Encuesta de Convivencia y Seguridad Ciudadana - ECSC - 2014

Carlos Ignacio Patiño (cpatinof@gmail.com)

Julio, 2015

Los microdatos y documentación sobre la Encuesta de Convivencia y Seguridad Ciudadana (ECSC), para 2014, se encuentran alojados en el siguiente link.

(Descripción de la encuesta, tomada de la documentación del DANE): La Política Nacional de Seguridad y Convivencia Ciudadana que se describe en el Plan Nacional de Desarrollo 2010-2014 «Prosperidad para todos», está "orientada a la protección del ciudadano frente a los riesgos y amenazas a su seguridad, permitiendo la convivencia y la prosperidad económica", para adoptar las medidas necesarias para brindar seguridad a los ciudadanos y garantizar sus derechos y libertades. Asimismo, dentro de los lineamientos de la estrategia se encuentra el mejorar el Sistema Nacional de Información del Delito, con el fin de mejorar la toma de decisiones en política pública.

En ese sentido, la ECSC es la herramienta de seguimiento, evaluación y monitoreo de políticas que permitirá realizar un diagnóstico de las principales ciudades del país en términos de convivencia y seguridad. Dada su importancia para el sector, la ECSC es un elemento fundamental del Plan Estadístico Sectorial que es un instrumento técnico permanente que identifica la producción de información estadística estratégica y los requerimientos de información estadística necesarios para tomar decisiones y formular política pública.

La encuesta además permite visibilizar los aspectos relacionados con la criminalidad en distintos contextos (en función del tipo de delito) y caracterizar la población afectada. Asimismo, establece los indicadores sobre el grado de violencia sufrida y la frecuencia de hechos violentos que hayan sido o no denunciados.

1 Cargue de datos

```
vivienda <- read.table("Datos de la vivienda/Datos de la vivienda.txt",
                       header=T, sep="\t")
personas <- read.table("Caracteristicas generales de las personas/personas.txt",
                       header=T, sep="\t")
persegco <- read.table("Percepcion de seg y conv/percepcionSegConv.txt",
                       header=T, sep="\t")
```

Se cargan tres tablas, la que contiene información sobre la vivienda, la que contiene información sobre las personas encuestadas en cada hogar y la que contiene 45 columnas referentes a la percepción de seguridad y convivencia.

2 Tabla Vivienda

De la tabla de vivienda nos interesan las siguientes columnas:

- `DIRECTORIO`: Directorio (identificador de vivienda)
- `DEPMUNI`: Municipio, 28 municipios.

Se crea la siguiente tabla, a partir de la información documentada por el DANE:

```
library(pander)
DEPMUNI <- read.table("DEPMUNI.txt", header=T, sep="\t")
pander(DEPMUNI)
```

DEPMUNI	Municipio
5001	Medellín
8001	Barranquilla
8758	Soledad
11001	Bogotá D.C.
13001	Cartagena
15001	Tunja
17001	Manizales
19001	Popayán
20001	Valledupar
23001	Montería
25754	Soacha
27001	Quibdó
41001	Neiva
44001	Riohacha
47001	Santa Marta
50001	Villavicencio
52001	Pasto
54001	Cúcuta
63001	Armenia
66001	Pereira
68001	Bucaramanga
70001	Sincelejo
73001	Ibagué
76001	Cali
76109	Buenaventura
76520	Palmira
76834	Tuluá
88001	San Andrés

Usaremos más adelante la información de esta tabla (`DEPMUNI`) para recodificar la columna correspondiente al municipio y poder hacer inferencia relacionada con niveles de seguridad (percepción) en cada municipio.

- `P5752S1`: Estrato para tarifa. En una etapa posterior, convertiremos esta columna al tipo "factor" con

el objetivo de usar también su información para efectos comparativos. En esta columna los valores 1 al 6 corresponden al estrato mientras que el 8 corresponde a la existencia de una conexión pirata o planta eléctrica y el 9 corresponde a "no sabe o no responde"

3 Tabla Personas

De la tabla de personas nos interesan las siguientes columnas:

- `DIRECTORIO`: Directorio (identificador de vivienda)
- `SECUENCIA_ENCUESTA`: Identifica hogar dentro de la vivienda (cada encuesta se aplica a un hogar)
- `SECUENCIA_P`: Identifica la persona en la encuesta (hogar)
- `ORDEN`: Identificador de la persona dentro del hogar
- `P220`: Género, 1 Masculino y 2 Femenino
- `P5785`: Edad. 0-107 (media 32)
- `P5501`: Parentesco con el jefe de hogar

Valor	Categoría
1	Jefe(a) del hogar
2	Cónyuge, compañero(a)
3	Hijo(a), hijastro(a).
4	Yerno, nuera.
5	Nietos(as).
6	Padre, madre, suegro(a).
7	Hermano(a).
8	Otro pariente.
9	Empleado del servicio doméstico.
10	Otros no parientes

Table 2: Códigos parentesco con jefe hogar

- `P6210`: Nivel educativo más alto.

Valor	Categoría
1	Ninguno
2	Preesoclar
3	Básica primaria (1-5)
4	Básica secundaria (6-9)
5	Media (10-13)
6	Superior o universitaria.
9	No sabe/No informa

Valor	Categoría

Table 3: Códigos nivel educativo individuo

- `P1366`: Estado civil (ver documentación DANE para tabla de códigos)
- `1402`: ¿Cuánto tiempo lleva viviendo en la actual vivienda?
- `1403`: ¿Cuánto tiempo lleva viviendo en el barrio? (1: Menos de un año, 2: Entre 1 y menos de 5 años, 3: Entre 5 y menos de 10 años, 4: 10 años o más)
- `1365`: Actividad predominante durante la semana pasada.

