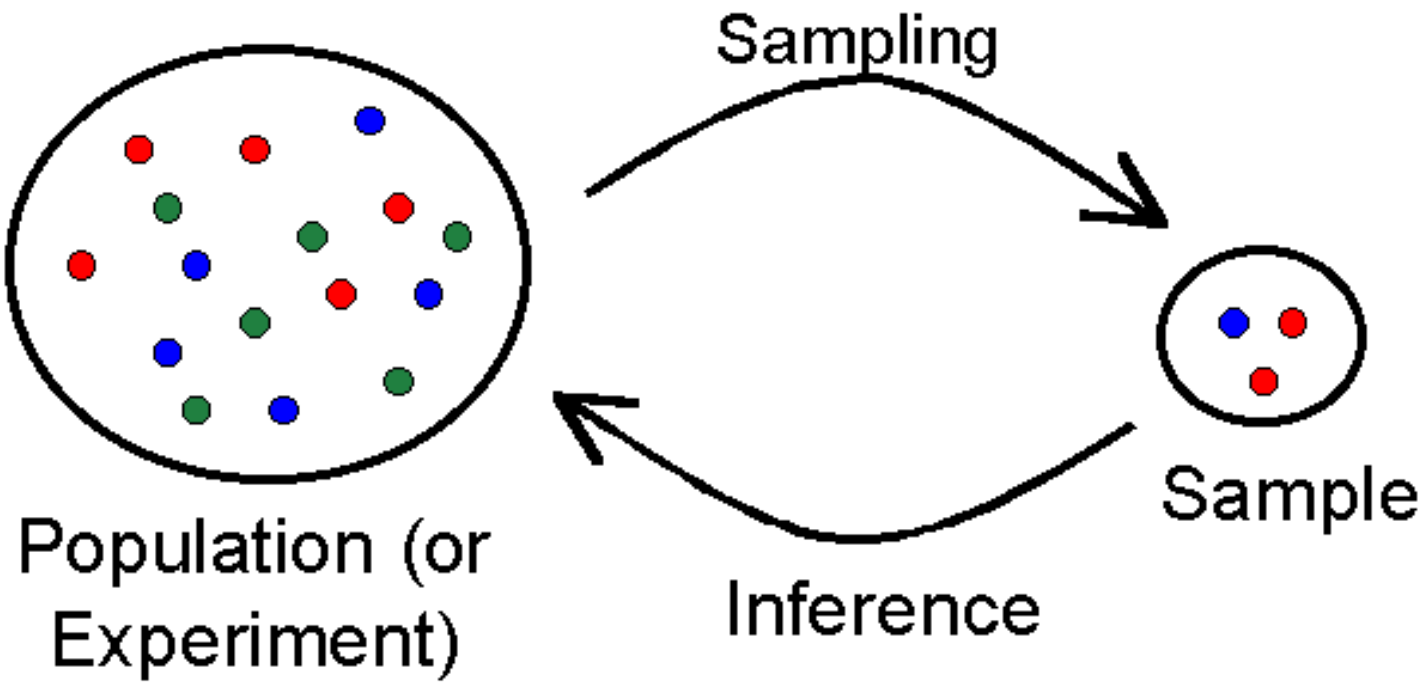
Conceptos de Inferencia

Olivier Nuñez

8 de junio de 2018


Sampling
Sample
Inference
Population (or
Experiment)

