Estadística Descriptiva: Soluciones sugeridas

Carlos Ignacio Patiño (cpatinof@gmail.com)

Julio 25, 2015

1 Objetivo

El objetivo del presente taller en clase (grupal) es permitir a los estudiantes practicar la aplicación de los conceptos de la estadística descriptiva, muy empleados en el análisis exploratorio de información.

2 ENUT

Los datos que serán empleados en el presente taller corresponden a la información de la Encuesta Nacional sobre Uso del Tiempo, llevada a cabo por el DANE durante el periodo 2012-2013. El archivo `enutVivienda.csv` contiene información relacionada con los estudiantes (de todos los niveles) encuestados que contestan de manera afirmativa a la pregunta de si durante la jornada de referencia (usualmente el día anterior a realizada la encuesta) dedicó algún tiempo, fuera de la jornada normal escolar, a estudiar o hacer tareas (relacionadas con sus estudios actuales) desde la casa.

En el archivo, las columnas corresponden a la siguiente información:

- `DIA_REFERENCIA_2`: corresponde a las 24 horas (de las 00:00 horas a las 23:59 horas) del día anterior al día de visita asignado Ej.: si se le ha asignado como día de visita al hogar el martes 14 de agosto el día de referencia será el lunes 13 de agosto de 2012 desde las 00:00 (o 12 de la noche) hasta las 23:59 (u 11:59 de la noche). Lunes (1) a Domingo (7), y (8) para día festivo.
- `P6040`: Edad. Si es menor de 1 año, el valor es 0.
- `P6020`: Género. (1: Masculio, 2: Femenino).
- `P6175`: Establecimiento (1: Público, 2: Privado).
- `P1158S1`: Nivel educativo: se refiere al nivel más alto de instrucción alcanzado por la persona, dentro del sistema formal de enseñanza, sea éste, educación preescolar, básica primaria, educación básica secundaria, superior o universitaria y postgrado.
- `P1161S1A1`: Horas de estudio (fuera de jornada escolar) en la VIVIENDA.
- `P1161S1A2`: Minutos de estudio (fuera de jornada escolar) en la VIVIENDA.

En todos los casos, las columnas han sido re-codificadas con el fin de tener en sus campos los valores descriptivos adecuados. Es decir, para el caso de la columna referente al género, en lugar de mantener los códigos 1 o 2, éstos se han recodificado a "Masculino" y "Femenino", con el objetivo de hacer más fácil la interpretación de cualquier resultado. Sin embargo, note que este tipo de estrategias puede ser poco adecuada para bases de datos muy grandes, en las que por eficiencia (no sólo por espacio sino también por velocidad en los procesos), se deben mantener estos códigos que ocupan menos espacio.

Igualmente, note que la información correspondiente al tiempo dedicado a estudiar en casa se encuentra separada en dos columnas diferentes. Esto no es una práctica muy recomendada en la gestión y análisis de bases de datos, por lo que el primer punto del presente taller le pedirá al estudiante que realice los pasos necesarios para obtener una base de datos apta para el análisis. En el proceso de análisis de información, los pasos necesarios para convertir una base de datos "cruda" en una base de datos analítica (o lista para el análisis) se denominan "Pasos de Procesamiento" o "Código de Procesamiento". En este caso, ya el instructor ha llevado a cabo este componente del proceso análitico y los estudiantes interesados en conocer dicha documentación, la podrán encontrar en el siguiente link.

3 Taller

1. (Limpieza de datos) Genere una nueva columna denominada `tiempoCasa` que combine las dos columnas correspondientes al tiempo dedicado al estudio desde la Vivienda. Esta columna puede ser expresada en horas o minutos.

```
# Creamos la nueva columna, agregando las columnas que contienen el tiempo
library(dplyr)
enut2 <- mutate(enut2, tiempoCasa=P1161S1A1+(P1161S1A2/60)) # Horas
```

2. (Estadística descriptiva) Caracterice el tiempo dedicado al estudio en la vivienda en Colombia. ¿Qué forma tiene su distribución? ¿Qué puede usted decir al respecto?

