Module 2. Exploratory and Descriptive Analytics in R

Unit 1.1 Exploratory and Descriptive Analytics in R

As a sports scientist, it is recommended that you are well versed in the fundamentals of statistics. As such, this module has been developed to provide you with a basic all around knowledge of statistical concepts that should be kept in mind when performing analytics on sports performance data. We will begin by covering the measures of central tendency, no pun intended, but many of the surrogate measures centre around them.

Measures of Central Tendency

In basic maths, we learned about the **average**, which is calculated by summing up all the values and dividing by the total count of that set of numbers.

The average is calculated as follows; you have a set of numbers, for example the following:

5 11 56 18 78

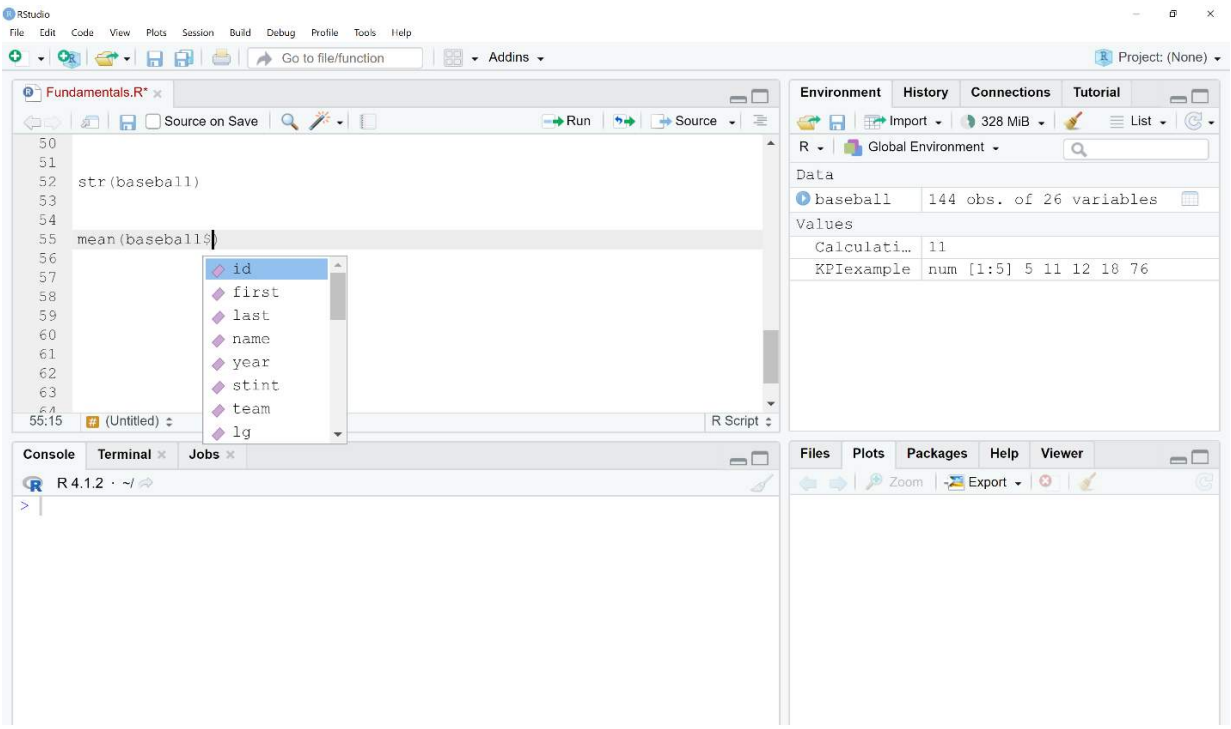
The 5 is summed up to 11; then summed up to 56; then summed up to 18, and then summed up 78, and then the summed up total of 168 is divided by 5, the final answer would be 168 divided by 5 = 33.6

It is also commonly referred to as the **mean**. In RStudio, we can calculate the mean of any variable by implementing the mean() as such:

- mean()

Let us verify this in R using the baseball dataset created in the last module and simply executing the line of code mean(). Now, this will differ from the previous module when we applied the summary() function on the entire data frame, as that provided the mean for all variables along with five other metrics. When applying a single statistic measure on a data frame, it is required that you specify on which variable. We do this by using the $ to gain access. The moment you type the mean function with baseball embedded within followed by the $, you will be displayed a drop-down menu with all the possible variables that you can choose from, as displayed in the figure below.

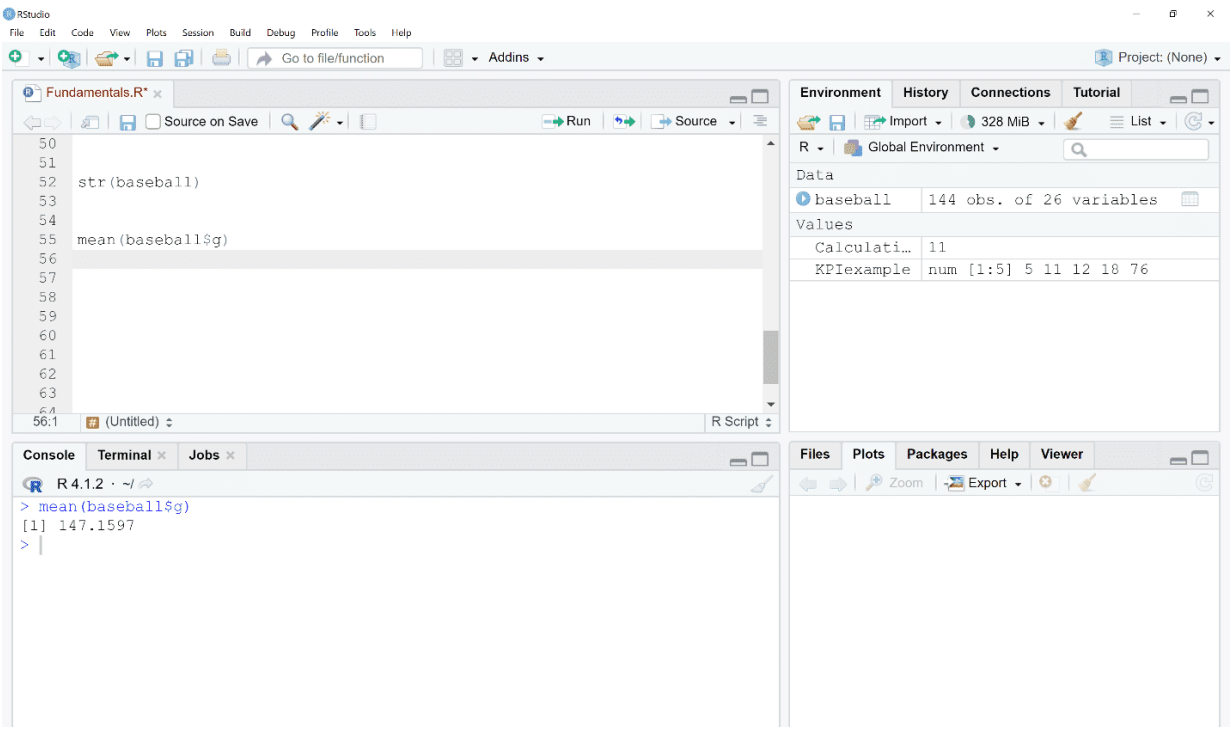
Figure 1: Drop-down menu of possible variables



Source: Screenshot by author from RStudio (RStudio, 2022).

Choose a variable, which in this case to follow along choose the variable called g within the baseball data frame, which represents the number of games played by each player. Then, you can see how the mean function can be applied to a variable within the data frame, as displayed in the figure below.

Figure 2: Mean function applied to a variable



Source: Screenshot by author from RStudio (RStudio, 2022).

The next measure of central tendency is the **mode**, and it is the most commonly appearing number, for instance, in a set of numbers as follows:

5, 11, 11, 11, 12, 18, 78

The mode would simply be 11.

