THE USE OF THE OLD FAITHFUL GEYSER DATA IN BOTH UNDERGRADUATE AND GRADUATE STATISTICS COURSES

By

DIDIER A. MURILLO FLOREZ

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

MATHEMATICS STATISTICS

UNIVERSITY OF PUERTO RICO
MAYAGÜEZ CAMPUS

May, 2019

Approved by:

Edgar Acuña Fernández, Ph.D
Member, Graduate Committee
Date

Edgardo Lorenzo González, Ph.D
Member, Graduate Committee
Date

Wolfgang A. Rolke, Ph.D
President, Graduate Committee
Date

Isabel Ríos, MBA
Representative of Graduate Studies
Date

Omar Colón, Ph.D
Chairperson of the Department
Date

Abstract of Dissertation Presented to the Graduate School
of the University of Puerto Rico in Partial Fulfillment of the
Requirements for the Degree of Master of Science

THE USE OF THE OLD FAITHFUL GEYSER DATA IN BOTH UNDERGRADUATE AND GRADUATE STATISTICS COURSES

By

DIDIER A. MURILLO FLOREZ

May, 2019

Chair: Wolfgang A. Rolke. PhD
Major Department: Mathematics

Many statistics teachers in line with the recommendations of several studies suggest using real-life context data during class lessons. The present work shows the use of the Old Faithful Geyser eruptions as a real-life data example for teaching statistics at a university. It can be adapted for basic courses to more advanced ones. During this study, we illustrated how to use the Old Faithful Geyser data to discuss concepts and methods related to descriptive statistics, inferences and curve fitting for an introductory statistics course. In the same way, we use the data to examine some uses in advanced class lessons that involve topics such as bootstrap method, inferences for the normal mixture models, goodness of fit tests, kernel density estimation (KDE), as well as some nonparametric regression methods, namely, kernel smoothing, LOWESS and smoothing splines.

Resumen de Disertación Presentado a Escuela Graduada
de la Universidad de Puerto Rico como requisito parcial de los
Requerimientos para el grado de Maestría en Ciencias

USO DE LOS DATOS DEL OLD FAITHFUL GÉISER EN CURSOS SUBGRADUADOS Y GRADUADOS DE ESTADÍSTICA

Por

DIDIER A. MURILLO FLOREZ

May, 2019

Consejero: Wolfgang A. Rolke. PhD
Departamento: Matemáticas

Muchos maestros de estadística, en línea con las recomendaciones de varios estudios, sugieren utilizar datos de la vida real durante las lecciones de clase. El presente trabajo muestra el uso de los datos de las erupciones en el Old Faithful Géiser como un ejemplo de datos de la vida real para la enseñanza de la estadística en una universidad. El uso de estos datos puede adaptarse desde cursos básicos hasta cursos más avanzados. Durante este trabajo, ilustramos como usar los datos del Old Faithful Géiser para discutir conceptos y métodos relacionados con estadísticas descriptivas, inferencias y ajuste de curvas para un curso de introducción a la estadística. De la misma manera, utilizamos los datos para mostrar algunos usos de estos en lecciones de clase avanzadas que involucran temas como el método bootstrap, inferencias en modelos de mezclas normales, pruebas de bondad de ajuste, estimación de densidad por kernel, así como también algunos métodos de regresión no paramétrica tales como, suavizamiento por kernel, regresión local y suavizamiento por splines.

My parents, Chiquinquirá y Francisco.

ACKNOWLEDGMENTS

To my thesis advisor, Professor Wolfgang A. Rolke, who has guided me through my studies.

To the Mathematics Departments faculty and staff for granting me the opportunity to do my Masters studies.

To my family, for the unconditional support.

To classmates and friends for their constant support in difficult times.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

KS	Kolmogorov-Smirnov.
KDE	Kernel Density Estimation.
SSE	Sum-of-Squares of Error.
LOWESS	Locally weighted regression scatterplot smoothing.

LIST OF SYMBOLS

μ	Population mean
$\bar{X}$	Sample mean
σ^2	Population variance
S_{XX}	Sum of the squares of the x's
S_{XY}	Sum of cross-products
R^2	R-squared

CHAPTER 1
INTRODUCTION

1.1 The Old Faithful Geyser

The Old Faithful is a famous geyser in Yellowstone National Park in the United States. Its eruptions are characterized by having a certain regularity; according to historical data it erupts every 43 to 96 minutes, with a duration of 1.5 to 5 minutes, and it can erupt around 20 times in a day. In this natural phenomenon there is a relationship between the duration and length of the waiting times, in general a large eruption is followed by a long waiting time and similarly a smaller eruption is followed by a short waiting time. But the Old Faithful geyser is not as faithful as the name suggests, because the time between eruptions and the length of each eruption is quite variable.

This geyser has a strange feature, eruptions with a duration in the interval between 2.8 to 3.4 minutes are not very frequent. Also, waiting times between 65 to 70 minutes do not happen often. This behavior can be explained by geological reasons and some characteristics related to water temperature (e.g. See [1], [2]). But we will not go into these aspects.

1.2 Background

Since the Old Faithful Geyser data set was collected in 1978 many statistical studies have been realized. For example, a study in [3] used this data set to create graphics based in a regression analysis. These graphs were used for educational purposes. A general analysis of data, including Markov chain and time series models, to model the pattern of waiting time and eruption duration of the geyser was realized by [4]. Also, W. Härdle used the data to illustrate various topics about nonparametric statistics [5].

A statistical learning-oriented research using the data to introduce students to statistical thinking and ideas about variability and prediction is presented in [6]. Also, an investigated about Rayleigh distributions statistical model to model eruptions can be consulted in [7]. For the Old Faithful Geyser in particular, some studies that can offer substantial insight into eruption dynamics, for example; (See [1], [2], [8]).

1.3 The data set

We will develop this work with a data set of 272 observations of eruptions of Old Faithful Geyser taken in 1978 [9]. This data set is formed of the duration of eruption and time between eruptions, both variables, waiting time and eruption duration, are given in minutes. This data set is available in `R` as a data frame with the name `faithful`.

Table 1–1: First ten observations of Old Faithful Geyser data set.

Observations	Eruptions	Waiting time
1	3.60	79
2	1.80	54
3	3.33	74
4	2.28	62
5	4.53	85
6	2.88	55
7	4.70	88
8	3.60	85
9	1.95	51
10	4.35	85

A header with the first 10 records of the data set is shown in the Table 1–1. The first record represents an eruption with a duration of 3.6 minutes, followed by a waiting time of 79 minutes. We can easily see that a short eruption would be followed by a short time interval until the next eruption and a longer eruption would be followed by a longer time interval.

1.4 Motivation

Often times in statistics education we do not have real-life data to explain or teach about particular methods, and thus it is necessary to use synthetic data. In fact, made up data can reinforce the perception in students that statistics is artificial and uninteresting [10]. Unlike using synthetic data, the use of real-life data is associated with cognitive and motivational aspects in the students learning experiences, as well as statistical thinking and reasoning [11]. In line with this, several studies recommend that statistics teachers incorporate real-life data into class lessons (e.g. See [12], [13]).

The Old Faithful Geyser data set has the advantage of being a real and still open problem, which has unique characteristics that involve variability and uncertainty. These qualities make the data ideal to teach statistics. The aim of our work is to present some uses of the Faithful Geyser data set in statistics education both graduate and undergraduate courses. The data can be used to discuss topics like descriptive statistics, inferences, simple linear and quadratic regression. Most of this work will be devoted to show the use of the data set in more advanced statistics courses that deal with topics such as normal mixture model distributions, parameter estimation, parametric and nonparametric bootstrap, nonparametric density estimation, and goodness of fit tests, as well as curve fitting, which we will approach using nonparametric methods.

This data set can be studied using a variety of statistical methods. It can be discussed in different levels of academic courses, from undergraduate to graduate level statistics.

1.5 Objectives

1.5.1 General Objective

Illustrate some uses of the Old Faithful Geyser eruption as a real-life data example in statistics education both graduate and undergraduate courses.

1.5.2 Specific Objectives

1. Use the Old Faithful Geyser data to illustrate graphs, concepts, and methods of descriptive statistics that are discussed in most introductory statistics courses.
2. Illustrate some inferences statistical and curve fitting methods on real-life context data for undergraduate statistics courses.
3. Exemplify some topics of graduate statistics courses like frequentist inferences statistical for normal mixture models, as well as a Bayesian analysis for these kind models on a real-life data context.
4. Show some uses of the Old Faithful Geyser eruption data to illustrate advanced topics such as kernel density estimate and nonparametric regression methods.

CHAPTER 2
INTRODUCTORY STATISTICS

In this chapter we will use the Old Faithful Geyser data to illustrate graphs, concepts, and methods that are discussed in most introductory statistics courses. Graphs such as the histogram, boxplot, scatterplot, and normal probability plot will be discussed using the data set; as well as basic concepts of descriptive statistics such as measures of central tendency and variability. Also, we will show how to to do point and interval estimation using the standard theory. Finally, we will illustrate the method of simple linear and polynomial regression.

2.1 Graphs

2.1.1 The Histogram

Histograms are an important and classic graph of descriptive statistics, they are uses when we have a set of continuous data. This graph, introduced by Karl Pearson [14], shows qualitative and quantitative information about the distribution of the data. A histogram can be used to get a first impression about the shape (e.g. skewness or symmetry) of the density function of the data by a simple visual check. Also, this graph permits to detect outliers.

An important step in the construction of histograms is the choice of the number of bins and their respective bandwidth. There is no single best method to determine the number of bins or classes, but there are some formulas that can suggest a number of bins. One of these is called Sturge's rule [15]. Following this rule, the data range should be split into k equally spaced bins. It is defined by following,

$$k = \lceil \log_2 n \rceil + 1. \tag{2.1}$$

In the expression 2.1 the ceiling operator means that the closest integer above the calculated value will be used. This equations can be derived from a binomial distribution [15]. If we apply the equation 2.1 we get,

$$k = \lceil \log_2(272) \rceil + 1 \simeq \lceil 8.08 \rceil + 1 = 10. \tag{2.2}$$

Above we obtained $k = 10$ bins, this means we should create a histogram with 10 bins equivalent to 10 bars. To make a histogram in `R` we can use the basic function `hist(.)`, however, in this work we use the `ggplot2` package to make all graphs. The `hist(.)` function computes, by default, the number of bins by using Sturge's method, but this number is only a suggestion. The recommended number of bins gets transfered to the `pretty(.)` function which can define a new set of intervals.To see more details about `pretty(.)` (e.g. See [16]). Figure 2–1 shows a frequency histogram of the waiting time variable.

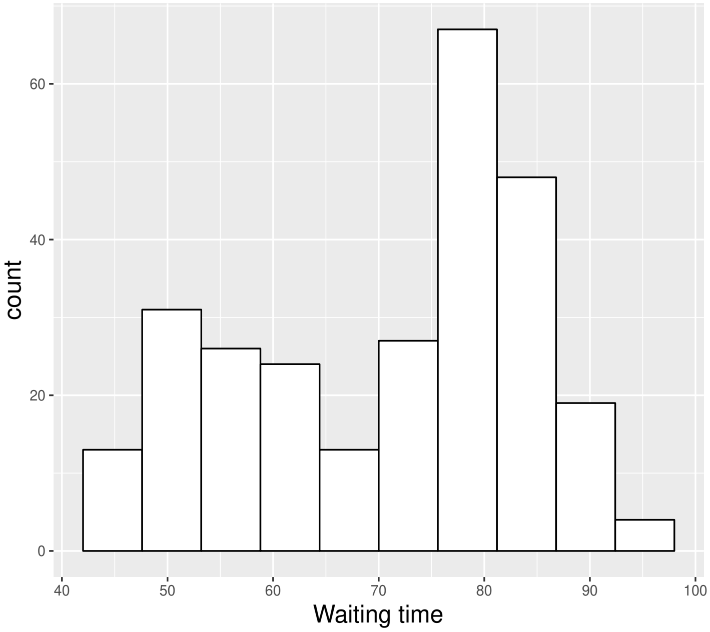

Figure 2–1: *Frequency histogram of waiting time with 10 bins by Sturge's rule.*

We can see that the form of this histogram is skewed to the left. Also, it seems to be split in two main parts and do not look bell-shaped. The center of distributions do not have many observations. In short, the histogram of the data shows that their distribution is bimodal. The form of the histogram can be justified by geological reasons (e.g See [1], [1]).

As we discussed earlier, if we use `hist(.)`, the number of bins can be different from the one recommended by Sturge's or another rule, because the `pretty(.)` function changes the number of bins to make the histogram visually appealing. But, if we want the histogram with the number of recommended bins, we must manually enter the values of the break points in the `breaks` option of the `hist(.)` function.

2.1.2 The Boxplot

The boxplot [17], is an ideal tool for conveying the measures of centrality, location, and variation of data sets. The box diagram shows the median and the quartiles of the data, and may also represent the outliers. The most frequent use of the boxplot is to compare groups. We will use two Old Faithful Geyser data sets, one from 1978 and the other from 2011, with a total of 272 observations. Figure 2–2 shows a boxplot graph of waiting time data sets.

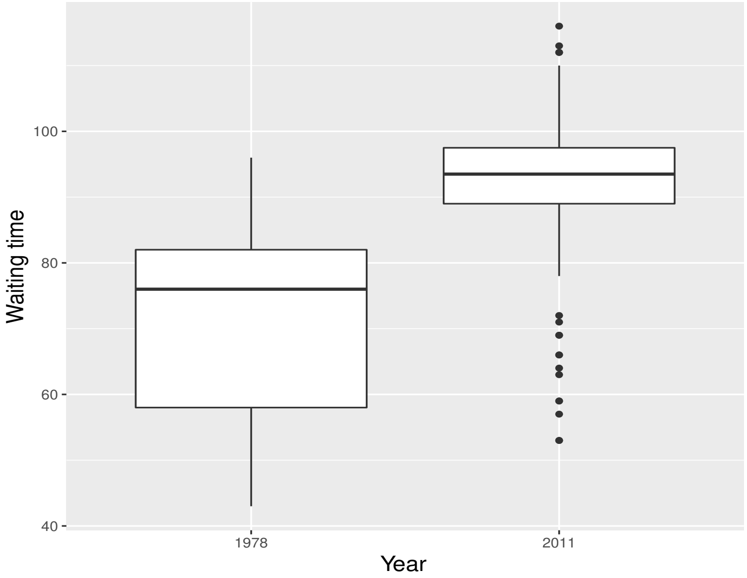

Figure 2–2: *Boxplot for the Old Faithful Geyser data sets of 1978 and 2011.*

In Figure 2–2, we can observe the changes in the behavior of the waiting time in the Old Faithful Geyser. The 2011 sample has longer waiting times and outliers. Unlike the 2011 sample, the Old Faithful Geyser data from 1978 has no outliers.

2.1.3 Normal Probability Plot

We can do a simple diagnostic about normality in the data by means of a normal probability plot [17]. If both sets of quantiles in a normality probability plot come from the normal distribution, we should see that the points form a straight line in the scatterplot. This test yield a visual representation of the data, but it is not a definitive test. Figure 2–3 depicts the normal probability plot of the waiting time variable.

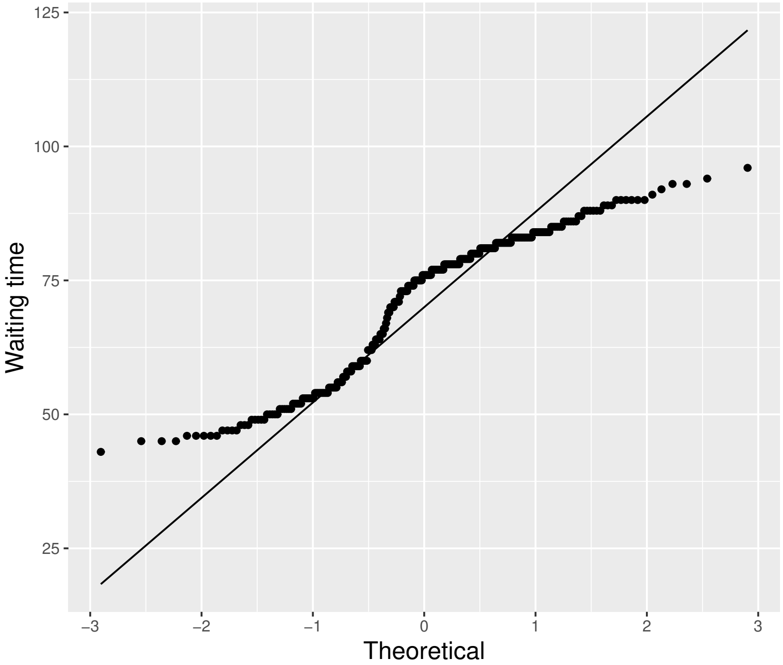

Figure 2–3: *Normal probability plot of the waiting time in the Old Faithful Geyser.*

Visually, in Figure 2–3 the probability plot of the waiting time variable does not depict the points as a linear-pattern. Also, in the center and the lower and upper extremes of the plot most points move away considerably from the straight line, which indicates that the normal distribution is not an appropriate model to describe the waiting time variable of the Old Faithful Geyser.

2.1.4 The Scatterplot

A scatterplot [17] uncovers relationships between two variables. So if we want to investigate about the possibility of the existence of relationships between the waiting time and the durations of eruptions, it is necessary to carry out a bivariate analysis of two joined variables. To describe the relationship we usually start by drawing a scatterplot. Figure 2–4 depicts a scatterplot of the waiting time and the duration of eruptions variables.

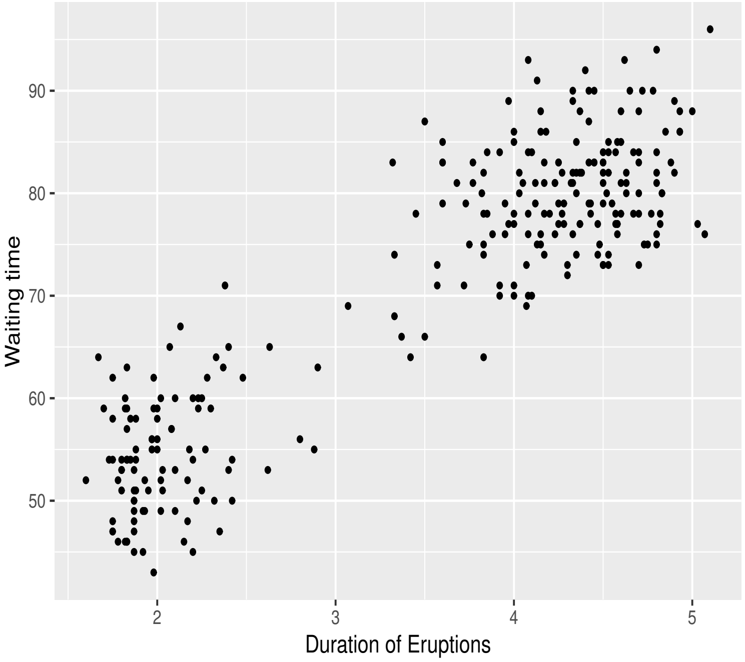

Figure 2–4: *Scatterplot for the waiting time and duration of eruptions variables of the Old Faithful Geyser data.*

In the scatterplot the x-axis and y-axis represent eruption durations and waiting time, respectively. We can observe that there is a linear relationship between the two variables. Also, we can se that the data is clumped in two distinct groups. This suggests that there are generally two types of eruptions, short-wait preceded by short-duration, and long-wait preceded by long-duration. In other words, we can see that the data segregate into two clusters. Rinehart explains the geological reasons of why the eruptions with durations in the 2.5 to 3.2 minute interval and waiting times between 60 and 70 minutes are atypical [1].

2.2 Summary statistics

2.2.1 Centrality Measures

Frequently, for a set of continuous data, the mean and the median are used as summary measures. The average waiting time in the Old Faithful Geyser, explain the trend, and centrality of this variable. We found that the average waiting time is 70.89 minutes, this means that we will expect that many waiting times will be close to 70.89 minutes. Instead of the average, we can choose the median as a centrality measure; we have that the median of the waiting times is 76 minutes, that is to say the 50% of waiting times have been less than or equal to 76 minutes. Table 2–1 summarizes these results.

Table 2–1: Summary of centrality measures of waiting time.

Statistics	Value
Mean	70.89
Median	76.00

We obtained a mean equal to 70.89 minutes, but if we look at the histogram in Figure 2–1 this measure yields misleading results. Actually, it is a value that is rarely seen in the Old Faithful Geyser. In this case, the mean is not appropriate statistic measure because the waiting time distribution is bimodal.

2.2.2 Variability Measures

Respect to the variability measures, the variance and standard deviation are usually used. Table 2–2 shows these measures to waiting time variable.

Table 2–2: Summary of waiting time variability measures for each group.

Statistics	Value
Variance	184.82
Deviation standard	13.60

The values in Table 2–2 do not contribute relevant information.

