Unidad 5: Análisis de asociación / Análisis Multivariable

Los estudios de género y su relación con la Ciencia Política son de gran relevancia para contrarrestar el sistema de discriminación existente en la cultura política hacia las mujeres y las poblaciones LGTBIQ de la sociedad. La participación de estos grupos humanos en la política debe ser tanto en función de representados como en función de representantes por lo que las estructuras informales e imaginarios que sitúan en una posición preferente al hombre deben ser analizadas y rebatidas. En este camino, la encuesta realizada por el Instituto de Opinión Pública de la Pontificia Universidad Católica del Perú nos permite apreciar el contexto en el que se ubica a la mujer peruana, las ideas que rodean los aspectos más básicos de su vida como el ámbito laboral, familiar y económico. Asimismo, los roles de género son abordados también en esta encuesta y nos permiten medir las percepciones que existen para ambos sexos. Para esta unidad aplicaremos las herramientas estadísticas a aprender en esta encuesta, intentando inferir afirmaciones sobre los roles de género en la sociedad peruana.

Análisis de Asociación

Para comprobar que existe una relación entre dos variables en la estadística se utiliza el *Análisis de* Asociación. Esta es una técnica inferencial de tipo bivariada; es decir, incluye dos variables. Nuestro objetivo al ejecutar este comando es ver si entre los casos de las variables mencionadas existe algún grado de relación. Para determinar esa situación, una serie de pruebas matemáticas han sido desarrolladas, estas se utilizan según el tipo de variables que queremos analizar. Los dos grandes ejes de este tema son la *asociación de variables nominales* y la *asociación para variables ordinales.* Para el primer tema se han desarrollado fórmulas como la *Prueba Chi,* la *Prueba Phi* o la *V de Crammer.* Para las variables ordinales, tenemos la *Prueba Gamma* y la *Prueba Tau B*, entre otras.

Para el software de R, al momento de ejecutar la asociación las variables tienen que estar distribuidas en una *tabla de contingencia*; por eso, como paso previo a la asociación vamos a explicar cómo se desarrolla esta herramienta en R.

Abriendo base de datos de temas de género del Instituto de Opinión Pública de la PUCP. Se utiliza el paquete *foreign* para exportar data de SPSS

```
install.packages("foreign")
library (foreign)
```

```
PROYECTO<- read.spss("IOPGENERO.sav",use.value.labels=TRUE, max.value.labels=Inf,
to.data.frame=TRUE) #leyendo data de genero
```

Tablas de contingencia:

Esta herramienta es una tabla de doble entrada, donde en el eje vertical se distribuyen las categorías de la primera variable y en el eje horizontal se distribuyen las categorías de la segunda variable. Esto nos permite visualizar de manera clara la distribución de los datos según las categorías conectadas de las variables. Para entender el método describamos las variables P4B y P4C.

P4B- Una pareja de mujeres lesbianas puede criar a un hijo tan bien como una pareja de hombre y mujer

```
describe(P4B)
```

```
     n missing  unique
  1203       0       7
```

Muy de acuerdo (12, 1%), De acuerdo (225, 19%)

En desacuerdo (612, 51%), Muy en desacuerdo (280, 23%)

Ni de acuerdo ni en desacuerdo (31, 3%)

No sabe (32, 3%), No contesta (11, 1%)

P4C – Una pareja de hombres homosexuales puede creer a un hijo tan bien como una pareja de hombre y mujer

```
describe(P4C)
```

```
     n missing  unique
  1203       0       7
```

Muy de acuerdo (10, 1%), De acuerdo (146, 12%)

En desacuerdo (637, 53%), Muy en desacuerdo (331, 28%)

Ni de acuerdo ni en desacuerdo (30, 2%)

No sabe (36, 3%), No contesta (13, 1%)

Ahora, para generar una *tabla de contingencia* se tiene que instalar el paquete *"descr"*.

```
Install.packages("descr")

library(descr) #paquete descr para generar una tabla de contingencia


table(P4B,P4C) #mostrar la tabla
```

