



Estadística para las Ciencias Sociales con



Notas de clase (v.1.0)

David Sulmont

**Pontificia Universidad Católica del
Perú**

Marzo 2015

ÍNDICE DE CONTENIDOS

Presentación.....	5
Conceptos Básicos.....	7
Objetos en R	9
Tipos de datos en R	9
Conocer las características de un objeto	11
Resumen	12
Manejo de Datos en R	13
Cargar datos en R.....	13
Manipular datos en R	17
Selección de casos en R	22
Seleccionar un subconjunto de variables.....	22
Seleccionar casos	23
Análisis Univariable: Tablas de frecuencias	24
Tablas de distribución de frecuencias.....	24
Recodificar una variable y hacer una tabla	25
Tablas parecidas al SPSS	26
Establecer intervalos de clase para una variable cuantitativa	27
Tablas bi-variables.....	28
Para exportar tablas	30
Gráficos.....	32
Gráficos para el análisis descriptivo	32
Criterios para elaborar gráficos	32
Entornos gráficos en el R.....	33
Representar una variable categórica: Barras	36
Representar una variable cuantitativa: Histograma	37
Amplitud de las barras en el histograma.....	41
Curva de densidad de Kernell	43
Gráficos con el paquete ggplot2	46
El paquete ggplot2	46
Capas o layers en ggplot2.....	46
Un gráfico simple: Gráfico de barras.....	47
Gráfico de barras múltiples.....	54
Usando facets	57
Histogramas	62
Curvas de Densidad de Kernell.....	64
Gráfico de líneas.....	67
Gráfico de puntos.....	70
Grabar el gráfico en un archivo	73
Estadísticos de Resumen en R	75
Estadísticos de resumen.....	75
Medidas de tendencia central	75

Medidas de dispersión	75
Cuantiles: Medidas de orden o posición	76
Manejo de valores perdidos.....	78
Resúmenes para grupos de casos.....	79
Boxplots	82
Boxplots o gráficos de cajas	82
Boxplot con ggplot.....	82
Inferencia Estadística y Distribuciones de Muestreo	85
Parámetros de la población	85
Muestreo, estadística inferencial y error muestral	86
Muestreo simple al azar en R.....	86
Distribuciones de muestreo empíricas	87
"LA" Distribución de muestreo	89
El Error Estándar e intervalos de confianza	92
Distribución de muestreo para proporciones	96
Margen de error de una muestra	99
Inferencia Estadística e Intervalos de Confianza	101
Inferencia estadística.....	101
Cálculo de intervalos de confianza para una media	102
Cálculo de intervalo de confianza para una proporción	104
Gráficar el intervalo de confianza	106
Pruebas de Hipótesis	112
Inferencia estadística y Pruebas de Hipótesis	112
Pruebas de hipótesis para medias de muestra única.....	113
Prueba de una y de dos colas.....	115
Prueba de hipótesis para medias de dos muestras independientes	117
Pruebas de hipótesis para medias de muestras relacionadas	119
Tablas de Contingencia	122
Tablas de contingencia como herramienta de análisis bivariante.....	122
Elaborar una tabla de contingencia.....	122
¿Cómo leer una tabla de contingencia?	124
Prueba de X ²	125
Medidas o coeficientes de asociación	131
Uso de medidas de asociación.....	131
Medidas de asociación para variables ordinales	132
Función tabla.cont	136
Funcion para generar tablas de contingencia.....	136
Análisis de la Varianza	141
Análisis de la varianza	141
Ejemplo: Rendimiento académico y nivel educativo de los padres.....	141
Modelo lineal general o de efectos aditivos.....	145
ANOVA y Suma de Cuadrados	146
Pasos de ANOVA	147
ANOVA con R	147
Prueba de diferencias entre grupos específicos	148
Otro ejemplo: Rendimiento según nivel educativo de la madre	150

Ejemplo: Horas dedicadas a las tareas según nivel educativo de la madre	152
Correlación y Regresión Simple.....	154
Pasos en el análisis de regresión.....	154
Cálculo de los coeficientes de regresión	159
Lógica de R ² : Primero el modelo de regresión.....	161
Prueba de significancia del coeficiente "r de Pearson"	170
Matriz de correlaciones.....	170

Presentación

Este texto contiene las notas de clase de “Estadística para las Ciencias Sociales con R” que dicto en la Pontificia Universidad Católica del Perú. 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.

Estas notas de clase no constituyen (todavía) un libro de texto de estadística social. Son parte de los materiales que he venido desarrollando para enseñar a utilizar el R como herramienta de análisis estadístico. Como texto de referencia de estadística para las ciencias sociales se recomienda revisar el siguiente libro:

Ritchey, Ferris J. 2008. *Estadística para las ciencias sociales*. México D.F.: McGraw Hill.

El R como herramienta de análisis estadístico es uno de los paquetes informáticos cuyo uso ha crecido de manera más acelerada en los últimos años. Sus principales ventajas son que se trata de un software libre y que cuenta con una amplia comunidad de usuarios expertos que continuamente introducen innovaciones, así como proveen de útiles materiales de consulta disponibles gratuitamente en internet.

Otra ventaja del R es que, siendo un lenguaje de programación, entrena al usuario a ordenar su estrategia de análisis de los datos cuantitativos. Si bien el proceso de aprendizaje del R puede ser al inicio algo complicado, en comparación con otros software que tienen interfaces más amigables, a la larga le permite al estudiante dominar herramientas analíticas más complejas, sobre todo aquellas que requieren algún tipo de programación para optimizar procesos de análisis de grandes o complejos conjuntos de datos.

El R puede descargarse desde la página web del “R Project for Statistical Computing” (<http://www.r-project.org/>) , donde además se puede encontrar una gran variedad de materiales de ayuda.

A lo largo de los ejercicios utilizaremos algunos datos generados por el Instituto de Opinión Pública de la PUCP. Estos datos pueden consultarse y descargarse de forma gratuita en la siguiente dirección: <http://iop-data.pucp.edu.pe>.

Las notas de clase han sido utilizadas para elaborar videos instructivos que le permitirán al estudiante seguir desde cualquier computadora las explicaciones y el desarrollo de los ejemplos y ejercicios propuestos. Todos estos materiales pueden accederse en la siguiente página web:
https://sites.google.com/a/pucp.pe/data_est/.

El texto, los videos, ejercicios prácticos y test de autoevaluación que constituyen el conjunto de materiales de enseñanza de este curso, han sido elaborados gracias al apoyo obtenido por el Primer Fondo Concursable de Innovación para la Docencia Universitaria, organizado por la Dirección Académica del Profesorado de la PUCP el 2014.

Debo reconocer y agradecer a César Córdova, licenciado en Psicología por la PUCP, por su apoyo en el desarrollo de estos materiales. César se ha desempeñado como jefe de prácticas del curso que he dictado en la facultad de Ciencias Sociales en el 2014-2. Asimismo, agradezco la paciencia y esfuerzo de los alumnos del pregrado y de la maestría de sociología matriculados en mis cursos el 2014-2, ellos me han permitido probar estos materiales en su calidad de involuntarios “conejillos de indias”.

En la medida que se trata de materiales que todavía están en etapa de desarrollo y prueba, me será de gran utilidad recibir comentarios, sugerencias y críticas que me permitan mejorarlo, para ello pueden comunicarse conmigo a: sulmont@pucp.pe. Todo aporte será muy bien recibido.

Espero que los estudiantes que utilicen estos materiales los encuentren útiles para aprender a dominar poco a poco herramientas necesarias para el oficio del sociólogo y de otras profesiones que deben trabajar con información de tipo cuantitativa.

David Sulmont
Profesor Principal
Departamento de Ciencias Sociales
Pontificia Universidad Católica del Perú

Conceptos Básicos

Estadística

Es el conjunto de procedimientos que nos permiten medir, resumir y analizar información cuantitativa adquirida sistemáticamente.

La estadística descriptiva nos permite organizar y describir las características de un conjunto de datos.

Unidad de análisis, población y muestra

- Unidad de análisis: es la unidad que estamos estudiando, observando o midiendo. Respecto de la cual se sacarán las conclusiones del análisis.
 - Ejemplo: hogares, electores, individuos
- Población: es la enumeración completa de todas las unidades de análisis
 - Ejemplo: todos los hogares del Perú; todos los electores inscritos en el padrón; todos los habitantes del país.
- Muestra: Es un subconjunto de unidades de análisis de la población que han sido seleccionadas para ser observadas en una investigación.

La estadística inferencial nos permite elaborar conclusiones sobre una población a partir del análisis de una muestra de datos de esa población.

Variables

Una variable es una representación de una propiedad de una unidad de análisis que estamos estudiando u observando.

Una variable posee diferentes valores o atributos, que son la forma específica en que se manifiesta esa propiedad en una unidad de análisis determinada.

Por ejemplo, si nuestra unidad de análisis son hogares, un atributo de los hogares es su tamaño. La variable "tamaño del hogar" puede medir ese atributo contando la cantidad de miembros del hogar. Los valores de esa variable pueden ser 1, 2, 3, etc., miembros del hogar. Otro atributo del hogar puede ser "nivel de pobreza del hogar", que puede medirse con una variable que posee los siguientes valores: "No pobre; pobre no extremo; pobre extremo".

Tipos de variables

Dependiendo de la forma en que se representan los valores de una variable, éstas puedes clasificarse en:

- Variables categóricas o cualitativas: Los valores son "cualidades", por ejemplo: "Hombre", "Mujer" para la variable sexo; "Primaria", "Secundaria", "Superior" para la variable nivel educativo.
- Variables cuantitativas: Los valores son puntuaciones numéricas, por ejemplo: 1, 2, 3 miembros de un hogar; 1590, 2450, 3220.3 soles mensuales de ingreso.

Niveles de medición: Variables categóricas

Para representar los valores de las variables, tenemos diferentes niveles de medición, dependiendo del tipo de variables:

- Nominal: Cuando las categorías expresan simplemente diferencias cualitativas. Por ejemplo el género (hombre o mujer); o el estado civil (soltero, casado, divorciado)
- Ordinal: Cuando las categorías pueden clasificarse de menos a más: por ejemplo el nivel educativo (primaria, secundaria, superior); o el grado de acuerdo con una afirmación (muy de acuerdo, de acuerdo, en desacuerdo, muy en desacuerdo)

Niveles de medición: Variables cuantitativas

Las variables cuantitativas pueden medirse usando escalas:

- Puntuaciones clasificadas ordinales: por ejemplo el orden de méritos en una clase (1ro, 2do, 3ro, 4to, 5to, etc)
- Escalas de intervalo:
 - Sin punto cero real: Cuando los valores tienen un punto cero arbitrario. Por ejemplo la temperatura (cero grados no equivale a ausencia de temperatura); o el año de nacimiento (haber nacido en el año cero significa haber nacido el mismo año que Cristo).
 - Escalas de razón: cuando los valores tienen un punto cero real que significa ausencia del atributo que se está midiendo. Por ejemplo la edad, el salario del mes, el peso en Kg, etc.

Importancia del nivel de medición

Es muy importante saber cómo se está midiendo una variable, ya que dependiendo de su tipo y nivel de medición podemos realizar algunas operaciones y otras no.

- En el caso de variables cuantitativas podemos realizar operaciones aritméticas entre sus valores: sumar o restar valores; dividirlos, multiplicarlos, etc.
- En el caso de variables categóricas no podemos hacer operaciones aritméticas con sus valores.

Objetos en R

Tipos de datos en R

En esta sección mostraremos cómo guarda el R la información de nuestras observaciones y variables.

El R soporta una variedad de tipos de datos y los almacena en una serie de objetos. En esta sección nos concentraremos en los siguientes tipos de objetos.

- Vectores
- Factores
- Data frames

Vectores

Podemos crear un vector numérico de la siguiente forma:

```
x <- c(10, 15, 14, 9.5, 18, 8)
```

Tenemos entonces un vector llamado "x" que contiene 6 observaciones.

```
x
## [1] 10.0 15.0 14.0 9.5 18.0 8.0
```

En el R, las variables cuantitativas se almacenan en un vector numérico.

Los vectores también pueden almacenar caracteres o expresiones lógicas (como verdadero / falso):

```
# Este es un vector de caracteres:
y <- c("José", "María", "Pedro", "Luis", "Elena",
      "Sofía")

# Este es un vector Lógico:
z <- c(FALSE, TRUE, TRUE, FALSE, TRUE, FALSE)
```

Los vectores nos permiten almacenar información, por ejemplo, los tres vectores que hemos creado pueden representar tres variables diferentes: las notas de un examen; el nombre de los estudiantes; si aprobaron o no el curso:

```
notas <- x
nom.est <- y
aprob <- z
notas

## [1] 10.0 15.0 14.0 9.5 18.0 8.0

nom.est

## [1] "José"  "María" "Pedro" "Luis"  "Elena" "Sofía"
```

aprob

```
## [1] FALSE TRUE TRUE FALSE TRUE FALSE
```

Factores

En el R, un factor representa una variable cualitativa o categórica. El factor almacena las categorías en la forma de un vector con números discretos integrales (1, 2, 3, 4, etc.) que son los códigos de los valores de la variable y otro vector de caracteres interno que contiene las etiquetas de esos códigos. Por ejemplo:

```
genero <- c("Masc", "Fem", "Masc", "Masc", "Fem", "Fem")
genero <- factor(genero)
genero

## [1] Masc Fem Masc Masc Fem Fem
## Levels: Fem Masc
```

El factor género en este caso es una variable categórica nominal y tiene 2 niveles o valores: Masc y Fem.

Internamente el R almacenará: Masc = 2; Fem = 1 (alfabeticamente). Eso puede verse al convertir el factor género en un vector de números enteros:

```
as.integer(genero)

## [1] 2 1 2 2 1 1
```

Factores ordinales

El factor como objeto también me permite almacenar variables categóricas ordinales. Supongamos que queremos registrar la evaluación que tienen los alumnos de sus profesores usando el siguiente esquema de codificación: Buena = 1; Regular = 2; Mala = 3. Entonces:

```
#Creamos un vector numérico que registra Los datos:
eval.prof <- c(1,3,2,2,1,3)
#Convertimos el vector en un factor
eval.prof <- factor(eval.prof)
#Asignamos los niveles al factor:
levels(eval.prof) <- c("Bueno", "Regular", "Malo")
#Indicamos que se trata de un factor ordinal:
eval.prof <- ordered(eval.prof)
eval.prof

## [1] Bueno Malo Regular Regular Bueno Malo
## Levels: Bueno < Regular < Malo
```

Data Frames o conjuntos de datos

Un data frame es un objeto que permite almacenar un conjunto de datos o una base de datos en la forma de una matriz de filas y columnas:

- Cada fila representa un registro, que corresponde a una unidad de análisis, observación o caso.
- Cada columna representa un campo o variable. Estas variables pueden ser vectores o factores.
- La primera fila contiene los nombres de las variables
- La primera columna contiene los identificadores de cada caso.

Podemos juntar los objetos que hemos creado a lo largo de esta presentación en un data frame:

```
#Creamos el data frame "mis.datos"
misdatos <- data.frame(nom.est, genero, notas, aprob,
                      eval.prof)
#Invocamos al data frame creado
misdatos

##   nom.est genero notas aprob eval.prof
## 1    José   Masc  10.0 FALSE    Bueno
## 2   María   Fem  15.0  TRUE     Malo
## 3   Pedro   Masc  14.0  TRUE   Regular
## 4    Luis   Masc   9.5 FALSE   Regular
## 5   Elena   Fem  18.0  TRUE    Bueno
## 6  Sofía   Fem   8.0 FALSE     Malo
```

Guardar un data frame

Podemos guardar un data frame. El archivo se guardará en el directorio de trabajo que hemos especificado al inicio de nuestra sesión:

```
save(misdatos, file="misdatos.Rda") # En formato de R
```

Si iniciamos una nueva sesión de R y queremos trabajar con los datos que hemos guardado, los cargamos usando el comando:

```
load("misdatos.Rda")
```

Para ello el archivo debe estar en el directorio de trabajo del R, de lo contrario debemos especificar la ruta completa del archivo entre los paréntesis y comillas.

Conocer las características de un objeto

Algunos comandos nos ayudan a conocer las características de un objeto. Por ejemplo usando los objetos que hemos creado hasta ahora:

```
length(genero) # número o cantidad de elementos del objeto
## [1] 6
str(genero)    # estructura del objeto
## Factor w/ 2 levels "Fem","Masc": 2 1 2 2 1 1
class(misdatos) # clase o tipo de objeto
```

```
## [1] "data.frame"

# nombres de Las variables (cuando el objeto es un data frame)
names(misdatos)

## [1] "nom.est"    "genero"     "notas"      "aprob"      "eval.prof"

# Niveles del factor (cuando el objeto es un factor)
levels(genero)

## [1] "Fem"       "Masc"
```

Ejemplo de cómo conocer la estructura de un data frame:

```
str(misdatos)

## 'data.frame':   6 obs. of  5 variables:
## $ nom.est : Factor w/ 6 levels "Elena","José",...: 2 4 5 3 1 6
## $ genero  : Factor w/ 2 levels "Fem","Masc": 2 1 2 2 1 1
## $ notas   : num  10 15 14 9.5 18 8
## $ aprob    : logi FALSE TRUE TRUE FALSE TRUE FALSE
## $ eval.prof: Ord.factor w/ 3 levels "Bueno"<"Regular"<...: 1 3 2 2
1 3
```

Resumen

Nuestras observaciones y variables pueden almacenarse en el R utilizando los siguiente objetos:

- Una variable cuantitativa puede almacenarse en un vector numérico.
- Una variable categórica puede almacenarse en un vector de caracteres o en un factor.
- Una variable categórica ordinal puede almacenarse en un factor ordinal.
- Una variable categórica dicotómica (tipo SI/NO; Verdadero/Falso), puede almacenarse en un vector lógico.
- Un conjunto de variables correspondientes al mismo grupo de casos u observaciones (registros) puede almacenarse en un Data Frame (o base de datos).

Manejo de Datos en R

En esta sección examinaremos los siguientes temas:

- Cargar datos en R
- Importar bases de datos de otros formatos al R
- Manipular variables en el R
 - Identificar valores perdidos
 - Recodificar variables
 - Calcular nuevas variables

Cargar datos en R

Para seguir los procedimientos mostrados en esta presentación, primero deberá descargar y descomprimir el archivo **r_import.zip** que encontrará en el siguiente enlace web:

https://sites.google.com/a/pucp.pe/data_est/archivos

Deberá guardar y descomprimir el contenido del archivo zip en su directorio de trabajo de R especificado al iniciar su sesión.

Si logró con éxito descomprimir los archivos contenidos en el enlace anterior, en su directorio de trabajo deberá encontrar los siguientes archivos:

- genero.rda : archivo en formato R
- genero.xls : archivo en formato Excel (97)
- genero.csv : archivo en formato csv
- genero.sav : archivo en formato SPSS

Estos son algunos de los formatos más usuales para guardar bases de datos con las cuales podemos realizar algunos análisis estadísticos. En las siguientes diapositivas les mostraremos cómo cargar datos desde estos formatos.

Cargar datos en formato R

Si tuvo éxito en el paso anterior, podrá cargar los datos en formato R usando el siguiente comando (recuerde que los archivos deben estar guardados en su directorio de trabajo, de lo contrario deberá especificar la ruta completa del archivo en su disco duro):

```
load("genero.rda")
```

Con ese comando usted está cargando un data frame en R que contiene la base de datos llamada "genero".

Explorar el contenido de la base de datos "genero"

Para explorar el contenido del archivo género puede usar:

```
head(genero)
```

```
##   NRO      SEXO EDAD edideal_muj edideal_hom      DOMINIO NSEGrup
## 1   1  Femenino   55       25       30 Lima-Callao    A/B
## 2   2 Masculino   62       26       30 Lima-Callao    A/B
## 3   3  Femenino   70       28       30 Lima-Callao    A/B
## 4   4  Femenino   20       30       35 Lima-Callao    A/B
## 5   5  Femenino   18       25       30 Lima-Callao    A/B
## 6   6 Masculino   25       25       30 Lima-Callao     C
```

Ello le muestra los seis primeros registros o casos de la base de datos o data frame.

Podemos ver el contenido la base de datos utilizando el comando:

```
str(genero)
```

```
## 'data.frame': 1203 obs. of 7 variables:
## $ NRO : num 1 2 3 4 5 6 7 8 9 10 ...
## $ SEXO : Factor w/ 2 levels "Masculino","Femenino": 2 1 2 2
## $ EDAD : num 55 62 70 20 18 25 46 38 52 53 ...
## $ edideal_muj: num 25 26 28 30 25 25 30 35 25 26 ...
## $ edideal_hom: num 30 30 30 35 30 30 35 38 25 28 ...
## $ DOMINIO : Factor w/ 5 levels "Lima-Callao",...: 1 1 1 1 1 1 1
## $ NSEGrup : Factor w/ 3 levels "A/B","C","D/E": 1 1 1 1 1 2 3 2
2 2 ...
```

Para más detalles sobre esta base de datos pueden acceder a la información completa a través del portal IOP-Data:

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

Las variables de la base de datos "genero" son:

- NRO: Número o ID del entrevistado
- SEXO: Sexo del entrevistado
- EDAD: Edad del entrevistado
- edideal_muj: Edad ideal para que una mujer se case
- edideal_hom: Edad ideal para que un hombre se case
- DOMINIO: Dominio geográfico de estudio
- NSEGrup: Nivel socioeconómico del hogar del entrevistado (agrupado)

Cargar datos en otros formatos: Excel

Para importar una base de datos en Excel, lo más recomendable es guardar el archivo Excel en un formato delimitado por comas (csv).

Entre los archivos que se les pidió descomprimir, hay una versión de la base de datos en Excel (genero.xls). Abra ese archivo en Excel y pida la opción "Guardar como" y seleccione como tipo de archivo "CSV (delimitado por comas)".

Al usar esta opción es importante tomar en cuenta cuál es el separador de decimales que está utilizando en su computadora. El formato estándar es usar el punto como separador de decimales, pero algunas personas prefieren usar las comas como separador de decimales. En este último caso, la conversión de Excel a CSV puede variar. En este ejemplo, estamos suponiendo que el estándar utilizado es la separación de decimales con el punto y que el sistema operativo empleado es el Windows.

Cargar archivos en otros formatos: CSV

El formato de archivo separado por comas es un estándar bastante difundido. Guarda las bases de datos en archivos donde en cada fila se almacena la información de un caso y los campos o variables están separados por comas. Estos archivos pueden leerse desde cualquier editor de textos.

Entre los archivos descomprimidos usted encontrará uno llamado "genero.csv". Para cargar el archivo al R deberá utilizar el siguiente comando:

```
genero2 <- read.csv("genero.csv")
```

Esto almacenará la base de datos en un objeto de R tipo data frame llamado "genero2". Le hemos puesto el sufijo "2" para diferenciarlo de "genero" creado previamente". Puede usar los comandos head(genero2) o names(genero2) para explorar el contenido de este data frame.

Cargar archivos en otros formatos: SPSS

Para importar un archivo de SPSS primero debemos cargar el paquete "foreign". Los paquetes en R son programas especiales de R que contienen funciones específicas, en este caso el paquete "foreign" sirve para importar datos en diversos formatos.

Los paquetes del R se almacenan en lo que se llama la "librería" (library) de R. El programa de base de R vienen con un conjunto de paquetes pre-instalados (entre ellos el "foreign"), sin embargo en algunos casos es necesario descargar un paquete desde un CRAN e instalarlo en el R. Para mayor información sobre cómo instalar paquetes en R recomendamos revisar el siguiente link:

<http://www.statmethods.net/interface/packages.html>

Vamos a importar el archivo "genero.sav" y guardarlo en un objeto tipo data frame que se llame "genero3":

```
library(foreign)
genero3 <- as.data.frame(read.spss("genero.sav"))
```

Comparar datos importados

```

head(genero, 2) # Base de datos en formato R

##   NRO      SEXO EDAD edideal_muj edideal_hom      DOMINIO NSEGrup
## 1   1   Femenino   55          25          30 Lima-Callao     A/B
## 2   2   Masculino   62          26          30 Lima-Callao     A/B

head(genero2, 2) # Base de datos importada desde csv

##   NRO SEXO EDAD edideal_muj edideal_hom DOMINIO NSEGrup
## 1   1   2   55          25          30      1      1
## 2   2   1   62          26          30      1      1

head(genero3, 2) # Base de datos importada desde spss

##   NRO      SEXO EDAD edideal_muj edideal_hom      DOMINIO NSEGrup
## 1   1   Femenino   55          25          30 Lima-Callao     A/B
## 2   2   Masculino   62          26          30 Lima-Callao     A/B

```

Invocar una variable específica

Para invocar una variable específica de un data frame debemos usar la expresión: `dataframe$variable`. Por ejemplo :

```

str(genero$DOMINIO)

##  Factor w/ 5 levels "Lima-Callao",...: 1 1 1 1 1 1 1 1 1 ...
class(genero2$DOMINIO)

## [1] "integer"

```

Ojo que el R sí diferencia entre mayúsculas y minúsculas.

Convertir un vector en factor

Vamos a convertir la variable "SEXO" del data frame "genero2" en un factor con su correspondientes etiquetas. Nótese que las etiquetas (labels) deben especificarse en el orden que corresponden a sus códigos.

```

genero2$SEXO <- factor(genero2$SEXO, labels=c("Femenino", "Masculino"))
)
str(genero2$SEXO)

##  Factor w/ 2 levels "Femenino","Masculino": 2 1 2 2 2 1 1 1 1 2 ...

```

Sigamos convirtiendo las demás variables categóricas del data frame "genero2" en factores con sus respectivas etiquetas:

```

genero2$DOMINIO <- factor(genero2$DOMINIO,
                           labels=c("Lima-Callao", "Norte", "Sur",
                                   "Centro", "Oriente"))
genero2$NSEGrup <- factor(genero2$NSEGrup,

```

```

labels=c("A/B", "C", "D/E"))
head(genero2)

##   NRO      SEXO EDAD edideal_muj edideal_hom      DOMINIO NSEGrup
## 1   1 Masculino   55        25          30 Lima-Callao     A/B
## 2   2 Femenino    62        26          30 Lima-Callao     A/B
## 3   3 Masculino   70        28          30 Lima-Callao     A/B
## 4   4 Masculino   20        30          35 Lima-Callao     A/B
## 5   5 Masculino   18        25          30 Lima-Callao     A/B
## 6   6 Femenino    25        25          30 Lima-Callao     C

```

Manipular datos en R

Valores perdidos

Cuando se registran datos en una investigación es posible que no se logre obtener la información de algunas variables para determinados casos. En las encuestas ello suele suceder cuando el entrevistado no quiere o no sabe responder una pregunta (los "no sabe / no responde").

En muchos casos se suele excluir este tipo de respuestas del análisis, registrándolas como "valores perdidos"

En el caso de la base de datos de "genero" que estamos usando, si revisan el cuestionario en IOP-Data, notarán que para la pregunta "Edad ideal para que una mujer se case", el código "99" identifica a las respuestas del tipo "No sabe / No responde". Al pedir una tabla de frecuencias de esta variable, podemos notar que hay 42 casos que tienen como respuesta el código "99".

```

table(genero$edideal_muj)

##
##   15   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32   33
34
##   1   12   3 101    6   37   31   40 393   72   71 101   22 207    1   12    7
1
##   35   36   38   40   60   99
##   34   3   1    4   1   42

```

En el R, el código para valores perdidos es "NA". En el caso de la variable de edad ideal para que una mujer se case, podemos indicar que el valor "99" es un valor perdido de la siguiente manera:

```

genero$edideal_muj[genero$edideal_muj==99] <- NA
table(genero$edideal_muj)

##
##   15   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32   33
34
##   1   12   3 101    6   37   31   40 393   72   71 101   22 207    1   12    7

```

```
1
## 35 36 38 40 60
## 34 3 1 4 1
```

Nótese que al pedir la tabla de frecuencias de la variable, ya no aparecen los casos que son "NA".

Podemos saber cuántos valores perdidos tiene una variable creando un vector lógico que idenfique a los valores perdidos:

```
perdidos.v1 <- is.na(genero$edideal_muj)
table(perdidos.v1)

## perdidos.v1
## FALSE TRUE
## 1161 42

# También podemos hacerlo de la siguiente manera:
table(is.na(genero$edideal_muj))

## 
## FALSE TRUE
## 1161 42
```

Hay 42 casos perdidos para la variable edideal_muj

Seleccionar o identificar casos y variables en un data frame

Si queremos seleccionar o identificar un grupo de casos de un data frame, una alternativa es la siguiente:

Ejemplo 1: Los datos del tercer registro de la base de datos genero

```
genero[3, ]
##   NRO      SEXO EDAD edideal_muj edideal_hom      DOMINIO NSEGru
## 3   3 Femenino    70          28            30 Lima-Callao     A/B
```

Ejemplo 2: Los seis primeros registros de la tercera variable de la base de datos:

```
head(genero[, 3])
## [1] 55 62 70 20 18 25
```

Ejemplo 3: El genero del 10º registro de la base de datos:

```
genero[10, 2]
## [1] Femenino
## Levels: Masculino Femenino
```

Podemos guardar el resultado de una selecciónn en un nuevo objeto, por ejemplo si queremos seleccionar todos los registros pero sólo de la variable edad, podemos:

```
edad <- genero[, 3]
head(edad)

## [1] 55 62 70 20 18 25
```

Recodificar variables

Vamos a recodificar la variable DOMINIO de la base de datos género en una nueva variable que tenga dos categorías: "Lima y Callao" y "Provincia"

```
class(genero$DOMINIO)
## [1] "factor"

levels(genero$DOMINIO)
## [1] "Lima-Callao" "Norte"       "Sur"         "Centro"      "Oriente"
e"
```

Para ello, primero vamos a convertir la variable dominio en un vector numérico temporal

```
dominio.t <- as.numeric(genero$DOMINIO)
```

Luego, crearemos un vector vacío donde vamos a recodificar los valores correspondientes a los códigos originales de la variable dominio. El nuevo vector será luego transformado en un factor.

```
dominio2 <- vector() # Se crea un vector vacío
dominio2[dominio.t == 1] <- "Lima y Callao"
dominio2[dominio.t > 1] <- "Provincia"
dominio2 <- as.factor(dominio2)
```

Finalmente, integraremos la nueva variable recodificada en el data frame

```
genero$dominio2 <- dominio2
table(genero$DOMINIO)

##
## Lima-Callao      Norte       Sur       Centro      Oriente
##        448          320        245        105         85

table(genero$dominio2)

##
## Lima y Callao     Provincia
##        448           755
```

Otra manera de recodificar la misma variable con los mismos resultados es la siguiente:

```
dominio.t2 <- as.numeric(genero$DOMINIO)
dominio3<- ifelse(dominio.t2 > 1, c("Provincia"), c("Lima y Callao"))
table(dominio3)
```

```
## dominio3
## Lima y Callao      Provincia
##          448          755
```

Recodificar vectores numéricicos en factores

Podemos convertir vectores numéricos en factores para agrupar los valores de una variable cuantitativa en intervalos de clase categóricos, por ejemplo, grupos de edad.

```
gedad <- vector()
gedad[genero$EDAD < 25] <- 1
gedad[genero$EDAD >= 25 & genero$EDAD < 35] <- 2
gedad[genero$EDAD >= 35 & genero$EDAD < 45] <- 3
gedad[genero$EDAD >= 45] <- 4
genero$gedad <- as.factor(gedad)
levels(genero$gedad) <- c("18 a 24", "25 a 34", "35 a 44", "45 a más")
table(genero$gedad)

##
## 18 a 24  25 a 34  35 a 44 45 a más
##       256      284      255      408
```

Otra forma de hacerlo es usando el comando "cut" para establecer puntos de corte para los intervalos

```
gedad2 <- cut(genero$EDAD, breaks=c(17, 24, 34, 44, 100),
               include.lowest=TRUE)
table(gedad2)

## gedad2
## [17,24]  (24,34]  (34,44] (44,100]
##       256      284      255      408

table(genero$gedad)

##
## 18 a 24  25 a 34  35 a 44 45 a más
##       256      284      255      408
```

Para mayores detalles sobre el comando "cut" puede solicitar la ayuda correspondiente mediante:

```
help(cut)
```

Calcular nuevas variables

En el R es simple calcular una nueva variable. Por ejemplo, queremos ver si para los entrevistados, la edad ideal para que un hombre se case es mayor o menor que la edad ideal para que una mujer se case. Esto puede hacerse de la siguiente manera:

```
dif.idealH_M <- genero$edideal_hom - genero$edideal_muj
```

En este caso la variable dif-idealH-M mide la diferencia entre la edad ideal para casarse de un hombre y de una mujer. Si el número es positivo, significa que el entrevistado opina que la edad ideal para que un hombre se case debe ser mayor que la de la mujer. Si el número es negativo, significa lo inverso.

Selección de casos en R

En ciertos casos nos interesa trabajar con una selección de variables, o una selección de casos de una base de datos.

En esta presentación veremos algunos procedimientos para trabajar con una selección de casos.

Cargamos los datos de trabajo

Base de datos para estos ejercicios: Familia y roles de género 2012, a descargar de:

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

Se sugiere descargar también el cuestionario para utilizarlo como referencia de libro de códigos. 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
```

Seleccionar un subconjunto de variables

Queremos trabajar con un grupo reducido de variables de la base de datos. Por ejemplo: SEXO, EDAD, P1 y P2.

Paso 1: Creamos un vector con los nombres de las variables que queremos seleccionar

```
misvars <- c("SEXO", "EDAD", "P1", "P2")
```

Paso 2: Seleccionamos las variables del data frame original y las guardamos en un nuevo data frame:

```
nuevo.df <- genero[misvars]
names(nuevo.df)

## [1] "SEXO" "EDAD" "P1"    "P2"

table(nuevo.df$SEXO)

##
## Masculino  Femenino
##      589       614
```

Podemos obtener el mismo resultado, usando la función "subset"

```
nuevo.df2 <- subset(genero, selec=c(SEXO, EDAD, P1, P2))
names(nuevo.df2)

## [1] "SEXO" "EDAD" "P1"    "P2"
```

```
table(nuevo.df2$SEXO)
```

```
##  
## Masculino Femenino  
##      589      614
```

Excluir variables

En ciertas ocasiones podríamos estar interesados en excluir alguna variable de un data frame. Por ejemplo, excluir la variable EDAD de uno de los nuevos DF que hemos creado.

```
var.fuera <- names(nuevo.df) %in% c("EDAD")  
nuevo.df3 <- nuevo.df[!var.fuera]  
names(nuevo.df3)  
## [1] "SEXO" "P1"   "P2"
```

Seleccionar casos

Podemos estar interesados en trabajar con un subconjunto de casos de una base de datos. Supongamos que queremos generar un data frame solo de los casos de Lima y Callao.

```
genero.s1 <- subset(genero, Ambito=="Lima-Callao")  
length(genero.s1$NRO)  
## [1] 448
```

Otra posibilidad es seleccionar sólo a los hombres de Lima y Callao

```
genero.s2 <- subset(genero, Ambito=="Lima-Callao"  
                     & SEXO=="Masculino")  
length(genero.s2$NRO)  
## [1] 218
```

Podemos hacer una selección por rango de edad: las mujeres de Lima y Callao entre 25 y 45 años

```
genero.s3 <- subset(genero, Ambito=="Lima-Callao"  
                     & SEXO=="Femenino" &  
                     (EDAD >= 25 & EDAD <= 45))  
length(genero.s3$NRO)  
## [1] 104
```

Para mayores detalles sobre el uso de operadores lógicos en el R, sugerimos revisar la siguiente página web:

<http://www.statmethods.net/management/operators.html>

Análisis Univariable: Tablas de frecuencias

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

Base de datos para estos ejercicios: Familia y roles de género 2012, a descargar de:

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

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
## 0.4896093 0.5103907

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

##
## Masculino Femenino
## 48.96093 51.03907
```

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
## 48.96093 51.03907
```

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





```

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
##                      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
```

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

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 4
0 41 42
## 55 43 39 28 38 32 21 32 25 20 37 40 40 18 25 28 19 31 32 22 26 18 3
2 15 29
## 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 6
5 66 67
## 19 31 34 22 17 22 15 29 13 21 14 12 8 14 16 8 9 20 5 15 11 6 1
0 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                  3.404255  2.297872
## De acuerdo                      18.297872 13.787234
## Ni de acuerdo ni en desacuerdo  2.553191  2.808511
## En desacuerdo                   21.191489 27.234043
## Muy en desacuerdo                3.319149  5.106383

# 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                  59.70149 40.29851
## De acuerdo                      57.02918 42.97082
## Ni de acuerdo ni en desacuerdo 47.61905 52.38095
## En desacuerdo                  43.76098 56.23902
## Muy en desacuerdo                39.39394 60.60606

# 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                  6.980803 4.485050
## De acuerdo                      37.521815 26.910299
## Ni de acuerdo ni en desacuerdo 5.235602 5.481728
## En desacuerdo                  43.455497 53.156146
## Muy en desacuerdo                6.806283 9.966777
```

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 = F
ELSE)
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   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.

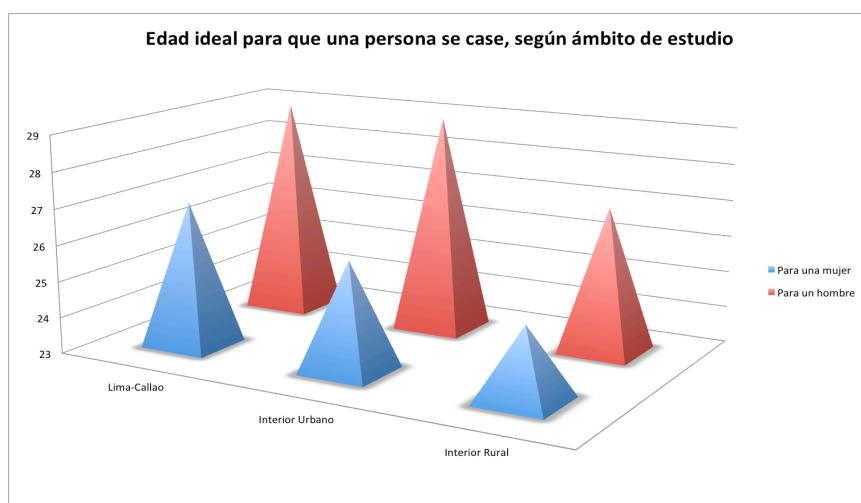
- Gráficos univariados para variables categóricas
 - Gráficos de barras
 - Gráficos de sector o "pies"
- Gráficos univariados para variables cuantitativas
 - Histogramas
 - Gráficos de líneas
 - Gráficos de cajas
- Gráficos bivariados
 - Variables categóricas: Barras múltiples
 - Variables cuantitativas: Líneas múltiples; gráficos de dispersión

Criterios para elaborar gráficos

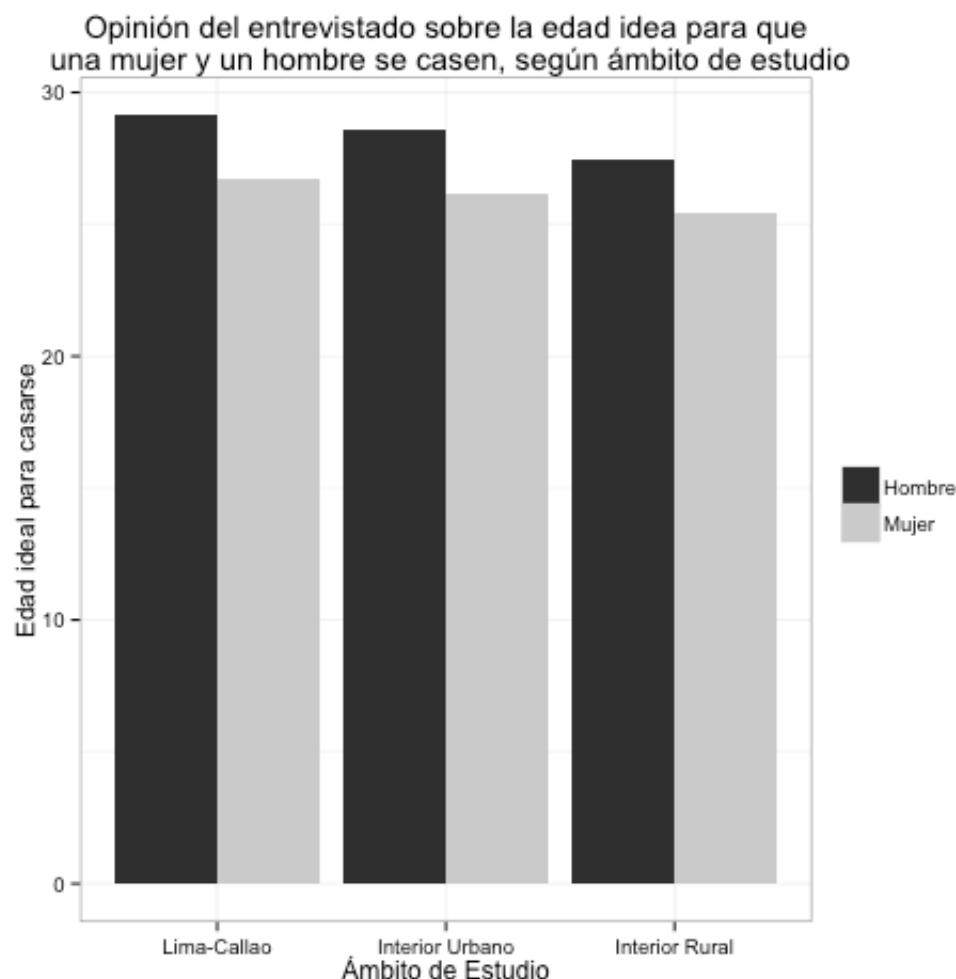
Un buen gráfico debe:

- 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

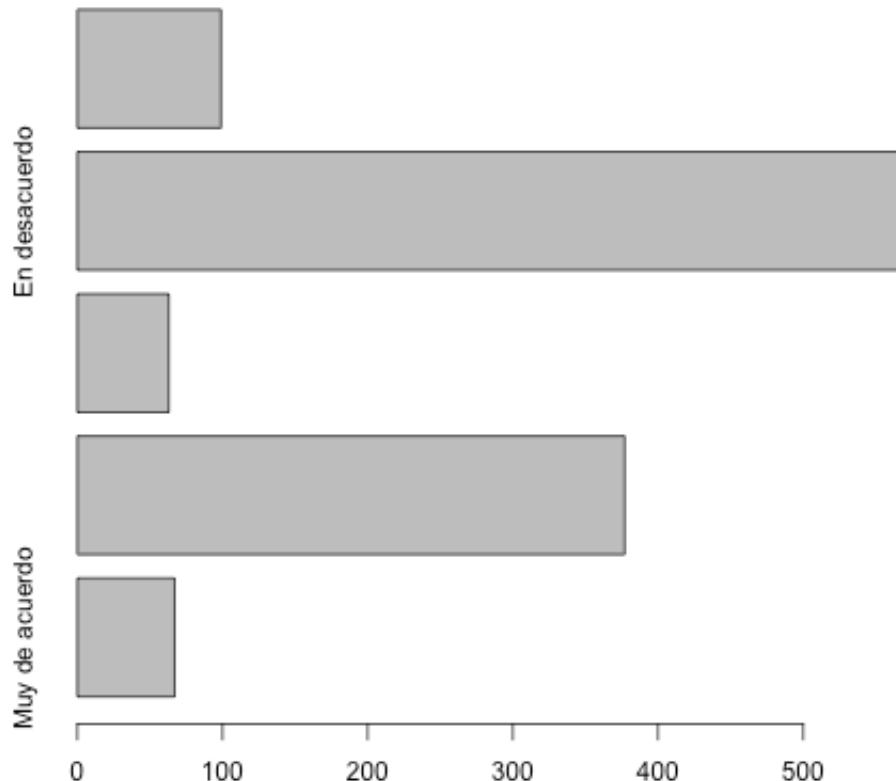
- El entorno gráfico básico
- El paquete lattice
- El paquete ggplot

