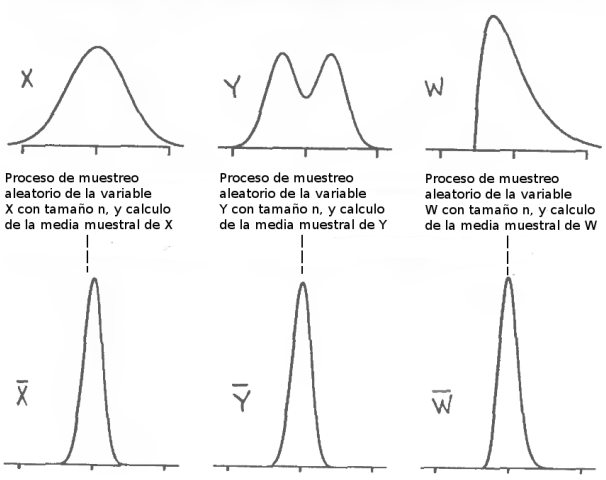
Bases téoricas-aplicadas para inferencia en muestreo estadístico



Autor: **Prof. Christian Salas Eljatib**

E-mail: **cseljatib@gmail.com**

Índice

1. Intervalos de confianza estadístico **2**

1.1. La distribución t 3

1.2. Intervalo de confianza para $\widehat{\mu}$ 6

1.3. Ejemplificando el concepto 7

1.4. Marco de muestreo 9

1.5. Efecto de diferentes niveles confidenciales 12

1.6. Efecto de diferentes tamaños de muestreo 13

2. Teorema central del límite y su relación con muestreo **15**

3. Tamaño muestral **19**

La inferencia hace relación a tratar de sacar conclusiones sobre los modelos estadísticos que hipotizamos en base a la muestra aleatoria con la cual los construimos.

1. Intervalos de confianza estadístico

- Estimación
- Incertidumbre
- Podríamos generar un intervalo[1] estimado para el parámetro, i.e., un intervalo aleatorio dentro del cual el parámetro de interés estaría incluido con un alto grado de éxito.
- Usualmente, un "alto grado de éxito" se denomina "coeficiente de confianza". Ejemplo, cuando el riesgo de fallar es $\alpha = 0{,}1$, el coeficiente de confianza (como porcentaje) es $(1-\alpha)100\,\% = 90$
- pregunta: ¿podemos generar un intervalo que incluya (cubra) el parámetro aproximadamente el 90 % del tiempo? Si.
- Un intervalo de confianza para cualquier estimador de un parámetro θ se calcula como

$$\widehat{\theta} \pm \sqrt{\mathrm{V}(\widehat{\theta})}\; t_{(1-\frac{\alpha}{2},df)}, \tag{1}$$

  donde:

  - $\widehat{\theta}$ es un estimador del parámetro θ;
  - $\mathrm{V}(\widehat{\theta})$ es la varianza del estimador $\widehat{\theta}$;
  - $t_{(1-\frac{\alpha}{2},df)}$ es el $(1-\frac{\alpha}{2})$ percentil de la distribución de t (o t de *Student*) con df grados de libertad, correspondiendo a una probabilidad α de dos colas.

- El error estándar de un estimador es

$$\mathrm{SE}(\widehat{\theta}) = \sqrt{\mathrm{V}(\widehat{\theta})} \tag{2}$$

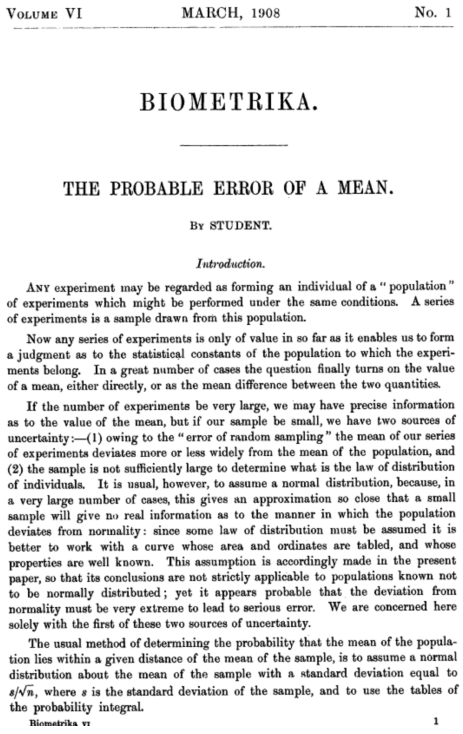
- Entonces el intervalo de confianza se puede también escribir como

$$\widehat{\theta} \pm \mathrm{SE}(\widehat{\theta})\; t_{(1-\frac{\alpha}{2},df)} \tag{3}$$

[1] Un intervalo es simplemente un rango, es decir, contiene dos números, el que indica el valor inferior del rango, y el valor superior del rango.

1.1. La distribución t

- William Gosset, un graduado en Química de Oxford University, con un minor en Matemáticas, contratado a los 23 años (en 1889) para trabajar en la cervecería Guinnes (la más grande del mundo en ese entonces).
- Experimentaban con distintas técnicas de fermentación, cevadas, etc, y costaba detectar si las diferencias eran solo por la variación natural y aun más con muestras tan pequeñas.
- Durante un año "sabático" se fue a estudiar con el Profesor Karl Pearson en el Laboratorio de Biometría de la University of London. Ahí público el siguiente paper: Student (1908).



VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information as to the value of the mean, but if our sample be small, we have two sources of uncertainty:—(1) owing to the "error of random sampling" the mean of our series of experiments deviates more or less widely from the mean of the population, and (2) the sample is not sufficiently large to determine what is the law of distribution of individuals. It is usual, however, to assume a normal distribution, because, in a very large number of cases, this gives an approximation so close that a small sample will give no real information as to the manner in which the population deviates from normality: since some law of distribution must be assumed it is better to work with a curve whose area and ordinates are tabled, and whose properties are well known. This assumption is accordingly made in the present paper, so that its conclusions are not strictly applicable to populations known not to be normally distributed; yet it appears probable that the deviation from normality must be very extreme to lead to serious error. We are concerned here solely with the first of these two sources of uncertainty.

The usual method of determining the probability that the mean of the population lies within a given distance of the mean of the sample, is to assume a normal distribution about the mean of the sample with a standard deviation equal to $s/\sqrt{n}$, where s is the standard deviation of the sample, and to use the tables of the probability integral.