Valor	Categoría
1	Trabajando
2	Buscando trabajo
3	Estudiando
4	Oficios del hogar
5	Incapacitado permanente para trabajar
6	Pensionado
7	Ocio
8	Otra actividad, ¿cuál?

Table 4: Códigos actividad individuo

4 Tabla Percepción de seguridad y convivencia

De la tabla con las respuestas sobre la percepción de seguridad nos interesan las siguientes columnas.

- `DIRECTORIO`: Directorio (identificador de vivienda)
- `SECUENCIA_ENCUESTA`: (por identificar)
- `SECUENCIA_P`: (por identificar)
- `ORDEN`: Identificador de la persona dentro del hogar
- `P1362`: Percepción de seguridad en el barrio. (1. Seguro, 2. Inseguro)
- `P1361S1` a `P1362S9`: Razones para sentir inseguridad en el barrio. (1. Si, 2. No). Familiares o amigos han sido víctimas de agresiones; información que ve en medios o escucha en la calle; poca presencia fuerza pública; presencia de delincuencia común, robos o agresiones; presencia de pandillas; existencia de lotes baldíos o vías públicas sin iluminación; existencia de expendios de drogas (ollas); existencia de basureros; presencia de guerrilla o bandas criminales
- `P1359`: Percepción de seguridad en la ciudad. (1. Seguro, 2. Inseguro)
- `P1358S1` a `P1358S10`: Razones para sentir inseguridad en la ciudad. (1. Si, 2. No). Familiares o amigos han sido víctimas de agresiones; información que ve en medios o escucha en la calle; poca presencia fuerza pública; presencia de delincuencia común, robos o agresiones; presencia de pandillas; existencia de lotes baldíos o vías públicas sin iluminación; existencia de expendios de drogas (ollas); existencia de basureros; presencia de guerrilla o bandas; falta de empleo
- `P564`: Cree que es posible ser víctima de un delito en los próximos 12 meses (1. Si, 2. No)
- `P1117`: Qué tan posible (1. Mucho, 2. Algo, 3, Poco)

- `P1356S1` a `P1356S7`: Como se siente en cuanto a seguridad en los siguientes lugares (1. Seguro, 2. Inseguro, 3. No frecuenta el sitio). Donde realiza su actividad principal; Parques públicos, espacios recreativos o deportivos; Plazas de mercado, calles comerciales; Transporte público; Cajeros automáticos en vía pública; Vía pública; Discotecas, bares o sitios de entretenimiento
- `P1116`: Principal medida que toma para su seguridad

Valor	Categoría
1	Cambia de rutina o de actividades
2	Evita salir de noche
3	Sale solamente a lo necesario, evita frecuentar sitios públicos
4	Evita salir solo
5	Evita hablar con desconocidos
6	Evita portar grandes cantidaddes de dinero u objetos de valor
7	Otra
8	Ninguna

Table 5: Códigos medidas de seguridad

- `P1115`: Qué haría si es testigo de un hecho delictivo, como hurto o agresión física? (1. Acude en ayuda de la persona, 2. Pide auxilio, 3. Huye, 4. No hace nada)

5 Integración y re-codificación de las variables de interés

A continuación integramos las tres tablas, filtrando las respectivas columnas de interés y posteriormente re-codificamos (o re-definimos los niveles de los factores) las columnas de tipo cualitativo (factores) con el fin de facilitar el análisis exploratorio de la base de datos. (Ver código completo en la versión completa del caso que no incluye solución sugerida)

6 Tutorial R: Intervalos de confianza y pruebas de hipótesis

Recordemos la distribución normal estandarizada, muy util para expresar variables aleatorias que siguen una distribución normal. La distribución normal estandarizada tiene una media igual a cero y una desviación estándar igual a 1. Recordemos también que el 68% de la distribución se encuentra entre 1 desviación estándar de la media. El 95% se encuentra entre 2 desviaciones y el 99.7% entre 3 desviaciones. De esta manera es facil calcular la probabilidad asociada a un rango de valores. Por ejemplo, supongamos que la presión sanguinea para hombres entre 35 y 44 años se distribuye normalmente con media 80 y desviación estándar 10. Si miramos a un hombre en ese rango de edad (aleatoriamente), ¿cuál es la probabilidad de que su presión esté por debajo de 70? Podemos aproximar la respuesta empleando la regla anterior. Sabemos que la media es 80. Tambien sabemos que el 68% de los datos se encuentran entre una desviación estándar, por lo que se puede decir que el 34% de los datos se encuentran entre la media y una desviación estándar por debajo de la media, o 70. Como la normal es simétrica, lo anterior significa que el 84% de los datos están por encima de 70. De esta manera, se estima que la probabilidad de observar una presión por debajo de 70 es igual a 16%. Verifiquemos esto empleando R.

```r
# La función pnorm() se emplea para encontrar la probabilidad
pnorm(70, mean = 80, sd = 10, lower.tail = TRUE)
```

```
[1] 0.1586553
```

El parámetro `lower.tail` se emplea para pedirle a R la probabilidad de una u otra cola. Por ejemplo, si en el caso anterior quisieramos conocer la probabilidad de observar un individuo con una presión por encima de 70, la función en R debería ser:

```r
# Cambiamos el parámetro lower.tail
pnorm(70, mean = 80, sd = 10, lower.tail = FALSE)
```

```
[1] 0.8413447
```

Supongamos ahora que tomamos una muestra aleatoria de 100 hombres entre 35 y 44 años, ¿cuál sería el intervalo de confianza para la presión medía de esta muestra, al 95% de confianza? Recordemos que el intervalo de confianza está definido como:

$$\bar{X} \pm z\frac{\sigma}{\sqrt{n}}$$

En R, la formula anterior puede ser implementada de la siguiente manera:

```r
80 + c(-1,1) * qnorm(0.975) * 10/sqrt(100)
```

```
[1] 78.04004 81.95996
```

Por lo que la media de una muestra de 100 individuos se encuentra entre 78 y 82 a un 95% de confianza. Lo anterior implica que si sacamos esa muestra muchas veces, el 95% de las veces la media muestral de la presión para esos 100 individuos se va a ubicar en el intervalo calculado previamente.

