Estadística para las Ciencias Sociales con 

Cuaderno de Ejercicios (v.1.0)

David Sulmont

César Córdova

Pontificia Universidad Católica del Perú

Marzo 2015

Presentación

El presente Cuaderno de Ejercicios contiene un conjunto de ejemplos y ejercicios destinados al aprendizaje y práctica de las técnicas y herramientas estadísticas que son parte del contenido de "Estadística para las Ciencias Sociales con R", diseñado por el profesor David Sulmont. Estos contenidos se ofrecen en los cursos de:

- "Estadística para el Análisis Sociológico 1", en la especialidad de Sociología de la Facultad de Ciencias Sociales
- "Técnicas de Análisis Sociológico", en la Maestría de Sociología de la Escuela de Posgrado.

Salvo que se indique expresamente en el texto otra fuente, las bases de datos de los ejercicios propuestos deberán descomprimirse desde el archivo que se descarga desde el siguiente enlace:

https://sites.google.com/a/pucp.pe/data_est/archivos/BDs_Workbook.zip?attredirects=0&d=1

Al igual que el texto "Notas de Clase", se trata de un material en desarrollo y que busca perfeccionarse continuamente con los comentarios y aportes de los estudiantes que llevan los cursos donde se desarrollan estos contenidos. En tal sentido agradecemos de antemano los aportes que puedan brindarnos.

David Sulmont y César Córdova[1]

Marzo de 2015

[1] David Sulmont es sociólogo y profesor principal del Departamento de Ciencias Sociales de la PUCP. César Córdova es psicólogo por la PUCP, analista en el Instituto de Opinión Pública de la PUCP y Jefe de Práctica en cursos de estadística de las especialidades de Sociología y de Psicología en los semestres 2014-2 y 2015-1.

1. Análisis univariado y gráficos

Ejemplos

Para esta sección, vamos a trabajar con la base de datos "CONFLICTOS.SAV", correspondiente al estudio sobre "Representación Política y Conflicto Social", realizado por el Instituto de Opinión Pública de la PUCP (IOP) en 2012. Los detalles de este estudio pueden verse en la plataforma IOP-Data. Se recomienda descargar el cuestionario de ese estudio para tenerlo como libro de códigos.

Para realizar los ejercicios usted deberá importar en el R la base de datos en SPSS, asignándolo a un objeto tipo data frame llamado "df" (si lo desea puede identificar al data frame usando otro nombre, pero para este ejercicio sugerimos usar el nombre "df"). Luego explore los nombres de las 40 primeras variables de la base de datos.

Carga de datos y vista de variables

```
# Código para cargar los datos y ver las variables de la BD
library(foreign)
df <- as.data.frame(read.spss("CONFLICTOS.sav"))
```

```
## re-encoding from latin1

head(names(df), 40) # Ver las primeras 40 variables
```

```
##  [1] "NUM"       "SEXO"      "EDAD"      "P1A"       "P1A_OTRO"  "P1B"
##  [7] "P1B_OTRO"  "P2"        "P3"        "P4A"       "P4B"       "P4C"
## [13] "P4D"       "P4E"       "P4F"       "P4G"       "P5"        "P6"
## [19] "P7A"       "P7B"       "P8A"       "P8B"       "P8C"       "P8D"
## [25] "P9A"       "P9B"       "P9C"       "P9D"       "P9E"       "P9F"
## [31] "P9G"       "P9H"       "P9I"       "P9J"       "P9K"       "P9L"
## [37] "P9M"       "P10"       "P11"       "P12A"
```

Recodificación y cálculo de un índice

La legitimidad de la protesta pública ha sido definida como el reconocimiento y aceptación por parte de los individuos de que otras personas tienen el derecho a reclamar o realizar algún tipo de protesta que pueda alterar el orden social establecido (Olsen y Baden, 1974).

Para evaluar el nivel de la legitimidad de la protesta pública se consideraron las preguntas P31A, P31B, P31C y P31D del cuestionario que corresponde al estudio "Representación política y Conflictos" (IOP, 2012).

Para poder trabajar con estas preguntas necesitamos hacer algunas transformaciones. Vamos a recodificar los valores de las variables P31A hasta la P31D, de forma tal que

los puntajes más altos (puntaje = 4) reflejen una opinión FAVORABLE hacia el derecho a protestar, y los puntajes más bajos reflejen una opinión DESFAVORABLE (puntaje = 1) hacia el derecho a protestar. En tal sentido, usted deberá convertir las variables P31A hasta la P31D (originalmente factores) en nuevas variables NUMÉRICAS (vectores numéricos), donde los valores numéricos (del 1 al 4) reflejen el esquema de codificación propuesto. En estas nuevas variables, las respuestas "NS / NR" deberán ser consideradas como valores perdidos (NA).

Recodificar la variable P31A:

```
library(car)
table(df$P31A)

##
##    Muy de acuerdo        De acuerdo     En desacuerdo Muy en desacu
erdo
##               240               708               189
14
##           NS / NR
##                52

p31a.r <- as.numeric(df$P31A)
table(p31a.r)

## p31a.r
##   1   2   3   4   5
## 240 708 189  14  52

df$p31a.r <- recode(p31a.r, "1=4; 2=3; 3=2; 4=1; 5=NA")
table(df$p31a.r)

##
##   1   2   3   4
##  14 189 708 240
```

Recodificar la variable P31B

```
table(df$P31B)

##
##    Muy de acuerdo        De acuerdo     En desacuerdo Muy en desacu
erdo
##               111               604               393
27
##           NS / NR
##                68

df$p31b.r <- recode(as.numeric(df$P31B), "5=NA")
table(df$p31b.r)

##
##   1   2   3   4
## 111 604 393  27
```

Pregunta: ¿Por qué estamos recodificando P31A y P31B de distinta manera?

Recodificamos las variables de distinta manera, porque mientras los valores de la variable más altos de la variable P31A indican un nivel más bajo de legitimidad de la protesta pública, en el caso de la variable P31D sucede lo contrario.

Hágalo usted mismo:

Recodifique ahora las preguntas P31C y P31D siguiendo el ejemplo anterior.

Construir un índice

Vamos a construir el índice de legitimidad de la protesta sumando los valores de las cuatro variables que lo componen y restando -4.

```
ind.1 <- (df$p31a.r + df$p31b.r + df$p31c.r + df$p31d.r) - 4
table(ind.1)

## ind.1
##   1   2   3   4   5   6   7   8   9  10  11  12
##   2   8  30  63 148 328 235 169  77  28   8   2
```

Pregunta: ¿Por qué restamos 4?, ¿en qué escala se mide la legitimidad de la protesta y qué valor representaría una legitimidad "media"?

Se resta 4 al puntaje total para que el valor mínimo de la escala sea cero.

Describir la variable "df$legit.prot"

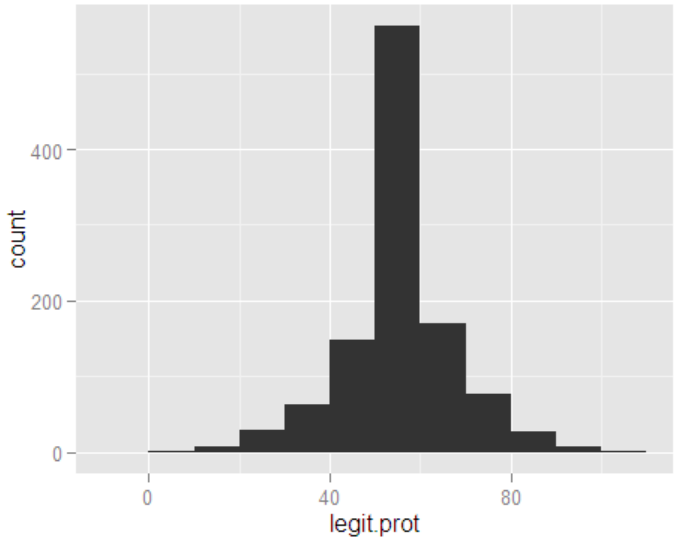
a) Use el comando "freq" del paquete "descr" para pedir una tabla de distribución de frecuencias de la variable "legit.prot", incluyendo la columna de frecuencias acumuladas.

```
library(descr)
freq(ordered(legit.prot), plot=F)

## ordered(legit.prot)
##                   Frequency  Percent Valid Percent Cum Percent
## 8.33333333333333          2   0.1663         0.1821      0.1821
## 16.6666666666667          8   0.6650         0.7286      0.9107
## 25                       30   2.4938         2.7322      3.6430
## 33.3333333333333         63   5.2369         5.7377      9.3807
## 41.6666666666667        148  12.3026        13.4791     22.8597
## 50                      328  27.2652        29.8725     52.7322
## 58.3333333333333        235  19.5345        21.4026     74.1348
## 66.6666666666667        169  14.0482        15.3916     89.5264
## 75                       77   6.4007         7.0128     96.5392
## 83.3333333333333         28   2.3275         2.5501     99.0893
## 91.6666666666667          8   0.6650         0.7286     99.8179
## 100                       2   0.1663         0.1821    100.0000
## NA's                    105   8.7282
## Total                  1203 100.0000       100.0000
```

Pregunta: ¿Qué porcentaje de los entrevistados tienen puntajes superiores a 30?

b) Genere un histograma de la variable "legit.prot" usando el paquete ggplot2, y fije la amplitud de las barras a 10 puntos. ¿Qué forma tiene el histograma?

