Statistics Review

Anthony Tricarico

Table of contents

Intro

This document is a collection of useful statistical concepts that will help better understand the topics of data analysis and forecasting. Depending on whether you have already taken statistics class you might find some concepts that you already know well, if that is the case feel free to skip forward or just skim through this document. Otherwise, I am sure that you will find this document helpful. Although your textbook is very good, it assumes you have prior knowledge about R and statistics and if that is not the case you might find it hard at times to fully grasp what is in it. For the purpose of this short document, topics will be divided into:

1. Descriptive statistics
2. Inferential statistics

Descriptive statistics

Descriptive statistics is the field of the discipline that deals with summarizing and making large quantities of data understandable highlighting their main features to prepare for analysis. It does so through basic algebra and visuals, which might help grasp specific features of the data that would otherwise be very hard to spot. With R you can generate summaries and plots of large quantities of data, but it is important to understand what these results convey and how you should interpret them.

Measures of central tendency

Most of the times, you are interested in understanding what value your observations (i.e., rows in your dataset) cluster around. In other words, you might be interested in knowing what are the most frequent values in your dataset. For that, you can use various *measures of central tendency*. Among these, the *mean* (a.k.a., average) is probably the most popular followed by the *median* and the *mode*. Let's take a look at them one by one.

Mean

The mean is defined as the sum of all values that a specific variable (i.e., a column in a dataset) in the dataset takes divided by the number of observations for your variable (i.e., the number of rows). Compactly, you can write this with the following notation.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The summation notation above indicates that you should sum all your observations (x_i) and then divide by n (or multiplying by $\frac{1}{n}$ which is equivalent!) the sum you obtained. The mean is usually a good measure of central tendency, but it is very sensitive to extreme values (both low and high), so you should be cautious when interpreting the mean for highly skewed variables in your dataset. An example is in order.

```
var <- c(1,1,1,2,2,2,3,4,5,6,7,8,9)
var_skewed <- c(1,1,1,1,2,3,4,10000000)                    ①

mean(var)                                                  ②
```

① creating a vector containing an extreme value
② mean of the non-skewed variable

```
[1] 3.923077
```

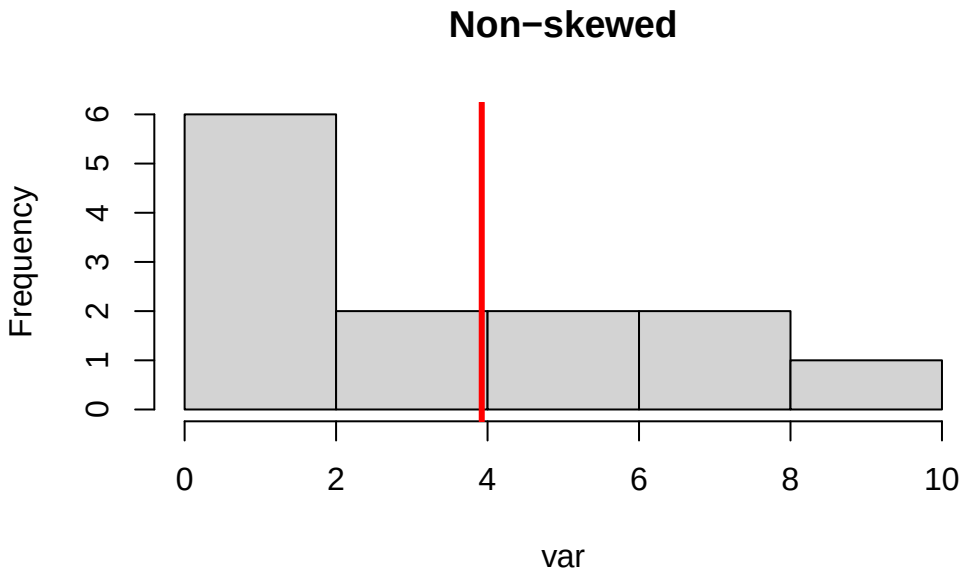
```
mean(var_skewed)
```

```
[1] 1250002
```