Finalmente, miremos la prueba de hipótesis usando R. Para mayor información sobre la implementación de pruebas de hipótesis en R, revisar el tutorial en linea de R Tutorial.

Primero que todo, recordemos que la pregunta de investigación debe ir siempre como hipótesis alterna. Igualmente, en la hipótesis alterna no se debe incluir el signo "=".

Prueba de cola inferior para la media poblacional (varianza conocida)

$$H_o : \mu \geq \mu_0$$
$$H_a : \mu < \mu_0$$

El estadístico z se define en términos de la media muestral, el tamaño de la muestra y la desviación estándar de la población (σ):

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

La hipótesis nula se rechaza en caso de que $z \leq -z_\alpha$, donde z_α corresponde al $100(1-\alpha)$ percentil de la distribución normal estandarizada.

Ejemplo: Supongamos que el fabricante señala que la vida promedio de sus bombillos es de más de 10.000 horas. En una muestra de 30 bombillos se determinó que éstos duraron solamente 9.900 horas, en promedio.

Suponiendo una desviación estándar poblacional de 120 horas, ¿es posible rechazar el señalamiento del fabricante? ($\alpha = 0.05$)

La hipótesis que nos interesa, desde el punto de vista de la investigación, es que la media es menor a 10.000 horas. De esta manera tenemos que $H_o : \mu \geq 10000$ y $H_a : \mu < 10000$. El código en R, que nos permite verificar esta hipótesis, se presenta a continuación:

```
xbar <- 9900            # media muestral
mu0 <- 10000            # valor hipótesis
sigma <- 120            # st. dev. población
n <- 30                 # tamaño muestra
z <- (xbar-mu0)/(sigma/sqrt(n))
z                       # estadístico de la prueba
```

```
[1] -4.564355
```

```
alpha <- 0.05
z.alpha <- qnorm(1-alpha)
-z.alpha                # valor crítico
```

```
[1] -1.644854
```

El estadístico de nuestra prueba se ubica en la zona de rechazo (debajo del valor crítico), por lo que rechazamos la hipótesis nula en favor de la alternativa. La vida media de un bombillo no se encuentra por encima de las 10.000 horas.

De manera alternativa, podemos calcular el valor p a partir del estadístico obtenido en el paso anterior. Así, no hay necesidad de calcular el valor crítico, ya que el valor p nos permite tomar una decisión para cualquier α de interés.

```
pvalue <- pnorm(z)
pvalue
```

```
[1] 2.505166e-06
```

Prueba de cola superior para la media poblacional (varianza conocida)

$$H_o : \mu \leq \mu_0$$
$$H_a : \mu > \mu_0$$

Ejemplo: La etiqueta de un paquete de galletas sugiere que cada galleta contiene por mucho 2 gramos de grasas saturadas. En una muestra de 35 galletas se encuentra que la cantidad de grasas saturadas es en promedio 2.1 gramos. Suponga una desviación estándar poblacional de 0.25 gramos. Al 5% de significancia, ¿es posible rechazar el señalamiento del fabricante?

```
xbar <- 2.1          # media muestral
mu0 <- 2             # valor hipótesis
sigma <- 0.25        # st. dev. población
n <- 35              # tamaño muestra
z <- (xbar-mu0)/(sigma/sqrt(n))
z
```

```
[1] 2.366432
```

No es necesario calcular el valor crítico ya que podemos usar el valor p para verificar la significancia.

```
pvalue <- pnorm(z, lower.tail=FALSE) # Note el parámetro lower.tail
pvalue
```

```
[1] 0.008980239
```

Dado el valor p es posible rechazar la hipótesis nula al nivel de significancia del 5%.

Prueba de dos colas para la media poblacional (varianza conocida)

$$H_o : \mu = \mu_0$$
$$H_a : \mu \neq \mu_0$$

Recuerde que el símbolo "=" nunca debe ir en la hipótesis alterna. En este caso, rechazamos la hipótesis nula cuando $z \leq -z_{\alpha/2}$ o cuando $z \geq z_{\alpha/2}$, donde $z_{\alpha/2}$ corresponde al $100(1-\alpha/2)$ percentil de la distribución normal estandarizada.

Ejemplo: Nos encontramos evaluando un programa de alimentación para niños de la primaria de un colegio de interés. El año pasado (año base), se levantaron métricas sobre la población objetivo (niños en el último año de primaria) y se obtuvo un peso promedio de 42Kg. Este año, a una muestra de 35 de niños en el último año de primaria se les calculó un peso promedio de 44Kg. Suponiendo una desviación estándar poblacional de 5Kg, ¿existe evidencia para asegurar que los pesos promedios no difieren año a año?

En este caso nuestra hipótesis alterna es $H_a : \mu \neq 42$.

```
xbar <- 44          # media muestral
mu0 <- 45           # valor hipótesis
sigma <- 5          # st. dev. población
n <- 35             # tamaño muestra
z <- (xbar-mu0)/(sigma/sqrt(n))
z
```

```
[1] -1.183216
```

Calculemos los valores críticos para generar nuestra conclusión.

```
alpha <- 0.05
z.alpha_2 <- qnorm(1-alpha/2)
c(-z.alpha_2, z.alpha_2)      # valor crítico
```

```
[1] -1.959964  1.959964
```

Dado que nuestro estadístico prueba se ubica dentro del rango crítico (o fuera de la zona de rechazo), no existe evidencia estadística suficiente para rechazar la hipótesis nula de que el peso promedio de los niños evaluados no difiere del peso reportado el año pasado (42Kg). El programa no ha tenido el impacto esperado.

