WEEK 3

Exploratory Data Analysis for Univariate Data: Numerical Summaries

Usually, one uses the word *Statistics* with three (maybe more?) meanings:

- Statistics is a Course, Scientific Subject name, topic we are studying at Universities
- Statistics is a numerical characteristic for a numerical dataset
- Statistics is a r.v. obtained from a random sample (will be discussed later)

Here we want to give some numerical summaries, characterizations for our data. The previous part of our course was devoted to graphical representation of the data, and here we want to talk about some numerical representation for our data.

Definition 3.1. *Given a numerical dataset $x_1, ..., x_n$, we will call a* ***Statistics*** *for that dataset any function of $x_1, ..., x_n$.*

Well, of course, this is a formal definition, and in the reality we will choose some meaningful and descriptive Statistics for our dataset, to get some useful information about our data.

Concerning the notations, I will use the following convention: in our Probability course, I was denoting by capital letters r.v.'s, say X was a r.v.. Here, in this section, by x I will denote a dataset - a collection of real numbers, values of some variable. Say, if I have a dataset $x_1, x_2, ..., x_n$, I will say that we have a dataset x. If y is another dataset, then y consists of some numbers (or 2D points) $y_1, y_2, ..., y_m$ etc.

In our Probability course, for a r.v. X we have defined some numerical characteristics: e.g.,

$\mathbb{E}(X)$ - the Expected Value or the Expectation of X;

$Var(X)$ - the Variance of X;

$SD(X)$ - the Standard Deviation,

and for a pair of r.v. we have defined

$Cov(X, Y)$ - the covariance between X and Y;

$Cor(X, Y)$ - the correlation coefficient between X and Y etc.

For numerical datasets we will define the analogous characteristics, and we will use small letters, and denote the corresponding quantities as $var(x)$, $sd(x)$ or $cov(x, y)$, say.

In this part we will talk only about univariate numerical data, about observations concerning one feature. In the next part, we will consider some numerical characteristics for the relationship of two numerical datasets.

3.1 Order Statistics (Ranks)

First, we want to introduce a useful notion and notation:

Definition 3.2. *For a dataset $x_1, x_2, ..., x_n$, let $x_{(j)}$ be the jth sample value, when our data is ordered in the increasing order. Then $x_{(j)}$ is called the* **j-th order statistics** *of our dataset.*

In other words, to obtain the j-th order statistics, one needs to arrange our data into the increasing order, and then calculate the j-th element in this arranged set[1].

So, by the definition, the dataset $x_{(1)}, x_{(2)}, ..., x_{(n)}$ coincides with $x_1, ..., x_n$, and

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

In particular,

$$x_{(1)} = \min\{x_1, x_2, ..., x_n\} \qquad \text{and} \qquad x_{(n)} = \max\{x_1, x_2, ..., x_n\}.$$

Example, Order Statistics: Here is a sample showing dayly number of informative emails (that I am reading and replying to) in my AUA acoount:

$$10, 10, 2, 5, 7, 5, 1, 2, 7, 2, 6, 8, 2, 4, 8, 6.$$

Now, in order to find, say 7th order statistic, we need to sort our data in the increasing order:

$$1, 2, 2, 2, 2, 4, 5, 5, 6, 6, 7, 7, 8, 8, 10, 10.$$

The 7th order statistic is the 7th element in this sorted list from the left: $x_{(7)} = 5$. Similarly, $x_{(1)} = 1$, $x_{(2)} = x_{(3)} = x_{(4)} = x_{(5)} = 2$, $x_{(6)} = 4$, $x_{(11)} = x_{(12)} = 7$, etc.

Computational Challenge: Given a numerical dataset $x_1, x_2, ..., x_n$,

- Sort the dataset
- for a given j, find the j-th order statistic of that dataset

R code, Order Statistics: In order to find the j-th order statistics in the dataset in **R**, one can use the *sort* command. Say, I want to find the 7th order statistics in the dataset above:

```
#Order Statistics, My Emails Data
my.emails <- c(10, 10, 2, 5, 7, 5, 1, 2, 7, 2, 6, 8, 2, 4, 8, 6)
my.emails.sorted <- sort(my.emails)
my.emails.sorted[7]
```

3.2 Statistics for the central tendency

Here we want to give some measures, estimates of location for the typical observed value, estimates for the location of the center of the data. We will give different statistics for that.

[1] counting from the left to the right $\ddot\smile$

3.2.1 The Sample Mean

Definition 3.3. *For a sample* $x_1, x_2, ..., x_n$, *the sample mean is*

$$\bar{x} = \text{mean}(x) = \frac{x_1 + x_2 + ... + x_n}{n}.$$

Geometric Interpretation: If we will put our data into the real line, and if we will put equal weights at that points, then the sample mean or the average of our set is the equilibrium position of that system, of that real line with the weights at our data points.

Example, Sample Mean: Consider the dataset x:

$$1, 1, 2, 3, -4, 5.$$

The Sample Mean for dataset is

$$\bar{x} = \text{mean}(x) = \frac{1 + 1 + 2 + 3 + (-4) + 5}{6} = \frac{4}{3}.$$

Example, Sample Mean: Looking at my emails dataset above, if I want to give someone the idea about the number of emails I am receiving daily, it will not be a good idea to give the daily email numbers, say, "one day I have received 10 emails, the other day I have received again 10 email, and for the next day the number of emails was 2" etc. Instead, if I will calculate the Sample Mean for my email number dataset:

$$\text{mean}(\text{emails}) = \frac{10 + 10 + 2 + 5 + 7 + 5 + 1 + 2 + 7 + 2 + 6 + 8 + 2 + 4 + 8 + 6}{16} = \frac{85}{16} = 5.3125$$

and now I can describe: the average number of my daily emails is 5.3.

By the way, sometimes, when stating some average numbers, that can be a little bit confusing. Say, one is stating that the average number of children in families in a country is 1.8. This can be confusing at the first sight, since the number of children cannot be non-integer. But you need to take this as the example above: the average number of emails is 5.3, but I will never get 5.3 emails (well, I will, if the internet connection will be interrupted during the 6th email download process $\ddot\smile$).

R code, Sample Mean, Daily Emails Dataset: Here is the **R** code to calculate the Sample Mean for the Daily Emails dataset:

```
#Sample Mean Calculation
x <- c(10, 10, 2, 5, 7, 5, 1, 2, 7, 2, 6, 8, 2, 4, 8, 6)
mean(x)
```

Remark, Idea Behind the Sample Mean: The idea of the Sample Mean is very intuitive, so maybe there is no necessity to give the idea behind that. But let me give an explanation to make a link between the Probability Theory and Sample Statistics. And you will find this type of links also in the rest of this part.

Assume we are given a numerical dataset $x_1, x_2, ..., x_n$. We then make a random variable X out of this dataset, by giving equal probability to each of this data points. We have made the same thing when considering the Empirical CDF. Say, if our dataset is $-1, 1, 1$, then we make a r.v. X with

Values of X	-1	1
$\mathbb{P}(X = x)$	$\frac{1}{3}$	$\frac{2}{3}$

Now, the Sample Mean of the dataset x is just the Expected Value of our r.v., $\mathbb{E}(X)$. Say, for the above example,

$$\mathbb{E}(X) = \frac{1}{3} \cdot (-1) + \frac{2}{3} \cdot 1 = \frac{-1+1+1}{3} = mean(x).$$

Now, if our dataset x is given by the frequency table: the frequency of the value x_k is f_k, for $k = 1, 2, ..., m$, then the Sample Mean of x can be calculated by

$$\bar{x} = mean(x) = \frac{f_1 \cdot x_1 + f_2 \cdot x_2 + ... + f_m \cdot x_m}{f_1 + f_2 + ... + f_m}.$$

To see why this formula is true, we just need to "recover" the whole dataset using the frequency table: our dataset will be

$$\underbrace{x_1, x_1, ..., x_1}_{f_1 \text{ times}}, \underbrace{x_2, x_2, ..., x_2}_{f_2 \text{ times}}, ..., \underbrace{x_m, x_m, ..., x_m}_{f_m \text{ times}}.$$

Example, Sample Mean by Frequencies: Considering again my emails dataset (in the sorted form),

$$1, 2, 2, 2, 2, 4, 5, 5, 6, 6, 7, 7, 8, 8, 10, 10$$

we can write it in the frequency table form:

No. of daily emails	1	2	4	5	6	7	8	10
Frequency	1	4	1	2	2	2	2	2

So the mean number of my daily emails can be calculated by the formula above:

$$mean(emails) = \frac{1 \cdot 1 + 4 \cdot 2 + 1 \cdot 4 + 2 \cdot 5 + 2 \cdot 6 + 2 \cdot 7 + 2 \cdot 8 + 2 \cdot 10}{1 + 4 + 1 + 2 + 2 + 2 + 2 + 2} = \frac{85}{16},$$

which gives, of course, the same result as above.

R code, Sample Mean by Frequencies: To calculate the Sample Mean by Frequencies, we can use the following two **R** codes:

```
#Sample Mean Calculation with Frequencies, v. 1
```

```
x.unique <- c(1, 2, 4, 5, 6, 7, 8, 10) #Unique x values
x.freq <- c(1, 4, 1, 2, 2, 2, 2, 2) #The corresponding Frequencies
mean_by_freq <- (x.unique%*%x.freq)/sum(x.freq) #The dot product over the sum of frequencies
mean_by_freq <- as.numeric(mean_by_freq) #Transforming the result to a number
```

Here the command $a\% * \%b$ calculates the dot product of a and b. In our case, we calculate the dot product of two vectors $x.unique$ and $x.freq$, and the result is a 1×1 matrix. Try to run the command $x.unique\% * \%x.freq$. To convert it to a number, we use the $as.numeric$ command.

The second version is more straightforward:

```
#Sample Mean Calculation with Frequencies, v. 2
x.unique <- c(1, 2, 4, 5, 6, 7, 8, 10)
x.freq <- c(1, 4, 1, 2, 2, 2, 2, 2)
mean_by_freq <- sum(x.unique*x.freq)/sum(x.freq)
```

Here the command $x.unique * x.freq$ returns a vector of corresponding elements products, in the above terms, it returns the vector $(f_1 \cdot x_1, f_2 \cdot x_2, ..., f_m \cdot x_m)$. We calculate the sum of the elements of this vector by $sum(x.unique * x.freq)$.

The Sample Mean is very easy to calculate, and, of course, usually serves as a good representation for the typical value, the location for the center for our data. But, unfortunately, it has a drawback, weak side: it is sensitive to extreme (very large or very small) values, to outliers.

Example, Sample Mean Sensitivity to Outliers: We have seen above that I can say that typically, I am replying to 5.3 emails daily. This will give a good sense of how busy am I[2] ☺ Now, assume that the next day of my above observation, I have received 100 emails. Now, if I will use this number too, then the number of daily emails will be

$$10, 10, 2, 5, 7, 5, 1, 2, 7, 2, 6, 8, 2, 4, 8, 6, 100,$$

and the Sample Mean will be

$$mean(\text{daily emails}) = \frac{10+10+2+5+7+5+1+2+7+2+6+8+2+4+8+6+100}{17} \approx 10.88$$

And if I will state that my daily email number is almost 11, that will not give the correct picture. And this is because of just one enormously email-busy day!