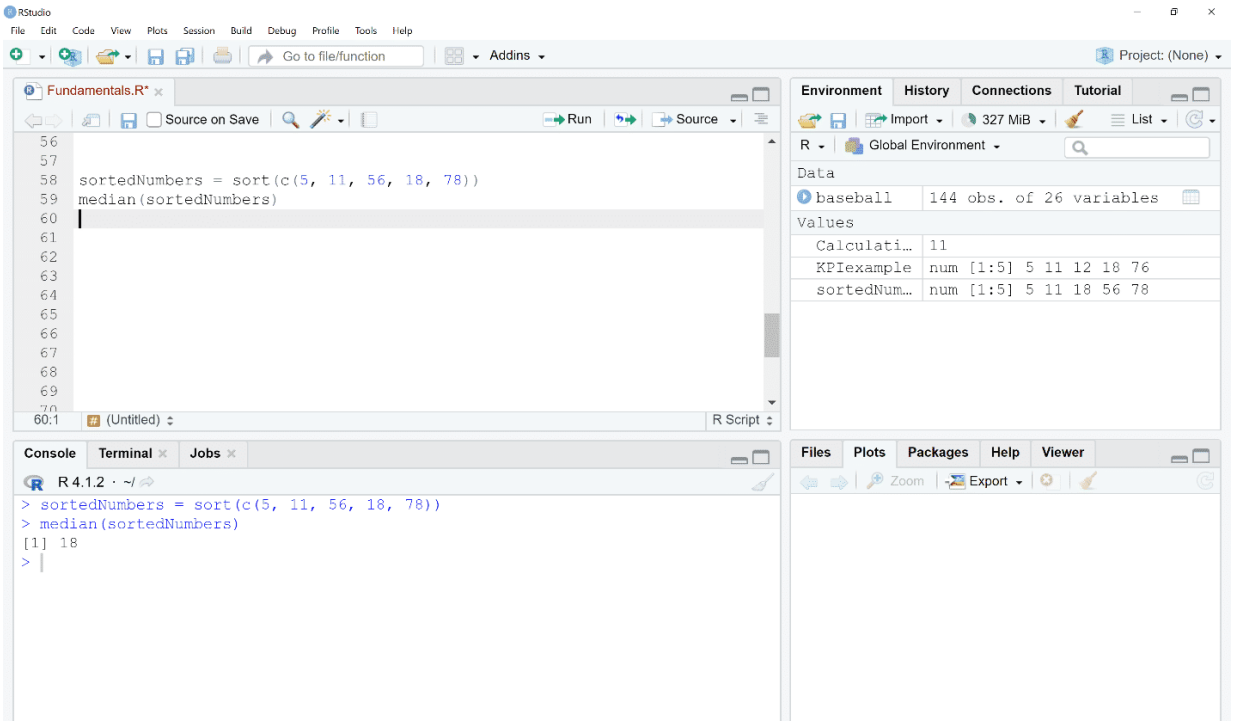
Finally, an extremely useful measure of central tendency would be the **median**, which is technically the 50th percentile. The median is often used instead of the average when you do not want to have so much of the influence of outliers, meaning that it is less sensitive to outliers.

Therefore, when working with outliers, assuming that they are not data entry errors, then we might want to include them and, therefore, work with the mean; however, if we think that the outlier is not representative of what is happening in the field, then the median would be the most appropriate measure of central tendency.

The median is calculated by first sorting the values and then using the following R code function median(). Let us first create an object called sortedNumbers that consists of the following numeric values: 56, 78, 18, 5, 11. After doing so, then you have to sort the numbers as shown on the figure below, followed by implementing the function median().

- sortedNumbers = sort(c(5, 11, 56, 18, 78))
- median(sortednumbers)

Figure 3: Example of R code to sort, create a new object, and calculate the median

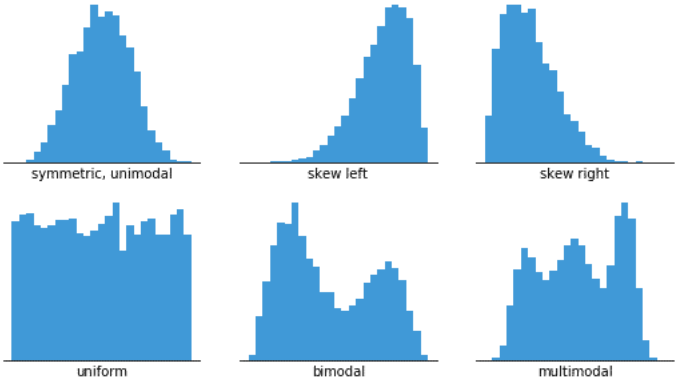


Source: Screenshot by author from RStudio (RStudio, 2022).

Also, keep in mind that as general best standards of practice for coding, when connecting several words in an object name, it is customary to either change the case, lower or upper as done in this example, or use a period or underscore in between.

Another way to examine these measures of central tendency is by visually looking at the distribution of the data to examine where the centre is and whether they are skewed or not, if there are any outliers, and how much dispersion there is of the data. (Please see the figure below that includes images of unimodal, bimodal, unimodal, skewed data, and outliers).

Figure 4: different distributions of data



Source: Yi, n.d., https://bit.ly/3RovwPe

Indicators to distinguish the central tendency:

- Mean, mode, median
- Shape – skewness
- Dispersion – outliers

Visually inspecting the data is an essential way to determine the shape of the data. It is important to identify whether the data is skewed or not, as this can have implications as to how you convey your findings, as well as whether you may want to gather additional samples of data to meet the criteria for the central limit theorem (CLT). More on the CLT a bit later on. But first, what exactly is skewness? What is left skewed, negatively skewed, right skewed, and positively skewed?

Skewness is a term defined by having the bulk of data at one end and the tail of the data on the opposite end, thereby nulling the possibility of having a normally, symmetrically distributed dataset. In other words, it seems as the bulk of scores of the KPI you are quantifying is grouping at one end or another, but not in a bell shaped format, rather it is being grouped at the low end (right and positively skewed) with few scores at the top end, typically denoted by the right side or vice versa.

Key pointers

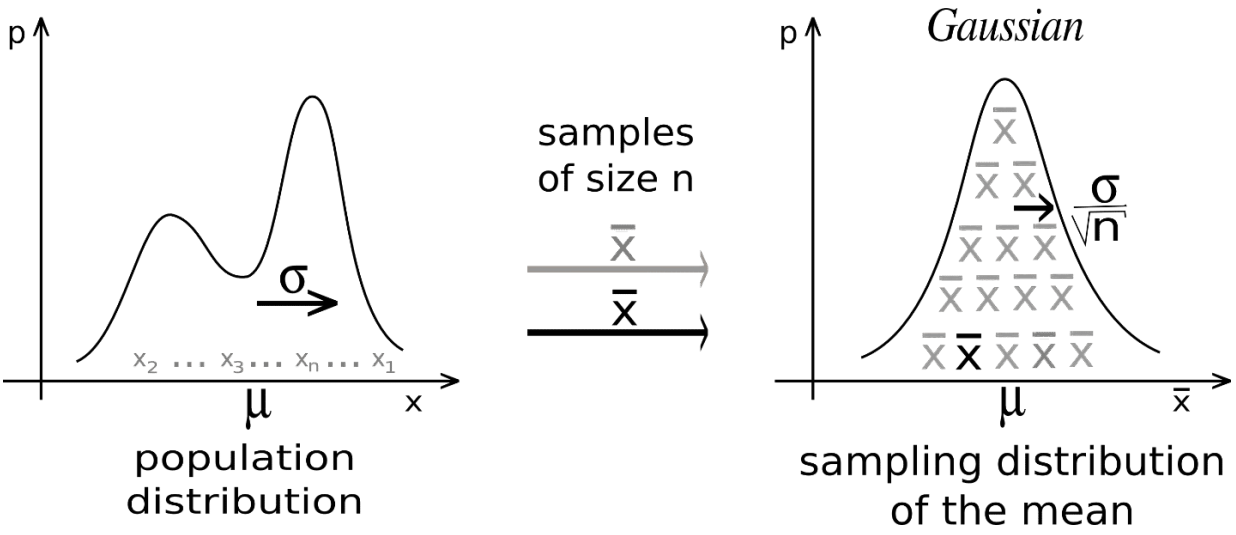
When the tail is longer on the right side, the distribution is positive skewed, or also referred to as right skewed. When the tail is longer on the left side and the bulk of the data is on the right side, the distribution is considered negatively

skewed or left skewed.

The reason it is important to examine whether your data is right or left skewed is because it will determine what types of analyses can be applied to this data, as well as whether you may need to collect more data so that it can meet the criteria for the normal distribution. How does this happen with more data? Without going into the mathematical proof, there is a theorem that states that when you have enough data, and you may be wondering what is enough data? In statistics, the rule of thumb is that if you have a sample with an n of at least 30, you can assume normality. Why is this important? Well, because when you assume normality and a bell shaped curve, there are statistical analyses that can be implemented, whereas otherwise, you may have to settle for non-parametric analyses (more on that later) that are not as granular in detail.