2.3 Statistical Inferences

2.3.1 Point estimation

To make a point estimation with the complete data set will be a mistake, because the results lack meaning in the context of the waiting time estimation for the Old Faithful Geyser. However, there is another option for this estimation. As we previously discussed, the data in Fig 2–4 is clumped in two distinct groups. Therefore, we can infer that the best way to analyze this data is to split it into two groups, named group 1 and 2. Group 1 (short times eruptions ≤ 3.2 minutes) will include waiting time values from 43 to 71 minutes, and Group 2 (long times eruptions > 3.2 minutes) equivalents to waiting time values over 64 to 96 minutes. Hence forth the terms group and sample will be used interchangeably. Figure 2–5 depicts the scatterplot with the groups.

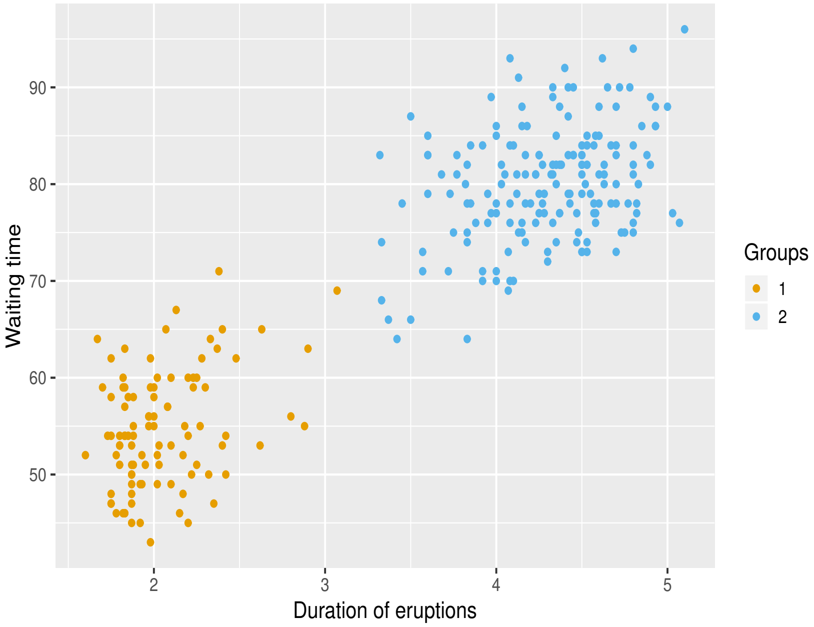

Figure 2–5: *Scatterer plot of waiting time and duration of eruptions with the groups.*

Visually, we can check the breakdown of the two groups for the waiting time in the Old Faithful Geyser data when the times eruptions are shorts or larges. Then, we have two samples with their respective population parameters.

Now, since the sample average $\bar{X}$ is an unbiased estimator of the population mean μ. We use as usual the statistic $\bar{X}$ as a point estimator of the parameter μ. In other words, we take the sample average waiting time as the point estimate of the time until the next eruption of each group. Before make this, we can make histograms for each sample just to get an idea of the shape of the underlying distributions. Figure 2–6 depicts the histograms for each group of the waiting time.

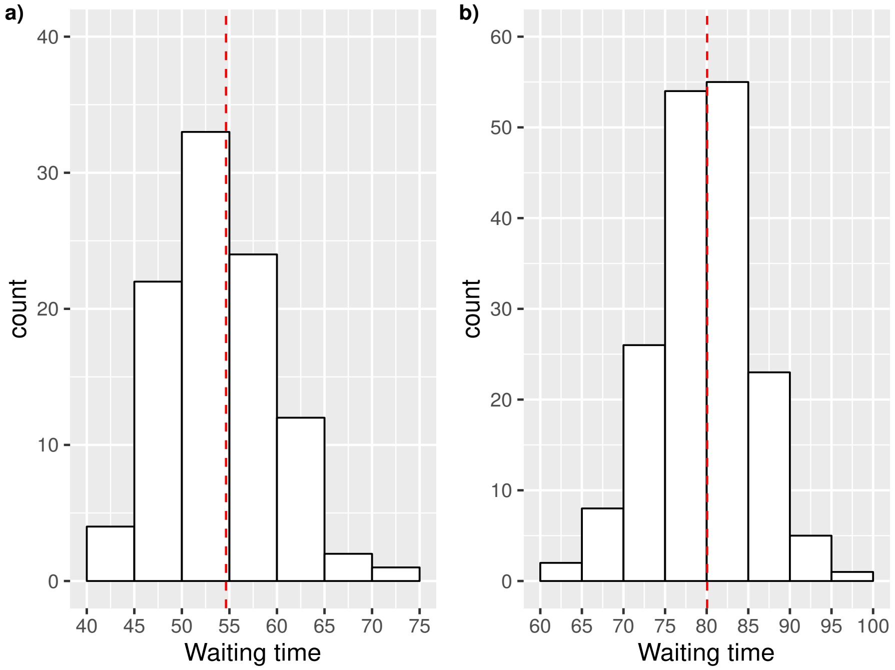

Figure 2–6: *a) Frequency histogram of the waiting time in group one. b) Frequency histogram of the waiting time in group two.*

The red dashed lines in the histograms represent the means for each group. Let $\bar{X}_1$ and $\bar{X}_2$ the mean for the two groups respectively. The point estimation for μ_1 and μ_2 are following by,

$$\mu_1 = \bar{X}_1 = 54.64 \text{ and } \mu_2 = \bar{X}_2 = 80.05 \tag{2.3}$$

The results for the means in 2.3 and for the median for each group are summarize in Table 2–3.

Table 2–3: Summary of waiting time centrality measures for each group.

Statistics	Group 1	Group 2
Mean	54.64	80.05
Median	54.00	80.00

The means estimated of each group are 54.64 minutes in group 1 and 80.05 minutes in group 2. In other words, we have that on average 54.64 minutes is the next predicted eruption if the length of duration is short and on average 80.05 minutes if the length of duration is large. The abstraction for the medians is similar. Therefore, the previous values correspond to what we would expect to occur in terms of expectation about the waiting times for each group, in this case these results have a greater significance compared with the values in Table 2–1.

Also, using point estimation we can estimate the variance and standard deviation of each group. Table 2–4 shows these estimates to both groups.

Table 2–4: Summary of waiting time variability measures for each group.

Statistics	Group 1	Group 2
Variance	36.0	34.8
Standard Deviation	6.0	5.9

The sample that has the most variability corresponds to group one with a standard variation of 6 minutes. In part this can be explained by larger sample size in group two.

2.3.2 Interval estimation

Thus far we have not yet discussed the distribution of samples 1 and 2, this means that we have not made any assumptions about the nature of their distribution. Now, to build confidence intervals from a parametric approach it is necessary to know the distribution of the data. In this case, is essential to make the assumption that the samples in both groups come from a normal distribution, that is to say that $X_1, ..., X_n$ and $X_1, ..., X_m$ were drawn from $N(\mu_i, \sigma_i)$, where $i = 1, 2$, and $n = 98$ and $m = 174$ are the length of samples 1 and 2.

2.3.3 Student t-interval

Since the population variance σ^2 is generally unknown, the construction of confidence intervals for the population mean μ will usually be based on Students t-distribution. Therefore, a $100(1-\alpha)\%$ confidence interval for the μ when the assumption about normality holds and the σ^2 is unknown is defined by,

$$\left[\bar{X} - t_{(\frac{\alpha}{2},n-1)}\frac{s}{\sqrt{n}}, \bar{X} + t_{(\frac{\alpha}{2},n-1)}\frac{s}{\sqrt{n}}\right]. \tag{2.4}$$

The expression in 2.4 is known as Student's t-interval. Following this equation we computed a 95% confidence interval of the population mean μ of the waiting time for each group. In group one the sample length is 98, therefore the value of the t-student distribution should be computed with $n-1$ degrees of freedom, this means 97. We can use the command `qt(0.975,98)` to compute the value of distribution t-student, in this case we obtained 1.98. Then,

$$\text{Lower} = 54.64 - 1.98 * \frac{6}{\sqrt{98}} = 53.44$$

$$\text{Upper} = 54.64 + 1.98 * \frac{6}{\sqrt{98}} = 55.84$$

Similarly, we can compute the confidence interval for group 2. Table 2–5 shows the results.

Table 2–5: A 95% confidence interval for average waiting time of both groups.

Group	Point estimation	Lower	Upper	Length sample
1	54.64	53.44	55.84	98
2	80.05	79.16	80.94	174

The 95% confidence intervals estimation of average waiting time is $(53.44, 55.84)$ minutes for group one, and $(79.16, 80.94)$ minutes for group two. We can see that the confidence interval in group 1 is wider than in group 2, the reason for this is that group 2's sample size is bigger, and the standard error is reduced when the sample size is larger. The confidence intervals in Table 2–5 can be computed using the function `t.test()`.

Sometimes the normality assumption in the sample is very restrictive and in many cases it does not hold. In fact, since the confidence interval in the expression 2.4 is theoretically constructed by taking that assumption, if it is not fulfilled, then the confidence interval is not totally true. We can do a simple diagnostic about normality in both samples by means of normal probability plot (See Figure 2–7). This test gives us a visual appreciation of the data, but it is not a definitive test.

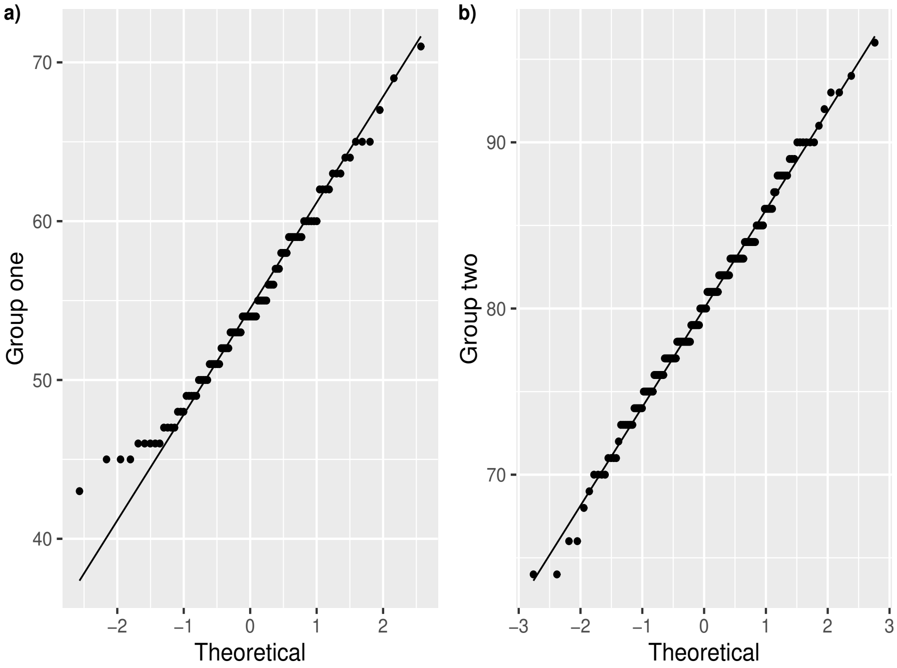

Figure 2–7: *a-b) Normal probability plot to samples one and two of waiting time.*

We can see that in both plots the majority of the points lay over the straight line of the center of distribution, with the exception some points in the left tail (Group 1). In this case we can consider the normality assumption acceptable in both cases. But, if we are unsure we could do a rigorous test like ShapiroWilk test. But, if definitively the normal assumption does not hold, a nonparametric method to do interval estimation like a Wilcoxon signed rank test can be a option.

2.4 Simple linear regression

Simple linear regression is a statistics method used to describe the relationship between two variables. In our case we are interested in describing the relationship between Y= Waiting time (response,dependent variable) and X= duration of eruptions (predictor, independent variable), and predicting the waiting times using the duration of eruptions. Given a sample n of ordered pairs (X_i, Y_i), then the model of linear regression is given by,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{2.5}$$

where $i = 1, ..., n$ and β_0 and β_1 are the $y-$intercept and the slope of the regression model respectively. ϵ_i is the random error, with the assumption that $\epsilon_i \sim Normal(0, \sigma^2)$. The simple linear regression takes the assumption that the expected value of y, $E[Y|X = x] = \beta_0 + \beta_1 x$ is linear with respect to the parameters (β_0, β_1), where β_0 and β_1 are unknown and should be estimated from the sample (X_i, Y_i). The usual method for estimating β_0 and β_1 is called least squares method. The estimates of the parameter values are frequently denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$.

In the least squared method the idea is to minimize the sum of the squares of the errors ϵ_i, with respect to β_0 and β_1. That is to say,

$$LS(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y - \beta_0 - \beta_1 x_i)^2. \tag{2.6}$$

The expression 2.6 is minimized when [9],

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}. \tag{2.7}$$

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}. \tag{2.8}$$

As proved by [9]. Finally, the estimated regression line, also called the fitted line by least squares, will be,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i. \tag{2.9}$$

When the regression line is fitted, the random error becomes an observed value and is called residual, which is represented by r_i.

The problem of predicting the waiting time in the Old Faithful Geyser can be represent by the following linear regression model,

$$\text{Waiting time} = \beta_0 + \beta_1 * \text{Duration}. \tag{2.10}$$

The linear relationship between waiting times and duration of eruptions allows carry out a linear regression fitting process. Figure 2–8 depicts the scatterplot of waiting time and duration of eruptions with the fitted line.

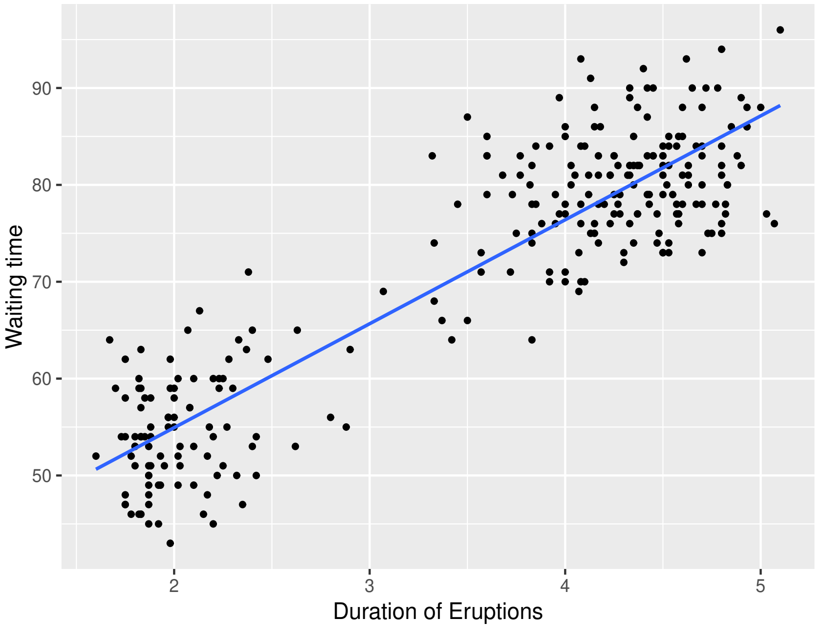

Figure 2–8: *Scatterer plot of waiting time and duration of eruptions with the fitted line.*

In `R` the function `lm` implements the least squares method. We used this function the follows way `lm(waiting ~ eruptions, data=faithful)` to fit the regression model 2.10. The output of `lm` showed that the estimate of the y-intercept is $\hat{\beta}_0 = 33.5$ and the slope is $\hat{\beta}_1 = 10.7$. With these results we can rewrite the model 2.10 as,

$$\text{Waiting Time} = 33.5 + 10.7 * \text{Duration}. \tag{2.11}$$

Equation 2.11 gives the fitted values of waiting time based on values of the duration of eruptions. In this case the interpretation about the y-intercept coefficient is nonsensical, because the duration of eruption can never have a value of zero. Now, for the slope $\hat{\beta}_1 = 10.7$; this means that the average waiting time increase 10.7 minutes when the average eruption duration increases by one minute.

To check the goodness of fit of a regression model we can use a measure called R-squared. It measures how close are the observed values of the regression line are. In practice, this value is interpreted as the amount of variance explained in the response variable by the predictor variable. In our regression model of the waiting time 2.11 we calculated an $\mathrm{R}^2 = 0.81$. This to say, the quantity of variance explained in the waiting time by the duration of eruptions is 81%. This indicate a “good” fit of the model.

2.5 Prediction

A linear regression model can be used to make predictions for the response variable for some fixed value of the predictor variable. Then, we will use the model 2.11 to make predictions of the waiting time, given a some value of the duration eruptions.

2.5.1 Confidence and prediction intervals

We can predict an individual value of the response variable as $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$, for instance, we can interested in predict the waiting time if the last eruption was 3 minutes, this is to say,

$$\text{Waiting time} = 33.5 + 10.7 * 3 = 65.6. \tag{2.12}$$

The result 2.12, indicate that given a value of duration eruptions of 3 minutes, the prediction of the average waiting time until the next eruptions be 65.6 minutes. This prediction is not bad, but it is more appropriate to predict it using an interval where the value of waiting time is expected to be in with a certain confidence level.

We can calculate confidence intervals of the average waiting time values based on values of eruption durations. Let us say we want to estimate an interval of the waiting time until the next eruption for the following values of eruption durations, 1.8, 3.2 and 4.5 minutes. We can find these intervals using the command `predict(.)`, with the argument `interval = confidence`. Table 2–6 shows the results.

Table 2–6: Linear regression 95% confidence intervals for average waiting time.

Duration	Point estimation	Lower	Upper
1.8	52.78	51.52	54.05
3.2	67.81	67.08	68.53
4.5	81.75	80.81	82.70

We can observe that the 95% confidence interval of the average waiting time for the 1.8 minute eruption duration falls within $(51.52, 54.05)$ minutes. For the 3.2 and 4.5 minute values it falls within $(67.08, 68.53)$ and $(67.08, 68.53)$ minutes respectively. Instead of calculating confidence intervals to the mean response, we want to predict an individual response of waiting time given a value of duration, this case is known as a prediction interval. Now, a similar process is done for the same values of eruption durations in Table 2–6 to calculate predictions intervals. Again, we can use the command `predict()`, but now the argument should be `interval = "prediction"`. The prediction intervals are showed in Table 2–7.

Table 2–7: Linear regression 95% prediction intervals for individual response of waiting time.

Duration	Point estimation	Lower	Upper
1.8	52.78	41.07	64.50
3.2	67.81	56.14	79.48
4.5	81.75	70.07	93.44

We can observe that a 95% confidence interval of the waiting time for the 1.8 minute eruption duration falls within $(41.07, 64.50)$ minutes. For the 3.2 and 4.5 minute values it falls within $(56.14, 79.48)$ and $(70.07, 93.44)$ minutes respectively.

2.6 Fitting Polynomial Regression

We can also considerer a quadratic regression model as an alternative to model the relationship between the duration of eruptions and waiting time. A quadratic model is a particular case of the polynomial regression when the order of polynomial is 2, with this approach is included a partially linear model with nonlinear terms in the explanatory variables. For example, given a sample n of ordered pairs (X_i, Y_i), the quadratic polynomial regression in one variable is,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i. \tag{2.13}$$

The expectation $E[Y|X = x]$ of the model 2.13 describes the quadratic function $\beta_0 + \beta_1 x + \beta_2 x^2$. In this model, β_1 is often called the linear effect parameter and β_2 the quadratic effect parameter. If the range of the data includes the zero. The parameter β_0 is is interpreted as the y-intercept. Otherwise, β_0 has no interpretation. The problem of predicting the waiting time of eruptions in the Old Faithful Geyser can be resolved using a quadratic regression model. This model is the follows,

$$\text{Waiting time} = \beta_0 + \beta_1 * \text{Duration} + \beta_2 * \text{Duration}^2. \tag{2.14}$$

Similar to the simple linear regression, we can estimate the unknown values of β_0, β_1 and β_2 through the least squared method using the `lm(.)` function. Table 2–8 shows the estimated values of β_0, β_1 and β_2.

Table 2–8: Parameter estimated for waiting time quadratic model.

Parameter	Estimation
β_0	17.25
β_1	22.17
β_2	-1.76

The least squares estimates are $\hat{\beta}_0 = 17.25$, $\hat{\beta}_1 = 22.17$, and $\hat{\beta}_2 = -1.76$.

Then, by substituting this values in 2.13, the quadratic regression model of the waiting time in the Old Faithful Geyser is,

$$\text{Waiting time} = 17.25 + 22.17 * \text{Duration} - 1.76 * \text{Duration}^2. \tag{2.15}$$

The model 2.15 can be visualized in Figure 2–9 by means of a scatterplot with the fitted curve.