	MA	DA	ED	MD	NAND	NS	NC
MA	7	0	2	2	1	0	0
DA	2	135	67	12	1	4	4
ED	0	11	550	48	2	0	1
MD	1	0	14	264	0	1	0
NAND	0	0	2	4	25	0	0
NS	0	0	0	1	0	31	0
NC	0	0	2	0	1	0	8

La tabla nos permite observar detalladamente cuantos casos existen para cada subconjunto existente. Por ejemplo, las personas que se encentran *muy de acuerdo* (MA) con que una pareja de lesbianas pueda criar a un hijo como una pareja heterosexual y las que están *muy de acuerdo* (MA) con que una pareja de homosexuales hombres pueda criar a un hijo como una pareja heterosexual son 10, esto se ve en la primera intersección de la tabla, estos son los subconjuntos que la *tabla de contingencia* crea.

También se puede apreciar que la suma final de una línea, ya sea vertical u horizontal, nos da el total de casos existentes para la categoría de una variable. Por ejemplo, la suma de la última línea vertical para la categoría *no contesta* (NC) de la variable P4C, nos da un total de 8 casos, estos son los mismos datos que la distribución de frecuencias de la variable. Esta situación se evidencia para la sumatoria de cada línea- vertical u horizontal-. La sumatoria de la línea final vertical y de la línea final horizontal- ubicada en la esquina inferior derecha y que en este caso no aparece- sería el total de casos para la variable.

```
prop.table(table(P4B,P4C),2)*100 #convertir la tabla a frecuencias
```

	MA	DA	ED	MD	NAND	NS	NC
MA	70.00	0.00	0.31	0.60	3.33	0.00	0.00
DA	20.00	92.46	10.51	3.62	3.33	11.11	30.76
ED	0.00	7.53	86.34	14.50	6.66	0.00	7.69
MD	10.00	0.00	2.19	79.75	0.00	2.77	0.00

NAND	0.00	0.00	0.31	1.20	83.33	0.00	0.00
NS	0.00	0.00	0.00	0.30	0.00	86.11	0.00
NC	0.00	0.00	0.31	0.00	3.33	0.00	61.53

Con este comando lo que se hace es distribuir la tabla según sus frecuencias. Luego de esto tenemos que guardar la tabla con un nombre de nuestra preferencia.

```
newtable<- table(P4B,P4C)#guardar la tabla en un nuevo elemento
```

Requisito Chi^2:

Para realizar la *asociación*, tanto para variables nominales, como para variables ordinales; se tiene que verificar la existencia de un requisito: Chi^2. Este requisito es una prueba de hipótesis, donde la hipótesis nula es que no existe asociación, mientras que la hipótesis alternativa señala que si existe asociación.

Al ser una prueba de significancia, el resultado es numérico, por lo que si tenemos establecido trabajar con una confianza del 95%, la significancia al ser mayor a 0.05 habrá pasado la hipótesis nula. Si la significancia es menor a 0.05, entonces se rechaza la hipótesis nula- la no existencia de asociación- y se acepta la hipótesis alternativa- existencia de asociación-. Al realizar una prueba Chi^2 se obtiene los siguientes resultados:

```
chisq.test(newtable)#verificar chi-square
```

```
        Pearson's Chi-squared test

data:  newtable

X-squared = 4224.816, df = 36, p-value < 2.2e-16

Mensajes de aviso perdidos

In chisq.test(newtable) : Chi-squared approximation may be incorrect
```

El Chi2 menor a 0.05 nos indica que debemos *rechazar* la H0; por lo tanto, rechazamos que no existe asociación y aceptamos la hipótesis alternativa: si existe asociación. La *asociación* nos permitirá ver en qué grado se encuentran asociadas estas variables. Un resultado por encima de 0.5 en una escala de 0 a 1, nos señala que es una asociación relevante.