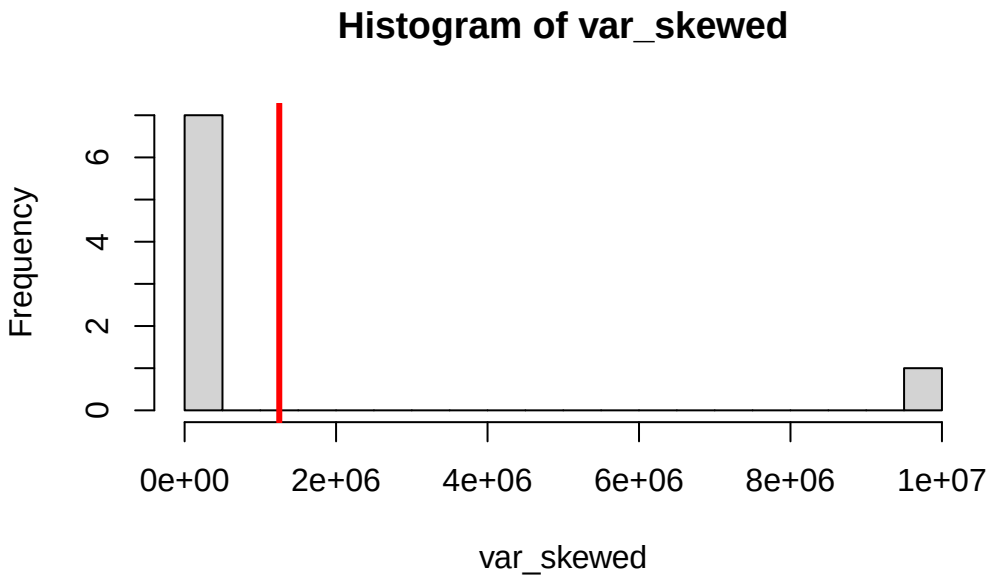
Notice how the vector containing the 10000000 has a much higher mean, illustrating how sensitive this measure is to extreme values. We can also plot this to show the difference in how those variables are *distributed*.

```r
hist(var, main = "Non-skewed")                                    ①
abline(v = mean(var), col="red", lwd=3)                           ②
```

① plotting an histogram of the *var* variable
② adding a vertical (v) line where the mean of the variable is located




```r
hist(var_skewed, breaks = 30)
abline(v = mean(var_skewed), col="red", lwd=3)
```



Notice how in the figure above the mean does not do a good job at identifying the value around which the data cluster. To improve on this, we introduce the *median*.

Median

The median.) is a measure of central tendency that is not sensitive to extreme values in the dataset (a.k.a., *outliers*). It corresponds to the middle point in an ordered set of data (when the total n is odd) or the mean of the two central middle observations (when the total n is even). To show how this measure works, let's take a look at how the median behaves when we plot it on the histogram of our data.