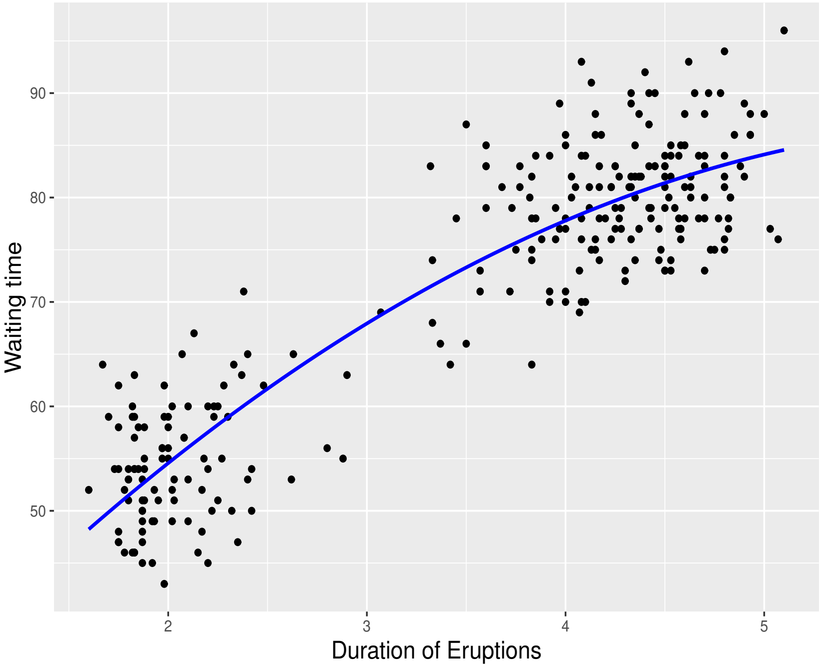

Figure 2–9: *Scatter plot of waiting time and duration of eruptions with fitted quadratic model.*

We can observe the waiting time model fitted to expression 2.15. We could think of high order polynomials as third, fourth, and fifth order, but we should remark the importance of keeping the order as low as possible. Although the higher order polynomials can be made to fit almost any pattern. However, interpretation of such models may be very difficult. Besides, the data does not merit a higher polynomial model. This models has an $R^2 = 0.82$, this means that the goodness of fit is quite better from linear regression model 2.11.

2.7 Model comparison

So far, we have discussed two ways to fit the model for the waiting time of the Old Faithful Geyser. In section 2.4 we utilized the simple linear regression. Next, we discussed the polynomial regression approach. As a consequence, we have two models to predict waiting time, and we need to choose the best model of the two. Thus, we have that the linear model, 2.11, is nested in the quadratic model, 2.15.

2.7.1 Model Comparison Using F-test

The Extra Sum of Squares Principle uses the Error Sum of Squares and F-test to compare two nested models. Where two models are nested if one of them is a particular case of the other one. In our case, the models of simple linear regression 2.11 and quadratic 2.15 for waiting time are nested since if we define $\beta_2 = 0$ in the quadratic model 2.15, it simplifies into the previous linear model 2.11. In this case, the linear model is called the reduced model and the quadratic model is named the complete model. Thus,

$$\textbf{Reduced Model:} \quad \text{Waiting time} = \beta_0 + \beta_1 * \text{Duration}$$

$$\textbf{Complete Model:} \quad \text{Waiting time} = \beta_0 + \beta_1 * \text{Duration} + \beta_2 * \text{Duration}^2$$

To test whether the quadratic terms should be included in the model, we test the null hypothesis,

$$\text{H}_0 : \beta_2 = 0$$

$$\text{H}_a : \beta_2 \neq 0$$

In short, the null hypothesis states that there is no difference between the models. To test the null hypothesis, we can use the command `anova(.)`. The p-value for this test is 0.00022, so we would reject the null hypothesis and conclude that the quadratic model is preferred over the linear model.

2.8 Residual Analysis

In linear regression, the main assumptions for the least squares regression are the homogeneity of residuals variance. On other head, the residuals should be normally distributed. To check the homogeneity of residuals variance, a commonly used graphical method is to plot the residuals versus fitted values. If the model is well-fitted, there should be no pattern to the residuals plotted against the fitted values. For assessing normality, an appropriate tool is the normal probability plot. Figure 2–10 depicts the fitted values against residuals plots and normal probability plots for both models for the waiting time (linear and quadratic).

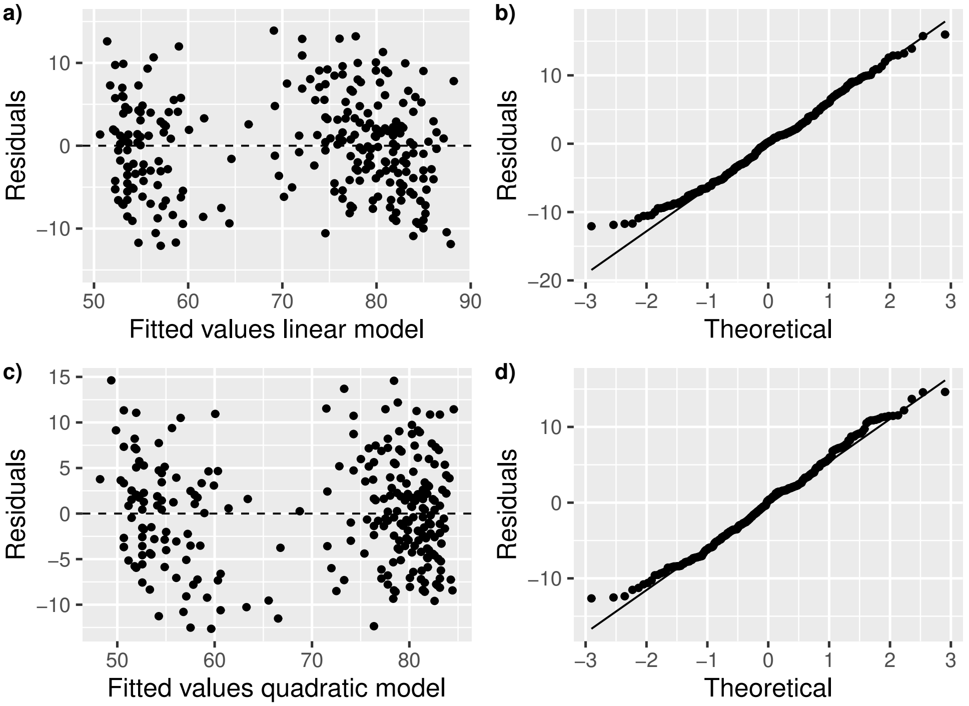

Figure 2–10: *a) Plot of fitted values linear model vs residuals. b) Normal probability plot of residuals linear model. c) Plot of fitted values quadratic model versus residuals. d) Normal probability plot of residuals quadratic model.*

Visually, the graphs show an appropriate appearance of the residuals versus the fitted values in both models (linear and quadratic), there is no pattern in the plot. This indicate that the models are well-fitted. In the same way, the assumption of normality of residuals is acceptable in both cases.

CHAPTER 3
STATISTICS GRADUATE LEVEL

In this chapter we will use the data set to discuss the confidence intervals by Bootstrap method. We will use the data to model the waiting time in the Old Faithful Geyser as a normal mixture model, as well as, show how to do point estimate in these kinds of models by Likelihood using the EM Algorithm. Also, we will apply the Kolmogorov-Smirnoff goodness of fit test for evaluating the fit of an univariate normal mixture model and how to use simulation to validate the results. Finally, we perform a Bayesian analysis for normal mixture models.

3.1 Empirical Bootstrap

Although the bootstrap idea existed already [18], it was Bradley Efron in 1979 [19] who made it popular and its usefulness as a statistical technique, thus widening its applicability and demonstrating how to implement the bootstrap effectively using computers. The main idea of the bootstrap is to construct re-samples of the observed data to estimate the variation of some statistic that is itself computed from the same data. Simply suppose we have n observations $X_1, X_2..., X_n$, which come from a unknown distribution F. Since an empirical bootstrap sample is a re-sample of the same size, this implies that $X_1^*, X_2^*..., X_n^*$ is a re-sample from the same sample. A statistic in the sample is denoted as u, but in the empirical bootstrap sample or re-sample as u^*. If the sample has a distribution F then the empirical bootstrap sample has a distribution F^*.

Theoretically, the bootstrap guarantees that $F^* \simeq F$. This means that although F is unknown, the true F and empirical distribution, F^*, are approximately equal

[20]. As a consequence, the bootstrap establishes is that the variation of the statistic u is well estimated using the variation of u^*.

In the bootstrap process, care must be taken to ensure constructing re-samples of the same size as the original sample, this is necessary because the variation of the statistic u will depend on the sample size.

3.1.1 Confidence intervals by empirical bootstrap method

The empirical bootstrap can be used to build confidence intervals with a $100(1-\alpha)\%$ confidence level around the any parameter of the population. The main goal in the process of estimating a confidence interval for any unknown parameter θ is to estimate the distribution of variation between the statistic T and true θ. These differences can be denoted as δ, and can be defined as follows,

$$\delta = T - \theta. \tag{3.1}$$

If we know the distribution of expression 3.1, then we can do a $100(1-\alpha)\%$ confidence interval for θ, that is to say that we can compute,

$$P(\delta_{1-\alpha/2} \leq T - \theta \leq \delta_{\alpha/2}) = 1 - \alpha,$$

which is equivalent to,

$$P(T - \delta_{1-\alpha/2} \geq \theta \geq T - \delta_{\alpha/2}) = 1 - \alpha.$$

Therefore, a confidence interval for θ with $100(1-\alpha)\%$ confidence level is given by,

$$\Big(T - \delta_{\alpha/2}, T - \delta_{1-\alpha/2}\Big). \tag{3.2}$$

The principle bootstrap abstraction gives a practical and simple way to estimate the distribution of δ, this means that it can be approximated through the distribution of δ^* as follows,

$$\delta^* = T^* - \theta, \tag{3.3}$$

where T^* in 3.3 is the statistics of an empirical bootstrap sample and θ is the parameter of interest. Therefore, the confidence interval for θ will be,

$$\left(T - \delta^*_{\alpha/2}, T - \delta^*_{1-\alpha/2}\right). \tag{3.4}$$

Finally, the expression 3.4 is the bootstrap $100(1-\alpha)\%$ confidence interval for the θ. In our case, we are interested for estimating the population mean of the waiting time in the Old Faithful Geyser for the two samples discussed in the section 2.3. In this case $\theta = \mu$. The methodology to find confidence intervals for μ using the empirical bootstrap principle simulation can be explained in five steps,

Confidence intervals empirical bootstrap principle simulation:

1. Given a sample $X_1, X_2, ..., X_n$ compute the expectation

$$\mu = \bar{X} = \frac{X_1 + X_2 + ... + X_n}{n}.$$

2. Generate B bootstrap samples $X_1^*, X_2^*, ..., X_3^*$ with the same size of original sample.
3. Compute the bootstrap sample mean $\bar{X}_i^*$ with $i = 1, 2, ..., B$.
4. Compute the centered sample mean of each bootstrap sample: $\delta_i^* = \bar{X}_i^* - \bar{X}$, with $i = 1, 2, ..., B$.
5. Calculate quantiles to probabilities $\alpha/2$ and $1 - \alpha/2$ into δ^*. Therefore, the confidence interval is,

$$\left(\bar{X} - \delta^*_{\alpha/2}, \bar{X} - \delta^*_{1-\alpha/2}\right),$$

   where $\delta^*_{\alpha/2}$ and $\delta^*_{1-\alpha/2}$ are the $(\alpha/2)$th and $(1-\alpha/2)$th percentiles respectively.

We computed bootstrap confidence intervals for population means in each group from waiting times of the Old Faithful Geyser eruptions. We have the samples $X_1, X_2, ..., X_{98}$ and $X_1, X_2, ..., X_{174}$ for the groups 1 and 2, respectively. We simulated $100,000$ bootstrap samples for each group.

Figure 3–1 depicts the histograms for re-sample means and differences δ^* in both groups.

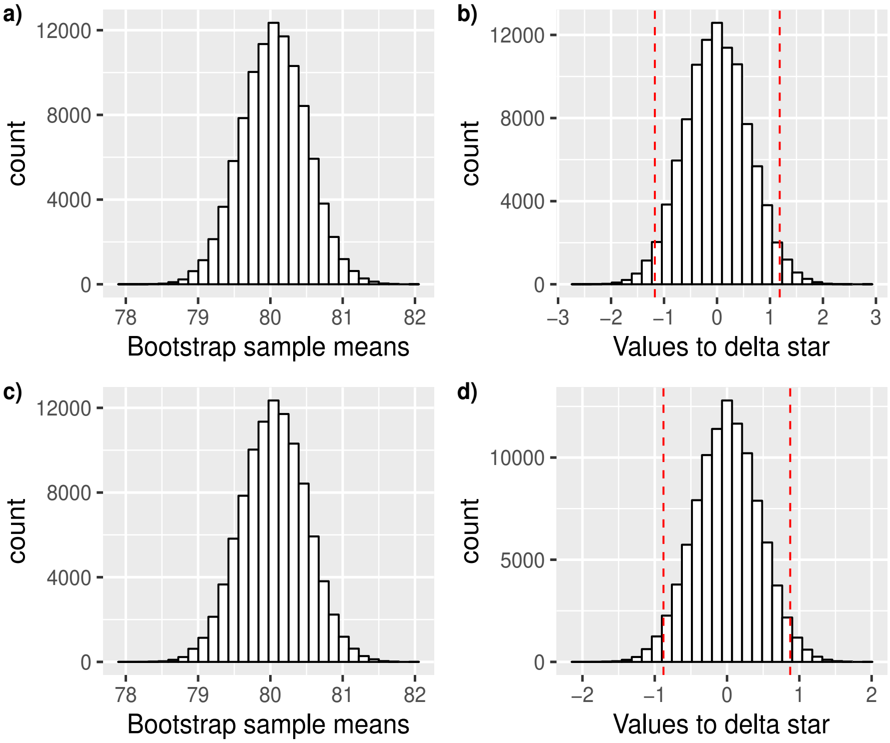

Figure 3–1: *a-b) Frequency histograms for* 100, 000 *bootstrap sample means and the differences* δ^* *in group 1. c-d) Frequency histograms for* 100, 000 *bootstrap sample means and the differences* δ^* *in group 2.*

The vertical dashed lines in Figure 3–1 refers to 2.5% and 97.5% of data, namely 0.025 and 0.975 percentiles. The summary of confidence intervals is shown in Table 3–1.

Table 3–1: Empirical bootstrap 95% confidence intervals for groups one and two.

Group	Point estimation	Lower	Upper	Bootstraps
1	54.64	53.45	55.81	100,000
2	80.05	79.17	80.93	100,000

Then, a 95% confidence interval for waiting time average in group 1 is $(53.45, 55.81)$ minutes and $(79.17, 80.93)$ minutes for group 2. If we compare these results with those obtained previously by means of Student t-interval in 2–5, we realize that they are very similar.

3.1.2 Percentile Method

As an alternative to computing the differences δ^*, the bootstrap percentile [19] method uses the distribution of the bootstrap sample statistic as a direct approximation of the data sample statistic, but this approach could be naive and misleading. It could work well, but it depends on the symmetry of the data's distribution.

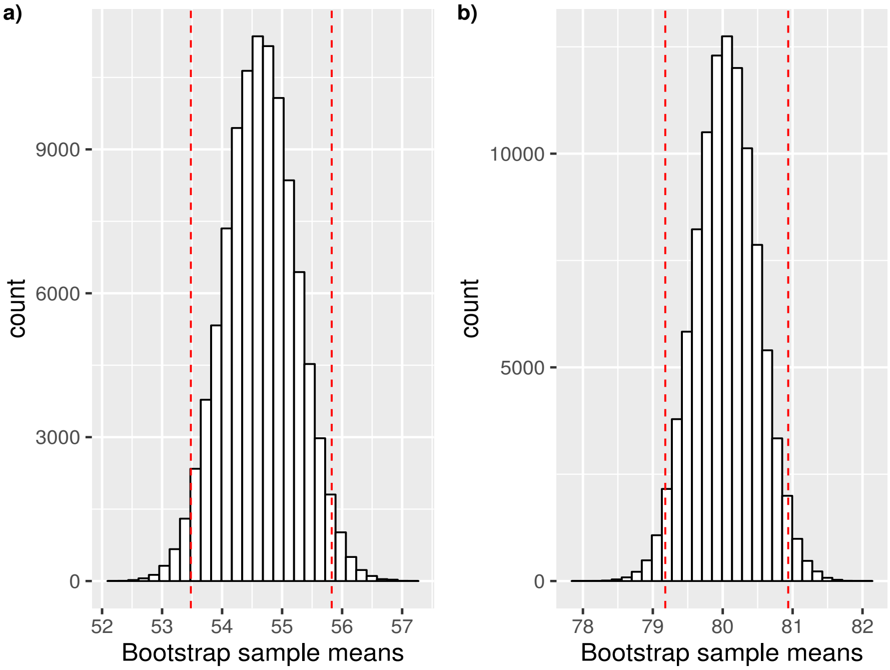

Figure 3–2: *a) Frequency histogram for* 100, 000 *bootstrap sample means in group 1. b) Frequency histogram for* 100, 000 *bootstrap sample means in group 2.*

Table 3–2 shows the confidence intervals for the average waiting time of group 1 and 2 using the percentile method.

Table 3–2: Percentile method 95% confidence intervals for groups 1 and 2.

Group	Point estimation	Lower	Upper	Bootstraps
1	54.64	53.45	55.82	100,000
2	80.05	79.17	80.93	100,000

Results shown in Table 3–2 were obtained by running 100, 000 bootstraps in each sample. Again, these results do not depart much from those obtained by *t*-interval from Table 2–5.

There are some advanced methods such as the bias-corrected (BC) bootstrap methods, but they are not part of this work, and therefore will not be discussed. In `R` the package `bootstrap` implements the bias-corrected-accelerated (BCa) [20] and ABC [20] methods through the functions `bcanon` and `abcnon`, respectively.

3.2 Waiting time as a univariate normal mixture model

As we discussed in Chapter 1, since the waiting time between eruptions has a bimodal distribution, a mixture of two normal distributions might represent the waiting time model. A normal mixture model assumes that the data is generated from k distinct normal distributions. Now, we will consider the data as only one data set and will make the assumption that the waiting time distribution is a univariate normal mixture model with $k = 2$ components. In this case, it is assumed that waiting time variables $X_1, X_2, ..., X_n$ are independent. Data will belong to either the short or long waiting time group, but we do not know which group they come from. Thus, we shall have $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$ normal distributions for each component. Therefore, we can use the following normal mixture model to estimate the waiting time eruptions of the Old Faithful Geyser,

$$f(x|\lambda, \mu_1, \sigma_1, \mu_2, \sigma_2) = \lambda N(\mu_1, \sigma_1) + (1 - \lambda)N(\mu_2, \sigma_2), \tag{3.5}$$

with $0 < \lambda < 1$. The Equation 3.5 is equivalent to,

$$f(x) = \lambda\varphi(x|\mu_1, \sigma_1) + (1 - \lambda)\varphi(x|\mu_2, \sigma_2). \tag{3.6}$$

It is clear that the function density $f(x)$ in 3.6 has five parameters, where λ is the percentage of each component and $\{\mu_i, \sigma_i\}_{i=1}^{2}$ are the means and standard deviations in each component. In this model, we have to estimate the five unknown parameters of the distribution above. For this task, we will illustrate the perform of the Maximum Likelihood Method.

3.3 Parameter Estimation

3.3.1 Point Estimation: Maximum Likelihood Estimator

We can estimate the five parameters in equation 3.5 using the maximum likelihood method. Then, as $X_1, X_2, ..., X_{272}$ are independent and normally distributed $N(\mu_i, \sigma_i)$ with $i = 1, 2$. The likelihood function is defined as follows,

$$L(\lambda, \mu_1, \sigma_1, \mu_2, \sigma_2|X) = \prod_{i=1}^{n} f(X_i) = \prod_{i=1}^{n} \left[\lambda\varphi(X_i|\mu_1, \sigma_1) + (1-\lambda)\varphi(X_i|\mu_2, \sigma_2)\right],$$

and the log-likelihood function is,

$$log(L) = l(\lambda, \mu_1, \sigma_1, \mu_2, \sigma_2|X) = \sum_{i=1}^{n} log\left[\lambda\varphi(X_i|\mu_1, \sigma_1) + (1-\lambda)\varphi(X_i|\mu_2, \sigma_2)\right]. \quad (3.7)$$

Let $\theta = (\lambda, \mu_1, \sigma_1, \mu_2, \sigma_2)$. Then, an estimate for $\hat{\theta}_{\text{MLE}}$ is provided by,

$$S(\hat{\theta}, x) = \frac{\partial l(\theta; x)}{\partial \theta} = 0. \quad (3.8)$$

The equation 3.8 is called the score function [21].

As we know, θ is a set of the five parameters, then the likelihood equation is a set of 5 simultaneous equations, defined by differentiating $l(\theta, x)$ with respect to all components of θ. We want to estimate the parameters, but unfortunately solving equations 3.7 and 3.8 is not trivial, because, it is a 5D problem and can not be maximized analytically.

