \ldots, x_{(10)}) = (6, 8, 9, 10, 11, 12, 14, 16, 19, 24)$$

Evaluate $q_{10}(0.85)$: Set $x_{(0)} \equiv 0$.

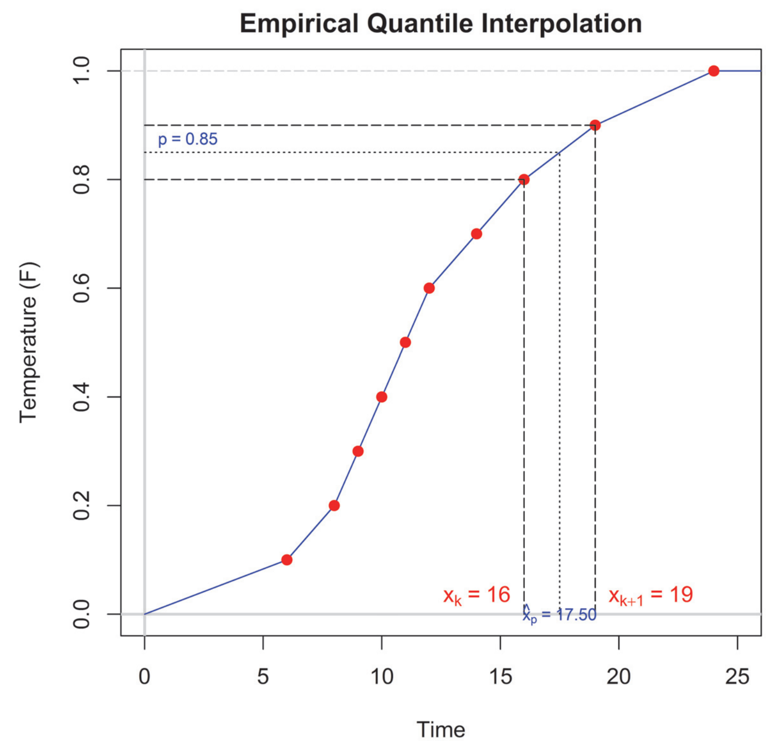
$$n = 10, p = 0.85 \text{ and } np = 8.5 \Rightarrow$$
$$8 \leq np < 9 \text{ and thus } k = 8.$$

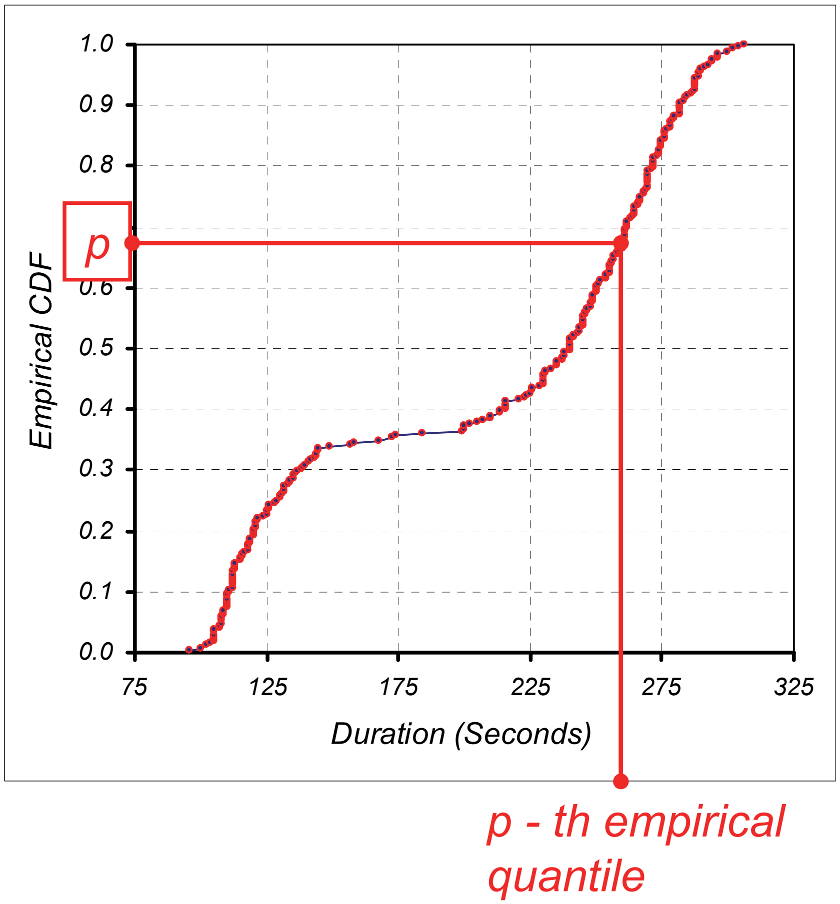
Thus:

$$\alpha = 8.5 - 8 = 0.5.$$

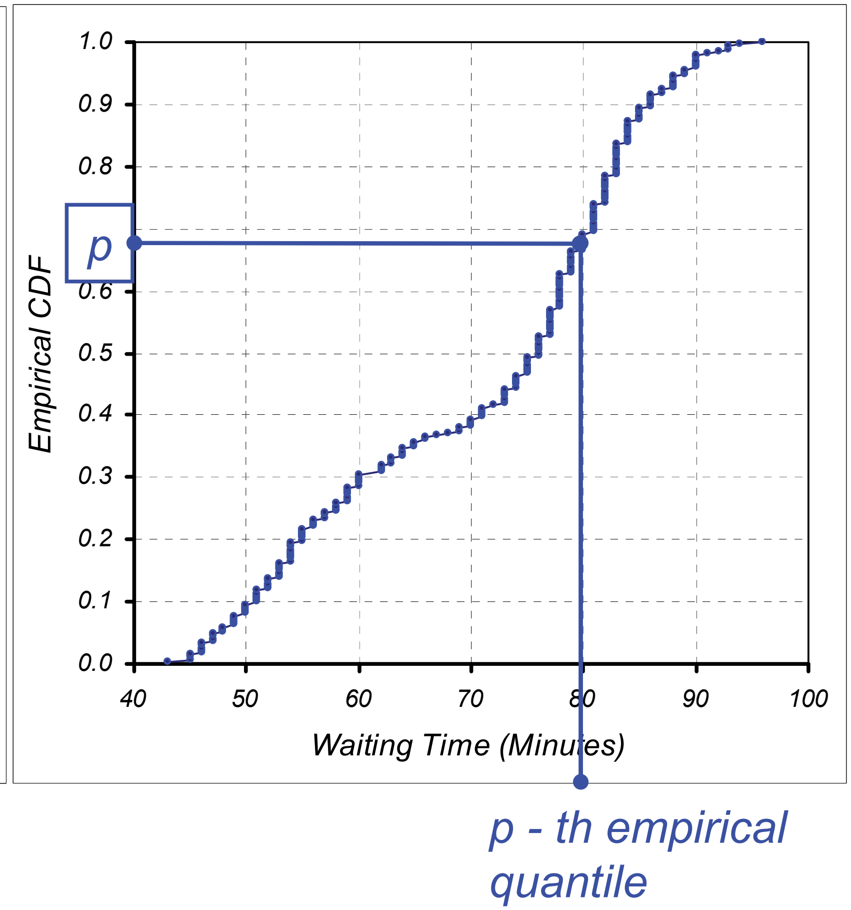
With $x_{(8)} = 16$ and $x_{(9)} = 19$:

$$q_{10}(0.85) = 0.5 \cdot (19 - 16) + 16$$
$$= 17.5$$



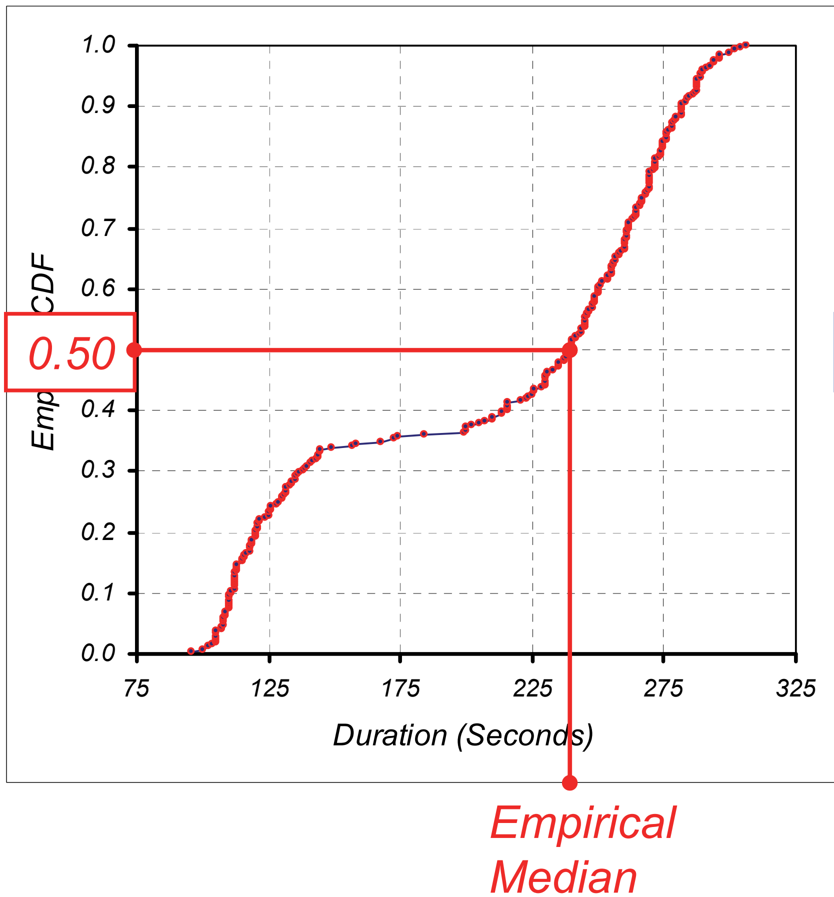


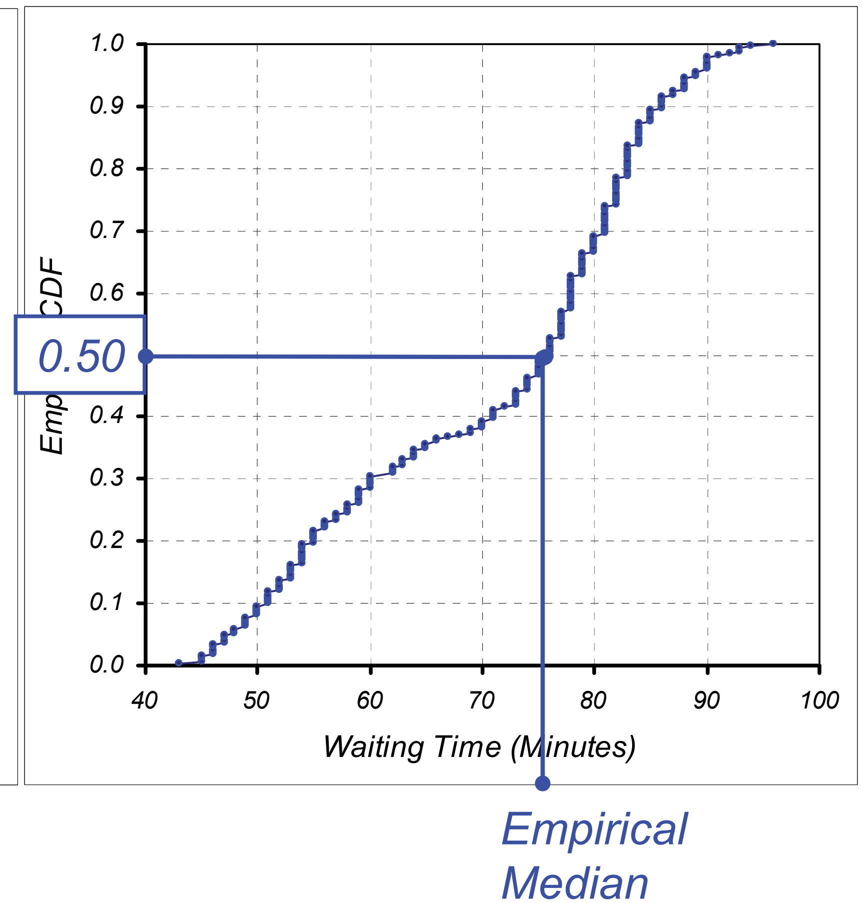
1.0
0.9
0.8
p
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Empirical CDF
75
125
175
225
275
325
Duration (Seconds)
p - th empirical quantile