Example, Sample Mean Sensitivity to Outliers: Assume you apply to a newly established programming company, and you ask about the mean salary in that company. And you get the response that the mean salary is $300,000$AMD. And you are happily thinking that if you will be their employee, then you will get something around 300,000. Well, plus/minus something, of course, based on the experience. But, let us see that the number can be misleading: assume the company has 21 employees, and 20 of them are receiving 165,000 and one gets 3,000,000. Then the mean will be, using the frequency form calculation,

$$mean(salary) = \frac{20 * 165,000 + 3,000,000}{20+1} = 300,000.$$

But the real picture is that almost all employees, except the boss, get 165,000.

So, when using the Sample Mean, please take an attention to outliers.

In fact, there are some methods trying to solve the problem of outliers sensitivity of the Sample Mean: the following methods are more robust in the sense of sensitiveness to extremes.

Trimmed (Truncated) Sample Mean: First we choose a natural number p, satisfying $2p < n$. Next we sort our data in the increasing order, we drop p lowest and p highest values[3], and then we calculate the sample mean of the rest of the data:

$$\text{trimmed sample mean}(x) = \bar{x}_{\text{trimmed}} = \frac{x_{(p+1)} + x_{(p+2)} + \ldots + x_{(n-p-1)} + x_{(n-p)}}{n-2p} = \frac{\sum\limits_{k=p+1}^{n-p} x_{(k)}}{n-2p}.$$

This trimmed means eliminates the influence of extreme values.

EXAMPLE, TRIMMED SAMPLE MEAN: Say, sometimes, when calculating the score of some international competition, one drops the highest and lowest values of the grades of judges, and then takes the trimmed mean (because one of the judges can give very high grades for participants from its own country or very low grades for "not-so-friendly" country participants).

See, for example, how the Diving Competition is scored at https://en.wikipedia.org/wiki/Diving#Scoring_the_dive.

EXAMPLE, TRIMMED SAMPLE MEAN: When doing grading for quizzes at AUA, we are usually dropping the lowest grade - this is some kind of half-trimming, since we are not dropping the highest grade (fortunately ?) $\ddot{\smile}$.

R CODE, TRIMMED SAMPLE MEAN: Here is the Code to calculate the Trimmed Mean:

```
#Trimmed Mean Calculation
x <- c(10, 10, 2, 5, 7, 5, 1, 2, 7, 2, 6, 8, 2, 4, 8, 6)
mean(x, trim = 0.2)
```

The parameter `trim` can take values from 0 to 0.5 (the default is 0), and shows the fraction of observations to be trimmed from each end of the dataset before the mean is computed.

Winsorized Sample Mean: Another variation is the Winsorized Mean - we sort our data in the increasing order, and, for a fixed number p, we replace the first (i.e., smallest) p numbers by $x_{(p)}$, and the last p numbers (i.e., largest p numbers) by $x_{(n-p)}$. Then we calculate the sample mean of the obtained dataset:

$$\text{winsorized mean}(x) = \frac{x_{(p)} + x_{(p)} + \ldots + x_{(p)} + x_{(p+1)} + x_{(p+2)} + \ldots + x_{(n-p)} + x_{(n-p)} + \ldots + x_{(n-p)}}{n} =$$

[3] Sometimes people take a number p from 1 to 100 and drop the lowest and highest $p\%$ of the data.

$$= \frac{p \cdot x_{(p)} + \sum\limits_{k=p+1}^{n-p-1} x_{(k)} + p \cdot x_{(n-p)}}{n}.$$

Weighted Sample Mean: Assume we want to calculate the mean of the dataset $x_1, x_2, ..., x_n$. We take nonnegative weights w_k's, such that $\sum_{k=1}^{n} w_k \neq 0$, and we calculate

$$\text{weighted sample mean} = \bar{x}_w = \frac{\sum_{k=1}^{n} w_k x_k}{\sum_{k=1}^{n} w_k}.$$

The weight of data x_k is then $\frac{w_k}{\sum_{i=1}^{n} w_i}$. This is to give more weight to some data, and give less weight to some others (if, say, we are unsure in the correctness of some that data). Say, if we are collecting data from different sources, and we trust some of our sources and not too much to others, we can give larger weights to the results from the trusted sources and small weights to the others (if we do not want to completely dismiss the results/observations of the latters). Another situation is, for example, when calculating the mean daily price of some stock, to make predictions about the price in the future, one can calculate the weighted mean of daily prices, giving more weights to recent information, and less weight to the old information. The idea is that we think that recent prices contain more information about the price, useful for prediction, rather than the old prices. But, of course, old prices do contain *some* information, so we do not want to dismiss them completely.

The Weighted Sample Mean is not helping us too much concerning the extremes sensitiveness, but is helping when we want to make a difference between the datapoints.

R code, Weighted Mean: To calculate the Weighted Sample Mean of a dataset x with weights w, one can use the **R**'s function $weighted.mean()$:

```
#Weighted Arithmetic Mean
x <- c(-1, 0, 3, 2, -2, 3, 2, 3, 2, 3)
w <- c( 2, 2, 2, 1, 1,  1, 2, 0, 1, 5)
weighted.mean(x, w)
```

Here **R** will produce 1.647059. To check the result is true, we can try

```
sum(x*w)/sum(w)
```

which will give the same result.

Remark, Theoretical and Sample Means as minimizers: Hope everybody remembers that the Theoretical Mean of a distribution (or a r.v. X with that distribution) is defined as the Expected Value of the r.v. X.

Then one can prove the following assertions:

a. m is the Mean of the r.v. X if and only if

$$m \in \operatorname*{argmin}_{a \in \mathbb{R}} \mathbb{E}\Big((X-a)^2\Big);$$

b. m is the mean of the dataset $x_1, x_2, ..., x_n$ if and only if

$$m \in \operatorname*{argmin}_{a \in \mathbb{R}} \frac{1}{n} \cdot \sum_{k=1}^{n} (x_k - a)^2.$$

Of course, here we assume that the expectation above is defined.

3.2.2 The Sample Median

The other measure of central tendency, "central value" of the dataset, more robust to outliers than the Sample Mean, is the **Median** of a dataset. The Median is a point (not necessarily from our dataset) such that half of the observations (datapoints) are less than or equal to that number, and half of the observations are greater than or equal to that number. In fact, different authors use different definitions, but the idea is the one above. The point is that when we have an odd number of datapoints, then we can find a unique datapoint dividing our dataset into two halves (that datapoint itself is included in both halves). So there is no ambiguity in the definition of the Median in this case.

Example, Median of a Dataset, odd number of elements: Assume our dataset is

$$1, 3, 2, 1, 2, 1, 2, 3, 4, 5, 2$$

First we sort our dataset:

$$1, 1, 1, 2, 2, 2, 2, 3, 3, 4, 5$$

Now, the sixth element in this sorted list (one of the 2's) divides our sorted dataset into two equal parts:

$$1, 1, 1, 2, 2, {\color{red}2}, 2, 3, 3, 4, 5$$

(there are 6 elements less than or equal to the red 2, and 6 elements larger than or equal to the same number). So the median is 2 for this dataset.

But for the case when we have an even number of datapoints, then we can have different approaches.

Example, Median of a Dataset, even number of elements: Assume our dataset is

$$1, 3, 2, 1$$

We want to find a point (number) such that half of the datapoints are less than or equal to that point, and half of the observations are greater than or equal to that point. Again, we sort our dataset:

$$1, 1, 2, 3$$

Now, any point between 1 and 2 will have the above property. Say, half of the datapoints are $\leqslant 1.3$ and half of the datapoints are $\geqslant 1.3$. So we can use 1.3 as a Median for this dataset. Similarly, 1.4 will work as well. Usually, people take the midpoint of these numbers, 1.5.

Note here that we do not require our Median to be a datapoint, to be in our dataset!

Now, one of the widely used definitions of the Median is the following (and we will use this one in the rest of the text):

Definition 3.4. *The Sample Median* $\mathrm{median}(x)$ *of the dataset* $x_1, x_2, \ldots, x_n$ *is the number dividing the sorted dataset*

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$$

to two equal parts, more precisely:

$$\textit{If } n \textit{ is odd,} \quad \mathrm{median}(x) = x_{\left(\frac{n+1}{2}\right)};$$

$$\textit{If } n \textit{ is even,} \quad \mathrm{median}(x) = \frac{1}{2} \cdot \left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)} \right).$$

Example, Sample Median: Continuing to consider the number of daily emails dataset,

$$10, 10, 2, 5, 7, 5, 1, 2, 7, 2, 6, 8, 2, 4, 8, 6.$$

let us calculate the median of it. To that end, we first sort our data in the increasing order:

$$1, 2, 2, 2, 2, 4, 5, 5, 6, 6, 7, 7, 8, 8, 10, 10.$$

Here we have $n = 16$ observations, an even number of observations, so, by our definition above,

$$\mathrm{median}(\text{daily emails}) = \frac{1}{2} \cdot \left(x_{(8)} + x_{(9)} \right) = \frac{1}{2} \cdot (5 + 6) = 5.5$$

So I can state that the "average" number of emails I am receiving daily is 5.5.

R code, Sample Median, Emails Dataset: Here is the code to calculate the Sample Median for the daily emails dataset:

```
#Sample Median Calculation
x <- c(10, 10, 2, 5, 7, 5, 1, 2, 7, 2, 6, 8, 2, 4, 8, 6)
median(x)
```

Once again, the Sample Median is a number dividing our set of observations into two equal-length parts. On the Fig. 3.1 you can see on red the Sample Median - we have 50 data points here (in Black), and 25 of them are to the left of the Median and 25 of them are to the right.