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)
```

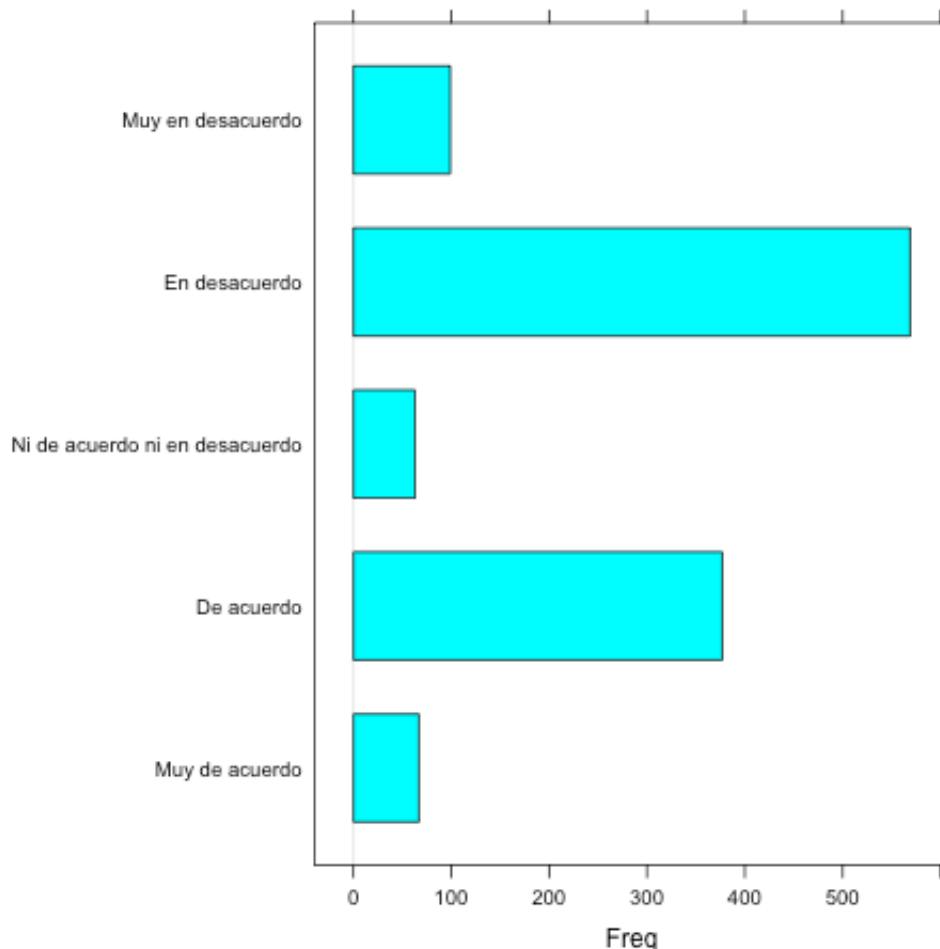
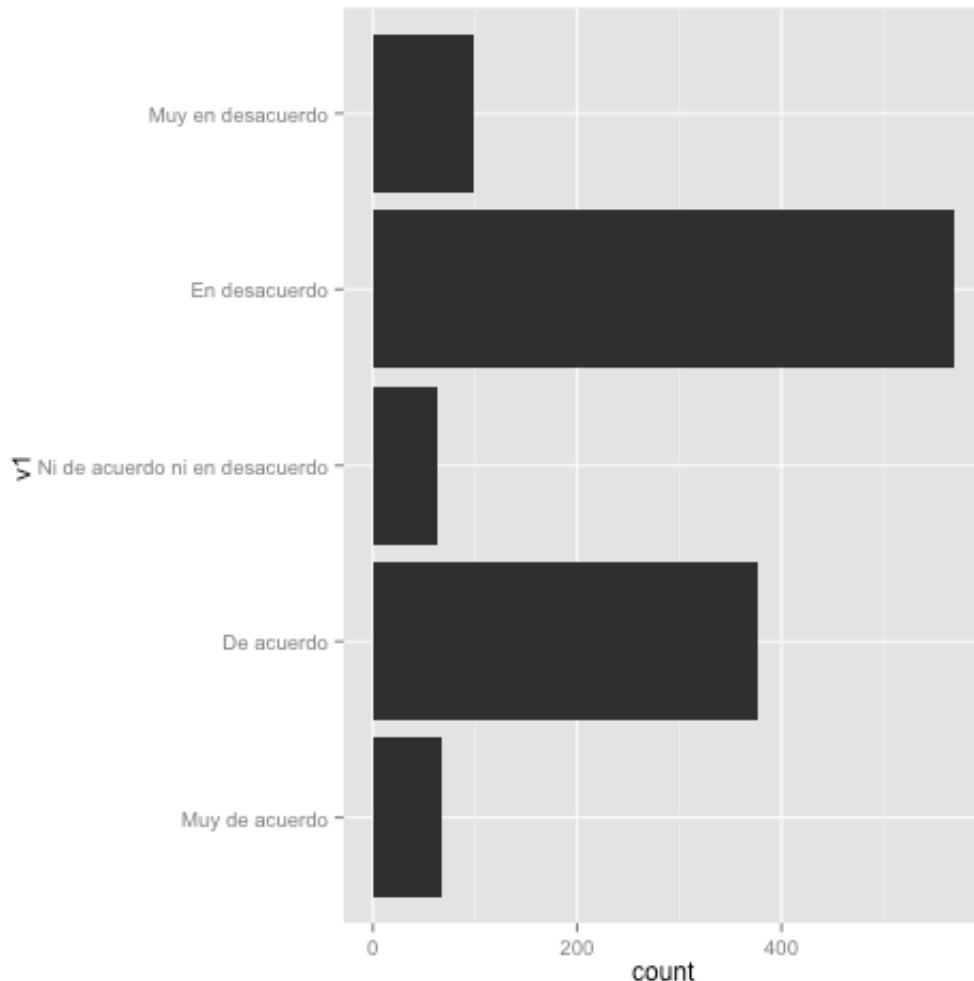


Gráfico usando el paquete ggplot

```
library(ggplot2)
v1 <- na.omit(genero$P3A.r)
graf5 <- qplot(v1, geom = "bar") + coord_flip()

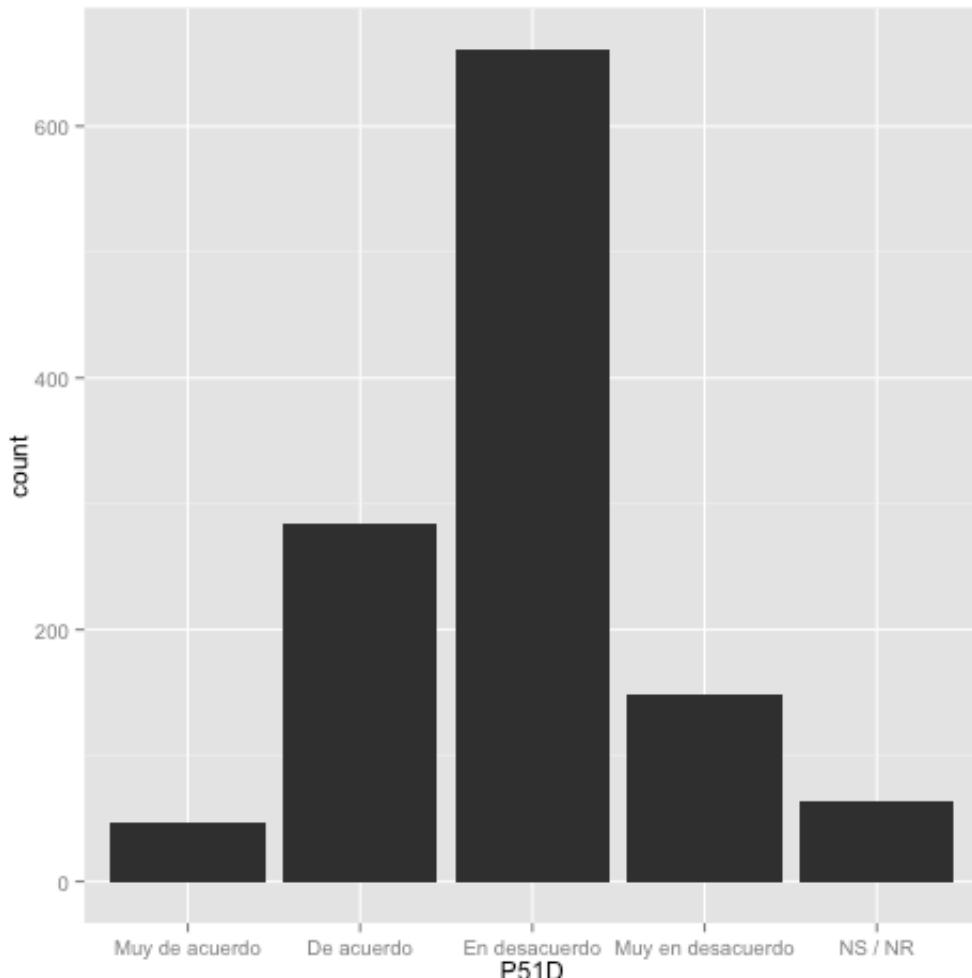
graf5
```



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:

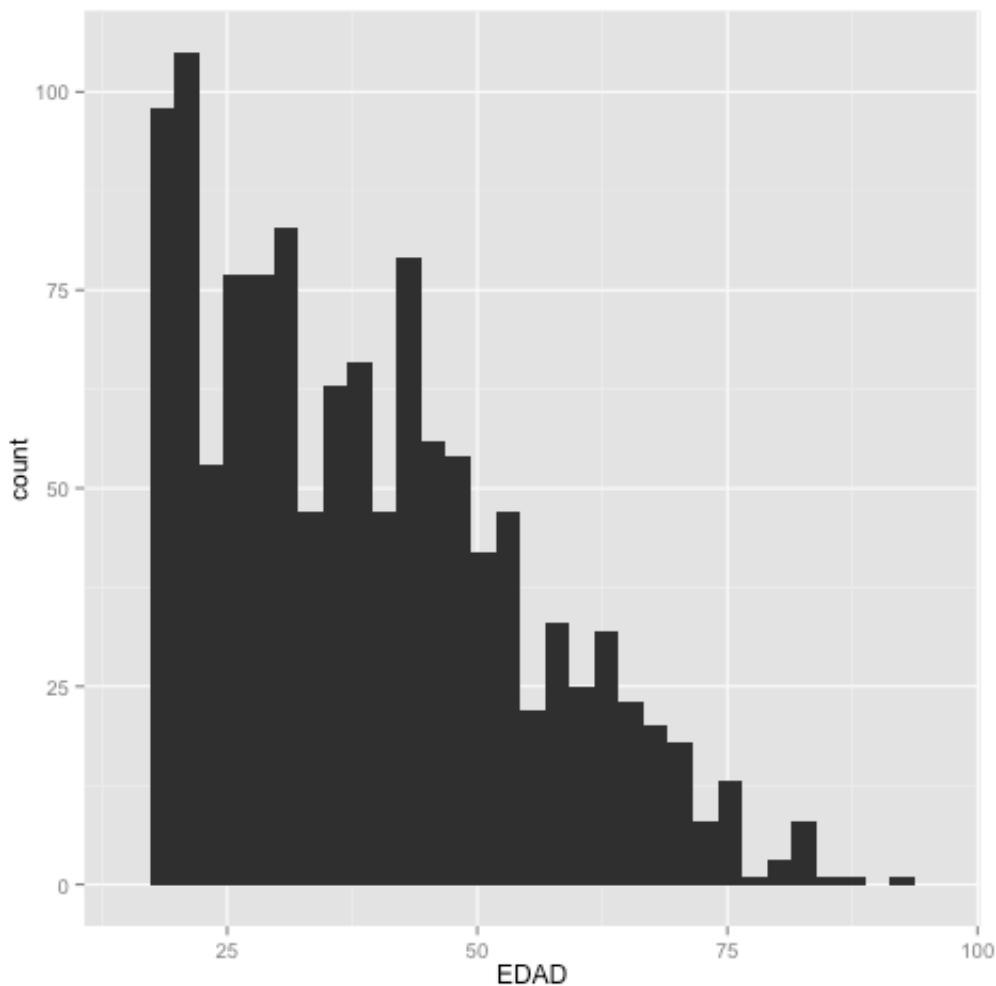
```
ggplot(genero, aes(P51D)) + geom_bar()
```



Representar una variable cuantitativa: Histograma

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

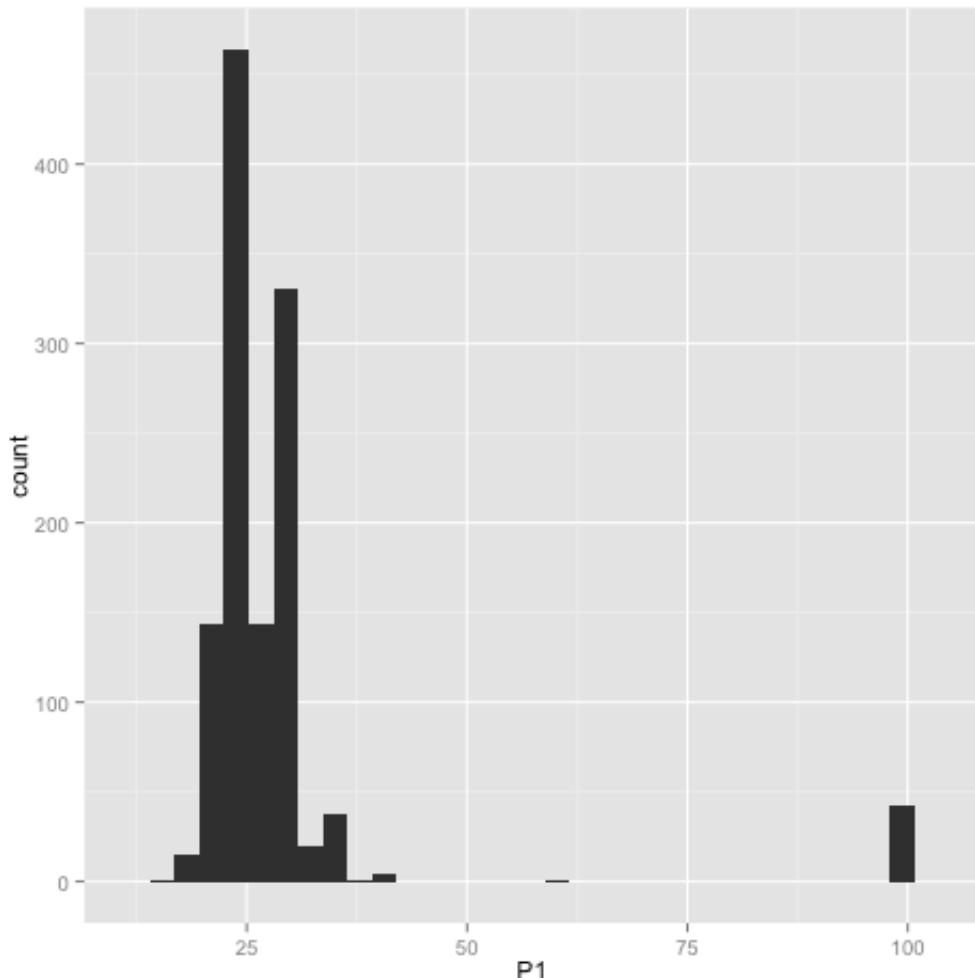
```
ggplot(genero, aes(EDAD)) + geom_histogram()
```



Utilidad del histograma: Casos atípicos

¿Qué está mal aquí?

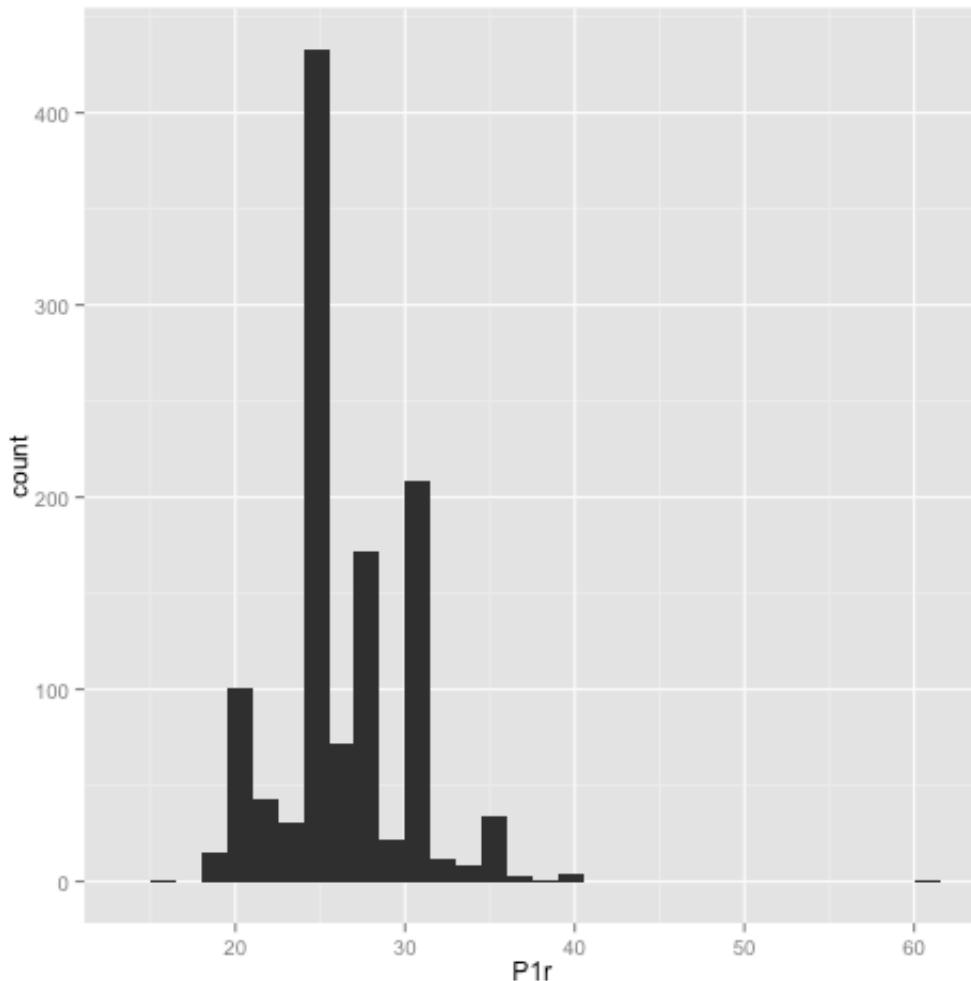
```
ggplot(genero, aes(P1)) + geom_histogram()
```



Corregimos:

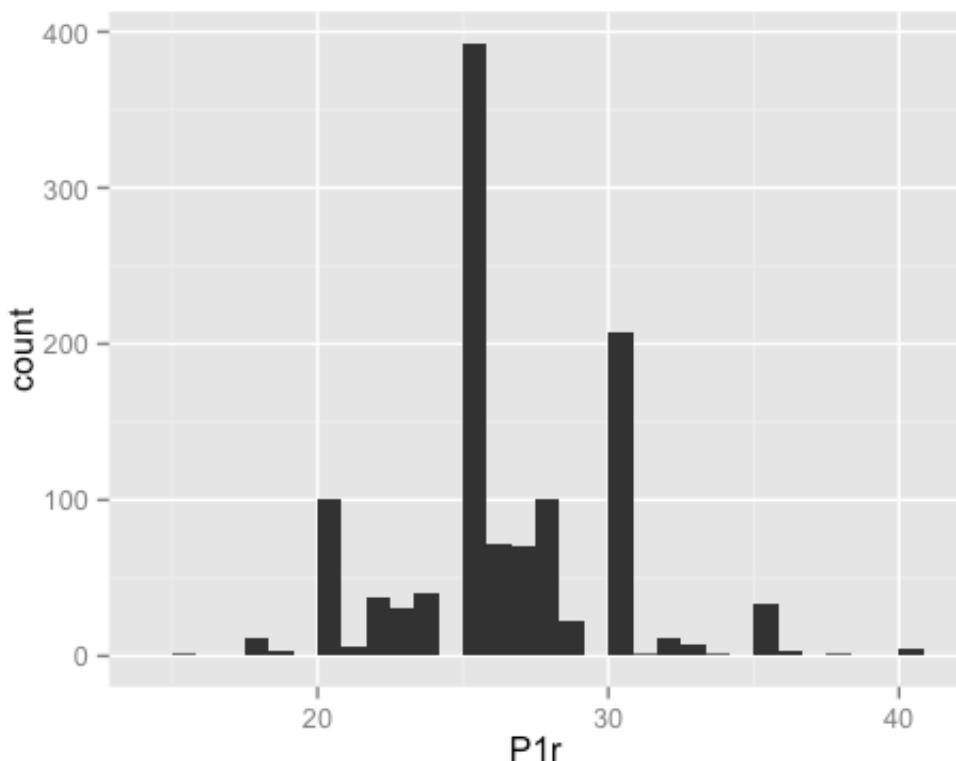
```
genero$P1r <- genero$P1
genero$P1r[genero$P1 == 99] <- NA # Designamos al código 99 como NA
hist3 <- ggplot(genero, aes(P1r)) + geom_histogram()

hist3
```



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()
```

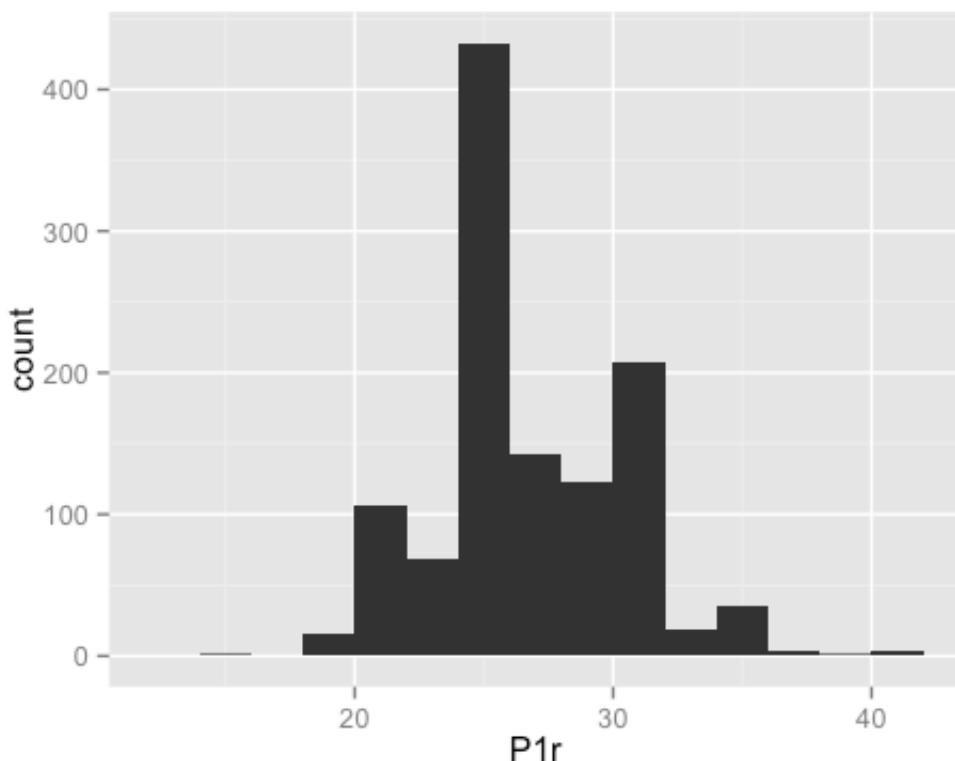


Amplitud de las barras en el histograma

- 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.

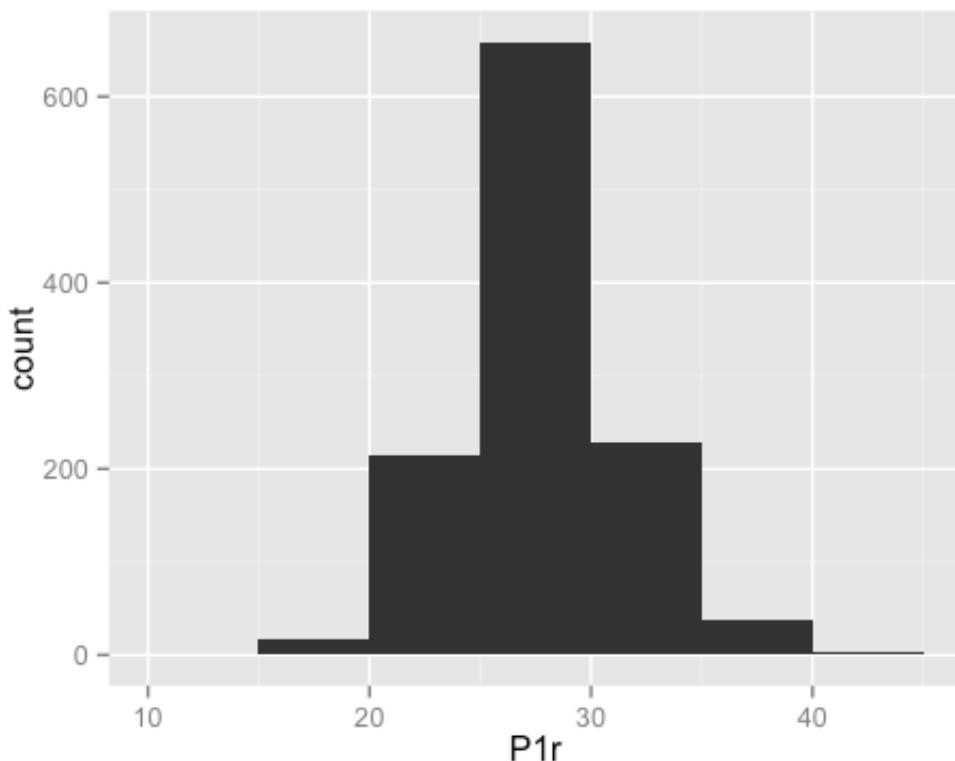
Cambiando la amplitud de las barras del histograma

```
#Amplitud 2
ggplot(genero, aes(P1r)) + geom_histogram(binwidth=2)
```



#Amplitud 5

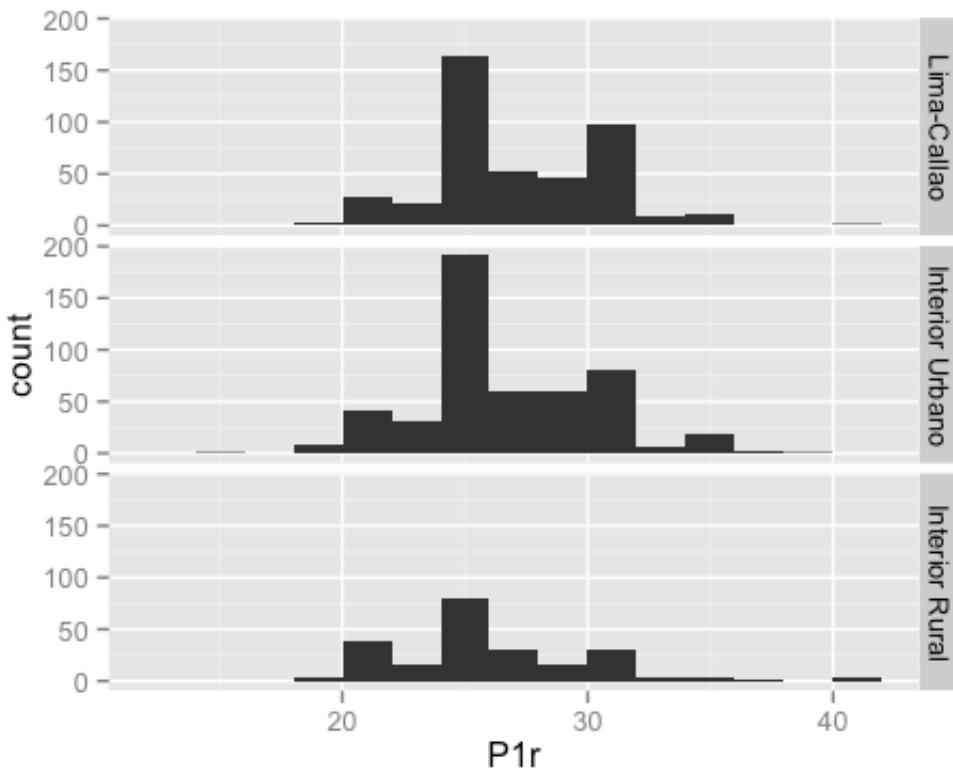
```
ggplot(genero, aes(P1r)) + geom_histogram(binwidth=5)
```



Comparar distribuciones

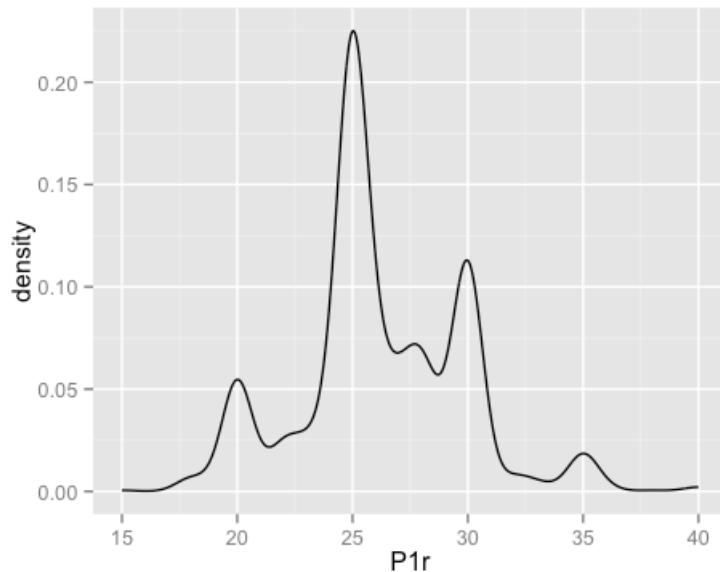
Con el histograma podemos comparar diferentes distribuciones de una variable cuantitativa:

```
ggplot(genero, aes(P1r)) + geom_histogram(binwidth=2) + facet_grid(Ambito ~.)
```

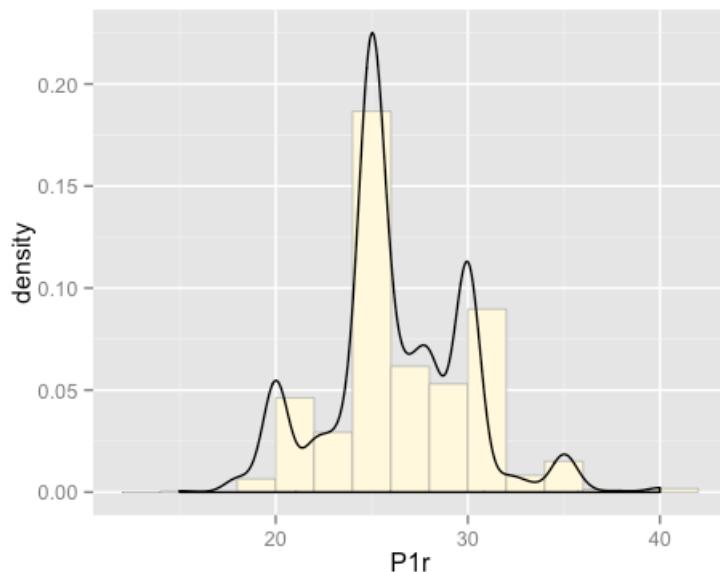


Curva de densidad de Kernel

La curva de densidad de Kernel nos permite observar gráficamente la distribución de una variable cuantitativa, "suavizando" los bordes de las barras del histograma. En el eje vertical se muestra la estimación de la probabilidad correspondiente a cada valor de la variable



Relación entre la curva de densidad y el histograma



También podemos comparar distribuciones con la curva de densidad:

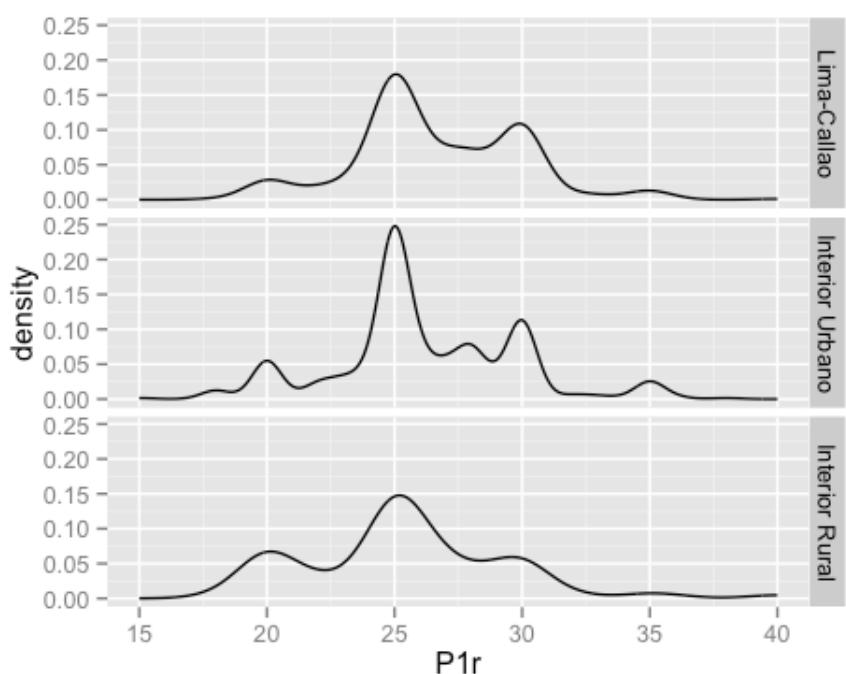
```
ggplot(genero, aes(P1r)) + geom_line(stat="density") + facet_grid(Ambi  
to ~.)
```



PUCP

ESTADÍSTICA PARA LAS CIENCIAS SOCIALES CON R

Profesor: David Sulmont



Gráficos con el paquete ggplot2

El paquete ggplot2

- Con el R es posible obtener el mismo resultado usando diferentes "caminos"
- El paquete ggplot es uno de los entornos gráficos del R
- Permite elaborar un gráfico a partir de un proceso de acumulación de **capas o layers**.
- Tiene un cierto nivel de complejidad pero se obtienen resultados muy profesionales.

Referencias bibliográficas

Textos disponibles en la biblioteca de CCSS de la PUCP para el uso de ggplot2

- Chang, Winston. 2012. *R Graphics Cookbook*. Sebastopol, CA: O'Reilly Media.
- Field, Andy P. 2012. *Discovering statistics using R*. London; Thousand Oaks, Calif: Sage.
- Wickham, Hadley. 2009. *Ggplot2: elegant graphics for data analysis*. New York: Springer.

Capas o layers en ggplot2

Un gráfico en ggplot2 puede tener varias capas, en su conjunto, las capas forman el gráfico al combinar:

- Un data frame y las variables a ser graficadas
- Una o varias capas indicando, entre otros:
 - El tipo de objeto a graficar o "geom" (barra, línea, punto, etc.)
 - Las transformaciones estadísticas a los datos
 - La posición de los objetos en el gráfico
- Una escala para cada variable a ser graficada
- Un sistema de coordenadas
- La especificación de "facetas" del gráfico

Cargamos los datos de trabajo

Base de datos para estos ejercicios: Familia y roles de género 2012, a descargar de:

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

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"))

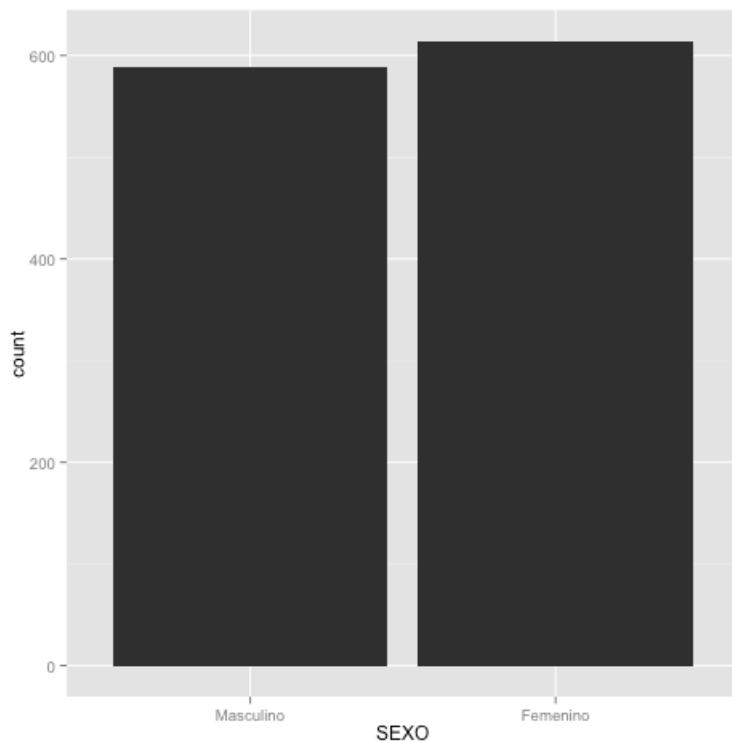
## re-encoding from UTF-8
```

Un gráfico simple: Gráfico de barras

El esquema básico de la gramática de ggplot es:

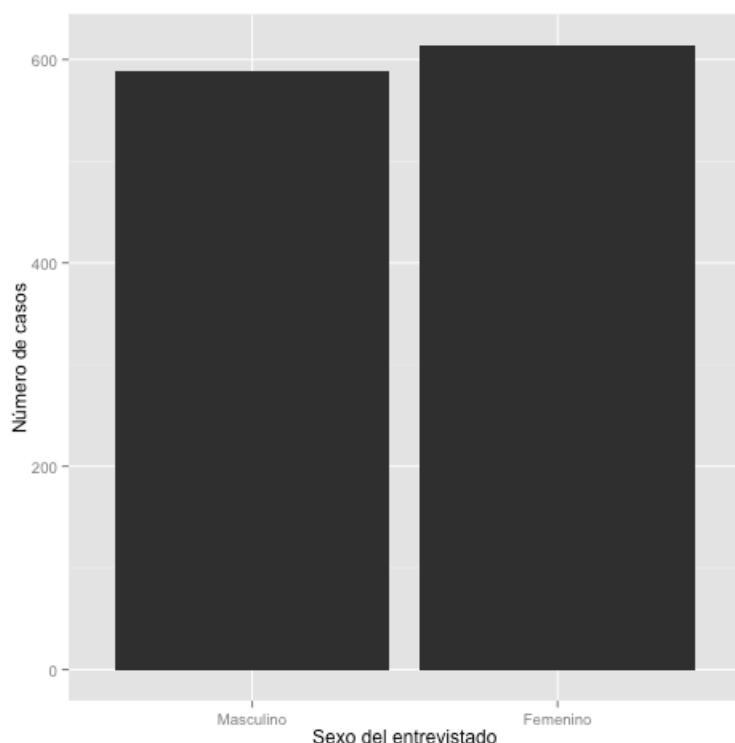
```
ggplot(data.frame, aes(x = variable)) + geom_forma()
```

```
library(ggplot2)
ggplot(genero, aes(x = SEXO)) + geom_bar()
```

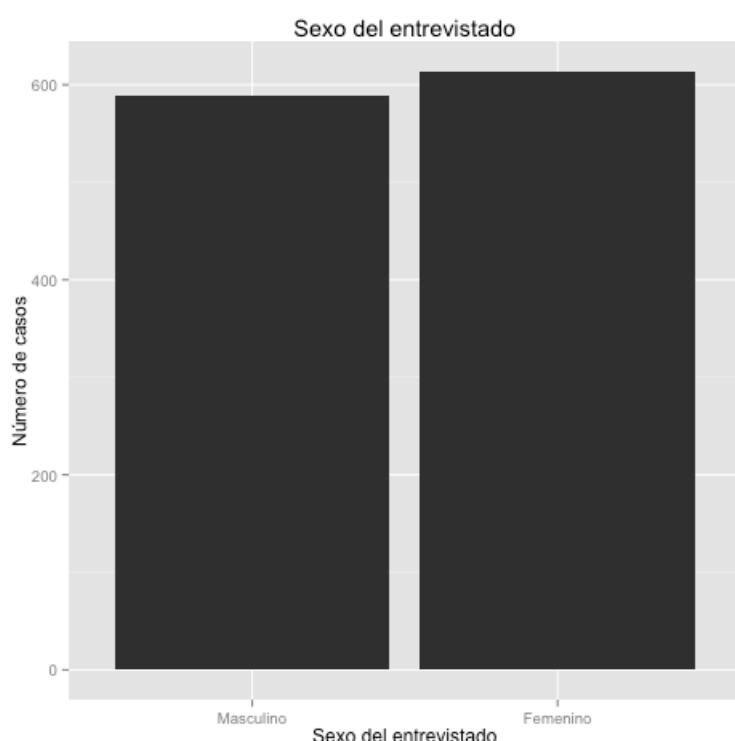


Podemos guardar el gráfico en un objeto y añadir más capas:

```
gr1 <- ggplot(genero, aes(x = SEXO)) + geom_bar()
gr1.1 <- gr1 + xlab("Sexo del entrevistado") + ylab ("Número de casos")
# etiquetas de los ejes
```

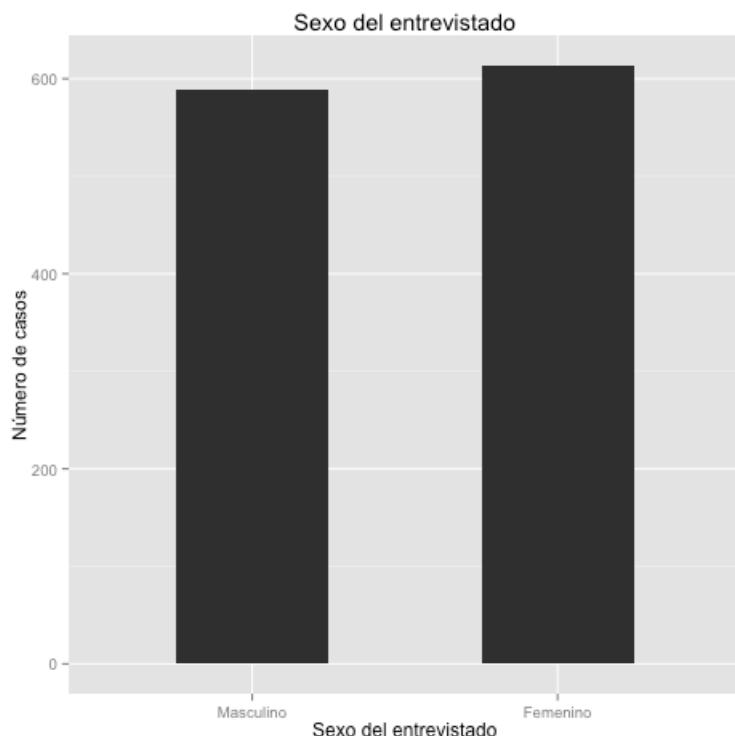


```
gr1.2 <- gr1 + xlab("Sexo del entrevistado") + ylab ("Número de casos")
) +
  ggtitle("Sexo del entrevistado")
gr1.2
```



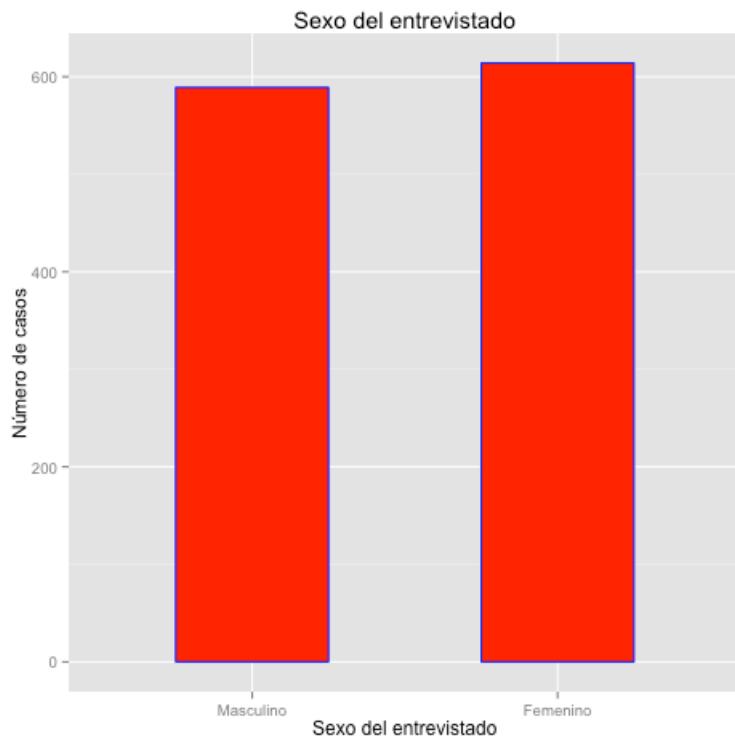
Podemos cambiar cómo se ven algunos elementos, por ejemplo, que las barras sean más pequeñas, en este caso la mitad (0.5) del tamaño por defecto:

```
gr1 <- ggplot(genero, aes(x = SEXO)) + geom_bar(width=0.5)
gr1.3 <- gr1 + xlab("Sexo del entrevistado") + ylab ("Número de casos")
) +
  ggtitle("Sexo del entrevistado")
gr1.3
```



Con la opción "colour" y "fill" en el comando **geom_bar** podemos cambiar el color del contorno de las barras y de su relleno

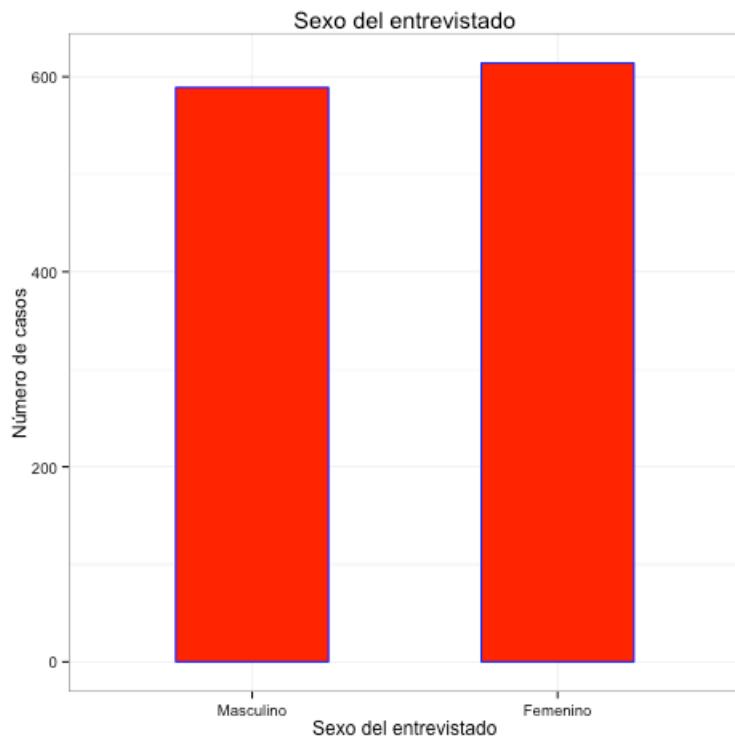
```
gr1 <- ggplot(genero, aes(x = SEXO)) + geom_bar(width=0.5, colour="blue",
  fill="red")
gr1.4 <- gr1 + xlab("Sexo del entrevistado") + ylab ("Número de casos")
) +
  ggtitle("Sexo del entrevistado")
gr1.4
```



Temas

Los temas (theme) son un conjunto de opciones predefinidas sobre la apariencia de los objetos en ggplot. El tema por defecto del ggplot dibuja el gráfico sobre un fondo gris. Podemos cambiarlo a blanco y negro añadiendo el comando `theme_bw()`

