R meets psychology

Sesión 1. Análisis Univariado

El ejercicio desarrollado en este material se ha preparado haciendo uso de la base de datos “Familia y roles de género 2012”, del IOP, que se encuentra para descargar en el siguiente enlace.

<http://iop-data.pucp.edu.pe/busqueda/encuesta/71>?

El IOP Data tiene varias bases de datos que puedes revisar, descargar, y desarrollar con esta plantilla.

Herramientas para el análisis descriptivo univariable

* Tablas de distribución de frecuencias
* Gráficos
* Resúmenes numéricos
  + Estadísticos de orden: Cuantiles y percentiles
  + Estadísticos de tendencia central: Moda, Media, Mediana
  + Estadísticos de dispersión: Rango, Varianza, Desviación Estándar, Rango intercuartil.

Tablas de distribución de frecuencias

Tablas de distribución de frecuencias

Descomprimir y grabar el archivo SPSS en el directorio de trabajo de R

# Importar la base de datos del SPSS a un data frame de R  
library(foreign)  
genero <- as.data.frame(read.spss("IOP\_1212\_01\_B.sav"))

## re-encoding from UTF-8

Tabla de frecuencias simple

# Una tabla de distribución de frecuencias del sexo del entrevistado:  
table(genero$SEXO)

##   
## Masculino Femenino   
## 589 614

Tabla de frecuencias relativas

prop.table(table(genero$SEXO)) # tabla en proporciones

##   
## Masculino Femenino   
## 0.4896 0.5104

prop.table(table(genero$SEXO))\*100 # tabla en porcentajes

##   
## Masculino Femenino   
## 48.96 51.04

Guardar tablas en objetos

tabla.1a <- table(genero$SEXO)  
tabla.1b <- prop.table(tabla.1a)\*100  
tabla.1a

##   
## Masculino Femenino   
## 589 614

tabla.1b

##   
## Masculino Femenino   
## 48.96 51.04

Redondear

Redondear o especificar espacios decimales (que sean significativos)

round(tabla.1b, digits = 0) # Para redondear a la unidad

##   
## Masculino Femenino   
## 49 51

Recodificar una variable y hacer una tabla

table(genero$P3A)

##   
## Muy de acuerdo De acuerdo   
## 67 377   
## En desacuerdo Muy en desacuerdo   
## 569 99   
## Ni de acuerdo ni en desacuerdo No sabe   
## 63 22   
## No contesta   
## 6

levels(genero$P3A)

## [1] "Muy de acuerdo" "De acuerdo"   
## [3] "En desacuerdo" "Muy en desacuerdo"   
## [5] "Ni de acuerdo ni en desacuerdo" "No sabe"   
## [7] "No contesta"

Recodificamos

library(car) # Cargar el paquete "car"  
p3a.t <- as.numeric(genero$P3A) # Creamos un objeto temporal  
table(p3a.t)

## p3a.t  
## 1 2 3 4 5 6 7   
## 67 377 569 99 63 22 6

genero$P3A.r <- recode(p3a.t, "3=4; 4=5; 5=3; 6:7=NA") # recodificamos  
table(genero$P3A.r)

##   
## 1 2 3 4 5   
## 67 377 63 569 99

Convertimos la variable recodificada en un factor y le asignamos niveles (etiquetas)

genero$P3A.r <- factor(genero$P3A.r)  
levels(genero$P3A.r) <- c("Muy de acuerdo", "De acuerdo",   
 "Ni de acuerdo ni en desacuerdo",   
 "En desacuerdo", "Muy en desacuerdo")  
tabla.2 <- round(prop.table(table(genero$P3A.r))\*100, digits = 2)  
tabla.2

##   
## Muy de acuerdo De acuerdo   
## 5.70 32.09   
## Ni de acuerdo ni en desacuerdo En desacuerdo   
## 5.36 48.43   
## Muy en desacuerdo   
## 8.43