R code, Sample Median: Fig. 3.1 is obtained by running the code:

```
#Sample Median calculation
x <- rnorm(50, mean = 2, sd = 10)
m <- median(x)
y <- rep(0,50)
plot(x,y, pch = 20, cex = 1.1)
points(m,0, pch = 16, col = "red", cex = 1.1)
```

Here the command $\mathrm{rnorm}(n, \mathrm{mean} = a, \mathrm{sd} = b)$ is generating a sample of size n from the Normal Distribution with a mean a and Standard Deviation b, i.e., we are getting 50 possible

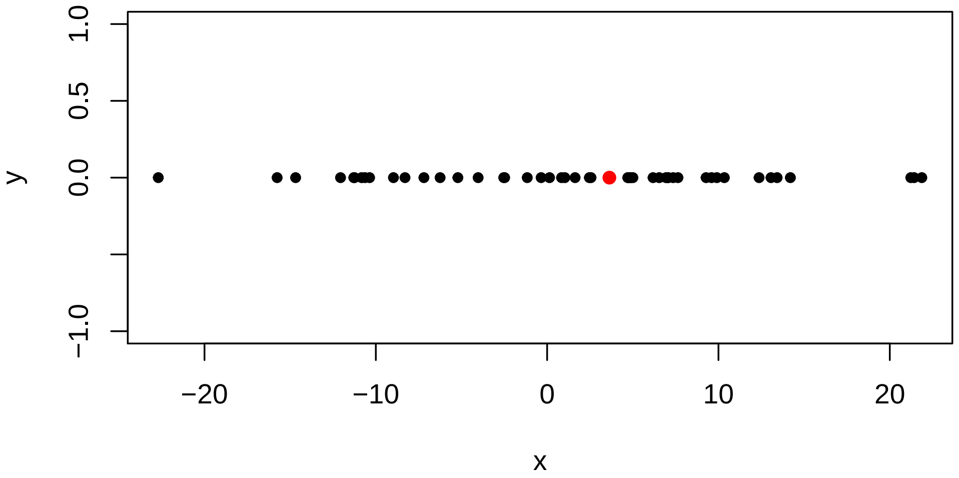

Fig. 3.1: Sample Median Example: Here we have a dataset of 50 points (in Black). The red point is the Median.

realizations of $X \sim \mathcal{N}(a, b^2)$. The command $\mathrm{plot}(x,y)$ runs as follows: if we have 2 vectors of the same size $x = (x_1, x_2, ...)$ and $y = (y_1, y_2, ...)$, it draws the corresponding points (x_i, y_i). Here we use $y < -\mathrm{rep}(0, 50)$: the command rep stands for "replicate", it copies 0 fifty times. So in the result, y will be the zero vector of size 50. Try to run, say, $\mathrm{rep}(c(1,2), 10)$ - this will copy 1,2 ten times. Going back to our goats, $\mathrm{plot}(x,y)$ will draw the points $(x_i, 0)$ for all i. pch is for point character - try changing the value to see the effect, and cex is for the point size - again play with the values to see the effect.

It can be seen, that Sample Median is not affected by outliers, by extreme observations.

Example, Sample Median: If we will consider the above example with fake email numbers dataset,

$$10, 10, 2, 5, 7, 5, 1, 2, 7, 2, 6, 8, 2, 4, 8, 6, 100,$$

or, in the sorted form,

$$1, 2, 2, 2, 2, 4, 5, 5, 6, 6, 7, 7, 8, 8, 10, 10, 100,$$

then the Sample Median will be (we have $n = 17$ observations)

$$\mathrm{median}(\text{daily emails}) = x_{\left(\frac{17+1}{2}\right)} = x_{(9)} = 6,$$

which is much more relevant (realistic ?) than the one obtained using the Sample Mean.

Here let us compare the Sample Median vs the Sample Mean: the good point of the Sample Median is that it is not affected by very large or small values, by outliers, but the Sample Mean is affected. The bad news is that Sample Median is not taking into account the actual values of our observations, rather than their ordering, but the Sample Mean is taking into account all values. So you need to take these pros/cons when using Sample Medians and Means to describe your dataset.

Remark, Sample Median Definition: Sometimes people use another definition of the sample median. As we have seen above, our definition of the Median can produce a number which is not in our dataset. In some cases this is not acceptable, so one uses another definition of the Median producing a datapoint, a number in our dataset. For example, one can define (for any case of n),

$$\mathrm{median}(x) = x_{(\lceil \frac{n}{2} \rceil)},$$

where $\lceil a \rceil$ is the smallest integer $\geqslant a$.

Remark, Theoretical Median: One can also define the theoretical median of a distribution. Assuming X is a r.v. with a CDF $F(x)$ (and, in the case if X is continuous, with a PDF $f(x)$), we call a number $m \in \mathbb{R}$ to be a median of that distribution, if

$$\mathbb{P}(X \leqslant m) \geqslant \frac{1}{2} \qquad \mathbb{P}(X \geqslant m) \geqslant \frac{1}{2}.$$

In other words, median is a number m such that the probability that $X \leqslant m$ and $X \geqslant m$ are not less that 50%.

It is easy to see that if our distribution is continuous with PDF $f(x)$, then m is a median if and only if

$$F(m) = \frac{1}{2} \qquad \text{or, equivalently,} \qquad \int_{-\infty}^{m} f(x)dx = \int_{m}^{+\infty} f(x)dx = \frac{1}{2}.$$

Example, Theoretical Median: Here let us find the Theoretical Median for the distribution with the PDF

$$f(x) = \begin{cases} 2x \cdot e^{-x^2}, & x \geqslant 0 \\ 0, & x < 0. \end{cases}$$

We have a continuous distribution here, and in this case we will calculate the CDF $F(x)$ (which is easy to do for this example) and solve $F(m) = \frac{1}{2}$ to find the median m. From the Probability course, we have

$$F(x) = \int_{-\infty}^{x} f(t)dt = \begin{cases} 0, & x < 0 \\ 1 - e^{-x^2}, & x \geqslant 0. \end{cases}$$

Clearly, we cannot have $F(m) = \frac{1}{2}$ for $m < 0$. So in our case, $m \geqslant 0$. Then we need to have

$$F(m) = 1 - e^{-m^2} = \frac{1}{2},$$

yielding $m = \sqrt{\ln 2}$.

If we will interpret this geometrically, then:

- For the PDF $f(x)$, m is a point on the OX axis such that the line $x = m$ divides the area under the graph of $y = f(x)$ into two equal parts: the area under f is exactly 0.5 for $x \in (-\infty, m]$ and exactly 0.5 for $x \in [m, +\infty)$, see Fig. 3.2

- For the CDF $F(x)$, m is an intersection point of the graph of $y = F(x)$ and the horizontal line $y = \frac{1}{2}$, see Fig. 3.3

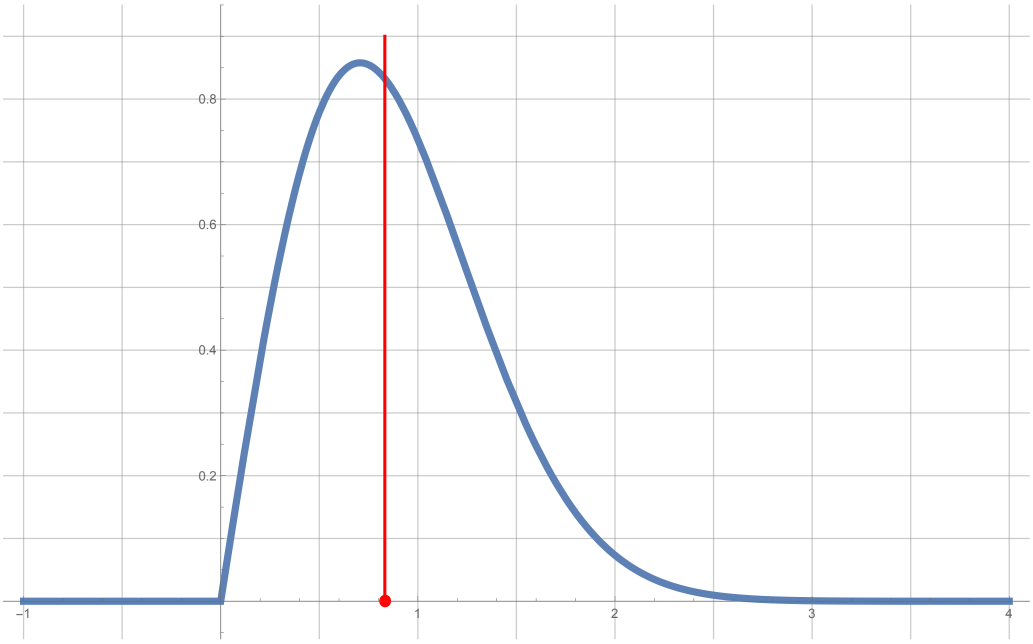

Fig. 3.2: Theoretical Median Example: Here we have the PDF graph. The line passing through the Median (red point on the OX axis) divides the area under the graph of PDF into two equal parts.

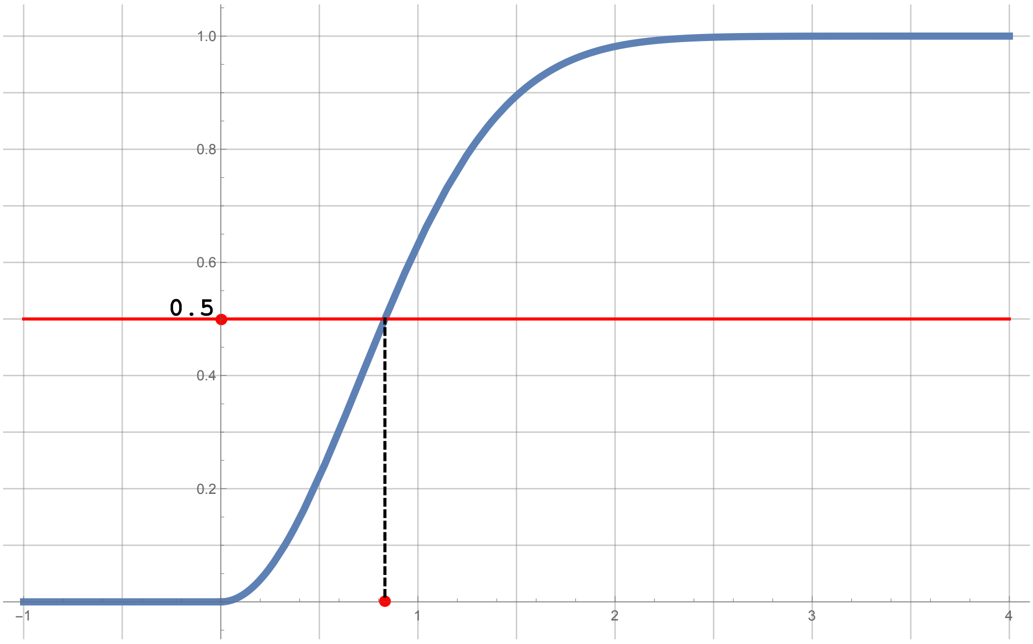

Fig. 3.3: Theoretical Median Example: Here we have the CDF graph. Median lies on the intersection of the graph of the CDF and the horizontal line $y = 0.5$.

Example, Theoretical Median: If we will try to find the Theoretical Median for the distribution with the PDF

$$f(x) = \begin{cases} 0.5, & x \in [-1,0] \cup [2,3] \\ 0, & \text{otherwise.} \end{cases}$$

then we will have infinitely many points m, such that the vertical line $x = m$ is dividing the area under the PDF into two equal-size parts. In fact, any point $m \in [0,2]$ will work. So in this case we have infinitely many Theoretical Medians, see Fig. 3.4.

This can be seen also by the graph of the CDF also: the intersection of CDF with the horizontal line $y = 0.5$ is the whole interval $[0,2]$, see Fig. 3.5.