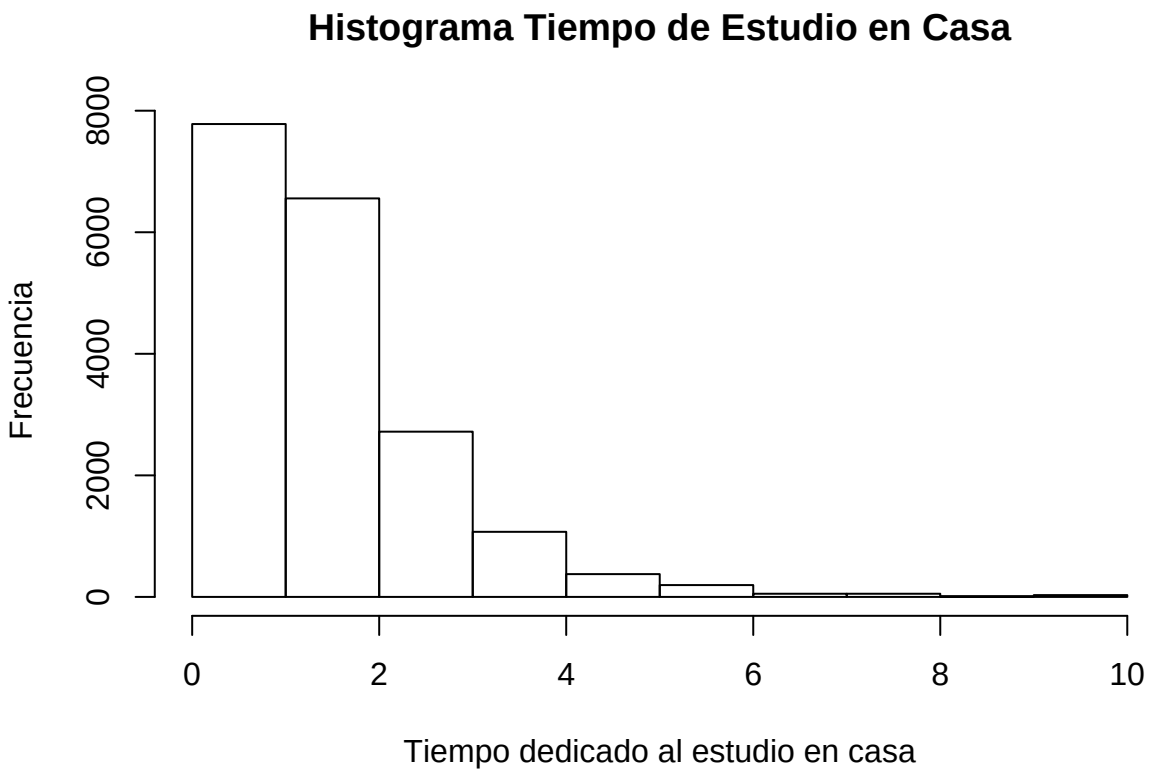
```
# Usamos la función summary() en R para generar estadísticas descriptivas
summary(enut2$tiempoCasa)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.01667  1.00000  2.00000  1.88800  2.00000 10.00000
```

Note que la mediana está por encima de la media, por lo que hay un (muy leve) sesgo hacia la izquierda en la distribución. En promedio, los estudiantes colombianos que reportan hacer tareas en sus casas (fuera de la jornada escolar), gastan cerca de 1.9 horas al día en dichas tareas. Se observa un máximo nivel de 10 horas, dato que se deba posiblemente a estudio durante un día de fin de semana.

Para inspeccionar la distribución, usamos la función base graficadora de R.

```
hist(enut2$tiempoCasa, main="Histograma Tiempo de Estudio en Casa",breaks=12,
     ylab="Frecuencia",xlab="Tiempo dedicado al estudio en casa")
```



Además del leve sesgo mencionado anteriormente, se observa que la mayoría de los estudiantes colombianos dedican entre 0 y 2 horas de estudio al día (en sus casas).

3. (Estadística descriptiva) ¿Existen diferencias en el comportamiento (tiempo de estudio en la Vivienda) entre hombres y mujeres?

Existen dos alternativas para responder a esta pregunta. Una numérica y la otra gráfica. Miremos primero la numérica.

