Tema 7. ESTIMACIÓN POR INTERVALO.

Objetivos

Conceptos:

- Conocer los siguientes modelos de probabilidad: Chi-cuadrado, t-Student, F-Snedecor. De cada uno de ellos:
  - Definición en función de otras variables aleatorias
  - Propiedades gráficas y de asimetría
- Comprender el significado del intervalo de confianza de un determinado parámetro
- Comprender el método general de obtención del intervalo de confianza de un parámetro
- Comprender cómo depende el intervalo del nivel de confianza y del tamaño de la muestra elegida.

Saber hacer:

- **Manejar las tablas** de los siguientes modelos de probabilidad: Chi-cuadrado, t-Student, F-Snedecor, para hallar percentiles con α=0.01, 0.025, 0.05,… ,0.95, 0.975, 0.99,…
- **Construir el intervalo de confianza** de un determinado parámetro[1],
  - sabiendo si se cumplen las condiciones para poder hallarlo,
  - eligiendo el pivote adecuado
  - y aplicando el método general de obtención de intervalos.
- Para un nivel de confianza dado:
  - dar una cota del error absoluto de una estimación
  - estimar un tamaño de la muestra que garantice una precisión determinada para un intervalo de confianza de la media de una población

Problemas de exámenes (web):

SIN: Febrero 2006: Problema 3 a)
Septiembre 2006: Problema 3
Febrero 2005: Problema 3 c)
Junio 2005: Problema 2 d) e)
Septiembre 2005: Problema 3
Junio 2004: Problema 3 c)
Septiembre 2004: Problema 3: 3.
Septiembre 2003: Problema 3 d)

CON: Febrero 2006: Parcial 1. a 5.
Febrero 2005: Modelo 3ª
Junio 2005: g) h)
Septiembre 2005: f) g)

[1] En este curso habrá que saber hallar los siguientes intervalos:
- Para poblaciones normales: media, varianza, diferencia de medias y cociente de varianzas
- Para poblaciones cualesquiera: media, proporción y diferencia de medias.

7.1.- Introducción

En el tema anterior comentamos las diversas formas de inferencia sobre los parámetros de una población:

- **Inferencia paramétrica:**
  - **Estimación puntual:** $\theta = \theta_0$
    $\lambda = 1.25641$ (como λ es la media de una Poisson, una posible estimación es la media muestral)
  - **Estimación por intervalo:** $\theta \in (a,b)$ **con un % de confianza**

    ```
    Confidence Intervals for Llamadas diarias
    ----------------------------------------
    ```
    95.0% confidence interval for mean: 1.25641 +/- 0.39853 [0.857878;1.65494]

    In practical terms we state with 95.0% confidence that the true mean Llamadas diarias is somewhere between 0.857878 and 1.65494.
  - **Contraste de hipótesis: aceptamos $\theta = \theta_0$ frente a $\theta \neq \theta_0$ (ó $\theta > \theta_0$ ó $\theta < \theta_0$), con nivel de significación α.**

En este tema, estudiaremos cómo hacer las **estimaciones por intervalo**, su fundamento teórico y su interpretación.

Veamos un ejemplo primero:

Consideramos la variable *X*: estatura de los alumnos del grupo GM23 (curso 2004/05), y queremos estimar su media.

Para ello consideramos la muestra ($n=39$) tomada al comienzo del curso.

Una ***estimación puntual*** vendría dada de forma natural por la media muestral: $\bar{X} = 174.615$.

Si queremos una ***estimación por intervalo***, una primera idea es buscar un intervalo tal que el valor de la media esté en dicho intervalo con una probabilidad determinada (por ejemplo, 0.95).

i) Para hallar probabilidades, **necesitamos una v.a. de la que conozcamos su distribución**.

En este caso, por la definición de la v.a. *X* podemos considerar que sigue una distribución normal $N(\mu,\sigma)$ (además, la muestra obtenida pasa el contraste de la normal con p-valor de 0.66). Por tanto, por las propiedades de la distribución normal, sabemos que el estimador de la media, $\bar{X}$, también sigue una distribución normal:

$$\bar{X} \sim N\left(\mu\ ,\ \sigma/\sqrt{n}\right)\ \text{, es decir}\ \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

ii) Para hallar el intervalo, podríamos **buscar valores** $K_1, K_2 \in \mathbb{R}$**, tales que** $P(K_1 \leq N(0,1) \leq K_2) = 0.95$**.**

En este caso, $K_1 = -1.96$ *y* $K_2 = 1.96$, y tendríamos que $-1.96 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96$ con probabilidad 0.95.

iii) Si ahora **despejamos** μ, tenemos que $\bar{X} - 1.96 \cdot \sigma/\sqrt{n} \leq \mu \leq \bar{X} + 1.96 \cdot \sigma/\sqrt{n}$ con probabilidad 0.95.

El intervalo al 95% sería $\left(\bar{X} - 1.96 \cdot \sigma/\sqrt{n} \ , \ \bar{X} + 1.96 \cdot \sigma/\sqrt{n} \right)$.

iv) A partir de los datos de la muestra realizada, se tiene que $\bar{X} = 174.615$ y n=39, pero desconocemos el valor de σ. Una solución, es aproximarlo por la cuasidesviación típica de la muestra S=6.243.

Con esos datos, **obtenemos un intervalo** (172.65 , 176.57) para μ con una confianza del 95%.

(El resultado **es aproximado, no es el correcto**. Al aproximar σ por la cuasidesviación típica, la distribución ya no es N(0,1), sino que seguirá otro modelo, que dará lugar a un intervalo diferente. Si lo hallamos con Statgraphics se obtiene: "95.0% confidence interval for mean: [172.59;176.639]", que es muy similar, pero más amplio. Comentaremos este aspecto más adelante.)

Observaciones:

- ¿De qué depende el intervalo de confianza?

  De la estimación del parámetro (en el ejemplo, el valor de $\bar{X}$), del nivel de confianza (pues determina los valores de K_1, K_2) y del tamaño de la muestra (n). Cómo influye cada uno, se comentará más adelante.

- ¿Por qué decimos **confianza** en lugar de probabilidad?

  ¿Es correcto decir que $P(172.65 \leq \mu \leq 176.57) = 0.95$? **No** es correcto hablar de probabilidades. Se pueden calcular probabilidades de v.a., pero μ **no es una v.a**. Por ello, cuando tenemos un intervalo obtenido a partir de los datos de una muestra, se habla de confianza.

- Si tomamos otra muestra de la misma población, ¿habríamos obtenido el mismo intervalo? ¿**es único el intervalo de confianza**?

  La **fórmula** para obtenerlo **sí es única**, pero los valores concretos que toma el intervalo dependen de los datos de la muestra y de su tamaño, por lo que para cada muestra se obtendrá un intervalo de confianza diferente.

Método de obtención de intervalos de confianza:

En general, para hallar **IC(θ)**: intervalo de confianza para un parámetro θ con un nivel de confianza de *1-*α, seguiremos los mismos pasos que en el ejemplo anterior. Formalizamos el método de la siguiente forma:

i) **Elección de un pivote** $X^* = X^*(X_1,\ldots,X_n;\theta)$ que:
- depende de los datos de la **muestra** y del parámetro θ: $X^* = X^*(X_1,\ldots,X_n;\theta)$
- con **distribución de probabilidad conocida que no depende de** θ.

ii) Encontramos **valores** $K_1, K_2 \in$ **R, tales que** $P(K_1 \le X^* \le K_2) = 1-\alpha$ **.**

iii) **Despejamos** θ en la expresión anterior para obtener $T_1 = T_1(X_1,\ldots,X_n)$, $T_2 = T_2(X_1,\ldots,X_n)$, tales que $P(T_1 \le \theta \le T_2) = 1-\alpha$.

iv) El intervalo de confianza buscado es **IC(θ)**: (T_1, T_2).

Observación:

Distinguimos entre la ***fórmula*** *para hallar el intervalo de confianza* (**el** intervalo de confianza), y el **intervalo** de números reales **obtenido a partir de los datos de una muestr**a (**un** intervalo de confianza).

Antes de obtener los datos (n)	**Después de obtener los datos (n)**
Como $T_1 = T_1(X_1,\ldots,X_n)$, $T_2 = T_2(X_1,\ldots,X_n)$ son **estadísticos** (y por tanto **v.a.**), tiene sentido interpretar $1-\alpha$ como una **probabilidad**: $P(T_1 \le \theta \le T_2) = 1-\alpha$.	A partir del resultado de la muestra $x_1,\ldots,x_n \in \mathbb{R}$, se tiene que los extremos del intervalo $t_i = T_i(x_1, \ldots, x_n) \in$ **R** (son **números reales**) por lo que no tiene sentido hablar de probabilidad $P(t_1 \le \theta \le t_2) = 1-\alpha$. Por eso, $1-\alpha$ se interpreta como **nivel de confianza**.
Se dice que (T_1, T_2) es **el** intervalo de confianza para θ, al $(1-\alpha)\cdot 100\%$ (también se habla del ***nivel de significación*** α, que se interpreta como la probabilidad de que θ esté fuera de dicho intervalo)	Se dice que (t_1, t_2) es **un** intervalo de confianza para θ, pues si se obtiene otra muestra, el resultado sería distinto y se tiene **otro** intervalo de confianza para θ. Una **confianza** del 95% significa que si obtenemos 100 intervalos de confianza para θ, el verdadero valor de θ estaría en 95 de ellos.