Coeficientes de Asociación para Variables Nominales:

Para ejecutar en R una asociación es necesario instalar el paquete *vcd* y utilizar el comando *assocstats*.

```
install.packages("vcd")#instalar el paquete vcd para asociación de nominales

library(vcd)#ejecutar paquete

assocstats(newtable)#pedir coeficientes de asociación
```

```
                    X^2 df P(> X^2)
Likelihood Ratio 1788.9 36       0
Pearson          4224.8 36       0
```

Phi-Coefficient   : NA

Contingency Coeff.: 0.882

Cramer's V   : 0.765

Para determinación el grado de asociación entre dos variables nominales suelen haber varias pruebas que nos darán el resultado que buscamos. En este caso, el comando nos da el resultado con la *Prueba phi,* con la *V de Cramer* y con el *Coeficiente de Contingencia*. El primero se suele utilizar para las variables nominales que tienen dos categorías por variable. La *V de Cramer* se utiliza con cualquier tipo de tabla, al igual que el *Coeficiente de Contingencia*. Esto significa que para nuestra tabla- 7x7- se puede utilizar tanto la *V de Cramer*, como el *Coeficiente de Contingencia* para determinar el nivel de asociación de las variables.

La *V de Cramer* siempre va a subestimar la relación; por eso, el resultado que obtenemos-0.765- es menor al que obtenemos con el *Coeficiente de Contingencia*-0.882-. En todo caso, ambos nos indican que existe un alto grado asociación entre la consideración **de que una pareja de homosexuales hombres y una pareja de homosexuales lesbianas puede criar a un niño tan bien como una pareja heterosexual.**

<u>Coeficientes de Asociación para Variables Ordinales:</u>

Para realizar la asociación, seguiremos los mismos pasos que para la asociación de variables ordinales. Primero, se debe recategorizar las variables GEDAD y P39 de nominal a ordinal, luego crear la tabla y guardarla como frecuencias; acto seguido, calculamos el Chi2 como requisito previo para la existencia de asociación. Por último, procedemos a pedir los estadísticos específicos para la asociación de ordinales.

Una diferencia entre ambas asociaciones es el paquete y el comando a utilizar, para las ordinales tenemos que utilizar el paquete *vcdExtra* y el comando *GKgamma(tabla)*.

```
install.packages("vdcExtra")#instalar vcdExtra para ordinales

library(vcdExtra)#ejecutar vcdExtra
```

GD- Grupos de edad

```
describe(GD)
```

```
      n missing  unique
   1203       0       3
45 a mÃ¡s (408, 34%), 30 a 44 (385, 32%)
18 a 29 (410, 34%)
```

P39- ¿Cuál es el grado de satisfacción que tiene Ud. con su trabajo (principal)?

```
describe(P39)
```

```
      n missing  unique
    889     314       7
Me siento completamente satisfecho/a (92, 10%)
Muy satisfecho/a (242, 27%)
Bastante satisfecho/a (303, 34%)
Ni satisfecho ni insatisfecho/a (212, 24%)
Bastante insatisfecho/a (26, 3%)
Muy insatisfecho/a (12, 1%)
Completamente insatisfecho/a (2, 0%)
```

```
library(descr)
```

```
table(GD,P39)
```

```
      P39
```

GD	CSat	MSat	Bsat	NSatNIn	BIns	MIns	CIns
45	36	77	101	71	8	5	2
30 a 44	31	87	105	81	12	5	0
18 a 29	25	78	97	60	6	2	0

```
TABordi<-table(GD,P39)
```

```
chisq.test(TABordi)
```

```
        Pearson's Chi-squared test

data:  TABordi

X-squared = 9.0725, df = 12, p-value = 0.6967
```

El Chi2 mayor a 0.05 nos indica que debemos *aceptar* la H0; por lo tanto, aceptamos que no existe asociación. La *asociación* nos permitiría ver en qué grado se encuentran asociadas estas variables. Un resultado por encima de 0.5 en una escala de 0 a 1, nos señala que es una asociación relevante. Sin embargo, al aceptar la inexistencia de asociación entre ambas variables veamos cómo se traduce esto en la prueba de asociación.

```
GKgamma(TABordi)#pedir coeficientes para ordinales
```

```
gamma      : -0.016

std. error  : 0.041

CI         : -0.096 0.064
```

Estas son los resultados que deberíamos tomar en cuenta en caso la tabla hubiera obtenido un Chi2 mayor a 0.05. En este caso, no existe asociación entre las variables ordinales. Además, vemos que en toda la base de datos no hay más variables ordinales por lo que no se pueden trabajar más casos de asociación a menos que se fuercen las variables nominales y se recatogoricen a nominales. Esta situación será muy común cuando trabajemos con otras datas, no siempre la información se va a adecuar a nuestros requerimientos; por eso, es necesario conocer más de una sola técnica para saber contrarrestar esa situación de desventaja.