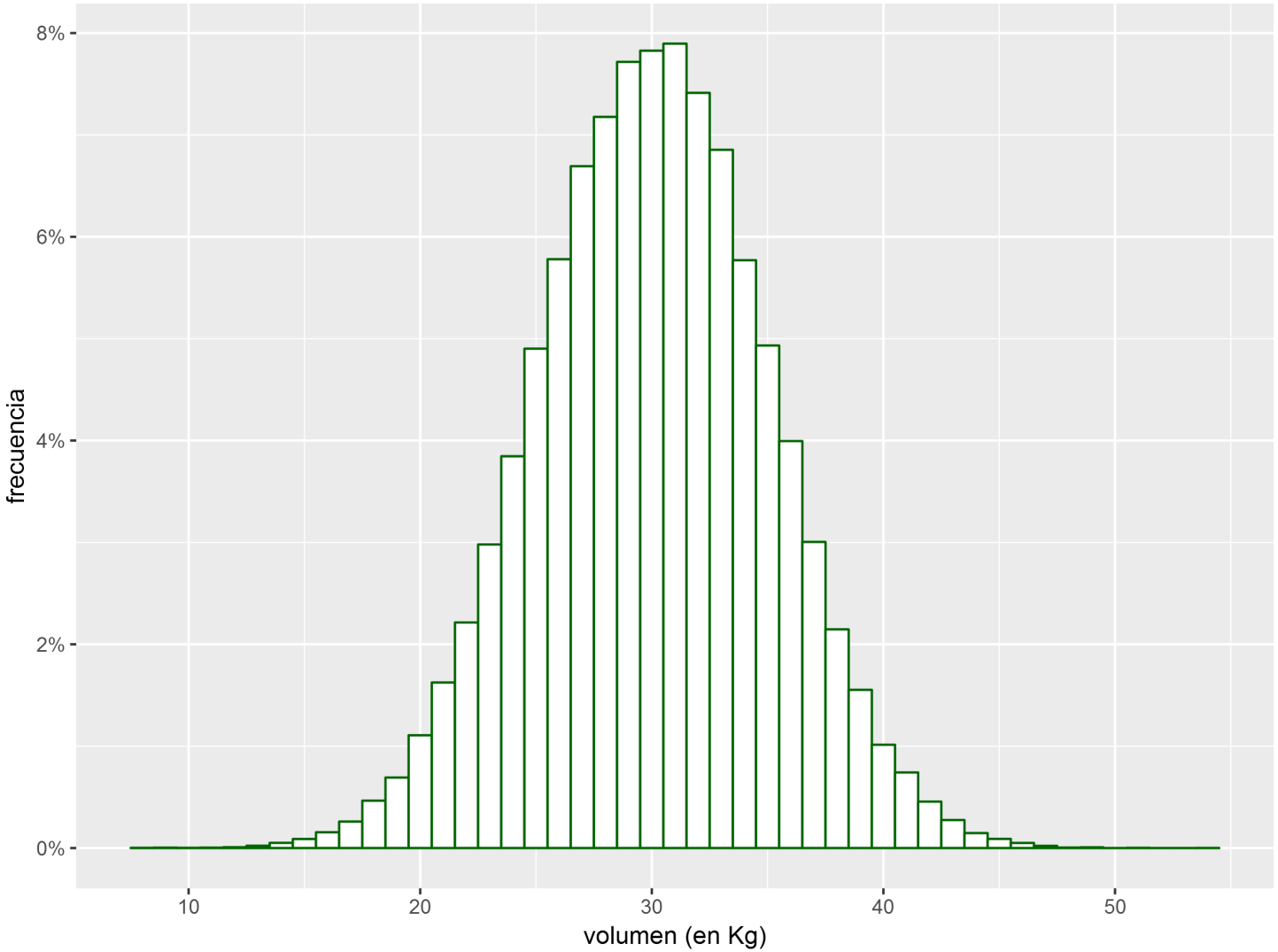
- La **Estadística descriptiva** aporta las técnicas para resumir y presentar la información extraída de una muestra. Sin embargo, rara vez nos interesa la muestra como tal, sino que interesa por su capacidad para aportar información acerca de otros sujetos o otras situaciones.

- La **Estadística Inferencial** aporta las técnicas para extraer conclusiones a partir de una muestra. En la inferencia distinguimos:
  - La Estimación: permite estimar los parámetros de la población a partir de la muestra (Ej.: ¿Qué volumen de envases recicla un hogar español?).
  - El Contraste de hipótesis: permite tomar una decisión sobre los parámetros de la población (Ej.: ¿Se recicla mejor que hace 10 años?)

- ¿Cuantos envases (en Kg) recicla un hogar al año?
- Este «volumen» de envases es una **variable** (X) porque varía de un hogar al otro.
- Es **aleatoria** porque no controlamos (del todo) sus variaciones; no podemos prever sus valores de manera deterministica.
- ¿Cómo describir una variable aleatoria?