Biometrika VI 1

- Gosset usó el pseudonimo de “Student”.
- pdf basada en la distribución Normal
- la base estadística proviene de dos variables aleatorias, una de ellas distribuyéndose normal estándar y la otra chi-cuadrada.

$$f(y;\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}} \left(1 + \frac{y^2}{\nu}\right)^{-\frac{1}{2}(\nu+1)} \tag{4}$$

donde ν es un parámetro denominado grados de libertad (“df”)

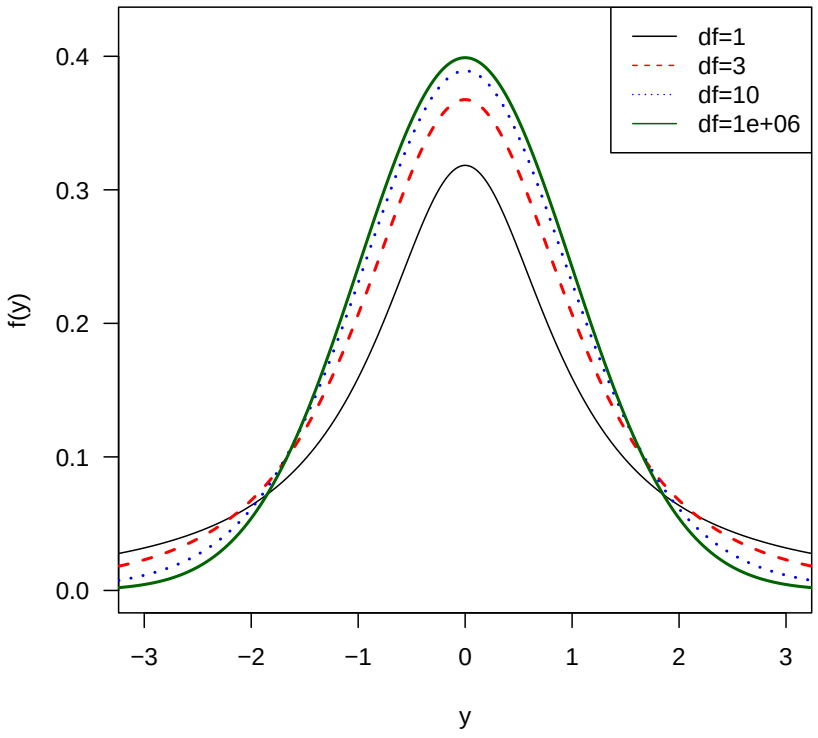
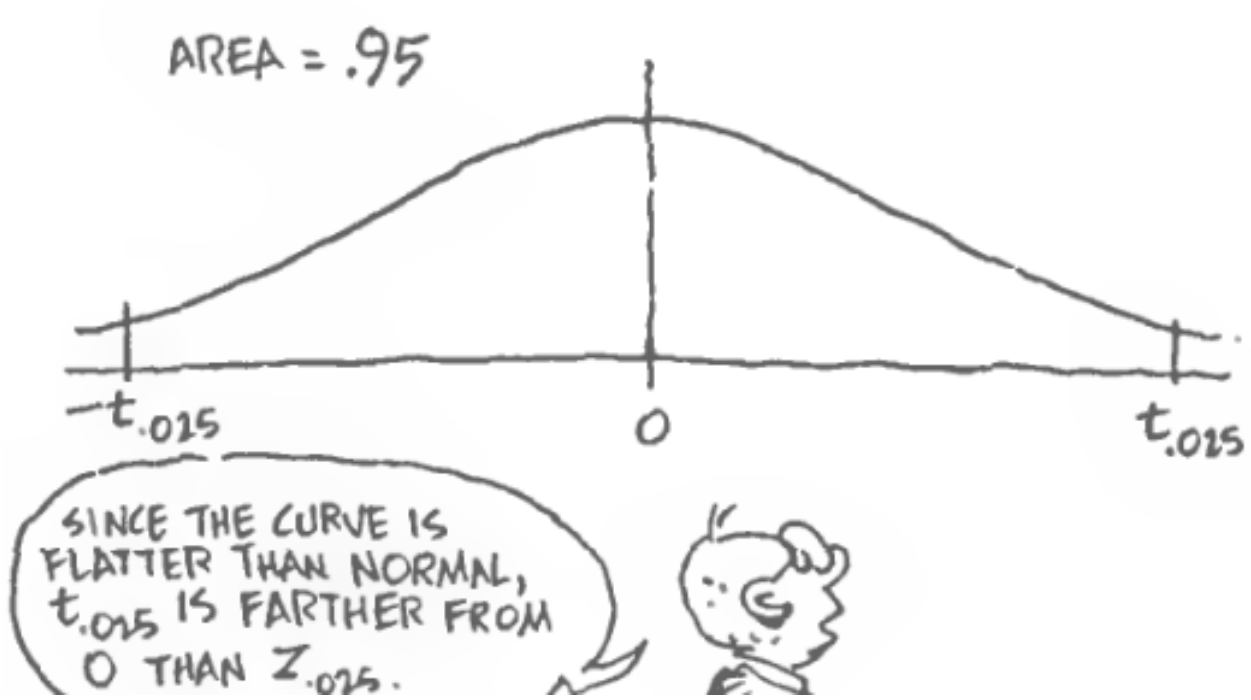


Figura 1: Pdf de t-student con diferentes grados de libertad para la variable aleatoria y. Note que a medida que los grados de libertad aumentan, la distribución se parece cada vez a una distribución normal estándar.

Cuantiles de la distribución t



- Necesitamos el valor de t correspondiente a $n-1$ grados de libertad y un valor de significancia α. ¿Cómo?

  a. Usando valores tabulados ("tablas de t") para valores fijos de grados de libertad $(n-1)$ y de significancia (α)

α-level in each tail	0.10	0.05	0.025	0.005
α-level in both tails	0.20	0.10	0.05	0.01
Degrees of freedom				
1	3.078	6.314	12.706	63.656
2	1.886	2.920	4.303	9.925
3	1.638	2.353	3.182	5.841
4	1.533	2.132	2.776	4.604
5	1.476	2.015	2.571	4.032
6	1.440	1.943	2.447	3.707
7	1.415	1.895	2.365	3.499
8	1.397	1.860	2.306	3.355
9	1.383	1.833	2.262	3.250
10	1.372	1.812	2.228	3.169

  b. Obtenerlo directo desde R usando la función `qt()` como sigue

```
> n<- 10; df<- n-1; conf<- 90
> alpha<- 1-(conf/100); p<- 1-(alpha/2)
>  ttab<- qt(p, df) #funcion qt()
> ttab

[1] 1.8331
```

entonces aquí hemos creado el objeto `ttab` $= t_{(1-\frac{0,1}{2},10-1)}$.

1.2. Intervalo de confianza para $\widehat{\mu}$

Un estimador del promedio poblacional μ es

$$\widehat{\mu}_y = \frac{1}{n}\sum_{i\in S} y_i = \overline{y}, \tag{5}$$

Para construir un intervalo confidencial (Ec. 3) necesitamos calcular el error estándar estimado del estimador $\overline{y}$, i.e. $\widehat{\mathrm{SE}}\,[\overline{y}]$, como sigue

$$\widehat{\mathrm{SE}}\,[\overline{y}] = \sqrt{\widehat{V}\,[\overline{y}]} = \sqrt{\frac{\widehat{\sigma}_y^2}{n}(1-f)} = \sqrt{\frac{\widehat{\sigma}_y^2}{n}(1-\frac{n}{N})}, \tag{6}$$

donde $\widehat{\sigma}_y^2$ es un estimador de la varianza poblacional de la variable y, el cual se calcula mediante

$$\widehat{\sigma}_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2 = s_y^2. \tag{7}$$

Nota aparte:
- En la literatura (no especializada) es común ver que $\widehat{\mathrm{SE}}\,[\overline{y}]$ (Ec. 6) se escribe simplemente como $\mathrm{SE}\,[\overline{y}]$, lo cual no es correcto, sin embargo, por simplicidad podría ser aceptado, siempre y cuando se entienda que se trata de un valor estimado.
- Es común también ver que $\sqrt{\widehat{\mathrm{Var}}(\overline{y})}$ es simbolizado por $\widehat{\sigma}_{\overline{y}}$, lo cual es correcto. Por favor no confundir con $\widehat{\sigma}_y$.

