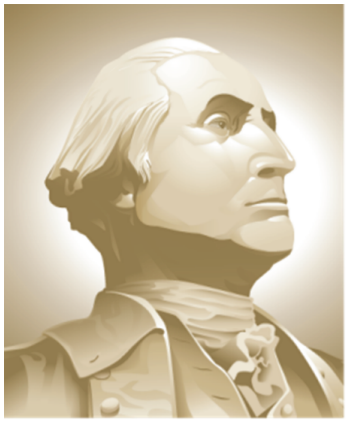
Lecture Notes EMSE 4765:
Data Analysis - Statistics Review

Chapter 15: Exploratory Data Analysis: Graphical Summaries

Version: 1/18/2021



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

15 Exploratory Data Analysis: Graphical Summaries

15.0 Introduction . . .

- **Probability models are used in practice to describe random phenomena.**
- **Confronted with a new (uncertain) phenomenon,** one conducts experiments and records observations concerning the phenomenon. **The set of observations is called a dataset.**
- **By exploring the dataset** we can **gain insight** into **what probability model might suit the phenomenon**. Exploration typically uses **visual graphical tools** to enhance **comprehension of general data set characteristics**.
- To graphically represent **univariate datasets**, consisting of repeated measurements of **one particular quantity**, we discuss **the classical histogram**, and **the empirical distribution function**.
- To graphically represent **a bivariate dataset**, which consists of repeated measurements of **two quantities**, we use **the scatterplot**.

15 Exploratory Data Analysis: Graphical Summaries

15.1 Example: The Old Faithfull Data . . .

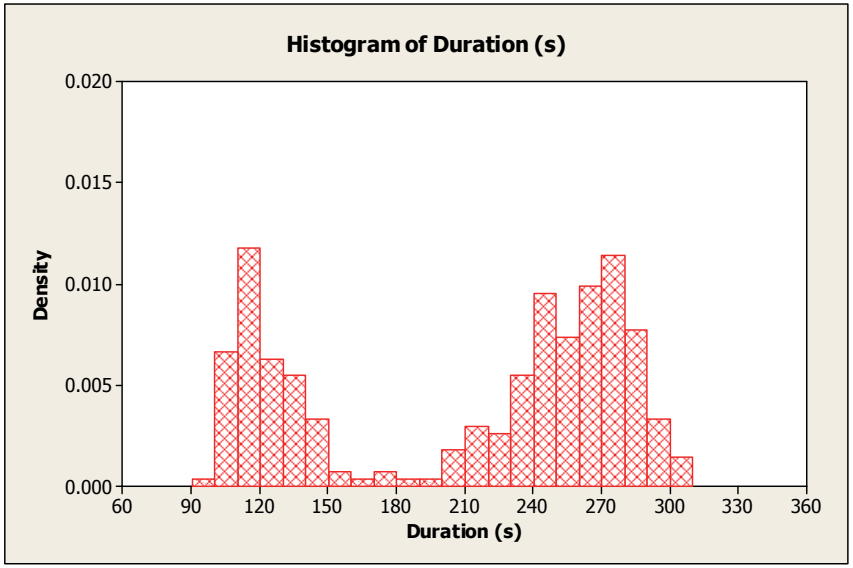
- **The Old Faithful geyser at Yellowstone National Park**, Wyoming, USA, was observed from August 1st to August 15th, 1985. During that time, data were collected on eruptions. **There were 272 durations of and time between eruptions observed**, of which 15 data points are listed below.

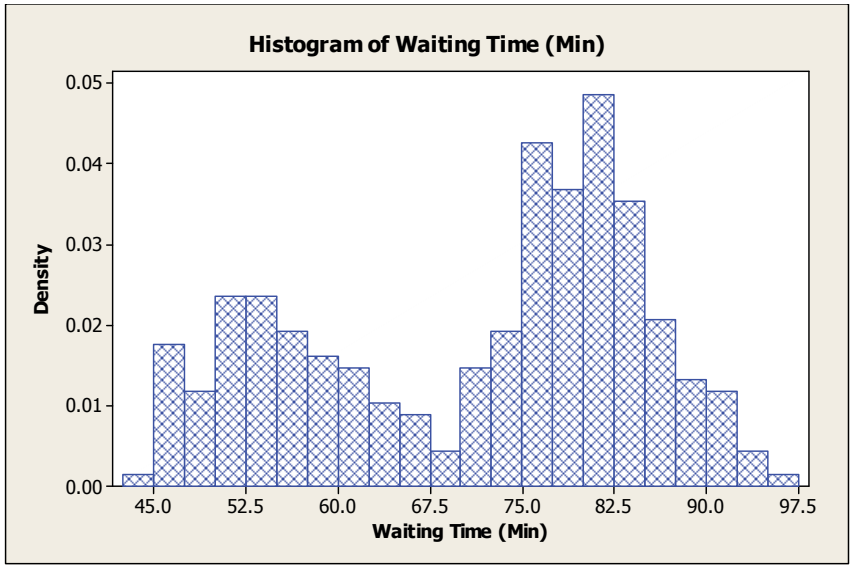
	Duration (s)	Waiting Time (Min)
1	119	43
2	112	45
3	115	45
4	132	45
5	107	46
6	109	46
7	110	46
8	110	46
9	129	46
10	105	47
11	105	47
12	112	47
13	141	47
14	105	48
15	112	48

15 Exploratory Data Analysis: Graphical Summaries

15.2 Histograms . . .

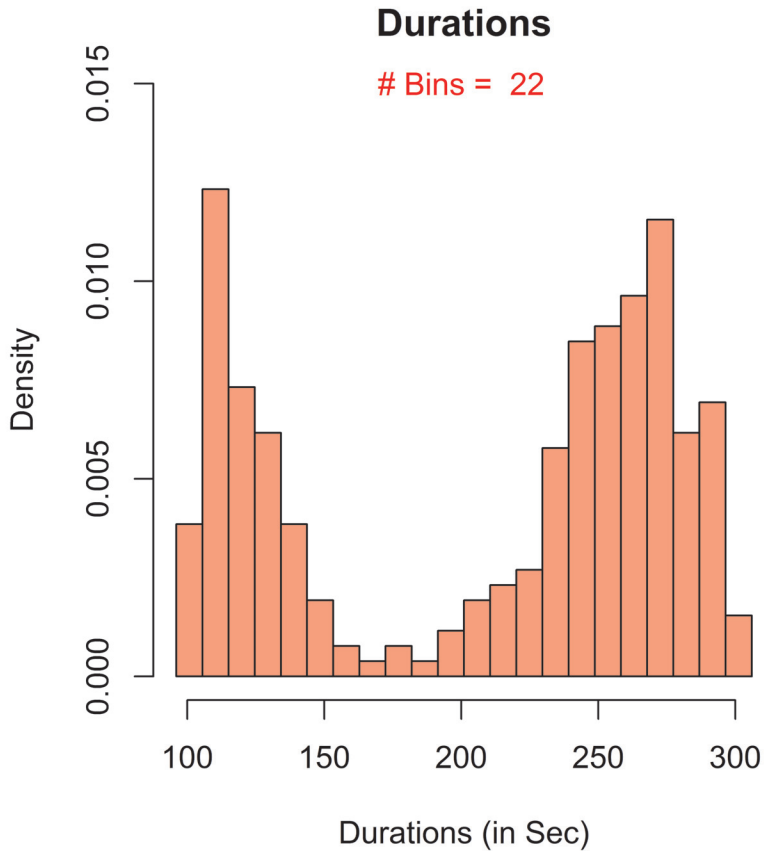
- **The classical method** to graphically represent data is **the histogram**, which probably dates from **the mortality studies of John Graunt in 1662.**