En este caso, el valor p se obtiene:

```
pvalue <- 2*pnorm(z, lower.tail=T) # x2 para dos colas
pvalue
```

```
[1] 0.2367236
```

Note que este valor es muy elevado, hecho que no permite rechazar la hipótesis nula ni siquiera a niveles de significancia más altos como 10% o incluso 15%.

Prueba de cola inferior para la media poblacional (varianza desconocida)

$$H_o : \mu \geq \mu_0$$
$$H_a : \mu < \mu_0$$

El estadístico t se define en términos de la media muestral, el tamaño de la muestra y la desviación estándar de la **muestra** (s):

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

La hipótesis nula se rechaza en caso de que $t \leq -t_\alpha$, donde t_α corresponde al $100(1-\alpha)$ percentil de la distribución t-Student con $n-1$ grados de libertad.

Ejemplo: Supongamos que el fabricante señala que la vida promedio de sus bombillos es de más de 10.000 horas. En una muestra de 30 bombillos se determinó que éstos duraron solamente 9.900 horas, en promedio. Suponiendo una desviación estándar muestral de 125 horas, ¿es posible rechazar el señalamiento del fabricante? ($\alpha = 0.05$)

La hipótesis que nos interesa, desde el punto de vista de la investigación, es que la media es menor a 10.000 horas. De esta manera tenemos que $H_o : \mu \geq 10000$ y $H_a : \mu < 10000$\$. El código en R, que nos permite verificar esta hipótesis, se presenta a continuación:

```
xbar <- 9900             # media muestral
mu0 <- 10000             # valor hipótesis
s <- 125                 # st. dev. muestra
n <- 30                  # tamaño muestra
t <- (xbar-mu0)/(s/sqrt(n))
t                       # estadístico de la prueba
```

```
[1] -4.38178
```

```
alpha <- 0.05
t.alpha <- qt(1-alpha, df=n-1)
-t.alpha                # valor crítico
```

```
[1] -1.699127
```

El estadístico de nuestra prueba se ubica en la zona de rechazo (debajo del valor crítico), por lo que rechazamos la hipótesis nula en favor de la alternativa. La vida media de un bombillo no se encuentra por encima de las 10.000 horas.

De manera alternativa, podemos calcular el valor p a partir del estadístico obtenido en el paso anterior. Así, no hay necesidad de calcular el valor crítico, ya que el valor p nos permite tomar una decisión para cualquier α de interés.

```
pvalue <- pt(t, df=n-1, lower.tail=T)
pvalue
```

```
[1] 7.035026e-05
```

Para cola superior y dos colas, se sigue la misma estrategia, pero haciendo uso del estadístico t.

Prueba de cola inferior para la proporción poblacional

Las hipótesis nula y alterna para una prueba de cola inferior para la proporción poblacional se expresan de la siguiente manera:

$$H_o : p \geq p_0$$
$$H_a : p < p_0$$

Donde p_0 es el límite inferior hipotético de la proporción poblacional verdadera p.

El estadístico z se define en términos de la proporción muestral y el tamaño de la muestra:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

La hipótesis nula se rechaza en caso de que $z \leq -z_\alpha$, donde z_α corresponde al $100(1-\alpha)$ percentil de la distribución normal estandarizada.

Ejemplo: Supongamos que en términos históricos, la proporción de personas que señala sentirse insegura en su barrio o vivienda es del 60%. En una encuesta reciente, se encontró que 51 de cada 100 personas encuestadas señalaron sentirse inseguras. A un nivel de significancia del 5%, ¿es posible rechazar la hipótesis nula de que la proporción de personas inseguras está por encima de 60% hoy en día?

En este ejemplo, la hipótesis nula es $H_o : p \geq 0.6$ y la hipótesis alterna es (lo que queremos investigar, es decir, que la proporción de individuos con percepción de inseguridad ha caído) $H_o : p < 0.6$. El primer paso es calcular el estadístico para la prueba:

```
pbar <- 51/100          # proporción muestral
p0 <- .6                # valor hipotético
n <- 100                # tamaño muestra
z <- (pbar-p0)/sqrt(p0*(1-p0)/n)
z                      # estadistico
```

```
[1] -1.837117
```

El valor crítico, para el nivel de significancia del 5% es:

```
alpha <- 0.05
z.alpha <- qnorm(1-alpha)
-z.alpha    # valor crítico
```

```
[1] -1.644854
```

El estadístico está por debajo del valor crítico (zona de rechazo) por lo que podemos rechazar la hipótesis nula en favor de la alterna. La proporción de individuos que reportan sentirse inseguros es inferior al histórico de 60%.

De manera alternativa, podemos calcular el valor p.

```
pval <- pnorm(z)   # lower.tail
pval
```

```
[1] 0.03309629
```

En R, podemos hacer pruebas para proporciones de manera sencilla, empleando la función `prop.test()`:

```
prop.test(51,100,p=0.6,alt="less",correct=FALSE)
```

```
    1-sample proportions test without continuity correction

data:  51 out of 100, null probability 0.6
X-squared = 3.375, df = 1, p-value = 0.0331
alternative hypothesis: true p is less than 0.6
95 percent confidence interval:
 0.000000 0.590873
sample estimates:
   p
0.51
```

Prueba de cola superior para la proporción poblacional

Usando la misma información del ejemplo anterior, probemos la hipótesis contraria. Es decir, nuestra hipótesis nula ahora es $H_o : p \leq 0.6$ y la alterna es $H_o : p > 0.6$. En este caso (cola superior) la zona de rechazo se encuentra donde $z \geq z_\alpha$.