Notación: En adelante, utilizaremos la siguiente notación:

$X_1, \ldots, X_n$: muestra aleatoria simple de la v.a. X.

$\bar{X}$: media muestral ($\bar{X} = \dfrac{X_1 + \ldots + X_n}{n}$)

S^2 : cuasivarianza muestral ($S^2 = \dfrac{(X_1 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2}{n-1}$).

$S = \sqrt{S^2}$: cuasidesviación típica muestral

7.2.- Intervalos de confianza en poblaciones normales.

En este apartado consideramos siempre que las variables estudiadas siguen una **distribución normal**. Estudiaremos:

- cómo hallar los intervalos de confianza para la **media y la varianza** de una sola variable y
- cómo hallar intervalos de confianza para **comparar las medias y las varianzas de dos variables**,

7.2.1.- Intervalos de confianza para los parámetros de una sola variable.

En este apartado consideramos $X \sim N(\mu,\sigma)$. Estudiaremos los siguientes intervalos:

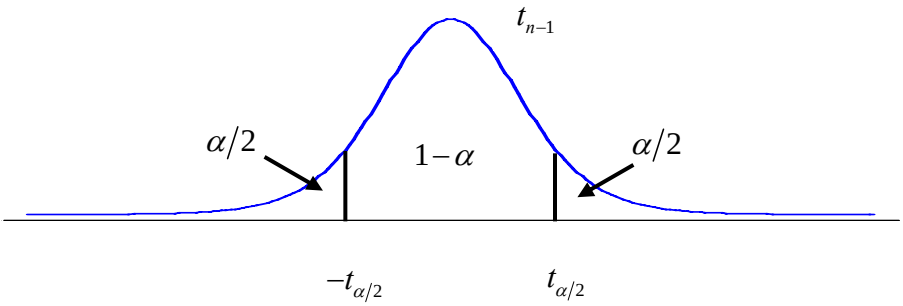
- IC(μ): Intervalo de confianza para la media
- IC(σ^2): Intervalo de confianza para la varianza

Para cada caso, damos los pivotes adecuados y la expresión final del intervalo.

IC(μ): Intervalo de confianza para la media

En la introducción vimos que si $X \sim N(\mu,\sigma)$, entonces $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, pero en la práctica no se suele conocer el valor de σ, por lo que utilizaremos la cuasidesviación típica *S*, dando lugar a otra expresión que sigue un modelo de distribución que no es N(0,1).

Pivote: $X^* = \dfrac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$ ([2]) ; **IC(μ):** $\bar{X} \pm t_{\alpha/2} S/\sqrt{n}$, donde $\mathrm{P}(t_{n-1} \le t_{\alpha/2}) = 1-\dfrac{\alpha}{2}$.



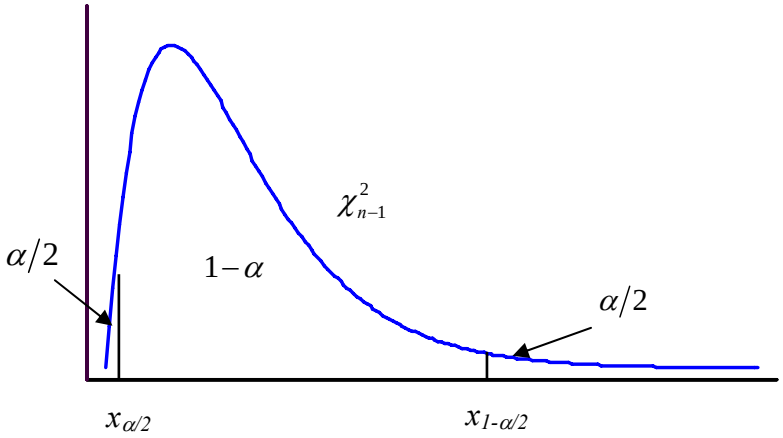
[2] Ver Anexo 1: Distribuciones χ^2_n y t_n.

$\chi^2_n \sim \sum_{i=1}^{n} X_i^2$, si $X_1,\ldots,X_n$ son v.a. independientes con distribución $N(0,1)$

$t_n \sim \dfrac{X}{\sqrt{\dfrac{Y}{n}}}$, si $X \sim N(0,1)$, $Y \sim \chi^2_n$ y son v.a. independientes

IC(σ^2): Intervalo de confianza para la varianza

Pivote: $X^* = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ ([3]) ; ***IC(σ^2):*** $\left(\dfrac{(n-1)S^2}{x_{1-\alpha/2}}, \dfrac{(n-1)S^2}{x_{\alpha/2}} \right)$, donde $\mathrm{P}\left(\chi^2_{n-1} \le x_p\right) = p$



Ejemplo:

Consideramos la variable *Peso de los estudiantes de la EUI*, y los datos obtenidos de la muestra tomada en el grupo SM22 (curso 2006/07). A partir de dichos datos, queremos obtener **un** intervalo de confianza, al 95% de confianza, para la estatura **media**, para su **varianza** y su **desviación típica**.

1. En primer lugar, **comprobamos** que es correcto suponer que dicha variable sigue una **distribución normal.**

Para ello, utilizamos la opción *Distribution Fitting* de Statgraphics, y obtenemos:
Approximate P-Value = 0,816771 **>0.3**, por lo que podemos aceptar que sigue una distribución normal, y podemos utilizar los pivotes vistos anteriormente para hallar los intervalos de confianza.

2. Intervalo de confianza para la media

Seguimos los pasos habituales:

i) **Elección del pivote X^*:**

Por ser I.C. de la media de una población normal, elegimos $X^* = \dfrac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$.

Como $n=49$, $X^* \sim t_{48}$.

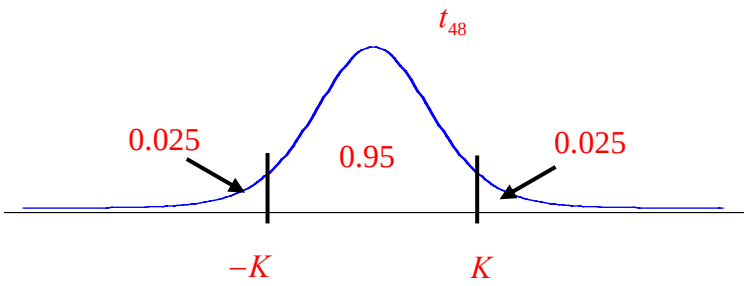
ii) Encontramos **valores $K_1, K_2 \in \mathbb{R}$, tales que $P\left(K_1 \le X^* \le K_2\right) = 1-\alpha$.**

En este caso, buscamos **$K_1, K_2 \in \mathbb{R}$, tales que $P\left(K_1 \le t_{48} \le K_2\right) = 0.95$.**

---

[3] Ver Anexo 1: $\chi^2_n \sim \sum_{i=1}^{n} X_i^2$, si $X_1, \ldots, X_n$ son v.a. independientes con distribución $N(0,1)$

Por la simetría de la *t-Student*, buscamos $K \in \mathbb{R}$, **tal que** $P(-K \le t_{48} \le K) = 0.95$.



Para obtener K, se pueden usar las tablas o Statgraphics, buscando K tal que $P(t_{48} \le K) = 0.975$.

En este caso, las tablas no incluyen el valor *n=48*, por lo que utilizamos Statg (*Describe/Distributions/Probability Distributions/Inverse CDF*):

Inverse CDF

Distribution: Student's t

CDF	Dist. 1 (t_{48})
0,025	-2,01064
0,975	2,01064

Por tanto, consideramos $K = 2.01064$.

iii) **Despejamos** μ en la expresión $-K \le X^* \le K$:

Partimos de $-K \le \dfrac{\bar{X} - \mu}{S/\sqrt{n}} \le K$, y se llega a $\bar{X} - K \cdot S/\sqrt{n} \le \mu \le \bar{X} + K \cdot S/\sqrt{n}$

iv) El intervalo de confianza al 95% es $\left(\bar{X} - K \cdot S/\sqrt{n} \ , \bar{X} + K \cdot S/\sqrt{n} \right)$.

Con los **datos de la muestra**, tenemos que: $n = 49, \bar{X} = 72.1837, S = 13.4485$. Además, $K = 2.01064$.

Sustituyendo esos valores en la expresión del intervalo, se tiene: **[68,3208; 76,0465]**.

