Máster en Matemática Computacional
Departamento de Matemáticas

Introducción al cálculo de tamaños muestrales, orientado a estudios bioestadísticos

Tesis de máster de: Carla Garí Peris
Supervisada por: María Victoria Ibáñez Gual

Este trabajo de investigación se presenta como **Tesis de Máster** dentro del programa de **Máster Universitario en Matemática Computacional** para optar al título de Máster.

Castellón, 11 de Noviembre de 2016

M. Victoria Ibáñez Gual

Índice general

Capítulo 1

Introducción

El tema del presente trabajo, el cálculo del tamaño muestral, se debe a la tarea realizada durante la estancia en prácticas del máster de matemática computacional. Dicha estancia fue realizada en la empresa Outcomes'10, situada en el Espaitec de la UJI, consultora especializada en Farmacoeconomía y en Investigación en Resultados en Salud. Más concretamente, sus servicios se basan en la concepción y desarrollo de proyectos de investigación destinados a obtener evidencia útil en las etapas de acceso al mercado y de empleo en la práctica clínica habitual de los productos farmacéuticos y de tecnologías sanitarias.

La tarea allí realizada consistió en la programación de una calculadora del tamaño muestral en Excel que realizase este cálculo para los casos más utilizados durante la realización de sus informes, así como un manual del usuario de la calculadora. (Disponible en el anexo I)

La primera fase de esta tarea consistió en la obtención de la información necesaria para la implementación de la calculadora. Es durante la realización de esta parte del trabajo cuando motivada por conocer todos los razonamientos matemáticos que hay detrás de cada una de las fórmulas que constituirían el programa final, decido llevar más allá el tema utilizándolo como base para el trabajo final del máster. Durante la realización de mis estudios de licenciatura así como de los estudios de este máster siempre hubo una asignatura personal pendiente, que es la aplicación de los conocimientos adquiridos a un problema real, a una situación cotidiana es por ello que este tema une tres partes fundamentales, la parte matemática y la parte computacional con su utilidad práctica, una utilidad, en este caso para mi, no sólo teórica, si no, una utilidad que veo reflejada a mi alrededor, en los empleados que en un futuro utilizarán la herramienta que he programado para facilitar, un poco, su trabajo diario.

En investigación, la finalidad de la estadística es utilizar datos obtenidos en una muestra de sujetos para realizar inferencias válidas para una población más amplia de individuos de características similares. La validez y utilidad de estas inferencias dependen de cómo el estudio ha sido diseñado y ejecutado, por lo que la estadística debe considerarse como una parte integrante del método científico.

Un aspecto fundamental en el diseño de estudios clínicos es la determinación del tamaño de muestra apropiado. Si el tamaño de muestra es muy pequeño, el estudio tendrá baja potencia estadística y en consecuencia, las estimaciones serán menos precisas y la probabilidad de encontrar diferencias significativas entre tratamientos o grupos será menor. Por otra parte, si el tamaño de muestra es muy grande, se estará haciendo un mal uso de recursos de investigación y sometiendo a pruebas a más pacientes de los estrictamente necesarios.

El presente trabajo consta de siete capítulos y un anexo. Los tres primero capítulos servirán para introducir los conceptos estadísticos y resultados necesarios a lo largo del desarrollo del texto. En el primer capítulo se describen conceptos como, población y muestra, variables aleatorias y función generatriz de momentos .En el capítulo segundo se describen algunos de los tipos de estimadores que se utilizarán posteriormente en la parte principal del trabajo.

En el capítulo tres estudiaremos distintas distribuciones y algunas de las propiedades de las mismas, este capítulo incluye: distribución Normal, Gamma, χ^2, F de Snedecor-Fisher, t-Student, Bernoulli y Binomial.

En el capítulo cuatro se inicia el cálculo del tamaño muestral, en este caso para la estimación de una media y de una proporción, también se mencionan las definiciones de intervalo de confianza y factor de corrección, para pasar en el capítulo cinco a desarrollar el cálculo del tamaño muestral mediante contrastes de hipótesis, tras una introducción este concepto desarrollaremos los siguientes casos: comparación de medias independientes, comparación de dos medias apareadas en una sola muestra, comparación de k medias, comparación de dos porporciones independientes, comparación de dos proporciones con población de referencia y comparación de k proporciones. En cada uno de los casos desarrollaremos una estructura similar que incluye: deducción del contraste, prueba de igualdad y prueba de no inferioridad/superioridad. Cuando se trata de las medias tendremos en cuenta la igualdad o no de la varianza así como si esta, es conocida o desconocida.

En el apartado seis desarrollaremos el cálculo del tamaño muestral para dos de los estudios epidemiológicos más utilizados : cohortes y casos y controles.

A lo largo de los tres apartados previos hemos desarrollado el cálculo del tamaño muestral suponiendo normalidad, es por ello que en el apartado siete introduciremos los estudios no paramétricos, para poder mostrar algunas de las alternativas existentes cuando no podemos suponer normalidad.

Finalmente, en el anexo, encontraremos un manual del usuario de la cálculadora del tamaño muestral realizada durante la estacia en prácticas, cuya realización sirvió para elegir el tema de este trabajo.

1.1. Población y muestra

Llamamos **población estadística** al conjunto de referencia del que extraemos las observaciones, es decir, el conjunto de todas las posibles unidades experimentales. En los estudios clínicos raramente es posible poder extraer los datos necesarios de todos los elementos de la población por ello es necesario introducir

el concepto de muestra.

Llamamos **muestra** a un suconjunto de elementos de la población que cumple los requisitos necesarios para la realización del estudio. El número de elementos que componen la muestra es lo que llamamos **tamaño muestral** y se suele representar por la letra minúscula n.

Por útimo definimos el **espacio muestral** como el conjunto de todos los resultados posibles de un experimento aleatorio. El espacio muestral se denota como Ω.

El objetivo final es llegar a conocer ciertas características de la población a partir de la muestra.

1.2. Variables aleatorias

Una **variable aleatoria** es una función con valores reales definida sobre el espacio muestral. Decimos que la variable tiene

- Una distribución discreta: si sólo puede tomar un número finito k de valores distintos, o a lo sumo, una sucesión infinita de valores distintos. En este caso, se define su **función de probabilidad**

  $$f(x) = P(X = x), \forall x \in \mathbb{R},$$

  y dado cualquier subconjunto A de la recta real,

  $$P(X \in A) = \sum_{x \in A} f(x).$$

- Una distribución continua: si existe una función no negativa f, definida sobre la recta real tal que, para cualquier intervalo A,

  $$P(X \in A) = \int_A f(x)dx.$$

  La función f se llama función de densidad de probabilidad (f.d.p.), y toda f.d.p debe satisfacer dos requisitos:

  - $f(x) \geq 0$ y
  - $\int_{-\infty}^{\infty} f(x)dx = 1$

Dadas dos variables aleatorias, X, Y definimos su **función de distribución conjunta** como:

- Sean X e Y variables aleatorias con distribución discreta, definimos su función de probabilidad conjunta f como:

  $$f(x, y) = P(X = x, Y = y)$$

  para cualquier punto $(x, y) \in \mathbb{R} \times \mathbb{R}$. Si (x, y) no es uno de los valores posibles del par de variables aleatorias (X, Y), entonces $f(x, y) = 0$. Además,

si la sucesión $(x_1, y_1), (x_2, y_2), ...$ incluye todos los posibles valores del par (X, Y), entonces

$$\sum_{i=1}^{\infty} f(x_i, y_i) = 1$$

Para cualquier subconjunto A del plano xy,

$$P[(X, Y) \in A] = \sum_{(x_i, y_i) \in A} f(x_i, y_i)$$

Si X e Y son variables aleatorias discretas independientes, la función de probabilidad conjunta viene dada por:

$$f(x, y) = P(X = x, Y = y) = P(X = x)P(Y = y)$$

- Sean X e Y variables aleatorias continuas, diremos que tienen función de distribución conjunta si existe una función f no negativa definida sobre todo $\mathbb{R}^2$ tal que para cualquier subconjunto A del plano,

$$P[(X, Y) \in A] = \int \int_A f(x, y)dxdy$$

La función f es la función de densidad de probabilidad conjunta de X e Y y debe satisfacer las condiciones siguientes:

$$f(x, y) \geq 0 \quad \text{para} \quad -\infty < x < \infty \quad -\infty < y < \infty$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)dxdy = 1$$

Si X e Y son variables aleatorias independientes continuas la función de densidad de probabilidad conjunta viene dada por:

$$P[(X, Y) \in A] = \int \int_A f(x, y)dxdy = \int_A f_X(x)dx \int_A f_Y(y)dy$$

Es decir, cuando tenemos variables aleatorias independientes, la f.d conjunta es el producto de las funciones de densidad de cada una de ellas.

Teorema 1.2.1 *Sea X una variable aleatoria continua con distribución de probabilidad $f_X(x)$. Si dada una función g, $Y = g(x)$ define una correspondencia uno a uno entre los valores de X y Y de tal forma que la ecuación $y = g(x)$ tenga su inversa $x = g^{-1}(y)$, entonces la función de densidad de probabilidad de Y, f_Y, es:*

$$f_Y(y) = f_X(g^{-1}(y))J$$

donde $J = \mid \frac{d}{dy} g^{-1}(y) \mid$ y recibe el nombre de jacobiano de la transformación.

Demostración
Estudiaremos para esta demostración dos casos:

1. $y = g(x)$ creciente
2. $y = g(x)$ decreciente

CASO 1. **Si** $y = g(x)$ **creciente**, escogemos dos puntos arbitrarios de y, por ejemplo a y b entonces:

$$\begin{aligned} P(a \leq Y \leq b) \quad &= P(Y \leq b) - P(Y \leq a) = \\ &= P(g(X) \leq b) - P(g(X) \leq a) = \\ &= P(X \leq g^{-1}(b)) - P(X \leq g^{-1}(a)) = \\ &= P[g^{-1}(a) \leq X \leq g^{-1}(b)] = \int_{g^{-1}(a)}^{g^{-1}(b)} f_X(x)dx \end{aligned}$$

Cambiamos ahora la variable de integración de x a y y utilizando la relación $x = g^{-1}(y)$ tenemos que:

$$dx = [g^{-1}(y)]dy$$

por tanto

$$P(a \leq Y \leq b) = \int_a^b f_X(g^{-1}(y))[g^{-1}(y)]dy$$

como a y b recorren todos los valores permisibles de y siempre que $a < b$ se tiene que

$$f_Y(y) = f_X(g^{-1}(y))[g^{-1}(y)] = f_X(g^{-1}(y))J$$

Se conoce a $J = [g^{-1}(y)]$ como el reciproco de la pendiente de la línea tangente a la curva de la función creciente $y = g(x)$ por la elección de a y b es evidente que $J = |J|$.

CASO 2. **Si** $y = g(x)$ **decreciente**, escogemos dos puntos arbitrarios de y, por ejemplo a y b entonces:

$$\begin{aligned} P(a \leq Y \leq b) \quad &= P(Y \leq b) - P(Y \leq a) = \\ &= P(g(X) \leq b) - P(g(X) \leq a) = \\ &= P(X \leq g^{-1}(b)) - P(X \leq g^{-1}(a)) = \\ &= P[g^{-1}(a) \leq X \leq g^{-1}(b)] = \int_{g^{-1}(a)}^{g^{-1}(b)} f_X(x)dx \end{aligned}$$

Cambiamos ahora la variable de integración de x a y y utilizando la relación $x = g^{-1}(y)$ tenemos que:

$$dx = [g^{-1}(y)]dy$$

por tanto

$$P(a \leq Y \leq b) = \int_a^b f_X(g^{-1}(y))[g^{-1}(y)]dy = -\int_a^b f_X(g^{-1}(y))[g^{-1}(y)]dy$$

como a y b recorren todos los valores permisibles de y siempre que $a < b$ se tiene que

$$f_Y(y) = f_X(g^{-1}(y))[g^{-1}(y)] = -f_X(g^{-1}(y))J$$

en este caso la pendiente de la curva es negativa, por tanto $J = -J$.

$$f_Y(y) = f_X(g^{-1}(y))J$$

Veremos ahora un teorema que nos permitirá calcular la función de densidad conjunta a partir de transformaciones.

Teorema 1.2.2 *Sean $X_1, \cdots, X_n$ una sucesión de v.a. continuas definidas sobre un espacio muestral S, con función de densidad de probabilidad conjunta $f(X_1, \cdots, X_n)$. Sean $Y_1, \cdots, Y_n$ otra sucesión de v.a. que se han obtenido a partir de transformaciones biyectivas $r_1, \cdots, r_n$ de las anteriores de la forma:*

$$\begin{aligned} Y_1 &= r_1(X_1, \cdots, X_n) \\ Y_2 &= r_2(X_1, \cdots, X_n) \\ &\vdots \\ Y_n &= r_n(X_1, \cdots, X_n) \end{aligned}$$

que quedarán definidas sobre un espacio muestral T. Al tratarse de transformaciones biyectivas, podemos encontrar las transformaciones inversas $s_1, \cdots, s_n$ tales que:

$$\begin{aligned} X_1 &= s_1(Y_1, \cdots, Y_n) \\ X_2 &= s_2(Y_1, \cdots, Y_n) \\ &\vdots \\ X_n &= s_n(Y_1, \cdots, Y_n) \end{aligned}$$

Suponiendo que $\partial s_i/\partial y_j$ existe $\forall i, j = 1, \cdots, n$, definimos J, el jacobiano de la transformación:

$$J = \begin{vmatrix} \partial s_1/\partial y_1 & \cdots & \partial s_1/\partial y_n \\ \partial s_2/\partial y_1 & \cdots & \partial s_2/\partial y_n \\ \vdots & & \vdots \\ \partial s_n/\partial y_1 & \cdots & \partial s_n/\partial y_n \end{vmatrix}$$

Entonces, a partir de los métodos de cálculo para cambio de variables en una integral múltiple se puede demostrar que la f.d.p. conjunta de las variables $Y_1, \cdots, Y_n$*, que denotaremos por* $g(Y_1, \cdots, Y_n)$ *es:*

$$g(Y_1, \cdots, Y_n) = \begin{cases} f(s_1(Y_1, \cdots, Y_n), \cdots, s_n(Y_1, \cdots, Y_n))J & \text{si } Y_i \in T,\ i \in \{1, \cdots, n\} \\ 0 & \text{en otro caso} \end{cases}$$

Dada una variable aleatoria X, se define su **función de distribución**, como $F(x) = P(X \leq x)$, y se define su **esperanza** como:

- $E(X) = \sum_x xf(x)$ si X es una variable con distribución discreta.
- $E(X) = \int_{-\infty}^{\infty} xf(x)dx$ si X es una variable con distribución continua.

Se dice que la esperanza existe

- Para una variable X con distribución discreta, si la suma es absolutamente convergente,i.e. síi
$$\sum_x \mid x \mid f(x) < \infty$$
.
- Para una variable X con distribución continua, si la integral es absolutamente convergente,i.e. síi
$$\int_{-\infty}^{\infty} \mid x \mid f(x)dx < \infty$$
.

Análogamente se define la **esperanza** de cualquier función $r(X)$ de la variable aleatoria como:

- $E(r(X)) = \sum_x r(x)f(x)$ si X es una variable con distribución discreta.
- $E(r(X)) = \int_{-\infty}^{\infty} r(x)f(x)dx$ si X es una variable con distribución continua.

Dada una variables aleatoria X definimos su varianza como $Var(X) = E(X^2) - E(X)^2$. En particular,

- $Var(X) = \sum_{x_i}(x_i - E(X))^2 f(x_i)$ si X es una variable con distribución discreta.
- $Var(X) = \int_{-\infty}^{\infty}(x_i - E(X))^2 f(x)dx$ si X es una variable con distribución continua.

1.3. Función generatriz de momentos(f.g.m)

Si X es una variable aleatoria, se define su momento de orden k como $E(X^k)$, siempre que la esperanza exista, y se define su función generatriz de momentos $\psi_X(t)$ como

$$\psi_X(t) = E(e^{tX}), \forall t \in R.$$

Si la variable aleatoria está acotada, $\psi_X(t)$ existirá para cualquier valor de t, pero si no lo está, puede existir para algunos valores de t y no existir para otros. De todas formas, para cualquier variable aleatoria X, $\psi_X(t)$ debe existir en $t = 0$ y $\psi_X(0) = E(1) = 1$.

Supongamos que existe la f.g.m de una variable aleatoria X para todos los valores de t en un intervalo alrededor del punto $t = 0$. Entonces se puede demostrar que existe la derivada $\psi'(t)$ en el punto $t = 0$ y que en ese punto la derivada de la esperanza de la ecuación $\psi(t) = E(e^{tX})$ debe ser igual a la esperanza de la derivada. Entonces,

$$\psi'(0) = \left[\frac{d}{dt}E(e^{tX})\right]_{t=0} = E\left[\left(\frac{d}{dt}e^{tX}\right)_{t=0}\right]$$

Pero, puesto que

$$\left(\frac{d}{dt}e^{tX}\right)_{t=0} = (Xe^{tX})_{t=0} = X$$

por tanto

$$\psi'(0) = E(X)$$

es decir, la derivada de la f.g.m $\psi(t)$ en el punto $t = 0$ es la media de X.

En general, si la f.g.m $\psi(t)$ de X existe para todos los valores de t en un intervalo alrededor del punto $t = 0$, entonces deben existir todos los momentos $E(X^k)$ de X $(k = 1, 2, ...)$. Para $n = 1, 2, ...$, la n-ésima derivada $\psi^{(n)}(0)$ en el punto $t = 0$ satisfará la relación siguiente:

$$\psi^{(n)}(0) = \left[\frac{d^n}{dt^n}E(e^{tX})\right]_{t=0} = E\left[\left(\frac{d^n}{dt^n}e^{tX}\right)_{t=0}\right] = E[(X^n e^{tX})_{t=0}] = E(X^n)$$

Entonces, $\psi'(0) = E(X), \psi''(0) = E(X^2), \psi'''(0) = E(X^3)$ y así sucesivamente.

Teorema 1.3.1 *Sea X una variable aleatoria cuya f.g.m es ψ_1; sea $Y = aX+b$, donde a y b son constantes cualesquiera; y sea ψ_2 la f.g.m de Y. Entonces, para cualquier valor de t tal que existe $\psi_1(at)$,*

$$\psi_2(t) = e^{bt}\psi_1(at)$$

Demostración

$$\psi_2(t) = E(e^{tY}) = E[e^{t(aX+b)}]e^{bt}E(e^{atX}) = e^{bt}\psi_1(at)$$

Teorema 1.3.2 *Si $X_1, ..., X_n$ son variables aleatorias independientes y $Y = X_1 + ... + X_n$, entonces*

$$\psi_Y(t) = \prod_{i=1}^{n} \psi_{X_i}(t)$$

donde $\psi_{X_i}(t)$ es el valor de la función generatriz de momentos de X_i en t.

Demostración [**?**]

Hacemos uso del hecho que las variables aleatorias son independientes y por tanto

$$f(x_1, ..., x_n) = f_1(x_1)...f_n(x_n)$$

Por tanto

$$\begin{aligned}
\psi_Y(t) \quad &= E(e^{YT}) = E\left[e^{(x_1+x_2+...x_n)t}\right] \\
&= \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} e^{(x_1+x_2+...x_n)t} f(x_1, ..., x_n)dx_1dx_2...dx_n \\
&= \int_{-\infty}^{\infty} e^{x_1t} f_1(x_1)dx_1 \int_{-\infty}^{\infty} e^{x_2t} f_2(x_2)dx_2... \int_{-\infty}^{\infty} e^{x_nt} f_n(x_n)dx_n \\
&= \prod_{i=1}^{n} \psi_{x_i}(t)
\end{aligned}$$

lo cual demuestra el teorema para el caso continuo, para demostrarlo para el caso discreto, sólo tenemos que reemplazar todas las integrales por sumas.

Teorema 1.3.3 *Teorema de Unicidad: Si las f.g.m de dos variables aleatorias X_1 y X_2 son iguales para todos los valores de t en un intervalo alrededor del punto $t = 0$, entonces las distribuciones de probabilidad de X_1 y X_2 son iguales.*

Sea $X = (X_1, ..., X_n)$ un vector aleatorio, definimos la **función generatriz de momentos conjunta** ψ_X como:

$$\psi_X(t_1, ..., t_n) = E\left[e^{\sum_{i=1}^{n} t_i X_i}\right]$$

si la esperanza existe para todo $t_1, ..., t_n$ tal que $t_i \in (-h, h)$para algún $h > 0, i = 1, ..., n$

Capítulo 2

Estimación

Como comentábamos en la introducción, un problema de inferencia estadística, es un problema en el cual se han de analizar datos que han sido generados de acuerdo con una distribución de probabilidad desconocida y en el que se debe realizar alguna inferencia acerca de dicha distribución.

A menudo, la distribución de probabilidad que generó los datos experimentales, se supone completamente conocida excepto por el valor de uno o más parámetros. En un problema de inferencia estadística, cualquier característica de la distribución que genera los datos experimentales que tenga un valor desconocido, como la media (μ) o la varianza (σ^2), se llama **parámetro** de la distribución. El conjunto Ω de todos los valores posibles del parámetro θ o de un vector de parámetros $(\theta_1, \cdots, \theta_k)$, se llama **espacio paramétrico**.

2.1. Estadísticos y estimadores

Supóngase que las variables aleatorias $X_1, ..., X_n$ constituyen una muestra aleatoria de una distribución con parámetro θ de valor desconocido. Un estadístico es cualquier función real $T = r(X_1, ...X_n)$ de las variables $X_1, ..., X_n$. Puesto que un estadístico T es una función de variables aleatorias, resulta que T es una variable aleatoria y su distribución puede, en principio, ser deducida de la distribución conjunta de $X_1, ..., X_n$. Esta distribución se denomina usualmente **distribución muestral** del estadístico T, porque se obtiene de la distribución conjunta de las observaciones de una muetra aleatoria.

Un **estimador** de θ es un estadístico, $\delta(X_1, ...X_n)$, que especifica el valor estimado de θ para cada conjunto posible de valores de $X_1, ..., X_n$. El requisito principal para tener un buen estimador, es que proporcione una estimación de θ que se aproxime lo máximo posible a su verdadero valor. Por ello, para comprobar la bondad de un estimador $\delta(X)$, se pueden definir distintas funciones de pérdida $L(\theta, \delta(X))$, cuyo valor aumenta a medida que aumenta la distancia entre el verdadero valor del parámetro θ y su estimación $\delta(X)$. Al estimador de un parámetro θ, habitualmente lo denotaremos por $\hat{\theta}$.

Podemos hablar de varios tipos de estimadores. Entre ellos:

- Estimador Bayes: Dada una muestra aleatoria $X = (X_1, ..., X_n)$ generada a partir de una distribución que involucra un parámetro θ que tiene un valor desconocido en un intervalo específico sobre la recta real Ω. Para cualquier función de pérdida $L(\theta, a)$, y cualquier f.d.p inicial $\chi(\theta)$, el estimador Bayes de θ, es el estimador $\delta(X)$ que satisface

$$E[L(\theta, \delta(X))|X] = \min_{a \in \Omega} E[L(\theta, a)|X]$$

para todo valor posible x de X. La forma del estimador Bayes dependerá tanto de la función de pérdida que se utilizó en el problema como de la distribución inicial que se asigna a θ.

- Estimador máximo verosímil: Supongamos que las variables aleatorias $X_1, ..., X_n$ constituyen una muestra aleatoria de una distribución discreta o una distribución continua cuya f.p o f.d.p es $f(X \mid \theta)$, donde el parámetro θ pertenece a un espacio paramétrico Ω. Aquí, θ puede ser una parámetro real o un vector de parámetros. Para cualquier vector observado $x = (x_1, ..., x_n)$ de la muestra, el valor de la f.p conjunta o f.d.p. conjunta, se denotará como $f_n(x \mid \theta)$. Cuando $f_n(x \mid \theta)$ se considera una función de θ para un vector concreto x, se denomina la **función de verosimilitud**.

  Para cada posible vector observado x, sea $\delta(x) \in \Omega$ un valor de $\theta \in \Omega$ cuya función de verosimilitud $f_n(x \mid \delta(x))$ es un máximo. Es decir, sea

$$\delta(x) = argmax_{\theta \in \Omega} f_n(x \mid \theta)$$

  Al estimador de θ definido de esta forma ($\hat{\theta} = \delta(X)$) se le denomina estimador máximo verosímil de θ.

Propiedad 2.1.1 *Los estimadores máximos verosímiles (EMV) de la media y la varianza de una distribución normal son la media muestral y la varianza muestral.*

Demostración

Supongamos que $X_1, ...X_n$ constituyen una muestra aleatoria de una distribución normal con media μ y varianza σ^2 desconocidas.Para cualesquiera valores observados $x_1, ...x_n$, la función de verosimilitud

$$f_n(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

Esta función se debe maximizar sobre todos los valores posibles de μ y de σ^2, donde $-\infty < \mu < \infty, \sigma^2 > 0$. En lugar de maximizar la función de verosimilitud, resulta más sencillo maximizar $log f_n(x|\mu, \sigma^2)$ y obtenemos:

$$\begin{aligned} L(\mu,\sigma^2) \quad &= \log f_n(x \mid \mu,\sigma^2) = \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \end{aligned} \tag{2.1}$$

Se deben obtener los valores de μ y σ^2 para los cuales $L(\mu,\sigma^2)$ sea máxima determinando los valores de μ y σ^2 que satisfacen las dos ecuaciones siguientes:

$$\frac{\partial L(\mu,\sigma^2)}{\partial \mu} = 0 \tag{2.2}$$

$$\frac{\partial L(\mu,\sigma^2)}{\partial \sigma^2} = 0 \tag{2.3}$$

De la ecuación (2.1) se obtiene la relación

$$\frac{\partial L(\mu,\sigma^2)}{\partial \sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-n\mu)^2$$

Por tanto de la ecuación(2.2) se obtiene que $\mu = \bar{x}_n$.

Además de la ecuación (2.1),

$$\frac{\partial L(\mu,\sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Cuando μ se reemplaza por el valor $\bar{x}_n$ que acabamos de obtener, de la ecuación (2.3) se obtiene que

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2$$

Así como $\bar{x}_n$ se denomina varianza muestral, el estadístico $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$ se denomina varianza muestral. Es la varianza de una distribución que asigna probabilidad $1/n$ a cada uno de los n valores observados $x_1, ...x_n$ de la muestra.

Por tanto, los EMV de μ y σ^2 son

$$\hat{\mu} = \bar{X}_n \qquad\qquad \hat{\sigma_2} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2$$

En muchos problemas en los que se debe estimar un parámetro θ, es posible determinar un estimador máximo verosímil o un estimador Bayes que sea apropiado. En algunos problemas, sin embargo, es posible que ninguno de estos estimadores sea apropiado. En estos casos utilizaremos un estadístico suficiente, cuya definición formal es:

Estadístico suficiente: Si T es un estadístico y t es un valor concreto de T, entonces la distribución conjunta condicional de $X_1, ..., X_n$, dado que $T = t$, se puede calcular a partir de la ecuación

$$f_n(x|\theta) = f(x_1|\theta)...f(x_n|\theta)$$

En general, esta distribución conjunta condicional dependerá del valor de θ. Por tanto, para cada valor de t, existirá una familia de distribuciones condicionales posibles que corresponden a los distintos valores posibles de $\theta \in \Omega$. Puede suceder, sin embargo, que para cada valor posible de t, la distribución conjunta condicional de $X_1, ...X_n$, dado que $T = t$, sea la misma para todos los valores de $\theta \in \Omega$ y , por tanto, realmente no depende del valor de θ. En este caso, se dice que T es un estadístico suficiente para el parámetro θ.

Por tanto, como un estimador de θ es un estadístico, en principio es posible deducir la distribución muestral de cualquier estimador de θ, entendiendo por distribución muestral a la distribución conjunta de las observaciones de una muestra aleatoria. Esto nos permitirá por ejemplo, calcular la probabilidad de que el estimador no difiera de θ más de un número específico de unidades o el E.C.M. de la estimación, antes de seleccionar la muestra y también permitirá calcular el tamaño muestral adecuado en un experimento concreto.

En la sección 3 vamos a recordar las distribuciones de probabilidad que aparecerán cuando hablemos de distribuciones muestrales de estadísticos.

Capítulo 3

Principales distribuciones en el muestreo

3.1. Distribución Normal

La distribución normal, es con mucho, la más importante de todas las distribuciones de probabilidad. Es una distribución de variable continua con un campo de variación de $]-\infty,\infty[$.

Debe su importancia a tres razones fundamentales, por un lado, un gran número de fenómenos reales se pueden modelizar con esta distribución, por otro lado, muchas de las distribuciones de uso frecuente tienden a aproximarse a la distribución normal bajo ciertas condiciones y por último en virtud del Teorema Central del Límite (que veremos a continuación), todas aquellas variables que puedan considerarse causadas por un gran número de pequeños efectos tienden a distribuirse con una distribución normal.

Se dice que una variable aleatoria X tiene una distribución normal con media μ y varianza σ^2 ($X \sim N(\mu,\sigma^2)$ con $-\infty < \mu < \infty, \sigma > 0$) si X tiene una distribución continua cuya función de densidad $f(x|\mu,\sigma^2)$ es la siguiente:

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{(2\pi)}\sigma} exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \quad \text{para} \quad -\infty < x < \infty \qquad (3.1)$$

La demostración de que esta función así definida es una f.d.p. puede encontrarse por ejemplo en [**?**]

Si $X \sim N(\mu,\sigma^2)$, su f.g.m

$$\psi_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} exp\left[tx - \frac{(x-\mu)^2}{2\sigma^2}\right] dx$$

Completando el cuadrado dentro de los paréntesis, se obtiene

$$tx - \frac{(x-\mu)^2}{2\sigma^2} = \mu t + \frac{1}{2}\sigma^2 t^2 - \frac{[x-\mu+\sigma^2 t]^2}{2\sigma^2}$$

Por tanto,

$$\psi_X(t) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} exp\left[\mu t + \frac{1}{2}\sigma^2 t^2 - \frac{[x-\mu+\sigma^2 t]^2}{2\sigma^2}\right] dx =$$

$$= exp[\mu t + \frac{1}{2}\sigma^2 t^2] \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} exp[-\frac{[x-(\mu+\sigma^2 t)]^2}{2\sigma^2}]dx$$

si llamamamos

$$C = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{[x-(\mu+\sigma^2 t)]^2}{2\sigma^2}} dx, \tag{3.2}$$

$$\psi_X(t) = exp[\mu t + \frac{1}{2}\sigma^2 t^2]C$$

Pero la ecuación (3.2) muestra la integral de la f.d.p. de una variable que sigue una ditribución $N(\mu + \sigma^2 t, \sigma^2)$, por lo que $C = 1$.

Por tanto, si $X \sim N(\mu, \sigma^2)$,

$$\begin{aligned} &\psi_X(t) = e^{(\mu t + \frac{1}{2}\sigma^2 t^2)} \quad \text{para} \quad -\infty < t < \infty \\ &E(X) = \psi'_X(0) = \mu \\ &Var(X) = \psi''_X(0) - [\psi'_X(0)]^2 = \sigma^2 \end{aligned} \tag{3.3}$$

Veamos ahora dos propiedades de la distribución normal:

Teorema 3.1.1 *Si X tiene una distribución normal con media μ y varianza σ^2 y si $Y = aX + b$ donde a y b son constantes y $a \neq 0$, entonces Y tiene una distribución normal con media $a\mu + b$ y varianza $a^2\sigma^2$*

Demostración La función generatriz de momentos ψ de X está dada por la ecuación

$$\psi_X(t) = exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right) \quad \text{para} \quad -\infty < t < \infty \tag{3.4}$$

En el teorema 1.3.1 hemos visto que si $Y = aX + b$, y ψ_Y es la f.g.m de Y, $\psi_Y(t) = e^{bt}\psi(at)$. Entonces

$$\psi_Y(t) = e^{bt}\psi_X(at) = exp\left[(a\mu + b)t + \frac{1}{2}a^2\sigma^2 t^2\right] \quad \text{para} \quad -\infty < t < \infty$$

Comparando esta expresión para ψ_Y con la f.g.m de una distribución normal (ecuación (3.3)), se observa que ψ_Y es la f.g.m de una distribución normal con media $a\mu + b$ y varianza $a^2\sigma^2$. Por tanto, Y debe tener esta distribución normal.

Teorema 3.1.2 *Si las variables aleatorias $X_1, ..., X_n$ son independientes y si X_i tiene una distribución normal con media μ_i y varianza $\sigma_i^2 (i = 1, ..., n)$, entonces la suma $X_1 + ... + X_n$ tiene una distribución normal con media $\mu_1 + ... + \mu_n$ y la varianza $\sigma_1^2 + ... + \sigma_n^2$*

Demostración Sea $\psi_i(t)$ la f.g.m de X_i para $i = 1, ..., n$ y sea $\psi(t)$ la f.g.m de $X_1 + ... + X_n$. Puesto que las variables $X_1, ..., X_n$ son independientes, entonces por el teorema 1.3.2 sabemos que

$$\begin{aligned}\psi(t) &= \prod_{i=1}^{n} \psi_i(t) = \prod_{i=1}^{n} exp\left(\mu_i t + \frac{1}{2}\sigma_i^2 t^2\right) \\ &= exp\left[\left(\sum_{i=1}^{n} \mu_i\right) t + \frac{1}{2}\left(\sum_{i=1}^{n} \sigma_i^2\right) t^2\right] \quad \text{para} \quad -\infty < t < \infty\end{aligned}$$

Esta f.g.m se puede identificar como la f.g.m de una distribución normal cuya media es $\sum_{i=1}^{n} \mu_i$ y cuya varianza es $\sum_{i=1}^{n} \sigma_i^2$. Por tanto, la distribución de $X_1 + ... + X_n$ debe ser esa distribución normal.

Definicion 1 *Convergencia en probabilidad: Una sucesión de variables aleatorias, $\{X_n\}_{n=1}^{\infty}$, converge en probabilidad a una variable aleatoria X (que puede degenerar en una constante K), y lo expresaremos como*

$$X_n \xrightarrow[n\to\infty]{} X$$

cuando se cumple que:

$$\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$$

Definicion 2 *Convergencia en distribución: Sea $X_1, X_2, ...$ una sucesión de variables aleatorias, y para n=1,2,..., sea F_n la función de distribución de X_n y sea X' otras variables aleatoria cuya función de distribución es F' continua sobre la recta real. Diremos que la sucesión $X_1, X_2, ...$ converge en distribución a la variable X' si*

$$\lim_{n\to\infty} F_n(x) = F'(x) \quad para \quad -\infty < x < \infty$$

Teorema 3.1.3 *Teorema central del límite: Sea $X_1, X_2, ..., X_n$ un conjunto de variables aleatorias, independientes e idénticamente distribuidas con media μ y varianza σ^2, $0 < \sigma^2 < \infty$. Sea:*

$$S_n = X_1 + ... + X_n$$

Entonces

$$[\lim_{n\to\infty} P(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq z) = \Phi(z)]$$

donde Φ es la función de distribución de una distribución normal con media 0 y desviación típica 1.

Demostración

Para esta demostración utilizaremos las propiedades de la función generatriz de momentos. Recordemos que si $Z \sim N(0, 1)$ entonces (ecuación (3.3))

$$\psi_Z(t) = e^{t^2/2}$$

Sea

$$Z_n = \frac{(\frac{1}{n}\sum_{i=1}^{n} x_i) - \mu}{\sqrt{\sigma^2/n}}$$

Demostraremos que

$$\psi_{Z_n}(t) \xrightarrow[n\to\infty]{} \psi_Z(t)$$

Esto nos garantiza que

$$\psi_{Z_n}(t) \xrightarrow[n\to\infty]{} \psi_Z(t) \quad \text{(converge en probabilidad)}$$

Ya que como hemos visto, cuando dos funciones tienen la misma función generatriz de momentos, siguen la misma distribución.

Escribimos $Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{x_i-\mu}{\sigma}$

$$\begin{aligned}\psi_{Z_n}(t) &= E\left(e^{tZ_n}\right) = E(e^{t/\sqrt{n}\sum_{i=1}^{n}\frac{x_i-\mu}{\sigma}})\Big) = \\ &= E\left(e^{t/\sqrt{n}(\frac{x_1-\mu}{\sigma})+\ldots+t/\sqrt{n}(\frac{x_n-\mu}{\sigma})}\right) = \\ &= \prod_{i=1}^{n} \psi_{\frac{x_i-\mu}{\sigma}}(\frac{t}{\sqrt{n}}) = \left[\psi_{\frac{x-\mu}{\sigma}}(\frac{t}{\sqrt{n}})\right]^n\end{aligned}$$

La última igualdad se produce porque para cada x_i la función generatriz de momentos es $\psi_{\frac{x_i-\mu}{\sigma}}(\frac{t}{\sqrt{n}})$, y aplicando que las variables están identicamente distribuidas obtenemos el resultado.