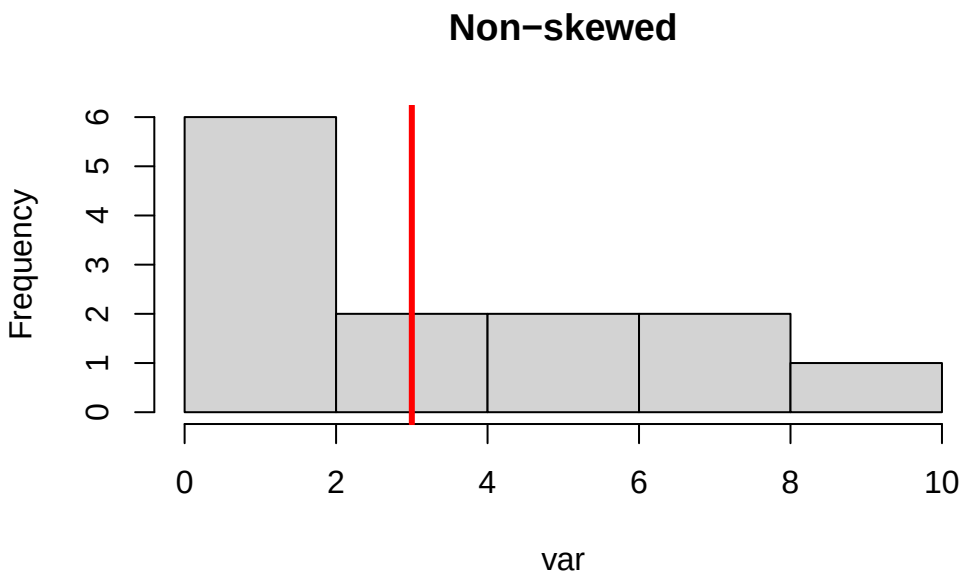
```
median(var)
```

```
[1] 3
```

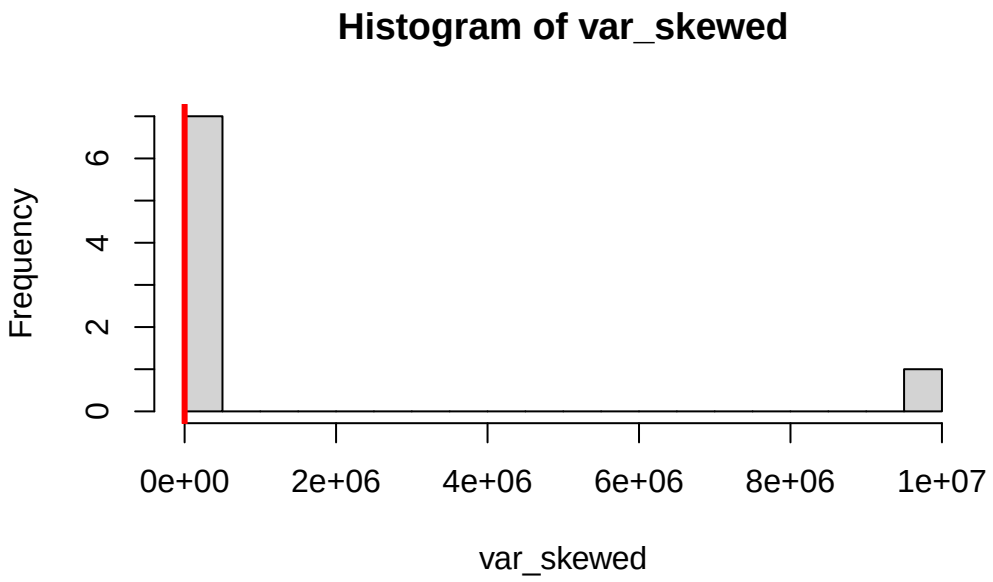
```
median(var_skewed)
```

```
[1] 1.5
```

```
hist(var, main = "Non-skewed")
abline(v = median(var), col="red", lwd=3)
```



```r
hist(var_skewed, breaks = 30)
abline(v = median(var_skewed), col="red", lwd=3)
```



from here you can clearly see that when data is highly skewed, the median is able to better pinpoint where most of the data will be located.

Mode

Finally, there is the mode which is simply the value of the observation that occurs most frequently. Let's have a look at the sample data we have to compute it. For that we can use the `table()` function in R and pass in the variable `var` as an argument to it which will return a *frequency table*showing how often each value appears in our variable.

```r
table(var)
```

```
var
1 2 3 4 5 6 7 8 9
3 3 1 1 1 1 1 1 1
```

The mode for the variable above is not unique (1,2). Indeed, the distribution of a variable might have one or more modes while others don't even have one.

```r
table(var_skewed)
```

```
var_skewed
    1     2     3     4 1e+07
    4     1     1     1     1
```

Instead, for the case above, the mode is unique and is the value 1.

> **Note**
>
> As a side note, consider the following variable X which can take values 1,2,3,4,5. Since each value appears with the same frequency (e.g., $\frac{1}{5}$), X does not have a mode!

Measures of dispersion

It is also useful to quantify by how much variables vary on average. These measures are often referred to as measures of dispersion and there are a few which are fundamental for any statistical application.

Variance

The variance of a variable measures the spread between its values and its mean.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$

The term $(X_i - \mu)$ is often referred to as *deviation from the mean*. When we add all those up (notice the $_i$ to the bottom right of the X) and we square them we have a measure of how much values of a variable are far from their mean. Since we want an average value, we divide the *squared sum of all the deviations* by the sample size N.

> **Note**
>
> μ is referred to as the population mean.
> Also, it is important to remember that the sum of all deviations from the mean is 0. This is why we square them so as to avoid negative terms in the sum and have a way to properly quantify the variance of a variable.