El peso medio está en el intervalo **[68,3208; 76,0465]** al 95% de confianza.

3. Intervalo de confianza para la varianza (y la desviación típica)

Seguimos los pasos habituales:

i) **Elección del pivote** X^***:**

Por ser I.C. de la varianza de una población normal, elegimos $X^* = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

Como *n=49*, $X^* \sim \chi^2_{48}$.

ii) Encontramos **valores** $K_1, K_2 \in \mathbb{R}$, **tales que** $P(K_1 \le X^* \le K_2) = 1 - \alpha$.

En este caso, buscamos $K_1, K_2 \in \mathbb{R}$, **tales que** $P\left(K_1 \le \chi^2_{48} \le K_2\right) = 0.95$.

Para obtener K_1, K_2, se pueden usar las tablas o Statgraphics, buscando K_1, K_2 tales que $P\left(\chi^2_{48} \le K_1\right) = 0.025, P\left(\chi^2_{48} \le K_2\right) = 0.975$.

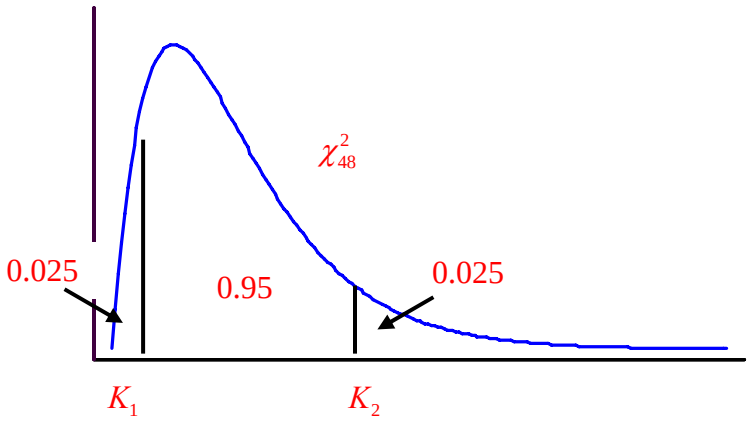
En este caso, las tablas tampoco incluyen el valor *n=48*, por lo que utilizamos Statg (*Describe/Distributions/Probability Distributions/Inverse CDF*):

Inverse CDF

Distribution: Chi-Square (48)

CDF	Dist. 1
0,025	30,7545
0,975	69,0226

Por lo que consideramos $K_1 = 30.7545,\ K_2 = 69.0226$.



iii) **Despejamos σ^2** en la expresión $K_1 \le X^* \le K_2$:

Partimos de $K_1 \le \dfrac{(n-1)S^2}{\sigma^2} \le K_2 \Leftrightarrow \dfrac{K_1}{(n-1)S^2} \le \dfrac{1}{\sigma^2} \le \dfrac{K_2}{(n-1)S^2} \underset{invertimos}{\Leftrightarrow} \dfrac{(n-1)S^2}{K_1} \ge \sigma^2 \ge \dfrac{(n-1)S^2}{K_2}$.

iv) El intervalo de confianza es $\left(\dfrac{(n-1)S^2}{K_2}, \dfrac{(n-1)S^2}{K_1}\right)$.

Con los **datos de la muestra**, tenemos que: $n = 49,\ S^2 = 180.861$ *(Variance)*.

Sustituyéndolos en la expresión del intervalo, junto con los valores de K_1, K_2, se tiene:

[125.775; 282.278].

La varianza del peso está en el intervalo **[125.775; 282.278]** al 95% de confianza.

Si queremos obtener el **intervalo de confianza para la desviación típica σ**, en el paso iii) **despejamos** σ en lugar de σ^2:

$$K_1 \le \frac{(n-1)S^2}{\sigma^2} \le K_2 \Leftrightarrow \ldots \Leftrightarrow \frac{(n-1)S^2}{K_2} \le \sigma^2 \le \frac{(n-1)S^2}{K_1} \underset{\text{raíces}}{\overset{\text{tomamos}}{\Leftrightarrow}} \sqrt{\frac{(n-1)S^2}{K_2}} \le \sigma \le \sqrt{\frac{(n-1)S^2}{K_1}}$$

Por tanto, el intervalo de confianza obtenido a partir de los datos de la muestra es:

$$\left[\sqrt{125.775};\sqrt{282.278}\right] = \textbf{[11.215; 16.801]}.$$

La desviación típica del peso está en el intervalo **[11.215; 16.801]** al 95% de confianza.

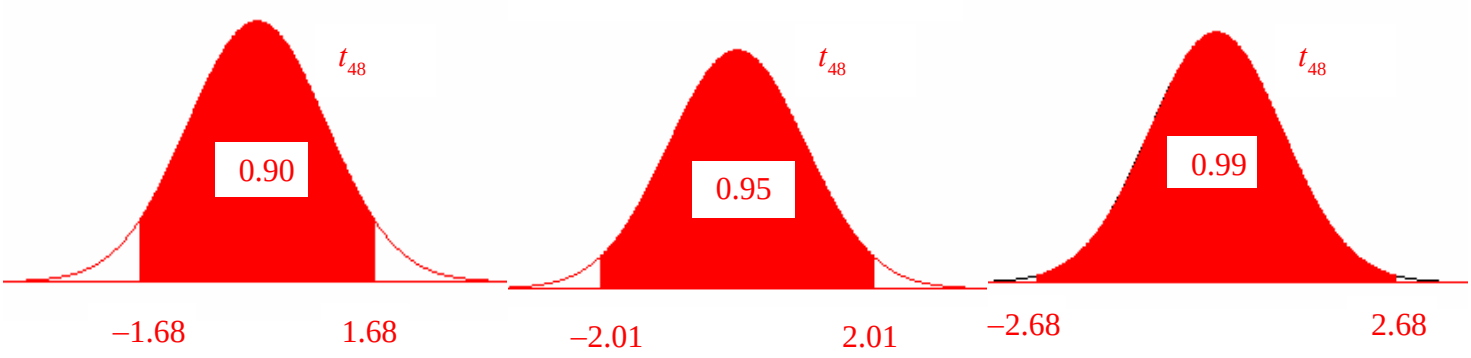
Observaciones.

1. ¿Cómo cambia el intervalo si cambia el nivel de confianza?

Si hallamos el intervalo de confianza para la media de la variable *Peso* con confianza del 90%, 95% y 99% se obtiene:

IC(μ) al 90%	IC(μ) al 95%	IC(μ) al 99%
[68,9614; 75,406]	[68,3208; 76,0465]	[67,0306;77,3368]

Observamos que **según aumenta el nivel de confianza, el intervalo se hace más amplio**.



Esto se justifica porque el nivel de confianza se utiliza en el paso ii), determinando los valores de $K_1, K_2 \in \mathbb{R}$, tales que $P(K_1 \le X^* \le K_2) = 1-\alpha$. Es claro que si aumenta el nivel de confianza, la amplitud del intervalo (K_1, K_2) será mayor, y por tanto también lo será la del intervalo de confianza. (Si consideramos un 100% de confianza, tendríamos el intervalo $(-\infty, +\infty)$)

2. Fijado un nivel de confianza, ¿cómo podemos hacer que el intervalo de confianza nos dé una información más precisa (sea más pequeño)?

Aumentando el tamaño de la muestra (*n*).

Por ejemplo, suponemos que la muestra elegida para estudiar la variable *Peso* es de tamaño *n=200*, y obtenemos el mismo valor de $S^2 = 180.861$.

Si hallamos el intervalo de confianza para la desviación típica, obtenemos **[12.27; 14.95]**[4], que es un intervalo más pequeño que el obtenido con *n=49* (**[11.215; 16.801]**), y nos da una información más precisa sobre los valores entre los que puede estar el verdadero valor de la desviación típica.

3. ¿Qué información podemos extraer de un intervalo de confianza?

Hemos visto que el peso medio está en el intervalo **[68,3208; 76,0465]** al 95% de confianza.
De ahí, con un nivel de confianza del 95%:

- podemos **afirmar** que el peso medio es mayor que 68 kg o bien que es menor que 77 Kg. (Como $\mu \in [68{,}3208;\ 76{,}0465]$, se tiene que **68 <** $68{,}3208 \le \mu \le 76{,}0465$ **< 77**),
- sin embargo **no podríamos asegurar**, al 95%, que el peso medio sea mayor que 69 kg, pues en el intervalo de posibles valores del peso medio hay valores que son menores que 69.
- es **aceptable** suponer que el peso medio pueda ser 70 Kg (en general, cualquier valor del interior del intervalo).

En general, si el $IC(\theta)$ al nivel de confianza $(1-\alpha)$ es (a,b), con nivel de confianza $(1-\alpha)$:

- se puede **asegurar** que $\theta > a$, y que cualquier valor menor que a,
- se puede **asegurar** que $\theta < b$, y que cualquier valor mayor que b,
- es **aceptable** suponer que $\theta = c$, siendo $c \in (a,b)$.