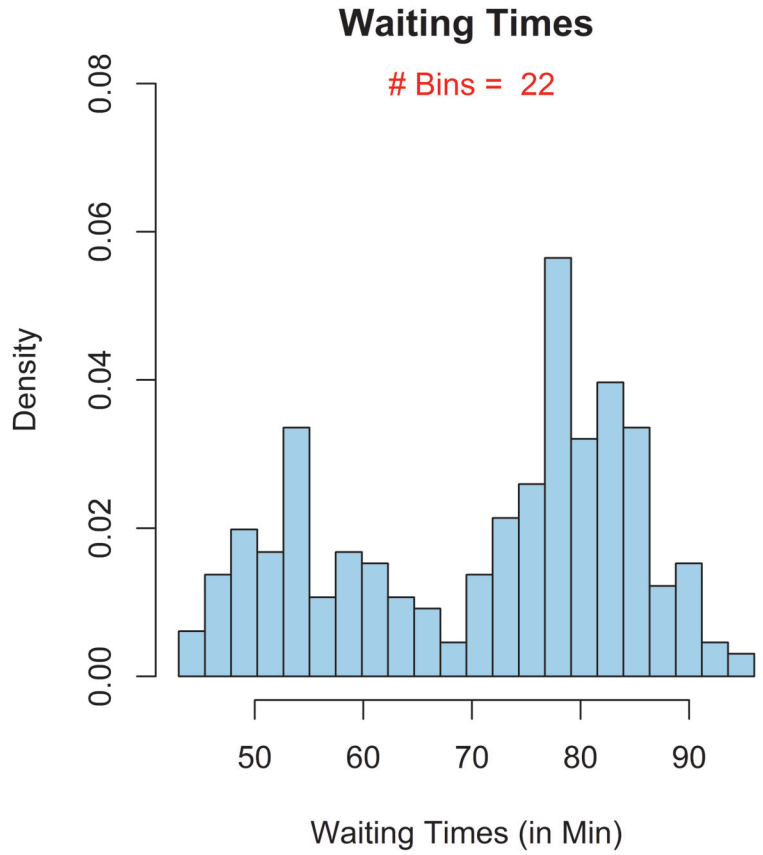




- Histograms reveals **the asymmetry of the datasets** and that the data accumulate somewhere **near 120s and 270s (52.5min and 80min)**.

Same analysis in file "OldFaithFul.R"





- Let us denote **a generic (univariate) dataset of size n by $x_1, x_2, \ldots, x_n$** and suppose we want to construct a histogram. Here we describe **the version of the histogram such that the total area under the curve is equal to one.**

- First **the data range is divided into m intervals. These intervals are called bins** and are denoted $B_1, \ldots, B_m$. The length of an interval $|B_i|$ **is called the bin width.** The bins do **not necessarily have the same width** (but often do).

- 1. We **determine empirically the probability** of an observation falling in B_i:

$$\widehat{P}(\text{Observation in } B_i) = \frac{\#\text{ Observations in } B_i}{n}$$

- 2. We **determine empirically the density value** in bin B_i with width $|B_i|$ as:

$$\widehat{f}(\text{Observation in } B_i) = \frac{\#\text{ Observations in } B_i}{n \cdot |B_i|}$$

Quick Exercise 15.2: From Old-Faithfull duration data count how many elements fall into each of the bins $(90, 120], (120, 150], \ldots, (300, 330]$ in Figure 15.1 and compute the height in each bin.

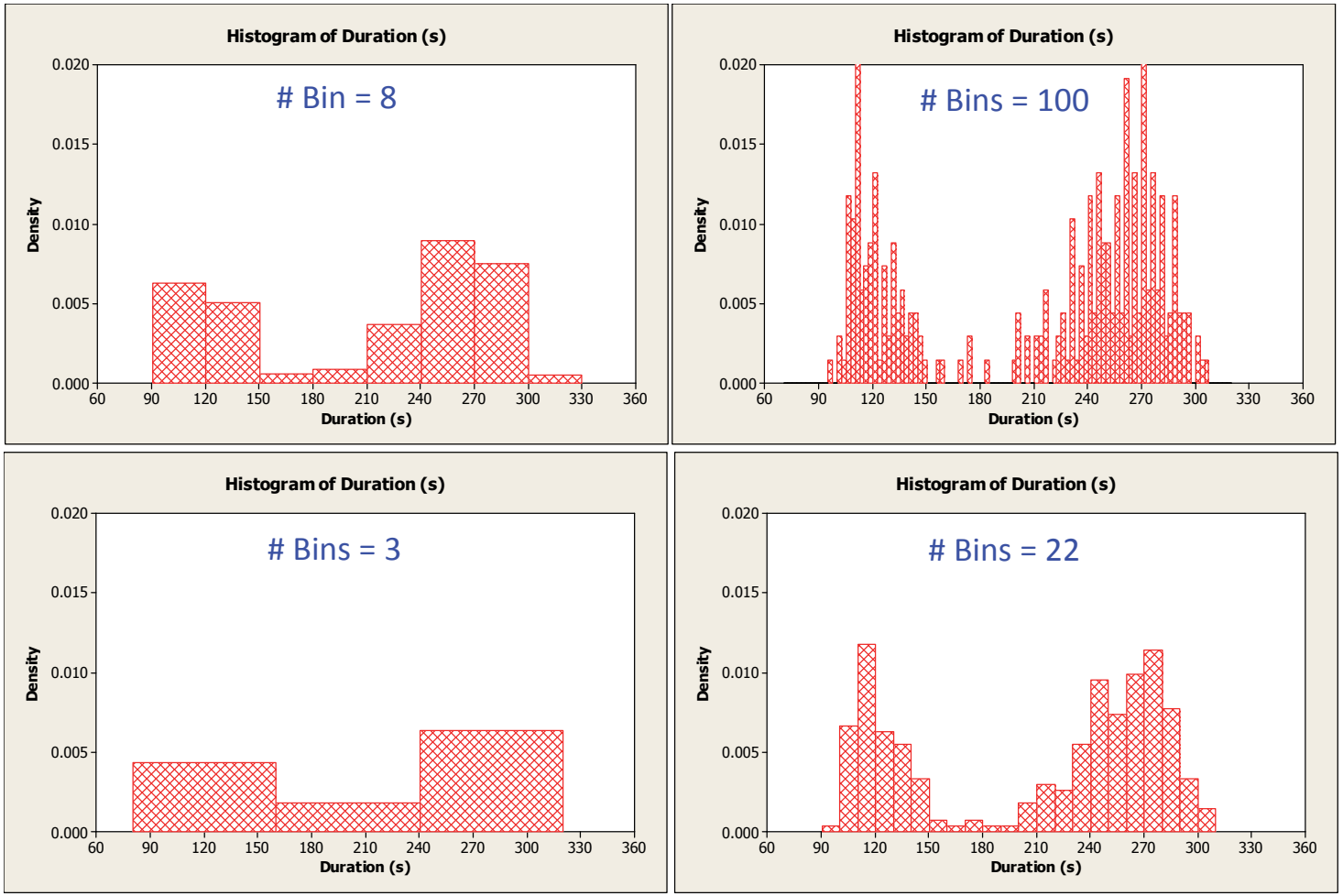
j	LB_j	UB_j	Counts	P(X in Bj)	f(x) in Bj
1	90	120	55	0.202	0.0067
2	120	150	37	0.136	0.0045
3	150	180	5	0.018	0.0006
4	180	210	9	0.033	0.0011
5	210	240	34	0.125	0.0042
6	240	270	75	0.276	0.0092
7	270	300	54	0.199	0.0066
8	300	330	3	0.011	0.0004
Total			272	1.000	