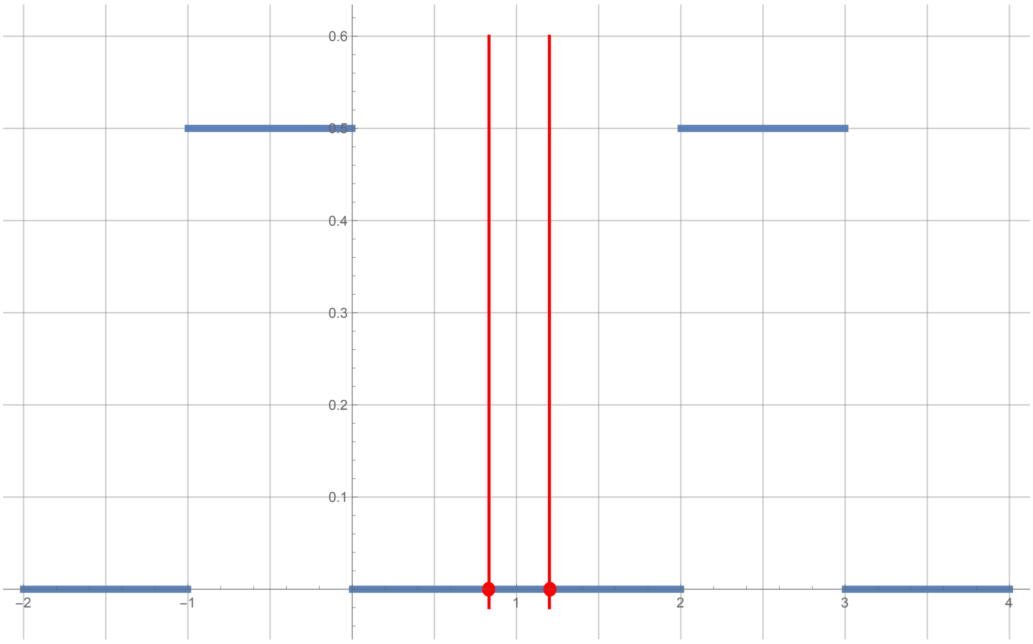

Fig. 3.4: Theoretical Median Example: Here we have the PDF graph. There are infinitely many points on the OX such that the vertical line passing through that points (say, 2 red points on the OX axis) divides the area under the graph of PDF into two equal parts.

Remark, Theoretical and Sample Medians as minimizers: Try to prove the following assertions:

a. m is a median of the r.v. X (i.w., of the theoretical distribution behind X) if and only if

$$m \in \operatorname*{argmin}_{a\in\mathbb{R}} \mathbb{E}(|X-a|)$$

b. If m is a median of the dataset $x_1, x_2, ..., x_n$, then

$$m \in \operatorname*{argmin}_{a\in\mathbb{R}} \frac{1}{n} \cdot \sum_{k=1}^{n} |x_k - a|. \tag{3.1}$$

Inversely, if m satisfies (3.1), then m divides our dataset into two equal-length parts.

Again, here we assume that the expectation above is defined.

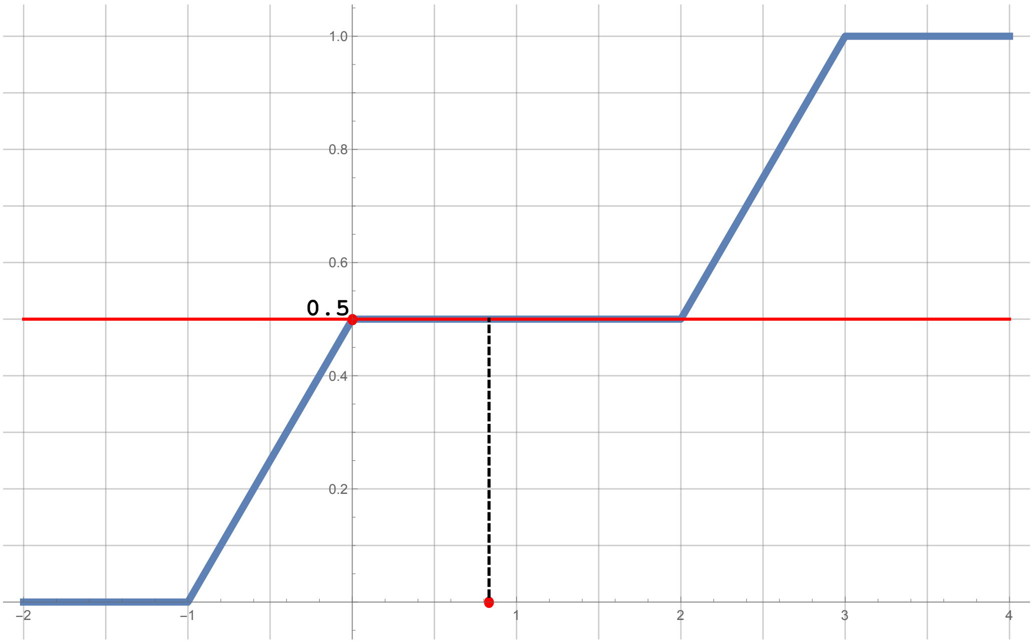

Fig. 3.5: Theoretical Median Example: Here we have the CDF graph. Median lies on the intersection of the graph of the CDF and the horizontal line $y = 0.5$. We have here that the intersection is the interval $[0, 2]$, and just one of the medians is shown

Remark, Sample Median and ECDF: Assume we have a dataset x, with data points $x_1, x_2, ..., x_n$. Then, as above, we can construct the r.v. X, taking the values x_k, with probabilities $\frac{1}{n}$ (if some x_k-s coincide, we add the corresponding probabilities).

Vnimanie, the question: Find the relation between the CDF of X, the ECDF of the dataset, the Theoretical Median of the distribution of X and the Sample Median.

Remark, Sample Median, use in DIP: See Gonzalez, Woods, Digital Image Processing, 3rd Ed, part 3.5.2, 5.3.2. Also find a lot of statistics in this books!

3.2.3 The Sample Mode

Another representative for the dataset is its most frequent element(s):

Definition 3.5. *The **Sample Mode** of the dataset is the value which occurs most frequently in our dataset.*

Example, Sample Mode: Let us consider the dataset of daily emails,

$$10, 10, 2, 5, 7, 5, 1, 2, 7, 2, 6, 8, 2, 4, 8, 6.$$

To find the mode, it is convenient first to form the frequency table:

No. of daily emails	1	2	4	5	6	7	8	10
Frequency	1	4	1	2	2	2	2	2

From this table it is clear that the most frequent value is 2 (the frequency is 4, the highest), so

$$mode(emails) = 2.$$

It is important that the Sample Mode can be non-unique. For example, the dataset $1, 2, 2, 3, 4, 4$ has two modes: 2 and 4.

In the case if we have a discrete dataset, where every value is unique, then we will say that all elements are modes of that dataset. For example, all elements of the dataset $1, 2, 3, 4$ are modes. Some authors define that in this case there is no mode at all. Also, if one is dealing with a continuous-valued dataset, then (since there is almost no chance that the values will be repeated) one first group the data into bins, calculate the frequencies for the bins, and take the bin with the highest frequency (or, usually, the midpoint of that bin) as the mode of the dataset. In other words, one calculates the $argmax$ for the frequency histogram, to find the mode of the continuous dataset.

Well, although the Mode can be far from the "center" of the data, from the "middle" of our data (and hence, not showing the central tendency), but sometimes it is used as a good representative for our data points. The most frequent data point is sometimes very important.

Remark, Unimodal, Bimodal Datasets: In Descriptive Statistics, if a dataset has a unique mode, then one calls that type of datasets Unimodal. In case we have 2 modes, one calls that type of datasets Bimodal. Multimodal datasets have more that 2 modes.

Remark, Theoretical Mode: If we have a distribution given by its PMF or PDF, then the Mode of that distribution is the $argmax$ for the PMF or PDF function. Please look at the textbooks and other papers for the Theoretical Mode, Unimodal and Bimodal distributions.

R code, Sample Mode: Unfortunately (or fortunately), **R**'s *mode(x)* command is not calculating the Sample Mode of x, rather it is giving the *type or storage mode* of x. Say, if x is a vector of numbers, for example,

```
x <- c(0,1,2,3)
```

then the result of *mode(x)* will be *"numeric"*. So let us write a function to calculate the Sample Mode of a dataset.

We create a function with a name *my.mode*:

```
#Sample Mode Calculation
my.mode <- function(x){
    y <- table(x)   #table of frequencies
    nam <- as.numeric(names(y)) # this will give unique values in x
    freq <- as.numeric(y)  #this will give the frequencies of the above values
    return(nam[freq == max(freq)]) #this will give all modes
}
```

Note that here the command *table(x)* returns a named array: names are the unique values and the elements are the corresponding frequencies. For example, if we will run

```
x <- c(0,4,9,0,9,0)
table(x)
```

Then the result will be:

```
x
0 4 9
3 1 2
```

Here the upper row is the vector of unique values in x: 1,2,3, and the bottom row show the corresponding frequencies: the frequency of 0 is 3, the frequency of 4 is 1, and the frequency of 9 is 2. Now, to extract the vector of names, we can use *names(table(x))*, and the result will be:

```
[1] "0" "4" "9"
```

These are not numbers, so in order to transform back to numbers we use *as.numeric(names(table(x)))*, and this will give the desired result. Obtaining frequencies is much easier: we use *as.numeric(table(x))* (since the actual elements of *table(x)* are the frequencies, and the values are just their names. We just need to convert that frequencies, stored as strings, back to numbers).

Next, try to make experiments/search Google to see what the code *x == max(x)* does with an array *x*, and the code *x(x == max(x))* is returning. Hope after that the code above will be clear.

Now, to test the code, we can run:

```
x <- c(10, 10, 2, 5, 7, 5, 1, 2, 7, 2, 6, 8, 2, 4, 8, 6)
my.mode(x)
```

or

```
x <- c(1,2,3,1,2,3,1,2,3,4)
my.mode(x)
```

Yeah, I know programming! Google, I am waiting for your offer!

Remark, MS Excel and Descriptive Stat: By the way, there are many Statistical functions implemented in MS Excel. For example, you can calculate the Mean, Median, Trimmed Mean, Mode a sample using Excel. And you can generate random numbers, and many more. Try to explore the real power of Excel!

3.2.4 Other Statistics for the Central Tendency

One can define also other Statistics to measure the central tendency, the "center" of the dataset. For example, given a dataset x:

$$x_1, x_2, ..., x_n,$$

one can calculate the MidRange[4]

$$\text{midrange}(x) = \frac{x_{(1)} + x_{(n)}}{2}.$$

[4] See, for example, https://en.wikipedia.org/wiki/Mid-range

Remark, MidRange as a Minimizer: We have seen above that the Sample Mean and the Sample Median are minimizers for some deviation measure. Here, for the MidRange, one can prove that if m is the MidRange of the dataset x, then

$$m \in \underset{a\in\mathbb{R}}{\operatorname{argmin}} \max_k |x_k - a|,$$

so it minimizes the maximum absolute deviation.