Entonces, un intervalo de confianza para $\widehat{\mu}_y$ se calcula como

$$\overline{y} \pm \widehat{\mathrm{SE}}\,[\overline{y}]\, t_{(1-\frac{\alpha}{2},n-1)}, \tag{8}$$

1.3. Ejemplificando el concepto

Población

Empleemos el archivo `eucaLeaf.csv` como nuestra población y entonces ahora exploremos esta dataframe (i.e., población)

```
> dim(dbase)
[1] 744   7
> head(dbase)
  leaf.no  time tree shoot   l  w    la
1       1 EARLY   11    SC 125 32  28.8
2       2 EARLY   11    SC 300 50 106.7
3       3 EARLY   11    SC 300 43  86.5
4       4 EARLY   11    SC 225 35  61.4
5       5 EARLY   11    SC 228 37  68.2
6       6 EARLY   11    SC 120 24  22.2
```

Entonces ahora podemos ver que el tamaño de la población es 744.
Para esta población, digamos que estamos interesados en la variable área foliar (la). Entonces la la simbolizamos como nuestra variable aleatoria de interes y la expresamos como la variable aleatoria y.
Ahora calculemos algunos parámetros[2]

```
      Min       Max        Mu  Variance
    0.500   146.600    55.056   867.986
    Total        CV
40962.000    53.512
```

De estos parámetros, estamos interesados en la media de la población, esto es decir, μ_y. Ahora bien, recuerde que solo como asumimos que estas 744 observaciones corresponden a una población, entonces todo lo que calculamos empleando todas esas observaciones corresponden a parámetros. Entonces $\mu_y = 55{,}056$.

[2] Recuerde que podemos sólo calcular parámetros cuando tenemos todos los datos de la población, por lo tanto conocemos y_is para cada i-ésimo elemento de la población de tamaño N, o tenemos la lista de datos $y_1, y_2, y_3, \cdots, y_N$.

Ahora realicemos un gráfico de distribución de la variable respuesta y

```
> hist(dbase$y,xlab='Variable y',main='')
> abline(v=mu.y, col='red')
> text((mu.y-6) , 92, expression(mu(y)))
```

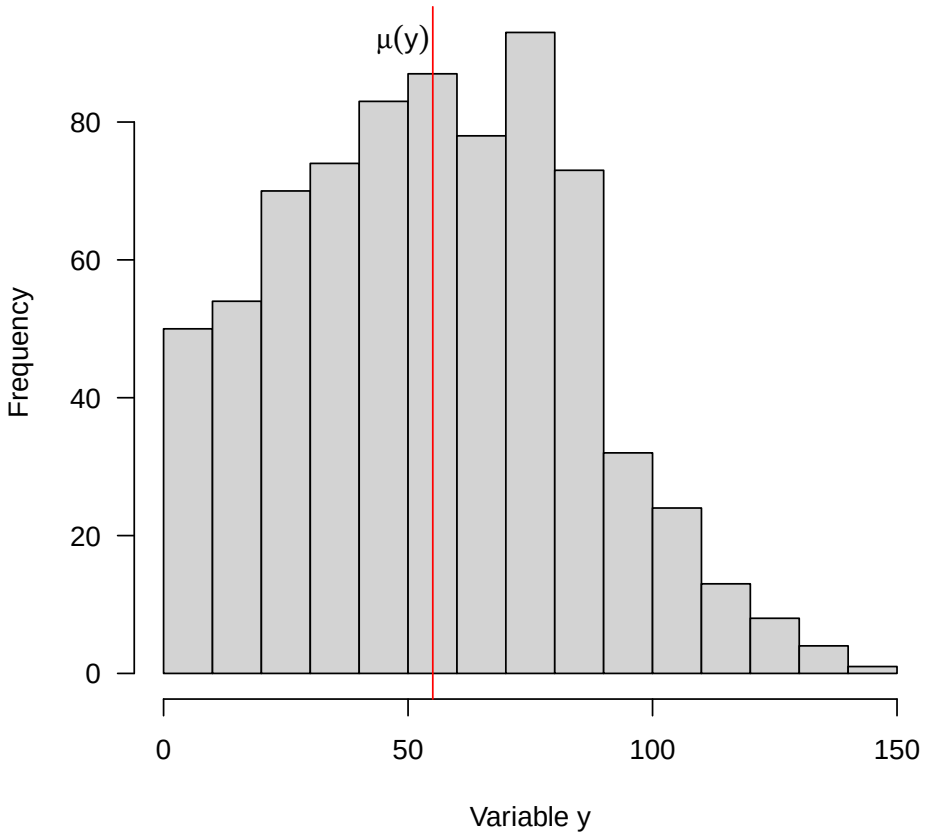


Figura 2: Histograma de la variable respuesta y. Note que en este gráfico también hemos representado el valor del parámetro μ_y (i.e., el parámetro de interés) como una línea vertical roja.

1.4. Marco de muestreo

Vamos a realizar muestreos aleatorios simples sin reemplazo (i.e., SRSwoR) de nuestra población para estimar el parámetro μ_y. Nuestro tamaño muestral (n) será 10 Emplearemos el siguiente estimador de μ

$$\widehat{\mu}_y = \frac{1}{n}\sum_{i\in S} y_i = \overline{y}, \tag{9}$$

y el error estándar como sigue

$$\widehat{\mathrm{SE}}\left[\overline{y}\right] = \sqrt{\frac{\widehat{\sigma}_y^2}{n}(1-\frac{n}{N})}, \tag{10}$$

donde $\widehat{\sigma}_y^2$ es la varianza poblacional estimada de y, siendo calculado como

$$\widehat{\sigma}_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2 = s_y^2. \tag{11}$$

Basado en lo expuesto anteriormente y al fijar un nivel de significancia de $\alpha = 0{,}1$, nosotros podemos construir un CI para μ_y como sigue

$$\overline{y} \pm \widehat{\mathrm{SE}}\left[\overline{y}\right] t_{(1-\frac{\alpha}{2},n-1)}, \tag{12}$$

La intensidad de muestreo, i.e., fracción de muestreo ($f = n/N$), es

$$f = \frac{n}{N} = 0{,}0134 \tag{13}$$

Dado que tanto $n = 10$ y $\alpha = 0{,}1$ son valores fijos, también lo será el valor $t_{(1-\frac{\alpha}{2},n-1)}$ como sigue

```
[1] 1.8331
```

Esto es $t_{(1-\frac{0,1}{2},10-1)} = t_{(0,95,9)} = 1{,}8331$

Simulando muestras

Realizaremos 100 simulaciones, y para cada simulación, nostros haremos lo siguiente :

a. Seleccionar una muestra aleatoria sin reemplazo (SRSwoR), de tamaño n,

b. Basado en la muestra, calcular los estadísticos $\bar{y}$, s_y, y $\widehat{\mathrm{SE}}\left[\overline{y}\right]$ mediante las formulas 9, 11, y 10, respectivamente

c. y calcular el CI (de acuerdo a Eq. 12), y graficarlo.

La figura 3 muestra los intervalos confidenciales para cada una de las 100 simulaciones empleando el procedimiento de muestreo explicado anteriormente.