- **Guide lines for number of bins (m) or bin widths (b) to use:**

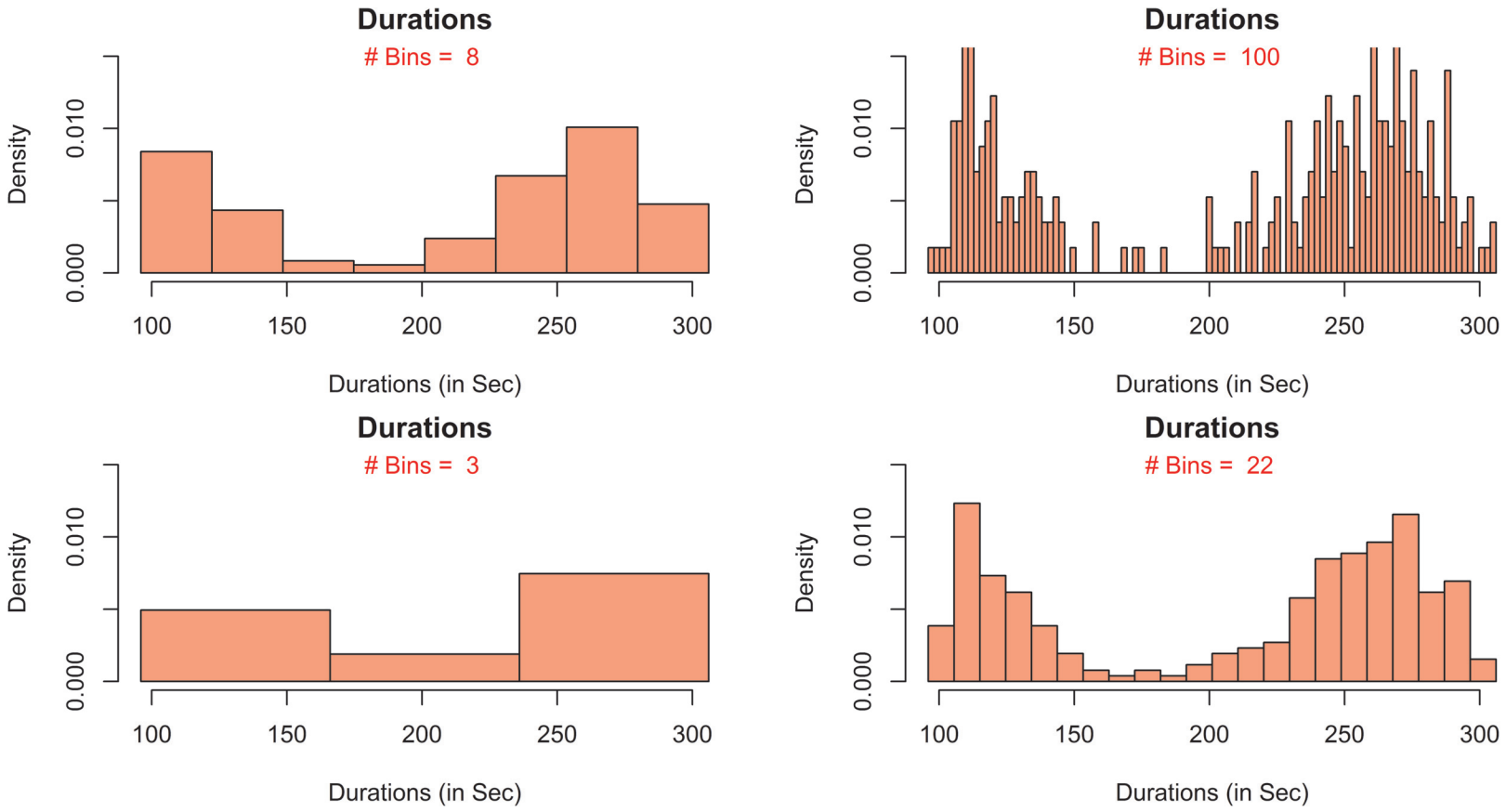
$$\begin{cases} m = 1 + 3.3\,{}^{10}\log n & \Rightarrow m \approx 9.03 \\ b = 3.49 s n^{-1/3} & \Rightarrow s = 68.48,\ b = 36.89,\ m = \frac{306-96}{36.89} \approx 5.69 \end{cases}$$

15 Exploratory Data Analysis: Graphical Summaries

15.2 Histograms . . .



Same analysis in file "OldFaithFul.R"



15 Exploratory Data Analysis: Graphical Summaries

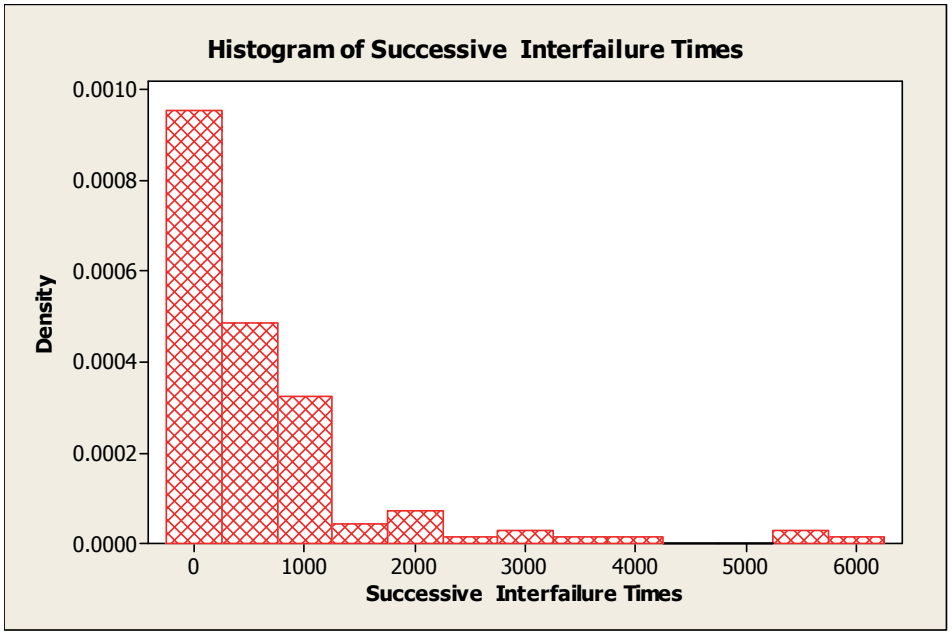
15.2 Histograms . . .

Example - Software Failure Times: In order to estimate the parameters of a software reliability model, failure data are collected. **The 136 failure times** are recorded, or equivalently, **the length of an interval between successive failures**. **Observed interfailure times in CPU seconds** for a certain software system.

Table 15.3. Interfailure times between successive failures.

30	113	81	115	9	2	91	112	15	138
50	77	24	108	88	670	120	26	114	325
55	242	68	422	180	10	1146	600	15	36
4	0	8	227	65	176	58	457	300	97
263	452	255	197	193	6	79	816	1351	148
21	233	134	357	193	236	31	369	748	0
232	330	365	1222	543	10	16	529	379	44
129	810	290	300	529	281	160	828	1011	445
296	1755	1064	1783	860	983	707	33	868	724
2323	2930	1461	843	12	261	1800	865	1435	30
143	108	0	3110	1247	943	700	875	245	729
1897	447	386	446	122	990	948	1082	22	75
482	5509	100	10	1071	371	790	6150	3321	1045
648	5485	1160	1864	4116					

Source: J.D. Musa, A. Iannino, and K. Okumoto. *Software reliability: measurement, prediction, application.* McGraw-Hill, New York, 1987; Table on page 305.



- Another way to graphically represent a dataset is **to plot the data in a cumulative manner.** This can be done using **the empirical cumulative distribution function** of the data.

$$F_n(x) = \frac{\text{number of data points } \leq x}{n}$$

Example:

Consider the dataset 4, 3, 9, 1, 7.

$$n = 5 \Rightarrow F_5(1) = \frac{1}{5},\ F_5(2) = \frac{1}{5},$$
$$F_5(3) = \frac{2}{5}, F_5(4) = \frac{3}{5},$$
$$F_5(7) = \frac{4}{5}, F_5(9) = \frac{5}{5}.$$