You can find a lot of other measures at https://en.wikipedia.org/wiki/Central_tendency. By the way, there is a nice relationship between the Sample Mean, Mode and Median for a unimodal distribution, see https://en.wikipedia.org/wiki/Central_tendency#Relationships_between_the_mean,_median_and_mode.

3.3 Statistics for the Spread/Variability

Of course, the Sample Mean or the Median are the first descriptors for a univariate numerical dataset. But, you can easily guess that this numbers are not enough to give some picture about the dataset. Say, if I will state that the mean salary for two persons is 200K, then that will not give the idea about their salaries: they both can get 200K, or one can get 150K and the other one - 250K, or, it can happen that one receives 0, and the other one - 400K. So knowing the center is not enough in most cases, we need to know how spread are the values about that center.

Now we want to give some measures for the concentration and spread, dispersion of our dataset. We will give again different measures for that.

Assume our observation, our dataset is $x_1, ..., x_n$.

3.3.1 Deviations

Definition 3.6. *If $\bar{x}$ is the Sample Mean of the dataset $x_1, ..., x_n$, then the differences*

$$x_k - \bar{x}, \qquad k = 1, ..., n$$

are called deviations of our dataset from the mean.

We can get a lot of information about the spread of our dataset using the deviations from the mean. Say, if the deviations are close to zero, then our dataset is concentrated about the mean, or if the deviations are symmetric about zero, then our dataset is symmetric about the mean. So deviations are good characteristics for our dataset spread, but, unfortunately, for a large dataset we will have a lot of numbers (deviations), so that will not help us to get the picture about the spread or variability. Instead, we want to describe the spread by just one number.

For example, we can try to find the mean of all the deviations, but, unfortunately, this will not give an information to us, because of the following Exercise:

Exercise: Prove that the sum of all deviations (and also the mean of all deviations) is 0.

Remark, Deviations from the Median: In fact, having a notion of the center of a dataset (say, Sample Mean, Median, Trimmed Mean or Weighted Mean), we can calculate the deviations from that center to estimate the variability and spread. For example, we can calculate deviations from the Sample Median of our dataset:

$$x_k - \mathrm{median}(x), \qquad k = 1, 2, ..., n.$$

Remark, Absolute Deviations: One is defining also the Absolute Deviations from the Mean as the dataset

$$|x_k - \bar{x}|, \qquad k = 1, 2, ..., n.$$

3.3.2 The Range

One of the simplest measures for the spread is the Range:

Definition 3.7. *The **range** of the dataset x is the difference*

$$\mathrm{Range}(x) = \big(\textit{the largest element in } x_k\big) \text{ - } \big(\textit{the smallest element in } x_k\big) = x_{(n)} - x_{(1)}.$$

R code, Remark On the *range* Function: Please note that in **R**, the range(x) function is returning the minimum and maximum values of the dataset x, but not the difference of that maximum and minimum. Here is an example:

```
#Range
x <- c(-2,1,3,2,4,2,5,3,0,-1,-2,-3)
range(x)
```

The result is:

```
[1] -3  5
```

To calculate the Range in our sense, we can use:

```
#Our Range, v1
x <- c(-2,1,3,2,4,2,5,3,0,-1,-2,-3)
r <- range(x)
my.range <-r[2]-r[1]
```

or just

```
#Our Range, v2
x <- c(-2,1,3,2,4,2,5,3,0,-1,-2,-3)
my.range <- max(x)-min(x)
```

Now, if we want to have a function to calculate the Range, we can write

```
#Our Range, v3, as a function
my.range <- function(x){
  return(max(x)-min(x))
}
```

and then try to use it for some dataset:

```
x <- c(2,3,9,0,1,0)
my.range(x)
```

Another method is:

```
#Our Range, v4, as a function, another version
my.range <- function(x){
  return(diff(range(x)))
}
```

And you can check by the above example dataset that this gives the same value.

3.3.3 The Sample Variance and Sample Standard Deviation

The above measure for the spread, the Sample Range, in fact, is not giving a good information for the spread, since we can have some outliers, and the rest of our dataset can be concentrated close to some point. One of the widely used measures for the spread or variability is the Sample Variance, and its square root, the Standard Deviation:

Definition 3.8. *The **Sample Variance** of our dataset is*

$$var(x) = s^2 = \frac{\sum\limits_{k=1}^{n}(x_k - \bar{x})^2}{n},$$

where $\bar{x}$ is the sample mean of our dataset:

$$\bar{x} = mean(x) = \frac{1}{n}\cdot\sum_{k=1}^{n} x_k.$$

In many Statistics books we will find also the following definition for the Sample Variance:

Definition 3.9. *The **Sample Variance** of our dataset is*

$$var(x) = s^2 = \frac{\sum\limits_{k=1}^{n}(x_k - \bar{x})^2}{n-1}.$$

There are some reasons why people prefer this last definition (later we will see that this is an *unbiased* estimate for the population variance[5]). In practice, if n is large, then these two definitions give very close results, so you can use either of them to calculate the Sample Variance.

Definition 3.10. *The **Sample Standard Deviation** is the square root of the Sample Variance:*

$$sd(x) = s = \sqrt{var(x)}.$$

[5] And sometimes taking $n-1$ instead of n in the Sample Variance calculation is referred to as Bessel's correction.

So, in fact, we will have 2 values for the Sample Standard Deviation - with either n and $n-1$ in the denominator of the Sample Variance.

Example, Sample Variance and SD: Let us calculate the Sample Variance and the Sample Standard Deviation for the dataset x:

$$-1, 2, 1, 3, 0, 2, 1$$

The number of observations here is $n = 7$.

First, we calculate the Sample mean:

$$\bar{x} = \frac{-1+2+1+3+0+2+1}{7} = \frac{8}{7}.$$

Then we calculate the Sample Variance with $n = 7$ in the denominator:

$$var(x) = \frac{1}{7} \cdot \left[\left(-1-\frac{8}{7}\right)^2 + \left(2-\frac{8}{7}\right)^2 + \left(1-\frac{8}{7}\right)^2 + \left(3-\frac{8}{7}\right)^2 + \left(0-\frac{8}{7}\right)^2 + \left(2-\frac{8}{7}\right)^2 + \left(1-\frac{8}{7}\right)^2\right] =$$

$$= \frac{532}{7^3} \approx 1.5510$$

and the Sample Standard Deviation is in this case

$$sd(x) = \sqrt{var(x)} = \sqrt{\frac{532}{7^3}} \approx 1.2454.$$

Now, if we want to calculate the Sample Variance and Standard Deviation with $n-1$ in the denominator, we just need to multiply the above var by $\frac{n}{n-1} = \frac{7}{6}$:

$$var(x) = \frac{532}{7^3} \cdot \frac{7}{6} = \frac{266}{3 \cdot 7^2} \approx 1.8095$$

and the Standard Deviation will be in this case

$$sd(x) = \sqrt{var(x)} = \sqrt{\frac{266}{3 \cdot 7^2}} \approx 1.3452.$$

Now, we give another formula to calculate the Sample Variance for the case when the denominator is n:

Proposition 3.1. *The Sample Variance (with the denominator n) can be calculated by the following formula*

$$var(x) = \frac{\sum\limits_{k=1}^{n} x_k^2}{n} - \left(\frac{\sum\limits_{k=1}^{n} x_k}{n}\right)^2 = \frac{\sum\limits_{k=1}^{n} x_k^2}{n} - \left(\bar{x}\right)^2.$$

We can write this, using an analogy with the r.v. Variance[6],

$$\mathrm{var}(x) = \mathrm{mean}(x^2) - \Big(\mathrm{mean}(x)\Big)^2,$$

where x^2 is the dataset $x_1^2, x_2^2, ..., x_n^2$. Just remember to use this in the case when the Sample Variance is with the denominator n !

The proof of the above Proposition is straightforward: one just need to do simple calculations. We leave the calculation joy to the interested reader $\ddot\smile$

REMARK, SAMPLE VARIANCE INTERPRETATION: Let us note here the link between the Sample Variance and a Variance of a r.v. . If we will define a r.v. X, which will take the values x_1, x_2, ..., x_n with the equal probabilities $\frac{1}{n}$, then the sample variance of the dataset $x_1, ..., x_n$, with the denominator n, will be exactly the variance of the r.v. X, i.e., we will have, for this X,

$$\mathrm{mean}(x) = \mathbb{E}(X) \qquad \text{and} \qquad \mathrm{var}(x) = \mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2).$$

In case we define the Sample Variance by taking $n-1$ in the denominator, then the formula above will look like:

$$\mathrm{var}(x) = \frac{\sum\limits_{k=1}^{n}(x_k - \bar{x})^2}{n-1} = \frac{\sum\limits_{k=1}^{n} x_k^2}{n-1} - \frac{\left(\sum\limits_{k=1}^{n} x_k\right)^2}{n(n-1)}.$$

R CODE, VARIANCE: In **R**, one can calculate the sample variance by using the command $\mathrm{var}(x)$ for a dataset x.

```
#Variance Calculation
x <- c(-2,1,3,2,4,2,5,3,0,-1,-2,-3)
var(x)
```

This will give the result:

```
[1] 6.727273
```

Note that **R** is using $n-1$ for the denominator. For example, if $n = 1$, i.e., we have only one observation, **R** will produce **NA**, i.e., Not Available. Try:

```
#Variance is calculated by (n-1) in the denominator
x <- c(2)
var(x)
```

The result is:

```
[1] NA
```

[6] Recall that for a r.v. X,
$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \Big(\mathbb{E}(X)\Big)^2.$$

R code, Implementation of the Variance: The following code will give the same result as $var(x)$:

```
#Variance by Sephakan Dzerqer
x <- c(-2,1,3,2,4,2,5,3,0,-1,-2,-3)
x.bar <- mean(x)
sum.of.deviations <-sum((x- x.bar)^2)
my.var <- sum.of.deviations/(length(x)-1)
```

We can also make a variance calculation as a function:

```
#Variance by Sephakan Dzerqer, v2
my.var <- function(x){
  return(sum((x-mean(x))^2)/(length(x)-1))
}
```

Now, to check the result, run

```
x <- c(-2,1,3,2,4,2,5,3,0,-1,-2,-3)
my.var(x)
```

Example, Comparing Datasets with the same mean and different variances: Now let us consider the following datasets:

$$x: \quad -3, -4, 2, 0, -1, 2, 3, 4, 8, 2, -2, -2$$

$$y: \quad 13, 19.5, 0, -30, -10, -25, -30, 40, 20, 10$$

It is easy to calculate that the Sample Mean for both of these datasets are the same:

$$mean(x) = mean(y) = 0.75.$$

But the Standard Deviations are not the same (we use the **R**'s sd, with $n-1$ in the denominator):

$$sd(x) \approx 3.41 \qquad \text{and} \qquad sd(y) \approx 23.96$$

Clearly, y is more spread out than x. See the Fig. 3.6. By the way, please note that the picture is somewhat misleading - we have different datapoints with the same values - they will represent by the same point on the graph.