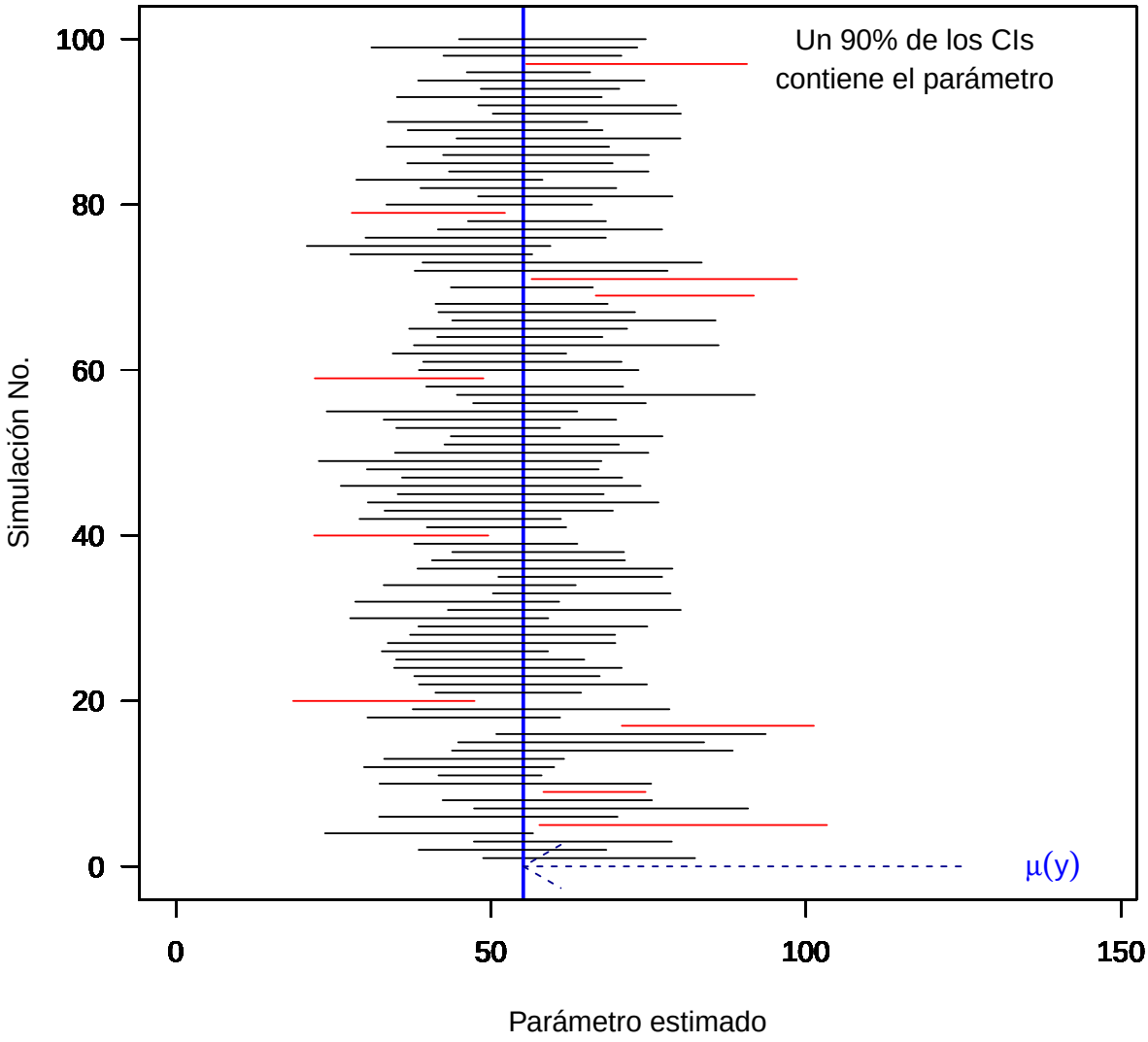


Figura 3: Esquema que representa el concepto de intervalos de confianza. Cada línea horizontal corresponde a un intervalo de confianza calculado a un nivel de significancia $\alpha = 0{,}1$ a partir de una muestra aleatoria de $n = 10$ elementos desde una población de $N = 744$ para el área foliar estimada. Se ha repetido este proceso en 100 ocasiones (simulaciones). En azul se ha marcado el valor del parámetro de interes, este es μ_y. Además aquellos intervalos confidenciales que no logran capturar al valor del parámetro son representados en rojo.

Ahora bien si se repite este proceso de simular 100 muestreos, en cuatro ocasiones (a, b, c, y d), se observa una pequeña variación en el porcentaje de simulaciones donde el intervalo de confianza cubre al parámetro (Figura 4). Sin embargo, si aumentaramos las simulaciones a 10000, este porcentaje va a tender a ser igual a $1 - \alpha\,\%$, es decir, el nivel de confianza estadístico empleado para determinar el valor del valor de la distribución de t-student.

