Estadística básica con R

Olivier Nuñez

2019-02-22

Índice

1. Preámbulo

Estadística descriptiva versus Estadística Inferencial

- La *Estadística descriptiva* aporta las técnicas para resumir y presentar la información incluida en una muestra.
- Sin embargo, rara vez nos interesa la muestra como tal, sino que interesa por su capacidad para aportar información acerca de una situación más general (el conjunto de la población) o nueva (futuro).
- La *Estadística Inferencial* aporta las técnicas para evaluar el error de una decisión/predicción sacada a partir de una muestra.

Error en la decisión

se puede distinguir dos tipos de errores en una decisión tomada a partir de una muestra:

1. El **error sitemático** (o de diseño), debido a una falta de representatividad de la muestra.
2. El **error muestral** que corresponde a la variabilidad de la decisión de una muestra a otra.

A continuación veremos mediante *tecnicas de muestreo* como conseguir evitar el primer error y como evaluar el segundo error en la *inferencia sobre medias*.

2. Muestreo

Aplicaciones

- **Estudios observacionales** (Muestreo versus Censo): en muchos casos, es demasiado caro y poco factible obtener datos de cada unidad de la población de estudio.
- **Evaluación del error** (bootstrap, cross-validation): evaluación empírica de la precisión de los métodos de inferencia y predicción.

¿Qué es el muestreo?

Algunas definiciones:

- *"Sampling consists of selecting some part of a population to observe so that one may estimate something about the whole population."*
  (**S. Thompson, 1992**).
- *"We are all accustomed to the idea of sampling in everyday life. The housewife visually samples the quality of the fruit she intends to buy. (...) If the greengrocer puts the best on display and sells us inferior qualities, we protest at the biased sample or change our supplier. (...) The notion of bias is not long the notion of sampling itself."*
  (**A. Stuart, 1983**).

Objetivo del muestreo

La Representatividad

- Evitar el sesgo de selección (exclusión sistema de parte de la población)
- Maximizar la representatividad (muestra = población a pequeña escala).
- Reducir los costes de recogida de datos.
- Controlar el **error muestral**.

Error muestral

Variación entre muestras

- Es consecuencia de nuestra observación parcial de la población.
- Corresponde a la variación de las conclusiones entre muestras.
- Es un error de naturaleza aleatoria que decrece a medida que aumenta el tamaño muestral.
- No debe ser confundido con el **sesgo** (o **error sistemático**)

Error sistemático

Sesgo de selección y de información

- Distorsión en las conclusiones que ocurre de manera sistemática (afecta a todas la muestras).
- Se puede clasificar según el sesgo ocurre a nivel del muestreo o en la medición:
  - **Sesgo de selección**: ocurre cuando la selección de los individuos está condicionada por la característica que queremos medir.
  - **Sesgo de información**: sesgo que ocurre en la medición de la característica de interés. Sesgo de memoria, efecto del entrevistador o del cuestionario, ...
- Dos ejemplos de sesgo de selección:

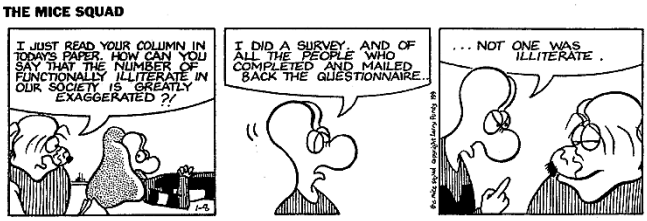


Figura 1: No respuesta

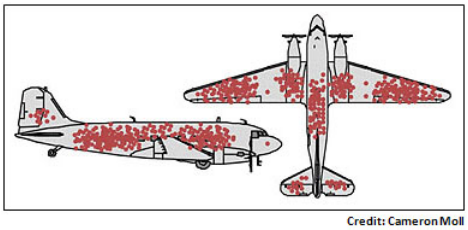


Figura 2: Durante la Segunda Guerra Mundial, el estadístico Abraham Wald fue invitado a ayudar a los británicos a decidir dónde añadir armadura a sus bombarderos. Después de analizar los aviones que volvieron, recomendó reforzar los sitios donde no hubo impacto!

- Una selección **aleatoria** de las unidades permite evitar el sesgo de selección.

Muestreo aleatorio simple

Es el diseño de muestreo más sencillo y se caracteriza por el hecho de que cada posible muestra de n unidades tiene la misma probabilidad de ser seleccionada.