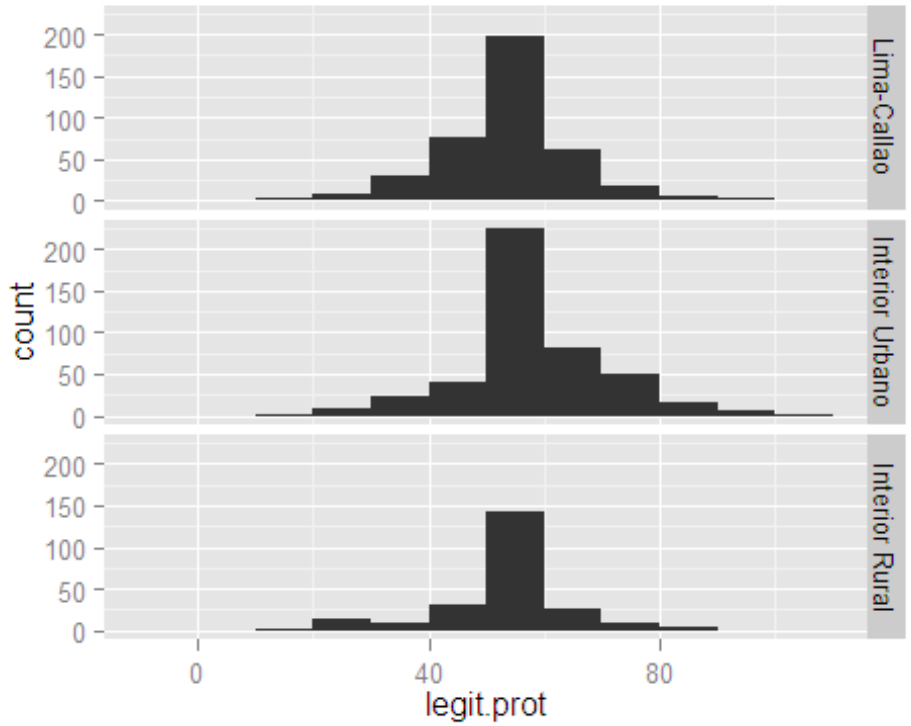
```
ggplot(df, aes(legit.prot)) + geom_histogram(binwidth = 10)
```



Pregunta: ¿Qué forma tiene el histograma?

c) Genere el mismo histograma, usando la opción de facetas (facets) para comparar ámbitos.

```
hist <- ggplot(df, aes(legit.prot)) + geom_histogram(binwidth = 10)
hist1 + facet_grid(ambito ~.)
```



Medidas de tendencia central

Vamos a pedir algunos estadísticos descriptivos de la variable "legit.prot" y luego, lo mismo, pero sólo para Lima y Callao.

```
# Para toda la muestra
mean(df$legit.prot, na.rm=TRUE)

## [1] 54.27

summary(df$legit.prot)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    8.33   50.00   50.00   54.30   66.70  100.00     105

# Para Lima-Callao
df.lc <- subset(df, AMBITO=="Lima-Callao")
mean(df.lc$legit.prot, na.rm=TRUE)

## [1] 52.32

summary(df.lc$legit.prot)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    8.33   41.70   50.00   52.30   58.30   91.70      39
```

Tabla de estadísticos

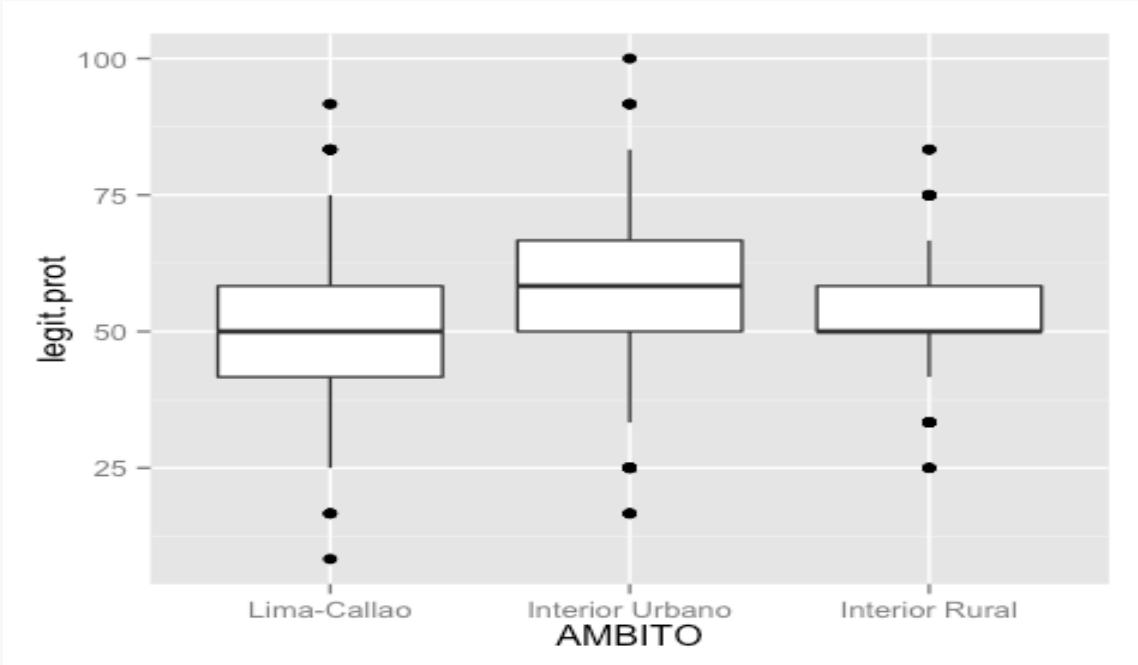
Podemos pedir una tabla con varios estadísticos de una variable por grupos. Por ejemplo, estadísticos de tendencia central y dispersión para "legit.prot", según Ambito de residencia.

```
t.media <- tapply(df$legit.prot, df$AMBITO, mean, na.rm=TRUE)
t.mediana <- tapply(df$legit.prot, df$AMBITO, median, na.rm=TRUE)
t.n <- tapply(df$legit.prot, df$AMBITO, length)

tablita <- cbind(media=t.media, mediana=t.mediana)
round(tablita, digits = 1)

##                 media mediana
## Lima-Callao      52.3    50.0
## Interior Urbano  55.9    58.3
## Interior Rural   54.3    50.0
```

Gráficos: El Boxplot

Veamos un boxplot de la variable legitimidad de la protesta, por Ambito Geográfico:

```
library(ggplot2)
ggplot(df, aes(x=AMBITO, y=legit.prot)) + geom_boxplot()
```



¿Qué diferencias observa entre los ámbitos?, ¿dónde hay mayor legitmidad?, ¿dónde las opiniones son más diversas o similares?

Ejercicios 1.A

Para esta sección, vamos a trabajar con la base de datos “CONFLICTOS.SAV”, correspondiente al estudio sobre “Representación Política y Conflicto Social”, realizado por el Instituto de Opinión Pública de la PUCP (IOP) en 2012. Los detalles de este estudio pueden verse en la plataforma IOP-Data.

Recodifique la variable P17 (auto ubicación del entrevistado en la escala izquierda – derecha) de forma tal que se convierta en un vector numérico que excluya los NS/NR como misssing values. La variable recodificada deberá llamarse p17r

Para ello recomendamos usar la siguiente sintaxis:

```
df$p17r <- as.numeric(df$P17)  -1
df$p17r[df$p17r==11] <- NA
```

(Nota: se resta -1 para que la escala vaya de 0 a 10).

1.A.1. Convierta ind.1 en una variable llamada "df$legit.prot" que mida la legitmidad de la protesta en una escala del 0 al 100. ¿Cómo lo harían?
1.A.2. Pida una tabla de distribución de frecuencias de la variable P17r y responda a la pregunta, ¿los peruanos son más izquierdistas que derechistas?
1.A.3. Genere un gráfico de curva de densidad en ggplot para la variable p17r, divido por facetas ubicadas en forma vertical, según la variable NSEGrup. ¿En qué NSE hay más izquierdistas?
1.A.4. Elabore la tabla necesaria para determinar cuánta gente de derecha votó por Keiko Fujimori en la 1ra y 2da vuelta del 2011.
1.A.5. Recodifique la variable P19 (ubicación del actual gobierno en la escala izquierda – derecha) de forma tal que se obtenga tres grupos: izquierda (de 0 a 3), centro (4 a 6) y derecha (7 a más). ¿Cuántas personas que aprueba la gestión del presidente Humala consideran que su gobierno es de izquierda?, ¿son más que los que lo desaprueban?
1.A.6. Genere un histograma con facets de legit.prot por Nivel Socioeconómico (NSEGrup)
1.A.7. Genere un histograma en ggplot2 que nos permita ver la distribución del índice de legitimidad de la protesta entre personas que se consideran de izquierda (de 0 a 3).
1.A.8. Genere una tabla que nos permita ver las medidas de centralidad de la variable P17r según nivel socioeconómico.
1.A.9. Aplique el comando "mean" y "summary" para "legit.prot" pero sólo para "Interior Rural"

1.A.10. Respecto de la variable autoposicionamiento ideológico (p17r), ¿ubique e interprete los siguientes percentiles: 10; 50; 80 y 90.
1.A.11. Aplique el comando "tapply" para obtener una tabla de la media, mediana y desviación estándar de legit.prot, según NSEGrup.
1.A.12. Elabore un boxplot de la variable legit.prot (legitimidad de la protesta) según auto-ubicación ideológica agrupada en 3 categorías (izquierda, centro, derecha). ¿Cómo interpretarlo?
1.A.13. Elabore un boxplot de la variable de autoposionamiento político del entrevistado (simple, no agrupada), según nivel socioeconómico (NSEGrup). ¿Cómo interpretarlo?
1.A.14. Genere un boxplot de legit.prot utilizando como variable de agrupación el Nivel Socioeconómico (NSEGrup)

Ejercicios 1.B

Para los siguientes ejercicios trabajaremos con los datos de la Encuesta Nacional sobre Familia y Roles de Género (GENERO.SAV) que realizó el Instituto de Opinión Pública de la PUCP en diciembre del 2012.

El primer paso será elaborar un índice de tolerancia a la homosexualidad (tolhomo) sobre la base de las preguntas P51D hasta la P51J que figuran en la base de datos y el cuestionario de la encuesta. Para elaborar el índice deberá recodificar las variables que lo componen de manera tal que se excluyan los valores perdidos.

La versión final del índice deberá medir la tolerancia a la homosexualidad en una escala de 0 a 100, donde 0 significa nula tolerancia y 100 alta tolerancia hacia la homosexualidad.

