Tema 3. ESTIMACIÓN Y CONTRASTE DE HIPÓTESIS.

3.1. INTRODUCCIÓN

Dentro del ámbito de problemas que se contemplan en la Gestión Costera, es frecuente encontrarse con la cuestión de cuál es el valor de los parámetros que caracterizan aquello que se está estudiando: concentración de contaminantes en un vertido; número de turistas que visitan anualmente un *resort;* tráfico de mercancías en un puerto; riesgo de ocurrencia de una catástrofe; etc. En general podemos suponer que las variables que se observan pueden describirse mediante una distribución de probabilidad conocida (*normal, binomial,* etc.), debiendo darse respuesta a la cuestión de cuáles son los valores de los parámetros que definen dicha distribución.

En este contexto, la *población* que se considera es en general el conjunto de *todos* los sujetos u objetos sobre los que interesa medir la variable en estudio. Una *muestra* de tamaño *n* de esta población es un conjunto de *n* observaciones escogidas al azar de la población.

Para *estimar un parámetro* es preciso disponer de un *estimador*. Un estimador es una función que, evaluada sobre la muestra, proporciona valores que se aproximan de algún modo al parámetro desconocido. El estimador es *puntual* si nos proporciona un único valor para el parámetro desconocido; el estimador es *por intervalo* si nos proporciona un intervalo dentro del cual podemos confiar en que se encuentra el parámetro. Es importante observar que *un estimador es una variable aleatoria*, ya que en cada muestra el estimador tomará un valor que puede ser distinto del que toma en otras muestras diferentes. No debe confundirse la función (*estimador*) con el valor que toma en cada muestra (*valor estimado*).

3.2. INFERENCIA ESTADÍSTICA: ESTIMACIÓN Y CONTRASTE DE HIPÓTESIS

Puede definirse la *inferencia estadística* como la colección de métodos existentes para extender o generalizar a una población las conclusiones o resultados obtenidos a partir de la información proporcionada por una muestra de la misma. Estas conclusiones sobre la población pueden ser de naturaleza diferente:

- Si se dispone de algún modelo del comportamiento de las variables en la población, los datos recogidos en la muestra se utilizarán entonces para obtener valores aproximados de los parámetros que caracterizan dicho modelo *(Problemas de estimación)*

- Si en el curso de una investigación se ha elaborado alguna hipótesis sobre la población, el objetivo del muestreo será contrastar si los datos obtenidos confirman dicha hipótesis o si, por el contrario obligan a descartarla *(Problemas de contraste de hipótesis)*.

Dado que el tamaño de la muestra es, en general, pequeño en relación con el tamaño de la población, *las conclusiones que se extraigan no podrán ser nunca seguras al 100%.* No obstante, los métodos de inferencia estadística permiten cuantificar, en términos de probabilidad, el grado de verosimilitud o certidumbre que podemos asignar a nuestras conclusiones.

Para que la información proporcionada por la muestra pueda emplearse aceptablemente para obtener conclusiones sobre la población *es necesario que la muestra sea representativa y que tenga un tamaño*

suficiente (en general, cuanto mayor sea la muestra más información proporcionará). El tamaño de muestra depende de cuál sea el problema que nos planteamos (estimación de parámetros o contraste de hipótesis), de las características de la población (en general, a mayor heterogeneidad de la población con respecto a la variable de interés, mayor habrá de ser el tamaño de la muestra) y de la magnitud de los errores que estamos dispuestos a cometer. La representatividad de la muestra se consigue cuando esta es aleatoria (todos los miembros de la población tienen la misma probabilidad de ser elegidos para formar parte de la muestra).

Supongamos que el objetivo de la investigación es estudiar cierta característica numérica X en una población (por ejemplo la concentración media de un contaminante en una franja costera) y que con este fin se elige una muestra aleatoria de objetos de la misma (los valores de concentración medidos en los puntos de una malla definida sobre la franja costera en estudio; aunque los puntos de la malla sean fijos, los valores de concentración del contaminante en dichos puntos son aleatorios, no se conocen mientras no sean medidos). Dado que, a priori, no se conocen los valores que tomará X en dichos objetos, X es una variable aleatoria y como tal, tendrá una función de distribución F. Existen entonces dos posibilidades:

(1) Se conoce la expresión funcional de F, pero no se conocen los valores de los parámetros que la caracterizan, y que denotaremos por $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$. Esto es lo que sucede si se sabe (o se sospecha) que los datos proceden, por ejemplo, de una distribución exponencial (de la que no conocemos el valor del parámetro λ) o de una Normal (de la que no sabemos lo que valen μ y σ).

(2) No se sabe nada de F salvo, quizás, si es continua o escalonada.

En el primer caso nos encontramos en un problema de *inferencia paramétrica.* Cualquier afirmación, en términos de probabilidad, sobre las características de X requiere conocer, aunque sea de modo aproximado, el valor del parámetro θ. El segundo caso corresponde a un problema de *inferencia no paramétrica.*

Claramente el primer problema de la inferencia paramétrica es *la obtención de valores aproximados de los parámetros (estimación).* Dos son los modos de estimación clásicos: la *estimación puntual* y la *estimación por intervalos*. En el primer caso se trata de obtener un valor aproximado del parámetro poblacional θ desconocido, sin indicar el grado de aproximación. En el segundo se obtiene un intervalo dentro del cual podemos tener cierta confianza en encontrar el valor de θ. La estimación por intervalos sí que proporciona una idea de la precisión conseguida: ya no se afirma: *"la concentración media del contaminante en la zona de estudio es un valor próximo a 57 ppm"* (estimación puntual), sino que *"con una confianza del 99% la concentración media del contaminante está en el intervalo [51,63]"*.

Un problema importante es la selección de la distribución que mejor modela la variable que estamos observando. Si se encuentra una distribución de probabilidad que representa bien los datos, se pueden aplicar métodos de inferencia paramétricos basados en dicha distribución.

Ejemplo: se desea estimar el tiempo medio μ que requiere una colonia bacteriana en alcanzar su volumen máximo a partir de un volumen inicial V_0. El procedimiento habitual consistirá en tomar una muestra aleatoria de n colonias de volumen inicial V_0, medir los tiempos $X_1, X_2, \ldots, X_n$, que tarda cada una en alcanzar su volumen máximo, calcular su media:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

y utilizar el valor $\overline{X}$ así obtenido como aproximación del verdadero valor de μ. Diremos en este caso que la media muestral es un estimador puntual de la media poblacional. Si en la muestra que finalmente hemos conseguido, la media muestral resulta ser de 1268 minutos, éste será el valor estimado de la media poblacional.

En este ejemplo, y en el problema general de estimación, podemos realizar varias observaciones importantes:

1. *La estimación se realiza a través de una función (estimador).*
   Como acabamos de ver en el ejemplo, un valor aproximado de la media de la población se obtiene mediante el cálculo de la media muestral que es, de hecho, una función de la muestra consistente en sumar todos sus valores y dividir por el número total de los mismos.

2. *La función (estimador) que se utiliza para estimar un parámetro es una variable aleatoria.*
   En nuestro ejemplo es evidente que hasta que no se obtenga una muestra de colonias bacterianas y se mida el tiempo que necesita cada una para alcanzar su volumen máximo, es imposible predecir con exactitud cuál va a ser la media de los tiempos observados. Por tanto *a priori,* antes de realizar el muestreo, la media muestral es una variable aleatoria.

3. *El valor estimado depende de la muestra elegida.*
   En el ejemplo anterior, es obvio que distintas muestras de colonias bacterianas darán lugar a distintas medias, y por tanto conducirán a distintas estimaciones de la media poblacional.

4. *¿Cómo obtener una función cuyo valor sea próximo al del parámetro que se pretende estimar?*
   En nuestro ejemplo parece razonable estimar la media poblacional mediante la media muestral. Sin embargo, en otros casos en que haya que estimar un parámetro arbitrario, puede no ser tan evidente qué función debe utilizarse para llevar a cabo la estimación. De hecho, cabe incluso la posibilidad de que haya dos o más funciones distintas que permitan estimar el mismo parámetro.

5. *¿Qué significa que el valor del estimador se aproxime al del parámetro que se pretende estimar?*
   Tal como acabamos de señalar, la estimación se realiza a través de una función (estimador) que, a priori, puede tomar muchos valores, dependiendo de la muestra que resulte elegida. Evidentemente, no todos los valores están igual de próximos al parámetro que se pretende estimar: la media de la población es un valor fijo determinado; algunas muestras tendrán medias más cercanas a este valor y otras tendrán medias más alejadas. Por tanto ¿qué se está diciendo realmente cuando se afirma que la media muestral (que no tiene un único valor fijo, sino que puede tomar muchos valores distintos) se aproxima a la media poblacional?

6. *¿Cómo evaluar el grado de proximidad conseguido entre el estimador y el parámetro?*
   En nuestro ejemplo la media μ de la población es desconocida. Cuando tomamos una muestra obtenemos una media muestral cuyo valor suponemos próximo a μ. Pero ... ¿cuán próximo? Si la media muestral que obtenemos es de 1268 minutos, ¿significa ésto que la media poblacional es como mucho 1300 minutos? ¿o podría ser de 1450 minutos? ¿quizás 1600? ¿2000 minutos? Si no tenemos ninguna referencia sobre el grado de proximidad alcanzado, el estimador puntual puede ser perfectamente inútil.

3.3. ESTUDIO DE SIMULACIÓN

Para fijar ideas, supongamos que el verdadero valor de μ (tiempo medio que tarda la colonia de bacterias en alcanzar el volumen máximo) es de 1200 minutos. Tomamos una muestra de 25 colonias

de idéntico volumen inicial V_0 y las dejamos crecer hasta que dejan de hacerlo (alcanzan su volumen máximo), obteniéndose los siguientes tiempos en cada colonia (en minutos):

Tiempos hasta alcanzar el volumen V_f								
1100	1652	1501	1878	1357	1897	1898	1079	1562
925	1075	299	818	945	758	1050	876	1137
714	2041	986	589	1838	705	610		

El tiempo medio en estas 25 colonias ha resultado ser $\overline{x} = 1171,65$ minutos. Repetimos el experimento con otras 25 colonias, obteniendo como resultado:

Tiempos hasta alcanzar el volumen V_f								
153	399	532	885	877	2786	413	2394	855
1658	157	1791	392	2396	324	512	211	1256
592	1002	304	1494	2141	1908	1426		

En este caso se obtiene la estimación de la media es $\overline{x} = 1074,29$, diferente de la anterior, como cabía esperar.

Imaginemos ahora que repetimos este proceso 10.000 veces, esto es, elegimos 10.000 muestras de 25 colonias cada una, y calculamos el tiempo medio que tardan las colonias de cada muestra en dejar de crecer. De esta forma obtenemos 10.000 estimaciones de la media poblacional. Evidentemente esto no lo podemos hacer en la realidad, pero sí es sencillo simularlo en el ordenador. La figura 3.1 muestra el histograma de las 10.000 medias que se han obtenido simulando este proceso.

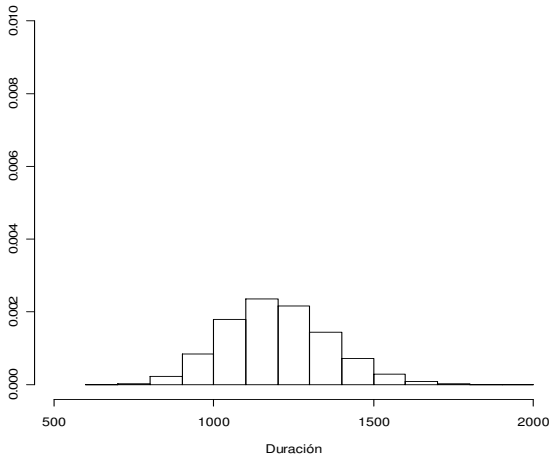


Figura 3.1: Histograma de los 10.000 tiempos medios obtenidos simulando 10.000 veces el crecimiento de 25 colonias bacterianas y observando cuánto tarda cada colonia en alcanzar el volumen máximo.

Como puede apreciarse, las 10.000 medias observadas oscilan alrededor de la verdadera media de la población (1200 minutos), habiéndose no obstante llegado a observar medias tan bajas como 642 minutos, o tan altas como 1884 minutos. Podemos medir la variabilidad de las medias obtenidas en todos estos muestreos mediante su desviación típica. En esta simulación, la media de todas las medias ha sido 1197,60 minutos y su desviación típica 162,76 minutos. Obsérvese en cualquier caso, como la media de todas las medias se acerca mucho a la media de la población (1200 minutos).

Imaginemos, por último, que repetimos este proceso de obtener 10.000 muestras, con muestras de tamaños 100, 250 y 500 colonias, calculando en cada muestra el tiempo medio que tardan las colonias

de bacterias en dejar de crecer. La figura 3.2 muestra los histogramas correspondientes a las 10.000 estimaciones de la media que se obtienen en cada caso.

Como puede apreciarse en el gráfico, cuanto mayor es el tamaño de las muestras, menor es la dispersión de sus medias. Así, las desviaciones típicas observadas han sido de 162,76 minutos para muestras de tamaño 25, de 81,45 minutos para muestras de tamaño 100, de 51,22 minutos para muestras de tamaño 250 y de 36,34 minutos para muestras de tamaño 500. Esto significa que a medida que crece el tamaño de la muestra, las medias muestrales observadas tienden a estar más agrupadas en torno a la media de la población, alejándose menos de dicha media.

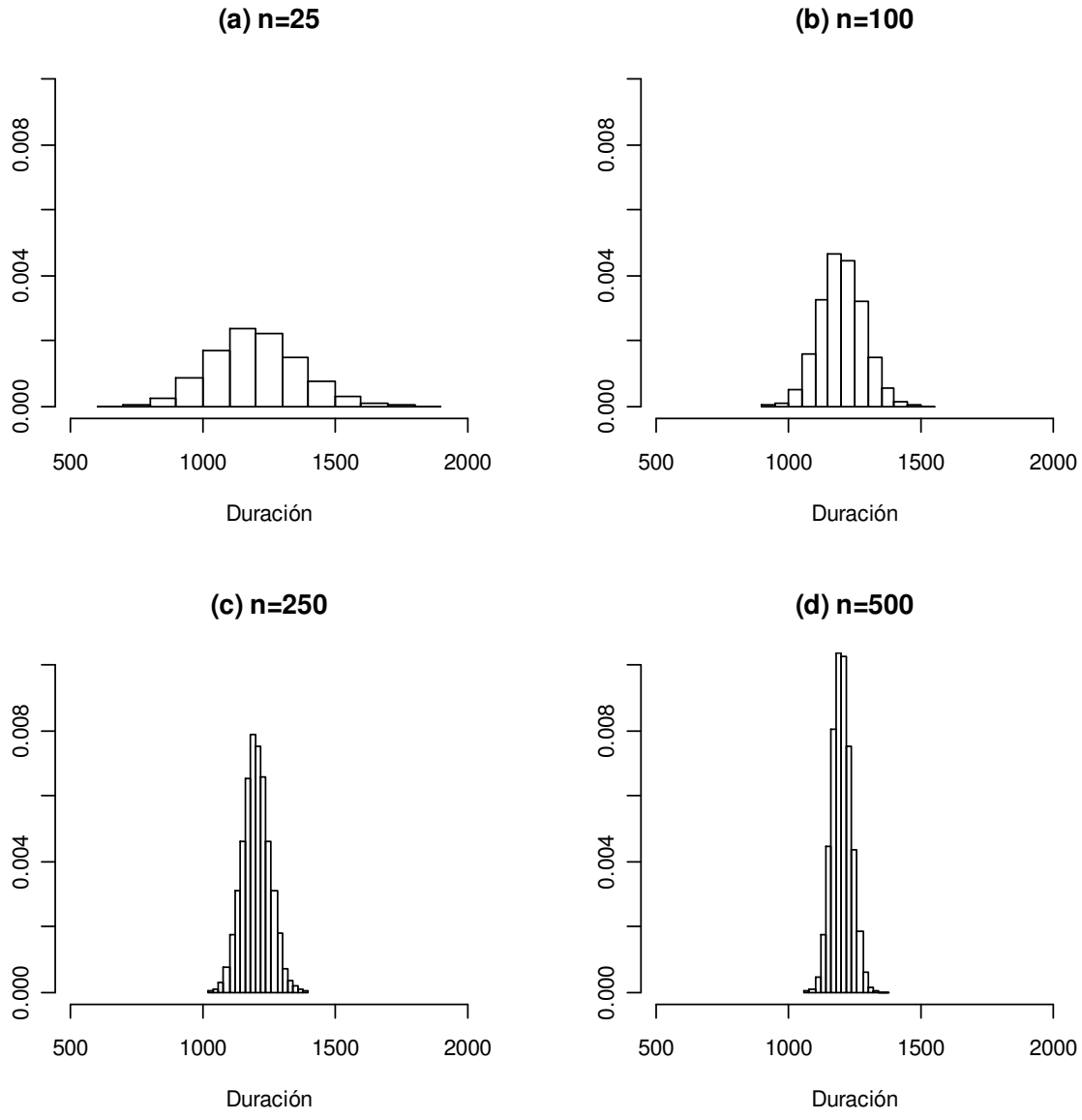


Figura 3.2: Histogramas de los 10.000 valores de duración media obtenidos simulando la obtención de 10.000 muestras **(a)** de 25 componentes **(b)** de 100 componentes **(c)** de 250 componentes **(d)** de 500 componentes.

¿Qué implicaciones prácticas tienen estas observaciones? En la práctica vamos a obtener una única muestra, no 10.000 como en estas simulaciones. Sin embargo, lo que estos experimentos nos enseñan es que si la muestra que tomamos es de tamaño 25, la media de nuestra *única muestra* puede alejarse mucho de la media de la población (recordemos que nos podían salir medias de entre 642 y 1884

minutos, prácticamente 600 minutos de diferencia con respecto a la media real). Y aún sin que nos salgan estos valores tan extremos, es muy fácil que la media de la muestra se diferencie en 200 o más minutos de la media de la población. Sin embargo, cuando la muestra es de tamaño 500, la peor estimación que llega a obtenerse tiene un error de unos 100 minutos (muy lejos de los 600 de la muestra de tamaño 25), y la inmensa mayoría de las estimaciones tiene un error inferior a 60 minutos. Por tanto, si nuestra muestra es de tamaño 500, lo más probable es que el error de estimación de la media sea pequeño. Evidentemente todas estas ideas habrán de ser cuantificadas de algún modo, pero lo que sí debería ser intuitivamente claro es que *si la única muestra disponible en la práctica es pequeña es muy fácil que el error de estimación sea grande; y si esa muestra es grande lo más probable es que el error sea pequeño.*

3.4. DEFINICIONES BÁSICAS.

A continuación presentamos algunas definiciones básicas para definir correctamente el problema de la estimación de parámetros.