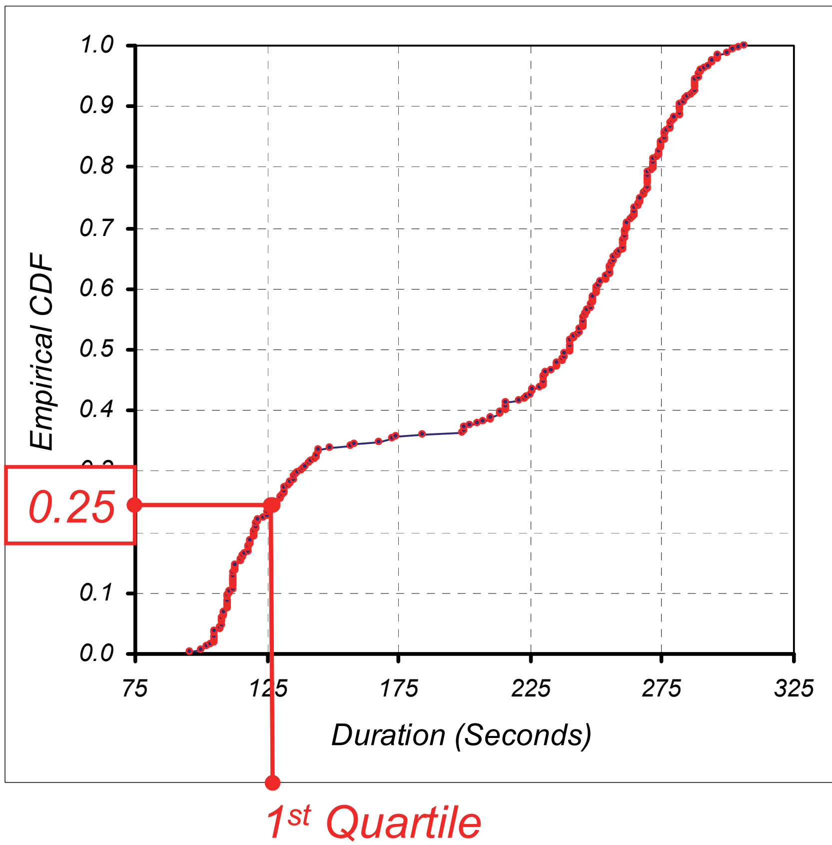


1.0
0.9
0.8
p
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Empirical CDF
40
50
60
70
80
90
100
Waiting Time (Minutes)
p - th empirical quantile

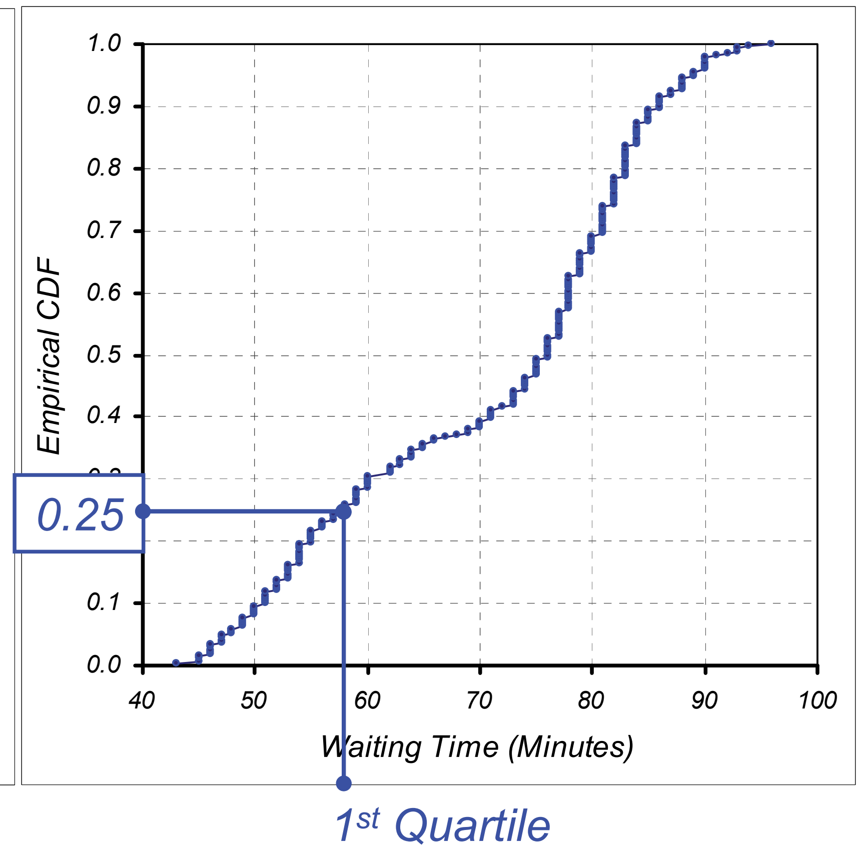
16.3 Empirical Quantiles, Quartiles and the Inter Quartile Range (IQR) . . .