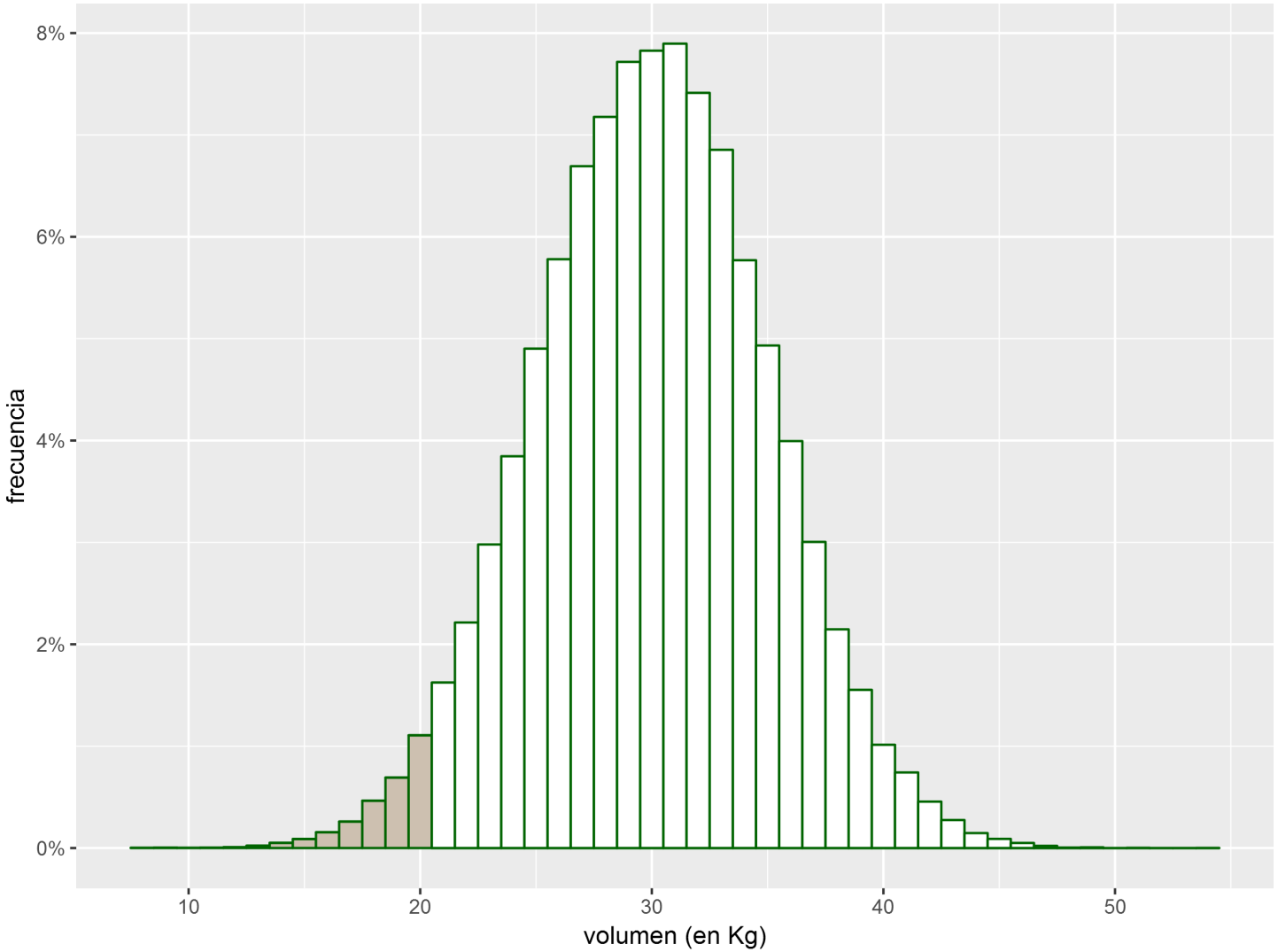
Descripción de la distribución de una variable