We can solve the equation 3.8 using computational methods such as Newton-Raphson [22]. One problem this method is the possibility of multiple local maxima since the likelihood will have multiple roots. Aside from some other computational issues [21]. Therefore, the convergence is not guarantied in this method; however, the implementation of the Newton-Raphson algorithm in this example is somewhat difficult. For the reasons mentioned above, we will illustrate the use of Expectation Maximization Algorithm for tasking of estimating parameters.

3.3.2 Expectation Maximization Algorithm

EM is an optimization algorithm developed in 1977 [23]. This algorithm has many applications, among them, data clustering, missing values estimation, and parameters estimation in mixed models and mixture distributions. Our case is the parameters estimation in a normal mixture distributions as we saw in the expression 3.5. The algorithm (EM) is preferred over the Newton-Raphson algorithm since the convergence in the latter strongly can be affected by the starts points of parameters to be estimated. To use the EM we need a latent variable Z_i that determines the component from which the observation originates. Then we need to redefine the log-likelihood function in 3.7 so that it looks like $l(\theta|X, Z)$.

Let us say,

$$Z \sim Bernoulli(\lambda),$$

and,

$$X_i|Z_i = 1 \sim N(\mu_1, \sigma_1) \quad \text{and} \quad X_i|Z_i = 0 \sim N(\mu_2, \sigma_2),$$

where,

$$P(Z_i = 1) = \lambda \quad \text{and} \quad P(Z_i = 0) = 1 - \lambda.$$

Then, the function $f(x, z)$ is defined as follows,

$$f(x, z) = \big[\lambda\varphi(x|\mu_1, \sigma_1)\big]^{z_i} \cdot \big[(1-\lambda)\varphi(x|\mu_2, \sigma_2)\big]^{1-z_i}. \tag{3.9}$$

Now, the maximum likelihood function of $f(x, z)$ is,

$$L(\theta|X, Z) = \prod_{i=1}^{n} f(x_i, z_i) = \prod_{i=1}^{n} \big[\lambda\varphi(x|\mu_1, \sigma_1)\big]^{z_i} \cdot \big[(1-\lambda)\varphi(x|\mu_2, \sigma_2)\big]^{1-z_i}.$$

Therefore, the log-likelihood function is,

$$\begin{aligned}
log(L(\theta|X,Z)) = l(\theta|X,Z))& \\
&= \sum_{i=1}^{n} log\Big\{\big[\lambda\varphi(x|\mu_1,\sigma_1)\big]^{z_i}\cdot\big[(1-\lambda)\varphi(x|\mu_2,\sigma_2)\big]^{1-z_i}\Big\} \\
&= \sum_{i=1}^{n}\Big[z_i(log(\lambda)+log(\varphi(x|\mu_1,\sigma_1))) \\
&+ (1-z_i)(log(1-\lambda)+log(\varphi(x|\mu_2,\sigma_2)))\Big], \qquad (3.10)
\end{aligned}$$

by distributing the sum in the expression 3.10 we obtain,

$$\begin{aligned}
l(\theta|X,Z)) = \sum_{i=1}^{n} z_i(log(\lambda)+log(\varphi(x|\mu_1,\sigma_1))) \\
+\sum_{i=1}^{n}(1-z_i)(log(1-\lambda)+log(\varphi(x|\mu_2,\sigma_2))). \qquad (3.11)
\end{aligned}$$

Finally, in 3.11 we derive the maximum log-likelihood expression to 3.9, and we can see that the equation 3.11 depends on x_i and z_i, this means that the EM algorithm maximizes this function and finds the parameters according to the latent variable Z_i. It is clear that EM is an optimization algorithm, but it is not a statistical inference principle. Therefore, it provides a maximum likelihood point estimate of the parameter.

In `R`, the package `mixtools` [24] has the function `normalmixEM`; this function implements the EM algorithm to estimate parameters for univariate normal mixture models. To use the `normalmixEM` we should specify some arguments, such as the sample, initial points for the mean, standard deviations, and first component percentage. For the mean's argument `mu` we specified the vector (54.64, 80.05), similarly for the `sigma` argument we chose standard deviations (6.00, 5.90). The percentage argument `lambda` was assigned using `lambda = 0.4`. If the `lambda = NULL`, then `normalmixEM` randomly draws the value from a uniform Dirichlet distribution [24].

The output of the function `normalmixEM` with the estimated parameters are shown in table 3–3.

Table 3–3: Estimated parameters of waiting time as mixtures of univariate normal distribution

Parameters	Component 1	Component 2
Lambda	0.36	0.64
mu	54.63	80.10
sigma	5.88	5.86

The estimated values of the percentage of components, means, and standard deviations for the mixture model 3.5. The first and second component represent 36% and 64% of the eruption waiting time data. The average waiting time in the first component (short times) is 54.63 and 80.10 minutes in the second (large times). Figure 3–3 depicts the fit for waiting time as a normal mixture model.

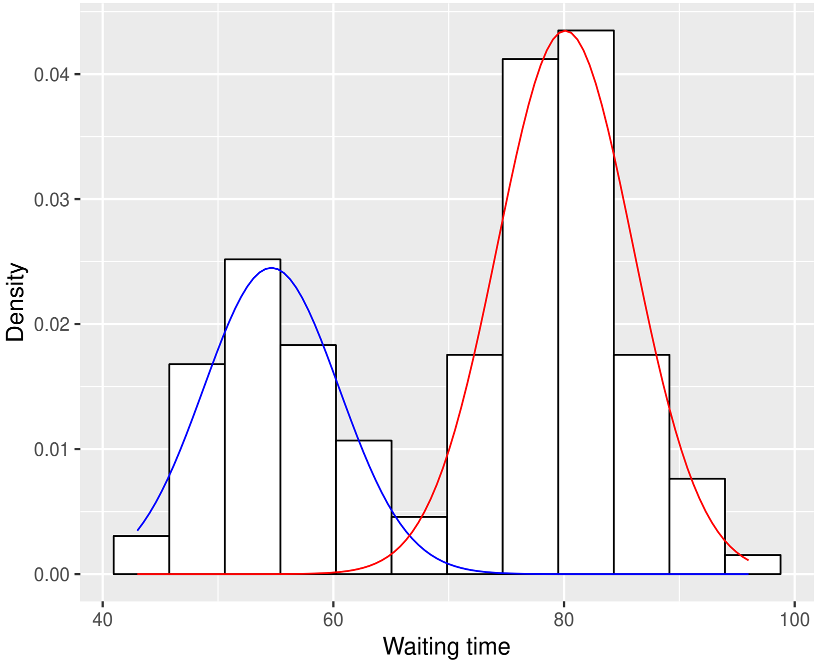

Figure 3–3: *Histogram of the two components with fitted normal curves.*

We can see the histogram for waiting times with the normal curves fitted with the parameters in Table 3–3. It means the red and blue lines indicate the two different fitted normal distributions.

According to the information above, the waiting time distribution as a mixture normal in the expression 3.5 can be rewritten with the respective estimated values of the mean, standard deviation, and percentage components as follows,

$$f(x) = 0.36N(54.63, 5.88) + 0.64N(80.10, 5.86). \tag{3.12}$$

The model in the expression 3.12 is the Maximum Likelihood estimate for the waiting time in the Old Faithful Geyser.

3.3.3 Interval Estimation

Interval estimation in mixture normal models can be done by many ways, for instance, bootstrap and maximum likelihood method. To find confidence intervals of the unknown $\boldsymbol{\theta}$ parameters in the normal mixture model of the waiting time of the Old Faithful Geyser we will discuss confidence intervals based on the maximum likelihood method. In our case, the waiting time model is a multi-parameter problem, where $\boldsymbol{\theta} = (\lambda, \mu_1, \sigma_1, \mu_2, \sigma_2)$. We assume that $X_1, ..., X_n$ are iid random variables with density function $f(x, \boldsymbol{\theta})$. Then, under some regularity conditions and the Slutsky's theorem [22], we have,

$$\sqrt{n}\big(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\big) \xrightarrow{d} Normal\big(0, \sqrt{I(\boldsymbol{\theta})^{-1}}\big),$$

where, I is the Fisher Information, which is defined by [22],

$$I(\theta)_{ij} = -E\left[\frac{\partial^i \partial^j}{\partial\theta^i \partial\theta^j} log f(x;\theta)\right]. \tag{3.13}$$

Since the Fisher Information is the negative of the expected value of the Hessian matrix of the log-likelihood function and by the law of large numbers,

$$\frac{1}{n}\mathbf{H} \to \boldsymbol{I}(\theta)_{ij}.$$

Then, a $100(1-\alpha)\%$ confidence interval for the i-th parameter is defined as,

$$\hat{\theta}_i \pm Z_{\frac{\alpha}{2}}\sqrt{I(\theta)^{-1}_{ij}}. \tag{3.14}$$

We need to find the second derivatives of each parameter θ_i, with $i = 1, ..., 5$, but this is somewhat difficult. As an alternative, we can find the Hessian matrix using some optimization functions in R. For this, we use the Non-Linear Minimization function nlm(.) with the option Hessian = TRUE. For the start points we entered the estimations of $\boldsymbol{\theta}$ found by the EM algorithm. This guarantees an even closer estimate of the true value of the parameters. Thus, we just need to solve $\mathbf{H}^{-1}$ and get $I(\theta)_{ij}^{-1}$ for each parameter.

Table 3–4 shows the estimates and 95% confidence intervals of $\boldsymbol{\theta}$ by the method of the Maximum Likelihood.

Table 3–4: Point estimations and 95% confidence interval for five parameters of the waiting time distribution.

Parameter	Point estimate	Lower	Upper
λ	0.36	0.29	0.42
μ_1	54.63	53.24	55.98
σ_1	5.87	4.81	6.92
μ_2	80.10	79.10	81.07
σ_2	5.86	5.08	6.65

In Table 3–4 we can observe that the point estimations for $\boldsymbol{\theta}$ are similar by the EM algorithm. Also, the parameter λ will be falls within (0.29, 0.42), μ_1 and μ_2 falls within (53.2 55.98) and (79.10 81.07) minutes respectively. Also, σ_1 and σ_2 falls within (4.81, 6.92) and (5.08, 6.65) minutes respectively.

3.4 Goodness of Fit

In this section we will illustrate the Kolmogorov-Smirnov goodness of fit test. We will use this tests to a formal check with a certain degree of confidence that waiting time data come from a mixture normal distribution with parameters shows in Table 3–3.

3.4.1 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov [25] is a test based on the empirical distribution function (ECDF) commonly used to decide if a sample comes from a population with a specific distribution. For example we say to have $X_1, ..., X_n$ which are continuous and independent random variable and we wish to test $H_0 : X_i \sim F$ vs $H_a : X_i \nsim F$ for all i. Thus, if the null hypothesis is true, then the empirical ECDF should be close to the true one, that is the distance between the two curves should be small. Figure 3–4 shows the fit between the ECDF of the waiting time sample and the theoretical CDF.

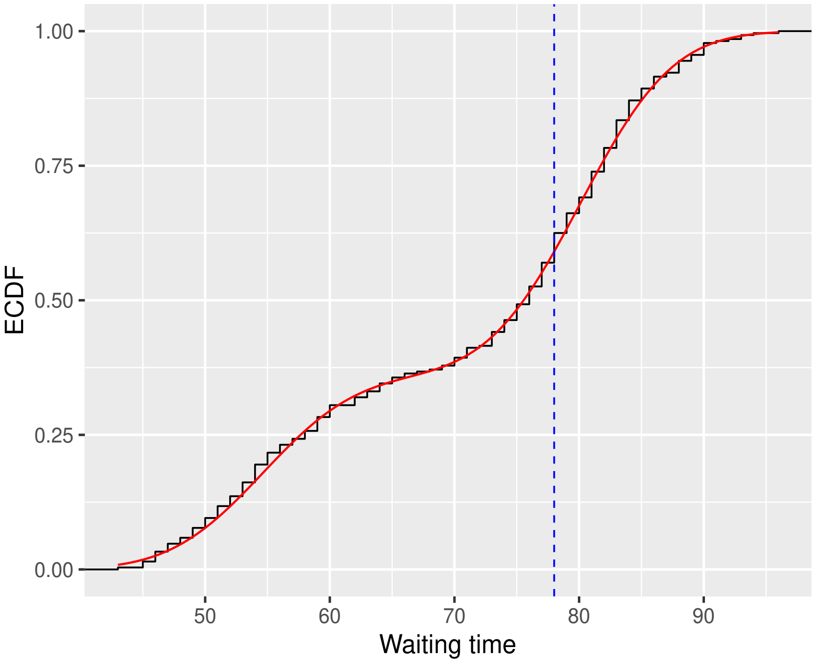

Figure 3–4: *ECDF plot for waiting time data and theoretical curve.*

We can observe that seems to perfectly fit the empirical distribution of the waiting time, where the jump points curve be refer to ECDF and the red curve to theoretical. The maximum difference occur when the waiting time is 78 minutes, according the blue vertical line. The distance D we can definite as the statistics test under of null hypothesis as following,

$$\text{D} = \sup\{|F(x) - \hat{F}(x_i)| : x \in \Re\}. \tag{3.15}$$

The expression 3.15 is called the Kolmogorov-Smirnov statistic. Then, all we need to do is find $F(X_i) - \hat{F}(X_i)$ for all i. The distribution of statistic D under the null hypothesis does not depend on F, it is known as a distribution-free statistic. The distribution of statistic D under the null hypothesis is difficult. Some literature about it (e.g. [26]).

The Kolmogorov-Smirnov goodness of fit test is implemented in `R` thought routine `ks.test(.)` of base package `stats`. In this function we should provide some arguments, as `x` refers to the sample, and `y` specifies the null hypothesis. An important remark is that in `ks.test(.)` the null hypothesis H_0 is that the data follow a specified distribution provided in the argument `y`.

Now, the respectively hypothesis to perform a K-S test are,

H_0 : *The sample data of waiting time are not significantly different than a normal mixture population.* ($\mathrm{H}_0 : X_i \sim F_{\hat{\theta}}$).

H_a : *The sample data of waiting time are significantly different than a normal mixture population.* ($\mathrm{H}_0 : X_i \nsim F_{\hat{\theta}}$).

The output of `ks.test(.)` with the waiting time data and `y = "pmnorm"`, where `"pmnorm"` is a function that compute the theoretical probabilities of data using the mixture normal distribution in 3.12 for the data is showed in Table 3–5.

Table 3–5: Output one-sample Kolmogorov-Smirnov test

Alternative hypothesis	D statistics	p-value
Two-sided	0.0337	0.9164

The value of tests statistics under null hypothesis is D = 0.0337. The p-value of this test 0.9164. With this result we do not reject the null hypothesis H_0 and we keep going with our main idea about the that waiting time distribution is a normal mixture like in 3.12. But formally, K-S test is inaccurate in this case, because the fact that we estimated the parameters from the same data that we are using to do the test, entails that be biasing the test towards failure to reject H_0 . In other words, the Kolmogorov-Smirnov goodness of fit test is not totally true when the population parameters are unknown and the sample is used to estimate them.

3.5 Simulation Studies: Adjusted p-value in K-S test

We can of course use simulation to implement such tests and try to overcome that latter bias to an extent via a parametric bootstrap.

3.5.1 Simulation of waiting time

Through of Expectation Maximization (EM) algorithm we found the five parameters to the waiting time modeled as a mixture normal distribution. Now, to simulate data drown from waiting time distribution in the equation 3.12 we should carry out the follows process or methodology.

Simulation of waiting time:

1. Generate $\text{U}_1 \sim Uniform(0, 1)$
2. If U_1 is less than λ, then generate $X_1 \sim N(\mu_1, \sigma_1)$
3. Otherwise generate $X_1 \sim N(\mu_2, \sigma_2)$
4. Repeat steps 1-3 n times.

Where λ is a percentage of first component, that means $\lambda = 0.36$ in this case, and $N(\mu_1, \sigma_1)$, $N(\mu_2, \sigma_2)$ are the normal distributions for each component with values to $\{\mu_i, \sigma_i\}_{i=1}^{2}$ showed in Table 3–3. With the process explained above, we get n random variables come from the normal mixture model for the waiting time specified in 3.12.

For example, we can generate $10,000$ observations of waiting time using this procedure and then carry out a graphical check, the first graph to validate the simulation, is just the relative-frequency histogram with the curve fit. With this, we can look if our simulation study is doing the right thing, that means actually it is generating data from 3.12 correctly. In other words, if the curve well-fit over relative-frequency histogram, then our algorithm is working very well. We can see the graph in the Figure 3–5.

Figure 3–5 is comparing the desired distribution with the simulated data. We can observe that the curve fits very well to the simulated data, where the curve

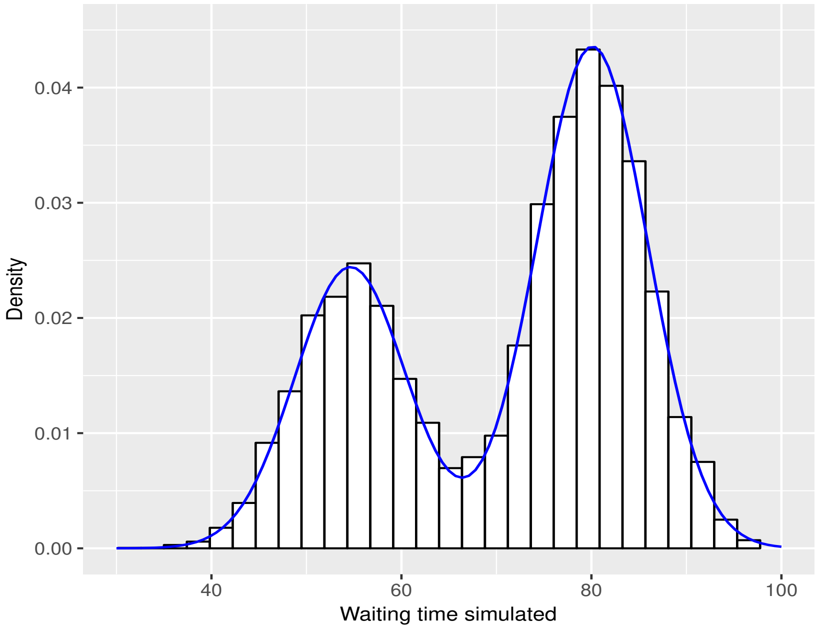

Figure 3–5: *Relative-frequency histogram of waiting time simulated data with fitted curve.*

corresponding to mixture normal distribution model of waiting time specified in 3.12.

3.5.2 Parametric bootstrap

We carried out a parametric bootstrap to try correct the latter bias in K-S test. For this purpose we can generate many samples from waiting time distribution in 3.12 using the methodology specified in the Section 3.5.1, then we estimated the parameters of the samples by EM algorithm using `normalmixEM` and calculated the K-S test for each sample using the estimated parameters. Under this construction, the null hypothesis is always true. The methodology of this simulation is describe below.

Adjusted p-values of K-S test:

1. Generate B samples from $0.36N(54.63, 5.88) + 0.64N(80.10, 5.86)$ of size 272.
2. Estimate the parameters in each sample by EM algorithm using `normalmixEM`.
3. Do the K-S test in each sample.
4. Compute the p-value as $\sum(D_{\text{sumulated}} > D_{\text{observed}})/B$.

We did ten thousand parametric bootstrap using the process above and found a p-value equal to 0.2511. Figure 3–6 shows the histogram of D statistics simulated.

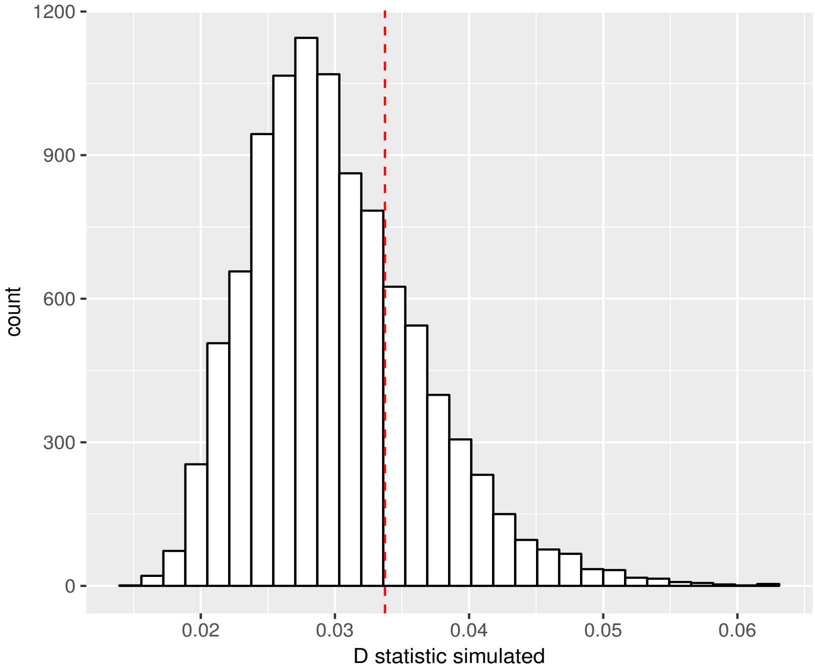

Figure 3–6: *Frequency histogram D statistics simulated.*

The dished vertical line represent the observed D statistic, hence all the values to right of the line represent the probability of finding a simulated D greater than the observed D.