El estadistico es igual al caso anterior (misma información):

```
pbar <- 51/100          # proporción muestral
p0 <- .6                # valor hipotético
n <- 100                # tamaño muestra
z <- (pbar-p0)/sqrt(p0*(1-p0)/n)
z                      # estadistico
```

```
[1] -1.837117
```

El valor crítico, para el nivel de significancia del 5% es:

```
alpha <- 0.05
z.alpha <- qnorm(1-alpha)
z.alpha    # valor crítico
```

```
[1] 1.644854
```

El estadístico está por debajo del valor crítico (zona de no rechazo) por lo que podemos no rechazar la hipótesis nula. La proporción de individuos que reportan sentirse inseguros es inferior al histórico de 60%.

De manera alternativa, podemos calcular el valor p.

```
pval <- pnorm(z, lower.tail=FALSE)   # upper.tail
pval
```

```
[1] 0.9669037
```

El valor p es consistente con el resultado anterior. Empleando la función `prop.test()`:

```
prop.test(51,100,p=0.6,alt="greater",correct=FALSE)
```

```
    1-sample proportions test without continuity correction

data:  51 out of 100, null probability 0.6
X-squared = 3.375, df = 1, p-value = 0.9669
alternative hypothesis: true p is greater than 0.6
95 percent confidence interval:
 0.4286002 1.0000000
sample estimates:
   p
0.51
```

Prueba de dos colas para la proporción poblacional

En este caso, las regiones de rechazo están definidas por: $z \leq -z_{\alpha/2}$ o $z \geq z_{\alpha/2}$, donde $z_{\alpha/2}$ corresponde al $100(1-\alpha/2)$ percentil de la distribución normal estandarizada.

Ejemplo: Supongamos que una moneda cae en cara en 12 de 20 lanzamientos. Al 5% de significancia, ¿es posible rechazar la hipótesis nula de que la moneda es "limpia"? $(H_0 : p = 0.5)$

```
pbar <- 12/20          # proporción muestral
p0 <- .5                # valor hipotético
n <- 20                # tamaño muestra
z <- (pbar-p0)/sqrt(p0*(1-p0)/n)
z                      # estadistico
```

```
[1] 0.8944272
```

Los valores críticos, para el nivel de significancia del 5% son:

```
alpha <- 0.05
z.alpha.2 <- qnorm(1-alpha/2)
c(-z.alpha.2, z.alpha.2)    # valor crítico
```

```
[1] -1.959964  1.959964
```

El estadístico se encuentra dentro del rango dado por los valores críticos (zona de no rechazo), por lo que no es posible rechazar la hipótesis nula. La moneda parece ser una moneda "limpia".

Recordemos que el valor p para este caso se define como:

```r
pval <- 2 * pnorm(z, lower.tail=FALSE)   # upper.tail
pval
```

```
[1] 0.3710934
```

O, de manera alternativa, usando la función `prop.test()` en R:

```r
prop.test(12,20,p=0.5,correct=FALSE)
```

```
    1-sample proportions test without continuity correction

data:  12 out of 20, null probability 0.5
X-squared = 0.8, df = 1, p-value = 0.3711
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3865815 0.7811935
sample estimates:
  p
0.6
```

Inferencia sobre dos poblaciones

A menudo nos encontramos con la necesidad de generar conclusiones acerca de la diferencia entre dos poblaciones, a partir de muestras para cada población.

Media poblacional entre dos muestras emparejadas

Hablamos de dos muestras emparejadas cuando los datos provienen de repetidas observaciones del mismo sujeto. Por ejemplo, es muy común ver este tipo de muestras en experimentos donde se selecciona aleatoriamente la mitad del total de individuos, se les aplica el tratamiento, y en una segunda instancia, se aplica el tratamiento a la segunda mitad, dejando la mitad inicial como control. En estos casos, se tienen dos observaciones para cada individuo, una bajo tratamiento y otra bajo placebo.

Supongamos que los 25 estudiantes de una clase presentaron un test donde el promedio de su nota fue 2.5. Ante los malos resultados, a los estudiantes se les reforzó el contenido a través de horas adicionales y talleres. En una segunda prueba, los estudiantes obtuvieron una nota promedio de 3.3. ¿Son estos promedios estadisticamente diferentes? En otras palabras, ¿fue efectivo el programa de horas adicionales de talleres para reforzar el contenido?

```r
head(df, n=3)  # Miremos los primeros registros de los datos
```

```
        pre     post
1 1.2387168 1.238717
2 4.5765390 4.576539
3 0.6167622 2.366149
```

```r
# Usamos la función t.test
t.test(df$post, df$pre, paired=T)
```

```
    Paired t-test
```

```
data:  df$post and df$pre
t = 5.2261, df = 24, p-value = 2.346e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.4723079 1.0888305
sample estimates:
mean of the differences
              0.7805692
```

La salida de la función `t.test()` nos muestra tanto la media de la diferencia entre las muestras, como el intervalo de confianza y el valor p asociado a la prueba. En este caso, observamos un valor p muy bajo, lo que nos permite rechazar la hipótesis nula, en favor de la alterna (la diferencia en medias es diferente de cero). Dado que el intervalo estimado se ubica en el rango positivo, es posible decir que el promedio de los estudiantes después del programa de mejora incrementó la nota promedio en el examen.

Media poblacional entre dos muestras independientes

Dos muestras son independientes si provienen de poblaciones no relacionadas y las muestras no se afectan entre si. Supongamos que en el ejemplo anterior fallamos en el momento de incorporar el hecho de que los registros están emparejados.

```
# Usamos la función t.test pero al argumento paired le asignamos F
t.test(df$post, df$pre, paired=F)
```

```
    Welch Two Sample t-test

data:  df$post and df$pre
t = 1.9706, df = 48, p-value = 0.05455
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01586459  1.57700304
sample estimates:
mean of x mean of y
 3.262418  2.481849
```

Notemos como en este caso no es posible rechazar la hipótesis nula por lo que se puede decir que la diferencia entre ambos promedios es cero. Al no considerar el hecho de que los registros están emparejados y dos notas corresponden al mismo estudiante (pre y post programa), la prueba incorpora la variabilidad de las dos muestras, por lo que no es posible rechazar la hipótesis.