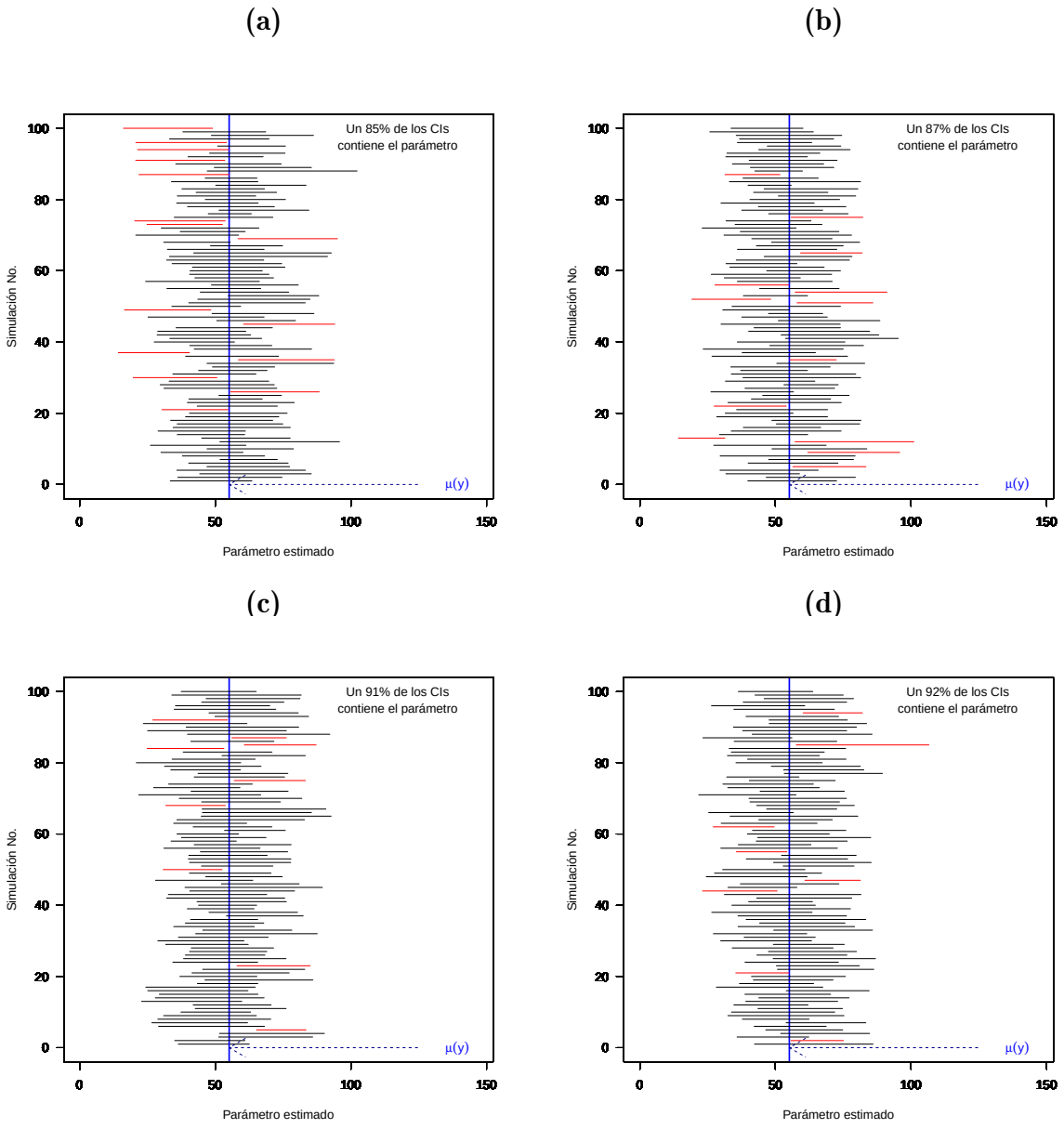


Figura 4: Cuatro simulaciones de 100 muestreos aleatorios y sus respectivos intervalos de confianza al 10 % de significancia estadística.

1.5. Efecto de diferentes niveles confidenciales

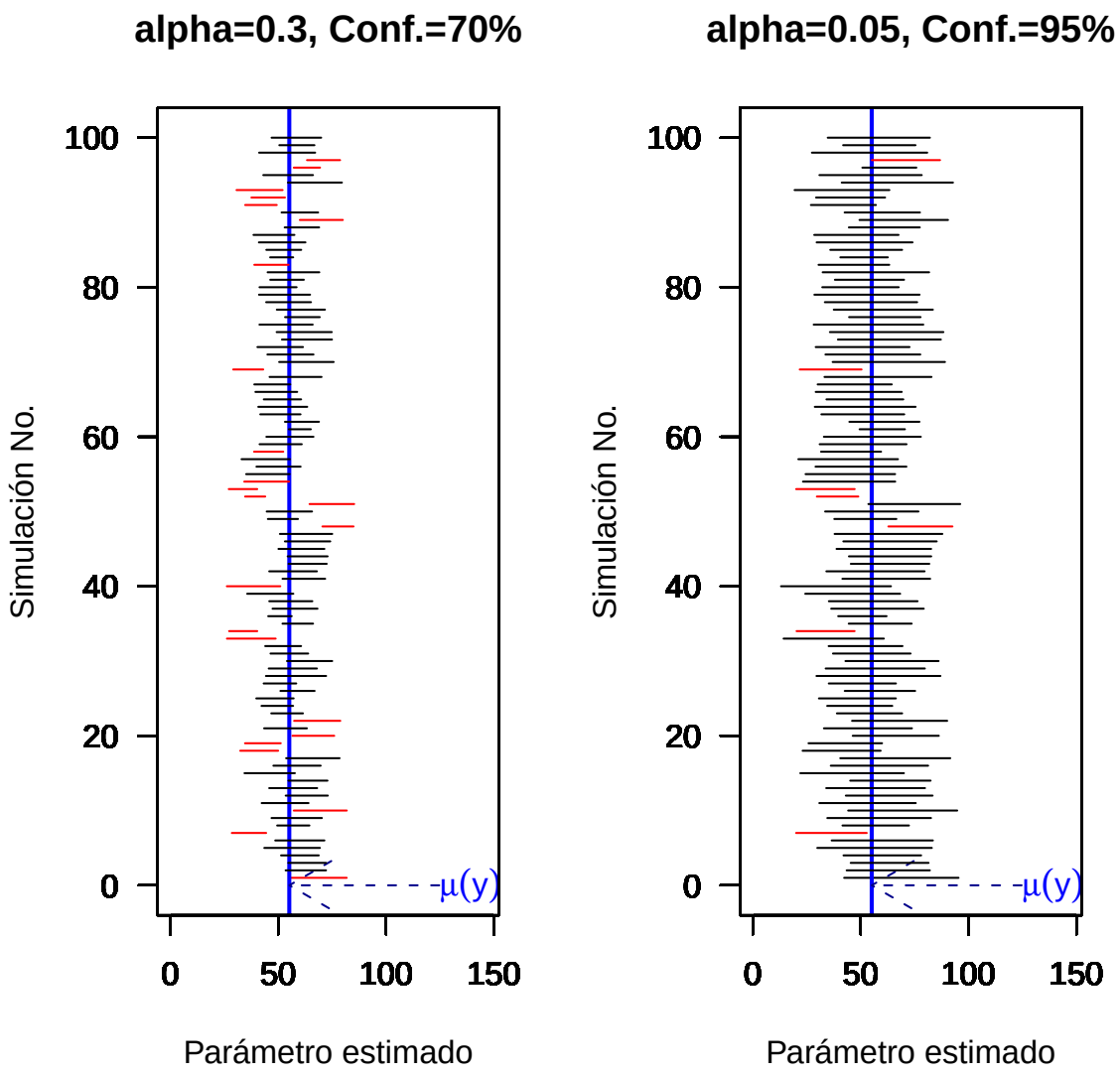


Figura 5: Efecto del nivel de significancia (α) en la construcción de intervalos confidenciales. Para una población de tamaño $N = 742$ se muestrea con $n = 10$, y para cada muestra se calcula el intervalo de confianza para un nivel de signficancia de un 10 % (izquierda) y de un 5 % (derecha).