The **Central Limit Theorem** states that the sample mean will approximately be normally distributed for sample sizes typically greater than an N of 30, regardless of the distribution from which we are sampling.

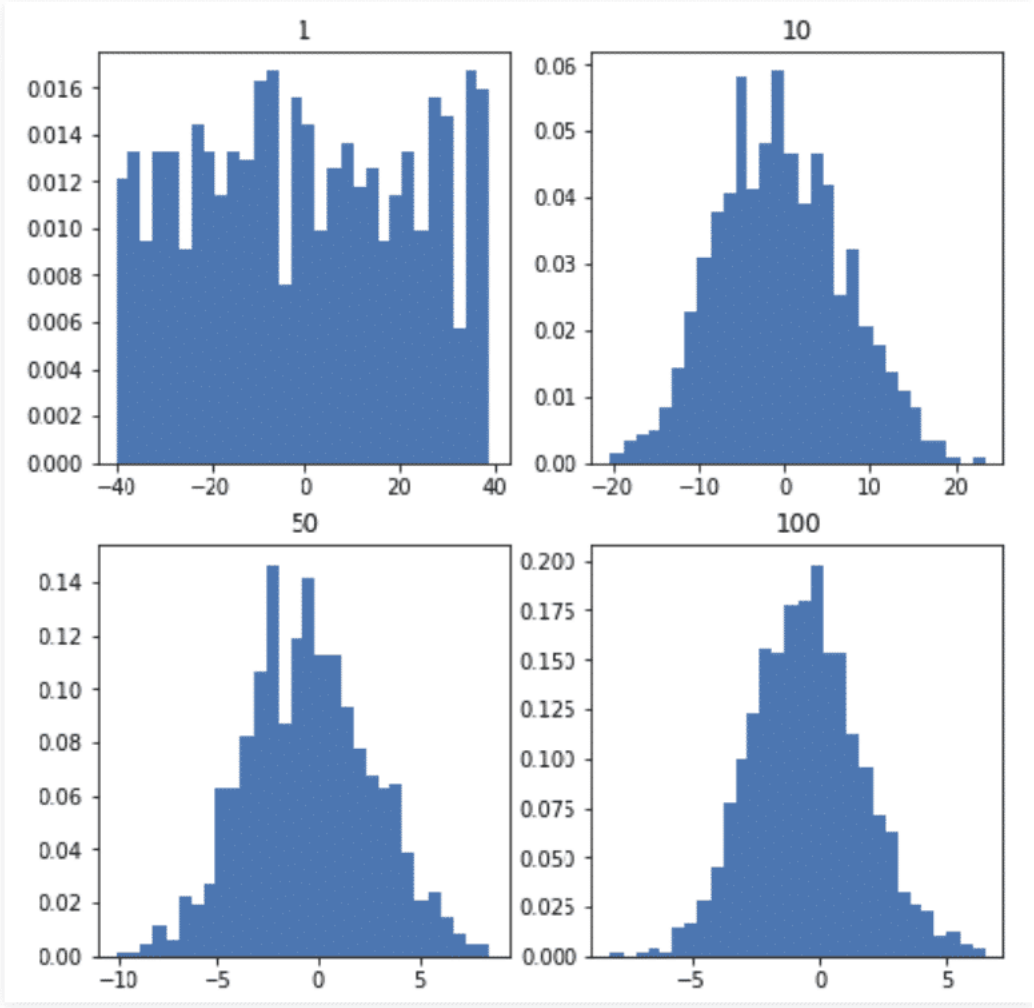
Figure 5: Central Limit Theorem



Source: Geeks for Geeks, 2021, https://bit.ly/3CnobRh

In the figure below, we can see that as the sample size increases from 1 to 100, the histogram tends to take the shape of a normal distribution.

Figure 6: Normal distribution



Source: Geeks for Geeks, 2021, https://bit.ly/3CnobRh

Outliers

Let us discuss outliers in the dataset, your dataset can be normally distributed or not, and still have outliers at either end of the curve.

Problem _

In professional sports, this is an everyday occurrence, and the sports scientist is left to decide whether to include the outliers in the final analysis or not.

Solution _

It is strongly recommended that you do not skip and that you are thorough in your approach to your analyses. Therefore, conduct an analysis with the outliers included in the data, as well as without them. When you convey the information to the key stakeholder interested in the finding; whether it be the coach, the GM, or other sports performance staff, be ready to provide context for why you conducted the analyses with and without them. The response to such a question could be related to whether you are analysing your star

player (who typically has outlier values) and comparing their values to the rest of the team or other players in general or whether you are interested in the average result of most of the players on the team (analysis without the outliers).

Finally, the data visualization will allow you to inspect for outliers as well. If you have the raw units labelled, then you will be able to determine if the outliers are considered unusual or extreme outliers, depending on whether there are 2 or 3 standard deviations above and below the mean. All that is needed to calculate this is the mean and the standard deviation.

We will be touching on standard deviation after we discuss variation, and then we will also discuss the commonly applied z score.

Measures of Variation

The most important concept in statistics is variation. Being able to fully comprehend variability will allow you to answer questions such as the following: Where do the data points land? How far are they away from each other and the mean? What is the lowest value and the highest value? All these questions are answered with measures of variance, this quantifies how the data is dispersed. Why is this important? Well, think of an example where you have a mean of 3 goals scored, and it is based on data where the teams scored between 2 and 4 goals consistently. But then think of a scenario where the mean of 3 goals is calculated from teams that scored 1 and 5 goals, yet the mean is 3. The point is that these are two very different types of scenarios and the context is key; this exemplifies why you have to examine the variance and the dispersion of the variables of interest.

How exactly is it quantified? There are several ways to measure variation and variability (interchangeable).

Variance – how do we calculate variance? Well, variance is calculated from measuring the distance between each data point and the mean of the sample.

This brings us to the definition of a **sample**, what is a sample? It is a group of randomly selected individuals (observations, data points, etc.) that are chosen in the hopes of inferring and generalising findings about the **population**. Key terms related to the population and sample are parameters and statistics. These glossary terms are used to describe the main characteristics of their respective domain; for instance, **statistics**, such as the mean, are denoted and pronounced as xbar and sample standard deviations are characteristics of sample data, whereas **parameters**, such as the population mean pronounced mu and sigma, are the metrics that characterise the population. It is also important to keep in mind that, when referring to the population, Greek letters are invoked, whereas when referring to the sample statistics, Roman letters are implemented.

Population Parameters

μ = population mean
σ = population standard deviation
π = population proportion

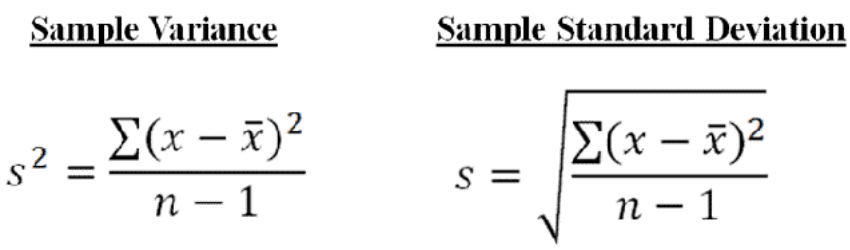
Sample Statistics

xbar = sample mean
s = sample standard deviation
p = sample proportion

Although variance sounds ideal to calculate, the difference from data point to mean, squaring them, and then summing up the numbers, it is the **standard deviation** that is most commonly relied upon. This is so for two main reasons: 1) It becomes a positive number no matter what. Originally, a data point that took on a value of -8 is now a + 64; 2) The reason we prefer the standard deviation is that we divide by total sample size and then get the square root, which brings us back to an area where we can cross interpret the data in relevant units.