7.2.2.- Intervalos de confianza para comparar parámetros de dos variables independientes con distribución normal

Como ejemplo, queremos estudiar las diferencias entre chicos y chicas de la EUI respecto de las variables *Peso* y *Estatura*.

Para ello, consideramos una muestra (grupo SM22, curso 2006/07), de tamaño *n=49*, y denotamos las variables de la siguiente forma:

P_X: peso de los chicos; P_Y: peso de los chicas; E_X: estatura de los chicos; E_Y: estatura de las chicas.

Se obtienen los siguientes datos:

Medias	$\overline{P}_X = 76.51 kg$	$\overline{P}_Y = 55.3 kg$	$\overline{E}_X = 179.67 m$	$\overline{E}_Y = 163.4 m$
Desviaciones típicas	$S_{P_X} = 11.26 kg$	$S_{P_Y} = 5.52 kg$	$S_{E_X} = 6.25 m$	$S_{E_Y} = 6.59 m$

A partir del estudio de intervalos de confianza, intentaremos dar respuesta a preguntas como:

- ¿Se puede afirmar que la diferencia entre los pesos medios es de más de 20 kg?
- ¿Se puede afirmar que la diferencia entre las estaturas medias es de menos de 20 cm?
- ¿Se puede afirmar que la variabilidad respecto de la media de la variable *Estatura* es igual en los chicos que en las chicas?
- ¿Se puede afirmar que la variabilidad respecto de la media de la variable *Peso* es igual en los chicos que en las chicas?

En este apartado utilizaremos la siguiente **notación**:

- *X* y *Y* son v.a. independientes con distribución normal: $X \sim N(\mu_X, \sigma_X)$, $Y \sim N(\mu_Y, \sigma_Y)$.
- $(X_1, \ldots, X_{nx})$, $(Y_1, \ldots, Y_{ny})$: m.a.s. de *X* e *Y*
- n_x : tamaño de la muestra de *X*
- n_y : tamaño de la muestra de *Y*

[4] Para la χ^2_{199}, al 95%, se obtienen unos valores de $K_1 = 161.822$, $K_2 = 239.962$. Sustituyendo en la expresión general del intervalo de confianza, junto con $n=200$ y $S^2 = 180.861$, se obtiene el intervalo [12.27; 14.95].

- $\overline{X}, \overline{Y}$: medias muestrales de *X* e *Y*
- S_X^2 , S_Y^2 : cuasivarianzas muestrales de *X* e *Y*

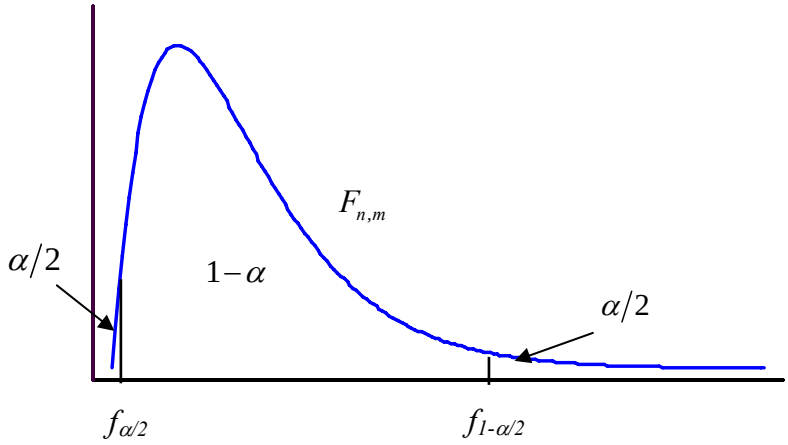
Estudiaremos los siguientes intervalos:

- IC(σ_X^2 / σ_Y^2): Intervalo de confianza para el **cociente de varianzas de v.a. independientes** con distribución normal.
- IC($\mu_X - \mu_Y$): Intervalo de confianza para la **diferencia de medias de v.a. independientes** con distribución normal.

IC(σ_X^2 / σ_Y^2): Intervalo de confianza para el cociente de varianzas de v.a. independientes con distribución normal

Pivote: $X^* = \dfrac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} = \dfrac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} \sim F_{n_X-1, n_Y-1}$, (*Ver Anexo 2: Distribución $F_{n,m}$* ([5])) ;

IC(σ_X^2 / σ_Y^2): $\left(\dfrac{S_X^2/S_Y^2}{f_{1-\alpha/2}} , \dfrac{S_X^2/S_Y^2}{f_{\alpha/2}} \right)$, donde $P\left(F_{n_X-1, n_Y-1} \le f_{1-\alpha/2}\right) = 1 - \dfrac{\alpha}{2}$ y $P\left(F_{n_X-1, n_Y-1} \le f_{\alpha/2}\right) = \dfrac{\alpha}{2}$



Criterio para decidir si se puede aceptar que las varianzas son iguales (nivel 1-α):

Si las varianzas fueran iguales, el valor de σ_X^2 / σ_Y^2 sería **1**. Por tanto, observamos si a partir de los datos de la muestra el 1 es uno de los posibles valores para dicho cociente. Si no está en el intervalo, significa que los datos de la muestra no corroboran la suposición de que sean iguales. Por tanto:

- Si $1 \in IC(\sigma_X^2 / \sigma_Y^2)$, podemos aceptar que son iguales.
- Si $1 \notin IC(\sigma_X^2 / \sigma_Y^2)$, podemos considerar que son diferentes.

[5] $F_{n,m} \sim \dfrac{X/n}{Y/m}$, si $X \sim \chi_n^2$, $Y \sim \chi_m^2$ y son v.a. independientes

IC($\mu_X - \mu_Y$): Intervalo de confianza para la diferencia de medias de v.a. independientes con distribución normal

- ***Se puede suponer que $\sigma_X^2 = \sigma_Y^2$ (igualdad de varianzas):***

Pivote: $X^* = \dfrac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{S_p\sqrt{\dfrac{1}{n_X}+\dfrac{1}{n_Y}}} \sim t_{n_X+n_Y-2}$ donde $S_p = \sqrt{\dfrac{(n_X-1)S_X^2+(n_Y-1)S_Y^2}{n_X+n_Y-2}}$;

IC($\mu_X - \mu_Y$): $(\overline{X}-\overline{Y}) \pm t_{\alpha/2} S_p \sqrt{\dfrac{1}{n_X}+\dfrac{1}{n_Y}}$, donde $\mathrm{P}\left(t_{n_X+n_Y-2} \le t_{\alpha/2}\right) = 1-\dfrac{\alpha}{2}$

- ***No se puede suponer que $\sigma_X^2 = \sigma_Y^2$ (varianzas distintas):***

Pivote: $X^* = \dfrac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{\sqrt{\dfrac{S_X^2}{n_X}+\dfrac{S_Y^2}{n_Y}}} \approx t_f$; donde el número de grados de libertad *f* viene dado por la aproximación de Welch: $f = \dfrac{\left(\dfrac{S_X^2}{n_X}+\dfrac{S_Y^2}{n_Y}\right)^2}{\dfrac{\left(\dfrac{S_X^2}{n_X}\right)^2}{n_X+1}+\dfrac{\left(\dfrac{S_Y^2}{n_Y}\right)^2}{n_Y+1}} - 2$.

IC($\mu_X - \mu_Y$): $(\overline{X}-\overline{Y}) \pm t_{\alpha/2}\sqrt{\dfrac{S_X^2}{n_X}+\dfrac{S_Y^2}{n_Y}}$, donde $\mathrm{P}\left(t_f \le t_{\alpha/2}\right) = 1-\dfrac{\alpha}{2}$

Criterio para decidir si se puede aceptar que las medias son iguales (confianza 1-α):

Si las medias fueran iguales, el valor de $\mu_X - \mu_Y$ sería **0**. Por tanto, observamos si a partir de los datos de la muestra el 0 es uno de los posibles valores para dicha diferencia. Si no está en el intervalo, significa que los datos de la muestra no corroboran la suposición de que sean iguales. Por tanto:

- Si $0 \in IC(\mu_X - \mu_Y)$, podemos aceptar que son iguales.
- Si $0 \notin IC(\mu_X - \mu_Y)$, podemos considerar que son diferentes.

Conclusión:

Para estudiar la igualdad de medias de dos v.a. aleatorias independientes con distribución normal, seguiremos los siguientes pasos:

a) Comprobar que ambas variables siguen una distribución normal.

b) **Estudiar** si se puede suponer o no ***igualdad de*** **varianzas**:
   - hallar IC(σ_X^2 / σ_Y^2) y
   - estudiar si contiene o no al 1.

c) **Estudiar** la ***igualdad de*** **medias**:
   - hallar el *IC*($\mu_X - \mu_Y$) que corresponda y
   - estudiar si contiene o no al 0.