Comparando dos proporciones

Usemos la base obtenida para este tutorial, sobre la encuesta de percepción de seguridad. Haciendo uso de la función `table()`, podemos sintetizar facilmente la información contenida en la base. Por ejemplo, si queremos saber la percepción de seguridad en la ciudad según hombres y mujeres, usamos el siguiente comando:

```
table(persegco_final$P220,persegco_final$P1359)
```

```
        Seguro Inseguro
  Hombre  43528    37550
  Mujer   48769    48596
```

Según esta tabla, 43.528 hombres reportan sentirse seguros en su ciudad mientras que 37.550 reportan sentirse inseguros. Para el caso de las mujeres, 48.769 se reportan seguras, mientras que una cantidad muy similar, 48.596 se reportan inseguras. Podemos usar la salida de la función `table()` para verificar si la diferencia en las proporciones de individuos que se sienten seguros entre hombres y mujeres es diferente de cero. La función mencionada anteriormente (`prop.test()`) también acepta como insumo tablas de una o dos dimensiones con las cuentas de "exitos" ("seguro" en este caso) y "fallas" ("inseguro" en este caso) para cada grupo, respectivamente. Así, la función puede ser utilizada incluyendo como insumo otra función `table()` que sintetiza y calcula las cuentas de cada incidencia de interés.

```
prop.test(table(persegco_final$P220,persegco_final$P1359))
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  table(persegco_final$P220, persegco_final$P1359)
## X-squared = 229.1752, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.03131370 0.04064095
## sample estimates:
##    prop 1    prop 2
## 0.5368657 0.5008884
```

El valor p muy bajo permite rechazar la hipótesis nula (al 5% de significancia) de que las proporciones son iguales. El orden de la resta de proporciones está definido por el orden en la tabla. En este caso, la proporción de hombres con percepción de seguridad va primero y a esta se le resta la proporción de mujeres seguras. El intervalo de confianza del 95% reporta entonces dicha diferencia. Según este, la diferencia de proporciones (hombres - mujeres) está entre 3% y 4% más o menos. Por lo tanto, consistente con el valor p, se puede decir que las mujeres reportan una menor proporción de individuos seguros. La percepción de seguridad es menor entre las mujeres.

Lo anterior se puede corroborar usando una prueba *Chi Cuadrado* para independencia. Recordemos que dos variables aleatorias son independientes si la distribución de probabilidad de una variable no se afecta por la presencia de la otra. En esta prueba se emplea también como insumo una tabla de contingencia (dos vías). Sus hipótesis están definidas como;

$$H_0 : \text{Las dos clasificaciones son independientes}$$
$$H_a : \text{Las dos clasificaciones son dependientes}$$

En R, la prueba de *Chi Cuadrado* se puede implementar con la función `chisq.test()`, así:

```
chisq.test(table(persegco_final$P220,persegco_final$P1359))
```

```
    Pearson's Chi-squared test with Yates' continuity correction

data:  table(persegco_final$P220, persegco_final$P1359)
X-squared = 229.1752, df = 1, p-value < 2.2e-16
```

El *valor p* nos permite rechazar la hipótesis nula de independencia al 5% de significancia. Género y percepción de seguridad no son dos variables independientes. En otras palabras, la distribucion de probabilidad de una u otra depende de la otra variable.

Recordemos que para más tutoriales (en inglés), el siguiente link es una muy buena fuente de información relacionada con estadística y R.

7 Caso de estudio: Intervalos de confianza y pruebas de hipótesis

Inicialmente, exploramos algunas de las columnas presentes en la base de datos procesada previamente en este caso. Para efectos prácticos, los cálculos en este caso los vamos a llevar a cabo sólo para los registros que corresponden a jefes de hogar (archivo `persegco_jefes.csv`). Si se está trabajando con el archivo completo (`persegco_final.csv`), necesitamos identificar la columna correspondiente que nos indique si el individuo es o no el jefe del hogar. La columna que buscamos es la `P5501` y el valor que nos interesa es 1 (Jefe(a) de hogar). Procedemos entonces a sacar un subconjunto de la base de datos que solo contenga jefes de hogar. Para cargar el subconjunto ya con los "jefes" filtrados, empleamos el siguiente código.

```
# Recuerde cargar el archivo de datos bajo el nombre jefes
jefes <- read.csv("persegco_jefes.csv",sep=";")
```

Ahora exploremos un poco la composición de los jefes de hogar encuestados. En total, hay 67928 individuos encuestados que reportan ser los jefes del hogar. La composición por género es la siguiente:

```
table(jefes$P220)
```

```
Hombre  Mujer
 38864  29064
```

En cuanto a la distribución de la edad de los jefes de hogar:

```
summary(jefes$P5785)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.00   36.00   48.00   48.19   59.00  104.00
```

A manera de calentamiento y exploración adicional de la base de datos empleada en este caso de estudio los estudiantes deberán presentar siguientes cuatro resultados.

1. Una tabla, en cualquier formato, que describa las características más relevantes de los jefes de hogar.

Edad Jefes de hogar

```
pander(summary(jefes$P5785))
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
15	36	48	48.19	59	104

Género Jefes de hogar

```
pander(summary(jefes$P220))
```

Hombre	Mujer
38864	29064

Nivel Educativo Jefes de hogar

```
pander(summary(jefes$P6210))
```

Media	Ninguno	No Sabe	Preescolar	Primaria
19420	2554	35	21	17838

Table 8: Table continues below

Secundaria	Superior
9748	18312

Actividad Jefes de hogar

```
pander(summary(jefes$P1365))
```

Buscando Trabajo	Estudiando	Hogar	Incapacitado	Ocio
1692	1221	11761	912	1437

Table 10: Table continues below

Otros	Pensionado	Trabajando
498	5052	45355

2. Un intervalo de confianza del 95% para la edad media de la población de jefes de hogar (en este caso debemos suponer que la muestra es representativa).