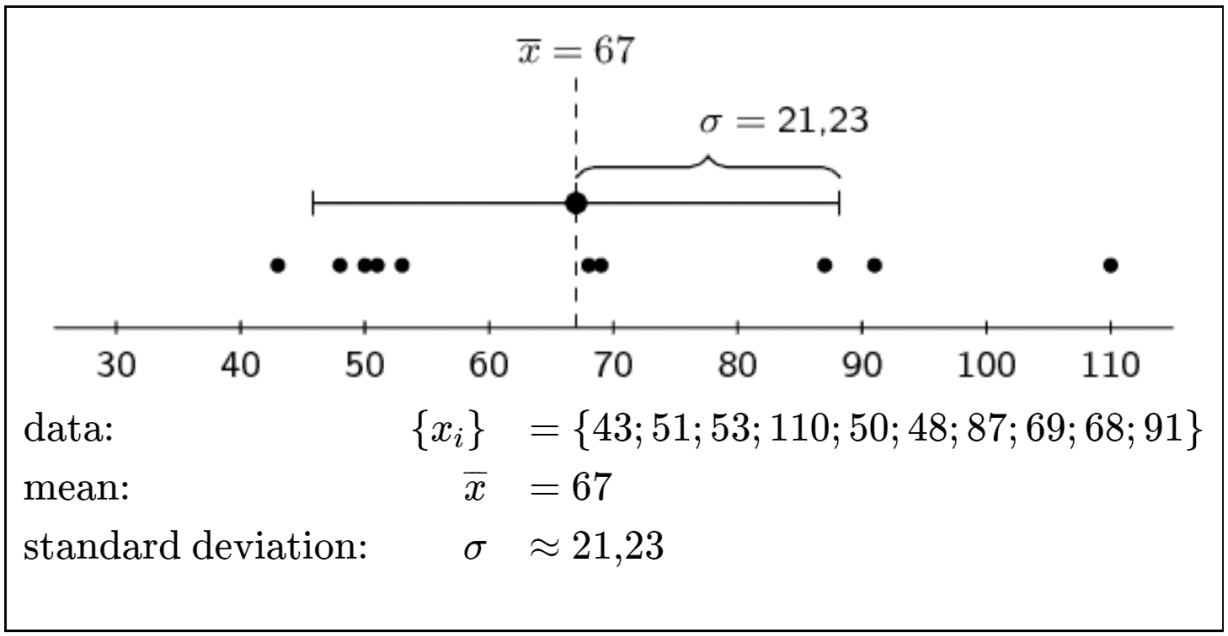
Below, is an image of the equation of variance and standard deviation.

Figure 7: Equation of variance and standard deviation



Source: Statistics Lectures, (n.d.), https://bit.ly/3SOXSue

Figure 8: The data, mean, variance and standard deviation



Source: Statistics Lectures, (n.d.), https://bit.ly/3SOXSue

The **standard deviation** of the population is called sigma, and a Greek symbol (pronounced sigma) is used to exemplify it. Sample standard deviation commonly referred to as s, or the Roman letter s is used to refer to the sample standard deviation and in R and RStudio, the code is sd().

The **range** is another option for measuring dispersion. However, the advantage is that it only needs and uses two values to calculate the range, the minimum value and the maximum value, but the disadvantage is that as well, that only two values are used to determine the dispersion. The calculation of the range is as follows: get the maximum value and subtract the minimum value = range.

Normal and Standard Normal Distribution

The **normal distribution** represents data that is symmetrical and mirrors itself on each side with the raw units of measurements.

The standard normal distribution looks similar to the normal distribution; however, it differs in that the raw values are now identified by being 1, 2, and 3 standard deviations above or below the mean implementing the z score.

- *Z score*

Recall early on, we were examining data to identify if there were outliers by determining if there were 2 or 3 standard deviations above or below the mean.

If raw units were implemented, then we assume that the data was based on a normally distributed sample dataset. However, if the data is converted to a mean of 0 and a standard deviation of 1, then we are working with the standard normal distribution. All this means is that the mean has now, in other words, shifted to a mean of 0 and a standard deviation of 1.

How would we calculate a z score? Get the data point of the athlete or player of interest in the variable of choice and subtract the mean of the sample, then divide by the sample standard deviation.

In R and RStudio, we would have to manually type the code as follows:

Zscore = x-xbar/s

And if preferable, we would create a function called z score that would automatically execute the function above when we call the code zscore().

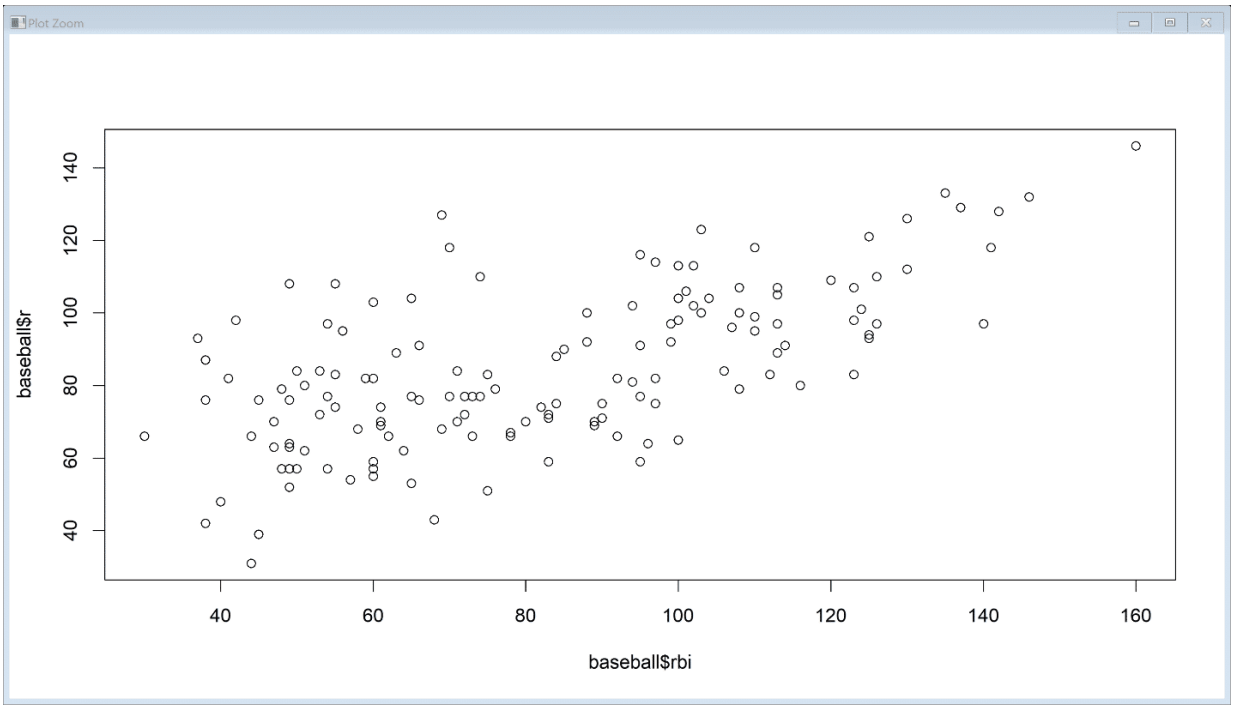
A term to know: **correlation**, it is a basic type of analysis that is very informative.

However, it is essential to know that two assumptions must be met to run a true and accurate correlational analysis:

- The two variables must be numeric.
- When plotting the x and y variables, there must be a visible linear relationship.

If either of these assumptions are not met, then running a correlational analysis will still yield a resulting correlation, but it would be inaccurate to extrapolate from the findings. If the assumptions were met and the correlational analysis was performed correctly, then it will determine the strength and the direction of the relationship between the variables. For instance, with the baseball data frame, we can easily inspect the two numeric variables r (runs) with rbi (runs batted in) by simply executing a plot command, this allows us to verify if there is linearity, which would then allow us to either decide to run the correlation officially or not. In this case, we can see that there does seem to be some sort of linearity, so we can then proceed to examine the actual correlation coefficient.