```
gr1 <- ggplot(genero, aes(x = SEXO)) + geom_bar(width=0.5, colour="blue", fill="red")
gr1.5 <- gr1 + xlab("Sexo del entrevistado") + ylab ("Número de casos")
+ ggttitle("Sexo del entrevistado") + theme_bw()
gr1.5
```



Complicando un poco más el asunto...

Hacer un gráfico de barras de sexo, pero con porcentajes en el eje vertical. Primero preparar los datos:

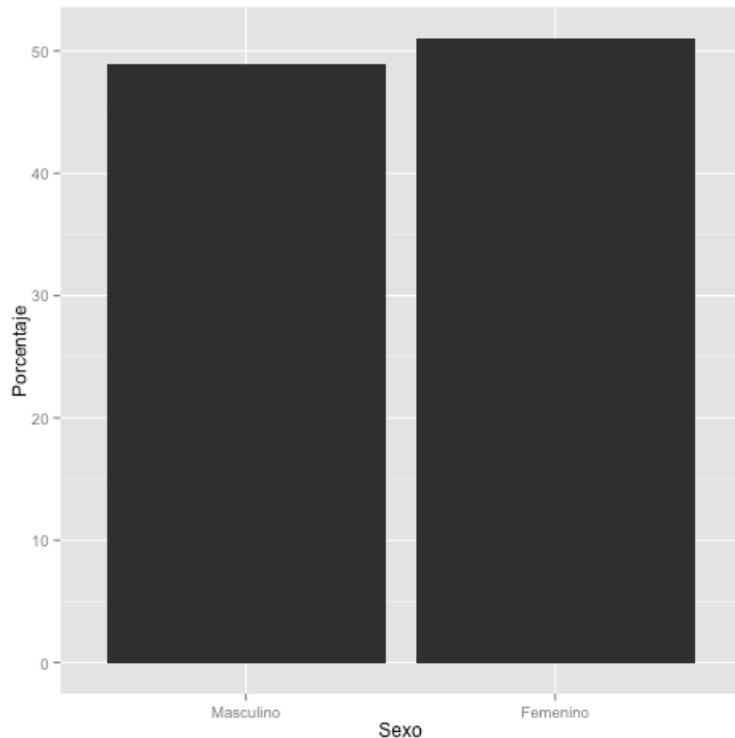
```
tab.sex1 <- as.data.frame(prop.table(table(genero$SEXO))*100)
tab.sex1

##      Var1     Freq
## 1 Masculino 48.96093
## 2 Femenino 51.03907

colnames(tab.sex1) <- c("Sexo", "Porcentaje")
```

Ahora el gráfico:

```
gr2 <- ggplot(tab.sex1, aes(x=Sexo, y=Porcentaje)) + geom_bar(stat="identity")
gr2
```

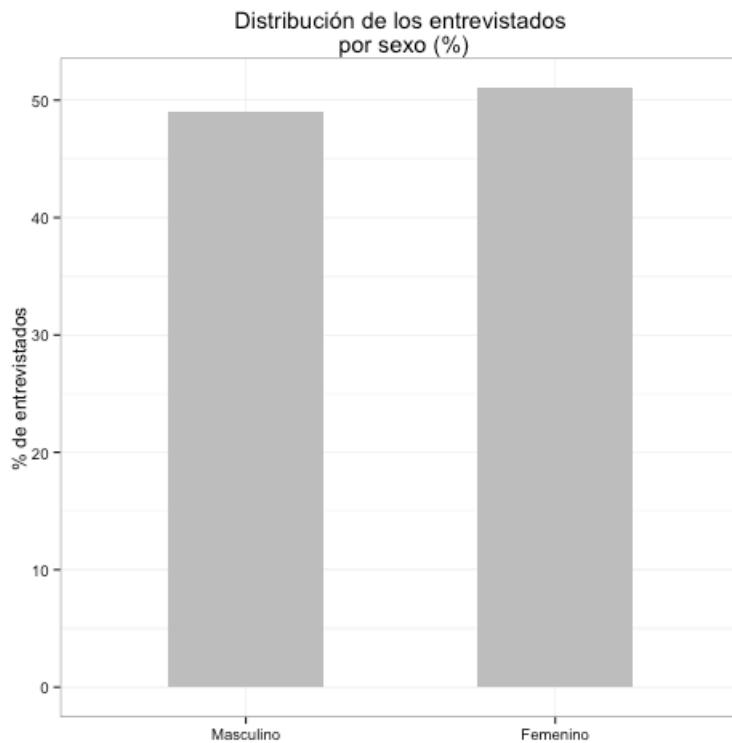


Afinamos el gráfico:

```
gr2 <- ggplot(tab.sex1, aes(x=Sexo, y = Porcentaje)) +
  geom_bar(stat="identity", width=0.5, fill = "grey")

gr2.1 <- gr2 + xlab(NULL) + ylab("% de entrevistados") +
  ggtitle("Distribución de los entrevistados\n por sexo (%)") + theme_bw()

gr2.1
```



Otra manera de llegar al mismo resultado

```
library(scales) # requiere instalar el paquete "scales"
gr2.a <- ggplot(genero, aes(SEXO)) +
  geom_bar(aes(SEXO,(..count..)/sum(..count..)), width=0.5, fill = "grey")

gr2.a2<- gr2.a + scale_y_continuous(labels=percent) + xlab(NULL) +
  ylab("% de casos") + ggtitle("Distribución de los\nn entrevistados po
r sexo") + theme_bw()

gr2.a2
```

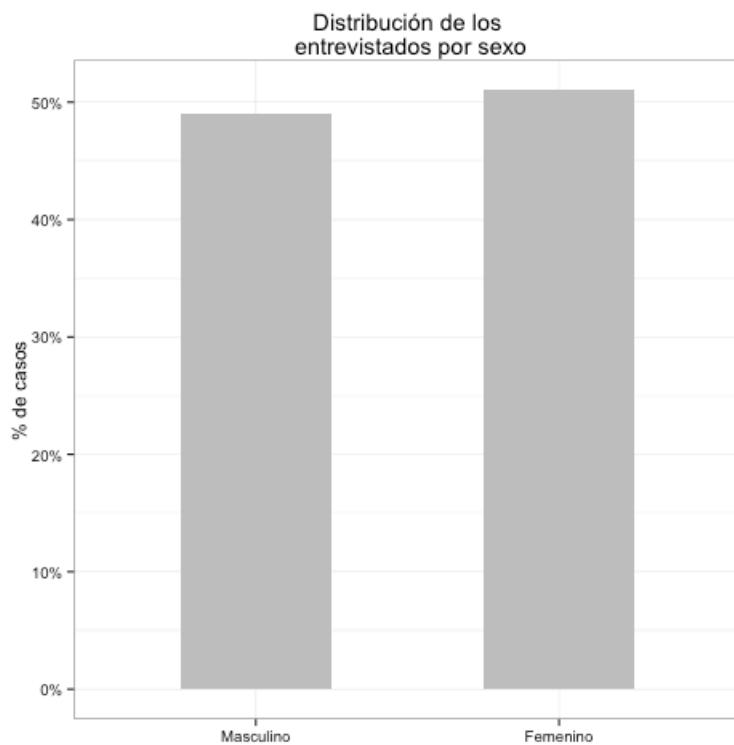
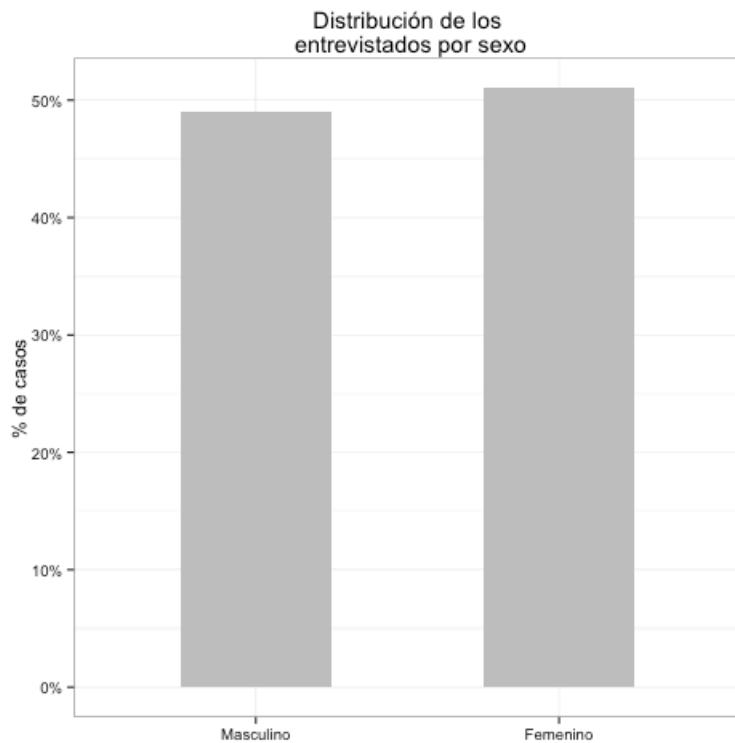


Gráfico de barras múltiples

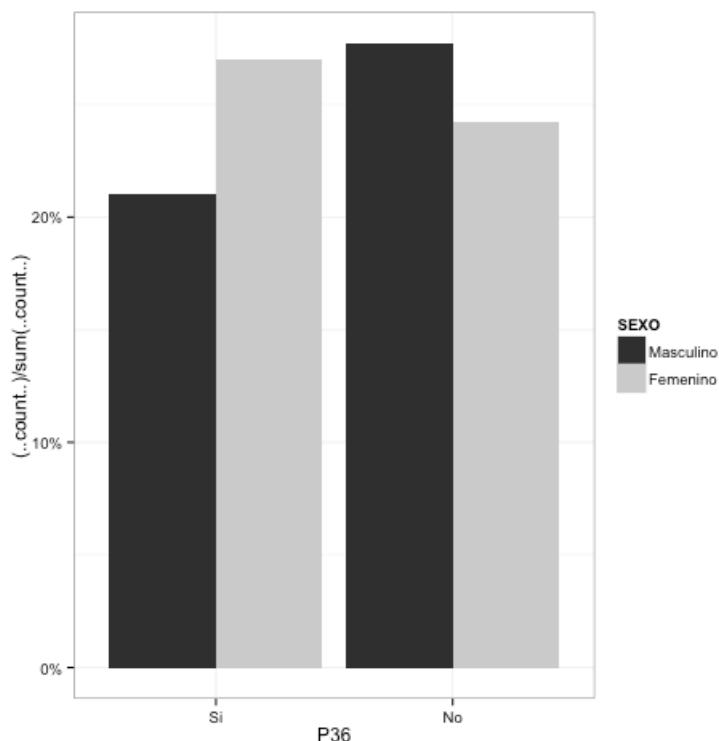
La distribución de frecuencias de la pregunta P36 según sexo

```
gr3 <- ggplot(genero, aes(P36, fill=SEXO)) + geom_bar(position="dodge")
gr3
```



¿Cómo quitamos "NS/NR" y establecemos que la escala del eje vertical sean %?

```
gr3.1 <- ggplot(genero[genero$P36!="NS/NR", ], aes(P36, fill=SEXO)) +  
  geom_bar(aes(P36,(..count..)/sum(..count..)), position="dodge") +  
  scale_y_continuous(labels=percent) + scale_fill_grey() + theme_bw()  
gr3.1
```

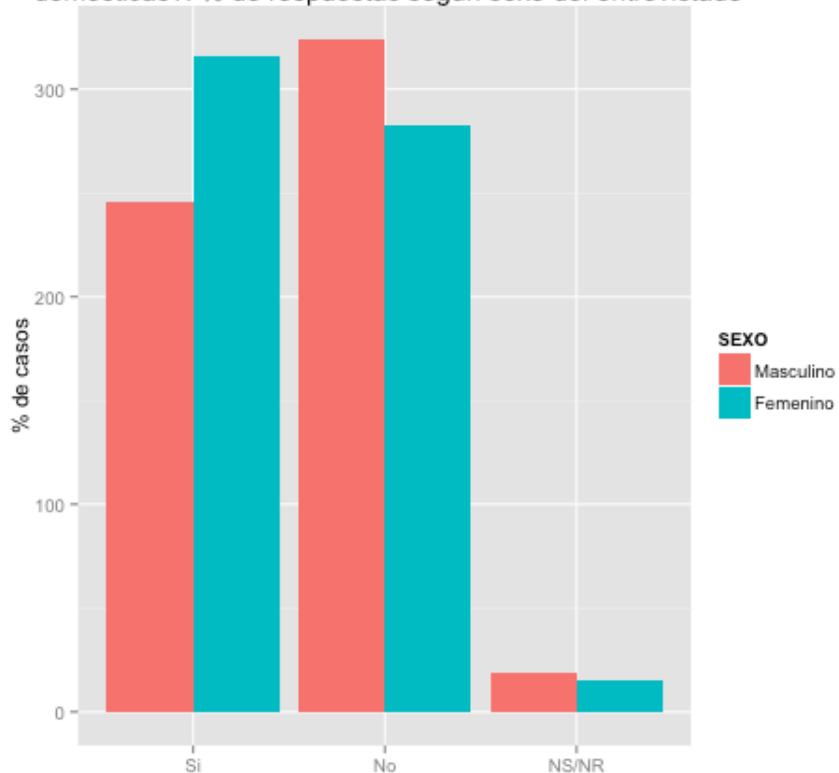


Completamos el gráfico

```
gr3.2 <- gr3 + xlab(NULL) + ylab("% de casos") + ggtitle("¿Los demás miembros del hogar deberían pagarle un sueldo o salario al miembro de 1 hogar que se encarga de las tareas domésticas?: % de respuestas según sexo del entrevistado")
```

```
gr3.2
```

¿Los demás miembros del hogar deberían pagarle un sueldo o salario al miembro del hogar que se encarga de las tareas domésticas?: % de respuestas según sexo del entrevistado



Usando facets

Las facetas o "facets" en ggplot me permiten reproducir el mismo gráfico en diferentes niveles de un factor. Hagamos un gráfico de la distribución en % de la pregunta P51D, para los diferentes dominios geográficos

```
# Primero preparamos Los datos en una tabla que convertimos en data frame
tab.5 <- as.data.frame(prop.table(table(genero$P51D, genero$DOMINIO),
  2)*100)
tab.5

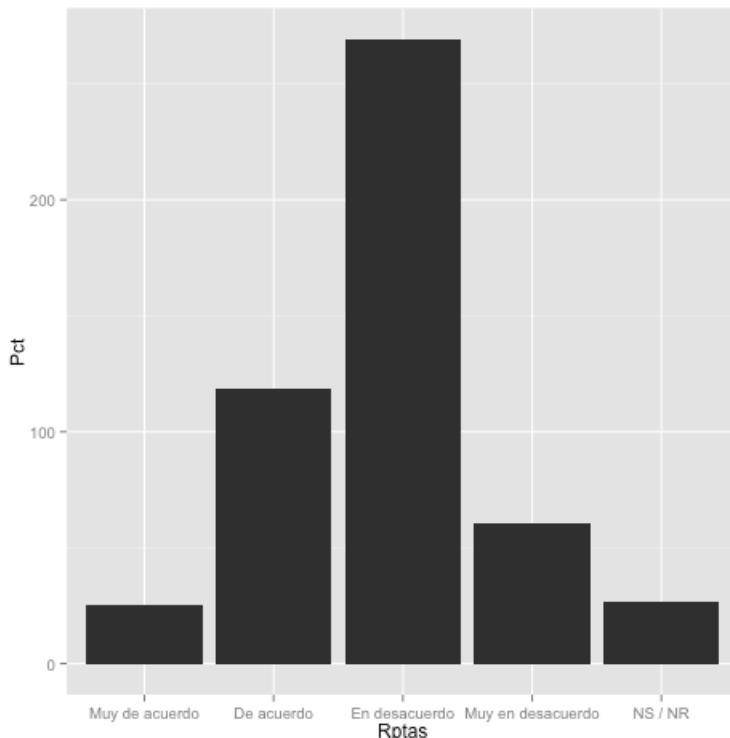
##                               Var1      Var2     Freq
## 1      Muy de acuerdo Lima-Callao 0.8928571
## 2          De acuerdo Lima-Callao 18.7500000
## 3      En desacuerdo Lima-Callao 62.0535714
## 4  Muy en desacuerdo Lima-Callao 13.1696429
## 5          NS / NR Lima-Callao  5.1339286
## 6      Muy de acuerdo        Norte 3.4375000
## 7          De acuerdo        Norte 25.9375000
## 8      En desacuerdo        Norte 55.3125000
## 9  Muy en desacuerdo        Norte 10.0000000
## 10         NS / NR        Norte  5.3125000
## 11      Muy de acuerdo       Sur  7.3469388
```

```
## 12      De acuerdo      Sur 31.0204082
## 13    En desacuerdo      Sur 43.6734694
## 14 Muy en desacuerdo      Sur 13.0612245
## 15          NS / NR      Sur 4.8979592
## 16 Muy de acuerdo Centro 13.3333333
## 17      De acuerdo Centro 22.8571429
## 18    En desacuerdo Centro 35.2380952
## 19 Muy en desacuerdo Centro 20.9523810
## 20          NS / NR Centro 7.6190476
## 21 Muy de acuerdo Oriente 0.0000000
## 22      De acuerdo Oriente 20.0000000
## 23    En desacuerdo Oriente 72.9411765
## 24 Muy en desacuerdo Oriente 3.5294118
## 25          NS / NR Oriente 3.5294118
```

```
colnames(tab.5) <- c("Rptas", "Dominio", "Pct")
```

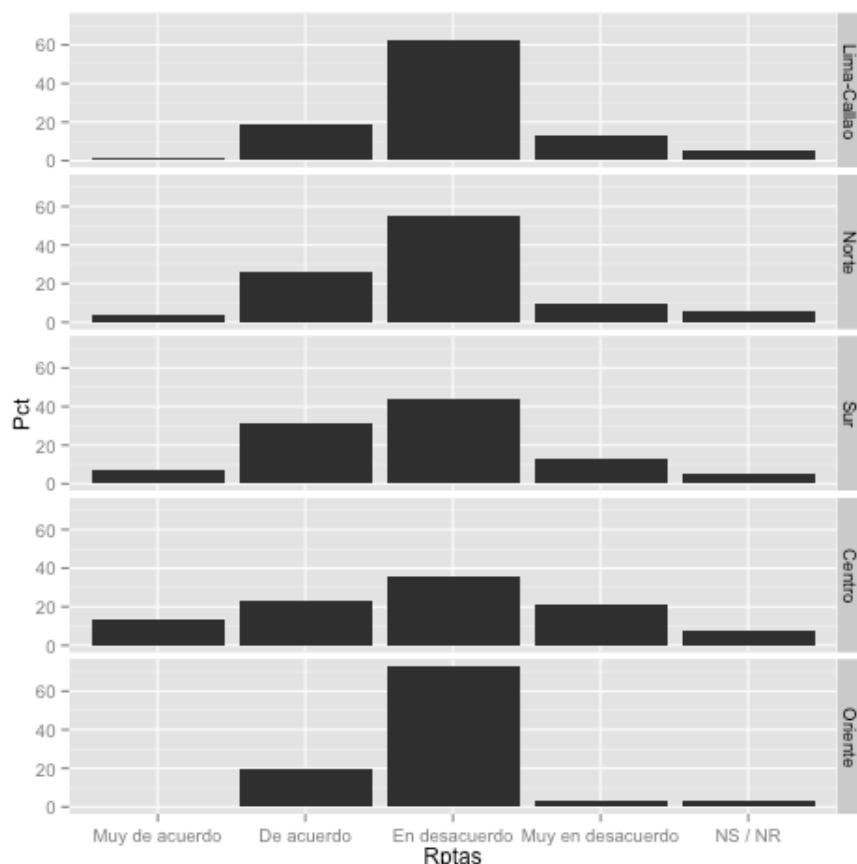
Este sería en gráfico de base:

```
gr4 <- ggplot(tab.5, aes(x=Rptas, y=Pct)) + geom_bar(stat="identity")
gr4
```



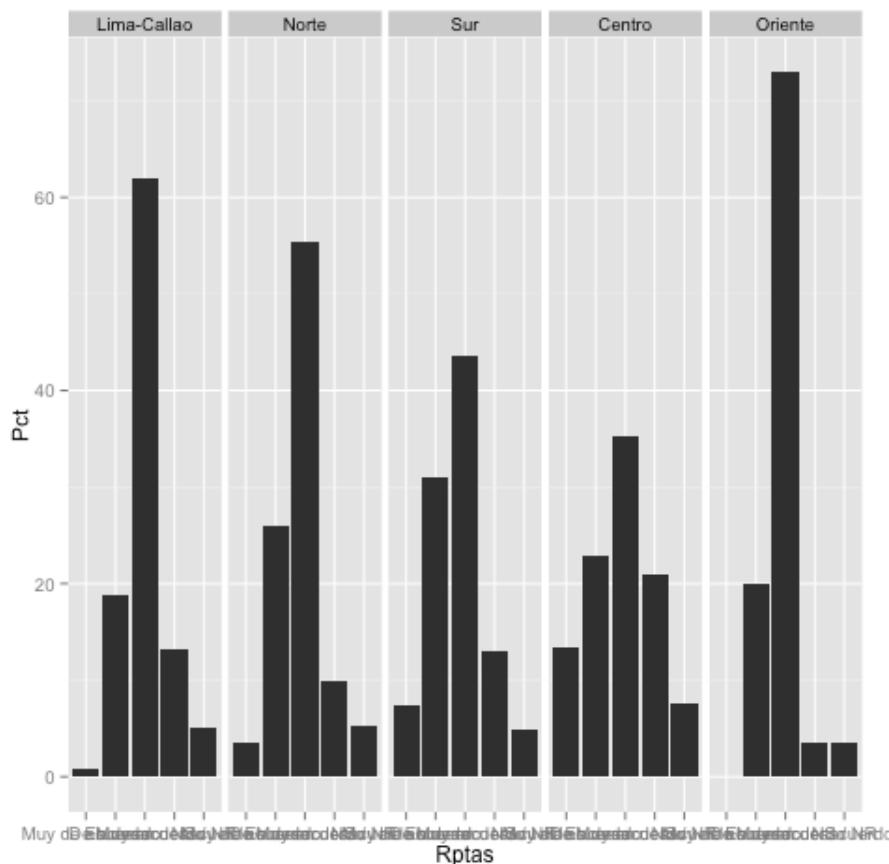
Hacemos las facetas verticales

```
gr4.v <- gr4 + facet_grid(Dominio ~.)
gr4.v
```



Si las queremos horizontales:

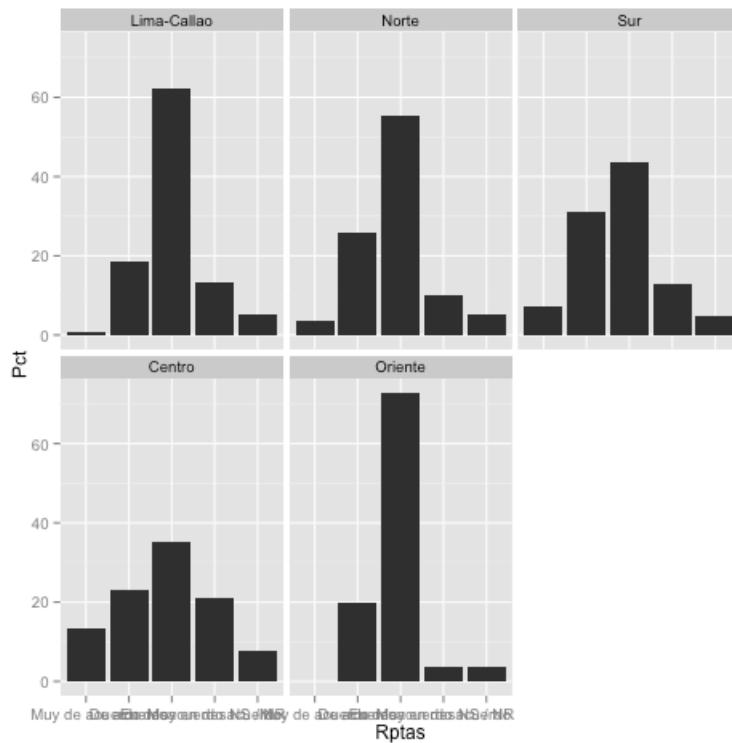
```
gr4.h <- gr4 + facet_grid(.~ Dominio)  
gr4.h
```



Si queremos que rotén a lo largo de columnas y filas:

```
gr4.r <- gr4 + facet_wrap(~ Dominio)
```

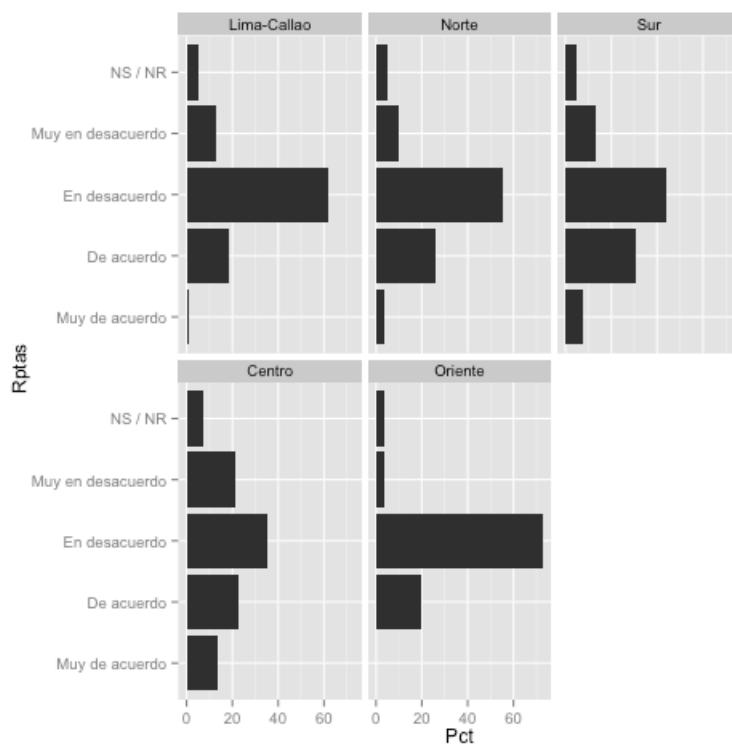
```
gr4.r
```



Podemos rotar el gráfico para que se aprecien mejor las etiquetas de respuestas

```
gr4.r2<- gr4 + coord_flip() + facet_wrap(~ Dominio)
```

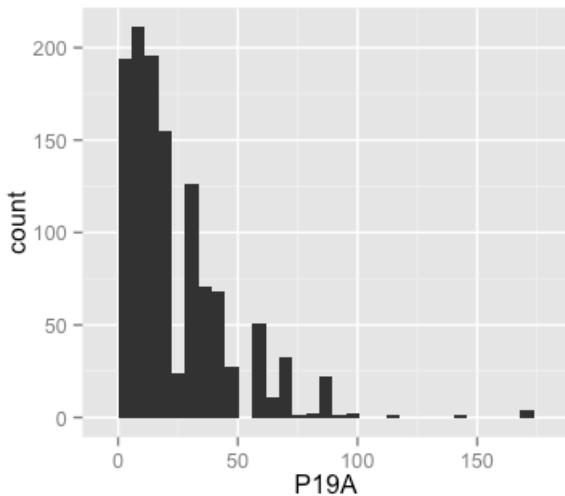
```
gr4.r2
```



Histogramas

Histograma básico de la pregunta P19A

```
ggplot(genero, aes(P19A)) + geom_histogram()
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

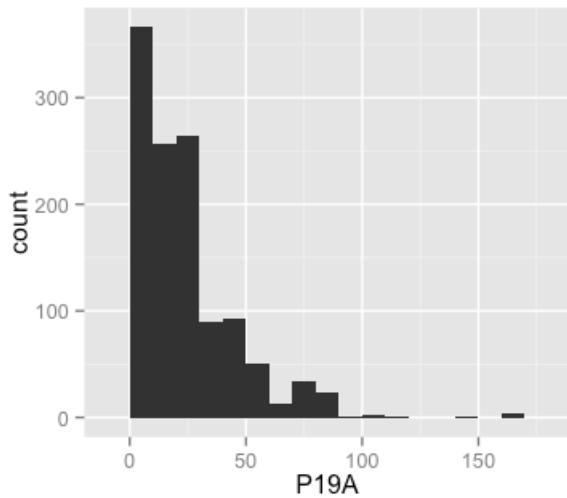


Amplitud de los intervalos del histograma

```
range(genero$P19A, na.rm=TRUE)
## [1] 0 168
168/30
## [1] 5.6
```

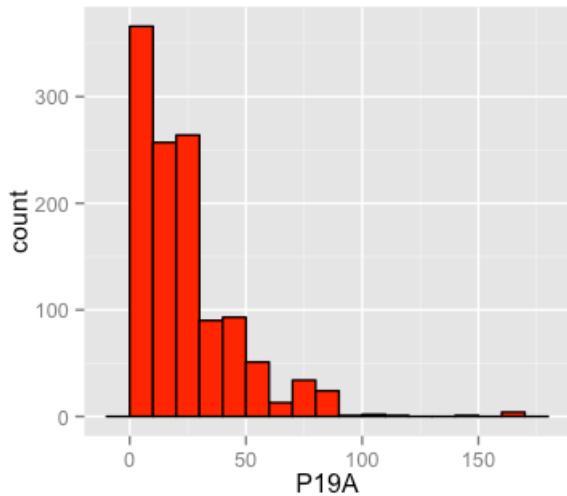
Cambiamos la amplitud del intervalo

```
ggplot(genero, aes(P19A)) + geom_histogram(binwidth = 10)
```



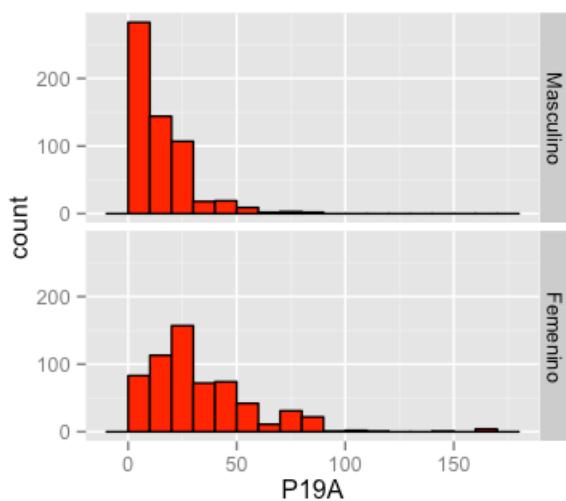
Algunas mejoras:

```
hist1 <- ggplot(genero, aes(P19A)) + geom_histogram(binwidth = 10,
                                                    fill="red", colour="black")
hist1
```



Podemos comparar el histograma de los hombres y de las mujeres

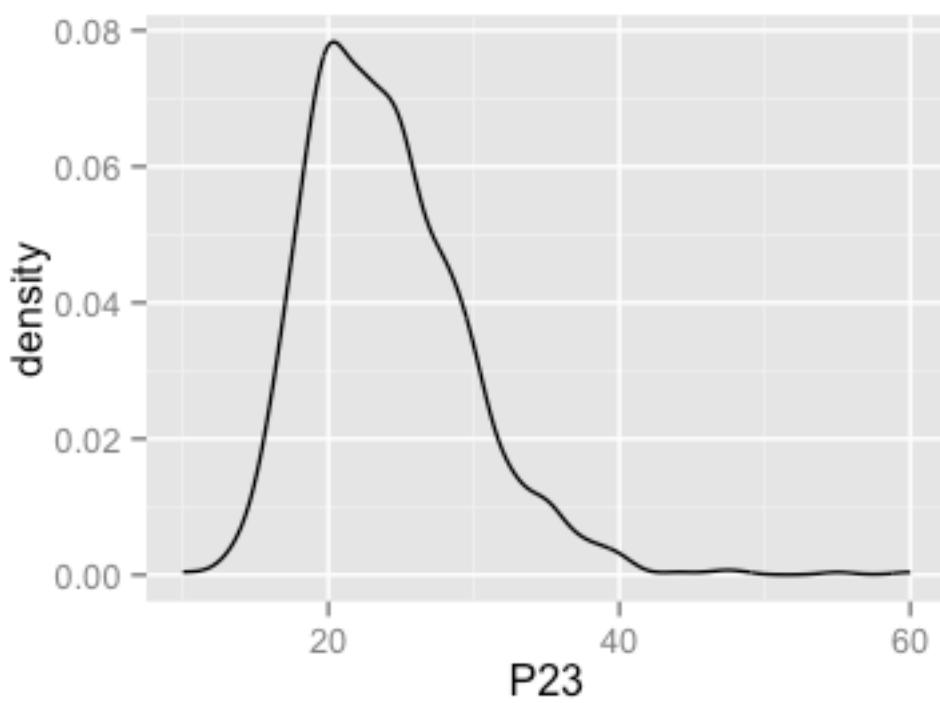
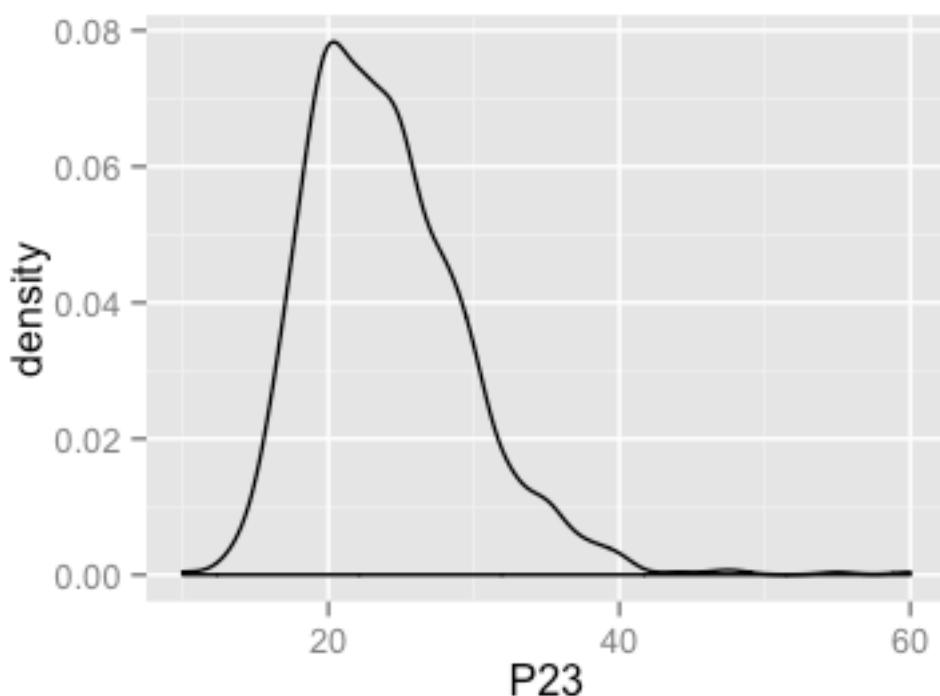
```
hist1 + facet_grid(SEXO ~.)
```



Curvas de Densidad de Kernel

Otra forma de ver la distribución de una variable cuantitativa, sobre la base del cálculo de densidades a partir del histograma:

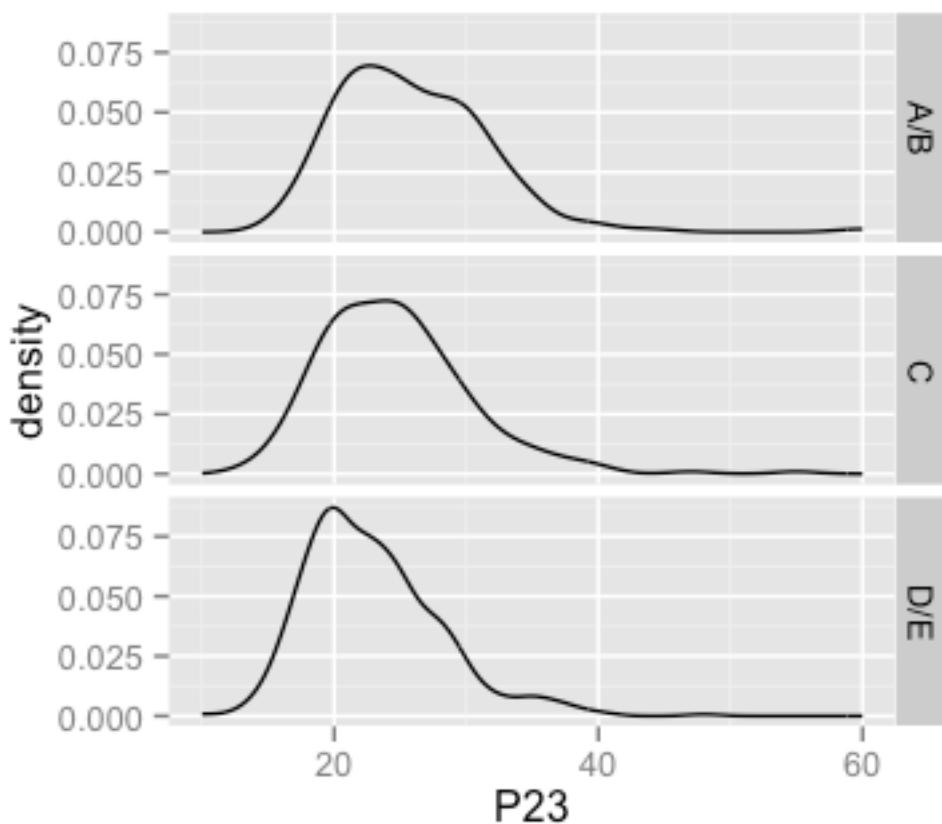
```
## Las sintaxis nos dan el mismo resultado. La segunda evita dibujar 1
a línea de abajo
ggplot(genero, aes(P23)) + geom_density()
ggplot(genero, aes(P23)) + geom_line(stat="density")
```



Curvas de densidad para grupos diferentes

Distribución de la edad en la que se casó según NSE

```
ggplot(genero, aes(P23)) + geom_line(stat="density") + facet_grid(NSEG ~.)
```



Distribución de la edad en la que se casó según NSE y Sexo

```
ggplot(genero, aes(P23)) + geom_line(stat="density") + facet_grid(NSEG  
rup ~ SEXO)
```

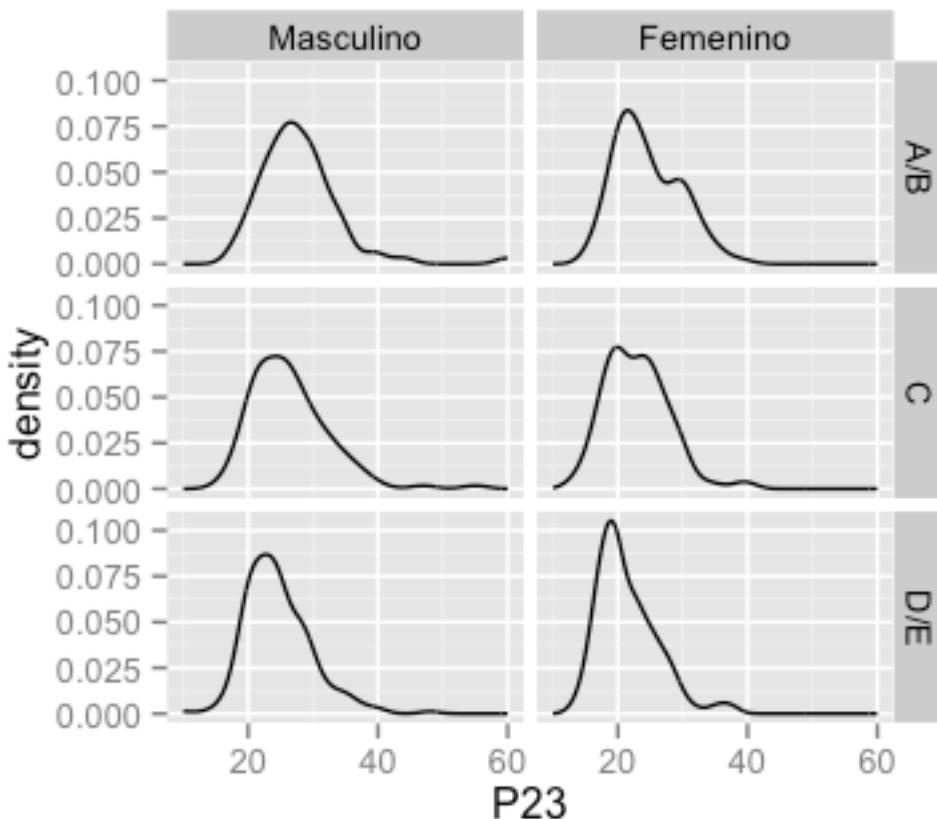


Gráfico de líneas

Objetivo: generar un gráfico de líneas que muestre cómo varía el % de personas que están en desacuerdo con la frase "Jamás tendría un amigo homosexual o una amiga lesbiana" (P51D)

Primero preparamos los datos:

Paso 1: Crear una variable categórica que agrupe las edades de los entrevistados en intervalos

```
edad <- genero$EDAD
gr.edad <- as.factor(cut(edad, breaks = c(18, 25, 35, 45, 55, 92),
                           include.lowest = TRUE))
table(gr.edad)

## gr.edad
## [18,25] (25,35] (35,45] (45,55] (55,92]
##     288      283      258      173      201
```

Paso 2: Recodificamos la variable P51D en dos grupos, lo que están de acuerdo, los que están en desacuerdo. Los NS/NR los pasamos como NA

```
table(genero$P51D)
```

```

## 
##      Muy de acuerdo      De acuerdo      En desacuerdo Muy en desacuer
erdo
##                      47                  284                  661
148
##          NS / NR
##                      63

p51dr <- as.numeric(genero$P51D)
table(p51dr)

## p51dr
##   1   2   3   4   5
##  47 284 661 148  63

library(car)
p51dr <- factor(recode(p51dr, "1:2=1; 3:4=2; 5=NA")) # recodificamos
levels(p51dr) <- c("De acuerdo", "En desacuerdo")

```

Paso 3: Generamos una tabla de % cruzados de respuestas a la pregunta P51D según grupos de edad y la convertimos en un data frame

```

tab1 <- prop.table(table(gr.edad, p51dr), 1)*100
tab1

##           p51dr
## gr.edad  De acuerdo En desacuerdo
##   [18,25]  21.58273  78.41727
##   (25,35]  26.56827  73.43173
##   (35,45]  29.33884  70.66116
##   (45,55]  35.22013  64.77987
##   (55,92]  37.89474  62.10526

```

Paso 4: Convertimos la tabla generada en un data frame

```

df.tab1 <- data.frame(tab1)
df.tab1

##   gr.edad      p51dr     Freq
## 1 [18,25]  De acuerdo 21.58273
## 2 (25,35]  De acuerdo 26.56827
## 3 (35,45]  De acuerdo 29.33884
## 4 (45,55]  De acuerdo 35.22013
## 5 (55,92]  De acuerdo 37.89474
## 6 [18,25] En desacuerdo 78.41727
## 7 (25,35] En desacuerdo 73.43173
## 8 (35,45] En desacuerdo 70.66116
## 9 (45,55] En desacuerdo 64.77987
## 10 (55,92] En desacuerdo 62.10526