Una vez que haya calculado este índice, al momento de solicitar estadísticos de resumen, usted deberá tener los siguientes resultados:

```
> summary(genero$tolhomo)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   38.10   52.38   51.36   61.90  100.00     252
```

Para producir un gráfico de barras que muestre los niveles de tolerancia hacia la homosexualidad según el sexo del entrevistado deberá recodificar el índice en cuatro categorías:

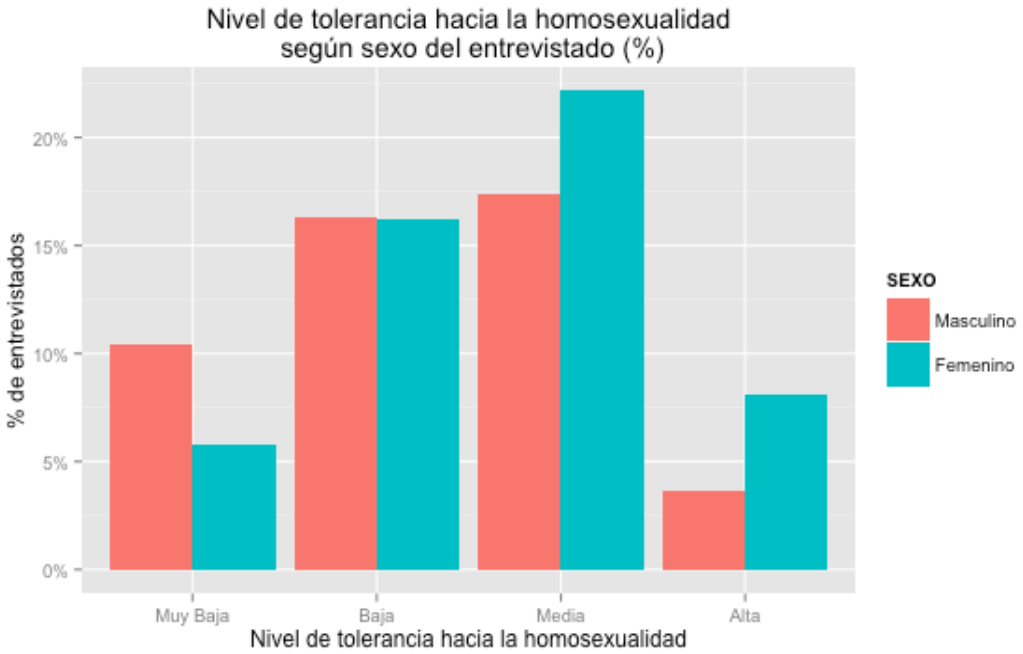
Categorías	Rango de valores
Muy baja	34 o menos
Baja	Entre 34 y 50
Media	Entre 50 y 67
Alta	Más de 67

La distribución de frecuencias de la variable agrupada en categorías debe ser la siguiente:

```
> table(genero$tol.g)
Muy Baja     Baja    Media     Alta
     154      309      376      112
```

1.B.1. Deberá crear un objeto en R llamado "grafico1". Al ser invocado este objeto deberá generar el siguiente gráfico:

>grafico

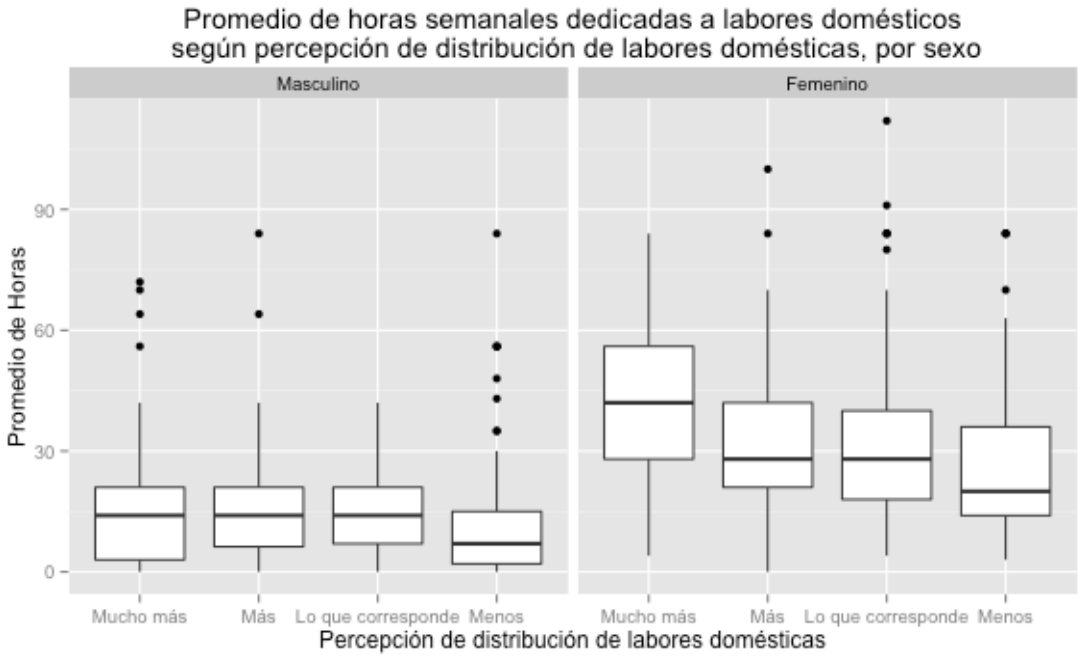


1.B.2. Deberá crear un objeto en R llamado “tabla1”. Al ser invocado este objeto deberá generar la siguiente tabla de estadísticos de tendencia central del índice de tolerancia hacia la homosexualidad:

```
>tabla1
##                Media  Desv.Std  NValid
## 18 a 29    54.54     19.43     339
## 30 a 44    51.10     17.37     309
## 45 a más 48.07     18.96     303
```

Luego de generar la tabla usted deberá darle un formato adecuado para ser incluida en un documento académico (títulos, rótulos, etc.)

1.B.3. Deberá crear un objeto en R llamado “grafico 2”. Al ser invocado este objeto deberá mostrar el siguiente gráfico.



Para producir el gráfico 2 primero es necesario recodificar la variable P32 (percepción de distribución de las labores domésticas en el hogar), agrupando las dos últimas categorías (“Hago menos de lo que me corresponde” y “Hago mucho menos de lo que me corresponde”) en una sola, además deberá asignar el código NA a los valores perdidos. Esta nueva variable recodificada deberá tener la siguiente distribución de frecuencias.

```
>library(descr)
>freq(p32r)
p32r
                    Frequency Percent Valid Percent
Mucho más                 147  12.219         22.21
Más                       141  11.721         21.30
Lo que corresponde        256  21.280         38.67
Menos                     118   9.809         17.82
NA's                      541  44.971
Total                    1203 100.000        100.00
```

En segundo lugar deberá recodificar la variable P19A (horas dedicadas al trabajo doméstico), para excluir los valores iguales o superiores a 140 asignándoles el código NA. Esta variable deberá tener los siguientes estadísticos descriptivos.

```
> summary(p19ar)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00    7.00   16.00   22.61   30.00  112.00       7
```

Finalmente para generar el gráfico se recomienda crear un data frame que contenga sólo las variables que figuran en el gráfico (para generar subconjuntos con selección de variables ver el video correspondiente) y donde se excluyan los NA, para esto último puede usar la función “na.omit”. Busque en la ayuda del R cómo funciona esta opción.

2. Inferencia estadística y prueba de hipótesis

Ejercicios 2.A: Muestras y estadísticos muestrales

Para desarrollar los ejercicios de esta sección deberá usar la base de datos "Trabajadores.sav" que contiene la información de un conjunto de empleados de una pequeá ciudad de la costa.

2.A.1. La base de datos salario tiene una variable "salariomes", que registra el salario mensual (en soles) que ganan los 541 trabajadores que viven en una pequeña ciudad de la costa. Pida un histograma de esta variable.
2.A.2. Para la variable "salariomes" pida los siguientes estadísticos descriptivos: mínimo, máximo, media y desviación estándar.
2.A.3. Extraer una muestra simple al azar (sin reemplazo) de 121 empleados. Luego, pedir los estadísticos descriptivos de dicha muestra (media y desviación estándar).
2.A.4. Calcular el error estándar de la muestra 1 y el intervalo de confianza al 95% de la misma.
2.A.5. Posteriormente proceda a realizar una muestra de 400 casos y calcule el error estándar y el intervalo de confianza de esta muestra. ¿Cuál es el error muestral de estas dos muestras? ¿Qué diferencias encuentra en los errores estándar e intervalos de confianza de ambas muestras?
2.A.6. La base de datos contiene la variable "segsalud" donde se registra la información respecto al seguro de salud al cual están afiliados los 514 trabajadores que estamos analizando. Pida una distribución de frecuencias para la variable "segsalud" en la base de datos original.
2.A.7. Pida las distribuciones de frecuencias para la misma variable en la base de datos que tienen las muestras de 121 y 400 trabajadores:
2.A.8. ¿De acuerdo con las muestra, cuál es la proporción de trabajadores que no están afiliados a un seguro de salud?
2.A.9. ¿Cuál es el error muestral del estadístico obtenido en esas muestras?
2.A.10. ¿Cuál es el error estándar de las muestras?, ¿podemos asumir que la distribución de muestreo es normal?
2.A.11. ¿Cuál es el intervalo de confianza al 95% de dichas muestras?

Ejercicios 2.B: Distribución de muestreo y cálculo de intervalos de confianza

En este ejercicio vamos a realizar 50 muestras diferentes de 200 trabajadores cada una. Para ello, utilice la siguiente función de muestro simple para que figura en la página 88 del texto de notas de clase del curso..

2.B.1. Anote los resultados de sus muestras: media y desviación estándar de la variable salario; proporción de trabajadores que no tienen seguro de salud (para anotar sus resultados puede usar una hoja de cálculo en Excel).
2.B.2. Calcule y registre el error estándar y el intervalo de confianza para cada uno de los estadísticos en las muestras tomadas.
2.B.3. Haga un histograma de los resultados de las 50 medias (puede usar el paquete ggplot2).
2.B.4. ¿Qué forma tiene la distribución de muestreo resultante?
2.B.5. ¿Cuántos de los intervalos de confianza incluyen al parámetro o proporción real de la población?

Ejercicios 2.C: Pruebas de hipótesis

Para estos ejercicios vamos a utilizar la base de datos "genero2" que se encuentra en el archivo "genero_v2.RData". Esta base de datos contiene un subconjunto de las variables de la encuesta sobre Familia y Roles de Género, realizada por el IOP en diciembre del 2012. Las variables en esta base de datos son:

- SEXO: Sexo del entrevistado
- NSEGrup: Nivel socioeconómico
- GEDAD:Grupo de edad
- Ambito: Ámbito de residencia
- p1: Edad ideal para que una mujer se case
- p2: Edad ideal para que un hombre se case
- p5 : Número ideal de hijos que una familia debería tener
- p19a: Horas que el entrevistado le dedica a labores domésticas en su hogar
- p29a: Horas que el cónyuge del entrevistado le dedica a labores domésticas en su hogar.
- p75r: Importancia de la religión en la vida del entrevistado (se considera sólo dos grupos: Muy importante; Poco o nada importante; los que consideraban la religión como "algo importante" han sido designados como NA – missing values).
- tolhomo: Escala de tolerancia hacia la homosexualidad (0 nada tolerante, 100 muy tolerante).