Tablas parecidas al SPSS

Si queremos hacer tablas parecidas al SPSS podemos usar la función "freq" del paquete "descr" (requiere que el paquete esté instalado en la librería del R)

library(descr)  
tabla.3 <- freq(genero$SEXO, plot = FALSE)  
tabla.3

## genero$SEXO   
## Frequency Percent  
## Masculino 589 48.96  
## Femenino 614 51.04  
## Total 1203 100.00

Tablas parecidas al SPSS

Cuando la tabla corresponde a una variable ordinal, podemos incluir las frecuencias acumuladas, indicando que estamos trabajando con un factor ordenado:

tabla.4 <- freq(ordered(genero$P3A.r), plot = FALSE)  
tabla.4

## ordered(genero$P3A.r)   
## Frequency Percent Valid Percent Cum Percent  
## Muy de acuerdo 67 5.569 5.702 5.702  
## De acuerdo 377 31.338 32.085 37.787  
## Ni de acuerdo ni en desacuerdo 63 5.237 5.362 43.149  
## En desacuerdo 569 47.298 48.426 91.574  
## Muy en desacuerdo 99 8.229 8.426 100.000  
## NA's 28 2.328   
## Total 1203 100.000 100.000

Establecer intervalos de clase para una variable cuantitativa, continua o discreta

Veamos la variable edad:

table(genero$EDAD)

##   
## 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42   
## 55 43 39 28 38 32 21 32 25 20 37 40 40 18 25 28 19 31 32 22 26 18 32 15 29   
## 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67   
## 19 31 34 22 17 22 15 29 13 21 14 12 8 14 16 8 9 20 5 15 11 6 10 13 10   
## 68 69 70 71 72 73 74 75 76 77 80 81 82 83 84 87 92   
## 5 5 12 6 4 4 3 8 2 1 2 1 4 4 1 1 1

Agrupar los valores en intervalos

En intervalos de 10 años:

edad <- genero$EDAD  
edad2 <- cut(edad, seq(from = 18, to = 98, by = 10), include.lowest=TRUE)  
table(edad2)

## edad2  
## [18,28] (28,38] (38,48] (48,58] (58,68] (68,78] (78,88] (88,98]   
## 370 281 239 150 104 45 13 1

edad2b <- cut(edad, seq(from = 15, to = 95, by = 10))  
table(edad2b)

## edad2b  
## (15,25] (25,35] (35,45] (45,55] (55,65] (65,75] (75,85] (85,95]   
## 288 283 258 173 114 70 15 2

Otra posibilidad:

edad2c <- cut(edad, breaks = c(18, 25, 35, 45, 55, 92),   
 include.lowest = TRUE)  
table(edad2c)

## edad2c  
## [18,25] (25,35] (35,45] (45,55] (55,92]   
## 288 283 258 173 201

La tabla anterior en formato similar al SPSS

tabla.5 <- freq(ordered(edad2c), plot = FALSE)  
tabla.5

## ordered(edad2c)   
## Frequency Percent Cum Percent  
## [18,25] 288 23.94 23.94  
## (25,35] 283 23.52 47.46  
## (35,45] 258 21.45 68.91  
## (45,55] 173 14.38 83.29  
## (55,92] 201 16.71 100.00  
## Total 1203 100.00

Tablas bi-variables

Una tabla bi-variable con las Rptas a la pregunta P3A en las filas y el sexo del entrevistado en las columnas: "P3A según sexo"

# Tabla de frecuencias cruzadas simples  
table(genero$P3A.r, genero$SEXO)

##   
## Masculino Femenino  
## Muy de acuerdo 40 27  
## De acuerdo 215 162  
## Ni de acuerdo ni en desacuerdo 30 33  
## En desacuerdo 249 320  
## Muy en desacuerdo 39 60

Tabla de frecuencias cruzadas en porcentajes calculados sobre el total de todos los casos de la tabla

prop.table(table(genero$P3A.r, genero$SEXO))\*100

##   
## Masculino Femenino  
## Muy de acuerdo 3.404 2.298  
## De acuerdo 18.298 13.787  
## Ni de acuerdo ni en desacuerdo 2.553 2.809  
## En desacuerdo 21.191 27.234  
## Muy en desacuerdo 3.319 5.106