Expresamos la f.g.m como una serie de potencias

$$\psi_{Z_n}(t) = \left[1 + \frac{t}{\sqrt{n}}E(\frac{x-\mu}{\sigma}) + \frac{t^2}{2n}E(\frac{x-\mu}{\sigma})^2 + \theta(\frac{t^2}{n})\right]^n$$

Como $E(\frac{x-\mu}{\sigma}) = 0$ y $E(\frac{x-\mu}{\sigma})^2 = 1$,

$$\psi_{Z_n}(t) = \left[1 + \frac{t^2}{2n} + \theta(\frac{t^2}{n})\right]^n = (1 + \frac{t^2}{2n})^n + \theta(\frac{t^2}{n}) \xrightarrow[n\to\infty]{} e^{t^2/2}$$

3.2. Distribución Gamma

Se dice que una variable aleatoria X tiene una distribución gamma con parámetros α y β ($\alpha > 0 \quad y \quad \beta > 0$) si X tiene una distribución continua cuya función de densidad viene dada por la siguiente expresión:

$$f(x|\alpha,\beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x} & si \quad x > 0 \\ 0 & si \quad x \leq 0 \end{cases}$$

Donde la función $\Gamma(\alpha)$ es la función Gamma de Euler que representa la siguiente integral:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx \tag{3.5}$$

que verifica que:

Propiedad 3.2.1 $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$

Demostración Aplicaremos el método de integración por partes a la integral de la ecuación (3.5). Si se define

$$u = x^{\alpha-1} \quad \text{y} \quad dv = e^{-x} dx$$

entonces

$$du = (\alpha - 1)x^{\alpha-2} dx \quad \text{y} \quad v = -e^{-x}$$

Por tanto,

$$\begin{aligned} \Gamma(\alpha) \quad &= \int_0^\infty u dv = [uv]_0^\infty - \int_0^\infty v du \\ &= [-x^{\alpha-1} e^{-x}]_0^\infty + (\alpha - 1) \int_0^\infty x^{\alpha-2} e^{-x} dx \\ &= 0 + (\alpha - 1)\Gamma(\alpha - 1) \end{aligned}$$

Propiedad 3.2.2 *Para cualquier entero positivo n,* $\Gamma(n) = (n - 1)!$

Demostración Aplicaremos, como en la demostración anterior, el método de integración por partes. Si definimos

$$u = x^{\alpha-1} \quad \text{y} \quad dv = e^{-x} dx$$

entonces

$$du = (\alpha - 1)x^{\alpha-2} dx \quad \text{y} \quad v = -e^{-x}$$

entonces

$$\Gamma(\alpha) = (\alpha - 1) \int_0^\infty x^{\alpha-2} e^{-x} dx = (\alpha - 1)\Gamma(\alpha - 1)$$

y sucesivamente

$$\begin{aligned} \Gamma(\alpha) \quad &= (\alpha - 1)(\alpha - 2)...\Gamma(1) \\ \Gamma(1) \quad &= 1 \\ \Gamma(\alpha + 1) \quad &= \alpha\Gamma(\alpha) \end{aligned}$$

La integral de esta función de densidad es 1, puesto que de la definición de la función gamma resulta que

$$\int_0^\infty x^{\alpha-1} e^{-\beta x} dx = \frac{\Gamma(\alpha)}{\beta^\alpha}$$

Si X tiene una distribución gamma con parámetros α y β, entonces los momentos de X se determinan a partir de las ecuaciones anteriores. Para $k = 1, 2, ...$, resulta que:

$$\begin{aligned} E(X^k) &= \int_0^\infty x^k f(x|\alpha,\beta)dx = \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty x^{\alpha+k-1}e^{\beta x}dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha+k)}{\beta^{\alpha+k}} = \frac{\Gamma(\alpha+k)}{\beta^k\Gamma(\alpha)} \\ &= \frac{\alpha(\alpha+1)...(\alpha+k-1)}{\beta^k} \end{aligned}$$

En particular:

$$\begin{aligned} E(X) &= \frac{\alpha}{\beta} \\ Var(X) &= \frac{\alpha(\alpha+1)}{\beta^2} \end{aligned}$$

3.3. Distribución χ^2

La distribución gamma con parámetros $\alpha = n/2$ y $\beta = 1/2$ para cualquier entero n positivo, se denomina distribución χ^2 con n grados de libertad, y se denota χ^2_n. Si una variable aleatoria X tiene una distribución χ^2 con n grados de libertad,de la ecuación de la función de densidad para una distribución Gamma obtenemos que la función de distribución de X es:

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)}x^{(n/2)-1}e^{-x/2}$$

Si una variable aleatoria X tiene una distribución χ^2 con n grados de libertad, resulta de las expresiones para la media y la varianza de la distribución gamma, que

$$\begin{aligned} E(X) &= n \\ Var(X) &= 2n \end{aligned}$$

Teorema 3.3.1 *Si las variables aleatorias $X_1, ..., X_k$ son independientes y si X_i tiene una distribución χ^2 con n_i grados de libertad $(i = 1, ..., k)$, entonces la suma $X_1 + ... + X_k$ tiene una distribución χ^2 con $n_1 + ... + n_k$ grados de libertad.*

Teorema 3.3.2 *Si las variables aleatoria $X_1, \cdots, X_k$ son i.i.d y cada una de ellas sigue una distribución $N(0, 1)$, entonces la suma de cuadrados $X_1^2 + \cdots + X_k^2$ sigue una distribución χ^2 con k grados de libertad.*

3.4. Distribución F de Snedecor-Fisher

Es una distribución de probabilidad de gran aplicación en la inferencia estadística, fundamentalmente en la contrastación de la igualdad de varianzas de dos poblaciones normales, y en el análisis de la varianza. La distribución F es una distribución continua de muestreo de dos variables aleatorias independientes con distribuciones χ^2, cada una de las cuales se divide entre sus grados de libertad.

Consideramos dos variables aleatorias independientes X e Y, tales que:

$$Y \sim \chi^2_m \quad \text{(con m grados de libertad)}$$
$$Z \sim \chi^2_n \quad \text{(con n grados de libertad)}$$

donde m y n son enteros y positivos. Si establecemos el cociente de ambas variables, divida cada una además por sus grados de libertad obtenemos:

$$X = \frac{\frac{Y}{m}}{\frac{Z}{n}} = \frac{nY}{mZ}$$

La distribución de la variable X se denomina distribución F con m y n grados de libertad, y se representa $F_{m,n}$.

Se demostrará a continuación que si la variable aleatoria X tiene una distribución F con m y n grados de libertad, entonces su f.d.p $f(x)$ es la siguiente:

$$f(x) = \frac{\Gamma\left[\frac{1}{2}(m+n)\right] m^{m/2} n^{n/2}}{\Gamma(\frac{1}{2}m)\Gamma(\frac{1}{2}n)} \frac{x^{(m/2)-1}}{(mx+n)^{(m+n)/2}} \quad \text{para } x > 0$$

Como las variables aleatorias Y y Z son independientes, su f.d.p conjunta $g(y,z)$ será el producto de sus f.d.p individuales. Además, puesto que Y y Z tienen distribuciones χ^2, $f(y,z)$ tiene la siguiente forma:

$$g(y,z) = cy^{(m/2)-1} z^{(n/2)-1} e^{-(y+z)/2} \quad \text{para } y > 0, z > 0 \tag{3.6}$$

donde

$$c = \frac{1}{2^{(m+n)/2}\Gamma(\frac{1}{2}m)\Gamma(\frac{1}{2}n)} \tag{3.7}$$

Realizamos un cambio de variable de Y y Z a X y Z, donde X esta definida como $X = \frac{nY}{mZ}$y $Y = (m/n)XZ$. Aplicando el teorema 1.2.1 la función de densidad conjunta $h(x,y)$ de X y Z la obtenemos reemplazando en la ecuación (3.6) por su expresión en función de x y z y multiplicando por el jacobiano de la transformación, $(m/n)z$, la función de densidad conjunta de X y Z será:

$$h(x,z) = c(\frac{m}{n})^{m/2} x^{(m/2)-1} z^{[(m+n)/2]-1} e^{-\frac{1}{2}(\frac{m}{n}x+1)z}$$

La constante c viene dada por la ecuación (3.7).

La f.d.p marginal $f(x)$ de X se puede obtener para cualquier valor de $x > 0$ a partir de la relación

$$f(x) = \int_0^\infty h(x, z)dz$$

Sabemos que:

$$\int_0^\infty x^{\alpha-1}e^{-\beta x}dx = \frac{\Gamma(\alpha)}{\beta^\alpha}$$

Por tanto

$$\int_0^\infty z^{[(m+n)/2]-1}e^{-\frac{1}{2}(\frac{m}{n}x+1)z}dz = \frac{\Gamma\left[\frac{1}{2}(m+n)\right]}{\left[\frac{1}{2}(\frac{m}{n}x+1)\right]^{(m+n)/2}}$$

De las ecuaciones anteriores se puede concluir que la f.d.p $f(x)$ tiene la forma:

$$f(x) = \frac{\Gamma\left[\frac{1}{2}(m+n)\right]m^{m/2}n^{n/2}}{\Gamma(\frac{1}{2}m)\Gamma(\frac{1}{2}n)}\frac{x^{(m/2)-1}}{(mx+n)^{(m+n)/2}} \quad \text{para } x > 0$$

3.5. Distribución t-Student

En cuanto a la estimación de medias y como más adelante veremos es necesario definir la distribución t-Student que utilizaremos en los casos en que la desviación típica poblacional sea desconocida y debamos trabajar con las desviación típica muestral.

La distribución t-Student es la distribución del cociente:

$$T = \frac{Z}{\sqrt{\frac{V}{\upsilon}}} = Z\sqrt{\frac{\upsilon}{V}} \tag{3.8}$$

donde

- Z es una variable aleatoria distribuida según una normal tipificada.
- V es una variable aleatoria que sigue una distribución χ^2 con υ grados de libertad.
- Z y V son independientes.

Veamos como obtener la función de densidad:

Supongamos que la distribución conjunta de Z y V es como hemos indicado en la definición de la distribución t. Entonces, puesto que Z y V son independientes, su función de densidad conjunta es igual al producto $f_1(z)f_2(\upsilon)$, donde $f_1(z)$ es la función de densidad de la distribución $N(0, 1)$ y $f_2(\upsilon)$ es la función de densidad de la distribución χ^2_n.

Sea T definida por la ecuación (3.8) y definimos variable auxiliar $W = V$. En primer lugar, determinaremos la función de densidad conjunta de T y W De las definiciones de T y W,

$$Z = \frac{1}{\sqrt{n}} T\sqrt{W} \quad \text{y} \quad V = W \tag{3.9}$$

El jacobiando de estas transformaciones de T y W a Z y V es $\sqrt{W/n}$.Aplicando el teorema 1.2.1 la función de densidad conjunta de $f(t, w)$ de T y W se puede obtener de la función de densidad conjunta $f_1(z)f_2(v)$ reemplazando z y v por las expresiones en (2) y multiplicando el resultado por $\sqrt{W/n}$. El valor de $f(t, w)$ para $-\infty < t < \infty$ y $w > 0$, obtenemos:

$$f(t, w) = cw^{(n-1)/2} exp\left[-\frac{1}{2}\left(1 + \frac{x^2}{n}\right) w\right] \tag{3.10}$$

donde

$$c = \left[2^{(n+1)/2}\sqrt{n\pi}\Gamma(\frac{n}{2})\right]^{-1}$$

La función de densidad marginal $g(t)$ de X se puede obtener de la ecuación (3) utilizando la relación
$g(t) = \int_0^\infty f(t, w)dw$
Se obtiene

$$g(t) = \frac{\Gamma((\upsilon + 1)/2)}{\sqrt{\upsilon\pi}\Gamma(\upsilon + 2)}(1 + t^2/\upsilon)^{-(\upsilon+1)/2} \quad \text{para} \quad -\infty < t < \infty$$

Teorema 3.5.1 *Teorema de Fisher: Sea $(X_1, ...X_n)$ una muestra aleatoria simple de tamaño n, procedente de una población $N(\mu, \sigma^2)$. Sea $\bar{X}_n = (X_1 + ... + X_n)/n$ la media muestral y sea $S_n^2(x) = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$ la cuasivarianza muestral. Entonces se verifica que:*

1. *Los estadísticos $\bar{X}_n$ y S_n^2 son independientes.*
2. *El estadístico*
$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}.$$
3. *El estadístico*
$$\frac{\bar{X} - \mu}{S_n/\sqrt{n}} \sim t_{n-1}.$$

Demostración [**?**]

1. Por los teoremas 3.1.1 y 3.1.2 sabemos que

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Para demostrar que los estadísticos media $\bar{X}_n$ y cuasivarianza muestral S_n^2, son independientes, demostraremos que $\bar{X}_n$ es independiente de $x_i - \bar{X}_n$ para

cada i, y procederemos directamente calculando la función generatriz de momentos conjunta de $\bar{X}_n$ y $x_i - \bar{X}_n$ y tenemos:

$$\begin{aligned}
\psi(t_1,t_2) &= E\left[e^{t_1\bar{X}_n + t_2(x_i-\bar{X}_n)}\right] = E\left[e^{t_2x_i+(t_1-t_2)\bar{X}_n}\right] = \\
&= E\left[e^{t_2x_i+(t_1-t_2)(\frac{x_1+\ldots+x_n}{n})}\right] = E\left[e^{(t_2+\frac{t_1-t_2}{n})x_i+\frac{(t_1-t_2)}{n}\sum_{i=1,j\neq i}^{n}x_j}\right] = \\
&= E\left[e^{(t_2+\frac{t_1-t_2}{n})x_i}\right]E\left[e^{\frac{(t_1-t_2)}{n}\sum_{i=1,j\neq i}^{n}x_j}\right] = \\
&= e^{(t_2+\frac{t_1-t_2}{n})\mu+\frac{1}{2}(t_2+\frac{t_1-t_2}{n})^2\sigma^2}e^{\frac{n-1}{n}(t_1-t_2)\mu+(\frac{t_1-t_2}{n})^2(n-1)\frac{\sigma^2}{2}} = \\
&= e^{t_1\mu+\frac{1}{2}\frac{t_1^2}{n}\sigma^2}e^{\frac{1}{2}t_2^2\frac{n-1}{n}\sigma^2}
\end{aligned}$$

que son las funciones generatrices de momentos correspondientes a una

$$N(\mu,\frac{\sigma^2}{n}) \quad \text{y} \quad N\left(0,\sigma^2\frac{n-1}{n}\right)$$

respectivamente, con lo cual hemos demostrado que:

$\bar{X}_n$ y $x_i - \bar{X}_n$ son independientes y en consecuencia también son independientes $\bar{X}_n$ y $\sum_{i=1}^{n}(x_i-\bar{X}_n)^2$ y por tanto $\bar{X}_n$ y S_n^2 son independientes.

2. Para demostrar que el estadístico $\frac{(n-1)S_n^2}{\sigma^2}$ sigue una χ^2_{n-1} partiremos del estadístico cuasivarianza muestral

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{X}_n)^2$$

de donde podemos escribir:

$$\begin{aligned}
(n-1)S_n^2 &= \sum_{i=1}^{n}(x_i-\bar{X}_n)^2 = \sum_{i=1}^{n}(x_i-\mu+\mu-\bar{X}_n)^2 = \\
&= \sum_{i=1}^{n}[(x_i-\mu)^2-(\bar{X}_n-\mu)^2] = \\
&= \sum_{i=1}^{n}[(x_i-\mu)^2-2(x_i-\mu)(\bar{X}_n-\mu)+(\bar{X}_n-\mu)^2] = \\
&= \sum_{i=1}^{n}(x_i-\mu)^2-2(\bar{X}_n-\mu)\sum_{i=1}^{n}(x_i-\mu)+n(\bar{X}_n-\mu)^2 = \\
&= \sum_{i=1}^{n}(x_i-\mu)^2-2(x_i-\mu)n(\bar{X}_n-\mu)+n(\bar{X}_n-\mu)^2 = \\
&= \sum_{i=1}^{n}(x_i-\mu)^2-n(\bar{X}_n-\mu)^2
\end{aligned}$$

y de aquí se tiene

$$\sum_{i=1}^{n}(x_i-\mu)^2=(n-1)S_n^2+n(\bar{X}_n-\mu)^2$$

dividiendo ambos miembros por la varianza poblacional resulta:

$$\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}=\frac{(n-1)S_n^2}{\sigma^2}+\frac{n(\bar{X}_n-\mu)^2}{\sigma^2}$$

O bien

$$\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2=\frac{(n-1)S_n^2}{\sigma^2}+\left(\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}\right)^2$$

Teniendo en cuenta la definición de χ_n^2 y su propiedad reproductiva resulta que:

$$\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2\sim\chi_n^2$$

pues tenemos una suma de variables aleatorias $N(0,1)$ independientes y elevadas al cuadrado.

Análogamente,

$$\left(\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}\right)^2\sim\chi_1^2$$

pues se trata de una variable aleatoria $N(0,1)$ y elevada al cuadrado.

Por tanto,

$$\frac{(n-1)S_n^2}{\sigma^2}\sim\chi_{n-1}^2$$

3. Sabemos que

$$\bar{X}_n\sim N(\mu,\frac{\sigma^2}{n})$$

y

$$\frac{(n-1)S_n^2}{\sigma^2}\sim\chi_{n-1}^2$$

y que los estadísticos $\bar{X}_n$ y S_n^2 son independientes. Tipificando la variable $\bar{X}_n$ se tiene

$$\left(\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}\right)^2\sim N(0,1)$$

pero incluye el parámetro σ desconocido que es conveniente eliminar.

Recordemos que la variable aleatoria t-Student estaba definida como un cociente entre una variable aleatoria $N(0,1)$ y la raíz cuadrada de una variable aleatoria χ^2 dividida por sus grados de libertad, ambas independientes, luego podemos escribir:

$$T = \frac{Z}{\sqrt{\frac{V}{(n-1)}}} = \frac{\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right)^2}{\sqrt{\frac{\frac{(n-1)S_n^2}{\sigma^2}}{(n-1)}}} = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim t_{n-1}$$

Teorema 3.5.2 *Relación entre la distribución t y la distribución normal*

Sea X una variable aleatoria con distribución t con n grados de libertad cuya función de densidad es $g(x) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n+2)}(1 + x^2/n)^{-(n+1)/2}$ se cumple que:

$$\lim_{n\to\infty} g(x) = \Phi(x)$$

donde $\Phi(x)$ es la fucnión de distribución normal tipificada.

Es decir, para cada valor $x, (-\infty < x < \infty)$ la f.d.p $g(x)$ converge a la f.d.p $\Phi(x)$. Por tanto, cuando n es grande, la distribución t con n grados de libertad se puede aproximar por la distribución normal tipificada.

Teorema 3.5.3 *(Relación entre las distribución t y la distribución F)*

Si una variable aleatoria $X \sim t_n$, entonces X^2 tendrá una distribución F con 1 y n grados de libertad.

Este resultado se deduce de la ecuación

$$X = \frac{Z}{\sqrt{\frac{V}{\upsilon}}} = Z\sqrt{\frac{\upsilon}{V}}$$

$\longrightarrow X^2 = \frac{Z^2}{\frac{V}{\upsilon}}$ donde $Z \sim N(0,1)$ y $V \sim \chi^2_\upsilon$.

Si $Z \sim N(0,1)$ por el teorema 3.3.2 sabemos que $Z^2 \sim \chi^2_1$, y aplicando la definición de distribución F tenemos el resultado que buscamos.

3.6. Distribución Bernoulli

Se dice que una variable aleatoria X tiene una distribución de Bernoulli con parámetro p $(0 \leq p \leq 1)$ si X puede tomar únicamente los valores 0 y 1 y las probabilidades son:

$$P(X = 1) = p \qquad P(X = 0) = 1 - p$$

Si definimos $q = 1 - p$ la función de probabilidad (f.p) de X se puede escribir como:

$$f(x|p) = \begin{cases} p^x q^{1-x} & si \qquad x = 0, 1 \\ 0 & en\ otro\ caso \end{cases}$$

Si X tiene una distribución de Bernoulli con parámetro p, entonces:

$$E(X) = 1 \cdot p + 0 \cdot q = p$$
$$E(X^2) = 1^2 \cdot p + 0^2 \cdot q = p$$
$$Var(X) = E(X^2) - [E(X)]^2 = pq$$

Si las variables aleatorias $X_1, X_2, ...$ son una sucesión infinita de variables identicamente distribuidas y si cada variable aleatoria X_i tiene una distribución Bernoulli con parámetro p, entonces se dice que las variables aleatorias $X_1, X_2, ...$ constituyen una sucesión infinita de pruebas Bernoulli con parámetro p. Análogamente, si n variables aleatorias $X_1, X_2, ..., X_n$ son identicamente distribuidas y cada una tiene una distribución de Bernoulli con parámetro p, entonces se dice que las variables $X_1, X_2, ...X_n$ constituyen n pruebas Bernoulli con parámetro p.

3.7. Distribución Binomial

La distribución binomial es una distribución de probabilidad discreta que cuenta el número de éxitos en una secuencia de n ensayos de Bernoulli independientes entre sí, con una probabilidad fija p de ocurrencia del éxito entre los ensayos. Un experimento de Bernoulli se caracteriza por ser dicotómico, a uno de los sucesos se le denomina éxito y tiene una probabilidad de ocurrencia p y al otro, fracaso, con una probabilidad $q = 1 - p$. La forma habitual de representar que una variable aleatoria sigue la distribución binomial se representa de la siguiente forma:

$$X \sim B(n, p)$$

donde n es el número de sucesos y p la probabilidad de éxito.

La función de probabilidad de la binomial es:

$$f(x) = \begin{cases} \binom{n,x}{p}^x (1-p)^{n-x} & \text{si} \quad x = 0, 1, ..., n \\ 0 & \text{en otro caso} \end{cases}$$

En [**?**] puede encontrarse la demostración de que esta función definida es una f.d.p.

Si las variables aleatorias $X_1, ..., X_n$ constituyen n pruebas de Bernoulli con parámetro p y si $X = X_1 + ... + X_n$, entonces X tiene una distribución binomial con parámetros n y p.

La esperanza y varianza de esta distribución son:

$$
\begin{aligned}
E(x) \quad &= \sum_{k=0}^{n} k \binom{n,k}{p}^{k} (1-p)^{n-k} = \sum_{k=1}^{n} k \binom{n,k}{p}^{k} (1-p)^{n-k} = \\
&= \sum_{k=1}^{n} k \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} = \sum_{k=1}^{n} k \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} = \\
&= np \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} = \\
&= np \sum_{k=1}^{n} \binom{n-1,k-1}{p}^{k-1} (1-p)^{(n-1)(k-1)} =_{(k-1)=j} np \sum_{k=1}^{n} \binom{n-1,j}{p}^{j} (1-p)^{n-1-j} = \\
&= np(p+(1-p))^{n-1} = np
\end{aligned}
$$

Recordemos que:

$$V(X) = E(X^2) - (E(X))^2 = E(X^2) - n^2 p^2$$

Por tanto:

$$
\begin{aligned}
E(X^2) \quad &= \sum_{k=0}^{n} k^2 \binom{n,k}{p}^{k} (1-p)^{n-k} = \sum_{k=0}^{n} (k(k-1)+k) \binom{n,k}{p}^{k} (1-p)^{n-k} = \\
&= \sum_{k=0}^{n} k(k-1) \binom{n,k}{p}^{k} (1-p)^{n-k} + \sum_{k=0}^{n} k \binom{n,k}{p}^{k} (1-p)^{n-k} = \\
&= \sum_{k=2}^{n} k(k-1) \binom{n,k}{p}^{k} (1-p)^{n-k} + E(X) = \\
&= \sum_{k=2}^{n} k(k-1) \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} + np = \sum_{k=2}^{n} \frac{n!}{(k-2)!(n-k)!} p^k (1-p)^{n-k} + np = \\
&= n(n-1)p^2 \sum_{k=2}^{n} \frac{(n-2)!}{(k-2)!(n-k)!} p^{k-2} (1-p)^{n-k} + np = \\
&=_{(k-2)=j} n(n-1)p^2 \sum_{j=0}^{n-2} \binom{n-2,j}{p}^{j} (1-p)^{n-2-j} + np = \\
&= n(n-1)p^2 (p+(1-p))^{n-2} + np = n(n-1)p^2 + np
\end{aligned}
$$

Finalmente:

$$V(X) = E(X^2) - (E(X))^2 = n(n-1)p^2 + np - n^2 p^2 = -np^2 + np = np(1-p)$$

Figura 3.1: Histograma y función de densidad

Teorema 3.7.1 *Teorema de Moivre-Laplace: Sea* $\{ X_i\}_{i=1}^{\infty}$ *una sucesión de v.a. de manera que cada una de ellas tenga una distribución* $X_i \sim B(n,p)$*. Entonces, la nueva sucesión*

$$\eta_i = \frac{X_i - E|X_i|}{\sqrt{Var(X_i)}} = \frac{X_n - np}{npq}$$

converge en probabilidad a una distribución $N(0,1)$

Del teorema de Moivre-Laplace se deduce que una distribución binomial puede aproximarse a una distribución normal de media np y desviación típica $\sqrt{npq}$ para un n grande.

3.7.1. Corrección por continuidad o corrección de Yates

Cuando aproximamos una distribución binomial mediante una normal, estamos aproximando una variable X discreta por una continua X'.

Supongamos el caso general de que queremos aproximar $f(x)$, la función de probabilidad de X, por una distribución continua con función de densidad de probabilidad $g(x)$. Si $g(x)$ proporciona una buena aproximación de la distribución de X, entonces para dos enteros cualesquiera a y b $(a < b)$ se puede aproximar la probabilidad

$$P(a \leq X \leq b) = \sum_{x=a}^{b} f(x)$$

por la integral

$$\int_a^b g(x)dx \tag{3.11}$$

Pero esta aproximación presenta problemas, ya que aunque $P(X \geq a)$ y $P(X > a)$ en general tendrán valores distintos para la distribución discreta, estas probabilidades siempres serán iguales para la distribución continua. Además, los valores de la probabilidad para valores fijos de la variable continua son cero. Para evitar este problema debemos introducir correcciones en la aproximación.

La función de probabilidad de X se puede representar mediante el histograma que se muestra en la figura 3.1.

Para cada entero x, la probabilidad de x representada por el área de un rectángulo cuya base se estiende desde $x - \frac{1}{2}$ hasta $x + \frac{1}{2}$ y cuya altura es $f(x)$, entonces el área del rectángulo cuya base está centrada en el entero x es $f(x)$. En la figura vemos representado $g(x)$ que es la distribución continua por la que queremos aproximar $f(x)$.

Desde este punto de vista se puede observar que $P(a \leq X \leq b)$, es la suma de las áreas de los rectángulos de la figura que están centrados en $a, a+1, ..., b$. Se puede observar también que la suma de estas áreas se aproxima por la integral

$$\int_{a-(1/2)}^{b+(1/2)} g(x)dx \tag{3.12}$$

El ajuste de la integral (3.11) se llama corrección por continuidad. Análogamente, si seguimos llamando X a la variable discreta cuya f.p. queremos aproximar por la f.d.p. de la variable continua X':

- $P(X = a) = P(a - 0{,}5 \leq X' \leq a + 0{,}5)$
- $P(X \leq a) = P(X' \leq a + 0{,}5)$ (para que contenga al punto a)
- $P(X < a) = P(X' \leq a - 0{,}5)$ (para que no contenga al punto a)
- $P(X > a) = P(X' \geq a + 0,5)$ (para que no contenga al punto a)
- $P(X \geq a) = P(X' \geq a - 0{,}5)$ (para que contenga al punto a)
- $P(a \leq X < b) = P(a - 0{,}5 \leq X' \leq b + 0{,}5)$ (para que contenga al punto a y no a b)

Capítulo 4

Intervalos de confianza

Un intervalo de confianza (IC) es un intervalo de extremos aleatorios que con un nivel de confianza determinado, contiene el verdadero valor del parámetro y debe presentarse junto a la estimación puntual de un parámetro, puesto que permite cuantificar la magnitud del error asociado a la estimación o error muestral. Veamos su definción para el caso general:

Definicion 3 *Dada una muestra aleatoria $X_1, \cdots, X_n$ de una distribución con parámetro θ desconocido, supongamos que podemos encontrar dos estadísticos $A(X_1, \cdots, X_n)$ y $B(X_1, \cdots, X_n)$ tales que*

$$P(A(X_1, \cdots, X_n) < \theta < B(X_1, \cdots, X_n)) = 1 - \alpha,$$

donde $1 - \alpha$ es una probabilidad fija $(0 < \alpha < 1)$. Si llamamos a y b a los respectivos valores observados de estos estimadores, diremos que el intervalo (a, b) es un intervalo de confianza para θ con un nivel de confianza $1 - \alpha$, o en otras palabras, que θ está en el intervalo (a, b) con una confianza $1 - \alpha$.

La amplitud del intervalo de confianza está directamente relacionada con el error muestral, y veremos que depende del tamaño de la muestra, por lo que el tamaño muestral mínimo estará en función del error máximo que se considere admisible.

En este apartado veremos el cálculo del tamaño muestral necesario para un estudio basándonos en la amplitud máxima admisible para el intervalo de confianza del parámetro que se pretende estimar. Por tanto, el tamaño muestral para la estimación de un parámetro depende de diversos factores:

- El nivel de confianza
- Precisión de la estimación: esta viene dada a través del error muestral. El máximo margen de error admisible en este caso fijado por el investigador ya que general tenemos que al aumentar la precisión aumenta también el tamaño muestral por lo que exigir una precisión muy elevada devuelve un tamaño muestral inviable para la realización del estudio.
- Otros elementos: estos serán definidos en cada caso.

4.1. I.C. y tamaño muestral para estimar la media de una distribución normal

Hemos visto (propiedad 2.1.1) que el estimador máximo verosímil de la media poblacional es la media muestral, y conocemos su distribución muestral por lo tanto en adelante trabajaremos con estos datos para realizar inferencias sobre μ y para determinar el tamaño muestral necesario para estimar μ bajo ciertas condiciones.

En los teoremas 3.1.1 y 3.1.2 hemos visto que dada una sucesión $X_1, ..., X_n$ de v.a. i.i.d donde cada $X_i \sim N(\mu, \sigma^2)$, entonces la media muestral $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$.

Además, por el teorema central del límite (Teorema 3.1.3) si $X_1, X_2, ..., X_n$ es un conjunto de variables aleatorias, independientes e idénticamente distribuidas con media μ y varianza σ^2, $0 < \sigma^2 < \infty$, entonces $[\lim_{n\to\infty} P(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq z) = \Phi(z)]$, siendo $S_n = X_1 + ... + X_n$ y $\Phi()$ la función de distribución de una distribución $N(0, 1)$.

Por lo tanto, tanto en el caso de tener observaciones que proceden de una distribución gaussiana con varianza conocida, como en el caso de tener tamaños muestrales grandes, podremos suponer que la media muestral sigue una distribución gaussiana.

En caso de no conocer la varianza poblacional de la distribución de los datos, y tener que trabajar con la varianza o con la cuasivarianza muestral, en el teorema 3.5.1 hemos probado que $\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim t_{n-1}$.

Por lo tanto, a la hora de calcular intervalos de confianza para la media de una distribución normal y calcular los tamaños muestrales adecuados para acotar los errores de estimación, deberemos distinguir dos casos, pero antes veremos un factor de correción para utilizar en caso de estar trabajando con poblaciones finitas.

4.1.1. Factor de corrección

Si se seleccionan muestras aleatorias de n observaciones independientes de una población con media μ y desviación estándar σ, entonces, cuando n es grande, la distribución muestral de medias tendrá aproximadamente una distribución normal con una media igual a μ y una desviación típica de $\frac{\sigma}{\sqrt{n}}$. La aproximación será cada vez más exacta a medida que n sea cada vez mayor. Denotamos σ_x como a la desviación típica de la distribución de media.

Si el muestreo se hace sin reemplazamiento en una población finita de tamaño N, las variables $X_1, X_2, ..., X_n$ no son independientes y en este caso:

$$E(\bar{X}) = \mu \qquad Var(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$$

Al término $\frac{N-n}{N-1}$ se le denomina factor de corrección para una población finita. Para trabajar en caso general sin distinguir entre poblaciones finitas e infinitas utilizaremos la notación σ_x para hablar de la desviación típica de la distribución de la media muestral, donde

Figura 4.1: Distribución normal tipificada

$$\sigma_x = \begin{cases} \frac{\sigma}{\sqrt{n}} & \text{caso poblaciones infinitas} \\ \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} & \text{caso poblaciones finitas de tamaño } N \end{cases}$$

4.1.2. Tamaño muestral necesario para la estimación de una media con desviación típica conocida (o tamaños muestrales grandes)

Partimos de que podemos suponer que $\bar{X}_n \sim N(\mu, \sigma_x^2)$, siendo $\bar{X}_n$ la media de n observaciones, entonces:

$$Z = \frac{\bar{X}_n - \mu}{\sigma_x} \sim N(0,1)$$

Si llamamos $Z_{\frac{\alpha}{2}}$ al percentil $1 - \frac{\alpha}{2}$ de la distribución $N(0,1)$ i.e. aquel valor t.q. $P(Z \geq Z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$ (ver figura 4.1), tenemos que:

$$\begin{aligned} &P\left(-Z_{\frac{\alpha}{2}} \leq \frac{\bar{X}_n - \mu}{\sigma_x} \leq Z_{\frac{\alpha}{2}}\right) = \\ &P\left(-Z_{\frac{\alpha}{2}}\sigma_x \leq \bar{X}_n - \mu \leq Z_{\frac{\alpha}{2}}\sigma_x\right) = \\ &P\left(-\frac{\alpha}{2}\sigma_x - \bar{X}_n \leq -\mu \leq Z_{\frac{\alpha}{2}}\sigma_x - \bar{X}_n\right) = \\ &P\left(\bar{X}_n - Z_{\frac{\alpha}{2}}\sigma_x \leq \mu \leq \bar{X}_n + Z_{\frac{\alpha}{2}}\sigma_x\right) = 1 - \alpha \end{aligned}$$

El intervalo de confianza para μ es:

$$\left[\bar{X}_n - Z_{\frac{\alpha}{2}}\sigma_x, \bar{X}_n + Z_{\frac{\alpha}{2}}\sigma_x\right]$$

Si definimos el error de estimación como la mitad de la amplitud del intervalo de confianza:

$$e = Z_{\frac{\alpha}{2}}\sigma_x$$

- En el caso de poblaciones infinitas, $\sigma_x = \sigma/\sqrt{n}$ y:

$$e = Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

Despejamos n y obtenemos:

$$\boxed{n = \frac{Z_{\frac{\alpha}{2}}^2\sigma^2}{e^2}}$$

Figura 4.2: Distribución t

- En el caso de poblaciones finitas, $\sigma_x = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ y:

$$e = Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Despejamos n y obtenemos:

$$\boxed{n = \frac{N Z^2_{\frac{\alpha}{2}} \sigma^2}{e^2(N-1) + Z^2_{\frac{\alpha}{2}} \sigma^2}}$$

Los estudios epidemiológicos, se realizan en general para poblaciones finitas, por tanto, deberemos introducir habitualmente el factor de corrección para este caso.

4.1.3. Tamaño muestral necesario para la estimación de una media con desviación típica desconocida

Veamos ahora la construcción del intervalo de confianza para estimar una media en caso de no conocer la desviación típica poblacional. En este caso, sabemos que

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim t_{n-1}$$

donde S_n es la cuasidesviación típica muestral. Al conocer la distribución de este estimador, podemos proceder de forma análoga a la anterior. Si llamamos $t_{(n-1,\frac{\alpha}{2})}$ al percentil $1-\frac{\alpha}{2}$ de la distribución t de Student con $n-1$ grados de libertad (ver figura 4.2), tenemos:

$$\begin{aligned}
&P\left(-t_{(n-1,\frac{\alpha}{2})} \leq \frac{\bar{X}_n - \mu}{\frac{S_n}{\sqrt{n}}} \leq t_{(n-1,\frac{\alpha}{2})}\right) = \\
&= P\left(-t_{(n-1,\frac{\alpha}{2})}\frac{S_n}{\sqrt{n}} \leq \bar{X}_n - \mu \leq t_{(n-1,\frac{\alpha}{2})}\frac{S_n}{\sqrt{n}}\right) = \\
&= P\left(-t_{(n-1,\frac{\alpha}{2})}\frac{S_n}{\sqrt{n}} - \bar{X}_n \leq \ -\mu \leq t_{(n-1,\frac{\alpha}{2})}\frac{S_n}{\sqrt{n}} - \bar{X}_n\right) = \\
&= P\left(\bar{X}_n - t_{(n-1,\frac{\alpha}{2})}\frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + t_{(n-1,\frac{\alpha}{2})}\frac{S_n}{\sqrt{n}}\right) = 1-\alpha
\end{aligned}$$

El intervalo de confianza para μ es

$$\left[\bar{X}_n - t_{(n-1,\frac{\alpha}{2})}\frac{S_n}{\sqrt{n}}, \bar{X}_n + t_{(n-1,\frac{\alpha}{2})}\frac{S_n}{\sqrt{n}}\right]$$

En este caso no es tan sencillo despejar n para obtener la fórmula del tamaño muestral. Si definimos de nuevo el error como la mitad de la amplitud del intervalo de confianza:

$$e = t_{n-1,-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}$$

$$n = t^2_{n-1,-\frac{\alpha}{2}}\frac{S_n^2}{e^2},$$

el tamaño muestral aparece también en los grados de libertad de la distribución *t-Student*, por lo que deberíamos tomar una aproximación (aproximar el percentil de la distribción t por el percentil de una $N(0,1)$) para poder obtener el valor de n.