Para usar "genero2.RData" debe grabar el archivo en su directorio de trabajo y cargarlo al R usando el comando:

```
>load("genero_v2.Rdata)
```

Para el desarrollo de los siguientes ejercicios deberá utilizar la función "summarySE" que se encuentra en el paquete "Rmisc" del R para el cálculo del error estándar y los intervalos de confianza.

2.C.1. Calcule el error estándar del índice de tolerancia hacia la homosexualidad según niveles socioeconómicos.
2.C.2. Calcule el intervalo de confianza al 99% del porcentaje de personas que consideran la religión como poco o nada importante (variable p75r).
2.C.3. Para cada grupo de edad, calcule el intervalo de confianza al 95% para el número ideal de hijos que una familia debería tener (variable p5).
2.C.4. Un investigador quiere saber si para las mujeres que viven en zonas rurales, el número de hijos que tienen es menor al que sus propias madres tuvieron a su misma edad. Diseñe la prueba necesaria ¿Cuál de las afirmaciones que se muestran sería
2.C.5. Realice la prueba de hipótesis correspondiente para determinar si, en promedio, los hombres que viven en zonas rurales consideran que la edad ideal para que una mujer se case es mayor que 25 años.
2.C.6. Realice la prueba de hipótesis necesaria para determinar si el nivel de tolerancia hacia la homosexualidad es diferente entre las personas que consideran que la religión es muy importante en sus vidas y aquellas que consideran que la religión es poco o nada importante.
2.C.7. Realice la prueba de hipótesis necesaria para determinar si el número ideal de hijos que una familia debería tener es diferente entre las personas que consideran que la religión es muy importante en sus vidas y aquellas que consideran que la religión es poco o nada importante.
2.C.8. Para el caso de los hombres, realice la prueba de hipótesis correspondiente para comparar las horas dedicadas a labores domésticas (p19a) de ellos mismos y de sus cónyuges (p28a).
2.C.9. Replique la prueba anteriorpara el caso de las mujeres.

Ejercicios 2.D:

Para los ejercicios propuestos de esta sección usted deberá trabajar con la base de datos “medios.RData”, que contiene un subconjunto de variables de la encuesta panel sobre Comunicación y Opinión Pública, realizada por el IOP en el marco del proceso revocatorio en Lima 2013.[2] Para la primera parte de la práctica vamos a trabajar con las variables NSE (NSEGrupMarco) e intención de voto antes del proceso electoral revocatorio del 2013 (P7_Panel1) y después de este proceso (P7_Panel3).

2.D.1. Realice un contraste de proporciones que le permita conocer qué postura política apoyaron en mayor medida los participantes de la muestra del estudio Panel después del proceso electoral (P7_Panel3). Marque su respuesta.

A. El mayor porcentaje de la muestra apoyó la postura política que representaba la permanencia de la alcaldesa de Lima (I.C.= 59.33 +/- 4.337768)

[2] Para mayores detalles ver: http://iop-data.pucp.edu.pe/busqueda/encuesta/84?

B. El mayor porcentaje de la muestra apoyó la postura política que representaba la vacancia de la alcaldesa de Lima (I.C.= 59.33 +/- 4.337768)
C. El mayor porcentaje de la muestra apoyó la postura política que representaba la permanencia de la alcaldesa de Lima, pero estas diferencias no fueron significativas
D. El mayor porcentaje de la muestra apoyó la postura política que representaba la permanencia de la alcaldesa de Lima I.C.= 35.36 +/- 4.225731
E. Ninguna de las anteriores

2.D.2. Realice un contraste de proporciones que le permita conocer qué estrato socioeconómico (NSEGrupMarco) apoyó en mayor medida la postura política que representaba la vacancia de la alcaldesa de la ciudad de Lima antes del proceso electoral (P7_Panel1). Marque su respuesta.

A. Si bien el porcentaje de personas del nivel NSE A/B que votaron por la vacancia de la alcaldesa es mayor que el de las personas del NSE D/E, estas diferencias no son significativas.
B. El porcentaje de personas del nivel NSE D/E que votaron por la vacancia de la alcaldesa es significativamente mayor que el de las personas de NSE A/B.
C. El porcentaje de personas del nivel NSE A/B que votaron por la vacancia de la alcaldesa es significativamente mayor que el porcentaje de personas que apoyó la postura contraria.
D. Si bien el porcentaje de personas del nivel NSE D/E que votaron por la vacancia de la alcaldesa es mayor que el de las personas del NSE A/B, estas diferencias no son significativas.
E. Ninguna de las anteriores

2.D.3. Si un investigador busca conocer diferencias significativas en los niveles de cinismo político (también denominado confianza política: promedio de las variables) entre los estratos socioeconómicos alto y bajo, cómo debería plantear su hipótesis de investigación. Marque la respuesta es correcta.

A. H0: No existen diferencias significativas en el nivel de cinismo político entre el NSE A/B y D/E. H1: Existen diferencias significativas en el nivel de cinismo político entre el NSE A/B y D/E.
B. H0: Existen diferencias significativas en el nivel de cinismo político entre el NSE A/B y D/E. H1: No existen diferencias significativas en el nivel de cinismo político entre el NSE A/B y D/E.

C. H0: No existen diferencias significativas en el nivel de cinismo político entre el NSE A/B y D/E. H1: El nivel de cinismo político es significativamente más alto en el NSE A/B si se le compara con el NSE D/E.
D. H0: No existen diferencias significativas en el nivel de cinismo político entre el NSE alto y bajo. H1: El nivel de cinismo político es significativamente más alto en el NSE D/E si se le compara con el NSE A/B.
E. Ninguna de las anteriores

2.D.4. Realice un contraste estadístico que le permita conocer si existen diferencias significativas en los niveles de cinismo político entre los estratos socioeconómicos alto (A/B) y bajo (D/E) antes (P42E_Panel1, P42F_Panel1, P42H_Panel1) y después (P42E_Panel3, P42F_Panel3, P42H_Panel3) del proceso de revocatoria a la alcaldía de Lima. Marque su respuesta.

A. Antes del proceso electoral el nivel de cinismo político es significativamente más alto en las personas que provienen del NSE A/B.
B. Antes del proceso electoral no se observan diferencias significativas entre los grupos de contraste, sin embargo, después del proceso si existen diferencias significativas.
C. Ni antes ni después del proceso se observan diferencias significativas entre los grupos de contraste.
D. Antes y después del proceso electoral el nivel de cinismo político es más alto en el NSE D/E.
E. Ninguna de las anteriores

2.D.5. Si un investigador busca conocer si existen diferencias significativas en los niveles de eficacia política interna en el NSE D/E antes y después del proceso de revocatoria, qué tipo de contraste estadístico debería utilizar. Marque la respuesta correcta.

A. Prueba T Student para dos muestras independientes
B. Prueba T Student para una muestra única
C. Prueba T Student muestras relacionadas
D. Contraste de proporciones para dos muestras independientes
E. Contraste de proporciones para una muestra única

2.D.6. Realice los contrastes estadísticos adecuados para conocer si existen diferencias significativas en los niveles de eficacia política interna en el NSE D/E antes (P42A_Panel1, P42B_Panel1) y después (P42A_Panel3, P42B_Panel3) del proceso electoral. Marque su respuesta.

A. Antes del proceso electoral los niveles de eficacia política interna fueron más altos en el NSE D/E.
B. Los niveles de eficacia política interna antes y después del proceso electoral no difieren significativamente
C. Después del proceso electoral los niveles de eficacia política interna en el fueron más altos en el NSE D/E.
D. Si bien los niveles de eficacia política interna antes proceso electoral son más altos en el NSE D/E, estas diferencias no significativas.
E. Ninguna de las anteriores

2.D.7. Al igual que en el caso anterior, realice los contrastes estadísticos adecuados para conocer si existen diferencias significativas en los niveles de cinismo político según sea la postura política (P7_Panel1) que se apoyó antes y después del proceso de revocatoria. Marque su respuesta.

A. Antes del proceso electoral los simpatizantes del "Sí" mostraban niveles de cinismo político más elevados que los simpatizantes del "NO". Después del proceso estas diferencias no son significativas.
B. Ni antes ni después del proceso electoral los simpatizantes del "Sí" mostraban niveles de cinismo político más elevados que los simpatizantes del "NO".
C. Después del proceso electoral los niveles de cinismo político en los simpatizantes del "Sí" fueron significativamente más altos que en los simpatizantes del "No".
D. Después del proceso electoral los niveles de cinismo político en los simpatizantes del "Sí" fueron significativamente más altos que en los simpatizantes del "No"; sin embargo estas diferencias no son significativas.
E. Ninguna de las anteriores

3. Tablas de contingencia y pruebas chi-cuadrado

En esta parte se trabajará con la base de datos "ecm" que está dentro de la carpeta "bases" archivo. Para cargar la base de datos utilice el comando:

```
#load("ecm.rdata")
```

Esta base de datos contiene los resultados de la encuesta mundial de valores de la ronda 2010 – 2014 que se aplicó a 59 países en aquella oportunidad.

Los valores de las variables sólo están expresados en códigos numéricos, para conocer las etiquetas de las variables deberá consultar el libro de códigos y el cuestionario. Los valores perdidos de las variables de esta base de datos están originalmente codificados como números negativos. Recuerde que deberá asignarles el código NA en el caso de las variables que vamos a utilizar en este análisis.

El libro de códigos se encuentra en el archivo Excel y el cuestionario aplicado (en Inglés) en el archivo PDF que acompañan la base de datos en el archivo comprimido. Es altamente recomendable que estudie detenidamente ese material para poder trabajar con la base de datos

Para mayor información sobre la Encuesta Mundial de Valores, puede consultar la siguiente página web: http://www.worldvaluessurvey.org/

Sobre la escala de postmaterialismo, se recomienda revisar el artículo que Catalina Romero y David Sulmont publicaron en la revista Debates en Sociología, No. 25-26 del 2001, que puede consultarse en el siguiente enlace:

http://revistas.pucp.edu.pe/index.php/debatesensociologia/article/view/7077/7258

Utilice las técnicas indicadas para analizar la relación entre la versión corta de la escala de postmaterialismo y los grupos de edad. Haga el análisis para los siguientes países:

- Perú
- Chile
- Ecuador
- Brasil
- Argentina

Para su análisis, genere un subconjunto de datos para cada país antes de elaborar la tabla de contingencia correspondiente. La variable que identifica los países es la "V2". Para conocer el código de cada país revise el cuestionario.