Variables with higher variance have values that are further away from their mean on average. The opposite is true. In R we use the `var()` function to compute the variance of a numeric variable passed as argument in the function.

```
var(var)                                                    ①
```

① here the outside `var` refers to the variance function while the *var* inside the parenthesis is the argument to the function, which in this case is our variable!

```
[1] 7.910256
```

```
var(var_skewed) #result is 12500000000000
```

```
[1] 1.25e+13
```

Think about the two results above and make sure they make sense before moving on! (Don't get confused by the scientific notation employed here by R to show the output)

> **Caution**
>
> When computing the variance of a variable you have to keep in mind that the resulting variance will be denominated by the square of the original unit of measure.
> For example, if my variable R representing the return on a stock is denominated in USD its variance ($Var[R]$) will be denominated in USD^2. This makes the results less interpretable because in real life there is no such thing as squared dollars!
> Next we learn how to address this problem.

Standard deviation

Standard deviation is defined as the square root of the variance. This solves the issue with the squared units of measures and brings everything back to unit scale so that the results become easier to interpret. In R you use `sd()` to compute the standard deviation of a variable.

```
sd(var)
```

```
[1] 2.812518
```

```
sd(var_skewed)
```

```
[1] 3535533
```

Additional resources

We've seen most of the basics by now, but in case you want to learn more about measures of dispersion here is a list of topics which you might find useful:

1. Quantiles and Quartiles
2. Interquartile range

Probability Distributions

Probability theory refers to the set of tools mathematicians developed to study and model uncertainty. Here we are just going to introduce elementary topics to not make things too complicated.

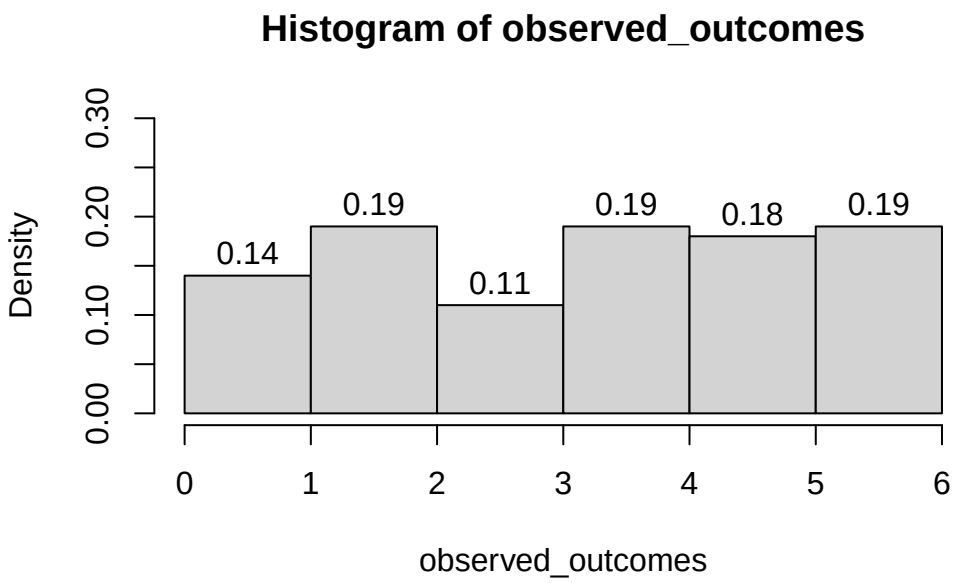
When we try to understand what are the chances that something happen or does not happen, we are basically questioning whether an *event* will or will not take place. An event is an observable outcome which can either happen or not. For instance, if my event A is defined as observing a 2 on a single die roll it can either happen (I roll the die and I get a 2) or it does not happen (I get a number other than 2). Now, assuming that the die is fair (i.e., every number as an equal chance to show), we can assign a probability to the event that we get a 2, denoted by $P(A) = \frac{1}{6}$ (meaning there is just a single 2, over a total of 6 possible numbers that I can get from a single die roll). In general, we can denote by X a random variable that takes on each roll the value of the observed number. Then, X is random because its value cannot be predetermined exactly *a priori* but is the result of the following die roll.