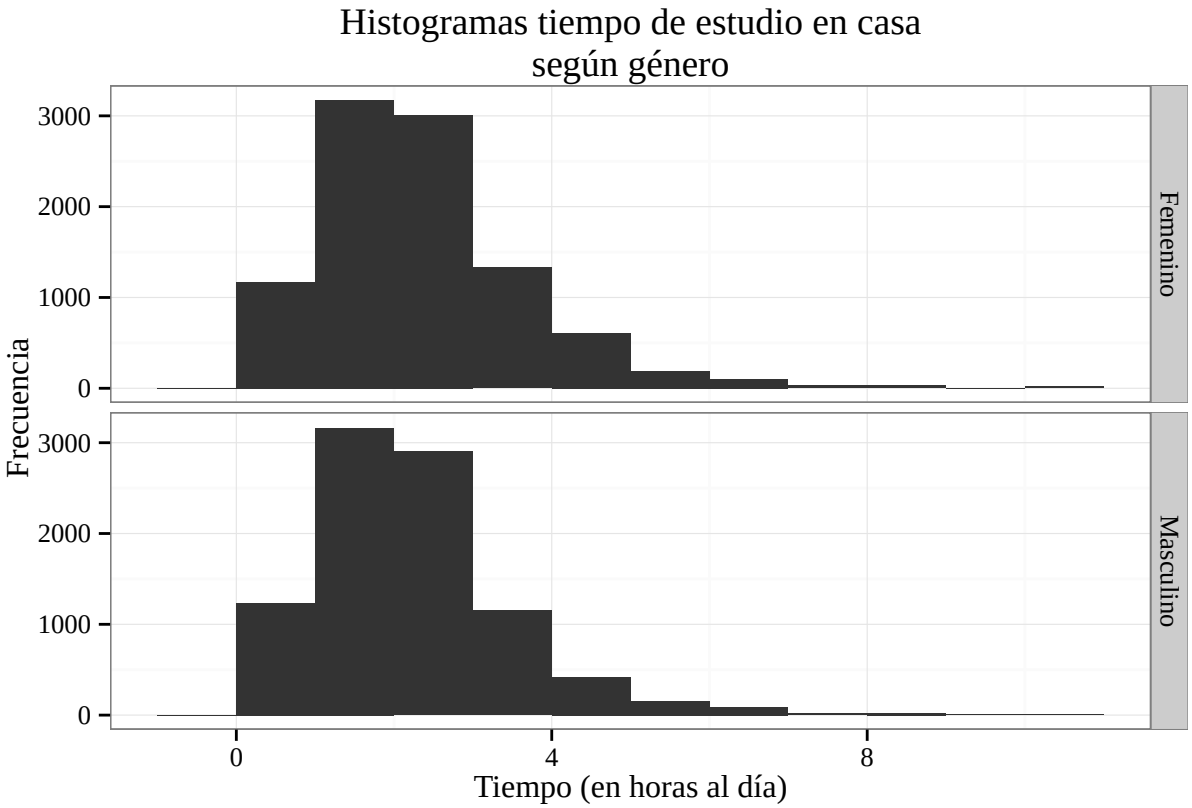
```
# Nuevamente usamos los verbos del paquete dplyr
enut_summ <- summarize(group_by(enut2,P6020),
                       tiempoMedio=mean(tiempoCasa,na.rm=T))
print(enut_summ)
```

```
## Source: local data frame [2 x 2]
##
##       P6020 tiempoMedio
## 1  Femenino    1.941038
## 2 Masculino    1.832307
```

Las mujeres, al parecer, dedican más tiempo al estudio en sus casas. Sin embargo, en las siguientes sesiones veremos como probar este tipo de hipótesis de manera estadística.

Ahora, miremos la manera gráfica.

```
library(ggplot2)
p <- ggplot(enut2, aes(x=tiempoCasa))
p + geom_histogram(binwidth=1) + facet_grid(P6020~.) +
        labs(title="Histogramas tiempo de estudio en casa\nsegún género") +
        labs(y="Frecuencia", x="Tiempo (en horas al día)") +
        theme_bw(base_family="Times", base_size=12)
```



Note que en análisis gráfico no es concluyente respecto a una diferencia significativa entre los tiempos de dedicación para hombres y mujeres.