```

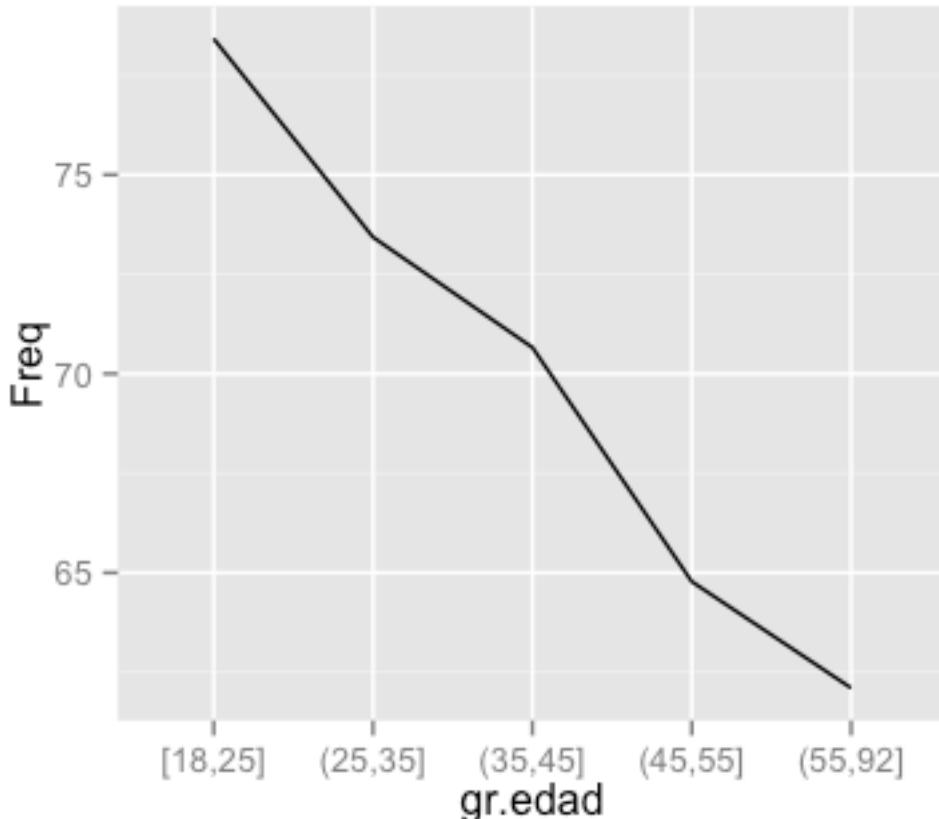
Nos interesa representar sólo los % en desacuerdo. Para ello seleccionamos del data frame únicamente los registros que corresponden a los "desacuerdo"

```
df.tab2 <- subset(df.tab1, p51dr=="En desacuerdo")
df.tab2

##      gr.edad      p51dr     Freq
## 6 [18,25] En desacuerdo 78.41727
## 7 (25,35] En desacuerdo 73.43173
## 8 (35,45] En desacuerdo 70.66116
## 9 (45,55] En desacuerdo 64.77987
## 10 (55,92] En desacuerdo 62.10526
```

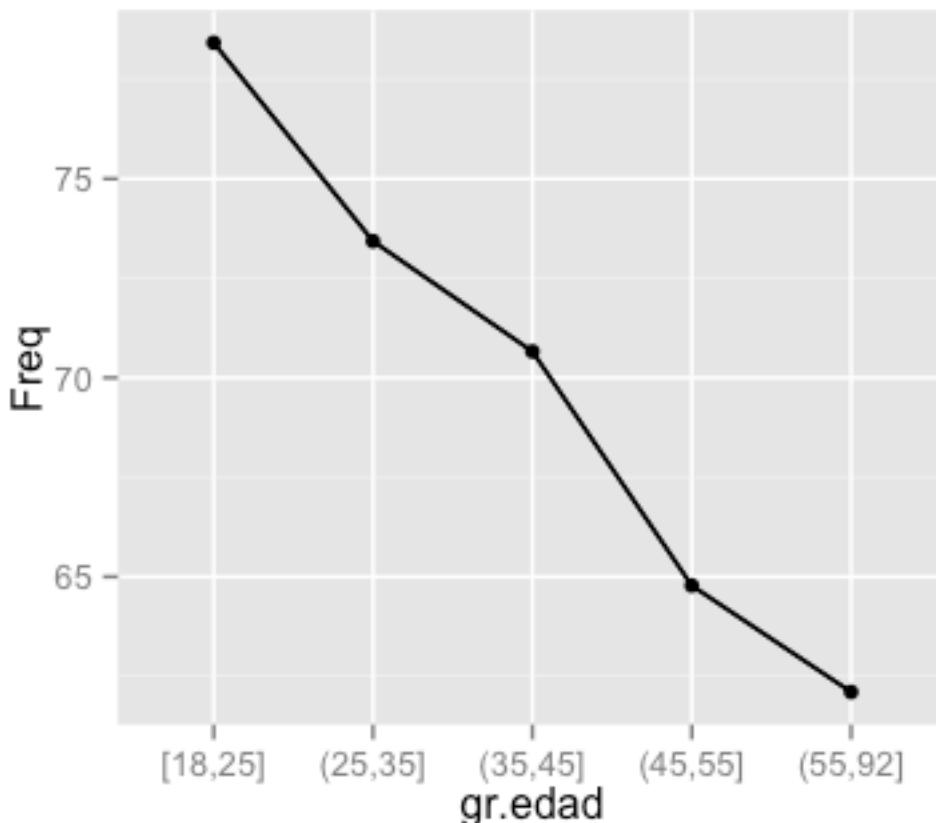
Ahora contamos con los datos para elaborar el gráfico de líneas:

```
graf.linea1 <- ggplot(df.tab2, aes(x=gr.edad, y=Freq, group=1)) + geom_line()
graf.linea1
```



Podemos añadir puntos al la línea para marcar mejor el dato correspondiente a los grupos de edad:

```
graf.linea2 <- graf.linea1 + geom_line() + geom_point()
graf.linea2
```



Podemos ajustar la escala del eje Y para evitar demasiadas distorsiones y añadir títulos a los ejes y al gráfico.

```
graf.linea2a <- graf.linea2 + ylim(0, 100) + xlab("Grupos de edad") +
ylab("% de casos") +
ggtitle("Porcentaje de personas en desacuerdo con la afirmación:\n" 
"Jamás tendría un amigo homosexual o una amiga lesbiana\",,\n" 
según grupo de edad")
```

graf.linea2a



Gráfico de puntos

Una alternativa a un gráfico de barras, puede ser un gráfico de puntos, conocido como "Cleveland Dot Plot". Observe este data frame que contiene el promedio de la edad considerada como ideal para que una mujer se case según dominio geográfico:

```
# Sintaxis para generar la tabla de datos
eideal.m <- genero$P1
eideal.m[genero$P1==99] <- NA
```

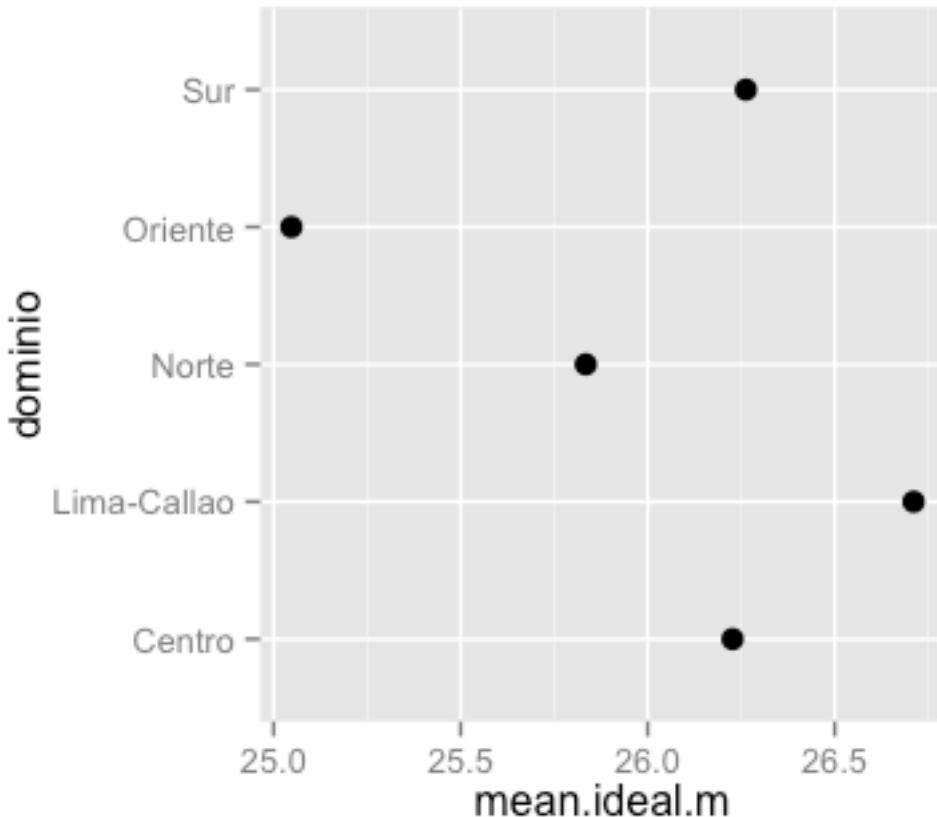
```
df2 <- data.frame(tapply(eideal.m, genero$DOMINIO, mean, na.rm = TRUE))
)
df2$dominio <- rownames(df2)
colnames(df2) <- c("mean.ideal.m", "dominio")

## Tabla de datos en forma de un data.frame
df2

##           mean.ideal.m      dominio
## Lima-Callao    26.71065 Lima-Callao
## Norte          25.83442     Norte
## Sur            26.26250       Sur
## Centro         26.22680    Centro
## Oriente        25.04762   Oriente
```

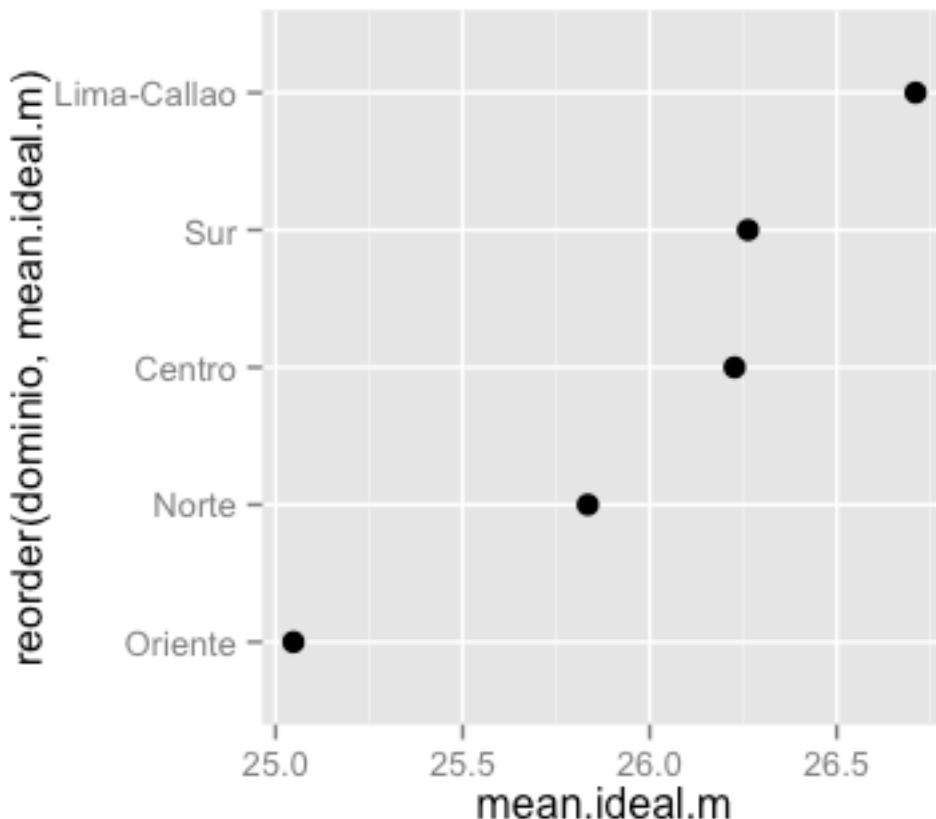
Aquí el grafico de puntos tipo "Cleveland":

```
ggplot(df2, aes(x=mean.ideal.m, y=dominio)) +
geom_point(size=3)
```



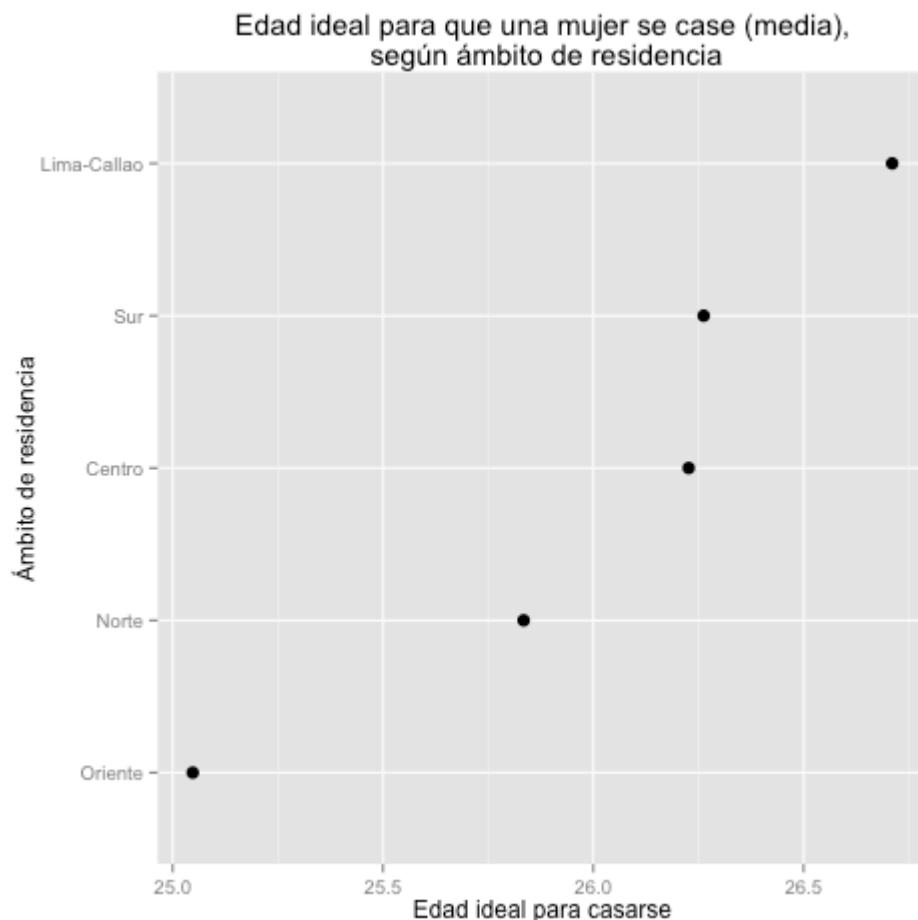
El mismo gráfico con los datos ordenados según la magnitud de la media

```
ggplot(df2, aes(x=mean.ideal.m, y=reorder(dominio, mean.ideal.m))) +
geom_point(size=3)
```



Mejoramos el gráfico

```
gclev <- ggplot(df2, aes(x=mean.ideal.m, y=reorder(dominio, mean.ideal.m))) +
  geom_point(size=3)
gclev1 <- gclev + xlab("Edad ideal para casarse") + ylab("Ámbito de residencia") +
  ggtitle("Edad ideal para que una mujer se case (media),\nsegún ámbito de residencia")
gclev1
```



Grabar el gráfico en un archivo

Los gráficos generados en R pueden guardarse en un archivo que luego puede usarse en otra aplicación (un documento en Word, en PDF o una página web, por ejemplo). Hay varios formatos gráficos disponibles: JPG, PNG, WMF, PDF, entre otros. Formatos adecuados para ser usados en Word son el PNG o el PDF. Por ejemplo, si queremos guardar el gráfico anterior podemos usar los siguientes comandos:

```
# En formato PNG
pgn(filename="graf_cleveland.pgn")
gclve1
dev.off()

# En formato pdf
pdf(filename="graf_cleveland.pdf")
gclve1
dev.off()
```

Ambos comandos graban sendos archivos en PNG y PDF en el directorio de trabajo del R. Si se quiere grabar el archivo en otra ubicación hay que proporcionar la ruta completa del directorio en su computadora.

Para más información sobre cómo grabar gráficos en archivos, se sugiere consultar los siguientes enlaces:

- <http://www.stat.berkeley.edu/~s133/saving.html>
- <http://blog.revolutionanalytics.com/2009/01/10-tips-for-making-your-r-graphics-look-their-best.html>

Estadísticos de Resumen en R

Estadísticos de resumen

Los estadísticos descriptivos son un número o categoría que nos proporciona información acerca algunos atributos importantes de una variable. Podemos clasificarlos en tres grandes tipos:

- Estadísticos de tendencia central: Moda, Mediana (M_d), Media (\bar{x})
- Estadísticos de dispersión: Mínimo, máximo, rango, varianza (σ^2, S^2), desviación estándar (σ, S), rango intercuartil.
- Cuantiles o estadísticos de orden o posición: percentiles, deciles, cuartiles, quintiles, etc.

Cargamos los datos de trabajo

Base de datos para estos ejercicios: Familia y roles de género 2012, a descargar de:

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

Se sugiere descargar también el cuestionario para utilizarlo como referencia de libro de códigos. 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
```

Medidas de tendencia central

```
median(genero$EDAD) # Mediana
## [1] 36

mean(genero$EDAD) # Media
## [1] 38.98088
```

Medidas de dispersión

```
range(genero$EDAD) # Rango
## [1] 18 92

var(genero$EDAD) # Varianza
## [1] 243.701

sd(genero$EDAD) # Desviación Estándar
## [1] 15.61092
```

Cuantiles: Medidas de orden o posición

Los cuantiles son medidas de orden o posición. Un cuantil es un valor de la variable por debajo o por encima del cual se encuentra una proporción determinada de casos de la distribución. Algunos cuantiles tienen nombres especiales: cuartiles, percentiles, deciles, etc. La mediana es el cuantil 0.5, también es el percentil 50, o el decil 5.

El comando "quantile" nos muestra los valores mínimo y máximo de la variable, así como los cuartiles.

```
quantile(genero$EDAD) # Cuartiles
```

```
##   0%  25%  50%  75% 100%
##  18   26   36   49   92
```

En este ejemplo tenemos que 49 años es el percentil 75 (o, lo que es lo mismo: el cuantil 0.75 o el tercer cuartil). Eso quiere decir que el 75% de los casos de esta distribución tienen entre 18 y 49 años de edad (puesto que 18 es la edad mínima). Eso también nos dice que el 25% de los casos tiene entre 49 y 92 años de edad.

Con los cuartiles podemos calcular el rango intercuartil: la diferencia entre el tercer cuartil (percentil 75) y el primer cuartil (percentil 25)

```
IQR(genero$EDAD)
```

```
## [1] 23
```

Si queremos otros cuantiles podemos crear un vector que contenga los cuantiles específicos y aplicar el comando condicionado a ese vector:

```
decil <- seq(0, 1, 0.1)
decil
##  [1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
quantile(genero$EDAD, decil)
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##  18   20   24   28   32   36   42   46   52   62   92
```

Otros ejemplos de lo anterior:

Quintiles:

```
quintil <- seq(0,1, 0.2)
quantile(genero$EDAD, quintil)
##   0%  20%  40%  60%  80% 100%
##  18   24   32   42   52   92
```

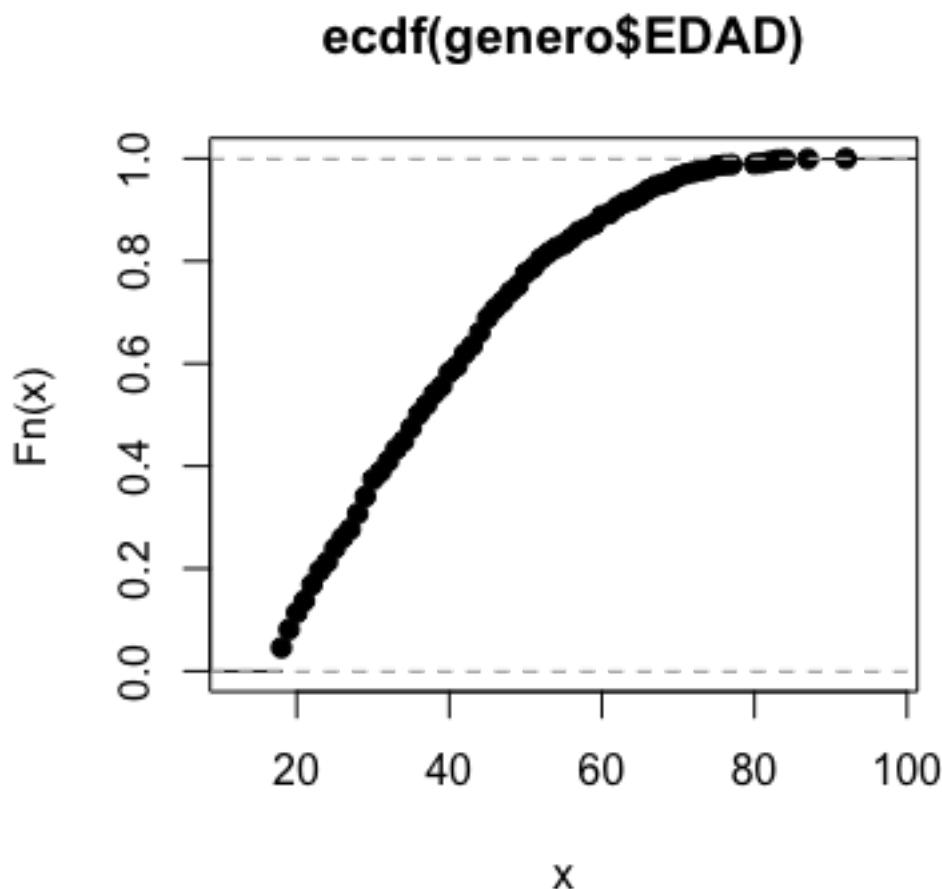
Cuantiles específicos

```
quantile(genero$EDAD, probs=c(.33, .44, .66, .855))
##   33%   44%   66% 85.5%
##   29    34    44    57
```

Distribución Empírica Acumulada

Otra forma de analizar una distribución es mediante la función de Distribución Empírica Acumulada (ecdf) que calcula el percentil correspondiente a cada valor de una distribución ordenada de una variable cuantitativa. Podemos observar el gráfico de una distribución empírica acumulada mediante el comando:

```
plot(ecdf(genero$EDAD))
```



Tambien podemos indagar acerca del percentil correspondiente a valores específicos de una variable. Por ejemplo:

```
Per <- ecdf(genero$EDAD)
Per(50) # Para La edad 50
## [1] 0.7763924
```

```
x <- c(25, 40, 65) # Para Las edades 25, 40 y 65
Per(x)

## [1] 0.2394015 0.5827099 0.9276808
```

Manejo de valores perdidos

Observen qué pasa cuando se pide lo siguiente:

```
mean(genero$P19A)

## [1] NA
```

No se puede computar la media porque hay valores perdidos. Podemos comprobar cuántos hay de la siguientes maneras:

```
table(is.na(genero$P19A)) # tabla

##
## FALSE TRUE
## 1201 2
```

Otra forma de comprobar los casos perdidos y válidos:

```
sum(is.na(genero$P19A)) # conteo de valores perdidos

## [1] 2

sum(!is.na(genero$P19A)) # conteo de casos válidos

## [1] 1201
```

Para indicar que se excluyan los casos perdidos (NA) del cómputo de un estadístico se usa la opción "na.rm = TRUE":

```
mean(genero$P19A, na.rm = TRUE)

## [1] 23.19234
```

Podemos usar el comando "summary" para producir algunos estadísticos descriptivos de una variable:

```
summary(genero$P19A)

##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.      NA's
##      0.00   7.00  16.00   23.19  30.00  168.00        2
```

O de una lista de variables que ponemos en un data frame:

```
mis.vars <- data.frame(genero$EDAD, genero$P1, genero$P19A)
summary(mis.vars)

##      genero.EDAD      genero.P1      genero.P19A
##      Min. :18.00      Min. :15.00      Min. : 0.00
##      1st Qu.:26.00    1st Qu.:25.00    1st Qu.: 7.00
##      Median :36.00    Median :25.00    Median :16.00
```

```
##   Mean :38.98   Mean :28.77   Mean : 23.19
## 3rd Qu.:49.00 3rd Qu.:30.00 3rd Qu.: 30.00
##  Max. :92.00   Max. :99.00   Max. :168.00
##                NA's :2
```

En los datos anteriores notamos que hay algunos problemas con las variables P1 y P19A: En P1, se considera como valor máximo 99 años (que es el código NS/NR). En P19A el valor máximo corresponde a una persona que dedica 24 horas diarias todos los días a las tareas domésticas (168/7). Si exploramos con más detalle ese dato (y en de la variable P19B), por ejemplo con el comando "table", notaremos que algunas personas dicen dedicarse a las tareas domésticas y de cuidado más de 20 horas al día todos los días, lo que es extraño. Podemos optar por excluir esos casos y marcarlos como valores perdidos usando el siguiente código:

```
genero$P1r <- genero$P1
genero$P1r[genero$P1==99] <- NA
genero$P2r <- genero$P2
genero$P2r[genero$P2==99] <- NA
genero$P19Ar <- genero$P19A
genero$P19Ar[genero$P19A >= 140] <- NA
genero$P19Br <- genero$P19B
genero$P19Br[genero$P19B >= 140] <- NA
```

Veamos algunos estadísticos de resumen de las variables modificadas

```
mis.var2 <- data.frame(genero$P1r, genero$P2r, genero$P19Ar, genero$P19Br)
summary(mis.var2)

##      genero.P1r      genero.P2r      genero.P19Ar      genero.P19Br
##  Min.   :15.00   Min.   :18.00   Min.   : 0.00   Min.   : 0.00
##  1st Qu.:25.00  1st Qu.:25.00  1st Qu.: 7.00  1st Qu.: 1.00
##  Median :25.00  Median :29.00  Median :16.00  Median :14.00
##  Mean   :26.22  Mean   :28.56  Mean   :22.61  Mean   :19.17
##  3rd Qu.:29.00 3rd Qu.:30.00 3rd Qu.:30.00 3rd Qu.:28.00
##  Max.   :60.00  Max.   :42.00  Max.   :112.00 Max.   :120.00
##  NA's   :42      NA's   :41     NA's   :7       NA's   :11
```

Resúmenes para grupos de casos

Con el comando "tapply" podemos solicitar estadísticos según los niveles de un factor. Por ejemplo, el tiempo dedicado a labores domésticas, según sexo:

```
tapply(genero$P19Ar, genero$SEXO, mean, na.rm=TRUE) # Media
## Masculino  Femenino
## 13.70187 31.19704

tapply(genero$P19Ar, genero$SEXO, median, na.rm=TRUE) # Mediana
## Masculino  Femenino
##      10      28
```

```
tapply(genero$P19Ar, genero$SEXO, sd, na.rm=TRUE) # Desviación est谩ndar
## Masculino Femenino
## 13.50444 21.35332

tapply(genero$P19Ar, genero$SEXO, length) # Número de casos
## Masculino Femenino
##      589      614
```

Podemos juntar los estadísticos en un mismo objeto:

```
Media <- tapply(genero$P19Ar, genero$SEXO, mean, na.rm=TRUE)
Mediana <- tapply(genero$P19Ar, genero$SEXO, median, na.rm=TRUE)
Desv <- tapply(genero$P19Ar, genero$SEXO, sd, na.rm=TRUE)
Ncasos <- tapply(genero$P19Ar, genero$SEXO, length)
cbind(Media, Mediana, Desv, Ncasos)

##           Media Mediana     Desv Ncasos
## Masculino 13.70187      10 13.50444    589
## Femenino  31.19704      28 21.35332    614
```

El problema es que no se consigna el número de casos válidos. Para ello una solución es:

```
df.val <- genero[is.na(genero$P19Ar)==FALSE, ]
Ninvalid <- tapply(df.val$P19Ar, df.val$SEXO, length)
cbind(Media, Mediana, Desv, Ncasos, Ninvalid)

##           Media Mediana     Desv Ncasos Ninvalid
## Masculino 13.70187      10 13.50444    589      587
## Femenino  31.19704      28 21.35332    614      609
```

Luego podemos

```
tab1 <- cbind(Media, Mediana, Desv, Ncasos, Ninvalid)
library(xtable)
print(xtable(tab1), type="html", file="resumen.html")
```

El comando "tapply" también se puede aplicar a "summary". Por ejemplo:

```
mujeres <- subset(genero, SEXO=="Femenino")
tapply(mujeres$P19Ar, mujeres$NSEGrup, summary)

## $`A/B`
##   Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
##   0.00   14.00  23.50  28.23  36.00  84.00     1
##
## $C
##   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
##   0.00   14.00  25.00  28.46  42.00  84.00
##
## $`D/E`
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
##	0.00	18.00	28.00	34.16	49.00	112.00	4

Boxplots

Boxplots o gráficos de cajas

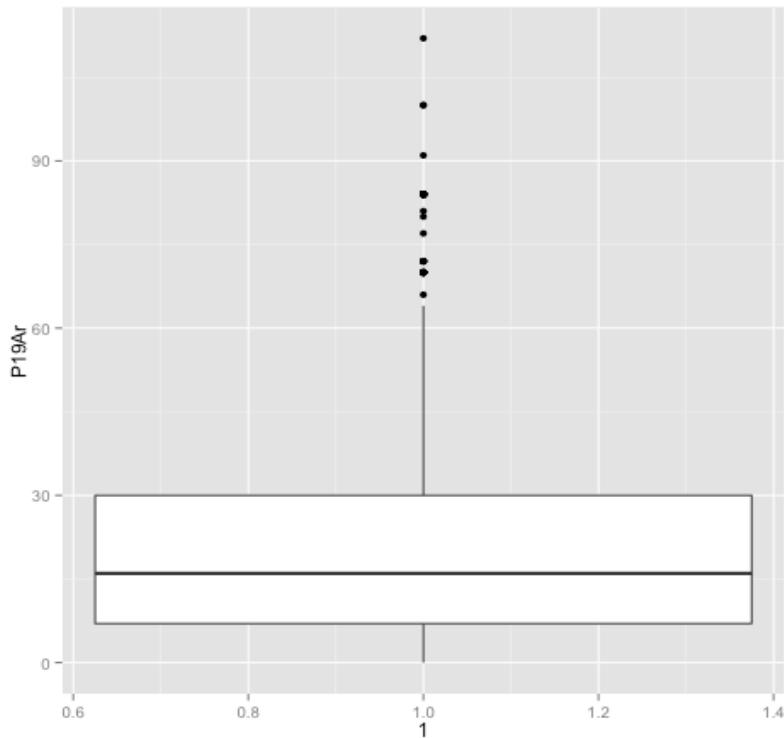
El boxplot o gráfico de cajas es una de las herramientas más útiles para el análisis descriptivo de la distribución de una variable, ya que nos presenta de manera suscinta:

- Una medida de tendencia central: La mediana
- Una medida de dispersión: El rango intercuartil
- Estadísticos de orden: Los cuartiles
- Identifica los casos atípicos o "outliers"

Boxplot con ggplot

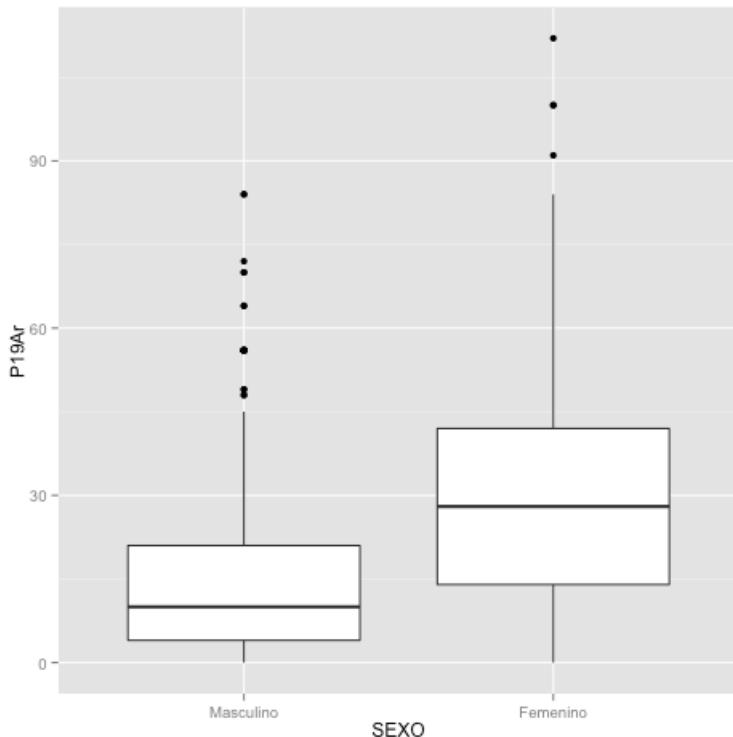
Un gráfico de cajas simple de una variable para un grupo único puede pedirse en ggplot con el siguiente comando:

```
# Para Las horas dedicadas a Labores domésticas:  
library(ggplot2)  
boxp1 <- ggplot(genero, aes(x=1, y=P19Ar)) + geom_boxplot()  
  
boxp1  
## Warning: Removed 7 rows containing non-finite values (stat_boxplot)  
.
```



Podemos solicitar un boxplot según grupos, en este caso sexo:

```
boxp2 <- ggplot(genero, aes(x=SEXO, y=P19Ar)) + geom_boxplot()
boxp2
## Warning: Removed 7 rows containing non-finite values (stat_boxplot)
.
```



Se pueden añadir facetas:

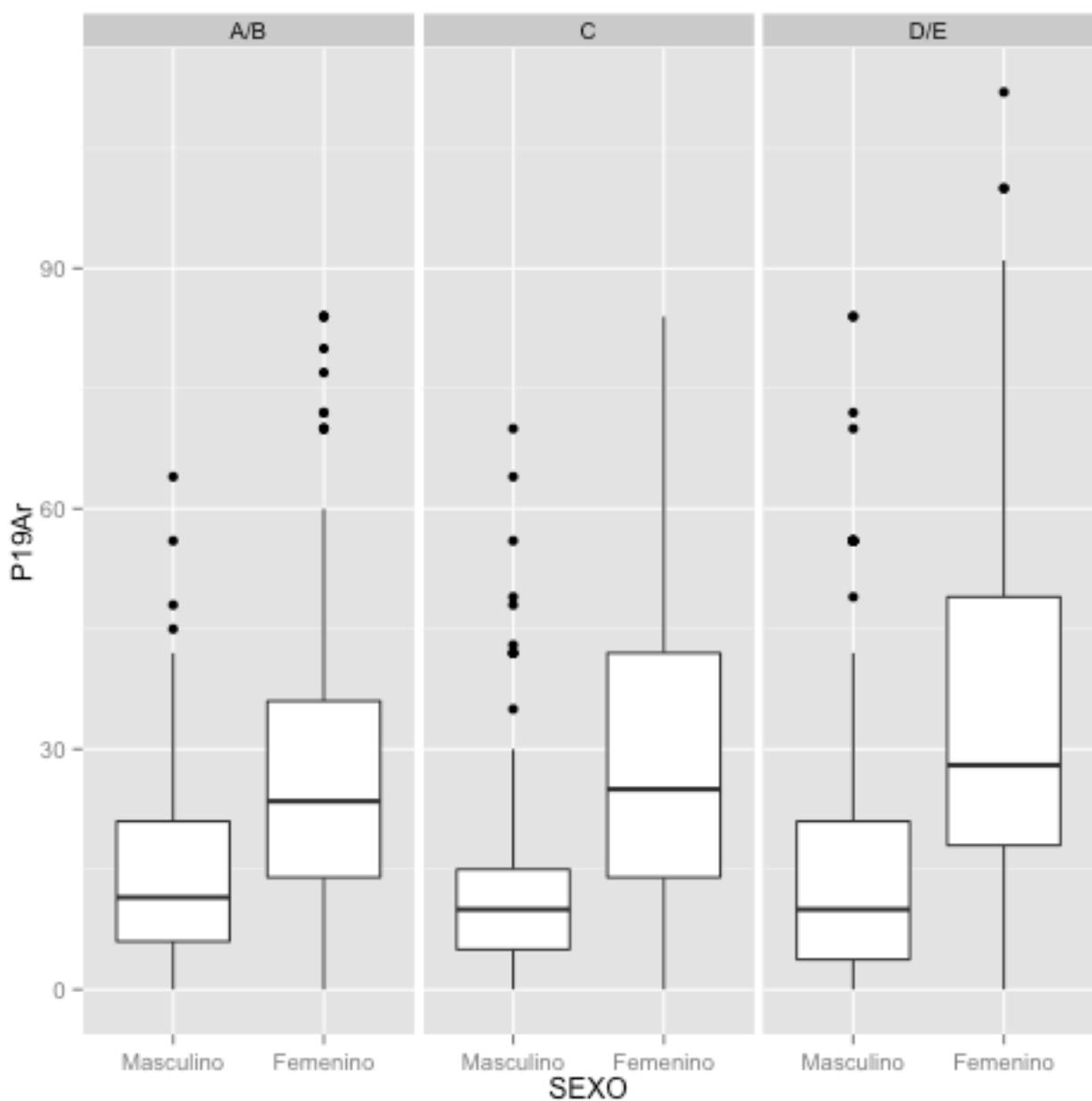
```
boxp3 <- ggplot(genero, aes(x=SEXO, y=P19Ar)) + geom_boxplot() + facet_grid(.~ NSEGrup)
boxp2
## Warning: Removed 2 rows containing non-finite values (stat_boxplot)
.
## Warning: Removed 1 rows containing non-finite values (stat_boxplot)
.
## Warning: Removed 4 rows containing non-finite values (stat_boxplot)
.
```



PUCP

ESTADÍSTICA PARA LAS CIENCIAS SOCIALES CON R

Profesor: David Sulmont



Inferencia Estadística y Distribuciones de Muestreo

Una población es el conjunto de todas las unidades de análisis que son nuestro objeto de estudio. Por ejemplo, si queremos estudiar el comportamiento electoral de los peruanos, la población estaría compuesta por todos los electores hábiles para votar en una elección, hoy en día (marzo 2015), esa población está compuesta por poco más de 21.3 millones de peruanos.

Cuando se estudia una población se busca observar ciertas características de las unidades de análisis que la componen. Un *censo* es una enumeración completa de todas las unidades de análisis de la población con el objeto de registrar los valores de las variables de estudio de todas ellas. Una *muestra* es una selección de un subconjunto de unidades de esa población con el objeto de observarlas para nuestro estudio.

Parámetros de la población

Para este ejemplo trabajaremos con la base de datos "enco" que contiene 704,720 entrevistados en la Encuesta Nacional Continua del INEI del 2006. En este caso pretendemos que esos entrevistados constituyen nuestro universo o población objetivo.

```
load("enco.Rdata")
names(enco)

## [1] "etnic"    "horas_w"
```

Como puede apreciarse, enco tiene dos variables: etnic, que es la autoidentificación étnica del entrevistado; y horas_w, que son la cantidad de horas trabajadas por el entrevistado la semana.

A continuación podemos ver los parámetros de ambas variables. Un *parámetro* es el valor del estadístico de resumen de una variable para el conjunto de la población o el universo.

```
prop.table(table(enco$etnic))

##
##      Quechua     Aymara   Amazonia      Afrod      Blanco     Mestizo
## 0.32192928 0.03914604 0.02239613 0.01334289 0.03430015 0.48140396
##      Otro
## 0.08748155

summary(enco$horas_w)

##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##      0.00   25.00  40.00  40.04   50.00 112.00
```

Como se aprecia, en nuestra población tenemos dos parámetros con los cuales vamos a trabajar:

- Media de las horas de trabajo semanal: $\mu = 40.04$

- Proporción de entrevistados que se autoidentifican como "Quechua": $\pi = 0.32$

Muestreo, estadística inferencial y error muestral

El objetivo de la inferencia estadística es estimar los parámetros poblacionales a partir de los estadísticos calculados sobre la base de muestras probabilísticas de esa población.

Como en muchos casos es muy costoso observar y medir a todos los elementos de una población, en la investigación social se trabaja por lo general con muestras. Una *muestra* es una selección de un subconjunto de elementos de la población.

Por ejemplo, en el caso de las elecciones presidenciales peruanas del 2011, la ONPE tenía registrada una población o universo de 19,949,915 electores hábiles. Sin embargo, para estimar las preferencias electorales, muchas empresas encuestadoras trabajaban con una muestra compuesta una mínima fracción de esa población, obteniendo resultados muy cercanos a la votación final. El 4 de junio del 2011, un día antes de la segunda vuelta de las elecciones presidenciales, IPSOS Apoyo realizó un simulacro de votación con una muestra de 4,000 electores, lo que supone apenas el 0.02% del total de la población. En esa medición Ollanta Humala obtuvo el 51.9% de los votos válidos. El resultado oficial de la ONPE le otorgó a Humala 51.4% de los votos válidos. En este ejemplo, podemos decir que la diferencia entre el parámetro poblacional y el estadístico muestral fue de -0.5%. Esta diferencia es lo que se conoce como el *error muestral*:

- Error muestral para el caso de una media: $e = \mu - \bar{x}$
- Error muestral para el caso de una proporción: $e = \pi - p$

Por lo general, antes de hacer una muestra no se conoce el parámetro de la población, pues precisamente la estimación de ese parámetro es uno de los objetivos de la investigación. Uno de los objetivos de la inferencia estadística es *estimar* cuál puede ser el margen de error de nuestros estadísticos muestrales con la finalidad de tener una idea de cuál puede ser el orden de magnitud de los parámetros poblacionales.

Muestreo simple al azar en R

Una muestra simple al azar es una selección aleatoria de un subconjunto casos de una población o universo, donde cada unidad de la población tiene una probabilidad conocida de ser seleccionada. El muestreo simple al azar es equivalente a realizar un sorteo. Hay dos tipos de muestreo simple:

- Muestreo sin reemplazo: cada elemento que es seleccionado de la población es extraído de la misma antes de seleccionar a otro elemento. En este caso la probabilidad de extraer el primer elemento de la población es igual a $1/N$, siendo N el tamaño de la población; el en caso del segundo elemento, la probabilidad es $1/(N-1)$; del tercero es $1/(N-2)$; y así sucesivamente.

- Muestreo con reemplazo: una vez que se extrae a un elemento de la población, éste es devuelto a la misma para proceder a otra selección. En este caso, en cada proceso de selección, cada elemento de la población tiene la misma probabilidad ($1/N$) de ser seleccionado.

Cuando se trata de muestras de poblaciones muy grandes, es muy poco probable que un mismo elemento sea seleccionado en más de una oportunidad (la probabilidad sería: $1/N * 1/N$), por lo que en la práctica puede considerarse que se trata de muestras con reemplazo.

Usando los datos de "enco" vamos a realizar una muestra simple al azar de 500 casos de nuestro universo. Trabajaremos en primer lugar con la variable "horas_w".

Para que los resultados de estos ejercicios sean reproducibles, primero fijaremos una semilla de aleatorización en el R. De lo contrario, la muestra que se obtenga en esta presentación no será la misma que la que los lectores obtengan al tratar de reproducir este ejemplo (para mayores detalles ver la ayuda del R sobre la función set.seed). Luego pediremos una muestra de 500 casos de la variable enco\$horas_w, que será almacenada en el objeto m1. Finalmente pediremos la media como estadístico muestral.

```
set.seed(200) # (el número 200 es completamente arbitrario)
m1 <- sample(enco$horas_w, 500, replace=T)
mean(m1)

## [1] 38.918
```

Puesto que contamos con el dato del parámetro poblacional, en este caso nos es posible calcular el error muestral de la muestra m1:

```
mean(enco$horas_w) - mean(m1)

## [1] 1.117928
```

Como hemos mencionado, por lo general el error muestral es una incógnita en nuestras investigaciones.

Distribuciones de muestreo empíricas

Una distribución de muestreo empírica es distribución de los resultados de muestras repetidas del mismo tamaño de una población.

Para generar una distribución empírica de muestreo, es necesario repetir la muestra que hemos tomado varias veces. Para ello podemos programar una función que haga esa tarea. En el R, una función es un pequeño programa que nos permite darle instrucciones al R respecto de un conjunto de procedimientos. En este ejemplo vamos a crear la función "muestra", que genera muestreos repetidos de una variable cuantitativa, y arroja como resultado las medias de la variable en cada una de las muestras obtenidas.

muestra es la función para obtener muestras múltiples y una media, donde:

v = variable a analizar

n = el tamaño de la muestra

r = la cantidad de muestras repetidas

```
muestra <- function(v, n, r){
  id <- 0
  m <- numeric()
  repeat{
    id <- id+1
    m1 <- mean(sample(v, n, replace=TRUE))
    m <- c(m, m1)
    if(id >= r) break
  }
  return(m)}
```

Por ejemplo, pidamos 10 muestras diferentes de 500 casos cada una de la variable horas_w de la población enco. El resultado será 10 medias diferentes, cada una corresponde al estadístico muestral de cada muestra obtenida:

```
set.seed(200)
muestra(enco$horas_w, 500, 10)

## [1] 38.918 39.120 39.510 39.244 42.068 40.250 39.986 39.406 40.096
## [6] 41.810
```

Como en este caso conocemos el parámetro, podemos calcular el error muestral de cada muestra de la siguiente manera:

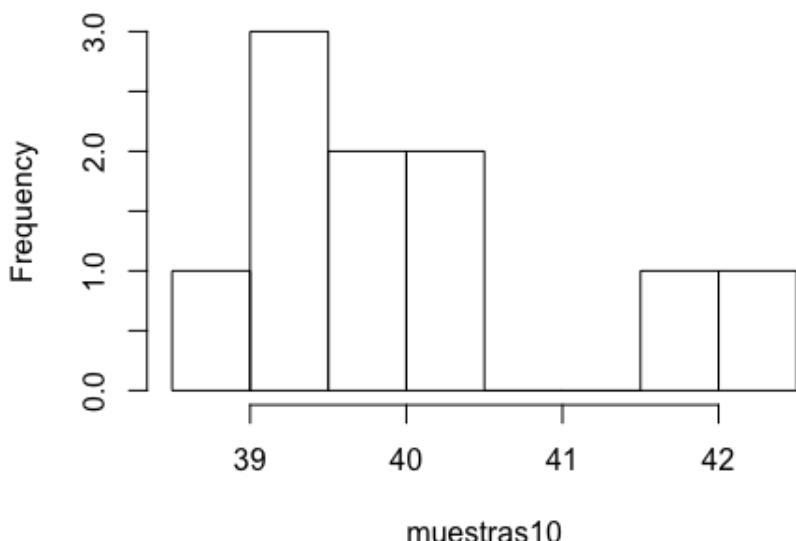
```
set.seed(200)
muestras10 <- muestra(enco$horas_w, 500, 10)
error1 <- mean(enco$horas_w) - muestras10
error1

## [1] 1.11792774 0.91592774 0.52592774 0.79192774 -2.03207226
## [6] -0.21407226 0.04992774 0.62992774 -0.06007226 -1.77407226
```

Como puede verse el error muestras o nivel de precisión de cada muestra es distinto: la muestra con mayor error es la 5ta muestra, que está sobreestimando el parámetro en 2.03 horas. La muestra más precisa es la 7ma muestra, que apenas subestima en parámetro en 0.0499 horas.