4.2. Estimar una proporción

En este apartado demostraremos la fórmula para el cálculo del tamaño muestral para el caso en que el parámetro que deseamos estimar sea una proporción. Aunque no lo hemos incluido en la memoria, es fácil demostrar que dada una distribución $B(n,p)$, la proporción muestral $\hat{p}$ es un estimador insesgado de p.

4.2.1. I.C y tamaño muestral para estimar una proporción

En numerosas ocasiones se plantea estimar una proporción o porcentaje, en estos dos casos la variable aleatoria toma solamente dos valores diferentes (éxito o fracaso), y para calcular una proporción nos interesa conocer la variable X "número total de éxitos", por lo tanto, $X \sim B(n,p)$. Cuando la extensión de la población es grande, podemos aproximar la distribución binomial $B(n,p)$ por la normal $N(np, npq)$ (ver Teorema 3.7.1). Por tanto, podemos suponer que para muestras de tamaño grande, la distribución muestral del estimador de una proporción (la porporción muestral) sigue una distribución normal donde el estimador puntual de p es:

$$\hat{p} = \frac{X}{n} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Por consiguiente,

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

Pero en vez de trabajar con esta variable, que complica mucho los cálculos, se define:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \approx \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

Bajo esta suposición, si llamamos $Z_{\frac{\alpha}{2}}$ al percentil $1 - \frac{\alpha}{2}$ de la distribución $N(0,1)$:

$$\begin{aligned}
&P\left(-Z_{\frac{\alpha}{2}} \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq Z_{\frac{\alpha}{2}}\right) = \\
&= P\left(-Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq \hat{p} - p \leq Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = \\
&= P\left(-Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} - \hat{p} \leq p \leq Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} - \hat{p}\right) = \\
&= P\left(\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha
\end{aligned}$$

El intervalo de confianza para p es:

$$\left[\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Si definimos el error de estimación como la mitad de la amplitud del intervalo de confianza, en este caso el error es:

$$e = Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Despejamos n y obtenemos:

$$\boxed{n = \frac{Z^2_{\frac{\alpha}{2}}\hat{p}\hat{q}}{e^2}}$$

En el caso de una población finita, añadimos el factor de corrección y siguiendo el mismo razonamiento obtenemos un error:

$$e = Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\sqrt{\frac{N-n}{N-1}}$$

Despejamos n y obtenemos:

$$\boxed{n = \frac{NZ^2_{\frac{\alpha}{2}}\hat{p}\hat{q}}{e^2(N-1) + Z^2_{\frac{\alpha}{2}}\hat{p}\hat{q}}}$$

En caso de no tener información sobre la proporción muestral, para estimar el valor del tamaño muestral se supone $\hat{p} = 0{,}5$ que maximiza el tamaño de la muestra si los demás elementos que intervienen en la fórmula, nivel de confianza y precisión del IC, están fijos.

Capítulo 5

Contraste de hipótesis

Supongamos que $X_1, X_2, ..., X_n$ es una muestra aleatoria de una distribución cuya función de densidad (o función de probabilidad) es $f(x|\theta)$, donde el parámetro θ es desconocido, pero debe pertenecer a un espacio paramétrico Ω. Supongamos además que podemos descompomer Ω en dos conjuntos disjuntos Ω_0 y Ω_1 tales que $\Omega_0 \cup \Omega_1 = \Omega$.

En todo procedimiento de contraste, se definirán dos hipótesis:

- $H_0 : \theta \in \Omega_0$
- $H_1 : \theta \in \Omega_1$

como los subconjuntos Ω_0 y Ω_1 constituyen una partición de Ω, exactamente una de las hipótesis debe ser cierta. La hipótesis H_0 se llama hipótesis nula, y la hipótesis H_1 se llama hipótesis alternativa.

Dada $X_1, X_2, ..., X_n$ una muestra aleatoria de $f(x|\theta)$, determinaremos el espacio muestral S del vector aleatorio n-dimensional $X = (X_1, X_2, ..., X_n)$, y el problema del contraste de hipótesis será encontrar un estadístico de contraste que nos permita dividir S en dos subconjuntos distintos: uno conteniendo los valores de X para los que no se rechazará H_0 y otro conteniendo el conjunto de valores de X para los que se rechazará H_0. A este último conjunto de valores se le llamará región de rechazo o región crítica del contraste.

Definicion 4 *Si llamamos C a la región crítica de un contraste, se define la función de potencia del contraste, $\pi(\theta)$ como la probabilidad de que el procedimiento concluya con el rechazo de H_0, i.e:*

$$\pi(\theta) = P(X \in C \mid \theta) \qquad \forall \theta \in \Omega$$

Para cualquier valor $\theta \in \Omega_0$, $\pi(\theta)$ es la probabilidad de que el estadístico tome una decisión incorrecta.

Si el cojunto Ω_i $(i = 0, 1)$ sólo puede contener un valor de θ, se dice entonces que la hipótesis H_i es una hipótesis simple. Si el conjunto Ω_i contiene más de una valor de θ diremos que la hipótesis H_i es una hipótesis compuesta.

5.0.1. Errores tipo I y tipo II

Cuando se lleva a cabo un contraste de hipótesis, podemos incurrir en dos tipos de error:

- Un error de tipo I, que se comete al rechazar una hipótesis nula H_0 correcta.
- Un error de tipo II, que se comete cuando no se rechaza la hipótesis nula H_0 siendo esta falsa.

Para cualquier procedimiento de contraste δ, se denotará por $\alpha(\delta)$ o simplemente α a la probabilidad de cometer un error de tipo I, y por $\beta(\delta)$ o simplemente β a la probabilidad de cometer un error de tipo II. Por tanto::

$$\pi(\theta \mid \theta \in \Omega_0) = \alpha$$
$$1 - \pi(\theta \mid \theta \in \Omega_1) = \beta.$$

En muchos problemas, un estadístico especificará una cota superior α_0 para la probabilidad de comenter un error de tipo I y considerará únicamente contrastes para los que $\pi(\theta \mid \theta \in \Omega_0) \leq \alpha_0$. A una cota superior α_0 así definida se la llama nivel de significación del contraste.

Al realizar el contraste de hipótesis se pueden dar las cuatro situaciones siguientes:

	H_0 es cierta	H_1 es cierta
Se escogió H_0	No hay error	Error de tipo II
Se escogió H_1	Error de tipo I	No hay error

Por tanto, es deseable encontrar un procedimiento de contraste δ para el cual las probabilidades de los dos tipos de error $\alpha(\delta)$ y $\beta(\delta)$ sean pequeñas. Es sencillo construir un procedimiento para el cual $\alpha(\delta) = 0$, simplemente aceptando siempre H_0 pero esto implica que $\beta(\delta) = 1$. Análogamente podemos construir un contraste para el que $\alpha(\delta) = 1$y $\beta(\delta) = 0$. A continuación veremos dos procedimientos para construir contrastes de hipótesis para minimizar $\alpha(\delta)$ y $\beta(\delta)$.

5.0.2. Contrastes de hipótesis simples

Supongamos que tenemos un contraste de hipótesis de la forma

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

Para $i = 0, 1$, se define $f_i(x)$ como la f.d.p. (o f.p.) conjunta de las observaciones de la muestra si la hipótesis i es cierta $(i = 0, 1)$.

Veamos procedimientos para construir contrastes de hipótesis para minimizar $\alpha(\delta)$ y $\beta(\delta)$ en este caso.

1. Minimización de una combinación lineal. Supongamos que a y b son constantes positivas específicas y que se desea hallar un procedimiento δ para el que $a\alpha(\delta) + b\beta(\delta)$ sea mínimo.

   Teorema 5.0.1 *Sea δ' un procedimiento de contraste tal que la hipótesis H_0 se acepta si $af_0(x) > bf_1(x)$ y la hipótesis H_1 se acepta si $af_0(x) < bf_1(x)$. Cualquiera de las dos hipótesis H_0 y H_1 puede ser aceptada si $af_0(x) = bf_1(x)$. Entonces, para cualquier otro procedimiento δ,*

   $$a\alpha(\delta') + b\beta(\delta') \leq a\alpha(\delta) + b\beta(\delta)$$

   Demostración

   Demostraremos este resultado para un problema en el que la muestra aleatoria $X_1, ..., X_n$ se seleccione de una distribución discreta. En este caso $f_i(x)$ representa la f.p conjunta de las observaciones de la muestra cuando H_i es cierta $(i = 1, 2)$.

   Si se define R como una región crítica de una procedimiento de contraste arbitrario δ, entonces R contiene los resultado muestrales x para los que δ especifica que H_0 debería ser rechazada y R^c contiene los resultados x para los que H_0 debería ser aceptada. Por tanto,

   $$\begin{aligned} a\alpha(\delta) + b\beta(\delta) &= a\sum_{x\in R} f_0(x) + b\sum_{x\in R^c} f_1(x) \\ &= a\sum_{x\in R} f_0(x) + b\left[1 - \sum_{x\in R} f_1(x)\right] \\ &= b + \sum_{x\in R}[af_0(x) - bf_1(x)] \end{aligned}$$

   De esta ecuación se deduce que el valor de la combinación lineal $a\alpha(\delta) + b\beta(\delta)$ será mínimo si la región crítica R se elige de forma que el valor de la útltima suma de la ecuación sera mínimo. Además, el valor de esta suma será mínimo si la suma incluye todos los puntos x pra los que $af_0(x) - bf_1(x) < 0$ y no incluye los puntos x para los que $af_0(x) - bf_1(x) > 0$, los puntos que verifiquen $af_0(x) - bf_1(x) = 0$ su pertenencia a R es irrelevate puesto que este término contribuye con cero a la suma que queremos minimizar. La decripción de la región crítica corresponde con la descripción del procedimiento de contraste δ' del ennunciado del teorema.

   Si la muestra proviene de una distribución continua,en cuyo caso $f_i(x)$ es una f.p.d conjunta, entonces cada una de las sumas que aparecerán en esta demostración se reemplazaría por la integral n-dimensional.

2. Minimización de la probabilidad de un error de tipo II. Supondremos en este caso que no se permite que la probabilidad $\alpha(\delta)$ es un error del tipo I sea mayor que un determinado nivel de significación y que se desea hallar un procedimiento δ para el cual $\beta(\delta)$ sea mínimo.

Lema 1 *Lema de Neyman-Pearson Supongamos que δ' es un procedimiento de contraste que tiene la siguiente forma para una constante $k > 0$: Se acepta la hipótesis H_0 si $f_0(x) > kf_1(x)$ y se acepta la hipótesis H_1 si $f_0(x) < kf_1(x)$. Cualquiera de las dos hipótesis, H_0 y H_1, puede ser aceptada si $f_0(x) = kf_1(x)$. Si δ es cualquier otro procedimiento de contraste tal que $\alpha(\delta) \leq \alpha(\delta')$, entonces resulta que $\beta(\delta) \geq \beta(\delta')$. Además, si $\alpha(\delta) < \alpha(\delta')$, entonces $\beta(\delta) > \beta(\delta')$.*

Demostración De la definición del procedimiendo δ' y del teorema 2.1, obtenemos que para cualquier otro procedimiento δ,

$$\alpha(\delta') + k\beta(\delta') \leq \alpha(\delta) + k\beta(\delta)$$

Si $\alpha(\delta) \leq \alpha(\delta')$, entonces de la desigualdad anterior resulta que $\beta(\delta) \geq \beta(\delta')$. Además, si $\alpha(\delta) < \alpha(\delta')$, entonces se deduce que $\beta(\delta) > \beta(\delta')$.

5.0.3. Contrastes uniformemente más potentes

Sea el contraste:

$$\begin{cases} H_0 & : & \theta \in \Theta_0 \\ H_1 & : & \theta \in \Theta_1 \end{cases}$$

donde Θ_1 contiene al menos dos valores distintos de θ, y donde la hipótesis nula puede ser simple o compuesta.

Definicion 5 *Un procedimiento de contraste δ' es un contraste uniformemente más potente (UMP) de las hipótesis anteriores al nivel de significación α_0 si $\alpha(\delta') \leq \alpha_0$ y, para cualquier otro procedimiento de contraste δ tal que $\alpha(\delta) \leq \alpha_0$, se verifica que*

$$\pi(\theta \mid \delta) \leq \pi(\theta \mid \delta') \qquad \forall \theta \in \Theta_1$$

donde $\pi(\theta \mid \delta)$ representa la función de potencia de un procedimiento de contraste δ.

5.1. Comparación de las medias de dos distribuciones normales

5.1.1. Deducción del contraste

Sean las variables X_{ij} $i = 1, 2, j = 1, ..., n_i$, j muestras aleatorias de n_i observaciones de dos distribuciones normales independientes con con medias μ_i y varianzas σ^2 desconocidas (la misma varianza para ambas distribuciones aunque desconocida).

Supóngase que queremos contrastar las siguientes hipótesis a un nivel de significación específico α_0 $(0 < \alpha_0 < 1)$

$$H_0 : \mu_1 \leq \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

Para cualquier procedimiento de contraste δ se define $\pi(\mu_1, \mu_2, \sigma^2 \mid \delta)$ como la función de potencia de δ. El objetivo es encontrar un procedimiento de contraste δ tal que:

- $\pi(\mu_1, \mu_2, \sigma^2 \mid \delta) \leq \alpha_0$ si $\mu_1 \leq \mu_2$
- $\pi(\mu_1, \mu_2, \sigma^2 \mid \delta)$ sea lo más grande posible si $\mu_1 > \mu_2$

Puede demostrarse (ver [**?**]), que no existe un contraste UMP para este caso, pero sí que podemos encontrar un procedimiento de contraste δ que verifique:

1. $\pi(\mu_1, \mu_2, \sigma^2 \mid \delta) = \alpha_0$ si $\mu_1 = \mu_2$
2. $\pi(\mu_1, \mu_2, \sigma^2 \mid \delta) < \alpha_0$ si $\mu_1 < \mu_2$
3. $\pi(\mu_1, \mu_2, \sigma^2 \mid \delta) > \alpha_0$ si $\mu_1 > \mu_2$
4. $\pi(\mu_1, \mu_2, \sigma^2 \mid \delta) \longrightarrow 0$ si $\mu_1 - \mu_2 \longrightarrow -\infty$
5. $\pi(\mu_1, \mu_2, \sigma^2 \mid \delta) \longrightarrow 1$ si $\mu_1 - \mu_2 \longrightarrow \infty$

Este procedimiento de contraste (ver [**?**]), define el estadístico

$$U = \frac{(n_1 + n_2 - 2)^{1/2}(\bar{X}_1 - \bar{X}_2)}{(\frac{1}{n_1} + \frac{1}{n_2})^{1/2}(S^2_{X_1} + S^2_{X_2})^{1/2}}$$

y especifica que se debería rechazar la hipótesis nula si $U > t_{n_1+n_2-2,\alpha}$, siendo $t_{n_1+n_2-2,\alpha}$ el percentil $1-\alpha$ de la distribución t de student con $n_1 + n_2 - 2$ grados de libertad.

- Este procedimiento de contraste se puede adaptar facilmente para contrastar las siguientes hipótesis a un nivel de confianza específico α_0

  $$H_0 : \mu_1 = \mu_2$$
  $$H_1 : \mu_1 \neq \mu_2$$

  Puesto que la hipótesis alternativa en este caso es bilateral, se puede probar (ver [**?**]) que el procedimiento de contraste sería definir de nuevo

  $$U = \frac{(n_1 + n_2 - 2)^{1/2}(\bar{X}_1 - \bar{X}_2)}{(\frac{1}{n_1} + \frac{1}{n_2})^{1/2}(S^2_{X_1} + S^2_{X_2})^{1/2}}$$

  y rechazar H_0 si $\mid U \mid > t_{n_1+n_2-2,\alpha/2}$.

- En caso de varianzas poblacionales conocidas e iguales, la adaptación del procedimiento de contraste para contrastar

  $$H_0 : \mu_1 = \mu_2$$
  $$H_1 : \mu_1 \neq \mu_2$$

se resumiría en definir el estadístico de contraste

$$U = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

y rechazar la hipótesis nula si $\mid U \mid > z_{\alpha/2}$, siendo $z_{\alpha/2}$ el percentil $1-\alpha/2$ de la distribución $N(0,1)$.

- En caso de varianzas poblacionales conocidas e iguales, la adaptación del procedimiento de contraste para contrastar

$$H_0 : \mu_1 \leq \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

se resumiría en definir el estadístico de contraste

$$U = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

y rechazar la hipótesis nula si $U > z_\alpha$.

5.1.2. Tamaño muestral. Prueba de no inferioridad / superioridad para la comparación de medias suponiendo varianzas poblacionales conocidas e iguales

El contraste de hipótesis que consideraremos en este apartado es el siguiente:

$$H_0 : \mu_1 \leq \mu_2 \qquad H_1 : \mu_1 > \mu_2$$

Por tanto, trabajaremos con el estadístico de contraste

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Las regiones de aceptación y crítica de este contraste son:

$$A = \{Z : Z \leq z_\alpha\} \quad C = \{Z : Z > z_\alpha\}$$

Rechazamos la hipótesis nula con un nivel de significación α si:

$$\frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > z_\alpha$$

Si queremos lograr una potencia de β,

$$P\left(\frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > z_\alpha \mid H_1\right) = 1 - \beta$$

Si H_1 es cierta, $\mu_1 > \mu_2$, por tanto

$$\begin{aligned} \bar{X}_i &\sim N(\mu_i, \sigma^2/n_i)\ i = 1, 2 \\ \bar{X}_1 - \bar{X}_2 &\sim N(\mu_1 - \mu_2, \sigma^2/n_1 + \sigma^2/n_2) \\ \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} &\sim N(0, 1) \end{aligned}$$

$$P\left(\frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > z_\alpha\right) = P\left(\frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} - \frac{(\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > z_\alpha - \frac{(\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right) = 1 - \beta$$

De donde podemos deducir que

$$z_\alpha - \frac{(\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = -z_\beta$$

Para poder obtener los tamaños muestrales:

- Aproximamos μ_1 y μ_2 por $\bar{X}_1$ y $\bar{X}_2$ respectivamente
- Suponemos que existe una relación de proporcionalidad entre los dos tamaños muestrales:

$$\boxed{\mathbf{n_1 = kn_2}}$$

- y obtenemos que podemos aproximar

$$\boxed{\mathbf{n_2 = \frac{(z_\alpha + z_\beta)^2\sigma^2(1 + 1/k)}{(\bar{X}_1 - \bar{X}_2)^2}}}$$

5.1.3. Tamaño muestral. Prueba de no inferioridad / superioridad para la comparación de medias suponiendo varianzas poblacionales desconocidas e iguales

Consideramos el siguiente contraste de hipótesis:

$$H_0 : \mu_1 \leq \mu_2 \qquad H_1 : \mu_1 > \mu_2$$

Como hemos comentado en el apartado anterior, el procedimiento de contraste de basa en el estadístico

$$T = (\bar{X}_1 - \bar{X}_2)\sqrt{\frac{n_1 + n_2 - 2}{(\frac{1}{n_1} + \frac{1}{n_2})(S^2_{X_1} + S^2_{X_2})}}$$

Las regiones de aceptación y crítica de este contraste son:

$$A = \{T : T \leq t_{n_1+n_2-2,\alpha}\} \quad C = \{T : T > t_{n_1+n_2-2,\alpha}\}$$

Rechazamos la hipótesis nula con un nivel de significación α si:

$$(\bar{X}_1 - \bar{X}_2)\sqrt{\frac{n_1+n_2-2}{(\frac{1}{n_1}+\frac{1}{n_2})(S^2_{X_1}+S^2_{X_2})}} > t_{n_1+n_2-2,\alpha}$$

Si exigimos una potencia de β, tendremos que

$$P\left((\bar{X}_1 - \bar{X}_2)\sqrt{\frac{n_1+n_2-2}{(\frac{1}{n_1}+\frac{1}{n_2})(S^2_{X_1}+S^2_{X_2})}} > t_{n_1+n_2-2,\alpha} \mid H_1\right) = 1-\beta$$

Si H_1 es cierta, $\mu_1 > \mu_2$, por tanto

$$\begin{aligned}
\bar{X}_i &\sim N\left(\mu_i, \frac{\sigma^2}{n_i}\right),\ i=1,2 \\
\bar{X}_1 - \bar{X}_2 &\sim N(\mu_1-\mu_2, \sigma^2(\frac{1}{n_1}+\frac{1}{n_2})) \\
\frac{(n_i-1)s^2_{X_i}}{\sigma^2} &\sim \chi^2_{n_i-1},\ i=1,2 \\
\frac{(n_1-1)s^2_{X_1}}{\sigma^2} + \frac{(n_2-1)s^2_{X_2}}{\sigma^2} &\sim \chi^2_{n_1+n_2-2} \\
\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s^2_{X_1}+s^2_{X_1}}{n_1+n_2-2}(\frac{1}{n_1}+\frac{1}{n_2})}} &\sim t_{n_1+n_2-2}
\end{aligned}$$

El tamaño muestral necesario para lograr una potencia de β viene dado por la siguiente ecuación:

$$\begin{aligned}
P\left((\bar{X}_1 - \bar{X}_2)\sqrt{\frac{n_1+n_2-2}{(\frac{1}{n_1}+\frac{1}{n_2})(S^2_{X_1}+S^2_{X_2})}} > t_{n_1+n_2-2,\alpha} \mid H_1\right) &= 1-\beta \\
1 - P\left((\bar{X}_1 - \bar{X}_2)\sqrt{\frac{n_1+n_2-2}{(\frac{1}{n_1}+\frac{1}{n_2})(S^2_{X_1}+S^2_{X_2})}} < t_{n_1+n_2-2,\alpha} \mid H_1\right) &= 1-\beta \\
P\left((\bar{X}_1 - \bar{X}_2)\sqrt{\frac{n_1+n_2-2}{(\frac{1}{n_1}+\frac{1}{n_2})(S^2_{X_1}+S^2_{X_2})}} < t_{n_1+n_2-2,\alpha} \mid H_1\right) &= \beta
\end{aligned}$$

Por tanto bajo H_1:

$$P\left((\bar{X}_1-\bar{X}_2)\sqrt{\frac{n_1+n_2-2}{(\frac{1}{n_1}+\frac{1}{n_2})(S^2_{X_1}+S^2_{X_2})}}<t_{n_1+n_2-2,\alpha}\right)=$$

$$=P\left((\bar{X}_1-\bar{X}_2)\sqrt{\frac{n_1+n_2-2}{(\frac{1}{n_1}+\frac{1}{n_2})(S^2_{X_1}+S^2_{X_2})}}-(\mu_1-\mu_2)\sqrt{\frac{n_1+n_2-2}{(\frac{1}{n_1}+\frac{1}{n_2})(S^2_{X_1}+S^2_{X_2})}}<\right.$$

$$\left.<t_{n_1+n_2-2,\alpha}-(\mu_1-\mu_2)\sqrt{\frac{n_1+n_2-2}{(\frac{1}{n_1}+\frac{1}{n_2})(S^2_{X_1}+S^2_{X_2})}}\right)=\beta$$

$$t_{n_1+n_2-2,\alpha}-(\mu_1-\mu_2)\sqrt{\frac{n_1+n_2-2}{(\frac{1}{n_1}+\frac{1}{n_2})(S^2_{X_1}+S^2_{X_2})}}=t_{n_1+n_2-2,\beta}$$

Para poder obtener el tamaño muestral, supondremos una relación de proporcionalidad entre los dos tamaños muestrales

$$\boxed{\mathbf{n_1=kn_2}}$$

$$\boxed{t_{n_2(k+1)-2,\alpha}-(\mu_1-\mu_2)\sqrt{\frac{n_2(1+k)-2}{(\frac{1}{kn_2}+\frac{1}{n_2})(S^2_{X_1}+S^2_{X_2})}}=t_{n_2(k+1)-2,\beta}}$$

En este caso una vez tomada esta relación de proporcionalidad, únicamente podemos calcular una aproximación del tamaño muestral puesto que n_2 forma parte de los grados de libertad de la distribución.

Para un tamaño muestral suficientemente grande, podemos aproximar la distribución t-Student mediante la distribución normal de este modo:

Si $n_2(k+1)>30$, podemos aproximar los percentiles de la distribución t por los percentiles de la distribución normal y escribir:

$$z_\alpha-(\mu_1-\mu_2)\sqrt{\frac{n_2(1+k)-2}{(\frac{1}{kn_2}+\frac{1}{n_2})(S^2_{X_1}+S^2_{X_2})}}=z_\beta$$

Por tanto el tamaño muestral viene dado por la fórmula:

$$z_\alpha-z_\beta=(\mu_1-\mu_2)\sqrt{\frac{n_2(1+k)-2}{(\frac{1}{kn_2}+\frac{1}{n_2})(S^2_{X_1}+S^2_{X_2})}}$$

$$\frac{(z_\alpha-z_\beta)^2(1+1/k)(S^2_{X_1}+S^2_{X_2})^2}{(\mu_1-\mu_2)^2}=n_2^2(1+k)-2n_2$$

$$\boxed{n_2=\frac{2\pm\sqrt{4-4\frac{(z_\alpha-z_\beta)^2(1+1/k)^2(S^2_{X_1}+S^2_{X_2})^2}{(\mu_1-\mu_2)^2}}}{2(1+k)}}$$

5.1.4. Tamaño muestral. Prueba de igualdad para la comparación de medias suponiendo varianzas poblacionales conocidas e iguales

El contraste de hipótesis que consideraremos en este apartado es el siguiente:

$$H_0 : \mu_2 - \mu_1 = 0 \qquad H_1 : \mu_2 - \mu_1 \neq 0$$

Por tanto trabajaremos con el estadístico de contraste

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Las regiones de aceptación y crítica de este contraste son:

$$A = \{Z : |Z| \leq z_{\frac{\alpha}{2}}\} \quad C = \{Z : |Z| > z_{\frac{\alpha}{2}}\}$$

Rechazamos la hipótesis nula con un nivel de significación α si:

$$\left| \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| > z_{\alpha/2}$$

Si queremos lograr una potencia de β,

$$P\left(\frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < -z_{\frac{\alpha}{2}} \mid H_1 \right) + P\left(\frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > z_{\alpha/2} \mid H_1 \right) = 1 - \beta$$

Si H_1 es cierta, $\mu_1 \neq \mu_2$, por tanto

$$\begin{aligned}
\bar{X}_1 &\sim N(\mu_1, \sigma^2/n_1) \\
\bar{X}_2 &\sim N(\mu_2, \sigma^2/n_2) \\
\bar{X}_1 - \bar{X}_2 &\sim N(\mu_1 - \mu_2, \sigma^2/n_1 + \sigma^2/n_2) \\
\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} &\sim N(0, 1)
\end{aligned}$$

$$P\left(\frac{\bar{X}_1-\bar{X}_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}-\frac{\mu_1-\mu_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}<-z_{\frac{\alpha}{2}}-\frac{\mu_1-\mu_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\mid H_1\right)+$$
$$+P\left(\frac{\bar{X}_1-\bar{X}_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}-\frac{\mu_1-\mu_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}>z_{\alpha/2}-\frac{\mu_1-\mu_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\mid H_1\right)=$$
$$P\left(\frac{\bar{X}_1-\bar{X}_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}-\frac{\mu_1-\mu_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}<-z_{\frac{\alpha}{2}}-\frac{\mu_1-\mu_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\mid H_1\right)+$$
$$+P\left(\frac{\bar{X}_1-\bar{X}_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}-\frac{\mu_1-\mu_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}<-z_{\alpha/2}+\frac{\mu_1-\mu_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\mid H_1\right)$$

Bajo la hipótesis alternativa ($H_1 : \mu_1 \neq \mu_2$) podemos distinguir dos casos:

- Si $\mu_1 > \mu_2$: $\mu_1 - \mu_2 = \mid \mu_1 - \mu_2 \mid$
- Si $\mu_1 < \mu_2$: $\mu_1 - \mu_2 = - \mid \mu_1 - \mu_2 \mid$

Pero en ambas situaciones:

$$P\left(\frac{\bar{X}_1-\bar{X}_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}-\frac{\mu_1-\mu_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}<-z_{\frac{\alpha}{2}}-\frac{\mu_1-\mu_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right)+$$
$$+P\left(\frac{\bar{X}_1-\bar{X}_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}-\frac{\mu_1-\mu_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}<-z_{\alpha/2}+\frac{\mu_1-\mu_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right)=$$
$$P\left(\frac{\bar{X}_1-\bar{X}_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}-\frac{\mu_1-\mu_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}<-z_{\frac{\alpha}{2}}-\frac{|\mu_1-\mu_2|}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right)+$$
$$+P\left(\frac{\bar{X}_1-\bar{X}_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}-\frac{\mu_1-\mu_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}<-z_{\alpha/2}+\frac{|\mu_1-\mu_2|}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right)$$

Dada $Z \sim N(0,1)$, como

$$-z_{\alpha/2}-\frac{|\mu_1-\mu_2|}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}<-z_{\alpha/2}$$

Se cumplirá que

$$P(Z<-z_{\alpha/2}-\frac{|\mu_1-\mu_2|}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}})<\alpha/2$$

Si consideramos este valor lo suficientemente pequeño como para desestimarlo, de los dos sumandos que tenemos, nos quedaremos únicamente con el segundo, y nos bastará con exigir que:

$$P\left(\frac{\bar{X}_1-\bar{X}_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}-\frac{\mu_1-\mu_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}<-z_{\alpha/2}+\frac{|\mu_1-\mu_2|}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right)=\beta$$

De donde podemos deducir que

$$-z_{\alpha/2}+\frac{|\mu_1-\mu_2|}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}=z_\beta$$

Para poder obtener los tamaños muestrales:

- Aproximamos μ_1 y μ_2 por $\bar{X}_1$ y $\bar{X}_2$ respectivamente
- Suponemos que existe una relación de proporcionalidad entre los dos tamaños muestrales:

$$\boxed{\mathbf{n_1=kn_2}}$$

- y obtenemos que podemos aproximar

$$\boxed{\mathbf{n_2}=\frac{(\mathbf{z_{\alpha/2}}+\mathbf{z}_\beta)^{\mathbf{2}}\boldsymbol{\sigma}^{\mathbf{2}}(\mathbf{1}+\mathbf{1/k})}{(\mathbf{\bar{X}_1}-\mathbf{\bar{X}_2})^{\mathbf{2}}}}$$

5.1.5. Tamaño muestral. Prueba de igualdad para la comparación de medias suponiendo varianzas poblacionales desconocidas e iguales

El contraste de hipótesis que consideraremos en este apartado es el siguiente:

$$H_0:\mu_2-\mu_1=0 \qquad H_1:\mu_2-\mu_1\neq 0$$

Como hemos comentado en el apartado anterior, el procedimiento de contraste de basa en el estadístico

$$T=(\bar{X}_1-\bar{X}_2)\sqrt{\frac{n_1+n_2-2}{(\frac{1}{n_1}+\frac{1}{n_2})(S^2_{X_1}+S^2_{X_2})}}$$

Las regiones de aceptación y crítica de este contraste son:

$$A=\{T:|T|\leq t_{\frac{\alpha}{2},n_1+n_2-2}\}\quad C=\{T:|T|>t_{\frac{\alpha}{2},n_1+n_2-2}\}$$

Rechazamos la hipótesis nula con un nivel de significación α si:

$$\left|(\bar{X}_1-\bar{X}_2)\sqrt{\frac{n_1+n_2-2}{(\frac{1}{n_1}+\frac{1}{n_2})(S^2_{X_1}+S^2_{X_2})}}\right| > t_{\frac{\alpha}{2},n_1+n_2-2}$$

Si exigimos una potencia de β, tendremos que

$$P\left((\bar{X}_1-\bar{X}_2)\sqrt{\frac{n_1+n_2-2}{(\frac{1}{n_1}+\frac{1}{n_2})(S^2_{X_1}+S^2_{X_2})}} < -t_{\frac{\alpha}{2},n_1+n_2-2} \mid H_1\right) +$$
$$P\left((\bar{X}_1-\bar{X}_2)\sqrt{\frac{n_1+n_2-2}{(\frac{1}{n_1}+\frac{1}{n_2})(S^2_{X_1}+S^2_{X_2})}} > t_{\frac{\alpha}{2},n_1+n_2-2} \mid H_1\right) = 1-\beta$$

Si H_1 es cierta, $\mu_1 > \mu_2$, por tanto

$$\begin{aligned}
\bar{X}_1 &\sim N\left(\mu_1, \frac{\sigma^2}{n_1}\right)\\
\bar{X}_2 &\sim N\left(\mu_2, \frac{\sigma^2}{n_2}\right)\\
\bar{X}_1-\bar{X}_2 &\sim \left(\mu_1-\mu_2, \sigma^2(\frac{1}{n_1}+\frac{1}{n_2})\right)\\
\frac{(n_1-1)s^2_{X_1}}{\sigma^2} &\sim \chi^2_{n_1-1}\\
\frac{(n_2-1)s^2_{X_2}}{\sigma^2} &\sim \chi^2_{n_2-1}\\
\frac{(n_1-1)s^2_{X_1}}{\sigma^2}+\frac{(n_2-1)s^2_{X_2}}{\sigma^2} &\sim \chi^2_{n_1+n_2-2}\\
\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s^2_{X_1}+s^2_{X_1}}{n_1+n_2-2}(\frac{1}{n_1}+\frac{1}{n_2})}} &\sim t_{n_1+n_2-2}
\end{aligned}$$

En este apartado utilizaremos el mismo razonamiento utilizado en el apartado 5.1.4 y por tanto tras desestimar un término con valor $< \alpha/2$, obtenemos:

$$P\left((\bar{X}_1-\bar{X}_2)\sqrt{\frac{n_1+n_2-2}{(\frac{1}{n_1}+\frac{1}{n_2})(S^2_{X_1}+S^2_{X_2})}} - |(\mu_1-\mu_2)|\sqrt{\frac{n_1+n_2-2}{(\frac{1}{n_1}+\frac{1}{n_2})(S^2_{X_1}+S^2_{X_2})}} <\right.$$
$$\left. < -t_{\frac{\alpha}{2},n_1+n_2-2} + |(\mu_1-\mu_2)|\sqrt{\frac{n_1+n_2-2}{(\frac{1}{n_1}+\frac{1}{n_2})(S^2_{X_1}+S^2_{X_2})}} \mid H_1\right) = 1-\beta$$

Por tanto:

$$-t_{\frac{\alpha}{2},n_1+n_2-2} + |(\mu_1-\mu_2)|\sqrt{\frac{n_1+n_2-2}{(\frac{1}{n_1}+\frac{1}{n_2})(S^2_{X_1}+S^2_{X_2})}} = -t_{\beta,n_1+n_2-2}$$

Para poder obtener el tamaño muestral:

- Aproximamos μ_1 y μ_2 por $\bar{X}_1$ y $\bar{X}_2$ respectivamente.
- supondremos una relación de proporcionalidad entre los dos tamaños muestrales

$$\boxed{\mathbf{n_1 = kn_2}}$$

Y obtenemos que podemos aproximar:

$$t_{\frac{\alpha}{2},n_2(1+k)-2} - t_{\beta,n_2(1+k)-2} = |(\mu_1 - \mu_2)|\sqrt{\frac{n_2(1+k)-2}{(\frac{1}{kn_2}+\frac{1}{n_2})(S^2_{X_1}+S^2_{X_2})}}$$

Para un tamaño muestral suficientemente grande, podemos aproximar los percentiles de la distribución t-Student por los percentiles de la distribución normal y de este modo calcular el tamaño muestral del modo siguiente:

$$z_{\frac{\alpha}{2}} - z_\beta = |(\mu_1 - \mu_2)|\sqrt{\frac{n_2(1+k)-2}{(\frac{1}{kn_2}+\frac{1}{n_2})(S^2_{X_1}+S^2_{X_2})}}$$

$$\frac{(z_{\frac{\alpha}{2}} - z_\beta)^2(1+1/k)(S^2_{X_1}+S^2_{X_2})^2}{(\mu_1-\mu_2)^2} = n_2^2(1+k) - 2n_2$$

$$\boxed{n_2 = \frac{2 \pm \sqrt{4 - 4\frac{(z_{\frac{\alpha}{2}} - z_\beta)^2(1+1/k)^2(S^2_{X_1}+S^2_{X_2})^2}{(\mu_1-\mu_2)^2}}}{2(1+k)}}$$

5.2. Comparación de medias de dos distribuciones normales asumiendo varianzas distintas

5.2.1. Deducción del contraste

Sean las variables X_{ij} $i = 1, 2, j = 1, ..., n_i$, muestras aleatorias de n_i observaciones de dos distribuciones normales independientes con medias μ_i y varianzas σ_i^2. Supóngase, además, que estos valores de μ_1, μ_2, σ_1^2 y σ_2^2 son desconocidos pero que $\sigma_2^2 = c\sigma_1^2$, donde c es una constante positiva conocida.