Acondicionamiento de datos

Para proceder al análisis de los datos de la encuesta mundial de valores, primero deberá acondicionar los datos de la siguiente manera:

Paso 1: Deberá agrupar la variable edad del entrevistado (V242) en 5 grupos de edad, de manera tal que la distribución resultante sea la siguiente:[3]

```
> table(ecm$gedad)
## [16, 24] (24, 34] (34, 44] (44, 54] (54, 99]
## 14228   18990   16631   14215   20853
```

Paso 2: La variable "nivel educativo" (V248) deberá recodificarse como factor con 5 categorías (Menos que secundaria completa, secundaria completa, técnica, universitaria incompleta, universitaria completa), de manera que la distribución resultante sea la siguiente:[4]

```
> table(ecm$educ.r)
## Sec.  Sec. Comp.        Tec.  Univ. Inc. Univ. Comp.
## 24779       15810       21950        6781       14920
```

Paso 3: La versión corta de la escala de postmaterialismo (Y002) deberá excluir los valores perdidos y deberá recodificarse como factor de tal forma que su distribución sea la siguiente:

```
> table(ecm$post1)
 ## Materialista      Mixto   Postmaterialista
 ##     28867            44234             7106
```

Paso 3: La versión larga de la escala de postmaterialismo (Y001) deberá recodificarse como un vector numérico con valores del 0 al 5, pero excluyendo los valores perdidos, de forma tal que su distribución sea:

```
> table(ecm$post2)
##    0        1        2          3         4        5
## 8827 19892 25314 17617  5643  1121
```

Paso 4: Deberá crear 5 vectores lógicos (con valores TRUE / FALSE), para poder representar:[5]

a. Las personas que tienen educación universitaria completa
b. Las personas que son postmaterialistas según la versión corta de la escala
c. Las personas que consideran que la religión es muy importante importante en sus vidas (V9)

[3] Recomendación: Usar el comando "cut"
[4] Recomendación: Usar el comando "recode" del paquete "car"
[5] Recomendación: Seguir el ejemplo del ejercicio que se hizo con los datos del proyecto LAPOP en la última clase teórica del curso.

d. Las personas que consideran que el tiempo libre es muy importante en sus vidas (V6)
e. Las personas que consideran que enseñarle a los hijos la cualidad e auto expresión es importante (V22)

Con estos vectores lógicos usted deberá poder hacer lo siguiente:

```
> table(ecm$univ.c) # Para universitaria completa
## FALSE  TRUE
## 69320 14920
> table(ecm$postmat) # Para postmaterialistas
## FALSE  TRUE
## 73101  7106
> table(ecm$rel.imp) # Para V9
## FALSE  TRUE
## 42528 41267
> table(ecm$tlib.imp) # Para V6
## FALSE  TRUE
## 53046 31071
> table(ecm$self.exp) # Para V22
## FALSE  TRUE
## 62727 22341
> mean(ecm$univ.c, na.rm=T)*100
## [1] 17.7113
> mean(ecm$postmat, na.rm=T)*100
## [1] 8.859576
> mean(ecm$rel.imp, na.rm=T)*100
## [1] 49.24757
> mean(ecm$tlib.imp, na.rm=T)*100
## [1] 36.93784
> mean(ecm$self.exp, na.rm=T)*100
## [1] 26.26252
```

Ejercicios de la sección 3

3.1. Utilizando la escala corta de postmaterialismo, indique: ¿qué porcentaje de los jóvenes menores de 24 años son postmaterialistas en PERU?
3.2. Utilizando la escala corta de postmaterialismo, indique: ¿qué porcentaje de los jóvenes menores de 24 años son postmaterialistas en CHILE?
3.3. Utilizando la escala corta de postmaterialismo, indique: ¿qué porcentaje de los jóvenes menores de 24 años son postmaterialistas en ECUADOR?
3.4. Utilizando la escala corta de postmaterialismo, indique: ¿qué porcentaje de los jóvenes menores de 24 años son postmaterialistas en BRASIL?
3.5. Utilizando la escala corta de postmaterialismo, indique: ¿qué porcentaje de los jóvenes menores de 24 años son postmaterialistas en ARGENTINA?
3.6. Considerando un nivel de significancia del 5%, el resultado de la prueba de Chi Cuadrado para la tabla de contingencia entre Grupo de Edad y Escala de

Postmaterialismo (versión corta) en el PERU ¿nos permite sostener que entre ambas variables existen una asociación estadísticamente significativa?

3.7. Considerando un nivel de significancia del 5%, el resultado de la prueba de Chi Cuadrado para la tabla de contingencia entre Grupo de Edad y Escala de Postmaterialismo (versión corta) en CHILE ¿nos permite sostener que entre ambas variables existen una asociación estadísticamente significativa?

3.8. Considerando un nivel de significancia del 5%, el resultado de la prueba de Chi Cuadrado para la tabla de contingencia entre Grupo de Edad y Escala de Postmaterialismo (versión corta) en ECUADOR ¿nos permite sostener que entre ambas variables existen una asociación estadísticamente significativa?

3.9. Considerando un nivel de significancia del 5%, el resultado de la prueba de Chi Cuadrado para la tabla de contingencia entre Grupo de Edad y Escala de Postmaterialismo (versión corta) en BRASIL ¿nos permite sostener que entre ambas variables existen una asociación estadísticamente significativa?

3.10. Considerando un nivel de significancia del 5%, el resultado de la prueba de Chi Cuadrado para la tabla de contingencia entre Grupo de Edad y Escala de Postmaterialismo (versión corta) en ARGENTINA ¿nos permite sostener que entre ambas variables existen una asociación estadísticamente significativa?

3.11. Considerando el nivel de medición de ambas variables, analice el coeficiente de asociación más apropiado para medir la relación entre el grupo de edad y la escala de postmaterialismo (versión corta) y responda: ¿en qué país diría usted que la relación es más fuerte?

4. Análisis de varianza

Ejemplos

Se trabajará con los datos de la encuesta de Roles de Género y Familia 2012 del IOP-PUCP

```
library(foreign)
genero <- as.data.frame(read.spss("IOP_1212_01_B.sav"))
```

Vamos a analizar los factores asociados a las horas que le dedican los entrevistados al trabajo doméstico por semana. Primero preparamos la variable dependiente retirando algunos casos atípicos:

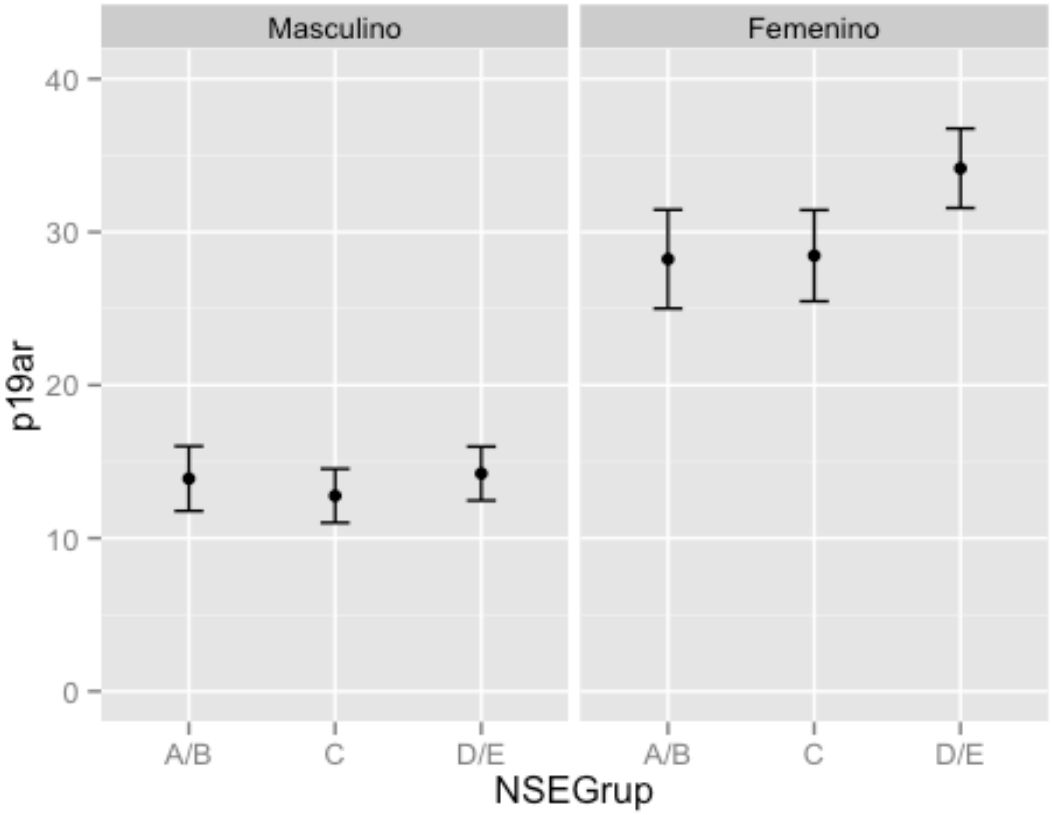
```
genero$p19ar <- genero$P19A
genero$p19ar[genero$P19A > 112] <- NA
```

¿Las diferencias en la dedicación a labores domésticas en el propio hogar según NSE son estadísticamente significativas según el nivel socioeconómico? Como sabemos que el sexo del entrevistado es un factor que provoca muchas diferencias, vamos a diferenciar el análisis por género. Primero veamos los estadísticos descriptivos:

```
library(Rmisc)
## Loading required package: lattice
## Loading required package: plyr
data <- genero
library(Rmisc)
est.des <- summarySE(data, measurevar="p19ar", groupvars=c("NSEGrup",
"SEXO"), na.rm=T)
est.des
##   NSEGrup      SEXO   N    p19ar       sd        se       ci
## 1     A/B Masculino 124 13.89516 11.91898 1.0703560 2.118704
## 2     A/B  Femenino 138 28.23188 19.18838 1.6334236 3.229983
## 3       C Masculino 183 12.77049 12.08540 0.8933786 1.762711
## 4       C  Femenino 173 28.45665 19.87445 1.5110264 2.982543
## 5     D/E Masculino 280 14.22500 14.97583 0.8949768 1.761765
## 6     D/E  Femenino 298 34.16107 22.74987 1.3178644 2.593535
```

Hagamos un gráfico de las medias con sus respectivos intervalos de confianza al 95% para observar mejor esas diferencias:

```
library(ggplot2)
grafico <- ggplot(est.des, aes(x=NSEGrup, y=p19ar)) + geom_point() + y
lim(0, 40) +
  geom_errorbar(aes(ymin=p19ar-ci, ymax=p19ar+ci), width = 0.2) +
  facet_grid(.~SEXO)
grafico
```



Pruebas de ANOVA: Para el caso de las mujeres

Generamos la tabla de ANOVA de horas semanales dedicadas al trabajo doméstico según NSE, para el caso de las mujeres:

```
data <- subset(genero, SEXO=="Femenino")
anova <- aov(data$p19ar~data$NSEGrup)
summary(anova)

##                Df Sum Sq Mean Sq F value  Pr(>F)
## data$NSEGrup    2   5131    2565   5.713 0.00348 **
## Residuals     606 272096     449
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 5 observations deleted due to missingness
```

Si tomamos en cuenta un $\alpha = 0.05$ y lo comparamos con la significancia del estadístico F de la prueba (Pr(>F)), podemos rechazar la hipótesis cero que sostiene que las medias de los grupos son iguales.