***Estadístico*:** Dada una muestra aleatoria $X_1, X_2, ..., X_n$, se llama *estadístico* a cualquier función de sus valores.

Estimador: Dado un parámetro θ característico de una población, y una muestra aleatoria $X_1, X_2, ..., X_n$ de la misma, se llama *estimador* de θ a cualquier estadístico $\hat{\theta} = \hat{\theta}(X_1, X_2, ..., X_n)$ cuyos valores se aproximen a θ.

Repetimos una vez más que un estimador es una *variable aleatoria* pues depende de los valores que tome la muestra aleatoria $X_1, X_2, ..., X_n$ (que no se conocen antes de realizar el muestreo) Las propiedades del estimador dependerán por tanto de su distribución de probabilidad.

En la sección 3.3 hemos planteado la cuestión de cómo medir el grado de proximidad entre un parámetro y un estimador suyo. Recordemos que el problema principal radica en que mientras que el parámetro tiene un único valor fijo, el estimador cambia de valor con la muestra.

Sesgo de un estimador: Dado que el estimador puede tomar muchos valores diferentes (según cual sea la muestra que se obtenga), una manera de medir la proximidad entre el estimador y el parámetro es mediante la distancia entre el valor esperado del estimador y el valor del parámetro. Dicha distancia recibe el nombre de *sesgo del estimador*:

$$Sesgo(\hat{\theta}) = E[\hat{\theta}] - \theta$$

Cuando el sesgo del estimador es cero (en cuyo caso $E[\hat{\theta}] = \theta$), el estimador se dice *insesgado* o *centrado.* En caso contrario el estimador es sesgado. Un estimador con sesgo positivo significa que sus valores, en media, están por encima del parámetro que pretende estimar y por tanto tiende a sobreestimarlo. De modo similar, los estimadores con sesgo negativo tienden a subestimar el parámetro.

Para entender el significado de $E\left[\hat{\theta}\right]$ recordemos el estudio de simulación de la sección anterior. Si el parámetro θ que se pretende estimar es la media μ de una población y utilizamos como estimador el valor $\hat{\theta} = \hat{\mu} = \bar{x}$, esto es, la media muestral, podemos interpretar $E\left[\hat{\theta}\right] = E\left[\hat{\mu}\right] = E\left[\bar{x}\right]$ como el valor medio de la media muestral, esto es, como la media de muchas medias obtenidas en distintos muestreos. Recordemos que en el estudio de simulación, con 10.000 muestras simuladas, la media de

las 10.000 medias obtenidas coincidía prácticamente con la media de la población, esto es $E[\bar{x}]=\mu$. Este es un resultado no solamente empírico, sino que se puede comprobar de forma teórica como veremos en el primero de los ejemplos que figuran a continuación.

Ejemplos:

- La media muestral es un estimador centrado de la media poblacional. En efecto:

$$E\left[\bar{X}\right]=E\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right]=\frac{1}{n}E\left[\sum_{i=1}^{n}X_i\right]=\frac{1}{n}\sum_{i=1}^{n}E\left[X_i\right]=\frac{1}{n}n\mu=\mu$$

- La varianza muestral es un estimador sesgado de la varianza poblacional. En efecto, como:

$$\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2=\sum_{i=1}^{n}\left(X_i-\mu+\mu-\bar{X}\right)^2=\sum_{i=1}^{n}\left(\left(X_i-\mu\right)-\left(\bar{X}-\mu\right)\right)^2=$$

$$=\sum_{i=1}^{n}\left(\left(X_i-\mu\right)^2-2\left(X_i-\mu\right)\left(\bar{X}-\mu\right)+\left(\bar{X}-\mu\right)^2\right)=$$

$$=\sum_{i=1}^{n}\left(X_i-\mu\right)^2-2\left(\bar{X}-\mu\right)\sum_{i=1}^{n}\left(X_i-\mu\right)+\sum_{i=1}^{n}\left(\bar{X}-\mu\right)^2=$$

$$=\sum_{i=1}^{n}\left(X_i-\mu\right)^2-2\left(\bar{X}-\mu\right)n\left(\bar{X}-\mu\right)+n\left(\bar{X}-\mu\right)^2=$$

$$=\sum_{i=1}^{n}\left(X_i-\mu\right)^2-n\left(\bar{X}-\mu\right)^2$$

y, por ser las X_i independientes:

$$E\left[\left(\bar{X}-\mu\right)^2\right]=Var\left(\bar{X}\right)=Var\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}Var\left(X_i\right)=\frac{1}{n^2}n\sigma^2=\frac{\sigma^2}{n}$$

para la varianza muestral se tiene enonces que:

$$E\left[s^2\right]=E\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right]=\frac{1}{n}E\left[\sum_{i=1}^{n}\left(X_i-\mu\right)^2-n\left(\bar{X}-\mu\right)^2\right]=$$

$$=\frac{1}{n}\left[\sum_{i=1}^{n}E\left[\left(X_i-\mu\right)^2\right]-nE\left[\left(\bar{X}-\mu\right)^2\right]\right]=\frac{1}{n}\left[n\sigma^2-n\frac{\sigma^2}{n}\right]=\frac{n-1}{n}\sigma^2$$

Por tanto:

$$Sesgo(s^2)=E\left[s^2\right]-\sigma^2=\frac{n-1}{n}\sigma^2-\sigma^2=-\frac{1}{n}\sigma^2$$

de donde se sigue que la varianza muestral subestima la varianza poblacional (aunque es verdad que a medida que el tamaño de la muestra *n* aumenta, el sesgo se hace más pequeño).

- La cuasivarianza muestral sí es un estimador centrado de la varianza poblacional:

$$E\left[S^2\right]=E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right]=\frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right]=$$

$$= \frac{1}{n-1} E\left[\sum_{i=1}^{n} (X_i - \mu)^2 - n(\bar{X} - \mu)^2 \right] = \frac{1}{n-1} \left[n\sigma^2 - \sigma^2 \right] = \sigma^2$$

Por esta razón, como estimador de la varianza poblacional, en la práctica se prefiere la cuasivarianza muestral.

- Si X es una variable aleatoria de Bernoulli de parámetro p, la proporción muestral de éxitos $\hat{p}$ es un estimador insesgado de la proporción poblacional p. En efecto, la proporción muestral de éxitos al observar una muestra aleatoria de tamaño n es:

$$\hat{p} = \frac{\text{Número de éxitos}}{\text{Número de observaciones}} = \frac{X}{n}$$

(donde el número de éxitos X sigue obviamente una distribución $B(n,p)$), y por tanto:

$$E[\hat{p}] = E\left[\frac{X}{n} \right] = \frac{1}{n} E[X] = \frac{1}{n} n \cdot p = p$$

Error estándar del estimador: así se llama a la desviación típica del estimador.

Tal como hemos visto, un estimador es una variable aleatoria cuyo valor cambia con la muestra. Si el estimador es centrado, ello indica que el centro de la distribución de valores del estimador coincide con el parámetro que se pretende estimar. Si embargo esto no nos informa de si dicha distribución tiene mucha o poca dispersión en torno al parámetro. Si la dispersión es grande, significa que habrá muestras que darán lugar a estimaciones muy alejadas del valor del parámetro. Si la dispersión es pequeña, aún en la peor de las muestras posibles, la estimación obtenida estará próxima al valor del parámetro. Por tanto, si se dispone de varios estimadores centrados del mismo parámetro, será preferible (producirá estimaciones más precisas del parámetro) aquél que tenga la menor dispersión. Dado que la dispersión se mide mediante la varianza del estimador, el mejor estimador centrado será el de mínima varianza (en caso de existir).

Error Cuadrático Medio de un estimador. Se define como:

$$ECM\left[\hat{\theta}\right] = E\left[\left(\hat{\theta} - \theta\right)^2 \right] = \left(Sesgo\left(\hat{\theta}\right) \right)^2 + Var\left(\hat{\theta}\right)$$

El ECM constituye una medida conjunta (de hecho es la suma) del sesgo y la varianza de un estimador. Es una medida que resulta útil cuando se debe elegir entre varios estimadores del mismo parámetro con características muy diferentes de sesgo y varianza. Así por ejemplo, puede ser más útil un estimador ligeramente sesgado pero con muy poca varianza (tal que, aunque sesgadas, todas las estimaciones están próximas al parámetro), que uno centrado pero con varianza mucho mayor (que puede dar lugar a muchas estimaciones muy alejadas del parámetro).

Consistencia de un estimador. Un estimador $\hat{\theta}$ de un parámetro θ es *consistente* si verifica que:

$$\lim_{n \to \infty} P\left(\left| \hat{\theta} - \theta \right| \leq \varepsilon \right) = 1 \quad \forall \varepsilon > 0$$

lo que significa que a medida que aumenta el tamaño de la muestra es cada vez más probable que el valor del estimador esté cada vez más próximo al valor del parámetro. En general es deseable que los estimadores que utilicemos sean consistentes.

Puede demostrarse que *la media muestral, la varianza muestral y la proporción muestral son estimadores insesgados, de mínima varianza y consistentes de sus parámetros respectivos*. En particular , si μ= E[X], el hecho de que:

$$\lim_{n\to\infty} P\left(\left|\overline{x}_n - \mu\right| \le \varepsilon\right) = 1 \quad \forall \varepsilon > 0$$

siendo $\overline{x}_n$ la media de una muestra de tamaño *n,* justifica que el concepto de esperanza de una variable aleatoria puede identificarse con el de media aritmética para grandes valores de *n*.

3.5. MÉTODOS DE OBTENCIÓN DE ESTIMADORES PUNTUALES

Otra pregunta que nos habíamos planteado al comienzo de este capítulo se refería a la manera en que podían obtenerse funciones cuyos valores se aproximaran al de un parámetro desconocido. Tres son los métodos que se emplean habitualmente para ello: el método de los momentos, el método de máxima verosimilitud y el método de los mínimos cuadrados.

5.7.1. ***Método de los momentos***

Dada una variable aleatoria X, se define el *momento de orden k* respecto al origen como:

$$\mu_k = E\left[X^k\right] = \begin{cases} \sum_{i\in E} x_i^k P(X=i) & \text{si X es discreta} \\ \int_{-\infty}^{\infty} x^k f(x)\,dx & \text{si X es continua} \end{cases}$$

Evidentemente $\mu = \mu_1$ y $\sigma^2 = \mu_2 - \mu_1^2$. De la misma forma que la media y la varianza se pueden poner en función de los momentos de primer orden, en general si una variable aleatoria X depende de unos parámetros desconocidos θ_1, θ_2, ...,θ_k, muchas veces será posible expresar estos parámetros como funciones de los momentos de primer orden, esto es, $\theta_j = g_j(\mu_1, \mu_2, ...)$, $j = 1,2,...,k$. El *método de los momentos* consiste en determinar estas funciones, estimar los momentos correspondientes mediante sus análogos muestrales:

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} (X_i)^2, \quad \ldots, \hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} (X_i)^k$$

y por último estimar los θ_j, mediante las funciones anteriores evaluadas en los momentos muestrales:

$$\hat{\theta}_j = g_j(\hat{\mu}_1, \hat{\mu}_2, ...), \quad j = 1,2,...,k$$

Este método tiene su fundamento en el teorema de Khintchine (que garantiza que los momentos muestrales convergen en probabilidad a los poblacionales) y en el hecho de que los momentos muestrales son estimadores insesgados de los momentos poblacionales.

Ejemplos:

1. Supongamos que se desea estimar el parámetro p de una variable X con distribución de Bernoulli. Recuérdese que un variable de Bernoulli toma sólo uno de dos posibles valores: 1 con probabilidad p y 0 con probabilidad 1-p. A modo de ejemplo p podría ser la proporción de visitantes que sufren accidentes en una urbanización turística. Si se toma una muestra aleatoria de n visitantes, la variable X_i para el visitante i-ésimo vale 1 si el sujeto sufre un accidente durante sus estancia en la urbanización y 0 si no lo sufre. Sea r el número total de

accidentados en la muestra. Sabemos que para la distribución de Bernoulli $\mu_1 = E[X] = p$. Por tanto, para estimar *p*, sustituimos μ_1 en esta ecuación por su estimador $\hat{\mu}_1 = \bar{x}$ y obtenemos como estimador de *p:*

$$\hat{p} = \bar{x}$$

Nótese que:

$$\hat{p} = \hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{r}{n}$$

es precisamente la proporción de visitantes accidentados en la muestra, ya que $r = \sum_{i=1}^{n} X_i$ es el número total de accidentados.

2. Supongamos que se desea estimar el parámetro *p* de una variable con distribución geométrica. En este caso $E[X] = \frac{1}{p}$ de donde $p = \frac{1}{E[X]} = \frac{1}{\mu_1}$. Por tanto, para estimar *p* por el método de los momentos sustituimos μ_l por su estimador $\hat{\mu}_1 = \bar{x}$, por lo que el estimador de *p* será:

$$\hat{p} = \frac{1}{\bar{x}}$$

De esta forma, para estimar *p*, deberemos tomar una muestra de tamaño *m*, calcular su media e invertirla (en el ejemplo, tomar una muestra de microcircuitos, ver para cada uno de ellos el número de fallos que se producen hasta que el chip ya no tiene arreglo, calcular la media de todos estos valores y estimar la probabilidad *p* de que ocurra un fallo irreparable mediante el inverso de la media).

3. Si $X \approx N(\mu,\sigma)$ y se desea estimar μ, basta observar que como $\mu = E[X] = \mu_1$, el estimador de es simplemente $\hat{\mu} = \bar{x}$.

4. En la misma situación anterior, si lo que interesa estimar es la varianza σ^2, bastará observar que:

$$\sigma^2 = E\left[X^2\right] - \left(E[X]\right)^2 = \mu_2 - \mu_1^2$$

por lo que el estimador de la varianza por el método de los momentos es, utilizando una muestra de tamaño *n*:

$$\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{x})^2$$

que es precisamente la varianza muestral.

5. Si $X \approx \Gamma(\alpha,\beta)$, para estimar los parámetros α y β por el método de los momentos, bastará tener en cuenta que para esta variable aleatoria:

$$E[X] = \frac{\alpha}{\beta} \qquad Var(X) = \frac{\alpha}{\beta^2}$$

De aquí se sigue que:

$$\left.\begin{array}{l}\alpha=\beta E[X]\\ \alpha=\beta^2 Var(X)\end{array}\right\}\Rightarrow 1=\frac{\beta E[X]}{\beta^2 Var(X)}=\frac{E[X]}{\beta Var(X)}\Rightarrow \beta=\frac{E[X]}{Var(X)}\Rightarrow$$

$$\Rightarrow \alpha=\beta E[X]=\frac{(E[X])^2}{Var(X)}$$

Por tanto, los estimadores por el método de los momentos serán:

$$\hat{\alpha}=\frac{\bar{X}^2}{S^2}\qquad \hat{\beta}=\frac{\bar{X}}{S^2}$$

6. Si $X\approx W(\alpha,\beta)$, para estimar α y β por el método de los momentos, al igual que en el caso anterior bastará con tener en cuenta que su esperanza y varianza son:

$$E[X]=\beta\Gamma\left(1+\frac{1}{\alpha}\right)\qquad Var(X)=\beta^2\left[\Gamma\left(1+\frac{2}{\alpha}\right)-\Gamma^2\left(1+\frac{1}{\alpha}\right)\right]$$

Si calculamos ahora el coeficiente de variación obtenemos que éste es función sólo del parámetro α:

$$CV(X)=\frac{\sqrt{Var(X)}}{E[X]}=\frac{\sqrt{\Gamma\left(1+\frac{2}{\alpha}\right)-\Gamma^2\left(1+\frac{1}{\alpha}\right)}}{\Gamma\left(1+\frac{1}{\alpha}\right)}=\sqrt{\frac{\Gamma\left(1+\frac{2}{\alpha}\right)}{\Gamma^2\left(1+\frac{1}{\alpha}\right)}-1}$$

Podemos entonces obtener un estimador de α igualando ahora este último término al coeficiente de variación de una muestra aleatoria de la distribución de Weibull. Obviamente no es posible despejar de aquí el valor de α explícitamente, pero es fácil construir un algoritmo numérico que resuelva el problema. Una vez obtenido $\hat{\alpha}$, se despeja β de:

$$\bar{x}=\hat{\beta}\Gamma\left(1+\frac{1}{\hat{\alpha}}\right)\Rightarrow\hat{\beta}=\bar{x}\Big/\Gamma\left(1+\frac{1}{\hat{\alpha}}\right)$$

5.7.2. ***Método de la máxima verosimilitud.***

Sea X una variable aleatoria cuya función de densidad depende uno o varios parámetros $\theta_1, \theta_2, ..., \theta_k$. Denotamos por $f_\Theta(x)$ a la función de densidad de X, siendo $\Theta=(\theta_1, \theta_2, ..., \theta_k)$. Dada una muestra aleatoria simple $x_1, x_2, ..., x_n$ de esta variable, se define la *función de verosimilitud* de la muestra como:

$$L(\theta_1,\theta_2,...,\theta_k)=f_\Theta(x_1)f_\Theta(x_2)...f_\Theta(x_n)$$

Esta función representa la densidad de probabilidad conjunta de las variables $X_1, X_2, ..., X_n$ en el punto $x_1, x_2, ..., x_n$. En otras palabras, si $X_1, X_2, ..., X_n$ son variables aleatorias discretas independientes, el valor de esta función es la probabilidad de observar los valores $x_1, x_2, ..., x_n$ cuando $\Theta=(\theta_1, \theta_2, ..., \theta_k)$. Análogamente, si las X_i son continuas, esta función nos da la densidad de probabilidad en un entorno de $x_1, x_2, ..., x_n$ cuando $\Theta=(\theta_1, \theta_2, ..., \theta_k)$.