Implementación

Se obtiene una muestra aleatoria simple de la siguiente manera:

1. Asignar un número de 1 a N a cada unidad de la población elegible.
2. Elegir n de estos números mediante el uso de algún proceso aleatorio (tablas o generador de números aleatorios)
3. Las unidades correspondientes a los números elegidos se toman como muestra.

```
##### Implementación del muestero aleatorio simple en R #####
state.name # la lista de estados americanos como Población
```

```
##  [1] "Alabama"        "Alaska"         "Arizona"        "Arkansas"
##  [5] "California"     "Colorado"       "Connecticut"    "Delaware"
##  [9] "Florida"        "Georgia"        "Hawaii"         "Idaho"
## [13] "Illinois"       "Indiana"        "Iowa"           "Kansas"
## [17] "Kentucky"       "Louisiana"      "Maine"          "Maryland"
## [21] "Massachusetts"  "Michigan"       "Minnesota"      "Mississippi"
## [25] "Missouri"       "Montana"        "Nebraska"       "Nevada"
## [29] "New Hampshire"  "New Jersey"     "New Mexico"     "New York"
## [33] "North Carolina" "North Dakota"   "Ohio"           "Oklahoma"
## [37] "Oregon"         "Pennsylvania"   "Rhode Island"   "South Carolina"
## [41] "South Dakota"   "Tennessee"      "Texas"          "Utah"
## [45] "Vermont"        "Virginia"       "Washington"     "West Virginia"
## [49] "Wisconsin"      "Wyoming"
```

```r
sample(state.name,size=5) #muestra aleatoria simple de 5 estados
```

```
## [1] "Oklahoma"     "Rhode Island" "Louisiana"    "Arizona"
## [5] "Montana"
```

```r
#sample(state.name,size=200) ## Error!!
muestra=sample(state.name,size=200,replace=TRUE) #muestreo con reposición
table(muestra)
```

```
## muestra
##        Alabama         Alaska        Arizona       Arkansas     California
##              6              5              3              6              3
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              8              2              5              3              6
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              2              3              9              3              1
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              4              6              3              2              2
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              5              4              4              1              3
##       Nebraska         Nevada  New Hampshire     New Jersey     New Mexico
##              5              4              1              3              5
##       New York North Carolina   North Dakota           Ohio       Oklahoma
##              7              2              7              5              8
##         Oregon   Pennsylvania   Rhode Island South Carolina      Tennessee
##              3              3              4              6              6
##          Texas           Utah        Vermont       Virginia     Washington
##              5              6              2              3              1
##  West Virginia      Wisconsin        Wyoming
##             10              3              2
```

Ventajas y limitaciones

- Procedimiento **equiprobabilístico**: todas las unidades de la población tienen la misma probabilidad de ser elegidas. Con lo cual, media y proporciones se estiman directamente mediante su equivalente muestral.
- Pero, es en general poco eficiente (algunas muestras pueden ser poco representativas).

Muestreo estratificado

- Se puede mejorar notablemente la representatividad de la muestra estratificando el muestreo.
- Si la afijación del tamaño muestral en cada estrato es proporcioinal a su tamaño, obtendremos una muestra equiprobabilistica.

```r
##### Muestreo de una base de datos #####
require(tidyverse)
require(readxl) #para leer fichero de excel en R
## Precios de la  gasolina 98 en España
gasolineras = read_excel("data/gasolineras.xls",skip=3) #Fuente: https://geoportalgasolineras.es
gasolineras
```

```
## # A tibble: 6,495 x 10
##    Provincia Localidad Dirección Margen `Toma de datos` Precio Rótulo
##    <chr>     <chr>     <chr>     <chr>  <chr>            <dbl> <chr>
##  1 SANTA CR~ GRANADIL~ CARRETER~ D      09/01/2019 11:~  0.928 PCAN
```

```
##  2 SANTA CR~ CUESTA, ~ AVENIDA ~ D      09/01/2019 09:~  0.934 PCAN
##  3 SANTA CR~ LA LAGUNA PLAZA SA~ N      09/01/2019 06:~  0.935 TGAS
##  4 SANTA CR~ LA LAGUNA GLORIETA~ N      09/01/2019 13:~  0.935 DISA ~
##  5 SANTA CR~ BALDIOS,~ CARRETER~ D      09/01/2019 00:~  0.935 TGAS ~
##  6 SANTA CR~ SAN CRIS~ AVENIDA ~ N      09/01/2019 13:~  0.935 SHELL~
##  7 SANTA CR~ SANTA CR~ CALLE SU~ D      09/01/2019 10:~  0.935 CANAR~
##  8 SANTA CR~ GUANCHA ~ CARRETER~ D      08/01/2019 06:~  0.937 TGAS ~
##  9 SANTA CR~ GUANCHA ~ AVENIDA ~ N      09/01/2019 11:~  0.937 PCAN
## 10 SANTA CR~ REALEJOS~ CARRETER~ N      10/01/2019 12:~  0.937 PCAN ~
## # ... with 6,485 more rows, and 3 more variables: `Tipo venta` <chr>,
## #   Rem. <chr>, Horario <chr>
```

```
summary(gasolineras$Precio) #resumen de la distribución del precio
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.928   1.313   1.339   1.327   1.365   1.549
```

```
n=250 # tamaño muestral
muestra <- gasolineras %>% sample_n(size=n)
summary(muestra$Precio) #distribución del precio en la muestra
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.935   1.310   1.339   1.327   1.359   1.539
```

```
#### Muestreo estratificado (proporcional) ####
muestra <- gasolineras %>% group_by(Provincia) %>% sample_frac(.1) # 1% en cada provincia
muestra %>% summarise(media=mean(Precio),n = n())
```

```
## # A tibble: 51 x 3
##    Provincia       media     n
##    <chr>           <dbl> <int>
##  1 ÁLAVA            1.33     4
##  2 ALBACETE         1.31     9
##  3 ALICANTE         1.35    29
##  4 ALMERÍA          1.33    10
##  5 ASTURIAS         1.36    14
##  6 ÁVILA            1.34     4
##  7 BADAJOZ          1.31     9
##  8 BALEARS (ILLES)  1.39    15
##  9 BARCELONA        1.36    54
## 10 BURGOS           1.33     8
## # ... with 41 more rows
```

Ejercicio 2.1 Estratificar la muestra por fecha de toma de datos (no tomar en cuenta la hora! Para ello, Utilizar la función `separate`).