R code, Comparing Datasets with the same mean and different variances: The code for the above example is the following:

```
#Comparing two datasets with the same mean and different SDs
x <- c(-3, -4, 2, 0, -1, 2, 3, 4, 8, 2, -2, -2)
xbar <- mean(x)
xsdev <- sd(x)
y <- c(13, 19.5, 0, -30, -10, -25, -30, 40, 20, 10)
ybar <- mean(y)
```

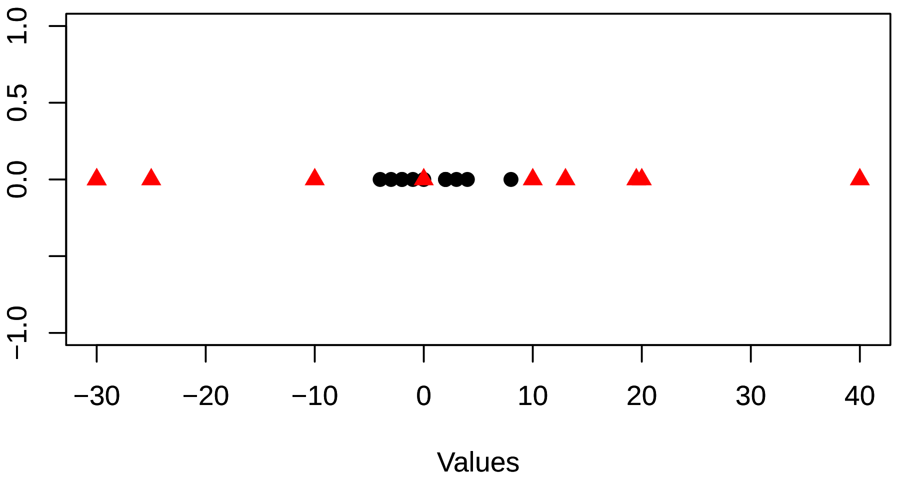

Fig. 3.6: The Datasets x (blue dots) and y (red triangles). They have the same Sample Mean, but the Sample Standard Deviation of y is larger than the x's one.

```
ysdev <- sd(y)
z1 <- rep(0, length(x))
z2 <- rep(0, length(y))
plot(x,z1, pch = 16, cex = 1.2, xlim = c(min(y), max(y)), xlab = "Values", ylab = "")
par(new = TRUE)
plot(y,z2, pch = 17, col = "red", cex = 1.2, xlim = c(min(y), max(y)),
 xlab = "Values", ylab = "")
```

Now, let us give some properties for the Sample Variance. They are the analogues for the properties of the variance for a r.v..

Proposition 3.2. *Assume x is the dataset $x_1, x_2, ..., x_n$, and $\alpha, \beta \in \mathbb{R}$ are constants. If we will denote by $\alpha \cdot x$ the dataset $\alpha \cdot x_1, \alpha \cdot x_2, ..., \alpha \cdot x_n$, and by $x + \beta$ the dataset $x_1 + \beta, x_2 + \beta, ..., x_n + \beta$, then*

a. $\mathrm{var}(x) \geqslant 0$;

b. $\mathrm{var}(x) = 0$ *if and only if* $x_k = x_j$ *for any* k, j;

c. $\mathrm{var}(\alpha \cdot x) = \alpha^2 \cdot \mathrm{var}(x)$;

d. $\mathrm{var}(x + \beta) = \mathrm{var}(x)$.

Proof. I suggest you to check these properties by yourself. □

The weak side of the Sample Variance is that it is very sensitive to outliers: if we will have an outlier, then the deviation will be large, and the squared deviation will be even larger. That's why people say that Sample Variance is not a **robust** Statistics for a spread, it is not resistant to outliers.

Remark, Sample Variance and a Minimum Problem: Above, when talking about the Sample Mean, we gave a remark that the Mean is the number minimizing the sum of squared distances. Namely,

a. m is the Mean of the r.v. X if and only if

$$m \in \underset{a\in\mathbb{R}}{\operatorname{argmin}}\, \mathbb{E}\Big((X-a)^2\Big);$$

b. m is the mean of the dataset $x_1, x_2, ..., x_n$ if and only if

$$m \in \underset{a\in\mathbb{R}}{\operatorname{argmin}}\, \frac{1}{n}\cdot\sum_{k=1}^{n}(x_k-a)^2.$$

Here we can add the following (we will use $var(x)$ with the denominator n):

a.

$$Var(X) = \min_{a\in\mathbb{R}} \mathbb{E}\Big((X-a)^2\Big);$$

b.

$$var(x) = \min_{a\in\mathbb{R}} \frac{1}{n}\cdot\sum_{k=1}^{n}(x_k-a)^2.$$

So, in fact, the Mean is the minimum point of the sum of squared distances, and the Variance is the minimum value of that sum of squared distances.

Btw, here again you can see the analogy for the r.v.'s and finite datasets: if we will make a r.v. from the dataset $x_1, ..., x_n$ by giving equal probabilities to each data point, then the discrete case can be obtained from the r.v. case.

Remark, Variance from a Median etc.: Note that if we have a measure for the central tendency, some descriptor for the typical element in our dataset (say, the mean, median, mode, weighted mean, trimmed mean,...), then we can define the Sample Variance from that typical element. For example, we can define

$$\text{Variance from the Median}(x) = \frac{1}{n}\cdot\sum_{k=1}^{n}\Big(x_k - Median(x)\Big)^2.$$

3.3.4 The Mean Absolute Deviation (MAD) from the Mean and Median

Another way of measuring the spread of a dataset is the Mean Absolute Deviation from the Mean (or Median). Of course, you can define, by analogy, the Mean Absolute Deviation from any measure of a location (say, Trimmed or Weighted Mean).

Definition 3.11. *The Mean Absolute Deviation (MAD) from the Mean for the dataset $x_1, ..., x_n$ is*

$$\mathrm{mad}(x) = \mathrm{mad}(x, \mathrm{mean}) = \frac{\sum_{k=1}^{n} |x_k - \bar{x}|}{n}.$$

This is, of course the mean of Absolute Deviations from the Mean. Analogously,

Definition 3.12. *The Mean Absolute Deviation (MAD) from the Median for the dataset $x_1, ..., x_n$ is*

$$\mathrm{mad}(x) = \mathrm{mad}(x, \mathrm{median}) = \frac{\sum_{k=1}^{n} |x_k - \mathrm{median}(x)|}{n}$$

Remark, Theoretical MAD: Clearly, the above definitions can be generalized for any theoretical distribution, for random variables. Hope you got the trick with making a r.v. from a dataset.

Say, the analogue for the MAD from the Mean for a r.v. X will be

$$MAD(X) = \mathbb{E}(|X - \mathbb{E}(X)|).$$

Remark, MAD in other Way: One can also define the *Median absolute deviations from the Mean or Median*. Try to give the definitions!

R code, MAD: Here we need to have **R** codes for functions to calculate MADs of a dataset. But we do not. Why? Do not ask! Write it by yourself ⌣̈

3.3.5 Some Experiments with a Sample Standard Deviation

Here, using **R**, I am giving some graphical plots for the Sample Standard Deviation. We will plot frequency histograms for some datasets, and show on the same graph the Sample Mean (using a thick vertical red line passing through the Sample Mean), and show the points

Sample Mean ± Sample Standard Deviation

using vertical thin red lines passing through that points.

R code, Sample Standard Deviation: Here are the codes:

```
#Sample Variance Experiments
set.seed(100)
x <- rnorm(1000, mean = 3, sd = 5)
bins <- seq(min(x)-1, max(x)+1, 1)
hist(x, breaks = bins)
xbar <- mean(x)
```

```
abline(v = xbar, col = "red", lwd = 3)
sdev <- sd(x)
abline(v = xbar - sdev, col = "red", lwd = 1)
abline(v = xbar + sdev, col = "red", lwd = 1)

set.seed(100)
x <- rnorm(1000, mean = 3, sd = 15)
bins <- seq(min(x)-1, max(x)+1, 1)
hist(x, breaks = bins)
xbar <- mean(x)
abline(v = xbar, col = "red", lwd = 3)
sdev <- sd(x)
abline(v = xbar - sdev, col = "red", lwd = 1)
abline(v = xbar + sdev, col = "red", lwd = 1)

set.seed(121)
x <- rnorm(1000, mean = 3, sd = 5)
bins <- seq(min(x)-1, max(x)+1, 1)
hist(x, breaks = bins)
xbar <- mean(x)
abline(v = xbar, col = "red", lwd = 3)
sdev <- sd(x)
abline(v = xbar - sdev, col = "red", lwd = 1)
abline(v = xbar + sdev, col = "red", lwd = 1)

set.seed(100)
x <- rexp(1000, rate = 2)
bins <- seq(min(x)-1, max(x)+1, 0.2)
hist(x, breaks = bins)
xbar <- mean(x)
abline(v = xbar, col = "red", lwd = 3)
sdev <- sd(x)
abline(v = xbar - sdev, col = "red", lwd = 1)
abline(v = xbar + sdev, col = "red", lwd = 1)
```

Here the command $set.seed(n)$ fixes, in some sense, the random number generator: every time you will run the code, you will get the same random numbers. I am suggesting you to read about $set.seed$ in the help documentation, and also to read about Random Number Generation methods.

The results are show in Fig. 3.7-3.10. If you will run the above codes, you will get *exactly* the same pictures.

3.3.6 Other Measures for the Spread/Variability

There are a lot of other measures for a Spread/Variability of a dataset. In the next section we will give the Inter-Quartile Range, which is one of the important Spread descriptors. Another one is the Winsorized Sample Variance and the Standard Deviation, which can be found in *Rand R. Wilcox, Basic Statistics: Understanding Conventional Methods and Modern Insights, Oxford University Press, 2009, p. 27-28.*

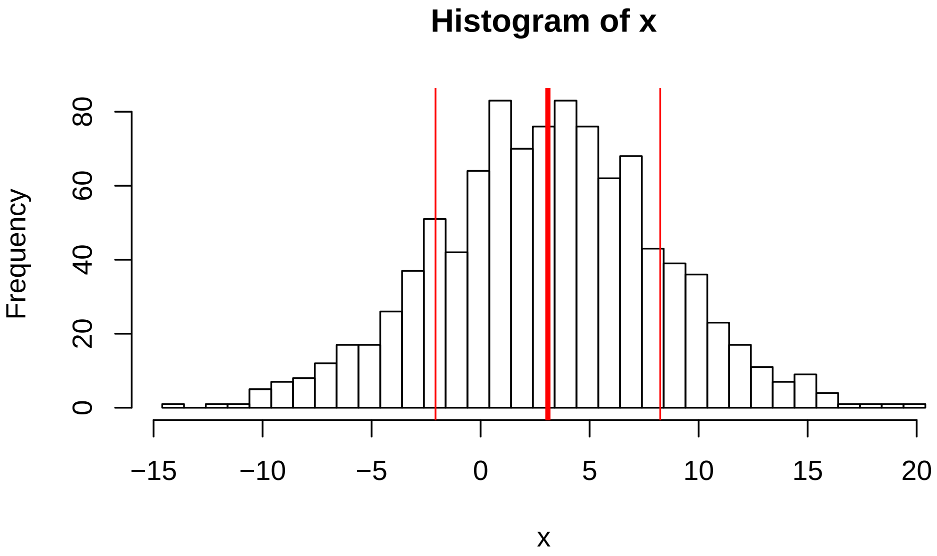

Fig. 3.7: 1000 samples from $\mathcal{N}(3, 5^2)$, from the seed 100. The Sample Mean is $\bar{x} \approx 3.084$, the Sample Standard Deviation is $\mathrm{sd}(x) \approx 5.153$.