# Tabla de porcentajes calculados sobre el total de cada fila  
prop.table(table(genero$P3A.r, genero$SEXO), 1)\*100

##   
## Masculino Femenino  
## Muy de acuerdo 59.70 40.30  
## De acuerdo 57.03 42.97  
## Ni de acuerdo ni en desacuerdo 47.62 52.38  
## En desacuerdo 43.76 56.24  
## Muy en desacuerdo 39.39 60.61

# Tabla de porcentajes calculados sobre el total de cada columna  
prop.table(table(genero$P3A.r, genero$SEXO), 2)\*100

##   
## Masculino Femenino  
## Muy de acuerdo 6.981 4.485  
## De acuerdo 37.522 26.910  
## Ni de acuerdo ni en desacuerdo 5.236 5.482  
## En desacuerdo 43.455 53.156  
## Muy en desacuerdo 6.806 9.967

Guardamos los resultados de la última tabla creada en un objeto llamado "tabla.6", con los números redondeados hasta 2 decimales

tabla.6 <- round(prop.table(table(genero$P3A.r, genero$SEXO), 2)\*100, digits = 2)  
tabla.6

##   
## Masculino Femenino  
## Muy de acuerdo 6.98 4.49  
## De acuerdo 37.52 26.91  
## Ni de acuerdo ni en desacuerdo 5.24 5.48  
## En desacuerdo 43.46 53.16  
## Muy en desacuerdo 6.81 9.97

Otra manera de tener una tabla de frecuencias cruzadas, usando el comando **"crosstab"** del paquete "descr". Es similar a la opción de tablas de contingencia o "Crosstabs" del SPSS

library(descr)  
tabla.7 <- crosstab(genero$P3A.r, genero$SEXO, prop.c = TRUE, plot = FALSE)  
tabla.7

Si en vez de calcular los % para las columnas los queremos para las filas, reemplazamos la opción **"prop.c = TRUE"** por **"prop.r = TRUE"**. Si los queremos calculados sobre el total de casos de la tabla, usamos **"prop.t = TRUE"**.

## Cell Contents   
## |-------------------------|  
## | Count |   
## | Column Percent |   
## |-------------------------|  
##   
## ==============================================================  
## genero$SEXO  
## genero$P3A.r Masculino Femenino Total  
## --------------------------------------------------------------  
## Muy de acuerdo 40 27 67  
## 6.981 4.485   
## --------------------------------------------------------------  
## De acuerdo 215 162 377  
## 37.522 26.910   
## --------------------------------------------------------------  
## Ni de acuerdo ni en desacuerdo 30 33 63  
## 5.236 5.482   
## --------------------------------------------------------------  
## En desacuerdo 249 320 569  
## 43.455 53.156   
## --------------------------------------------------------------  
## Muy en desacuerdo 39 60 99  
## 6.806 9.967   
## --------------------------------------------------------------  
## Total 573 602 1175  
## 48.766 51.234  
## ==============================================================

Para exportar tablas

Para enviar tablas a otros programas, por ejemplo el Excel. Podemos usar la función "xtable", del paquete con el mismo nombre (requiere ser instalado en la librería del R) Para más información ver:

<http://cran.r-project.org/web/packages/xtable/xtable.pdf>

<http://cran.r-project.org/web/packages/xtable/vignettes/xtableGallery.pdf>

tabla.2 # Correspondiente a la pregunta P3A

##   
## Muy de acuerdo De acuerdo   
## 5.70 32.09   
## Ni de acuerdo ni en desacuerdo En desacuerdo   
## 5.36 48.43   
## Muy en desacuerdo   
## 8.43

library(xtable)  
print(xtable(tabla.2, caption = "P3A"), type = "html",   
 file = "tabla2.html")

Ubique el archivo "tabla2.html" en su directorio de trabajo de R y ábralo desde el

Hagamos lo mismo con otras tablas y vea como quedan abriendo los archivos HTML desde el Excel.

print(xtable(tabla.3, caption ="Sexo del entrevistado"),   
 type = "html", file = "tabla3.html")  
  
print(xtable(tabla.4, caption ="P3A"),  
 type = "html", file = "tabla4.html")  
  
print(xtable(tabla.5, caption ="Grupos de edad"),   
 type = "html", file = "tabla5.html")  
  
print(xtable(tabla.6, caption = "P3A según sexo"),  
 type = "html", file = "tabla6.html")  
  
print(xtable(tabla.7, caption = "P3A según sexo"),  
 type = "html", file = "tabla7.html")

Gráficos

Gráficos para el análisis descriptivo

Una de las mejores herramientas para realizar un análisis estadístico descriptivo es el uso de gráficos estadísticos.