4. (Estadística descriptiva) ¿Cuáles medidas de tiempo presentan mayor variabilidad? ¿Las asociadas a los estudiantes de Primaria, Secundaria y Media? O las asociadas a los estudiantes de Pregrado y programas Técnicos y Tecnológicos?

```
# Agrupamos los niveles de interés
primaria <- c("Primaria", "Secundaria o Media")
terciaria <- c("Universitario", "Técnico", "Tecnológico")
enut2$GrupoNivel <- "Otros"
enut2$GrupoNivel[enut2$P1158S1 %in% primaria] <- "Primaria/Secundaria"
enut2$GrupoNivel[enut2$P1158S1 %in% terciaria] <- "Universidad/Técnico"

# Revisamos la desviación estándar de cada grupo
summarize(group_by(enut2,GrupoNivel), StDev=sd(tiempoCasa,na.rm=T))
```

```
## Source: local data frame [3 x 2]
##
##            GrupoNivel    StDev
## 1               Otros 1.216540
## 2 Primaria/Secundaria 1.062848
## 3 Universidad/Técnico 1.627491
```

Se observa claramente que el tiempo dedicado por parte de los estudiantes universitarios y técnicos tiene una mayor variabilidad. Su desviación estándar es de 1.62, mientras que la del tiempo dedicado por estudiantes de colegio es de 1.06.

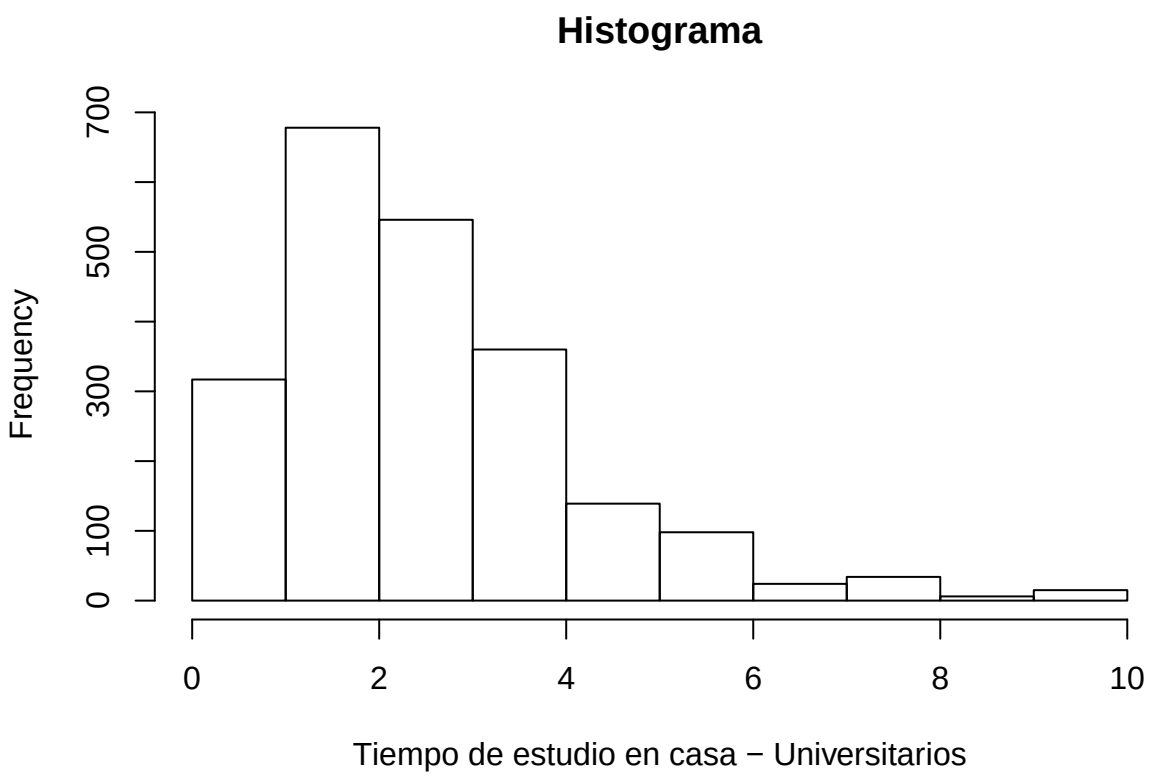
5. (Regla Empírica) Un estudiante de Pregrado reporta que durante el día de referencia estudió en su vivienda durante 7 horas. ¿Qué tan probable es dicho registro? ¿Se trata esto de un dato "normal", o de un posible dato atípico?