El histograma



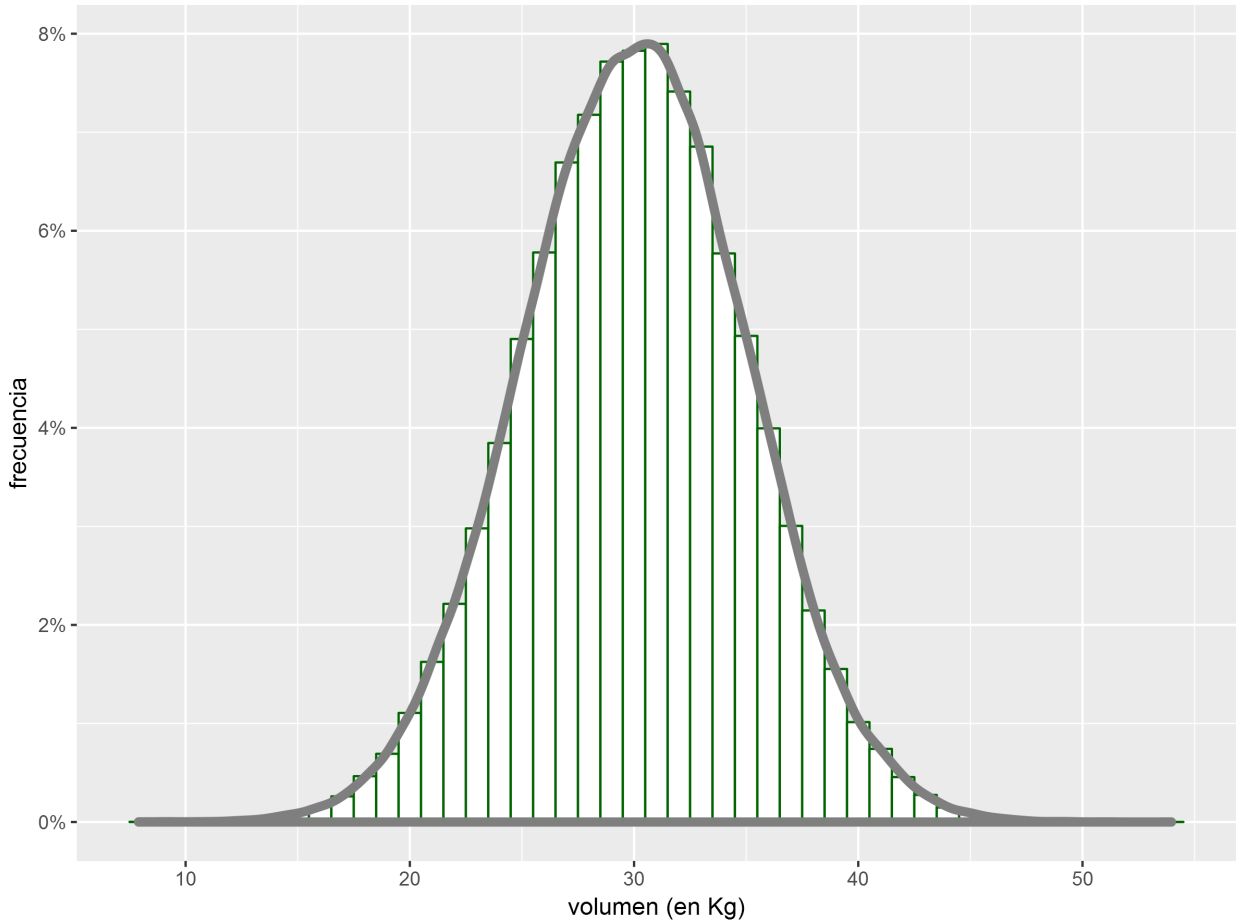
Descripción de la distribución de una variable

¿Qué proporción de hogares reciclan menos de 20 Kg?