1.6. Efecto de diferentes tamaños de muestreo

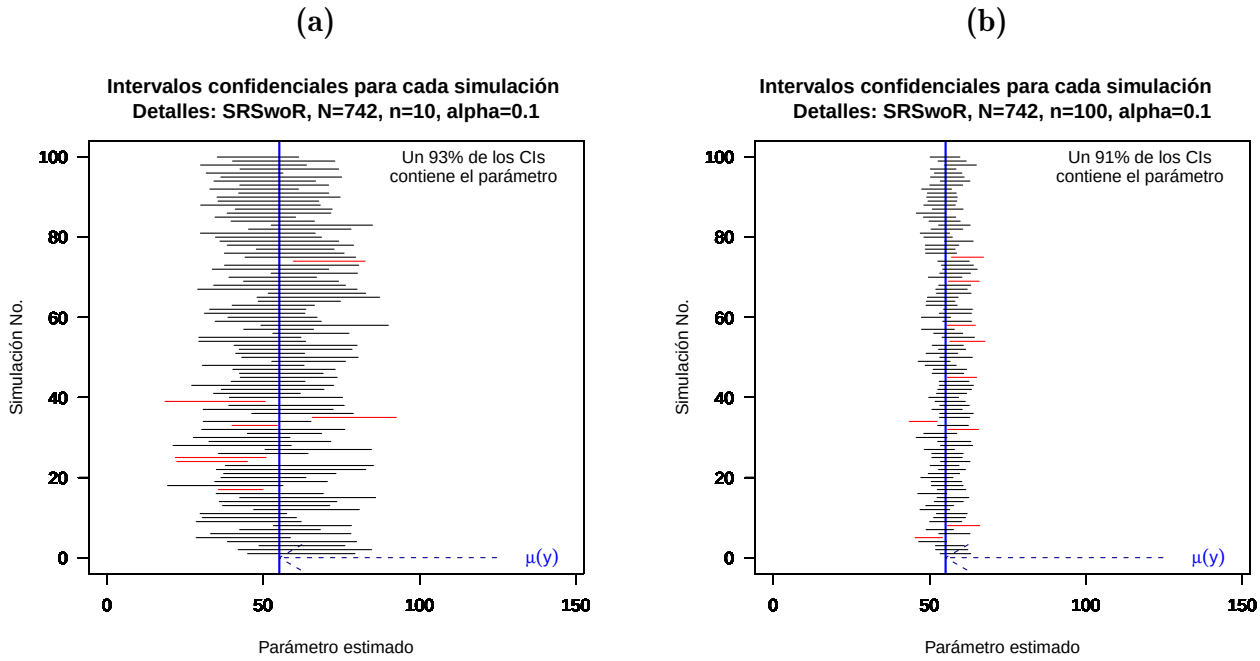


Figura 6: Efecto del tamaño muestral en la construcción de intervalos confidenciales. Para una población de tamaño $N = 742$ y empleando un nivel de signficancia de un 10 % ($\alpha = 0{,}1$), en (a) se muestrea la población con un tamaño muestral (n) de 10, mientras que en (b) $n = 100$.

Para recordar entonces

¿Qué NO es un intervalo de confianza?

- Un intervalo de confianza estadístico de $1-\alpha\,\%$ **No** indica que el parámetro tiene una $1-\alpha\,\%$ probabilidad de estar en ese intervalo de confianza.
- Esto es la interpretación más frecuente sobre CI, sin embargo, es incorrecta. No repetirla!!
- **No** podemos hacer declaraciones probabilísticas sobre parámetros, recuerden, ellos son constantes!

¿Qué SI es un intervalo de confianza?

- El valor de t es un número que multiplica el error estandar, lo cual define la distancia a ambos lados de $\widehat{\theta}$ dentro del cual el parámetro θ estaría incluido la mayoría del tiempo.
- ¿Qué significa esto de "mayoría del tiempo"? El coeficiente de confianza!: muestras diferentes generan diferentes intervalos. El coeficiente de confianza es el porcentaje de estas muestras que cubrirían θ.

Nota aparte:
- Los intervalos confidenciales son sumamente ocupados en estadística y en el mundo científico-técnico, por lo tanto, es vital entender su base e interpretación.
- Les recomiendo leer el artículo de Ellison y Dennis (1996).

2. Teorema central del límite y su relación con muestreo

- Recien se ha revisado lo que es un intervalo de confianza.
- Se ocupa el valor de t que se asemeja a una distribución normal a medida de que sus grados de libertad se tornen cada vez más grandes
- mmmmmhhh, pero ¿Qué pasa con las distribuciones de las variables de interes? ¿Son todas normales?.... (ejemplos en Figura 7).

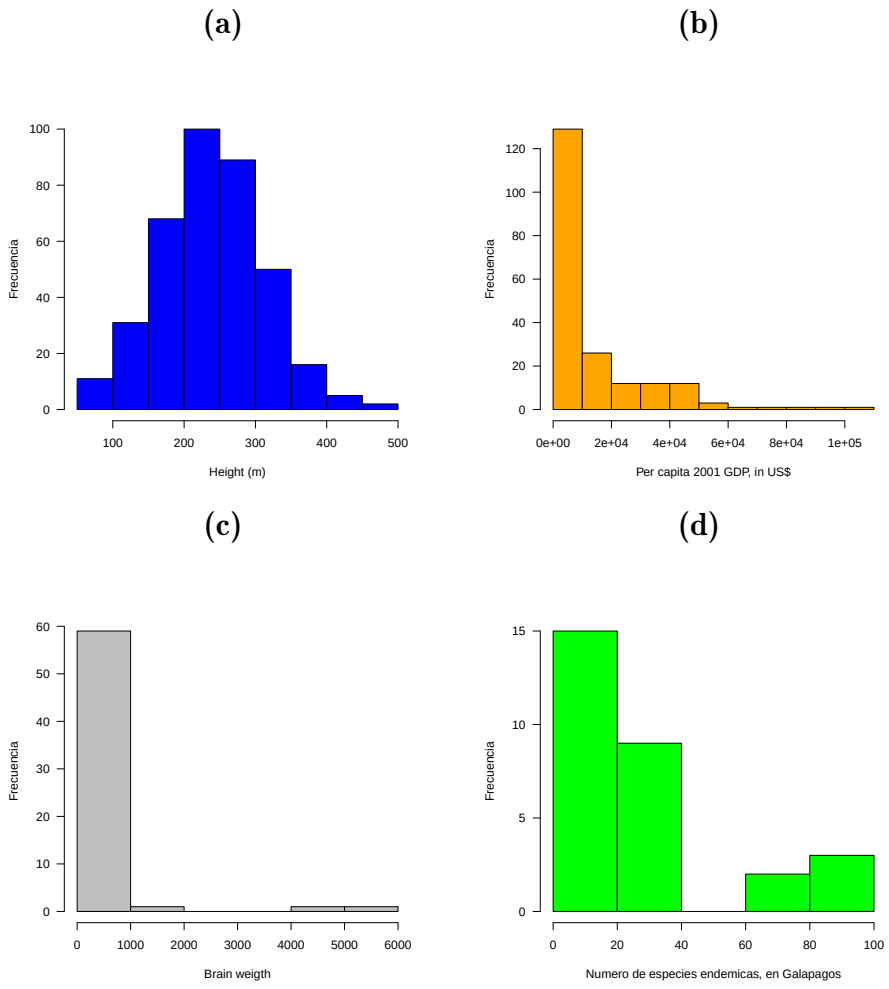


Figura 7: Histogramas de diferentes variables con datos obtenidos de la librería `alr3` de R. Las variables son: (a) altura de árboles, (b) ingreso per capita de países, (c) peso de cerebros de 62 species de mamiferos, y (d) número de especies endémicas en las islas Galapagos; dichas variables provienen de las dataframes `ufc`, `UN`, `brains`, y `galapagos`.

- ¿Importa la distribución de la variable aleatoria de interes?