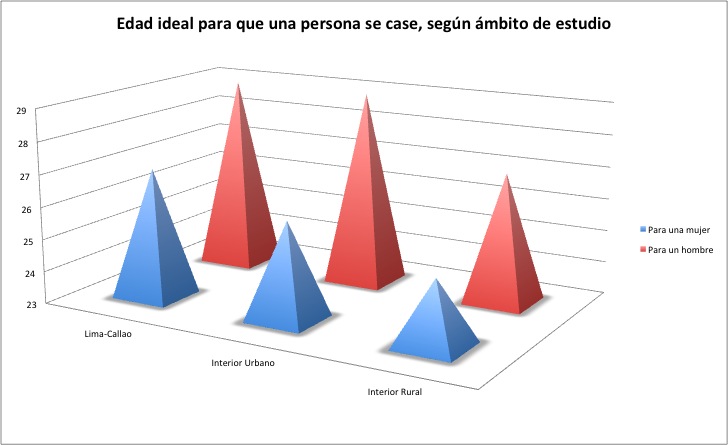
* Gráficos univariables para variables categóricas
  + Gráficos de barras
  + Gráficos de sector o "pies"
* Gráficos univariables para variables cuantitativas
  + Histogramas
  + Gráficos de líneas
  + Gráficos de cajas
* Gráficos bivariables
  + Variables categóricas: Barras múltiples
  + Variables cuantitativas: Líneas múltiples; gráficos de dispersión

Criterios para elaborar gráficos

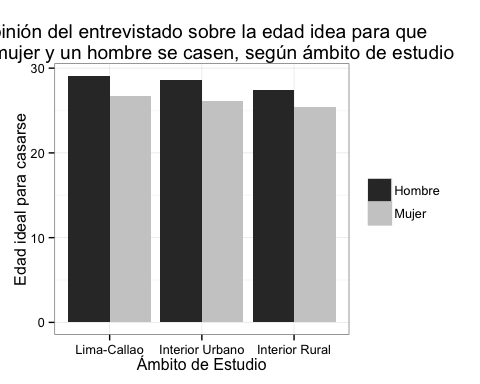
Un buen gráfico debe:

* Tener como objetivo mostrar la estructura de los datos
* Evitar distorsionar los datos
* Mostrar muchos datos y números con el mínimo de tinta posible
* Darle coherencia a grandes conjuntos de datos

Ejemplo de un mal gráfico



Ejemplo de un mejor gráfico (hecho en R)



Entornos gráficos en el R

El R ofrece diferentes posibilidades para realizar gráficos

* El entorno gráfico básico
* El paquete lattice
* El paquete ggplot

Veamos cómo se ven los resultados de la siguiente tabla usando los tres entornos gráficos seleccionados:

##   
## Muy de acuerdo De acuerdo   
## 67 377   
## Ni de acuerdo ni en desacuerdo En desacuerdo   
## 63 569   
## Muy en desacuerdo   
## 99

Gráfico de barras horizontales en el entorno básico

barplot(table(genero$P3A.r), horiz=TRUE)

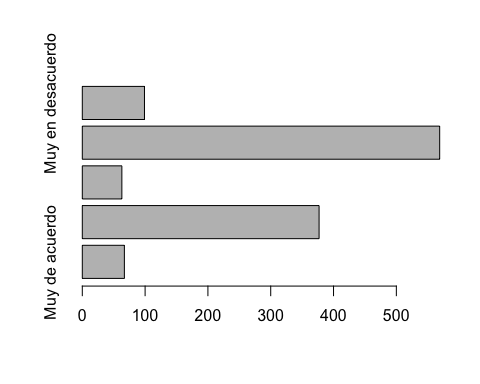


Gráfico usando el paquete lattice

library(lattice)  
barchart(genero$P3A.r)