Ejemplo 1: Estudiamos las diferencias entre chicos y chicas de la EUI respecto de la variable *Peso.* (95% de confianza)

a) Comprobamos que las variables P_X (peso de los chicos) y P_Y (peso de las chicas), siguen una distribución normal.

Con la opción *Distribution Fitting* de Statgraphics obtenemos:

P_X : Approximate P-Value = 0,406201 > 0.3 (aceptamos que es normal)
P_Y : Approximate P-Value = 0,908182 > 0.3 (aceptamos que es normal)

Luego ambas variables pueden considerarse normales.

b) Estudiamos si las varianzas se pueden considerar iguales o no (95% de confianza).

Hallamos el intervalo de confianza al 95% para el cociente de varianzas según el método habitual:

i) El pivote es $X^* = \dfrac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} = \dfrac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} \sim F_{n_X-1,n_Y-1}$.

Consideramos $\dfrac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} \sim F_{38,9}$, pues $n_X=39$ y $n_Y=10$.

ii) Hallamos $K_1, K_2 \in \mathbb{R}$, tales que $P\left(K_1 \le F_{38,9} \le K_2\right) = 0.95$.

Para obtener K_1, K_2, se pueden usar las tablas[6] o Statgraphics, buscando K_1, K_2 tales que $P\left(F_{38,9} \le K_1\right) = 0.025, P\left(F_{38,9} \le K_2\right) = 0.975$.

En este caso, las tablas no incluyen el valor $n=38$, por lo que utilizamos Statg (*Describe/Distributions/Probability Distributions/Inverse CDF*):

Distribution: F (variance ratio) (n=38, m=9)

CDF	Dist. 1
0,025	0,404701
0,975	3,51423

[6] Ver en el Anexo 2 cómo buscar valores en las tablas de la $F_{n,m}$.

Por lo que consideramos $K_1 = 0.404701,\ K_2 = 3.51423$.

iii) Despejamos *el cociente de varianzas* en la expresión $K_1 \le X^* \le K_2$:

Partimos de $K_1 \le \dfrac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} \le K_2 \Leftrightarrow \dfrac{K_1}{S_X^2/S_Y^2} \le \dfrac{1}{\sigma_X^2/\sigma_Y^2} \le \dfrac{K_2}{S_X^2/S_Y^2} \underset{invertimos}{\Leftrightarrow} \dfrac{S_X^2/S_Y^2}{K_1} \ge \sigma_X^2/\sigma_Y^2 \ge \dfrac{S_X^2/S_Y^2}{K_2}$.

iv) El intervalo de confianza del cociente σ_X^2/σ_Y^2 es $\left(\dfrac{S_X^2/S_Y^2}{K_2}, \dfrac{S_X^2/S_Y^2}{K_1}\right)$.

Sustituyendo los valores de K_1, K_2, y los de $S_{P_X} = 11.26\,kg$ y $S_{P_Y} = 5.52\,kg$, tenemos el intervalo: **(1.18 ,10.28)**.

Como **$1 \notin$ (1.18 ,10.28)**, **no podemos considerar iguales las varianzas**.

c) Comparamos las medias.

Hallamos el intervalo de confianza (95%) teniendo en cuenta que **no** podemos suponer igualdad de varianzas.

i) En este caso el pivote es $X^* = \dfrac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\dfrac{S_X^2}{n_X} + \dfrac{S_Y^2}{n_Y}}} \sim t_f$.

Hallamos $f = \dfrac{\left(\dfrac{S_X^2}{n_X} + \dfrac{S_Y^2}{n_Y}\right)^2}{\dfrac{\left(\dfrac{S_X^2}{n_X}\right)^2}{n_X+1} + \dfrac{\left(\dfrac{S_Y^2}{n_Y}\right)^2}{n_Y+1}} - 2 = \dfrac{\left(\dfrac{11.26^2}{39} + \dfrac{5.52^2}{10}\right)^2}{\dfrac{\left(\dfrac{11.26^2}{39}\right)^2}{39+1} + \dfrac{\left(\dfrac{5.52^2}{10}\right)^2}{10+1}} - 2 = 33.79$

ii) Por la simetría de la t-Student, hallamos $K \in \mathbb{R}$, tal que $P(-K \le t_f \le K) = 0.95$.

Se pueden usar las tablas o Statgraphics, buscando K tal que $P(t_f \le K) = 0.975$.

En este caso, ni las tablas ni Statgraphics admiten el valor *n=33.79*, por lo que redondeamos a *n=34* y utilizamos Statg (*Describe/Distributions/Probability Distributions/Inverse CDF*):

Distribution: Student's t (n=34)

CDF	Dist. 1
0,975	2,03225

Por lo que consideramos $K = 2.03225$.

iii) Despejamos $(\mu_X - \mu_Y)$ en la expresión $-K \le X^* \le K$, y obtenemos la expresión del intervalo de confianza: $(\overline{X} - \overline{Y}) \pm K\sqrt{\dfrac{S_X^2}{n_X} + \dfrac{S_Y^2}{n_Y}}$

Sustituyendo los valores de $K = 2.03225$, y los de n_X=*39*, n_Y=*10*, $\overline{P}_X = 76.51\,kg$, $\overline{P}_Y = 55.3\,kg$ $S_{P_X} = 11.26\,kg$ y $S_{P_Y} = 5.52\,kg$, tenemos el intervalo: **(16.11 ,26.31)**.

Como era de esperar, **0 ∉ (16.11 ,26.31)**, **no podemos considerar iguales las medias**.

Una vez que tenemos los intervalos de la diferencia de medias y el cociente de varianzas, sí podemos contestar a las preguntas planteadas al principio:

- ¿Se puede afirmar que la diferencia entre los pesos medios es de más de 20 kg?

  Con un 95% de confianza, no se puede garantizar que la diferencia sea de más de 20 kg, pues el posible valor de dicha diferencia está en el intervalo **(16.11 ,26.31)**, que incluye valores menores a 20, por lo que dicha diferencia podría ser menor que 20.

  Sí se puede garantizar que la diferencia del peso medio es mayor de 15 kg. ($15<16.11<\mu_X-\mu_Y$).

- ¿Se puede afirmar que la variabilidad respecto de la media de la variable *Peso* es igual en los chicos que en las chicas?

  Como se ha visto al hallar el intervalo del cociente de varianzas, no se pueden suponer iguales. Además, como IC(σ_X^2/σ_Y^2) es **(1.18 ,10.28)**, se tiene que $\sigma_X^2/\sigma_Y^2>1.18>1$, y por tanto $\sigma_X^2>\sigma_Y^2$.

  Es decir, se puede asegurar, al 95% de confianza, que los chicos tienen una mayor variabilidad en el peso que las chicas.

Ejemplo 2: Estudiamos las diferencias entre chicos y chicas de la EUI respecto de la variable *Estatura.* (95% de confianza)

a) Comprobamos que las variables E_X (estatura de los chicos) y E_Y (estatura de las chicas), siguen una distribución normal.

Con la opción *Distribution Fitting* de Statgraphics obtenemos:

E_X : Approximate P-Value = 0,630571 > 0.3 (aceptamos que es normal)
E_Y : Approximate P-Value = 0,85326 > 0.3 (aceptamos que es normal)

Luego ambas variables pueden considerarse normales.

b) Estudiamos si las varianzas se pueden considerar iguales o no (95% de confianza).

Hallamos el intervalo de confianza al 95% para el cociente de varianzas según el método habitual.

Con Statgraphics se obtiene: Ratio of Variances: **[0,256645;2,22858]**

Como **1** ∈ [0,256645; 2,22858], **sí podemos considerar iguales las varianzas**.

c) Comparamos las medias.

Hallamos el intervalo de confianza (95%) teniendo en cuenta que **sí** podemos suponer igualdad de varianzas.

i) En este caso el pivote es $X^* = \dfrac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{S_p\sqrt{\dfrac{1}{n_X}+\dfrac{1}{n_Y}}} \sim t_{n_X+n_Y-2}$ donde

$$S_p = \sqrt{\frac{(n_X-1)S_X^2+(n_Y-1)S_Y^2}{n_X+n_Y-2}}$$

ii) Por la simetría de la t-Student, hallamos $K \in \mathbb{R}$, tal que $P(-K \le t_{47} \le K) = 0.95$.

Buscando K tal que $P(t_{47} \le K) = 0.975$, utilizamos Statg (*Describe/Distributions/Probability Distributions/Inverse CDF*):

Distribution: Student's t (n=47)

CDF	Dist. 1
0,975	2,01174

Por lo que consideramos $K = 2.01174$.

iii) Despejamos $(\mu_X - \mu_Y)$ en la expresión $-K \le X^* \le K$, y obtenemos la expresión del intervalo de confianza:

$$(\overline{X}-\overline{Y}) \pm K \cdot \sqrt{\frac{(n_X-1)S_X^2+(n_Y-1)S_Y^2}{n_X+n_Y-2}} \cdot \sqrt{\frac{1}{n_X}+\frac{1}{n_Y}},$$

Sustituyendo los valores obtenidos en la muestra, se tiene:

$$(179.67-163.4) \pm 2.01174 \cdot \sqrt{\frac{(39-1)6.25^2+(10-1)6.59^2}{39+10-2}} \cdot \sqrt{\frac{1}{39}+\frac{1}{10}},$$

que nos da el intervalo: **(11.76, 20.77).**

Como era de esperar, **0 ∉ (11.76, 20.77)**, **no podemos considerar iguales las medias**.