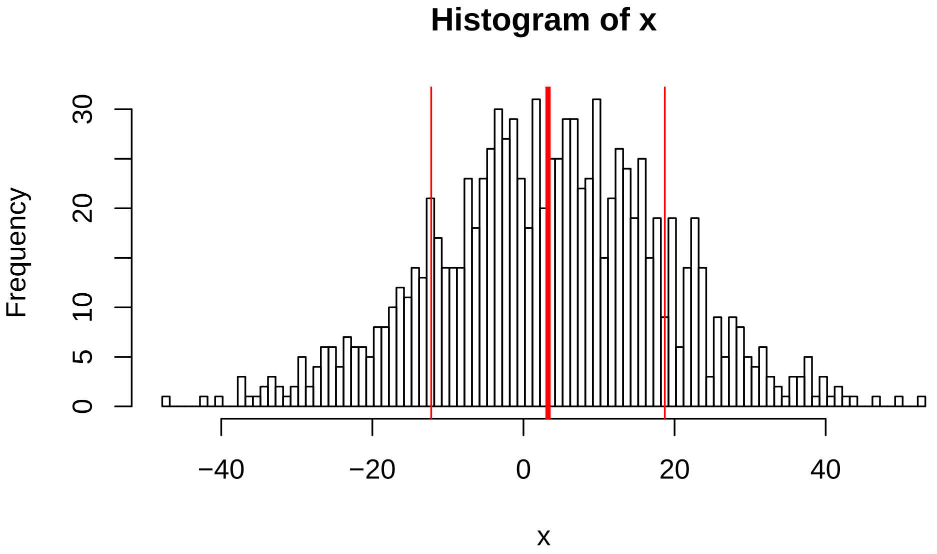

Fig. 3.8: 1000 samples from $\mathcal{N}(3, 15^2)$, from the seed 100. The Sample Mean is $\bar{x} \approx 3.252$, the Sample Standard Deviation is $\mathrm{sd}(x) \approx 15.459$.

You can find others at `https://en.wikipedia.org/wiki/Deviation_(statistics)` and at `https://en.wikipedia.org/wiki/Statistical_dispersion`.

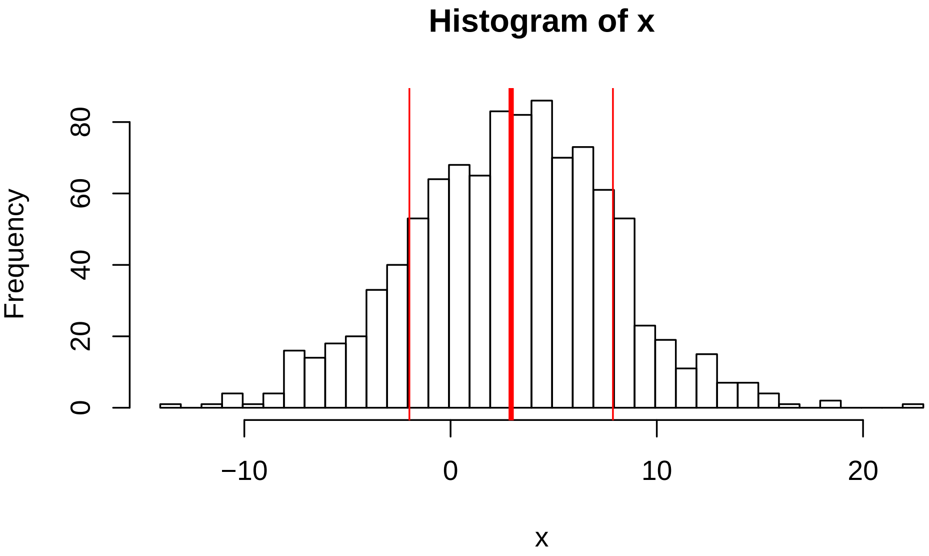

Fig. 3.9: 1000 samples from $\mathcal{N}(3, 5^2)$, from the seed 121. The Sample Mean is $\bar{x} \approx 2.939$, the Sample Standard Deviation is $\mathrm{sd}(x) \approx 4.934$.

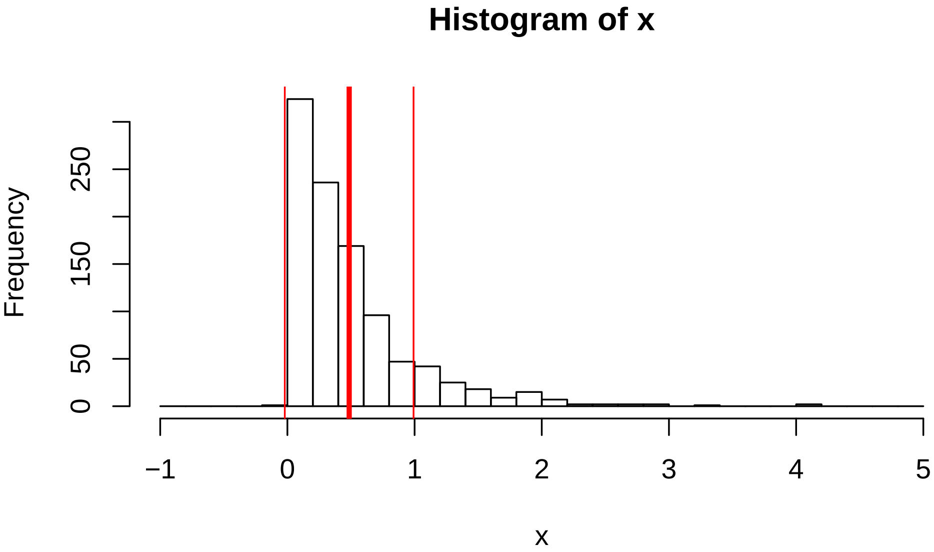

Fig. 3.10: 1000 samples from $\mathrm{Exp}(2)$, from the seed 100. The Sample Mean is $\bar{x} \approx 0.486$, the Sample Standard Deviation is $\mathrm{sd}(x) \approx 0.506$.

3.4 Other Numerical Summaries

Besides describing a dataset through the location and variability, people also give other numerical summaries. Widely used summaries are the shape parameters: Sample Kurtosis and Sample Skewness. See, for example, https://en.wikipedia.org/wiki/Kurtosis and https://en.wikipedia.org/wiki/Skewness.

Week 4

Exploratory Data Analysis for Univariate Data: Quantiles and BoxPlots

4.1 Quantiles

The idea of quantile is a generalization of the Median idea. The idea of the Sample Median was to give a number dividing the dataset into to parts, such that half of the data is to the left (or equal to) of that number, and half of the data - to the right (or equal to).

The idea of a Sample Quantile is a straightforward generalization of the Median idea: if we want to define the α-order quantile, or the α-quantile, for $\alpha \in (0,1)$, then we want to find a number that will divide our dataset into the proportions α and $1-\alpha$, i.e., we want to find a number such that $100\alpha\%$ of our datapoints will be to the left of that number (or equal to), and the rest, i.e., $100(1-\alpha)\%$ of the datapoints will be to the right (or equal to) of that number.

First we define the quantiles (or percentiles) for a dataset and for a distribution.

Theoretical Quantiles, Quantiles for Distributions Assume we have a CDF $F(x)$ for some distribution.

Definition 4.1. *For $\alpha \in (0,1)$, the α-th quantile q_α is defined by*

$$q_\alpha = q_\alpha^F = \inf\{x \in \mathbb{R} : F(x) \geqslant \alpha\}.$$

If F is strictly increasing and continuous, then the α-th quantile q_α is defined to be the unique solution to

$$F(q_\alpha) = \alpha.$$

In other words, if q_α is the α-th quantile for F, which is continuous and strictly increasing, and if X is a r.v. with CDF $F(x)$, then

$$\mathbb{P}(X \leqslant q_\alpha) = \alpha \qquad \text{and} \qquad \mathbb{P}(X > q_\alpha) = 1-\alpha.$$

So for the α-th quantile q_α, and for r.v. X with CDF $F(x)$, we will have that with probability α the values of X are to the less than or equal to q_α, and with the probability $1-\alpha$, the values of X are larger than q_α.

Geometrically, this is the point where the graph of our CDF $F(x)$ crosses or jumps over the value α.

Example: Give geometric Example:

```
#Quantile, Geometrically
plot(pnorm, xlim = c(-5,5))
par(new = T)
abline(0.3, 0, xlim = c(-5,5), lwd = 3)
```

Quantiles for a dataset Now, if we have a dataset x, then the q-th quantile is the "point" for which "exactly" $100 \cdot q\%$ of data is below that "point": say, if $q = 0.3$, then the 0.3-quantile is the "point" below which we will have 30% of our observations, and above which will be 70% of all observations.

For example, the $q = \frac{1}{2}$-quantile of a dataset x is the Median of x, and Median divides our sorted list of observations into two equal-length parts. The $\frac{1}{4}$ and $\frac{3}{4}$-quantiles are our first and third quartiles Q_1 and Q_3, and we have talked that the 25% of observations are to the left of Q_1 and the rest are to the right of Q_1. In other words,quartiles Q_1, $Q_2 = Median$, Q_3 divide our sorted dataset into four equal-length parts.

In fact, one defines also **deciles**: $D_1 < D_2 < ... < D_9$, dividing our sorted dataset into 10 equal-length parts. There is a notion of **percentiles**: $P_1 < P_2 < ... < P_{99}$, dividing our sorted dataset into 100 equal-length parts. So 1% of data is to the left of P_1, 99% of data is to the right[1]. 2% of data is to the left of P_2, 98% of data is to the right to P_2 etc.

One continues this way to define for any $\alpha \in (0, 1)$ the α-th quantile of a dataset (or the order α quantile). Please note that there are different definitions of a sample quantile, and they give slightly different values. For example, you can read the help file of R package to find the description of 9 types of quantiles.

Here is one of the definitions (given by Arnak in his slides):

Definition 4.2. *For a dataset x and $\alpha \in (0, 1)$, the quantile of order α is defined by*

$$q_\alpha = q_\alpha^x = x_{([\alpha \cdot n])}.$$

Another possible definition is (from Wesserman's book):

Definition 4.3. *For a dataset x and $\alpha \in (0, 1)$, the quantile of order α is defined by*

$$q_\alpha = q_\alpha^x = \inf\{x : \widehat{F}_n(x) \geqslant \alpha\}.$$

4.1.1 Interquartile Range, IQR

Let

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

be the sorted dataset obtained from $x_1, ..., x_n$. Assume Med is the Sample Median of our dataset.