Supóngase que queremos contrastar las siguientes hipótesis a un nivel de significación específico α_0 $(0 < \alpha_0 < 1)$

$$H_0 : \mu_1 \leq \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

El objetivo es encontrar un procedimiento de contraste δ tal que:

- $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \mid \delta) \leq \alpha_0$ si $\mu_1 \leq \mu_2$
- $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \mid \delta)$ sea lo más grande posible si $\mu_1 > \mu_2$

Puede demostrarse (ver [**?**]), que no existe un contraste UMP para este caso, pero sí que podemos encontrar un procedimiento de contraste δ que verifique:

1. $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \mid \delta) = \alpha_0$ si $\mu_1 = \mu_2$
2. $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \mid \delta) < \alpha_0$ si $\mu_1 < \mu_2$
3. $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \mid \delta) > \alpha_0$ si $\mu_1 > \mu_2$
4. $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \mid \delta) \longrightarrow 0$ si $\mu_1 - \mu_2 \longrightarrow -\infty$
5. $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \mid \delta) \longrightarrow 1$ si $\mu_1 - \mu_2 \longrightarrow \infty$

Este procedimiento de contraste (ver [**?**]), define el estadístico

$$U = \frac{(n_1 + n_2 - 2)^{1/2}(\bar{X}_1 - \bar{X}_2)}{(\frac{1}{n_1} + \frac{c}{n_2})^{1/2}(S^2_{X_1} + \frac{S^2_{X_2}}{c})^{1/2}}$$

y especifica que se debería rechazar la hipótesis nula si $U > t_{n_1+n_2-2,\alpha}$, siendo $t_{n_1+n_2-2,\alpha}$ el percentil $1-\alpha$ de la distribución t de student con $n_1 + n_2 - 2$ grados de libertad.

- Este procedimiento de contraste se puede adaptar facilmente para contrastar las siguientes hipótesis a un nivel de confianza específico α_0

  $$H_0 : \mu_1 = \mu_2$$
  $$H_1 : \mu_1 \neq \mu_2$$

  Puesto que la hipótesis alternativa en este caso es bilateral, se puede probar (ver [**?**]) que el procedimiento de contraste sería definir de nuevo

  $$U = \frac{(n_1 + n_2 - 2)^{1/2}(\bar{X}_1 - \bar{X}_2)}{(\frac{1}{n_1} + \frac{c}{n_2})^{1/2}(S^2_{X_1} + \frac{S^2_{X_2}}{c})^{1/2}}$$

  y rechazar H_0 si $\mid U \mid > t_{n_1+n_2-2,\alpha/2}$.

- En caso de varianzas poblacionales conocidas e iguales, la adaptación del procedimiento de contraste para contrastar

  $$H_0 : \mu_1 = \mu_2$$
  $$H_1 : \mu_1 \neq \mu_2$$

  se resumiría en definir el estadístico de contraste

  $$U = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

  y rechazar la hipótesis nula si $\mid U \mid > z_{\alpha/2}$, siendo $z_{\alpha/2}$ el percentil $1 - \alpha/2$ de la distribución $N(0, 1)$.

- En caso de varianzas poblacionales conocidas y distintas, la adaptación del procedimiento de contraste para contrastar

$$H_0 : \mu_1 \leq \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

se resumiría en definir el estadístico de contraste

$$U = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

y rechazar la hipótesis nula si $U > z_\alpha$.

5.2.2. Tamaño muestral. Prueba de no inferioridad / superioridad para la comparación de medias suponiendo varianzas poblacionales conocidas y distintas

El contraste de hipótesis que consideraremos en este apartado es el siguiente:

$$H_0 : \mu_1 \leq \mu_2 \qquad H_1 : \mu_1 > \mu_2$$

Por tanto, trabajaremos con el estadístico de contraste

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Las regiones de aceptación y crítica de este contraste son:

$$A = \{Z : Z \leq z_\alpha\} \quad C = \{Z : Z > z_\alpha\}$$

Rechazamos la hipótesis nula con un nivel de significación α si:

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_\alpha$$

Si queremos lograr una potencia de β,

$$P\left(\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_\alpha \mid H_1\right) = 1 - \beta$$

Si H_1 es cierta, $\mu_1 > \mu_2$, por tanto

$$\begin{aligned}
\bar{X}_1 &\sim N(\mu_1, \sigma_1^2/n_1) \\
\bar{X}_2 &\sim N(\mu_2, \sigma_2^2/n_2) \\
\bar{X}_1 - \bar{X}_2 &\sim N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2) \\
\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} &\sim N(0, 1)
\end{aligned}$$

$$P\left(\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}>z_\alpha\right)=P\left(\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}-\frac{(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}>z_\alpha-\frac{(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right)=1-\beta$$

De donde podemos deducir que

$$z_\alpha-\frac{(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}=-z_\beta$$

Para poder obtener los tamaños muestrales:

- Aproximamos μ_1 y μ_2 por $\bar{X}_1$ y $\bar{X}_2$ respectivamente
- Suponemos que existe una relación de proporcionalidad entre los dos tamaños muestrales:
$$\mathbf{n_1=kn_2}$$
- Suponemos que existe una relación de proporcionalidad entre las dos varianzas poblacionales:
$$\boldsymbol{\sigma_1^2=(1/c)\sigma_2^2}$$
- y obtenemos que podemos aproximar

$$\boxed{\mathbf{n_2}=\frac{(\mathbf{z}_\alpha+\mathbf{z}_\beta)^\mathbf{2}\boldsymbol{\sigma}_\mathbf{2}^\mathbf{2}(\frac{\mathbf{1}}{\mathbf{c}}+\mathbf{k})}{(\bar{\mathbf{X}}_\mathbf{1}-\bar{\mathbf{X}}_\mathbf{2})^\mathbf{2}}}$$

5.2.3. Tamaño muestral. Prueba de no inferioridad / superioridad para la comparación de medias suponiendo varianzas poblacionales desconocidas y distintas

El contraste de hipótesis que consideraremos en este apartado es el siguiente:

$$H_0:\mu_1\leq\mu_2 \qquad\qquad H_1:\mu_1>\mu_2$$

Por tanto, el estadístico con el que trabajaremos es

$$T=\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

Las regiones de aceptación y crítica de este contraste son:

$$A=\{T:T\leq t_\alpha\}\quad C=\{T:T>t_\alpha\}$$

Debido a que hemos supuesto que las varianzas son distintas utilizaremos la prueba de Welch-Satterthwaite, basada en la estadístico:

$$T=\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{S_{X_1}^2}{n_1}+\frac{S_{X_2}^2}{n_2}}}$$

La diferencia entre las varianzas dificulta en gran medida el cálculo de la función de distribución de T . Sin embargo, ya Welch ([**?**]), Welch ([**?**]) y Satterthwaite ([**?**]) han ofrecido aproximaciones que se consideran satisfactorias para el uso práctico.

Satterthwaite ([**?**]) define un estimador complejo de varianza como una combinación lineal de cuadrados medios independientes. Welch ([**?**]) ya había demostrado antes que la distribución de este tipo de estimadores puede aproximarse con la distribución χ^2. Específicamente, si MS_i son cuadrados medios independientes con r_i grados de libertad, $i = 1, 2, ..., k$, y si $\hat{V_s} = \sum_{i=1}^{k} \frac{1}{n_i} MS_i$ es un estimador complejo de varianza basado en ellos, los grados de libertad de la aproximación χ^2 son

$$r_s = \frac{\left(\sum_{i=1}^{k} \frac{1}{n_i} E(MS_i)\right)^2}{\sum_{i=1}^{k} \frac{(\frac{1}{n_i} E(MS_i))^2}{r_i}}$$

Los $E(MS_i)$ son desconocidos pero Satterthwaite ([**?**]) verifica, para varios casos, que se pueden reemplazar por los cuadrados medios sin generar mayores inconvenientes en la aproximación a la distribución χ^2 con grados de libertad dados por:

$$\hat{r_s} = \frac{\left(\sum_{i=1}^{k} \frac{1}{n_i} MS_i\right)^2}{\sum_{i=1}^{k} \frac{(\frac{1}{n_i} MS_i)^2}{r_i}}$$

Para el caso de dos muestras independientes, la diferencia de medias, $\mu_1 - \mu_2$, se estima por $\bar{X}_1 - \bar{X}_2$. Su varianza, $\frac{\sigma_{X_1}^2}{n_1} + \frac{\sigma_{X_2}^2}{n_2}$, se estima por $\frac{S_{X_1}^2}{n_1} + \frac{S_{X_2}^2}{n_2}$. Este es el estimador complejo de varianza con $k = 2$. Para la primera muestra,

$$MS_1 = MS_{X_1} = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (X_{1i} - \bar{X}_1)^2,$$
$$r_1 = n_1 - 1 \text{ y } E(MS_{X_1}) = \sigma_{X_1}^2.$$

Para la segunda muestra,

$$MS_2 = MS_{X_2} = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (X_{2i} - \bar{X}_2)^2,$$
$$r_2 = n_2 - 1 \text{ y } E(MS_{X_2}) = \sigma_{X_2}^2.$$

Por tanto:

$$\hat{r_s} = \frac{\left(\frac{S_{X_1}^2}{n_1} + \frac{S_{X_2}^2}{n_2}\right)^2}{\frac{(\frac{S_{X_1}^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{S_{X_2}^2}{n_2})^2}{n_2 - 1}}$$

Por tanto, tenemos que:

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S^2_{X_1}}{n_1} + \frac{S^2_{X_2}}{n_2}}} \sim t_{\hat{r_s}}$$

El tamaño muestral necesario para obtener una potencia β viene dado por la ecuación:

$$P\left(\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} - \frac{(\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} > t_{\hat{r_s},\alpha} - \frac{(\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}\right) = 1 - \beta$$

Para calcular el tamaño muestral, supondremos una relación de proporcionalidad entre los dos tamaños muestrales:

$$\mathbf{n_1 = kn_2}$$

En este caso una vez tomada esta relación de proporcionalidad, únicamente podemos calcular una aproximación del tamaño muestral puesto que n_1 forma parte de los grados de libertad de la distribución.

5.2.4. Tamaño muestral. Prueba de igualdad para la comparación de medias suponiendo varianzas poblacionales conocidas y distintas

El contraste de hipótesis que consideraremos en este apartado es el siguiente:

$$H_0 : \mu_2 - \mu_1 = 0 \qquad H_1 : \mu_2 - \mu_1 \neq 0$$

Por tanto, trabajaremos con el estadístico de contraste

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Las regiones de aceptación y crítica de este contraste son:

$$A = \{Z : |Z| \leq z_{\frac{\alpha}{2}}\} \quad C = \{Z : |Z| > z_{\frac{\alpha}{2}}\}$$

Rechazamos la hipótesis nula con un nivel de significación α si:

$$\left|\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right| > Z_{\alpha/2}$$

Si queremos lograr una potencia β,

$$P\left(\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < -z_{\frac{\alpha}{2}} \mid H_1\right) + P\left(\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_{\alpha/2} \mid H_1\right) = 1 - \beta$$

Si H_1 es cierta, $\mu_1 \neq \mu_2$, por tanto

$$\begin{aligned} \bar{X}_1 &\sim N(\mu_1, \sigma_1^2/n_1) \\ \bar{X}_2 &\sim N(\mu_2, \sigma_2^2/n_2) \\ \bar{X}_1 - \bar{X}_2 &\sim N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2) \\ \frac{|\bar{X}_1 - \bar{X}_2| - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} &\sim N(0,1) \end{aligned}$$

$$P\left(\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} - \frac{(\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < -z_{\frac{\alpha}{2}} - \frac{(\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \mid H_1\right) +$$
$$+P\left(\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} - \frac{(\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_{\alpha/2} - \frac{(\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \mid H_1\right) = 1 - \beta$$

Utilizando el mismo razonamiento que en el apartado 5.1.4, podemos desestimar una factor cuyo valor es $< \alpha/2$ y obtenemos:

$$P\left(\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} - \frac{(\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < -z_{\alpha/2} + \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \mid H_1\right) = \beta$$

De donde podemos deducir que

$$-z_{\alpha/2} + \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = z_\beta$$

Para poder obtener los tamaños muestrales:

- Aproximamos μ_1 y μ_2 por $\bar{X}_1$ y $\bar{X}_2$ respectivamente
- Suponemos que existe una relación de proporcionalidad entre los dos tamaños muestrales:
$$\mathbf{n_1 = kn_2}$$
- Suponemos que existe una relación de proporcionalidad entre las dos varianzas poblacionales:
$$\boldsymbol{\sigma_1^2 = (1/c)\sigma_2^2}$$
- y obtenemos que podemos aproximar

$$\boxed{\mathbf{n_2} = \frac{(\mathbf{z}_{\frac{\alpha}{2}} + \mathbf{z}_\beta)^2(\frac{1}{\mathbf{c}} + \mathbf{k})\boldsymbol{\sigma_2^2}}{(\bar{\mathbf{X}}_\mathbf{1} - \bar{\mathbf{X}}_\mathbf{2})^2}}$$

5.2.5. Tamaño muestral. Prueba de igualdad para la comparación de medias suponiendo varianzas poblacionales desconocidas y distintas

El contraste de hipótesis que consideraremos en este apartado es el siguiente:

$$H_0 : \mu_2 - \mu_1 = 0 \qquad H_1 : \mu_2 - \mu_1 \neq 0$$

Por tanto, el estadístico:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Las regiones de aceptación y crítica de este contraste son:

$$A = \{T : |T| \leq t_{\frac{\alpha}{2}}\} \quad C = \{T : |T| > t_{\frac{\alpha}{2}}\}$$

En este apartado utilizaremos el razonamiento demostrado anteriormente en el apartado 5.2.3 para los grados de libertad de la distribución t.

Por tanto tenemos:

$$\left| \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \right| > t_{\frac{\alpha}{2},\hat{r}_s}$$

Si queremos obtener una potencia β,

$$P\left(\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} > t_{\frac{\alpha}{2},\hat{r}_s} \right) + P\left(\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} < -t_{\frac{\alpha}{2},\hat{r}_s} \right)$$

Si H_1 es cierta, $\mu_1 \neq \mu_2$, por tanto

$$\begin{aligned}
\bar{X}_1 &\sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \\
\bar{X}_2 &\sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right) \\
\bar{X}_1 - \bar{X}_2 &\sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \\
\frac{(n_1 - 1)s_{X_1}^2}{\sigma_1^2} &\sim \chi^2_{n_1 - 1} \\
\frac{(n_2 - 1)s_{X_2}^2}{\sigma_2^2} &\sim \chi^2_{n_2 - 1} \\
\frac{(n_1 - 1)s_{X_1}^2}{\sigma_1^2} + \frac{(n_2 - 1)s_{X_2}^2}{\sigma_2^2} &\sim \chi^2_{n_1 + n_2 - 2} \\
\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_{X_1}^2 + s_{X_1}^2}{n_1 + n_2 - 2}(\frac{1}{n_1} + \frac{1}{n_2})}} &\sim t_{n_1 + n_2 - 2}
\end{aligned}$$

$$P\left(\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}-\frac{\mu_1-\mu_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}>t_{\frac{\alpha}{2},,\hat{r}_s}-\frac{\mu_1-\mu_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}\right)+$$
$$+P\left(\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}-\frac{\mu_1-\mu_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}<-t_{\frac{\alpha}{2},\hat{r}_s}-\frac{\mu_1-\mu_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}\right)$$

Utilizando el mismo razonamiento que en apartado 5.1.4, podemos desestimar un término $< \alpha/2$ y de este modo obtenemos:

$$P\left(\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}-\frac{\mu_1-\mu_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}<-t_{\frac{\alpha}{2},\hat{r}_s}+\frac{|\mu_1-\mu_2|}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}\right)=1-\beta$$

De donde podemos deducir que:

$$-t_{\frac{\alpha}{2},\hat{r}_s}+\frac{|\mu_1-\mu_2|}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}=-t_{\beta,\hat{r}_s}$$

Como podemos observar el tamaño muestral que se desea aproximar es uno de los elementos que aparecen en los grados de libertad de la distribución por lo que únicamente podemos obtener una aproximación y en el caso $\hat{r}_s > 30$ aproximar los percentiles de la distribución t por los percentiles de una distribución normal para de este modo poder realizar el cálculo.

5.3. Tamaño muestral para la comparación de dos medias apareadas

Sea x_{ji} la respuesta observada de el sujeto i-ésimo, $i = 1, ..., n$ en el instante (o ante la variable) j, $j = 1, 2$. Este caso se utiliza a menudo para la comparación de los resultados previos y posteriores a la realización de un tratamiento, y en lugar de trabajar con las variables originales, se trabaja con la variable $d_i = x_{1i} - x_{2i}$, $\forall i = 1, ..., n$. Podemos asumir que d_i son variables independientes e idénticamente distribuidas con distribución normal.

En todo este apartado trabajaremos con la variable aleatoria $D = X_1 - X_2 \sim N(\mu, \sigma^2)$

5.3.1. Deducción del contraste

Supongamos que las variables $X_1, ..., X_n$ constituyen una muestra aleatoria de una distribución normal con media μ y varianza σ^2 desconocidas. Supongamos que se desean contrastar las siguientes hipótesis con un nivel de significación $\alpha_0 (0 < \alpha_0 < 1)$

$$H_0 : \mu \leq \mu_0$$

$$H_1 : \mu > \mu_0$$

Para cualquier procedimiento de contraste δ se define $\pi(\mu, \sigma^2|\delta)$ como función de potencia de δ. El objetivo es encontrar un procedimiento de contraste δ tal que:

- $\pi(\mu, \sigma^2 \mid \delta) \leq \alpha_0$ para todo punto $(\mu, \sigma^2) \in \Omega_0$
- $\pi(\mu, \sigma^2 \mid \delta) \leq \alpha_0$ debería ser lo más grande posible para todo punto $(\mu, \sigma^2) \in \Omega_1$

Podemos demostrar ([**?**]) que para cualquier nivel de significación específico $\alpha_0(0 < \alpha_0 < 1)$, existe un contraste UMP de estas hipótesis.

La función de potencia $\pi(\mu, \sigma^2 \mid \delta)$ del contraste UMP es:

$$\pi(\mu, \sigma^2 \mid \delta) = P(\text{Rechazar}H_0 \mid \mu) = P(\bar{X}_n \geq \mu_0 + (n^{1/2}(c - \mu_0)/\sigma)\sigma n^{-1/2}|\mu)$$

donde $c = \mu_0 + (n^{1/2}(c - \mu_0)/\sigma)\sigma n^{-1/2}$

El contraste UMP δ rechaza H_0 cuando $\bar{X}_n \leq c$ donde

$$c = \mu_0 - (n^{1/2}(c - \mu_0)/\sigma)\sigma n^{-1/2}$$

La función potencia $\pi(\mu, \sigma^2 \mid \delta)$ será:

$$\pi(\mu, \sigma^2 \mid \delta) = P(\bar{X}_n \leq c \mid \mu)$$

Para el contraste de hipótesis:

$$H_0 : \mu = \mu_0 \qquad\qquad H_1 : \mu \neq \mu_0$$

no existe un contraste de hipótesis UMP pero sí que podemos encontrar un procedimiento de contraste que verifíque:

1. $\pi(\mu, \sigma^2 \mid \delta) = \alpha_0$ si $\mu = \mu_0$
2. $\pi(\mu, \sigma^2 \mid \delta) < \alpha_0$ si $\mu < \mu_0$
3. $\pi(\mu, \sigma^2 \mid \delta) > \alpha_0$ si $\mu > \mu_0$
4. $\pi(\mu, \sigma^2 \mid \delta) \longrightarrow 0$ si $\mu \longrightarrow -\infty$
5. $\pi(\mu, \sigma^2 \mid \delta) \longrightarrow 1$ si $\mu \longrightarrow \infty$

5.3.2. Tamaño muestral.Prueba de no inferioridad / superioridad para la comparación de medias apareadas suponiendo varianza poblacional conocida

El contraste de hipótesis que consideraremos en este apartado es el siguiente:

$$H_0 : D \leq 0 \qquad H_1 : D > 0$$

El estadístico con el que trabajaremos en este apartado es:

$$Z = \frac{\bar{D}}{\sigma/\sqrt{n}}$$

Las regiones de aceptación y rechazo de este contraste son:

$$A = \{Z : Z \leq z_\alpha\} \quad C = \{Z : Z > z_\alpha\}$$

Rechazamos la hipótesis nula con un nivel de significación α si:

$$\frac{\bar{D}}{\sigma/\sqrt{n}} > z_\alpha$$

Si queremos lograr una potencia de β,

$$P\left(\frac{\sqrt{n}D}{\sigma} > z_\alpha\right) = 1 - \beta$$

Si H_1 es cierta, $D > 0$, por tanto

$$\begin{aligned} \bar{D} \qquad & \sim N(\mu, \sigma^2/n) \\ \frac{\bar{D} - (\mu - \mu_0)}{\sigma/\sqrt{n}} \qquad & \sim N(0,1) \end{aligned}$$

Por tanto

$$P\left(\frac{\sqrt{n}D}{\sigma} - \frac{(\mu - \mu_0)}{\sigma/\sqrt{n}} > z_\alpha - \frac{(\mu - \mu_0)}{\sigma/\sqrt{n}}\right) = 1 - \beta$$

El tamaño muestral necesario para lograr un poder de $1 - \beta$ viene dado por la siguiente ecuación:

$$z_\alpha - \frac{(\mu - \mu_0)}{\sigma/\sqrt{n}} = -z_\beta$$

Para poder obtener el tamaño muestral:

- Aproximamos $\mu - \mu_0$ por $\bar{D}$
- y obtenemos que podemos aproximar

$$\boxed{\mathbf{n} = \frac{(\mathbf{Z}_\alpha + \mathbf{Z}_\beta)^\mathbf{2}\boldsymbol{\sigma}^\mathbf{2}}{\bar{\mathbf{D}}^\mathbf{2}}}$$

5.3.3. Tamaño muestral.Prueba de no inferioridad / superioridad para la comparación de medias apareadas suponiendo varianza poblacional desconocida

El contraste de hipótesis que consideraremos en este apartado es el siguiente:

$$H_0 : D \leq 0 \qquad H_1 : D > 0$$

El estadístico con el que trabajaremos es:

$$T = \frac{\bar{D}}{S_D/\sqrt{n}}$$

Las regiones de aceptación y rechazo de este contraste son:

$$A = \{T : T \leq t_{\alpha,n-1}\} \quad C = \{T : T > t_{\alpha,n-1}\}$$

Rechazamos la hipótesis nula con un nivel de significación α si:

$$\frac{\bar{D}}{S_D/\sqrt{n}} > t_{\alpha,n-1}$$

Si queremos lograr una potencia β,

$$P\left(\frac{\bar{D}}{S_D/\sqrt{n}} > t_{\alpha,n-1}\right)$$

Si H_1 es cierta, $D > 0$, por tanto

$$\begin{aligned}
\bar{D} &\sim N(\mu, \sigma^2/n) \\
\frac{\bar{D} - (\mu - \mu_0)}{\sigma/\sqrt{n}} &\sim N(0,1) \\
\frac{(n-1)S_D^2}{\sigma^2} &\sim \chi^2_{n-1} \\
\frac{\bar{D} - (\mu - \mu_0)}{S_D/\sqrt{n}} &\sim t_{n-1}
\end{aligned}$$

El tamaño muestral viene dado por la siguiente ecuación:

$$P\left(\frac{\bar{D}}{S_D/\sqrt{n}} - \frac{(\mu - \mu_0)}{S_D/\sqrt{n}} > t_{\alpha,n-1} - \frac{(\mu - \mu_0)}{S_D/\sqrt{n}}\right) = 1 - \beta$$

De donde podemos deducir:

$$t_{\alpha,n-1} - \frac{(\mu - \mu_0)}{S_D/\sqrt{n}} = -t_{\beta,n-1}$$

En este caso, únicamente podemos calcular una aproximación al tamaño muestral, aunque si este es suficientemente grande podemos aporximar la distribución t-Student a la distribución normal para de esto modo calcularlo.Suponemos $n > 30$ y la aproximación es la siguiente:

$$z_\alpha - \frac{(\mu - \mu_0)}{S_D/\sqrt{n}} = -z_\beta$$

Para poder obtener el tamaño muestral:

- Aproximamos $\mu - \mu_0$ por $\bar{D}$
- y obtenemos que podemos aproximar

$$\boxed{\mathbf{n} = \frac{(\mathbf{z_{\alpha/2}} + \mathbf{z_\beta})^\mathbf{2}\mathbf{S_D^2}}{\mathbf{\bar{D}^2}}}$$

5.3.4. Tamaño muestral.Prueba de igualdad para la comparación de medias apareadas suponiendo varianza poblacional conocida

El contraste de hipótesis que consideraremos en este apartado es el siguiente:

$$H_0 : D = 0 \qquad H_1 : D \neq 0$$

El estadístico con el que trabajaremos es:

$$Z = \frac{\bar{D}}{\sigma/\sqrt{n}}$$

Las regiones de aceptación y rechazo para este contraste son:

$$A = \{Z : |Z| \leq z_{\frac{\alpha}{2}}\} \quad C = \{Z : |Z| > z_{\frac{\alpha}{2}}\}$$

Rechazamos la hipótesis nula con un nivel de significación α si:

$$\left|\frac{\bar{D}}{\sigma/\sqrt{n}}\right| > Z_{\alpha/2}$$

Si queremos obtener una potencia β,

$$\begin{aligned} &P\left(\left|\frac{\bar{D}}{\sigma/\sqrt{n}}\right| > z_{\alpha/2}\right) = \\ &= P\left(\frac{\bar{D}}{\sigma/\sqrt{n}} < -z_{\alpha/2}\right) + P\left(\frac{\bar{D}}{\sigma/\sqrt{n}} > z_{\alpha/2}\right) = \\ &= P\left(\frac{\bar{D}}{\sigma/\sqrt{n}} < -z_{\alpha/2}\right) + P\left(\frac{-\bar{D}}{\sigma/\sqrt{n}} < -z_{\alpha/2}\right) = 1 - \beta \end{aligned}$$

Si H_1 es cierta, $D > 0$, por tanto

$$\begin{aligned} \bar{D} \qquad &\sim N(\mu, \sigma^2/n) \\ \frac{\bar{D} - (\mu - \mu_0)}{\sigma/\sqrt{n}} \qquad &\sim N(0, 1) \end{aligned}$$

Por tanto

$$P\left(\frac{\bar{D}}{\sigma/\sqrt{n}} - \frac{(\mu-\mu_0)}{\sigma} < -z_{\alpha/2} - \frac{(\mu-\mu_0)}{\sigma/\sqrt{n}}\right) + P\left(\frac{-\bar{D}}{\sigma/\sqrt{n}} - \frac{(\mu-\mu_0)}{\sigma/\sqrt{n}} < -z_{\alpha/2} - \frac{(\mu-\mu_0)}{\sigma/\sqrt{n}}\right) = 1-\beta$$

Utilizando el mismo razonamiento que hemos desarrollado en el apartado 5.1.4, podemos desestimar un término cuyo valor es $< \alpha/2$ y obtenemos:

$$P\left(\frac{\bar{D}}{\sigma/\sqrt{n}} - \frac{(\mu-\mu_0)}{\sigma/\sqrt{n}/\sqrt{n}} < -z_{\alpha/2} + \frac{|\mu-\mu_0|}{\sigma/\sqrt{n}}\right) = \beta$$

El tamaño muestral necesario para lograr un poder de $1-\beta$ viene dado por la siguiente ecuación:

$$-z_{\alpha/2} + \frac{|\mu-\mu_0|}{\sigma/\sqrt{n}} = z_\beta$$

Para poder obtener el tamaño muestral:

- Aproximamos $\mu - \mu_0$ por $\bar{D}$
- y obtenemos que podemos aproximar

$$\boxed{\mathbf{n} = \frac{(\mathbf{z}_{\alpha/2} + \mathbf{z}_\beta)^2\sigma^2}{\bar{\mathbf{D}}^2}}$$

5.3.5. Tamaño muestral. Prueba de igualdad para la comparación de medias apareadas suponiendo varianza poblacional desconocida

El contraste de hipótesis que consideraremos en este apartado es el siguiente:

$$H_0 : D = 0 \qquad H_1 : D \neq 0$$

El estadístico es:

$$T = \frac{\bar{D}}{S_D/\sqrt{n}}$$

Las regiones de aceptación y rechazo son:

$$A = \{T : |T| \leq t_{\frac{\alpha}{2},n-1}\} \quad C = \{T : |T| > t_{\frac{\alpha}{2},n-1}\}$$

Rechazamos la hipótesis nula con un nivel de significación α si:

$$\left|\frac{\bar{D}}{S_D/\sqrt{n}}\right| > t_{\frac{\alpha}{2},n-1}$$

Si queremos obtener una potencia β,

$$P\left(\left|\frac{\bar{D}}{S_D/\sqrt{n}}\right| > t_{\frac{\alpha}{2},n-1}\right) =$$
$$= P\left(\frac{\bar{D}}{S_D/\sqrt{n}} < -t_{\frac{\alpha}{2},n-1}\right) + P\left(\frac{\bar{D}}{S_D/\sqrt{n}} > t_{\frac{\alpha}{2},n-1}\right) = 1-\beta$$

Si H_1 es cierta, $D \neq 0$, por tanto

$$\begin{aligned} \bar{D} &\sim N(\mu, \sigma^2/n) \\ \frac{\bar{D}-(\mu-\mu_0)}{\sigma/\sqrt{n}} &\sim N(0,1) \\ \frac{(n-1)S_D^2}{\sigma^2} &\sim \chi^2_{n-1} \\ \frac{\bar{D}-(\mu-\mu_0)}{S_D/\sqrt{n}} &\sim t_{n-1} \end{aligned}$$

$$= P\left(\frac{\bar{D}}{S_D/\sqrt{n}} - \frac{(\mu-\mu_0)}{S_D/\sqrt{n}} < -t_{\frac{\alpha}{2},n-1} - \frac{(\mu-\mu_0)}{S_D/\sqrt{n}}\right) +$$
$$+P\left(\frac{\bar{D}}{S_D/\sqrt{n}} - \frac{(\mu-\mu_0)}{S_D/\sqrt{n}} > t_{\frac{\alpha}{2},n-1} - \frac{(\mu-\mu_0)}{S_D/\sqrt{n}}\right) = 1-\beta$$

Utilizando un razonamiento análogo al realizado en el apartado 5.1.4, podemos desestimar un término cuyo valor es $< \alpha/2$, obteniendo:

$$P\left(\frac{\bar{D}}{S_D/\sqrt{n}} - \frac{(\mu-\mu_0)}{S_D/\sqrt{n}} < -t_{\frac{\alpha}{2},n-1} + \frac{|\mu-\mu_0|}{S_D/\sqrt{n}}\right) = \beta$$

El tamaño muestra se obtiene de la ecuación:

$$-t_{\frac{\alpha}{2},n-1} + \frac{|\mu-\mu_0|}{S_D/\sqrt{n}} = t_{\beta,n-1}$$

Para un n suficientemente grande, podemos aproximar la distribución t a la distribución normal, por tanto la ecuación que obtendríamos es de la forma:

$$-z_{\frac{\alpha}{2}} + \frac{|\mu-\mu_0|}{S_D/\sqrt{n}} = z_\beta$$

Para poder obtener el tamaño muestral:

- Aproximamos $\mu - \mu_0$ por $\bar{D}$
- y obtenemos que podemos aproximar

$$\boxed{\mathbf{n} = \frac{(\mathbf{z}_\beta + \mathbf{z}_{\frac{\alpha}{2}})^2 \mathbf{S_D^2}}{(\bar{\mathbf{D}})^2}}$$

5.4. Tamaño muestral para la comparación de más de dos medias

5.4.1. Deducción del contraste F

Consideremos un problema de contraste de hipótesis que utiliza la distribución F. Sean $X_1, ...X_m$ variables aleatorias que constituyen una muestra aleatoria de m observaciones de una distribución normal con media μ_1 y varianza σ_1^2 desconocidas y sean las variables aleatorias $Y_1, ..., Y_n$ que constituyen una muestra aleatoria independiente de n observaciones de otra distribución normal con media μ_2 y varianza σ_2^2 desconocidas.

Supongamos que se van a contrastar las siguientes hipótesis a un nivel de significación específico $\alpha_0 (0 < \alpha_0 < 1)$:

$$H_0 : \sigma_1^2 \leq \sigma_2^2 \tag{5.1}$$

$$H_1 : \sigma_1^2 > \sigma_2^2$$

Para cualquier procedimiento de contraste δ se define $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \mid \delta)$ como la función de potencia de δ. El objetivo es encontrar un procedimiento de contraste δ tal que:

- $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \mid \delta) \leq \alpha_0$ si $\sigma_1^2 \leq \sigma_2^2$
- $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \mid \delta)$ sea lo más grande posible si $\sigma_1^2 > \sigma_2^2$

No existe un contraste UMP que verifique las hipótesis (5.1), pero en la práctica es común utilizar un procedimiento particular, denominado contraste F. El contraste F, cuya deducción podemos encontrar en [**?**], tiene un nivel de significación específico α_0 y además tiene las cinco propiedades siguientes:

1. $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \mid \delta) = \alpha_0$ si $\sigma_1^2 = \sigma_2^2$
2. $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \mid \delta) < \alpha_0$ si $\sigma_1^2 < \sigma_2^2$
3. $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \mid \delta) > \alpha_0$ si $\sigma_1^2 > \sigma_2^2$
4. $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \mid \delta) \longrightarrow 0$ si $\sigma_1^2/\sigma_2^2 \longrightarrow -\infty$
5. $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \mid \delta) \longrightarrow 1$ si $\sigma_1^2/\sigma_2^2 \longrightarrow \infty$

Este procedimiento de contraste (ver [**?**]), define el estadístico

$$V = \frac{S_X^2/(m-1)}{S_Y^2/(n-1)}$$

Sabemos que la variables aleatoria S_X^2/σ_1^2 tiene una distribución χ^2 con $m-1$ grados de libertad y la variable aleatoria S_Y^2/σ_2^2 tiene una distribución χ^2 con $n-1$ grados de libertad. Además estas dos variables son independientes puesto

que son calculadas de dos muestras distintas. Por tanto la siguiente variables aleatoria tendrá una distribución F con $m-1$ y $n-1$ grados de libertad.

$$V' = \frac{S_X^2/[(m-1)\sigma_1^2]}{S_Y^2/[(n-1)\sigma_2^2]} \sim F_{m-1,n-1}$$

5.4.2. Análisis de la varianza

Supongamos que para $i = 1, 2, \cdots, k$, $X_{i1}, \cdots, X_{in_i}$ constituyen una muestra aleatoria de n_i observaciones de una variable aleatoria que sigue una distribución normal con media μ_i y varianza σ^2 desconocidas (la misma varianza para todas distribuciones aunque desconocida). Se define $n = \sum_{i=1}^{k} n_i$ y se supone que las n observaciones son independientes.

Supóngase que queremos contrastar las siguientes hipótesis a un nivel de significación específico α_0 $(0 < \alpha_0 < 1)$

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_1 : \text{existen diferencias en las medias}$$

5.4.3. Deducción del contraste

Antes de desarrollar un procedimiento de contraste adecuado necesitamos un poco de álgebra.

Para $i = 1, \cdots, k$, definimos $\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$, que es un EMV de μ_i, y definimos

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$$

que es un EMV para σ^2.