Table 3–6 summary of the simulation's results.

Table 3–6: Parametric bootstrap for the Kolmogorov-Smirnov test

Simulations	D observed	p-value
10,000	0.0337	0.2511

Now, the p-value in Table 3–6 is correct and we can use it to make a decision about H_0. Therefore, the decision keep going be not reject the null hypothesis with a confidence level of 95%. In conclusion, a normal mixture is a suitable model for the waiting time in the Old Faithful Geyser.

3.6 Bayesian analysis

Let us illustrate a Bayesian analysis to find point and interval estimates for the normal mixture model parameters of the waiting time in the Old Faithful. This means we will treat the parameters $\lambda, \mu_1, \sigma_1, \mu_2$, and σ_2 as random variables.

3.6.1 Prior and Posterior Distribution

Given a sample $\mathbf{X} = (X_1, X_2, ..., X_n)$ iid random variables drawn from some probability density function $f(x; \theta)$, where θ is a value parameter lying in Θ. A Bayesian analysis begins by specifying a prior distribution $\pi(\theta)$ to observe $\boldsymbol{X}$. This prior distribution is supposed to encrypt our knowledge of the parameter before an experiment is done. Let us denote the sampling distribution by $f(\mathbf{X}|\theta)$, then the joint probability distribution function of X and θ is given by,

$$f(x, \theta) = f(\boldsymbol{X}|\theta)\pi(\theta). \tag{3.16}$$

The marginal of the distribution of $\boldsymbol{X}$ is $m(\mathbf{X}) = \displaystyle\int_\Theta f(\mathbf{X})\pi(t)\mathrm{d}t$ and the posterior distribution is the conditional distribution of θ given the sample $\boldsymbol{X}$ defined as follows,

$$\pi(\theta|\mathbf{X}) = \frac{f(\mathbf{X})\pi(\theta)}{m(\mathbf{X})}. \tag{3.17}$$

When θ is a vector of parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$, the posterior distributions in 3.17 can be generalized as follows,

$$\pi(\boldsymbol{\theta}|\mathbf{X}) = \frac{f(\mathbf{X})\pi(\boldsymbol{\theta})}{\displaystyle\int \cdots \int_\Theta f(\mathbf{X}, t)\pi(t)\mathrm{d}t}. \tag{3.18}$$

For the case of the waiting time model, we have $\boldsymbol{\theta} = (\lambda, \mu_1, \sigma_1, \mu_2, \sigma_2)$. The marginal posterior density of any single parameter in θ can be obtained by integrating the joint posterior density over all the other parameters, but this is difficult for the waiting time model. For instance, to find a credible interval for μ_1, implies knowing the marginal $\pi(\mu_1|X)$. To overcome this problem, we can use the

Metropolis-Hastings algorithm (see [27], [28]), which is common for posterior distribution simulation [21]. Then, with the sampling from the marginal posterior density of each parameter in $\boldsymbol{\theta}$, we can find credible intervals for $\boldsymbol{\theta}$ using the sample quantiles.

3.6.2 Posterior distribution by Metropolis-Hastings algorithm.

Fortunately, we do not need to know the marginal distribution to employ the Metropolis-Hastings algorithm. We just need a proposal distribution $q(\theta^*; \theta)$. The decision about whether we accept a value, θ^*, from this proposal density will be based on the acceptance ratio,

$$a(\theta^*; \theta) = \frac{q(\theta; \theta^*)\pi(\theta^*)}{q(\theta^*; \theta)\pi(\theta)} \tag{3.19}$$

With the formal notation defined, we can now construct a Metropolis-Hastings algorithm for mixture models [29].

Algorithm: Metropolis-Hastings algorithm.

The steps to find the posterior distribution using the Metropolis-Hastings algorithm are [29],

1. Choose initial values $\theta^{(0)}$. Let t = 0.
2. Sample θ^* from $q(\theta^*; \theta^{(t)})$.
3. Generate $u \sim Unif(0, 1)$, such that Unif$(0, 1)$ is taken to mean the uniform distribution over the interval $(0, 1)$.
4. Set
$$\theta^{(t+1)} = \begin{cases} \theta^{(t+1)} & a(\theta^*; \theta^{(t)}) > u \\ \theta^{(t)} & a(\theta^*; \theta^{(t)}) \leq u \end{cases}$$
5. Increment t and repeat steps 2 through 4.

Once we have a sample from the posterior distribution, we can carry out inference on the parameters such as point estimations and credible intervals. The priors will be $\lambda \sim Unif(0, 1)$, $\mu_1 \sim 1$, $\mu_2 \sim 1$, $\sigma_1 \sim 1/\sigma_1$ and $\sigma_2 \sim 1/\sigma_2$. Figure 3–7 shows histograms for the marginals posteriors of the parameters $\lambda, \mu_1, \sigma_1, \mu_2$ and σ_2 .

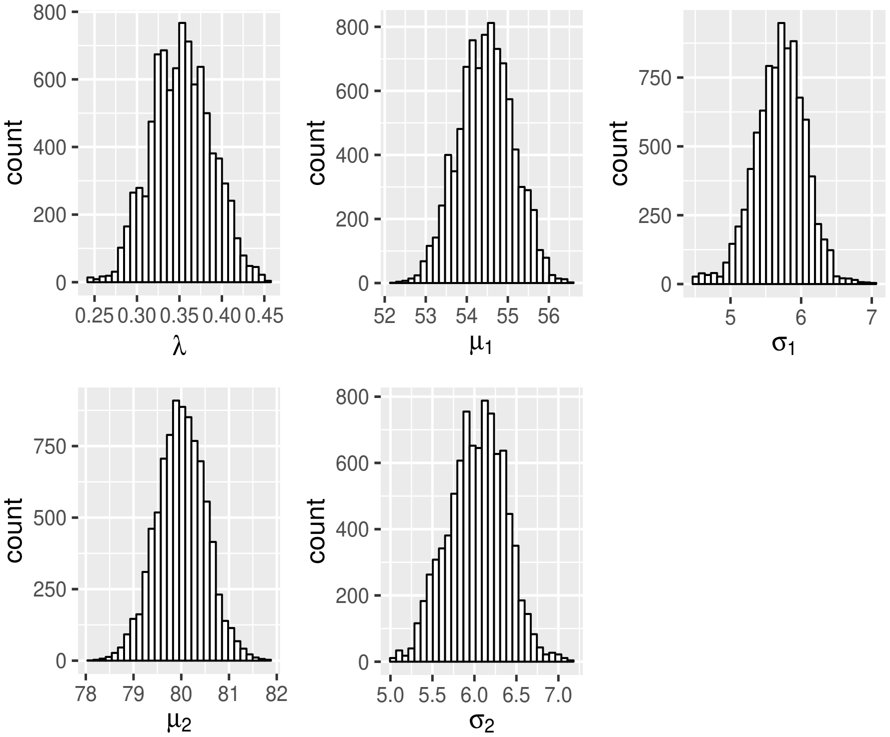

Figure 3–7: *Frequency histograms for the marginal posteriors.*

By means of the histograms, we can observe that these distributions are not skewed; therefore, it is reasonable find the mean of the posterior distribution to get an estimate of $\boldsymbol{\theta}$ for a given $X = x$. Table 3–7 shows the estimates and 95% credible intervals of $\boldsymbol{\theta}$.

Table 3–7: Estimations and 95% credible interval for parameter of waiting time model.

Parameter	Posteriori mean	Lower	Upper
λ	0.36	0.28	0.42
μ_1	54.46	53.16	55.72
σ_1	5.70	4.96	6.37
μ_2	80.00	78.96	81.00
σ_2	6.03	5.36	6.65

The estimations for $\boldsymbol{\theta}$, using the posterior mean are close from the likelihood estimates by frequentist statistics. Similarly, the boundaries for credible intervals look-alike with the confidence intervals showed in Table 3–4.

CHAPTER 4
OTHER TOPICS GRADUATE LEVEL

In this chapter, we will use Old Faithful Geyser eruption data to illustrate the kernel density estimate (KDE) method in a real-life data context. The most common selection methods for the bandwidth parameter in KDE will also be discussed. In addition, we will use the waiting time estimate problem to illustrate the most typical curve fitting nonparametric methods. We will start with the use of data to exemplify the performance of kernel smoothing, such as the Nadaraya-Watson (See [30],[31]), as well as the Locally Weighted Scatterplot Smoothing (LOWESS) [32], and finally spline smoothing (See [33], [34]). Moreover, we will illustrate how to use the cross validation method to choose the smoothing parameter in nonparametric regression.

4.1 Nonparametric density estimation

In the Section 3.2 we used the EM algorithm to find the maximum-likelihood fit for the eruption waiting time. We obtained the following model,

$$f(x) = 0.36N(54.63, 5.88) + 0.64N(80.10, 5.86). \tag{4.1}$$

The normal mixture model in 4.1 is effective for estimating the waiting time distribution. However, the choice of the normal distributions for the two components in the mixture model is open for debate, as other models that have a positive support, like the Gamma mixture model, can be use to estimate the waiting time. Now, let us carry out the estimate of the waiting time by using an estimation method that does not make any assumptions about the shape of the density. We shall use the kernel density estimate (KDE), which is a nonparametric density estimation method that uses kernel functions.

4.2 Assumptions on the data

In our problem we have n observations $X_1, X_2, ..., X_n$ of the eruption waiting time. We assume that these observations configure a random sample from a continuous population. That is to say, the random variables X_i are iid with density function f. But, we do not know the density function f of where the data comes from, however this information is not necessary.

4.3 Kernel Density Estimation

4.3.1 The histogram

A histogram is the first way to estimate the density function f. In this approach the idea is to aggregate the data in intervals of the form $[x_0, x_0+h)$ and then use their relative frequency to estimate the density at x in these intervals. For this, an origin point a_0 and a bandwidth $h > 0$ are necessary. Therefore, the histogram builds a piecewise constant function using the data within the intervals. These intervals are defined as follows,

$$\{[a_k, a_{k+1}); a_k = a_0 + (k-1)h,\ h > 0,\ k = 1, ..., j\} \tag{4.2}$$

We will refer to these intervals in 4.2 as bins, where k represents the kth bin. Then, the f estimated by means of the histogram is,

$$\hat{f}_{\text{hist}}(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathrm{I}_{[a_k, a_{k+1})}(X_i), \tag{4.3}$$

In 4.3 each sample point in a bin adds $\dfrac{1}{nh}$ to the height of the estimate for all points x within $[a_k, a_{k+1})$. An advantage of this method is that it is easy to compute, but a disadvantage is its annoying dependence on a_0. Also, it is not a smooth estimator of f, because $\hat{f}_{\text{hist}}$ is continuous to pieces, but in reality f is a continuous function.

4.3.2 Centered histogram

This method tries to overcome the non-smoothness problem in the estimate obtained by the histogram in 4.3. This method takes bins with fixed width $2h$, these bins will be centered on the point at which the density estimation is desired. With this approach, the estimation of f changes to,

$$\hat{f}_{\text{hist}}(x) = \frac{1}{2nh}\sum_{i=1}^{n} \mathrm{I}_{[x-h,x+h)}(X_i). \tag{4.4}$$

An advantage of centered histograms is that there is an exact computation and plot of estimate for all X in the sample. Also, we only depend on the parameter h, because in this type of histogram the points are considered to be near x by being no more than $h/2$ away from x. Figure 4–1 shows the centered histogram for the eruption waiting time.

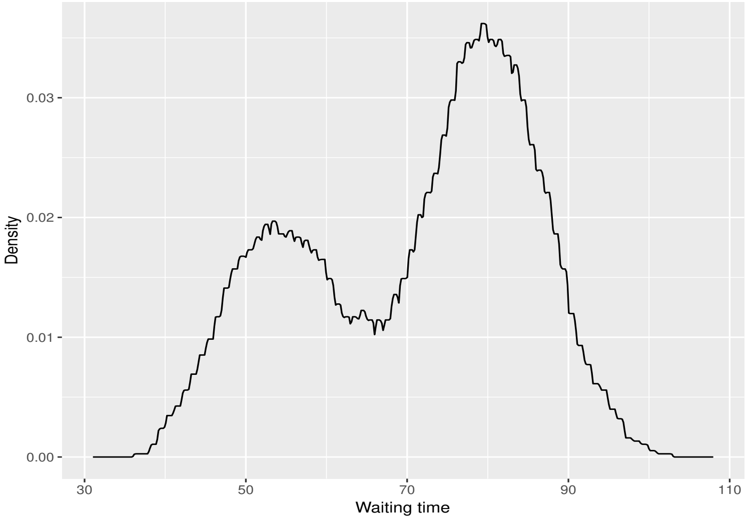

Figure 4–1: *Centered histogram estimate for eruption waiting time with bandwidth $h = 4$.*

We can see that the resulting graphs of this method are not as blocky as the estimation in 4.3, but the shape is still discontinuous. This method continues to be a non-smooth estimator of the waiting time density.

We can rewrite the estimate 4.4 in the following way,

$$\hat{f}_{\text{hist}}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\Big(\frac{x - X_i}{h}\Big), \tag{4.5}$$

where, $K(z) = \frac{1}{2}\mathrm{I}_{(-1,1]}(z)$.

According to 4.5 the estimate of f is assembled by placing a box of width $2h$ and height $\frac{1}{2nh}$ on each observation and then summing to obtain the estimate [35]. The estimation above will still be non-smooth, this problem can be solved by adding a smooth function. We can use the function `density(data,kernel="r")` to build the centered histogram in `R`.

4.3.3 Kernels

The idea is to replace in 4.4 the indicator function by a smooth function, centered at the midpoint of the bin and symmetric. Then the expression 4.3 keep have the same form, but now K is a continuous function,

$$\hat{f}_n(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big). \tag{4.6}$$

In this case, K is a positive function called kernel (smooth function). The density estimation using Kernel is actually a weighted average by the distance of the observations to the point to be estimated. The greater the distance from the point to an element of the observations, the lower its weight in the estimate. The weight will be determined by the chosen Kernel and the value of h. The greater the value of h, the greater the weight of those elements from the observations that are far from the point, so a h is usually called bandwidth. A kernel function must fulfill the following properties,

(i) $K(-x) = K(x)$

(ii) $\displaystyle\int_{-\infty}^{\infty} K(x)\mathrm{d}x = 1$

Above we see that the Kernel K is a nonnegative function and the integrate should be 1. In fact, the property (ii) implies that,

$$\int_{-\infty}^{\infty} xK(x)\mathrm{d}x = 0 \qquad \text{and} \qquad \sigma_K^2 = \int_{-\infty}^{\infty} x^2K(x)\mathrm{d}x > 0.$$

There are several kinds of kernels, but in this work we just show the estimate of density the following,

$$\begin{array}{lll} \textbf{Rectangular Kernel:} & K(x) = \frac{1}{2} & -1 \leq x < 1 \\ \textbf{Gaussian Kernel:} & K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} & -\infty < x < \infty \\ \textbf{Epanechnikov Kernel:} & K(x) = \frac{3(1-x^2)}{4} & -1 \leq x < 1 \\ \textbf{Triangle Kernel:} & K(x) = 1 - |x| & -1 \leq x < 1 \end{array}$$

The Rectangular Kernel yields the centered histogram as a particular case, but the other kernels produce a continuous estimate of f. Figure 4–2 shows the KDE of the waiting time with the Rectangular, Epanechnikov, Gaussian and Triangular Kernel.

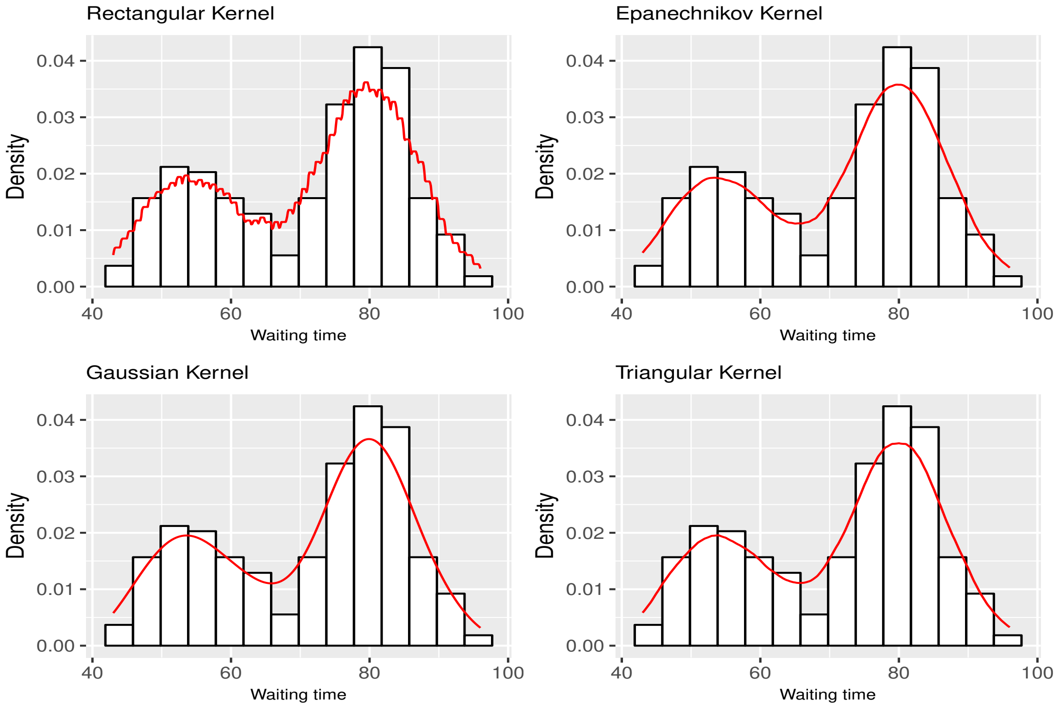

Figure 4–2: *Frequency histograms with kernel density estimates of waiting time using the Kernels: Rectangular, Epanechnikov, Gaussian and Triangular.*

According to [36], the kernel choice does not have a significant impact on the statistical properties of kernel density estimates, since all of these work fine. Also, the Epanechnikov Kernel seems like an appropriate choice because of the ease of calculation. It should be noted that the choice of the kernel type is only important in exceptional cases, for example if the underlying distribution has the Boundary Problem [37]. Unlike the choice of the kernel, the choice of bandwidth is a crucial part of kernel density estimation.

4.4 Bandwidth selection

Since the KDE method has a strong dependence on the bandwidth h, we need to choose the value of h properly. There are several bandwidth selector methods to achieve this. The goal is to estimate f, but this must be done with the best approximation to the true function. Figure 4–3 depicts the different KDE of the waiting time using four values to bandwidth.