Una manera de visualizar una distribución empírica de muestreo es pedir un histograma con las medias de cada una de las muestras:

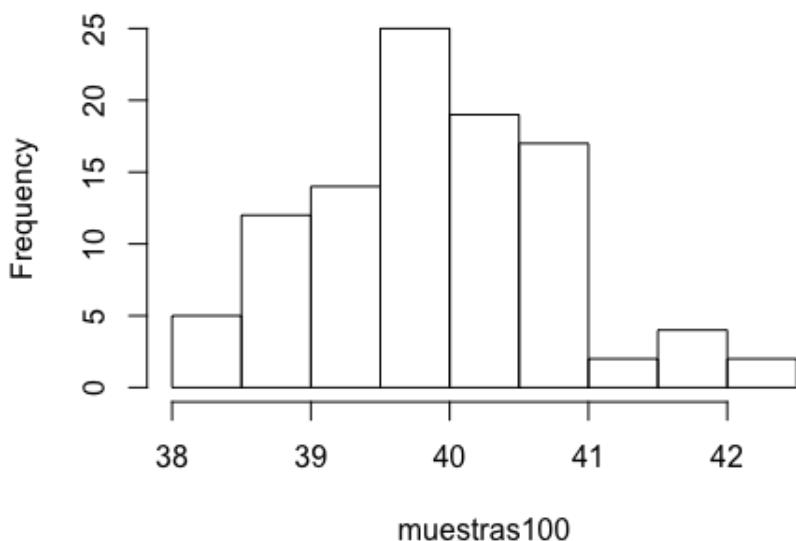
```
hist(muestras10)
```

**Histogram of muestras10**

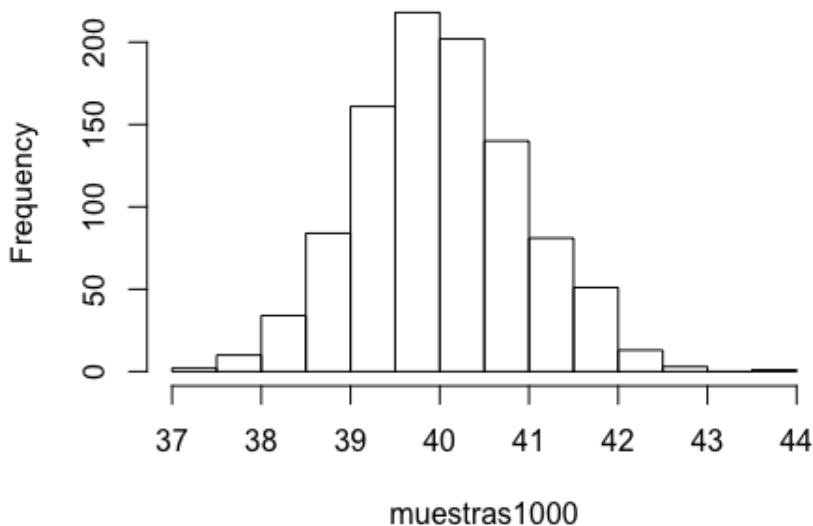
"LA" Distribución de muestreo

¿Qué pasa si en vez de pedir 10 muestras pedimos 100?, ¿y si pedimos 1000?, ¿10000?, ¿cómo se verían los resultados de esos muestreos?

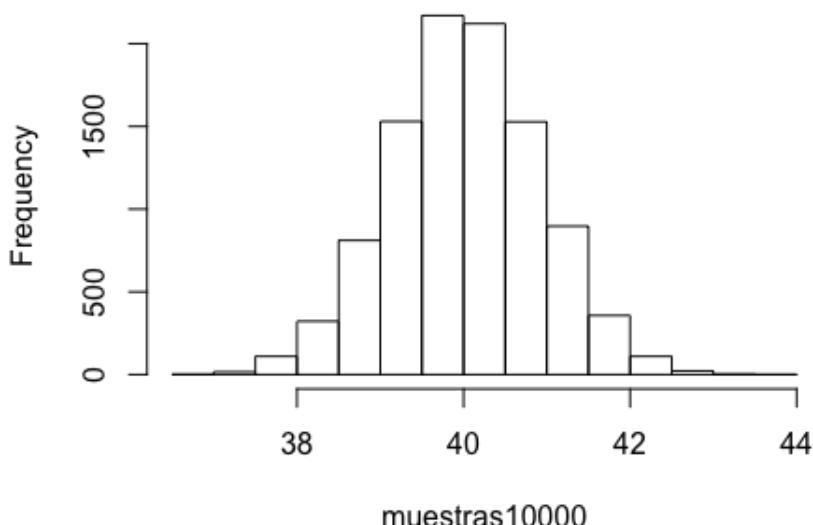
```
set.seed(200)
muestras100 <- muestra(enco$horas_w, 500, 100)
hist(muestras100)
```

Histogram of muestras100

```
set.seed(200)
muestras1000 <- muestra(enco$horas_w, 500, 1000)
hist(muestras1000)
```

Histogram of muestras1000


```
set.seed(200)
muestras10000 <- muestra(enco$horas_w, 500, 10000)
hist(muestras10000)
```

Histogram of muestras10000


Como puede apreciarse, cuantas más muestras solicitamos, el histograma de la distribución empírica de muestreo tiende a hacerse cada vez más normal, y siempre centrado alrededor del parámetro poblacional. Ojo que lo que estamos incrementando es la cantidad de muestras que tomamos, no el tamaño de cada muestra. Cada muestra el del mismo tamaño: 500 entrevistados. En la práctica, un investigador sólo tiene recursos para hacer *una* muestra de una población grande. Este ejercicio de hacer muestreos repetidos solo es posible si tenemos recursos ilimitados, o si, como en este caso, lo hacemos mediante una simulación por computadora.

Supongamos que obtenemos *todas* las muestras posibles de 500 casos de una población de poco más de 700 mil personas (como la población de "enco") y para cada muestra calculamos la media. Ese es un ejercicio teórico, y el resultado sería la distribución de muestreo de todas las medias muestrales posibles de ese tipo de muestra. En este caso *la* distribución de muestreo es un objeto teórico, y es la representación matemática de todos los muestreos repetidos de tamaño n de una población de tamaño N .

Como hemos visto, cuantos más muestreos repetidos realizamos, las distribuciones de muestreo tienen a hacerse normales. En el caso de *la* distribución de muestreo, se trata de una distribución normal. Este fenómeno tiene que ver con lo que se conoce como el *teorema del límite central*. El teorema del límite central nos dice que:

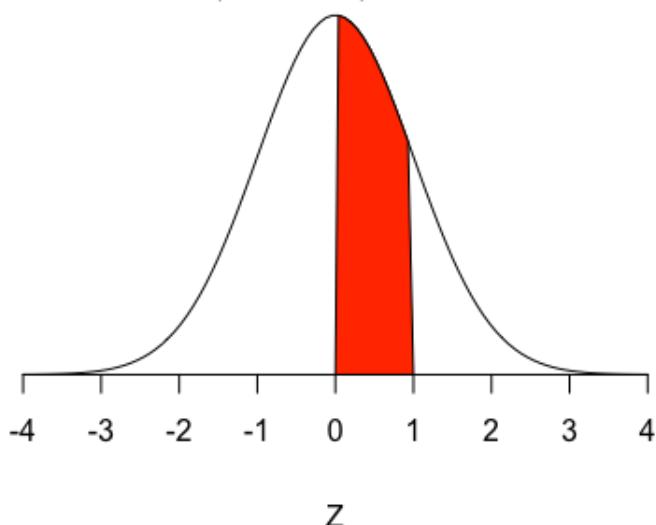
"Sin importar la forma de una puntuación bruta de una variable de intervalo o razón, su distribución muestral será normal cuando el tamaño de la muestra, n , sea mayor que 121 casos y se centrará en la media de la población verdadera" (Ritchey 2008: 214)

Estas propiedades son muy importantes. Como hemos visto, la distribución normal estándar tiene propiedades bastante específicas:

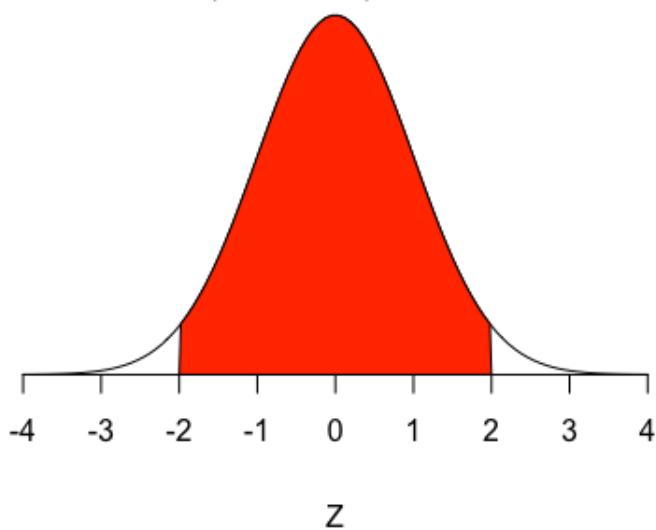
- Es simétrica: es decir, la media y la mediana coinciden
- La media es igual a 0 (cero).
- La desviación estándar es igual a 1
- Existe una proporción fija y conocida de casos entre dos valores de la distribución normal, por ejemplo:
 - Entre la media y una desviación estándar se encuentra el 34.1% del área total de la distribución normal.
 - Entre dos desviaciones estándar por debajo y por encima de la media se encuentra el 95.4% del área total de la distribución.

**Distribución Normal**

$$P(0 < Z < 1) = 0.341$$

**Distribución Normal**

$$P(-2 < Z < 2) = 0.954$$

**El Error Estándar e intervalos de confianza**

Como hemos visto, el error muestral es la diferencia entre el parámetro poblacional y el estadístico muestral. El error muestral es una propiedad de una muestra en particular, y es una incógnita si no se conoce el parámetro poblacional.

En el caso de la distribución de muestreo, de acuerdo con el teorema del límite central, la media de la distribución es igual al parámetro poblacional. Como toda distribución de una variable cuantitativa, es posible calcular la desviación estándar de la distribución de muestreo:

$$\sigma_{\bar{x}} = \sqrt{\frac{\sum(\mu - \bar{x})^2}{N}}$$

La desviación estándar de la distribución de muestreo se conoce como el **error estándar** y puede interpretarse como el error muestral típico de una media de una muestra de tamaño n de nuestra población.

Recordemos algunos conceptos y principios claves ligados al teorema del límite central:

- La distribución de muestreo, es decir, la distribución de todas las medias de muestras posibles de tamaño 500 de "enco" es una distribución normal
- La desviación estándar de la distribución de muestreo es el error estándar.
- Siendo una distribución normal estándar, sabemos que el 95% del área de la distribución de muestreo se encuentra entre 1.96 desviaciones estándar por encima y por debajo de la media poblacional. En otras palabras, es posible afirmar que el 95% de todos los resultados muestrales tiene un error muestral que no es mayor a 1.96 desviaciones estándar del parámetro poblacional.

En una de nuestras simulaciones de muestras, pedimos 10,000 muestras de tamaño 500 de "enco". Comparemos el resultado de la media de enco con la media de la distribución de muestreo de 10,000 muestras de tamaño 500:

```
mean(enco$horas_w)
## [1] 40.03593
mean(muestras10000)
## [1] 40.02028
```

Como ven la media de muestras10000 es casi igual a la media poblacional. Asumamos por ahora que muestras10000 es la distribución muestral. De acuerdo con ese supuesto, el error estándar debería ser:

```
sd(muestras10000)
## [1] 0.8953256
```

Aplicando los principios antes vistos, deberíamos encontrar que el 95% de los resultados en muestras10000 deberían estar a una distancia no mayor a 1.96 veces la desviación estándar de muestras10000. Veamos qué sucede en realidad:

```
linf <- mean(muestras10000) - (1.96*sd(muestras10000)) # Límite inferior
lsup <- mean(muestras10000) + (1.96*sd(muestras10000)) # Límite superior
```

or

```
# Vector Lógico que nos dice si una muestra cae dentro de esos Límites
:
dentro <- muestras10000 >= linf & muestras10000 <=lsup

# Número de muestras que cae dentro de esos Límites:
table(dentro)

## dentro
## FALSE TRUE
## 497 9503

# % de muestras que cae dentro de esos Límites:
prop.table(table(dentro))*100

## dentro
## FALSE TRUE
## 4.97 95.03
```

Como puede verse, en una distribución empírica de muestreo de 10,000 muestras de tamaño 500 de "enco", tan solo el 4.97% de todas las muestras tiene un error mayor a 1.96 veces el error estándar (ya sea hacia arriba o hacia abajo). El 95.03% de las muestras entás dentro de los límites del intervalo fijado, un resultado casi idéntico al pronosticado por la teoría.

Cálculo del intervalo de confianza para una sola muestra

Como investigadores casi siempre disponemos de una sola muestra para hacer nuestras inferencias. Usemos en este caso la muestra 1.

En la práctica no puede calcularse directamente el error estándar ya que se desconoce μ . Sin embargo una buena aproximación para calcular el error estándar de una media de una muestra simple al azar consiste en aplicar la siguiente fórmula:

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Donde s es la desviación estándar de la media muestral, y n es el tamaño de la muestra. En el cálculo del estimado del error estándar podemos usar la desviación estándar ya que, más allá de la diferencia puntual, que puede haber entre la media muestras y el parámetro, la **forma** o el patrón de dispersión de datos de la distribución de la muestra tiende a ser muy similar a la forma de la distribución poblacional.

Veamos cómo se aplica el concepto y el cálculo del error estándar para la primera muestra que hemos calculado en esta sección:

```
set.seed(200)
m1 <- sample(enco$horas_w, 500, replace=T)
mean(m1)
```

```
## [1] 38.918
sd(m1)
## [1] 18.9001
media1 <- mean(m1)
desv1 <- sd(m1)
error.est1 <- desv1/sqrt(500)
error.est1
## [1] 0.8452381
```

De acuerdo con los datos de la primera muestra simple al azar que hemos obtenido en esta sección, podemos estimar que el error estándar de la media es de 0.845 horas por semana.

Con esta información podemos calcular un intervalo de confianza de la siguiente forma:

$$IC_{1-\alpha} = \bar{x} \pm Z\sigma_{\bar{x}}$$

Donde Z es el valor de la distribución normal estándar que representa nuestro nivel de confianza.

Si queremos calcular un **intervalo de confianza al 95%** podemos usar un Z = 1.96, ya que a más/menos 1.96 desviaciones estándar de la media de la curva normal, se encuentra el 95% del área de la distribución. En la práctica muchas personas suelen redondear 1.96 a 2 para simplificar los cálculos.

Por lo general, los intervalos de confianza más utilizados son:

- IC al 95%, donde Z = 1.96
- IC al 99% donde Z = 2.57

Siguiendo con nuestro ejemplo, el intervalo de confianza al 95% de la media de la muestra 1 sería:

```
linf <- media1 - (1.96*error.est1) # Límite inferior del IC
lsup <- media1 + (1.96*error.est1) # Límite superior del IC

linf
## [1] 37.26133
lsup
## [1] 40.57467
```

En conclusión podemos afirmar que: "**tenemos un 95% de confianza que la media poblacional se encuentra en algún lugar entre 37.26 y 40.57 horas semanales de trabajo**". Esta afirmación resulta ser correcta ya que, como hemos visto en este ejemplo, la media poblacional o parámetro es 40.036 horas.

Si queremos ampliar el intervalo de confianza, reemplazamos Z por el valor correspondiente. Por ejemplo, si queremos calcular un intervalo de confianza al 99% debemos hacer:

```
media1 - (2.57*error.est1) # Límite inferior del IC al 99%
## [1] 36.74574
media1 + (2.57*error.est1) # Límite superior del IC al 99%
## [1] 41.09026
```

Un error común de interpretación de los intervalos de confianza es afirmar que tenemos un % de confianza en que la media poblacional o parámetro **varía** entre los límites del intervalo. Ello es totalmente incorrecto, ya que para una población y variable en un momento dado, el parámetro es uno sólo. Lo que puede variar es la media muestral ya que eso depende del azar y del resultado de cada proceso de selección aleatoria de miembros de la población.

Distribución de muestreo para proporciones

En el caso de proporciones (o porcentajes si las multiplicamos por 100), el razonamiento es similar al de la distribución de muestreo de medias. La diferencia radica en el tipo de estadístico que estamos estimando.

La siguiente función nos permite obtener muestras repetidas del porcentaje de casos que cae dentro de una de las categorías de una variable cualitativa:

Función para muestras múltiples y el % de una categoría de la variable, donde:

v = variables de análisis; n = tamaño de la muestra; c = categoría de análisis; r = la cantidad de muestras

```
muestra.p <- function(v, n, c, r){
  id <- 0
  m <- numeric()
  repeat{
    id <- id+1
    m0 <- sample(v, n, replace=TRUE)
    m1 <- mean(ifelse(m0==c, 1, 0))
    m <- c(m, m1)
    if(id >= r) break
  }
  return(m)}
```

Ejemplo de distribuciones empíricas de muestreo para proporciones

En la población, el parámetro para la proporción de personas que se identifican como "Quechua" es 0.322 ó 32.2%:

```
prop.table(table(enco$etnic))
```

```
##  
##      Quechua      Aymara     Amazonia      Afrod      Blanco     Mestizo  
## 0.32192928 0.03914604 0.02239613 0.01334289 0.03430015 0.48140396  
##      Otro  
## 0.08748155
```

Usando la función anterior podemos seleccionar una muestra de 500 casos y obtener el porcentaje para la categoría que estamos analizando:

```
set.seed(300)  
muestra.p(enco$etnic, 500, "Quechua", 1)  
  
## [1] 0.314
```

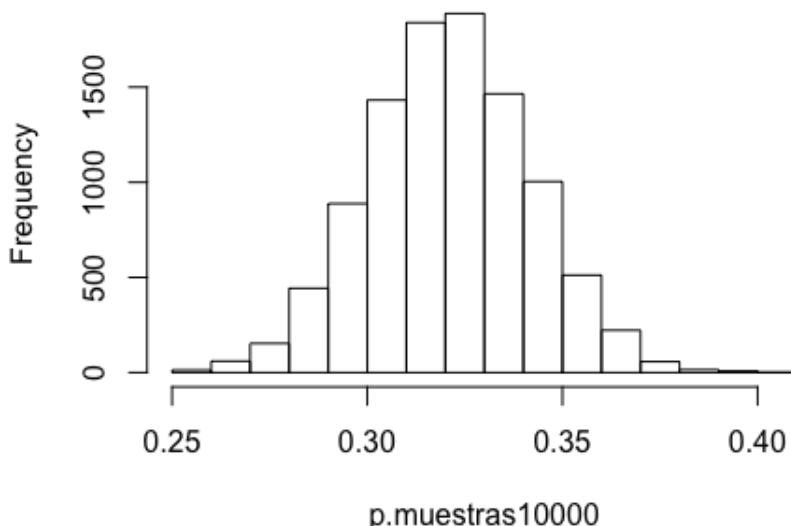
Como se aprecia, en esta primera muestra, el 0.314 de las personas se autoidentifican como "Quechua". Esto nos da un error muestral de:

$$e = \pi - p = 0.322 - 0.314 = 0.008$$

Podemos pedir 10,000 muestras repetidas y ver que la forma de su distribución empírica de muestreo se aproxima a una distribución normal estándar.

```
set.seed(300)  
p.muestras10000 <- muestra.p(enco$etnic, 500, "Quechua", 10000)  
hist(p.muestras10000)
```

Histogram of p.muestras10000



Error estándar e intervalo de confianza para proporciones

Al igual que en el caso de las medias, el error estándar de la distribución de muestreo de porcentajes es la desviación estándar de todos los resultados de las

muestras. Para estimar el error estándar de una proporción se usa la siguiente fórmula:

$$\sigma_p = \sqrt{\frac{pq}{n}}$$

Donde:

- p es la proporción de la categoría de análisis en nuestra muestra
- q es igual a 1 - p

En el caso de la primera muestra que hemos tomado, el error estándar sería:

```
set.seed(300)
prop1 <- muestra.p(enco$etnic, 500, "Quechua", 1)
error.est.p1 <- sqrt((prop1*(1-prop1))/500)
error.est.p1

## [1] 0.02075591

## En porcentajes sería:
error.est.p1*100

## [1] 2.075591
```

Para calcular el intervalo de confianza de la proporción, se emplea también la misma lógica que en el caso de las medias:

$$IC_{1-\alpha} = p \pm Z\sigma_p$$

Intervalo de confianza al 95% para la proporción de la muestra 1:

```
prop1 - 1.96*error.est.p1 # Límite inferior
## [1] 0.2733184

prop1 + 1.96*error.est.p1 # Límite superior
## [1] 0.3546816
```

Nuestra conclusión sería que tenemos 95% de confianza que la proporción de personas que se identifican como "Quechua" en la población objetivo está en algún punto entre 0.27 y 0.35, ó, si queremos expresarlo en %, entre 27% y 35%.

Observen que el margen de error, expresado en %, para este resultado muestral es de:

```
(1.96*error.est.p1)*100 # Margen de error al 95% de confianza
## [1] 4.068159
```

Si queremos cambiar el nivel de confianza, simplemente hay que indicar el valor de Z correspondiente.

Margen de error de una muestra

Muchas veces nos interesa saber cuál sería el marge de error de una muestra para poder hacer algunas estimaciones. Como se desprende de las fórmulas del error estándar, el margen de error depende de:

- El tamaño de la muestra (N)
- La heterogeneidad o nivel de complejidad de la variable de análisis, expresado por la desviación estándar (s) en el caso de medias, o por $p * q$ en el caso de proporciones.

La heterogeneidad de la variable es una característica de la población objetivo, por lo tanto no es un dato que dependa del investigador. Lo que está en las manos del investigador es definir el tamaño de la muestra, que generalmente depende de los recursos con los que se cuente para hacer el estudio.

Con las fórmulas para el cálculo del error estándar y del intervalo de confianza, podemos tener elementos para estimar el marge de error de una muestra. En el caso de proporciones, el margen de error puede estimarse de la siguiente manera:

$$e = \sigma_p * Z$$

Para estimar el error estándar de una proporción necesitamos el valor de p , sin embargo, este valor es una incógnita antes de haber realizado la investigación. A pesar de ello podemos formularlos una hipótesis para responder a esta pregunta: ¿cuál sería el valor de p en el caso de mayor complejidad? La respuesta es:

$$p = 0.5$$

Ya que eso haría que el término

$$p * q$$

de la ecuación del error estándar adquiera el mayor valor posible.

Entonces para estimar el error estándar, podemos ponernos en la hipótesis de mayor heterogeneidad, donde $p = q$. Suponiendo que sólo tuviéramos recursos para hacer una muestra de 400 casos, podemos estimar con un 95% de confianza que el error estimado para una variable categórica dicotómica (donde sólo hay dos resultados posibles), sería igual a:

$$e = \left(\sqrt{\frac{0.5^2}{400}} \right) * 1.96 = 0.049$$

Lo que expresado en % vendría a ser: 4.9%

Un ejemplo de una variable categórica dicotómica es: aprueba o desaprueba una afirmación. Sólo hay dos resultados posibles. Cualquier categoría de una variable cualitativa puede representarse por una variable dicotómica.

Otro ejemplo: Error estimado para una muestra de 1,000 casos, con un nivel de confianza del 95% y bajo el supuesto de máxima heterogeneidad:

```
sqrt((0.5^2)/1000) * 1.96
```

```
## [1] 0.03099032
```

Inferencia Estadística e Intervalos de Confianza

Inferencia estadística

En esta parte del curso examinaremos las siguientes herramientas de inferencia estadística

- Cálculo de intervalo de confianza
- Representación gráfica de intervalos de confianza

Cargamos los datos de trabajo

Base de datos para estos ejercicios: Familia y roles de género 2012, a descargar de:

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

Se sugiere descargar también el cuestionario para utilizarlo como referencia de libro de códigos. 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
```

Preparar las variables de trabajo

En estos ejemplos trabajaremos con las siguientes variables:

- Edad ideal para que una mujer se case (P1)
- Edad ideal para que un hombre se case (P2)
- Edad en la que el entrevistado se casó o empezó a convivir (P23)
- Número de hijos (P44A)
- Horas dedicadas al trabajo doméstico (P19A)
- Horas que la pareja le dedica al trabajo doméstico (P28A)

Antes de utilizar estas variables es necesario evaluar si necesitan algún tipo de acondicionamiento (identificar valores perdidos o "raros", recodificar, etc)

Acondicionamiento de las variables

Las siguientes transformaciones nos permitirán acondicionar las variables para utilizarlas en el análisis:

```
genero$p1r <- genero$P1
genero$p2r <- genero$P2
genero$p1r[genero$P1 > 40] <- NA
genero$p2r[genero$P2 == 99] <- NA
```

```
genero$p19ar <- genero$P19A
genero$p19ar[genero$P19A >= 140] <- NA

genero$p28ar <- genero$P28A
genero$p28ar[genero$P28A > 120] <- NA
```

Cálculo de intervalos de confianza para una media

Para calcular el intervalo de confianza de la media de una variable que proviene de una muestra simple al azar usamos la siguiente fórmula:

$$IC_{1-\alpha} = \bar{x} \pm Z\sigma_{\bar{x}}$$

Donde:

- $1 - \alpha$ es el nivel de confianza deseado
- Z es el valor de la distribución normal estándar que representa el nivel de confianza
- $\sigma_{\bar{x}}$ es el error estándar de la media que se obtiene mediante:

$$\sigma_{\bar{x}} = \frac{S_x}{\sqrt{n}}$$

Cálculo de intervalo de confianza al 95% para una media

Entonces para calcular el intervalo de confianza para una media podemos proceder de la siguiente manera:

```
media <- mean(na.omit(genero$p19ar)) # Pedimos La media
desv <- sd(na.omit(genero$p19ar)) # La desviación estándar
N <- length(na.omit(genero$p19ar)) # El tamaño válido de La muestra
error.est <- desv/sqrt(N) # Calculamos el error estándar
error <- 2*error.est # Fijamos Z=2 para indicar un nivel de confianza de 95%
lim.inf <- media-error # Límite inferior del intervalo
lim.sup <- media+error # Límite superior del intervalo

# Guardamos todos los datos generados en un objeto data frame
resultado1 <- data.frame(media, desv, N, error.est, error, lim.inf, lim.sup)
resultado1

##      media      desv      N error.est      error    lim.inf    lim.sup
## 1 22.61037 19.94936 1196  0.5768506 1.153701 21.45667 23.76407
```

Cálculo del intervalo de confianza al 95% usando la distribución de t

```
media <- mean(na.omit(genero$p19ar))
desv <- sd(na.omit(genero$p19ar))
N <- length(na.omit(genero$p19ar))
```

```

error.est <- desv/sqrt(N)
error <- qt(0.975, df= N-1) * error.est # Usar el cuantil 0.975 de t
lim.inf <- media-error
lim.sup <- media+error
resultado2 <- data.frame(media, desv, N, error.est, error, lim.inf, li
m.sup)
resultado2

##      media      desv      N error.est      error    lim.inf    lim.sup
## 1 22.61037 19.94936 1196 0.5768506 1.131753 21.47862 23.74212

resultado1

##      media      desv      N error.est      error    lim.inf    lim.sup
## 1 22.61037 19.94936 1196 0.5768506 1.153701 21.45667 23.76407

```

Uso de una función ad-hoc para el cálculo del intervalo de confianza

Una función es un pequeño programa en R que sirve para automatizar algunos procedimientos. En este caso la función calcula los estadísticos necesarios para generar el intervalo de confianza de una media.

Generamos la función:

```

int.conf <- function(x, ic = 95) {
  y <- na.omit(x)
  Media <- mean(y)
  Desv.Est <- sd(y)
  N <- length(y)
  Error.Est <- Desv.Est/sqrt(N)
  ci.y <- 1-((100-ic)/100)/2
  Error <- Error.Est * qt(ci.y, df = N-1)
  Lim.Inf <- Media - Error
  Lim.Sup <- Media + Error
  result <- data.frame(Media, Desv.Est, N, Error.Est, Error, Lim.Inf,
  Lim.Sup)
  return(result)
}

```

Usamos la función para calcular un intervalo de confianza al 95% (ic por defecto de la función)

```

int.conf(genero$p19ar)

##      Media Desv.Est      N Error.Est      Error    Lim.Inf    Lim.Sup
## 1 22.61037 19.94936 1196 0.5768506 1.131753 21.47862 23.74212

```

En vez de un intervalo de 95% podemos usar otros niveles de confianza:

```

int.conf(genero$p19ar, ic=99) # Intervalo de 99% de confianza

##      Media Desv.Est      N Error.Est      Error    Lim.Inf    Lim.Sup
## 1 22.61037 19.94936 1196 0.5768506 1.488246 21.12212 24.09861

```

```
int.conf(genero$p19ar, ic=90) # Intervalo de 90% de confianza
##      Media Desv.Est   N Error.Est   Error Lim.Inf Lim.Sup
## 1 22.61037 19.94936 1196 0.5768506 0.949571 21.6608 23.55994
```

Cálculo de intervalo de confianza para una proporción

También podemos calcular un intervalo de confianza para una proporción a partir de una muestra grande.

$$IC_{1-\alpha} = p \pm Z\sigma_p$$

Donde:

- $1 - \alpha$ es el nivel de confianza deseado
- Z es el valor de la distribución normal estándar que representa el nivel de confianza
- σ_p es el error estándar de la proporción que se obtiene mediante:

$$\sigma_p = \sqrt{\frac{pq}{n}}$$

Donde:

- p es la proporción de la categoría de análisis en nuestra muestra
- q es igual a $1 - p$

Ejemplo

Vamos a usar la variable P32

```
library(descr)
freq(genero$P32, plot = FALSE)

## genero$P32
##                                     Frequency Percent Valid P
ercent
## Hago mucho más de lo que me corresponde          147 12.219
21.304
## Hago algo más de lo que me corresponde          141 11.721
20.435
## Hago más o menos lo que me corresponde          256 21.280
37.101
## Hago algo menos de lo que me corresponde          65  5.403
9.420
## Hago mucho menos de lo que me corresponde          53  4.406
7.681
## No contesta                                         28  2.328
4.058
## NA's                                              513 42.643
## Total                                         1203 100.000
00.000
```

Preparamos los datos para el cálculo del intervalo de confianza, en este caso vamos a seleccionar una categoría: las personas que a la pregunta P32 responden "Hago más o menos lo que me corresponde" en la distribución de las tareas domésticas en su hogar.

```
p32r <- na.omit(genero$P32)
cat <- ifelse(p32r=="Hago más o menos lo que me corresponde", 1, 0)
prop.table(table(cat))

## cat
##          0           1
## 0.6289855 0.3710145

p <- mean(cat)
p # Esta es la proporción de personas que respondieron esta opción

## [1] 0.3710145
```

Preparamos el resto de la información

```
n <- length(cat) # Tamaño de la muestra
error.est.p <- sqrt((p*(1-p))/n) # Error estándar de la proporción
error.p <- 2 * error.est.p # Usamos Z = 2 para indicar un nivel de confianza del 95%
lim.inf.p <- p - error.p
lim.sup.p <- p + error.p
result.p <- data.frame(p, n, error.est.p, error.p, lim.inf.p, lim.sup.p)
result.p

##          p      n error.est.p   error.p lim.inf.p lim.sup.p
## 1 0.3710145 690  0.0183904 0.0367808 0.3342337 0.4077953
```

Función para cálculo de intervalo de confianza de proporción

Generamos la función

```
int.conf.p <- function(x, cat, ic = 95){
  vcat <- na.omit(x)
  p <- mean(ifelse(vcat==cat, 1, 0))
  n <- length(vcat)
  error.est.p <- sqrt((p*(1-p))/n)
  beta <- 1-((100-ic)/100)/2
  error <- error.est.p * qt(beta, df = n-1)
  lim.inf.p <- p - error
  lim.sup.p <- p + error
  result.p <- data.frame(p, n, error.est.p, error, lim.inf.p, lim.sup.p)
  return(result.p)
}
```

Usamos la función

```

int.conf.p(genero$P32, "Hago más o menos lo que me corresponde")# IC95%
##           p   n error.est.p      error lim.inf.p lim.sup.p
## 1 0.3710145 690  0.0183904 0.03610795 0.3349065 0.4071224

int.conf.p(genero$P32, "Hago más o menos lo que me corresponde", ic=99)
## IC99%
##           p   n error.est.p      error lim.inf.p lim.sup.p
## 1 0.3710145 690  0.0183904 0.04750211 0.3235124 0.4185166

int.conf.p(genero$P32, "Hago más o menos lo que me corresponde", ic=90)
## IC90%
##           p   n error.est.p      error lim.inf.p lim.sup.p
## 1 0.3710145 690  0.0183904 0.03029025 0.3407242 0.4013047

freq(genero$P10, plot=FALSE)

## genero$P10
##             Frequency   Percent
## Si            1070 88.94431
## No             132 10.97257
## No contesta       1 0.08313
## Total          1203 100.00000

int.conf.p(genero$P10, "Si")
##           p   n error.est.p      error lim.inf.p lim.sup.p
## 1 0.8894431 1203 0.009041058 0.01773801 0.871705 0.9071811

```

Gráficar el intervalo de confianza

Podemos crear un gráfico del intervalo de confianza de la media de una variable. En este ejemplo usaremos el gráfico de puntos del paquete ggplot2. Vamos a representar gráficamente el promedio de horas semanales dedicadas a labores domésticas que le dedican los hombres y las mujeres.

Primero vamos a generar un data frame que contenga los estadísticos que necesitamos graficar (la media, los límites inferior y superior del intervalo de confianza). Para ello usaremos la función "summarySE" del paquete "Rmisc" (que debe instalarse en su sistema). Esta función summarySE hace algo parecido a la función que hemos creado en una de las diapositivas anteriores.

```

library(Rmisc)

## Loading required package: lattice
## Loading required package: plyr

df <- summarySE(genero, measurevar="p19ar", groupvars="SEXO", na.rm=T)
df

```

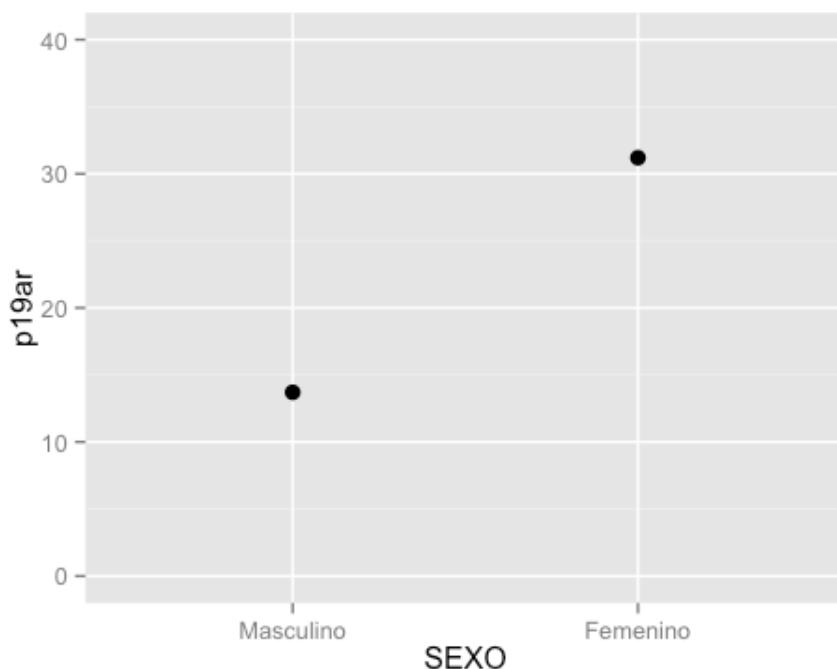
```
##          SEXO   N    p19ar      sd      se      ci
## 1 Masculino 587 13.70187 13.50444 0.5573881 1.094722
## 2 Femenino  609 31.19704 21.35332 0.8652803 1.699301
```

Generar un gráfico de puntos:

Un gráfico de puntos que muestra las medias de p19ar según sexo

```
library(ggplot2)
graf.punto1 <- ggplot(df, aes(x=SEXO, y=p19ar)) +
  geom_point(size = 3) + ylim(0, 40)

graf.punto1
```



Ahora añadimos las barras que indican el intervalo de confianza, con títulos

```
graf.punto <- graf.punto1 + geom_errorbar(aes(ymin=p19ar-ci, ymax=p19
ar+ci), width = 0.2) + ylab("Horas semanales") +
  ggtitle("Intervalo de confianza al 95% para la media de las horas se
  manales\n dedicadas al trabajo doméstico, según sexo")

graf.punto
```

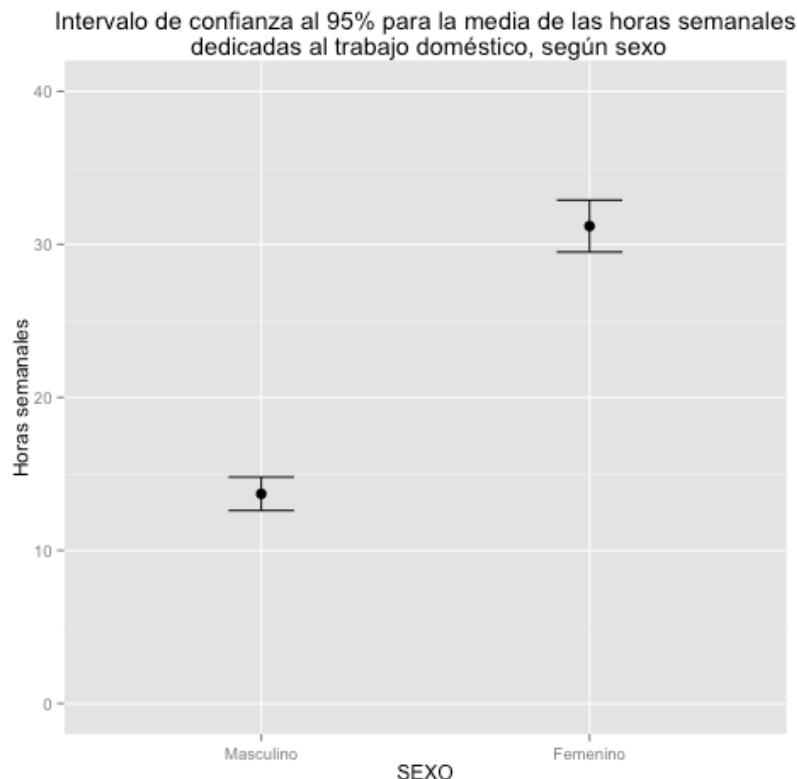
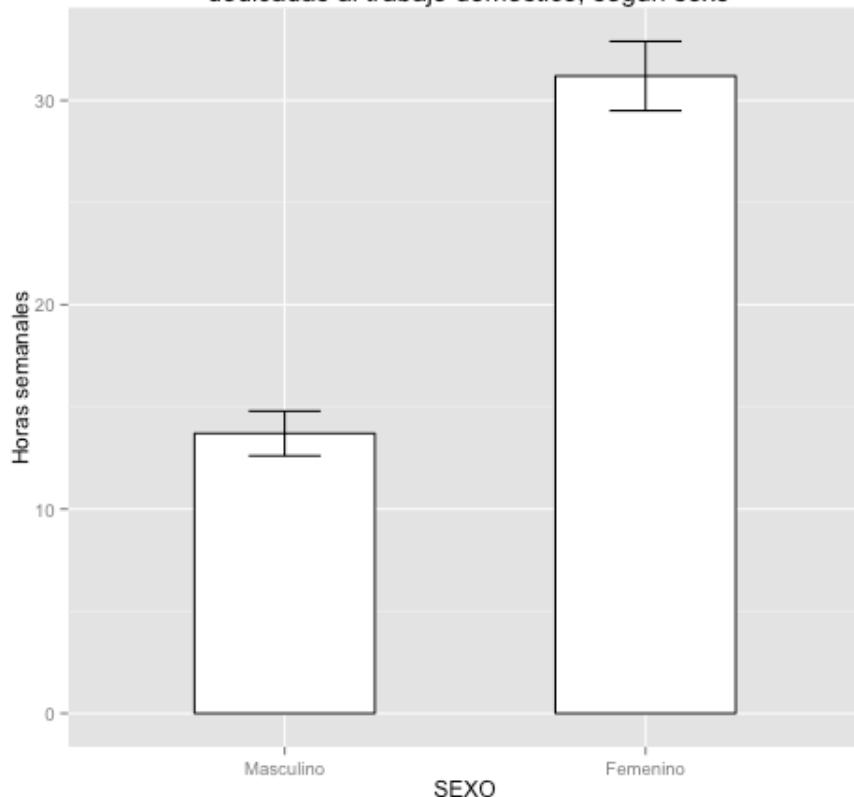


Gráfico de barras para intervalo de confianza

Podemos graficar el mismo concepto con barras:

```
graf.barra1 <- ggplot(df, aes(x=SEXO, y=p19ar)) +
  geom_bar(width=0.5, fill="white", colour="black", stat="identity") +
  geom_errorbar(aes(ymin=p19ar-ci, ymax=p19ar+ci), width = 0.2) +
  ylab("Horas semanales") + ggtitle("Intervalo de confianza al 95% para la media de las horas semanales\ndedicadas al trabajo doméstico, según sexo")
graf.barra1
```

Intervalo de confianza al 95% para la media de las horas semanales dedicadas al trabajo doméstico, según sexo



Algunas variaciones

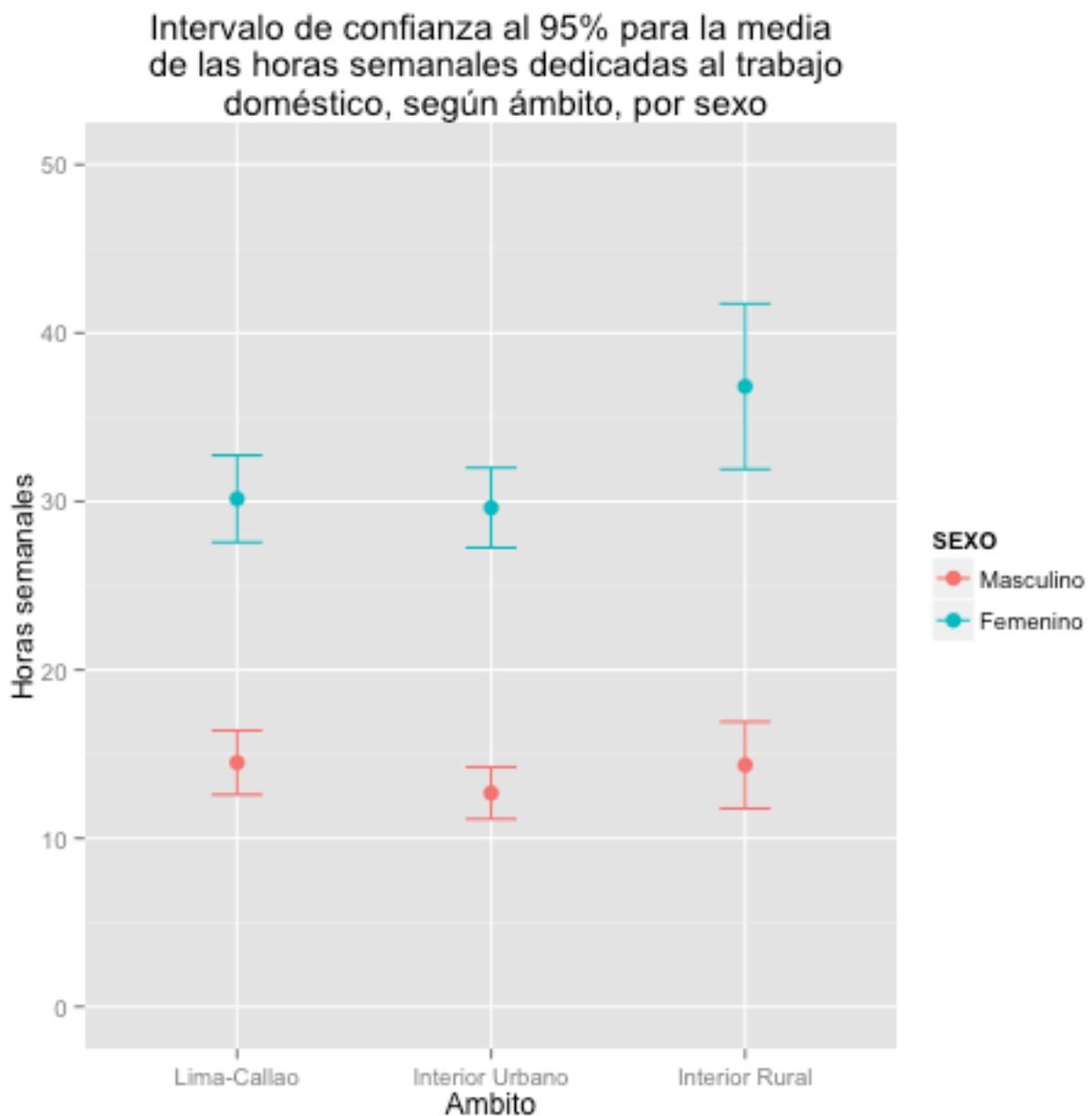
Según ámbito y sexo del entrevistado