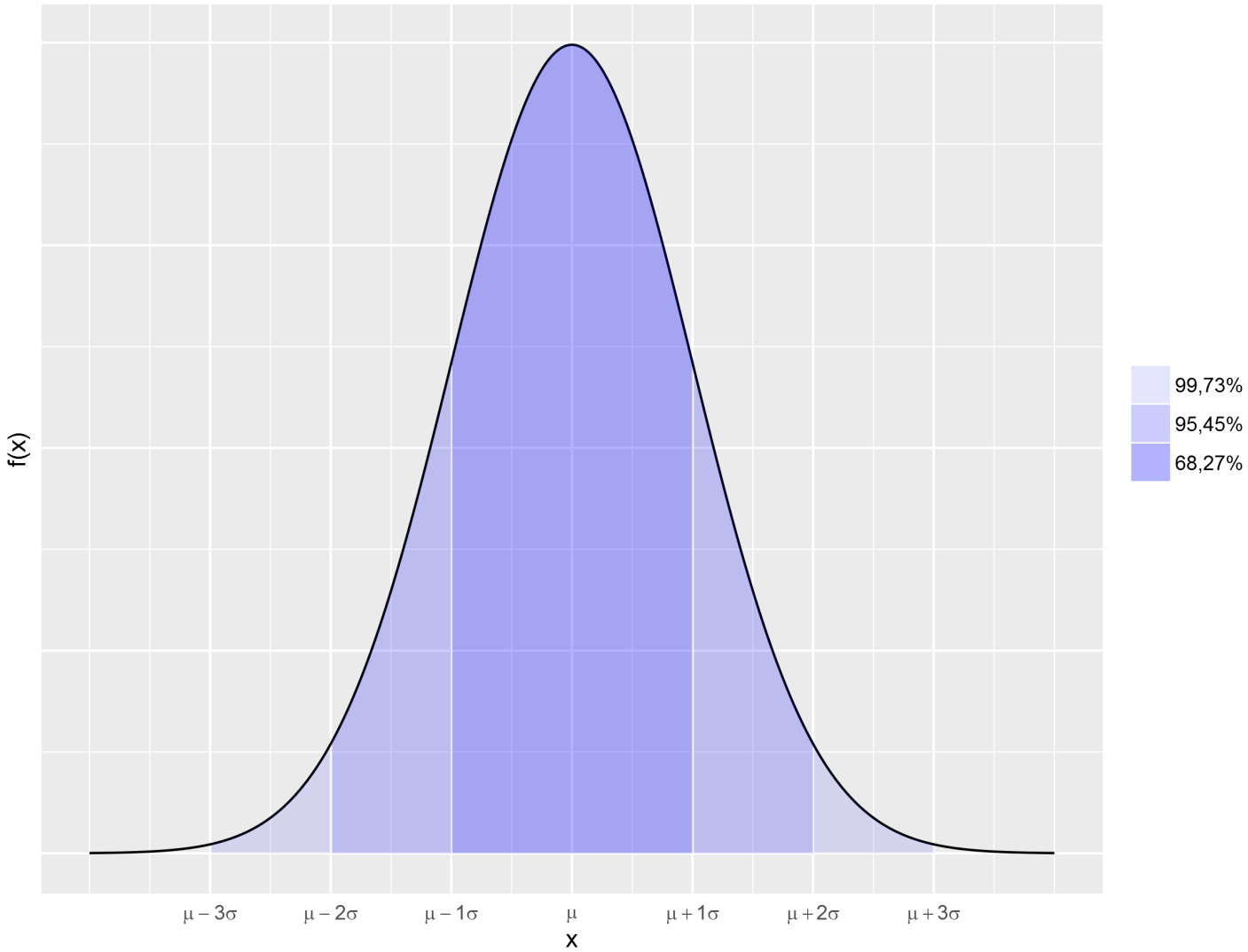
- Si X tiene muchos valores posibles, describir su distribución mediante un histograma puede ser laborioso.
- La mayoría de los modelos de distribución se caracterizan por el conocimiento de su media y su varianza (Ej.: la distribución Normal).
- La media y la desviación típica son respectivamente medidas de **localización** y de **dispersión** de la variable.

Si X es continua y su distribución es aproximadamente simétrica, la distribución normal suele ser un buen modelo para su distribución:



Distribución Normal

La campana de Gauss



- Una forma natural de estimar un parámetro poblacional consiste en utilizar su equivalente muestral.
- La media muestral $\overline{y}$ es la media de los valores de la muestra:

$$\overline{y} = \frac{1}{n} \sum_{i \in M} y_i$$

- Con un muestreo equiprobabilistico, $\overline{y}$ es un estimador insesgado de la media poblacional

$$\mu = \frac{1}{N} \sum_{i=1}^{N} y_i$$

Incertidumbre en la estimación

¿Cómo varía nuestra estimación de una muestra a otra?