Análisis de Correlación

El correlato del *análisis de asociación* para variables escalares es el *análisis de correlación,* esta herramienta se ha diseñado para identificar la relación entre dos variables numéricas. El resultado nos permite apreciar la intensidad de la correlación y si esta es una relación directa o una indirecta. Sin embargo, es importante recalcar que esta herramienta no nos permite señalar una relación de causalidad- dirección- entre las variables.

En esta oportunidad, utilizaremos las variables escalares **P1** y **P2** y el comando *cor(variable1,variable2, use="complete.obs")*. Pero primero, es necesario estandarizar las variables

```
P1Z<-scale(P1)#convertir a la puntuación Z
```

```
summary(P1Z)
```

```
V1
 Min.   :-3.0060
 1st Qu.:-0.3280
 Median:-0.3280
 Mean   : 0.0000
 3rd Qu.: 0.7432
 Max.   : 9.0451
 NA's   :42
```

```
P2Z<- scale(P2)
```

```
summary(P2Z)
```

```
V1
 Min.   :-2.7270
 1st Qu.:-0.9185
 Median: 0.1149
 Mean   : 0.0000
 3rd Qu.: 0.3733
```

Max. : 3.4737

NA's :41

```
cor(P1Z,P2Z, use="complete.obs")#obtener puntuación de correlación
```

[1] 0.7471912

El resultado nos indica que entre la variable **P1** y la variable **P2** existe una correlación de 0.74. Sin embargo, una muestra más exacta nos proporciona la significa de la prueba de correlaciones: el comando *cor.test(P1,P2, use="complete.obs")*.

```
cor.test(P1Z,P2Z, use="complete.obs")#obtener correlación con significancia
```

Pearson's product-moment correlation

data: P1 and P2

t = 38.2247, df = 1156, **p-value < 2.2e-16**

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.7206014 0.7715865

sample estimates:

cor

0.7471912

Con la significancia que este comando nos da, aceptamos o rechazamos la hipótesis nula- H0-: la inexistencia de correlación entre las variables. Al igual que para la asociación, la hipótesis alternativa- H1- es la existencia de correlación entre las variables. Esta, al ser menor a 0.05, nos indica que se rechaza la H0 y se acepta la H1; es decir, rechazamos que no existe correlación: existe correlación. Además, esta correlación es alta- por ser mayor a 0.7- y es directa- por ser de signo positivo-.

Análisis Multivariable

La estadística le permite encontrar estas diferencias con la prueba ANOVA. Esta es una herramienta que busca indicar si existe igualdad en las medias de tres grupos o más con respecto a una misma variable. Para esto, requiere de dos variables: una numérica y una nominal politómica y el único pre-requisito para ejecutarla es la normalidad en la variable escalar.

La hipótesis nula para ANOVA es la existencia de igualdad de medias entre los grupos segmentados de la variable nominal; por lo tanto, la hipótesis alternativa para la prueba ANOVA es la inexistencia de igualdad de medias. Para ejemplificar esta situación veamos las variables **EDAD** y **P3A** de la data de temas de Género del IOP.

```
describe(EDAD)
```

```
EDAD
       n missing  unique    Info    Mean     .05     .10     .25     .50     .75     .90     .95
    1203       0      67       1   38.98      19      20      26      36      49      62      68

lowest : 18 19 20 21 22, highest: 82 83 84 87 92
```

```
describe(P3A)
```

```
P3A
       n missing  unique
    1203       0       7
```

Muy de acuerdo (67, 6%), De acuerdo (377, 31%)

En desacuerdo (569, 47%), Muy en desacuerdo (99, 8%)