Muestreo de una población teorica

- Existen modelos (o arquetipos) de distribución que simplifican considerablemente el análisis de los fenómenos aleatorios.
- Estos modelos de distribución descansan sobre dos hipótesis básicas:
  - **Homogeneidad** (los sujetos de la población tienen la misma probabilidad de enfermar)
  - **Independencia** (la enfermedad ocurre de manera independiente en cada sujeto)
- A pesar de que en practica estas hipótesis no se cumplen de manera estricta, estos modelos
  - proporcionan en muchas ocasiones una buena aproximación.

- permiten definir una marco teórico (o población teórica) a contrastar con la realidad.

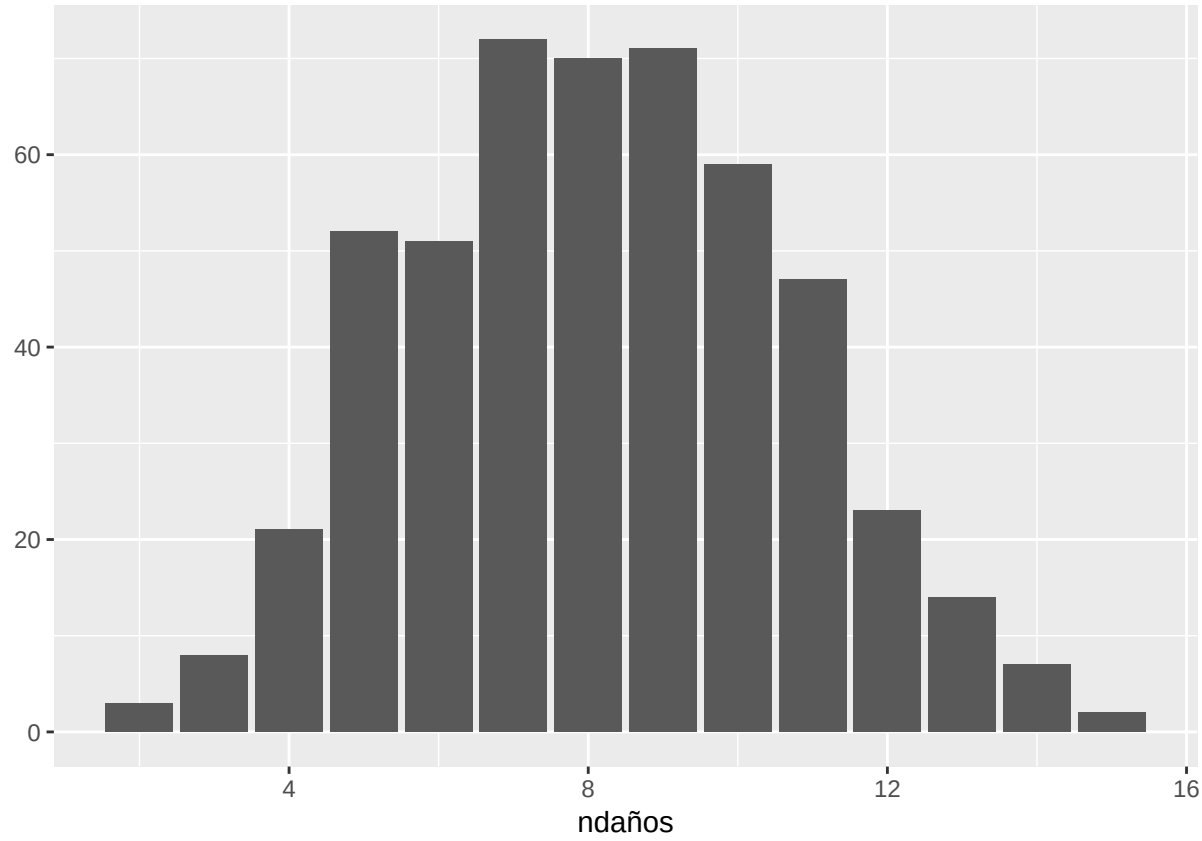
Distribución Binomial

Así, podríamos suponer que todas las manzanas que llegan a un supermercado tienen la misma probabilidad p de dañarse después de un día en cámara fría y que estos daños ocurren de manera independientes (¡dudoso!). Entonces, el número de frutas dañadas en una caja de n manzanas se distribuye como una distribución binomial $B(n, p)$.

```
n=40
p=.2 #probabilidad de daño igual al 20%
ndaños=rbinom(500,n,p) #simulación del número de frutas dañadas en 500 cajas de n manzanas
ndaños
```

```
##   [1] 13 11  7  4  7  8  9  5 13  6  9  6  8  4 10  7 11 10  6  6  6  9  7
##  [24]  7  6  6  5  8  9  5  4  7  7  8  6  7 10  7  7 10  5  6  5  4  5  9
##  [47]  5  9  6  9  8  5 10  9 14  6  7  5 13  7  6  5 11 10  6 10  7  9 11
##  [70]  8  7  2 15 10  8  4 10  7  6  8 12 10  6 10  9  9  8  8  7  7 11  5
##  [93]  4  6 11  4 10 12  3  8  6  5 10 11 10  8 13 14  8  9  9 11  7 11  5
## [116] 