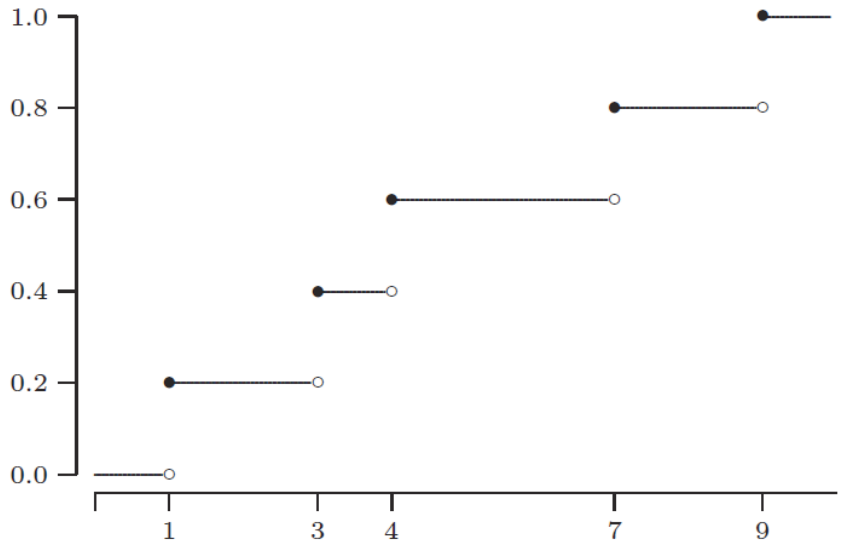
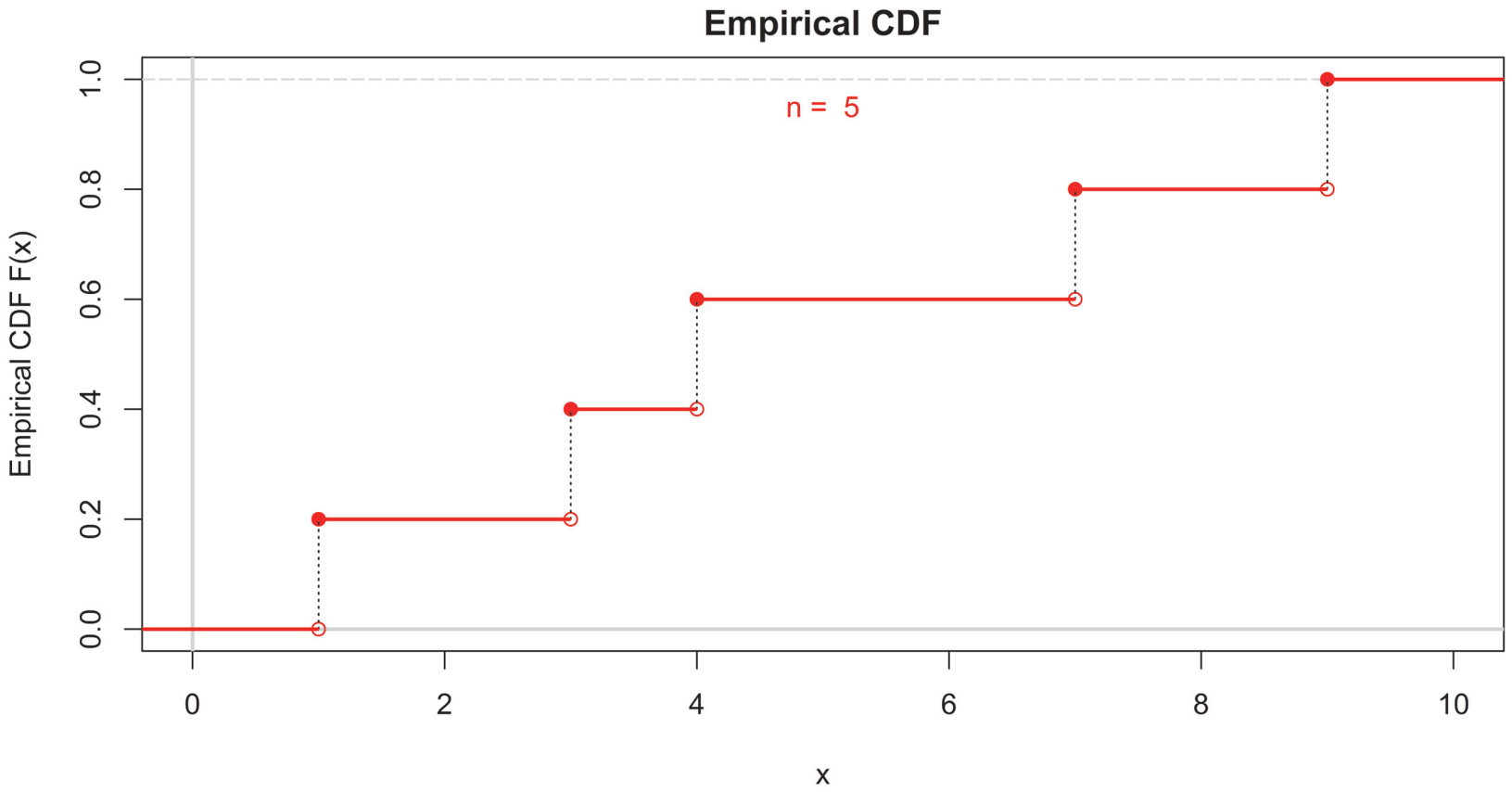


Fig. 15.9. Empirical distribution function.

Same analysis in file "ECDF_example.R"



15 Exploratory Data Analysis: Graphical Summaries

15.3 Empirical Distribution Function . . .

- **Ordered data:** Given a data set $x_1, x_2, \ldots, x_n$ we denote:

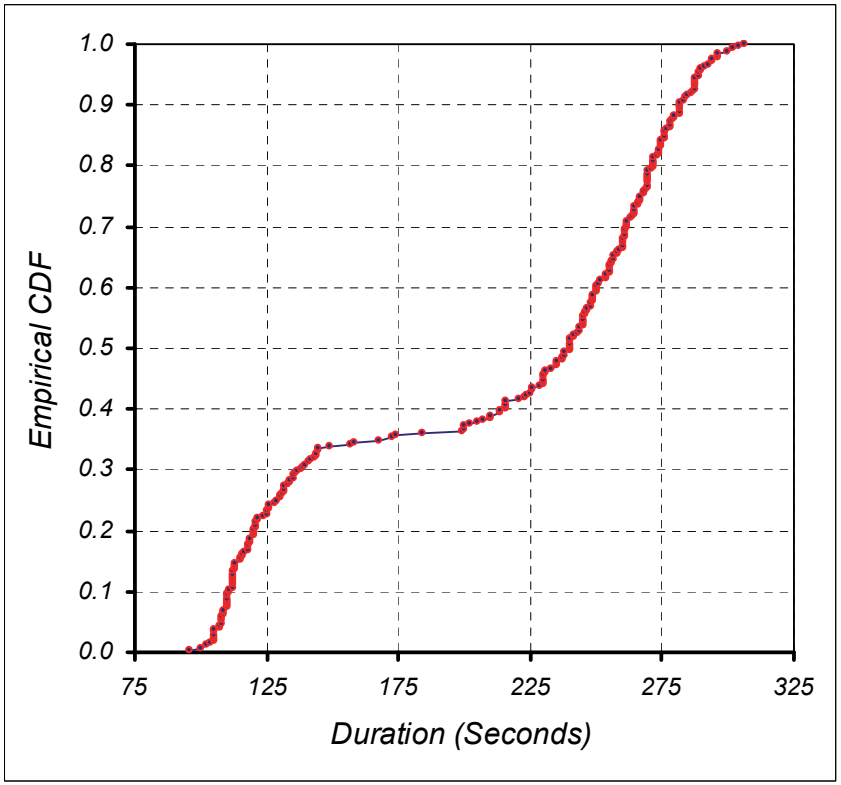
$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)} \Rightarrow F_n(x_{(j)}) = \frac{j}{n}$$

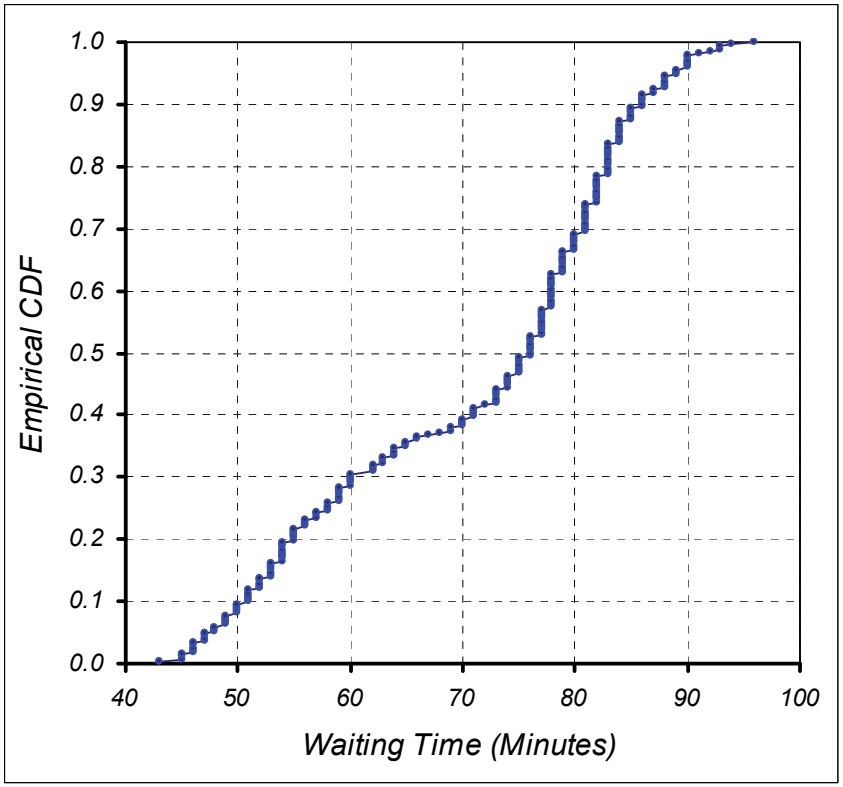
Table: Duration in seconds and Waiting Time of 272 eruptions of the Old Faithful geyser.