Obviamente, el valor de la función de verosimilitud depende de lo que valgan los parámetros $(\theta_1, \theta_2, ..., \theta_k)$. La ocurrencia de unos valores determinados $X_1=x_1, X_2=x_2, ..., X_n=x_n$ en la muestra puede ser más o menos probable dependiendo de lo que valgan los θ_i. Ello nos permite interpretar la función de verosimilitud como una medida de la certidumbre o confianza que podemos tener en que los

parámetros desconocidos tomen unos valores concretos. De esta forma, si en repetidas observaciones de una variable aleatoria hay determinados valores que ocurren con mucha frecuencia, será razonable suponer que es porque dichos valores son muy probables. Aunque los parámetros θ_i sean desconocidos, lo más verosímil es que sus valores sean tales que la densidad de X asigne mucha probabilidad a esas observaciones muy frecuentes. Este es el *principio de la máxima verosimilitud:* estimar los θ_i mediante aquellos valores que maximizan la función de verosimilitud. En la práctica estos valores se pueden obtener resolviendo el sistema de ecuaciones formado por las derivadas parciales de $L(\theta_1, \theta_2, ..., \theta_k)$ respecto de los parámetros $\theta_1, \theta_2, ..., \theta_k$ igualadas a cero:

$$\frac{\partial L(\theta_1,\theta_2,...,\theta_k)}{\partial \theta_j} = 0, \quad \forall j = 1,2,...,k$$

La función de verosimilitud será en general un producto de funciones, $L(\theta_1,\theta_2,...,\theta_k) = f_\Theta(x_1) f_\Theta(x_2) ... f_\Theta(x_n)$ lo que en muchas ocasiones da lugar a que la evaluación de las derivadas anteriores sea una tarea compleja. Por esta razón, la función de verosimilitud suele sustituirse por su logaritmo (log-verosimilitud):

$$l(\theta_1,\theta_2,...,\theta_k) = \log\left(L(\theta_1,\theta_2,...,\theta_k)\right) = \sum_{i=1}^{n} \log\left(f_\Theta(x_i)\right)$$

Por ser el logaritmo una función monónota, la verosimilitud y la log-verosimilitud alcanzan su máximo en los mismos puntos. Por tanto, *maximizar la función de verosimilitud es equivalente a maximizar su logaritmo (log-verosimilitud)*, siendo en general el problema de maximizar la log-verosimilitud mucho más sencillo de resolver.

A continuación enunciamos algunas propiedades de los estimadores de máxima verosimilitud:

1. Son consistentes.
2. Son invariantes frente a transformaciones biunívocas, es decir, si $\hat{\theta}_{MV}$ es el estimador máximo verosímil de un parámetro θ, y $g(\theta)$ es una función biunívoca de θ, entonces $g\left(\hat{\theta}_{MV}\right)$ es el estimador máximo verosímil de $g(\theta)$.
3. Son asintóticamente normales.
4. Son asintóticamente eficientes, es decir, entre todos los estimadores consistentes de un parámetro θ, los de máxima verosimilitud son los de varianza mínima.
5. No siempre son insesgados.

Ejemplos:

1. Supongamos que $X \approx exp(\theta)$. En este caso $f_\theta(x) = \frac{1}{\theta} e^{-\frac{1}{\theta}x}$. Dada una muestra $X_1=x_1$, $X_2=x_2$, ..., $X_n=x_n$, la función de verosimilitud y su derivada respecto a θ serán:

$$L(\theta) = f(x_1)\cdot f(x_2)\cdot ... \cdot f(x_n) = \frac{1}{\theta} e^{\frac{-x_1}{\theta}} \cdot \frac{1}{\theta} e^{\frac{-x_2}{\theta}} \cdot ... \cdot \frac{1}{\theta} e^{\frac{-x_n}{\theta}} = \left(\frac{1}{\theta}\right)^n e^{\frac{-1}{\theta}(\sum x_i)}$$

$$\frac{\partial L(\theta)}{\partial \theta} = -n\theta^{-n-1} e^{\frac{-1}{\theta}(\sum x_i)} + \theta^{-n-2}\left(\sum x_i\right) e^{\frac{-1}{\theta}(\sum x_i)} = 0$$

Despejando θ en esta ecuación obtenemos el estimador máximo-verosimil de este parámetro:

$$\theta^{-n-2} e^{\frac{-1}{\theta}(\sum x_i)} \left(-n\theta + \sum x_i\right) = 0 \Rightarrow \left(-n\theta + \sum x_i\right) = 0 \Rightarrow \hat{\theta} = \frac{\sum x_i}{n} = \bar{X}$$

2. Supongamos que se desea estimar el parámetro p de una variable X con distribución de Bernoulli por el método de máxima verosimilitud. Si se ha observado la muestra $X_1=k_1$, $X_2=k_2$, ..., $X_m=k_m$,(donde las k_i son 0 ó 1) la verosimilitud asociada es:

$$L(p)=f_p(k_1,k_2,...,k_m)=P(X_1=k_1)P(X_2=k_2)...P(X_m=k_m)$$
$$=\left[P(X=1)\right]^r\left[P(X=0)\right]^{n-r}=p^r(1-p)^{n-r}$$

(hemos supuesto que se ha observado r veces el 1 y n-r veces el cero)

Obtener la derivada de esta expresión respecto a p es complicado. No obstante, podemos observar que si se toma logaritmo de esta expresión la derivada es mucho más sencilla. Tomando logaritmos en la expresión anterior obtenemos la log-verosimilitud:

$$l(p)=\ln L(p)=r\ln(p)+(n-r)\ln(1-p)$$

Ahora es fácil derivar respecto a p e igualar a 0:

$$l(p)=\ln L(p)=r\ln(p)+(n-r)\ln(1-p)$$
$$l'(p)=\frac{r}{p}-\frac{n-r}{1-p}=0\Rightarrow(1-p)r-p(n-r)=0\Rightarrow$$
$$\Rightarrow r-pn=0\Rightarrow p=\frac{r}{n}$$

Como puede apreciarse en este caso el estimador de máxima verosimilitud coincide con el que ya obtuvimos por el método de los momentos, aunque en general no tiene por qué ocurrir así.

3. Supongamos ahora que tomamos una muestra de observaciones de una variable con distribución de Weibull de parámetros α y β. Para estimar estos parámetros por máxima verosimilitud, primero calculamos la función de verosimilitud, suponiendo que se ha observado la muestra $X_1=x_1$, $X_2=x_2$, ..., $X_n=x_n$,:

$$L(\alpha,\beta)=f(x_1)\cdot f(x_2)\cdot...\cdot f(x_n)=\left(\frac{\alpha}{\beta^\alpha}\right)^n(x_1\cdot x_2\cdot....\cdot x_n)^{\alpha-1}\exp\left(-\sum_{i=1}^n(x_i/\beta)^\alpha\right)$$

Como el máximo de esta función se encuentra en el mismo lugar que el máximo de su logaritmo, calculamos su logaritmo (log-verosimilitud):

$$l(\alpha,\beta)=\log\left(L(\alpha,\beta)\right)=n\log(\alpha)-\alpha n\log(\beta)+(\alpha-1)\sum_{i=1}^n\log(x_i)-\sum_{i=1}^n(x_i/\beta)^\alpha$$

Para hallar ahora los valores de α y β que maximizan esta expresión, calculamos las derivadas parciales e igualamos a 0:

$$\frac{\partial l(\alpha,\beta)}{\partial\alpha}=0\qquad\frac{\partial l(\alpha,\beta)}{\partial\beta}=0$$
$$\frac{\partial l(\alpha,\beta)}{\partial\alpha}=\frac{n}{\alpha}-n\log(\beta)+\sum_{i=1}^n\log(x_i)-\sum_{i=1}^n(x_i/\beta)^\alpha\log(x_i/\beta)=0$$
$$\frac{\partial l(\alpha,\beta)}{\partial\beta}=-\frac{\alpha n}{\beta}+\frac{\alpha}{\beta}\sum_{i=1}^n(x_i/\beta)^\alpha=0$$

De la segunda ecuación se obtiene:

$$\frac{1}{\beta^{\alpha}}\sum_{i=1}^{n}(x_i)^{\alpha}=n\Rightarrow\beta=\left(\frac{1}{n}\sum_{i=1}^{n}(x_i)^{\alpha}\right)^{1/\alpha}$$

y sustituyendo en la primera ecuación:

$$\frac{n}{\alpha}-n\log(\beta)+\sum_{i=1}^{n}\log(x_i)-\frac{1}{\beta^{\alpha}}\sum_{i=1}^{n}x_i^{\alpha}\left(\log(x_i)-\log(\beta)\right)=0$$

$$\frac{n}{\alpha}-n\log(\beta)+\sum_{i=1}^{n}\log(x_i)-\frac{1}{\beta^{\alpha}}\sum_{i=1}^{n}x_i^{\alpha}\log(x_i)+\frac{\log(\beta)}{\beta^{\alpha}}\sum_{i=1}^{n}x_i^{\alpha}=0$$

$$\frac{n}{\alpha}-n\log\left(\left(\frac{1}{n}\sum_{i=1}^{n}x_i^{\alpha}\right)^{1/\alpha}\right)+\sum_{i=1}^{n}\log(x_i)-\frac{1}{\frac{1}{n}\sum_{i=1}^{n}x_i^{\alpha}}\sum_{i=1}^{n}x_i^{\alpha}\log(x_i)+\frac{\log\left(\left(\frac{1}{n}\sum_{i=1}^{n}x_i^{\alpha}\right)^{1/\alpha}\right)}{\frac{1}{n}\sum_{i=1}^{n}x_i^{\alpha}}\sum_{i=1}^{n}(x_i)^{\alpha}=0$$

$$\frac{n}{\alpha}-n\log\left(\left(\frac{1}{n}\sum_{i=1}^{n}x_i^{\alpha}\right)^{1/\alpha}\right)+\sum_{i=1}^{n}\log(x_i)-n\frac{\sum_{i=1}^{n}x_i^{\alpha}\log(x_i)}{\sum_{i=1}^{n}x_i^{\alpha}}+n\log\left(\left(\frac{1}{n}\sum_{i=1}^{n}x_i^{\alpha}\right)^{1/\alpha}\right)=0$$

$$\frac{n}{\alpha}+\sum_{i=1}^{n}\log(x_i)-n\frac{\sum_{i=1}^{n}x_i^{\alpha}\log(x_i)}{\sum_{i=1}^{n}x_i^{\alpha}}=0;\qquad\frac{1}{\alpha}+\frac{\sum_{i=1}^{n}\log(x_i)}{n}-\frac{\sum_{i=1}^{n}x_i^{\alpha}\log(x_i)}{\sum_{i=1}^{n}x_i^{\alpha}}=0$$

De donde podemos despejar:

$$\alpha=\left[\frac{\sum_{i=1}^{n}x_i^{\alpha}\log(x_i)}{\sum_{i=1}^{n}x_i^{\alpha}}-\frac{\sum_{i=1}^{n}\log(x_i)}{n}\right]^{-1}$$

Esta última ecuación no tiene una solución explícita, debiendo resolverse numéricamente. Una vez que se obtenga de esta manera el valor estimado de α, se sustituye en la ecuación para β, obteniéndose de este modo el estimador máximo verosímil de este parámetro.

3.6. ESTIMACIÓN POR INTERVALOS DE CONFIANZA

Hasta ahora hemos visto cómo podemos obtener un estimador puntual para un parámetro de una distribución de probabilidad. Si se dan las condiciones adecuadas (error cuadrático medio pequeño), sabemos que el estimador al ser evaluado sobre distintas muestras produce valores próximos al valor del parámetro que se pretende estimar. Ahora bien: ¿cuánto se parece el valor estimado al verdadero valor del parámetro?. O dicho de otra forma: ¿cuál es el grado de precisión alcanzado en la estimación? El concepto de *intervalo de confianza* permite dar respuesta a estas preguntas.

Dado un parámetro desconocido θ, característico de una población determinada, y dada una muestra aleatoria $X = \{X_1, X_2, ..., X_n\}$ de dicha población, diremos que el intervalo $[\theta_1(X), \theta_2(X)]$, (donde $\theta_1(X)$ y $\theta_2(X)$ son variables aleatorias que dependen de la muestra) es un *intervalo de confianza* a nivel $1-\alpha$ para el parámetro θ si la probabilidad de que intervalo contenga a dicho parámetro es 1-α, esto es:

$$P\left(\theta \in \left[\theta_1(X), \theta_2(X)\right]\right) = 1-\alpha$$

De esta forma, si disponemos de un intervalo de confianza para un parámetro θ desconocido, ya no nos limitaremos a decir que θ tiene un valor parecido a $\hat{\theta}$ (su estimador puntual), sino que además podremos afirmar con una confianza del 100(1-α)% que θ vale como mínimo $\theta_1(X)$ y como máximo $\theta_2(X)$. Ello nos da una idea aproximada de la precisión conseguida en la estimación. Nótese que en la definición de intervalo de confianza, los extremos $\theta_1(X)$ y $\theta_2(X)$ son variables aleatorias ya que son funciones de la muestra y ésta es aleatoria. Ello significa que dos muestras distintas de la misma población producirán intervalos de confianza distintos.

3.7. INTERVALOS DE CONFIANZA PARA LOS PARÁMETROS DE UNA DISTRIBUCIÓN NORMAL

Intervalo de confianza para la media de una distribución normal

Si se dispone de una muestra $X_1, X_2, \ldots, X_n$ de observaciones completas independientes de una distribución normal $N(\mu,\sigma)$, se tiene que:

$$\sum_{i=1}^{n} X_i \approx N\left(\sum_{i=1}^{n}\mu_i, \sqrt{\sum_{i=1}^{n}\sigma_i^2}\right) = N\left(n\mu, \sqrt{n\sigma^2}\right) = N\left(n\mu, \sigma\sqrt{n}\right)$$

Asimismo, utilizando las propiedades de la esperanza y varianza de una variable aleatoria ante un cambio de escala:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \Rightarrow \begin{cases} E\left[\bar{X}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}n\mu = \mu \\ Var\left(\bar{X}\right) = Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n} \end{cases}$$

y por tanto:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

de donde:

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \approx N(0,1)$$

Utilizando tablas de la distribución normal (o cualquier software que maneje esta distribución[1]) es posible encontrar el percentil $z_{\alpha/2}$ tal que, si $Z \approx N(0,1)$, entonces $P\left(-z_{\alpha/2} \le Z \le z_{\alpha/2}\right) = 1-\alpha$

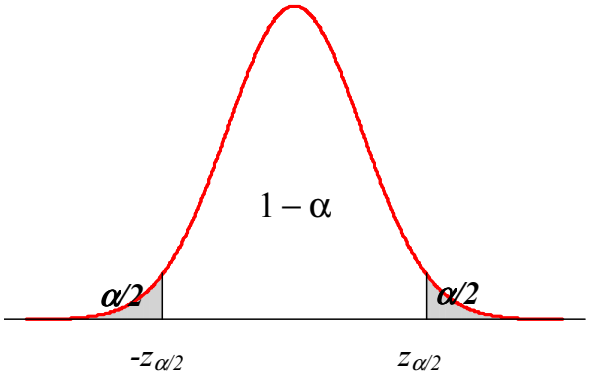
[1] EXCEL o R calculan estos percentiles de manera muy sencilla. Asimismo muchas calculadoras científicas incluyen la opción del cálculo de percentiles de las distribuciones de probabilidad más comunes.

Así pues:

$$P\left(-z_{\alpha/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) = 1-\alpha \Rightarrow$$

$$\Rightarrow P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \bar{X}-\mu \le z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \Rightarrow$$

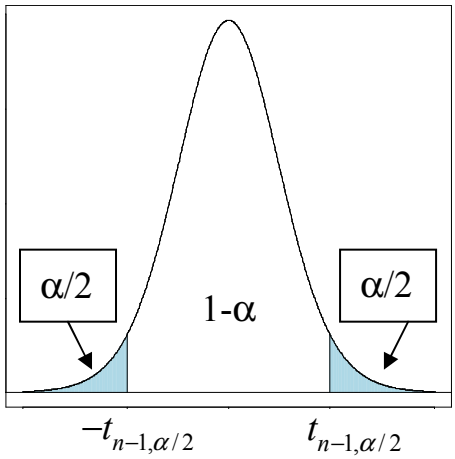
$$\Rightarrow P\left(\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$



De esta forma, si conociéramos la varianza σ^2 de la población, el resultado anterior nos indica que el intervalo de confianza a nivel 1-α para la media de la población es:

$$\left[\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Este intervalo, en la práctica resulta de poca utilidad, toda vez que normalmente la varianza poblacional σ^2 es desconocida. Afortunadamente, es posible demostrar que si $X_1, X_2, \ldots, X_n$ es una muestra de observaciones independientes de una distribución normal $N(\mu,\sigma)$, siendo μ y σ desconocidas, entonces:



$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \approx t_{n-1}$$

siendo $s = \sqrt{\dfrac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1}}$ la desviación típica de la muestra. Al igual que en el caso anterior, es posible utilizar las tablas de la *t* de Student (o cualquier software equipado al efecto) para encontrar el percentil $t_{n-1,\alpha/2}$ de esta distribución, de tal forma que $P\left(-t_{n-1,\alpha/2} \le t_{n-1} \le t_{n-1,\alpha/2}\right) = 1-\alpha$. Podemos escribir entonces:

$$P\left(-t_{n-1,\alpha/2} \le \frac{\bar{X}-\mu}{S/\sqrt{n}} \le t_{n-1,\alpha/2}\right) = 1-\alpha$$

de donde, operando en el interior del intervalo:

$$P\left(\bar{X}-\frac{S}{\sqrt{n}}t_{n-1,\alpha/2} \le \mu \le \bar{X}+\frac{S}{\sqrt{n}}t_{n-1,\alpha/2}\right) = 1-\alpha$$

o, expresado de otra forma:

$$P\left(\mu \in \left[\bar{X}-\frac{S}{\sqrt{n}}t_{n-1,\alpha/2}, \bar{X}+\frac{S}{\sqrt{n}}t_{n-1,\alpha/2}\right]\right) = 1-\alpha$$

Ejemplo: Se dispone de una muestra de 20 barcos pesqueros que han descargado en puerto a lo largo de una semana. De cada barco se mide el peso total de las capturas, obteniéndose un peso medio de 4.8 toneladas con desviación típica de 0.5 toneladas. Podemos calcular un intervalo de confianza para el peso medio descargado por la población de barcos sustituyendo estos valores en la fórmula anterior. Fijando 1-α=0.95 (esto es α=0.05) obtenemos:

$$\left[\bar{X}-\frac{S}{\sqrt{n}}t_{n-1,\alpha/2},\bar{X}+\frac{S}{\sqrt{n}}t_{n-1,\alpha/2}\right]=\left[4.8-\frac{0.5}{\sqrt{20}}t_{19,0.025},4.8+\frac{0.5}{\sqrt{20}}t_{19,0.025}\right]=$$

$$=\left[4.8-\frac{0.5}{\sqrt{20}}2.09,4.8+\frac{0.5}{\sqrt{20}}2.09\right]=\left[4.8-0.234,4.8+0.234\right]=\left[4.566,5.034\right]$$

Así pues, podemos concluir que, con una confianza del 95%, el peso medio de las capturas descargadas por la población de barcos en ese puerto durante esa semana se encuentra en el intervalo [4.56, 5.03]; o dicho de otro modo, podemos afirmar con una confianza del 95% que el peso medio de las capturas por barco es de 4.8 toneladas con un margen de error (precisión) de ±0.234 toneladas.