Una vez que tenemos los intervalos de la diferencia de medias y el cociente de varianzas, sí podemos contestar a las preguntas planteadas al principio:

- ¿Se puede afirmar que la diferencia entre las estaturas medias es de menos de 20 cm?

  Con un 95% de confianza, no se puede garantizar que la diferencia sea de menos de 20 cm, pues el posible valor de dicha diferencia está en el intervalo **(11.76, 20.77)**, que incluye valores mayores a 20, por lo que dicha diferencia podría ser mayor que 20. (Si hacemos el intervalo con un **89%** de confianza, se obtiene [12.6176, 19.9157], y sí podríamos afirmar, con ese nivel de confianza, que la diferencia de las estaturas medias es de menos de 20 cm.)

  Al 95% sí se puede garantizar que la diferencia de las estaturas medias es de más de 10 cm ($10 < 11.76 < \mu_X - \mu_Y$) y de menos de 21 cm ($\mu_X - \mu_Y < 20.77 < 21$).

- ¿Se puede afirmar que la variabilidad respecto de la media de la variable *Estatura* es igual en los chicos que en las chicas?

  Como se ha visto al hallar el intervalo del cociente de varianzas, sí se pueden suponer varianzas iguales. Es decir, se puede suponer que la estatura de los chicos tiene la misma variabilidad que la de las chicas, o al menos no se puede asegurar que sean distintas.

7.2.3.- Intervalo de confianza para la diferencia de medias de v.a. con distribución normal y muestras pareadas.

Ejemplo:

1) Se quiere estudiar los efectos de una dieta al cabo de 6 meses; para ello se mide el peso de las personas antes de iniciar la dieta y el peso de esas mismas personas al cabo de 6 meses de dieta. En este caso, ambos pesos **no** son variables independientes, pues es claro que el peso al cabo de 6 meses dependerá del peso que se tenía anteriormente.
2) Se quiere comparar la velocidad de ejecución de dos algoritmos. Se podría hacer utilizando cada uno de los algoritmos sobre dos conjuntos aleatorios de problemas independientes. Pero en ese caso, el experimento podría verse afectado por el tipo de problemas que hayan sido elegidos para cada algoritmo. Esto puede evitarse si a cada tipo de problema se le aplica un algoritmo y después el otro.

En este apartado, se estudia el caso en el que *X* e *Y* son v.a. tales que $X \sim N(\mu_X, \sigma_X)$, $Y \sim N(\mu_Y, \sigma_Y)$, que representan características diferentes de la **misma población**.
Para contrastar sus diferencias, conviene tomar ***muestras pareadas***:

- se obtiene el valor de *X* e *Y* sobre los **mismos** individuos de la población
- el **tamaño de la muestra (*n*) es igual para *X* e *Y***
- se define la **v.a. *diferencia* *D* = *X* – *Y***, sigue un modelo de distribución normal $N(\mu_D, \sigma_D)$, donde $\mu_D = \mu_X - \mu_Y$
- $D_i = X_i - Y_i$ es una m.a.s. de la variable diferencia $D = X - Y$.
- $\bar{D}$ es la media muestral de $D = X - Y$ ($\bar{D} = \bar{X} - \bar{Y}$).
- S_D^2 es la cuasivarianza muestral de $D = X - Y$.
  ($S_D^2 \neq S_X^2 + S_Y^2$, pues no hay independencia de las variables.)

Ejemplo:

X: peso de una persona antes de comenzar la dieta
Y: peso de una persona a los 6 meses de comenzar la dieta
D: diferencia de peso al cabo de 6 meses de comenzar la dieta

Se toma una muestra de 15 personas y se obtienen los siguientes datos:

X	78	86	98	78	86	65	70.5	76	99	102	64	88	67.5	92	101
Y	76	85	99	73	85	63	65	75	96.5	90	61	85.5	68	90	95
D	2	1	-1	5	1	2	5.5	1	2.5	12	3	2.5	-0.5	2	6

Con esos datos se obtiene que: $\bar{X} = 83.4$, $\bar{Y} = 80.4667$, $\bar{D} = 2.9333$, $S_D = 3.20639$. De hecho, se trabajará sólo con los datos de la variable *D*.

Para estudiar la diferencia de medias ($\mu_X - \mu_Y = \mu_D$), podemos hallar el **intervalo de confianza para μ_D** , es decir un intervalo de confianza para la media de una población normal (ver 7.2.1). Tenemos:

Pivote: $X^* = \dfrac{\bar{D} - \mu_D}{S_D/\sqrt{n}} \sim t_{n-1}$; ***IC(μ):*** $\bar{D} \pm t_{\alpha/2} S_D/\sqrt{n}$, donde $\mathrm{P}\left(t_{n-1} \le t_{\alpha/2}\right) = 1 - \dfrac{\alpha}{2}$.

Para decidir si se puede considerar que las medias μ_X y μ_Y son iguales (nivel de confianza 1-α):

- Si 0 ∈ *IC*(μ_D), podrían ser iguales.
- Si 0 ∉ *IC*(μ_D), no podemos considerar que son iguales.

En el ejemplo, para un nivel de confianza del 95% se obtiene el intervalo [1.15769, 4.70898], por lo que no podemos considerar que son iguales, y **hay una diferencia de peso significativa**.

¿Sería correcto garantizar que se adelgazan 3 kg? ¿Y que se adelgaza al menos 1 kg?

7.3.- Intervalos de confianza en poblaciones no normales.

Cuando las variables que estudiamos no siguen una distribución normal, nos encontramos con la dificultad de conocer cuál es la distribución del estimador.

Sin embargo, cuando queremos estimar la **media de la población**, podemos aplicar el **Teorema Central del Límite**[3], que nos permite aproximar la distribución de la media o de la diferencia de medias por una distribución normal.

En esta sección estudiaremos cómo hallar los intervalos de confianza para:

- la **media de una variable**
- la **proporción** de individuos de una población que siguen determinada característica (es un caso particular de la media)
- la **diferencia de medias de dos variables**

considerando siempre que el **tamaño de la muestra** es **suficientemente grande**, de forma que estemos en las condiciones en que se puede aplicar el Teorema Central del Límite. (En general, se trabajará con muestras de tamaño mayor o igual a 100.)

NOTA: Para poblaciones cualesquiera (no normales) no podemos hallar intervalos de confianza para la varianza, ni para comparar varianzas, ya que no se tiene un pivote para estudiar la varianza con distribución conocida.

7.3.1.- Intervalos de confianza para la media de una variable.

En este apartado consideramos la v.a. X con **media** θ, y un **tamaño de la muestra, n, suficientemente grande**.

IC(θ): Intervalo de confianza para la media

Pivote: $X^* = \dfrac{\bar{X}-\theta}{S/\sqrt{n}} \approx t_{n-1}$; **IC(θ):** $\bar{X} \pm t_{\alpha/2} S/\sqrt{n}$, donde $\mathrm{P}\left(t_{n-1} \le t_{\alpha/2}\right) = 1-\dfrac{\alpha}{2}$.

Observación:

El que el **tamaño de la muestra sea suficientemente grande** es importante para:

- poder aplicar el TCL,
- poder considerar S como una buena aproximación de la desviación típica de la población; por ello, se aproxima la distribución del pivote por una t_{n-1}.
- Cuando $n>120$, la *t-Student* se aproxima por la $N(0,1)$, por lo que también se podría considerar que $X^* = \dfrac{\bar{X}-\theta}{S/\sqrt{n}} \approx N(0,1)$.

[3] Tema 5, sección 5.4: Sean $X_1, \ldots, X_n$ v.a.i.i.d. (variables aleatorias **independientes con idéntica distribución**). Si denotamos $E(X_i)=\mu$ y $V(X_i)=\sigma^2$, **cuando n → +∞**, se tiene que $\bar{X} = \dfrac{X_1 + \ldots + X_n}{n} \approx N(\mu, \dfrac{\sigma}{\sqrt{n}})$.

Ejemplo:

Queremos estudiar la media de la variable *Libros no académicos leídos en un año por los estudiantes de la EUI*. Para ello, hallamos un intervalo de confianza al 95%, a partir de los datos obtenidos en el grupo SM22 (curso 2006/07): $n = 48;\ \bar{X} = 3.58333;\ S = 4.08335$.