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

$$\operatorname{se}(\overline{y}) = \frac{\sigma}{\sqrt{n}}$$

- Además, cuando n y $(N - n)$ son «grandes», tenemos que

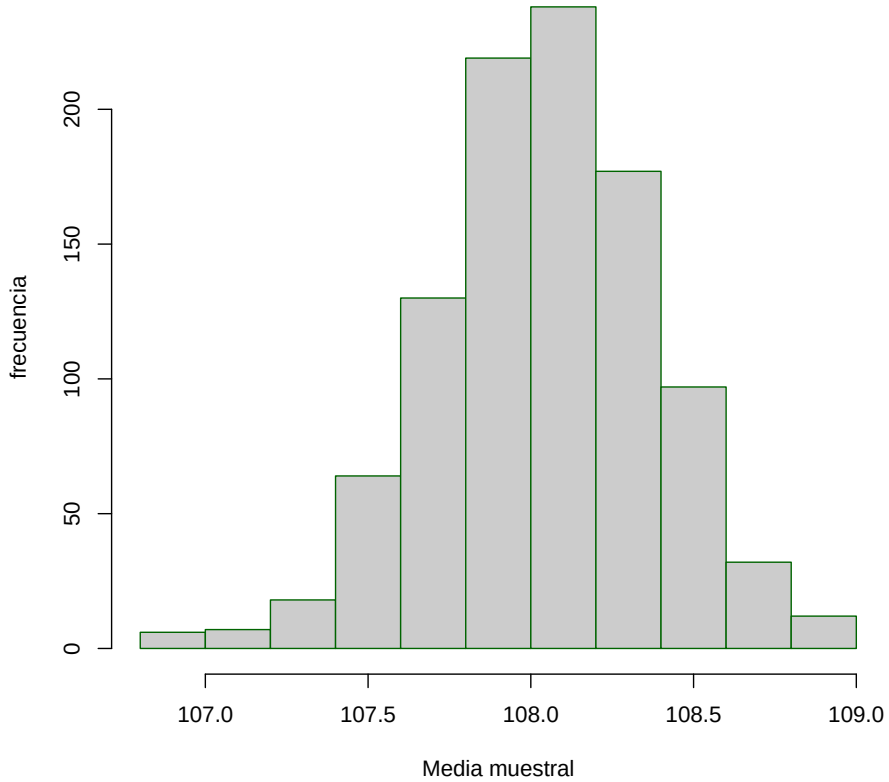
$$\overline{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

donde $N(\mu, \sigma)$ denota la distribución normal con media μy desviación típica σ.

Incertidumbre en la estimación

Ejemplo: estimación de la altura de las niñas de 5 años

Distribución de 1000 medias muestrales de tamaño $n = 200$ sacadas de una distribución normal $N(108, 5)$:



- El **intervalo de confianza** permite reflejar la incertidumbre en la estimación.
- Típicamente, un intervalo de confianza (del 95 %) de la media poblacional tiene la forma:

$$\overline{y} \pm 1{,}96 \times \mathrm{se}\left(\overline{y}\right)$$

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.

Estimación del error estándar

La variabilidad de los datos en la poblacional suele ser desconocida

- En practica, la varianza poblacional σ^2 es desconocida y requiere ser estimada para construir el intervalo de confianza.

- Para un muestreo equiprobabilistico, la varianza muestral

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2$$

es un estimador insesgado de la varianza poblacional

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \mu)^2$$

- Al sustituir σ por su estimador s 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.

- Cabe mencionar que si $n \simeq 200$, la distribución de Student $t_{(n-1)}$ es muy próxima a la normal y tenemos que $t_{(n-1),97{,}5\,\%} \simeq 1{,}96$

Determinación del tamaño muestral

Precisión y confianza compiten

- Para un nivel de confianza y un margen de error dados, ¿Cómo de grande ha de ser el tamaño muestral n?
- El margen de error del intervalo de confianza $(1-\alpha)$ % es por definición (aproximación normal):

$$e = z_{1-\frac{\alpha}{2}} \times \text{se}\left(\overline{y}\right) = z_{1-\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$$

donde $z_{1-\frac{\alpha}{2}}$ es el cuantil $\left(1-\frac{\alpha}{2}\right)$ de la distribución normal estándar.