```
# La función t.test realiza prueba de hipótesis y genera a su vez intervalos
# de confianza al nivel de confianza seleccionado
res <- t.test(jefes$P5785,mu=48.2,conf.level=0.95,alternative="two.sided")
# El intervalo de confianza (95%) puede ser accedido con:
res$conf.int
```

```
[1] 48.06982 48.30473
```

```
attr(,"conf.level")
[1] 0.95
```

3. Respuesta a la pregunta: ¿Existen diferencias entre las percepciones de seguridad entre hombres y mujeres jefes de hogar? (usar intervalos de confianza).

```
# La función prop.test realiza prueba de hipótesis y genera a su vez intervalos
# de confianza al nivel de confianza seleccionado
res <- prop.test(table(jefes$P1359[jefes$P220=="Hombre"]),p=0.5,
                 conf.level=0.95,alternative="two.sided")
# El intervalo de confianza (95%) puede ser accedido con (recuerde que corresponde
# a la proporción de hombres que tienen percepción de inseguridad):
res$conf.int
```

```
[1] 0.4826426 0.4926068
attr(,"conf.level")
[1] 0.95
```

```
# Para el caso de las mujeres:
res <- prop.test(table(jefes$P1359[jefes$P220=="Mujer"]),p=0.5,
                 conf.level=0.95,alternative="two.sided")
res$conf.int
```

```
[1] 0.5050041 0.5165317
attr(,"conf.level")
[1] 0.95
```

El porcentaje de hombres que perciben inseguridad en la ciudad está entre 48.3% y 49.3%, mientras que el de mujeres está entre 50.5% y 51.7%. Dado que no hay un *overlap* entre ambos intervalos de confianza, existen indicios para pensar que hay una diferencia significativa entre las percepciones de seguridad de hombres y mujeres.

4. Evalúe la hipótesis "los jefes de hogar en Cali tienen una edad media de 45 años" (use alfa igual a 0.05).

```
# Usamos funcion t.test nnuevamente, pero esta vez revisamos todo su output
res <- t.test(jefes$P5785[jefes$Municipio=="Cali"],
              mu=45,
              conf.level=0.95,
              alternative="two.sided")
res
```

```
    One Sample t-test

data:  jefes$P5785[jefes$Municipio == "Cali"]
t = 12.9535, df = 2489, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 45
95 percent confidence interval:
 48.49331 49.73962
sample estimates:
mean of x
 49.11647
```

El grupo de trabajo se encuentra actualmente haciendo una consultoría para una ONG, en la cual se busca analizar la percepción de seguridad de los jefes de hogar para los estratos 1, 2 y 3. Según el director de la ONG, la población objetivo de los programas que la ONG está dispuesta a ofrecer en el territorio nacional es la de mujeres jefes de hogar entre 25 y 40 años de edad. Empleando las herramientas vistas en clase hasta la fecha, presente su recomendación al director de la ONG en cuanto a la estrategia actual de la organización. ¿Debería ésta continuar su enfoque en la población objetivo? Asegúrese de reportar gráficos y tablas que sustenten su respuesta. Asegúrese también de plantear su hipótesis y la conclusión obtenida a partir de la información en la base de datos.

Enfoque sugerido

Comparemos entonces la percepción de seguridad (o inseguridad) para el grupo objetivo y las personas que se encuentran fuera del rango de edad establecido por la ONG como población objetivo. Si encontramos diferencias significativas, que muestran que efectivamente la percepción de inseguridad del grupo objetivo es mayor a la del resto de la población, se puede decir que el programa, desde ese punto de vista, esta bien focalizado. Dado el calentamiento previo, ya sabemos que las mujeres presentan una mayor percepción de inseguridad, por lo que podemos suponer que en cuanto a género, el programa (dejando todo lo demás constante) apunta a una población que sí tiene una percepción mayor de inseguridad. Por esto, dejaremos esa dimensión por fuera de este análisis y nos enfocaremos en la edad.

Generemos primero nuevas columnas con la información sobre la población objetivo.

```
jefes.mujeres <- subset(jefes, P220=="Mujer") # subset mujeres
jefes.mujeres$edadObjetivo <- "No"
jefes.mujeres$estratoObjetivo <- "No"
jefes.mujeres$edadObjetivo[jefes.mujeres$P5785>=25 & jefes.mujeres$P5785<=40] <- "Si"
jefes.mujeres$estratoObjetivo[jefes.mujeres$Estrato %in% c(1,2,3)] <- "Si"
jefes.mujeres$poblacionObjetivo <- "No"
jefes.mujeres$poblacionObjetivo[jefes.mujeres$edadObjetivo=="Si" &
                                    jefes.mujeres$estratoObjetivo=="Si"] <- "Si"
jefes.mujeres <- jefes.mujeres[,c("P1359","edadObjetivo",
                                  "estratoObjetivo","poblacionObjetivo")]
```

Ahora, verifiquemos que el *data frame* sea correcto, empleando la función `table`

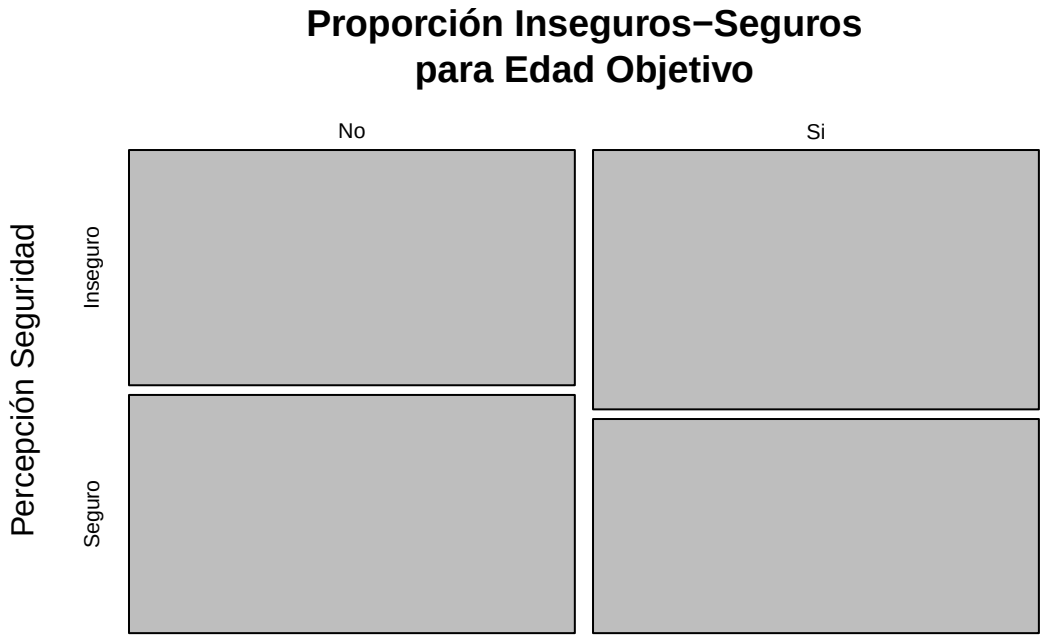
```
table(jefes.mujeres$poblacionObjetivo,jefes.mujeres$P1359)
```