Ni de acuerdo ni en desacuerdo (63, 5%)

No sabe (22, 2%), No contesta (6, 0%)

Pruebas de Normalidad

El primer paso para ejecutar ANOVA es probar la normalidad de la variable escalar con la que queremos trabajar. Como se mencionó en la Unidad 3, para probar la normalidad se tiene que aceptar la hipótesis nula verificando que la significancia sea mayor a 0.05. Las reglas son las mismas que se mencionaron líneas arriba: para una población mayor a 50, se utiliza el *test Kolmogorov-Smirnov* y para una población menor a 50, se utiliza el *test de Shapiro-Wilk*. Asimismo, la significancia debe ser mayor a 0.05 para aceptar la hipótesis nula: la existencia de normalidad en la variable.

```
shapiro.test(EDAD)$p.value #pedir test de Shapiro-Wilk
```

[1] **2.174357e-20**

```
library(nortest)#ejecutar paquete nortest
```

```
lillie.test(EDAD)$p.value #pedir test de Kolmogorov-Smirnov
```

[1] **1.179652e-26**

En este caso, se rechaza la H0 de normalidad puesto que la significancia es menor a 0.05. De esta forma, para las variables **EDAD** y **P3A** NO SE PODRÍA REALIZAR ANOVA. Si se revisan las demás variables escalares de la data **PROYECTO**, descubriremos que ningún pasa la prueba normalidad. Es común que esto suceda debido a que esta es una base de datos real a nivel nacional. De esta forma, no se puede forzar que los datos recogidos en un muestreo nacional se establezcan conforme a una prueba estadística. Sin embargo, con miras a establecer los procedimientos para ejecutar ANOVA, se continuarán los demás procesos con un recordatorio de que Anova no es una herramienta aplicable en este contexto.

Test de Levene

Una vez confirmado que la variable escalar es normal, se procede a verificar la homogeneidad de varianzas con la prueba de Levene. En este caso la hipótesis nula es la existencia de igualdad de varianzas; por lo tanto, la hipótesis alternativa es la inexistencia de igualdad de varianzas. Así, cuando la significancia sea mayor a 0.05, se procede a aceptar la hipótesis nula. Para esto es necesario instalar el paquete *lawstat* y utilizar el comando *levene.test*.

```
install.packages("lawstat")#instalar paquete lawstat
```

```
library(lawstat)#ejecutar paquete lawstat
```

```
levene.test(EDAD, P3A, location=c("median", "mean", "trim.mean"), trim.alpha=0.25,bootstrap = "false", num.bootstrap=1000, kruskal.test=FALSE, correction.method=c ("none", "correction.factor", "zero.removal","zero.correction"))#pedir test de levene
```

modified robust Brown-Forsythe Levene-type test based on the absolute

deviations from the median

data: EDAD

Test Statistic = 2.1563, **p-value = 0.04482**

La homogeneidad o heterogeneidad de varianzas nos dirá qué estadísticos usar para describir las diferencias entre los grupos de la variable nominal. Debido a que la significancia es menor a 0.05 (0.044) se rechaza la H0 de la prueba de homogeneidad de varianzas; es decir, las varianzas son distintas. Esto nos permitirá, más adelante, qué estadísticos usar para describir las diferencias.