Prueba de comparaciones múltiples de Tukey

Al rechazar la H0, podemos proceder a una prueba Post-Hoc para identificar entre qué grupos existen diferencias estadísticamente significativa. Para ello emplearemos la prueba de Diferencias Significativas de Tukey (TukeyHSD):

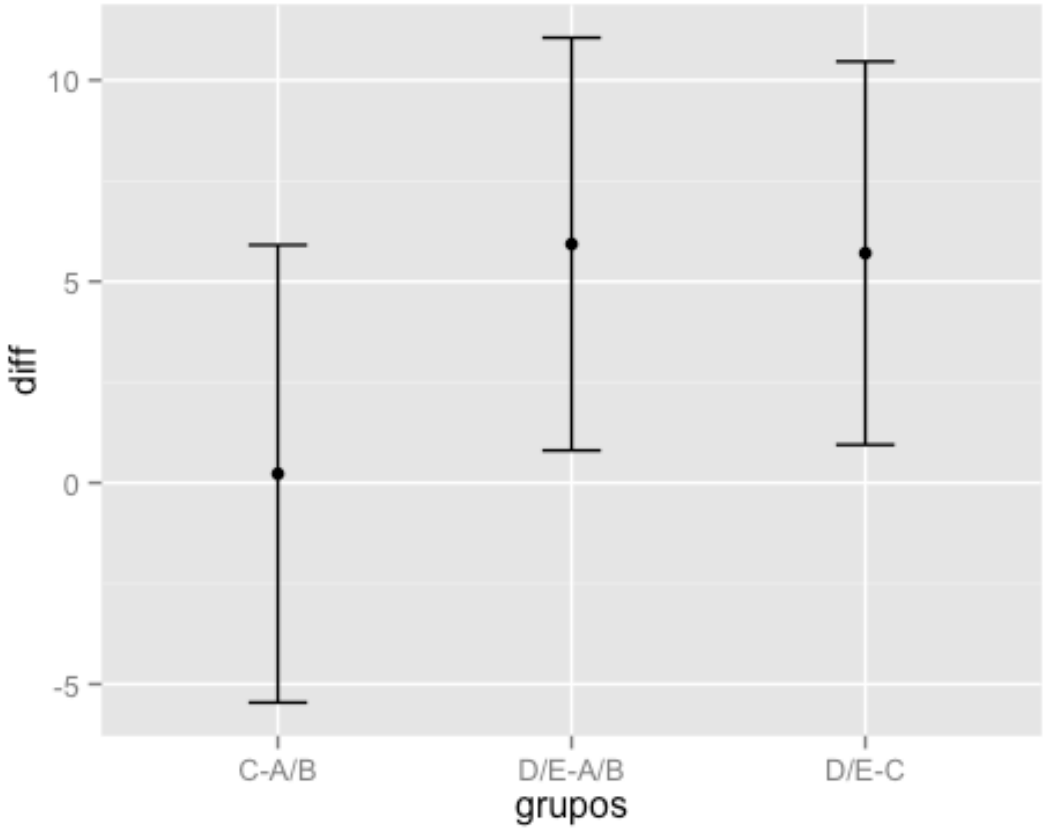
```
TukeyHSD(anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = data$p19ar ~ data$NSEGrup)
##
## $`data$NSEGrup`
##               diff        lwr       upr     p adj
## C-A/B    0.2247633 -5.4574106  5.906937 0.9952496
## D/E-A/B 5.9291898  0.8030258 11.055354 0.0185326
## D/E-C    5.7044264  0.9458593 10.462994 0.0138709
```

El test de comparaciones múltiples nos muestra el intervalo de confianza al 95% de las diferencias de medias entre los tres pares de comparaciones. La "probabilidad ajustada" (p adj) nos ayudará a decidir si aceptamos o rechazamos la H0 de que las diferencias no son estadísticamente significativas. Si tomamos en cuenta un $\alpha = 0.05$ concluimos que, en el caso de las mujeres el grupo que se distingue de los demás es el del NSE D/E: en este NSE las mujeres trabajan significativamente más horas en labores domésticas que en los otros dos. No se encuentran diferencias estadísticamente significativas entre los NSE A/B vs C.

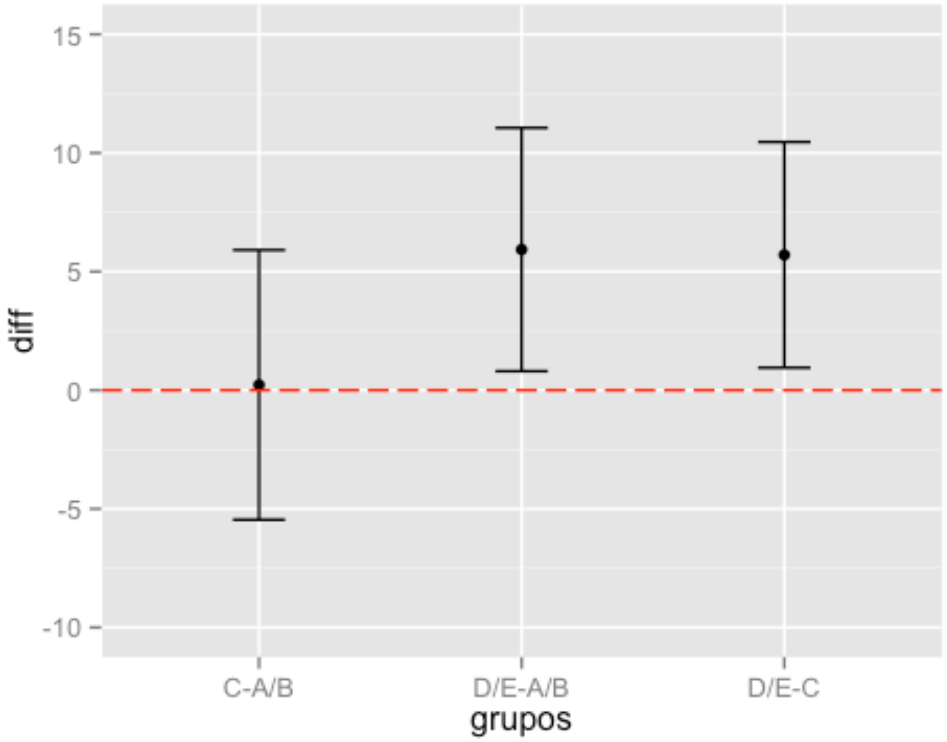
Para mostrar nuestras conclusiones, podemos graficar los intervalos de confianza de las diferencias de medias entre los grupos:

```
tab <- TukeyHSD(anova) # se crea un objeto tipo lista a partir del res
ultado de Tukey
tab.dif <- as.data.frame(tab[[1]]) # se selecciona la tabla dentro de
esa lista
tab.dif$grupos <- row.names(tab.dif) # se añade una columna con los no
mbres

graf.dif <- ggplot(tab.dif, aes(x=grupos, y = diff)) + geom_point() +
  geom_errorbar(aes(ymin=lwr, ymax=upr), width=0.2)
graf.dif
```



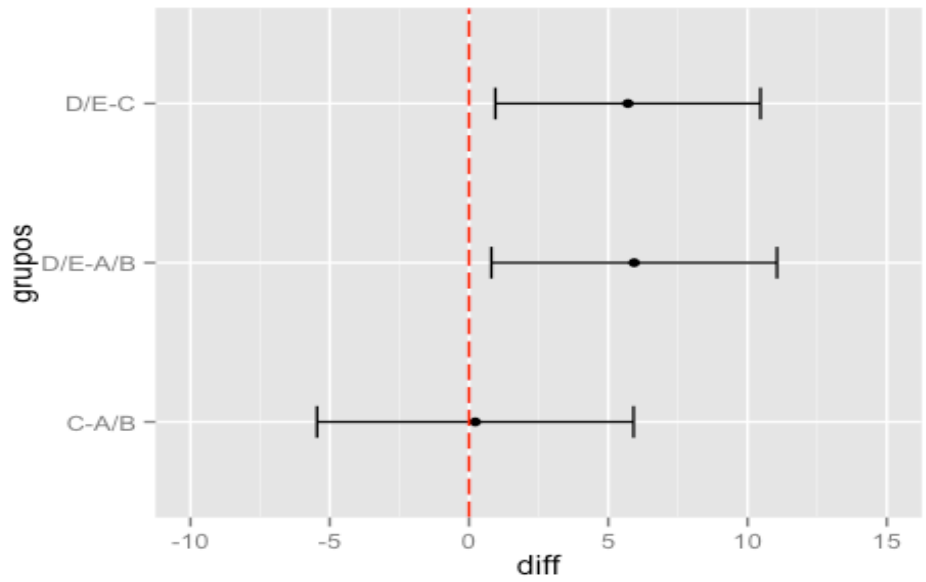
Mejoramos un poco el gráfico:

```
graf.dif <- graf.dif + ylim(-10, 15) +
  geom_hline(yintercept = 0, col="red", linetype = "longdash")
graf.dif
```



Le damos la vuelta al gráfico:

```
graf.dif + coord_flip()
```



Pruebas de ANOVA: Para el caso de los hombres

Generamos la tabla de ANOVA de horas semanales dedicadas al trabajo doméstico según NSE, para el caso de los hombres:

```
data <- subset(genero, SEXO=="Masculino")
anova <- aov(data$p19ar~data$NSEGrup)
summary(anova)

##                Df Sum Sq Mean Sq F value Pr(>F)
## data$NSEGrup    2    240   120.0   0.657  0.519
## Residuals     584 106629   182.6
## 2 observations deleted due to missingness
```

En este caso, el p-value (Pr(>F)) del estadístico de la prueba es mayor que el nivel de significancia, lo que nos lleva a aceptar la H0. Por lo tanto no se puede decir que en el caso de los hombres existan diferencias estadísticamente significativas en las horas dedicadas a labores domésticas entre diferentes NSE.