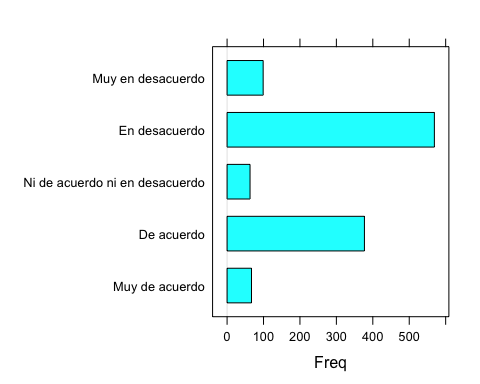
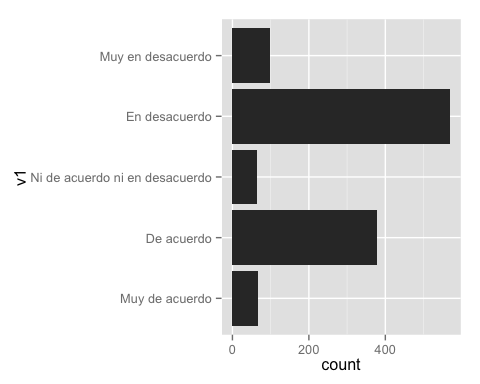


Gráfico usando el paquete ggplot

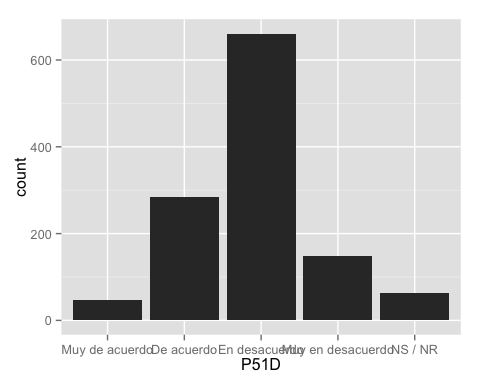
library(ggplot2)  
 v1 <- na.omit(genero$P3A.r)  
qplot(v1, geom = "bar") + coord\_flip()



Representar una variable categórica: Barras

El gráfico de barras es la mejor opción para representar variables categóricas. Por ejemplo, si queremos hacer un gráfico de distribución de las respuestas a la pregunta P51D:

ggplot(genero, aes(P51D)) + geom\_bar()



Representar una variable categórica: Pie o Sectores

absoluta <- table(p51D) #tabla de frecuencias absolutas

pie(absoluta)

Representar una variable cuantitativa: Boxplot

boxplot(Edad)

Empezando a darle forma

boxplot(Edad, main="Edad", horizontal=TRUE, range=0.5)

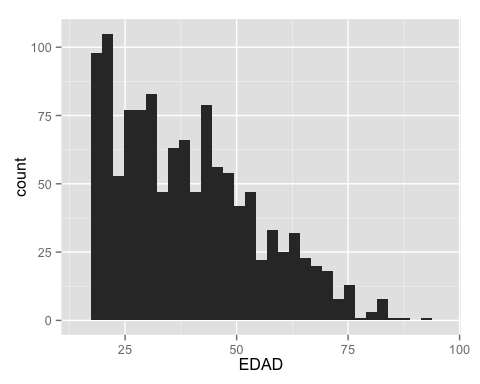
Boxplot por categorías

boxplot(Edad ~ Sexo, main="Edad según Sexo", xlab= "Sexo", ylab="Edad")

Representar una variable cuantitativa: Histograma

El histograma es la primera opción para observar la distribución de una variable cuantitativa:

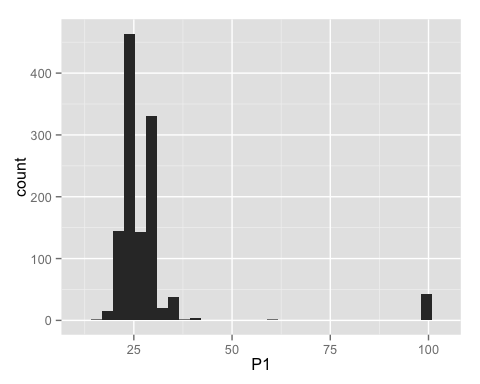
ggplot(genero, aes(EDAD)) + geom\_histogram()