Anova

Esta prueba nos confirmará la existencia de igualdad de medias entre los distintos grupos de la variables politómica respecto de la variable escalar. Para esto, observamos la significancia de la Prueba F. Si se acepta la H0: existencia de igualdad de medias, se deduce que las medias entre los grupos son iguales; por lo tanto, no varían significativamente. Para esto, la significancia tiene que ser mayor a 0.05. Si la significancia es menor a 0.05, se rechaza la H0 y se acepta la H1: la inexistencia de igualdad de medias entre los grupos. Para esto, es necesario instalar el paquete *multcomp* y utilizar el comando *fit.* Es necesario señalar que para la conformación de ANOVA la variable escalar se establece como la variable dependiente y la variable politómica se establece como el factor. Instalemos el paquete *multcomp.*

```
install.packages("multcomp")#instalar paquete multcomp

library(multcomp)#ejecutar multcomp


fit <- aov(EDAD ~ P3A)#ejecutar ANOVA y guardar en un elemento


summary(fit)#describir ANOVA
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
P3A	6	1261	210.1	0.861	**0.523**
Residuals	1196	291668	243.9		

Debido a que la significancia del *estadístico F* es mayor a 0.05 (0.523) se acepta la H0: la existencia de igualdad de medias. Este paso nos permite concluir con la prueba ANOVA. Sin embargo, se señaló que la variable escalar no pasaba la prueba de normalidad; por lo tanto, no SE PUEDE EJECTUAR ANOVA. En caso se hubiera aceptado la normalidad, la prueba de ANOVA estaría concluida y se hubiera aceptado la igualdad de medias en los grupos de la variable politómica.

Pruebas Post-Hoc: TukeyHSD

En caso exista diferencia entre las medias de los grupos de la variable nominal politómica con respecto a la variable escalar, una serie de estadísticos se utilizan para señalar en qué grupos existe diferencia. Si ANOVA nos muestra que existen diferencias entre las medias de los grupos; es decir, se rechace H0 y se acepte H1 y haya homogeneidad de varianzas: la prueba *Tukey* es uno de los muchos estadísticos que se utilizan para encontrar estas diferencias.

```
TukeyHSD(fit)#pedir Post-Hoc
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = EDAD ~ P3A)
```

	diff	lwr	upr	p adj
DA - MA	0.38	-5.73	6.49	**0.99**
ED - MA	-0.62	-6.57	5.33	**0.99**
MD - MA	0.11	-7.17	7.41	**1.00**
NAND - MA	3.81	-4.27	11.91	**0.80**
NS - MA	1.27	-10.06	12.60	**0.99**
NC - MA	2.42	-17.23	22.07	**0.99**
ED - DA	-1.00	-4.06	2.05	**0.96**
MD - DA	-0.26	-5.46	4.94	**0.99**
NAND - DA	3.43	-2.83	9.71	**0.67**
NS - DA	0.89	-9.22	11.00	**0.99**
NC - DA	2.04	-16.93	21.01	**0.99**
MD - ED	0.74	-4.28	5.76	**0.99**
NAND - ED	4.44	-1.68	10.56	**0.32**
NS - ED	1.89	-8.12	11.91	**0.99**
NC - ED	3.04	-15.88	21.97	**0.99**
NAND - MD	3.69	-3.73	11.13	**0.76**
NS - MD	1.15	-9.71	12.02	**0.99**
NC - MD	2.30	-17.08	21.69	**0.99**
NS - NAND	-2.54	-13.96	8.87	**0.99**
NC - NAND	-1.39	-21.10	18.30	**0.99**

NC - NS	1.15	-20.09	22.39	**0.99**

Vemos que TukeyHSD nos muestra todas las comparaciones poblaciones entre los grupos de la variable politómica. Nos explica, la diferencia de medias, un valor máximo, uno mínimo y una prueba de significancia. Dado que la prueba de normalidad no permite que realicemos ANOVA, tampoco podemos leer los Post-HOC. Si el caso fuera contrario, dado que la prueba ANOVA estableció que existe igualdad de medias, deducimos que no hay grandes diferencias entre las medias de los grupos; por lo tanto, las pruebas Post- HOC tampoco son aplicables en este contexto.

Ahora es tu turno

Ejercicios

- Usted se encuentra planteando una investigación sobre la inseguridad ciudadana y la participación de la Policía Nacional en la reducción del crimen. Para esto está utilizando el Registro Nacional de Delitos del INEI. Analice si existe algún tipo de asociación entre las siguientes variables: IH203, IH201 e IH204 y ensaye una hipótesis que explique esta asociación.

- La variable central de su investigación sobre la influencia de los sobre el Gobierno Central del Perú es un conteo sobre el número de víctimas de crímenes en el año 2013. Aplique Anova para analizar el grado de intervención de las dependencias policiales sobre estos delitos. Utilice las variables IH212 y IH201.