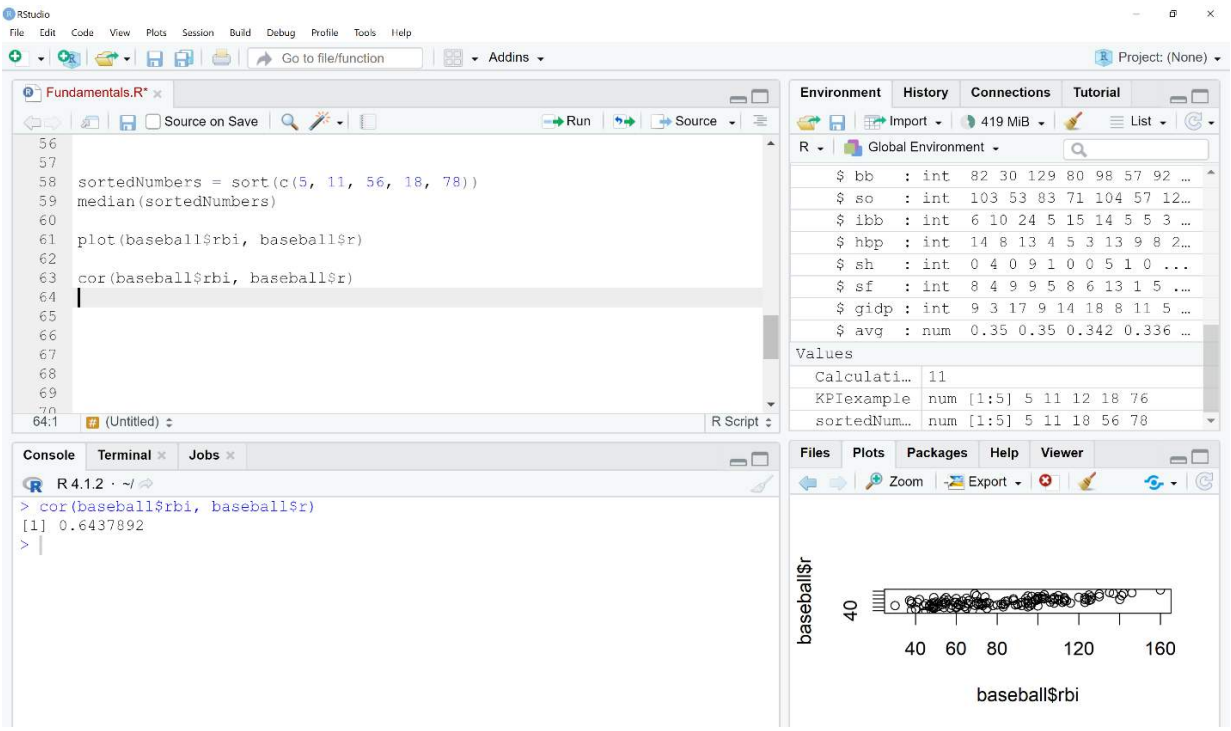
Figure 9: Basic plot of two numeric variables, r and rbis



Source: Screenshot by author from RStudio (RStudio, 2022).

Then, after visually inspecting for linearity, we can execute the line of code for the correlations' coefficient, as shown in the figure below, where the correlation coefficient is .64, meaning that there is a moderately positive correlation between runs and rbis in this sample baseball data frame.

Figure 10: Plot and cor command yielding correlation coefficient



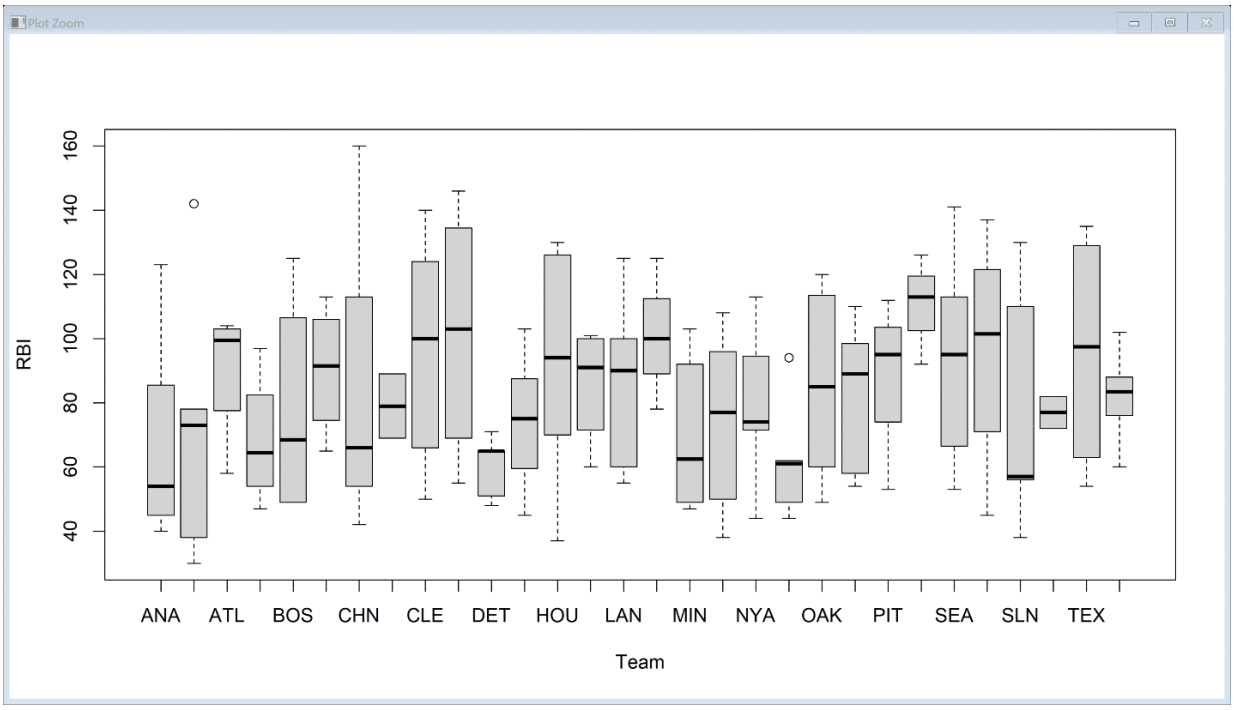
Source: Screenshot by author from RStudio (RStudio, 2022).

What to do if the variables are not numeric?

If the data you are working with is categorical, then a bar or column chart would be a better choice for a visual. In commonly used statistics textbooks, the main difference between a column and bar chart is only the orientation of the bars, whether they are placed vertically or horizontally.

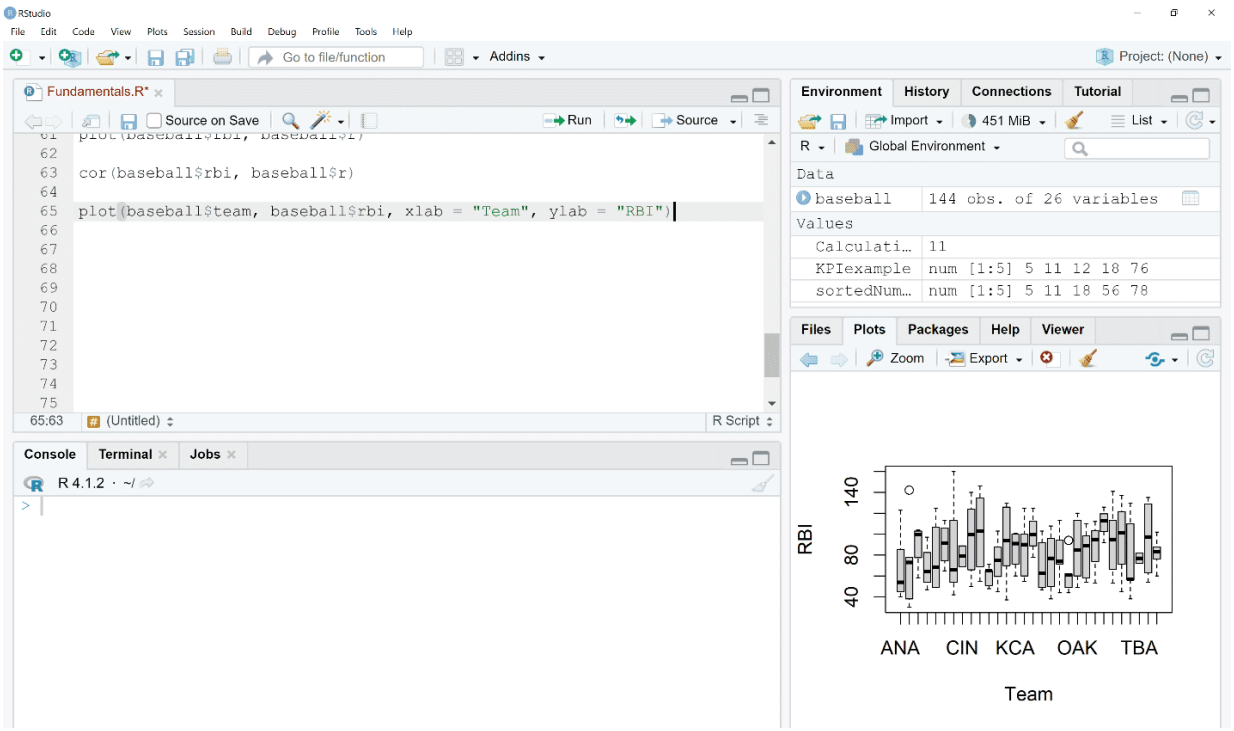
In R and RStudio, in module 4, we will cover both the bar and horizontal bar charts and how the main difference is that the bar, which is coded using geom_bar(), is meant to visualise one variables' distribution, meaning it will be plotted by default on the x-axis and the y-axis will be the frequency or count. Whereas, when running the command to run a column chart using the geom_col(), the difference is that the y-axis will take the form of the y variable rather than the count or frequency (more on this in module 4 regarding data visualisation in R and RStudio). However, for simplicity, since ggplot will be introduced in module 4, we will execute the basic plot command which will auto-identify to place the variables in bars based on their data type.