Es facil comprobar que

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{(X_{ij}-\mu_i)^2}{\sigma^2} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{(X_{ij}-\bar{X}_i)^2}{\sigma^2} + \sum_{i=1}^{k}\frac{n_i(\bar{X}_i-\mu_i)^2}{\sigma^2}$$

Si definimos:

$$Q_1 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{(X_{ij}-\mu_i)^2}{\sigma^2}$$
$$Q_2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{(X_{ij}-\bar{X}_i)^2}{\sigma^2}$$
$$Q_3 = \sum_{i=1}^{k}\frac{n_i(\bar{X}_i-\mu_i)^2}{\sigma^2}$$

vemos que

- Q_1 tiene una distribución χ^2 con $\sum_{i=1}^{k} n_i = n$ grados de libertad, y puede verse como una medida de la variación total de las observaciones alrededor de sus medias.
- Q_2 tiene una distribución χ^2 con $\sum_{i=1}^{k}(n_i-1) = n-k$ grados de libertad, y puede verse como una medida la dispersión de los valores de cada muestra con respecto a sus correspondientes medias muestrales.
- Q_3 tiene una distribución χ^2 con k grados de libertad, y puede verse como una medida de la variación total de las medias muestrales alrededor de las medias reales.

A su vez Q_3 puede descomponerse como:

$$Q_3 = \sum_{i=1}^{k} \frac{n_i(\bar{X}_i - \mu_i)^2}{\sigma^2} = \sum_{i=1}^{k} \frac{n_i(\bar{X}_i - \bar{X} - \alpha_i)^2}{\sigma^2} + \frac{(\bar{X} - \mu)^2}{\sigma^2}$$

donde:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{X}_i$$

$$\mu = \frac{1}{n}\sum_{i=1}^{k} n_i \mu_i$$

$$\alpha_i = \mu_i - \mu$$

Por tanto:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \frac{(X_{ij} - \mu_i)^2}{\sigma^2} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \frac{(X_{ij} - \bar{X}_i)^2}{\sigma^2} + \sum_{i=1}^{k} \frac{n_i(\bar{X}_i - \bar{X} - \alpha_i)^2}{\sigma^2} + \frac{(\bar{X} - \mu)^2}{\sigma^2}$$

donde los tres sumandos siguen distribuciones χ^2 con $n-k$, $k-1$ y 1 grado de libertad respectivamente.

El parámetro α_i se denomina efecto de la i-ésima distribución, y el contraste original equivaldría a plantear el contraste

$$H_0 : \alpha_i = 0 \text{ para } i = 1, \cdots, p$$
$$H_1 : \text{la hipótesis nula no es cierta}$$

Si llamamos

$$Q_4 = \sum_{i=1}^{k} \frac{n_i(\bar{X}_i - \bar{X} - \alpha_i)^2}{\sigma^2} \sim \chi^2_{k-1}$$

$$Q_5 = \frac{(\bar{X} - \mu)^2}{\sigma^2} \sim \chi^2_1$$

$Q_1 = Q_2 + Q_4 + Q_5$. Bajo H_0, Q_4 tiene la forma

$$Q_4^0 = \sum_{i=1}^{k} \frac{n_i(\bar{X}_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{k-1}$$

Como Q_2 y Q_4^0 son independientes, podemos definir la variable aleatoria

$$F = \frac{Q_4^0/(k-1)}{Q_2/(n-k)} \sim F_{k-1,n-k}$$

y se puede probar ([**?**]) que el procedimiento del cociente de verosimilitudes para este contraste de hipótesis especifica el rechazo de H_0 cuando

$$F > F_{\alpha,k-1,n-k}$$

donde $F_{\alpha,k-1,n-k}$ es el percentil $1-\alpha$ de la distribución F con $k-1$ y $n-k$ grados de libertad.

Bajo la hipótesis alternativa F se distribuye como una χ^2 no centrada con $k-1$ grados de libertad y con un parámetro de no centralidad, $\lambda = n\Delta$, donde

$$\Delta = \frac{1}{\sigma^2} \sum_{i=1}^{k} (\mu_i - \bar{\mu})^2$$

Por tanto, el tamaño muestral necesario para un poder de $1-\beta$ se obtiene resolviendo:

$$\chi^2_{k-1}(\chi^2_{\alpha,k-1}|\lambda) = \beta$$

donde $\chi^2_{\alpha,k-1}(.|\lambda)$ es la función de distribución acumulativa no centrada de la distribución χ^2 con $k-1$ grados de libertad y con parámetro de no centralidad λ. Dado un valor inicial de Δ y obteniendo λ de la tabla

	$1-\beta=0,80$		$1-\beta=0,90$	
k	$\alpha=0,01$	$\alpha=0,05$	$\alpha=0,01$	$\alpha=0,05$
2	11,68	7,85	14,88	10,51
3	13,89	9,64	17,43	12,66
4	15,46	10,91	19,25	14,18
5	16,75	11,94	20,74	15,41
6	17,87	12,83	22,03	16,47
7	18,88	13,63	23,19	17,42
8	19,79	14,36	24,24	18,29
9	20,64	15,03	25,22	19,09
10	21,43	15,65	26,13	19,83
11	22,28	16,25	26,99	20,54
12	22,89	16,81	27,80	21,20
13	23,57	17,34	28,58	21,84
14	24,22	17,85	29,32	22,44
15	24,84	18,34	30,04	23,03
16	25,44	18,82	30,73	23,59
17	26,02	19,27	31,39	24,13
18	26,58	19,71	32,04	24,65
19	27,12	20,14	32,66	25,16
20	27,65	20,16	33,27	25,66

El tamaño muestral es:

$$\boxed{\mathbf{n}=\frac{\lambda}{\mathbf{\Delta}}}$$

5.4.4. Comparación por parejas

El contraste de hipótesis que utilizaremos en este apartado para las hipótesis de interés son los siguientes:

$$H_0:\mu_i=\mu_j \qquad\qquad H_1:\mu_i\neq\mu_j$$

para algunos pares (i,j). Bajo las hipótesis anteriores, hay $k(k-1)/2$ posibles comparaciones. Sabemos que las comparacione múltiples augmentan el rror de tipo I, como resultado se sugiere un ajuste para controlar el error de tipo I y obtener el nivel de significación deseado. Asumiremos que hay τ comparaciones de interés, donde $\tau\leq /k(k-1)/2$. Rechazamos la hipótesis H_0 con un nivel de significación α si:

$$\left|\frac{\sqrt{n}(\bar{x_i}-\bar{x_j})}{\sqrt{2\hat{\sigma}}}\right|>t_{\alpha/(2\tau),k(n-1)}$$

El poder de esta prueba viene dado por

$$1 - P\left(\frac{\sqrt{n}\epsilon_{ij}}{\sqrt{2\hat{\sigma}}} < t_{\alpha/(2\tau),k(n-1)}\right) + P\left(\frac{\sqrt{n}\epsilon_{ij}}{\sqrt{2\hat{\sigma}}} < -t_{\alpha/(2\tau),k(n-1)}\right)$$

$$\approx 1 - P\left(\frac{\sqrt{n}|\epsilon_{ij}|}{\sqrt{2\hat{\sigma}}} < t_{\alpha/(2\tau),k(n-1)}\right)$$

donde $\epsilon_{ij} = \mu_i - \mu_j$. Por tanto, el tamaño muestral necesario para conseguir un poder $1-\beta$ para detectar una diferencia clinicamente significativa entre μ_i y μ_j es:

$$n = \text{máx}\{n_{ij}, \text{para todas las comparaciones significativas}\}$$

donde n_{ij} se calcula mediante

$$P\left(\frac{\sqrt{n}|\epsilon_{ij}|}{\sqrt{2\hat{\sigma}}} < t_{\alpha/(2\tau),k(n_{ij}-1)}\right) = \beta$$

Cuando el tamaño muestral es suficientemente grande, podemos utilizar la fórmula:

$$\mathbf{n} = \frac{\mathbf{2}(\mathbf{Z}_{\alpha/(\mathbf{2}\tau)} + \mathbf{Z}_{\beta})^{\mathbf{2}}\sigma^{\mathbf{2}}}{\epsilon_{\mathbf{ij}}^{\mathbf{2}}}$$

5.5. Tamaño muestral para la comparación de dos proporciones independientes

En este apartado estudiaremos la comparación de dos proporciones independientes distinguiendo tres casos, que vendrán dados por el tipo de contraste de hipótesis.

Al trabajar con proporciones, trabajaremos sobre variables binarias. Sea x_{ij} la respuesta binaria observada sobre el j-ésimo sujetos en el i-ésimo grupo de tratamiento, $i = 1, 2$, $j = 1, ..., n_i$. Fijado un i, podemos asumir que las variables X_{ij} están identicamente distribuidas con $P(X_{ij} = 1) = p_i$, en la práctica estimaremos p_i por el valor observado de la proporción en el grupo i-ésimo de tratamiento :

$$\hat{p_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$$

5.5.1. Deducción del contraste

Sean las variables X_{ij} $i = 1, 2, j = 1, ..., n_i$, las variables descritas anteriormente.

Supóngase que queremos contrastar las siguientes hipótesis a un nivel de significación específico α_0 $(0 < \alpha_0 < 1)$

$$H_0 : p_1 \leq p_2$$
$$H_1 : p_1 > p_2$$

Para cualquier procedimiento de contraste δ se define $\pi(p_1, p_2 \mid \delta)$ como la función de potencia de δ. El objetivo es encontrar un procedimiento de contraste δ tal que:

- $\pi(p_1, p_2 \mid \delta) \leq \alpha_0$ si $p_1 \leq p_2$
- $\pi(p_1, p_2 \mid \delta)$ sea lo más grande posible si $p_1 > p_2$

Puede demostrarse, que no existe un contraste UMP para este caso, pero sí que podemos encontrar un procedimiento de contraste δ que verifique:

1. $\pi(p_1, p_2 \mid \delta) = \alpha_0$ si $p_1 = p_2$
2. $\pi(p_1, p_2 \mid \delta) < \alpha_0$ si $p_1 < p_2$
3. $\pi(p_1, p_2 \mid \delta) > \alpha_0$ si $p_1 > p_2$
4. $\pi(p_1, p_2 \mid \delta) \longrightarrow 0$ si $p_1 - p_2 \longrightarrow -\infty$
5. $\pi(p_1, p_2 \mid \delta) \longrightarrow 1$ si $p_1 - p_2 \longrightarrow \infty$

5.5.2. Tamaño muestral.Prueba de igualdad para la comparación de dos proporciones

El contraste de hipótesis que consideraremos en este apartado se utiliza para ver si existe diferencia entre los grupos y es:

$$H_0 : p_1 = p_2 \qquad H_1 : p_1 \neq p_2$$

Por tanto:

$$Z = \frac{(\hat{p_1} - \hat{p_2})}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}}$$

Las regiones de aceptación y crítica para este contraste son:

$$A = \{Z : |Z| \leq z_{\alpha/2}\} \qquad C = \{Z : |Z| > z_{\alpha/2}\}$$

Rechazamos la hipótesis nula con un nivel de significación α si:

$$\left| \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p_1}(1-\hat{p_1})/n_1 + \hat{p_2}(1-\hat{p_2})/n_2}} \right| > z_{\alpha/2}$$

Si queremos obtener una potencia β,

$$P\left(\frac{(\hat{p_1}-\hat{p_2})}{\sqrt{\hat{p_1}(1-\hat{p_1})/n_1+\hat{p_2}(1-\hat{p_2})/n_2}} > z_{\alpha/2}\right) + \\ +P\left(\frac{(\hat{p_1}-\hat{p_2})}{\sqrt{\hat{p_1}(1-\hat{p_1})/n_1+\hat{p_2}(1-\hat{p_2})/n_2}} < -z_{\alpha/2}\right) = 1-\beta$$

Si H_1 es cierta, $p_1 \neq p_2$, por tanto como podemos aproximar una distribución binomial a una distribución normal (apartado **??**) tenemos:

$$\begin{aligned} \hat{p_1} \qquad & \sim N(p_1, \frac{\hat{p_1}(1-\hat{p_1})}{n_1}) \\ \hat{p_2} \qquad & \sim N(p_2, \frac{\hat{p_2}(1-\hat{p_2})}{n_2}) \\ \frac{(\hat{p_1}-\hat{p_2})-(p_1-p_2)}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}} \qquad & \sim N(0,1) \end{aligned}$$

Aplicamos el razonamiento desarrollado en el apartado 5.1.4 y tras desestimar un término cuyo valor es $< \alpha/2$ obtenemos:

$$P\left(\frac{(\hat{p_1}-\hat{p_2})-(p_1-p_2)}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}} - \frac{p_1-p_2}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}} < -z_{\alpha/2} + \frac{|p_1-p_2|}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}}\right) = 1-\beta$$

Podemos deducir que el tamaño muestral necesario para lograr obtener un poder de $1-\beta$ viene dado por la ecuación:

$$-z_{\alpha/2} + \frac{|p_1-p_2|}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}} = z_\beta$$

Para poder obtener el tamaño muestral:

- Aproximamos $p_1 - p_2$ por $\hat{p_1} - \hat{p_2}$
- Suponemos que existe una relación de proporcionalidad entre los dos tamaños muestrales:

$$\mathbf{n_1 = kn_2}$$

- y obtenemos que podemos aproximar

$$\boxed{\mathbf{n_2} = \frac{(\mathbf{z_{\alpha/2}} + \mathbf{z_\beta})^2[\frac{\hat{\mathbf{p_1}}(\mathbf{1}-\hat{\mathbf{p_1}})}{\mathbf{k}} + \hat{\mathbf{p_2}}(\mathbf{1}-\hat{\mathbf{p_2}})]}{(\hat{\mathbf{p_1}} - \hat{\mathbf{p_2}})^2}}$$

5.5.3. Caso particular para el cálculo del tamaño muestral.Prueba de igualdad para la comparación de proporciones

En este apartado veremos otra aproximación para el cálculo del tamaño muestral visto en el apartado anterior.

El contraste de hipótesis que consideraremos en este apartado es el siguiente:

$$H_0 : p_1 - p_2 = 0 \qquad H_1 : p_1 - p_2 \neq 0$$

El estadístico con el que trabajaremos es:

$$Z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{p}(1-\hat{p})}}$$

donde

$$\hat{p} = \frac{n_1\hat{p_1} + n_2\hat{p_2}}{n_1 + n_2}$$

Las regiones de aceptación y crítica de este contraste son:

$$A = \{Z : |Z| \leq z_{\alpha/2}\} \qquad C = \{Z : |Z| > z_{\alpha/2}\}$$

Rechazamos la hipótesis nula con un nivel de significación α si:

$$\left| \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{p}(1-\hat{p})}} \right| > z_{\alpha/2}$$

Si queremos obtener una potencia β,

$$P\left(\left| \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{p}(1-\hat{p})}} \right| > z_{\alpha/2} \right) =$$

$$= P\left(\frac{\hat{p_1} - \hat{p_2}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{p}(1-\hat{p})}} < -z_{\alpha/2} \right) + P\left(\frac{\hat{p_1} - \hat{p_2}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{p}(1-\hat{p})}} > z_{\alpha/2} \right) = 1 - \beta$$

Aproximamos $\left(\frac{1}{n_2} + \frac{1}{n_1}\right)\hat{p}(1-\hat{p})$ del siguiente modo([**?**]):

$$\left(\frac{1}{n_2}+\frac{1}{n_1}\right)\hat{p}(1-\hat{p}) \approx \left(\frac{1}{n_2}+\frac{1}{n_1}\right)p(1-p) \geq$$
$$\geq \frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2} \approx$$
$$\approx \frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}$$

Si H_1 es cierta, $p_1 \neq p_2$,por tanto como podemos aproximar una distribución binomial a una distribución normal (apartado **??**) tenemos:

$$\begin{aligned} \hat{p_1} \quad & \sim N(p_1, \frac{\hat{p_1}(1-\hat{p_1})}{n_1}) \\ \hat{p_2} \quad & \sim N(p_2, \frac{\hat{p_2}(1-\hat{p_2})}{n_2}) \\ \frac{(\hat{p_1}-\hat{p_2})-(p_1-p_2)}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}} \quad & \sim N(0,1) \end{aligned}$$

Por tanto tenemos

$$= P\left(\frac{\hat{p_1}-\hat{p_2}}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\hat{p}(1-\hat{p})}} - \frac{p_1-p_2}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}} <\right.$$
$$\left. < -z_{\alpha/2}\frac{\sqrt{(1/n_1+1/n_2)\hat{p}(1-\hat{p})}}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1+\hat{p}_2(1-\hat{p}_2)/n_2}} - \frac{p_1-p_2}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}}\right) +$$
$$+P\left(\frac{\hat{p_1}-\hat{p_2}}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\hat{p}(1-\hat{p})}} - \frac{p_1-p_2}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}} >\right.$$
$$\left. > z_{\alpha/2}\frac{\sqrt{(1/n_1+1/n_2)\hat{p}(1-\hat{p})}}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1+\hat{p}_2(1-\hat{p}_2)/n_2}} - \frac{p_1-p_2}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}}\right) = \beta$$

Realizando un razonamiento análogo al del apartado 5.1.4, podemos desestimar un término cuyo valor es $\alpha/2$ y obtenemos:

$$P\left(\frac{\hat{p_1}-\hat{p_2}}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\hat{p}(1-\hat{p})}}-\frac{p_1-p_2}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}}<\right.$$
$$\left.<-z_{\alpha/2}\frac{\sqrt{(1/n_1+1/n_2)\hat{p}(1-\hat{p})}}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1+\hat{p}_2(1-\hat{p}_2)/n_2}}+\frac{|p_1-p_2|}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}}\right)=\beta$$

Podemos deducir que el tamaño muestral necesario para lograr obtener un poder de $1-\beta$ viene dado por la ecuación:

$$-z_{\alpha/2}\frac{\sqrt{(1/n_1+1/n_2)\hat{p}(1-\hat{p})}}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1+\hat{p}_2(1-\hat{p}_2)/n_2}}+\frac{|p_1-p_2|}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}}=z_\beta$$

Para poder obtener el tamaño muestral:

- Aproximamos p_1-p_2 por $\hat{p_1}-\hat{p_2}$
- Suponemos que existe una relación de proporcionalidad entre los dos tamaños muestrales:
$$\mathbf{n_1=kn_2}$$
- y obtenemos que podemos aproximar

$$\boxed{\mathbf{n_2}=\frac{[\mathbf{z}_{\alpha/2}\sqrt{(1+1/k)\hat{p}(1-\hat{p}))}+\mathbf{z}_\beta\sqrt{\hat{\mathbf{p}}_1(1-\hat{\mathbf{p}}_1)/k+\hat{\mathbf{p}}_2(1-\hat{\mathbf{p}}_2)}]^2}{(\hat{\mathbf{p}_1}-\hat{\mathbf{p}_2})^2}}$$

5.5.4. Tamaño muestral.Prueba de No inferioridad/Superioridad para la comparación de proporciones

El contraste de hipótesis que consideraremos en este apartado es el siguiente:

$$H_0: p_1-p_2\leq 0 \qquad H_1: p_1-p_2>0$$

El estadístico con el que trabajaremos es:

$$Z=\frac{(\hat{p_1}-\hat{p_2})}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}}$$

Las regiones de aceptación y crítica de este contraste son:

$$A=\{z: z\leq z_\alpha\} \qquad C=\{z: z>z_\alpha\}$$

Rechazamos la hipótesis nula con un nivel de significación α si:

$$\frac{(\hat{p_1} - \hat{p_2})}{\sqrt{\hat{p_1}(1-\hat{p_1})/n_1 + \hat{p_2}(1-\hat{p_2})/n_2}} > z_\alpha$$

Si queremos obtener una potencia β,

$$P\left(\frac{(\hat{p_1} - \hat{p_2})}{\sqrt{\hat{p_1}(1-\hat{p_1})/n_1 + \hat{p_2}(1-\hat{p_2})/n_2}} > z_\alpha\right) = 1 - \beta$$

Si H_1 es cierta, $p_1 \neq p_2$, por tanto como podemos aproximar una distribución binomial a una distribución normal (apartado **??**) tenemos:

$$\begin{aligned} \hat{p_1} \qquad & \sim N(p_1, \frac{\hat{p_1}(1-\hat{p_1})}{n_1}) \\ \hat{p_2} \qquad & \sim N(p_2, \frac{\hat{p_2}(1-\hat{p_2})}{n_2}) \\ \frac{(\hat{p_1} - \hat{p_2}) - (p_1 - p_2)}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}} \qquad & \sim N(0,1) \end{aligned}$$

Por tanto

$$P\left(\frac{(\hat{p_1} - \hat{p_2}) - (p_1 - p_2)}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}} > z_\alpha - \frac{p_1 - p_2}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}}\right) = 1 - \beta$$

El tamaño muestral necesario para lograr un poder de $1 - \beta$ viene dado por la siguiente ecuación:

$$z_\alpha - \frac{p_1 - p_2}{\sqrt{\hat{p_1}(1-\hat{p_1})/n_1 + \hat{p_2}(1-\hat{p_2})/n_2}} = -z_\beta$$

Para poder obtener el tamaño muestral:

- Aproximamos $p_1 - p_2$ por $\hat{p_1} - \hat{p_2}$
- Suponemos que existe una relación de proporcionalidad entre los dos tamaños muestrales:

$$\mathbf{n_1 = kn_2}$$

- y obtenemos que podemos aproximar

$$\boxed{\mathbf{n_2} = \frac{(\mathbf{z}_\alpha + \mathbf{z}_\beta)^\mathbf{2}[\frac{\hat{\mathbf{p_1}}(\mathbf{1}-\hat{\mathbf{p_1}})}{\mathbf{k}} + \hat{\mathbf{p_2}}(\mathbf{1} - \hat{\mathbf{p_2}})]}{(\hat{\mathbf{p_1}} - \hat{\mathbf{p_2}})^\mathbf{2}}}$$

5.6. Tamaño muestral para la comparación de dos proporciones, con población de referencia

En este apartado como en el anterior tenemos la comparación de dos proporciones, aunque pueda parecer que se trata del mismo caso, una de las proporciones es la obtenida de una población de referencia, por ello no calcularemos dos tamaños muestrales, uno para cada grupo, únicamente calcularemos el tamaño muestral para el grupo sometido al nuevo tratamiento puesto que para la población de referencia no es posible tomar datos nuevos. Tomaremos p como la respuesta al nuevo fármaco y p_0 el valor de referencia.

Dados $x_i, i = 1, ..., n$ respuestas binarias del sujeto ith. Podemos asumir que x_i's están identicamente distribuidas con $P(x_i = 1) = p$, en la práctica estimaremos p por el valor observado de la proporción en el sujeto ith de tratamiento :

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n_i} x_i$$

5.6.1. Deducción del contraste

Sean las variables $X_1, ..., X_n$, las variables descritas anteriormente. Supongamos que se desean contrastar las siguientes hipótesis con un nivel de significación $\alpha_0 (0 < \alpha_0 < 1)$

$$H_0 : p \leq p_0$$
$$H_1 : p > p_0$$

Para cualquier procedimiento de contraste δ se define $\pi(p \mid \delta)$ como función de potencia de δ. El objetivo es encontrar un procedimiento de contraste δ tal que:

- $\pi(p \mid \delta) \leq \alpha_0$ para todo punto $p \in \Omega_0$
- $\pi(p \mid \delta) \leq \alpha_0$ debería ser lo más grande posible para todo punto $p \in \Omega_1$

Puede demostrarse, que no existe un contraste UMP para este caso, pero sí que podemos encontrar un procedimiento de contraste δ que verifique:

1. $\pi(p \mid \delta) = \alpha_0$ si $p = p_0$
2. $\pi(p \mid \delta) < \alpha_0$ si $p < p_0$
3. $\pi(p \mid \delta) > \alpha_0$ si $p > p_0$
4. $\pi(p \mid \delta) \longrightarrow 0$ si $p \longrightarrow -\infty$
5. $\pi(p \mid \delta) \longrightarrow 1$ si $p \longrightarrow \infty$

5.6.2. Tamaño muestral.Prueba de igualdad para la comparación dos proporciones con población de referencia.

El contraste de hipótesis que consideramos en este apartado es el siguiente:

$$H_0 : p - p_0 = 0 \qquad\qquad H_1 : p - p_0 \neq 0$$

El estadístico con el que trabajaremos es:

$$Z = \frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{\hat{p}(1-\hat{p})}}$$

Rechazamos la hipótesis nula con un nivel de significación α si:

$$\left| \frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{\hat{p}(1-\hat{p})}} \right| > z_{\alpha/2}$$

Si queremos obtener una potencia β,

$$P\left(\left| \frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{\hat{p}(1-\hat{p})}} \right| > z_{\alpha/2} \right) =$$
$$= P\left(\frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{\hat{p}(1-\hat{p})}} < -z_{\alpha/2} \right) + P\left(\frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{\hat{p}(1-\hat{p})}} > z_{\alpha/2} \right) = 1 - \beta$$

Si H_1 es cierta, $p - p_0 \neq 0$,por tanto como podemos aproximar una distribución binomial a una distribución normal (apartado **??**) tenemos:

$$\begin{aligned} \hat{p} \qquad & \sim N(p, \frac{\hat{p}(1-\hat{p})}{n}) \\ \frac{(\hat{p} - p_0) - (p - p_0)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \qquad & \sim N(0,1) \end{aligned}$$

$$= P\left(\frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{\hat{p}(1-\hat{p})}} - \frac{(p - p_0)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < -z_{\alpha/2} - \frac{(p - p_0)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \right) +$$
$$+P\left(\frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{\hat{p}(1-\hat{p})}} - \frac{(p - p_0)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} > z_{\alpha/2} - \frac{(p - p_0)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \right) = 1 - \beta$$

Utilizando un razonamiento análogo al realizado en el apartado 5.1.4, podemos desestimar un término de tamaño $< \alpha/2$ y obtenemos:

$$P\left(\frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{\hat{p}(1-\hat{p})}} - \frac{(p - p_0)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < -z_{\alpha/2} + \frac{|p - p_0|}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \right) = \beta$$

El tamaño muestral necesario para lograr un poder de $1-\beta$ viene dado por la siguiente ecuación:

$$-z_{\alpha/2}+\frac{|p-p_0|}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}=z_\beta$$

Para obtener el tamaño muestral, aproximamos p por $hatp$ y obtenemos:

$$\boxed{\mathbf{n}=\frac{(\mathbf{z_{\alpha/2}}+\mathbf{z}_\beta)^{\mathbf{2}}\hat{\mathbf{p}}(\mathbf{1}-\hat{\mathbf{p}})}{(\hat{\mathbf{p}}-\mathbf{p_0})^{\mathbf{2}}}}$$

5.6.3. Tamaño muestral.Caso particular de la prueba de igualdad para la comparación dos proporciones con población de referencia.

El contraste de hipótesis que consideramos en este apartado es el siguiente:

$$H_0: p-p_0=0 \qquad\qquad H_1: p-p_0\neq 0$$

El estadístico con el que trabajaremos es:

$$Z=\frac{\sqrt{n}(\hat{p}-p_0)}{\sqrt{p_0(1-p_0)}}$$

Rechazamos la hipótesis nula con un nivel de significación α si:

$$\left|\frac{\sqrt{n}(\hat{p}-p_0)}{\sqrt{p_0(1-p_0)}}\right|>z_{\alpha/2}$$

Si queremos obtener una potencia β,

$$P\left(\left|\frac{\sqrt{n}(\hat{p}-p_0)}{\sqrt{p_0(1-p_0)}}\right|>z_{\alpha/2}\right)=$$
$$=P\left(\frac{\sqrt{n}(\hat{p}-p_0)}{\sqrt{p_0(1-p_0)}}<-z_{\alpha/2}\right)+P\left(\frac{\sqrt{n}(\hat{p}-p_0)}{\sqrt{p_0(1-p_0)}}>z_{\alpha/2}\right)=1-\beta$$

Aproximamos $\left(\frac{1}{n_2}+\frac{1}{n_1}\right)\hat{p}(1-\hat{p})$ del siguiente modo([**?**]):

$$\frac{\sqrt{n}(\hat{p}-p_0)}{\sqrt{\hat{p}(1-\hat{p})}}\approx\frac{\sqrt{n}(p-p_0)}{\sqrt{p(1-\hat{p})}}\geq$$
$$\geq\frac{\sqrt{n}(p-p_0)}{\sqrt{p_0(1-p_0)}}\approx\frac{\sqrt{n}(\hat{p}-p_0)}{\sqrt{p_0(1-p_0)}}$$

Si H_1 es cierta, $p - p_0 \neq 0$, por tanto como podemos aproximar una distribución binomial a una distribución normal (apartado **??**) tenemos:

$$\begin{aligned} \hat{p} \qquad & \sim N(p, \frac{\hat{p}(1-\hat{p})}{n}) \\ \frac{(\hat{p}-p_0)-(p-p_0)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \qquad & \sim N(0,1) \end{aligned}$$

Por tanto

$$P\left(\frac{\sqrt{n}(\hat{p}-p_0)}{\sqrt{p_0(1-p_0)}} - \frac{(p-p_0)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < -z_{\alpha/2}\frac{\sqrt{p_0(1-p_0)}}{\sqrt{\hat{p}(1-\hat{p})}} - \frac{(p-p_0)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}\right) +$$

$$+P\left(\frac{\sqrt{n}(\hat{p}-p_0)}{\sqrt{p_0(1-p_0)}} - \frac{(p-p_0)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} > z_{\alpha/2}\frac{\sqrt{p_0(1-p_0)}}{\sqrt{\hat{p}(1-\hat{p})}} - \frac{(p-p_0)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}\right) = 1-\beta$$

Utilizando un razonamiento análogo al realizado en el apartado 5.1.4, podemos desestimar un término de tamaño $< \alpha/2$ y obtenemos:

$$P\left(\frac{\sqrt{n}(\hat{p}-p_0)}{\sqrt{\hat{p}(1-\hat{p})}} - \frac{(p-p_0)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < -z_{\alpha/2}\frac{\sqrt{p_0(1-p_0)}}{\sqrt{\hat{p}(1-\hat{p})}} + \frac{(|p-p_0|}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}\right) = \beta$$

El tamaño muestral necesario para lograr un poder de $1-\beta$ viene dado por la siguiente ecuación:

$$-z_{\alpha/2}\frac{\sqrt{p_0(1-p_0)}}{\sqrt{\hat{p}(1-\hat{p})}} + \frac{|p-p_0|}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} = z_\beta$$

Para obtener el tamaño muestral debemos aproximar p por $\hat{p}$ y obtenemos

$$\boxed{\mathbf{n} = \frac{[\mathbf{z}_{\alpha/2}\sqrt{\mathbf{p_0(1-p_0)}} + \mathbf{z}_\beta\sqrt{\hat{\mathbf{p}}(\mathbf{1}-\hat{\mathbf{p}})}]^{\mathbf{2}}}{(\hat{\mathbf{p}}-\mathbf{p_0})^{\mathbf{2}}}}$$

5.6.4. Tamaño muestra.Prueba de No inferioridad/Superioridad para la comparación de dos proporciones con población de referencia

El contraste de hipótesis que consideraremos en este apartado es el siguiente:

$$H_0 : p - p_0 \leq 0 \qquad H_1 : p - p_0 > 0$$

El estadístico con el que trabajaremos es:

$$Z = \frac{\sqrt{n}(\hat{p}-p_0)}{\sqrt{\hat{p}(1-\hat{p})}}$$

Las regiones de aceptación y crítica para este contraste son:

$$A = \{Z : Z \leq z_\alpha\} \qquad C = \{Z : Z > z_\alpha\}$$

Rechazamos la hipótesis nula con un nivel de significación α si:

$$\frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{\hat{p}(1-\hat{p})}} > z_\alpha$$

Si queremos obtener una potencia β,

$$P\left(\frac{\sqrt{n}((\hat{p} - p_0))}{\sqrt{\hat{p}(1-\hat{p})}} > z_\alpha\right) = 1 - \beta$$

Si H_1 es cierta, $p - p_0 > 0$, por tanto como podemos aproximar una distribución binomial a una distribución normal (apartado **??**) tenemos:

$$\begin{aligned} \hat{p} \qquad & \sim N\left(p, \frac{\hat{p}(1-\hat{p})}{n}\right) \\ \frac{(\hat{p} - p_0) - (p - p_0)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \qquad & \sim N(0,1) \end{aligned}$$

Por tanto podemos deducir que:

$$P\left(\frac{\sqrt{n}((\hat{p} - p_0))}{\sqrt{\hat{p}(1-\hat{p})}} - \frac{(p - p_0)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} > z_\alpha - \frac{(p - p_0)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}\right) = 1 - \beta$$

El tamaño muestral necesario para lograr un poder de $1 - \beta$ viene dado por la siguiente ecuación:

$$z_\alpha - \frac{\sqrt{n}((p - p_0))}{\sqrt{\hat{p}(1-\hat{p})}} = -z_\beta$$

Siendo

$$\boxed{\mathbf{n} = \frac{(\mathbf{z}_\alpha + \mathbf{z}_\beta)^\mathbf{2}\hat{\mathbf{p}}(\mathbf{1} - \hat{\mathbf{p}})}{(\hat{\mathbf{p}} - \mathbf{p_0})^\mathbf{2}}}$$

5.7. Comparación de más de dos proporciones

Para la comparación de más de dos proporciones utilizaremos la prubea de la χ^2, es una de las pruebas más frecuentes utilizadas para el contraste de variables cualitativas, aplicandose para comparar si dos caracteríticas cualitativas están relacionadas entre sí, si varias muestras de carácter cualitativo proceden de igual población o si los datos observados siguen una determinada disribución teórica.

Para su cálculo se calculan las frecuencias esperadas (las que deberían haberse observado si la hipótesis de independencia fuese cierta), para compararlas con las observadas en la realidad. Se calcula el valor del estadístico χ^2 como:

$$\chi^2 = \sum \frac{|O_{ij} - E_{ij}|^2}{E_{ij}} \sim \chi^2_{(f-1)(c-1)}$$

donde

- O_{ij} corresponden a las frecuencias observadas dentro de la casilla de la fila i y columna j.
- E_{ij} corresponden a las frecuencias esperadas o teóricas.
- f es el número de filas y c el número de columnas.
- $(f-1)*(c-1)$ corresponden a los grados de libertad de la distribución del estadístico de contraste

El primer paso consiste en construir la tabla de contingencia asociada a las variables a analizar. A partir de ella se calculan las frecuencias esperadas en cada casilla bajo la suposición de que las variables sean independientes.

En el caso de una tabla de contingencia de f filas y c columnas, las frecuencias esperadas se pueden obtener de manera similar, como se describe en la siguiente tabla f x c:

	A_1	A_2	...	A_c	TOTAL
Y_1	$E_{11} = \frac{f_1 * f_1}{f}$	$E_{12} = \frac{f_1 * f_2}{f}$		$E_{1c} = \frac{f_1 * f_c}{f}$	$f_{1.}$
Y_2	$E_{21} = \frac{f_2 * f_1}{f}$	$E_{22} = \frac{f_2 * f_2}{f}$		$E_{2c} = \frac{f_2 * f_c}{f}$	$f_{2.}$
...					
Y_f	$E_{f1} = \frac{f_f * f_1}{f}$	$E_{f2} = \frac{f_f * f_2}{f}$		$E_{fc} = \frac{f_f * f_c}{f}$	$f_{f.}$
TOTAL	$f_{,1}$	$f_{,2}$	...	$f_{.c}$	$f_{..}$

Cuadro 5.1: Tabla de contingencia

Para obtener el valor de la χ^2 las frecuencias observadas se comparan con los valores observados. Así, cuando mayor sea la diferencia entre los valores esperados y los observados mayor será el valor del estadístico, existiendo en este caso asociación entre las variables comparadas. El hecho de que las diferencias se eleven al cuadrado convierte cualquier diferencia en positiva, lo que indica si existe o no relación entre los factores pero no en que sentido se produce tal asociación.

Cuando el tamaño muestral no es demasiado grande, puede introducirse algún sesgo en los cálculos, ya que estos contrastes aproximan una distribución discreta por una continua por lo que podemos utilizar la corrección de Yates.

Veremos ahora el test χ^2 de Pearson para obtener la fórmula del tamaño muestral. Consideraremos el estadístico:

$$T = \sum_{i=1}^{c} \sum_{j=1}^{f} \frac{n(E_{ij} - E_{i.}E.j)^2}{E_i.E_{.j}}$$

Bajo la hipótesis nula, Y y A son independientes, T se distribuye asintóticamente como una χ^2 con $(f-1)(c-1)$ grados de libertad.

Bajo la alternativa local con

$$\lim_{n\to\infty} \sum_{i=1}^{c} \sum_{j=1}^{f} \frac{n(E_{ij} - E_{i.}Ej.)^2}{E_i.E_{j.}} = \delta$$

donde $E_ij = P(Y = y_i, A = a_j)$,$E_{i.} = P(Y = y_i)$ y $E_{.j} = P(A = a_j)$

Para un α dado, si deseamos obtener una potencia de β, δ puede obtenerse resolviendo:

$$\chi^2_{(f-1)(c-1)}(\chi_{\alpha,(f-1)(c-1)} \mid \delta) = 1 - \beta$$

Sea $\delta_{\alpha,\beta}$ la solución, el tamaño muestral necesario para lograr una potencia β viene dado por:

$$\boxed{\mathbf{n} = \delta_{\alpha,\beta} \left[\sum_{\mathbf{i=1}}^{\mathbf{c}} \sum_{\mathbf{j=1}}^{\mathbf{f}} \frac{\mathbf{n(E_{ij} - E_{i.}E.j)^2}}{\mathbf{E_i.E_{.j}}} \right]^{-1}}$$

Capítulo 6

Estudios epidemiológicos

En este apartado veremos el cálculo del tamaño muestral para dos tipos de estudios que aunque sus fórmulas son casos particulares de comparación de proporciones su estudio resulta de interés ya que son dos de los estudios más utilizados.