```
    Inseguro Seguro
  No    10827  10888
  Si     4018   3331
```

Gráficos: Edad Objetivo y Estrato Objetivo

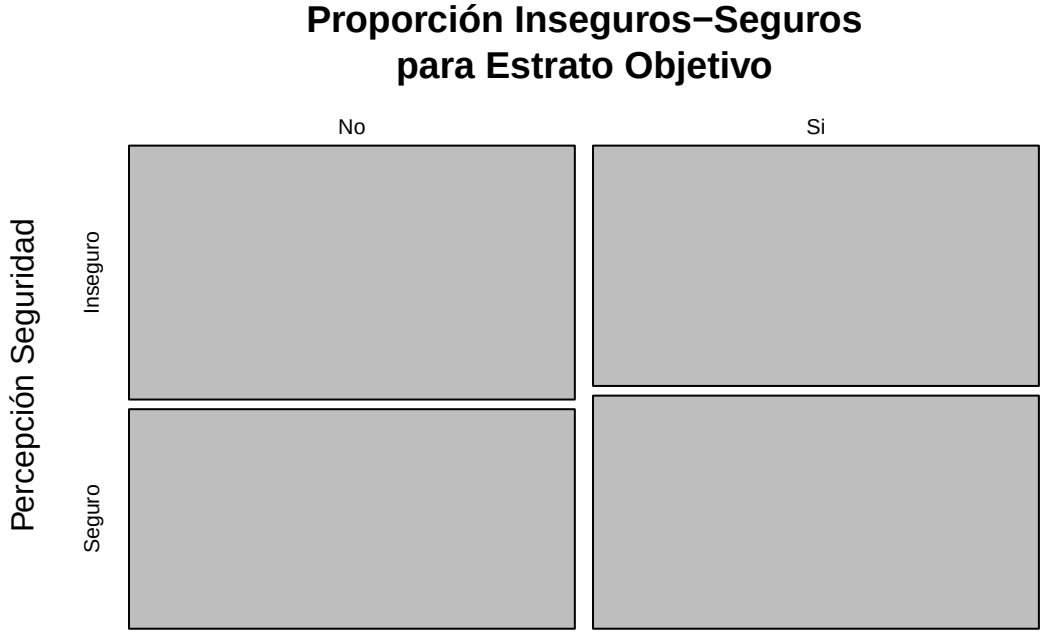
```
propTbl <- prop.table(table(jefes.mujeres$edadObjetivo,jefes.mujeres$P1359),
                      margin=1)

plot(propTbl, main="Proporción Inseguros-Seguros\npara Edad Objetivo",
     xlab="Edad Objetivo?", ylab="Percepción Seguridad")
```




```
propTbl <- prop.table(table(jefes.mujeres$estratoObjetivo,jefes.mujeres$P1359),
                      margin=1)

plot(propTbl, main="Proporción Inseguros-Seguros\npara Estrato Objetivo",
     xlab="Estrato Objetivo?", ylab="Percepción Seguridad")
```



En el caso de la edad se observa que el grupo objetivo presenta una mayor percepción de inseguridad. Sin embargo, en el caso del estrato, los estratos 1 al 3 presentan una menor percepción de inseguridad. Miremos

ahora ambos efectos combinados.

Prueba: A continuación empleamos una prueba de hipótesis sobre dos muestras (proporciones), con el objetivo de comparar ambas proporciones (percepción de inseguridad) para las poblaciones objetivo y no objetivo.

```
tbl <- table(jefes.mujeres$poblacionObjetivo,jefes.mujeres$P1359)
prop.test(tbl)
```

```
    2-sample test for equality of proportions with continuity
    correction

data:  tbl
X-squared = 50.7415, df = 1, p-value = 1.054e-12
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.06141860 -0.03487262
sample estimates:
   prop 1    prop 2
0.4985954 0.5467411
```

La proporción 2 corresponde a la proporción de personas dentro de la población objetivo que señalan tener una percepción de inseguridad en sus ciudades. Es decir, el 54.7% de las mujeres entre 25 y 40 años (inclusive) pertenecientes a los estratos 1, 2 y 3 señalan sentirse inseguras en sus respectivas ciudades. Según la prueba de proporciones para dos muestras, esta proporción es significativamente diferente a la proporción registrada en el resto de la población (50%).