Módulo 2. Análisis exploratorio y descriptivo en R

Como científico del deporte, se recomienda que conozcas bien los fundamentos de la estadística. Como tal, este módulo se ha desarrollado para proporcionarte un conocimiento básico general de los conceptos estadísticos que debes tener en cuenta al realizar análisis de los datos de rendimiento deportivo. Comenzaremos cubriendo las medidas de tendencia central, muchas de las medidas sustitutas giran en torno a ellas.

Medidas de tendencia central

En matemáticas básicas, aprendimos que el **promedio** se calcula sumando todos los valores y dividiendo entre el total de valores en ese conjunto de números.

El promedio se calcula de la siguiente manera; tienes un conjunto de números, por ejemplo, los siguientes:

5 11 56 18 78

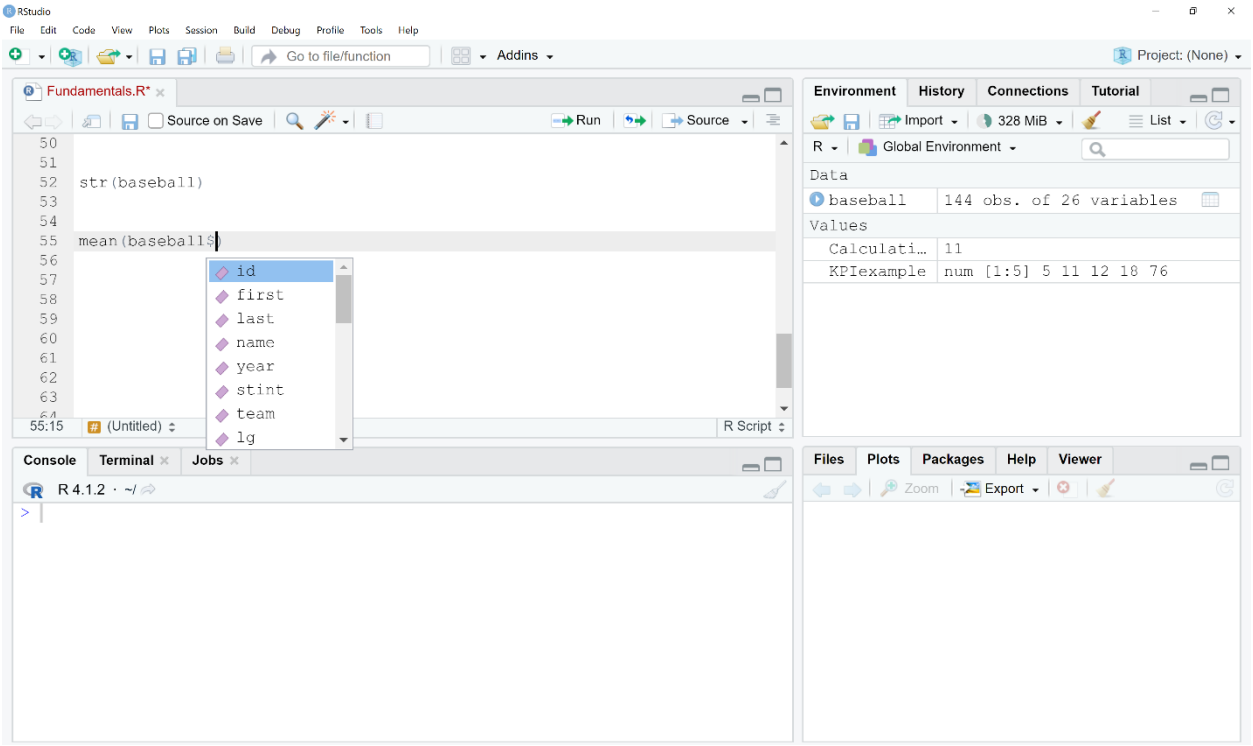
El 5 se suma al 11; luego se suma al 56; luego se suma al 18 y luego se suma al 78, y después el total sumado de 168 se divide por 5, y la respuesta final sería 168 dividido por 5 = 33,6.

Comúnmente, también se le llama la media. En RStudio, podemos calcular la media de cualquier variable implementando la función mean() de la siguiente manera:

- mean()

Verifiquemos esto en R usando el conjunto de datos de béisbol creado en el último módulo y simplemente ejecutando la línea de código mean(). Esto será diferente al módulo anterior cuando aplicamos la función summary() en todo el data frame ya que proporcionaba el promedio para todas las variables junto con otras cinco métricas. Al aplicar una sola medida estadística en un data frame, es necesario que especifiquemos en qué variable. Hacemos esto usando el símbolo $ para acceder. En el momento en que escribas la función mean con béisbol incrustado y luego el símbolo $, se te mostrará un menú desplegable con todas las posibles variables entre las que puedes elegir, como se muestra en la imagen a continuación.

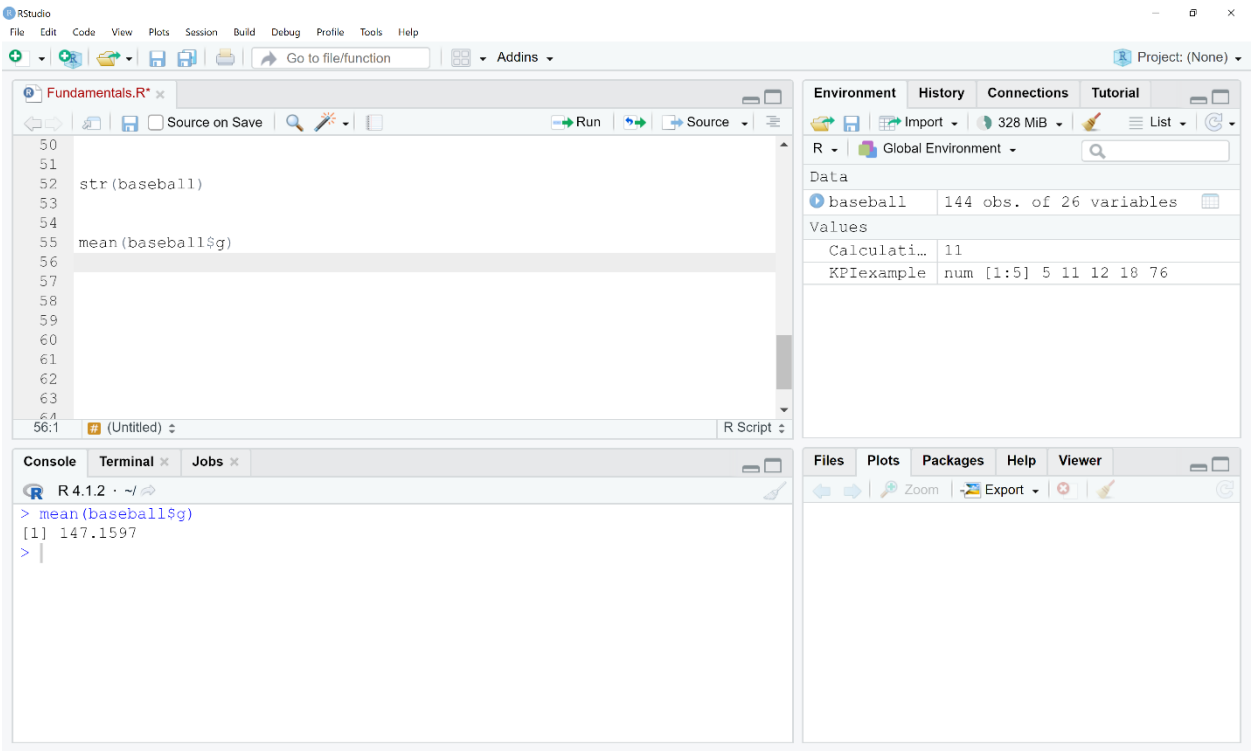
Imagen 1: Menú desplegable con las posibles variables



Fuente: Captura de pantalla de RStudio realizada por la autora (RStudio, 2022).

Elige una variable, que, en este caso, para seguir el ejemplo sería la variable "g" dentro del marco de datos "béisbol", que representa el número de partidos de cada jugador. Luego, puedes ver cómo se puede aplicar la función mean a una variable dentro del data frame como se muestra en la imagen a continuación.

Imagen 2: Función mean aplicada a una variable



Fuente: Captura de pantalla de RStudio realizada por la autora (RStudio, 2022).

La siguiente medida de tendencia central es la **moda**, que es el número que más veces aparece en un conjunto de números, por ejemplo:

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

La moda sería: 11.