6.1. Estudios de cohortes

En este tipo de estudio los individuos son identificados en función de la presencia o ausencia de exposición a un determinado factor. En este momento todos están libres de la enfermedad de interés y son seguidos durante un período de tiempo para observar la frecuencia de aparición del fenómeno que nos interesa. Si al finalizar el período de observación la incidencia de la enfermedad es mayor en el grupo de expuestos, podremos concluir que existe una asociación estadística entre la exposición a la variable y la incidencia de la enfermedad.

Los estudios de cohorte pretenden evaluar una posible relación causa-efecto sin embargo, una de sus principales limitaciones es la imposibilidad del investigador de controlar la exposición del factor de riesgo a diferencia de lo que ocurre en los ensayos clínicos. Es este sentido son estudios observacionales. Otra característica que presentan los estudios de cohortes, es que son longitudinales por lo que es posible comprobar que la presencia del factor de riesgo antecede al evento, algo que es difícil demostrar en los estudios transversales y que resulta fundamental para confirmar asociaciones de causalidad. Para más información consultar por ejemplo [**?**].

La cuantificación de esta asociación la podemos calcular construyendo una razón entre la incidencia del fenómeno en los expuestos a la variable y la incidencia del fenómeno en los no expuestos . Esta razón entre incidencias se conoce como riesgo relativo (RR)y su cálculo se estima como:

Sean

a: N^o de personas NO expuestas al factor de riesgo que NO desarrollan la enfermedad.

b: Nº de personas expuestas al factor de riesgo que NO desarrollan la enfermedad.

c: Nº de personas NO expuestas al factor de riesgo que desarrollan la enfermedad.

d: Nº de personas expuestas al factor de riesgo que desarrollan la enfermedad.

I_ne: $c/(a+c)$: Incidencia en el grupo de personas no expuestas.

I_e: $d/(b+d)$: Incidencia en el grupo de personas expuestas.

Riesgo relativo:

$$RR = \frac{I_e}{I_ne} = \frac{d/(b+d)}{c/(a+c)}$$

Entre las ventajas y desventajas de estos estudios podemos destacar las siguientes:

Ventajas

- Estiman incidencia directamente. Se puede estimar la incidencia de la enfermedad en los grupos expuestos y no expuestos, así como en diferentes exposiciones a la vez.
- Existe una secuencia temporal entre la exposición del factor de riesgo y la enfermedad.
- Se pueden estudiar exposiciones poco frecuentes.
- Se pueden estudiar enfermedades con largos periodos de latencia.

Limitaciones

- Suelen tener un coste elevado dada su complejidad en cuanto al diseño de la cohorte y además requieren generalmente un tamaño muestral elevado.
- No son útiles en enfermedades raras y poco frecuentes, siendo preferible un diseño de casos-control.
- Pueden requerir periodos de seguimiento muy largos, con lo que aumenta la posibilidad de pérdidas de individuos durante el seguimiento.
- El paso del tiempo puede introducir cambios en los métodos y criterios diagnósticos.
- La exposición no es asignada aleatoriamente, a diferencia de un ensayo clínico.

Los individuos de la cohorte pueden salir de ella porque la abandonen, mueran, se pierdan del estudio, o simplemente porque se presente la enfermedad o evento de interés. Dependiendo del momento en el que se inicie el estudio respecto a la ocurrencia del evento, los podemos clasificar como prospectivos o retrospectivos. En los estudios prospectivos, en el momento de iniciar el estudio

aún no ha ocurrido el evento de interés o enfermedad, mientras que en los retrospectivos, al iniciar el estudio sabemos si se ha producido o no la enfermedad y reconstruimos hacia atrás el pasado para evaluar la presencia del factor de riesgo.

Si el riesgo relativo es igual a uno significa que no hay asociación entre las variables, es decir la cantidad de veces que un evento ocurra va a ser igual con o sin la presencia del factor, la relación es 1 : 1, es por ello que tendremos en cuenta el siguiente contraste de hipótesis para realizar el cálculo del tamaño muestral:

$$H_0 : RR = 1 \qquad\qquad H_1 : RR \neq 1$$

si el objetivo es probar que el RR es estadísticamente diferente de 1 se deberá çonocer":

a) Dos de los siguientes elementos:

- Probabilidad de enfermar en personas expuestas al factor de interés P_1
- Probabilidad de enfermar en personas no expuestas al factor de interés: P_2
- Riesgo Relativo: RR

b) Nivel de confianza: $100(1-\alpha)\,\%$

c) Potencia del test: $100(1-\beta)\,\%$

d) Cantidad de no expuestos por cada expuesto: r

La fórmula del tamaño muestral se obtiene realizando un razonamiento análogo al realizado en el apartado(5.5.3) ya que si $p_1 = p_2 \rightarrow RR = 1$ y es la siguiente:

$$\boxed{\mathbf{n} = \frac{\mathbf{z}_{\frac{\alpha}{2}}\sqrt{(\mathbf{r}+\mathbf{1})\mathbf{p}(\mathbf{1}-\mathbf{p})} - \mathbf{z}_{\beta}\sqrt{\mathbf{r}\mathbf{p_1}(\mathbf{1}-\mathbf{p_1}) + \mathbf{p_2}(\mathbf{1}-\mathbf{p_2})}}{\mathbf{r}(\mathbf{p_1}-\mathbf{p_2})^{\mathbf{2}}}}$$

donde $p = (p_1 + rp_2)/(r+1)$

Los estudios de cohorte en función de su diseño pueden presentar diferentes formas:

- Cohorte única: Corresponde a un grupo de individuos que en el pasado fueron sometidos a una exposición, si bien en el presente no lo están.
- Dos Cohortes: Es el diseño mas habitual, en el que dos grupos de individuos libres de la enfermedad uno de ellos expuestos al factor de riesgo y el otro no, son seguidos a lo largo del tiempo. Posteriormente se mide en cada uno de ellos la incidencia de la enfermedad.

- Cohortes múltiples: Se crean varios grupos con diferentes grados de exposición y posteriormente se compara la incidencia de la enfermedad con un grupo control en donde su exposición al factor de riesgo ha sido muy baja o casi inexistente. Este tipo de estudios permite evaluar una relación dosis-respuesta.

- Casos y controles anidados: Concluido el periodo de seguimiento e identificados los pacientes con la enfermedad, estos son seleccionados (casos) y comparados con un grupo de individuos de la cohorte elegidos aleatoriamente y que no han desarrollado la enfermedad (controles).

Algo que hay que tener en cuenta en los estudios de cohorte, es que pueden tener tiempos de seguimiento muy largos abarcando muchos años, lo que conlleva necesariamente perdidas de seguimiento. Además, el grado de exposición al factor de riesgo puede ser cambiante. Un problema añadido, es que el diagnóstico de la enfermedad e incluso la propia definición la misma, puede variar con el paso de los años por lo que debe ser tenido en cuenta en el momento del análisis. No debemos olvidar que las técnicas diagnósticas mejoran con el paso del tiempo y como consecuencia de ello, la sensibilidad en la detección de la enfermedad aumenta progresivamente.

6.2. Estudio de casos y controles

Este tipo de estudio identifica a personas con una enfermedad (u otra variable de interés) que estudiemos y los compara con un grupo control apropiado que no tenga la enfermedad. La relación entre uno o varios factores relacionados con la enfermedad se examina comparando la frecuencia de exposición a éste u otros factores entre los casos y los controles.

A este tipo de estudio que es de los más utilizados en la investigación médica se le podría describir como un procedimiento epidemiológico analítico, no experimental con un sentido retrospectivo, ya que partiendo del efecto, se estudian sus antecedentes, en el que se seleccionan dos grupos de sujetos llamados casos y controles según tengan o no la enfermedad. Para más información, consultar por ejemplo [**?**].

Entre las ventajas y desventajas de estos estudios podemos destacar las siguientes:

Ventajas

- Útiles en enfermedades raras o con periodos de latencia largos.

- Suelen ser mas sencillos y menos costosos que los estudios de cohortes prospectivo.

- Se pueden estudiar simultáneamente diferentes factores etiológicos (fruto de la causalidad).

- Suelen tener menos errores en la clasificación de la enfermedad.

- En algunas circunstancias, pueden servir como estimadores del Riesgo Relativo.

Limitaciones

- Muchas veces no existe una secuencia temporal clara entre la exposición del factor de riesgo y la enfermedad.
- No sirve para valorar exposiciones raras o poco frecuentes.
- No se puede calcular directamente la incidencia de la enfermedad entre expuestos y no expuestos.
- La calidad de la información recogida sobre la exposición del factor de riesgo puede ser distinta en los pacientes enfermos que en los sanos.

Si la frecuencia de exposición a la causa es mayor en el grupo de casos de la enfermedad que en los controles, podemos decir que hay una asociación entre la causa y el efecto. La medida de asociación que permite cuantificar esta asociación se llama odds ratio (razón de productos cruzados, razón de disparidad, proporción de desigualdades ...) que se calcula del siguiente modo:

Sean

a: Nº de personas ENFERMAS (casos) CON el factor de riesgo.
b: Nº de personas ENFERMAS (casos) SIN el factor de riesgo.
c: Nº de personas SANAS (controles) CON el factor de riesgo.
d: Nº de personas SANAS (controles) SIN el factor de riesgo.

$$OR = \frac{a * d}{b * c}$$

Un factor importante en estos estudios es además de la selección de los pacientes, el tamaño de la muestra (tema que nos ocupa) ya que de ello dependerá la posibilidad de comprobar la hipótesis de asociación entre un factor de riesgo y una enfermedad (o relación causa-efecto).

Si la Odds ratio es igual a uno significa que no hay asociación entre las variables, es decir la cantidad de veces que un evento ocurra va a ser igual con o sin la presencia del factor, la relación es $1:1$, es por ello que tendremos en cuenta el siguiente contraste de hipótesis para realizar el cálculo del tamaño muestral:

$$H_0 : OR = 1 \qquad\qquad H_1 : OR \neq 1$$

Además para realizar el cálculo debemos conocer:

a) Dos de los siguientes elementos:

- Probabilidad de la exposición al factor en individuos enfermos P_1
- Probabilidad de la exposición en individuos sanos P_2
- Razón de Odds OR

b) Nivel de confianza: $100(1-\alpha)\,\%$

c) Potencia del test: $100(1-\beta)\,\%$

Notemos que si conocemos el valor de P_1 y OR, podemos calcular, P_2 mediante:

$P_2 = \frac{P_1}{OR(1-P_1)+P_1}$ análogamente podemos obtener OR por:

$OR = \frac{P_1/(1-P_1)}{P_2/(1-P_2)}$ y P_1 por:

$P_1 = \frac{P_2}{(1-P_2)/OR+P_2}$

La fórmula del tamaño muestral se obtiene realizando un razonamiento análogo al realizado en el apartado(5.5.3) ya que si $p_1 = p_2 \rightarrow OR = 1$ y es la siguiente:

$$\boxed{\mathbf{n} = \frac{\mathbf{z_{\frac{\alpha}{2}}}\sqrt{\mathbf{(r+1)p(1-p)}} - \mathbf{z_{\beta}}\sqrt{\mathbf{rp_1(1-p_1) + p_2(1-p_2)}}}{\mathbf{r(p_1-p_2)^2}}}$$

donde $p = (p_1 + rp_2)/(r+1)$

Otros diseños de estudios de casos y controles son:

- Casos y controles anidados: Consiste en seleccionar los individuos que forman los casos y los controles a partir de un estudio de cohortes. Supongamos que de un estudio de cohortes se seleccionan como casos todas aquellas personas que presentan la enfermedad, y como controles una muestra aleatoria de personas que no la tienen. Una característica de este tipo de estudios, es que un mismo individuo puede ser caso y control. Si una persona está libre de la enfermedad puede ser seleccionada como control, sin embargo, si años después desarrolla la enfermedad, podría formar parte de los casos. De cualquier forma esto no invalida el estudio, ya que se trata de medir su exposición al factor de riesgo en al momento de realizar el análisis, da igual si esta persona fue elegida como control en un análisis anterior.

- Casos y controles emparejados: Una forma de controlar el efecto de la confusión entre la exposición y la enfermedad consiste en elegir para cada caso uno o más controles de similares características en aquellas variables que pensamos pudieran ser confusoras y de este modo mejorar la eficiencia del estudio. Por ejemplo, podemos obtener para cada caso, un control del mismo sexo y grupo de edad. Sin embargo, a veces puede ser complicada la elección de los controles cuando se trata de emparejar por múltiples variables ya que hay que identificar los individuos que cumplen todas las características incrementando el coste del estudio. Además hay que señalar que el emparejamiento es un proceso irreversible y que requiere de un análisis estadístico concreto para datos emparejados.

- Casos y controles cruzados: Este diseño de estudio podría considerarse como una variante del estudio emparejado con la peculiaridad de que cada caso sirve también como su propio control.Se suelen utilizar cuando una exposición corta o infrecuente provoca un evento agudo a corto plazo.

Capítulo 7

Pruebas paramétricas y no paramétricas

Las pruebas paramétricas hacen la suposición de conocimiento previo de que los datos se distribuyen normalmente. Varias pruebas pueden llevarse a cabo para determinar si es o no es una suposición válida. Si los datos no están normalmente distribuidos, pueden transformarse de diversas maneras para que las pruebas paramétricas se puedan seguir utilizando. Como alternativa, se pueden utilizar los análisis no paramétricos. Las pruebas no paramétricas no hacen suposiciones sobre la distribución de los datos [**?**].

Las pruebas paramétricas realizan inferencias sobre parámetros que modelizan un conjunto de datos que se distribuyen normalmente. La media, la varianza, la desviación estándar y la asimetría son ejemplos. Estos parámetros se utilizan para hacer inferencias en las pruebas paramétricas. Por el contrario, las pruebas no paramétricas se centran sobre la media y la varianza de la distribución.

Hasta ahora en todos los apartados del cálculo del tamaño muestral hemos utilizado pruebas paramétricas para la comparación tanto de medias como de proporciones. A continuación veremos una breve introducción de algunas pruebas no paramétricas que deberían utilizarse para los distintos casos, si como hemos dicho no podemos transformar los datos para poder suponer normalidad y poder aplicar las pruebas paramétricas.

7.1. Pruebas no paramétricas con dos variables relacionadas

7.1.1. Prueba de Wilcoxon

La prueba de Wilcoxon es aplicable a variables medibles en al menos una escala ordinal relacionadas. Consideramos un contraste de hipótesis donde la hipótesis nula del contraste postula que las muestras proceden de la misma dis-

tribución de probabilidad y la alternativa establece que hay diferencias respecto a la tendencia central de las poblaciones.

La prueba consiste en calcular las diferencias entre las puntuaciones de los elementos de cada par asociados y ordenarlas de menor a mayor por valor absoluto. Una vez ordenadas las diferencias, se numeran de 1 a n, siendo n el número de elementos de la muestra; al número asignado se le denomina rango. El rango 1 se asigna a la mínima diferencia observada en valor absoluto, y así sucesivamente hasta n, cuyo rango corresponde a la máxima diferencia. Si hay dos iguales, se asigna a cada diferencia igual la media de los rangos implicados en el empate.

Una vez ordenados los datos, se suman los rangos de las diferencias positivas, $W+$, y las negativas,$W-$ y se elige el menor de los dos. Los casos en que la diferencia es cero se ignoran.

La prueba se basa en que, si la hipótesis nula es cierta y las dos tienen el mismo valor central, los rangos deben estar repartidos de forma homogénea, y tan probable es encontrar un rango grande positivo como negativo. Por lo tanto, si se suman los rangos correspondientes a diferencias positivas, $W+$, y los rangos correspondientes a diferencias negativas, $W-$, deben ser similares y se encontrarán pequeñas diferencias debidas al azar. Si las diferencias entre las suma de rangos son grandes, indica que entre las variables hay diferencias debidas a causas distintas al azar.

Las hipótesis en la prueba de Wilcoxon se pueden enunciar también de la manera siguiente:

$$H_0 : W(+) = W(-) \qquad\qquad H_1 : W(+) \neq W(-)$$

.

El estadístico para la prueba de Wilcoxon es el siguiente

$$T^+ = \sum_{i=1}^{n} R_i \psi_i$$

donde R_i es la suma de los rangos R_i correspondientes a los valores positivos de $z_i = y_i - x_i$ para n pares de observaciones, denominadas (x_i, y_i) y

$$\psi_i = \begin{cases} 1 & \text{si} \quad z_i > 0 \\ 0 & \text{si} \quad z_i > 0 \end{cases}$$

El contraste se resuelve para muestras pequeñas, consultando las tablas de Wilcoxon

n	.005 (una cola) .01 (dos colas)	.01 (una cola) .002 (dos colas)	.025 (una cola) .05 (dos colas)	.05 (una cola) .010 (dos colas)
5	*	*	*	1
6	*	*	1	2
7	*	0	2	4
8	0	2	4	6
9	2	3	6	8
10	3	5	8	11
11	5	7	11	14
12	7	10	14	17
13	10	13	17	21
14	13	16	21	26
15	16	20	25	30
16	19	24	30	36
17	23	28	35	41
18	28	33	40	47
19	32	38	46	54
20	37	43	52	60
21	43	49	59	68
22	49	56	66	74
23	55	62	73	83
24	61	69	81	92
25	68	77	90	101
26	76	85	98	110
27	84	93	107	120
28	92	102	117	130
29	100	111	127	141
30	109	120	137	152

Cuadro 7.1: Valores críticos de T para la prueba de rangos con signos de Wilcoxon

en las que se representan las máximas o mínimas sumas de rangos consideradas aceptables. Para muestras mayores que 30 se puede hacer una aproximación a la normal.

7.1.2. Test de Mcnemar

Este test se utiliza cuando se trata de comparar dos proporciones observadas en dos muestras relacionadas, por ejemplo, en el mismo grupo de individuos en dos ocasiones distintas de tiempo (antes y después de algún estímulo). Se pretende comparar si se produce algún cambio significativo entre ambas mediciones. Clasificamos un grupo de individuos entre dos categorías mutuamente excluyentes, indicadas por + (positivo) y – (negativo). Pasado un estímulo o

intervención es posible que alguno de estos individuos cambie de categoría, de manera que la tabla de frecuencias que se obtendría sería la siguiente:

		Después		
		Positivo	Negativo	Total
Antes	Positivo	a	b	a+b
	Negativo	c	d	c+d
	Total	a+c	b+d	n

Cuadro 7.2: Tabla general de contingencia para dos proporciones observadas en un mismo grupo en dos ocasiones distintas de tiempo

La proporción de individuos con la característica positiva antes sería $p_1 = \frac{a+b}{n}$ y después sería $p_2 = \frac{a+c}{n}$. Nos interesa contrastar si la diferencia entre estas dos proporciones es cero (hipótesis nula) frente a que p_1 y p_2 sean diferentes ($p_1 - p_2 = \frac{b-c}{n} \neq 0$) Para ello,nos podemos centrar en las celdas b y c que son las que muestran discordancia entre las dos mediciones, contrastando si el número de individuos que tras la intervención han dejado de presentar la característica+(b) es el mismo que el número de individuos que tras la intervención han realizado el cambio inverso (c), es decir han dejado de presentar la característica $-$. El error estándar para la diferencia entre dos proporciones es:

$EED = \frac{1}{n}\sqrt{b + c - \frac{(b-c)^2}{n}}$ que bajo la hipótesis nula ($H_1 : b - c = 0$) se reduce a $EED = \frac{1}{n}\sqrt{b+c}$

El estadístico de contraste que sigue una distribución Normal (0,1) se calcula como:

$$Z = \frac{p_1 - p_2}{EED} = \frac{\frac{b-c}{n}}{\frac{1}{n}\sqrt{b+c}} = \frac{b-c}{\sqrt{b+c}}$$

También se puede considerar el estadístico de contraste: $\chi^2 = \frac{(b-c)^2}{b+c}$ que sigue una distribución Ji-cuadrado con 1 grado de libertad. Como en el caso de la χ^2, si las frecuencias son pequeñas puede utilizarse la corrección de Yates:

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c}$$

7.2. Pruebas no paramétricas para dos muestras independientes

7.2.1. Prueba de Mann-Whitney

Esta prueba es aplicable para comparar los valores de dos variables cuantitativas independientes, también se puede aplicar a variables ordinales, es la versión no paramétrica de la habitual prueba t de Student, por lo que podemos aplicarla para la comparación de medias. La dos muestras pueden tener tamaños

distintos. Es la prueba no paramétrica considerada más potente para comparar los valores de dos variables cuantitativas independientes.

El procedimiento es el siguiente: se agrupan los datos de las dos muestras en un sólo grupo, se ordenan los datos de menor a mayor, asignándole a cada dato el rango correspondiente a su orden. Si no hay diferencias entre las dos variables, se espera que los rangos estén uniformemente repartidos entre los dos grupos; por el contrario, si hay diferencias entre las dos variables, se espera que los rangos menores se asocien con una de las muestras y los mayores con la otra.

Las hipótesis pueden enunciarse de la manera siguiente:

H_0 : No hay diferencias entre las variables $\qquad$ H_1 : Hay diferencias entre las variables

Si existen diferencias mayores de las esperadas por efecto del azar entre los vaores de las variables, los detectaría la prueba propuesta por Mann-Whitney, basada en la suma de los rangos correspondientes a cada muestra.

Se dispone de datos cuatitativos correspondientes a dos muestras aleatorias, con tamaños n_1 y n_2; la suma de lso rangos correspondientes a cada grupo se denotan mediantes R_1 y R_2. Los estadísticos U_1 y U_2 se obtienen mediante las expresiones siguientes:

$$U_1 = n_1 n_2 + \left[\frac{n_1(n_1+1)}{2}\right] - R_1$$
$$U_2 = n_1 n_2 + \left[\frac{n_2(n_2+1)}{2}\right] - R_2$$

Una vez calculados los parámetros anteriores, se elige el menor; a este valor se le denomina U y, mediante las tablas:

U	$n_2 = 3$		
n_1	1	2	3
0	0,250	0,100	0,050
1	0,500	0,200	0,100
2	0,750	0,400	0,200
3		0,600	0,300
4			0,500
5			0,650

U	$n_2 = 4$			
n_1	1	2	3	4
0	0,200	0,067	0,028	0,014
1	0,400	0,133	0,057	0,029
2	0,600	0,267	0,114	0,057
3		0,400	0,200	0,100
4		0,600	0,314	0,171
5			0,429	0,243
6			0,571	0,343
7				0,443
8				0,557

U	$n_2 = 5$				
n_1	1	2	3	4	5
0	0,167	0,047	0,018	0,008	0,004
1	0,333	0,095	0,036	0,016	0,008
2	0,500	0,190	0,071	0,032	0,016
3	0,667	0,286	0,125	0,056	0,028
4		0,429	0,196	0,095	0,048
5		0,571	0,286	0,143	0,075
6			0,393	0,206	0,111
7			0,500	0,278	0,155
8			0,607	0,365	0,210
9				0,452	0,274
10				0,548	0,345
11					0,421
12					0,500
13					0,579

U	$n_2 = 6$					
n_1	1	2	3	4	5	6
0	0,143	0,036	0,012	0,002	0,001	
1	0,286	0,071	0,024	0,004	0,002	
2	0,428	0,143	0,048	0,009	0,004	
3	0,571	0,214	0,083	0,015	0,008	
4		0,321	0,131	0,026	0,013	
5		0,429	0,190	0,041	0,021	
6		0,571	0,274	0,063	0,032	
7			0,357	0,089	0,047	
8			0,452	0,123	0,066	
9			0,548	0,165	0,090	
10				0,214	0,120	
11				0,268	0,155	
12				0,331	0,197	
13				0,396	0,242	
14				0,465	0,294	
15				0,535	0,350	
16					0,409	
17					0,469	
18					0,531	

se comprueba si las diferencias entre los valores de las variables son estadísticamente significativas. Cuando las muestras tienen más de 20 casos, se consigue una buena aproximación a una distribución normal.

7.2.2. Prueba de Kolmogorov-Smirnov

Esta prueba no paramétrica es válida para comparar dos variables independientes, las variables deben ser cuantitativas. La prueba pretende comprobar si las distribuciones poblacionales de las dos variables son iguales o distintas. La prueba de dos colas es sensible a diferencias en tendencia central, dispersión y colocación. La hipótesis se pueden enunciar de la manera siguiente:

H_0 : Las distribuciones son iguales $\qquad$ H_1 : Las distribuciones son distintas

El estadístico de contraste es D, que es la máxima diferencia entre las frecuencis relativas acumuladas calculadas para cada valor. El parámetro D se puede calcular mediante la expresión:

$$D = \text{máx}[F_1(x) - F_2(x)]$$

donde F_1 es la frecuencia relativa acumulada de valores de la primera muestra, que son iguales o menores que x, F_2 es la proporción de valores de la segunda muestra que son iguales o menores que x. La diferencia anterior se calcula para todos los valores y el valor de la diferencia máxima es el parámetro D.

El parámetro D está tabulado y, consultando las correspondientes tablas, se puede comprobar si las diferencias son o no estadísticamente significativas.

7.2.3. Test exacto de Fisher

El test exacto de Fisher permite analizar la asociación entre dos variables dicotómicas cuando no se cumplen las condiciones necesarias para la aplicación del test de la X^2.Para aplicar la prueba de la X^2 se exige que el 80 % de las celdas de la tabla de contingencia presenten frecuencias esperadas superiores a 5. Así, en las tablas 2x2 es necesario que se verifique en todas sus celdas, aunque en la práctica se permite que una de ellas se muestre ligeramente por debajo. El test de Fisher se aplica también cuando alguno de los valores esperados es inferior a 2.

Esta prueba se basa en el cálculo de la probabilidad exacta de las frecuencias observadas. Evalúa la probabilidad asociada a cada una de las tablas 2x2 que se pueden formar manteniendo los mismos totales de filas y columnas que los de la tabla observada. La probabilidad exacta de observar un conjunto concreto de frecuencias a, b, c y d en una tabla 2x2, cuando se asume independencia y los totales de filas y columnas se consideran fijos, viene dada por una distribución hipergeométrica:

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

Esta probabilidad se calcula para todas las tablas de contingencia que puedan formarse con los mismos totales que en la tabla observada, utilizándolos para calcular el valor de la p asociado al test de Fisher. El valor de p puede calcularse sumando aquellas probabilidades inferiores a la probabilidad de la tabla observada. Si el valor de p es pequeño ($p < 0,05$) se debe rechazar la hipótesis nula de independencia, asumiendo que ambas variables están asociadas estadísticamente

7.3. Pruebas no paramétricas para k variables relacionadas

7.3.1. Prueba de Friedman

En estadística la prueba de Friedman es una prueba no paramétrica desarrollado por el economista Milton Friedman. Equivalente a la prueba ANOVA para medidas repetidas en la versión no paramétrica, el método consiste en ordenar los datos por filas o bloques, reemplazándolos por su respectivo orden. Al ordenarlos, debemos considerar la existencia de datos idénticos.

Esta prueba puede utilizarse en aquellas situaciones en las que se seleccionan n grupos de k elementos de forma que los elementos de cada grupo sean lo más parecidos posible entre sí, y a cada uno de los elementos del grupo se le aplica

uno de entre k "tratamientos", o bien cuando a cada uno de los elementos de una muestra de tamaño n se le aplican los k "tratamientos".

La hipótesis nula que se contrasta es que las respuestas asociadas a cada uno de los "tratamientos" tienen la misma distribución de probabilidad o distribuciones con la misma mediana, frente a la hipótesis alternativa de que por lo menos la distribución de una de las respuestas difiere de las demás. Para poder utilizar esta prueba las respuestas deben ser variables contínuas y estar medidas por lo menos en una escala ordinal.

Los datos se disponen en una tabla en la que en cada fila se recogen las respuestas de los k elementos de cada grupo a los k tratamientos.

A las observaciones de cada fila se les asignan rangos de menor a mayor desde 1 hasta k; a continuación se suman los rangos correspondientes a cada columna, siendo R_j la suma correspondiente a la columna j-ésima. Si la hipótesis nula es cierta, la distribución de los rangos en cada fila se debe al azar, y es de esperar que la suma de los rangos correspondientes a cada columna sea aproximadamente igual a $n(k+1)/2$. La prueba de Friedman determina si las R_j observadas difieren significativamente del valor esperado bajo la hipótesis nula.

$$H_0 : R_1 = R_2 = ... = R_j \qquad H_1 : R_i \neq R_j \text{para algún } i, j$$

Para resolver este contraste de hipótesis Friedman propuso un estadístico que se distribuye como una χ^2 con $k-1$ grados de libertad, siendo k el número de variables relacionadas.

$$\chi^2_{FR} = \frac{12}{nK(J+1)} \sum_{i=1}^{k} R_i^2 - 3n(K+1)$$

donde n representa el número de elementos o bloques, k el número de variables relacionadas y R_i representa la suma de rangos de la i-ésima variable.

7.3.2. Q de Cochran

Esta prueba es válida para evaluar si la respuesta de un grupo de elementos ante un conjunto de características es homogénea, o por el contrario existen diferencias entre los elementos estudiados y tiene una respuesta dicotómica y permite estudiar si las diferencias entre las caracterísicas son estadísticamente significativas. Pues utilizarse para comparar proporciones de dos o más grupos apareados.

Las hipótesis se pueden enunciar como:

$$H_0 : \text{No hay diferencias entre las características}$$

$$H_1 : \text{Hay diferencias entre las características}$$

El estadístico de contraste es el siguiente:

$$Q = \frac{K(K-1)\sum_{i=1}^{k}[(K-1)(\sum_{j=1}^{n} S_j)^2]}{K\sum_{j=1}^{n} S_j - \sum_{j=1}^{n} S_j^2}$$

donde K es el número de pruebas o características, n es el número de casos, S_j es la suma de las puntuaciones otorgadas para cada caso y T_i es la suma de las puntuaciones de cada prueba.

7.4. Pruebas no paramétricas para k variables independientes

7.4.1. Test de Kruskall-Wallis

Esta prueba es válida para comparar simultáneamente los valores de K variables cuantitativas u ordinales.

Las hipótesis son:

$$H_0 : \text{Los valores de las k variables son similares}$$

$$H_1 : \text{Los valores de las k variables son diferentes}$$

La prueba se basa en agrupar los datos de K variables en un solo grupo, ordenando de menor a mayor, asignando a cada dato el correspondiente rango. Si los valores son similares, los datos de las K variables se repartirán de manera homogénea en el grupo común ordenado, y la suma de los rangos asignados a cada grupo tendrá valores próximos. Por el contrario, si los valores son distintos son de esperar diferencias entre las sumas de rangos más grandes que las aplicables por el azar.

El estadístico de contraste para esta prueba se puede calcular mediante la siguiente expresión:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

En la expresión anterior, K es el número de grupos, n_i es el número de casos del i-ésimo grupo y N es el número total de sujetos que intervienen en la prueba.

Para muestras pequeñas, la significación de los valores de H está tabulada. Según aumenta el tamaño de la muestra, H se aproxima a una distribución χ^2 con $k-1$ grados de libertad. La aproximación a la χ^2 puede hacerse para muestras de más de ocho elementos.

En caso de empates (hay dos o más datos con los mismos valores), se resuelven asignando a cada dato implicado en el empate el rango medio correspondiente a todos los rangos implicados en dicho empate. En caso de empates, el estaístico H debe ser corregido y se calcula mediantes la expresión:

$$H = \frac{\frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)}{1 - \left[\frac{\sum_{s=1}^{r} (t_s^3 - t_s)}{(N^3 - N)} \right]}$$

donde s indica el s-ésimo empate y r es el número total de empates;t_s es el número de sujetos empatados en el s-ésimo empate.

Apéndice A

Anexo I: Calculadora del tamaño muestral

En este apartado veremos el manual del usuario de la calculadora del tamaño muestral realizada durante la estacia en prácticas. En el que se detalla la estructura de las diferentes pestañas con las que podrá trabajar el usuario así como la solución de algunos problemas con los que pueda encontrarse durante su uso.

A.1. Introducción

El presente documento describe la herramienta creada con Microsoft Excel® para la estimación del tamaño muestral. En los siguientes apartados se presenta la descripción general de la herramienta, así como una guía para su correcto uso en la que se incluyen imágenes de la calculadora. Todas las tablas que aparecerán con los distintos valores de α y β han sido sacados de [**?**]

A.2. Introducción a la herramienta

A.2.1. Descripción

La herramienta que se describe a continuación incluye el cálculo del tamaño muestral para siete casos diferentes:

- Estimar una proporción
- Estimar una media
- Comparación de dos proporciones independientes
- Comparación de una proporción observada con una población de referencia
- Comparación de dos medias independientes

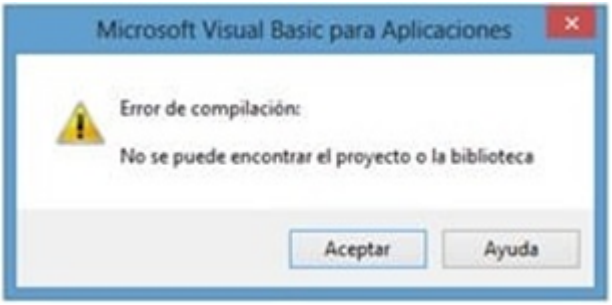
- Comparación de dos medias apareadas en un solo grupo
- Comparación de dos medias apareadas en dos grupos

Las características de estas funciones serán explicadas en los apartados correspondientes a cada una de ellas.

A.2.2. Aspectos generales de la herramienta

Al abrir la calculadora

Al abrir la calculadora aparecerá un aviso solicitando habilitar las macros. El usuario deberá pulsar Aceptar y automáticamente se visualizará la portada con el título y un botón "Entrar" que lleva al cuerpo de la calculadora. Tras habilitar las macros puede aparecer la siguiente advertencia:



Este error se debe a que el proyecto contiene una referencia a una biblioteca que esta desactivada. Para poder ejecutar la calculadora, las bibliotecas que deben estar activadas son las siguientes:

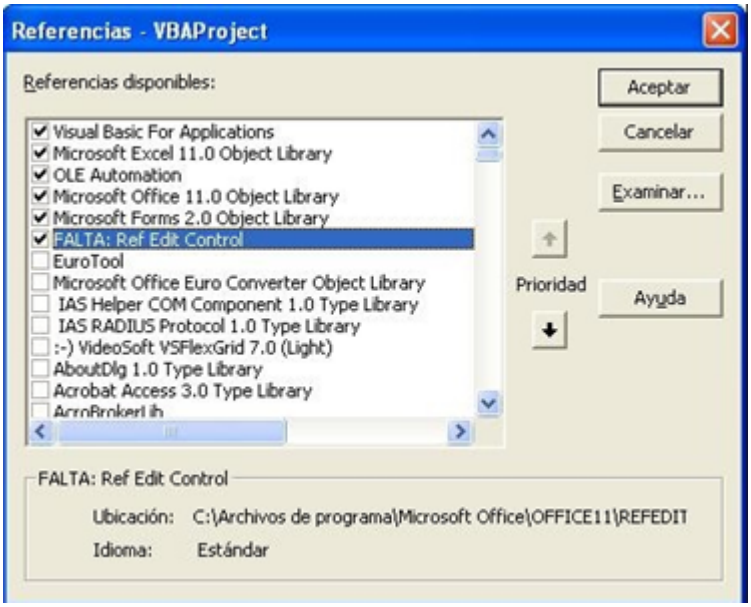
- Visual Basic for aplications
- Microsoft Excel 16.0 Object Library o Microsoft Excel 15.0 Object Library
- OLE Automation
- Microsoft Office 16.0 Object Library o Microsoft Office 15.0 Object Library
- Microsoft Forms 2.0 Object Library
- Microsoft Windows Common Controls 6.0
- Microsoft Outlook 16.0 Object Library o Microsoft Outlook 15.0 Object Library

Para comprobar cual o cuales de las anteriores bibliotecas están desactivadas y activarlas, el usuario debe seguir el procedimiento siguiente:

1. Ir al apartado Desarrollador de la barra de tareas de Excel
2. Pulsar el botón Visual Basic

3. En la barra de tareas ir a Herramientas → Referencias

La biblioteca o bibliotecas que producen el error aparecerán indicadas como vemos en la imagen siguiente, precedidas por la palabra FALTA.