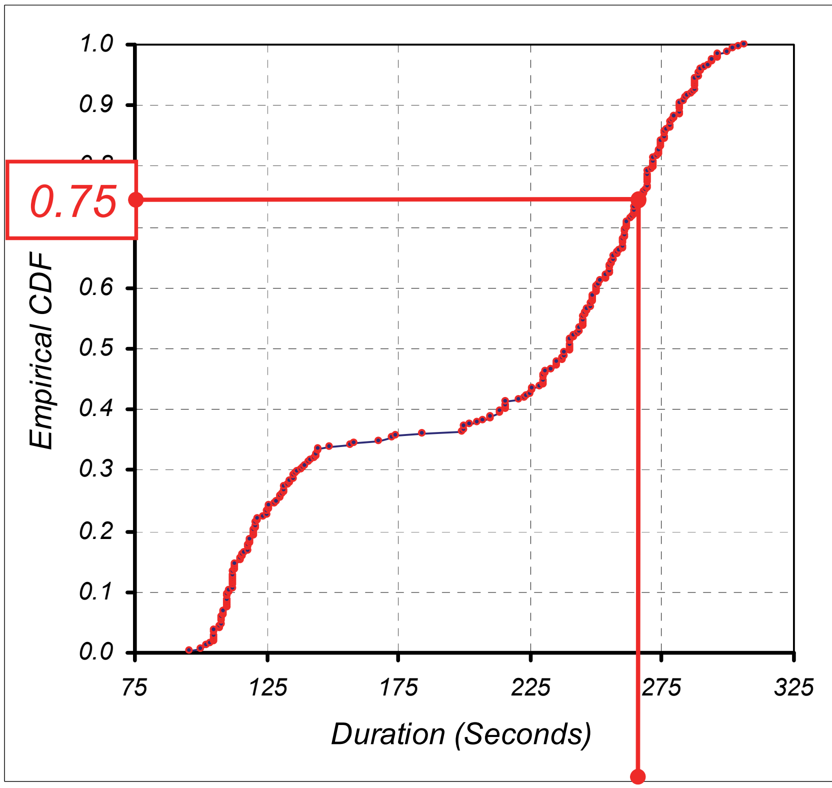


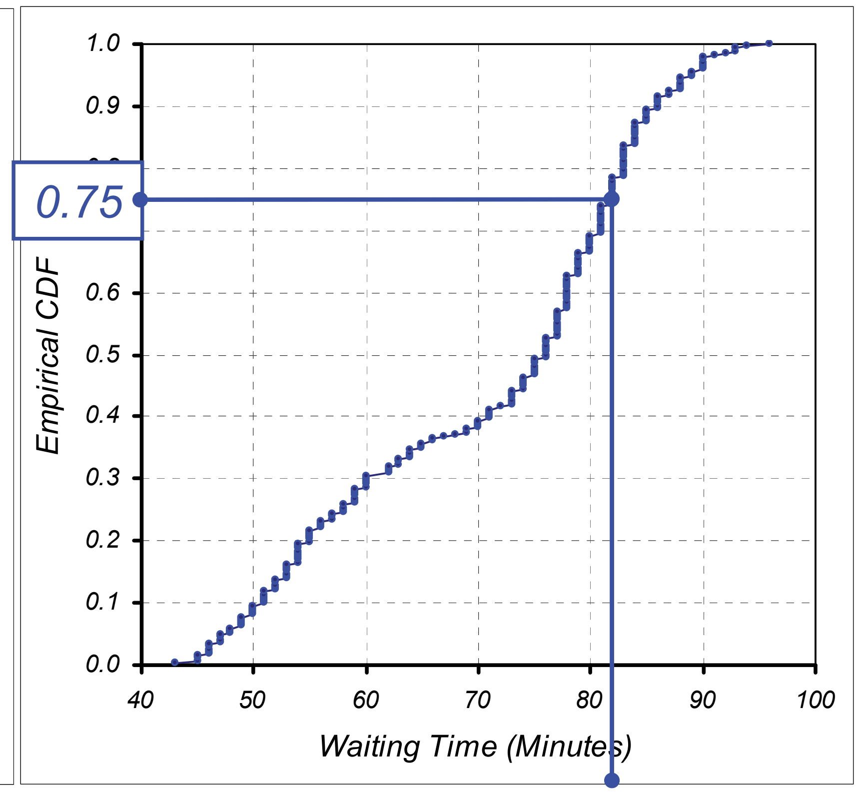


Empirical CDF
Duration (Seconds)
0.25
1st Quartile



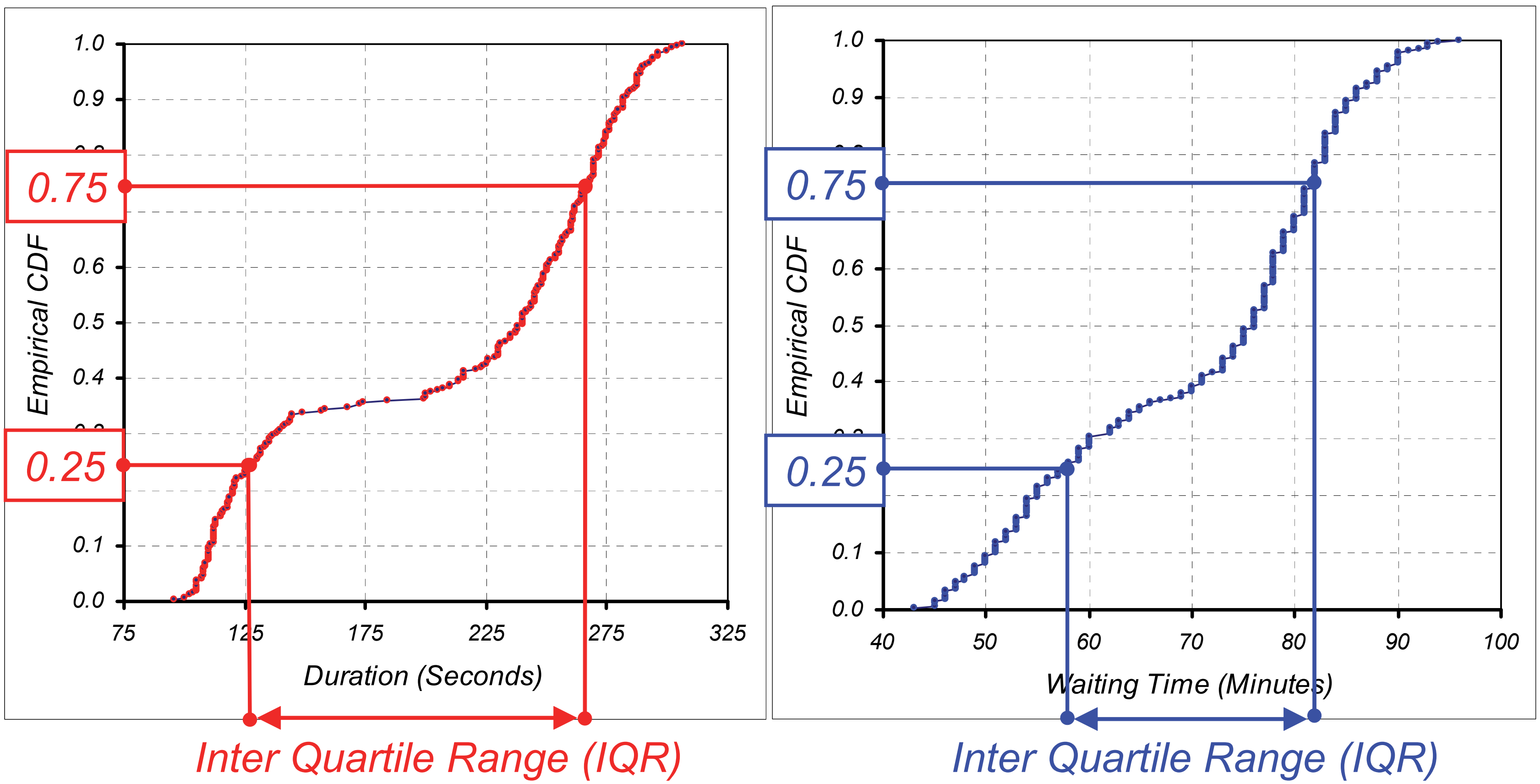
Empirical CDF
Waiting Time (Minutes)
0.25
1st Quartile