Determinación del tamaño muestral

Precisión y confianza compiten

- El margen de error del intervalo de confianza $(1-\alpha)$ % es por definición (aproximación normal):

$$e = z_{1-\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$$

- Despejando n en la ecuación, obtenemos que n ha de verificar

$$n \geq \frac{z^2_{1-\frac{\alpha}{2}}}{e^2} \times s^2$$

y este umbral crece con la precisión, el nivel de confianza y la variabilidad de los datos.

Contraste de hipótesis

Tomar decisiones en situación de incertidumbre

- En ocasiones, el interés de la investigación se centra no tanto en estimar el valor concreto de un parámetro poblacional, sino en determinar si es superior o inferior a un valor determinado.
- Los contrastes de hipótesis parten de una hipótesis nula, H_0. Esta hipótesis se aceptará a no ser que los datos muestrales aporten evidencia en contra. En este caso, H_0 se rechazaría en favor de una hipótesis alternativa H_1.

- Ejemplo: las botellas de coca-cola que estamos produciendo contienen al menos 330ml de bebida?

$$H_0 : \mu \geq 330 \text{ versus } H_1 : \mu < 330$$

Errores en la decisión

Error de tipo I y II

En el contraste de hipótesis existen dos tipos potenciales de error:

- El riesgo de equivocarnos al rechazar H_0, se denota α y se llama riesgo o **error de tipo I**, denominado también nivel de significación del contraste. Es el riesgo del fabricante en fiabilidad (riesgo de equivocarse al dar la alarma).
- El riesgo de equivocarnos al aceptar H_0, se denota β y se llama riesgo o **error de tipo II** (riesgo del cliente en fiabilidad).

		Realidad	
		H_0	H_1
Decisión	H_0	Acierto	Error de tipo II(β)
	H_1	Error de tipo I (α)	Acierto

- Un regla de decisión natural para contrastar la hipótesis $H_0 : \mu \geq \mu_0$, consiste en rechazar esta hipótesis si $\overline{y}$ es «pequeño». ¿Pero cuanto de pequeño?

- El primer riesgo que se quiere controlar es α, el riesgo de equivocarse al rechazar H_0. Para ello, se rechazará H_0 si el intervalo de confianza

$$\left[-\infty, \overline{y} + t_{(n-1),95\,\%} \times \frac{s}{\sqrt{n}}\right]$$

no contiene μ_0.

Regla de decisión

Contraste bilateral

- De manera similar, se rechazará la hipótesis $H_0 : \mu = \mu_0$, si el intervalo de confianza

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

no contiene μ_0.

- Esta regla de decisión tiene también un error de tipo I igual a α.

Contrastar diferencias de medias

Vocabulario

- En el caso de la comparación de las medias de dos grupos, la hipótesis nula más natural es la ausencia de diferencia entre las medias, $H_0 : \mu_1 = \mu_2$.
- La regla de decisión consistirá en rechazar H_0 si el intervalo de confianza de la diferencia de las dos medias $\mu_1 - \mu_2$ no contiene 0.
- Para un nivel de significación α, este intervalo de confianza tiene la forma

$$(\overline{y}_1 - \overline{y}_2) \pm t_{(n_1+n_2-2),1-\alpha} \times s \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

donde las medias muestrales ($\overline{y}_1$ y $\overline{y}_2$) y los tamaños muestrales (n_1 y n_2) corresponden a muestras extraídas de los dos grupos. La varianza muestral s^2 se obtiene aquí juntando las dos muestras (pooled variance).

Acerca del p-valor

Medida de la credibilidad de H_0

- Una manera sencilla de medir la credibilidad de H_0 consiste en calcular la probabilidad de observar algo que discrepe más de H_0 que lo observado, en caso de que H_0 fuese cierto.

- Esta probabilidad se llama el **p-valor** y se puede interpretar como una medida de credibilidad de H_0. Si es inferior al nivel de significación establecido α, rechazamos la hipótesis nula.

- Ejemplo: supongamos que observamos que en una muestra de 10 botellas el volumen medio de coca-cola es 327 ml. Si realmente la hipótesis $H_0 : \mu \geq 330$ fuese cierta y la precisión de la maquina fuese $\sigma = 5$ ml, la probabilidad de observar algo aún más alejado de H_0 sería:

$$p = P\left(\overline{y} < 327 | H_0 \text{ cierta}\right) = \text{pnorm}\left(327, 330, \frac{5}{\sqrt{10}}\right) \simeq 3\,\%$$