Universidad Finis Terrae
Facultad de Ingeniería
2do Semestre 2023

FACULTAD DE INGENIERÍA
UNIVERSIDAD FINIS TERRAE

Laboratorio 3 - Anova

TÓPIC AVANZADO EN DATA SCIENCE

Sección:
76802

Integrantes:
Nicolas Navarro
Carlos Ramirez
Alexandra Vargas

Profesor:
Darío Rojas

Fecha de entrega:
24 de Septiembre del 2023

Santiago, Chile

Tabla de Contenidos

1. Requisitos para ANOVA en R

Para determinar si los supuestos para llevar a cabo un análisis de varianza (ANOVA) se cumplían, se llevaron a cabo las siguientes investigaciones y averiguaciones.

1.1) Normalidad de la variable dependiente.

La normalidad de la variable dependiente es una de las supuestas claves para llevar a cabo un análisis de varianza (ANOVA) paramétrico. La normalidad implica que los datos siguen en una distribución normal, lo que significa que la mayoría de los valores se encuentran cerca de la media y la distribución es simétrica [1]

Para determinar si se cumple el supuesto de normalidad de la variable dependiente en un ANOVA, hay varias formas de realizarlo. Algunas opciones que se pueden utilizar para evaluar la normalidad de los datos:

- *Gráficos de histogramas y gráficos Q-Q:* Puedes crear un histograma de tus datos y verificar si sigue una distribución aproximadamente normal. Además, los gráficos Q-Q (Quantile-Quantile) te permiten comparar la distribución de tus datos con una distribución normal teórica. Si los puntos en el gráfico forman una línea aproximadamente recta, indica una buena aproximación a la normalidad.

- *Prueba de normalidad:* Puedes utilizar pruebas estadísticas como la prueba de normalidad de Kolmogorov-Smirnov o la prueba de Shapiro-Wilk para evaluar la normalidad de tus datos. Estas pruebas comparan tus datos con una distribución normal y generan un valor p, que indica la probabilidad de que tus datos provengan de una distribución normal. Un valor de $p\ alto\ (>\ 0.05)$ indica que los datos se distribuyen aproximadamente normalmente.

- *Prueba de asimetría y curtosis:* La asimetría y la curtosis son medidas que describen la forma de la distribución de los datos. Un valor de asimetría cercano a cero y una curtosis cercana a 3 indican una buena aproximación a la distribución normal.

1.2) Homocedasticidad.

La homocedasticidad significa que las varianzas de las poblaciones subyacentes son aproximadamente iguales para todos los grupos o niveles de la variable independiente. Esto es importante porque ANOVA asume que las muestras provienen de poblaciones con varianzas iguales, y si este supuesto no se cumple, los resultados de ANOVA pueden ser sesgados y poco confiables, el supuesto de homogeneidad de varianzas, también conocido como supuesto de homocedasticidad, considera que la varianza es constante (no varía) en los diferentes niveles de un factor, es decir, entre diferentes grupos.[2]

A la hora de realizar contrastes de hipótesis o intervalos de confianza, cuando los tamaños de cada grupo son muy distintos ocurre que:

- Si los grupos con tamaños muestrales pequeños son los que tienen mayor varianza, la probabilidad real de cometer un error de tipo I en los contrastes de hipótesis será menor de lo que se obtiene al hacer el test. En los intervalos, los límites superior e inferior reales son menores que los que se obtienen. La inferencia será por lo general más conservadora.

- Si por el contrario, son los grupos con tamaños muestrales grandes los que tienen mayor varianza, entonces se tendrá el efecto contrario y las pruebas serán más liberales. Es decir, la probabilidad real de cometer un error de tipo I es mayor que la devuelta por el test y los intervalos de confianza verdaderos serán más amplios que los calculados.

Formas de determinar si se cumple el supuesto de homocedasticidad en un conjunto de datos antes de realizar un ANOVA:

- *Gráfico de dispersión de residuos versus valores ajustados:* Al ejecutar un ANOVA, puede crear un diagrama de dispersión de los residuos (la diferencia entre los valores observados y ajustados) versus los valores ajustados. Si la dispersión de los puntos en el gráfico es aproximadamente constante dentro del rango de valores adecuados, esto indica homogeneidad. Si el rango se amplía o se estrecha a medida que aumentan los valores de aptitud, esto puede indicar una falta de uniformidad.