Revisemos primero las estadísticas descriptivas para el tiempo dedicado por los estudiantes universitarios:

```
summary(enut2$tiempoCasa[enut2$P1158S1=="Universitario"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.25    2.00    3.00    2.97    4.00   10.00
```

```
hist(enut2$tiempoCasa[enut2$P1158S1=="Universitario"], main="Histograma",
     xlab="Tiempo de estudio en casa - Universitarios")
```



La función `summary()` en R no genera automáticamente la desviación estándar, por lo que es necesario que la calculemos directamente.

```
sd(enut2$tiempoCasa[enut2$P1158S1=="Universitario"],na.rm=T)
```

```
## [1] 1.649396
```

Recordemos la regla empírica:

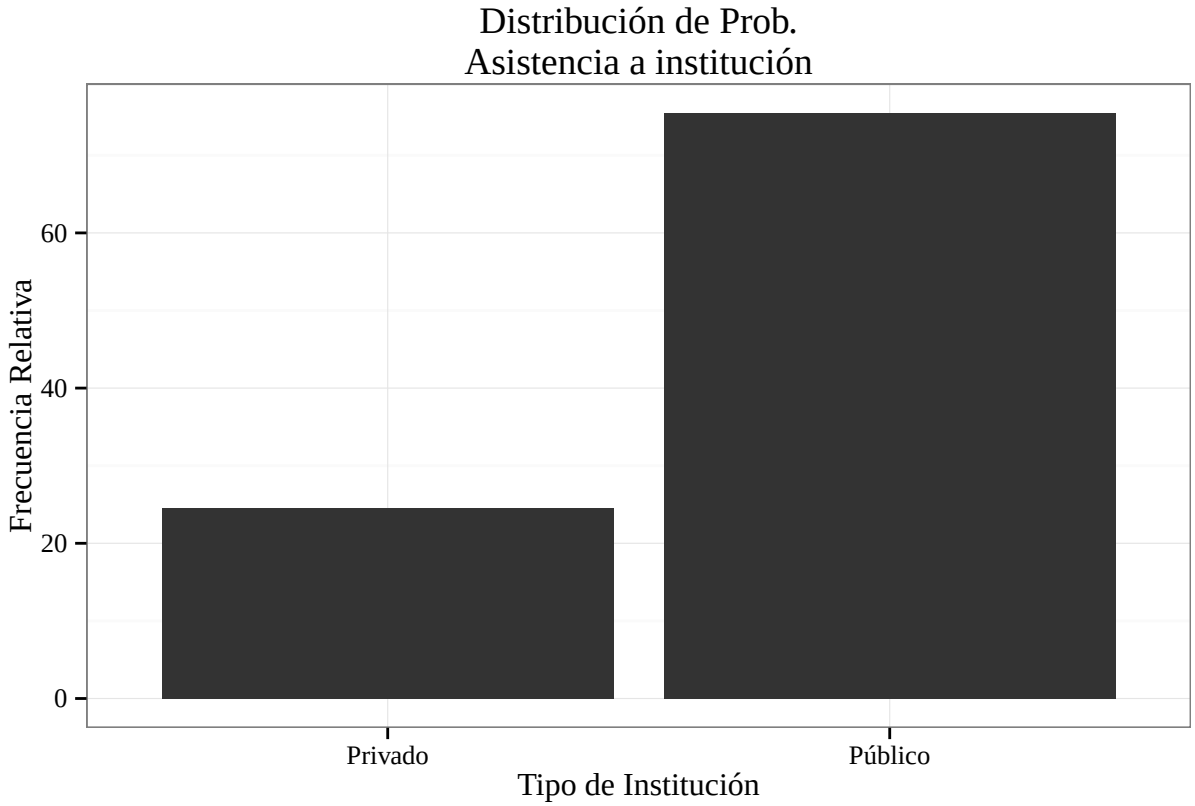
- Aplica sólo para distribuciones simétricas y con forma de "montaña"
- Aproximadamente el 68% de los valores está ubicado dentro de 1 desviación estándar
- Aproximadamente el 95% de los valores está ubicado dentro de 2 desviaciones estándar
- Aproximadamente el 99.7% de los valores está ubicado dentro de 3 desviaciones estándar

Dado que en nuestro caso la media es igual a la mediana, la distribución aparentemente es simétrica.