- **Teorema central del límite (TCL)**:
  "Si tomamos muestras aleatorias de tamaño n de una variable aleatoria Y desde una población con media μ y desviación estándar σ, entonces a medida que n aumente, la media muestral $\overline{y}$ se aproxima a una distribución normal con media μ y desviación estándar $\frac{\sigma}{\sqrt{n}}$"

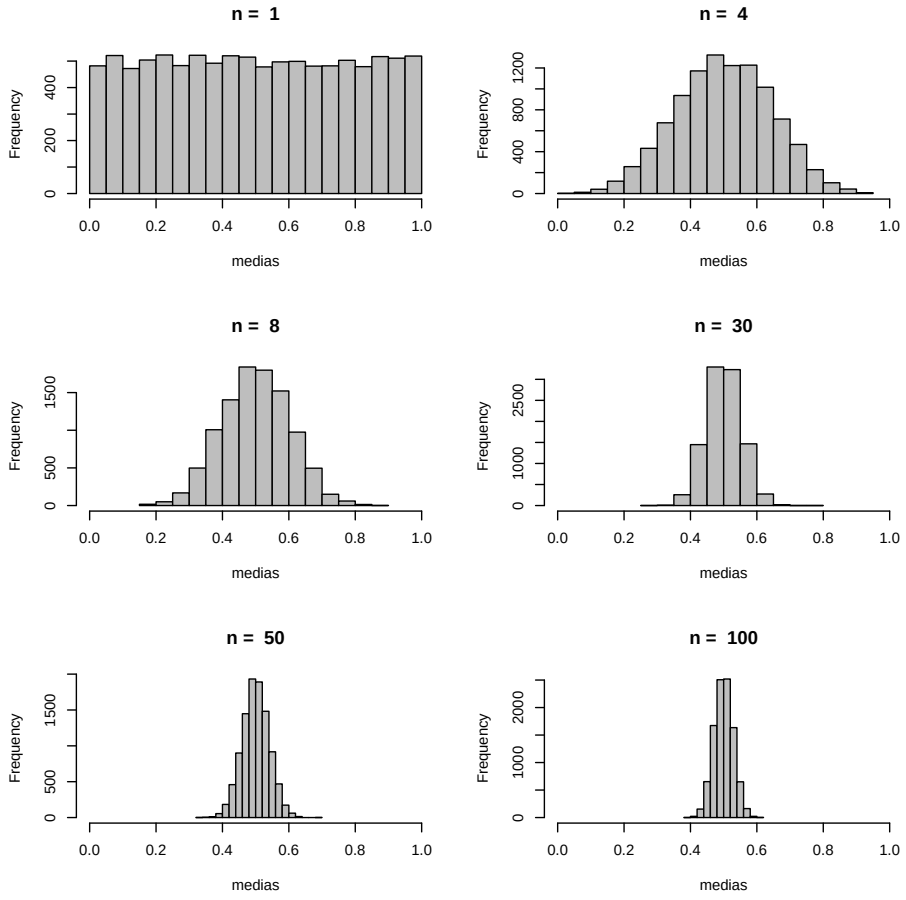


Figura 8: Representación del teorema central del límite. Muestreo aleatorio de tamaño n desde una población de una variable aleatoria Y de distribución uniforme, luego calculamos la media aritmética ($\overline{y}$) para cada muestreo y representamos la distribución de este estimador.

- A medida que n aumente nos acercamos más a μ y además la variación del estimador disminuye.

- Además el TCL tiene otra característica respecto a las distribuciones de las variables aleatorias.
- Analizemos la Figura 9

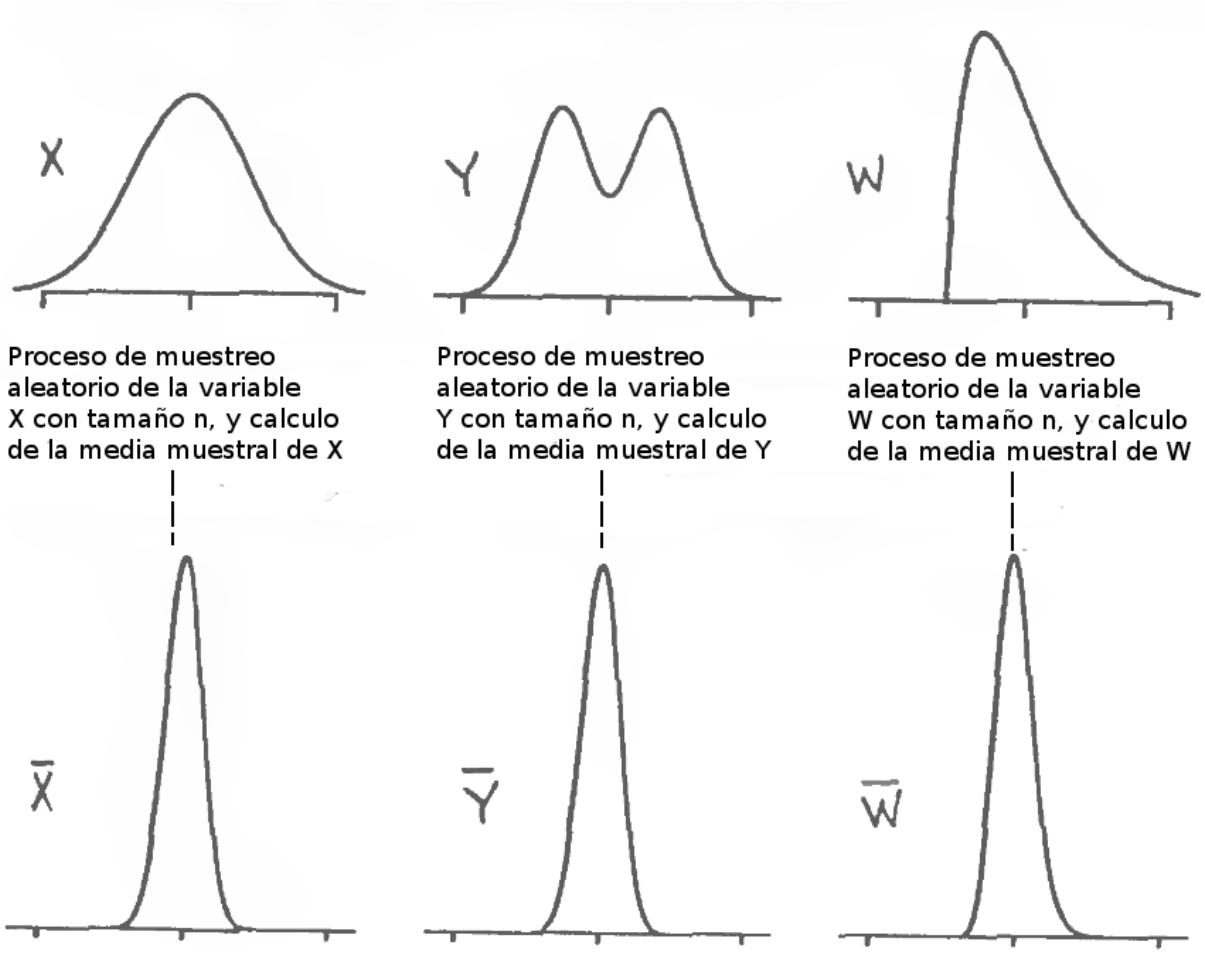


Figura 9: Representación del teorema central del límite. Las tres variables aleatorias de la primera fila de la figura (X, Y, W) poseen la misma media poblacional μ pero diferentes distribuciones, sin embargo, si muestreamos cada una de sus respectivas poblaciones con una muestra de tamaño n, y calculamos la media muestral $(\overline{y}, \overline{x}, \overline{w})$ de cada una de ellas, la distribución de este estimador tendra una distribución normal.

- No importa la distribución de la variable aleatoria, la media siempre tenderá a aproximarse a una distribución normal a medida que n aumente.
- **¿De verdad?** 

- Veamos una población real. Datos de la variable aleatoria volumen de 14.443 árboles de *Pinus taeda* en el sur de Estados Unidos. Estos datos son descritos y empleados en el estudio de Salas y Gregoire (2010).