Intervalo de confianza para la varianza σ^2 de una población normal

Ya hemos visto que la cuasi-varianza muestral $S^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$ es un estimador centrado de la varianza poblacional. Si la muestra $\{X_1, X_2,\ldots, X_n\}$ está formada por *n* observaciones independientes de una distribución $N(\mu,\sigma)$, el teorema de Fisher prueba que:

$$\frac{(n-1)S^2}{\sigma^2}\approx\chi^2_{n-1}$$

Por tanto, utilizando una tabla de la distribución χ^2_{n-1} (o cualquier software que disponga de esta distribución) podemos encontrar los percentiles $\chi^2_{n-1,1-\alpha/2}$ y $\chi^2_{n-1,\alpha/2}$ para los que:

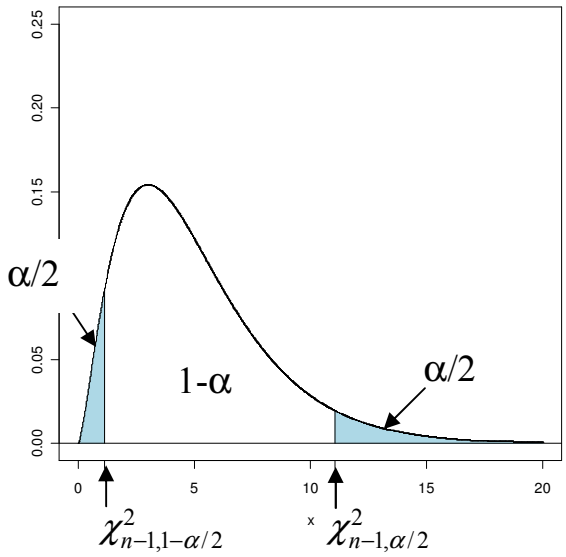
$$P\left(\chi^2_{n-1,1-\alpha/2}\le\frac{(n-1)S^2}{\sigma^2}\le\chi^2_{n-1,\alpha/2}\right)=1-\alpha$$

Operando en el interior del intervalo podemos despejar σ^2:

$$P\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\le\sigma^2\le\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right)=1-\alpha$$

De esta forma, el intervalo de confianza a nivel 1-α para la varianza de una población normal es:

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}},\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right)$$



3.8. INTERVALO DE CONFIANZA PARA UNA PROPORCIÓN

En general consideraremos que se dispone de una población de objetos, cada uno de los cuales, independientemente de los demás, puede tener o no cierta característica de interés. Llamaremos π a la proporción poblacional de objetos con esa característica. Nuestro objetivo será estimar el valor de π, y dar un intervalo de confianza para la estimación. Para ello supondremos que se dispone de una muestra aleatoria de n objetos de la población, para cada uno de los cuales se mide la variable X_i, definida como:

$$X_i = \begin{cases} 0 & \text{si el objeto } i \text{ NO tiene la característica de interés} \\ 1 & \text{si el objeto } i \text{ SÍ tiene la característica de interés} \end{cases}$$

En estas condiciones la variable X_i es obviamente una variable de Bernoulli de parámetro π. El número total de objetos en la muestra que tienen esta característica es $Y = \sum_{i=1}^{n} X_i$. Tal como hemos visto, la proporción muestral:

$$\hat{\pi} = \frac{\text{nº de objetos con la característica de interés}}{\text{nº de objetos en la muestra}} = \frac{Y}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

es un estimador insesgado de π, esto es, $E[\hat{\pi}] = \pi$. Además, como $Y = \sum_{i=1}^{n} X_i$ es una suma de variables de Bernoulli independientes, su distribución es binomial $B(n,\pi)$.

> **Ejemplo:** Se ha tomado una muestra de 60 pequeñas depuradoras instaladas en hoteles, comprobándose que 37 han tenido algún fallo antes de los 3 años de uso. Se desea estimar la proporción de depuradoras que duran más de tres años sin fallar
>
> Dado que 37 de las 60 depuradoras han fallado antes de los 3 años, el resto, 23, han tenido el primer fallo después de ese momento. Por tanto, podemos estimar la proporción de las que duran más de tres años como:
>
> $$\hat{\pi} = \frac{23}{60} = 0.3833 = 38.33\%$$

Para calcular un intervalo de confianza para el verdadero valor de π existen varios métodos.

Método de Wilson.

La varianza del estimador $\hat{\pi}$ puede calcularse teniendo en cuenta que $Y = \sum_{i=1}^{n} X_i \approx B(n,\pi)$:

$$Var(\hat{\pi}) = Var\left(\frac{Y}{n}\right) = \frac{1}{n^2} Var(Y) = \frac{1}{n^2} n\pi(1-\pi) = \frac{\pi(1-\pi)}{n}$$

Si el valor de n es suficientemente grande (en la práctica si $n\hat{\pi} > 5$ y $n(1-\hat{\pi}) > 5$), puede probarse que la distribución del estimador de la proporción es aproximadamente normal:

$$\hat{\pi} \approx N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$$

y por tanto $\dfrac{\hat{\pi}-\pi}{\sqrt{\dfrac{\pi(1-\pi)}{n}}} \approx N(0,1)$. De aquí se sigue que:

$$P\left(-z_{a/2} \le \frac{\hat{\pi}-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \le z_{\alpha/2}\right) = 1-\alpha \Rightarrow P\left(\left|\frac{\hat{\pi}-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}\right| \le z_{\alpha/2}\right) = 1-\alpha \Rightarrow$$

$$\Rightarrow P\left(\left(\frac{\hat{\pi}-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}\right)^2 \le z_{\alpha/2}^2\right) = 1-\alpha \Rightarrow P\left((\hat{\pi}-\pi)^2 \le z_{\alpha/2}^2 \frac{\pi(1-\pi)}{n}\right) = 1-\alpha \Rightarrow$$

$$\Rightarrow P\left(n(\hat{\pi}-\pi)^2 \le z_{\alpha/2}^2 \pi(1-\pi)\right) = 1-\alpha \Rightarrow P\left(\left(n+z_{\alpha/2}^2\right)\pi^2 - \left(2n\hat{\pi}+z_{\alpha/2}^2\right)p + n\hat{\pi}^2 \le 0\right) = 1-\alpha$$

Por tanto el intervalo de confianza a nivel 1-α estará formado por los valores de π para los que se verifica la desigualdad $\left(n+z_{\alpha/2}^2\right)\pi^2 - \left(2n\hat{\pi}+z_{\alpha/2}^2\right)\pi + n\hat{\pi}^2 \le 0$. Para hallar estos valores basta tener en cuenta que la función $g(\pi) = \left(n+z_{\alpha/2}^2\right)\pi^2 - \left(2n\hat{\pi}+z_{\alpha/2}^2\right)\pi + n\hat{\pi}^2$ corresponde a una parábola con los brazos abiertos hacia arriba por lo que la desigualdad anterior se verificará para los valores de π comprendidos entre los dos puntos en que la parábola corta al eje de abcisas. Estos puntos son las soluciones de la ecuación $\left(n+z_{\alpha/2}^2\right)\pi^2 - \left(2n\hat{\pi}+z_{\alpha/2}^2\right)\pi + n\hat{\pi}^2 = 0$, que se obtienen fácilmente como:

$$\pi = \frac{\left(2n\hat{\pi}+z_{\alpha/2}^2\right) \pm \sqrt{\left(2n\hat{\pi}+z_{\alpha/2}^2\right)^2 - 4\left(n+z_{\alpha/2}^2\right)n\hat{\pi}^2}}{2\left(n+z_{\alpha/2}^2\right)} = \frac{\left(2n\hat{\pi}+z_{\alpha/2}^2\right) \pm \sqrt{4nz_{\alpha/2}^2\hat{\pi}(1-\hat{\pi})+z_{\alpha/2}^4}}{2\left(n+z_{\alpha/2}^2\right)} =$$

$$= \frac{\left(n\hat{\pi}+z_{\alpha/2}^2/2\right)}{\left(n+z_{\alpha/2}^2\right)} \pm \frac{z_{\alpha/2}\sqrt{n}}{\left(n+z_{\alpha/2}^2\right)}\sqrt{\hat{\pi}(1-\hat{\pi})+z_{\alpha/2}^2/4n}$$

Nótese que como $\hat{\pi} = \dfrac{\sum_{i=1}^{n} X_i}{n} = \dfrac{Y}{n}$, entonces $n\hat{\pi} = Y =$"*número de objetos en la muestra que tienen la característica de interés*". Por tanto el intervalo de confianza a nivel 1-α para π es :

$$\pi \in \left[\frac{\left(Y+z_{\alpha/2}^2/2\right)}{\left(n+z_{\alpha/2}^2\right)} \pm \frac{z_{\alpha/2}\sqrt{n}}{\left(n+z_{\alpha/2}^2\right)}\sqrt{\hat{\pi}(1-\hat{\pi})+z_{\alpha/2}^2/4n}\right]$$

Ejemplo: Se desea calcular un intervalo de confianza a nivel 1-α = 95% para la proporción de depuradoras que duran más de 3 años sin fallar en el ejemplo anterior. Como α/2 = 0.025, entonces $z_{0.025} = 1.96$. El número de equipos que han durado más de 3 años sin fallar han sido Y = 23 de entre n = 60. El estimador de la proporción es $\hat{\pi} = 23/60 = 0.3833$. Como $n\hat{\pi} = 60 \cdot 23/60 = 23 > 5$, $n(1-\hat{\pi}) = 60(1-23/60) = 37 > 5$, se cumplen las condiciones para la aplicación del método de Wilson. Sustituyendo estos valores en la expresión anterior obtenemos el intervalo:

$$\pi \in [\, 0.39035 \pm 0.11947] = [\, 0.27088,\, 0.50982]$$

Método de Agresti-Coull

Este método proporciona un intervalo de confianza para la proporción con una expresión algo más sencilla que la anterior, si bien requiere tamaños muestrales mayores que 40. En estas condiciones se puede utilizar la aproximación:

$$\hat{\pi} \approx N\left(\pi, \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right)$$

o lo que es lo mismo: $\dfrac{\hat{\pi}-\pi}{\sqrt{\dfrac{\hat{\pi}(1-\hat{\pi})}{n}}} \approx N(0,1)$. Por tanto:

$$P\left(-z_{a/2} \le \frac{\hat{\pi}-\pi}{\sqrt{\dfrac{\hat{\pi}(1-\hat{\pi})}{n}}} \le z_{\alpha/2}\right) = 1-\alpha \Rightarrow P\left(-z_{a/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \le \hat{\pi}-\pi \le z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right) = 1-\alpha \Rightarrow$$

$$\Rightarrow P\left(\hat{\pi} - z_{a/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \le \pi \le \hat{\pi} + z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right) = 1-\alpha$$

De aquí se sigue que el intervalo de confianza a nivel 1-α para π es:

$$\pi \in \left[\hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right] \quad \textit{(Intervalo de Wald)}$$

Este intervalo tiene, no obstante, mal comportamiento para muy diversos valores de n y π, por lo que su uso es desaconsejable. Agresti y Coull[2] han propuesto una modificación de este intervalo que resuelve estos problemas. La modificación consiste en definir:

$$\tilde{Y} = Y + z_{\alpha/2}^2/2, \quad \tilde{n} = n + z_{\alpha/2}^2, \quad \tilde{\pi} = \tilde{Y}/\tilde{n}$$

y recalcular el intervalo de confianza de Wald sustituyendo $\hat{\pi}$ por $\tilde{\pi}$, y n por $\tilde{n}$. El intervalo de confianza a nivel 1-α es entonces:

$$\pi \in \left[\tilde{\pi} \pm z_{\alpha/2}\sqrt{\frac{\tilde{\pi}(1-\tilde{\pi})}{\tilde{n}}}\right] \quad \textit{(Intervalo de Agresti y Coull)}$$

Ejemplo: Dado que n >40 podemos utilizar este intervalo para los datos del ejemplo anterior. En este caso, el resultado que se obtiene es:

$$\pi \in [\, 0.39035 \pm 1.96 \cdot 0.06105] = [\, 0.39035 \pm 1.96 \cdot 0.11964] = [\, 0.27069,\, 0.51002]$$

[2] Puede encontrarse una excelente discusión sobre los distintos métodos de estimación del parámetro de una distribución binomial en el artículo "*Interval estimation for a binomial proportion*" publicado en *Statistical Science* en 2001 y disponible en la red en http://www-stat.wharton.upenn.edu/~tcai/paper/Binomial-StatSci.pdf

que como puede apreciarse es muy similar al obtenido por el método de Wilson (los extremos se diferencian en menos de una milésima). De hecho, a medida que *n* aumenta los métodos de Agresti y Coull, y Wilson tienden a producir el mismo intervalo.

Método de Clopper y Pearson

En el caso de que el tamaño *n* de la muestra o el valor de la proporción estimada $\hat{\pi}$ sean tan pequeños que no se dan las condiciones para aplicar los métodos de Wilson o Agresti y Coull, puede probarse que el siguiente intervalo garantiza un nivel de confianza de *al menos* 1-α para la estimación del parámetro π:

$$\left(\frac{Y}{(n-Y+1)F_1+Y}, \frac{(Y+1)F_2}{(n-Y)+(Y+1)F_2} \right) \text{ (Intervalo de Clopper-Pearson)}$$

donde:

$$F_1 = F_{2(n-Y+1),2Y,\alpha/2},\ F_2 = F_{2(Y+1),2(n-Y),\alpha/2}$$

son percentiles de la distribución *F* de Fisher.

Conviene señalar que al ser un intervalo que garantiza que la confianza es *al menos* 1-α, en muchas ocasiones el nivel de confianza real será mayor, por lo cual este intervalo resulta en general más amplio y por tanto más impreciso que los anteriores, y sólo debe emplearse si no se dan las condiciones para utilizar alguno de aquéllos.

Ejemplo: Si con los datos del ejemplo anterior calculamos el intervalo de Clopper-Pearson, obtenemos:

$$F_1 = F_{2(60-23+1),2\cdot 23,0.025} = F_{76,46,0.025} = 1.71636,\ F_2 = F_{2(23+1),2(60-23),0.025} = F_{48,74,0.025} = 1.65605$$

$$\pi \in \left(\frac{23}{(60-23+1)1.71636+23}, \frac{(23+1)\cdot 1.65605}{(60-23)+(23+1)\cdot 1.65605} \right) = (0.26071,\ 0.51789)$$

Como puede apreciarse este intervalo es similar a los anteriores, aunque algo más amplio. Esta mayor amplitud se debe a que el nivel de confianza de este intervalo es algo mayor que el 95%.

3.9. ¿POR QUÉ EL TÉRMINO "CONFIANZA"?

1. Si $[\theta_1(X), \theta_2(X)]$ es un intervalo de confianza para un parámetro desconocido θ, entonces $\theta_1(X)$ y $\theta_2(X)$ son funciones de la muestra aleatoria $X = \{X_1, X_2, ..., X_n\}$ que se utiliza para estimar θ. Por tanto, $\theta_1(X)$ y $\theta_2(X)$ son *variables aleatorias* que tomarán *distintos valores* según cual sea la muestra que se elija.

   A modo de ejemplo, en el caso particular de la estimación de la media de una distribución normal estas variables son $\theta_1(X_1,...,X_n) = \bar{X} - \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}$, $\theta_2(X_1,...,X_n) = \bar{X} + \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}$

2. Por tanto *antes de tomar la muestra no sabemos cuáles son los valores que van a tomar $\theta_1(X)$ y $\theta_2(X)$*. Como el valor del parámetro θ es una *cantidad fija desconocida, en ese momento tiene sentido hablar de la probabilidad de que el intervalo* $[\theta_1(X),\theta_2(X)]$ *contenga al parámetro* θ.

3. Una vez que se ha tomado la muestra, las variables X_1, X_2, ..., X_n toman valores concretos $X_1=x_1$, $X_2=x_2$, ..., $X_n=x_n$, por lo que las funciones $\theta_1(X)$ y $\theta_2(X)$ toman también valores concretos:

$$\hat{\theta}_1 = \theta_1(X) = \theta_1(x_1, x_2, ..., x_n)$$
$$\hat{\theta}_2 = \theta_2(X) = \theta_1(x_1, x_2, ..., x_n)$$

con lo que también el intervalo de confianza queda particularizado al intervalo concreto $\left[\hat{\theta}_1, \hat{\theta}_2\right]$.

En este momento, el valor θ o bien pertenece a ese intervalo o bien no pertenece al mismo. Ya no tiene sentido hablar de la probabilidad de pertenencia.

Es una cuestión en cierto modo parecida a preguntarse por la probabilidad de sacar cara al lanzar una moneda equilibrada; *antes* de lanzar la moneda esta probabilidad es ½. Lanzamos la moneda y la tapamos para no ver el resultado; en este momento no cabe preguntarse por la probabilidad de sacar cara en esta tirada, puesto que la tirada ya se ha realizado y tiene un resultado concreto; como mucho podremos valorar en ½ nuestra *confianza* en que haya salido cara.
De ahí que estos intervalos reciban el nombre de intervalos de confianza: como el procedimiento de construcción de los mismos garantiza a priori una probabilidad 1-α de contener al parámetro, a posteriori podemos tener una confianza 1-α de haberlo "*capturado*" en nuestro intervalo.

4. El valor de confianza de un intervalo puede entenderse también del siguiente modo: si obtuviésemos un número grande de muestras, de tamaño n cada una de ellas, y para cada una calculásemos el intervalo de confianza correspondiente, del hecho de que $P\left(\theta \in \left[\theta_1(X), \theta_2(X)\right]\right) = 1-\alpha$ se sigue que, de todos los intervalos obtenidos, una proporción $1-\alpha$ contendrán al valor del parámetro, y una proporción α no lo contendrán.