3rd Quartile

3rd Quartile


1.0
0.9
0.75
0.6
0.5
0.4
0.25
0.1
0.0
Empirical CDF
75
125
175
225
275
325
Duration (Seconds)
Inter Quartile Range (IQR)
1.0
0.9
0.75
0.6
0.5
0.4
0.25
0.1
0.0
Empirical CDF
40
50
60
70
80
90
100
Waiting Time (Minutes)
Inter Quartile Range (IQR)

- **The distance between the upper and lower quartiles** is called **the interquartile range**, or IQR:

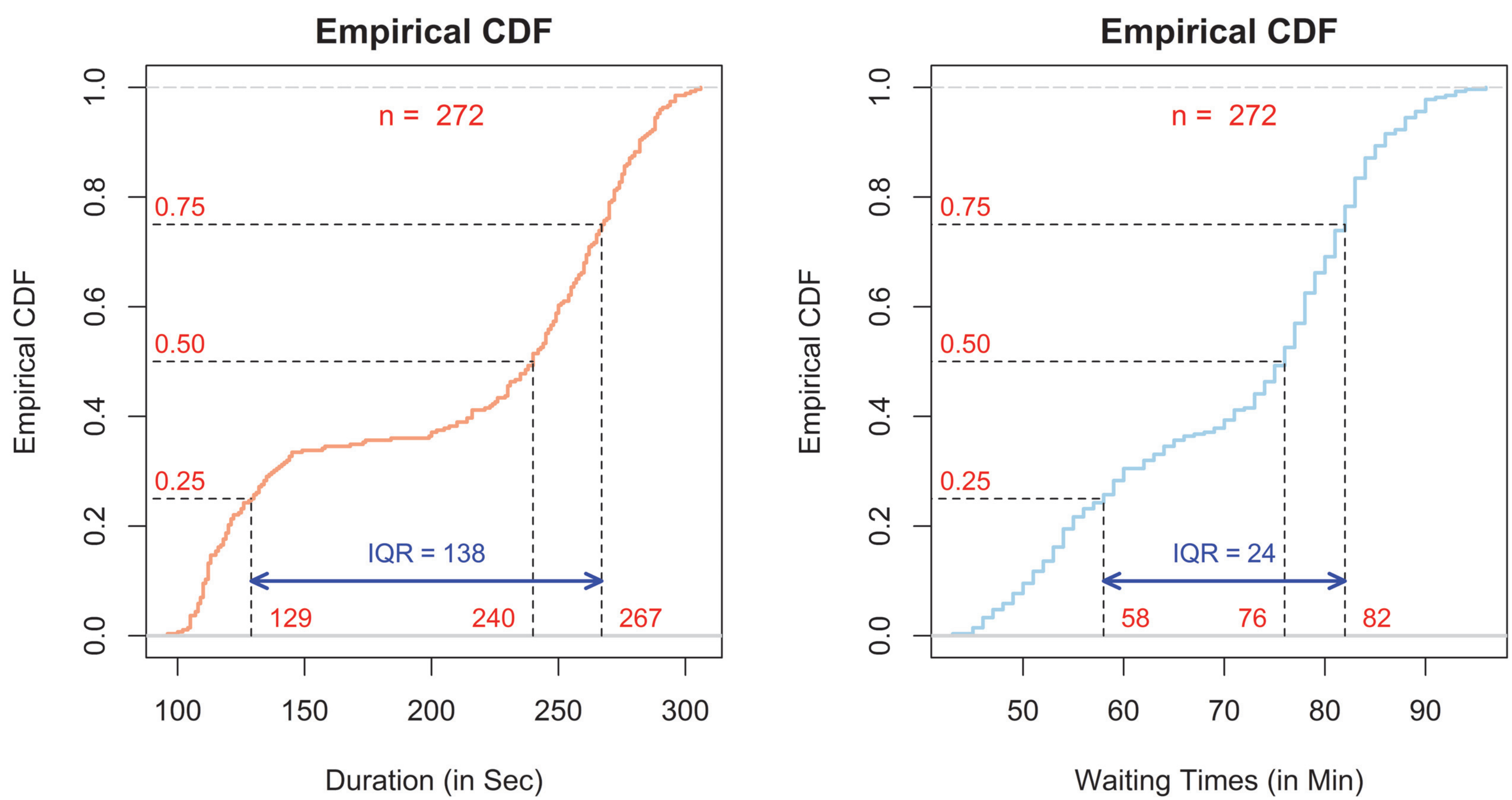
$$IQR = q_n(0.75) - q_n(0.25)$$

Example: Old Faithfull Duration Data

n **272**

	Min	Lower Quartile	Median	Upper Quartile	Max	IQR
	96	129	240	267	306	138
p	0.0037	0.25	0.5	0.75	1	
np	1	68	136	204	272	
k	1	68	136	204	272	
α	0	0	0	0	0	
$x_{(k)}$	96	129	240	267	306	
$x_{(k+1)}$	100	130	240	268	Infin	

Same analysis in file "OldFaithFul_IQR.R"



- Tukey (1977) proposed visualizing the five-number summary discussed in the previous section by a so-called box-and-whisker plot, briefly boxplot.