10  7  5  8 12 11 11 10 13  8  8  7  6  9 12 11  7  9  6  7  5 10  7
## [139]  8 11 11 11  8  6  8  7 13  7  5 10  9  4  7  7  7  8  6 10  6  8 10
## [162] 10 12  4  7  8 13 10  7 11  8  8  5 10 12  7  7 11  7  8  9  7 10  3
## [185] 10 11  7  7 12  9  6  8  9 14  8 11  9  5 13 13 10  8 11  6 10  8  7
## [208]  7  4  8  7 11  3 10  4  8 11 12  9 11  7  6  4  7  4  7  6  8 11 10
## [231]  6  5 10  7  9 10  8  5  7 14 12 11  7  8  9 14  7  9  5 13 12  5  9
## [254] 10  7 10  9  8  9  7 13  8 10 14  5 10 12 10  8  8  9  6 11 10  9  9
## [277]  8  9  9  9 12  5  8  8 10  5  5  6 11 10 13  9  5  7 11  9  5  7  7
## [300]  4  8  6  4 12 10  8  7  5  7  7  9 11  6 10  7 10  9  4  6  6  8  6
## [323] 13  8 11  6  3 11  9  9  9  5 10  5  6 10  9 11  8 12  5  4  7 12  7
## [346] 10  8  6  7 11  5  8  6  9 12 11  8  7  6  3  9  4 11  9 11  8  9  9
## [369]  9  4 11 10 11  5  5  9  8  5  9  8  6  6  6  9  2  5 12  8  5  7  9
## [392]  9  8  6  4  6 15  5  9 11  9  8  8 10  7 12  5  8  9  9 13 10  6  9
## [415]  8 10 12 11  8  7  3  9  7  9  2 10 10  9  8 10  9  9  8  5 11  7 11
## [438]  9 10 10  7  9  5 12  9 10 12  9  7  8  9  6  9  3  5  5  5  4  8  8
## [461]  6  8  6  6  3  5  8 10 10 11 12  7  9  5 11  5  8  9  5 11  6  6  5
## [484]  7 12  7 11 10 14  9  8 10  7 11  5  5  8  7  7  8
```

```
qplot(ndaños,geom="bar") #distribución empirica del número de frutas dañadas
```



Distribución Normal

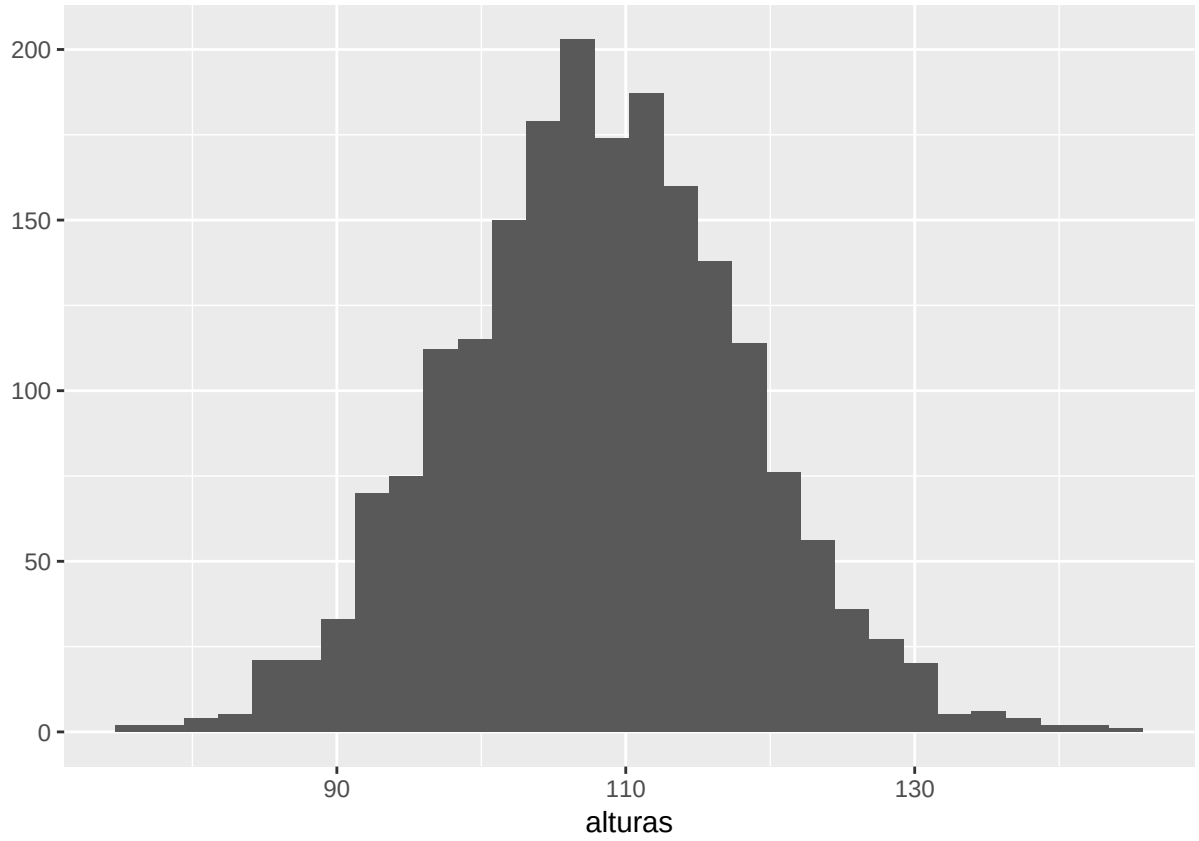
La distribución normal (o gaussiana) es la más utilizada para modelizar las variaciones de una variable continua. Razones principales de este éxito:

- Su simplicidad (sólo depende de la media y de la desviación típica de la variable)
- El **Teorema Central del Limite** que afirma que el promedio de un gran numero de observaciones independientes tiene aproximadamente un comportamiento normal.
- Apropiada para modelizar fenómenos naturales sometidos a numerosos efectos aleatorios de pequeña magnitud (Ej.:errores de medida y medidas antropometricas en población homogénea).

```
####### Altura de niñas españolas de 5 años
mu=108 #altura media de las niñas (en cm)
sigma=10 #variación promedia de la altura entre niñas (en cm)
alturas=rnorm(2000,mu,sigma) #alturas de 2000 niñas seleccionadas al azar en la población
summary(alturas)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   76.16  101.27  108.06  108.09  114.81  144.91
```

```
qplot(alturas) #histograma de la distribución de la altura
```




```
mean(alturas>115) #proporción muestral de niñas con una altura superior a 115 cm
```

```
## [1] 0.243
```

```
pnorm(115,mu,sigma,FALSE) #probabilidad teorica (poblacional) de que una niña mida más de 115 cm
```

```
## [1] 0.2419637
```

Ejercicio 2.2 En una fabrica, la maquina que rellena las botellas de coca-cola está calibrada para que en promedio tengan un volumen $\mu = 330$ ml. Se sabe que la maquina tiene una precisión $\sigma = 5$ ml (desviación típica del volumen de las botellas). Supongamos que en una muestra de 25 botellas sacada de la producción observamos que en promedio el volumen es igual a $\bar{y} = 327{,}5$ ml. ¿Podemos concluir que la maquina se ha desajustado? Asumiendo que la maquina tiene errores normales, realizar un experimento de simulación para contrastar esta hipótesis.

3. Inferencia sobre medias

En estadística, la *Inferencia* es el conjunto de técnicas para extraer conclusiones sobre la población a partir de una muestra (metáfora de la sopa).

Test sobre una media

Planteamiento

Queremos contrastar si la media poblacional $\mu = 330$ ml a partir de la observación del volumen medio $\overline{y} = 327{,}5$ ml en una muestra de 25 botellas.

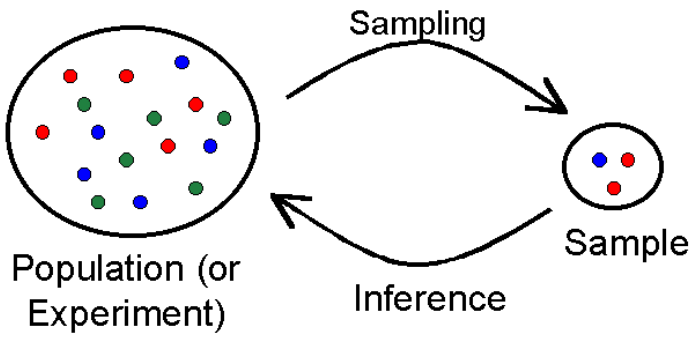


Figura 3: Muestreo versus Inferencia

Hasta que punto podemos confiar en el valor observado del volumen medio? O en otras palabras, ¿cómo varia $\overline{y}$ de una muestra a otra?

Teorema central del limite

- El error estándar (o error de estimación) de $\overline{y}$ es proporcional a la desviación típica σ de los datos:

$$\text{se}\left(\overline{y}\right) = \frac{\sigma}{\sqrt{n}}$$

- Además, cuando n es "grande"", tenemos que

$$\overline{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

  donde $N\left(\mu, \sigma\right)$ denota la distribución normal con media μ y desviación típica σ.