	Duration (s)	Waiting Time (Min)	Ordered Duration (s)		Emprical CDF	
1	119	43	96	$x_{(1)}$	0.0037	$\hat{F}(x_{(1)}) = 1/272$
2	112	45	100		0.0074	
3	115	45	102		0.0110	
4	132	45	104		0.0147	
5	107	46	105		0.0184	
6	109	46	105		0.0221	
7	110	46	105		0.0257	
8	110	46	105		0.0294	
9	129	46	105		0.0331	
10	105	47	105	$x_{(10)}$	0.0368	$\hat{F}(x_{(10)}) = 10/272$
11	105	47	107		0.0404	
12	112	47	107		0.0441	
13	141	47	108		0.0478	
14	105	48	108		0.0515	
15	112	48	108		0.0551	
⋮	⋮	⋮	⋮		⋮	

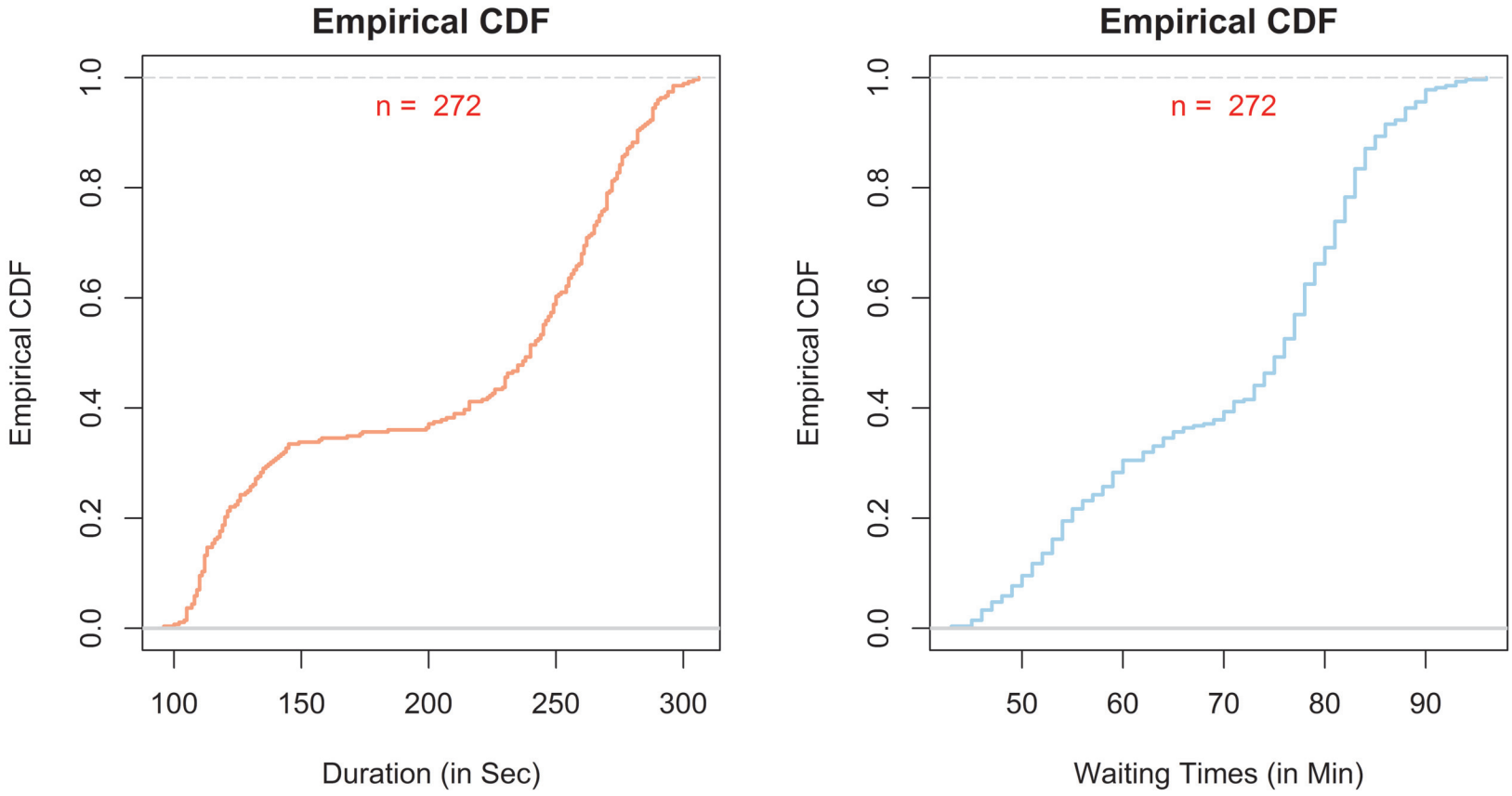
Source: W. H¨ardle. *Smoothing techniques with implementation in S*. 1991; Table 3, page 201. Springer New York.

Example: Empirical cumulative distribution functions for duration (Sec.) and waiting times (Min.) for the old-faithfull data (272 data points).





Same analysis in file "OldFaithFull.R"



Example - Software Failure Times: **Recall the 136 recorded failure times**, i.e. **observed interfailure times in CPU seconds** for a certain software system.

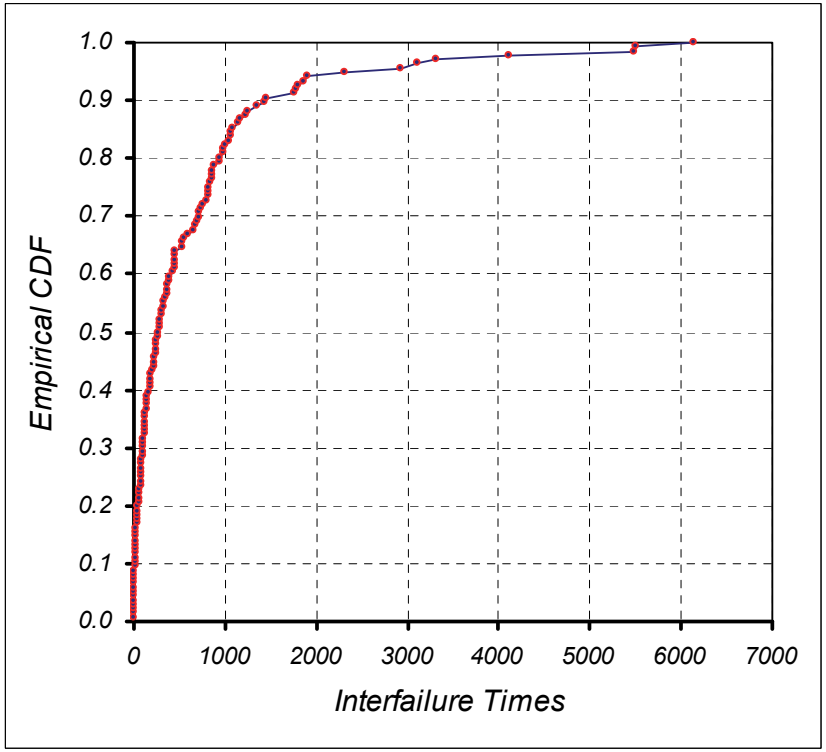
	Successive Interfailure Times	Order Failure Times	Empirical CDF
1	30	0	0.0074
2	113	0	0.0147
3	81	0	0.0221
4	115	2	0.0294
5	9	4	0.0368
6	2	4	0.0441
7	91	6	0.0515
8	112	8	0.0588
9	15	9	0.0662
10	138	10	0.0735
11	50	10	0.0809
12	77	10	0.0882
13	24	12	0.0956
14	108	15	0.1029
15	88	15	0.1103
16	670	16	0.1176
17	120	21	0.1250
18	26	22	0.1324
19	114	24	0.1397
20	325	26	0.1471
⋮	⋮	⋮	⋮