4. Deseleccionar la biblioteca precedida por FALTA

5. Buscar esta biblioteca en la lista y seleccionarla

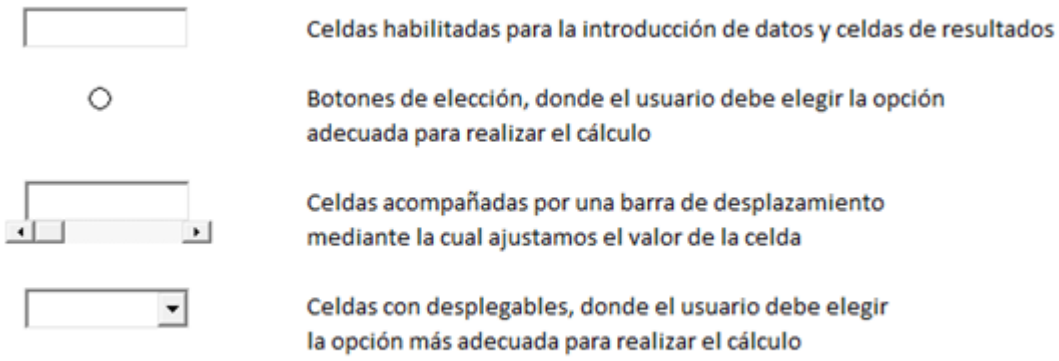
6. Pulsar Aceptar

Navegación a través de la calculadora

Todas las pestañas (excepto la portada y la introducción) disponen de un botón en la parte inferior para volver a la pestaña de introducción. Adicionalmente, se han habilitado botones en la parte inferior de cada pestaña que, una vez calculado el tamaño muestral, generan un documento de Microsoft Word con los valores de los parámetros y el resultado final.

Tipos de celdas

A lo largo de la herramienta el usuario encontrará cuatro tipos de celda:



A.3. Uso de la herramienta

A continuación, se describen las pestañas que componen la herramienta junto con imágenes de la visualización que encontrará el usuario en cada una de ellas.

A.3.1. Introducción a la calculadora

La primera pestaña es la *Introducción* (Figura A.1) la cual contiene una breve explicación del objetivo de la calculadora y los diferentes casos para los que el usuario puede realizar el cálculo del tamaño muestral, estos están divididos en dos bloques, dependiendo de si el usuario quiere calcular el tamaño muestral en estudios para determinar parámetros o quiere calcular el tamaño muestral en estudios para contraste de hipótesis. Cada una de las opciones está precedida por un botón de elección mediante el cual el usuario tendrá acceso a la pestaña correspondiente para realizar el cálculo elegido.

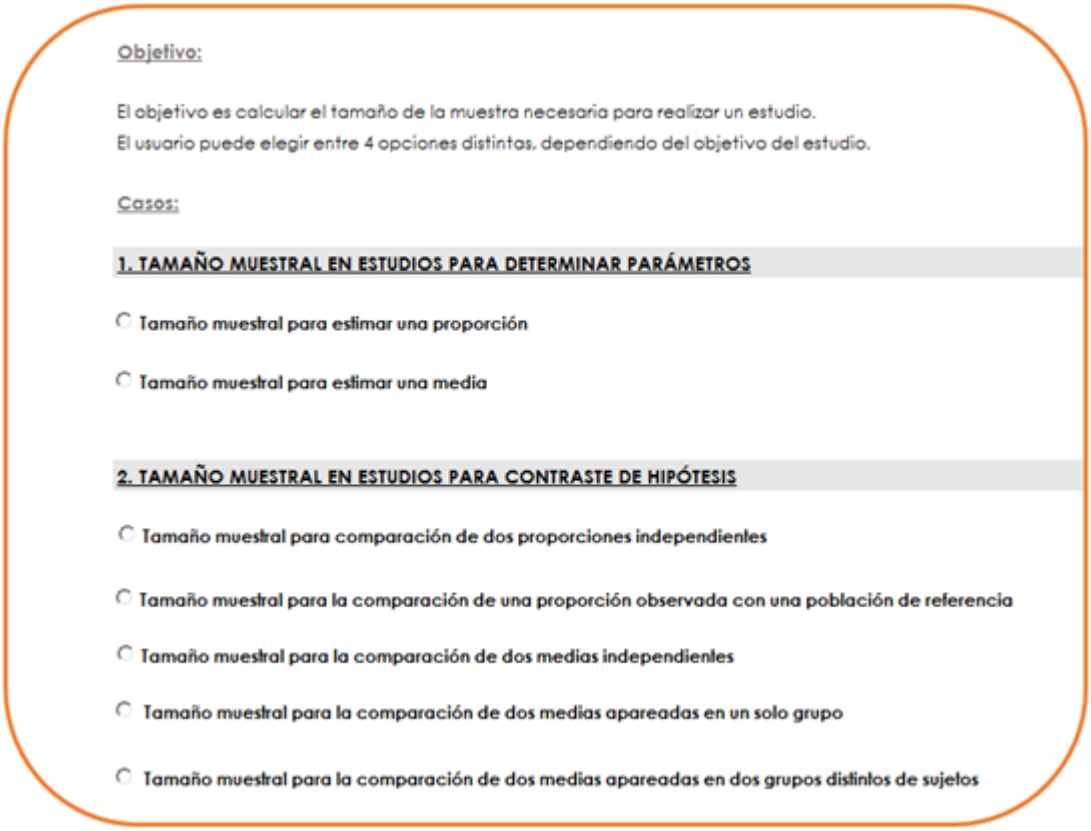


Figura A.1: Pestaña de introducción

A.3.2. Tamaño muestral para estimar una proporción

Definición e introducción de los parámetros

La función Tamaño muestral para estimar una proporción (Figura A.2) permite calcular el tamaño de la muestra requerido para realizar un estudio cuyo objetivo principal sea la estimación de una proporción. Los parámetros que intervienen en este cálculo son los que se mencionan a continuación:

- Total de la población (N): es la población que cumple las condiciones requeridas para el estudio sobre la que debe ser tomada la muestra.
- Nivel de confianza o seguridad (α): corresponde a la probabilidad de que la estimación efectuada se ajuste a la realidad. Este nivel de confianza da lugar a un coeficiente z_α. Las probabilidades entre las cuales va a poder elegir el usuario y sus correspondientes coeficientes son los siguientes :

  Por defecto en la calculadora aparecerá un 95 %.

Nivel de confianza	α	z_α
90 %	0,1	1,645
95 %	0,05	1,960
97'5 %	0,025	2,240
99 %	0,01	2,576

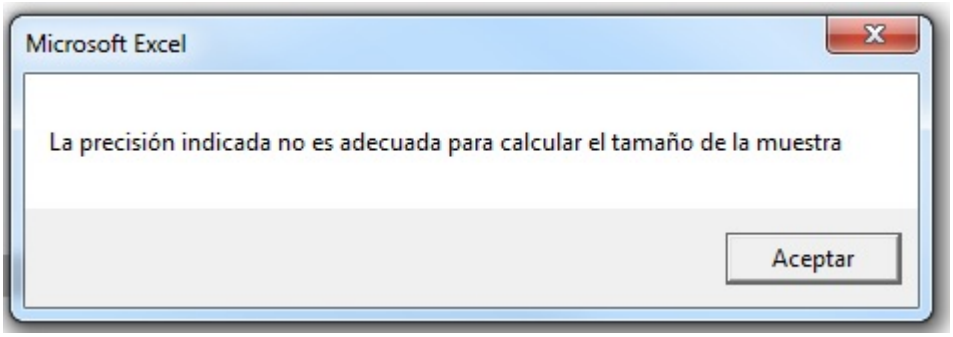
Cuadro A.1: Tabla de valores del nivel de confianza

Si el usuario desea modificar el nivel de confianza encontrará un desplegable con las distintas probabilidades indicadas anteriormente.

- Proporción esperada (p): es una idea del valor aproximado del parámetro que se quiere medir (en este caso una proporción). Esta idea se puede obtener revisando la literatura o mediante estudio pilotos previos. En caso de no tener dicha información, se aconseja utilizar el valor $p = 50\,\%$ ya que este valor maximiza el tamaño muestral. El usuario encontrará dos opciones a elegir, o bien introducir la proporción mediante una barra de desplazamiento, o bien utilizar el valor 50 % que se inserta automáticamente al indicar que no sé conoce la proporción.

- Precisión deseada (d): precisión que se desee para el estudio, no se recomienda utilizar un valor superior al 10 % ya que no es adecuada para el cálculo del tamaño muestral. Por defecto en la calculadora aparecerá un 5 %.

  El usuario podrá modifica este valor, escribiendo en la celda correspondiente la precisión deseada, en caso de introducir un número mayor que 10 % aparecerá el siguiente aviso:



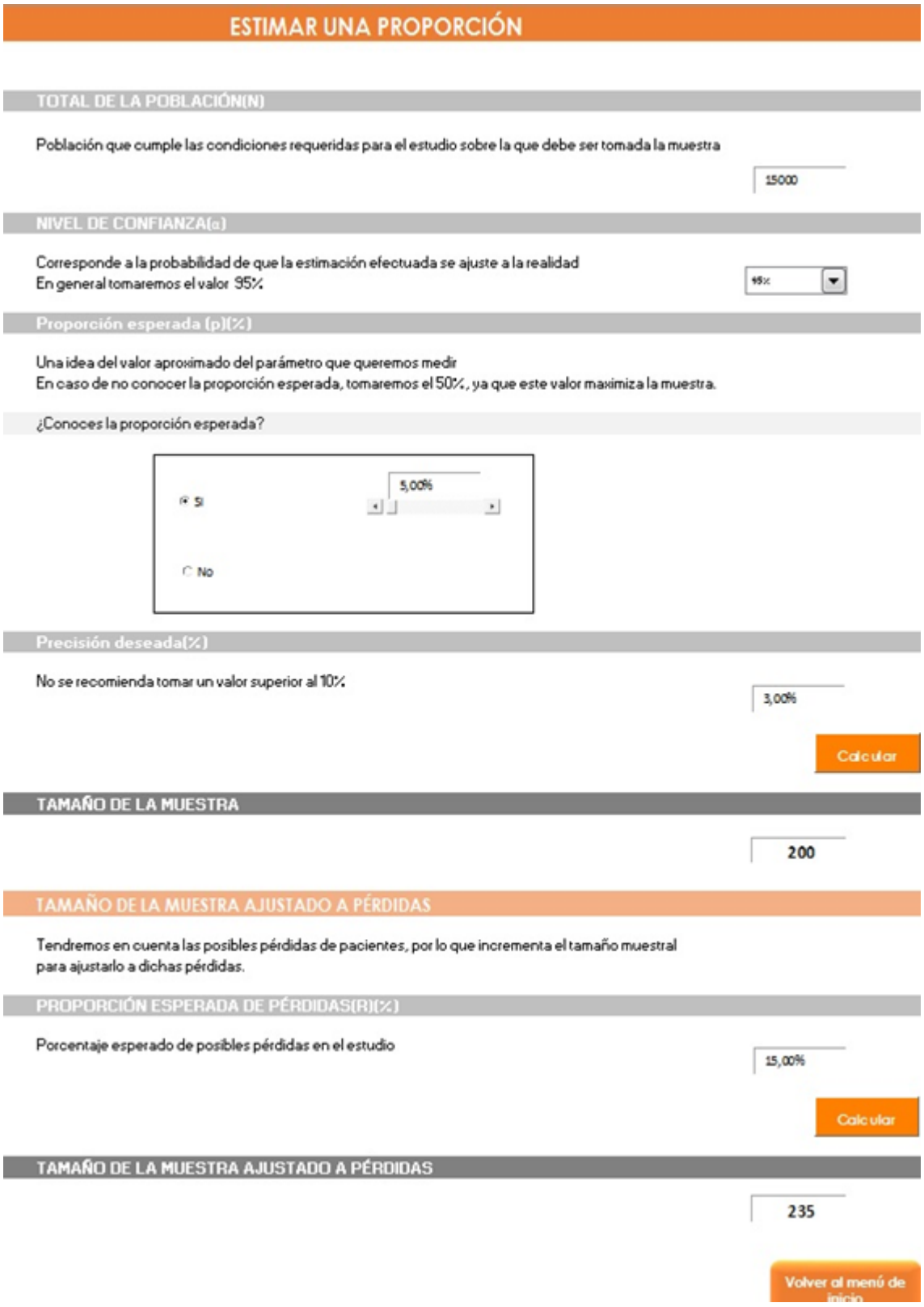


Figura A.2: Pestaña 'Estimar una proporción'

Cálculo del tamaño muestral

Una vez cumplimentados todos los datos requeridos el usuario debe pulsar el botón *Calcular*, que le devolverá el cálculo realizado tras la aplicación de la siguiente fórmula en la celda correspondiente al tamaño muestra :

$$n = \frac{N * z_\alpha^2 * p * q}{d^2 * (N-1) + z_\alpha^2 * p * q}$$

Para realizar este cálculo tomaremos $p/100$ y $d/100$ ya que son introducidos en las celdas en forma de porcentaje y $q = 1 - p$.

Tras calcular el tamaño muestral, la calculadora ofrece la posibilidad de ajustar el tamaño muestral a las posibles pérdidas de pacientes por razones diversas (pérdida de información, abandono, no respuesta. . .) por lo que se debe incrementar el tamaño muestral respecto a dichas pérdidas.

Para realizar este cálculo el usuario debe introducir:

- Proporción esperada de pérdidas(R): porcentaje esperado de posibles pérdidas en el estudio.

El usuario introducirá el valor del porcentaje en la celda mediante una barra de desplazamiento.

Una vez cumplimentado este dato pulsando el botón Calcular le devolverá el cálculo realizado mediante la fórmula:

$$n * \left(\frac{1}{(1-R)}\right)$$

Donde n es el tamaño muestral calculado previamente.

Exportar los datos a un documento Word

En esta pestaña el usuario encontrará el botón Exportar datos a un documento Word, en la parte inferior. Tras pulsarlo, se genera automáticamente un documento (Figura A.3) cumplimentado con los datos introducidos en la pestaña de Excel.

Tamaño muestral para estimar una proporción

Tabla1. Estimación del tamaño muestral

Parámetros	Valores
Total de la población	15000
Nivel de confianza	95,00%
Proporción esperada	5,00%
Precisión deseada	3,00%
Tamaño muestral	200
Proporción de pérdidas	15,00%
Tamaño muestral ajustado a pérdidas	235

Figura A.3: Documento Word del tamaño muestral para estimar una proporción

Finalmente, el usuario encontrará en la parte inferior el botón *Volver al menú de inicio*, que le devolverá a la pestaña *Introducción*.

A.3.3. Tamaño muestral para estimar una media

Definición e introducción de los parámetros

La función Tamaño muestral para estimar una media (Figura A.4) permite calcular el tamaño de la muestra requerido para realizar un estudio cuyo objetivo principal sea la estimación de una media. Los parámetros que intervienen en este cálculo son los que se mencionan a continuación:

- Total de la población (N): es la población que cumple las condiciones requeridas para el estudio sobre la que debe ser tomada la muestra.

- Nivel de confianza o seguridad (α): corresponde a la probabilidad de que la estimación efectuada se ajuste a la realidad. Este nivel de confianza da lugar a un coeficiente z_α. Las probabilidades entre las cuales va a poder elegir el usuario y sus correspondientes coeficientes son los siguientes :

Nivel de confianza	α	z_α
90 %	0,1	1,645
95 %	0,05	1,960
97'5 %	0,025	2,240
99 %	0,01	2,576

Cuadro A.2: Tabla de valores del nivel de confianza

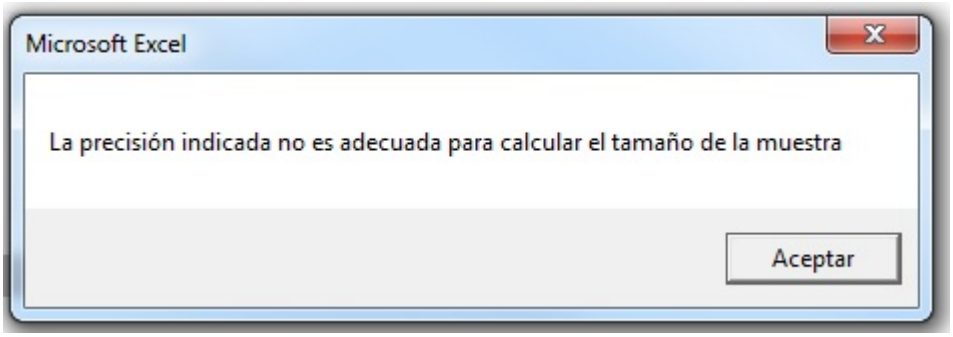
Por defecto en la calculadora aparecerá un 95 %.

Si el usuario desea modificar el nivel de confianza encontrará un desplegable con las distintas probabilidades indicadas anteriormente.

- Desviación (σ): es una idea del valor de la desviación de la distribución de la variable cuantitativa que se supone existe en la población.

- Precisión deseada (d): precisión que se desee para el estudio, no se recomienda utilizar un valor superior al 10 % ya que no es adecuada para el cálculo del tamaño muestral. Por defecto en la calculadora aparecerá un 5 %.

  El usuario podrá modifica este valor, escribiendo en la celda correspondiente la precisión deseada, en caso de introducir un número mayor que 10 % aparecerá el siguiente aviso:



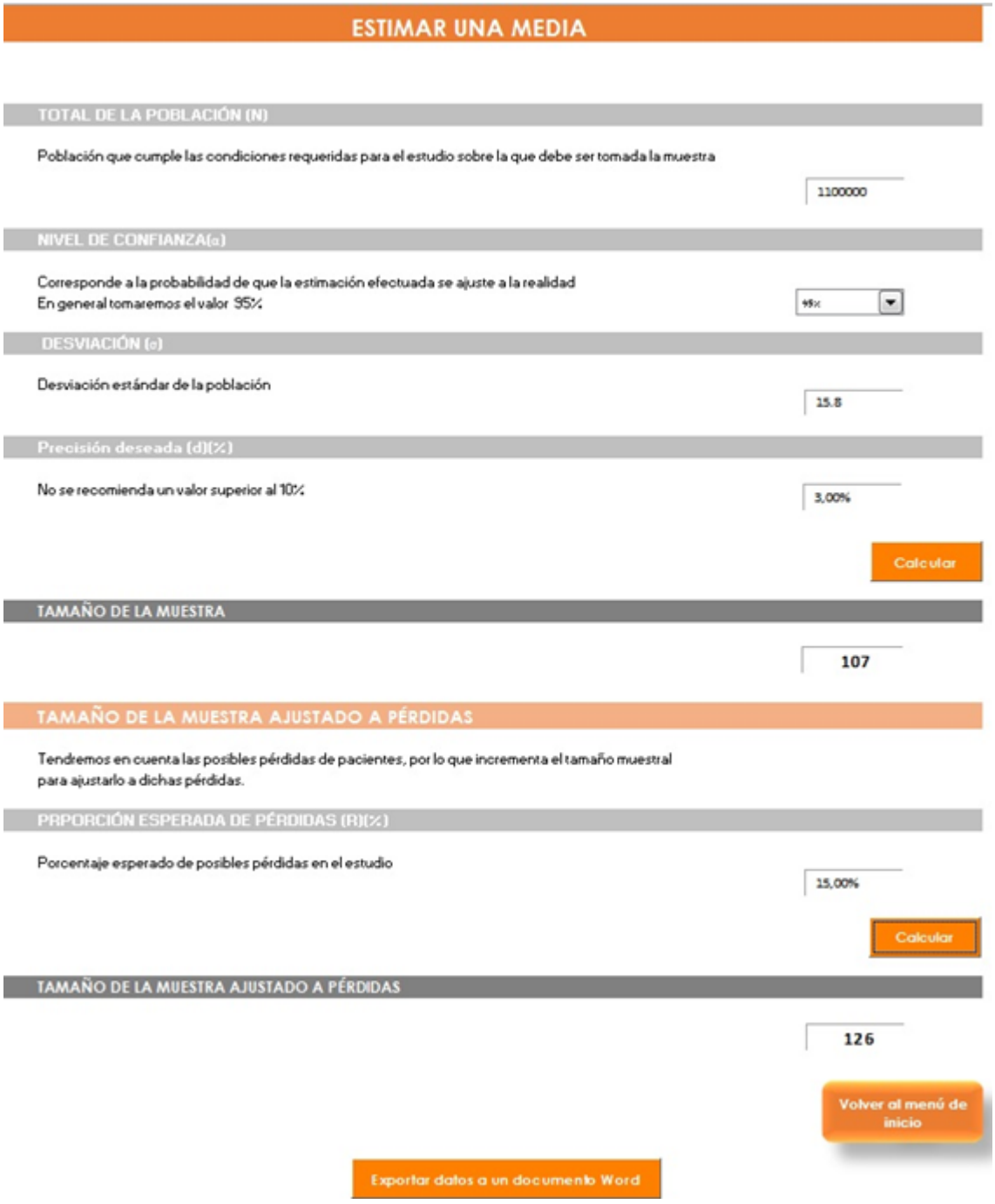


Figura A.4: Pestaña 'Estimar una media'

Cálculo del tamaño muestral

Una vez cumplimentados todos los datos requeridos el usuario debe pulsar el botón Calcular, que le devolverá el cálculo realizado mediante la fórmula:

$$n = \frac{N * z_{\alpha}^2 * \sigma^2}{d^2 * (N-1) + z_{\alpha}^2 * \sigma^2}$$

en la celda correspondiente al tamaño muestral.

Tras calcular el tamaño muestral, la calculadora ofrece la posibilidad de ajustar el tamaño muestral a las posibles pérdidas de pacientes por razones diversas (pérdida de información, abandono, no respuesta...) por lo que se debe incrementar el tamaño muestral respecto a dichas pérdidas.

Para realizar este cálculo el usuario debe introducir:

- Proporción esperada de pérdidas(R): porcentaje esperado de posibles pérdidas en el estudio.

El usuario introducirá el valor del porcentaje en la celda mediante una barra de desplazamiento.

Una vez cumplimentado este dato pulsando el botón Calcular le devolverá el cálculo realizado mediante la fórmula:

$$n * \left(\frac{1}{(1-R)} \right)$$

Donde n es el tamaño muestral calculado previamente.

Exportar los datos a un documento Word

En esta pestaña el usuario encontrará el botón Exportar datos a un documento Word, en la parte inferior. Tras pulsarlo, se genera automáticamente un documento (Figura A.5) cumplimentado con los datos introducidos en la pestaña de Excel.

Tamaño muestral para estimar una media

Tabla1. Estimación del tamaño muestral

Parámetros	Valores
Total de la población	1100000
Nivel de confianza	95,00%
Desviación	15.8
Precisión deseada	3,00%
Tamaño muestral	107
Proporción de pérdidas	15,00%
Tamaño muestral ajustado a pérdidas	126

Figura A.5: Documento Word del tamaño muestral para estimar una media

Finalmente, el usuario encontrará en la parte inferior el botón *Volver al menú de inicio*, que le devolverá a la pestaña *Introducción*.

A.3.4. Tamaño muestral para la comparación de dos proporciones independientes

Definición e introducción de los parámetros

La función Tamaño muestral para la comparación de dos proporciones independientes (Figura A.6) permite calcular el tamaño de la muestra por grupo requerido para realizar un estudio cuyo objetivo principal sea la comparación de

dos proporciones independientes. Los parámetros que intervienen en este cálculo son los que se mencionan a continuación:

- Nivel de confianza o seguridad (α): corresponde a la probabilidad de que la estimación efectuada se ajuste a la realidad. Este nivel de confianza da lugar a un coeficiente z_α. Las probabilidades entre las cuales va a poder elegir el usuario y sus correspondientes coeficientes son los siguientes :

Nivel de confianza	α	z_α
90 %	0,1	1,645
95 %	0,05	1,960
97'5 %	0,025	2,240
99 %	0,01	2,576

Cuadro A.3: Tabla de valores del nivel de confianza

  Por defecto en la calculadora aparecerá un 95 %.

  Si el usuario desea modificar el nivel de confianza encontrará un desplegable con las distintas probabilidades indicadas anteriormente.

- Potencia estadística (β): La potencia estadística da lugar a un coeficiente z_β. Las distintas potencias estadísticas entes las cuales va a poder elegir el usuario y sus correspondientes coeficientes son las siguiente:

Potencia estadística	β	z_β
80 %	0,2	0,842
85 %	0,15	1,036
90 %	0,10	1,282
95 %	0,05	1,645
99 %	0,01	2,326

Cuadro A.4: Tabla de valores de la potencia estadística

  Por defecto en la calculadora aparecerá un 95 %. Si el usuario desea modificar la potencia estadística encontrará un desplegable con las distintas probabilidades indicadas anteriormente.

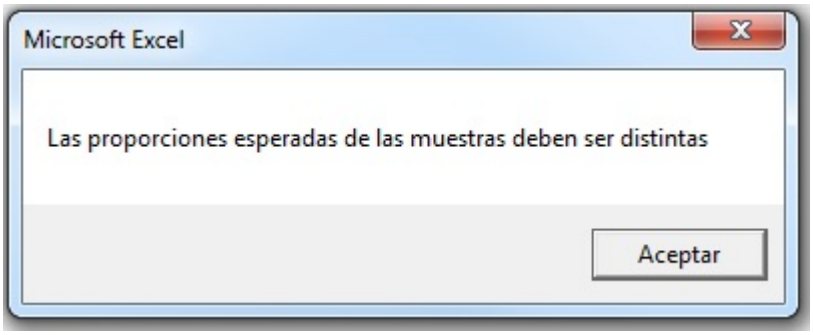
- Proporción esperada (p1): es el valor de la proporción en el grupo de referencia, control o tratamiento habitual.

  El usuario encontrará una celda acompañada por una barra de desplazamiento mediante la cual debe ajustar el valor deseado.

- Proporción esperada (p2): es el valor de la proporción en el grupo del nuevo tratamiento, intervención o técnica.

  El usuario encontrará una celda acompañada por una barra de desplazamiento mediante la cual debe ajustar el valor deseado.

Las dos proporciones descritas anteriormente deben ser valores distintos, de no serlo la calculadora mostrará el siguiente aviso:



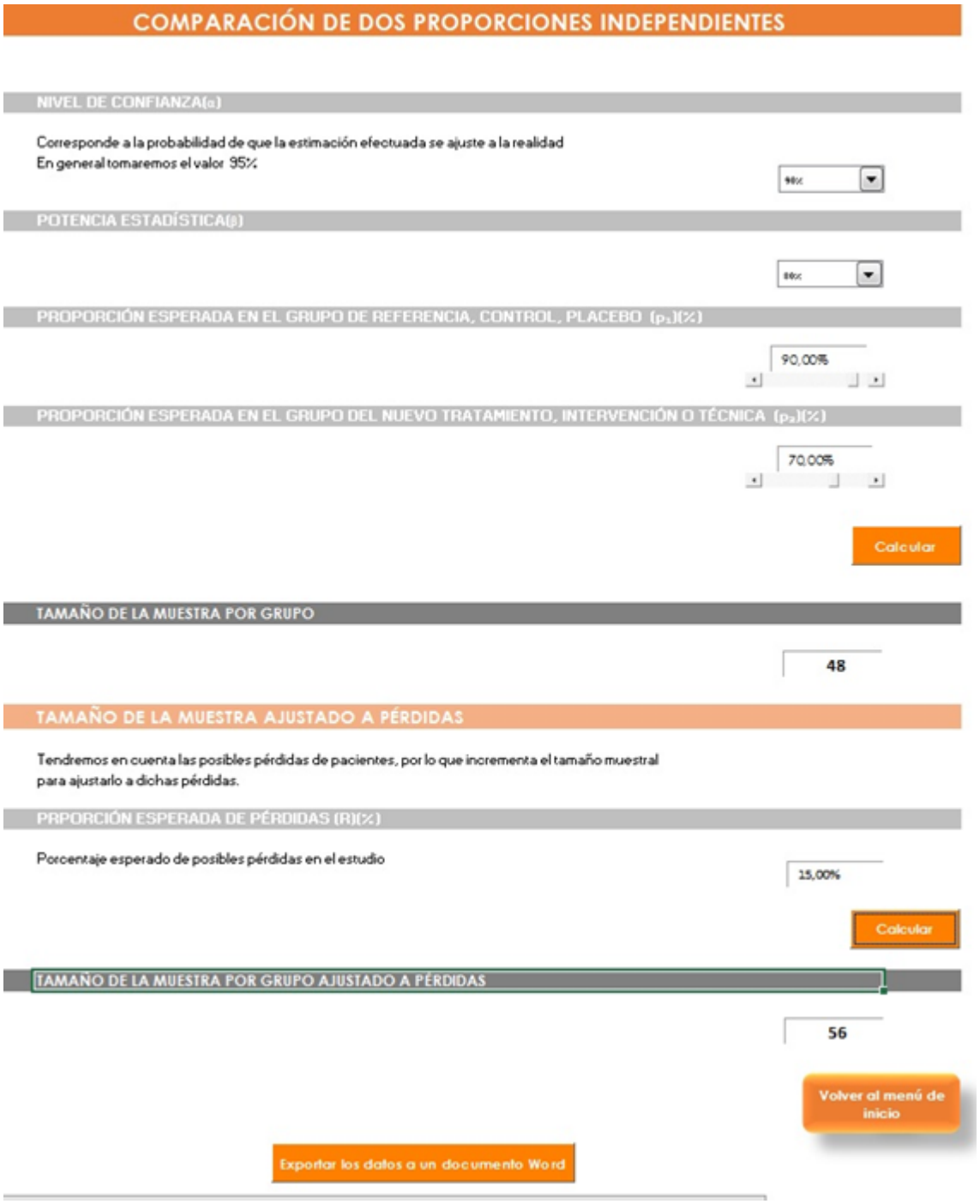


Figura A.6: Pestaña 'Comparación de dos proporciones independientes'

Cálculo del tamaño muestral

Una vez cumplimentados todos los datos requeridos el usuario debe pulsar el botón calcular, que le devolverá el cálculo realizado mediante la fórmula:

$$n = \frac{z_\alpha * \sqrt{(2p(1-p))} + z_\beta \sqrt{(p_1(1-p_1) + p_2(1-p_2))}}{(p_1 - p_2)^2}$$

en la celda correspondiente al tamaño muestral. Para realizar este cálculo tomaremos $p1/100$ y $p2/100$ ya que has sido introducidos en forma de porcentaje y $p = (p1 + p2)/2$ Tras calcular el tamaño muestral, la calculadora nos ofrece la posibilidad de ajustar el tamaño muestral a las posibles pérdidas de pacientes por razones diversas (pérdida de información, abandono, no respuesta...) por lo que se debe incrementar el tamaño muestral respecto a dichas pérdidas.

Para realizar este cálculo el usuario debe introducir:

- Proporción esperada de pérdidas(R): porcentaje esperado de posibles pérdidas en el estudio.

El usuario introducirá el valor del porcentaje en la celda mediante una barra de desplazamiento.

Una vez cumplimentado este dato pulsando el botón Calcular le devolverá el cálculo realizado mediante la fórmula:

$$n * \left(\frac{1}{(1-R)} \right)$$

Donde n es el tamaño muestral calculado previamente.

Exportar los datos a un documento Word

En esta pestaña el usuario encontrará el botón Exportar datos a un documento Word, en la parte inferior. Tras pulsarlo, se genera automáticamente un documento (Figura A.7) cumplimentado con los datos introducidos en la pestaña de Excel.

Tamaño muestral para la comparación de dos proporciones independientes

Tabla1. Estimación del tamaño muestral

Parámetros	Valores
Nivel de confianza	90,00%
Potencia estadística	80,00%
Proporción esperada en el grupo de referencia o control	90,00%
Proporción esperada en el grupo del nuevo tratamiento	70,00%
Tamaño muestral por grupo	48
Proporción de pérdidas	15,00%
Tamaño muestral por grupo ajustado a pérdidas	56

Figura A.7: Documento Word del tamaño muestral para la comparación de dos proporciones independientes

Finalmente, el usuario encontrará en la parte inferior el botón *Volver al menú de inicio*, que le devolverá a la pestaña *Introducción*.

A.3.5. Tamaño muestral para la comparación de una proporción observada con una población de referencia

Definición e introducción de los parámetros

La función Tamaño muestral para la comparación de una proporción observada con una población de referencia (Figura A.8) permite calcular el tamaño de la muestra por grupo requerido para realizar un estudio cuyo objetivo principal sea la comparación de una proporción observada con una población de referencia. En cuanto a la comparación de proporciones, el apartado A 3.4 es el más utilizado puesto que la opción más frecuente es aquella en la que una vez calculado el tamaño muestral por grupo, podemos realizar el estudio para ambos grupos. Para diferenciar este cálculo del realizado en el apartado anterior lo ilustraremos este apartado con un ejemplo. Supongamos que introducimos ciertos cambios en la unidad de cuidados intensivos de un hospital y queremos ver si realmente existe una mejoría respecto a la esperanza de vida de los pacientes. En este caso nuestros datos de la población de referencia serían los recogidos antes de realizar los cambios, por tanto una vez calculado el tamaño muestral necesario, este solo podría utilizarse para recoger los datos de los pacientes de la nueva unidad de cuidados intensivos, ya no sería posible recoger nuevamente datos de la unidad antigua. Los parámetros que intervienen en este cálculo son los siguientes:

- Nivel de confianza o seguridad (α): corresponde a la probabilidad de que la estimación efectuada se ajuste a la realidad. Este nivel de confianza da lugar a un coeficiente z_α. Las probabilidades entre las cuales va a poder elegir el usuario y sus correspondientes coeficientes son los siguientes :

Nivel de confianza	α	z_α
90 %	0,1	1,645
95 %	0,05	1,960
97′5 %	0,025	2,240
99 %	0,01	2,576

Cuadro A.5: Tabla de valores del nivel de confianza

  Por defecto en la calculadora aparecerá un 95 %.

  Si el usuario desea modificar el nivel de confianza encontrará un desplegable con las distintas probabilidades indicadas anteriormente.

- Potencia estadística (β): La potencia estadística da lugar a un coeficiente z_β. Las distintas potencias estadísticas entes las cuales va a poder elegir el usuario y sus correspondientes coeficientes son las siguiente:

  Por defecto en la calculadora aparecerá un 95 %. Si el usuario desea modificar la potencia estadística encontrará un desplegable con las distintas probabilidades indicadas anteriormente.

- Proporción esperada (p):): es el valor de la proporción esperada para la población de referencia.

  El usuario encontrará una celda acompañada por una barra de desplazamiento mediante la cual debe ajustar el valor deseado.

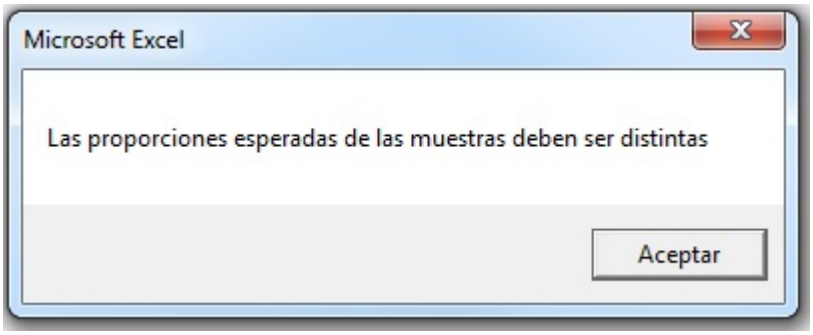
Potencia estadística	β	z_β
80 %	0, 2	0, 842
85 %	0, 15	1, 036
90 %	0, 10	1, 282
95 %	0, 05	1, 645
99 %	0, 01	2, 326

Cuadro A.6: Tabla de valores de la potencia estadística

- Proporción esperada (p_e): es el valor de la proporción en el grupo expuesto al nuevo tratamiento, intervención o técnica.

  El usuario encontrará una celda acompañada por una barra de desplazamiento mediante la cual debe ajustar el valor deseado.