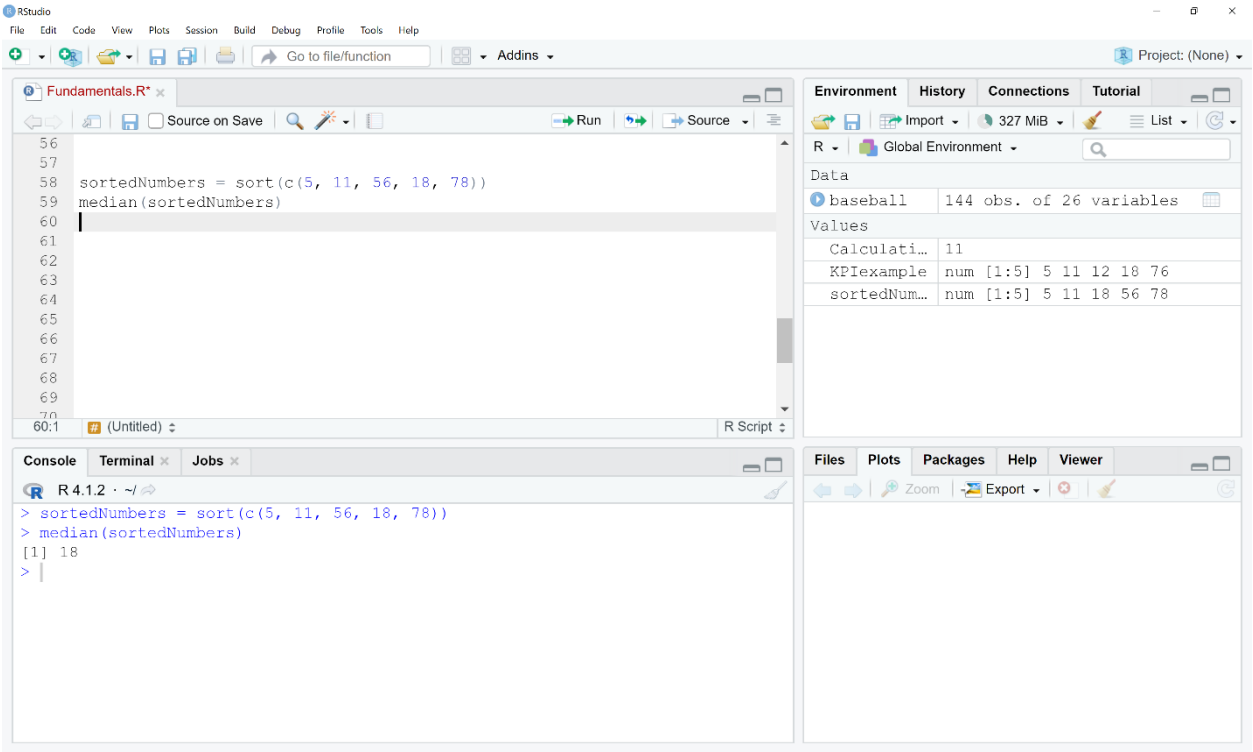
Finalmente, una medida de tendencia central muy útil es la **mediana**, que técnicamente es el percentil 50. La mediana, generalmente, se utiliza en lugar del promedio cuando no se desea tener tanta influencia de valores atípicos, lo que significa que es menos sensible a estos valores

Por lo tanto, al trabajar con valores atípicos, asumiendo que no son errores de entrada de datos, es posible que deseemos incluirlos y, por lo tanto, trabajar con la media; sin embargo, si creemos que el valor atípico no representa lo que está sucediendo en el campo, entonces la mediana sería la medida de tendencia central más apropiada.

La mediana se calcula primero ordenando los valores y luego utilizando la siguiente función de código en R: median(). Primero crearemos un objeto llamado "sortedNumbers" que consiste en los siguientes valores numéricos: 56, 78, 18, 5, 11. Después de hacerlo, debes ordenar los números como se muestra en la imagen a continuación y luego implementar la función median().

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

Imagen 3*:* Ejemplo de código en R para ordenar, crear un objeto nuevo y calcular la mediana

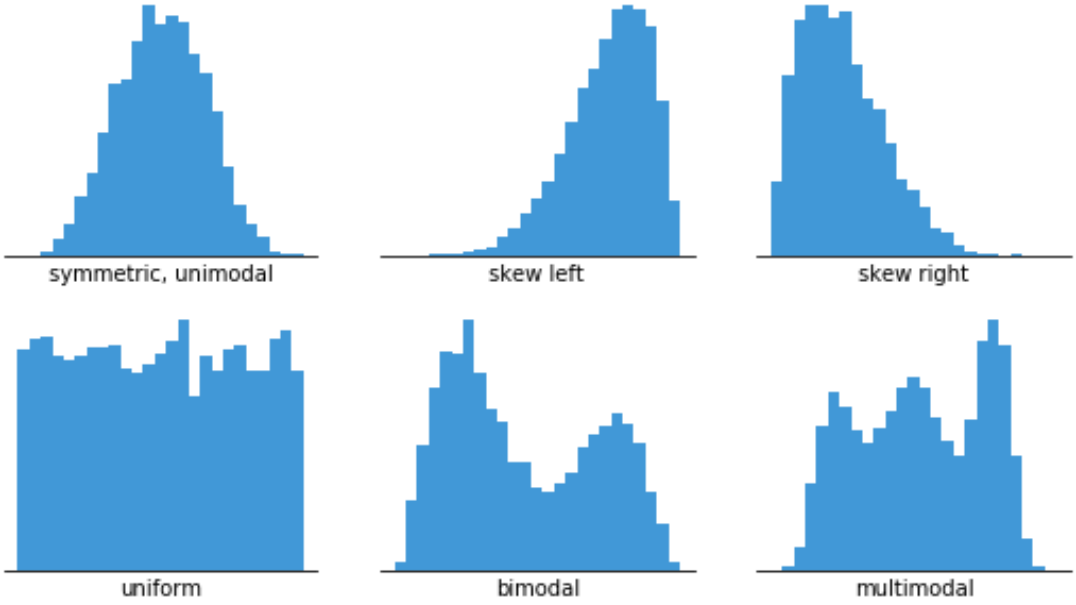


Fuente: Captura de pantalla de RStudio realizada por la autora (RStudio, 2022).

- Además, tengamos en cuenta que, como estándar general de práctica para la codificación, al conectar varias palabras en el nombre de un objeto, es costumbre cambiar la letra ya sea en minúsculas o mayúsculas, como se hace en este ejemplo, o usar un punto o guión bajo entre ellas.

Otra forma de examinar estas medidas de tendencia central es observar la distribución de los datos para examinar dónde está el centro y si están sesgados o no, si hay valores atípicos y cuánta dispersión hay en los datos. (Consulten la imagen a continuación, que incluye imágenes de distribuciones unimodales, bimodales, sesgadas y valores atípicos).

Imagen 4: diferentes distribuciones de datos



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

Symmetric, unimodal	Simétrico, unimodal
Skew left	Sesgado a la izquierda
Skew right	Sesgado a la derecha
Uniform	Uniforme
Bimodal	Bimodal
Multimodal	Multimodal

Indicadores para diferenciar la tendencia central

- Promedio, moda, mediana
- Forma - sesgo
- Dispersión - valores atípicos

Inspeccionar visualmente los datos es una manera esencial de determinar la forma de los datos. Es importante identificar si los datos están sesgados o no, ya que esto puede tener implicaciones en cómo transmitimos nuestros hallazgos, así como si es posible que queramos recolectar muestras adicionales de datos para cumplir con los criterios del teorema del límite central (TLC). Hablaremos más sobre el TLC un poco más adelante. Pero primero, ¿qué es exactamente la asimetría? ¿Qué significa estar sesgado a la izquierda, sesgado negativamente, sesgado a la derecha y sesgado positivamente?

La asimetría es un término definido por tener la mayor parte de los datos en un extremo y la cola de los datos en el extremo opuesto, anulando así la posibilidad de tener un conjunto de datos normalmente distribuido de manera simétrica. En otras palabras, parece que la mayoría de las puntuaciones del indicador clave de rendimiento (KPI, por sus siglas en inglés) que estamos cuantificando se agrupa en un extremo o en otro, pero no en forma de campana; más

bien, se agrupa en el extremo bajo (sesgado a la derecha y positivo) con pocas puntuaciones en el extremo superior, normalmente indicado por el lado derecho o viceversa.

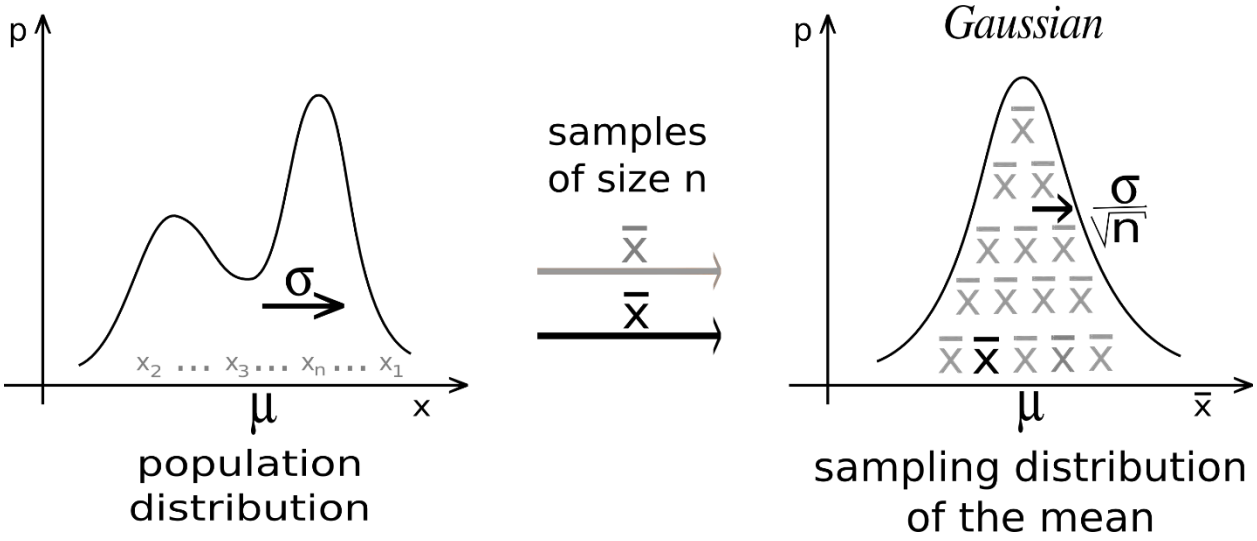
Puntos clave

Cuando la cola es más larga del lado derecho, la distribución está sesgada positivamente, también conocida como sesgada a la derecha. Cuando la cola es más larga del lado izquierdo y la mayor parte de los datos está en el lado derecho, la distribución se considera sesgada negativamente o sesgada a la izquierda.