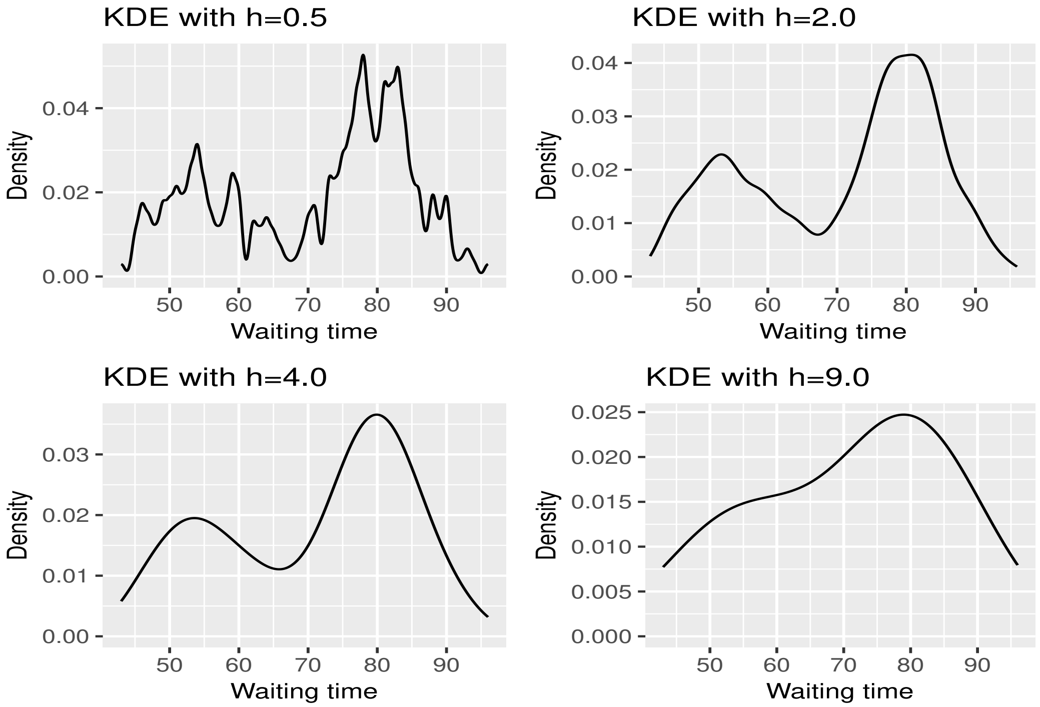

Figure 4–3: *KDE of waiting time using different values of the bandwidth;* $h = 0.5, 2.0, 4.0$ *and* 9.0*.*

We can observe that when modifying the bandwidth, considerably different estimates of waiting time distribution are obtained. So a bandwidth of 0.5 produces an evident under-smoothing, while a bandwidth of 9.0 yields $\hat{f}$ over smoothing. Then, we could think that a bandwidth of 4.0 is reasonable, but this choice must be done in such a way that minimizes the estimation error. In summary, small values of h lead to very spiky estimates (not much smoothing) while larger h values lead to over-smoothing. In the process of selection, the optimal bandwidth for the estimation of $\hat{f}$ at point x, we could consider criteria, such as the Mean Square Error

(MSE) and the Mean Integrate Square Error (MISE), which are defined below,

$$\mathrm{MSE}[\hat{f}(\cdot;h)] = E\Big(\hat{f}(x;h) - f(x)\Big)^2. \tag{4.7}$$

$$\mathrm{MISE}[\hat{f}(\cdot;h)] = E\left(\int \Big(\hat{f}(x;h) - f(x)\Big)^2\right). \tag{4.8}$$

In accordance with [38], the result of solving the (MISE) of KDE in 4.8 is as follows,

$$\mathbf{MISE}[\hat{f}(\cdot;h)] = \frac{1}{nh}\|K\|_2^2 + \frac{1}{4}h^4\mu_2^2\|f''\|_2^2 + O(nh)^{-1} + O(h^2). \tag{4.9}$$

In equation 4.9, the higher order terms are ignored [38], and the resulting set is called Asymptotic Mean Integrated Squared Error (AMISE),

$$\mathbf{AMISE}[\hat{f}(\cdot;h)] = \frac{1}{nh}\|K\|_2^2 + \frac{1}{4}h^4\mu_2^2(K)\|f''\|_2^2. \tag{4.10}$$

Using 4.10, we can determine the optimal bandwidth h_0 which minimizes this function with respect to the parameter h.

$$h_0 = \left[\frac{\|K\|_2^2}{n\mu_2^2(K)\|f''\|_2^2}\right]^{1/5}. \tag{4.11}$$

This final result seems appropriate, but the presence of the f'' makes it difficult to solve for h, so we do not know f, let alone f''.

4.4.1 Rule-of-thumb

If we assume a parametric form for the density f, we have a simple solution to 4.11. Therefore, we can find an h that is optimal for that particular shape, for example if we know that the data come from a normal distribution $N(\mu, \sigma)$. So with this assumption and using a Gaussian kernel, Silverman in 1986, obtained the most popular method by a plug-in called **rule-of-thumb** [35].

$$h = \left[\frac{\|K\|_2^2}{n\mu_2^2(K)\|f''\|_2^2}\right]^{1/5}\hat{\sigma}. \tag{4.12}$$

For estimating the standard deviation of X, Silverman suggests employing the minimum of both quantities [35],

$$\hat{\sigma} = \mathbf{min}\left(s, \frac{\mathrm{IQR}(x)}{1.34}\right). \tag{4.13}$$

In other hand, the Gaussian kernel choice implies that $\mu_2(K) = 1$ and $\|K\|_2^2 = \frac{1}{2\sqrt{\pi}}$. Therefore, the **rule-of-thumb** [35] is,

$$h = \left(\frac{4}{3}\right)^{1/5} \hat{\sigma} n^{-1/5} \simeq 1.06 \hat{\sigma} n^{-1/5} \tag{4.14}$$

The rule-of-thumb [35], equation 4.14, is implemented by the function `bw.nrd(.)`. Also, Scott in 1992 proposed a variation of rule-of-thumb by [39], where a factor of 0.9 is used instead of 1.06, obtained by the function `bw.nrd0(.)` in `R`. Figure 4–4 illustrates the KDE of the waiting time using the two versions of rule-of-thumb and reference density (MLE).

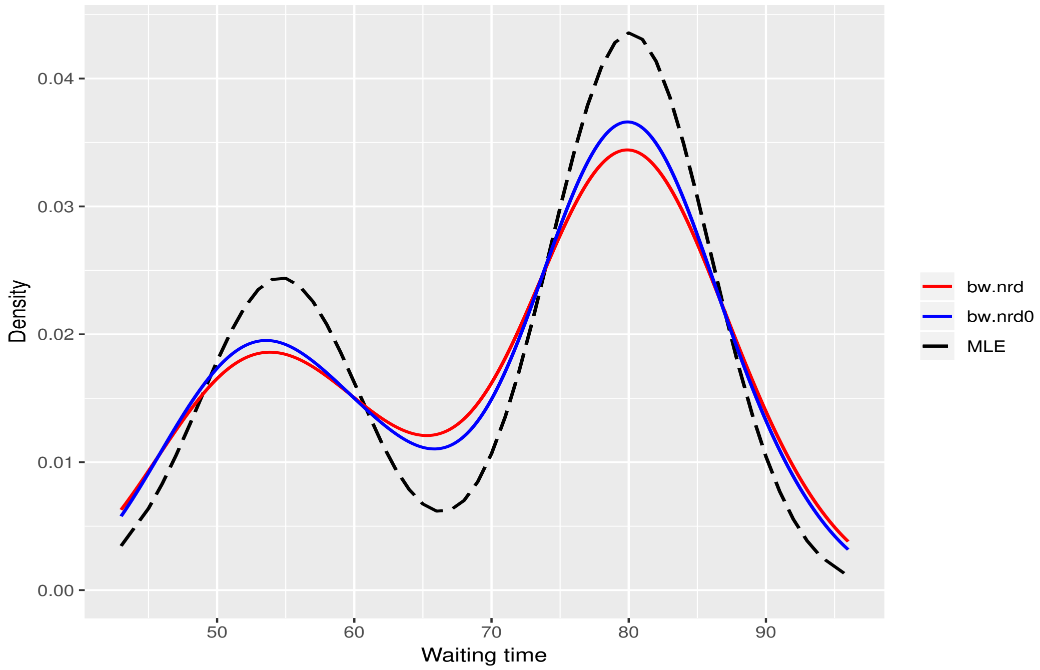

Figure 4–4: *Reference density and KDE of waiting time with the two rules-of-thumb and reference density (MLE).*

Visually, the estimates of f using these methods differ practically only in the maximum and minimum. Moreover, these estimates differ completely with respect to reference density (MLE) of waiting time. It is evident that the parametric assumption of normality makes the KDE very dependent on the sample and for obvious reasons, these bandwidth selection methods do not work well with the bimodal distribution of the waiting time in the Old Faithful Geyser. The `density()` function does KDE with the second version's rule-of-thumb, `bw = "bw.nrd0"`, by default.

4.4.2 Cross-validation

Unbiased cross-validation (UCV), also called Least Squared Cross validation (LSCV) [38] is a classic method that estimates the bandwidth in a different way from rule-of-thumb. With this method, the main task is to try to minimize the MISE using the sample in two stages: one for finding the KDE and other for checking its performance on estimating f. To avoid the dependence on the sample, the data used for computing the KDE is not used for its evaluation. Starting by remembering the MISE,

$$\begin{aligned} \text{MISE}[\hat{f}(\cdot;h)] &= E\left(\int \left(\hat{f}(x;h) - f(x)\right)^2 \mathrm{d}x\right) \\ &= E\left(\int \hat{f}(x;h)^2 \mathrm{d}x - 2\int \hat{f}(x;h) f(x)\mathrm{d}x + \int f(x)^2 \mathrm{d}x\right). \end{aligned} \tag{4.15}$$

To minimize MISE in (4.15), the term $E\left(\int f(x)^2 \mathrm{d}x\right)$ is ignored (see [38]) because it does not depend on h. Then, we can make an unbiased estimate by [38],

$$\text{UCV}(h) = \int \hat{f}(x;h)^2 \mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{(-i)}(X_i;h). \tag{4.16}$$

The term $\hat{f}_{(-i)}(X_i;h)$ in (4.16) refers to leave-one-out KDE and is based on the sample with X_i removed. Where it is defined by,

$$\hat{f}_{(-i)}(X_i;h) = \frac{1}{n-1}\sum_{\substack{j=1 \\ i\neq j}}^{n} K_h(x - X_j).$$

Finally, the unbiased cross-validation (UCV) selection method is defined as,

$$\hat{h}_{\text{UCV}} = \underset{h>0}{\operatorname{argmin}}\{\text{UCV}\}. \tag{4.17}$$

To find $\hat{h}_{\text{UCV}}$, it is necessary to make use of numerical optimization, that is, a very different method to the rule of thumb (plug-in selectors). In `R`, the function `bw.ucv()` finds the value of h by unbiased cross-validation (UCV). Figure 4–5 shows the plot of UCV and the KDE of the waiting time using the UCV selection method.

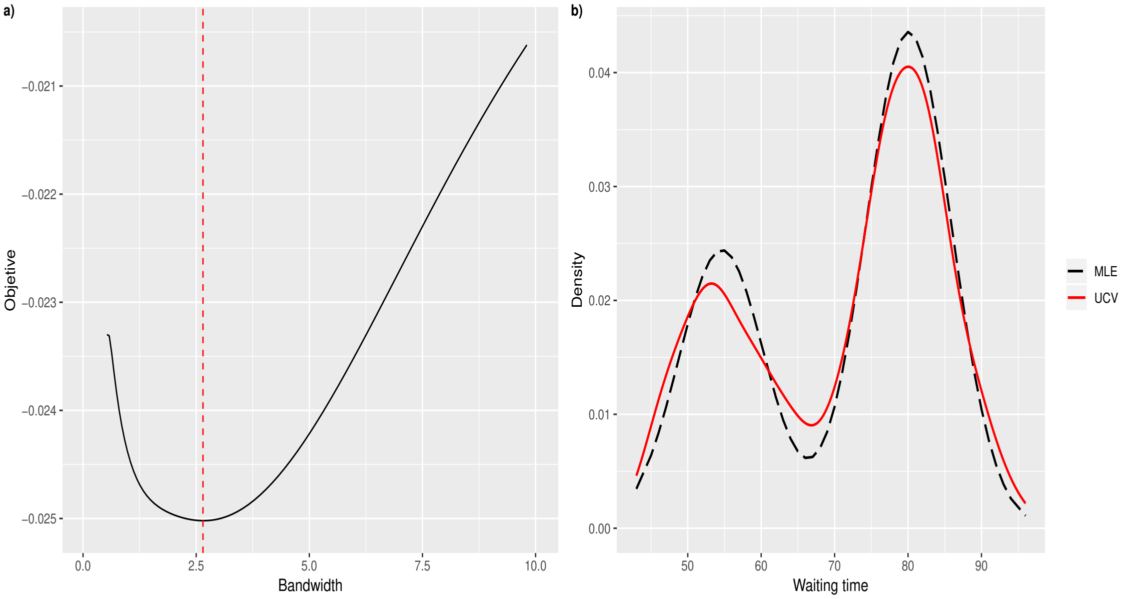

Figure 4–5: *a) Plot of the UCV selector. b) KDE of waiting time with the UCV selector method of bandwidth and reference density (MLE).*

In Figure 4–5 a), we can observe the plot of UCV bandwidth selector, where the UCV is minimized when $h = 2.65$, as indicated by the vertical line. On the other hand, in Figure 4–5 b) we can observe the KDE of waiting time using the bandwidth UCV selection method. There is a significant difference between the KDE and the reference distribution (MLE) in the maximums and minimums, but we can see a better fit with respect to the reference distribution, if it is compared to the fit found by the rule-of-thumb. Nonetheless, the estimates are quite close.

Another cross validation selection method is based on biased cross-validation (BCV) [38]. The BCV selector presents a mixture procedure that combines plug-in and cross validation ideas. The attractive property of h_{BCV} is that it has a considerably smaller variance compared to h_{UCV}, but this reduction in variance comes at the price of an increased bias (See [38]). This method is implemented in R through the function `bw.bcv(.)`. Figure 4–6 shows the biased cross-validation plot and the KDE of waiting time with this bandwidth selector.

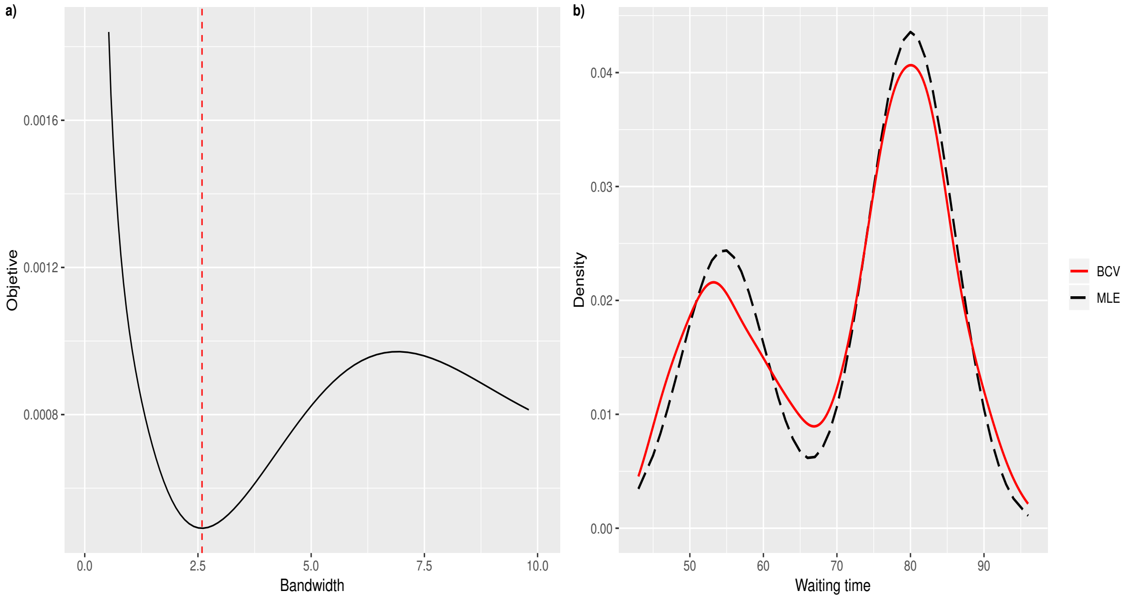

Figure 4–6: *a) Plot of the biased cross-validation selection method. b) KDE of waiting time with the BCV method and reference density (MLE).*

In Figure 4–6 a), the plot of UCV bandwidth selection method is illustrated, where the UCV is minimized when $h = 2.60$, as indicated by the vertical line. In Figure 4–5 b), however, we can observe the KDE of waiting time using the bandwidth BCV selector method. Again there is a significant difference between the KDE and the reference distribution (MLE) in the maximums and minimums. However, the estimates are also quite close.

In summary, bandwidth selector methods based on cross-validation offer smaller values of h compared to plugin-in methods, such as based on the rule-of-thumb.

There are other selector methods based on direct plug-in (DPI) (see [40]). Also, bootstrap methods can be used to select the bandwidth in KDE (see [41]).

4.5 Shape of waiting time density

With the aim of inquiring about some changes through the years on the properties of the waiting times distribution of the Old Faithful Geyser, we will do KDE of waiting time with the classic data used in this work [4] and two more sets, one from the year 2011 [42] and one current set from March 2019 [42]. We ensured that the three data sets had the same length, $n = 272$ observations.

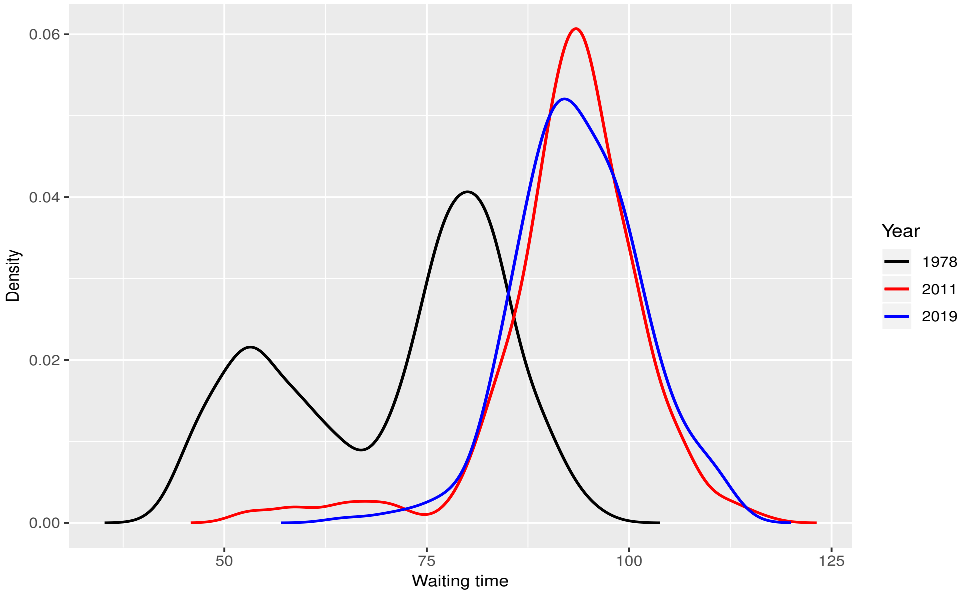

Figure 4–7: *KDE of waiting time sets 1978, 2011 and 2019.*

We can see in Figure 4–7 that the shape of waiting time distribution is not bimodal, as in the past. Now, the short waiting times do not occur. Currently these times are between 65 and 125 minutes. According to the National Park Service (See [43]), these alterations are attributed of changes in circulation that resulted from 1983 Borah Peak earthquake, as well as other local and smaller earthquakes. Then, as consequence, the average interval between eruptions has been lengthening during the last several decades. However, after a local earthquake in 1998, Old Faithfuls eruptions are more often of the long duration, long interval type.

4.6 Nonparametric Regression Methods

Given a set of data points of the form (X_i, Y_i) with $i = 1, ..., n$. The relationship between X_i and Y_i can be modeled by a nonparametric regression function,

$$y_i = m(x_i) + \epsilon_i, \quad i = 1, 2, ..., n \tag{4.18}$$

Essentially, the smoothing of the set (X_i, Y_i) discovers an estimate of m in the the nonparametric regression function showed in equation 4.18. In this equation, we refer to m as the **mean or average response curve** $E[Y|X = x]$ and to ϵ as the random error. We assume that the ϵ is independent and identically distributed with a mean of zero. With this approach, m is smooth, flexible, but unknown. The shape of m and the distribution of the errors are determined using the input data. In this sense, we will use the Old Faithful Geyser data eruptions and the problem to estimate the expectation of the waiting time to illustrate the performance of nonparametric regression, starting by the Kernel smoothing.

4.7 Kernel Smoothing

4.7.1 Nadaraya-Watson Estimator

Nadaraya [30] and Watson [31] proposed simultaneously, in 1964, an estimator of the function m as a locally weighted average, using a kernel as a weighting function. We assume that the observations (X_i, Y_i) are independent and identically distributed. To understand the idea of the Nadaraya-Watson estimator, we can check the definition of the conditional expectation estimate,

$$m(x) = E[Y|X = x] = \int y \frac{f(x, y)}{f_X(x)} \mathrm{d}y \tag{4.19}$$

In Section 4.1, we showed how to estimate $f_X(x)$ in the denominator of equation 4.19 using the kernel density estimate (KDE). Using the same idea, we can estimate the joint density $f(x, y)$ by the multiplicative kernel (e.g. see [5]),

$$\hat{f}(x, y; h_1, h_2) = n^{-1} \sum_{i=1}^{n} K_{h1}(x - X_i) * K_{h2}(y - Y_i) \tag{4.20}$$

In kernel smoothing, the kernel acts as a weighting function, which gives large weight to points near where the curve is smooth and a low weight to points that are far from it. As we saw in Chapter 4.1, there are many kernel functions that can be considered, but in this section we will only use the Gaussian kernel.

According to Nadaraya [30] and Watson [31] we can factorize the numerator in expression 4.19 as follows,

$$\int y\hat{f}(x, y; h_1, h_2)\mathrm{d}x = n^{-1} \sum_{i=1}^{n} K_h(x - X_i)Y_i$$

Note that $\hat{f}$ initially depends on two parameters in h (h_1 and h_2), but finally only a single h is needed for x and y. Now, if we use the result above and combine it with the expression 4.19, we get the Nadaraya and Watson estimator $\hat{m}(x)$, which is defined as follows,

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K_h(x - X_i)Y_i}{\sum_{j=1}^{n} K_h(x - X_j)} = \sum_{i=1}^{n} W_{hi}(x)Y_i, \tag{4.21}$$

where W is the weighting function and can be broken down as shown below,

$$W_{hi}(x) = \frac{K_h(x - X_i)}{\sum_{j=1}^{n} K_h(x - X_j)} = \frac{h^{-1}K(\frac{x - X_i}{h})}{\hat{f}(x; h)}. \tag{4.22}$$

About the results in 4.21 and 4.22, according to [5], the weights $W_{hi}(X)$ depend only on the whole sample $X_1, ..., X_n$ through the kernel density estimate $\hat{f}(x; h)$. Also, the observations Y_i obtain more weight in those areas where the corresponding X_i are sparse. Given the case that the denominator equals zero, then the numerator is also equal to zero, so we set the estimate to 0.