  Las dos proporciones descritas anteriormente deben ser valores distintos, de no serlo la calculadora mostrará el siguiente aviso:



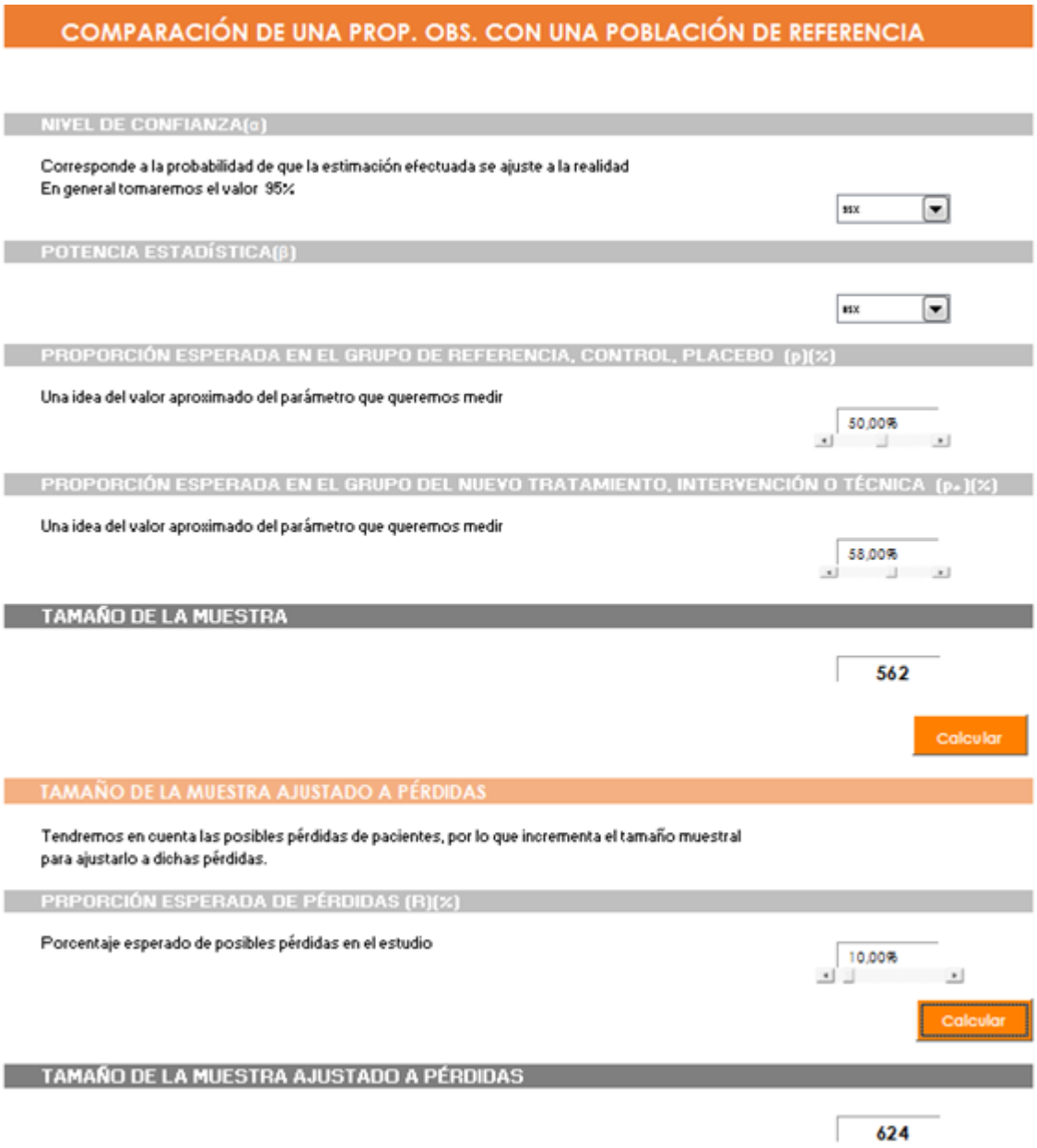


Figura A.8: Pestaña 'Comparación proporciones (2)'

Cálculo del tamaño muestral

Una vez cumplimentados todos los datos requeridos el usuario debe pulsar el botón calcular, que le devolverá el cálculo realizado mediante la fórmula :

$$n = \frac{z_\alpha * \sqrt{(2p(1-p)} + z_\beta \sqrt{(p_e(1-p_e))}}{(p-p_e)^2}$$

en la celda correspondiente al tamaño muestral. Para realizar este cálculo tomaremos $p/100$y $p_e/100$ ya que han sido introducidos en forma de porcentaje.

Tras calcular el tamaño muestral, la calculadora nos ofrece la posibilidad de ajustar el tamaño muestral a las posibles pérdidas de pacientes por razones diversas (pérdida de información, abandono, no respuesta. . .) por lo que se debe incrementar el tamaño muestral respecto a dichas pérdidas.

Para realizar este cálculo el usuario debe introducir:

- Proporción esperada de pérdidas(R): porcentaje esperado de posibles pérdidas en el estudio.

El usuario introducirá el valor del porcentaje en la celda mediante una barra de desplazamiento.

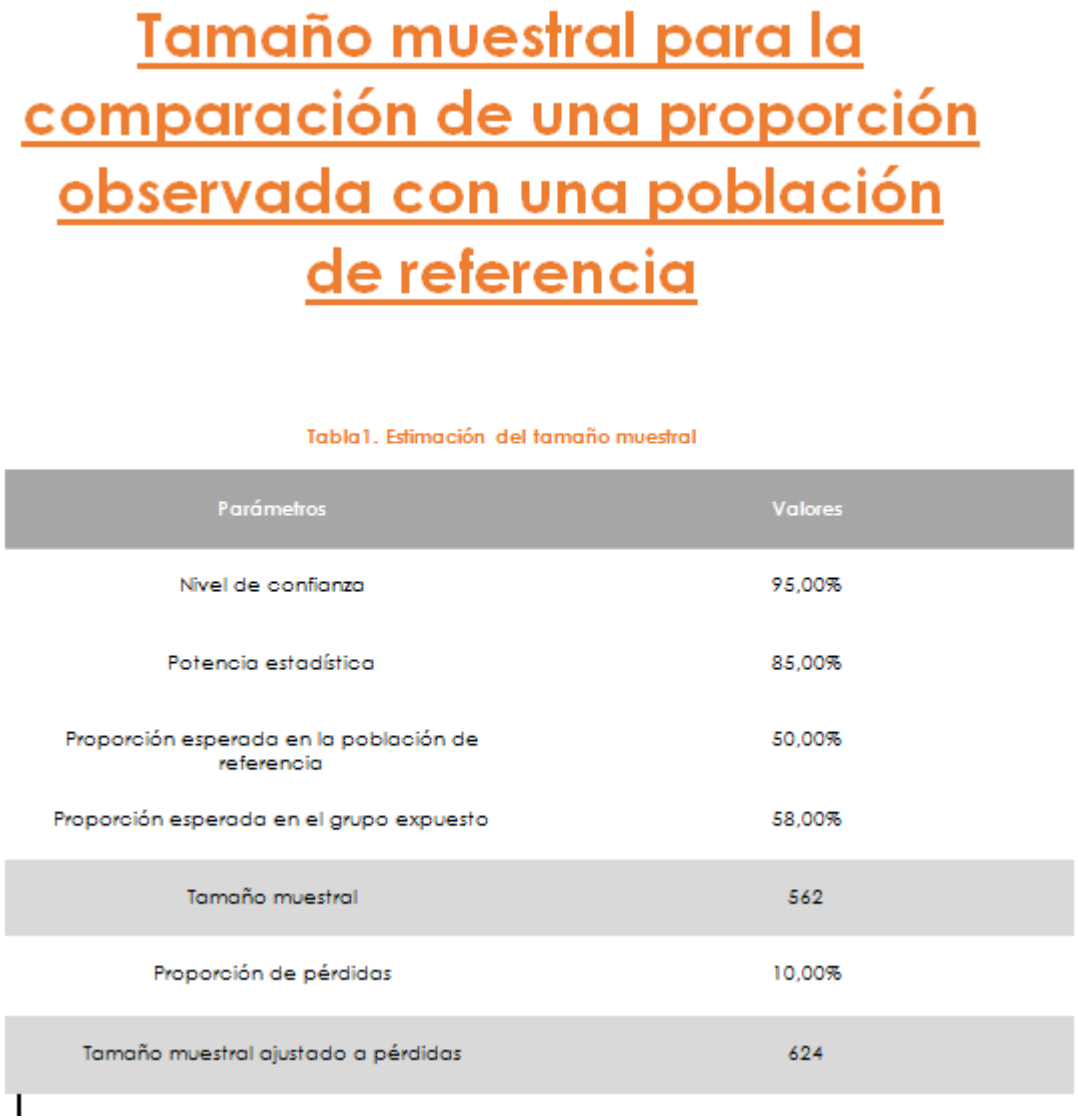
Una vez cumplimentado este dato pulsando el botón Calcular le devolverá el cálculo realizado mediante la fórmula:

$$n * \left(\frac{1}{(1-R)}\right)$$

Donde n es el tamaño muestral calculado previamente.

Exportar los datos a un documento Word

En esta pestaña el usuario encontrará el botón Exportar datos a un documento Word, en la parte inferior. Tras pulsarlo, se genera automáticamente un documento (Figura A.9) cumplimentado con los datos introducidos en la pestaña de Excel.



Tamaño muestral para la comparación de una proporción observada con una población de referencia

Tabla1. Estimación del tamaño muestral

Parámetros	Valores
Nivel de confianza	95,00%
Potencia estadística	85,00%
Proporción esperada en la población de referencia	50,00%
Proporción esperada en el grupo expuesto	58,00%
Tamaño muestral	562
Proporción de pérdidas	10,00%
Tamaño muestral ajustado a pérdidas	624

Figura A.9: Documento Word del tamaño muestral para la comparación de una proporción observada con una población de referencia

Finalmente, el usuario encontrará en la parte inferior el botón *Volver al menú de inicio*, que le devolverá a la pestaña *Introducción*.

A.3.6. Tamaño muestral para la comparación de dos medias independientes

Definición e introducción de los parámetros

La función Tamaño muestral para la comparación de dos medias independientes (Figura A.10) permite calcular el tamaño de la muestra por grupo requerido para realizar un estudio cuyo objetivo principal sea la comparación de dos medias independientes. Los parámetros que intervienen en este cálculo son los que se mencionan a continuación:

- Nivel de confianza o seguridad (α): corresponde a la probabilidad de que la estimación efectuada se ajuste a la realidad. Este nivel de confianza da lugar a un coeficiente z_α. Las probabilidades entre las cuales va a poder elegir el usuario y sus correspondientes coeficientes son los siguientes :

Nivel de confianza	α	z_α
90 %	0,1	1,645
95 %	0,05	1,960
97'5 %	0,025	2,240
99 %	0,01	2,576

Cuadro A.7: Tabla de valores del nivel de confianza

  Por defecto en la calculadora aparecerá un 95 %.

  Si el usuario desea modificar el nivel de confianza encontrará un desplegable con las distintas probabilidades indicadas anteriormente.

- Potencia estadística (β): La potencia estadística da lugar a un coeficiente z_β. Las distintas potencias estadísticas entes las cuales va a poder elegir el usuario y sus correspondientes coeficientes son las siguiente:

Potencia estadística	β	z_β
80 %	0,2	0,842
85 %	0,15	1,036
90 %	0,10	1,282
95 %	0,05	1,645
99 %	0,01	2,326

Cuadro A.8: Tabla de valores de la potencia estadística

  Por defecto en la calculadora aparecerá un 95 %. Si el usuario desea modificar la potencia estadística encontrará un desplegable con las distintas probabilidades indicadas anteriormente.

- Desviación (d): es la desviación de la variable cuantitativa que tiene el grupo de control, placebo o referencia. El usuario debe introducir el valor correspondiente en la celda habilitada para ello.

- Diferencia de medias (d): es el valor mínimo de la diferencia que se desea detectar. El usuario debe introducir el valor correspondiente en la celda habilitada para ello.

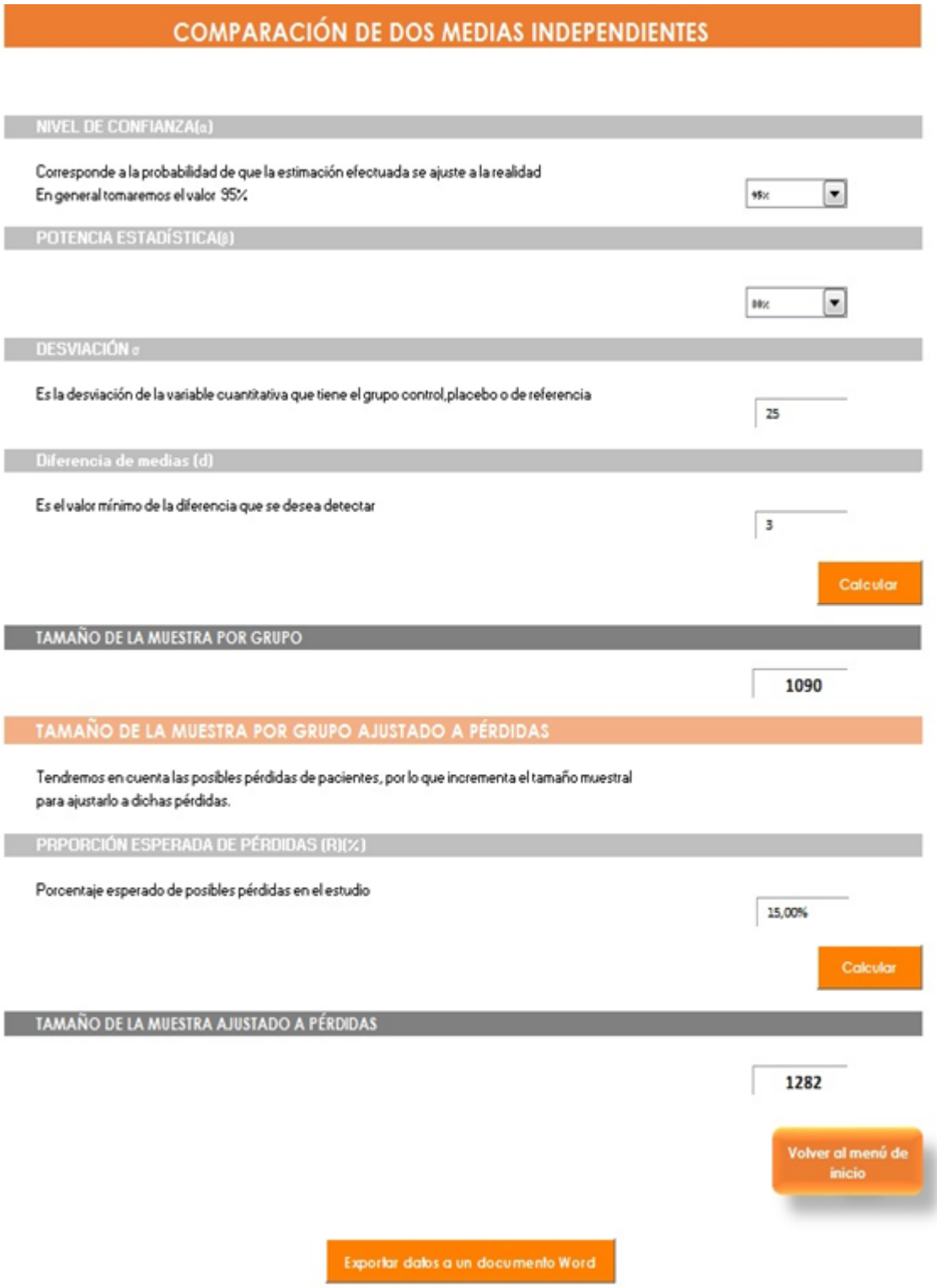


Figura A.10: Pestaña 'Comparación de dos medias independientes'

Cálculo del tamaño muestral

Una vez cumplimentados todos los datos requeridos el usuario debe pulsar el botón calcular, que le devolverá el cálculo realizado mediante la fórmula:

$$n = \frac{2(z_\alpha + z_\beta)^2\sigma^2}{d^2}$$

en la celda correspondiente al tamaño muestral.

Tras calcular el tamaño muestral, la calculadora nos ofrece la posibilidad de ajustar el tamaño muestral a las posibles pérdidas de pacientes por razones diversas (pérdida de información, abandono, no respuesta...) por lo que se debe incrementar el tamaño muestral respecto a dichas pérdidas.

Para realizar este cálculo el usuario debe introducir:

- Proporción esperada de pérdidas(R): porcentaje esperado de posibles pérdidas en el estudio.

El usuario introducirá el valor del porcentaje en la celda mediante una barra de desplazamiento.

Una vez cumplimentado este dato pulsando el botón Calcular le devolverá el cálculo realizado mediante la fórmula:

$$n * \left(\frac{1}{(1-R)}\right)$$

Donde n es el tamaño muestral calculado previamente.

Exportar los datos a un documento Word

En esta pestaña el usuario encontrará el botón Exportar datos a un documento Word, en la parte inferior. Tras pulsarlo, se genera automáticamente un documento (Figura A.11) cumplimentado con los datos introducidos en la pestaña de Excel.

Tamaño muestral para la comparación de dos medias independientes

Tabla1. Estimación del tamaño muestral

Parámetros	Valores
Nivel de confianza	95,00%
Potencia estadística	80,00%
Desviación dela variable cuantitativa del grupo de control	25
Diferencia de medias	3
Tamaño muestral por grupo	1090
Proporción de pérdidas	15,00%
Tamaño muestral por grupo ajustado a pérdidas	1282

Figura A.11: Documento Word del tamaño muestral para la comparación de dos medias independientes

Finalmente, el usuario encontrará en la parte inferior el botón *Volver al menú de inicio*, que le devolverá a la pestaña *Introducción*.

A.3.7. Tamaño muestral para la comparación de dos medias apareadas en un solo grupo

Definición e introducción de los parámetros

La función Tamaño muestral para la comparación de dos medias apareadas en un solo grupo (Figura A.12) permite calcular el tamaño de la muestra requerido para realizar un estudio cuyo objetivo principal sea la comparación de dos medias apareadas en un solo grupo. Los parámetros que intervienen en este cálculo son los que se mencionan a continuación:

- Nivel de confianza o seguridad (α): corresponde a la probabilidad de que la estimación efectuada se ajuste a la realidad. Este nivel de confianza da lugar a un coeficiente z_α. Las probabilidades entre las cuales va a poder elegir el usuario y sus correspondientes coeficientes son los siguientes :

Nivel de confianza	α	z_α
90 %	0,1	1,645
95 %	0,05	1,960
97'5 %	0,025	2,240
99 %	0,01	2,576

Cuadro A.9: Tabla de valores del nivel de confianza

Por defecto en la calculadora aparecerá un 95 %.

Si el usuario desea modificar el nivel de confianza encontrará un desplegable con las distintas probabilidades indicadas anteriormente.

- Potencia estadística (β): La potencia estadística da lugar a un coeficiente z_β. Las distintas potencias estadísticas entes las cuales va a poder elegir el usuario y sus correspondientes coeficientes son las siguiente:

Potencia estadística	β	z_β
80 %	0,2	0,842
85 %	0,15	1,036
90 %	0,10	1,282
95 %	0,05	1,645
99 %	0,01	2,326

Cuadro A.10: Tabla de valores de la potencia estadística

Por defecto en la calculadora aparecerá un 95 %. Si el usuario desea modificar la potencia estadística encontrará un desplegable con las distintas probabilidades indicadas anteriormente.

- Desviación (d): tomaremos la desviación basal o de inicio

  El usuario debe introducir el valor correspondiente en la celda habilitada para ello.

- Diferencia de medias (d): la media de las diferencias entre los valores basales y posteriores. (El usuario debe introducir el valor correspondiente en la celda habilitada para ello.)

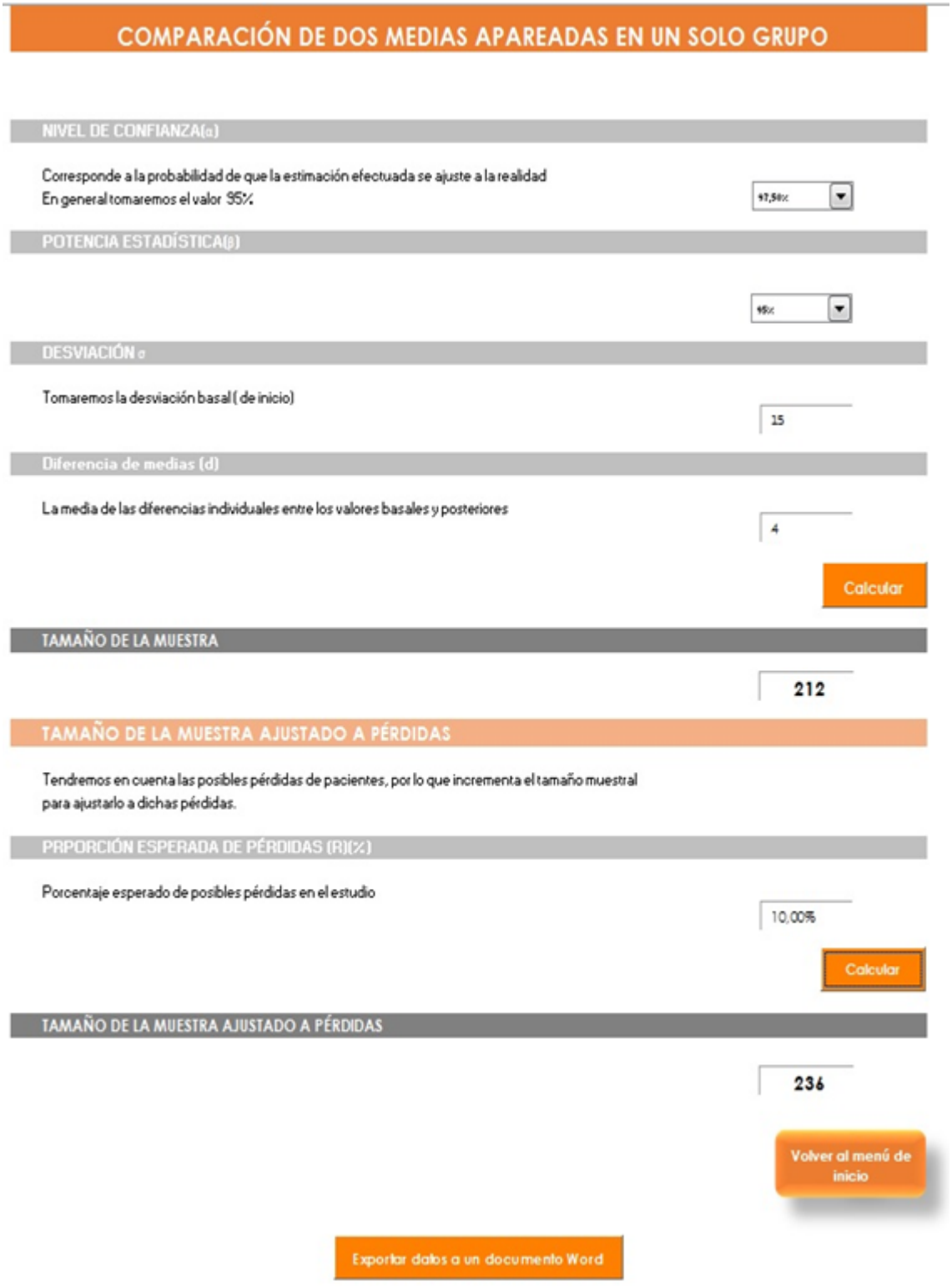


Figura A.12: Pestaña 'Medias apareadas un grupo'

Cálculo del tamaño muestral

Una vez cumplimentados todos los datos requeridos el usuario debe pulsar el botón calcular, que le devolverá el cálculo realizado mediante la fórmula:

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2)}{d^2}$$

en la celda correspondiente al tamaño muestral.

Tras calcular el tamaño muestral, la calculadora nos ofrece la posibilidad de ajustar el tamaño muestral a las posibles pérdidas de pacientes por razones diversas (pérdida de información, abandono, no respuesta...) por lo que se debe incrementar el tamaño muestral respecto a dichas pérdidas.

Para realizar este cálculo el usuario debe introducir:

- Proporción esperada de pérdidas(R): porcentaje esperado de posibles pérdidas en el estudio.

El usuario introducirá el valor del porcentaje en la celda mediante una barra de desplazamiento.

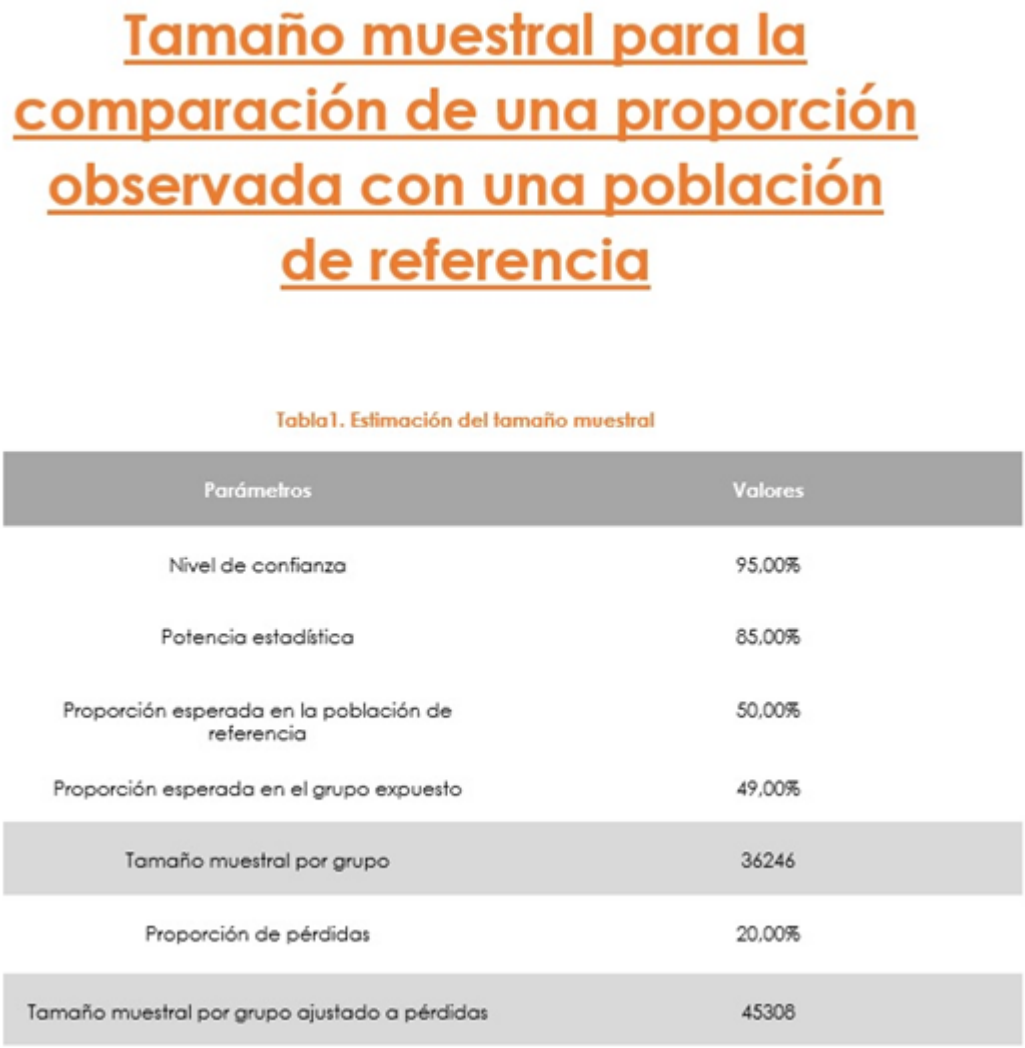
Una vez cumplimentado este dato pulsando el botón Calcular le devolverá el cálculo realizado mediante la fórmula:

$$n * \left(\frac{1}{(1 - R)} \right)$$

Donde n es el tamaño muestral calculado previamente.

Exportar los datos a un documento Word

En esta pestaña el usuario encontrará el botón Exportar datos a un documento Word, en la parte inferior. Tras pulsarlo, se genera automáticamente un documento (Figura A.13) cumplimentado con los datos introducidos en la pestaña de Excel.



Tamaño muestral para la comparación de una proporción observada con una población de referencia

Tabla1. Estimación del tamaño muestral

Parámetros	Valores
Nivel de confianza	95,00%
Potencia estadística	85,00%
Proporción esperada en la población de referencia	50,00%
Proporción esperada en el grupo expuesto	49,00%
Tamaño muestral por grupo	36246
Proporción de pérdidas	20,00%
Tamaño muestral por grupo ajustado a pérdidas	45308

Figura A.13: Documento Word del tamaño muestral para la comparación de dos medias apareadas en un solo grupo

Finalmente, el usuario encontrará en la parte inferior el botón *Volver al menú de inicio*, que le devolverá a la pestaña *Introducción*.

A.3.8. Tamaño muestral para la comparación de dos medias apareadas en dos grupos

Definición e introducción de los parámetros

La función Tamaño muestral para la comparación de dos medias apareadas en dos grupos (Figura A.14) permite calcular el tamaño de la muestra por grupo requerido para realizar un estudio cuyo objetivo principal sea la comparación de dos medias apareadas en dos grupos. Los parámetros que intervienen en este cálculo son los que se mencionan a continuación:

- Nivel de confianza o seguridad (α): corresponde a la probabilidad de que la estimación efectuada se ajuste a la realidad. Este nivel de confianza da lugar a un coeficiente z_α. Las probabilidades entre las cuales va a poder elegir el usuario y sus correspondientes coeficientes son los siguientes :

Nivel de confianza	α	z_α
90 %	0,1	1,645
95 %	0,05	1,960
97'5 %	0,025	2,240
99 %	0,01	2,576

Cuadro A.11: Tabla de valores del nivel de confianza

Por defecto en la calculadora aparecerá un 95 %.

Si el usuario desea modificar el nivel de confianza encontrará un desplegable con las distintas probabilidades indicadas anteriormente.

- Potencia estadística (β): La potencia estadística da lugar a un coeficiente z_β. Las distintas potencias estadísticas entes las cuales va a poder elegir el usuario y sus correspondientes coeficientes son las siguiente:

Potencia estadística	β	z_β
80 %	0,2	0,842
85 %	0,15	1,036
90 %	0,10	1,282
95 %	0,05	1,645
99 %	0,01	2,326

Cuadro A.12: Tabla de valores de la potencia estadística

Por defecto en la calculadora aparecerá un 95 %. Si el usuario desea modificar la potencia estadística encontrará un desplegable con las distintas probabilidades indicadas anteriormente.

- Desviación (d):tomaremos la desviación basal o de inicio.

  El usuario debe introducir el valor correspondiente en la celda habilitada para ello.

- Diferencia de medias (d): la media de las diferencias entre los valores basales y posteriores.

  El usuario debe introducir el valor correspondiente en la celda habilitada para ello.

- Coeficiente de correlación (ρ): es el coeficiente de correlación entre la media basal y la final. Debe tomar valores entre -1 y 1.

  El usuario podrá modifica este valor, escribiendo en la celda correspondiente el coeficiente de correlación en caso de introducir un número que no se encuentre en el intervalo indicado aparecerá el siguiente aviso:

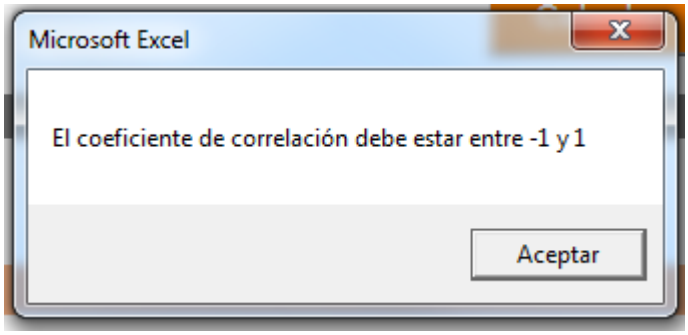


Figura A.14: Pestaña 'Medias apareadas dos grupos'

Cálculo del tamaño muestral

Una vez cumplimentados todos los datos requeridos el usuario debe pulsar el botón calcular, que le devolverá el cálculo realizado mediante la fórmula:

$$n = \frac{2(z_\alpha + z_\beta)^2(1 - p\rho)\sigma^2}{d^2}$$

en la celda correspondiente al tamaño muestral.

Tras calcular el tamaño muestral, la calculadora nos ofrece la posibilidad de ajustar el tamaño muestral a las posibles pérdidas de pacientes por razones diversas (pérdida de información, abandono, no respuesta...) por lo que se debe incrementar el tamaño muestral respecto a dichas pérdidas.

Para realizar este cálculo el usuario debe introducir:

- Proporción esperada de pérdidas(R): porcentaje esperado de posibles pérdidas en el estudio.

El usuario introducirá el valor del porcentaje en la celda mediante una barra de desplazamiento.

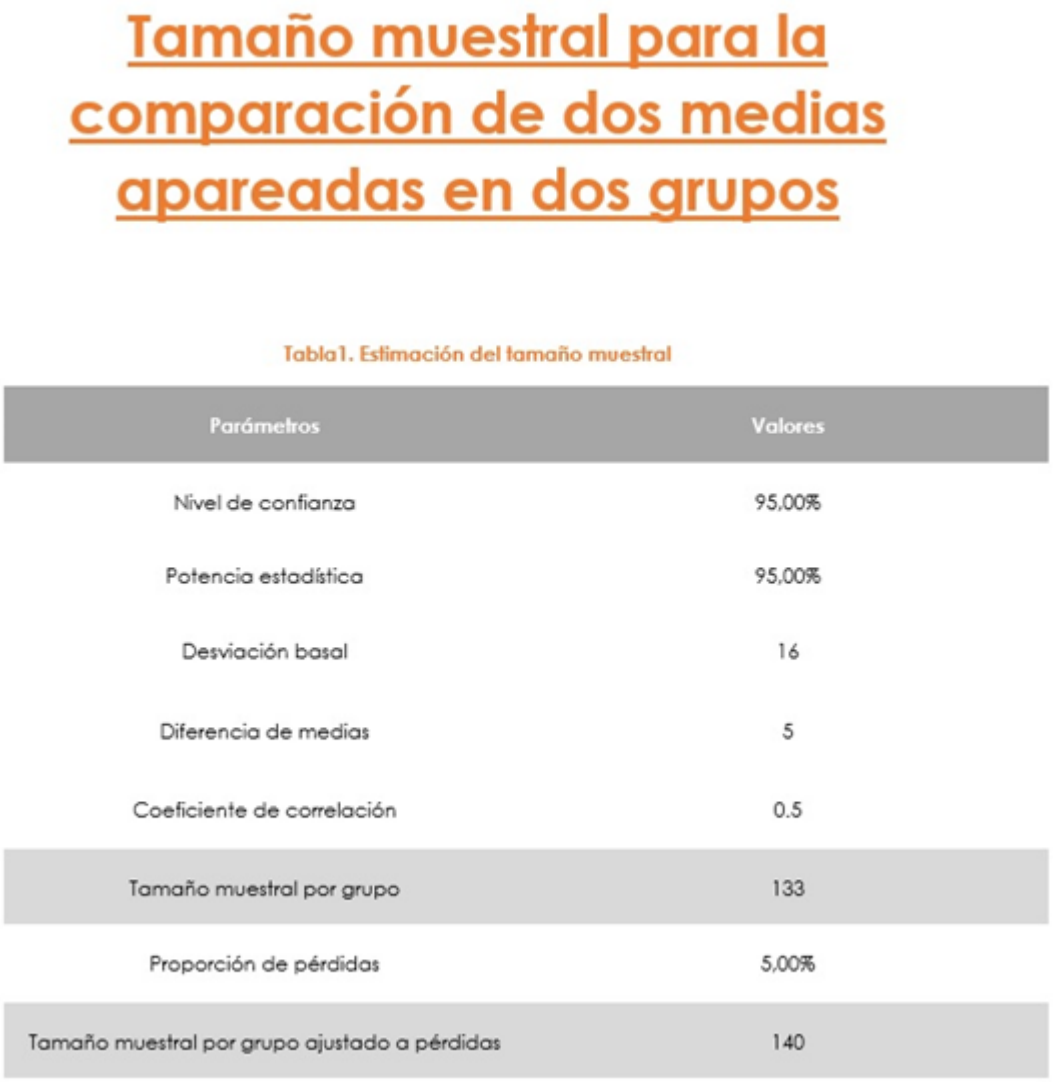
Una vez cumplimentado este dato pulsando el botón Calcular le devolverá el cálculo realizado mediante la fórmula:

$$n * \left(\frac{1}{(1 - R)}\right)$$

Donde n es el tamaño muestral calculado previamente.

Exportar los datos a un documento Word

En esta pestaña el usuario encontrará el botón Exportar datos a un documento Word, en la parte inferior. Tras pulsarlo, se genera automáticamente un documento (Figura A.15) cumplimentado con los datos introducidos en la pestaña de Excel.



Tamaño muestral para la comparación de dos medias apareadas en dos grupos

Tabla1. Estimación del tamaño muestral

Parámetros	Valores
Nivel de confianza	95,00%
Potencia estadística	95,00%
Desviación basal	16
Diferencia de medias	5
Coeficiente de correlación	0.5
Tamaño muestral por grupo	133
Proporción de pérdidas	5,00%
Tamaño muestral por grupo ajustado a pérdidas	140

Figura A.15: Documento Word del tamaño muestral para la comparación de dos medias apareadas en dos grupos

Finalmente, el usuario encontrará en la parte inferior el botón *Volver al menú de inicio*, que le devolverá a la pestaña *Introducción*.