```
df2 <- summarySE(genero, measurevar="p19ar", groupvars=c("SEXO", "Ambito"),
  na.rm=T)
df2

##          SEXO            Ambito     N      p19ar       sd      se      c
#i
## 1 Masculino      Lima-Callao 218 14.49083 14.27131 0.9665754 1.90507
## 2 Masculino Interior Urbano 246 12.68293 12.26298 0.7818591 1.54002
## 3 Masculino Interior Rural 123 14.34146 14.42552 1.3007051 2.57487
## 4 Femenino       Lima-Callao 229 30.15721 19.83378 1.3106528 2.58254
## 5 Femenino Interior Urbano 264 29.62500 19.59105 1.2057460 2.37414
## 6 Femenino Interior Rural 116 36.82759 26.70557 2.4795498 4.91151
```

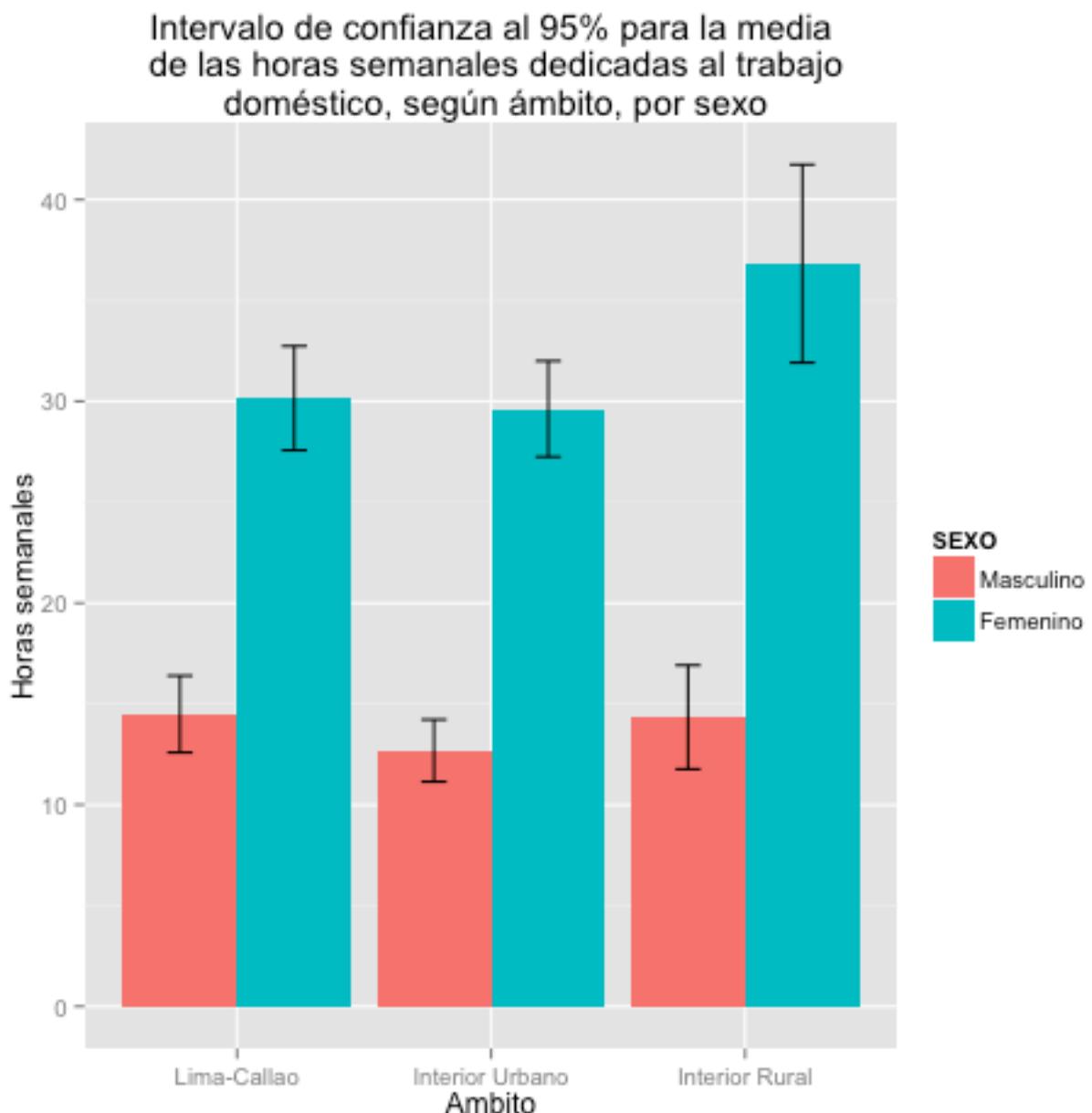
El gráfico de puntos

```
graf.punto2 <- ggplot(df2, aes(x=Ambito, y=p19ar, colour=SEXO)) +
  geom_point(size = 3) + ylim(0, 50) +
  geom_errorbar(aes(ymin=p19ar-ci, ymax=p19ar+ci), width = 0.2) +
  ylab("Horas semanales") +
  ggtitle("Intervalo de confianza al 95% para la media\n de las horas semanales dedicadas al trabajo\n doméstico, según ámbito, por sexo")
graf.punto2
```



El gráfico de barras

```
graf.bar2 <- ggplot(df2, aes(x=Ambito, y=p19ar, fill=SEXO)) +
  geom_bar(position="dodge", stat = "identity") +
  geom_errorbar(aes(ymin=p19ar-ci, ymax=p19ar+ci), width = 0.2,
  position=position_dodge(.9)) + ylab("Horas semanales") +
  ggtitle("Intervalo de confianza al 95% para la media\n de las horas semanales dedicadas al trabajo\n doméstico, según ámbito, por sexo")
graf.bar2
```



Pruebas de Hipótesis

Inferencia estadística y Pruebas de Hipótesis

En esta parte del curso examinaremos las siguientes herramientas de inferencia estadística

- Pruebas para medias de muestra única
- Pruebas para medias de dos muestras independientes
- Pruebas para medias de dos muestras relacionadas

Cargamos los datos de trabajo

Base de datos para estos ejercicios: Familia y roles de género 2012, a descargar de:

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

Se sugiere descargar también el cuestionario para utilizarlo como referencia de libro de códigos. 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
```

Preparar las variables de trabajo

En estos ejemplos trabajaremos con las siguientes variables:

- Edad ideal para que una mujer se case (P1)
- Edad ideal para que un hombre se case (P2)
- Edad en la que el entrevistado se casó o empezó a convivir (P23)
- Número de hijos (P44A)
- Horas dedicadas al trabajo doméstico (P19A)
- Horas que la pareja le dedica al trabajo doméstico (P28A)

Antes de utilizar estas variables es necesario evaluar si necesitan algún tipo de acondicionamiento (identificar valores perdidos o "raros", recodificar, etc)

Acondicionamiento de las variables

Las siguientes transformaciones nos permitirán acondicionar las variables para utilizarlas en el análisis:

```
genero$p1r <- genero$P1
genero$p2r <- genero$P2
genero$p1r[genero$P1 > 40] <- NA
genero$p2r[genero$P2 == 99] <- NA
```

```
genero$p19ar <- genero$P19A
genero$p19ar[genero$P19A >= 140] <- NA

genero$p28ar <- genero$P28A
genero$p28ar[genero$P28A > 120] <- NA
```

Pruebas de hipótesis para medias de muestra única

Las pruebas de hipótesis para medias de muestra única nos sirven para comparar un estadístico muestral con un parámetro o un valor de referencia establecido para la comparación.

En la base de datos de género se preguntó acerca del número de hijos que ha tenido una mujer. La media de esta variable fue de 2.207.

```
genero.f <- subset(genero, SEXO=="Femenino")
summary(genero.f$P44A)

##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##    0.000   1.000   2.000   2.207   3.000  16.000
```

De acuerdo con el censo de población del 2007, el número promedio de hijos que han tenido las mujeres de 15 años o más era de 2.63. ¿El resultado que hemos obtenido en la muestra de la encuesta del 2012 es significativamente diferente al del censo?

Pasos para una prueba de hipótesis

De acuerdo con el texto de Ritchey (2006) en toda prueba de hipótesis es necesario considerar 5 pasos:

1. Definir claramente la Hipótesis Cero y la Hipótesis Alternativa
2. Seleccionar la distribución de muestreo para la Hipótesis Cero
3. Definir el nivel de significancia o α
4. Calcular el estadístico de la prueba
5. Decidir si se acepta o rechaza la Hipótesis Cero

Paso 1: Formular las hipótesis

$$H_0: \bar{x} - \mu = 0$$

$$H_1: \bar{x} - \mu \neq 0$$

Paso 2: Se utilizará una distribución de t de Student

Paso 3: Trabajaremos con un nivel de significancia del 5% ó un $\alpha = 0.05$

Paso 4: El estadístico de la prueba se calcula de la siguiente manera:

$$t = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

```
t.test(genero.f$P44A, mu=2.63)

##
##  One Sample t-test
##
## data: genero.f$P44A
## t = -5.0558, df = 613, p-value = 5.669e-07
## alternative hypothesis: true mean is not equal to 2.63
## 95 percent confidence interval:
##  2.042469 2.371211
## sample estimates:
## mean of x
##  2.20684
```

Paso 5: Tomar una decisión

- Si el p-value del estadístico de la prueba es menor al nivel de significancia, se rechaza la Hipótesis Cero y se adopta la hipótesis 1.
- Si el p-value del estadístico de la prueba es mayor o igual al nivel de significancia, se acepta la Hipótesis Cero

En este caso se **rechaza** la Hipótesis Cero. Existen diferencias estadísticamente significativas entre el estadístico muestral y un parámetro igual a 2.63.

Otro ejemplo

¿Las mujeres le dedican a realizar labores domésticas en su hogar más de 30 horas semanales?

```
t.test(genero$p19ar[genero$SEXO=="Femenino"], mu=30)

##
##  One Sample t-test
##
## data: genero$p19ar[genero$SEXO == "Femenino"]
## t = 1.3834, df = 608, p-value = 0.167
## alternative hypothesis: true mean is not equal to 30
## 95 percent confidence interval:
##  29.49774 32.89635
## sample estimates:
## mean of x
##  31.19704
```

Si consideramos un $\alpha = 0.05$, notamos que el p-value para el estadístico de la prueba es menor que nuestro nivel de significancia. Por lo tanto en este caso se **acepta** la Hipótesis Cero. No existen diferencias estadísticamente significativas entre la media muestral y un valor de comparación igual a 30.

Prueba de una y de dos colas

Las pruebas de t que hemos visto son de tipo **bidireccional** o de **dos colas**.
Cuando:

$$H_0: \bar{x} - \mu = 0$$

$$H_1: \bar{x} - \mu \neq 0$$

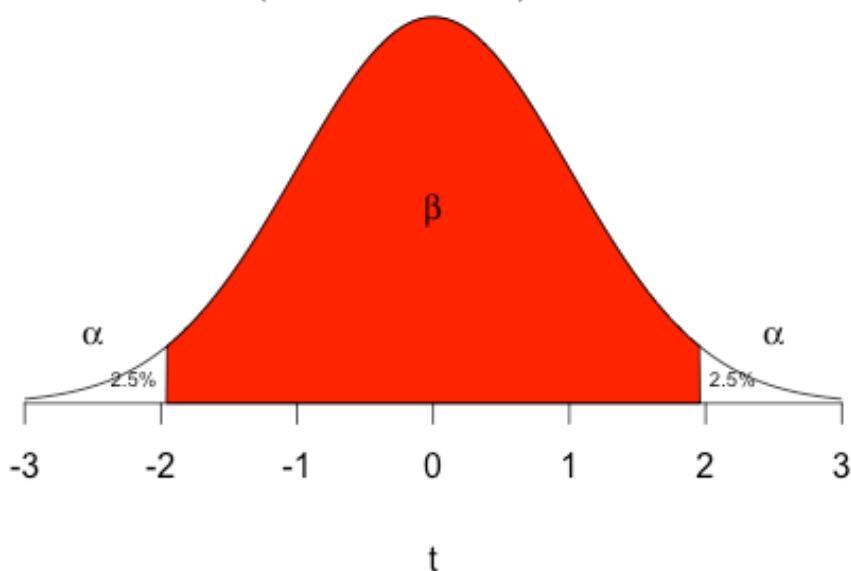
No importa hacia qué lado de la distribución de muestreo cae el valor del estadístico de la prueba, por lo que se consideran ambas colas para la región α .

Zonas α y β en prueba de dos colas

Prueba de dos colas con nivel de significancia = 0.05

Distribución de t para N > 121

$$P(-1.96 < t < 1.96) = 0.95$$



Prueba de una cola o unidireccional

En cambio si:

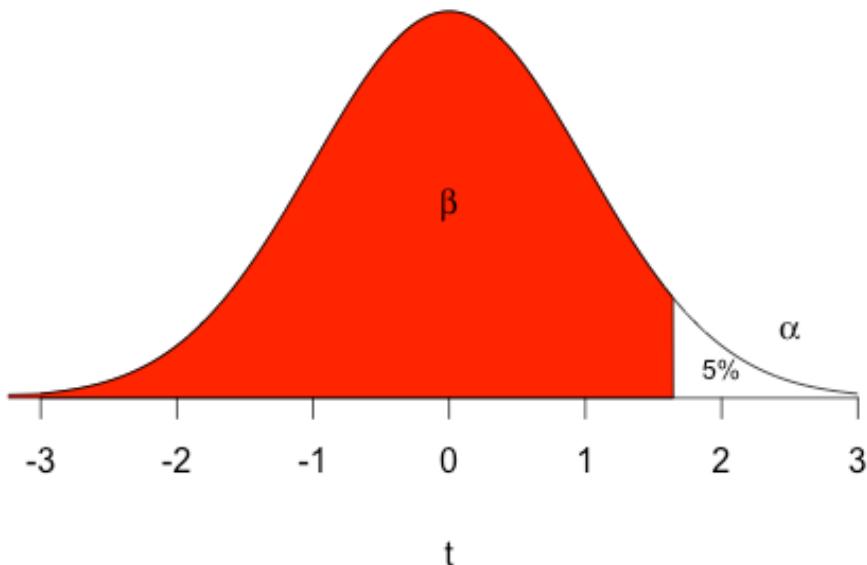
$$H_0: \bar{x} - \mu \leq 0$$

$$H_1: \bar{x} - \mu > 0$$

Prueba de una cola con nivel de significancia = 0.05

Distribución de t para N > 121

$$P(t < 1.645) = 0.95$$



Ejemplo:

¿Las mujeres le dedican a realizar labores domésticas en su hogar más de 30 horas semanales?

Prueba de dos colas:

```
t.test(genero$p19ar[genero$SEXO=="Femenino"], mu=30)

##
## One Sample t-test
##
## data: genero$p19ar[genero$SEXO == "Femenino"]
## t = 1.3834, df = 608, p-value = 0.167
## alternative hypothesis: true mean is not equal to 30
## 95 percent confidence interval:
## 29.49774 32.89635
## sample estimates:
## mean of x
## 31.19704
```

Prueba de una cola, asumiendo una hipótesis alternativa positiva ("mayor que")

```
t.test(genero$p19ar[genero$SEXO=="Femenino"], mu=30, alternative = "greater")

##
## One Sample t-test
##
```

```
## data: genero$p19ar[genero$SEXO == "Femenino"]
## t = 1.3834, df = 608, p-value = 0.08352
## alternative hypothesis: true mean is greater than 30
## 95 percent confidence interval:
## 29.77161      Inf
## sample estimates:
## mean of x
## 31.19704
```

Prueba de hipótesis para medias de dos muestras independientes

Nos sirven para comparar las medias muestrales de dos grupos independientes. Por ejemplo, ¿la edad en la que una personas se casó o empezó a convivir es la misma entre hombres y mujeres?

En este caso las hipótesis se formulan de la siguiente manera:

$$H_0: \bar{x}_1 - \bar{x}_2 = 0$$

$$H_1: \bar{x}_1 - \bar{x}_2 \neq 0$$

Se siguen los mismos pasos que en las pruebas de hipótesis de una muestra. En este caso se usa también la distribución de t de Student como distribución de muestreo. Para el cálculo del estadístico de la prueba se debe estimar el error estándar de la **diferencia de medias**:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)\sigma_{x_1}^2 + (n_2 - 1)\sigma_{x_2}^2}{n_1 + n_2 - 2} \sqrt{\frac{n_1 + n_2}{n_1 n_2}}}}$$

Ejemplo

En el ejemplo propuesto, vamos a comparar la edad promedio en la que se casaron o empezaron a convivir hombres y mujeres.

- La hipótesis cero, sostiene que no hay diferencias entre hombres y mujeres
- La hipótesis alternativa sostiene que sí hay diferencias (se trata de una prueba de dos colas)
- La distribución de muestreo para H_0 será la distribución de t de Student
- Utilizaremos un nivel de significancia del 5% ó $\alpha = 0.05$

Para calcular los resultados de la prueba usamos el comando:

```
t.test(genero$P23~genero$SEXO)
```

```
##  
## Welch Two Sample t-test  
##  
## data: genero$P23 by genero$SEXO  
## t = 8.9404, df = 803.979, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 2.527755 3.949971  
## sample estimates:  
## mean in group Masculino mean in group Femenino  
## 25.73464 22.49578
```

Como se aprecia, el p-value del estadístico de la prueba resulta ser menor al nivel de significancia seleccionado, por lo tanto se **rechaza** la hipótesis cero. Los hombres empezaron a convivir o se casaron a una edad mayor que las mujeres y esas diferencias son estadísticamente significativas con un $\alpha = 0.05$.

Otro ejemplo:

¿Los hombres y las mujeres tienen la misma opinión respecto de cuál es la edad ideal para que una mujer se case?. Hagamos la prueba considerando un $\alpha = 0.05$

```
t.test(genero$p1r~genero$SEXO)  
  
##  
## Welch Two Sample t-test  
##  
## data: genero$p1r by genero$SEXO  
## t = -2.3819, df = 1156.439, p-value = 0.01738  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.91486807 -0.08844004  
## sample estimates:  
## mean in group Masculino mean in group Femenino  
## 25.93794 26.43960
```

¿Y respecto de la edad ideal para que un hombre se case?

```
t.test(genero$p2r~genero$SEXO)  
  
##  
## Welch Two Sample t-test  
##  
## data: genero$p2r by genero$SEXO  
## t = -3.081, df = 1159.999, p-value = 0.002112  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.1398995 -0.2529314  
## sample estimates:  
## mean in group Masculino mean in group Femenino  
## 28.19788 28.89430
```

Veamos cómo se ven las diferencias en las horas dedicadas a labores domésticas entre las mujeres de Lima-Callao vs las de las ciudades del interior del país.

```
genero.s2 <- subset(genero, SEXO=="Femenino" & Ambito!="Interior Rural")
t.test(genero.s2$p19ar~genero.s2$Ambito)

##
## Welch Two Sample t-test
##
## data: genero.s2$p19ar by genero.s2$Ambito
## t = 0.2988, df = 479.493, p-value = 0.7652
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -2.967144 4.031554
## sample estimates:
## mean in group Lima-Callao mean in group Interior Urbano
## 30.15721 29.62500
```

Pruebas de hipótesis para medias de muestras relacionadas

Este tipo de pruebas nos sirve para comparar las puntuaciones de dos variables en el mismo grupo de personas. Por ejemplo, si queremos determinar si existen diferencias entre la edad considerada como ideal para que una mujer se case y para que un hombre se case.

En este caso lo que se compara es la diferencia entre las puntuaciones de ambas variables:

$$\bar{D} = \frac{\sum(x_1 - x_2)}{n}$$

Las hipótesis se formulan en los siguientes términos:

$$H_0: \bar{D} = 0$$

$$H_1: \bar{D} \neq 0$$

Para el caso de muestras relacionadas, el estadístico de la prueba se calcula de la siguiente manera:

$$t = \frac{\bar{D}}{\sigma_{\bar{D}}}$$

Donde:

$$\sigma_{\bar{D}} = \frac{\sigma_D}{\sqrt{n}}$$

Calculamos el estadístico de la prueba para el ejemplo propuesto. Definimos como nivel de significancia un $\alpha = 0.05$

```
genero.s3 <- na.omit(data.frame(p1 = genero$p1r, p2= genero$p2r))
summary(genero.s3)

##          p1            p2
##  Min.   :15.0   Min.   :18.00
##  1st Qu.:25.0   1st Qu.:25.00
##  Median :25.0   Median  :29.00
##  Mean   :26.2   Mean   :28.57
##  3rd Qu.:29.0   3rd Qu.:30.00
##  Max.   :40.0   Max.   :42.00

t.test(genero.s3$p1, genero.s3$p2, paired=TRUE)

##
##  Paired t-test
##
##  data:  genero.s3$p1 and genero.s3$p2
##  t = -30.3124, df = 1157, p-value < 2.2e-16
##  alternative hypothesis: true difference in means is not equal to 0
##  95 percent confidence interval:
##  -2.528495 -2.221073
##  sample estimates:
##  mean of the differences
##                      -2.374784
```

El resultado de la prueba nos lleva a rechazar la hipótesis cero. Por lo tanto, con un nivel de significancia del 5% (incluso menor), podemos decir que la edad considerada ideal para que un hombre se case es mayor que la de una mujer.

Otro ejemplo:

En el caso de los hombres, ¿la edad considerada como ideal para que un hombre se case es muy diferente que la edad en que se empezó a convivir o se casó?

```
genero.s4 <- na.omit(data.frame(p2=genero$p2r, p23=genero$P23, sexo=genero$SEXO))
genero.s4 <- subset(genero.s4, sexo=="Masculino")
summary(genero.s4)

##          p2            p23           sexo
##  Min.   :18.00   Min.   :10.00   Masculino:390
##  1st Qu.:25.00   1st Qu.:22.00   Femenino  : 0
##  Median :28.00   Median  :25.00
##  Mean   :27.98   Mean   :25.76
##  3rd Qu.:30.00   3rd Qu.:29.00
##  Max.   :40.00   Max.   :60.00
```

Solicitamos la prueba correspondiente:

```
t.test(genero.s4$p2, genero.s4$p23, paired=TRUE)
```

```
##  
## Paired t-test  
##  
## data: genero.s4$p2 and genero.s4$p23  
## t = 6.8822, df = 389, p-value = 2.369e-11  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 1.586165 2.854860  
## sample estimates:  
## mean of the differences  
## 2.220513
```

Considerando un nivel de significancia del 5% se rechaza la hipótesis cero y se concluye que los hombres se casaron o empezaron a convivir a una edad menor a la que se considera ideal para que ello ocurra.

Tablas de Contingencia

Tablas de contingencia como herramienta de análisis bivariante

Uno de los objetivos del análisis estadístico es evaluar la existencia de relaciones entre dos variables.

Se dice que dos variables están relacionadas cuando la distribución de los valores de una variable se modifica o presenta cambios asociados a la distribución de los valores de otra variable.

En un análisis bivariante se puede distinguir entre:

- Variable dependiente o de análisis (Y): Es la variable cuya distribución se quiere explicar.
- Variable independiente (X): Es la variable que nos puede ayudar a comprender por qué varía la distribución de la variable dependiente

Cuando trabajamos con variables categóricas o cualitativas (nominales y ordinales), una de las herramientas de análisis son las tablas de contingencia.

Una tabla de contingencia es una tabla de frecuencias cruzadas, que nos permite observar la distribución de los casos en categorías cruzadas de dos variables.

Ejemplo

Para este ejemplo vamos a trabajar con la base de datos de la encuesta sobre "Representación Política y Conflictos" realizada por el Instituto de Opinión Pública de la PUCP en noviembre del 2012.

```
library(foreign)
conf <- as.data.frame(read.spss("IOP_1112_01_B.sav"))
```

En esta base de datos podemos identificar dos variables:

- P2: Interés en la Política
- NSEGrup: Nivel Socioeconómico del Entrevistado (agrupado)

La pregunta que vamos a tratar de responder es si existe alguna relación entre el NSE y el Interés en la Política. En este ejemplo, el interés en la política será nuestra variable dependiente y el NSE la variable independiente.

Elaborar una tabla de contingencia

El primer paso es mostrar la distribución de frecuencias de nuestras variables:

```
library(descr)
freq(conf$P2, plot=FALSE)

## conf$P2
## Frequency Percent
```

```

## Muy interesado      81  6.733
## Algo interesado   268 22.278
## Poco interesado   538 44.722
## Nada interesado   299 24.855
## NS/NR              17  1.413
## Total               1203 100.000
    
```

Se aprecia que hay algunos casos que son NS/NR que vamos a recodificar como "missing values"

```

conf$p2r <- conf$P2
conf$p2r[conf$P2=="NS/NR"] <- NA
conf$p2r <- factor(conf$p2r)
freq(conf$p2r, plot=FALSE)

## conf$p2r
##             Frequency Percent Valid Percent
## Muy interesado      81  6.733      6.83
## Algo interesado    268 22.278     22.60
## Poco interesado    538 44.722     45.36
## Nada interesado    299 24.855     25.21
## NA's                 17  1.413
## Total                1203 100.000     100.00

freq(conf$NSEGrup, plot=FALSE)

## conf$NSEGrup
##             Frequency Percent
## A/B            257  21.36
## C              365  30.34
## D/E            581  48.30
## Total          1203 100.00
    
```

Las frecuencias de cada una de nuestras variables se conocen en el análisis de tablas de contingencia como **frecuencias marginales**.

Elaborar una tabla de contingencia (2)

Una forma rápida y simple para realizar una tabla de contingencia es usar el comando:

```

table(conf$p2r, conf$NSEGrup)

##
##             A/B   C D/E
## Muy interesado 29  29  23
## Algo interesado 72  97  99
## Poco interesado 101 153 284
## Nada interesado 53  83 163
    
```

Reglas para elaborar y analizar una tabla de contingencia

Para poder ser analizada adecuadamente, una tabla de contingencia debe elaborarse tomando en cuenta las siguientes reglas:

- Las frecuencias cruzadas deben expresarse en porcentajes
- Los porcentajes de cada celda de frecuencias cruzadas deben calcularse en función de las categorías de la variable **independiente**.
 - Si las categorías de la variable independiente están en las columnas, cada columna debe sumar 100%
 - Si las categorías de la variable independiente están en las filas, cada fila debe sumar 100%

Elaborar una tabla de contingencia (3)

Caso en que las categorías de la variable independiente están en las columnas:

```
tabla1 <- prop.table(table(conf$p2r, conf$NSEGrup), 2)*100
tabla1

##
##          A/B        C        D/E
## Muy interesado 11.372549 8.011050 4.042179
## Algo interesado 28.235294 26.795580 17.398946
## Poco interesado 39.607843 42.265193 49.912127
## Nada interesado 20.784314 22.928177 28.646749
```

Elaborar una tabla de contingencia (4)

Caso en que las categorías de la variable independiente están en las filas:

```
prop.table(table(conf$NSEGrup, conf$p2r), 1)*100

##
##      Muy interesado Algo interesado Poco interesado Nada interesad
## 0
##  A/B      11.372549      28.235294      39.607843      20.78431
## 4
##  C       8.011050      26.795580      42.265193      22.92817
## 7
##  D/E      4.042179      17.398946      49.912127      28.64674
## 9
```

¿Cómo leer una tabla de contingencia?

Para leer una tabla de contingencia, la regla es comparar los porcentajes de las celdas en el sentido contrario al que fueron calculados. El objetivo es identificar si la distribución de la variable dependiente cambia cuando nos movemos entre las categorías de la variable independiente. Si eso ocurre es un indicador de que ambas variables pueden estar asociadas.

```
##          A/B      C      D/E
## Muy interesado 11.37 8.01 4.04
## Algo interesado 28.24 26.80 17.40
## Poco interesado 39.61 42.27 49.91
## Nada interesado 20.78 22.93 28.65
```

Análisis bivariable

Cuando se realiza un análisis bivariable utilizando tablas de contingencia es necesario evaluar los siguientes elementos:

- Describir cómo cambia la distribución de la variable dependiente entre diversas categorías de la variable independiente.
- Determinar si la asociación observada es una asociación estadísticamente significativa.
- Medir la fuerza de la relación entre ambas variables.

Para el segundo paso se usa el estadístico y la prueba de X^2

Prueba de X^2

La prueba de X^2 es una prueba de hipótesis no paramétrica. Compara la tabla de contingencia que hemos elaborado con una tabla **hipotética** de frecuencias **esperadas**. Esta tabla hipotética representa el supuesto de cómo sería la distribución de frecuencias cruzadas **si no existiese una relación** entre las dos variables que están en la tabla de contingencia.

Tabla de frecuencias observadas

Esta es la tabla de frecuencias observadas del ejemplo que estamos trabajando. A la tabla le hemos añadido las frecuencias marginales:

```
t1 <- table(conf$p2r, conf$NSEGrup)
addmargins(t1)

##          A/B      C      D/E   Sum
## Muy interesado 29     29     23    81
## Algo interesado 72     97     99   268
## Poco interesado 101    153    284   538
## Nada interesado 53     83     163   299
## Sum            255    362    569  1186
```

Las frecuencias esperadas

Las frecuencias esperadas (f_e) de una celda de una tabla de contingencia se calculan de la siguiente manera:

$$f_e = (f_c \times f_f) / f_t$$

Donde:

- f_c es la frecuencia marginal de la columna
- f_f es la frecuencia marginal de la fila
- f_t es el total de casos de la tabla

Tabla de frecuencias esperadas

En nuestro ejemplo, esta es la tabla de frecuencias observadas:

	A/B	C	D/E	Sum
Muy interesado	29	29	23	81
Algo interesado	72	97	99	268
Poco interesado	101	153	284	538
Nada interesado	53	83	163	299
Sum	255	362	569	1186

Y esta sería la tabla de frecuencias esperadas:

	A/B	C	D/E
Muy interesado	17.41568	24.72344	38.86088
Algo interesado	57.62226	81.80101	128.57673
Poco interesado	115.67454	164.21248	258.11298
Nada interesado	64.28752	91.26307	143.44941

Tabla de porcentajes de la tabla de frecuencias esperadas

Si calculamos los porcentajes de las columnas de la tabla de frecuencias esperadas, esta se vería así:

	A/B	C	D/E
Muy interesado	6.82968	6.82968	6.82968
Algo interesado	22.59696	22.59696	22.59696
Poco interesado	45.36256	45.36256	45.36256
Nada interesado	25.21079	25.21079	25.21079

Lo que muestra una tabla **hipotética** donde NO hay asociación entre ambas variables.

La prueba de X^2

X^2 es una prueba de hipótesis, y como tal se siguen los 5 pasos de toda prueba de hipótesis:

Paso 1

Formular la Hipótesis cero y la Hipótesis alternativa. En este caso la hipótesis cero sostiene que NO hay asociación entre las variables de la tabla, y la hipótesis uno sostiene lo contrario:

$$H_0: f_e = f_o$$

$$H_1: f_e \neq f_o$$

Paso 2

Seleccionar una distribución de muestreo para H_0 , en este caso la de X^2

Paso 3

Seleccionar un nivel de significancia para la prueba. Los usuales son un $\alpha = 0.05$ ó un $\alpha = 0.01$

Paso 4

Cálculo del estadístico de la prueba. El estadístico de la prueba de X^2 se calcula de la siguiente manera:

$$X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Paso 5

Decidir si se acepta o rechaza la H_0

Se **rechaza** la hipótesis cero cuando:

- El estadístico de la prueba cae en la zona α de la distribución de X^2
- La significancia o el p-value del estadístico de la prueba es menor a la significancia o valor α que hemos fijado en el paso 3.

Cálculo de X^2 en R

En el R para calcular y solicitar la prueba de X^2 se procede de la siguiente manera:

Se pide la tabla de frecuencias observadas en números absolutos:

```
tabla1 <- table(conf$p2r, conf$NSEGrup)
tabla1

##
##          A/B   C D/E
## Muy interesado 29  29 23
## Algo interesado 72  97 99
## Poco interesado 101 153 284
## Nada interesado 53  83 163
```

Se ejecuta la función sobre la tabla creada

```
chisq.test(tabla1)
```

```
##  
## Pearson's Chi-squared test  
##  
## data: tabla1  
## X-squared = 38.7519, df = 6, p-value = 8.005e-07
```

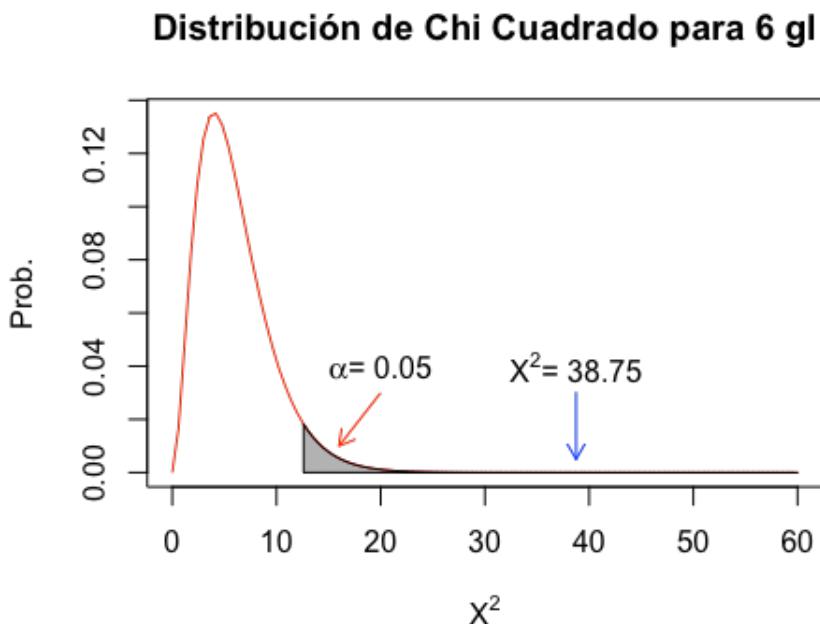
El valor de X^2 para la tabla1 es 38.75 con 6 grados de libertad (df)

En X^2 los grados de libertad son: $df = (col - 1) \times (fil - 1)$

Luego se compara el valor de X^2 de la tabla con el valor crítico de X^2 para una tabla con 6 grados de libertad y para una prueba con un nivel de significancia $\alpha = 0.05$ ó de 5%. Para hallar ese valor crítico se puede pedir:

```
qchisq(.95, df=6)  
## [1] 12.59159
```

En el siguiente gráfico podemos comparar el valor crítico de $X^2 = 12.59$ para seis grados de libertad y un $\alpha = 0.05$ con el valor calculado de $X^2 = 38.75$ para nuestra tabla:



```
chisq.test(tabla1)  
##  
## Pearson's Chi-squared test  
##  
## data: tabla1  
## X-squared = 38.7519, df = 6, p-value = 8.005e-07
```

La otra forma de tomar una decisión es comparar el p-value o significancia del estadístico de la prueba con el nivel de significancia que hemos establecido para la misma. Notamos que el p-value es mucho menor que nuestro nivel de significancia (0.05), lo que nos lleva a **rechazar H₀**.

Si rechazamos H₀, nuestra conclusión es que las variables de nuestra tabla tienen una asociación estadísticamente significativa.

Otro ejemplo

Interés en la política según sexo del entrevistado:

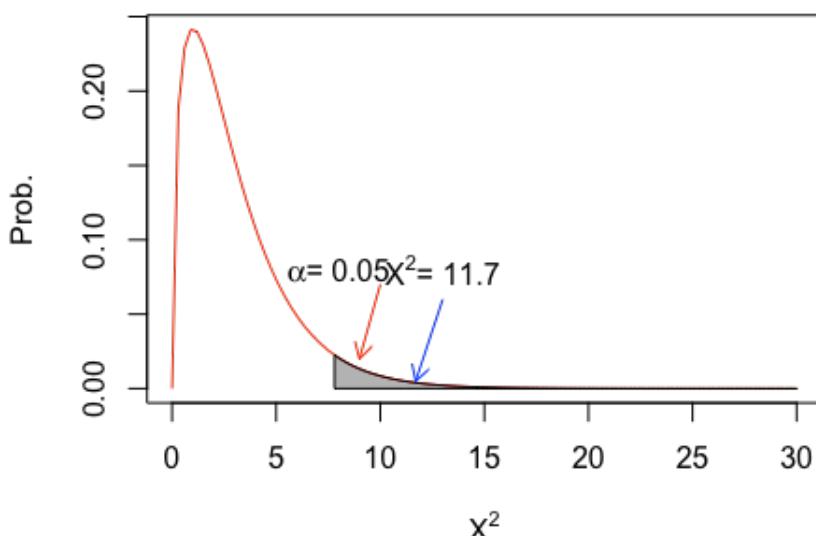
```
tabla2 <- table(conf$p2r, conf$SEXO)
prop.table(tabla2, 2)*100

##
##          Masculino   Femenino
## Muy interesado 8.561644 5.149502
## Algo interesado 25.171233 20.099668
## Poco interesado 41.952055 48.671096
## Nada interesado 24.315068 26.079734

chisq.test(tabla2)

##
## Pearson's Chi-squared test
##
## data: tabla2
## X-squared = 11.7437, df = 3, p-value = 0.008315
```

Distribución de Chi Cuadrado para 3 gl



```
qchisq(.95, df=3) # Valor crítico de Chi2 para sig. 0.05
```

```
## [1] 7.814728
```

Otro ejemplo

¿La aprobación presidencial está asociada con el sexo del entrevistado?

```
tabla3 <- table(conf$P7A, conf$SEXO)
prop.table(tabla3, 2)*100

##
##          Masculino Femenino
##  Aprueba    41.116751 37.418301
##  Desaprueba 48.730964 53.104575
##  NS/NR      10.152284  9.477124

chisq.test(tabla3)

##
##  Pearson's Chi-squared test
##
##  data: tabla3
##  X-squared = 2.3166, df = 2, p-value = 0.314
```

Requisitos para usar la prueba de X^2

Esta prueba sólo puede usarse cuando:

- Por lo menos el 80% de las celdas tienen frecuencias esperadas mayores a 5
- Ninguna de las celdas tiene frecuencias observadas iguales a cero

Si no se cumplen estos requisitos, el resultado de X^2 no tiene ningún sentido y no se puede interpretar. En estos casos es necesario reducir las dimensiones de la tabla (agrupar categorías), hasta que se cumplan los requisitos de la tabla.

Medidas o coeficientes de asociación

El tercer paso del análisis de tablas de contingencia es medir la fuerza de la asociación entre dos variables categóricas.

Una herramienta para ello son las medidas de asociación. A partir de las pruebas de X^2 podemos calcular tres tipos de medidas de asociación para variables categóricas:

- El coeficiente Phi (ϕ) : Para tablas 2x2
- El coeficiente de contingencia: Para tablas de cualquier dimensión
- La V de Cramer : Para tablas mayores a 2x2

Estos coeficientes miden la asociación entre dos variables categóricas en una escala que va de 0 a 1. Cuanto mayor es el valor del coeficiente, mayor será la asociación.

Uso de medidas de asociación

Veamos las tabla 1 y 2:

```
prop.table(tabla1, 2)*100

##
##          A/B      C      D/E
## Muy interesado 11.372549 8.011050 4.042179
## Algo interesado 28.235294 26.795580 17.398946
## Poco interesado 39.607843 42.265193 49.912127
## Nada interesado 20.784314 22.928177 28.646749

prop.table(tabla2, 2)*100

##
##          Masculino  Femenino
## Muy interesado 8.561644 5.149502
## Algo interesado 25.171233 20.099668
## Poco interesado 41.952055 48.671096
## Nada interesado 24.315068 26.079734
```

Las medidas de asociación basadas en X^2 se piden con el siguiente comando:

```
library(vcd)
assocstats(tabla1)

##
##          X^2 df   P(> X^2)
## Likelihood Ratio 38.758 6 7.9833e-07
## Pearson         38.752 6 8.0053e-07
##
## ## Phi-Coefficient : 0.181
## Contingency Coeff.: 0.178
## Cramer's V       : 0.128
```

```
assocstats(tabla2)

##                      X^2 df  P(> X^2)
## Likelihood Ratio 11.793 3 0.0081277
## Pearson          11.744 3 0.0083146
##
## Phi-Coefficient   : 0.1
## Contingency Coeff.: 0.099
## Cramer's V        : 0.1
```

Medidas de asociación para variables ordinales

Cuando las DOS variables de una tabla de contingencia son ordinales, se puede usar el coeficiente de asociación de Gamma de Goodman y Kruskal. Como en el caso de los coeficientes basados en X^2 , Gamma mide la fuerza de la asociación en una escala de 0 a 1. El signo de Gamma nos indica si la relación es directa (+) o inversa (-).

Ejemplo del uso de gamma con la tabla 1

```
prop.table(tabla1, 2)*100

##
##                      A/B         C         D/E
## Muy interesado 11.372549 8.011050 4.042179
## Algo interesado 28.235294 26.795580 17.398946
## Poco interesado 39.607843 42.265193 49.912127
## Nada interesado 20.784314 22.928177 28.646749

library(vcdExtra)
GKgamma(tabla1)

## gamma       : 0.211
## std. error   : 0.038
## CI           : 0.136 0.285
```

Ejemplo de relación inversa

Gusto por partidos de izquierda (P13J) según posición izquierda - derecha del entrevistado (P17)

- Gusto por partidos de izquierda (X): 1 = Nada - 10 = Mucho
- Posicion izquierda - derecha (Y) : 0 = Extrema izquieda - 10 = Extrema Derecha

Se supone que cuanto mayor sea el valor de la variable X, menor será el valor de la variable Y.

Primero hay que transformar las variables

```
#Gusto por partidos de izquieda
library(car)
```

```

p13j <- as.numeric(conf$P13J)
p13j[p13j > 10] <- NA
p13j.r <- recode(p13j, "1:2=1; 3:4=2; 5:6=3; 7:8=4; 9:10=5")
p13j.r <- factor(p13j.r)
levels(p13j.r) <- c("1-2", "3-4", "5-6", "7-8", "9-10")
conf$part.izq <- p13j.r

# Posicion izquierda - derecha del entrevistado
p17r <- as.numeric(conf$P17)
p17r[p17r > 11] <- NA
p17r <- p17r-1
p17r2 <- recode(p17r, "0:2 = 1; 3:4 = 2; 5 = 3; 6:7=4; 8:10=5")
p17r2 <- factor(p17r2)
levels(p17r2) <- c("IZQ", "C-IZQ", "CEN", "C-DER", "DER")
conf$izde <- p17r2

#Distribución de gusto por partidos de izquierda (recodificada)
prop.table(table(conf$part.izq))*100

##
##      1-2      3-4      5-6      7-8      9-10
## 54.076739 21.223022 17.146283  4.916067  2.637890

#Distribución de posición del entrevistado en La escala I-D (recodificada)
prop.table(table(conf$izde))*100

##
##      IZQ      C-IZQ      CEN      C-DER      DER
## 9.808343 13.641488 39.120631 22.435175 14.994363

```

Tabla de contingencia de gusto por partidos de izquierda según posición izquierda - derecha del entrevistado y coeficiente de asociación Gamma de Goodman y Kruskal:

```

tab <- table(conf$part.izq, conf$izde)
prop.table(tab,2)*100

##
##      IZQ      C-IZQ      CEN      C-DER      DER
## 1-2  45.945946 27.884615 55.252918 54.000000 71.681416
## 3-4  16.216216 36.538462 22.178988 26.666667  9.734513
## 5-6  18.918919 26.923077 15.953307 18.000000 10.619469
## 7-8  10.810811  4.807692  5.058366  1.333333  5.309735
## 9-10  8.108108  3.846154  1.556420  0.000000  2.654867

GKgamma(tab)

## gamma       : -0.252
## std. error   : 0.046
## CI           : -0.341 -0.162

```

Ojo: El signo del coeficiente Gamma depende de la codificación de las categorías

Ejemplo: Interés en la política según nivel educativo del entrevistado. Primero preparamos la variable Nivel Educativo, para agrupar los casos

```

##                                     Ninguno      Inicial o primaria incompleta
##                               22                      91
##          Primaria completa      Secundaria incompleta
##                               106                     143
##          Secundaria completa      Superior técnica incompleta
##                               367                      87
##          Superior técnica completa Superior universitaria incompleta
##                               160                     108
##          Superior universitaria completa      Post grado
##                               102                      17
##                               NS/NR
##                               0

educ <- as.numeric(conf$DG4)
table(educ)

## educ
##   1   2   3   4   5   6   7   8   9   10
##  22  91 106 143 367  87 160 108 102  17

educ2 <- recode(educ, "1:4=1; 5:6=2; 8=2; 7=3; 9:10=4")
table(educ2)

## educ2
##   1   2   3   4
## 362 562 160 119

educ3 <- factor(educ2)
levels(educ3) <- c("Menos que Sec. Comp.", "Sec. Comp.", "Tec. Comp.",
"Univ. Comp.")
conf$educ <- educ3

tabla4 <- table(conf$p2r, conf$educ)
prop.table(tabla4, 2)*100

##
##                                     Menos que Sec. Comp. Sec. Comp. Tec. Comp. Univ.
## Comp.
## Muy interesado                  3.418803  6.822262  7.547170 15.9
## 66387
## Algo interesado                17.948718 22.082585 27.044025 32.7
## 73109
## Poco interesado                45.868946 46.678636 45.283019 37.8
## 15126
## Nada interesado                32.763533 24.416517 20.125786 13.4
## 45378

```

GKgamma(tabla4)

```
## gamma      : -0.236
## std. error  : 0.036
## CI          : -0.307 -0.165
```

Función tabla.cont

Funcion para generar tablas de contingencia

Requiere los paquetes "vcd", "vcdExtra" y "data.table"

```
library(vcd)
## Loading required package: grid

library(vcdExtra)
## Loading required package: gnm

library(data.table)
```

Uso:

```
tabla.cont(df, x, y, pc = "s", asoc = FALSE)
```

Produce una tabla de frecuencias cruzadas de dos variables categóricas, excluyendo los NA. Además, calcula los porcentajes verticales u horizontales; los estadísticos de Chi Cuadrado de la tabla; las medidas de asociación nominal basadas en Chi Cuadrado (Phi, Coeficiente de Contingencia y V de Cramer); y el coeficiente Gamma de Goodman y Kruskal.