Ejemplo: Hemos simulado que se realiza 100 veces la experiencia de obtener una muestra de tamaño 20 de temperaturas de microcomponentes electrónicos en régimen de funcionamiento normal. En la simulación se han generado datos correspondientes a microcomponentes cuya temperatura media poblacional en régimen de funcionamiento normal es de 25 grados. La figura 6.1 muestra los 100 intervalos de confianza obtenidos. Como puede apreciarse cuatro de los mismos, que hemos marcado en rojo, no contienen al valor medio de la población, 25. Ello era de esperar, pues al calcular los intervalos al 95% cabe esperar que del orden del 5% (en este caso han sido cuatro intervalos) de los mismos no contengan al parámetro.

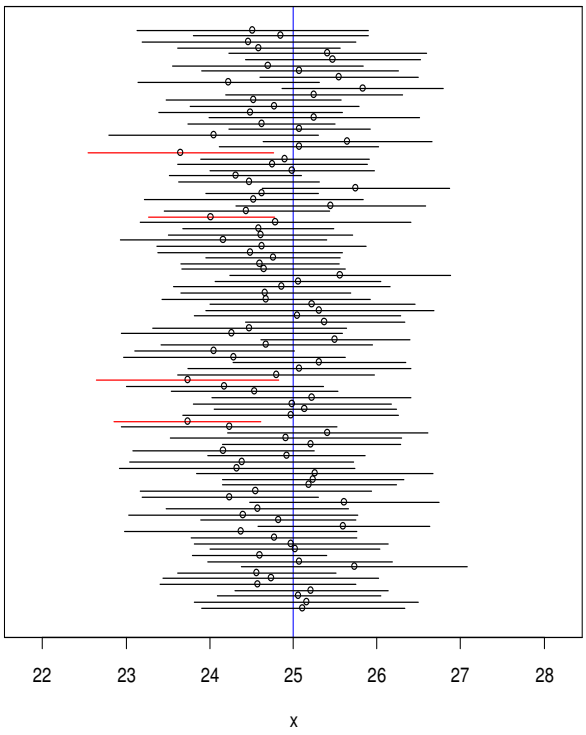


Figura 6.1: 100 intervalos de confianza calculados, respectivamente, a partir de 100 muestras de tamaño 20

Este es un experimento simulado en el que hemos repetido 100 veces la realización de 20 medidas. Ahora bien, *como en la práctica dispondremos realmente de una única muestra, y por tanto de un único intervalo, no podemos estar seguros de si es uno de los que contienen al parámetro o no. Lo más que podemos hacer es confiar en que sea uno de los que sí contienen al parámetro.* Esta confianza la valoramos en el 95% porque *a priori* esta era la probabilidad de que el intervalo contuviese al parámetro.

3.10. MÉTODO GENERAL DE CONSTRUCCIÓN DE INTERVALOS DE CONFIANZA

El procedimiento de construcción de un intervalo de confianza para un parámetro θ sigue en líneas generales los pasos dados en la sección 6.2, en que se obtuvieron intervalos de confianza para la media μ de una población normal de varianza desconocida, o en la sección 6.3 en la aplicación del método de Wilson para obtener el intervalo de confianza para una proporción.

En general deberemos disponer de una *función pivote* $T(\theta,X)$, (donde $X = \{X_1, X_2, ..., X_n\}$ es la muestra aleatoria), que dependa de θ y de los datos, tal que su distribución de probabilidad sea conocida y no dependa de θ. De esta forma, sin conocer θ, siempre será posible calcular dos valores $\tau_I(\alpha)$ y $\tau_S(\alpha)$ tales que:

$$P\left(\tau_I(\alpha) \le T(\theta,X) \le \tau_S(\alpha)\right) = 1-\alpha$$

Si además la función $T(\theta,X)$ es monótona en θ siempre se podrán obtener los valores de θ que resuelven las ecuaciones:

$$T(\theta,X) = \tau_I(\alpha)$$
$$T(\theta,X) = \tau_S(\alpha)$$

cuyas soluciones respectivas $\theta_I(X,\alpha)$ y $\theta_S(X,\alpha)$ verifican:

$$P\left(\theta_I(X,\alpha) \le \theta \le \theta_S(X,\alpha)\right) = 1-\alpha$$

proporcionando así el intervalo de confianza buscado.

Ejemplo:

Así, en el ejemplo 1, en el que el parámetro a estimar era μ, la función pivote utilizada fue:

$$T(\mu,X) = \frac{\bar{X}-\mu}{S/\sqrt{n}}$$

que depende de μ, y cuya distribución de probabilidad es siempre una *t* de Student con *n*-1 grados de libertad (independientemente del valor de μ). En este caso, $\tau_I(\alpha) = -t_{n-1,\alpha/2}$ y $\tau_S(\alpha) = t_{n-1,\alpha/2}$, y los extremos del intervalo son $\theta_I(X,\alpha) = \bar{X} - \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}$ y $\theta_S(X,\alpha) = \bar{X} + \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}$.

3.11. INTERVALO DE CONFIANZA PARA EL PARÁMETRO DE UNA DISTRIBUCIÓN EXPONENCIAL

Vamos a ocuparnos ahora del cálculo del intervalo de confianza para estimar el parámetro λ de una distribución exponencial. Para ello supondremos que se dispone de una muestra $\{X_1, X_2, \ldots, X_n\}$ de observaciones independientes. Sabemos que si $X \approx exp(\lambda)$, entonces $E[X] = 1/\lambda$. Por tanto $\lambda = 1/E[X]$, lo que sugiere el siguiente estimador por el método de los momentos para λ:

$$\hat{\lambda} = \frac{1}{\bar{X}}$$

Hemos visto en el tema 4 que si $X_1, X_2, \ldots, X_n$ son n variables exponenciales de parámetro λ, su suma $Y = \sum_{i=1}^{n} X_i$ sigue una distribución $Gamma(n,\lambda)$. Consideremos ahora la variable:

$$V = 2\lambda Y = 2\lambda \sum_{i=1}^{n} X_i = 2\lambda n \frac{\sum_{i=1}^{n} X_i}{n} = 2n\lambda \bar{X}$$

V se obtiene a partir de Y por un simple cambio de escala (ya que resulta de multiplicar Y por la constante 2λ). Por tanto, la forma de la distribución de probabilidad de V sigue siendo una gamma, con los parámetros alterados por el cambio de escala, esto es, $V \approx Gamma(\alpha,\beta)$. Para determinar α y β observemos primero que:

$$\mu_V = \frac{\alpha}{\beta} \qquad \sigma_V^2 = \frac{\alpha}{\beta^2}$$

de donde se sigue fácilmente que:

$$\alpha = \frac{\mu_V^2}{\sigma_V^2} \qquad \beta = \frac{\mu_V}{\sigma_V^2}$$

Ahora bien, al ser $Y \approx Gamma(n,\lambda)$, se tiene que:

$$\mu_Y = \frac{n}{\lambda} \qquad \sigma_Y^2 = \frac{n}{\lambda^2}$$

y como $V = 2\lambda Y$, por las propiedades de la media y la varianza de variables aleatorias frente a cambios de escala:

$$\mu_V = 2\lambda\mu_Y = 2\lambda\frac{n}{\lambda} = 2n \qquad \sigma_V^2 = (2\lambda)^2\sigma_Y^2 = (2\lambda)^2\frac{n}{\lambda^2} = 4n$$

Por tanto:

$$\alpha = \frac{\mu_V^2}{\sigma_V^2} = \frac{4n^2}{4n} = n \qquad \beta = \frac{\mu_V}{\sigma_V^2} = \frac{2n}{4n} = \frac{1}{2}$$

Así pues:

$$V = 2\lambda n\bar{X} \approx Gamma\left(n, \frac{1}{2}\right) = \chi^2_{2n}$$

V es una función pivote en el sentido que se ha descrito en 6.5, ya que:

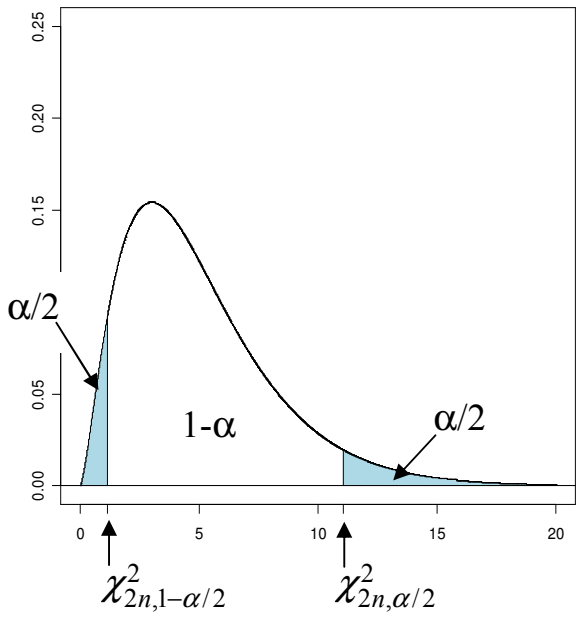
- V depende del parámetro desconocido λ,
- La distribución de probabilidad de V no depende de dicho parámetro ya que, como acabamos de ver, la distribución de V es siempre una variable χ^2 con $2n$ grados de libertad cualquiera que sea el valor de λ..

La tabla de la distribución χ^2 nos permite obtener los percentiles $\chi^2_{2n,1-\alpha/2}$ y $\chi^2_{2n,\alpha/2}$ de forma que:

$$P\left(\chi^2_{2n,1-\alpha/2} \le V \le \chi^2_{2n,\alpha/2}\right) = 1-\alpha$$

Por tanto:

$$P\left(\chi^2_{2n,1-\alpha/2} \le 2n\lambda\bar{X} \le \chi^2_{2n,\alpha/2}\right) = 1-\alpha$$



Dividiendo todos los términos del interior del intervalo por $2n\bar{X}$:

$$P\left(\frac{\chi^2_{2n,1-\alpha/2}}{2n\bar{X}} \le \lambda \le \frac{\chi^2_{2n,\alpha/2}}{2n\bar{X}}\right) = 1-\alpha$$

o lo que es lo mismo:

$$P\left(\lambda \in \left[\frac{\chi^2_{2n,1-\alpha/2}}{2n\bar{X}}, \frac{\chi^2_{2n,\alpha/2}}{2n\bar{X}}\right]\right) = 1-\alpha$$

De esta forma el intervalo de confianza a nivel 1-α para el parámetro λ de una distribución exponencial es:

$$\left[\frac{\chi^2_{2n,1-\alpha/2}}{2n\bar{X}}, \frac{\chi^2_{2n,\alpha/2}}{2n\bar{X}}\right]$$

Ejemplo: en una instalación eléctrica, cada vez que se funde un fusible, es reemplazado por otro de iguales características. El tiempo entre reemplazamientos se supone exponencial. A partir de los datos de los últimos 20 fusibles que se han reemplazado, se ha obtenido un tiempo medio entre reemplazamientos de 23 días. Se desea estimar el valor del parámetro λ, así obtener como un intervalo de confianza al 95% para dicho parámetro.

El estimador de λ es simplemente $\hat{\lambda} = \frac{1}{\bar{X}} = \frac{1}{23} = 0.0435$. En la tabla de la distribución χ^2 encontramos los valores $\chi^2_{40,0.975} = 24.433$, $\chi^2_{40,0.025} = 59.342$. Por tanto el intervalo de confianza al 95% es:

$$\left[\frac{\chi^2_{2n,1-\alpha/2}}{2n\bar{X}}, \frac{\chi^2_{2n,\alpha/2}}{2n\bar{X}}\right] = \left[\frac{24.433}{2\cdot 20\cdot 23}, \frac{59.342}{2\cdot 20\cdot 23}\right] = [0.0266, 0.0645]$$

3.12. INTERVALO DE CONFIANZA PARA EL PARÁMETRO DE UNA DISTRIBUCIÓN DE POISSON

Otra situación frecuente en la práctica es que los datos disponibles procedan de una distibución de Poisson de parámetro λ. Recuérdese que la distribución de Poisson contaba el número de sucesos discretos observados en un determinado periodo de tiempo: número de pequeñas averías diarias en una planta eléctrica, número de correos electrónicos que se reciben por hora, etc. Si $X \approx P(\lambda)$, el hecho de que E[X]=λ sugiere estimar λ mediante $\hat{\lambda} = \bar{x}$. Si se dispone de una muestra {X_1, X_2, …, X_n} de observaciones independientes de la variable de Poisson, puede demostrarse que el siguiente intervalo garantiza un nivel de confianza de al menos 1 - α para la estimación del parámetro:

$$\lambda \in \left[\frac{1}{2n}\chi^2_{n_1,1-\alpha/2}, \frac{1}{2n}\chi^2_{n_2,\alpha/2}\right], \quad n_1 = 2n\bar{x}, \quad n_2 = 2(n\bar{x}+1)$$

Ejemplo: Se realiza un estudio del número de pequeñas averías diarias que ocurren diariamente en una planta industrial. Para ello se han seleccionado al azar n=40 días del último año y se ha contado el número de pequeñas averías cada día. En total en ese periodo se

observaron 134 de estas averías. Suponiendo que el número de pequeñas averías diarias sigue una distribución de Poisson se desea estimar su parámetro con un intervalo de confianza del 95%.

Podemos estimar $\hat{\lambda} = \bar{x} = \frac{134}{40} = 3.35$. Para obtener el intervalo de confianza utilizando la expresión anterior obtenemos:

$$n_1 = 2n\bar{x} = 2 \cdot 40 \cdot \frac{134}{40} = 268, \qquad n_2 = 2\left(40 \cdot \frac{134}{40} + 1\right) = 270$$

$$\chi^2_{268,0.975} = 224.5465, \qquad \chi^2_{270,0.025} = 317.4092$$

Por tanto, el intervalo de confianza al 95% es:

$$\lambda \in \left[\frac{1}{80}224.5465, \frac{1}{80}317.4092\right] = [2.80683,\ 3.96762]$$

3.13. OTROS INTERVALOS DE CONFIANZA

3.13.1. Intervalo de confianza para el cociente de varianzas de poblaciones normales

Hemos visto que, dadas dos muestras de tamaños respectivos n1 y n2 de dos poblaciones normales, el cociente de sus varianzas sigue la distribución:

$$\frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \approx F_{n_1-1,n_2-1}$$

Ahora, como:

$$P\left(F_{n_1-1,n_2-1,1-\alpha/2} \le F_{n_1-1,n_2-1} \le F_{n_1-1,n_2-1,\alpha/2}\right) =$$

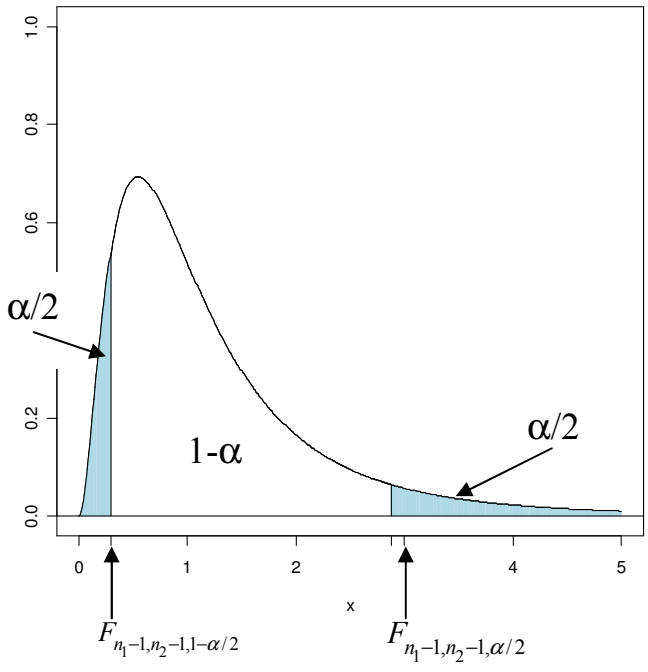
$$= 1-\alpha$$

Se tiene que:

$$P\left(F_{n_1-1,n_2-1,1-\alpha/2} \le \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \le F_{n_1-1,n_2-1,\alpha/2}\right) = 1-\alpha$$

De donde, despejando el cociente de varianzas poblacionales en el centro del intervalo:

$$P\left(\frac{S_2^2}{S_1^2}F_{n_1-1,n_2-1,1-\alpha/2} \le \frac{\sigma_2^2}{\sigma_1^2} \le \frac{S_2^2}{S_1^2}F_{n_1-1,n_2-1,\alpha/2}\right) = 1-\alpha$$



y si tenemos en cuenta que:

$$F_{n,m,\alpha} = \frac{1}{F_{m,n,1-\alpha}},$$

podemos expresar el intervalo de confianza a nivel 1-α para el cociente de varianzas de poblaciones normales como:

$$\frac{\sigma_2^2}{\sigma_1^2} \in \left(\frac{S_2^2/S_1^2}{F_{n_2-1,n_1-1,\alpha/2}}, \frac{S_2^2/S_1^2}{F_{n_2-1,n_1-1,1-\alpha/2}} \right)$$

3.13.2. Intervalos de confianza para la diferencia de medias de poblaciones normales.

Supondremos que se dispone de una muestra de tamaño n_1 de una variable aleatoria $N(\mu_1,\sigma_1)$, y de otra muestra de tamaño n_2 de otra variable $N(\mu_2,\sigma_2)$, en las que se han estimado las medias muestrales respectivas, $\bar{x}_1$ y $\bar{x}_2$, así como las varianzas muestrales respectivas s_1^2, s_2^2

3.13.2.1. Muestras Independientes: Varianzas conocidas.

Sabiendo que:

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx N(0,1)$$

podemos proceder de modo análogo al caso del intervalo de confianza para la media de una población normal con varianza conocida obtenemos el intervalo a nivel 1-α:

$$\left((\bar{X}_1 - \bar{X}_2) \mp z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

3.13.2.2. Muestras Independientes: Varianzas desconocidas e iguales.

En este caso

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \approx t_{n_1+n_2-2}, \text{ siendo } s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

Procediendo de modo análogo al caso del intervalo de confianza para la media de una población normal con varianza desconocida obtenemos el intervalo a nivel 1-α:

$$\left((\bar{X}_1 - \bar{X}_2) \mp t_{n_1+n_2-2,\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