- *La prueba de Levene:* Es una prueba estadística que compara la varianza de grupos y evalúa si son significativamente diferentes. Los valores de p pequeños en la prueba de Levene (típicamente <0,05) sugieren evidencia de falta de homogeneidad. En este caso, podría considerar utilizar una versión modificada de ANOVA, como el ANOVA de Welch, que ciertamente no es uniforme.

- *La prueba de Bartlett:* Es otra prueba estadística que evalúa si las varianzas de los grupos son iguales. De manera similar a la prueba de Levene, un valor p pequeño en la prueba de Bartlett indica una falta de homogeneidad.

- *Células en forma de caja y bigotes:* Puedes crear un diagrama de caja para cada grupo y comparar la longitud de los bigotes. Si la longitud de los bigotes es la misma en todos los grupos, es un signo de homogeneidad. Si notas una diferencia significativa en la longitud de la barba, esto puede indicar una falta de uniformidad. Prueba de relación de varianza (ratio F): esta prueba compara las diferencias más grandes y más pequeñas entre grupos. Si la relación F es significativamente diferente de 1, puede indicar una falta de uniformidad.

1.3) Independencia de observaciones.

La independencia de las observaciones se refiere a la suposición de que las observaciones dentro de cada grupo o categoría son independientes entre sí, es decir, que la medición o el resultado para un individuo no está influenciado por el resultado de otro individuo [3].

Algunas formas de determinar si se cumple el supuesto de independencia de las observaciones en un diseño de ANOVA son las siguientes:

- *Diseño experimental:* Es importante diseñar el estudio de tal manera que las observaciones sean independientes entre sí. Esto implica asignar al azar los participantes a los diferentes grupos o niveles de la variable independiente y evitar la influencia de factores no controlados. Al utilizar un diseño experimental adecuado, se puede minimizar la posibilidad de que las observaciones estén relacionadas entre sí.

- *Análisis de residuos:* Después de realizar el ANOVA, se puede examinar los residuos para verificar si existe estructura o patrón que indique falta de independencia de las observaciones. Los residuos son las diferencias entre los valores observados y los valores predichos por el modelo. Si hay alguna autocorrelación o patrón en los residuos, puede haber violación del supuesto de independencia.

- *Investigación previa:* Además de los métodos anteriores, es importante revisar la literatura existente en el área de estudio para determinar si otros estudios similares han encontrado alguna evidencia de dependencia o autocorrelación en las observaciones. Esto puede proporcionar información adicional sobre la independencia de las observaciones en tu propio estudio.

Para asegurar el cumplimiento del supuesto de independencia de observaciones en un análisis de varianza (ANOVA), es esencial seguir un diseño experimental apropiado, analizar los residuos y revisar exhaustivamente la literatura relevante en el área de estudio. Estas estrategias se combinan para evaluar y confirmar la independencia de las observaciones, fundamental para mantener la validez de los resultados obtenidos a través del ANOVA.

2. ANOVA en R

2.1) Normalidad de la variable dependiente

La función 'shapiro.test()' es una función estadística utilizada en el software de programación R. Esta función se utiliza para realizar una prueba de normalidad en un conjunto de datos. En otras palabras, evalúa si una muestra de datos proviene de una distribución normal (también conocida como distribución gaussiana).

La prueba de Shapiro-Wilk es una de las pruebas de normalidad más comunes y se utiliza para determinar si los datos siguen una distribución normal. La hipótesis nula (H0) de esta prueba es que los datos siguen una distribución normal, mientras que la hipótesis alternativa (H1) es que los datos no siguen una distribución normal.

```
data:  Edad
W = 0.91559, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  VCu
W = 0.81782, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  VCd
W = 0.91, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  VCt
W = 0.91133, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  EducacionM
W = 0.86971, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  VCu
W = 0.81782, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  VCd
W = 0.91, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  VCt
W = 0.91133, p-value < 2.2e-16
```

```
data:  NotasGu
W = 0.98554, p-value = 4.934e-06


        Shapiro-Wilk normality test

data:  VCu
W = 0.81782, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  VCd
W = 0.91, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  VCt
W = 0.91133, p-value < 2.2e-16
```

```
        Shapiro-Wilk normality test

data:  NotasGd
W = 0.96167, p-value = 5.583e-12


        Shapiro-Wilk normality test

data:  VCu
W = 0.81782, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  VCd
W = 0.91, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  VCt
W = 0.91133, p-value < 2.2e-16
```

```
data:  NotasGt
W = 0.92598, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  VCu
W = 0.81782, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  VCd
W = 0.91, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  VCt
W = 0.91133, p-value < 2.2e-16
```

Vemos que en los diferentes casos el valor de probabilidad (p) es muy superior a nuestro nivel elegido (0.05), por lo que no rechazamos la hipótesis nula.