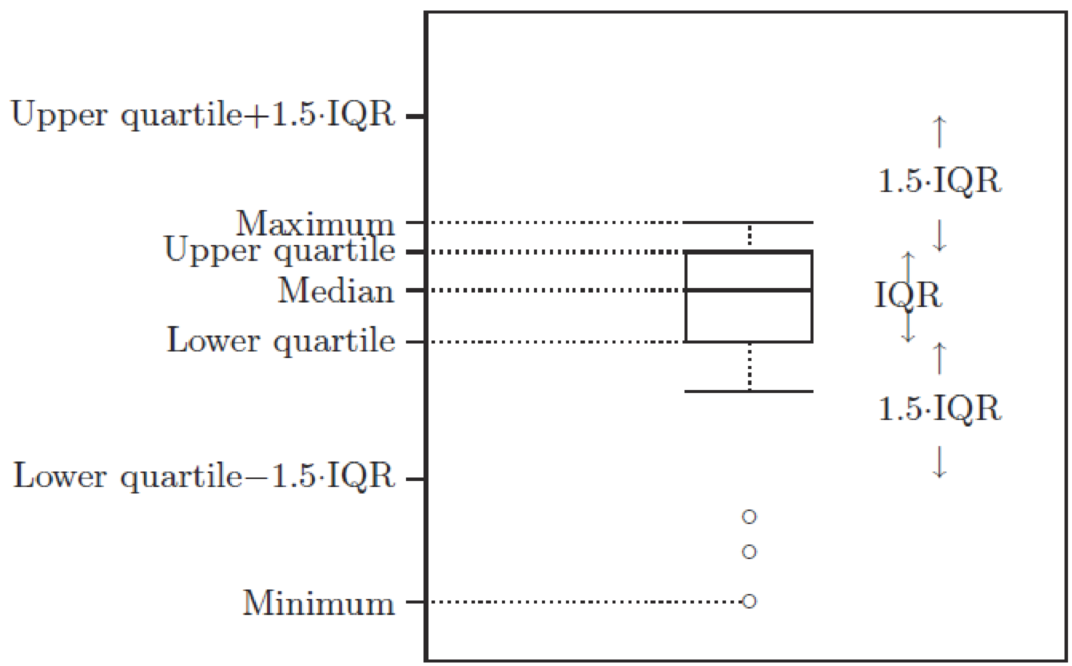


Fig. 16.3. A boxplot.