3.13.2.3. Muestras Independientes: Varianzas desconocidas y distintas.

Como $\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx t_n$, siendo $n = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\left(\frac{s_1^2}{n_1} \right)^2 \frac{1}{n_1 - 1} + \left(\frac{s_2^2}{n_2} \right)^2 \frac{1}{n_2 - 1}}$

el intervalo de confianza a nivel 1-α será:

$$\left(\left(\bar{X}_1 - \bar{X}_2 \right) \mp t_{n,\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$

3.13.2.4. Muestras emparejadas.

En este caso:

$$\frac{\left(\bar{X}_1 - \bar{X}_2 \right) - \left(\mu_1 - \mu_2 \right)}{s_d / \sqrt{n}} \approx t_{n-1}, \text{ siendo } s_d = \sqrt{\frac{\sum_{i=1}^{n} \left(\left(X_{1i} - X_{2i} \right) - \left(\bar{X}_1 - \bar{X}_2 \right) \right)^2}{n-1}} = \sqrt{s_1^2 + s_2^2 - 2 r s_1 s_2}$$

y r el coeficiente de correlación entre X_1 y X_2. El intervalo de confianza a nivel 1-α es:

$$\left(\left(\bar{X}_1 - \bar{X}_2 \right) \mp t_{n-1,\alpha/2} \frac{s_d}{\sqrt{n}} \right)$$

3.13.3. Intervalo de confianza para la diferencia de proporciones en poblaciones independientes.

Supondremos que se dispone de una muestra de tamaño n_1 de una variable aleatoria de Bernoulli $b(p_1)$ y de otra muestra de tamaño n_2 de otra variable también de Bernoulli $b(p_2)$, habiéndose extraído ambas muestras en poblaciones independientes (por ejemplo p_1 puede ser la proporción de machos en cierta especie de tortugas, y p_2 puede ser dicha proporción entre las tortugas de otra especie). Sean n_{E1} y n_{E2} el número de éxitos observados, respectivamente en cada una de las muestras. Las proporciones poblacionales pueden estimarse entonces como:

$$\hat{p}_1 = \frac{n_{E1}}{n_1} \qquad \hat{p}_2 = \frac{n_{E2}}{n_2}$$

Si se cumple que:

$$\hat{p}_1 > 0.05,\ (1-\hat{p}_1) > 0.05,\ n_1 \hat{p}_1 > 20,\ n_1 (1-\hat{p}_1) > 20$$
$$\hat{p}_2 > 0.05,\ (1-\hat{p}_2) > 0.05,\ n_2 \hat{p}_2 > 20,\ n_2 (1-\hat{p}_2) > 20$$

de acuerdo con el teorema central del límite:

$$\hat{p}_1 \cong N\left(p_1, \sqrt{\frac{\hat{p}_1 (1-\hat{p}_1)}{n_1}} \right), \quad \hat{p}_2 \cong N\left(p_2, \sqrt{\frac{\hat{p}_2 (1-\hat{p}_2)}{n_2}} \right)$$

y por tanto:

$$\hat{p}_1 - \hat{p}_2 \cong N\left(p_1 - p_2, \sqrt{\frac{\hat{p}_1 (1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2 (1-\hat{p}_2)}{n_2}} \right)$$

De esta forma:

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

con lo que el intervalo de confianza a nivel 1-α para la diferencia de proporciones será:

$$\left(\hat{p}_1-\hat{p}_2\pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p})_1}{n}+\frac{\hat{p}_2(1-\hat{p})_2}{n}}\right)$$

Este intervalo mejora añadiendo una corrección por continuidad:

$$\left(\hat{p}_1-\hat{p}_2\pm\left[z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}+\frac{1}{4}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right]\right)$$

3.13.4. Intervalo de confianza para el cociente de proporciones en poblaciones independientes.

Una manera habitual de comparar dos proporciones medidas en poblaciones independientes es mediante el *riesgo relativo*, definido como:

$$\rho=\frac{p_1}{p_2}$$

Si se dispone de sendas muestras en las poblaciones de referencia, el riesgo relativo puede estimarse como:

$$\hat{\rho}=\frac{\hat{p}_1}{\hat{p}_2}$$

Puede probarse que:

$$\operatorname{var}\left(\log\hat{\rho}\right)\cong\frac{1-\hat{p}_1}{n_1\hat{p}_1}+\frac{1-\hat{p}_2}{n_2\hat{p}_2}$$

y que, aproximadamente:

$$\log\rho\approx N\left(\log\hat{\rho},\sqrt{\frac{1-\hat{p}_1}{n_1\hat{p}_1}+\frac{1-\hat{p}_2}{n_2\hat{p}_2}}\right)$$

por lo que el intervalo de confianza para *log* ρ a nivel 1-α se obtiene sencillamente como:

$$\log\rho\in\left[\log\hat{\rho}\pm z_{\alpha/2}\sqrt{\frac{1-\hat{p}_1}{n_1\hat{p}_1}+\frac{1-\hat{p}_2}{n_2\hat{p}_2}}\right]$$

y en la escala original en que se mide el cociente de proporciones:

$$\rho\in\exp\left[\log\hat{\rho}\pm z_{\alpha/2}\sqrt{\frac{1-\hat{p}_1}{n_1\hat{p}_1}+\frac{1-\hat{p}_2}{n_2\hat{p}_2}}\right]$$

3.14. CONTRASTES DE HIPÓTESIS

Hasta ahora, en el ámbito de la **inferencia estadística**, nos hemos ocupado solamente del problema de la estimación de parámetros: cómo utilizar una muestra para obtener un valor próximo al de un parámetro característico de una población de interés, y cómo hallar un intervalo dentro del cual podamos estar *razonablemente seguros* (con un nivel de confianza predeterminado) de que se encuentra el valor real de dicho parámetro.

Sin embargo muchos de los problemas con que se encuentra un científico al investigar un fenómeno o situación de interés **se plantean en términos de tener que decidir sobre la veracidad o falsedad de alguna conjetura o hipótesis**. Esta hipótesis:

- Puede haber sido sugerida por una serie de observaciones repetidas de un fenómeno o situación de interés.
- Puede haber sido deducida de un modelo teórico.
- Puede deberse a la intuición del investigador, que trata con ello de explicar algunos hechos observados.
- Puede simplemente reflejar las condiciones en que debe desarrollarse cierto proceso (natural o artificial) y que hay que comprobar si se cumplen o no.
- ....

Por ejemplo:

- Se ha diseñado un nuevo método de depuración de agua, cuyas características físico-químicas inducen a suponer que reducirán la concentración de ciertos contaminantes biológicos con mayor eficiencia que el método que se venía usando hasta ahora. ¿Será verdad esta suposición?
- Se cree que cierto compuesto químico actúa sobre los peces que se crían en tanques de cultivo, reduciendo los niveles de estrés que presentan estos animales al tener que compartir un espacio reducido con un elevado número de congéneres. ¿Es cierta esta conjetura?
- Diversos estudios argumentan que la distribución espacial de cierta especie vegetal se encuentra fuertemente asociada a la presencia de determinados rumiantes. ¿Existe realmente dicha asociación?
- Un método de análisis químico A es mucho más caro que otro método B, pero ¿es realmente mucho más preciso?
- ¿Ha aumentado realmente el gasto medio por turista en las Islas Canarias con respecto a hace diez años?
- ¿La tasa de mortalidad en cultivos marinos realizados en tanques cerrados es superior a la que se produce en cultivos en mar abierto?

Todos estos ejemplos se caracterizan por describir situaciones en las que es imposible realizar un experimento u observación que nos confirme o desmienta *de una manera absolutamente segura* la hipótesis planteada. De ahí que los procedimientos para tomar decisiones sobre la veracidad o falsedad de estas hipótesis hayan de ser necesariamente procedimientos estadísticos, que permitan calcular la probabilidad de que se tomen decisiones erróneas.

Una **hipótesis estadística** es una afirmación o conjetura con respecto a una o más poblaciones. Se llamará **hipótesis nula (H_0)** a la hipótesis de partida, que se desea poner a prueba para decidir sobre su validez, e **hipótesis alternativa (H_1)** a la hipótesis que será aceptada en caso de que se rechace H_0.

Las hipótesis estadísticas pueden plantearse de muy diversas formas:

- En función de los **parámetros** de una o varias poblaciones: ¿el valor medio de cierta variable en una población es cero?, ¿son iguales las medias de dos poblaciones?, ¿la proporción de sujetos con cierta característica supera el 70% de la población?.
- En términos de la **forma de la distribución** de la variable de interés: ¿se distribuye una variable de igual forma en dos poblaciones?, ¿es normal la distribución de una variable?.
- En términos de **características de asociación:** ¿son dos variables independientes?, ¿la relación entre dos variables es lineal?

Un **contraste o test de hipótesis estadístico** es un método que nos permita decidirnos por H_0 o por H_1 en función de la evidencia disponible y del margen de error que estemos dispuestos a aceptar.

Tipos de hipótesis estadísticas

Una hipótesis es **simple** cuando propone un valor concreto para un parámetro. En caso contrario se dice que es **compuesta.**

Ejemplo 1:

$$\left.\begin{array}{l} H_0 : p = 0.4 \\ H_1 : p \neq 0.4 \end{array}\right\} \begin{array}{l} Simple \\ Comp. \end{array}$$

Ejemplo 2:

$$\left.\begin{array}{l} H_0 : \mu \leq 15 \\ H_1 : \mu > 15 \end{array}\right\} \begin{array}{l} Comp. \\ Comp. \end{array}$$

Ejemplo 3:

$$\left.\begin{array}{l} H_0 : p = 0.4 \\ H_1 : p = 0.7 \end{array}\right\} \begin{array}{l} Simple \\ Simple \end{array}$$

Ejemplo 4:

$$\left.\begin{array}{l} H_0 : p = 0.4 \\ H_1 : p > 0.4 \end{array}\right\} \begin{array}{l} Simple \\ Comp. \end{array}$$

Tipos de Error en los contrastes de hipótesis.

La verdad o falsedad de una hipótesis estadística no puede conocerse con absoluta certeza a menos que examinemos toda la población. Dado que observar toda la población es usualmente imposible, en la práctica deberemos conformarnos con la información que nos aporte una muestra de la misma. Dado que además esta información será en general incompleta, debemos ser conscientes de que todo contraste de hipótesis estadístico es susceptible de cometer dos tipos de error:

- **ERROR TIPO I: Rechazar una hipótesis nula que es verdadera** (y por tanto aceptar una hipótesis alternativa falsa)
- **ERROR TIPO II: Aceptar una hipótesis nula que es falsa** (y por tanto rechazar una hipótesis alternativa verdadera)

En general, llamaremos:

$$\alpha = P(\text{Error Tipo I}) = P(\text{Rechazar } H_0 / H_0 \text{ es cierta})$$
$$\beta = P(\text{Error Tipo II}) = P(\text{Aceptar } H_0 / H_0 \text{ es falsa})$$

Así pues, las situaciones posibles al realizar un contraste de hipótesis son:

	H_0 cierta	**H_0 falsa**
Aceptar H_0	Decisión correcta	Error tipo II
Rechazar H_0	Error tipo I	Decisión correcta

3.15. CONTRASTES DE SIGNIFICACIÓN

Aunque existe una teoría general debida a Neyman y Pearson para el planteamiento y resolución de contrastes de hipótesis estadísticos de manera óptima, en este curso nos limitaremos a estudiar el caso particular de los **contrastes o pruebas de significación**, desarrollados por Fisher, que dan lugar a los métodos de contraste de hipótesis más extendidos en la práctica científica habitual.

El procedimiento general de las pruebas de significación es el siguiente:

1. Fijar las hipótesis nula (H_0) y alternativa (H_1) y la probabilidad de error de tipo I, α.
2. Determinar un **estadístico de contraste** $\theta(X_1,X_2, ...,X_n)$ que mida de algún modo la discrepancia existente entre lo que especifica en H_0 y lo que se observe en una eventual muestra aleatoria $X_1,X_2, ...,X_n$. La distribución de probabilidad de θ ha de ser conocida cuando H_0 es cierta.
3. Determinar el conjunto R_A de valores más probables de θ cuando H_0 es cierta (**región de aceptación**), de tal forma que:
$$P(\theta \in R_A / H_0 \text{ es cierta}) = 1 - \alpha$$
4. Obtener una muestra representativa de la población a la que se refieren las hipótesis, y evaluar el valor θ_{exp} que toma $\theta(X_1,X_2, ...,X_n)$ en dicha muestra.
5. Si $\theta_{exp} \in R_A$ aceptar H_0. En caso contrario rechazar H_0 y aceptar H_1.

Nótese que, con la regla de decisión especificada en el punto 5, rechazar H_0 equivale a que $\theta_{exp} \notin R_A$. Así, la región de aceptación especificada en 3 implica que:

$$P(\text{Rechazar } H_0 / H_0 \text{ cierta}) = P(\theta_{exp} \notin R_A / H_0 \text{ cierta}) = \alpha,$$

y por tanto este método garantiza que la probabilidad de cometer un error de tipo I es como mucho α.

La lógica de este método es evidente: si H_0 fuera cierta, lo más probable sería que el valor de θ observado (el θ_{exp}) cayera en la región de aceptación. Si una vez tomada la muestra es justo eso lo que ocurre, entonces debemos concluir que no hay evidencia de que la hipótesis nula sea falsa, ya que se observa algo que era probable que ocurriese si fuera cierta. Por tanto se acepta H_0. Si por el contrario, al tomar la muestra el valor de θ_{exp} cae fuera de la región de aceptación, se estaría observando un valor

muy improbable de θ en caso de ser H_0 cierta. Ello significaría que hemos encontrado evidencia de que H_0 es probablemente falsa y por tanto se rechaza.

Ejemplo:

Las algas de cierta especie que se cultivan con fines farmacológicos son muy sensibles al pH del agua. Se ha observado que el desarrollo de estas algas es óptimo cuando el pH promedio es 1, y diariamente se realizan controles con el objetivo de aplicar medidas correctoras (añadir aditivos químicos al agua) si el pH se aparta de este valor. Estos controles consisten en tomar 5 muestras de agua y evaluar el pH en cada una. En un día en que el pH medio de las cinco muestras es de 1.2 con una desviación típica de 0.4.¿sería preciso aplicar alguna medida correctora? (se supone que la distribución del pH es normal)

1. Si llamamos μ al pH medio real del agua, el problema puede plantearse como una decisión entre las hipótesis:

$$\begin{cases} H_0 : \mu = 1 \\ H_1 : \mu \neq 1 \end{cases}$$

2. Dado que el mejor estimador de la media poblacional μ es la media muestral $\bar{X}$, podemos medir la discrepancia entre la hipótesis nula (μ=1) y lo observado en la muestra mediante la diferencia $|\bar{X}$ -1|. En efecto, si $\bar{X}$ es parecida a 1, la muestra es poco discrepante con H_0; cuanto más lejos esté $\bar{X}$ de 1 mayor será esta discrepancia. Pero ¿cómo de lejos debe estar $\bar{X}$ de 1 para rechazar H_0 ?

   Para responder a esta pregunta observemos que aunque $|\bar{X}$ -1| no tenga directamente una distribución de probabilidad conocida, sabemos que si H_0 es cierta, entonces $\dfrac{\bar{X}-1}{s/\sqrt{5}} \approx t_4$.

   Obviamente este estadístico es también una medida de discrepancia entre la muestra y H_0, ya que su valor absoluto aumenta cuanto más lejos esté la media muestral de la media hipotética μ=1. Por tanto, nuestro estadístico de contraste para este problema será:

$$\theta(X_1, X_2, ..., X_n) = \left| \frac{\bar{X}-1}{s/\sqrt{5}} \right|$$

3. Para determinar el conjunto de valores más probables de este estadístico cuando H_0 es cierta, si elegimos α=0.05, bastará determinar el δ tal que:

$$P\left(\left| \frac{\bar{X}-1}{s/\sqrt{5}} \right| \leq \delta \Big/ H_0 \text{ es cierta} \right) = 0.95$$

   Cuando H_0 es cierta $\Rightarrow \dfrac{\bar{X}-1}{s/\sqrt{5}} \approx t_4$, y por tanto $\delta = t_{4,0.025} = 2.776$

   Así pues, la región de aceptación para nuestro contraste es:

$$R_A = \{\theta / |\theta| \leq 2.776\} = [-2.776, 2.776]$$

Dicho de otra forma, si H_0 es cierta, lo más probable (con probabilidad 0.95) es que el valor que se observe en el estadístico de contraste caiga en el intervalo [-2.776, 2.776]. Al mismo tiempo garantizamos que la probabilidad de cometer un error tipo I es como mucho 0.05.

4. En nuestra muestra:

$$\theta_{\text{exp}} = \frac{\bar{X} - 1}{s/\sqrt{5}} = \frac{1.2 - 1}{0.4/\sqrt{5}} = 1.11$$

5. Como $\theta_{\text{exp}} \in R_A$ aceptamos H_0. El valor observado está dentro de lo que era muy probable observar si H_0 fuera cierta. Por tanto no existe evidencia para rechazar H_0.

Otra forma alternativa y equivalente de plantear las pruebas de significación es la siguiente:

1. Fijar las hipótesis nula (H_0) y alternativa (H_1), y la probabilidad de error de tipo I, α.
2. Determinar un **estadístico de contraste** $\theta(X_1, X_2, \ldots, X_n)$ que mida la discrepancia entre la muestra y H_0, y cuya distribución de probabilidad sea conocida cuando H_0 es cierta.
3. Obtener una muestra representativa de la población a la que se refieren las hipótesis, y evaluar el valor θ_{exp} que toma $\theta(X_1, X_2, \ldots, X_n)$ en dicha muestra.
4. Calcular la probabilidad de obtener un valor tanto o más extraño que θ_{exp} si H_0 fuera cierta (esta probabilidad suele denominarse **p-valor**).
5. Si esta probabilidad es alta (**p-valor ≥ α**) se acepta H_0; en caso contrario (**p-valor < α**) se rechaza H_0 y se acepta la alternativa H_1.

En la práctica para aplicar las pruebas de significación de esta segunda forma suele ser preciso el uso del ordenador, ya que el cálculo de la probabilidad señalada en el punto 4 habitualmente requiere el cálculo de sumatorias o integrales no inmediatas.