De acuerdo con la regla empírica, el 68% de los registros de tiempo debe estar ubicado dentro de una desviación estándar, es decir, entre 1.3 y 4.6 horas al día. El 95% de los registros está entre dos desviaciones, es decir, entre -0.3 y 6.3 horas al día. Y el 99.7% entre -2 y 7.9 horas al día. Según lo anterior, un registro de 7 horas, si bien es alto, se encuentra dentro de los rangos "normales" según la distribución observada.

6. (Probabilidad y variables aleatorias) Grafique la distribución de probabilidad de la asistencia a una institución pública o privada.

```
inst <- summarize(group_by(enut2,P6175), estudiantes=n())
inst <- mutate(inst, estudiantes_per=(estudiantes/dim(enut2)[1])*100)
p <- ggplot(inst, aes(x=P6175, y=estudiantes_per))
p + geom_bar(stat="identity") +
        labs(title="Distribución de Prob.\nAsistencia a institución") +
        labs(y="Frecuencia Relativa", x="Tipo de Institución") +
        theme_bw(base_family="Times", base_size=12)
```



7. (Probabilidad) ¿Cuál es $P(P/E)$?, donde P es igual a asistir a una institución pública y E es igual a ser un estudiante que asegura haber estudiado desde su casa durante el día de referencia?

La probabilidad que buscamos es la de pertenecer a una IE pública, dado que el individuo afirma haber hecho tareas o estudiado algún tiempo durante el día de referencia. Por definición, nuestra base de datos sólo incuye estudiantes que afirman haber realizado tareas en su casa, por lo que la **condición** de la probabilidad condicional ya se encuentra implícita en la distribución de probabilidad graficada en la pregunta anterior. De esta manera, la probabilidad de que un estudiante (que reporta estudiar desde su casa) asista a una institución pública es el 75.4%.

8. (Regla de Bayes -Opcional-) Empleando la respuesta anterior, calcule $P(E/P)$. (Ayuda: 18.844 estudiantes reportan haber estudiado desde la casa durante el día de referencia. El total de estudiantes encuestados es 40.753. De los estudiantes que reportan no haber estudiando en la vivienda, 16.381 asisten a una IE Pública).

Según la regla de Bayes:

$$P(E/P) = \frac{P(E)P(P/E)}{P(E)P(P/E) + P(E^c)P(P/E^c)}$$

Donde ya conocemos $P(P/E)$ la cual es igual a 75.4% (Probabilidad de asistir a IE pública, dado que el estudiante reporta haber hecho alguna labor académica fuera de la jornada escolar durante el día de referencia, desde su casa). La otra probabilidad que necesitamos es $P(E)$ la cual es la probabilida de que el estudiante estudie o haga otras labores académicas desde su vivienda. Esta probabilidad la podemos calcular fácilmente a partir de la información brindada como ayuda en la pregunta. Si tenemos 40.753 estudiantes encuestados, de los cuales 18.844 reportan estudiar desde sus viviendas, la probabilidad en mención es entonces igual a 18.844/40.753 o 46%. Por otro lado, $P(E^c)$ es la probabilidad del complemento del evento E, es decir, 1-0.46, o 54%. Finalmente, necesitamos calcular $P(P/E^c)$. Esta es la probabilidad de pertenecer a una IE pública dado que el estudiante reporta no haber estudiado desde su vivienda. Sabemos que en total hay 40.753 estudiantes, de los cuales 18.844 reportan estudiar en casa. Por lo tanto, la diferencia (21.909) reportan no haber estudiado en casa. Según la información, de este últmo grupo 16.381 asisten a una IE pública. Así, $P(P/E^c) = 16.381/21.909 = 0.75$, o 75%. Con estas cuatro piezas de información podemos calcular la probabilidad condicional requerida.

$$P(E/P) = \frac{(0.46 * 0.75)}{(0.46 * 0.75) + (0.54 * 0.75)} = 0.46$$

De esta manera, la probabilidad de estudiar desde la vivienda, dado que el individuo asiste a una IE pública es igual a 46%.