2.2) Homocedasticidad Prueba de Breusch-Pagan

La prueba de Breusch-Pagan se utiliza para determinar si la heteroscedasticidad está presente o no en un modelo de regresión.

La prueba utiliza las siguientes hipótesis nulas y alternativas :

- ➢ Hipótesis nula (H0): la homocedasticidad está presente (los residuos se distribuyen con la misma varianza)
- ➢ Hipótesis alternativa (HA): Heteroscedasticidad está presente (los residuos no se distribuyen con la misma varianza)

Si el valor p de la prueba es menor que algún nivel de significancia (es decir, $\alpha = 0.5$), entonces se rechaza la hipótesis nula y se concluye que la heterocedasticidad está presente en el modelo de regresión.

```
> #Analisis Homocedasticidad  Prueba de Breusch-Pagan
> resultadou <- bptest(Edad ~ VCu); resultadou

        studentized Breusch-Pagan test

data:  Edad ~ VCu
BP = 0.072161, df = 1, p-value = 0.7882

> resultadoud <- bptest(Edad ~ VCd); resultadoud

        studentized Breusch-Pagan test

data:  Edad ~ VCd
BP = 1.1389, df = 1, p-value = 0.2859

> resultadout <- bptest(Edad ~ VCd); resultadout

        studentized Breusch-Pagan test

data:  Edad ~ VCd
BP = 1.1389, df = 1, p-value = 0.2859
```

```
> #Analisis Homocedasticidad  Prueba de Breusch-Pagan
> resultadouu <- bptest(EducacionM ~ VCu); resultadouu

        studentized Breusch-Pagan test

data:  EducacionM ~ VCu
BP = 1.6383, df = 1, p-value = 0.2006

> resultadodd <- bptest(EducacionM ~ VCd); resultadodd

        studentized Breusch-Pagan test

data:  EducacionM ~ VCd
BP = 0.4145, df = 1, p-value = 0.5197

> resultadodt <- bptest(EducacionM ~ VCd); resultadodt

        studentized Breusch-Pagan test

data:  EducacionM ~ VCd
BP = 0.4145, df = 1, p-value = 0.5197
```

```
> #Analisis Homocedasticidad  Prueba de Breusch-Pagan
> resultadotu <- bptest(NotasGu ~ VCu); resultadotu

        studentized Breusch-Pagan test

data:  NotasGu ~ VCu
BP = 0.003941, df = 1, p-value = 0.9499

> resultadotd <- bptest(NotasGu ~ VCd); resultadotd

        studentized Breusch-Pagan test

data:  NotasGu ~ VCd
BP = 0.62799, df = 1, p-value = 0.4281

> resultadott <- bptest(NotasGu ~ VCd); resultadott

        studentized Breusch-Pagan test

data:  NotasGu ~ VCd
BP = 0.62799, df = 1, p-value = 0.4281
```

```
> #Analisis Homocedasticidad  Prueba de Breusch-Pagan
> resultadocu <- bptest(NotasGd ~ VCu); resultadocu

        studentized Breusch-Pagan test

data:  NotasGd ~ VCu
BP = 0.54538, df = 1, p-value = 0.4602

> resultadocd <- bptest(NotasGd ~ VCd); resultadocd

        studentized Breusch-Pagan test

data:  NotasGd ~ VCd
BP = 0.17264, df = 1, p-value = 0.6778

> resultadoct <- bptest(NotasGd ~ VCd); resultadoct

        studentized Breusch-Pagan test

data:  NotasGd ~ VCd
BP = 0.17264, df = 1, p-value = 0.6778
```

```
> #Analisis Homocedasticidad  Prueba de Breusch-Pagan
> resultadosu <- bptest(NotasGt ~ VCu); resultadosu

        studentized Breusch-Pagan test

data:  NotasGt ~ VCu
BP = 0.17889, df = 1, p-value = 0.6723

> resultadosd <- bptest(NotasGt ~ VCd); resultadosd

        studentized Breusch-Pagan test

data:  NotasGt ~ VCd
BP = 1.4319, df = 1, p-value = 0.2315

> resultadost <- bptest(NotasGt ~ VCd); resultadost

        studentized Breusch-Pagan test

data:  NotasGt ~ VCd
BP = 1.4319, df = 1, p-value = 0.2315
```

2.3) Independencia de observación, por correlación de Pearson

La prueba de correlación de Pearson , entrega los niveles de relación que poseen las variables estudiadas, intentando comparar con valores que poseen cierto significado, por ejemplo.