La razón por la que es importante examinar si los datos están sesgados hacia la derecha o hacia la izquierda es porque determinará qué tipos de análisis se pueden aplicar, y si necesitamos recopilar más datos para que cumplan con los criterios de la distribución normal ¿Cómo es esto con más datos? Sin entrar en la prueba matemática, existe un teorema que establece que cuando tienes suficientes datos, es posible que nos preguntemos ¿cuántos datos son suficientes? En estadística, la regla general es que si tenemos una muestra con un tamaño n de al menos 30, podemos asumir la normalidad. ¿Por qué esto es importante? Bueno, porque cuando asumes la normalidad y una curva en forma de campana, hay análisis estadísticos que se pueden implementar, mientras que, de lo contrario, es posible que debamos conformarnos con análisis no paramétricos (más sobre eso después) que no son tan detallados.

El **Teorema del límite central** establece que la media de la muestra aproximadamente seguirá una distribución normal para tamaños generalmente mayores a un N de 30, independientemente de la distribución de la cual estamos tomando la muestra.

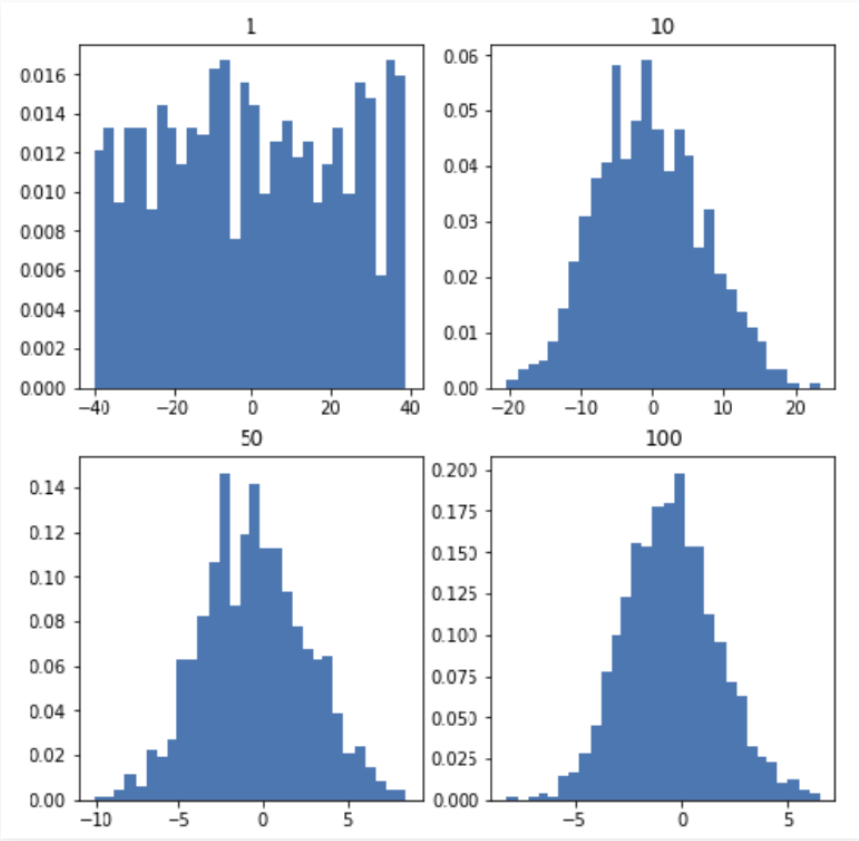
Imagen 5: Teorema del límite central



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

Samples of size n	Muestras de tamaño n
Population distribution	Distribución de la población
Sampling distribution of the mean	Distribución de la muestra de la media
Gaussian	Gaussiana

En la figura a continuación, podemos ver que a medida que el tamaño de la muestra aumenta de 1 a 100, el histograma tiende a tomar la forma de una distribución normal.
Imagen 6: Distribución normal



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

Valores atípicos

Hablemos de los valores atípicos en el conjunto de datos. Nuestro conjunto de datos puede estar distribuido normalmente o no, y aun así tener valores atípicos en cualquiera de los extremos de la curva.

Problema: En el deporte profesional, esto es algo común y el científico del deporte debe decidir si incluir los valores atípicos en el análisis final o no.

Solución: Se recomienda no saltear este paso y ser minuciosos en el enfoque de nuestro análisis. Por lo tanto, realizamos un análisis con los valores atípicos incluidos en los datos, así como sin ellos. Cuando comuniquemos la información al principal interesado en los hallazgos, ya sea el entrenador, el gerente general u otro personal de rendimiento deportivo, debemos prepararnos para proporcionar contexto sobre por qué realizamos los análisis con y sin valores atípicos. La respuesta a una pregunta así podría estar relacionada con si estamos analizando a nuestro jugador estrella (quien generalmente tiene valores atípicos) y comparando sus valores con el resto del equipo u otros jugadores en general, o si estamos interesado en el resultado promedio de la mayoría de los jugadores en el equipo (análisis sin los valores atípicos).

Finalmente, la visualización de datos también nos permitirá inspeccionar los valores atípicos. Si tenemos las unidades sin procesar etiquetadas, podremos determinar si los valores atípicos se consideran inusuales o valores atípicos extremos, según si están a 2 o 3 desviaciones estándar por encima y por debajo de la media. Todo lo que necesitamos para calcular esto es la media y la desviación estándar.

Abordaremos la desviación estándar después de analizar la variación, y luego también analizaremos la puntuación z comúnmente aplicada.

Medida de variación

El concepto más importante en estadística es la variación. Ser capaz de comprender completamente la variabilidad nos permitirá responder preguntas como las siguientes: ¿Dónde se ubican los puntos de datos? ¿Qué tan lejos están entre sí y de la media? ¿Cuál es el valor más bajo y el valor más alto? Todas estas preguntas se responden con medidas de variación que cuantifican cómo se dispersan los datos. ¿Por qué esto es importante? Bueno, pensemos en un ejemplo donde tenemos una media de 3 goles anotados, y se basa en datos donde los equipos anotaron consistentemente entre 2 y 4 goles. Pero luego pensemos en un escenario donde la media de 3 goles se calcula a partir de equipos que anotaron 1 y 5 goles, aun así, la media es 3. El punto es que estos son dos tipos de escenarios muy diferentes y el contexto es clave. Esto ejemplifica por qué debemos examinar la variación y la dispersión de las variables de interés.

¿Cómo se cuantifica exactamente? Hay varias formas de medir la variación y la variabilidad (intercambiables).

Varianza: ¿cómo calculamos la varianza? La varianza se calcula midiendo la distancia entre cada punto de datos y la media de la muestra.

Esto nos lleva a la definición de una **muestra**, ¿qué es una muestra? Es un grupo de individuos seleccionados al azar (observaciones, puntos de datos, etc.) que se eligen con la idea de inferir y generalizar hallazgos sobre la **población**. Términos clave relacionados con la población y la muestra son parámetros y estadística. Estos términos del glosario se utilizan para describir las principales características de su respectivo dominio; por ejemplo, las **estadísticas**, como la media, se representan como xbar y las desviaciones estándar de la muestra son características de los datos de muestra, mientras que los **parámetros**, como la media de la población pronunciada mu y sigma, son las métricas que caracterizan a la población. También es importante tener en cuenta que, al referirse a la población, se utilizan letras griegas, mientras que al referirse a las estadísticas de la muestra, se implementan letras romanas.

Parámetros de la Población

μ = media de la población

σ = desviación estándar de la población

π = proporción de la población

Estadística de la muestra

xbar = media de la muestra

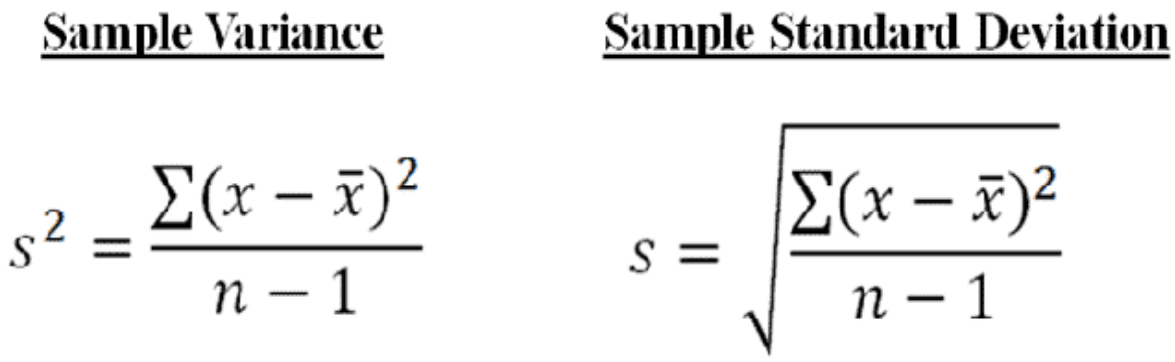
s = desviación estándar de la muestra

p = proporción de la muestra

Aunque la varianza suena ideal para calcularla, tomando la diferencia entre el punto de datos y la media, elevándolos al cuadrado y luego sumando los números, es la **desviación estándar** en la que más confiamos. Esto se debe a dos razones principales: 1) Se convierte en un número positivo sin importar qué. Originalmente, un punto de datos que tenía un valor de -8 ahora es +64; 2) La razón por la que preferimos la desviación estándar es que la dividimos por el tamaño total de la muestra y luego sacamos la raíz cuadrada, lo que nos devuelve a un área donde podemos interpretar los datos en unidades relevantes.