Con la opción *Distribution Fitting* de Statgraphics, contrastamos si es normal y obtenemos:

Approximate P-Value = 0,0622957 **<** 0.3 (**no** podemos aceptar que es **normal**).

Podemos considerar $n=48$ como suficientemente grande para utilizar el TCL y considerar la aproximación: $X^* = \dfrac{\bar{X} - \theta}{S/\sqrt{n}} \approx t_{n-1}$.

A partir de ella, obtenemos el intervalo de confianza por el método habitual:

- Hallamos $K \in \mathbb{R}$, tal que $P(-K \le t_{47} \le K) = 0.95$: $K = 2.01174$ (visto antes).
- **IC(θ):** $\bar{X} \pm t_{\alpha/2} \dfrac{S}{\sqrt{n}} \Rightarrow 3.58333 \pm 2.01174 \cdot \dfrac{4.08335}{\sqrt{48}} \Rightarrow IC = (2.39765; 4.76902)$

Con ese resultado, al 95% de confianza, podemos:

- asegurar que los estudiantes de la EUI leen, en media, más de 2 libros no académicos al año,
- asegurar que los estudiantes de la EUI leen, en media, menos de 5 libros no académicos al año,
- aceptar como posible el que los estudiantes de la EUI leen, en media, 4 libros no académicos al año.

7.3.2. IC(p): Intervalo de confianza para la proporción

Una situación muy común es querer obtener información sobre la **proporción** de una población que verifica una determinada característica. Para ello, buscamos un modelo matemático apropiado que nos permitirá hallar un intervalo de confianza utilizando los resultados vistos anteriormente. Lo explicamos con un ejemplo.

Supongamos que queremos hallar un **intervalo de confianza para la proporción** de estudiantes de la EUI a los que las matemáticas les provocan sensaciones negativas.

Consideramos datos obtenidos en diversos cursos. Con una muestra de tamaño $n=86$, tenemos que hay 51 estudiantes que tenían sensaciones que se pueden considerar negativas ante las matemáticas. Con esos datos, una **estimación puntual de la proporción** sería $\hat{P} = \dfrac{51}{86} = 0.5930$.

Para hallar un **intervalo de confianza**, planteamos el problema de la siguiente forma:

i) Definimos la v.a. *X*, que toma los valores:

- 1 si un estudiante tiene sensaciones negativas ante las matemáticas
- 0 en caso contrario.

ii) Consideramos que $X \sim B(1, \boldsymbol{p})$, siendo ***p*** la probabilidad de que un estudiante tenga sensaciones negativas ante las matemáticas, o bien, siendo ***p*** la **proporción** de estudiantes con sensaciones negativas ante las matemáticas.

Como queremos hallar un intervalo para ***p*** y *E(X)*= ***p***, es equivalente a hallar el intervalo para la **media** de la variable *X*.

iii) Por tanto, teniendo en cuenta lo visto en *7.3.1* (intervalo para la media de una población cualquiera), consideramos el pivote $X^* = \dfrac{\bar{X} - p}{S/\sqrt{n}} \approx t_{n-1}$.

En este caso, se tiene que $\bar{X} = \hat{P} = \dfrac{r}{n}$, siendo $r = X_1 + \ldots + X_n$ el número de estudiantes que tienen la característica estudiada (sensación negativa ante las matemáticas, en este caso).

Además, por ser $X \sim B(1,p)$, se tiene que $V(X) = p(1-p)$, y se puede comprobar que $S^2 = \dfrac{n}{n-1}\hat{P}(1-\hat{P})$ [7].
Por tanto, sustituyendo la expresión anterior en el pivote obtenemos[8]:

$$X^* = \frac{\hat{P} - p}{\sqrt{\dfrac{\hat{P}(1-\hat{P})}{n-1}}} \approx t_{n-1}.$$

Como conclusión, para hallar el **intervalo de confianza para la proporción** de individuos de una población que cumplen determinada característica, elegimos:

$$\textbf{Pivote: } X^* = \frac{\hat{P} - p}{\sqrt{\dfrac{\hat{P}(1-\hat{P})}{n-1}}} \approx t_{n-1} \ ;$$

$$\textbf{IC(p):}\ \hat{P} \pm t_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n-1}} \quad \text{, donde } P\left(t_{n-1} \le t_{\alpha/2}\right) = 1 - \frac{\alpha}{2}.$$

En el ejemplo, para hallar un intervalo al 95%:

- Hallamos $K \in \mathbb{R}$, tal que $P(-K \le t_{85} \le K) = 0.95 \Leftrightarrow P(t_{85} \le K) = 0.975 \Leftrightarrow K = 1.98827$ (Statgraphics).
- **IC(p):** $0.5930 \pm 1.98827\sqrt{\dfrac{0.5930(1-0.5930)}{86-1}} \Rightarrow IC(p): (0.487; 0.699)$

---

[7] Cuando $X \sim B(1,p)$, se tiene que $\sum X_i^2 = \sum X_i$, pues la variable sólo puede tomar los valores 0 y 1. Por tanto, se tiene que la varianza muestral es $V = \dfrac{\sum X_i^2}{n} - \bar{X}^2 = \dfrac{\sum X_i}{n} - \bar{X}^2 = \bar{X} - \bar{X}^2 = \bar{X}(1-\bar{X}) = \hat{P}(1-\hat{P})$. Como $S^2 = \dfrac{n}{n-1}V$, se tiene que $S^2 = \dfrac{n}{n-1}\hat{P}(1-\hat{P})$.

[8] $X^* = \dfrac{\hat{P}-p}{S/\sqrt{n}} = \dfrac{\hat{P}-p}{\sqrt{S^2/n}} = \dfrac{\hat{P} - p}{\sqrt{\dfrac{n}{n-1}\hat{P}(1-\hat{P})/n}} = \dfrac{\hat{P} - p}{\sqrt{\dfrac{\hat{P}(1-\hat{P})}{n-1}}}$

Con ese resultado, al 95% de confianza, podemos:

- asegurar que más del 48% de los estudiantes de la EUI tienen sensaciones negativas ante las matemáticas,
- aceptar como posible que 2/3 de los estudiantes de la EUI tienen sensaciones negativas ante las matemáticas.

7.3.3.- Intervalos de confianza para la diferencia de medias de dos variables

En este apartado utilizaremos la misma **notación** que en 7.2.2:

- X y Y son v.a. independientes, con $E(X)=\mu_X$, $E(Y)=\mu_Y$, $V(X)=\sigma_X^2$, $V(Y)=\sigma_Y^2$.
- $(X_1, \dots, X_{nx})$, $(Y_1, \dots, Y_{ny})$: m.a.s. de X e Y
- n_x : tamaño de la muestra de X
- n_y : tamaño de la muestra de Y
- $\overline{X}, \overline{Y}$: medias muestrales de X e Y
- S_X^2, S_Y^2 : cuasivarianzas muestrales de X e Y

Observación:

Para **tamaños muestrales suficientemente grandes**:

- Se puede considerar que $\overline{X}_1, \overline{X}_2$ siguen distribuciones normales (TCL).
- Sin embargo, como las poblaciones **no** son normales, **los resultados para comparar las varianzas no son válidos**.
- Como no se puede asegurar nada sobre la igualdad de varianzas, se utiliza como pivote el caso más general visto en 7.2.2 (intervalo para la diferencia de medias de poblaciones normales con varianzas distintas), aproximando la distribución t_f por la N(0,1).

Por tanto:

IC($\mu_1 - \mu_2$): Intervalo de confianza para la diferencia de medias de v.a. independientes

Pivote: $$X^* = \frac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{S_X^2}{n_X}+\frac{S_Y^2}{n_Y}}} \approx N(0,1)\ ;$$

IC($\mu_X - \mu_Y$): $(\overline{X}-\overline{Y}) \pm z_{\alpha/2}\sqrt{\frac{S_X^2}{n_X}+\frac{S_Y^2}{n_Y}}$, donde $\mathrm{P}\left(Z \le z_{\alpha/2}\right)=1-\frac{\alpha}{2}$.