Figure 11: Displaying a categorical variable on the x-axis (team) with a numeric variable on the y-axis (rbi)



Screenshot by author from RStudio (RStudio, 2022).

Figure 12: The code used to generate the plot, including the lines of code for the labels

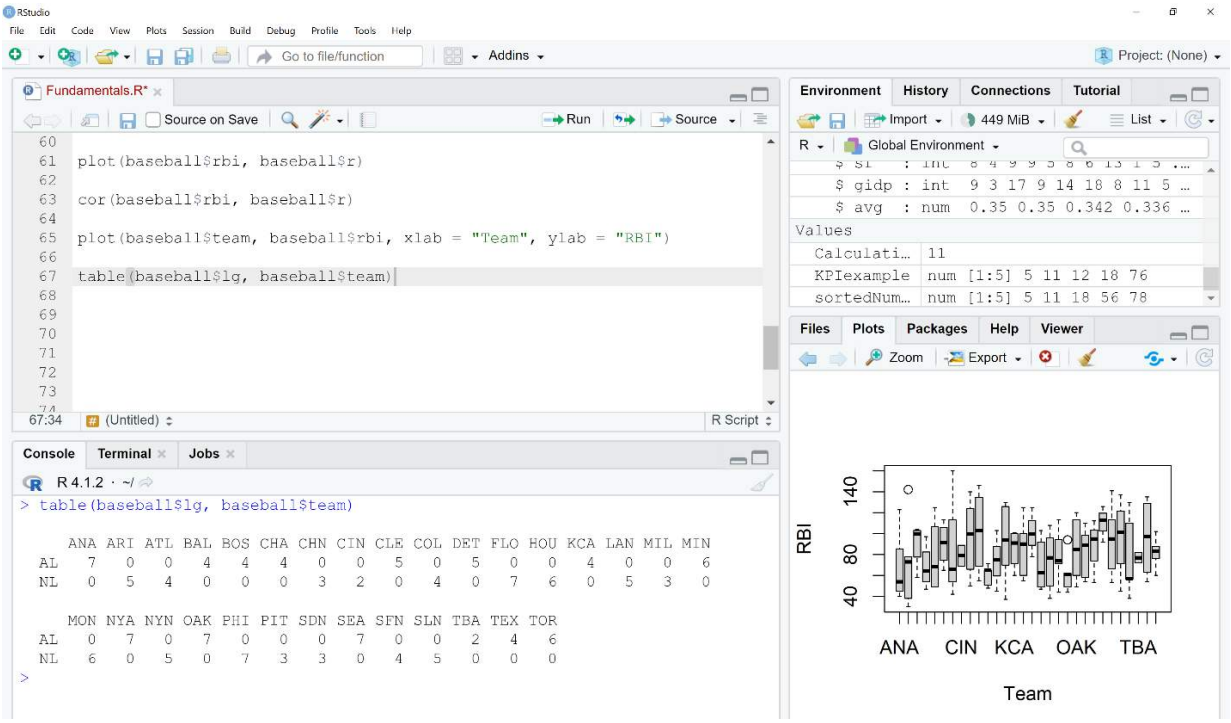


Source: Screenshot by author from Rstudio (Rstudio, 2022).

How about if both variables are categorical?

A table can be the optimal display for this type of data, as shown in the figure below, by simply implementing a function called table() in R and RStudio.

Figure 13: Displaying table command and output in console of the contingency table



Source: Screenshot by author from RStudio (RStudio, 2022).

The basics covered up to this point provide a strong fundamental for exploring your sports dataset. Ranging from examining the variability within your sample data set as well as the central tendency, and further evaluation of any association among numeric variables, as well as a brief introduction into visually inspecting different types of data with basic R functions. In the latter modules, you will learn how to create more advanced and visually aesthetically pleasing data visualisations in R using ggplot2.

Statistical Analyses and Modelling in R

There are a multitude of statistical analyses and models that can be conducted to analyse data. Here, we will cover the most commonly used analyses and models conducted in professional sports. It is primordial to collect data that is as accurate as possible, and then to identify what is the goal of the analyses and modelling. For instance, is it to compare a certain player's athleticism to a benchmark? Is the question rather to compare a team to another team on a key performance indicator? Is it to predict performance from a key performance indicator? Identify the question,

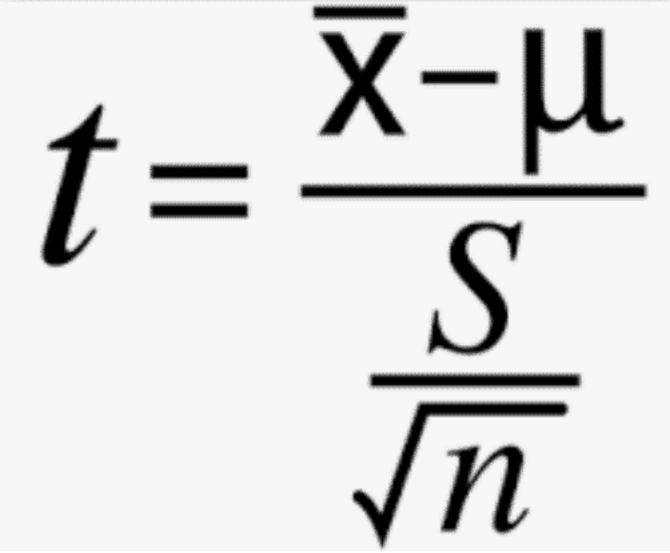
then revert to the data to examine what data you have acquired and if the data types will allow for the question of interest to be answered.

Comparisons to benchmarks

- One sample t-test

A t-test is based on the t distribution, which is similar to the z distribution; however, it has wider tails and less peaked mode compared to the z distribution, when the population's standard deviation, sigma, is not known.

Figure 14: t-test



Source: Graph Pad, (n.d.), https://bit.ly/3ThGQ7H

- One sample z-test

The one sample z-test is used to compare an athlete's score to compare to a benchmark using the z distribution when the population standard deviation, sigma, is given or known.

Figure 15: z-test

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Source: Statistics Lectures, (n.d.), https://bit.ly/3CSt1rc

Statistical modelling can be implemented to examine group differences, whether those differences of interest are between teams, divisions, or player positions. When group differences is the topic of interest, it is recommended to use t-tests, analyses of variance, or clustering, which will be discussed in the latter modules.

If there are only two groups that need to be compared, it is strongly encouraged to conduct an independent t-test. The most important criteria to be met to be able to execute an independent t-test accurately is to be cognizant that there must be one variable that is categorical with two levels and another variable that is continuous. Typically, the categorical variable is identified as the independent variable and the continuous variable is recognized as the dependent variable.

In addition to the criteria of the format of the data types, there are a few assumptions that need to be met to satisfy the requirements to conduct a t-test:

- Assumption of normality of the dependent variable – whereby the (theoretically the residuals) dependent variable is assumed to be approximately normally distributed within each of the groups.
    - This can be tested by conducting a Shapiro-Wilks test of normality or with visual inspection of a Q-Q Plot.
    - If the data shows that the assumption of normality is violated, then it is strongly recommended that the non-parametric version, the Mann Whitney U test, be implemented, as it does not require the assumption of normality to be met; otherwise there is another alternative which is to transform the data; however, transformed data is typically much more difficult to interpret and to then communicate the findings to key stakeholders.
    - The Mann Whitney U test is based on distinguishing between the two groups distribution through analyses of medians and mean ranks.

- Assumption of homogeneity of variance – whereby the variances of the two groups are assumed to be equal in both groups.

    - This can be tested by implementing the Levene's test of Equality. We can assume homogeneity of variance if, when the Levene's test of Equality is conducted, the p value surpasses 0.05 (keep in mind that it is typically the opposite of what we want to find in terms of significance when conducting the actual analysis, not the assumption tests).
    - If Levene's test of Equality is violated, then in R and RStudio, you can simply type in the line of code, var.equal = F, which means the variance is not equal in both groups.

- Reporting the result of an independent t-test – the t statistic value, the degrees of freedom (df) and the significance value of the test (p-value) should be included.