A continuación, hay una imagen de la ecuación de la varianza y la desviación estándar.

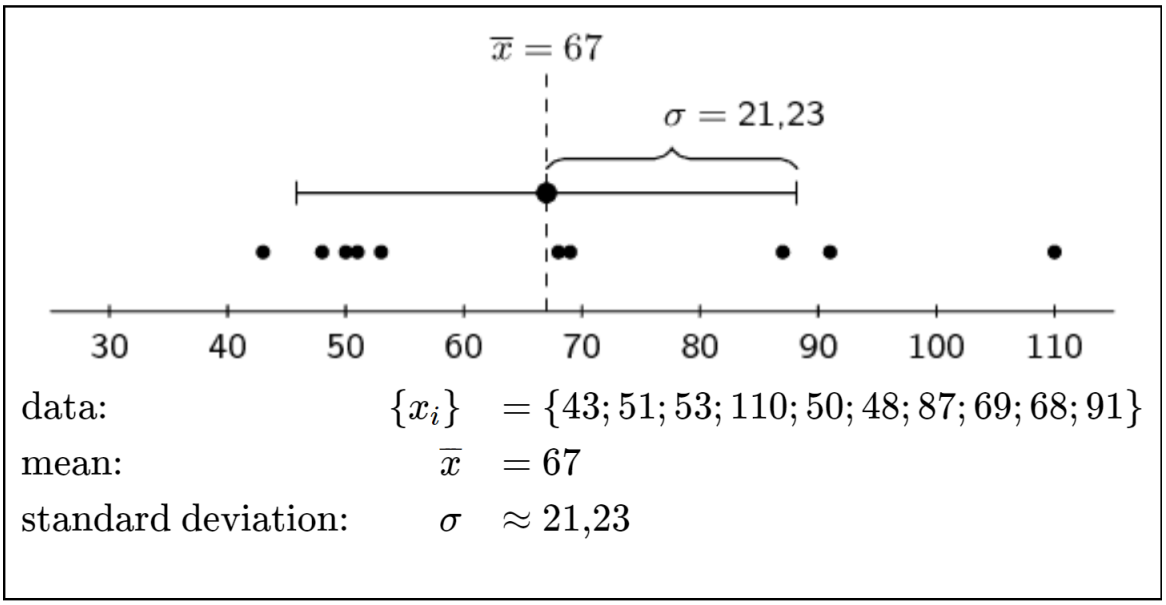
Imagen 7: Ecuación de la varianza y la desviación estándar



Fuente: Statistics Lectures, (s.f.), https://bit.ly/3SOXSue

Sample variance	Varianza de la muestra
Sample standard deviation	Desviación estándar de la muestra

Imagen 8: Los datos, media, varianza y desviación estándar



Fuente: Statistics Lectures, (s.f.), https://bit.ly/3SOXSue

Data	Datos
Mean	Media
Standard deviation	Desviación estándar

La **desviación estándar** de la población se llama sigma, y se utiliza un símbolo griego σ (pronunciado sigma) para representarlo. La desviación estándar de la muestra comúnmente se denomina *s*, o se utiliza la letra romana s para referirse a la desviación estándar de la muestra, y en R y Rstudio el código es sd().

El **rango** es otra opción para medir la dispersión. Sin embargo, la ventaja es que solo necesita y utiliza dos valores para calcular el rango, el valor mínimo y el valor máximo, pero la desventaja es que, así como con la desviación estándar, solo se utilizan dos valores para determinar la dispersión. El cálculo del rango es el siguiente: obtenemos el valor máximo y le restamos el valor mínimo = rango.

Distribución normal y distribución normal estándar

La **distribución normal** representa datos que son simétricos y se reflejan en cada lado con las unidades de medida sin procesar.

La distribución normal estándar se ve similar a la distribución normal; sin embargo, difiere en que los valores sin procesar ahora se identifican por estar 1, 2 y 3 desviaciones estándar por encima o por debajo de la media utilizando la puntuación z.

- ***Puntuación z***

Recordemos al principio que estábamos examinando datos para identificar si había valores atípicos al determinar si había 2 o 3 desviaciones estándar por encima o por debajo de la media.

Si se implementaran unidades sin procesar, asumiríamos que los datos se basaban en un conjunto de datos de muestra distribuido normalmente. Sin embargo, si estos se convierten a una media de 0 y una desviación estándar de 1, entonces estamos trabajando con la distribución normal estándar. Todo esto significa que la media ahora, en otras palabras, se ha desplazado a una media de 0 y una desviación estándar de 1.

¿Cómo calcularíamos una puntuación z? Obtenemos el punto de datos del deportista o jugador de interés en la variable elegida y restamos la media de la muestra, luego dividimos por la desviación estándar de la muestra.

En R y RStudio, tendríamos que escribir el código manualmente de la siguiente manera:

PuntuacionZ = x - xbar / s

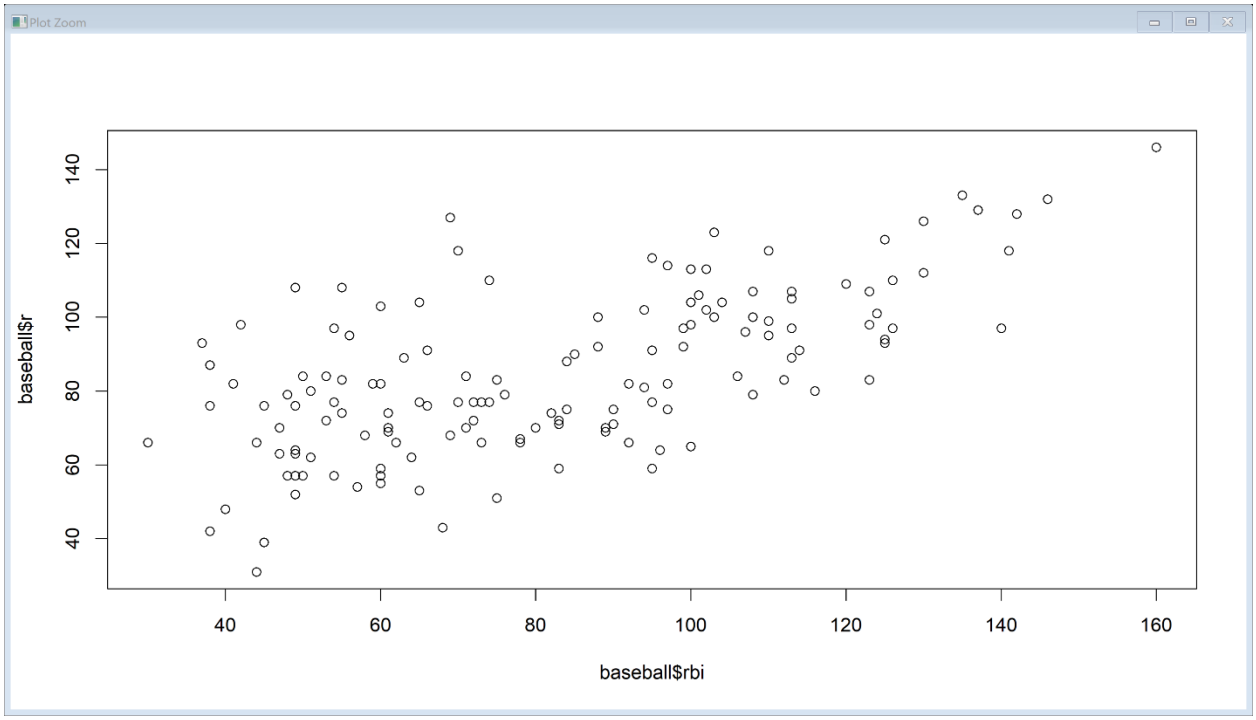
Y si se prefiere, podríamos crear una función llamada "puntuación z" que ejecutaría automáticamente la función anterior cuando llamamos al código "puntuacionZ()".

Un término importante para conocer es la **correlación**, que es un tipo de análisis básico muy informativo.
Sin embargo, es esencial saber que se deben cumplir dos supuestos para realizar un análisis de correlación verdadero y preciso:

- Las dos variables deben ser numéricas.
- Al trazar las variables x e y, debe haber una relación lineal visible.

Si alguno de estos supuestos no se cumple, entonces realizar un análisis de correlación aún producirá un coeficiente de correlación resultante, pero sería inexacto extrapolar a partir de los hallazgos. Si se cumplieron los supuestos y se realizó el análisis de correlación correctamente, eso determinará la fuerza y dirección de la relación entre las variables. Por ejemplo, con el conjunto de datos de béisbol, podemos inspeccionar fácilmente las dos variables numéricas r (carreras) y rbi (carreras impulsadas) simplemente ejecutando un comando de trazado. Esto nos permite verificar si existe linealidad, lo que nos permitiría decidir si ejecutar la correlación de manera oficial o no. En este caso, podemos ver que parece haber cierta linealidad, por lo que podemos proceder a examinar el coeficiente de correlación real.