- Si el valor entregado por la prueba de Pearson es $|r| \approx 1$, implica que la correlación lineal entre las variables estudiadas es fuerte y hay cierta influencia entre ellas
- Si el valor de $|r| \approx 0$, indica que no hay correlación aparente entre las variables estudiadas.
- Si el valor exacto de $r = -1$ implica que la correlación es perfectamente negativa y son proporcionalmente inversas, es decir, si una baja, la otra naturalmente subirá.
- Si el valor exacto de $r = 1$ indica que la correlación es perfectamente positiva y ambas aumentan de manera proporcional.
- Si el valor de $r = 0$, indica que existe una falta de correlación entre las variables estudiadas y no se puede inferir que existe una relación entre ambos grupos de datos,

```
> #Independencia de Observaciones Por correlacion de Pearson
> correlationu <- cor(Edad, VCu, method = "pearson");correlationu
[1] -0.02055946
> correlationd <- cor(Edad, VCd, method = "pearson");correlationd
[1] -0.004910259
> correlationt <- cor(Edad, VCt, method = "pearson");correlationt
[1] 0.1128046
```

```
> #Independencia de Observaciones Por correlacion de Pearson
> correlationud <- cor(EducacionM, VCu, method = "pearson");correlationud
[1] 0.02442057
> correlationdd <- cor(EducacionM, VCd, method = "pearson");correlationdd
[1] -0.0196863
> correlationtd <- cor(EducacionM, VCt, method = "pearson");correlationtd
[1] 0.009536494
```

```
> #Independencia de Observaciones Por correlacion de Pearson
> correlationut <- cor(NotasGu, VCu, method = "pearson");correlationut
[1] 0.0487946
> correlationdt <- cor(NotasGu, VCd, method = "pearson");correlationdt
[1] -0.09449661
> correlationtt <- cor(NotasGu, VCt, method = "pearson");correlationtt
[1] -0.07405263
```

```
> #Independencia de Observaciones Por correlacion de Pearson
> correlationuc <- cor(NotasGd, VCu, method = "pearson");correlationuc
[1] 0.08958778
> correlationdc <- cor(NotasGd, VCd, method = "pearson");correlationdc
[1] -0.1066779
> correlationtc <- cor(NotasGd, VCt, method = "pearson");correlationtc
[1] -0.07946919
```

```
> #Independencia de Observaciones Por correlacion de Pearson
> correlationus <- cor(NotasGt, VCu, method = "pearson");correlationus
[1] 0.06336113
> correlationds <- cor(NotasGt, VCd, method = "pearson");correlationds
[1] -0.1227049
> correlationts <- cor(NotasGt, VCt, method = "pearson");correlationts
[1] -0.08764072
```

Como se puede observar, a través del cálculo de la correlación de Pearson con cierto grupo de variables a estudiar, se puede generalizar que los casos que más se repiten de los mencionados al inicio del ítem, son los valores que trabajan con el valor entregad "r", con valor absoluto, entregando cierta información sobre la correlación existente con los grupos de variables trabajados.

3. Conclusión

En base a los datos obtenidos y calculados a lo largo del análisis realizado a la base de datos correspondiente, se puede concluir que las variables tratadas, poseen cierta influencia unas con otras, además de que existe también cierto grado de correlación donde se pueden inferir casos como la proporción inversa de estos datos y la proporción directa entre ellos.

Par finalizar, se puede inferir que los datos estudiados poseen cierta relación unos con otros y genera cierto interés conocer el grado de correlación, el cual fue obtenido a través de pruebas como la correlación de Pearson, la cual entrega el grado exacto de relación existente y también la dirección de la relación.

4. Referencias Bibliográficas

[1] de La Varianza Con Un Factor, 1-Análisis. (s/f). *ESTADÍSTICA APLICADA. PRÁCTICAS CON SPSS. TEMA 2*. Ugr.es. Recuperado el 25 de septiembre de 2023, de https://www.ugr.es/~metcuant/asignaturas/docencia/estadistica%20aplicada/SPSS/SPSS_T2.pdf

[2]Rodrigo, J. A. (s. f.). *Análisis de la homogeneidad de varianza (homocedasticidad) con R*. https://cienciadedatos.net/documentos/9_homogeneidad_de_varianza_homocedasticidad

[3] Peláez, I. M. (s/f). *Comparación de medias*. Revistaseden.org. Recuperado el 25 de septiembre de 2023, de https://www.revistaseden.org/files/12-CAP%2012.pdf