Ejemplo:

Podemos estudiar si las "nuevas generaciones" hablan más tiempo por teléfono. Una aproximación, podría ser comparar el *tiempo máximo en una llamada de teléfono* en dos poblaciones distintas. Consideramos:

X: Tiempo máximo en una llamada de teléfono en alumnos de Estadística EUI ***(curso 2004/05)***
Y: Tiempo máximo en una llamada de teléfono en alumnos de Estadística EUI ***(curso 2006/07)***

A partir de los datos recogidos en los respectivos años, se tiene:

X	$n_X = 39$	$\bar{X} = 64.1$ *minutos*	$S_X = 42.31$ *minutos*
Y	$n_Y = 49$	$\bar{Y} = 95.39$ *minutos*	$S_Y = 83.63$ *minutos*

Con la opción *Distribution Fitting* de Statgraphics obtenemos:

X : Approximate P-Value = 0,00401001 **<** 0.3 (**no** podemos suponer que es normal)
Y : Approximate P-Value = 0,0216398 **<** 0.3 (**no** podemos suponer que es normal)

Considerando que los tamaños muestrales son suficientemente grandes[9] podemos considerar como pivote: $X^* = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1^2}{n} + \dfrac{S_2^2}{m}}} \approx N(0,1)$, y a partir de él, obtener el intervalo de confianza al 95% por el método habitual:

- Hallamos $K \in \mathbb{R}$, tal que
  $P(-K \le N(0,1) \le K) = 0.95 \Leftrightarrow P(N(0,1) \le K) = 0.975 \Leftrightarrow K = 1.96$.
- ***IC*($\mu_X - \mu_Y$):** $(\bar{X} - \bar{Y}) \pm z_{\alpha/2}\sqrt{\dfrac{S_X^2}{n_X} + \dfrac{S_Y^2}{n_Y}} \Rightarrow (64.1 - 95.39) \pm 1.96\sqrt{\dfrac{42.31^2}{39} + \dfrac{83.63^2}{49}}$

Y se obtiene el intervalo: **(−58.21, −4.37)**.

Con ese resultado, al 95% de confianza, podemos:

- asegurar que **hay diferencias significativas** entre los tiempos medio de los estudiantes de el curso actual y los de dos cursos anteriores: como $0 \notin (-58.21, -4.37)$, no puede considerarse que las medias sean iguales.
- asegurar que la diferencia de medias es **negativa**, pues todos los valores del intervalo de confianza lo son. Por tanto, si $\mu_X - \mu_Y < 0$, se tiene que $\mu_X < \mu_Y$; es decir, la media de los estudiantes del curso 2006/07 se mayor que la de los estudiantes del curso 2004/05… lo que es indicativo de que las "nuevas generaciones" están más tiempo "colgadas al teléfono".

9.4.- Error absoluto. Tamaño de la muestra.

Si se quiere dar una estimación de un parámetro θ y se utiliza un estimador $\hat{\theta}$, se denomina **error absoluto de la estimación** a $\left|\hat{\theta} - \theta\right|$.

Como se desconoce el valor exacto de θ, no se puede conocer el valor exacto del error, pero sí **podemos obtener cotas superiores** del mismo con un determinado nivel de confianza.

Por ejemplo, en la deducción del intervalo de confianza para la media, se tiene la expresión

$$t_{\alpha/2}\, S/\sqrt{n} \le \bar{X} - \mu \le t_{\alpha/2}\, S/\sqrt{n},$$

[9] En general, para garantizar una buena aproximación se necesita un tamaño de la muestra mayor, preferiblemente mayor que 100. En este caso, hacemos la suposición de que son "suficientemente grandes" para utilizar datos de estudiantes de la EUI.

de donde se deduce que

$$\left|\bar{X}-\mu\right| \le t_{\alpha/2}\, S/\sqrt{n},$$

que nos da una **cota del error absoluto** de la estimación de la media.

Ejemplos:

1) En el ejemplo del estudio de la proporción de *estudiantes de la EUI a los que las matemáticas les provocan sensaciones negativas*, tenemos que:

- la estimación puntual de la proporción es $\hat{\boldsymbol{P}} = \frac{51}{86} = 0.5930$.
- ***IC(p):*** $\hat{P} \pm t_{\alpha/2}\sqrt{\frac{\hat{P}\ (1-\hat{P}\ )}{n-1}} \Rightarrow 0.5930 \pm 1.98827\sqrt{\frac{0.5930\ (1-0.5930\ )}{86-1}} \Rightarrow IC(p):(0.487;0.699)$

En este caso, el **error absoluto de la estimación** sería $\left|\hat{\boldsymbol{P}} - \boldsymbol{p}\right|$ y la **cota superior** del error vendría dada por:

$$\left|\hat{P}-p\right| \le t_{\alpha/2}\sqrt{\frac{\hat{P}\ (1-\hat{P}\ )}{n-1}} \Rightarrow \left|\hat{P}-p\right| \le 1.98827\sqrt{\frac{0.5930\ (1-0.5930\ )}{86-1}} \Rightarrow \left|\hat{P}-p\right| \le 0.106$$

Se puede decir que el porcentaje de estudiantes de la EUI a los que las matemáticas les provocan sensaciones negativas es del 59.3% con un error máximo del ±10.6% (95% de confianza).

2) En el ejemplo del estudio del *peso de los estudiantes de la EUI*, si nos centramos en la **desviación típica**, teníamos que:

- la estimación puntual de la desviación típica era $S = 13.4485\ kg$.
- $IC(\sigma): \sqrt{\frac{(n-1)S^2}{K_2}} \le \sigma \le \sqrt{\frac{(n-1)S^2}{K_1}}$ = (11.215; 16.801)

En este caso, el **error absoluto de la estimación** sería $\left|\boldsymbol{S}-\boldsymbol{\sigma}\right|$ y la **cota superior** del error vendría dada por la máxima diferencia entre ***S*** y los extremos del intervalo de confianza:

$$\left|S-\sigma\right| \le máx\left\{\left|S-\sqrt{\frac{(n-1)S^2}{K_2}}\right|, \left|S-\sqrt{\frac{(n-1)S^2}{K_1}}\right|\right\} = máx\left\{\left|13.4485-11.215\right|, \left|13.4485-16.801\right|\right\} \Rightarrow$$

$$\left|S-\sigma\right| \le 3.3525$$

Se puede decir que la desviación típica del peso de los estudiantes de la EUI es de 13.4485 *kg* con un error máximo de ±3.4 kg (95% de confianza).

En general, la cota del error dependerá del tamaño de la muestra y por ello tiene interés **estimar cuál es un tamaño de muestra adecuado para garantizar un error menor a un cierto valor**.

En el siguiente ejemplo, se muestra cómo hacer esta estimación para las estimaciones de la **media.**

Ejemplo:

Queremos obtener una estimación del peso medio de los estudiantes de la EUI con un error menor de 1kg (95%). Hallar el tamaño de la muestra necesario para obtenerla.

En este caso, el **error absoluto de la estimación** sería $\left|\bar{X}-\mu\right|$ y la **cota superior** del error vendría dada por $\left|\bar{X}-\mu\right| \le K \cdot S/\sqrt{n}$, siendo $\boldsymbol{P}\left(\boldsymbol{t}_{n-1} \le \boldsymbol{K}\right) = 0.975$.

Con los datos de la muestra obtenida ($n=49, K=2.01064, S=13.4485$), tenemos que $\left|\bar{X}-\mu\right| \le 3.8629$, es decir hay un error máximo de ± 3.9 kg (95%).

Para reducir el error, buscamos *n* tal que $\left|\bar{X}-\mu\right| \le K \cdot S/\sqrt{n} \le 1 \ kg$.

Pero hay que tener en cuenta que tanto *K* como *S* también dependen de *n* ($\boldsymbol{P}\left(\boldsymbol{t}_{n-1} \le \boldsymbol{K}\right) = 0.975$ y $S^2 = \dfrac{\sum\left(X_i-\mu\right)^2}{n-1}$).

Por ello, para estimar *n* seguiremos los siguientes pasos:

i) Suponemos que el **valor de *S* es fijo**, tomando como tal el obtenido en alguna muestra. En este caso $S=13.4485$ (obtenido con la muestra de n=49).

ii) Hallamos una **estimación de *K*** suponiendo que $\dfrac{\bar{X}-\mu}{S/\sqrt{n}} \approx N(0,1)$:

$\boldsymbol{P}\left(\boldsymbol{N}(0,1) \le \boldsymbol{K}\right) = 0.975 \Rightarrow \boldsymbol{K} = 1.96$.

iii) Hallamos una **estimación de *n*** imponiendo $\left|\bar{X}-\mu\right| \le K \cdot S/\sqrt{n} \le 1 \ kg$. En este caso, resolvemos $1.96 \cdot 13.4485/\sqrt{n} \le 1 \Rightarrow \sqrt{n} \ge 1.96 \cdot 13.4485 \Rightarrow n \ge \left(1.96 \cdot 13.4485\right)^2 \Rightarrow n \ge 694.80$. Por ejemplo *n=695.*

Si hallamos una estimación del peso con una muestra de tamaño mayor o igual a 695 (700 por ejemplo) se garantizaría un error menor de 1 kg (95%).

Observación: Si el valor obtenido para *n* es menor que 120 (valor a partir del cual $t_n \approx N(0,1)$), se puede obtener una **segunda estimación de *K* (*K´*)** usando que $\dfrac{\bar{X}-\mu}{S/\sqrt{n}} \approx t_{n-1}$: $\boldsymbol{P}\left(\boldsymbol{t}_{n-1} \le \boldsymbol{K}'\right) = 1-\dfrac{\alpha}{2}$, y luego una **segunda estimación de *n*** imponiendo $K' \cdot S/\sqrt{n} \le error$.