Imagen 9: Gráfico básico de dos variables numéricas, r y rbis

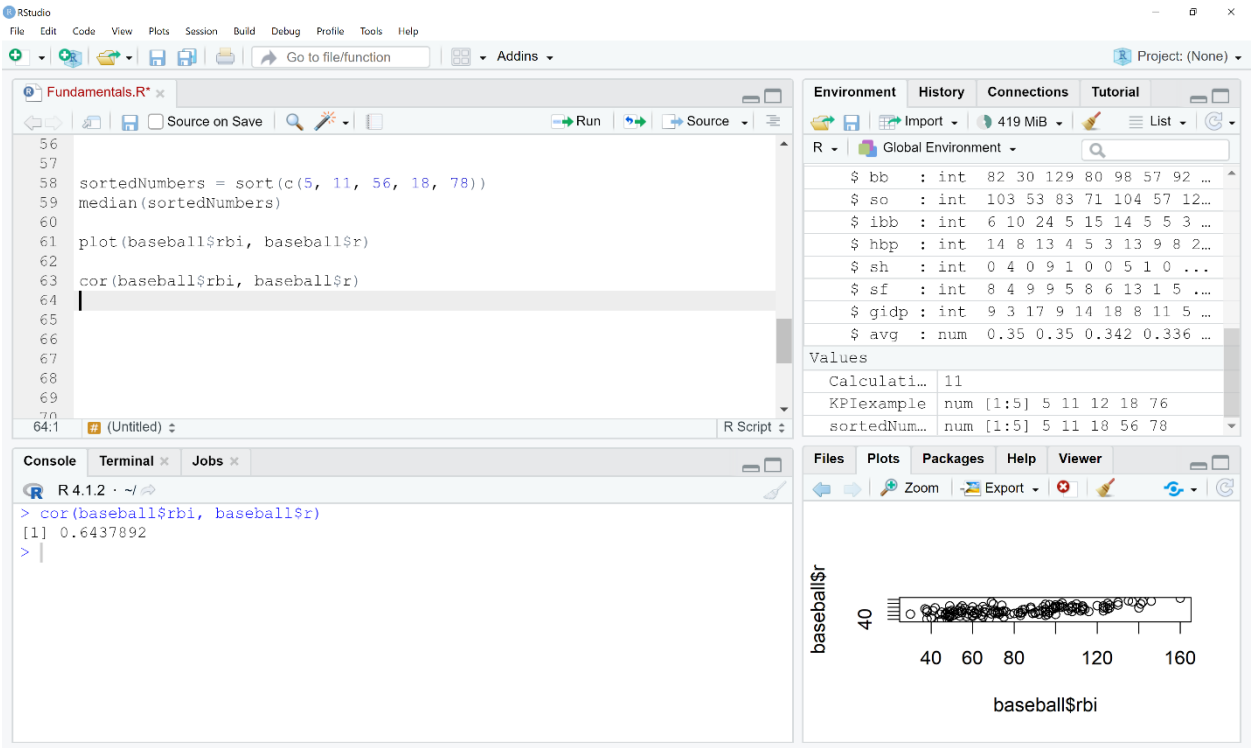


Fuente: Captura de pantalla de RStudio realizada por la autora (RStudio, 2022).

Baseball$rbi	Baseball$rbi

Luego, después de inspeccionar visualmente la linealidad, podemos ejecutar la línea de código para el coeficiente de correlación, como se muestra en la figura a continuación, donde el coeficiente de correlación es 0,64, lo que significa que hay una correlación positiva moderada entre carreras y carreras impulsadas en este conjunto de datos de muestra de béisbol.

Imagen 10: Gráfico y comando 'cor' que produce el coeficiente de correlación



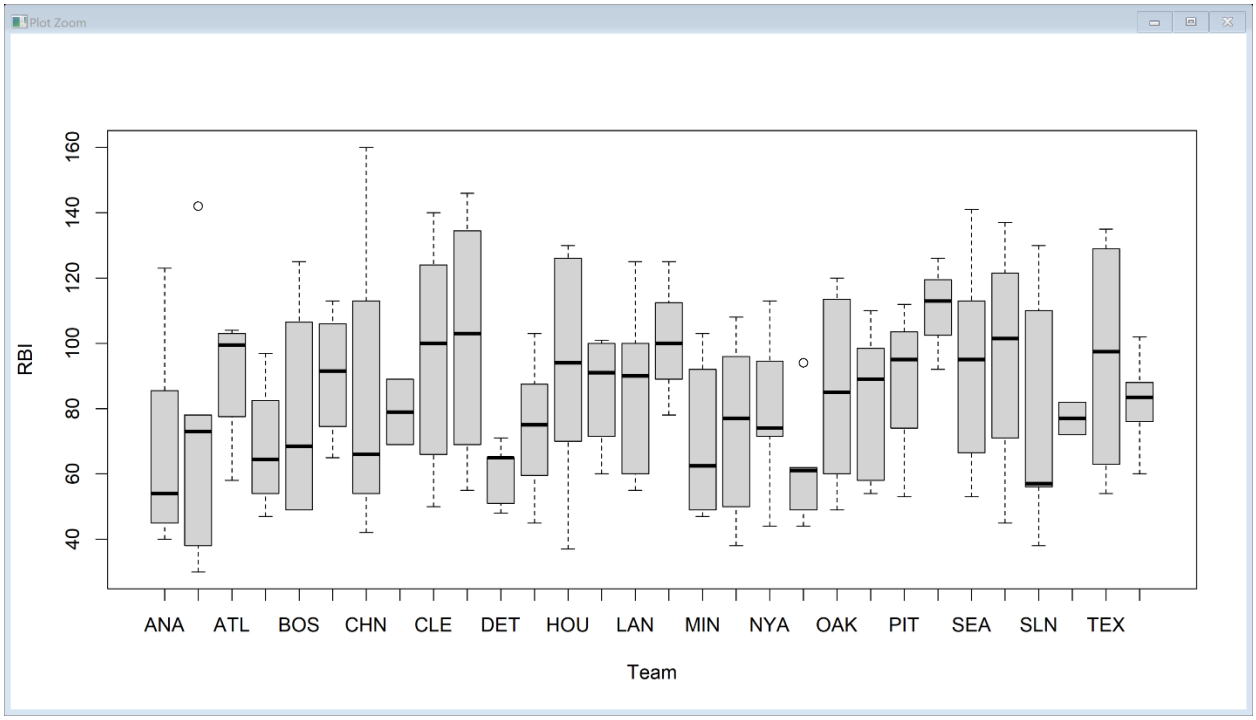
Fuente: Captura de pantalla de RStudio realizada por la autora (RStudio, 2022).

¿Qué hacer si las variables no son numéricas?

Si los datos con los que estás trabajando son categóricos, entonces un gráfico de barras o columnas sería una mejor elección para una representación visual. En los libros de estadísticas comúnmente utilizados, la principal diferencia entre un gráfico de columnas y uno de barras es solo la orientación de las barras, ya sea que estén colocadas vertical u horizontalmente.

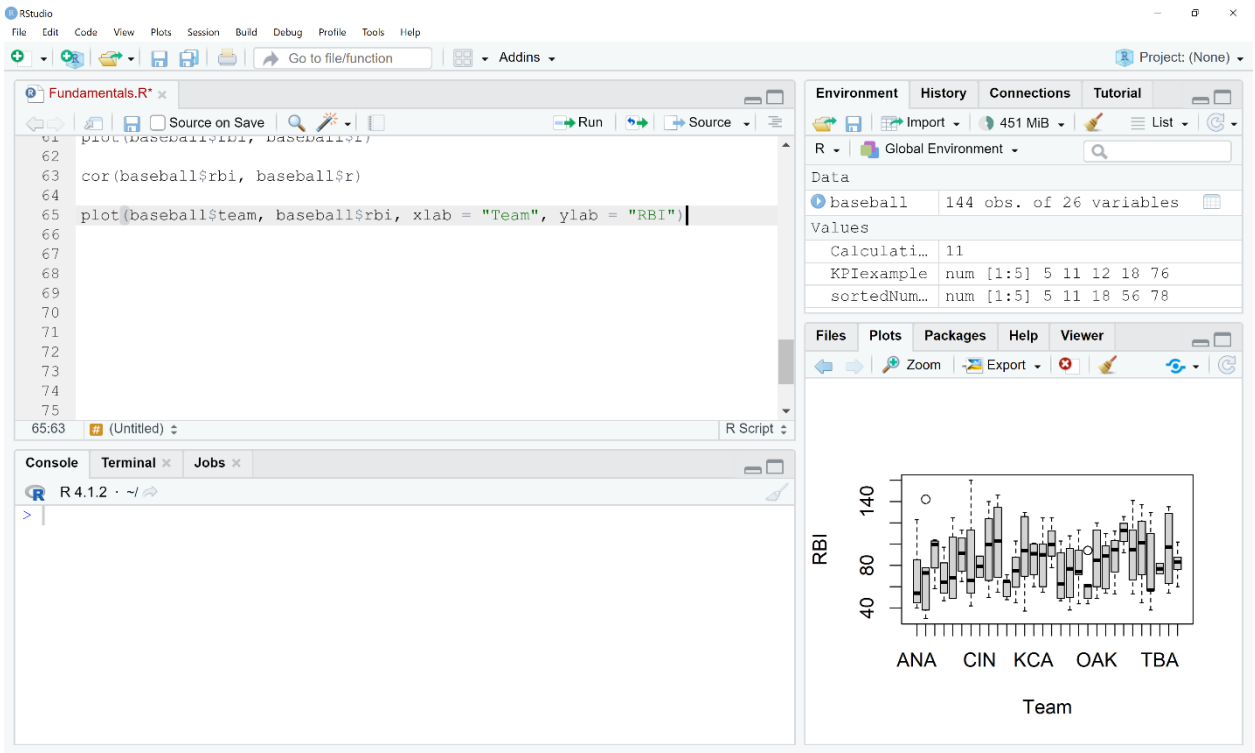
En R y RStudio, en el módulo 4, cubriremos tanto los gráficos de barras como los de columnas y cómo la diferencia principal es la orientación de las barras, ya sea vertical u horizontal. Cuando se utiliza el comando geom_bar() para un gráfico de barras, se visualizará la distribución de una variable en el eje x de manera predeterminada, y el eje y representará la frecuencia o el conteo. Por otro lado, al utilizar el comando geom_col() para un gráfico de columnas, la diferencia está en que el eje y representará la variable y, en lugar del conteo o frecuencia (más detalles en el módulo 4 sobre visualización de datos en R y RStudio). Sin embargo, para simplificar, ya que se presentará ggplot en el módulo 4, ejecutaremos el comando básico de trazado que automáticamente identificará cómo colocar las variables en las barras en función de su tipo de datos.