Utilidad del histograma: Casos atípicos

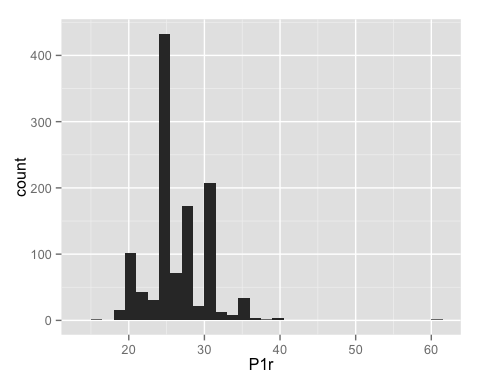
¿Qué está mal aqui?

ggplot(genero, aes(P1)) + geom\_histogram()



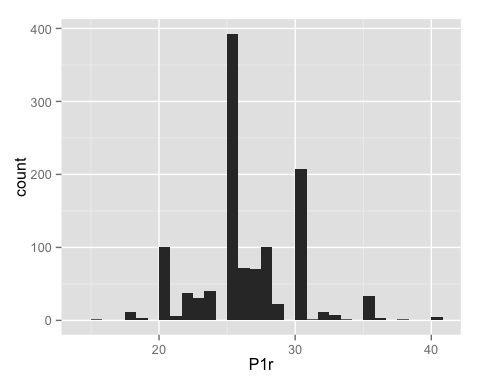
Corregimos:

genero$P1r <- genero$P1  
genero$P1r[genero$P1 == 99] <- NA # Designamos al código 99 como NA  
ggplot(genero, aes(P1r)) + geom\_histogram()



Observe que en el histograma anterior hay otro dato "raro", pruebe con quitarlo y rehacer nuevamente el histograma. Debería quedar como esto:

genero$P1r <- genero$P1  
genero$P1r[genero$P1 >= 60] <- NA # Designamos al código 99 como NA  
ggplot(genero, aes(P1r)) + geom\_histogram()

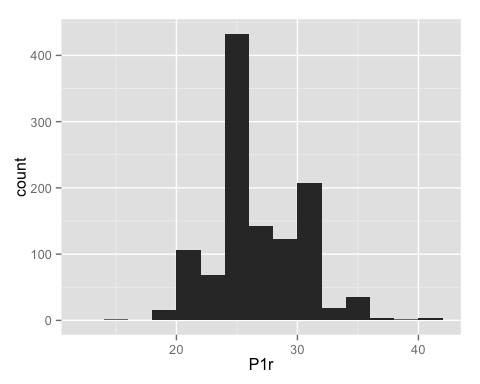


Amplitud de las barras en el histograma

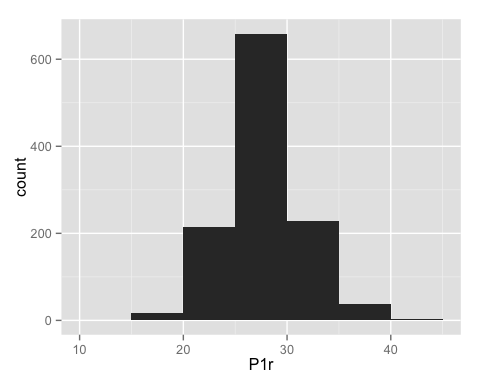
* En un histograma, la amplitud de las barras representa un intervalo de clase.
* En el histograma anterior, la amplitud de cada intervalo es = 1.5, ya que por defecto la función geom\_histogram divide el rango entre 30.
* Cambiando la amplitud de las barras podemos modificar cómo se presentan los datos.

Cambiando la amplitud de las barras del histograma

#Amplitud 2  
ggplot(genero, aes(P1r)) + geom\_histogram(binwidth=2)



#Amplitud 5  
ggplot(genero, aes(P1r)) + geom\_histogram(binwidth=5)



Comparar distribuciones

Con el histograma podemos comparar diferentes distribuciones de una variable cuantitativa:

ggplot(genero, aes(P1r)) + geom\_histogram(binwidth=2) + facet\_grid(Ambito ~.)