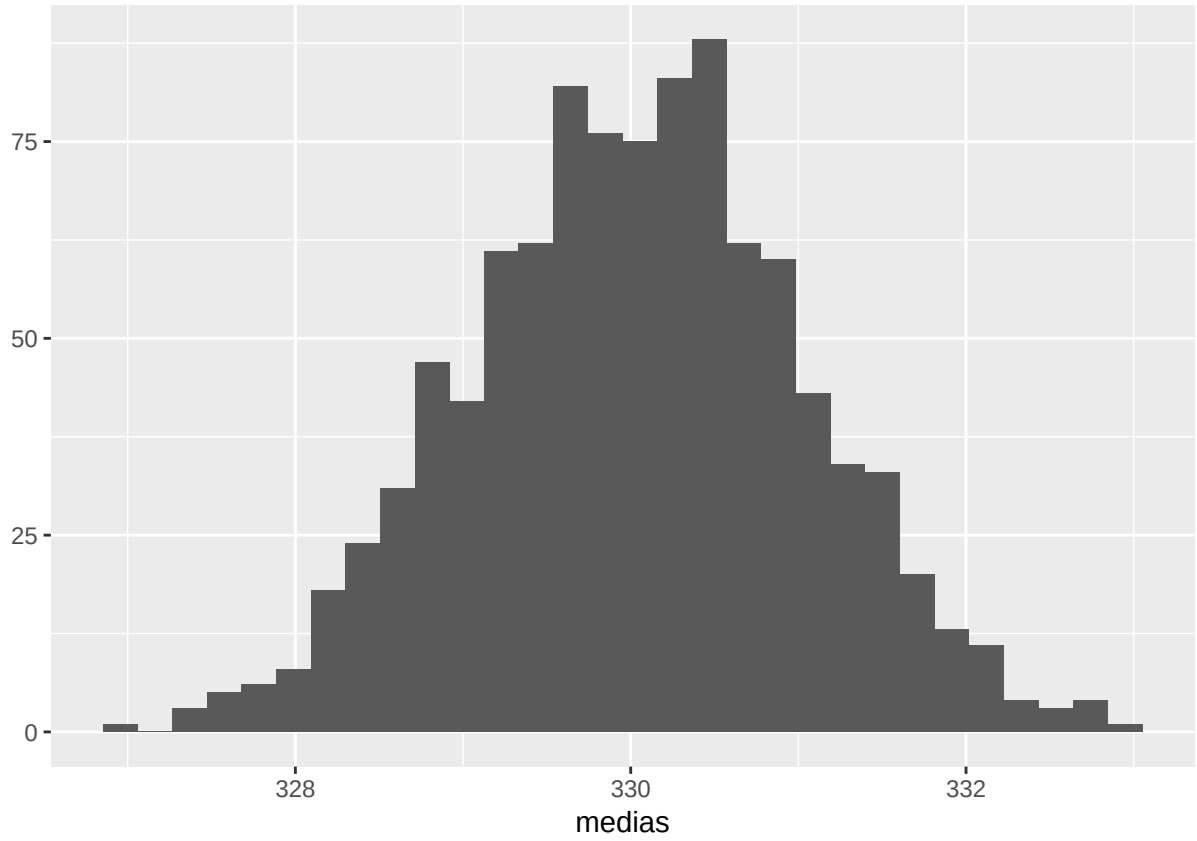
```
## Distribución de la media muestral en muestras de 25 botellas
K=1000 #número de muestras
medias=replicate(K,mean(rnorm(25,330,5)))
summary(medias) #distribución de las medias
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   327.0   329.3   330.0   330.0   330.7   333.0
```

```
sd(medias) #error estándar de la media muestral (teoricamente = sigma/sqrt(n)=5/sqrt(25)=1)
```

```
## [1] 0.9978727
```

```
qplot(medias)
```



Una horquilla para la media poblacional

El *intervalo de confianza* es una horquilla que permite reflejar la precisión en la estimación.

Típicamente, un intervalo de confianza (del 95 %) de la media poblacional tiene la forma:

$$\overline{y} \pm 1{,}96 \times \text{se}\,(\overline{y})$$

donde 1,96 corresponde al cuantil 97.5 % de la distribución normal $N\,(0, 1)$.

La confianza del intervalo corresponde a la proporción de muestras (entre todas la muestras posibles de tamaño n) para las cuales el intervalo así construido contiene la media poblacional μ.

Test de Student

Sin embargo, en practica la desviación típica σ es desconocida y requiere ser estimada para construir el intervalo de confianza. Al sustituir σ por su estimación en la muestra $s =$`sd(y)` en la expresión del error estándar de $\overline{y}$, se obtiene el siguiente intervalo de confianza (del 95 %):

$$\overline{y} \pm t_{(n-1),97{,}5\,\%} \times \frac{s}{\sqrt{n}}$$

donde $t_{(n-1),97{,}5\,\%}$ es el cuantil 97.5 % de la distribución de student con $n - 1$ grados de libertad.