Definition 4.4.

- *The first (or lower) quartile, Q_1, is the Median of the ordered dataset of all observations to the left of Med_x (including Med_x, if it is a data point)*
- *The second (or middle) quartile, Q_2, is the Median of our dataset;*
- *The third (or upper) quartile, Q_3, is the Median of the ordered dataset of all observations to the right of Med_x (including Med_x, if it is a data point);*
- *The InterQuartile Range, $IQR = Q_3 - Q_1$.*

!!! See https://en.wikipedia.org/wiki/Quartile for different methods to calculate the quartiles

Roughly, 25% of all observations are less than or equal to Q_1, and 25% of all observations are greater than or equal to Q_3. So quartiles divide our sorted dataset into four equal parts. Also, about 50% of observations lie in $[Q_1, Q_3]$. In some sense, IQR is the most central interval containing 50% of observations. The length of IQR shows how spread is the central portion of our data.

The following R code gives a simple summary statistics for a data:

[1] I am not specifying whether "left" or "right" to some value include the value itself. One needs to be rigorous, but... ⌣̈

```
#summary
x<-c(1,1,1,2,3,1,3,4,1,5,1,6)
summary(x)
fivenum(x)
```

Definition 4.5. *We will say that x_i is an **outlier**, if*

$$x_i \notin \left[Q_1 - \frac{3}{2}\mathrm{IQR}, Q_3 + \frac{3}{2}\mathrm{IQR}\right].$$

There are other definitions of outliers, but we will not touch this topic.

Question: Will the set of outliers be changed if we will change the scale, change the units?

Question: Why $\frac{3}{2}$ in the definition of the IQR and outliers? This comes from the Normal Distribution. Explain!!

4.2 BoxPlot

There is another convenient way of visualizing our data: boxplots. Assume $x_1, ..., x_n$ is our dataset.

To obtain the BoxPlot, we calculate:

- M, the Median of our dataset, and the lower and upper quartiles Q_1, Q_3;
- the lower and upper fences $W_1 = \min\{x_i : x_i \geqslant Q_1 - 1.5 \cdot \mathrm{IQR}\}$ and $W_2 = \max\{x_i : x_i \leqslant Q_3 + 1.5 \cdot \mathrm{IQR}\}$, i.e., the first and last observations lying in $\left[Q_1 - \frac{3}{2}\mathrm{IQR}, Q_3 + \frac{3}{2}\mathrm{IQR}\right]$; the line joining that fences to corresponding quartiles are the whiskers;
- the set of all outliers $O = \left\{x_i : x_i \notin \left[Q_1 - \frac{3}{2}\mathrm{IQR}, Q_3 + \frac{3}{2}\mathrm{IQR}\right]\right\}$

Then we draw the points W_1, Q_1, M, Q_3, W_2 on the real line and add all outliers, and make a box over $[Q_1, Q_3]$:

The R code:

```
#Boxplot
x <- rnorm(20, 1, 1)
x <- c(x, 3)
boxplot(x)
boxplot(x, horizontal = TRUE)
boxplot.stats(x)
```

and

```
#Boxplot Example 2
x <- rnorm(20)
par(mfrow=c(1,2))
boxplot(x, horizontal = TRUE)
y <- rep(0, 20)
plot(x,y)
```

and

```
#Boxplot Example 3
mtcars
boxplot(mpg~cyl,data=mtcars, main="Car Milage Data", xlab="Number of Cylinders", ylab="Miles Pe
```

Example: Plot different examples: Histogram vs BoxPlot (symmetric, skewed left, right etc)

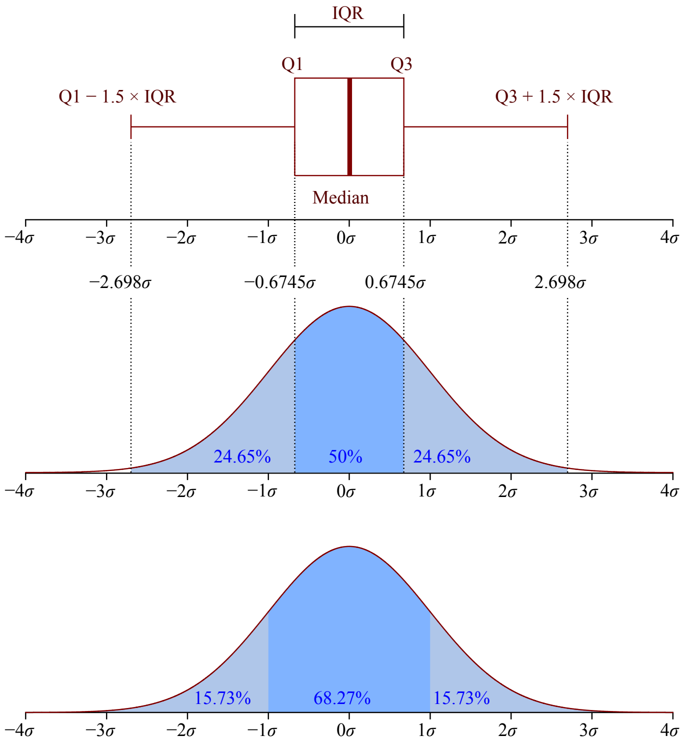

Fig. 4.1: Boxplot vs PDF for the Normal Distribution

Graphical Example:

4.3 Outliers

We are given a dataset $x_1, x_2, ..., x_n$. We want to separate atypical elements in our dataset, the outliers. Other term for describing the outliers is anomalies.

There are different methods to define outliers:

Definition 4.6. *We will say that x_i is an outlier, if*

Classical 1 -

$$|x_i - \bar{x}| \geqslant 2 \cdot sd(x)$$

Classical 2 -

$$|x_i - \bar{x}| \geqslant 3 \cdot sd(x)$$

BoxPlot Method -

$$x_i < Q_1 - 1.5 \cdot IQR \qquad \text{or} \qquad x_i > Q_3 + 1.5 \cdot IQR,$$

that is, if

$$x_i \notin \Big[Q_1 - 1.5\cdot \mathrm{IQR},\ Q_3 + 1.5\cdot \mathrm{IQR}\Big].$$

The idea behind these definitions is the following. It is easy to calculate, using some Math software, that if $X \sim \mathcal{N}(\mu, \sigma^2)$, i.e., if X is r.v. with a Normal Distribution with mean μ and standard deviation σ, then[2]

$$\mathbb{P}(|X-\mu| \leqslant 2\sigma) \xlongequal{Z=\frac{X-\mu}{\sigma}} \mathbb{P}(|Z| \leqslant 2) \approx 0.954499,$$

so

$$\mathbb{P}(|X-\mu| > 2\sigma) \approx 0.04550026.$$

R code, Normal Distribution:

```
#We calculate the probability that Z\in[-2,2]
pnorm(2)-pnorm(-2)
```

This means that if X is normally distributed with a mean μ and standard deviation σ, then the probability that X will be in $[\mu-2\sigma, \mu+2\sigma]$ is 95.4%

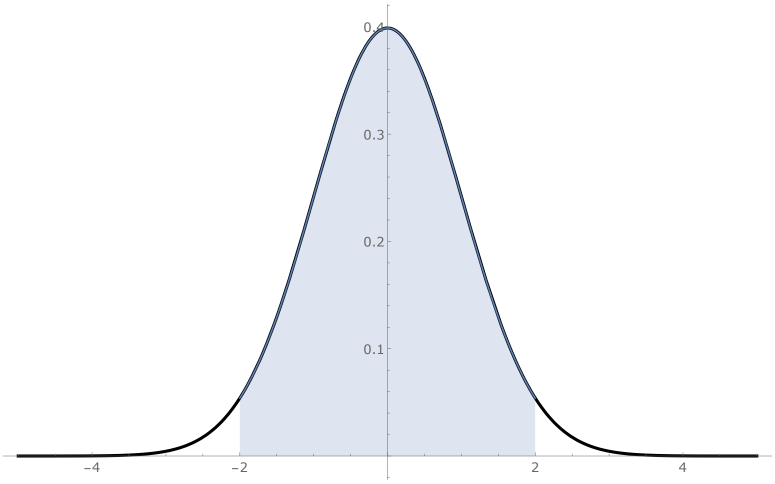

Fig. 4.2: The area under the graph of the Standard Normal Distribution from -2 to 2, the shaded region area, is 0.954

Similarly, for $X \sim \mathcal{N}(\mu, \sigma^2)$,

$$\mathbb{P}(|X-\mu| \leqslant 3\sigma) \approx 0.9973,$$

so with more that 99.7% probability, X will be in $[\mu-3\sigma, \mu+3\sigma]$. Now, if **we will assume that our data points, our observations x_k, come from the Normal Distribution**[3], then the chance that x_k will be more than 3σ away from μ is veeeery small, only 0.3%. So if x_k is not lying in $[\mu-3\sigma, \mu+3\sigma]$ we can call it an outlier (or a black swan, or a white raven,...), for sure. Please note that in the definition above we are using $\bar{x}$ and $\mathrm{sd}(x)$, but not μ and σ, because we do not have that μ and σ, and, moreover, we even do not know that our data comes from the Normal Distribution.

[2] Recall the Normalization or the Z-score!

[3] Veeeery strong assumption.

Similarly, for $X \sim \mathcal{N}(0,1)$, the lower and upper (theoretical) quartiles are

$$Q_1 \approx -0.6744898 \qquad \text{and} \qquad Q_3 \approx 0.6744898,$$

so

$$IQR = Q_3 - Q_1 \approx 1.34898.$$

Now,

$$\mathbb{P}(X \in [Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR]) \approx 0.9930234$$

meaning that if our data comes from a Normal Standard Distribution (you can translate this, by using again the normalization argument, to Normal r.v.s with other mean and variance), then the chances that a datapoint will be out of the interval $[Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR]$ is only 0.7%. So if it is not in this interval, it is an outlier.

R code, Probabilities for Standard Normal r.v.:

```
Q1=qnorm(0.25) #the 25% quantile, i.e., the lower quartile for Standard Normal RV
Q3=qnorm(0.75) #the 75% quantile, i.e., the upper quartile for Standard Normal RV
IQR = Q3 - Q1  #the IQR
pnorm(Q3+1.5*IQR)-pnorm(Q1-1.5*IQR)
```

4.4 Numerical Summaries

R: fivenum, summary,...

4.5 Interesting Things

Problem from AMM, No 11962: *Proposed by Elton Hsu, Northwestern University, Evanston, IL.* Let $\{X_n\}$, $n \geqslant 1$ be a sequence of independent and identically distributed random variables each taking the values ± 1 with probability $1/2$. Find the distribution of the random variable

$$\sqrt{\frac{1}{2} + \frac{X_1}{2}\sqrt{\frac{1}{2} + \frac{X_2}{2}\sqrt{\frac{1}{2} + \dots}}}$$