Now, to illustrate the kernel-based smoothing Nadaraya-Watson estimator with Gaussian kernel, we will attempt to estimate the waiting time until the next eruption in the Old Faithful Geyser. In this case, the function `ksmooth` from `R` base package `stats` solves this type of kernel regression. Figure 4.7.1 shows two fits of the Old Faithful Geyser data set with kernel-based smoothing Nadaraya-Watson with the Gaussian kernel.

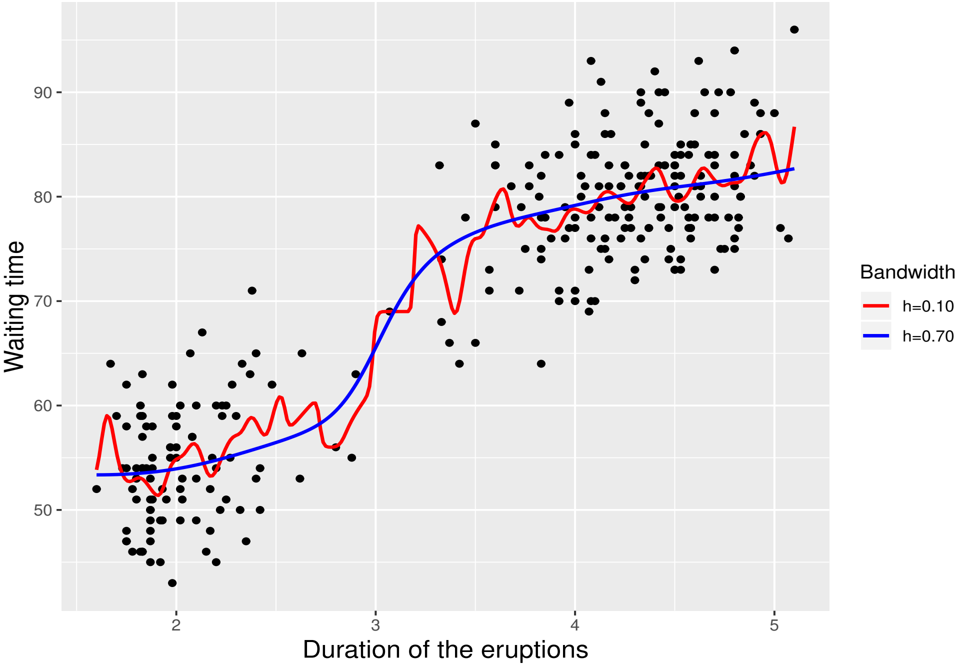

Figure 4–8: *Nadaraya-Watson estimate of the expectation of the waiting time using the bandwidths;* $h = 0.10, 0.70$.

We can visualize two estimates of the expectation of the waiting time using two different values for the bandwidth parameter h (Smoothing parameter). In the first estimation (red line and $h = 0.10$), we can observe that the estimate is not completely smooth, since there are some spiky estimates of $m(x)$. This differs from the other estimate (blue line and $h = 0.70$), which is a smooth curve.

In short, the parameter bandwidth h works as a regulator for the amount of smoothing, just like the on the kernel density estimator (KDE). Therefore, the kernel regression also suffers from the **bias-variance trade-off**. That to say, when h is small, the variability is large but the bias is small; when h is large, the variability is small but the bias is large. In nonparametric regression, the best smoothing bandwidth parameter should equilibrium the bias and variability. Let us take a look at some estimates of the expected waiting time for each of the h values.

Table 4–1: Nadaraya-Watson fitted values of the expectation of waiting time for bandwidths $h = 0.10, 0.70$.

	Bandwidth	
Duration	$h = 0.10$	$h = 0.70$
1.8	52.95	53.48
3.2	76.36	72.30
4.5	79.58	80.91

In Table 4–1, for instance, given a duration of eruption equal to 3.2 minutes, the kernel smoothing prediction would be $\hat{y} = 76.36$ minutes when $h = 0.10$, but for the same value of the duration and when $h = 0.70$, the fitted value is $\hat{y} = 72.30$ minutes. We cannot certainly determine which is the best estimate, but we can select h with some criterion that makes it optimal. Next, we present a tasteful approach called cross-validation that allows us to select the smoothing bandwidth subject to certain optimality criteria.

4.7.2 Optimum Bandwidth: Cross-Validation

There are several methods to find the optimal bandwidth for a smoother parameter, unfortunately, none of them is fully satisfactory. In order to find an optimal solution, like on the KDE, we start by considering the Mean Integrated Square Error (MISE),

$$\text{MISE} = E\Bigg(\int \Big(\hat{m}(x;h) - m(x)\Big)^2 \mathrm{d}x\Bigg) \tag{4.23}$$

As usual, since the MISE depends on the unknown function m, it is necessary to use the data for estimating m. The procedure used for this purpose is cross-validation, specifically, the leave-one-out cross-validation (LOOCV). This method can be computationally expensive if it is employed with large data sets. Even though its difficulties, this method allows to choose the smoothing bandwidth subject to certain optimality criterion. This method has the advantage that the generality of its definition allows it to be applied to quite a wide variety of settings. In the present case, the idea is to choose h to minimize,

$$CV = \sum_{i=1}^{n} \left\{ y_i - \hat{m}_{(-i)}(x) \right\}^2 \tag{4.24}$$

In Equation 4.24, the curve fitted at x_i is constructed from the remainder of the data, excluding x_i. The aim then is to evaluate the amount of smoothness through the extent to which each observation is predicted from the smooth curve produced by the rest of the data. It is expected that the value of h which minimizes Equation 4.24 should provide an adequate level of smoothing. Figure 4–9 shows the plot of the leave-one-out cross-validation (LOOCV) for the Nadaraya-Watson estimator.

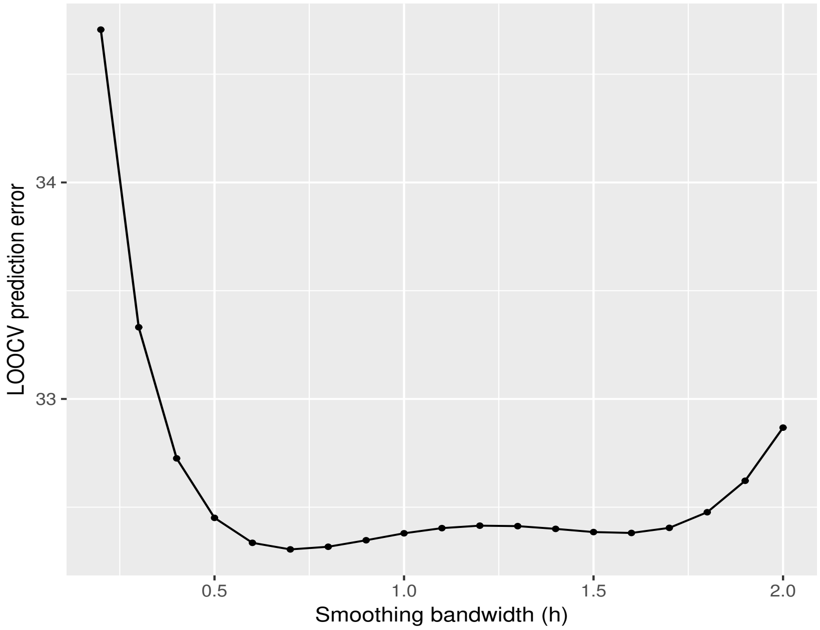

Figure 4–9: *Plot of the leave-one-out cross-validation (LOOCV) for the Nadaraya-Watson estimator in Kernel smoothing.*

The curve in Figure 4–9 represents how the smoothing bandwidth affects the quality of prediction. In this case, the smoothing bandwidth parameter h that minimizes this LOOCV error is 0.70. The Nadaraya-Watson fit using this value of h is represented by the blue curve in Figure 4.7.1.

Also, the bandwidth parameter can be optimized by generalized cross-validation (GCV). This is an approximation of Equation 4.24, and according to [44] it is more computationally efficient. Other variant of the cross-validation is the k-fold cross-validation (k-CV) (e.g. see [45]).

4.8 Locally weighted scatterplot smoothing (LOWESS)

Locally weighted regression scatterplot smoothing (LOWESS), originally proposed by Cleveland in 1979 [32] and further developed by Cleveland and Devlin in 1988 [46], is a nonparametric method used to model a relationship between X_i and Y_i, when the linear relationship is inconvenient. This type of fitting procedure y_0 is estimated using only x values close to x_0, or at least the contribution of x values close to x_0 is weighted more heavily than others. LOWESS regression fits a line to a bivariate scatter of points in a series of iterations using the following steps (see [32] and [47]),

- Step 1: Find the k nearest neighbors of x_0, which constitute a neighborhood $N(x_0)$. The number of neighbors k is specified as a percentage of the total number of points in the data set. This percentage is called the span and is a tuning parameter of the method.
- Step 2: Calculate the largest distance $D(x_0)$ between x_0 and another point in the neighborhood.
- Step 3: Assign weights to each point in $N(x_0)$ using the weight function:

$$W\left(\frac{x_0 - x_1}{\Delta(x_0)}\right) \tag{4.25}$$

The expression 4.25 is called tricube weight function [32], where,

$$W(u) = \begin{cases} (1-u^3)^3 \quad \text{for} & 0 \leq u \leq 1 \\ 0 & \text{Otherwise} \end{cases}$$

- Step 4: Calculate the weighted least squares fit of x_0 on the neighborhood $N(x_0)$.

As in the Kernel smoothing method, for the case of LOWESS regression, there is also a smoothing parameter called span. The span parameter represents the proportion of the total number of points that contribute to each local fitted value. To carry out LOWESS regression in `R` we can use the function `loess()` from `R` base package `stats`. This function uses `span=0.75` by default.

Next, a LOWESS regression is performed for the Old Faithful Geyser with span values of `0.25` and `0.65`. Figure 4–10 shows the LOWESS fits for the Old Faithful Geyser data using span values of 0.25 and 0.65.

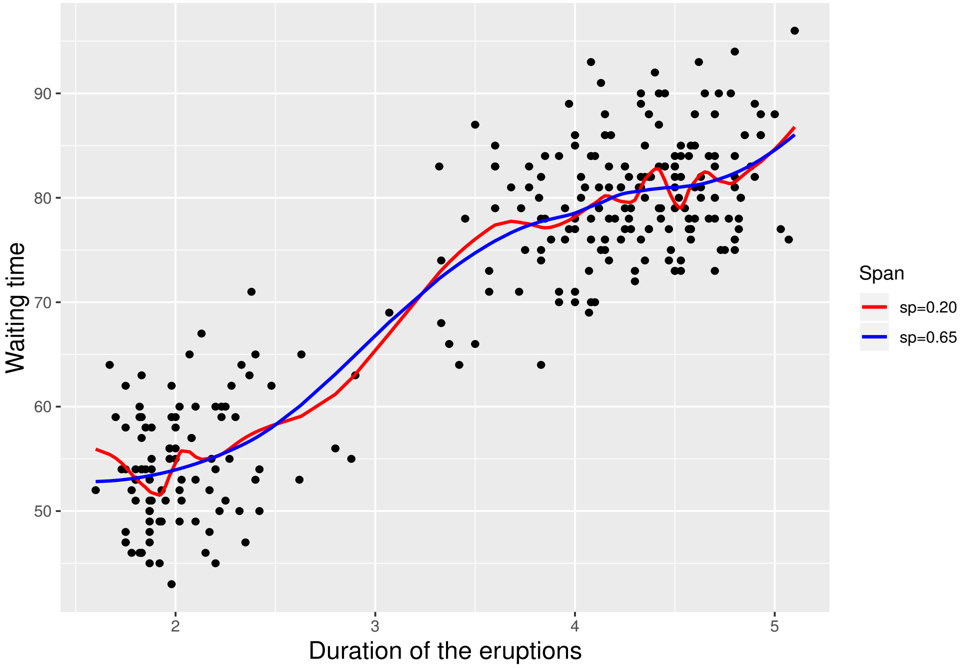

Figure 4–10: *LOWESS regression for the expectation of the waiting time using the span=0.25, 0.65.*

The LOWESS regression using span values of `0.20` and `0.65`, equivalent to the red and blue curves, respectively. This means that each neighborhood consists of 20% and 65% of the observations for each LOWESS fit. Notice that for the small value of span (`0.20`) we get a very rugged curve that follows the observations closely, so we have a curve with a high fit but also high variance. In this case, the optimal value for the span was chosen with the cross-validation method discussed above.

We found that the optimal span value for LOWESS regression with the Old Faithful Geyser data is `0.65`, represented by the blue curve in Figure 4–10.

4.8.1 Interval estimation in LOWESS

It is possible to do statistical inference on the fitted values in LOWESS regression. This allows for computing prediction intervals. To get an estimate of the standard error by including `se=TRUE` argument within the `predict()` function. The command `predict()` does not have an option for interval estimation but we can manually calculate confidence or prediction intervals using output information of the fit.

As we know, for the linear regression, the standard errors for confidence and predict intervals are the following,

$$Se.fit = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}} \tag{4.26}$$

$$Se.pre = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}} \tag{4.27}$$

We can manipulate Equations 4.26 and 4.27, in such a way that we can find useful expressions to build prediction intervals with the information given by the `loess()` function. We can start by Equation 4.26,

$$Se.fit^2 = \hat{\sigma}^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right).$$

$$\frac{Se.fit^2}{\hat{\sigma}^2} = \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}.$$

Now, using Equation 4.27,

$$Se.pre = \hat{\sigma}\sqrt{1 + \frac{Se.fit^2}{\hat{\sigma}^2}} = \sqrt{\hat{\sigma}^2 + Se.fit^2}. \tag{4.28}$$

In Equation 4.28, we have the standard error for prediction intervals in LOWESS regression, where $\hat{\sigma}^2$ is the variance of errors. For instance, we can compute prediction intervals for a set of values of the duration of eruptions.

Table 4–2: LOWESS 95% prediction intervals for average waiting.

Duration	Point estimation	Lower	Upper
1.8	53.00	43.64	62.36
3.2	70.45	61.01	79.89
4.5	81.11	71.81	90.42

A 95% prediction interval of the average waiting time for the 1.8 minute eruption duration falls within $(43.64, 62.36)$ minutes. For the 3.2 and 4.5 minute values, it falls within $(61.01, 79.89)$ and $(71.81, 90.42)$ minutes, respectively.

According to [32] the residual plots can provide an analyst with useful guidance for controlling the LOWESS fitting process. With this graph, is possible to detect some dependence of the scale of the errors in the fitted values range. Figure 4–11 shows the residual plot from the original LOWESS curve that was fitted to the Old Faithful Geyser data.

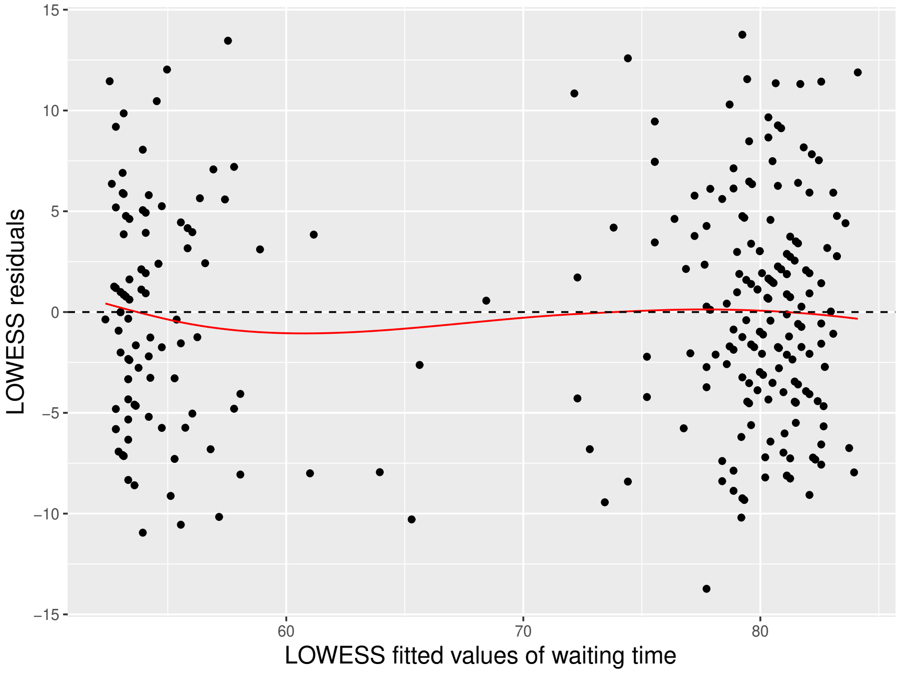

Figure 4–11: *Residual plot from the LOWESS curve fitted to the Old Faithful Geyser data*

We can see a constant scale since there is little change in the smoothed points. The red curve represents the LOWESS fit for the residuals, which is fairly straight. It indicates that there is no strange pattern between the fitted values and the residuals.

This result provides strong evidence that the LOWESS fit with `span=0.65` depicted in Figure 4–10 does provide an adequate representation of the Old Faithful Geyser data.

4.9 Smoothing Splines

Given real numbers $x_1, x_2, ..., x_n$ on some interval $[a, b]$, a spline (Schoenberg, 1964) of order p with knots $t_1, ...t_k$ on the interval $[a, b]$ is a function S that can be obtained by splitting the interval into k subintervals of the form $(a, x_1), (x_1, x_2), ..., (x_k, b)$. Then we can use a polynomial of a degree less than or equal to p in each of the subintervals to smoothly join each of the knots. The spline $S(x)$ is defined by,

$$S(x) = \beta_0 + \beta_1 x + ... + \beta_p x^p + \sum_{j=1}^{k} \beta_{p+j}(x - t)_+^p. \tag{4.29}$$

Where $\beta_0 + \beta_1 + ... + \beta_{p+1} + ... + \beta_k$ are unknown constants. The case when $p = 3$, $S(x)$ is called **cubic spline**.

Now, in the smoothing spline approach the idea is to find a function $m(x)$ in the nonparametric regression function 4.18 that fits the data well but will also be smooth. In that case, we can attempt to reconstitute the function $m(x)$ by constructing a spline function $S(x)$ which minimizes the value of the penalized least squares criterion (e.g. See [48]),

$$\sum_{i}^{n}(y_i - m(x_i))^2 + \lambda \int [m''(x)]^2 \mathrm{d}x, \tag{4.30}$$

where $\lambda > 0$ is the smoothing parameter and $\int [m''(x)]^2 \mathrm{d}x$ is a roughness penalty. Also, when m is rough, the penalty is large, but when m is smooth, the penalty is small. Thus the two parts of the criterion balance fit against smoothness. There is a single minimum of 4.30, which is a natural cubic spline with knots at the data points [49]. The `smooth.spline(.)` function from `R` package base `stats` performs cubic smoothing splines. In this routine, λ is a function that depends on `spar`; the amount of smoothness is regulated by this argument. Figure 4–12 shows the cubic smoothing spline fit of the expectation of the waiting time using two `spar` values.

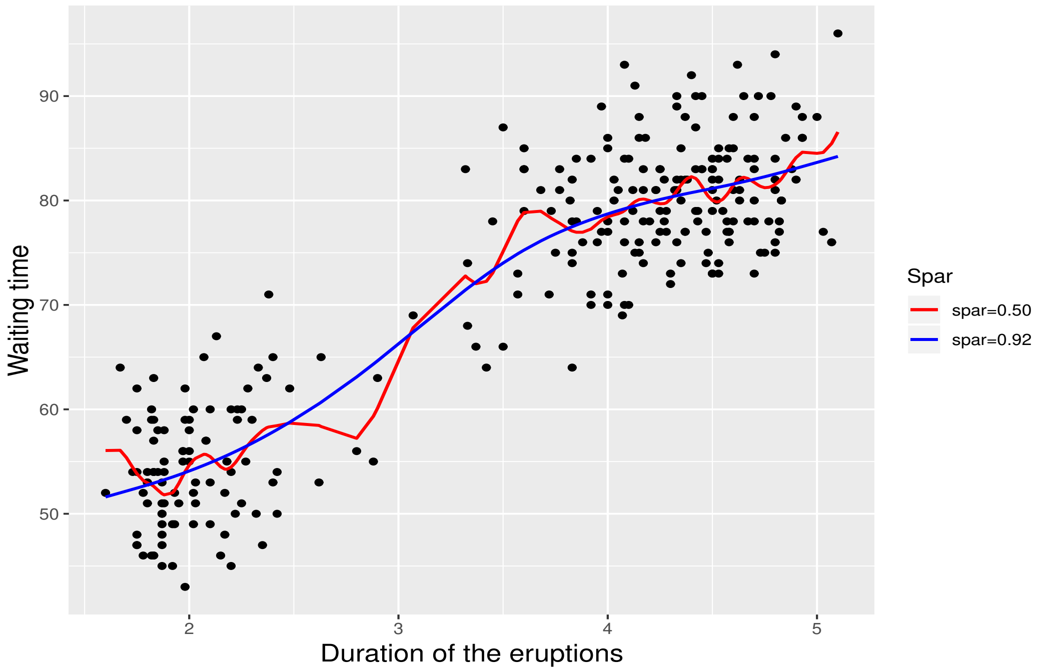

Figure 4–12: *Smoothing spline fits for the expectation of the waiting time under the values of spar=0.50, 0.92.*

In Figure 4–12 we can observe the smoothing splines fits for the expectation of the waiting time with `spar=0.50` (red curve) and `spar=0.92` (blue curve). The value of 0.92 was selected using cross-validation. It can be observed that this model is smooth and fits the data well.