Ejemplo:

En el ejemplo anterior:

$$\theta(X_1, X_2, \ldots, X_n) = \left|\frac{\bar{X} - 1}{s/\sqrt{5}}\right|, \quad \theta_{\text{exp}} = 1.11$$

Entonces:

$$P(\theta > \theta_{\text{exp}} / H_0 \text{ cierta}) = P\left(\left|\frac{\bar{X} - 1}{s/\sqrt{5}}\right| > 1.11\right) = P(|t_4| > 1.11) = 2P(t_4 > 1.11) =$$
$$= 2 \cdot 0.164 = 0.328$$

Por tanto, cuando H_0 es cierta es bastante probable (0.328) que el estadístico de contraste valga al menos 1.11. Así pues, no hay evidencia para rechazar la hipótesis nula.

Relación entre los dos tipos de error en los contrastes de hipótesis.

A la probabilidad α de cometer un error de tipo I, rechazar la hipótesis nula cuando es cierta, se la llama **nivel de significación** del contraste. Este error se fija a priori, y depende de la gravedad o importancia que tenga el rechazar una hipótesis nula que es verdadera.

A la probabilidad $1-\beta = P\left(\theta_{\exp} \notin R_A / H_0 \text{ falsa}\right)$ se la llama **potencia** del contraste, y representa la probabilidad de rechazar la hipótesis nula cuando es falsa. La probabilidad β, complementaria de la potencia, representa la probabilidad de cometer un error de tipo II, esto es, aceptar una hipótesis nula falsa.

En general, el comportamiento de ambos tipos de error es inverso: **para un tamaño de muestra fijo, disminuir el valor de α significa un incremento de β; y viceversa, una disminución en β lleva aparejado un incremento en α.**

La razón de este comportamiento es bastante intuitiva: disminuir α, debido al modo de construcción del contraste, significa tomar una región de aceptación más grande; y si la región de aceptación se hace más grande será más fácil también aceptar H_0 cuando sea falsa, esto es, cometer un error de tipo II, por lo que β será mayor.

Si se desea que tanto α como β disminuyan, será en general necesario incrementar el tamaño de la muestra (en otras palabras, aumentar la cantidad de información disponible).

Ejemplo:

En el ejemplo anterior, para el contraste:

$$\begin{cases} H_0 : \mu = 1 \\ H_1 : \mu \neq 1 \end{cases}$$

hemos aceptado la hipótesis nula (μ=1) aún cuando la media muestral era 1.2. ¿Cuál es la probabilidad β de cometer un error de tipo II en este contraste, esto es, cuál es la probabilidad de aceptar H_0 siendo falsa?

Para calcular esta probabilidad hemos de observar que si H_0 es falsa es porque el valor medio poblacional del pH es un valor μ^* **distinto de 1**. Por tanto el valor de β depende de cuanto valga μ^* o, dicho de otra forma, **β es una función del verdadero valor μ^* de la media poblacional**. Además hemos de tener en cuenta que si la verdadera media de la población es μ*≠1, el estadístico $\frac{\bar{X}-1}{s/\sqrt{5}}$ ya no sigue una distribución *t* de Student, y sí lo hará el estadístico $\frac{\bar{X}-\mu^*}{s/\sqrt{5}}$.

Por tanto:

$$\beta(\mu^*) = P(\text{Aceptar } H_0 / H_0 \text{ falsa porque } \mu = \mu^*) =$$

$$= P\left(\left|\frac{\bar{X}-1}{s/\sqrt{5}}\right| \le 2.776 \Big/ \mu = \mu^*\right) = P\left(-2.776 \le \frac{\bar{X}-1}{s/\sqrt{5}} \le 2.776 \Big/ \mu = \mu^*\right) =$$

$$= P\left(-2.776 \le \frac{\bar{X}-\mu^*+\mu^*-1}{s/\sqrt{5}} \le 2.776 \Big/ \mu = \mu^*\right) =$$

$$= P\left(-2.776 - \frac{\mu^*-1}{s/\sqrt{5}} \le \frac{\bar{X}-\mu^*}{s/\sqrt{5}} \le 2.776 - \frac{\mu^*-1}{s/\sqrt{5}} \Big/ \mu = \mu^*\right) =$$

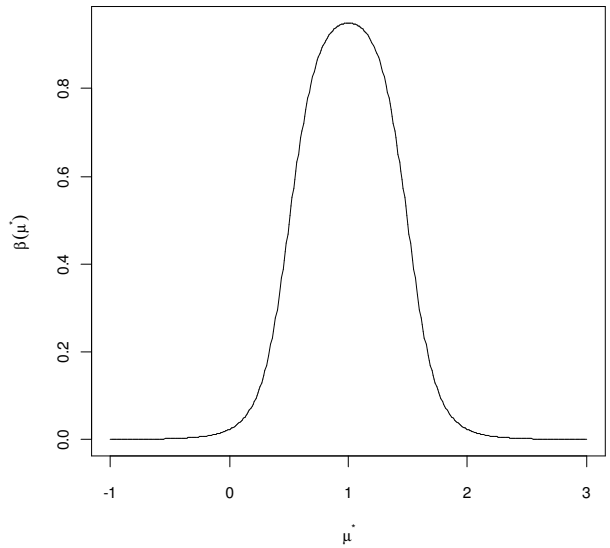
$$= P\left(-2.776 - \frac{\mu^*-1}{s/\sqrt{5}} \le t_4 \le 2.776 - \frac{\mu^*-1}{s/\sqrt{5}}\right) =$$

$$= P\left(t_4 \ge -2.776 - \frac{\mu^*-1}{s/\sqrt{5}}\right) - P\left(t_4 \ge 2.776 - \frac{\mu^*-1}{s/\sqrt{5}}\right)$$

En la tabla siguiente mostramos los valores de β(μ*) para diversos valores de μ^* (con s=0.4):

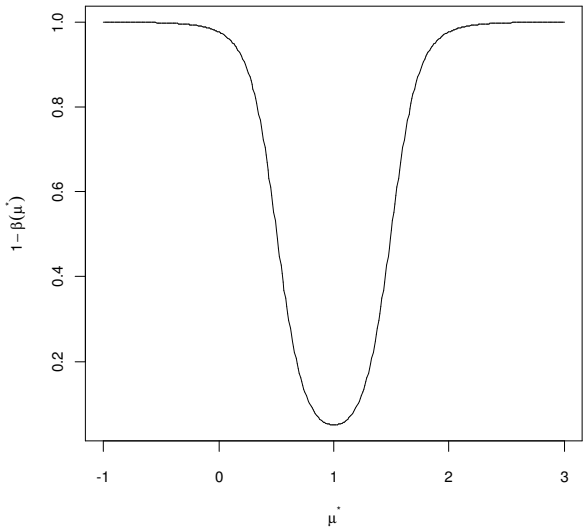
μ*	0.0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
β(μ*)	0.0235	0.0816	0.295	0.687	0.905	0.95	0.905	0.687	0.295	0.0816	0.0235

Gráficamente:



Como puede apreciarse, el error β(μ*) es tanto mayor cuanto más próximo esté μ^* a 1, alcanzando su máximo cuando μ^* coincide con el valor especificado en la hipótesis nula (μ^*=1), en cuyo caso se tiene que β(1)=1-α=0.95. La razón de este comportamiento de β es bastante intuitiva: en nuestro contraste estamos tratando de decidir si la verdadera media de la población es 1; será más fácil equivocarse aceptando que es 1 cuando realmente es 1.05 o 1.1 (el verdadero valor μ^* está cerca de 1) que cuando la verdadera media es un valor más alejado de 1, como el 2 ó el 3.

Del mismo modo que hemos hecho un gráfico del error tipo II, podemos hacer un gráfico de la potencia de este test, su capacidad para rechazar H_0 cuando es falsa porque la verdadera media es $\mu^* \neq 1$:



Aquí vemos que la potencia es mínima en $\mu^*=1$, y va aumentando a medida que μ^* se aleja de este valor.

Otra forma de entender la relación entre α y β en este problema (y también en casos más generales) consiste en observar que si H_0 es cierta, entonces:

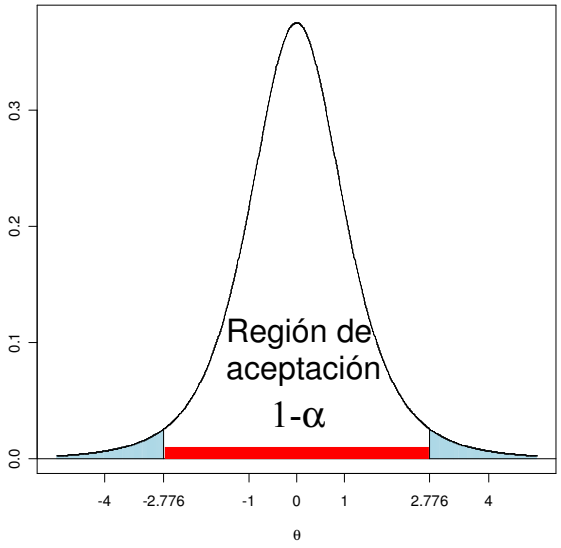
$$\theta(X_1, X_2, ..., X_n) = \frac{\bar{X} - 1}{s/\sqrt{5}} \approx t_4$$

y como la región de aceptación debe obtenerse de forma que:

$$P\left(\theta(X_1, X_2, ..., X_n) \in R_A / H_0 \text{ cierta}\right) = 1 - \alpha$$

resulta que para α=0.05:

$$R_A = \left[-t_{4,\alpha/2}, t_{4,\alpha/2}\right] = \left[-2.776, 2.776\right]$$

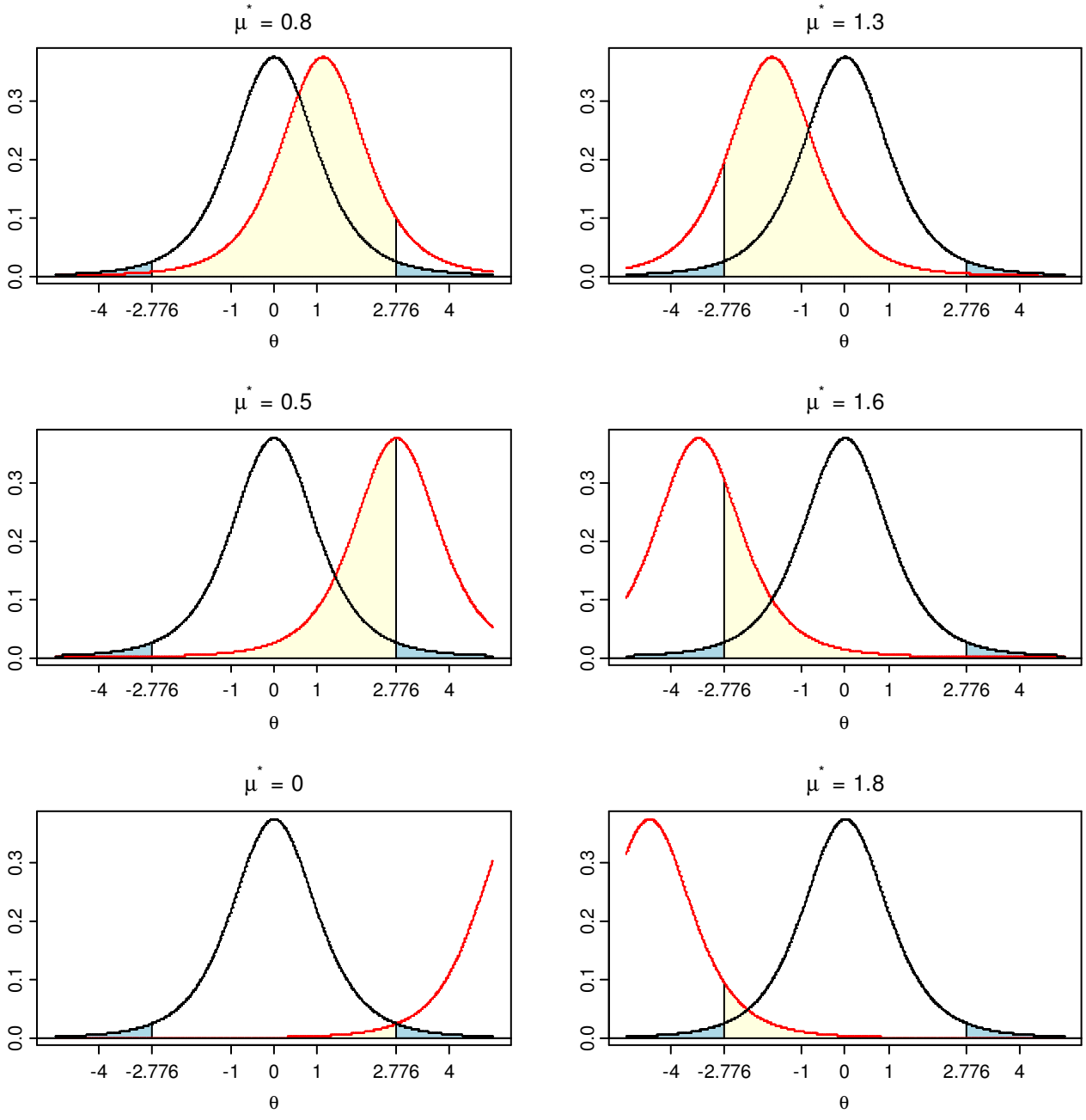


Ahora bien, si H_0 es falsa, la probabilidad de un error tipo II cuando la verdadera media es μ^* es:

$$\beta(\mu^*) = P\left(-2.776 - \frac{\mu^* - 1}{s/\sqrt{5}} \le t_4 \le 2.776 - \frac{\mu^* - 1}{s/\sqrt{5}}\right) = P\left(-2.776 \le t_4 + \frac{\mu^* - 1}{s/\sqrt{5}} \le 2.776\right)$$

esta probabilidad corresponde al área entre –2.776 y 2.776 bajo la curva de la t de Student con 4 grados de libertad desplazada en una cantidad $\frac{\mu^* - 1}{s/\sqrt{5}}$.

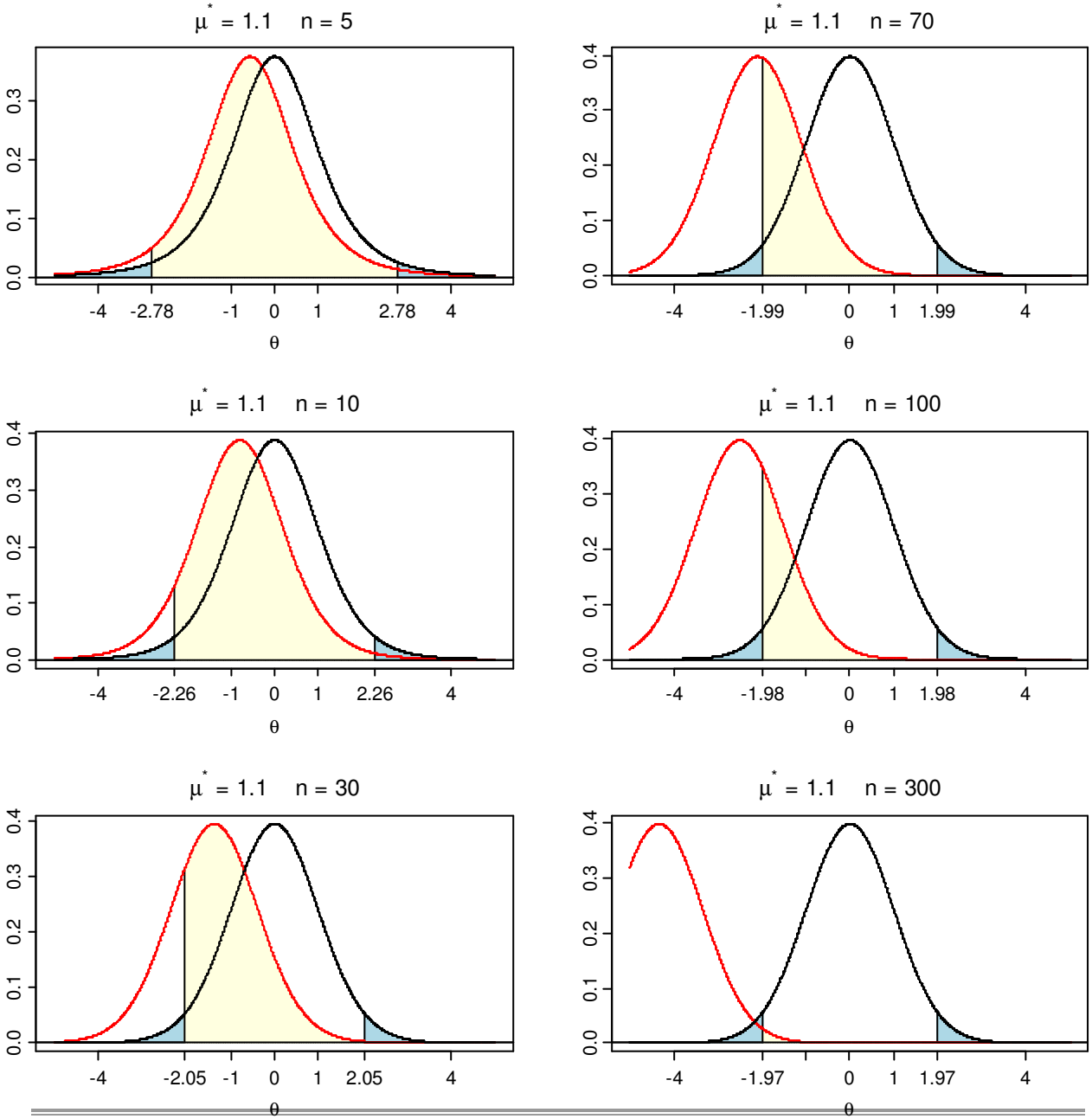
En las gráficas siguientes hemos dibujado esta área para varios valores de la alternativa μ^*. Como puede apreciarse, cuanto más próxima está la alternativa a 1, mayor es el área (mayor es el error tipo II). Cuánto más se aleja, más disminuye el área:



Para apreciar ahora la relación entre los dos tipos de error y el tamaño de la muestra, notemos que la probabilidad de error II para una muestra de tamaño *n*, y para un error tipo I con valor α fijo es:

$$\beta(\mu^*) = P\left(-t_{n-1,\alpha/2} \le t_{n-1} + \frac{\mu^* - 1}{s/\sqrt{n}} \le t_{n-1,\alpha/2}\right)$$

que corresponde al área entre $-t_{n-1,\alpha/2}$ y $t_{n-1,\alpha/2}$ bajo la curva de la t de Student con *n-1* grados de libertad desplazada en una cantidad $\frac{\mu^* - 1}{s/\sqrt{n}} = \frac{\mu^* - 1}{s}\sqrt{n}$. Por tanto, tal como se aprecia en las gráficas, para un valor de α fijo cuanto mayor es *n,* mayor es el desplazamiento y menor es el área β (en amarillo en las gráficas) correspondiente al error II:



Tratamiento asimétrico de las hipótesis en un contraste

1. El tratamiento de las hipótesis nula H_0 y alternativa H_1 no es simétrico: **cuando la hipótesis nula H_0 se acepta no es porque dicha hipótesis sea verdadera, sino porque no existe una evidencia clara de que sea falsa.**

2. Asimismo, cuando H_0 se rechaza es porque hay una fuerte evidencia en su contra. Por tanto, si H_1 especifica lo contrario de H_0, **H_1 se acepta con una fuerte evidencia en su favor.**

3. Por esta razón, cuando en un contraste se acepta la hipótesis nula se dice que el contraste ha resultado **no significativo.** En caso contrario el contraste es **significativo.**

Esta asimetría debe tenerse muy en cuenta al plantear un contraste de hipótesis, sobre todo si el tamaño de la muestra no es demasiado grande. Así, por ejemplo, en los dos siguientes contrastes, aparentemente equivalentes:

Contraste 1	Contraste 2
$\begin{cases} H_0 : \mu < \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$	$\begin{cases} H_0 : \mu > \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$

(la hipótesis nula del primer contraste es la alternativa del segundo y viceversa), en caso de aceptar la hipótesis nula en el primer contraste, lo haríamos por no haber evidencia en su contra; la misma hipótesis, en caso de ser aceptada como alternativa en el segundo contraste, lo sería por tener fuerte evidencia a su favor.