Binomial distribution

Knowing this, we can use a probability distribution to compute the probability that we observe a specific number of 2's over a series of n die rolls. This is an example of a binomial experiment. If you are interested you can learn more about it but for now it suffices to say that we can model this using a binomial distribution. Plotting the distribution we observe that rolling the die 100 times yields the following.

```
set.seed(126)
possible_outcomes <- seq(1,6)
observed_outcomes <- sample(possible_outcomes, size = 100,
                            replace = T, prob = c(1/6,1/6,1/6,1/6,1/6,1/6))

hist(observed_outcomes, breaks = c(0,1,2,3,4,5,6), freq = F,
     labels = T, ylim = c(0,.3))
```




```
table(observed_outcomes)
```

```
observed_outcomes
 1  2  3  4  5  6
14 19 11 19 18 19
```

In this case 19% of all rolls (meaning 19 rolls) produced had as an outcome a 2. We could have computed this probability before conducting the experiment so as to form an expectation of what the outcome of the die roll would more likely be by multiplying the probability of observing a 2 on a die roll ($P(A) = \frac{1}{6}$) by the number of times the die is rolled ($n = 100$) and the expected numbers of 2 would have been approximately 17, not that far from what we observed!

Normal distribution

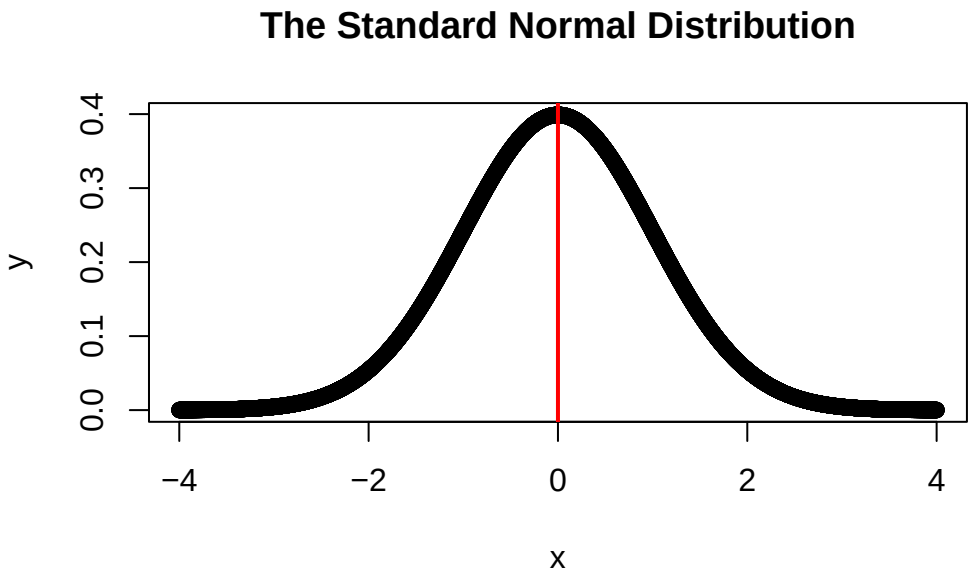
If you had to learn about only one distribution, let that distribution be the Normal Distribution. There are so many important things there are to say about this distribution, but let's keep it simple:

1. it is used to model many real-life phenomena
2. it depends on only two parameters its mean ($ $) and its variance ($ ^2$)
3. given some assumptions, any distribution can be approximated by the Normal Distribution! This is a result of the Central Limit Theorem.
4. it has many nice properties and is used all over statistics and forecasting, along with its sample counterpart being the T distribution

Let's plot it out

```
x <- seq(-4,4,.001)
y <- dnorm(x)

plot(x,y, main = "The Standard Normal Distribution")
abline(v=mean(x), col = "red", lwd=2)
```



Right, there is a special case of the normal distribution which occur when $\mu = 0$ and $\sigma^2 = 1$. This is called the standard normal distribution (represented by Z). This means that one can *standardize*any normal random variable to bring it back to this standard scale. To do that, you simply need to subtract the *mean* of the variable you are standardizing and divide by its *standard deviation*.

$$Z = \frac{X - \mu}{\sigma}$$