> **Nota**: Cabe mencionar que si $n \simeq 200$, la distribución de Student $t_{(n-1)}$ es muy próxima a la normal y tenemos que $t_{(n-1),97{,}5\,\%} \simeq 1{,}96$
>
> **¿Cuando usar el test de Student?** Si las observaciones son independientes y casi normales. Si estas condiciones no se cumplen se puede utilizar el test (no-parametrico) de Wilcoxon (más detalles en `?wilcox.test`).

```r
### Utilización del Test de Student en R ####
muestra=c(326.4,324.5,319.2,332.9,328.9,331.3,327.4,323.6,322.7,323.9,325.3,329.6,338,328.8,327.6,333,3
t.test(muestra,mu=330) #contrasta si mu=330
```

```
##
##  One Sample t-test
##
## data:  muestra
## t = -2.4522, df = 24, p-value = 0.02185
## alternative hypothesis: true mean is not equal to 330
## 95 percent confidence interval:
##  325.3959 329.6041
## sample estimates:
## mean of x
##     327.5
```

```r
t.test(muestra,mu=330,alternative="less") #contrasta si mu>330
```

```
##
##  One Sample t-test
##
## data:  muestra
## t = -2.4522, df = 24, p-value = 0.01093
## alternative hypothesis: true mean is less than 330
## 95 percent confidence interval:
##      -Inf 329.2442
## sample estimates:
## mean of x
##     327.5
```

```r
mean(muestra)+sd(muestra)/sqrt(length(muestra))*qt(.95,24) #cálculo a mano del limite superior del IC (
```

```
## [1] 329.2442
```

El p-valor

- Una manera de medir la credibilidad de la hipótesis $H_0 : \mu \geq 330$ ml, consiste en calcular la probabilidad (en caso de que H_0 fuese cierta) de observar algo que discrepe más de esta hipótesis que lo que observamos.
- Si es inferior al nivel de significación establecido ($\alpha = 5\,\%$), rechazamos la hipótesis nula H_0.
- En el ejemplo anterior observamos $\overline{y} = 327{,}5$ ml, por lo tanto la probabilidad de observar algo aún más alejado de la hipótesis $H_0 : \mu \geq 330$ sería:

$$p = P\left(\overline{y} < 327{,}5 | H_0 \text{ cierta}\right) = \mathsf{pt}\left(-2{,}45, 24\right) \simeq 1{,}1\,\%$$

Comparando medias

En el ejemplo anterior, comparamos la población de estudio con una población de referencia o teórica (las botellas producidas con las botellas que esperamos observar si la maquina está bien calibrada).

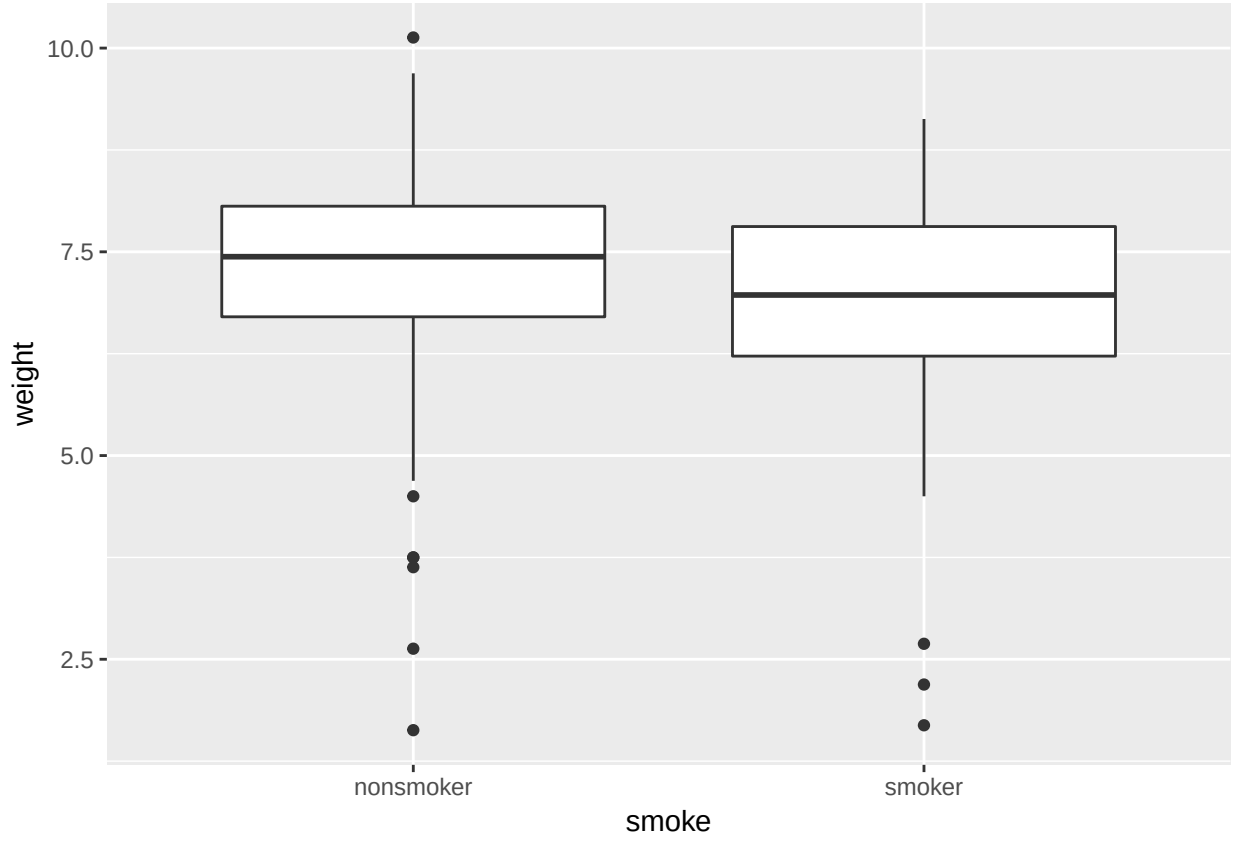
Supongamos que queramos ahora comparar dos poblaciones en estudio (chicas versus chicos, grupo de tratamiento versus grupo placebo, ...). En este caso, observaríamos dos muestras a partir de las cuales pretendemos decidir si hay diferencias entre las dos poblaciones de donde fueron sacadas.