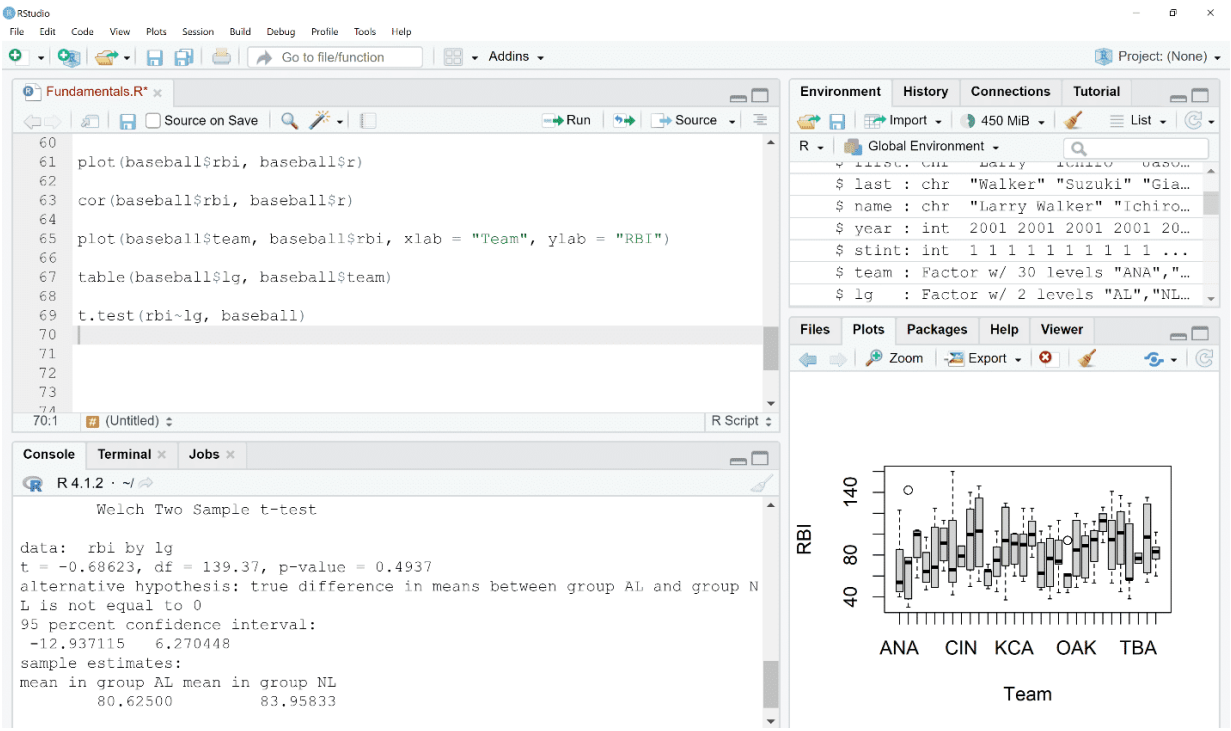
For more information on these criteria, you can visit the following link: https://statistics.laerd.com/statistical-guides/independent-t-test-statistical-guide.php

Now, there are several different scenarios for which mildly different types of t-tests should be conducted; for instance, when you have two different groups, then you should run a t-test with the following code.

Below is the sample structure of the code that should be implemented to assess group differences. The t.test() function is the line of code that will run the t-test, with the first argument within the parentheses being the dependent variable for which, in a t-test, is continuous and numeric, followed by an ~ which means by the following argument, which is the IV that it represents, independent variable which would be a variable indicative of the groups that have the two player positions, teams, or leagues in question embedded within that IV.

The following line of code is the template structure for a t-test; however, keep in mind that, if the variance is not set to equal, then it will not assume equal variances and thus will result in a Welch's Two Sample t-test as displayed in the figure below; where in line 69, the function t.test() is followed by the dependent variable, rbi, followed by the ~, followed by the independent variable that is categorical with two levels, which in this case is lg (league), followed by the argument that is the name of the dataset.

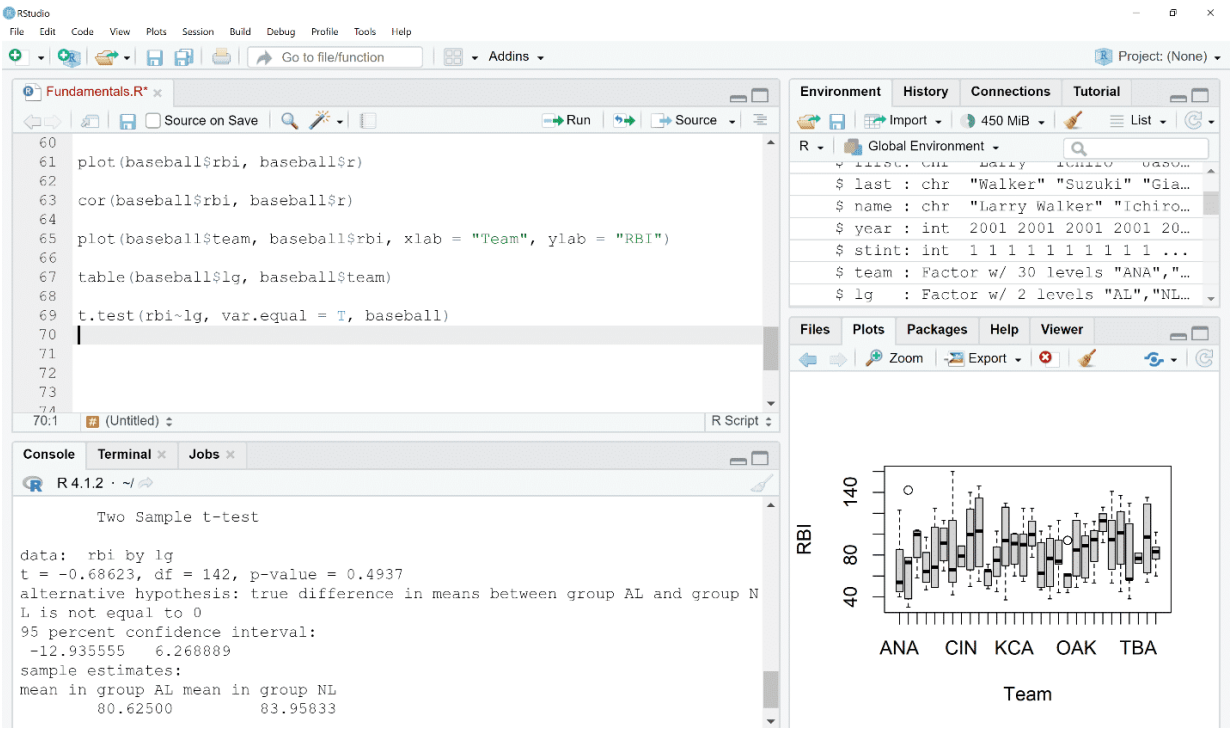
Figure 16: The Welch's Two Sample t-test



Source: Screenshot by author from RStudio (RStudio, 2022).

Now, to execute the official independent t-test, include the code that will state that the assumption of equal variance among both groups is met, as shown below in the figure.