Los resultados se muestran en un objeto tipo lista que contiene 3 objetos: a) la tabla; b) los coeficientes de Chi Cuadrado; c) el Gamma de G&K

Opciones:

- df : Data frame que contiene las variables (factores) de análisis
- x : Variable (factor) en las columnas, debe estar entre comillas
- y : Variable (factor) en las filas, debe estar entre comillas.
- pc : "s" sin porcentajes; "col" porcentaje en las columnas; "fila" porcentaje en las filas
- asoc : si es TRUE produce los estadísticos de Chi Cuadrado y el Gamma de Goodman y Kruskal

Sintaxis de la función

```

tabla.cont <- function(df, x, y, pc="s", asoc = FALSE){
  library(vcd)
  library(vcdExtra)
  library(data.table)
  as <- c()
  gam <- c()
  misvars <- c(y, x)
  data <- na.omit(df[misvars])
  tab.0 <- table(data)
  asoc.0 <- assocstats(tab.0)
  gam.0 <- GKgamma(tab.0)
  if(pc=="s"){
    tab.1 <- addmargins(tab.0, 1, FUN = list(list(TOTALc = sum)))
    tab.1 <- addmargins(tab.1, 2, FUN = list(list(TOTALf = sum)))
    if(asoc==TRUE){
      as <- asoc.0
      gam <- gam.0}
    return(list(tab.1, as, gam))}
  if (pc=="col"){
    tab.2 <- addmargins(tab.0, 2, FUN = list(list(TOTAL = sum)))
    tab.2 <- round(prop.table(tab.2, 2)*100,2)
    tab.2 <- round(addmargins(tab.2, 1, FUN = list(list(TOTAL = sum))))
  ,1)
    tab.0 <- tab.0
    tab.0 <- addmargins(tab.0, 2, FUN = list(list(TOTAL = sum)))
    tab.0 <- addmargins(tab.0, 1, FUN = list(list(Nvalid = sum)))
    tab.2 <- rbind(tab.2, tail(tab.0, 1))
    if(asoc==TRUE){
      as <- asoc.0
      gam <- gam.0
    }
    return(list(tab.2, as, gam))}
  if (pc=="fila"){
    tab.2 <- addmargins(tab.0, 1, FUN = list(list(TOTAL = sum)))
    tab.2 <- round(prop.table(tab.2, 1)*100,2)
    tab.2 <- round(addmargins(tab.2, 2, FUN = list(list(TOTAL = sum))))
  ,1)
    tab.0 <- tab.0
    tab.0.1 <- addmargins(tab.0, 1, FUN = list(list(TOTAL = sum)))
    tab.0.1 <- addmargins(tab.0.1, 2, FUN = list(list(Nvalid = sum)))
    tab.2 <- cbind(tab.2, Nvalid=tab.0.1[, "Nvalid"])
    if(asoc==TRUE){
      as <- asoc.0
      gam <- gam.0
    }
    return(list(tab.2, as, gam))}
}

```

Ejemplo de uso:

Preparamos los datos

```
library(foreign)
conf <- as.data.frame(read.spss("IOP_1112_01_B.sav"))
conf$p2r <- conf$P2
conf$p2r[conf$P2=="NS/NR"] <- NA
conf$p2r <- factor(conf$p2r)
```

Uso de la función:

a) Tabla de frecuencias absolutas

```
tabla.cont(conf, "NSEGrup", "p2r")
## [[1]]
##                               NSEGrup
##   p2r          A/B     C   D/E TOTALf
##   Muy interesado  29   29   23    81
##   Algo interesado 72   97   99   268
##   Poco interesado 101  153  284   538
##   Nada interesado 53   83   163   299
##   TOTALc          255  362  569  1186
##
## [[2]]
## NULL
##
## [[3]]
## NULL
```

b) Tabla con % calculados en las columnas

```
tabla.cont(conf, "NSEGrup", "p2r", pc="col")
## [[1]]
##          A/B     C   D/E TOTAL
## Muy interesado 11.4  8.0  4.0  6.8
## Algo interesado 28.2 26.8 17.4 22.6
## Poco interesado 39.6 42.3 49.9 45.4
## Nada interesado 20.8 22.9 28.6 25.2
## TOTAL          100.0 100.0 100.0 100.0
## Nvalid         255.0 362.0 569.0 1186.0
##
## [[2]]
## NULL
##
## [[3]]
## NULL
```

c) Tabla con % calculados en las filas

```
tabla.cont(conf, "NSEGrup", "p2r", pc="fila")

## [[1]]
##          A/B     C   D/E TOTAL Nvalid
## Muy interesado 35.8 35.8 28.4   100     81
## Algo interesado 26.9 36.2 36.9   100    268
## Poco interesado 18.8 28.4 52.8   100    538
## Nada interesado 17.7 27.8 54.5   100    299
## TOTAL           21.5 30.5 48.0   100   1186
##
## [[2]]
## NULL
##
## [[3]]
## NULL
```

d) Tabla con % calculados en las columnas y con coeficientes de asociación

```
tabla.cont(conf, "NSEGrup", "p2r", pc="col", asoc=T)

## [[1]]
##          A/B     C   D/E TOTAL
## Muy interesado 11.4   8.0   4.0   6.8
## Algo interesado 28.2  26.8  17.4  22.6
## Poco interesado 39.6  42.3  49.9  45.4
## Nada interesado 20.8  22.9  28.6  25.2
## TOTAL           100.0 100.0 100.0 100.0
## Nvalid          255.0 362.0 569.0 1186.0
##
## [[2]]
##          X^2 df  P(> X^2)
## Likelihood Ratio 38.758  6 7.9833e-07
## Pearson          38.752  6 8.0053e-07
##
##          Phi-Coefficient : 0.181
##          Contingency Coeff.: 0.178
##          Cramer's V       : 0.128
##
## [[3]]
##          gamma      : 0.211
##          std. error  : 0.038
##          CI         : 0.136 0.285
```

Si queremos guardar e imprimir la tabla

Se guarda el resultado en un objeto que, en este caso, llamamos tabla1

```
tabla1 <- tabla.cont(conf, "NSEGrup", "p2r", pc="col", asoc=T)
```

Seleccionamos el primer objeto de esa tabla

```
tabla1 <- tabla1[[1]]
```

Guardamos el resultado usando xtable

```
library(xtable)
print(xtable(tabla1, caption = "Resultado de tabla.cont"), type = "html"
1",
      file = "tabla1.html")
```

Podemos inserter el archivo generado en un document en word, el resultado se vería así (luego se puede editar para que se vea mejor):

	A/B	C	D/E	TOTAL
Muy interesado	11.40	8.00	4.00	6.80
Algo interesado	28.20	26.80	17.40	22.60
Poco interesado	39.60	42.30	49.90	45.40
Nada interesado	20.80	22.90	28.60	25.20
TOTAL	100.00	100.00	100.00	100.00
Nvalid	255.00	362.00	569.00	1186.00

Resultado de tabla.cont

Análisis de la Varianza

Análisis de la varianza

Es una técnica para analizar la relación entre una variable dependiente métrica (cuantitativa, medida en escala de intervalo) y una variable independiente categórica llamada “Factor” (cualitativa, nominal u ordinal)

- Variable dependiente: Métrica (intervalo o razón)
- Variable independiente (Factor): Categórica (nominal u ordinal)

Analiza la relación entre estas variables identificando si existen diferencias significativas entre las medias de la variable dependiente en diferentes niveles del factor independiente.

Ejemplo: Rendimiento académico y nivel educativo de los padres

¿El rendimiento educativo de un estudiante está asociado de alguna forma con el nivel educativo de sus padres?

- Variable dependiente: Rendimiento en pruebas de razonamiento matemático
- Variable independiente (Factor): Nivel educativo del padre del alumno

Los datos provienen de las pruebas de rendimiento de alumnos de 4to de secundaria realizadas el 2001 a descargar de la siguiente dirección:

https://sites.google.com/a/pucp.pe/data_est/archivos/educ4sec.rda?attredirects=0&d=1

El libro de códigos del archivo puede verse en el siguiente archivo:

https://sites.google.com/a/pucp.pe/data_est/archivos/DIC_educ4sec.doc?attredir=0&d=1

```
load("educ4sec.rda")
educ <- educ4sec
```

Estadísticos descriptivos por grupo

Veamos los estadísticos descriptivos de la variable dependiente según el grupo:

```
library(descr)
compmeans(educ$r.mat, educ$educ.pa, plot = FALSE)

## Mean value of "educ$r.mat" according to "educ$educ.pa"
##               Mean   N Std. Dev.
## Sec. inc. o menos 697.7469 284  41.02232
## Secundaria        708.0465 205  47.50560
## Superior         724.3295 301  52.22912
## Total            710.5479 790  48.55295
```

Gráfico de medias

Podemos graficar las medias y sus respectivos intervalos de confianza:

```

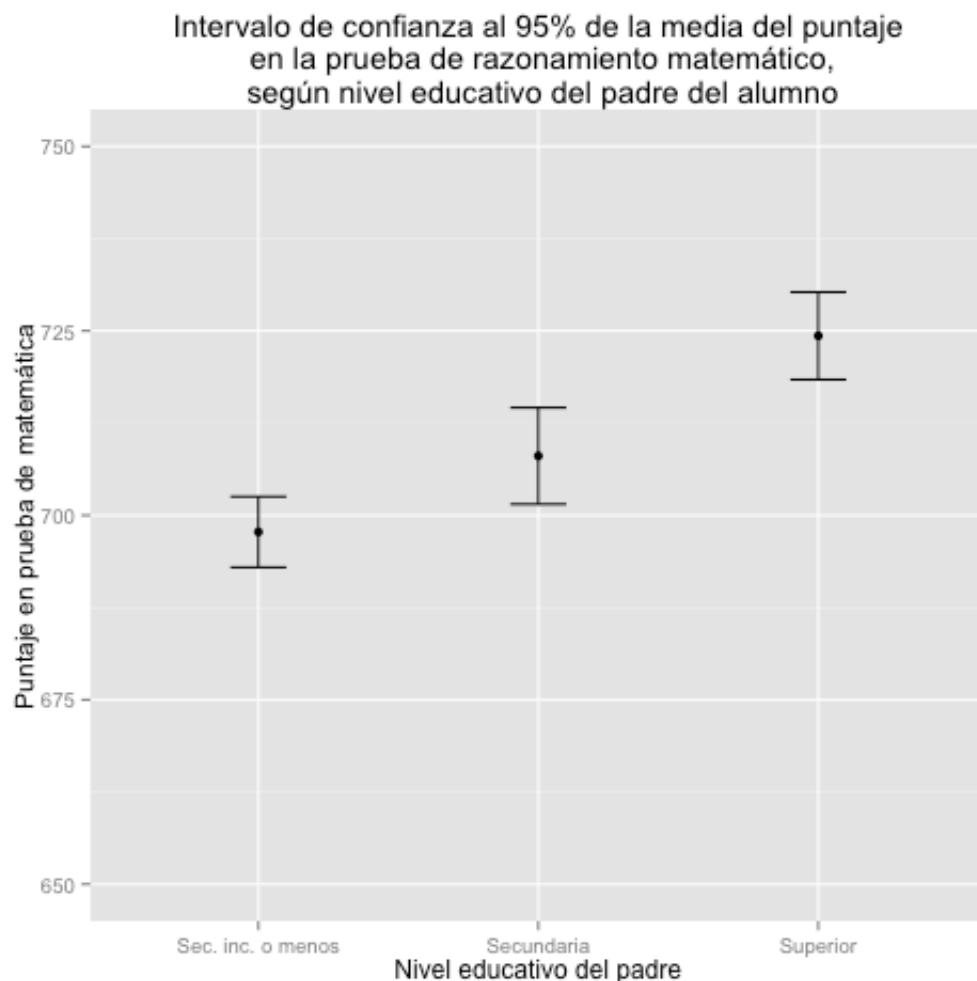
library(Rmisc)
tabla.des <- summarySE(educ, measurevar="r.mat", groupvars="educ.pa",
na.rm=T)
tabla.des

##             educ.pa   N     r.mat      sd      se      ci
## 1 Sec. inc. o menos 284 697.7469 41.02232 2.434227 4.791489
## 2           Secundaria 205 708.0465 47.50560 3.317935 6.541843
## 3          Superior 301 724.3295 52.22912 3.010437 5.924247

library(ggplot2)
graf.m <- ggplot(tabla.des, aes(x=educ.pa, y=r.mat)) + geom_point() +
ylim(650, 750) +
  geom_errorbar(aes(ymin=r.mat-ci, ymax=r.mat+ci), width=0.2) +
  xlab("Nivel educativo del padre") + ylab("Puntaje en prueba de matemática") +
  ggtitle("Intervalo de confianza al 95% de la media del puntaje\nen la prueba de razonamiento matemático,\nsegún nivel educativo del padre del alumno")

graf.m

```



Planteamiento de la prueba de hipótesis:

Paso 1: Formular la hipótesis cero y la hipótesis uno

$$H_0 : \bar{y}_1 = \bar{y}_2 = \bar{y}_3 = \bar{y}_g$$

$$H_1: \bar{y}_1 \neq \bar{y}_2 \neq \bar{y}_3 \neq \bar{y}_g$$

Se compara la media de cada grupo con la **media global (g)**. Las diferencias entre la media de cada grupo y la media global se conoce como **efectos principales o efectos de la prueba (EP)**

$$EP_{y_i} = \bar{y}_i - \bar{y}_g$$

	m.grupo	m.global	Ef.Pr
## Sec. inc. o menos	697.75	710.5	-12.75
## Secundaria	708.05	710.5	-2.45
## Superior	724.33	710.5	13.83

Puntuación de desviación y componentes de la varianza

Tomemos el caso del alumno 185, quien obtuvo un puntaje de 765.47 en la prueba de matemáticas y su padre tenía educación superior. En este caso:

- Su puntuación de desviación respecto de la media global es:

$$D_{y_{185}} = y_{185} - \bar{y}_g$$

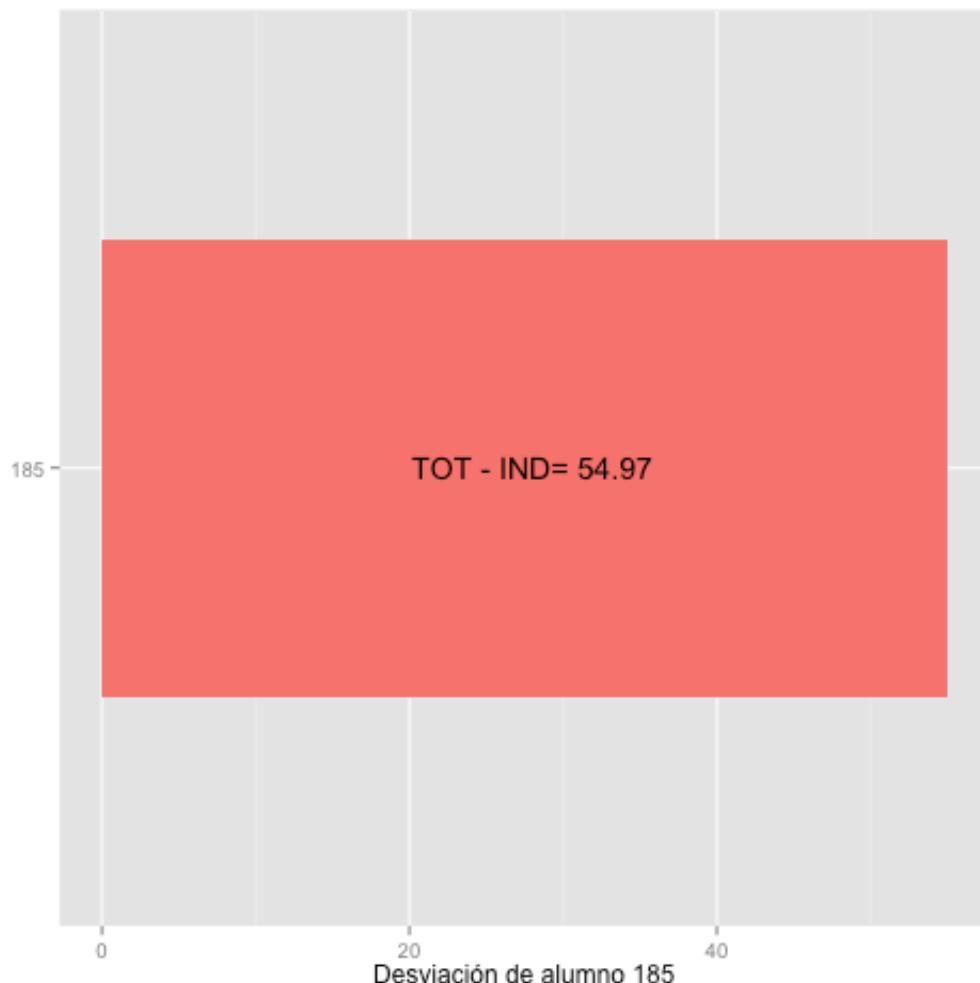
$$D_{y_{185}} = 765.47 - 710.5 = 54.97$$

- Su puntuación de desviación respecto de la media de su grupo es:

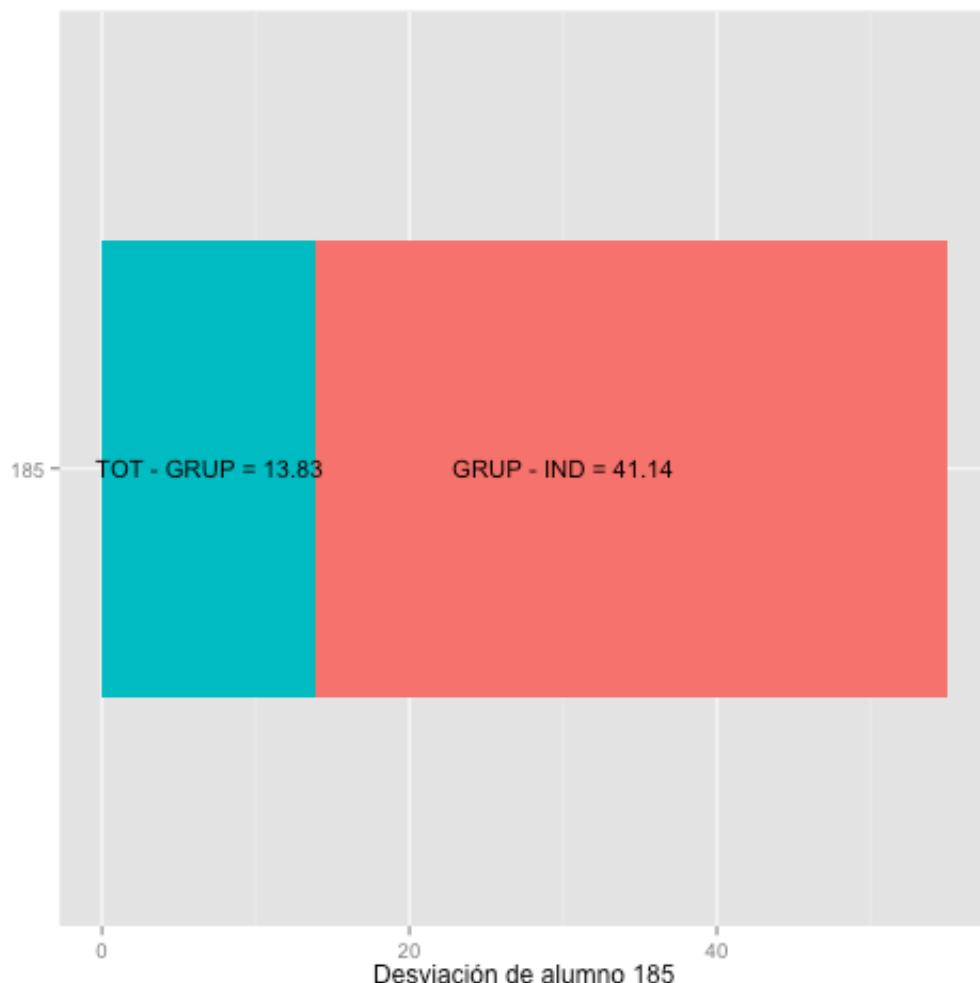
$$D_{y_{185}} = y_{185} - \bar{y}_{sup}$$

$$D_{y_{185}} = 765.47 - 724.33 = 41.14$$

Desviación del alumno 185



Desviación del alumno 185



En el caso del alumno 185 podemos decir que su puntuación es igual a:

$$y_{185} = \bar{y}_g + (y_{185} - \bar{y}_g)$$

$$y_{185} = \bar{y}_g + (\bar{y}_{sup} - \bar{y}_g) + (y_{185} - \bar{y}_{sup})$$

$$765.47 = 710.5 + 13.83 + 41.14$$

El puntaje del alumno 185 es igual a la media global + la desviación explicada entre grupos + la desviación no explicada dentro del grupo

Modelo lineal general o de efectos aditivos

La mejor predicción del puntaje de una variable Y será \bar{Y} más los efectos de una variable independiente X

$$Y = \bar{Y} + bX + e$$

El modelo lineal general descompone cada puntuación de Y en tres partes:

- La cantidad de Y explicada por la media global \bar{Y}
- La cantidad de su desviación explicada por X (el efecto principal)
- La cantidad de su desviación no explicada por X, es decir el error.

ANOVA y Suma de Cuadrados

ANOVA busca determinar cuánto de Y puede ser explicado por X. Hasta qué punto la varianza explicada por X es mayor a la varianza no explicada por X. Para ello descompone la varianza total de Y en dos partes, aplicando lo que se conoce como el método de la Suma de Cuadrados:

- Suma total de cuadrados o varianza total:

$$SS_y = \sum_{i=1}^n (y_i - \bar{y})^2$$

- Suma de cuadrados "explicada" o varianza entre grupos:

$$SS_x = \sum_{j=1}^c n (\bar{y}_j - \bar{y})^2$$

- Suma de cuadrados no explicada o varianza dentro de los grupos:

$$SS_{error} = \sum_{j=1}^c \sum_{i=1}^n (y_{ij} - \bar{y}_j)^2$$

Por tanto:

$$SS_y = SS_x + SS_{error}$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{j=1}^c n (\bar{y}_j - \bar{y})^2 + \sum_{j=1}^c \sum_{i=1}^n (y_{ij} - \bar{y}_j)^2$$

Cuadrados medios y estadística de F

- Cuadrado medio entre grupos:

$$CM_E = \frac{\sum_{j=1}^c n (\bar{y}_j - \bar{y})^2}{K - 1}$$

Donde k = al número de grupos; y k-1 son los grados de libertad de la varianza entre grupos

- Cuadrado medio dentro de los grupos:

$$CM_D = \frac{\sum_{j=1}^c \sum_{i=1}^n (y_{ij} - \bar{y}_j)^2}{n - K}$$

donde n = número total de casos; y n-k son los grados de libertad de la varianza dentro de los grupos.

- Estadística de F: razón entre la varianza explicada y la varianza no explicada

$$F = \frac{CM_E}{CM_D} = \frac{\sum_{j=1}^c n (\bar{y}_j - \bar{y})^2 / (K - 1)}{\sum_{j=1}^c \sum_{i=1}^n (y_{ij} - \bar{y}_j)^2 / (n - 1)}$$

Pasos de ANOVA

Paso 2: Seleccionar la distribución de muestreo de la hipótesis cero

Se utiliza la distribución de F.

Paso 3: Seleccionar un nivel de significancia

Puede ser $\alpha = 0.05$ ó $\alpha = 0.01$, por ejemplo

Paso 4: Se calcula el estadístico de la prueba

Es este caso será un valor de F, que deberá compararse con un valor crítico de F dados los grados de libertad. F tiene dos tipos de grados de libertad, los del numerador ($k-1$) y los del denominador ($n-k$):

$$F = \frac{CM_E}{CM_D} = \frac{\sum_{j=1}^c n (\bar{y}_j - \bar{y})^2 / (K - 1)}{\sum_{j=1}^c \sum_{i=1}^n (y_{ij} - \bar{y}_j)^2 / (n - 1)}$$

Paso 5: Se decide si se acepta o rechaza H0

Se acepta H0 cuando:

- el estadístico de F es menor que el valor crítico de F ó
- cuando el p-value o significancia del estadístico de la prueba es mayor que el nivel de significancia de la prueba.

Caso contrario, se rechaza H0

ANOVA con R

Se utiliza la función "summary(aov(x~f))" para generar una tabla de ANOVA que muestra los diferentes componentes de la varianza de Y

```
summary(aov(educ$r.mat~educ$educ.pa))

##               Df  Sum Sq Mean Sq F value    Pr(>F)
## educ$educ.pa   2 104990   52495   23.54 1.18e-10 ***
## Residuals     787 1754989     2230
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En este caso los resultados de la prueba nos lleva a rechazar H0 y afirmar que sí existen diferencias estadísticamente significativas en los promedios de las pruebas

de matemáticas entre los grupos analizados. El efecto principal de X en Y es estadísticamente significativo.

Nótese que la varianza total es:

$$S^2 = (SS_x + SS_{error})/(n - 1) = (104990 + 1754989)/787 = 2363.38$$

Prueba de diferencias entre grupos específicos

ANOVA sólo nos dice que el efecto principal es estadísticamente significativo (las medias de los grupos no son iguales). Pero no nos dice entre qué grupos se encuentran las diferencias.

Para identificar las diferencias específicas entre grupos se utiliza el método de las diferencias altamente significativas inventado por John Tukey (TukeyHSD):

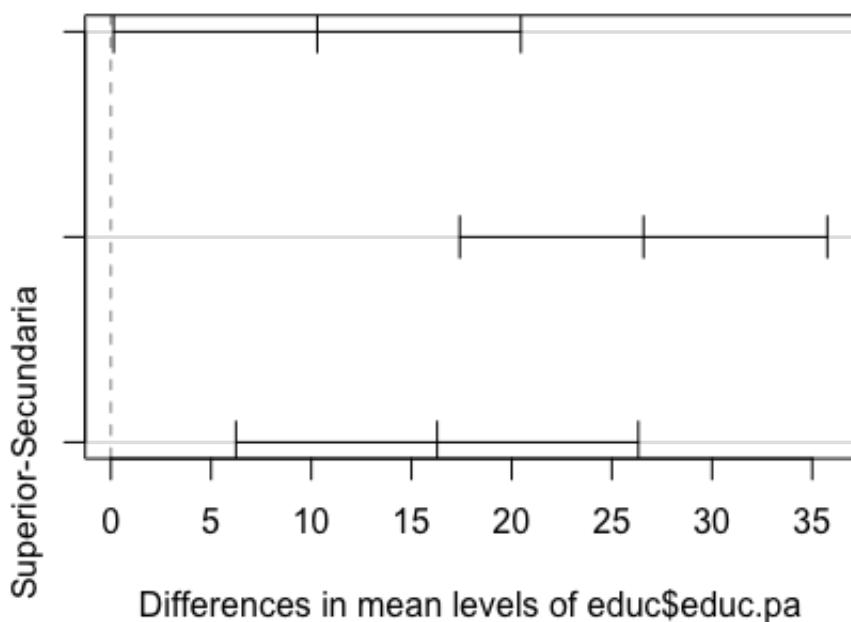
```
m <- aov(educ$r.mat~educ$educ.pa)
TukeyHSD(m)

##    Tukey multiple comparisons of means
##      95% family-wise confidence level
##
## Fit: aov(formula = educ$r.mat ~ educ$educ.pa)
##
## $`educ$educ.pa`
##                               diff      lwr      upr      p adj
## Secundaria-Sec. inc. o menos 10.29964  0.1372835 20.46199 0.0461402
## Superior-Sec. inc. o menos  26.58263 17.4096332 35.75563 0.0000000
## Superior-Secundaria        16.28300  6.2416745 26.32432 0.0004435
```

Graficar las diferencias significativas

```
plot(TukeyHSD(m))
```

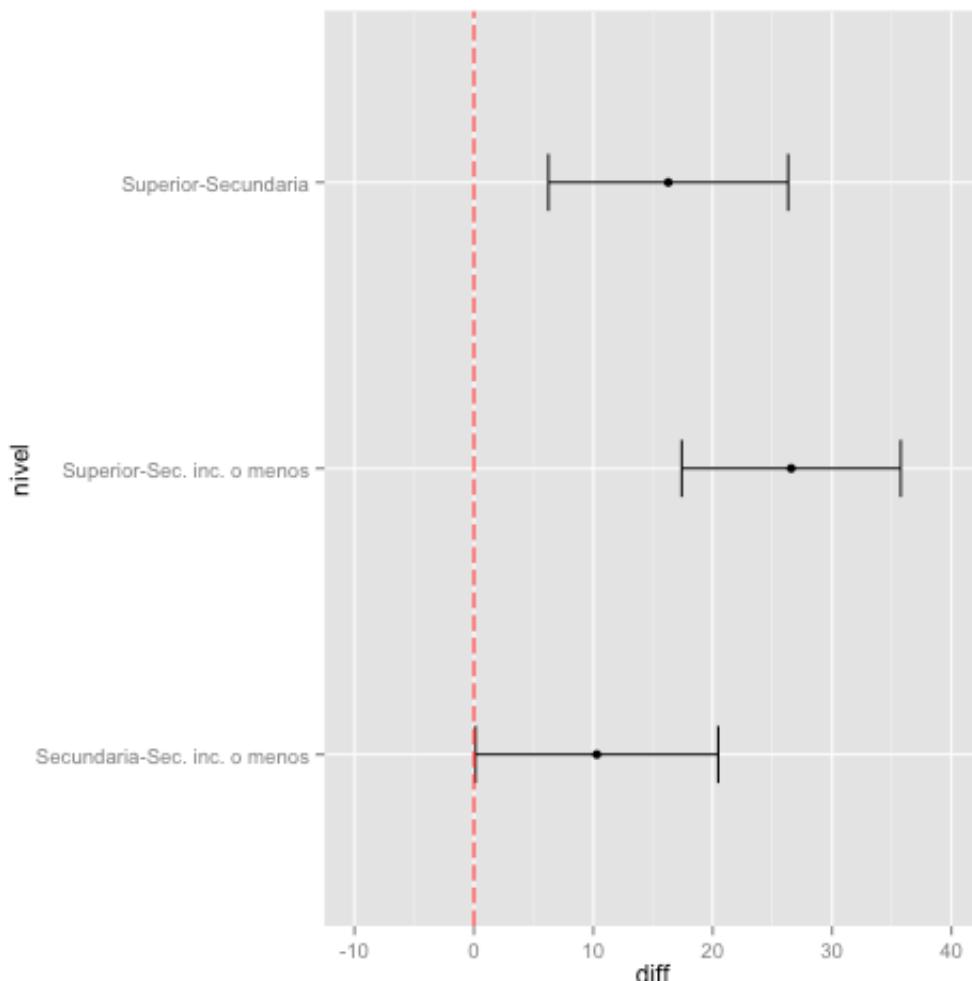
95% family-wise confidence level



Un mejor gráfico en ggplot

```
p <- TukeyHSD(m)
p1 <- p[[1]]
p1 <- as.data.frame(p1)
p1$nivel <- row.names(p1)
graf.dif <- ggplot(p1, aes(x=nivel, y=diff)) + geom_point() +
  geom_errorbar(aes(ymin=lwr, ymax=upr), width=0.2) +
  ylim(-10, 40) + geom_hline(yintercept=0, col="red", linetype = "long
  dash") + coord_flip()

graf.dif
```

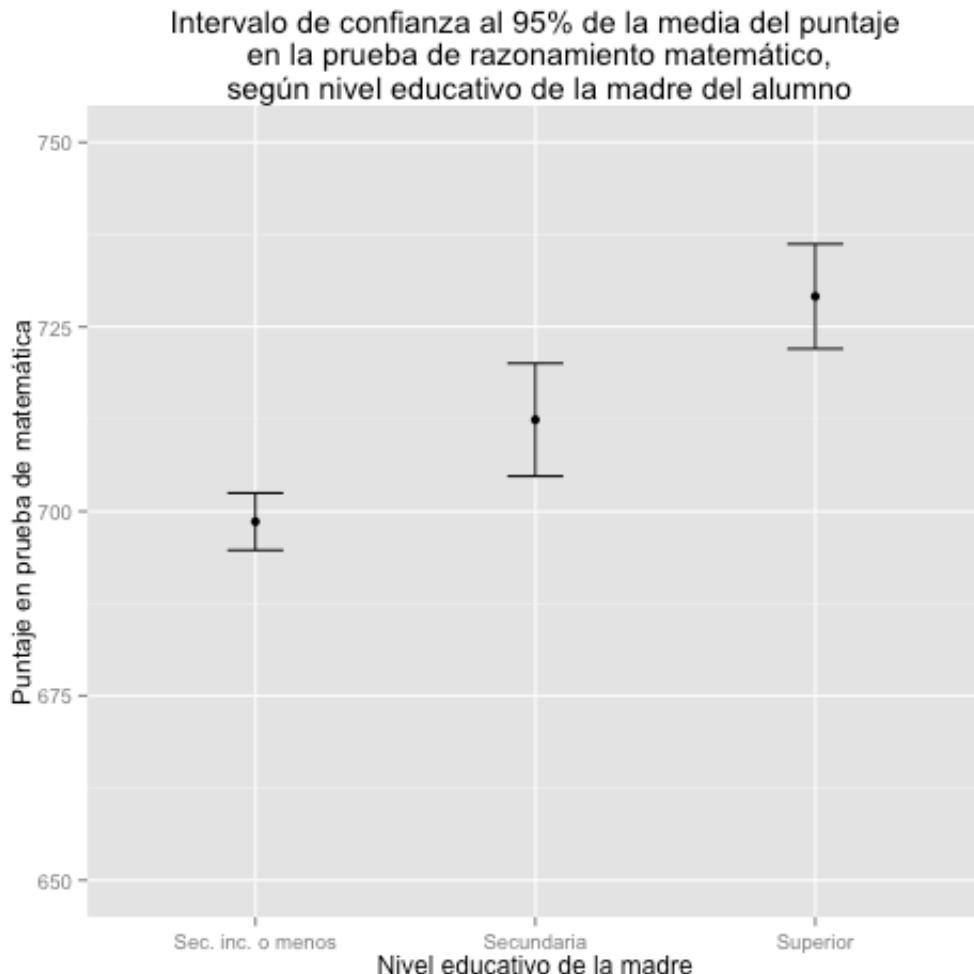


Otro ejemplo: Rendimiento según nivel educativo de la madre

```
compmeans(educ$r.mat, educ$educ.mam, plot = FALSE)

## Mean value of "educ$r.mat" according to "educ$educ.mam"
##               Mean   N Std. Dev.
## Sec. inc. o menos 698.6035 379  38.35287
## Secundaria        712.4066 186  52.87332
## Superior          729.1312 225  54.04877
## Total             710.5479 790  48.55295
```

Gráfico de medias



Prueba de ANOVA

```
m2 <- aov(educ$r.mat~educ$educ.mam)
summary(m2)

##                   Df  Sum Sq Mean Sq F value    Pr(>F)
## educ$educ.mam   2  132415   66208   30.16 2.39e-13 ***
## Residuals      787 1727565     2195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

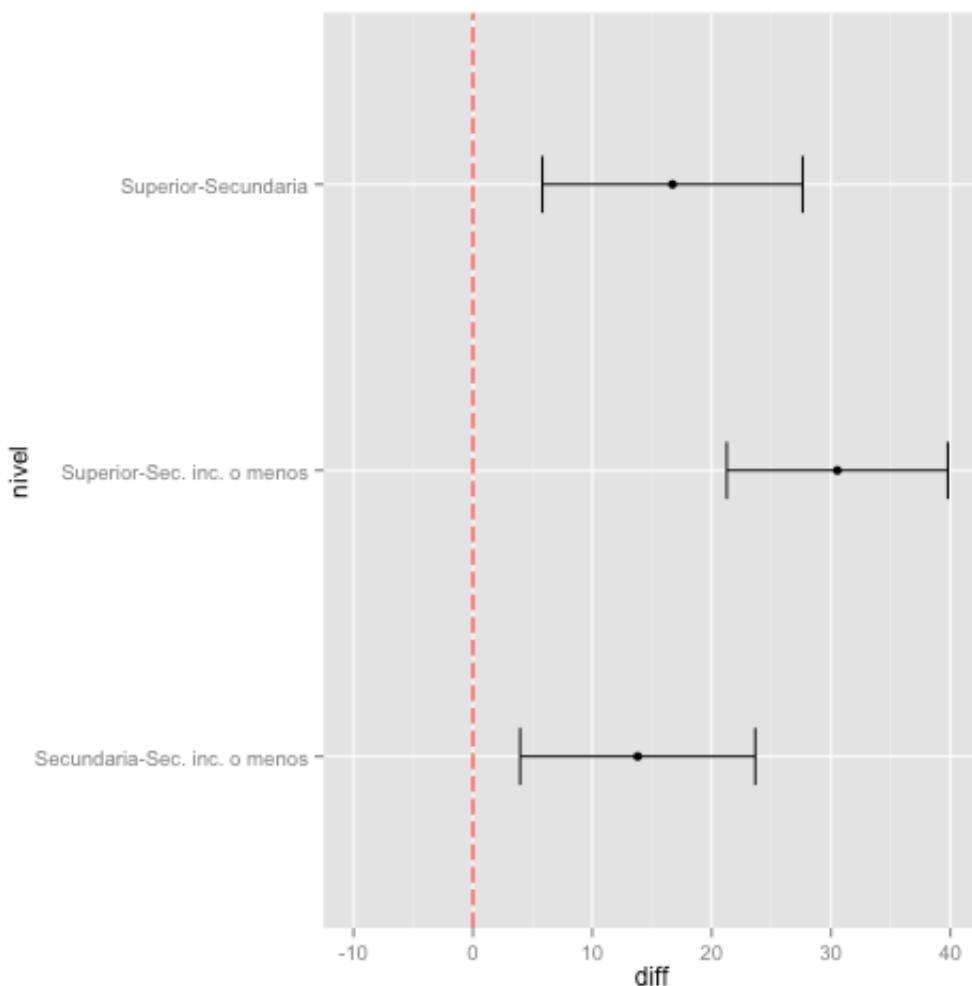
Prueba de HSD de Tukey

```
TukeyHSD(m2)

##    Tukey multiple comparisons of means
##      95% family-wise confidence level
##
## Fit: aov(formula = educ$r.mat ~ educ$educ.mam)
```

```
## 
## $`educ$educ.mam` 
##          diff      lwr      upr      p adj
## Secundaria-Sec. inc. o menos 13.80308  3.953808 23.65234 0.0029944
## Superior-Sec. inc. o menos   30.52767 21.268687 39.78666 0.0000000
## Superior-Secundaria        16.72460  5.822024 27.62717 0.0009773
```

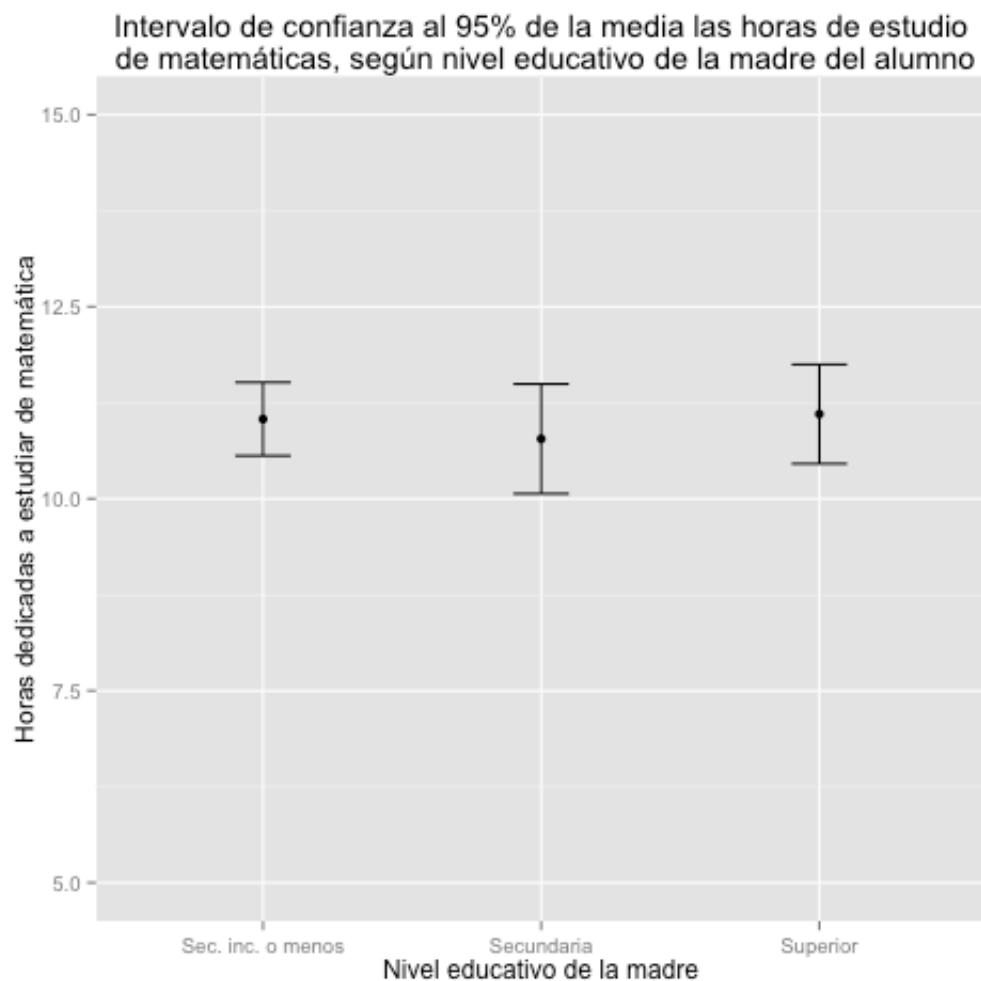
Gráfico de diferencias significativas de Tukey



Ejemplo: Horas dedicadas a las tareas según nivel educativo de la madre

```
compmeans(educ$h.tareas.mat, educ$educ.mam, plot = FALSE)

## Mean value of "educ$h.tareas.mat" according to "educ$educ.mam"
##               Mean   N Std. Dev.
## Sec. inc. o menos 11.03694 379  4.727350
## Secundaria       10.77957 186  4.938521
## Superior         11.10222 225  4.922003
## Total            10.99494 790  4.828882
```



```
summary(aov(educ$h.tareas.mat~educ$educ.mam))

##           Df Sum Sq Mean Sq F value Pr(>F)
## educ$educ.mam    2     12   5.943   0.254  0.775
## Residuals      787  18386  23.362
```

Correlación y Regresión Simple

En análisis de correlación y de regresión simple son técnicas para identificar y modelar o representar la relación que existe entre dos variables **cuantitativas** medidas en escala de intervalo o razón.

Se dice que dos variables están **correlacionadas** cuando se aprecia un cambio sistemático en sus respectivas puntuaciones.