Examples: Old Faithfull Duration and Software Data

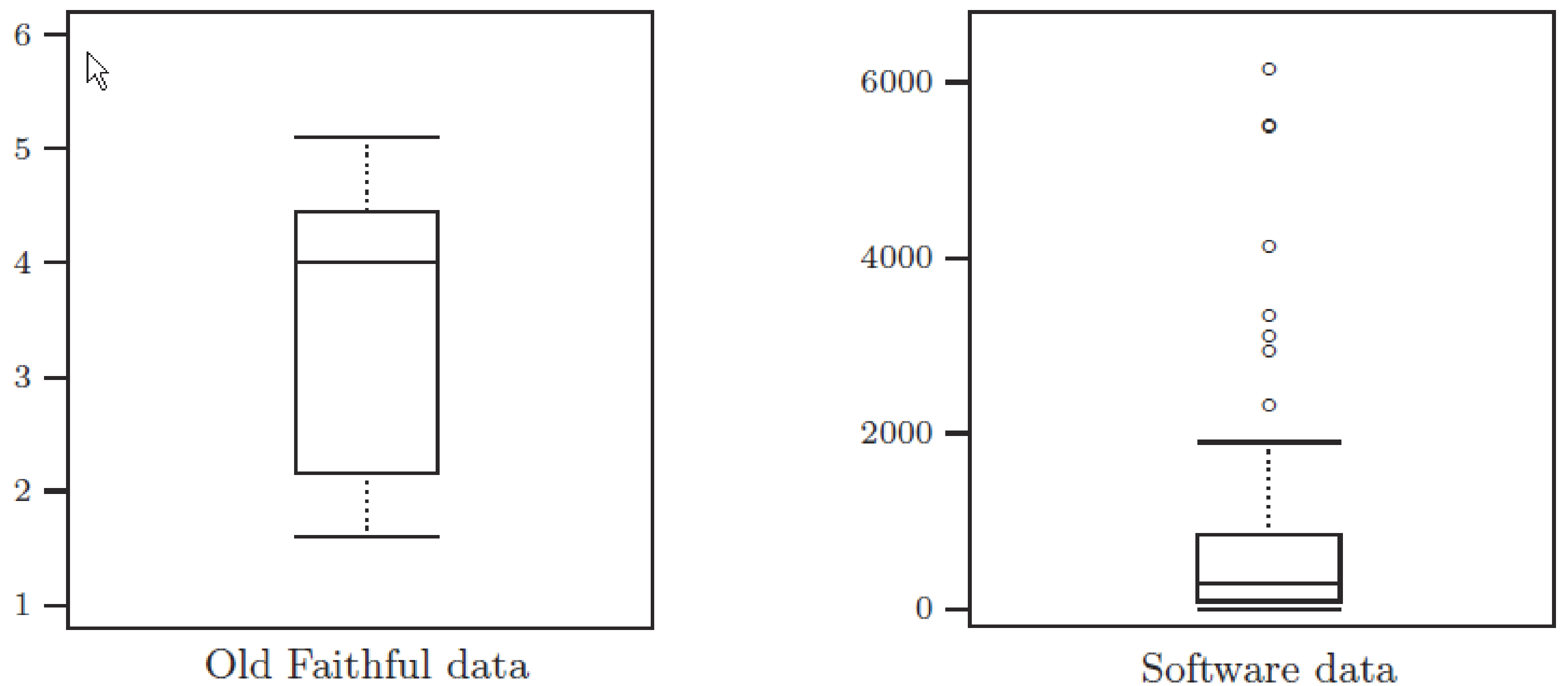
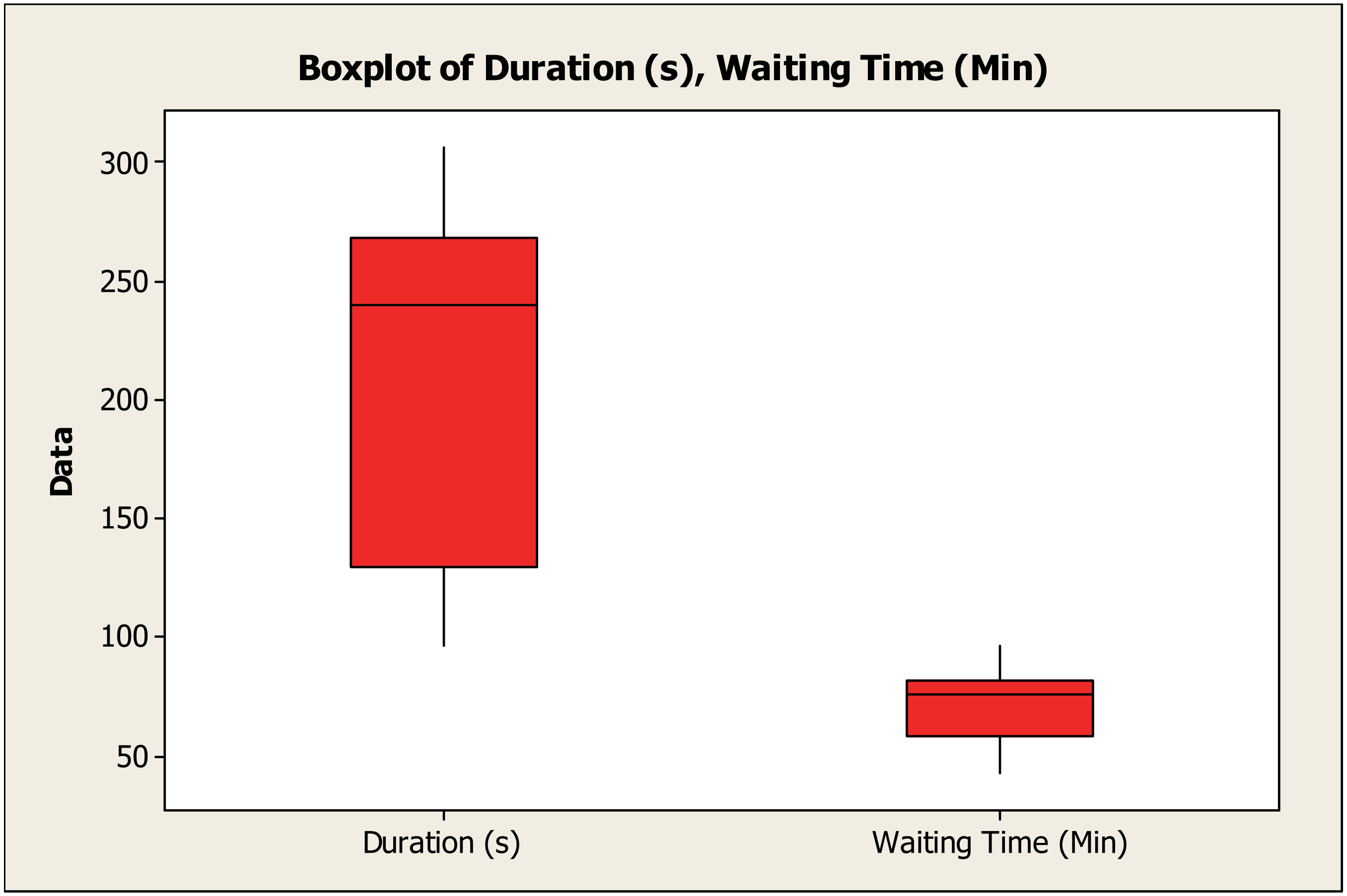
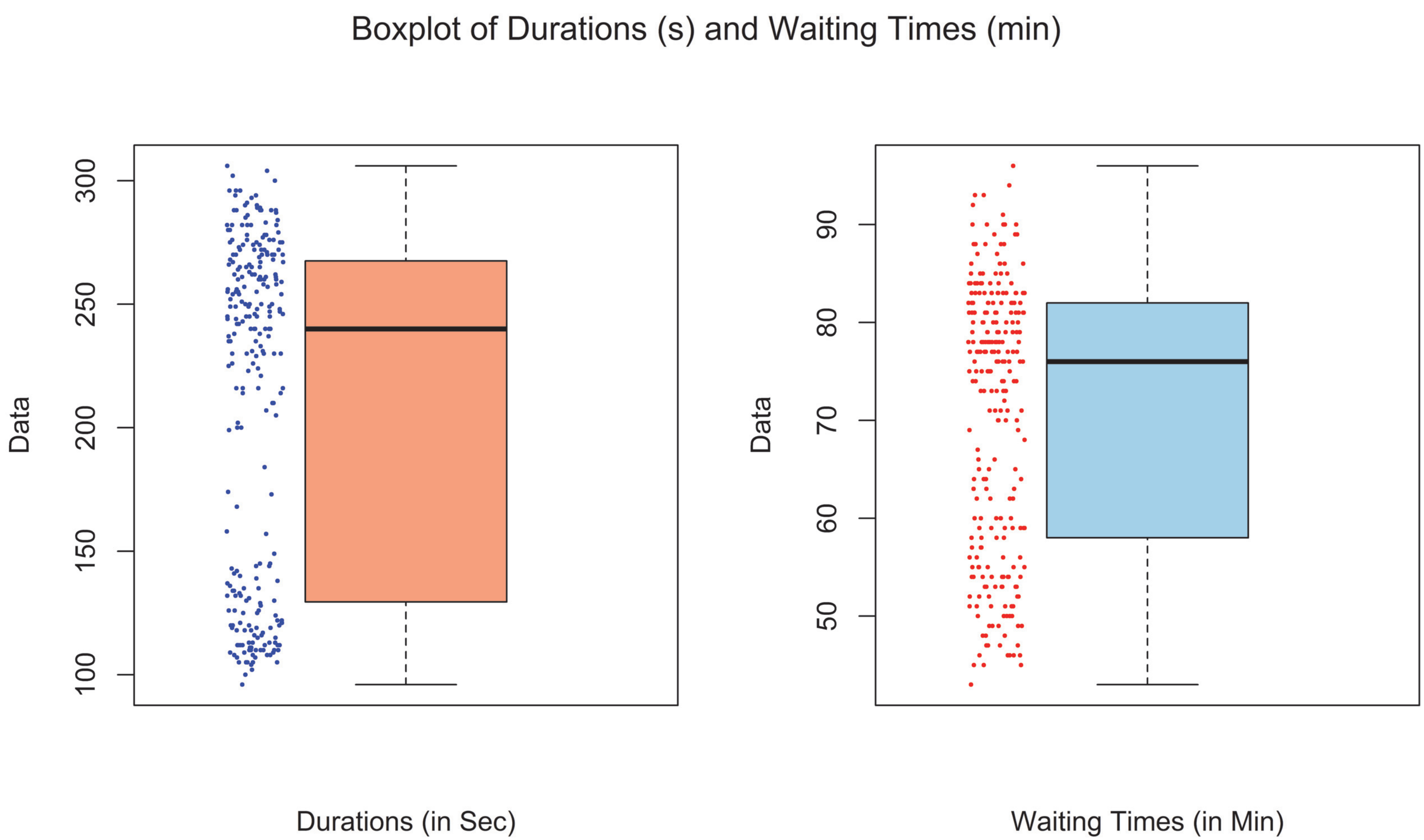