$x_{(3)}$

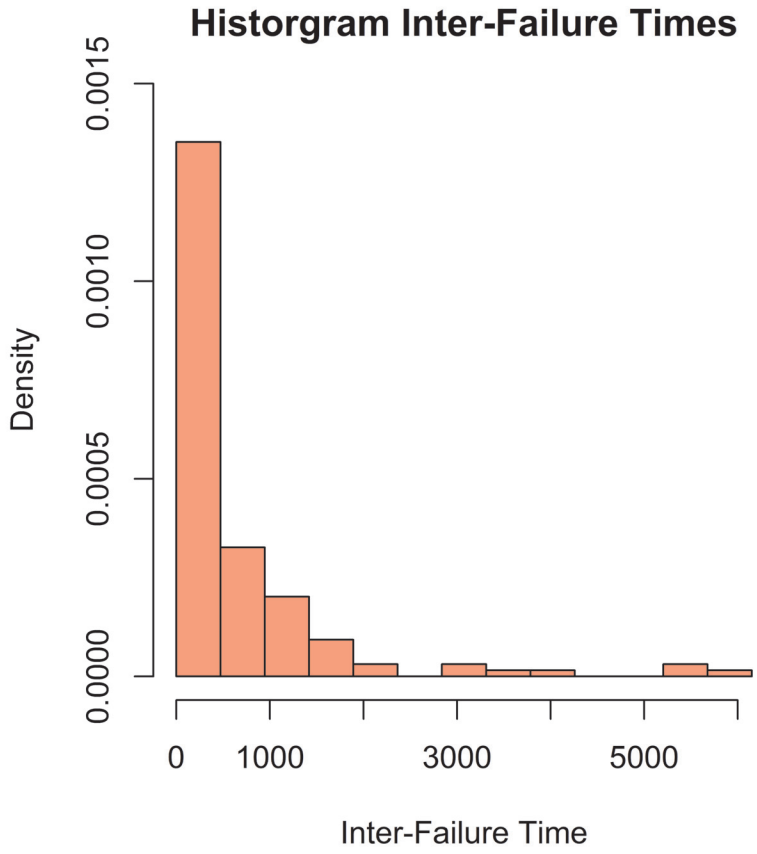
$\hat{F}(x_{(3)}) = 3/136$

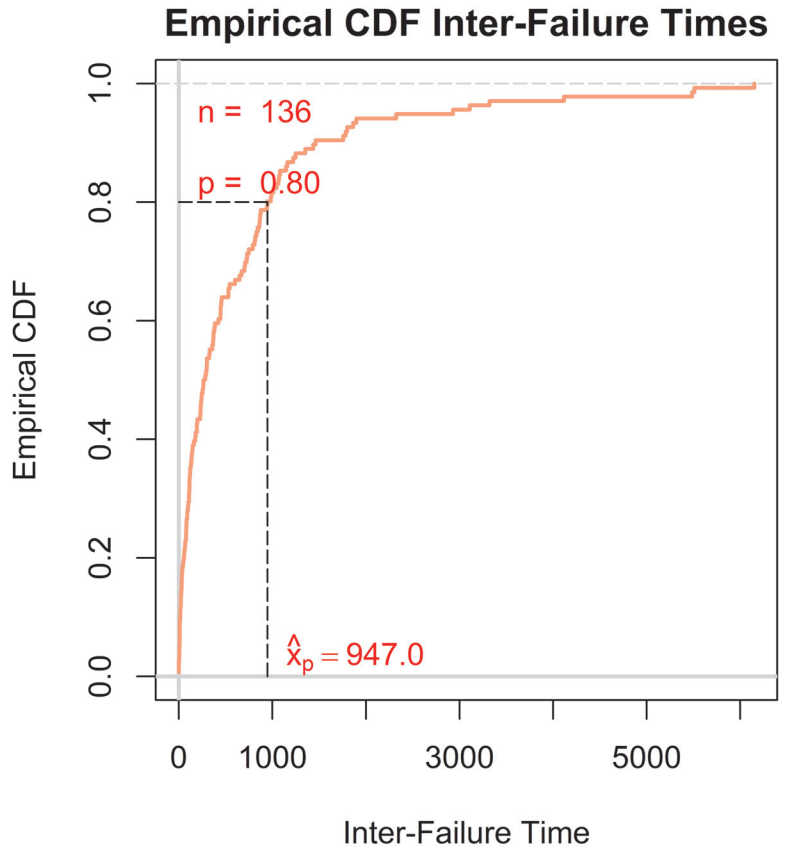
$x_{(17)}$

$\hat{F}(x_{(17)}) = 17/136$



Same analysis in file "SoftwareFailure.R"





- In some situations one wants to investigate **the relationship between two or more variables**. **In the case of two variables x and y**, the dataset consists of pairs of observations:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

We call such a dataset **a bivariate dataset** as opposed to **a univariate dataset.**

Typical question: Does the variable y depend on the variable x?

- **A first step** is to take a look at the data **in a scatter plot**, i.e., to plot the points (x_i, y_i) for $i = 1, 2, \ldots, n$ in **a cartesian coordinate system.**

Example: Drilling in Rock

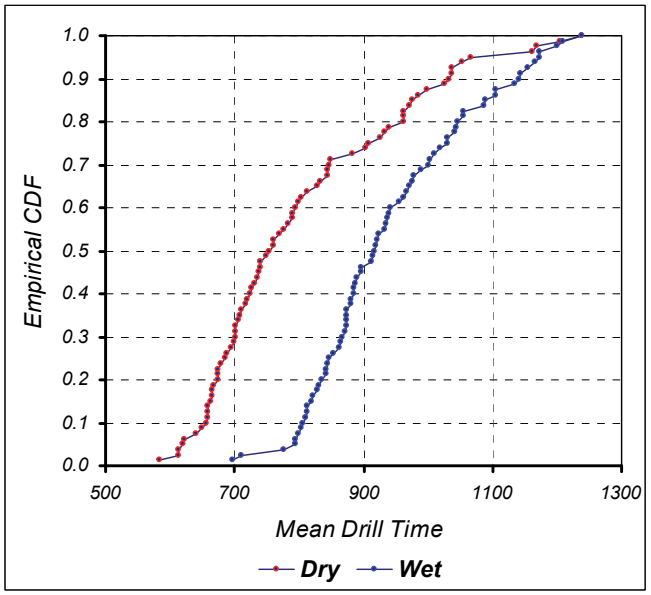
During a study about **"dry" and "wet" drilling in rock**, six holes were drilled, three corresponding to each process. In a dry hole compressed air down the drill flushes the cutting and the drive hammer, whereas in a wet hole one forces water.

15 Exploratory Data Analysis: Graphical Summaries

15.4 Scatter Plots . . .

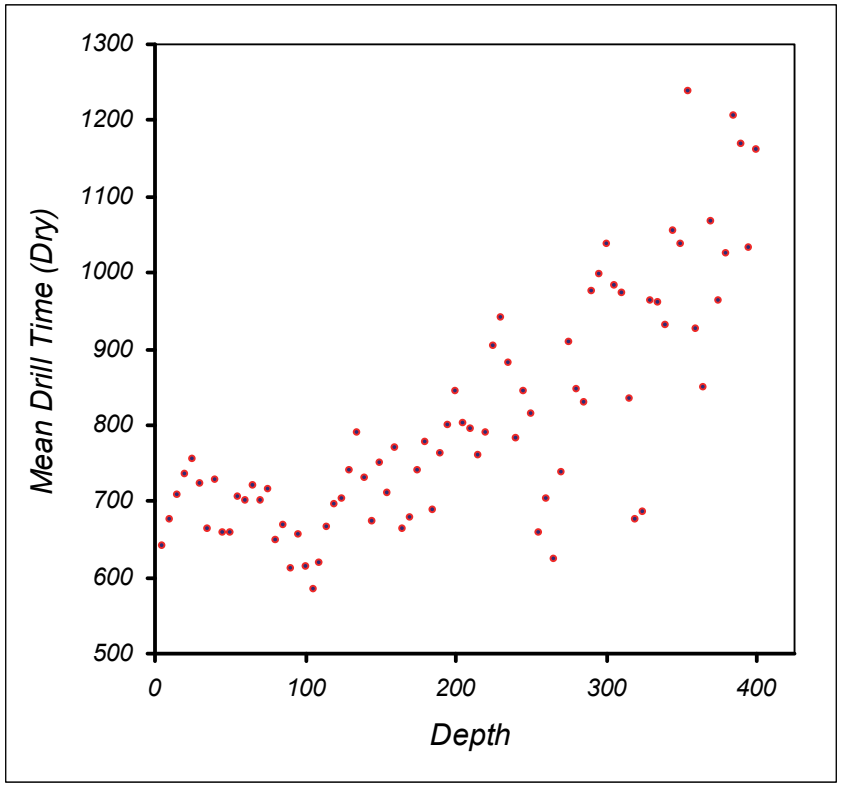
As the hole gets deeper, one has to add a rod of 5 feet length to the drill. **In each hole, the time was recorded to advance 5 feet to a total depth of 400 feet.**