10. (Distribución Normal Estandarizada) ¿Cuál es la probabilidad de que un estudiante que atiende a una IE en el nivel de primaria, dedique más de 2 horas al día estudiando y haciendo tareas en su vivienda? (Chequee primero la normalidad de la distribución)

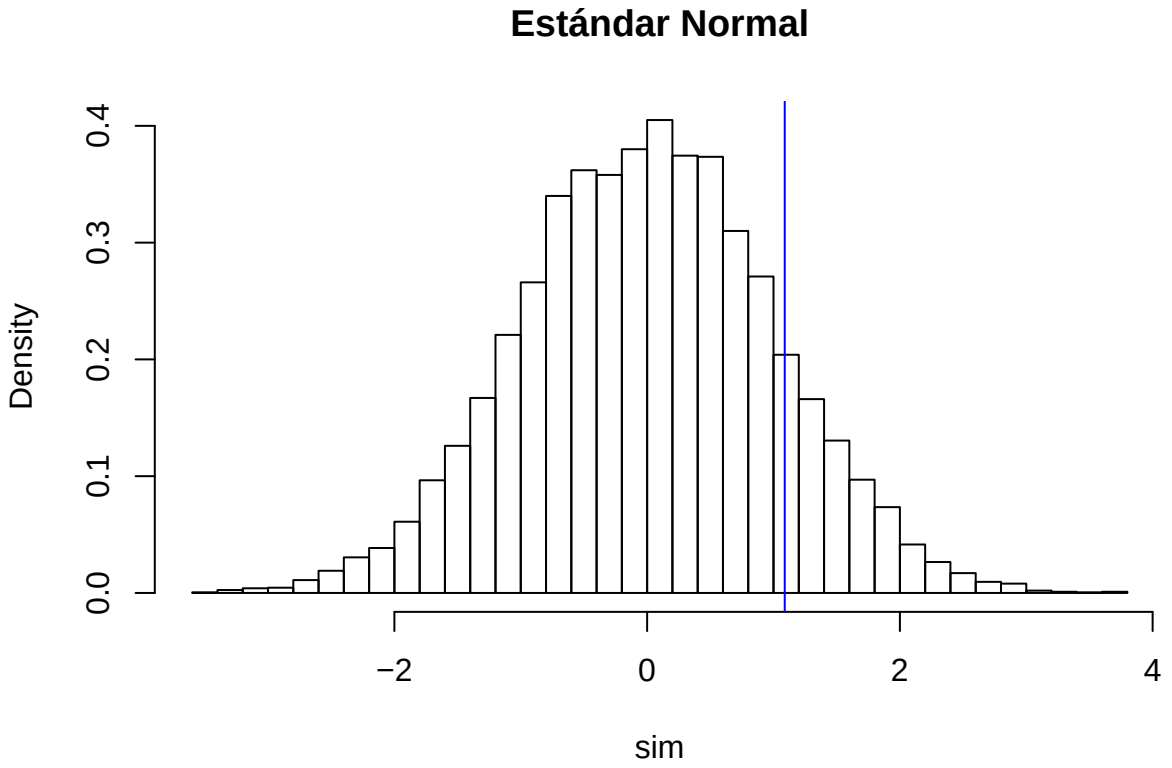
Para obtener dicha probabilidad, es necesario hacer uso de la distribución normal estandarizada. Se procede entonces a calcular el valor z, empleando la media y la desviación estándar. Note que la desviación estándar calculada por R corresponde a la muestral, por lo que es necesario ajustar el cálculo para obtener la poblacional. No obstante, dado el tamaño de la muestra (más de 7000), la diferencia entre las dos desviaciones estándar no va a ser significativa.

```
mu <- sd(enut2$tiempoCasa[enut2$P1158S1=="Primaria"],na.rm=T)
sigma <- sd(enut2$tiempoCasa[enut2$P1158S1=="Primaria"],na.rm=T)
sigma_pob <- sigma*sqrt((table(enut2$P1158S1)[["Primaria"]]-1)/
                                (table(enut2$P1158S1)[["Primaria"]]))
```

Con una media poblacional igual a 0.96 y una desviación estándar poblacional igual a 0.96, calculamos el valor z:

```
z <- (2-mu)/sigma_pob
```

De esta manera, lo que buscamos es $P(z > 1.09)$. Gráficamente:



Lo que buscamos es el area debajo de la curva a la derecha de la linea azul. Según la tabla, esta probabilidad es igual a 0.1381214. De esta manera, la probabilidad de que un estudiante del nivel primaria estudie más de dos horas al día es no más del 14%.