Imagen 11: Mostrar una variable categórica en el eje x (equipo) con una variable numérica en el eje y (rbi)



Fuente:

Imagen 12: Código utilizado para generar el gráfico, incluidas las líneas de código para las etiquetas

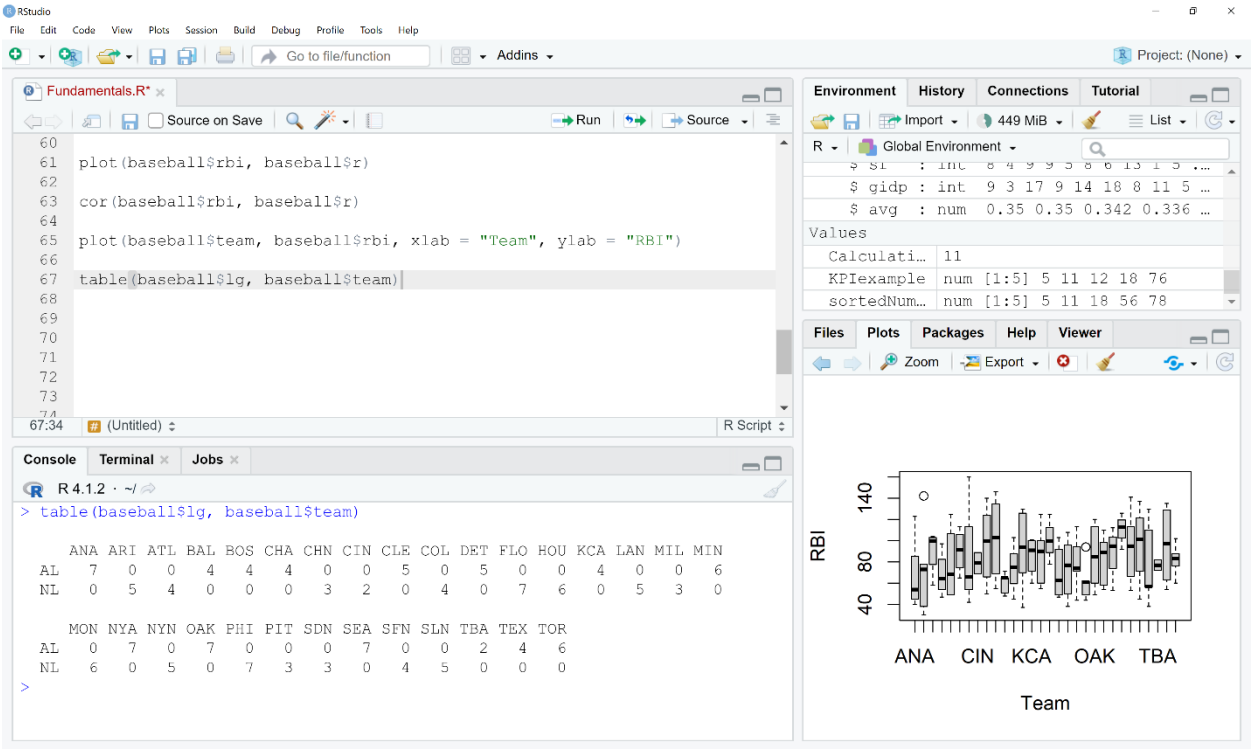


Fuente: Captura de pantalla de RStudio realizada por la autora (RStudio, 2022).

¿Y si ambas variables son categóricas?

Una tabla puede ser la forma óptima de mostrar este tipo de datos, como se muestra en la figura a continuación, simplemente implementando una función llamada table() en R y RStudio.

Imagen 13: Mostrar el comando de la tabla y la salida en la consola de la tabla de contingencia



Fuente: Captura de pantalla de RStudio realizada por la autora (RStudio, 2022).

Los conceptos básicos cubiertos hasta este punto brindan una base sólida para explorar tus conjuntos de datos deportivos. Desde examinar la variabilidad dentro de tu conjunto de datos de muestra hasta la tendencia central y la evaluación de cualquier asociación entre variables numéricas, así como una breve introducción a la inspección visual de diferentes tipos de datos con funciones básicas de R. En los módulos posteriores, aprenderás a crear visualizaciones de datos más avanzadas y estéticamente agradables en R utilizando ggplot2.

Análisis y modelado estadístico en R

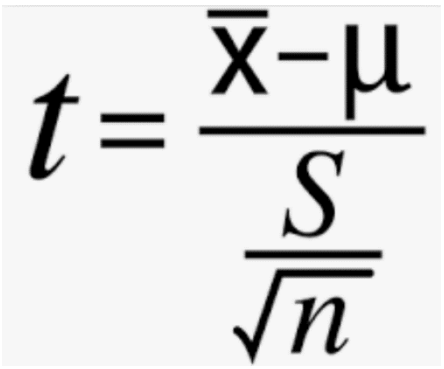
Existen una multitud de análisis estadísticos y modelos que se pueden realizar. Aquí, cubriremos los análisis y modelos más comúnmente utilizados en el deporte profesional. Es primordial recopilar datos lo más precisos posible y luego identificar cuál es el objetivo de los análisis y el modelado. Por ejemplo, ¿se trata de comparar la destreza atlética de un jugador en particular con un punto de referencia? ¿La pregunta se hace para comparar a un equipo con otro en un indicador clave de rendimiento? ¿Es para predecir el rendimiento a partir de un indicador clave de rendimiento? Identifica la pregunta y luego consulta los datos para examinar cuáles tenemos y si te permitirán responder a la pregunta de interés.

Comparaciones con puntos de referencia

- Prueba t de una muestra

Una prueba t se basa en la distribución t, que es similar a la distribución z; sin embargo, tiene colas más anchas y una moda menos pronunciada en comparación con la distribución z, cuando se desconoce la desviación estándar de la población, sigma.

Imagen 14: prueba t

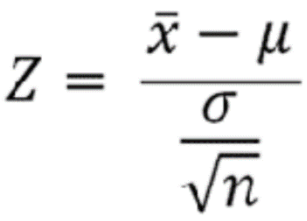


Fuente: Graph Pad, (n.d.), https://bit.ly/3ThGQ7H

- Prueba z de una muestra

La prueba z de una muestra se utiliza para comparar el puntaje de un deportista con un punto de referencia utilizando la distribución z cuando se desconoce la desviación estándar de la población, sigma.

Imagen 15: prueba z



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

El modelado estadístico se puede implementar para examinar las diferencias entre grupos, ya sean entre equipos, divisiones o posiciones de jugadores. Cuando las diferencias entre grupos son el tema de interés, se recomienda utilizar pruebas t, análisis de varianza o clustering, que se discutirán en los módulos posteriores.

Si solo hay dos grupos que deben compararse, se recomienda realizar una prueba t independiente. El criterio más importante que debe cumplirse para poder ejecutar esta prueba

con precisión es ser consciente de que debe haber una variable que sea categórica con dos niveles y otra variable que sea continua. Por lo general, la variable categórica se identifica como la variable independiente y la variable continua se conoce como la variable dependiente.

Además de los criterios del formato de los tipos de datos, hay algunas suposiciones que deben cumplirse para satisfacer los requisitos para llevar a cabo una prueba t:

- Suposición de normalidad de la variable dependiente: se asume que la variable dependiente (teóricamente los residuos) está aproximadamente distribuida de manera aproximadamente normal dentro de cada uno de los grupos.
  - Esto se puede probar mediante la realización de una prueba de normalidad de Shapiro-Wilk o mediante la inspección visual de un gráfico Q-Q.
  - Si los datos muestran que se viola la suposición de normalidad, se recomienda implementar la versión no paramétrica, la prueba de Mann-Whitney U, ya que no requiere que se cumpla dicha suposición, de lo contrario, hay otra alternativa que es transformar los datos; sin embargo, los datos transformados suelen ser mucho más difíciles de interpretar y luego comunicar los hallazgos a las partes interesadas clave.
  - La prueba de Mann-Whitney U se basa en distinguir entre la distribución de los dos grupos mediante análisis de medianas y rangos medios.