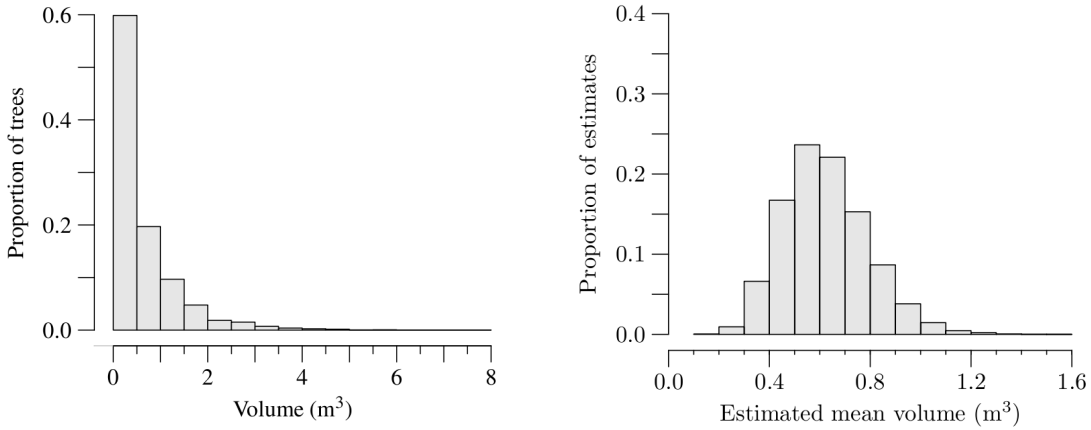


Figura 10: (a) Distribución de la variable aleatoria volumen, para una población compuesta por 14.443 árboles de *Pinus taeda* en el sur de Estados Unidos. (b) Distribución estimada de la media aritmética de volumen de árboles basado en 25.000 muestreas aleatorias simples de 20 árboles cada una (i.e., $n = 20$).

3. Tamaño muestral

- Hasta ahora, hemos ejercitado con un diseño de muestreo para un tamaño muestral n fijado por nosotros
- En torno a la determinación de n varias dudas pueden surgir
  a. ¿Fue ese n suficiente?
  b. ¿Debimos muestrear con una mayor intensidad?
  c. ¿Debimos muestrear con una menor intensidad?
  d. ¿Cómo podemos determinar el n a emplear?
- En general podemos ocupar la siguiente fórmula (ver el origen al final de esta sección)

$$n = \frac{t^2_{(1-\frac{\alpha}{2},df)}\, CV^2}{EMA^2}, \tag{14}$$

  donde CV es el coeficiente de variación de la variable aleatoria de interes, expresado en %; $t_{(1-\frac{\alpha}{2},df)}$ es el $(1-\frac{\alpha}{2})$ percentil de la distribución de t (o t de *Student*) con *df* grados de libertad, para un nivel de significancia α; y EMA es el error de muestreo o error máximo admisible, en %.
- Por ejemplo: Supongamos que queremos estimar un parámetro de una variable aleatoria (i.e., Y) con un nivel de confianza estadístico del 95 % con un nivel de error máximo admisible de 10 %. Supongamos también que conocemos el CV de la variable Y, ya sea por un premuestreo o por estudios similares, y es de 35 %.
- El valor de t depende de n, asumamos n que tiende a infinito primero, esto es

```
> ttab <- qt(.975,10000000000)
> ttab
[1] 1.96
```

- Con este valor del t calculado, comenzamos a iterar hasta encontrar un n que establize el resultado.

  Primero definimos los valores dados de variabilidad y error máximo admisible

```
> N <- 400
> cv.p <- 35
> ema <- 10
```

Posteriormente, calculo los valores para n según la fórmula de la expresión (14), como sigue

```
> #primer calculo
> n.aste <- (ttab^2*cv.p^2)/(ema^2)
> n.aste <- round(n.aste,0)
> n.aste

[1] 47

> #segundo calculo
> ttab.2 <- qt(.975, n.aste-1)
> ttab.2

[1] 2.0129

> n.aste2 <- (ttab.2^2*cv.p^2)/(ema^2)
> n.aste2<- round(n.aste2,0)
> n.aste2

[1] 50

> #tercer calculo
> ttab.3 <- qt(.975, n.aste2-1)
> ttab.3

[1] 2.0096

> n.aste3 <- (ttab.3^2*cv.p^2)/(ema^2)
> n.aste3 <- round(n.aste3,0)
> n.aste3

[1] 49

> #cuarto calculo
> ttab.4 <- qt(.975, n.aste3-1)
> ttab.4

[1] 2.0106
```

```
> n.aste4 <- (ttab.4^2*cv.p^2)/(ema^2)
> n.aste4 <- round(n.aste4,0)
> n.aste4

[1] 50
```

luego de iterar los valores respectivos para n se ha llegado a una valor estabilizado de $n = 50$.

- Para poblaciones finitas existe una corrección de la formula (14), como sigue

$$n^* = \frac{n}{1 + \frac{n}{N}}, \tag{15}$$

```
> n.f <-  n.aste4/(1+(n.aste4/N))
> n.f

[1] 44.444
```

o lo que es equivalente a modificar la formula (14) y reescribirla considerando la corrección de (15) como

$$n^* = \frac{N\ t^2_{(1-\frac{\alpha}{2},df)}\ CV^2}{N\ EMA^2 + t^2_{(1-\frac{\alpha}{2},df)}\ CV^2}, \tag{16}$$

¿Cómo se llega a la fórmula para calcular el tamaño muestral?:

- Todo parte con la definición del intervalo de confianza
  Estimación ± margen de error
- Entonces la pregunta es ¿Cómo determinar ese margen de error?
- Ya sabemos por lo explicado anteriormente que

$$E = \text{SE}(\widehat{\theta})\ t_{(\alpha,n)}, \tag{17}$$

$$= \frac{\sigma_y}{\sqrt{n}}\ t_{(\alpha,n)} \tag{18}$$

y desde aquí se despeja n, como sigue

$$\sqrt{n} = \frac{\sigma_y}{E}\ t_{(\alpha,n)}, \tag{19}$$

$$n = \left(\frac{\sigma_y}{E}\ t_{(\alpha,n)}\right)^2, \tag{20}$$

donde σ_y y el margen de error E se expresan en las mismas unidades de medición que la variable aleatoria Y de interes.

Una forma más simple de representar la variabilidad de la variable aleatoria y es usando el coeficiente de variación en términos porcentuales, y así también el error, por lo tanto, arrivando a

$$n = \left(\frac{CV_y}{ME}\ t_{(\alpha,n)}\right)^2, \tag{21}$$

donde: CV_y es el coeficiente de variación (en %) de la variable aleatoria Y en la población; ME es el margen de error en %, y $t_{(\alpha,n)}$ es el valor de la distribución de t-student para un nivel α de significancia y un tamaño muestral n. Esta expresión es la misma dada anteriormente en Ec. (14).

Referencias

Ellison AM, B Dennis. 1996. Paths to statistical fluency for ecologists. *Frontiers in Ecology and the Environment* 8(7):362–370.

Salas C, TG Gregoire. 2010. Statistical analysis of ratio estimators and their estimators of variances when the auxiliary variate is measured with error. *Eur. J. For. Res.* 129(5):847–861.

Student. 1908. The probable error of a mean. *Biometrika* 6(1):1–25.