El análisis de regresión simple busca identificar y representar una **relación lineal** entre dos variables de intervalo o razón. Una relación lineal puede representarse mediante una **ecuación lineal** o una **recta de regresión**:

$$\hat{Y} = b_0 + b_1 X$$

Esta ecuación nos dice que el **valor esperado de Y** (\hat{Y}) será igual a una constante b_0 más el puntaje de X multiplicado por un coeficiente b_1 . En tal sentido, por cada cambio de $X + 1$ se espera que \hat{Y} cambie en b_1 . Por otro lado, si $X = 0$, entonces $\hat{Y} = b_0$

Datos

Indicadores sociodemográficos para 207 países del mundo en 1998. Los datos se encuentran en el archivo "BD_Mundo.zip" que puede descargarse desde:

http://sites.google.com/a/pucp.pe/data_est/archivos

```
load("mundo98.rda")
## Cargamos además algunos paquetes que usaremos en esta clase:
library(ggplot2)
library(grid)
library(scales)
```

Vamos a trabajar con dos variables:

- **educationFemale**: Promedio de años de educación formal de las mujeres
- **contraception**: Porcentaje de mujeres que utilizan métodos anticonceptivos

```
myvars <- c("educationFemale", "contraception")
data <- na.omit(mundo98[myvars])
```

Pasos en el análisis de regresión

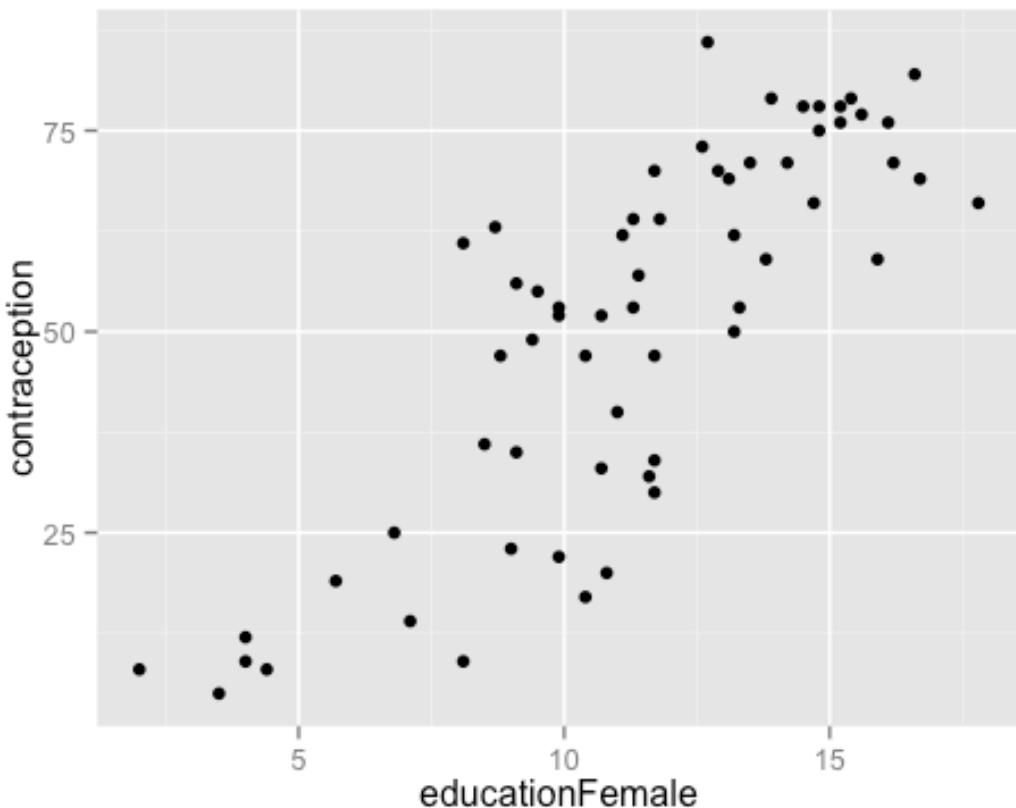
En el análisis de regresión usualmente debemos seguir los siguientes pasos:

1. Inspeccionar visualmente la relación entre las variables mediante un diagrama de dispersión.
2. Estimar la ecuación o recta de regresión.
3. Interpretar los coeficientes de regresión.

4. Calcular los coeficientes de correlación y de determinación
5. Evaluar la significancia de los coeficientes de regresión y correlación

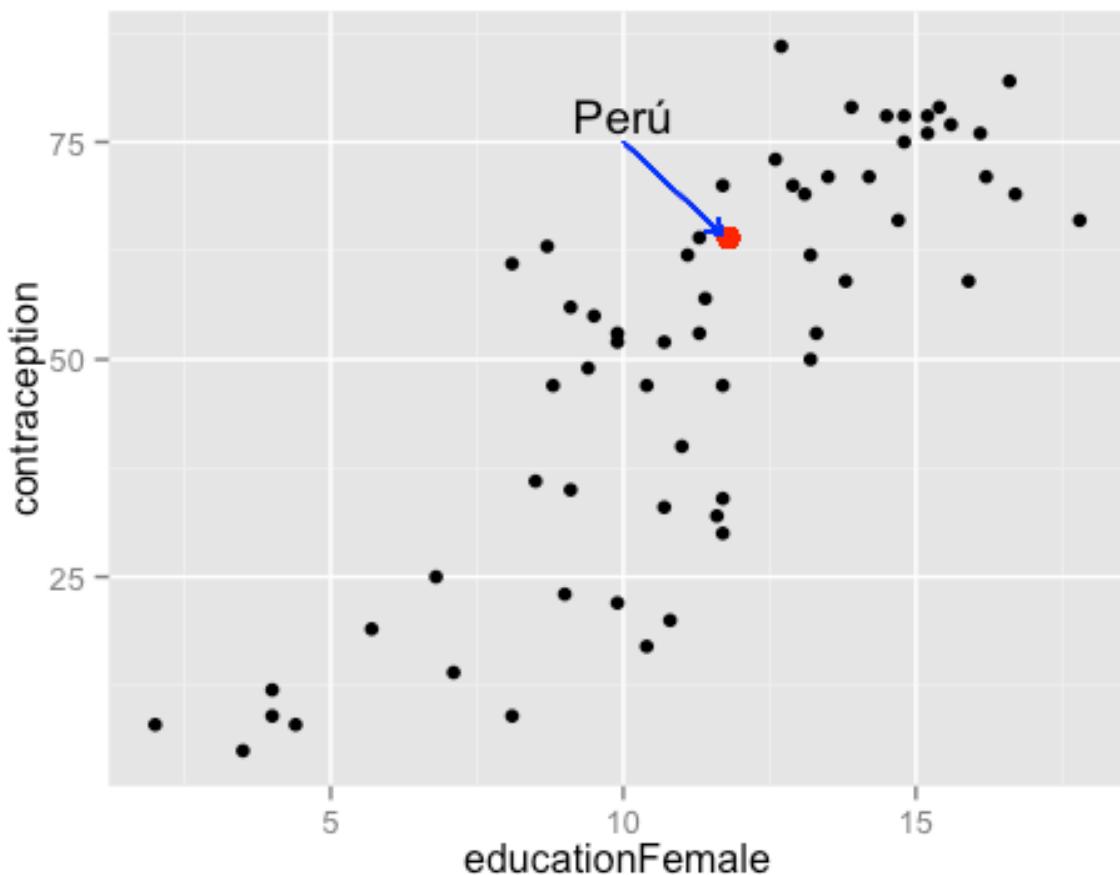
Paso 1: Diagrama de dispersión

```
p <- ggplot(data, aes(x=educationFemale, y=contraception)) + geom_point()  
p
```



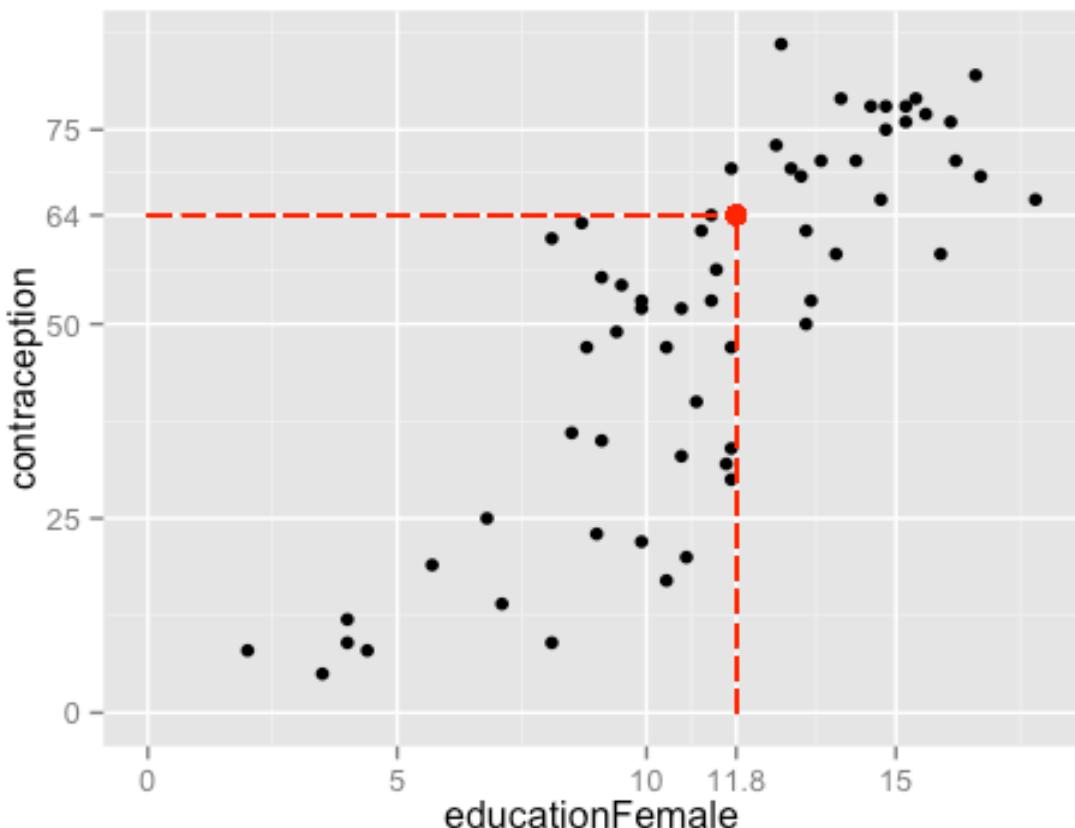


Veamos dónde está Perú en este diagrama:





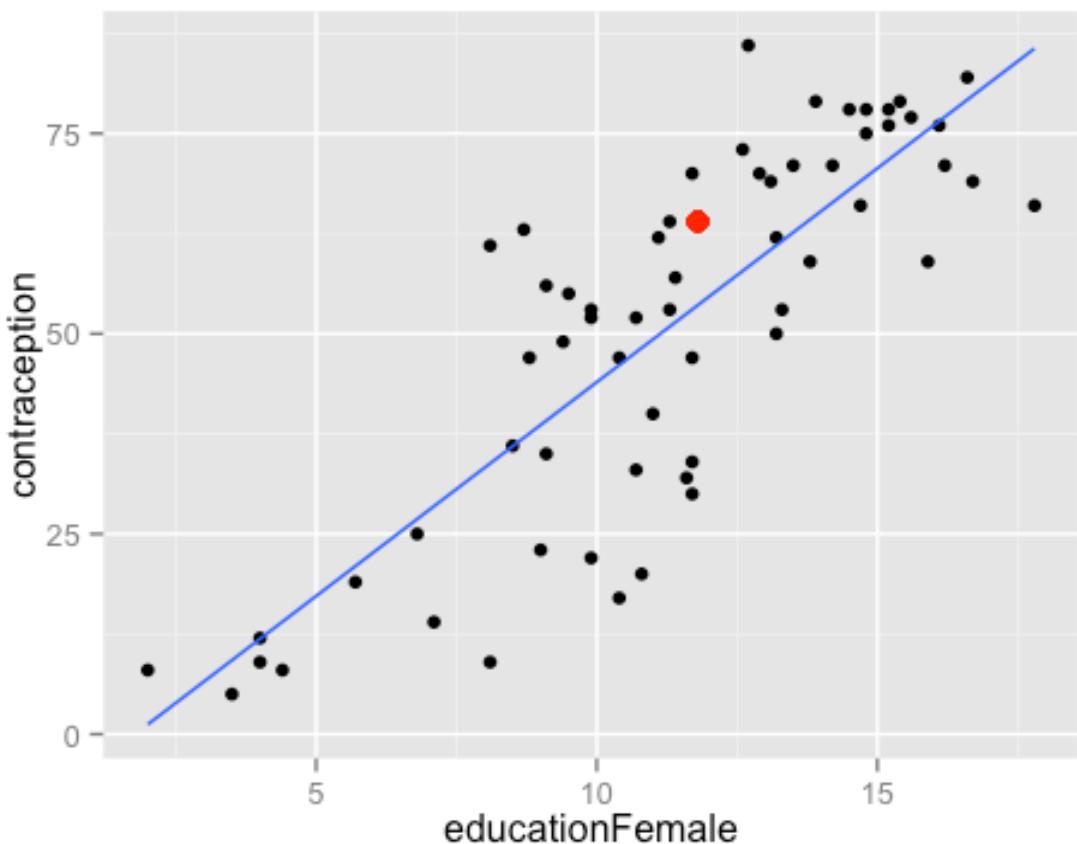
Perú



Paso 2: Recta de regresión

La recta o ecuación de regresión es una línea recta que pasa en medio de los puntos del diagrama de dispersión y que representa la relación entre ambas variables.

Para calcularla se emplea el criterio de los mínimos cuadrados



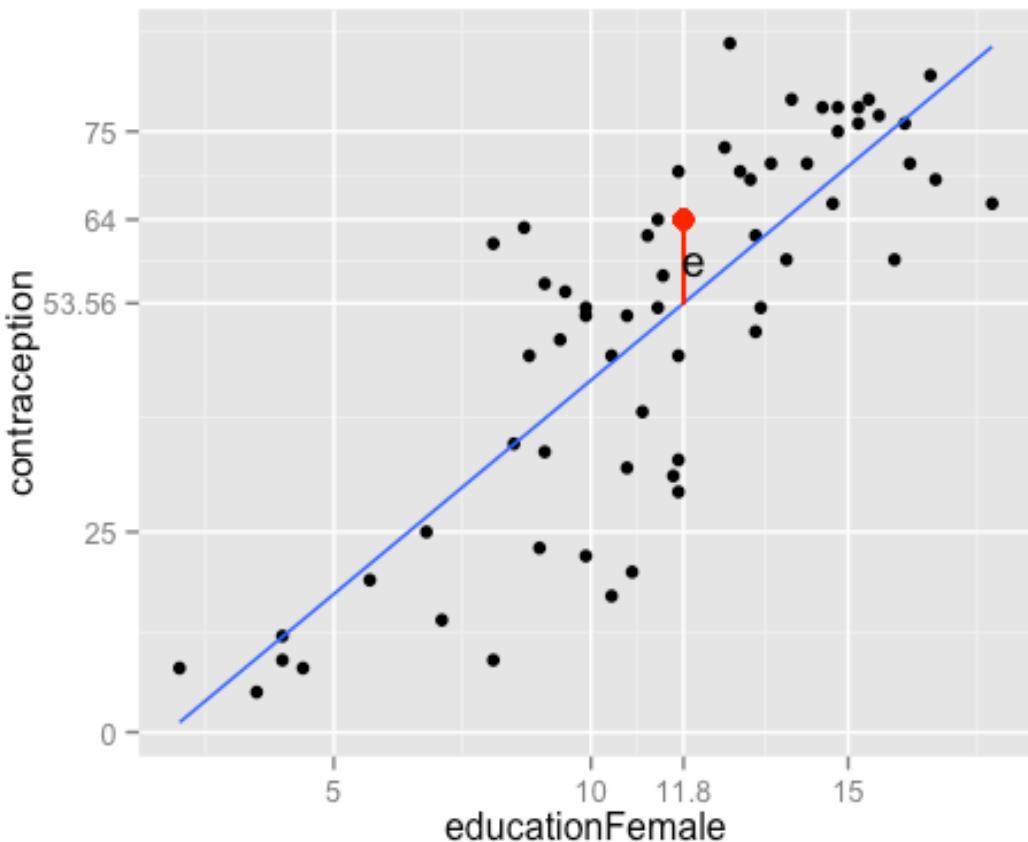
Criterio de los mínimos cuadrados

El método de los mínimos cuadrados implica determinar los valores de los coeficientes de regresión b_0 y b_1 que minimizan la suma de cuadrados de las desviaciones entre los valores observados de Y y sus respectivos valores estimados \hat{Y}

$$\sum e^2 = \sum(Y_i - \hat{Y}_i)^2$$

Error para un caso:

$$e_{Peru}^2 = (Y_i - \hat{Y}_i)^2 = (64 - 53.56)^2 = (10.44)^2 = 108.99$$



Cálculo de los coeficientes de regresión

La recta que minimiza los errores cuadrados se puede calcular con las siguientes ecuaciones:

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

En el R con la siguiente sintaxis se puede obtener los coeficientes de regresión b_0 y b_1

```
lm(contraception~educationFemale, data=data)

##
## Call:
## lm(formula = contraception ~ educationFemale, data = data)
##
## Coefficients:
## (Intercept) educationFemale
## -9.461         5.341
```

Por lo tanto: $\hat{Y} = -9.461 + 5.341(X)$

Paso 3: Interpretar los coeficientes de regresión

De acuerdo con los resultados del cálculo de los coeficientes, nuestro modelo o recta de regresión simple nos dice que:

$$\hat{Y} = -9.461 + 5.341(X)$$

Eso significa que:

- Por cada incremento de 1 año en los años promedio de estudio de las mujeres en un país, se espera que el % de mujeres que usan anticonceptivos se incremente en 5.341%.
- Si el promedio de años de estudio de las mujeres en un país fuese = 0, se esperaría un % de uso de anticonceptivos de -9.461% (es un valor teórico)

Prueba para un caso: En el Perú, en 1998 en promedio las mujeres estudiaban 11.8 años. De acuerdo con el modelo calculado, dado ese valor de X se esperaría que el porcentaje de mujeres que usan anticonceptivos en el Perú fuese de 53.56%:

$$\hat{Y} = -9.461 + (5.341)(11.8) = 53.56$$

Paso 4a: El Coeficiente de correlación "r de Pearson"

El coeficiente de correlación bivariante **r de Pearson** mide la fuerza y dirección de la relación entre dos variables cuantitativas en una escala que varía entre -1 y +1. Cuanto más alejado del 0 sea el valor del coeficiente, más fuerte será la relación. El signo nos indica si se trata de una relación directa o inversa.

El coeficiente r de Pearson se calcula utilizando la siguiente fórmula

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

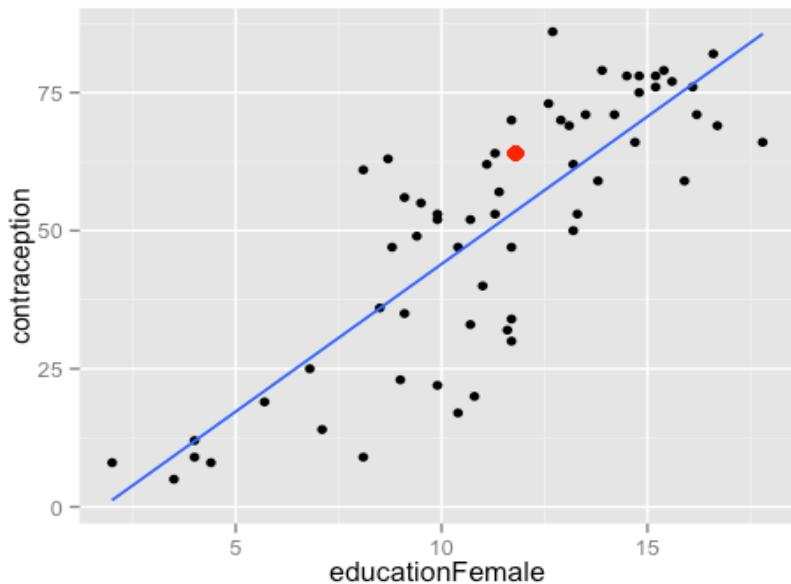
En el R, el coeficiente de correlación se pide usando el siguiente comando:

```
cor(data$contraception, data$educationFemale)
## [1] 0.8192834
```

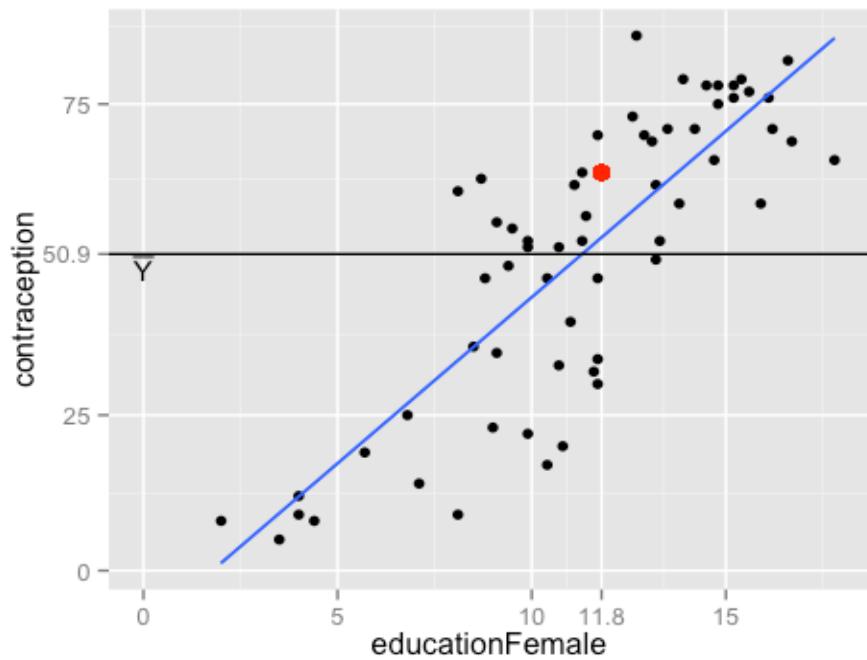
Paso 4b: El coeficiente de Determinación o R^2

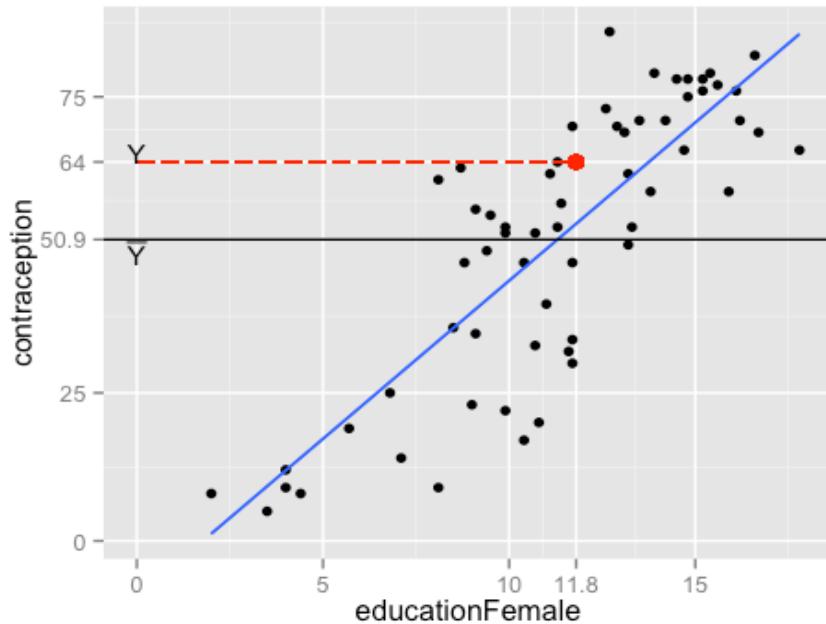
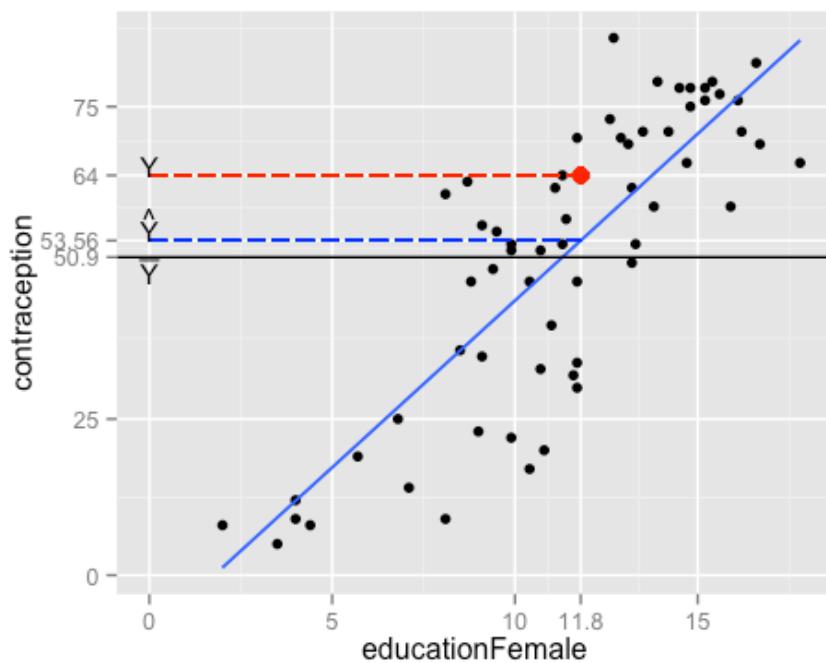
- El R^2 mide la "bondad de ajuste" del modelo de regresión a los datos analizados.
- Nos indica qué tan bien la recta de regresión es capaz de predecir los valores de Y
- R^2 varía en una escala de 0 a 1, cuanto mayor es su valor, mayor poder predictivo tiene el modelo de regresión.
- Puede interpretarse como la proporción de la varianza total de Y que es "explicada" por el modelo de regresión (similar a ANOVA)

Lógica de R^2 : Primero el modelo de regresión...

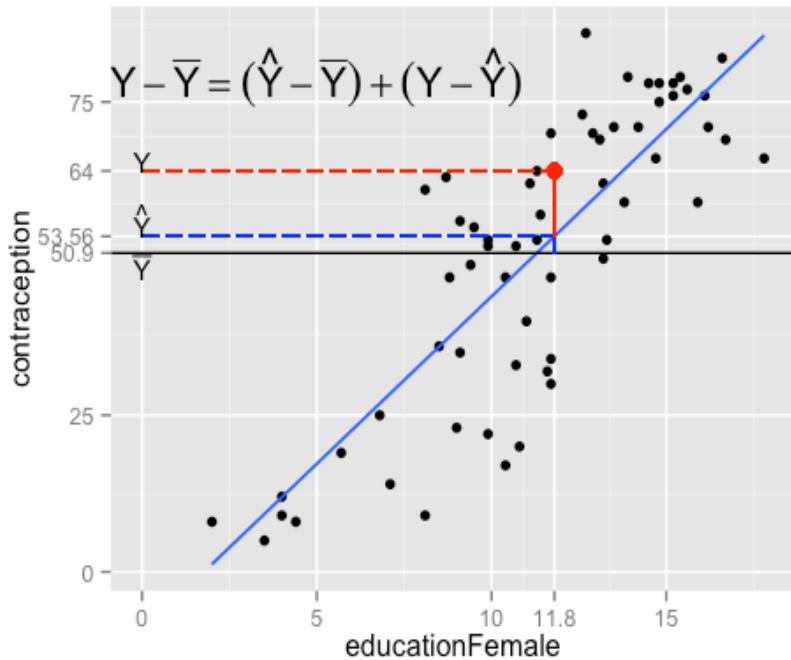


Añadimos la línea que representa \bar{Y}

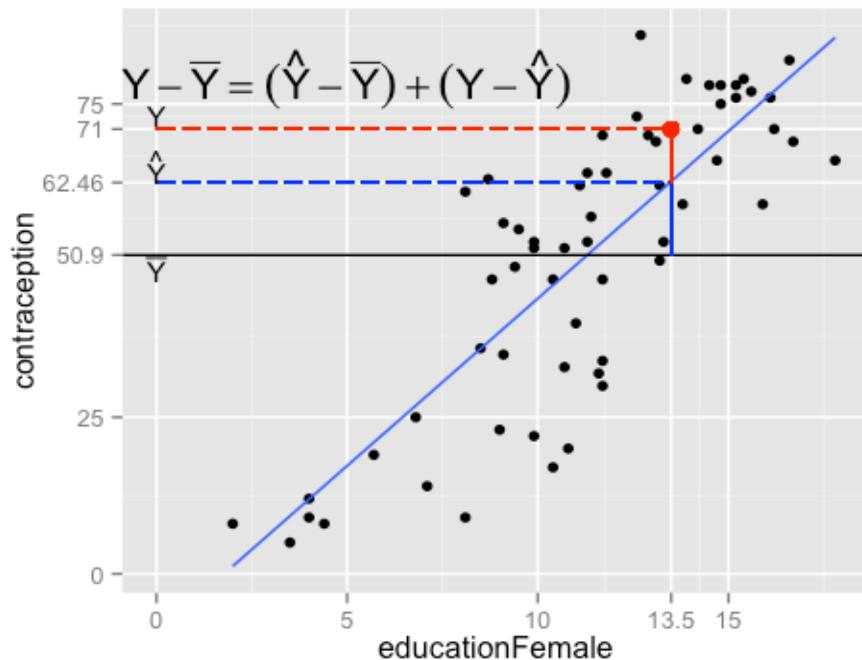


Mostramos el valor de Y (para Perú)

Mostramos el valor esperado de Y para Perú (\hat{Y})


Descomponiendo la variación total de Y (Perú)



Otro país: Suiza



Descomponiendo la variación total:

En el caso de Perú

$$Y - \bar{Y} = (\hat{Y} - \bar{Y}) + (Y - \hat{Y})$$

$$64 - 50.9 = (53.56 - 50.9) + (64 - 53.56)$$

$$13.1 = 2.66 + 10.44$$

En el caso de Zuiza

$$71 - 50.9 = (62.46 - 50.9) + (71 - 62.46)$$

$$20.1 = 11.56 + 8.54$$

Si aplicamos la misma lógica a **todos los casos** tenemos:

$$\sum(Y - \bar{Y})^2 = \sum(\hat{Y} - \bar{Y})^2 + \sum(Y - \hat{Y})^2$$

La suma total de cuadrados SS_y es igual a la suma de cuadrados de la regresión SS_x más la suma de los errores cuadrados SS_e .

Cálculo de R^2

R^2 se calcula:

$$R^2 = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2}$$

Por lo tanto representa la proporción de la varianza total de Y que es "captada" o explicada por el modelo de regresión (X).

El R^2 forma parte del conjunto de estadísticos de resumen de un modelo de regresión.

Resumen del modelo de regresión

```
modelo1 <- lm(contraception~educationFemale, data=data)
summary(modelo1)

##
## Call:
## lm(formula = contraception ~ educationFemale, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -29.088  -7.949   1.531   8.560  27.628 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -9.4606     5.7131  -1.656   0.103    
## educationFemale 5.3412     0.4826  11.068 3.99e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Residual standard error: 13.39 on 60 degrees of freedom
## Multiple R-squared:  0.6712, Adjusted R-squared:  0.6657
## F-statistic: 122.5 on 1 and 60 DF,  p-value: 3.992e-16
```

Paso 5: Evaluar la significancia de los coeficientes de regresión

Para cada coeficiente de regresión se puede calcular un error estándar del coeficiente.

$$\sigma_{b_0} = \sqrt{\frac{\sigma_e^2 \sum x^2}{n \sum (x - \bar{x})^2}}$$

$$\sigma_{b_1} = \sqrt{\frac{\sigma_e^2}{\sum (x - \bar{x})^2}}$$

Donde:

$$\sigma_e^2 = \frac{\sum e^2}{(n - 1 - k)} = \frac{\sum (y - \hat{y})^2}{(n - 1 - k)}$$

Siendo $(n - 1 - k)$ los grados de libertad del error o residuales, donde k es el número de variables independientes del modelo.

Intervalo de confianza del modelo de regresión

Con los errores estándar se puede calcular el intervalo de confianza de los coeficientes del modelo de regresión:

$$\beta_{1-\alpha} = b \pm t_{\alpha/2} \sigma_b$$

En el modelo calculado tenemos que:

$$\sigma_{b_0} = 5.71$$

$$\sigma_{b_1} = 0.48$$

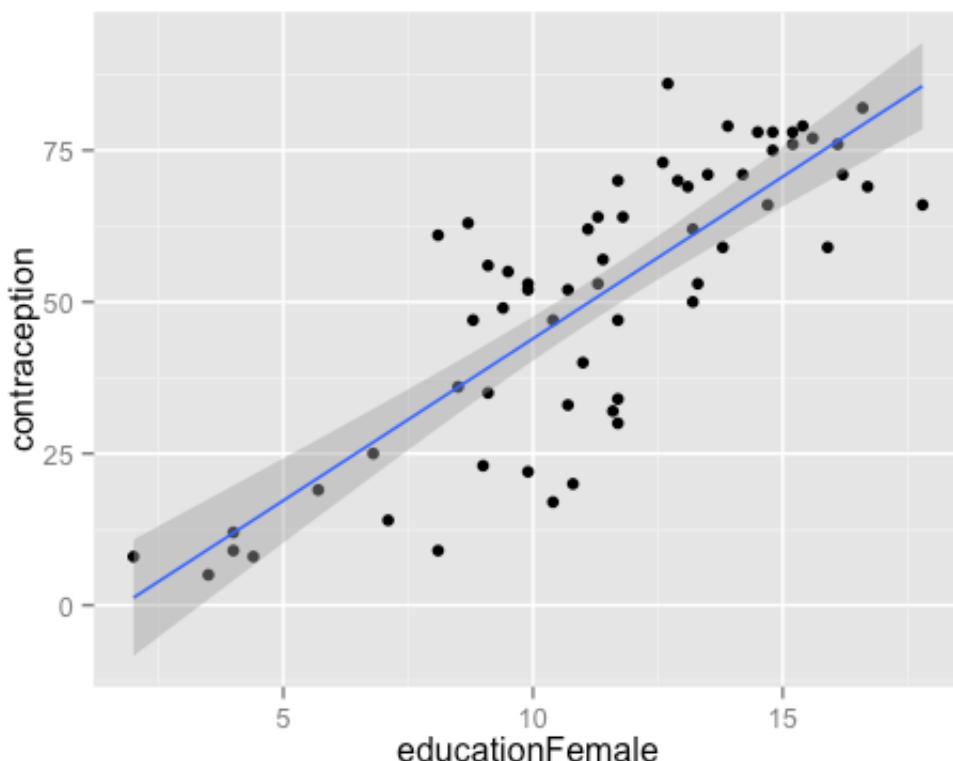
Se puede usar para ello la siguiente función para un intervalo de confianza del 95%

```
confint(modelo1)
```

```
##                   2.5 %   97.5 %
## (Intercept) -20.888497 1.967228
## educationFemale 4.375853 6.306491
```

Graficamos el intervalo de confianza del modelo:

```
ggplot(data, aes(x=educationFemale, y=contraception)) + geom_point() +
    geom_smooth(method=lm)
```



Prueba de hipótesis para la significancia de los coeficientes

Con el error estándar de los coeficientes se puede calcular una prueba de hipótesis (prueba de t) para determinar si el coeficiente de regresión es estadísticamente significativo. En este caso las hipótesis se formulan:

$$H_0: b = 0$$

$$H_1: b \neq 0$$

El estadístico de la prueba es un valor de t que se calcula:

$$t = \frac{b}{\sigma_b}$$

En nuestro ejemplo, para:

- b_0 tenemos que $t = -1.656$ con una significancia ó p-value de 0.103
- b_1 tenemos que $t = 11.068$ con una significancia ó p-value de 3.99e-16

Podemos decidir si rechazamos H_0 comparando el nivel de significancia con el α que hayamos fijado para la prueba.

Significancia de R^2

Para evaluar la significancia del coeficiente de determinación utilizamos la tabla de ANOVA del modelo:

```
anova(modelo1)

## Analysis of Variance Table
##
## Response: contraception
##             Df Sum Sq Mean Sq F value    Pr(>F)
## educationFemale  1  21972  21971.5   122.5 3.992e-16 ***
## Residuals      60  10762    179.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Error estándar del modelo

El error estándar del modelo (o error residual estándar) es el error típico que se cometería al tratar de predecir el valor de Y usando la ecuación de regresión:

$$\sigma = \sqrt{\frac{\sum(y - \hat{y})^2}{(n - 1 - k)}}$$

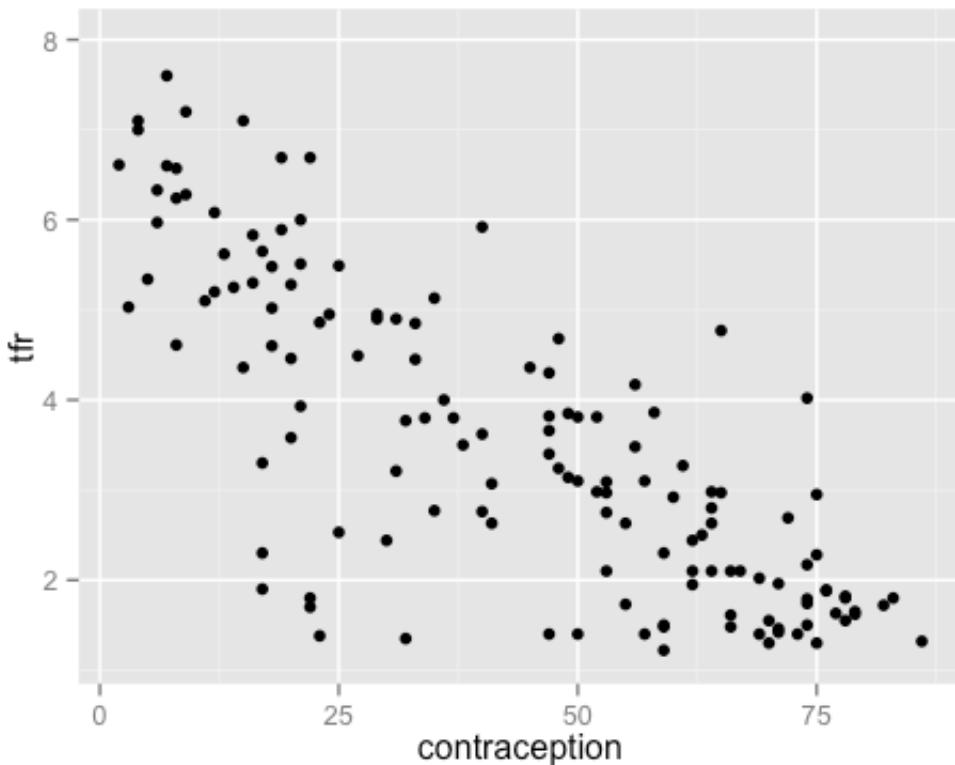
En nuestro ejemplo, el error estándar del modelo es 13.39. Si queremos predecir el % de mujeres que usan anticonceptivos a partir de una regresión sobre la base de los años de educación promedio que tienen las mujeres en un país, el error típico sería de $\pm 13.39\%$.

```
##
## Call:
## lm(formula = contraception ~ educationFemale, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.088 -7.949  1.531  8.560 27.628
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.4606    5.7131 -1.656   0.103
## educationFemale 5.3412    0.4826 11.068 3.99e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.39 on 60 degrees of freedom
## Multiple R-squared:  0.6712, Adjusted R-squared:  0.6657
## F-statistic: 122.5 on 1 and 60 DF,  p-value: 3.992e-16
```

Otro ejemplo:

Paso 1: Diagrama de dispersión de la Tasa de fecundidad según uso de anticonceptivos

```
ggplot(mundo98, aes(x=contraception, y=tfr)) + geom_point()
```



Paso 2 y 3: Los coeficientes de regresión

```

modelo2 <- lm(tfr~contraception, data=mando98)
summary(modelo2)

##
## Call:
## lm(formula = tfr ~ contraception, data = mundo98)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.3000 -0.4562  0.0617  0.6620  2.4620
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.978989  0.186531  32.05  <2e-16 ***
## contraception -0.056477  0.003778 -14.95  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.069 on 139 degrees of freedom
## (66 observations deleted due to missingness)
## Multiple R-squared:  0.6165, Adjusted R-squared:  0.6138
## F-statistic: 223.5 on 1 and 139 DF,  p-value: < 2.2e-16

```

Paso 4: Cálculo del coeficiente de correlación

```
misvars <- c("tfr", "contraception")
data2 <- na.omit(mundo98[misvars])
cor(data2)

##                      tfr  contraception
## tfr                 1.000000     -0.785205
## contraception      -0.785205     1.000000
```

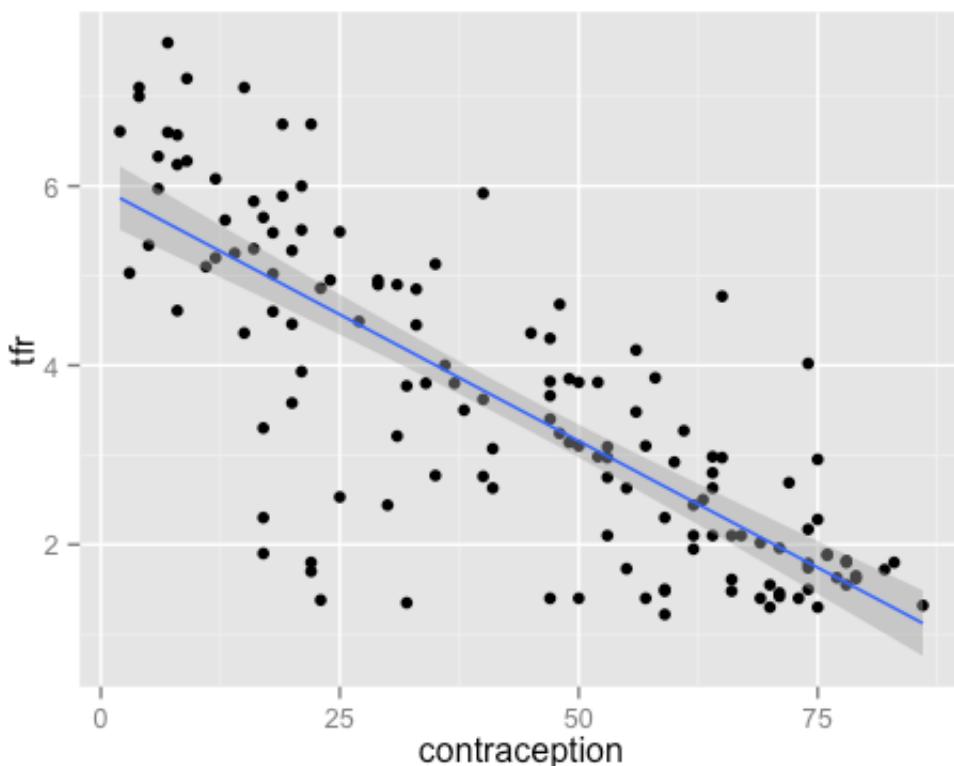
Paso 5: Evaluar la significancia de los coeficientes

```
summary(modelo2)

##
## Call:
## lm(formula = tfr ~ contraception, data = mundo98)
##
## Residuals:
##   Min     1Q Median     3Q    Max
## -3.3000 -0.4562  0.0617  0.6620  2.4620
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.978989  0.186531  32.05  <2e-16 ***
## contraception -0.056477  0.003778 -14.95  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.069 on 139 degrees of freedom
##   (66 observations deleted due to missingness)
## Multiple R-squared:  0.6165, Adjusted R-squared:  0.6138
## F-statistic: 223.5 on 1 and 139 DF,  p-value: < 2.2e-16
```

Gráfico con intervalos de confianza

```
ggplot(data2, aes(x=contraception, y=tfr)) + geom_point() +
    geom_smooth(method=lm)
```



Prueba de significancia del coeficiente "r de Pearson"

```
misvars <- c("tfr", "contraception")
data2 <- na.omit(mundo98[misvars])
cor.test(data2$tfr, data2$contraception)

##
## Pearson's product-moment correlation
##
## data: data2$tfr and data2$contraception
## t = -14.9498, df = 139, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.8413116 -0.7123600
## sample estimates:
##       cor
## -0.785205
```

Matriz de correlaciones

Podemos evaluar rápidamente cómo se presentan las correlaciones entre un conjunto de variables utilizando una matriz de correlaciones:

```
misvars <- c("tfr", "contraception", "economicActivityFemale")
data3 <- na.omit(mundo98[misvars])
cor(data3)
```

```
##                                     tfr contraception economicActivityFem
ale
## tfr                         1.0000000 -0.7490240      -0.2567
071
## contraception            -0.7490240  1.0000000       0.1455
538
## economicActivityFemale -0.2567071   0.1455538      1.0000
000
```

Matriz de correlaciones con la función "rcorr"

```
library(Hmisc)
rcorr(as.matrix(data3))

##                                     tfr contraception economicActivityFemale
## tfr                         1.00          -0.75      -0.26
## contraception            -0.75           1.00      0.15
## economicActivityFemale -0.26           0.15      1.00
##
## n= 116
##
##
## P
##                                     tfr     contraception economicActivityFemale
## tfr                         0.0000        0.0054
## contraception            0.0000        0.1190
## economicActivityFemale 0.0054        0.1190
```