- Suposición de homogeneidad de varianza: se asume que las varianzas de los dos grupos son iguales en ambos grupos.
  - Esto se puede probar implementando la prueba de igualdad de Levene. Podemos asumir la homogeneidad de varianza si, cuando se realiza la pueba de igualdad de Levene, el valor p supera 0,05 (hay que tener en cuenta que típicamente es lo contrario de lo que queremos encontrar en términos de relevancia cuando se realiza el análisis real, no las pruebas de suposición).
  - Si se viola la prueba de igualdad de Levene, entonces en R y RStudio, simplemente podemos escribir la línea de código var.equal = F, lo que significa que la varianza no es igual en ambos grupos.

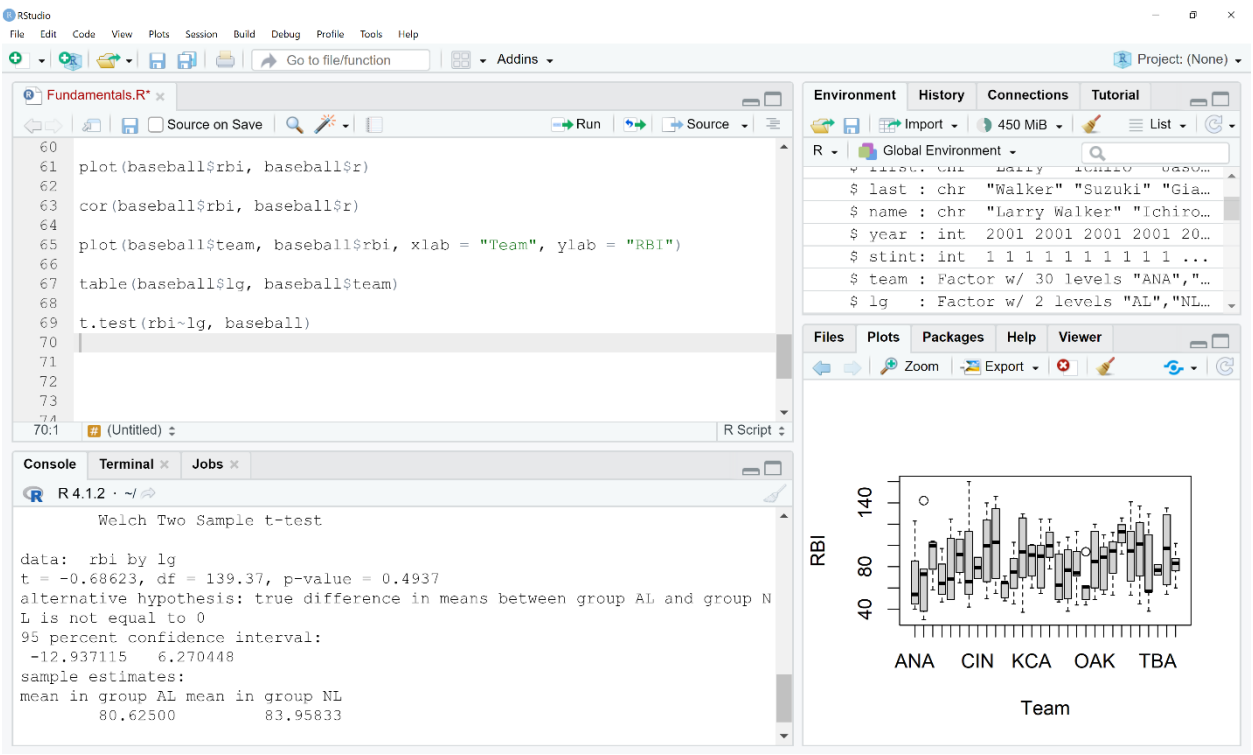
- Informe del resultado de una prueba t independiente: se debe incluir el valor estadístico t, los grados de libertad (df) y el valor de significación de la prueba (valor p).

Para obtener más información sobre estos criterios, pueden visitar el siguiente enlace: https://statistics.laerd.com/statistical-guides/independent-t-test-statistical-guide.php
Hay varios escenarios diferentes para los cuales se deben realizar tipos de pruebas t ligeramente distintas por ejemplo, cuando tienes dos grupos diferentes, debes ejecutar una prueba t con el siguiente código.

A continuación, encontrarán la estructura de muestra del código que debe implementarse para evaluar las diferencias entre grupos. La función t.test() es la línea de código que ejecutará la prueba t, siendo el primer argumento dentro de los paréntesis la variable dependiente que, en una prueba t, es continua y numérica, seguido de un ~ que significa que por el siguiente argumento, que es la variable independiente (IV) que representa, se entiende la variable que indica los grupos que tienen las dos posiciones de jugador, equipos o ligas en cuestión incrustadas en esa IV.

La siguiente línea de código es la estructura de plantilla para una prueba t; sin embargo, hay que tener en cuenta que, si la varianza no se establece como igual, no asumirá varianzas iguales y, por lo tanto, resultará en una prueba t de dos muestras de Welch como se muestra en la figura a continuación; donde en la línea 69, la función t.test() es seguida por la variable dependiente, rbi, seguida por el ~, seguida por la variable independiente que es categórica con dos niveles, que en este caso es lg (liga), seguida por el argumento que es el nombre del conjunto de datos.

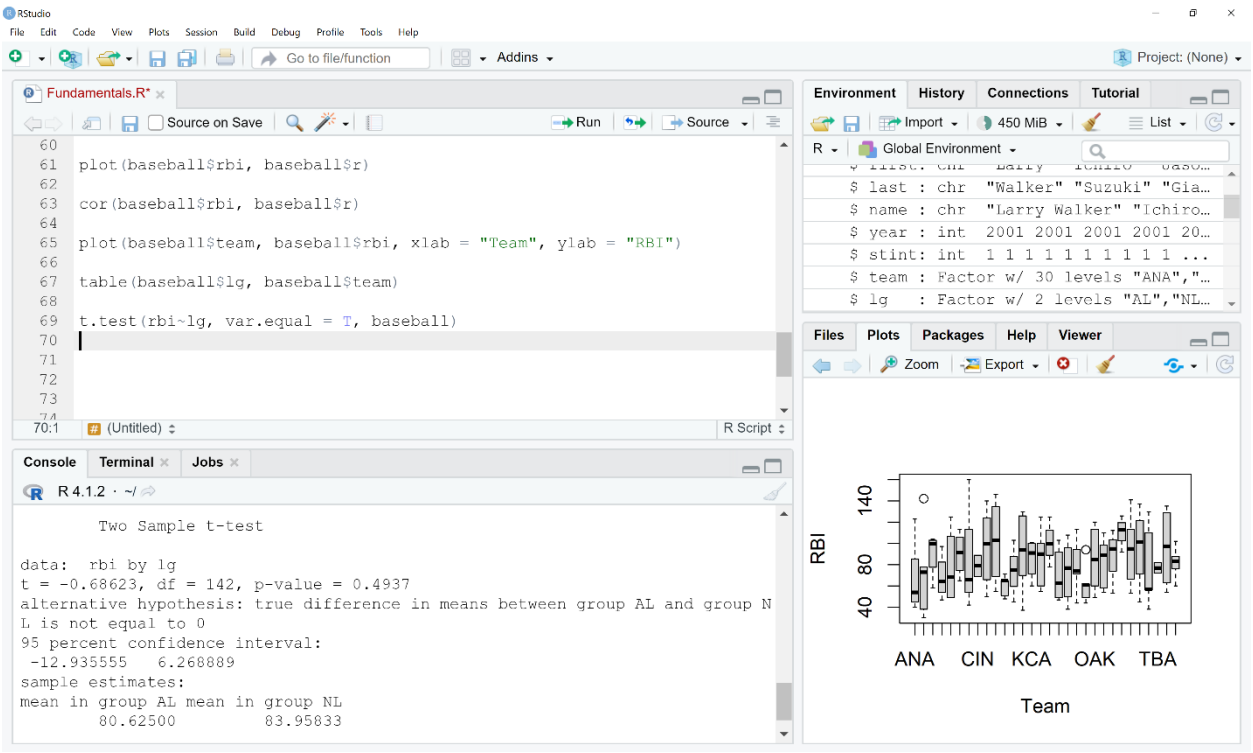
Imagen 16: Prueba t de Welch para dos muestras



Fuente: Captura de pantalla de RStudio realizada por la autora (RStudio, 2022).

Ahora, para ejecutar la prueba t independiente oficial, incluimos el código que indique que se cumple la suposición de igualdad de varianzas entre ambos grupos, como se muestra a continuación en la figura.