Ejercicios 4.A

Proceda a realizar el mismo tipo de comparaciones que mostramos en los ejemplos anteriores. Para cada género, compare las horas dedicadas a labores domésticas entre:

- Grupos de Edad (GEDAD)
- Ámbito de estudio (Ambito)

Ejercicios 4.B

Algunos estudios sugieren que hay una asociación entre la cantidad de horas que los hombres le dedican a labores domésticas y la satisfacción con la vida familiar: los hombres más satisfechos son aquellos que colaboran más con labores domésticas (o colaborar con labores domésticas puede tener sus "recompensas"). ¿Los datos de esta encuesta son consistentes con esa teoría?

Primero recodificaremos la variable P40 de la siguiente manera:

```
library(descr)
freq(ordered(genero$P40), plot=FALSE)

## ordered(genero$P40)
##                                      Frequency   Percent Cum Percent
## Me siento completamente satisfecho/a       161  13.38321       13.38
## Muy satisfecho/a                           406  33.74896       47.13
## Bastante satisfecho/a                      409  33.99834       81.13
## Ni satisfecho ni insatisfecho/a            175  14.54697       95.68
## Bastante insatisfecho/a                     25   2.07814       97.76
## Muy insatisfecho/a                          11   0.91438       98.67
## Completamente insatisfecho/a                 4   0.33250       99.00
## No sabe                                      1   0.08313       99.09
## No contesta                                 11   0.91438      100.00
## Total                                     1203 100.00000

satisf.fam <- as.numeric(genero$P40)
library(car)
satisf.fam <- recode(satisf.fam, "6:7 = 5; 8:9=NA")
satisf.fam <- factor(satisf.fam)
levels(satisf.fam) <- c("Comp. Satisf","Muy Satisf.", "Bast. Satisf.",
"Ni sat. ni insat.", "Insatif.")
genero$satisf.fam <- satisf.fam
freq(ordered(genero$satisf.fam), plot=FALSE)

## ordered(genero$satisf.fam)
##                    Frequency  Percent Valid Percent Cum Percent
## Comp. Satisf             161  13.3832        13.518       13.52
## Muy Satisf.              406  33.7490        34.089       47.61
## Bast. Satisf.            409  33.9983        34.341       81.95
## Ni sat. ni insat.        175  14.5470        14.694       96.64
## Insatif.                  40   3.3250         3.359      100.00
## NA's                      12   0.9975
## Total                   1203 100.0000       100.000
```

Con la nueva variable proceda a generar el gráfico de medias con sus respectivos intervalos de confianza; la prueba de ANOVA; si corresponde, la prueba de Tukey.

Ejercicios 4.C:

En esta parte se trabajará con la base de datos "ecm" que está dentro de la carpeta "bases" archivo.

4.C.1. Realice una prueba de ANOVA con sus correspondientes pruebas de comparaciones múltiples (TukeyHSD) para evaluar si existen diferencias estadísticamente significativas entre las medias de la versión larga de la escala de postmaterialismo según los diferentes niveles educativos del entrevistado. Haga su análisis por separado para cada uno de los países que se mencionaron en la sección 3 de este cuaderno de ejercicios

4.C.2. De los 5 países de América Latina que estamos considerando, a cuál de ellos corresponde el siguiente valor de F para la prueba de ANOVA que considera la

escala de postmaterialismo (versión larga) como variable dependiente y el nivel educativo como factor. F=6.433.

4.C.3. Luego de la prueba de ANOVA, realice una prueba de Tukey para determinar entre qué grupos del factor "Nivel Educativo" existen diferencias estadísticamente significativas en la media de la escala de postmaterialismo (versión larga) en PERU. Luego de evaluar los resultados de las comparaciones y considerando un alpha de 0.05, ¿diría usted que la media del grupo de personas con educación universitaria completa es significativamente mayor que la del grupo de personas con educación secundaria completa?

4.C.4. Luego de la prueba de ANOVA, realice una prueba de Tukey para determinar entre qué grupos del factor "Nivel Educativo" existen diferencias estadísticamente significativas en la media de la escala de postmaterialismo (versión larga) en CHILE. Luego de evaluar los resultados de las comparaciones y considerando un alpha de 0.05, ¿diría usted que la media del grupo de personas con educación universitaria completa es significativamente mayor que la del grupo de personas con educación secundaria completa?

4.C.5. Luego de la prueba de ANOVA, realice una prueba de Tukey para determinar entre qué grupos del factor "Nivel Educativo" existen diferencias estadísticamente significativas en la media de la escala de postmaterialismo (versión larga) en ECUADOR. Luego de evaluar los resultados de las comparaciones y considerando un alpha de 0.05, ¿diría usted que la media del grupo de personas con educación universitaria completa es significativamente mayor que la del grupo de personas con educación secundaria completa?

4.C.6. Luego de la prueba de ANOVA, realice una prueba de Tukey para determinar entre qué grupos del factor "Nivel Educativo" existen diferencias estadísticamente significativas en la media de la escala de postmaterialismo (versión larga) en BRASIL. Luego de evaluar los resultados de las comparaciones y considerando un alpha de 0.05, ¿diría usted que la media del grupo de personas con educación universitaria completa es significativamente mayor que la del grupo de personas con educación secundaria completa?

4.C.7. Luego de la prueba de ANOVA, realice una prueba de Tukey para determinar entre qué grupos del factor "Nivel Educativo" existen diferencias estadísticamente significativas en la media de la escala de postmaterialismo (versión larga) en ARGENTINA. Luego de evaluar los resultados de las comparaciones y considerando un alpha de 0.05, ¿diría usted que la media del grupo de personas con educación universitaria completa es significativamente mayor que la del grupo de personas con educación secundaria completa?

4.C.8. Considerando las pruebas de comparaciones múltiples de Tukey, ¿en cuál de los siguientes países la diferencia de medias en la escala de postmaterialismo (versión larga) entre el grupo de universitaria completa y el de secundaria completa es la más importante?

5. Correlación y regresión

Ejemplo1: Regresión simple

Para la realización de estos ejercicios, utilizaremos las bases de datos de indicadores sociodemográficos de países del mundo 1998 t 2005. Los datos pueden descargarse desde:

https://sites.google.com/a/pucp.pe/data_est/archivos/BD_Mundo.zip?attredirects=0&d=1

Revise el libro de códigos que viene con las bases de datos. Descomprima y guarde los archivos en su directorio de trabajo de R y proceda a cargarlos:

```
load("mundo98.rda")
load("mundo2005.rda")
```

Además necesitará instalar y cargar los siguientes paquetes de R:

```
library(ggplot2)
library(grid)
library(scales)
library(Hmisc)
```

OJO: Observen que en el caso de mundo2005, tenemos la variable "alfab_f" que es % de alfabetismo femenino, mientras que en mundo98 tenemos la variable "illiteracyFemale" que es % de analfabetismo femenino. Tomen en cuenta estas diferencias al momento de interpretar los resultados de los modelos de regresión.

```
Verifique los nombres de las variables en las bases de datos:

names(mundo98)

##  [1] "region"                 "tfr"
##  [3] "contraception"          "educationMale"
##  [5] "educationFemale"        "lifeMale"
##  [7] "lifeFemale"             "infantMortality"
##  [9] "GDPperCapita"           "economicActivityMale"
## [11] "economicActivityFemale" "illiteracyMale"
## [13] "illiteracyFemale"

names(mundo2005)

##  [1] "Region"      "Pob"         "tgf_75"      "tgf_05"      "antic
onc"
##  [6] "mortInf"     "pbi"         "evida_masc"  "evida_f"     "alfab
_f"
## [11] "alfab_masc"  "matric_fem"  "matric_masc" "pea_fem"     "pea_f
em2"
## [16] "pbiPc05"
```

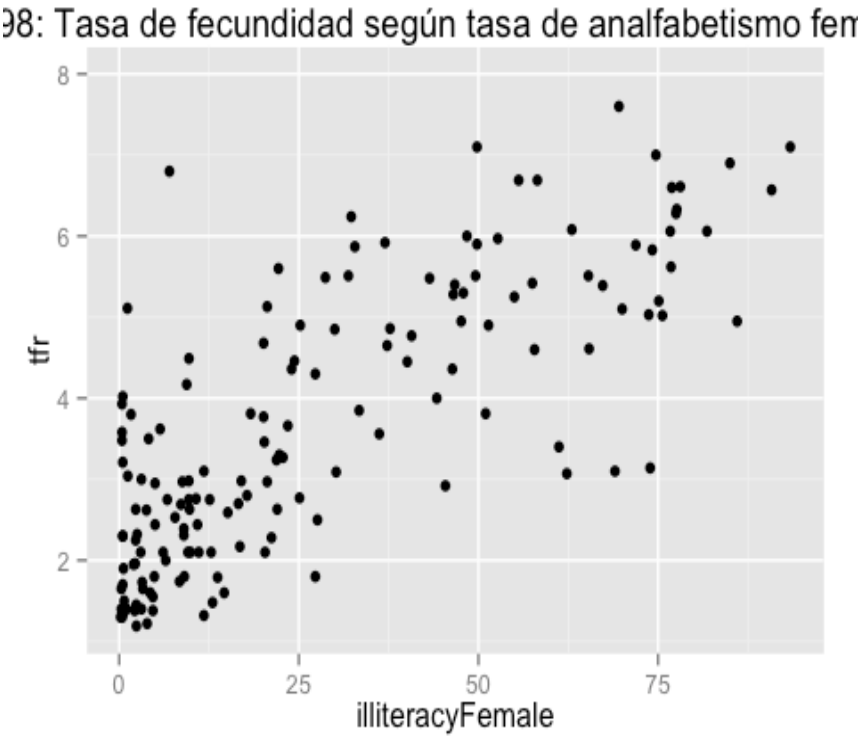
Factores asociados a la fertilidad: 1998