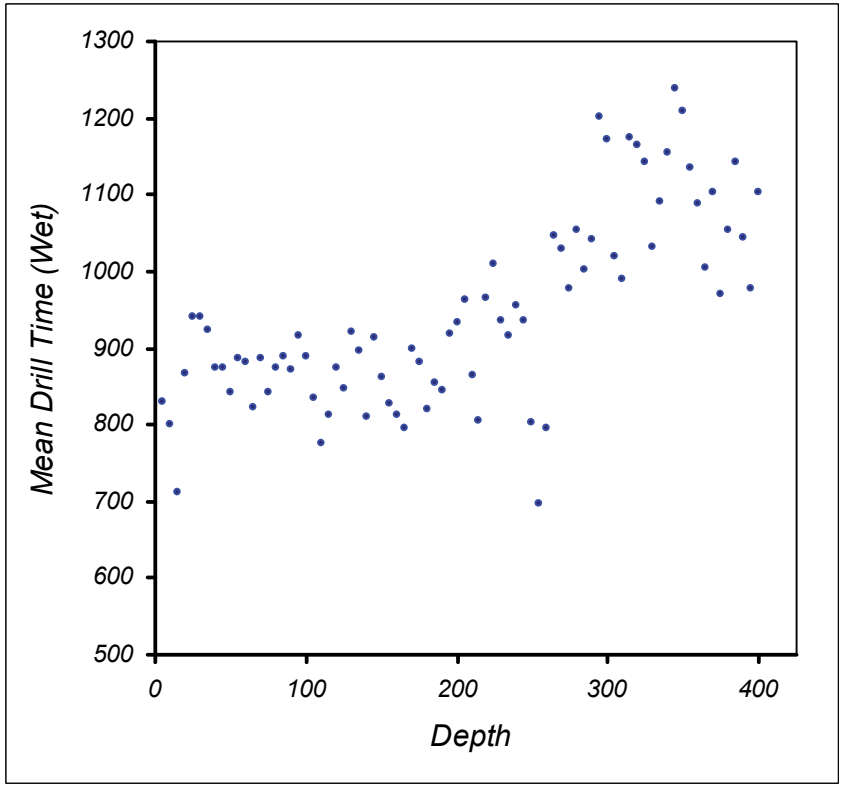
	Depth	Mean Drill Time Dry	Mean Drill Time Wet	Ordered Mean Drill Time Dry	Ordered Mean Drill Time Wet	Empirical CDF
1	5	640.67	830	584	697.33	0.0125
2	10	674.67	800	612.67	711.33	0.025
3	15	708	711.33	614	776.33	0.0375
4	20	735.67	867.67	619.67	795.67	0.05
5	25	754.33	940.67	623.5	795.67	0.0625
6	30	723.33	941.33	640.67	800	0.075
7	35	664.33	924.33	649.67	803.67	0.0875
8	40	727.67	873	656.67	805.67	0.1
9	45	658.67	874.67	658	810.33	0.1125
10	50	658	843.33	658	811.67	0.125
11	55	705.67	885.67	658.67	812.67	0.1375
12	60	700	881.67	663	819.67	0.15
13	65	720.67	822	664.33	822	0.1625
14	70	701.33	886.33	666	828	0.175
15	75	716.67	842.5	667.33	830	0.1875
16	80	649.67	874.67	674	835.33	0.2
17	85	667.33	889.33	674.67	842.5	0.2125



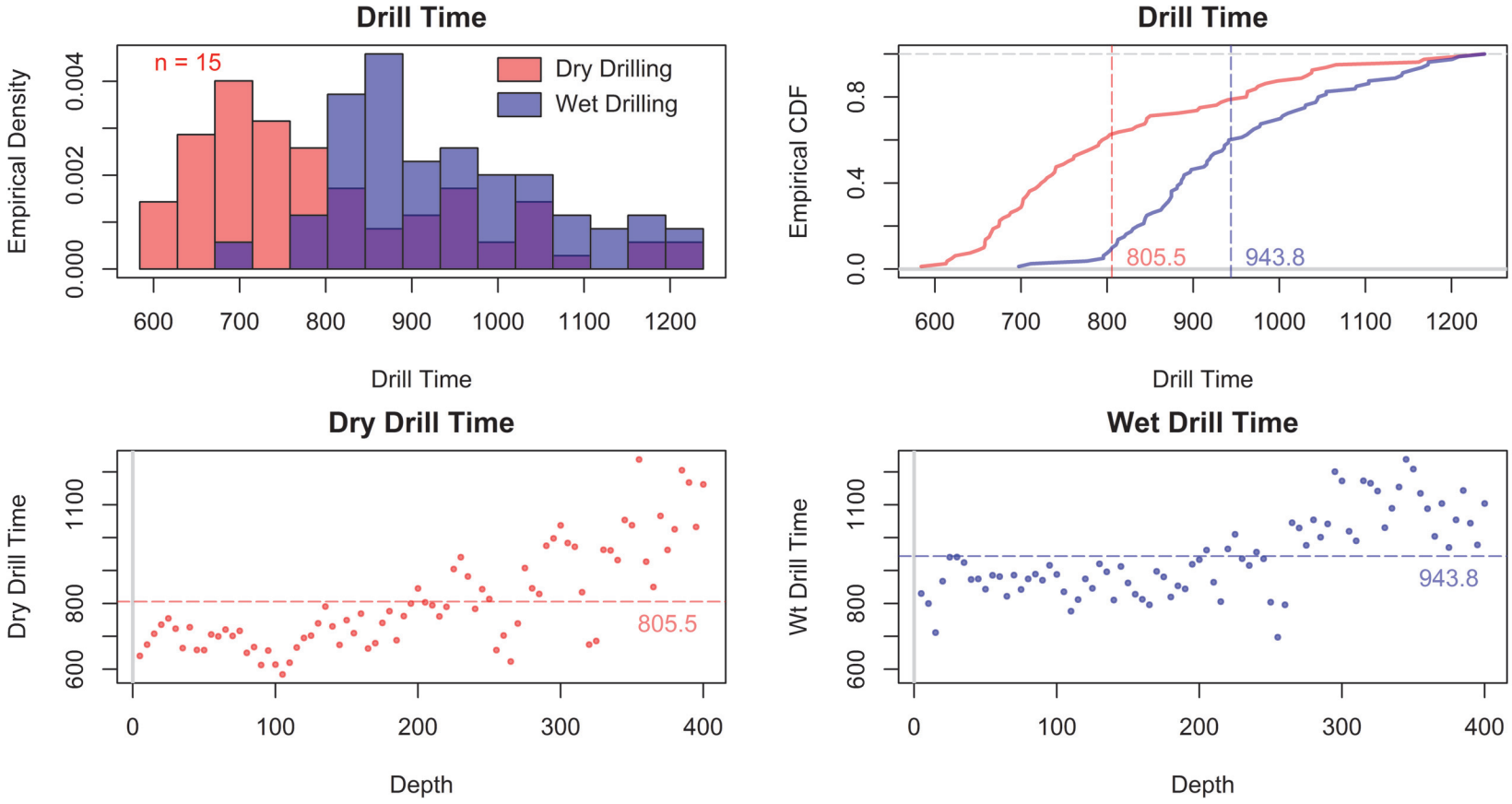
- From empirical cdf's of mean drill time it follows that **dry drilling is quicker on average than wet drilling.**

- **However, does drill time depend on the depth at which one is drilling?**





Same analysis in file "DrillTime.R"