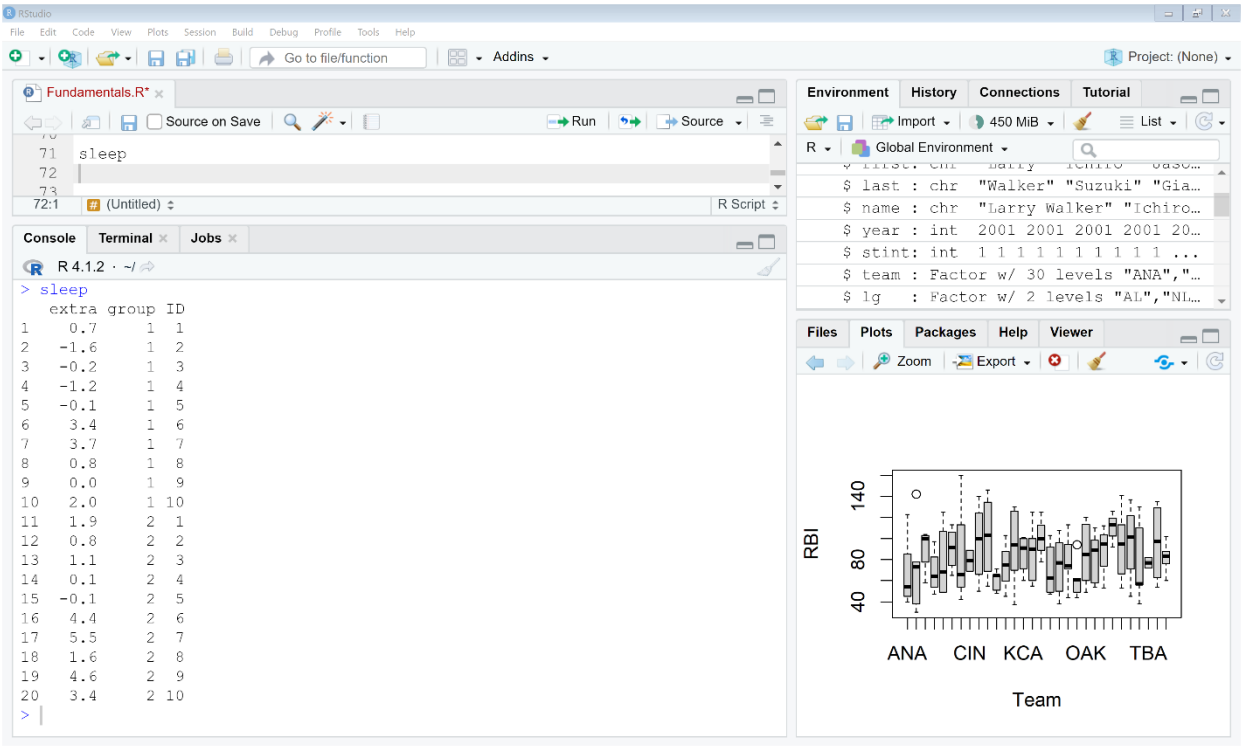
Imagen 17: Prueba t independiente asumiendo igual varianza



Fuente: Captura de pantalla de RStudio realizada por la autora (RStudio, 2022).

Si estamos evaluando dos grupos en términos de cómo les fue al comienzo de la temporada y al final de la temporada, o de antes a después, entonces el análisis apropiado es la prueba t pareada. Para este ejemplo, examinaremos el conjunto de datos "sleep» que viene con R. Supongamos que este conjunto de datos incluye tres variables llamadas: extra, group e ID, que representan lo siguiente: ID de los jugadores que van del 1 al 10; group que está etiquetado como 1 o 2, cada uno representando si fue un valor de pretemporada (1) o un valor de postemporada (2); y extra es el indicador clave de rendimiento. Como se puede ver en los datos que se muestran en la consola cuando escribes "sleep" en el script de R y luego presionan CTRL + Enter, los ID se repiten dos veces, esto se debe a que este conjunto de datos está destinado a mostrar que cada ID ha sido evaluado en ambos grupos o, en este caso, en dos momentos, antes y después de la temporada.

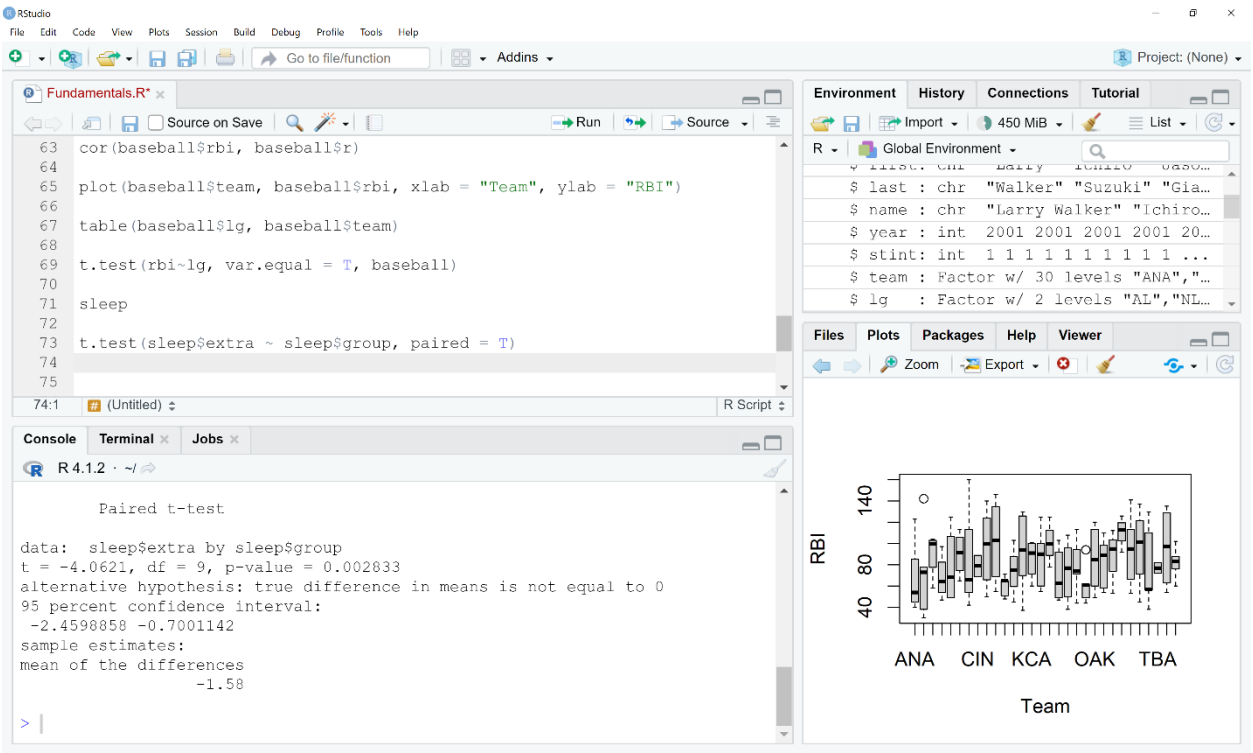
Imagen 18: Mostrar la información del conjunto "sleep"



Fuente: Captura de pantalla de RStudio realizada por la autora (RStudio, 2022).

En este caso, implementaremos lo que comúnmente se conoce como una prueba t emparejada, dependiente o pareada utilizando el siguiente código (hay que tener en cuenta que el argumento del conjunto de datos se puede usar por sí solo o, cuando llamas a la variable, puedes listarla seguida de $ y luego el nombre de la variable), como se muestra en la figura a continuación.

Imagen 19: Prueba t emparejada, dependiente o pareada



Fuente: Captura de pantalla de RStudio realizada por la autora (RStudio, 2022).

Como podemos ver en las tres variaciones diferentes de pruebas t analizadas en este módulo, se pueden implementar fácilmente con una sola línea de código. Además de saber cómo ejecutar el código correcto para implementar la prueba estadística que deseamos realizar, es fundamental que entendamos por qué estamos eligiendo una de estas opciones. El elemento más importante es asegurarse de que estamos seleccionando el análisis más apropiado e interpretando y comunicando los resultados con cuidado a las partes interesadas.

¿Qué ocurre cuando deseamos comparar más de dos grupos?

Cuando hay un escenario en el que, por ejemplo, el entrenador pretende comparar el número de asistencias entre delanteros, mediocampistas y defensores, entonces el análisis de varianza (ANOVA) es el modelo estadístico de elección.

Sin embargo, antes de llevar a cabo este tipo de análisis estadístico, es importante tener en cuenta que, al igual que en la prueba t, hay suposiciones que deben cumplirse para realizar un análisis de varianza.

Las suposiciones para llevar a cabo un ANOVA incluyen lo siguiente:

- Suposición de que la variable dependiente es continua/numérica y de nivel intervalo o de razón.

- Suposición de que la variable independiente es categórica y tiene tres o más niveles.

- Suposición de independencia de las observaciones, lo que significa que no hay relación entre las observaciones en cada grupo ni entre los propios grupos.

- Suposición de que no hay valores atípicos significativos.

- Suposición de distribución normal, en la que la variable dependiente debe estar distribuida de manera aproximadamente normal para cada categoría de la variable independiente.

- Suposición de homogeneidad de varianzas, que se puede probar con la prueba de Levene para la homogeneidad de varianzas.

Para obtener más información, consulten el siguiente link: https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php

Algo importante a tener en cuenta es que el ANOVA solo indicará si hay diferencias significativas entre los grupos, pero no exactamente dónde se encuentra la diferencia. Para identificar dónde se encuentra la diferencia significativa, debemos seguir el ANOVA con un análisis post hoc, para el cual hay muchos tipos diferentes, algunos más robustos que otros.

Recordemos para ejecutar un ANOVA, la variable independiente debe ser de tipo categórico/factor y debe haber más de dos niveles (de lo contrario, una prueba t habría sido suficiente) y por último, la variable dependiente debe ser de tipo numérico continuo.

Referencias

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

Geeks for Geeks. (2021.29.. – Teorema del límite central. https://www.geeksforgeeks.org/python-central-limit-theorem/

Graph Pad. (s.f.). Prueba t de una muestra. https://www.graphpad.com/quickcalcs/oneSampleT1/

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

Statistics Lectures. (s.f.). 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/