Con los datos de mundo98 vamos a calcular un modelo de regresión simple considerando a la tasa global de fecundidad (nacimientos por mujer, variable "trf") como variable dependiente y el % de analfabetismo femenino como variable independiente (illiteracuFemale)

Diagrama de dispersión

```
disp.1 <- ggplot(mundo98, aes(x=illiteracyFemale, y=tfr)) + geom_point
() +
  ggtitle("1998: Tasa de fecundidad según tasa de analfabetismo femeni
no")
disp.1
```

```
## Warning: Removed 50 rows containing missing values (geom_point).
```

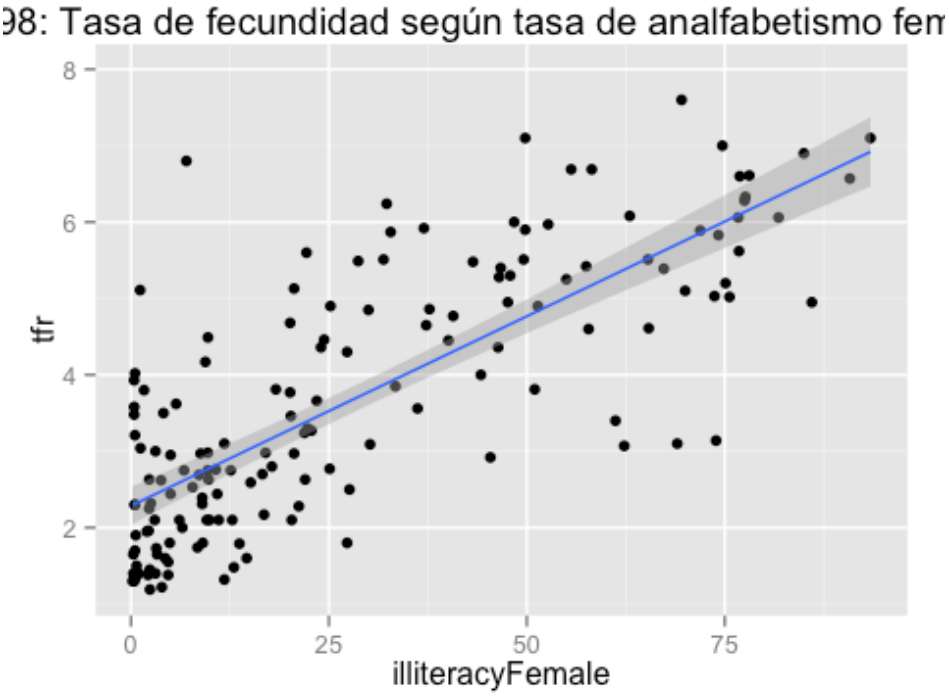


Modelo de regresión

```
modelo1 <- lm(tfr~illiteracyFemale, data=mundo98)
summary(modelo1)

##
## Call:
## lm(formula = tfr ~ illiteracyFemale, data = mundo98)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8126 -0.7613 -0.1325  0.6595  4.1668
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.285893   0.125278   18.25   <2e-16 ***
## illiteracyFemale 0.049618   0.003246   15.29   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.08 on 155 degrees of freedom
##   (50 observations deleted due to missingness)
## Multiple R-squared:  0.6012, Adjusted R-squared:  0.5986
## F-statistic: 233.7 on 1 and 155 DF,  p-value: < 2.2e-16
```

Diagrama de dispersión con recta de regresión e intervalo de confianza al 95%

```
disp.1 + geom_smooth(method=lm)

## Warning: Removed 50 rows containing missing values (stat_smooth).

## Warning: Removed 50 rows containing missing values (geom_point).
```



Ejercicios 5A:

Vamos a calcular 4 modelos de regresión simple considerando a la tasa global de fecundidad (nacimientos por mujer, variable "trf") como variable dependiente y a las siguientes variables como variables independientes:

- illiteracyFemale: % de Mujeres de 15 años o más que son analfabetos
- contraception: Tasa de prevalencia anticonceptiva (% de mujeres en edad fértil que utilizan (o cuya pareja utiliza) un método anticonceptivo en un determinado momento)
- economicActivityFemale: PEA Femenina (% de Mujeres económicamente activas)
- GDPperCapita: PBI per cápita en US$

Para cada modelo:

- Elaborar y analizar el diagrama de dispersión correspondiente
- Calcular e interpretar los coeficientes de regresión (b_0, b_1, R^2)
- Dibujar la línea de regresión en el diagrama de dispersión.
- Evaluar la significancia de los coeficientes de regresión y analizar el error estándar del modelo

Luego de realizar los cálculos vamos a comparar los resultados de los modelos, tratando de responder a las preguntas, ¿cuál de las variables independientes tiene un

efecto más importante en la fecundidad?; ¿qué modelo tiene un error más pequeño en la predicción del valor de la variable dependiente?; ¿qué tan bien representa la línea calculada por el modelo de regresión a la relación entre las variables el diagrama de dispersión?

Ejercicios 5B:

Repita los ejercicios anteriores, esta vez con los datos del 2005 y las variables equivalente. Compare los modelos entre ambos años. ¿Qué diferencias observa entre ellos?

Ejemplo 2: Matriz de correlaciones

En esta parte vamos a examinar las correlaciones entre algunos indicadores sobre la situación de las mujeres, específicamente:

- Esperanza de vida al nacer de las mujeres
- Tasa de analfabetismo / alfabetismo
- Participación de la mujer en la PEA
- PBI per cápita

Para ello utizaremos como herramienta una matriz de correlaciones, utilizando la base de datos mundo98

En primer lugar vamos a generar un subconjunto de datos que contenga únicamente las cuatro variables que serán analizadas y los países con información completa para todas ellas (excluiremos los "NA"). Ello nos dará como resultado un data frame con 4 variables y 127 países.

```
misvars98 <- c("lifeFemale", "illiteracyFemale", "economicActivityFema
le", "GDPperCapita")
data98 <- na.omit(mundo98[misvars98])
length(row.names(data98)) # para ver el número de casos del data frame

## [1] 127
```

Seguidamente utilizaremos la función "rcorr" del paquete "Hmisc" para generar la matriz de correlaciones y los p-value de las correspondientes pruebas de significancia estadística.

```
rcorr(as.matrix(data98))

##                          lifeFemale illiteracyFemale economicActivity
Female
## lifeFemale                     1.00            -0.79
-0.11
## illiteracyFemale             -0.79             1.00
-0.13
## economicActivityFemale       -0.11            -0.13
1.00
## GDPperCapita                  0.52            -0.32
-0.06
##                          GDPperCapita
## lifeFemale                       0.52
## illiteracyFemale                -0.32
## economicActivityFemale          -0.06
## GDPperCapita                     1.00
##
## n= 127
##
##
## P
##                          lifeFemale illiteracyFemale economicActivity
Female
## lifeFemale                           0.0000           0.2097
## illiteracyFemale         0.0000                       0.1427
## economicActivityFemale   0.2097      0.1427
## GDPperCapita             0.0000      0.0002           0.4840
##                          GDPperCapita
## lifeFemale               0.0000
## illiteracyFemale         0.0002
## economicActivityFemale   0.4840
## GDPperCapita
```

¿Cómo están correlacionadas estas variables?, ¿en qué casos las relaciones son:

- Más fuertes?
- Directas o inversas?
- Estadísticamente significativas?

Ejercicios 5C:

Replique el análisis precedente con las variables correspondientes a "mundo2005". Compare los resultados de ambos años.

Ejercicios 5D:

Con la misma base dela encuesta mundial de valores (ecm), genere una base de datos con los siguientes indicadores agregados a nivel de país de las siguientes variables:

- % de personas que son postmaterialistas (acorde con la versión corta de la escala)
- % de personas con educación universitaria completa

- % de personas que consideran que la religión es muy importante en sus vidas
- % de personas que consideran que el tiempo libre es muy importante en sus vidas
- % de personas que consideran enseñar la autoexpresión a los niños

Para ello deberá usar la función data.table :

```
dt <- data.table(ecm)

dt.out <- dt[, list(postmat  = mean(postmat, na.rm=T)*100, univ.c = me
an(univ.c, na.rm=T)*100, rel.imp = mean(rel.imp, na.rm=T)*100, tlib.im
p = mean(tlib.imp, na.rm=T)*100, self.exp = mean(self.exp, na.rm=T)*10
0), by = c("V2")]

dt.out
```

La base de datos agregada final tiene 59 países y deberá verse se la siguiente manera[6] (aquí se muestran los 6 primeros registros):

```
> head(dt.out)

V2   postmat    univ.c  rel.imp tlib.imp self.exp

## 1: 12  6.697248  9.758132 91.19866 38.40948 17.75000

## 2: 32 10.805301  5.631068 22.82609 36.91406 86.11650

## 3: 51  3.732057 25.683060 59.78162 28.21229 20.27273

## 4: 36 22.938323 40.684411 17.04225 42.43490 35.40961

## 5: 31  4.100000 34.530938 36.92615 24.35130 53.19361

## 6: 48 21.116928 11.166667 40.65109 21.91667 12.83333
```

Resolver

5.D.1. Deberá realizar sendos gráficos de dispersión con la variable “postmat” como variable dependiente (en el eje Y) y las demás variables como variables independientes.

5.D.2. Deberá calcular una matriz de correlación entre las variables de esta base de datos (se recomienda usar la función “rcorr” del paquete “Hmisc”).

5.D.3. Deberá estimar e interpretar tres modelos de regresión:

- Modelo 1: postmat según rel.imp
- Modelo 2: rel.imp según univ.c
- Modelo 3: rel.imp según self.exp

[6] La variable V2 corresponde al código del país

5.D.4. Tomando en cuenta el % de personas que consideran a la religión como muy importante en nuestro país ¿qué porcentaje de personas se esperaría que sean postmaterialistas en el Perú?

5.D.5. Tomando en cuenta el % de personas que tienen educación universitaria completa en nuestro país ¿qué porcentaje de personas se esperaría que le den mucha importancia a la religión en sus vidas en el Perú?

5.D.6. Tomando en cuenta el % de personas que consideran que enseñarles a auto expresarse a los niños es importante para su educación en nuestro país ¿qué porcentaje de personas se esperaría que le den mucha importancia a la religión en sus vidas en el Perú?

5.D.7. De los modelos elaborados, ¿cuál de ellos tiene mayor "poder predictivo" de la variable dependiente?