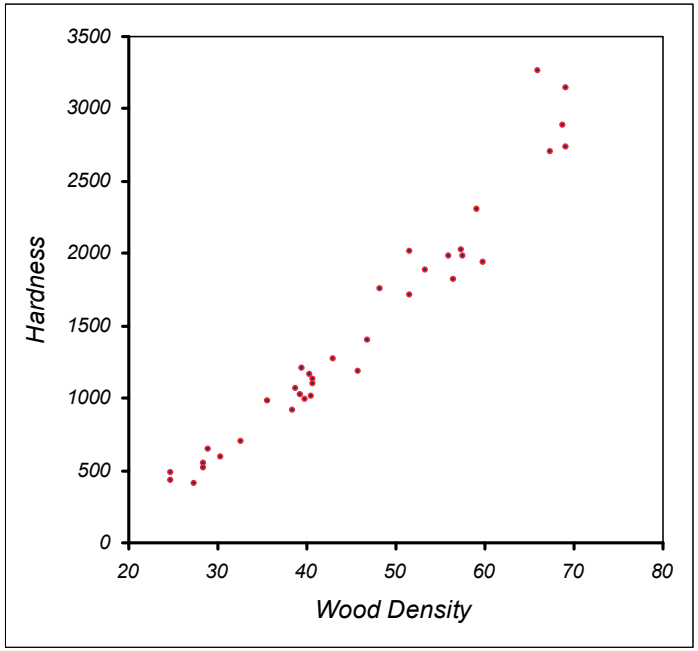
Example: Predicting hardness of Australian Timber

The Janka hardness test measures the hardness of wood. **To measure Janka hardness directly is difficult.** However, **it is related to the density of the wood**, which is **comparatively easy to measure**. In Table 15.5 **a bivariate dataset** is given of density (x) and Janka hardness (y) of 36 Australian eucalypt hardwoods.

Table 15.5. Density and hardness of Australian timber.

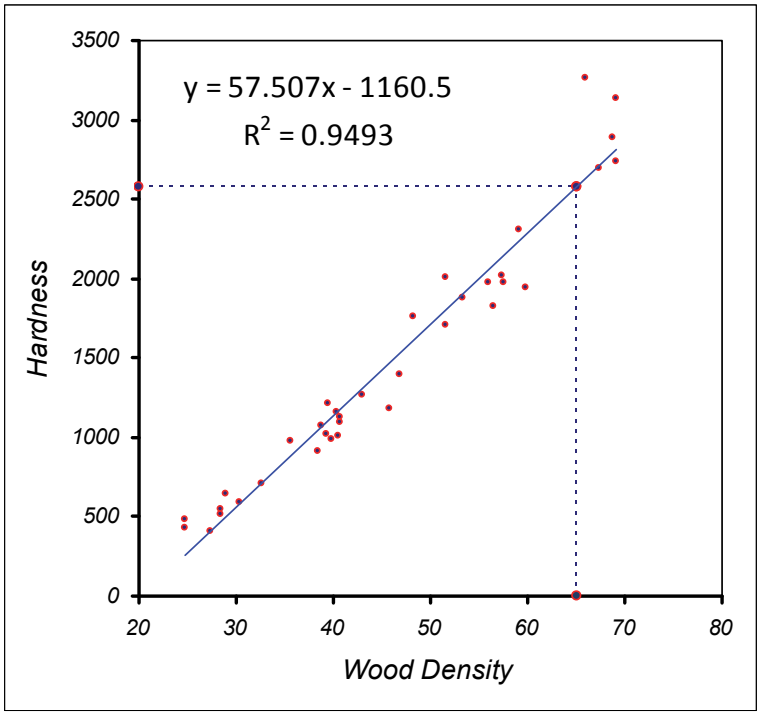
Density	Hardness	Density	Hardness	Density	Hardness
24.7	484	39.4	1210	53.4	1880
24.8	427	39.9	989	56.0	1980
27.3	413	40.3	1160	56.5	1820
28.4	517	40.6	1010	57.3	2020
28.4	549	40.7	1100	57.6	1980
29.0	648	40.7	1130	59.2	2310
30.3	587	42.9	1270	59.8	1940
32.7	704	45.8	1180	66.0	3260
35.6	979	46.9	1400	67.4	2700
38.5	914	48.2	1760	68.8	2890
38.8	1070	51.5	1710	69.1	2740
39.3	1020	51.5	2010	69.1	3140

Source: E.J. Williams. *Regression analysis.* John Wiley & Sons Inc., New York, 1959; Table 3.1 on page 43.



Exercise: Suppose we have a eucalypt hardwood tree with density 65. What would your prediction be for the corresponding Janka hardness?

Answer:



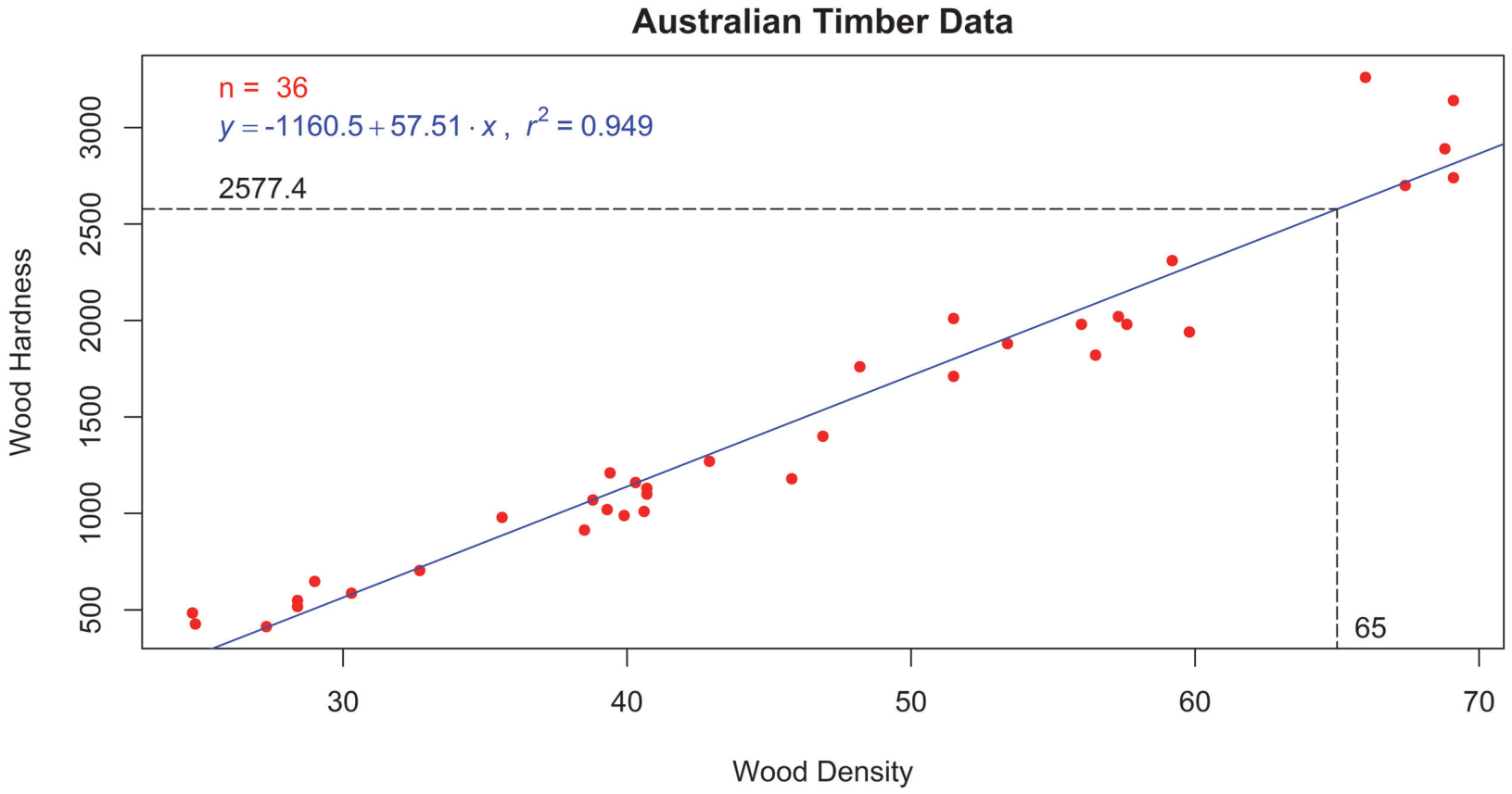
Step 1: Add **linear trendline** in MicroSoft Excel to find equation estimate:

$$y = 57.707x - 1160.5$$

Step 2:

$$x = 65 \Rightarrow y \approx 2577$$

Same analysis in file "AustralianTimber.R"



Same analysis in file "OldFaithful.R"