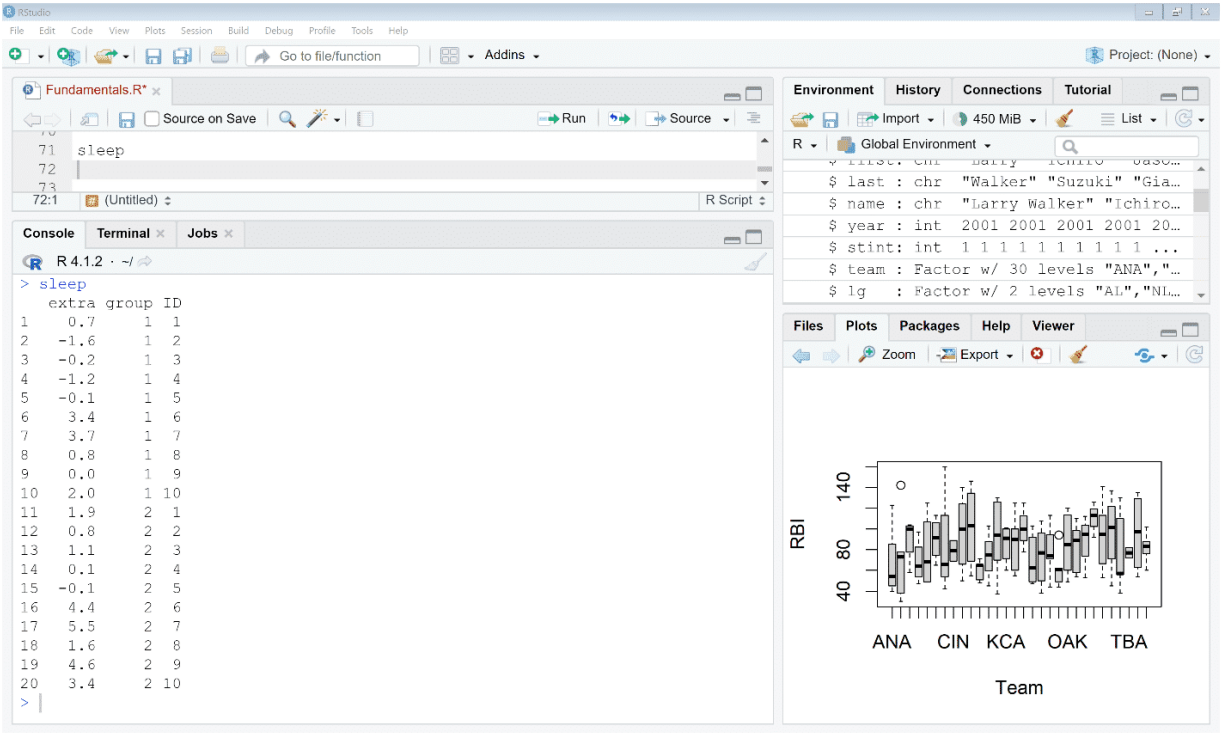
Figure 17: Independent t-test assuming equal variance



Source: Screenshot by author from RStudio (RStudio, 2022).

Now, if you are assessing two groups in terms of how they performed at the start of the season and at the end of the season, or pre to post, then the paired t-test is the appropriate analysis. For this example, we will examine the sleep dataset that comes with R. Let us presume that this dataset includes three variables called: extra, group, and ID, which represent the following: ID of the players that span from 1 to 10; Group which is labelled as 1 or 2, each representing whether it was a preseason value (1) or a post season value (2); and extra being the key performance indicator. As you can tell from the data that is displayed in the console when you type sleep in the R script and then press CTRL + Enter, the IDs repeat themselves twice, this is because this is a dataset that is meant to display that each ID has been tested in both groups or, in this case at two times, pre- and post-season.

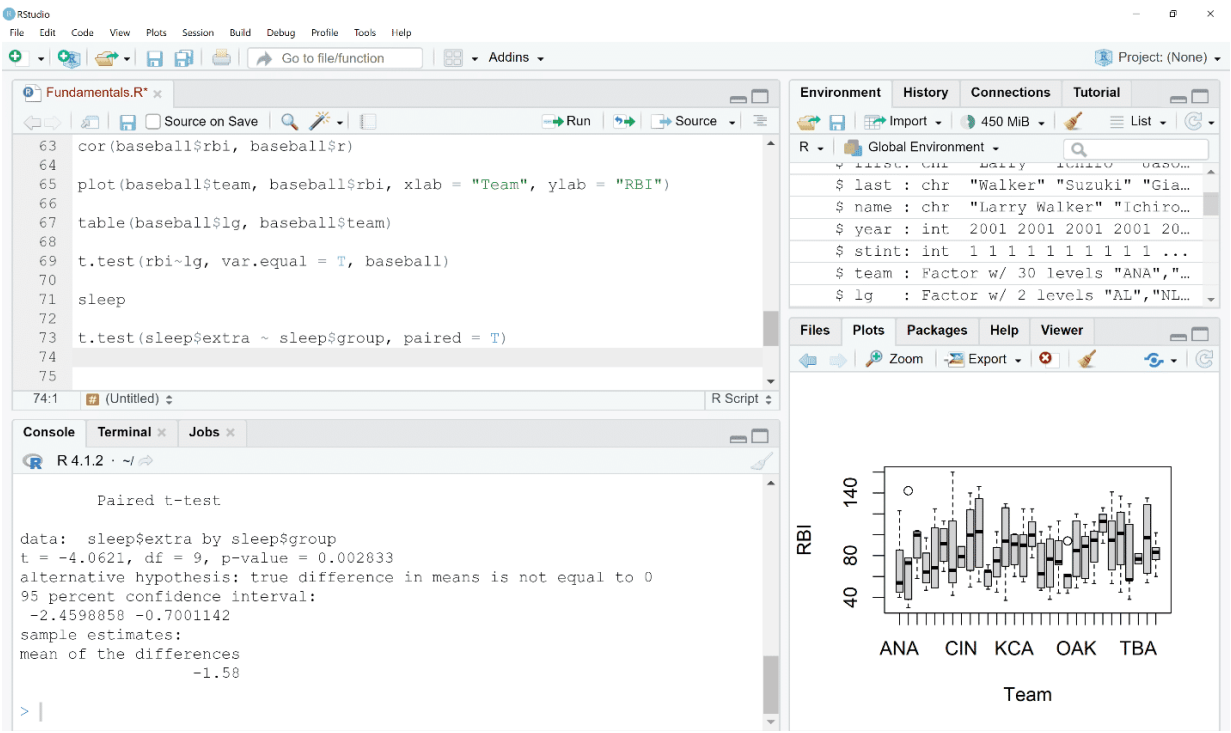
Figure 18: Displaying the data from the sleep dataset



Source: Screenshot by author from RStudio (RStudio, 2022).

In this case, we will implement what is commonly referred to as a matched, paired, or dependent t-test using the following code (keep in mind that the dataset argument can be used on its own, or when you call the variable it can be listed followed by the $ then by the variable name), as displayed in the figure below.

Figure 19: Matched, paired, or dependent t-test



Source: Screenshot by author from RStudio (RStudio, 2022).

As you can see from the three different variations of t-test discussed in this module, it can easily be implemented with a single line of code. Aside from knowing how to run the correct code to implement the statistical test you want to run, it is fundamental that you understand why you are choosing any of these as your choice. The most important item is to ensure that you are choosing the most appropriate analysis and are interpreting and communicating the findings with care to the interested stakeholders.

What about when you would like to compare more than two groups?

When there is a scenario, for instance, where the coach intends to compare the number of assists by forwards, midfielders, and defenders, then the Analysis of Variance (ANOVA) is the statistical model of choice.

However, before conducting this type of statistical analysis, be cognizant that, like the t-test, there are assumptions that need to be met to perform an analysis of variance.

Assumptions for conducting an ANOVA include the following:

- Assumption that the dependent variable is continuous/numeric and of either interval or ratio level.
- Assumption that the independent variable is categorical and has three or more levels.

- Assumption of independence of observations, which means that there is no relationship between the observations in each group or between the groups themselves.
- Assumption that there are no significant outliers.
- Assumption of normal distribution, whereby the dependent variable should be approximately normally distributed for each category of the independent variable.
- Assumption of homogeneity of variances, which can be tested with Levene's test for homogeneity of variances.

For more information, visit the following link:

Laerd Statistics (s. f.). One-way ANOVA in SPSS Statistics. https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php

Now, an important thing to keep in mind is that the ANOVA will only yield whether there are significant differences between groups, but not exactly where the significant difference lies. To identify where the significant difference lies, we must follow up the ANOVA with a post hoc analysis, for which there are many different types, some more robust than others.

To reiterate, to execute an ANOVA, the independent variable has to be of categorical/factor type and there should be more than two levels (otherwise a t-test would have sufficed) and lastly, the dependent variable should be of continuous numeric type.

CONTINUE

References

Allaire, J. J. (2022). R 4.2.1 [Computer Software]. RStudio, Inc. https://cran.r-project.org/index.html

Geeks for Geeks. (2021, May 29). Python – Central Limit Theorem. https://www.geeksforgeeks.org/python-central-limit-theorem/

Graph Pad. (n.d.). One Sample t-test. https://www.graphpad.com/quickcalcs/oneSampleT1/

Statistics Lectures. (n.d.). Variance and Standard Deviation of a Sample. http://www.statisticslectures.com/topics/variancesample/

Statistics Lectures. (n.d.). One Sample z-Test. http://www.statisticslectures.com/topics/onesamplez/

Yi, M. (n.d.). A Complete Guide to Histograms. *Chartio*. https://chartio.com/learn/charts/histogram-complete-guide/

CONTINUE

Download

PDF

Module 2. Exploratory and Descriptive Analytics in R.pdf

4.6 MB