To find the optimal parameter for `spar`,use the cross validation method, which is already included in the function `smooth.spline()`. If we do not specify a bandwidth parameter (`spar` or `df`) and the argument `cv=FALSE` then, the function chooses the bandwidth by default using generalized cross-validation (GCV). In contrast, if `cv=TRUE`, then `spar` is chosen using classical cross-validation (LOOCV).

Finally, Figure 4–13 shows fits for the expected waiting times of the Old Faithful Geyser using Kernel smoothing, LOWESS and smoothing spline.

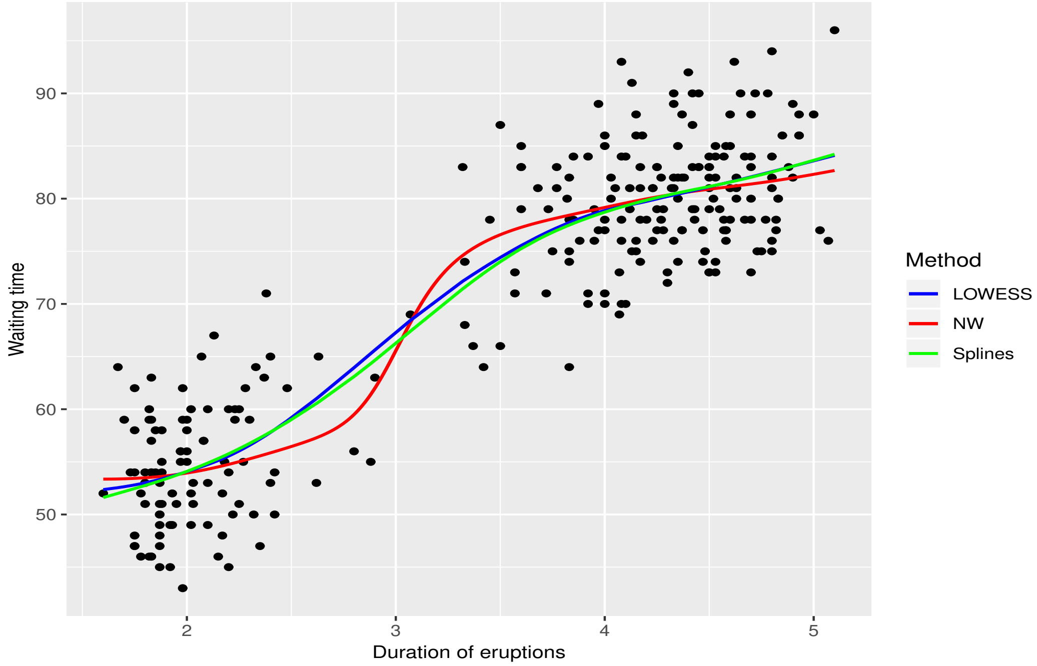

Figure 4–13: *Nadaraya-Watson, LOWESS, and smoothing spline fits of the expectation of the waiting time for the next eruption based on the duration of the eruptions.*

We can see in Figure 4–13, that the fits of the expected waiting time by LOWESS and spline smoothing are very similar. Nadaraya-Watson yields a similar yet different fit. In each case, the smooth parameter was chosen using cross-validation.

CHAPTER 5
DISCUSSION

This work presented some uses of the Old Faithful Geyser data to teach statistics on different levels of complexity. The basic approach uses the Old Faithful Geyser eruptions as a real-life data example to teach statistics in an introductory university course. The more advanced approach employs the same data for teaching statistics in graduate courses. Figure 5–1 depicts the relevant topics for teaching statistics in an introductory statistics course.

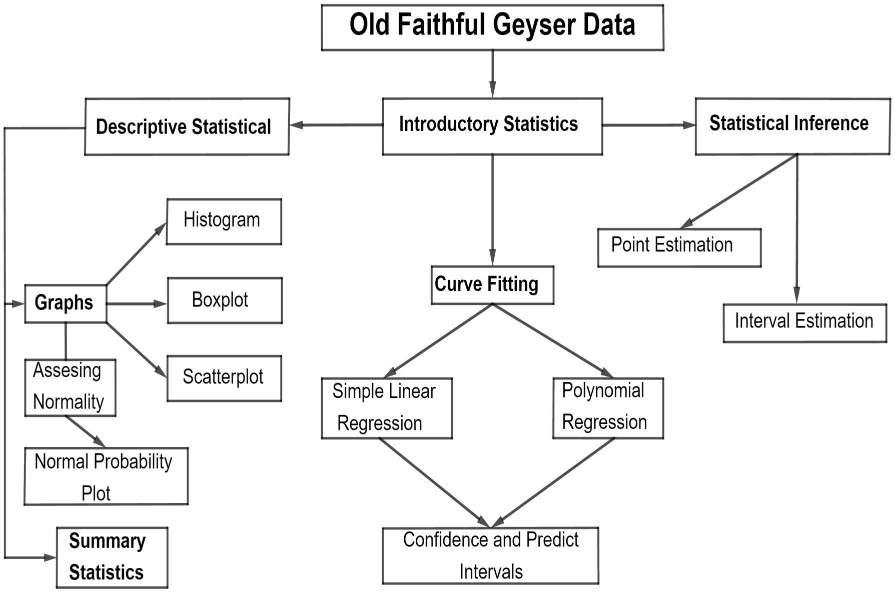

Figure 5–1: *Topics that can be explained in an introductory statistics course using the Old Faithful Geyser data.*

The main topics discussed are Descriptive Statistics, Statistical Inferences, and Curve Fitting; they can be seen in Figure 5–1. Descriptive statistics can be discussed by using the data to create graphs (e.g. histogram, boxplot and scatterplot) and to discuss their meaning. The data can also be used to find summary statistics, including the mean, median, variance, and standard deviation. Additionally, the data can be used to illustrate the normal probability plot as a simple visual test to assess normality and demonstrate that the Old Faithful Geyser data does not follow a normal distribution.

Teaching point and interval estimation with the complete data set will yield misleading results. Instead, we can consider that a large eruption is followed by a long waiting time, and similarly, a shorter eruption is followed by a short waiting time. The data can be split into two parts and statistical inferences can be realized for each of these groups. This strategy guarantees better estimations of waiting time. Moreover, as in the estimation by the Student-t interval, it is necessary to consider the assumption that the population is normally distributed. Again, we can use the data to show the performance of a graphical test like the normal probability plot for checking normality in each group. This exercise will help students learn point and interval estimation and develop their statistical thinking and reasoning skills.

A more advanced method for estimating and predicting the waiting time is the simple linear regression method. Here, the duration of eruptions is used to model the average waiting time by a straight line. The data illustrates concepts such as the relationship between two quantitative variables, as well as variability and uncertainty.

Finally, use of the quadratic regression model is discussed using the data. In a regression analysis, the data can be modeled by a straight or a curvilinear. For our data, the quadratic (i.e curvilinear) model yields a better fit. Also, since the linear and quadratic model are nested, the F-test can be introduced to compare them and draw conclusions of which is better.

Next we can use the Old Faithful Geyser data in more advanced statistics courses. Figure 5–2 depicts the relevant topics for teaching statistics in graduate level courses.

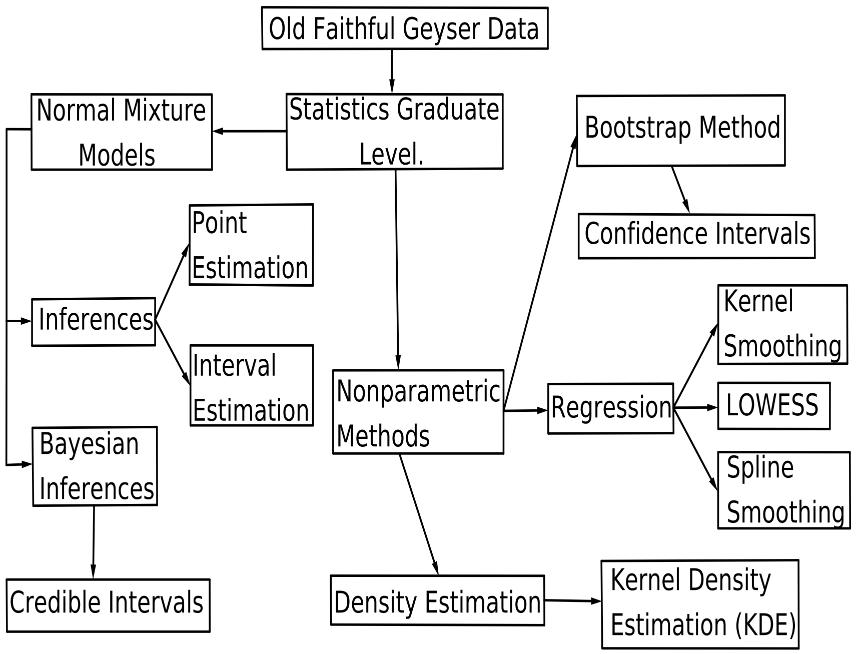

Figure 5–2: *Topics that can be explained in graduate level course using the Old Faithful Geyser data.*

According to Figure 5–2 we can discuss the Old Faithful Geyser data with advanced statistics students as an application for two-component normal mixture models. By using this approach we can make inferences, through either point or interval estimation, on the parameters of this model. In this stage, the Newton-Raphson and the Expectation Minimization Algorithms play an important role in parameter estimation .

Bayesian analysis is another way to make statistical inference in normal mixture models. In this way, the five parameters are considered as random variables and their posterior distributions are approximated by the Metropolis-Hastings algorithm. Finally, the point estimates are obtained using the mean of the posterior distributions and the interval estimations using the quantiles.

A part of this thesis uses of the Old Faithful Geyser data to illustrate how some nonparametric methods work. Starting with the famous bootstrap method; again we can discern that it is more effective to split the data into two parts for point and interval estimation using the bootstrap method. A useful nonparametric method for estimating densities is the kernel density estimation (KDE). Here the bimodal distribution of the waiting time is a good choice to illustrate the method, as well as the most common methods of bandwidth choice. The estimation of the waiting time offers other important applications for data, for instance, the nonparametric curve fitting method such as kernel smoothing, LOWESS, and smoothing splines. These methods can be discussed with advanced students with the objective to estimate the waiting time in the Old Faithful Geyser data.

CHAPTER 6
CONCLUSIONS AND FUTURE WORKS

6.1 Conclusions

Using real-life data during class lessons increases the interest, motivation, and engagement in students learning experiences (e.g. see [11], [12] and [13]]). This work uses Old Faithful Geyser eruptions data to use these advantages while teaching in university-level statistics courses. Throughout the chapters, we illustrated how to discuss concepts and methods related to descriptive statistics, inferences, and curve fitting for students enrolled in introductory statistics courses by applying real-life data. Moreover, we found that these data are also a good example to teach the bootstrap method, inferences of the normal mixture models, kernel density estimations, and some nonparametric methods for curve fitting to advanced statistics students.

After having performed this work, we can conclude that the Old Faithful Geyser real-life data example is an ideal case study for teachers to illustrate statistical concepts in both a theoretical framework and in an applied context. The students, in turn, master basic statistic concepts and principles, practice calculations, and learn to use statistical software. In short, the Old Faithful Geyser data is an excellent real-life example that provides both beginner and advanced students context for problems in statistics. This way, we can demonstrate that statistics can be both an attractive science and crucial for solving practical problems.

Basic prediction of the waiting time of Old Faithful depends upon the duration of the previous eruption. These eruptions can be short or large. Since there are physical causes, the underlying geological that make the variation in the Old Faithful data is not totally random. So we can introduce an analytical approach of statistics results for the attempt to discover explanations, investigate causes, make predictions and look inside the data, addressing the variability in the relationship eruption and waiting time. Also, taking the pattern of the eruptions as part of the variability.

6.2 Future work

Lastly, the Old Faithful data can be used to further develop this work. Including addition of topics such as Bayesian analysis for regression, classification, and cluster regression. Also, an application can be designed with shiny-RStudio to facilitate the teaching process of these subjects and the ones discussed in this thesis.

APPENDICES

APPENDIX A
USEFUL R COMMANDS

Table A–1: Useful R commands.

Package	Function	Use
stats (base)	hist()	Computes a histogram.
stats (base)	boxplot()	Computes a boxplot
stats (base)	qqnorm();qqline()	Computes a Normal probability plot
stats (base)	t.test()	Find Student t-interval
stats (base)	lm()	Computes linear and polynomial regression
stats (base)	anova()	Performs a F-test for nested models.
mixtools	normalmixEM()	EM algorithm for finite mixture models.
stats (base)	ks.test()	Kolmogorov-Smirnov test
stats (base)	nlm()	Minimization of the function f.
stats (base)	density()	Kernel Density Estimation
stats (base)	bw.nrd();bw.nrd0()	Bandwidth based on rule-of-thumb.
stats (base)	bw.ucv()	Bandwidth based unbiased cross-validation.
stats (base)	bw.bcv()	Bandwidth based biased cross-validation.
stats (base)	ksmooth()	Fit a Nadaraya-Watson kernel regression.
stats (base)	loess()	Fit a LOWESS regression.
stats (base)	smooth.spline()	Fit a Smoothing Spline.

REFERENCE LIST

[1] J.S. Rinehart. Earth tremors generated by old faithful. *Science*, 150:494496, 1965.

[2] J.S. Rinehart. Thermal and seismic indications of old faithful geysers inner working. *Journal of Geophysical Research*, 74:566573, 1969.

[3] Lorraine Denby and Daryl Pregibon. An example of the use of graphics in regression. *The American Statistician*, 41(1):33–38, 1987.

[4] A. Azzalini and A. W. Bowman. A look at some data on the old faithful geyser. *Applied Statistics*, 39(3):357365, 1990.

[5] W. Hardle. *Smoothing Techniques with Implementation in S.* Springer, first edition, 1990.

[6] J. Michael Shaughnessy and Maxine Pfannkuch. Statistical thinking: A story of variation and prediction. *The National Council of Teachers of Mathematics*, 95(4):252–259, 2002.

[7] Kieran D. OHara and E.K. Esawi. Model for the eruption of the old faithful geyser, yellowstone national park. *GSA TODAY*, 23(6):4–9, 2013.

[8] Kieffer. S.W. Seismicity at old faithful geyser: An isolated source of geothermal noise and possible analogue of volcanic seismicity. *Journal of Volcanology and Geothermal Research*, 22:5995, 1984.

[9] Sanford Weisberg. *Applied Linear Regression.* Addison-Wesley Publishing Company, second edition, 1985.

[10] J. D. Singer and J. B Willett. Improving the teaching of applied statistics: Putting the data back into data analysis. *The American Statistician*, 44(3):223230, 1996.

[11] Michelle Hood David L. Neumann and Michelle M. Neumann. Using real-life data when teaching statistics: Student perceptions of this strategy in an introductory statistics course. *Statistics Education Research Journal*, 12(2):59–70, 2013.

[12] T. E. Bradstreet. Teaching introductory statistics courses so that nonstatisticians experience statistical reasoning. *The American Statistician*, 50(1):6978, 1996.

[13] J. Garfield and D. Ben-Zvi. *Developing students statistical reasoning: Connecting research and teaching practice*. Dordrecht, The Netherlands: Springer, first edition, 2008.

[14] Daniel Ria no Rufilanchas. On the origin of karl pearsons term histogram. *Estadística Española*, 59(192):1–7, 2017.

[15] Herbert A. Sturges. The choice of a class interval. *Journal of the American Statistical Association*, 21(153):65–66, 1926.

[16] Allan R Wilks R. A. Becker, J. M. Chambers. *New S Language*. Chapman & Hall/CRC, first edition, 1998.

[17] John M Chambers. *Graphical methods for data analysis*. Belmont, Calif. : Wadsworth International Group ; Boston : Duxbury Press, first edition, 1983.

[18] Maurice H Quenouille. Notes on bias in estimation. *Biometrika*, 43(3-4):353–360, 1979.

[19] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[20] B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. CHAPMAN & HALL CRC, first edition, 1993.

[21] G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, New York, second edition, 2000.

[22] Keith Knight. *Mathematical statistics*. CHAPMAN & HALL CRC, first edition, 1999.

[23] A. P. Dempster; N. M. Laird; D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[24] T. Benaglia and D. Chauveau. An r package for analyzing finite mixture models. *Journal of Statistical Software*, 32:6, 2009.

[25] Frank J. Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

[26] Z. W. Birnbaum. Numerical tabulation of the distribution of kolmogorov's statistic for finite sample size. *Journal of the American Statistical Association*, 47(259):425–441, 1952.

[27] Metropolis et al. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[28] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[29] Derek S. Young. An overview of mixture models. *Statistics Surveys*, 0, 2008.

[30] E. A. Nadaraya. On estimating regression. *Theory of probability and its applications*, 9(1):141–142, 1964.

[31] G. S. Watson. Smooth regression analysis. *The Indian Journal of Statistics (Series A)*, 26(4):359–372, 1964.

[32] W.S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836, 1979.

[33] Grace Wahba. Smoothing noisy data with spline functions. *Springer-Verlag*, 24:383–393, 1975.

[34] P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. CHAPMAN & HALL CRC,

first edition, 1994.

[35] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. CHAPMAN & HALL CRC, 1986.

[36] V. A. Epanechnikov. Nonparametric estimation of a multidimensional probability density. *Theory Probab. Appl.*, 14(14:1):153158, 1969.

[37] M. C. Jones. Simple boundary correction for kernel density estimation. *Statistics and Computing*, 3:135146, 1993.

[38] David W. Scott and George R. Terrell. Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, 82(400):1131–1146, 1987.

[39] David W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley Series in Probability and Statistics, first edition, 1992.

[40] Artur Gramacki. *Nonparametric Kernel Density Estimation and Its Computational Aspects*. Springer, first edition, 2018.

[41] Julian J. Faraway and Myoungshic Jhun. Bootstrap choice of bandwidth for density estimation. *Journal of the American Statistical Association*, (85:412):1119–1122, 1990.

[42] GeyserTimes. Eruptions of Old Faithful Geyser, November 2011-March 2019. https://geysertimes.org, 2011,2019.

[43] National Park Service. Old Faithful Geyser Frequently Asked Questions. https://www.nps.gov/yell/learn/nature/oldfaithfulgeyserfaq.htm, 2019.

[44] Oohn Rice. Bandwidth choice for nonparametric regression. *The analysis of statistics*, 32(4):1215–1230, 1984.

[45] Stone M. Cross-validatory choice and assessment of statistical predictions. *J. Royal Stat. Soc.*, 36(2):111147, 1974.

[46] William S. Cleveland and Susan J. Devlin. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 83(403):596–610, 1988.

[47] Wolfgang A. Rolke. Local Regression Models - Smoothing. http://academic.uprm.edu/wrolke/esma6665/locreg.htm, 2019.

[48] Julian J. Faraway. *Extending the Linear Model with R.* Addison-Wesley Publishing Company (Texts in Statistical Science), first edition, 2005.

[49] CHRISTIAN H. REINSCH. Smoothing by spline functionse. *Numerische Mathematik*, 10:177–183, 1967.

THE USE OF THE OLD FAITHFUL GEYSER DATA IN BOTH UNDERGRADUATE AND GRADUATE STATISTICS COURSES

DIDIER A. MURILLO FLOREZ

Department of Mathematics

Chair: Wolfgang A. Rolke. PhD

Degree: Master of Science

Graduation Date: May, 2019

Many statistics teachers in line with the recommendations of several studies suggest using real-life context data during class lessons. The present work shows the use of the Old Faithful Geyser eruptions as a real-life data example for teaching statistics at a university. It can be adapted for basic courses to more advanced ones. During this study, we illustrated how to use the Old Faithful Geyser data to discuss concepts and methods related to descriptive statistics, inferences and curve fitting for an introductory statistics course. In the same way, we use the data to examine some uses in advanced class lessons that involve topics such as bootstrap method, inferences for the normal mixture models, goodness of fit tests, kernel density estimation (KDE), as well as some nonparametric regression methods, namely, kernel smoothing, LOWESS and smoothing splines.