Fig. 16.4. Boxplot of the Old Faithful data and the software data.

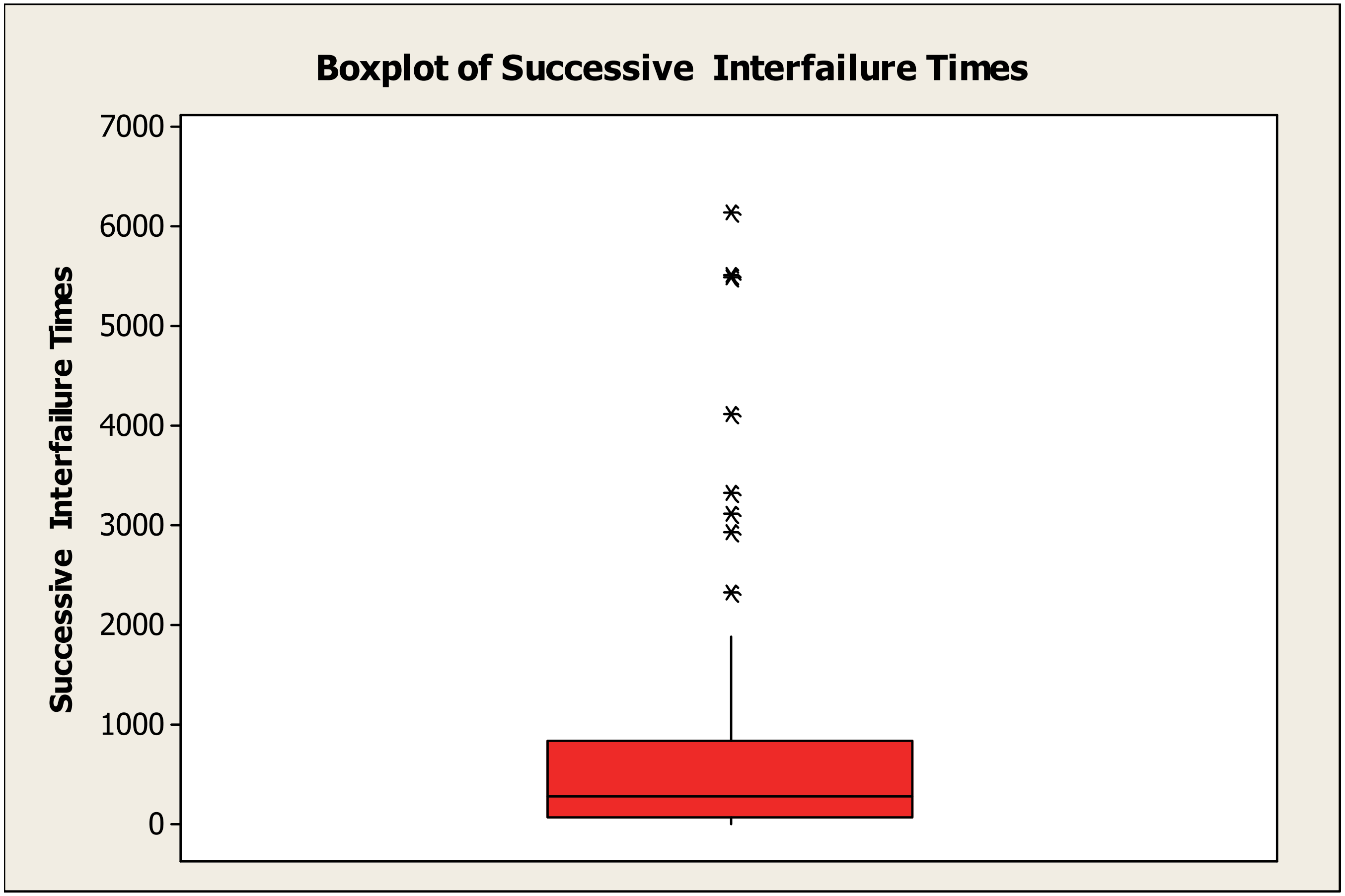
Example: Minitab box-plot Old-Faithfull data



Same analysis in file "OldFaithFul_BoxPlot.R"



Example: Minitab box-plot software data



Same analysis in file "SoftwareFailure_BoxPlot.R"