Por tanto, si al poner a prueba una hipótesis queremos, en caso de aceptarla, hacerlo con mucha evidencia a su favor, deberemos plantear tal hipótesis como hipótesis alternativa del contraste.

Si observamos que:

Contraste 1	$\alpha_1 = P(\text{Rechazar } \mu < \mu_0 / \text{Realmente es } \mu < \mu_0)$ $\beta_1 = P(\text{Aceptar } \mu < \mu_0 / \text{Realmente es } \mu > \mu_0)$
Contraste 2	$\alpha_2 = P(\text{Rechazar } \mu > \mu_0 / \text{Realmente es } \mu > \mu_0)$ $\beta_2 = P(\text{Aceptar } \mu > \mu_0 / \text{Realmente es } \mu < \mu_0)$

nos daremos cuenta que la causa de lo anterior es que, aunque α esté controlado (se fija siempre a priori en un valor pequeño), tal como ya hemos visto la probabilidad β de aceptar la hipótesis nula siendo falsa puede ser alta; por tanto la hipótesis que se acepta como nula puede llegar a tener una probabilidad alta de ser falsa. Sin embargo, si la hipótesis nula se rechaza (la alternativa se acepta), la probabilidad de que sea verdadera (la alternativa falsa) es α que, por construcción del contraste, es una cantidad pequeña (habitualmente 0.01, 0.05 ó 0.1).

Ejemplo:

Si en nuestro ejemplo, las algas se desarrollan bien si $\mu > 1$, pero mueren si $\mu \leq 1$, el contraste de hipótesis adecuado sería:

$$\begin{cases} H_0 : \mu \leq 1 \\ H_1 : \mu > 1 \end{cases}$$

ya que en caso de aceptar H_0 tomaríamos alguna medida correctora para garantizar que $\mu \geq 1$, y en caso de rechazar H_0, estaríamos bastante seguros de que las algas van a desarrollarse bien.

Si el tamaño de la muestra es lo suficientemente grande como para que α y β sean pequeños, es indiferente cual hipótesis se coloca como nula y cual como alternativa.

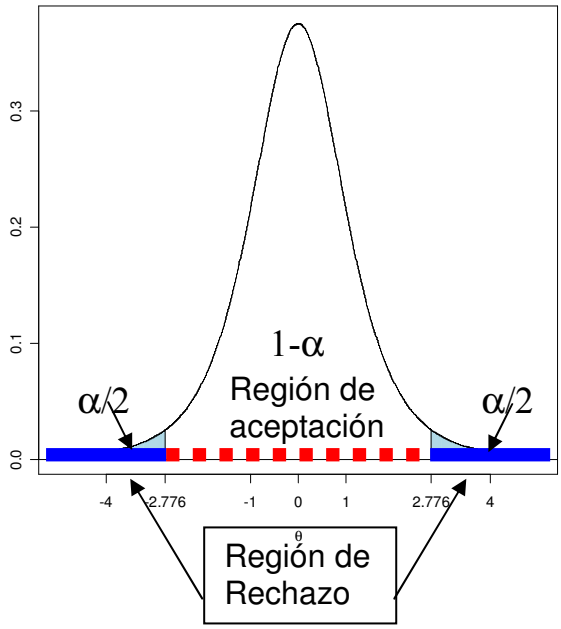
Contrastes unilaterales y contrastes bilaterales.

Al conjunto complementario de la región de aceptación en un contraste de hipótesis se le llama **región de rechazo** o **región crítica**.

En el contraste de hipótesis que hemos estado utilizando hasta ahora como ejemplo:

$$\begin{cases} H_0 : \mu = 1 \\ H_1 : \mu \neq 1 \end{cases}$$

la región de aceptación era [-2.776, 2.776]. Por tanto, la región crítica es $(-\infty, -2.776) \cup (2.776, \infty)$. Los contrastes cuya región crítica es de esta forma (dos intervalos correspondientes a las colas de una distribución) se denominan **contrastes bilaterales.**



Cuando la región crítica se sitúa a un solo lado, el contraste se denomina **unilateral**.

Los contrastes unilaterales corresponden a contrastes de hipótesis del tipo:

$$\begin{cases} H_0 : \mu \le 1 \\ H_1 : \mu > 1 \end{cases}$$

Ejemplo:

Si suponemos que las algas de nuestro ejemplo se desarrollan bien si μ > 1, pero mueren si μ ≤ 1, siendo μ el pH medio del agua del tanque de cultivo, el contraste de hipótesis adecuado sería el anterior. Si en 7 análisis de agua hemos obtenido un valor medio de pH $\bar{x}$ = 1.1, con desviación típica 0.3, ¿hay evidencia suficiente para rechazar H_0?

Ahora como medida de discrepancia entre la hipótesis nula y la muestra observada podemos elegir el valor de $\bar{X} - 1$. *Cuanto mayor y más positivo sea este valor, más discrepante será la muestra con* H_0. *El mismo argumento puede aplicarse a la cantidad:*

$$\theta(X_1, X_2, ..., X_n) = \frac{\bar{X} - 1}{s/\sqrt{n}}$$

que en el caso particular de que la hipótesis nula sea verdadera porque μ = 1, seguirá una distribución *t* de Student con n-1 (en este caso 6) grados de libertad.

La región de aceptación será el conjunto de valores de θ más probables cuando H_0 es cierta (menos discrepantes con H_0). Por tanto esta región habrá de comprender los valores de θ menores que un cierto θ_α (ya que los valores grandes son los más discrepantes), esto es:

$$R_A = \{\theta / \theta \le \theta_\alpha\}$$

donde el valor de θ_α se obtiene de:

$$1 - \alpha = P(\theta \le \theta_\alpha / H_0 \text{ cierta}) = P\left(\frac{\bar{X} - 1}{s/\sqrt{n}} \le \theta_a \Big/ H_0 \text{ cierta} \right) = P(t_{n-1} \le \theta_\alpha) \Rightarrow \theta_\alpha = t_{n-1,\alpha}$$

En concreto, con n=7, si α=0.05 se tiene que $t_{6,\,0.05}$ = 1.943. Por tanto la región de aceptación es

$$R_A = [-\infty, 1.943]$$

Por tanto, en este contraste se aceptrrá H_0 si:

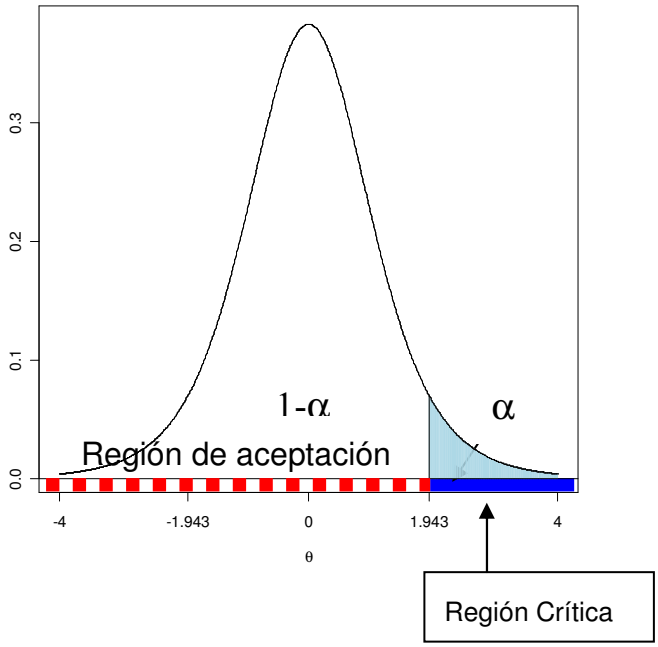
$$\theta(X_1, X_2, ..., X_n) = \frac{\bar{X} - 1}{s/\sqrt{n}}$$

es menor o igual que 1.943, y se rechazará en caso contrario. Si sustituimos los valores obtenidos en la muestra:

$$\frac{\bar{X} - 1}{s/\sqrt{n}} = \frac{1.1 - 1}{0.3/\sqrt{7}} = 0.88$$

Por tanto el valor del estadístico de contraste cae en la región de aceptación, lo que significa que aceptamos H_0 (que μ ≤ 1). Por tanto, aunque la media muestral haya salido ligeramente mayor que 1

ello no constituye evidencia suficiente de que la media poblacional no sea menor que 1. El que la media muestral haya salido algo mayor que 1 puede deberse al simple efecto del azar en el muestreo.



Nótese que el cálculo de la región de aceptación lo hemos llevado a cabo bajo la hipótesis de que $\mu = 1$, aunque la hipótesis nula especifica que $\mu \leq 1$. Es fácil darse cuenta que la región de aceptación así obtenida da lugar a que para cualquier otro valor de μ conforme con H_0, la probabilidad de error tipo I sea menor que α (ya que implicaría un desplazamiento de la curva de la *t* hacia la izquierda, lo que redunda en una reducción de α)

Tamaño de la muestra para la realización de un contraste con nivel de significación y potencia prefijados.

La formulación vista hasta ahora de los contrastes de significación permite calcular a priori el tamaño que ha de tener una muestra si se desea tomar la decisión con unas probabilidades de error I y error II prefijadas.

Ejemplo:

Veamos el método a aplicar a través de nuestro ejemplo inicial. Supongamos que nuestras algas son muy sensibles al pH del agua, y que el valor medio del mismo debe estar en torno a 1, resultando dañino tanto el pH superior a 1.2 como el inferior a 0.8. Si todos los días se desea realizar un control del pH del agua, ¿cuántas muestras hay que tomar para que tanto la probabilidad de realizar intervenciones innecesarias como la probabilidad de no intervenir cuando hace falta sean ambas inferiores al 3%?

El contraste a realizar es:

$$\begin{cases} H_0 : \mu = 1 \\ H_1 : \mu \neq 1 \end{cases}$$

Ya hemos visto que el estadístico de contraste es:

$$\frac{\bar{X}-1}{s/\sqrt{n}}$$

siendo la región de aceptación $R_A = \left[-t_{n-1,\alpha/2}, t_{n-1,\alpha/2}\right]$

Por tanto, para cualquier tamaño de muestra *n*, la regla de decisión consistente en aceptar H_0 si:

$$\frac{\bar{X}-1}{s/\sqrt{n}} \in \left[-t_{n-1,\alpha/2}, t_{n-1,\alpha/2}\right]$$

y rechazarla en caso contrario, nos garantiza que la probabilidad de cometer un error tipo I es α.

Asimismo, la probabilidad de cometer un error tipo II cuando la verdadera media es μ^* es:

$$\beta(\mu^*) = P\left(-t_{n-1,\alpha/2} \le t_{n-1} + \frac{\mu^*-1}{s/\sqrt{n}} \le t_{n-1,\alpha/2}\right)$$

Por tanto, para conseguir una probabilidad de error tipo II inferior al 3%, deberemos determinar el tamaño de la muestra *n* a partir de:

$$\beta(\mu^*) = P\left(-t_{n-1,\alpha/2} \le t_{n-1} + \frac{\mu^*-1}{s/\sqrt{n}} \le t_{n-1,\alpha/2}\right) \le 0.03$$

Obviamente esta cantidad depende de μ^*. Por tanto primero debemos decidir para qué valores de μ^* nos interesa que esta probabilidad sea inferior a 0.03. Dado que del enunciado del problema se sigue que los valores de pH problemáticos son los superiores a 1.2 ó los inferiores a 0.8, serán éstos los valores de μ^* que usaremos para determinar *n*. De hecho da igual utilizar uno ú otro dado que ambos se encuentran a la misma distancia del 1, y por ser simétrica la *t* de Student $\beta(1.2) = \beta(0.8)$. Así, llamando $\delta = \mu^* - 1 = 1.2 - 1 = 0.2$, y $\beta = 0.03$ tenemos:

$$P\left(-t_{n-1,\alpha/2} \le t_{n-1} + \frac{\delta}{s/\sqrt{n}} \le t_{n-1,\alpha/2}\right) = \beta \Leftrightarrow$$

$$\Leftrightarrow P\left(-t_{n-1,\alpha/2} - \frac{\delta}{s/\sqrt{n}} \le t_{n-1} \le t_{n-1,\alpha/2} - \frac{\delta}{s/\sqrt{n}}\right) = \beta \Leftrightarrow$$

$$\Leftrightarrow P\left(t_{n-1} \ge -t_{n-1,\alpha/2} - \frac{\delta}{s/\sqrt{n}}\right) - P\left(t_{n-1} \ge t_{n-1,\alpha/2} - \frac{\delta}{s/\sqrt{n}}\right) = \beta \Leftrightarrow$$

$$\underset{(aprox.)}{\Leftrightarrow} 1 - P\left(t_{n-1} \ge t_{n-1,\alpha/2} - \frac{\delta}{s/\sqrt{n}}\right) = \beta \Leftrightarrow$$

$$\Leftrightarrow P\left(t_{n-1} \ge t_{n-1,\alpha/2} - \frac{\delta}{s/\sqrt{n}}\right) = 1-\beta \Leftrightarrow t_{n-1,\alpha/2} - \frac{\delta}{s/\sqrt{n}} = t_{n-1,1-\beta} = -t_{n-1,\beta}$$

$$\Leftrightarrow \frac{\delta}{s/\sqrt{n}} = t_{n-1,\alpha/2} + t_{n-1,\beta} \Leftrightarrow n = \left(\frac{\left(t_{n-1,\alpha/2} + t_{n-1,\beta}\right)s}{\delta}\right)^2$$

Hemos obtenido así una fórmula general para el cálculo del tamaño de la muestra que garantiza que los errores tipo I y II tienen como mucho probabilidades α y β de ocurrir, siempre que la diferencia entre el valor de la media especificado en la hipótesis nula y el primer valor con el que no desea confundirse sea δ.

Si suponemos que *n* va a salir mayor que 30, podemos aproximar los valores de la *t* de Student por los de la N(0,1). En ese caso:

$$n = \left(\frac{\left(z_{\alpha/2} + z_{\beta} \right) s}{\delta} \right)^2$$

Para conseguir $\alpha = \beta = 0.03$ con $\delta = 0.2$, si de experiencias anteriores hemos obtenido que s = 0.4, necesitaremos una muestra de tamaño:

$$n = \left(\frac{\left(z_{0.015} + z_{0.03} \right) 0.4}{0.2} \right)^2 = \left(\frac{\left(2.17 + 1.88 \right) 0.4}{0.2} \right)^2 = 65.63 \cong 66$$

El valor de δ merece alguna aclaración más. Tal como hemos calculado el tamaño de la muestra, la probabilidad de aceptar que μ es 1 cuando realmente es 1.05 es superior al 3%. Si aceptamos que μ es 1 cuando realmente es 1.05 estamos de hecho cometiendo un error de tipo II; sin embargo este error no nos preocupa (y por tanto nos da igual que su probabilidad sea alta) porque las algas sólo tienen problemas cuando el pH difiere en más de $\delta = 0.2$ unidades de 1. Esta cantidad δ es lo que se denomina la **mínima diferencia relevante. No nos interesa detectar diferencias inferiores a ésta**. Obviamente, cuanto más pequeña sea δ más difícil será detectarla y más información necesitaremos. De hecho, podemos observar que el tamaño de la muestra es inversamente proporcional a δ.

Nótese que si el contraste anterior lo realizamos con una muestra de tamaño 300, y obtenemos una media muestral de 1.05, el contraste nos obligaría a rechazar la hipótesis nula de que μ=1. Se diría entonces que el contraste ha resultado significativo, o que la diferencia detectada de 0.05 unidades es **estadísticamente significativa**. **No debe confundirse la significación estadística con la significación o relevancia práctica**. Si en la práctica la diferencia 0.05 es irrelevante, no tiene importancia ninguna el hecho de que sea estadísticamente significativa. De hecho, con una muestra suficientemente grande cualquier diferencia es estadísticamente significativa, por muy irrelevante que pueda resultar en la práctica.

Contrastes de Hipótesis e intervalos de confianza.

Cuando se concluye la realización de un contraste de hipótesis es interesante presentar un intervalo de confianza para el verdadero valor del parámetro:

- Si se ha aceptado H_0 el intervalo resulta útil para tener una idea del grado de veracidad de H_0. Si se acepta que $\mu = 1$ y el intervalo de confianza para μ es [0.996, 2.031] ello apunta a que esa hipótesis se ha aceptado "por los pelos", y hay indicios de que se puede estar cometiendo un error tipo II aceptando una H_0 falsa. Si esta cuestión es importante conviene tomar una muestra mayor que reduzca este error.

- Si se ha rechazado H_0 el intervalo resulta útil para saber en torno a que valores se encuentra el verdadero valor del parámetro.

- Si el contraste es bilateral conviene obtener un intervalo de confianza bilateral. Si es unilateral habrá de calcularse un intervalo de confianza de una cola en la dirección de H_1 si rechazamos H_0, y en la dirección contraria de H_1 si aceptamos H_0.