A partir de la base de datos `births` de la librería `openintro` se pretende contrastar el efecto del habito de la madre con el tabaco sobre el peso del neonato.

```r
require(openintro)
glimpse(births) #?births para más detalles
```

```
## Observations: 150
## Variables: 9
## $ fAge      <int> 31, 34, 36, 41, 42, 37, 35, 28, 22, 36, 27, 35, 25, ...
## $ mAge      <int> 30, 36, 35, 40, 37, 28, 35, 21, 20, 25, 19, 34, 19, ...
## $ weeks     <int> 39, 39, 40, 40, 40, 40, 28, 35, 32, 40, 32, 40, 41, ...
## $ premature <fct> full term, full term, full term, full term, full ter...
## $ visits    <int> 13, 5, 12, 13, NA, 12, 6, 9, 5, 13, 5, 15, 13, 10, 1...
## $ gained    <int> 1, 35, 29, 30, 10, 35, 29, 15, 40, 34, 32, 20, 47, 2...
## $ weight    <dbl> 6.88, 7.69, 8.88, 9.00, 7.94, 8.25, 1.63, 5.50, 2.69...
## $ sexBaby   <fct> male, male, male, female, male, male, female, female...
## $ smoke     <fct> smoker, nonsmoker, nonsmoker, nonsmoker, nonsmoker, ...
```

```r
qplot(smoke,weight,data=births,geom="boxplot")
```



Se aprecia una diferencia entre los dos grupos, siendo los hijos de madres no fumadoras los que suelen tener mayor peso al nacer. Sin embargo, esta diferencia podría ser fruto del azar (del muestreo). Para evaluar si esta diferencia es significativo se puede aplicar el test de Student:

```r
t.test(weight~smoke,data=births)
```

```
##
##  Welch Two Sample t-test
##
## data:  weight by smoke
```

```
## t = 1.4967, df = 89.277, p-value = 0.138
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1311663  0.9321663
## sample estimates:
## mean in group nonsmoker    mean in group smoker
##                  7.1795                  6.7790
```

Para un nivel de significación del 5 %, la diferencia observada no es significativa (el p-valor $p = 9{,}31\,\% > 5\,\%$). Sin embargo, se trata de un efecto importante:0,40 libras, es decir una diferencia promedia de unos 200 gramos entre los dos grupos de neonatos, y merece mayor investigación. Acordarse del aforismo "Absence of evidence is not evidence of absence!".

Ejercicio 3.1 Muestrear la base de datos `babies` del paquete `openintro` de manera que los tamaños muestrales de los dos grupos (madres fumadoras y no fumadoras; quitar valores ausentes) sean sucesivamente $n = 10, 50, 100, 250$. Contrastar el efecto del tabaco sobre el peso del neonato para estos distintos tamaños muestrales.

> **Nota**
>
> Por defecto, se supone que las dos muestras son independientes (la opción `paired=FALSE`en la función `t.test`). Pero, si los dos grupos corresponden a medidas realizadas sobre los mismos individuos, esta condición no se cumple y el test de Student se ha de utilizar con la opción `paired=TRUE`.

```
######### Simulación de un ejemplo con muestras dependientes
nn=100
col0=rnorm(nn,200,50) #valor basal de colesterol
efecto = rnorm(nn,-10, 2) #efecto del tratamiento con estatina
col1=col0 + efecto #valor del colesterol despues del tratamiento
datos=data_frame(antes=col0,despues=col1) %>% gather(visita)

t.test(value~visita,data=datos) #test de Student para muestras independientes
```

```
##
##  Welch Two Sample t-test
##
## data:  value by visita
## t = 1.5607, df = 198, p-value = 0.1202
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.564551 22.028591
## sample estimates:
##   mean in group antes mean in group despues
##               195.0637              185.3317
```

```
t.test(value~visita,data=datos,paired=TRUE) #test de Student para datos emparejados
```

```
##
##  Paired t-test
##
## data:  value by visita
## t = 46.277, df = 99, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   9.314738 10.149301
## sample estimates:
```

```
## mean of the differences
##                 9.73202
```