Capítulo 6

Inferencia con muestras grandes

[0] Apuntes realizados por Ismael Sánchez. Universidad Carlos III de Madrid.

6.1. Introducción

En este tema seguiremos avanzando en la presentación de herramientas para la inferencia estadística. Es decir, herramientas para poder conocer cómo es una población a partir de la información limitada de una muestra. En el tema anterior presentamos elementos importantes de la inferencia como son:

- El concepto de estadístico, como operación realizada con una muestra. Al depender el resultado de la muestra, un estadístico es una variable aleatoria de la que debemos conocer sus propiedades. En este tema aprenderemos a utilizar la distribución muestral de los estadísticos para poder extrapolar la información de la muestra a la población.

- El concepto de estimación y estimador. Un estimador es un estadístico que se emplea para asignar un valor a un parámetro poblacional desconocido. Como el valor del estimador depende de la muestra, un estimador es una variable aleatoria. La estimación es el valor que toma el estimador en nuestra muestra.

- Presentamos algunos criterios para comparar estimadores alternativos para un parámetro: sesgo, varianza y error cuadrático medio.

- Presentamos varios métodos de estimación, como el método de los momentos o el de máxima verosimilitud

- Presentamos un método para evaluar la idoneidad de un modelo de probabilidad para representar a la población de la que proceden nuestros datos. Mediante este método, denominado test de la chi-cuadrado, podemos medir a través del denominado **p-valor** lo verosímil que resulta un modelo para representar a toda una población, usando sólo la información de una muestra.

El p-valor que se utiliza el test de bondad de ajuste es una herramienta de inferencia muy útil, pues nos permite introducir una medida de la incertidumbre a las conclusiones que extraemos de la muestra. No es lo mismo concluir que la distribución de nuestra población puede ser una lognormal, con un p-valor de 0.75 que con un p-valor de 0.10. En este segundo caso, la seguridad que nos proporciona la muestra es mucho menor, y debemos tenerlo en cuenta a la hora de tomar decisiones prácticas basadas en este resultado.

En este tema ampliaremos esta idea de **acompañar conclusiones sobre la población basadas en la muestra con medidas de su precisión**. En general, sólo diremos que estamos haciendo inferencia sobre una población (a partir de una muestra) si acompañamos nuestras conclusiones con medidas de incertidumbre de este tipo.

En este tema introduciremos herramientas de inferencia que son independientes de la distribución de la población. Por tanto, pueden emplearse para cualquier tipo de distribución de nuestra variable aleatoria de interés X. El precio que vamos a pagar por tener herramientas tan generales es que tenemos que utilizar muestras grandes ($n > 30$). Cuanto menor sea la muestra, menos fiables serán los resultados que expongamos en este tema.

6.2. Intervalos de confianza para μ con muestras grandes

6.2.1. Introducción

Cuando queremos **estimar** el valor de un parámetro poblacional θ a partir de la información de una muestra $X_1, X_2, ..., X_n$ utilizamos un **estimador** $\hat{\theta}$. Dicho estimador aplicado a una muestra nos proporcionará un valor numérico, que se denomina **estimación** de θ. La precisión de ese estimador está relacionada con la probabilidad de que $\hat{\theta}$ nos proporcione un valor próximo a θ. Esa precisión viene determinada por las propiedades estadísticas del estimador; es decir, por su distribución en el muestreo.

En esta sección extenderemos las propiedades de la media muestral $\bar{X}$ como estimador de la media poblacional μ. El objetivo es que en el ejercicio de estimación no sólo proporcionemos el valor estimado obtenido con la muestra, sino una medida de la incertidumbre de dicho valor como estimación de μ. La incertidumbre procede de haber utilizado sólo una muestra de tamaño finito. Por tanto, con otras muestras hubiéramos obtenido estimaciones diferentes que serían igual de válidas que las que hemos proporcionado con nuestra muestra. Si tuviésemos toda la población, no tendríamos incertidumbre sobre la media poblacional.

Lo que haremos es dar una medida de esa incertidumbre utilizando las propiedades de la media muestral que ya vimos en el tema anterior. Vamos a recordar esas propiedades. Supongamos que tenemos una población representada por la variable aleatoria X tal que $E(X) = \mu$ y $\text{var}(X) = \sigma^2$, y utilizamos la media muestral como estimador de μ, es decir,

$$\hat{\mu} = \bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Entonces, como demostramos en el tema anterior, se cumple que

$$E(\hat{\mu}) = \mu; \text{Var}(\hat{\mu}) = \frac{\sigma^2}{n},$$

por lo que la media muestral es un estimador insesgado cuya precisión aumenta con el tamaño muestral. Además, por el teorema central del límite, **si n es grande** tenemos que, **independientemente de cómo sea la distribución de** X,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right). \tag{6.1}$$

Por ejemplo, si X es una Poisson de parámetro λ se tiene que como $E(X) = \lambda$, $\text{Var}(X) = \lambda$. Entonces, si n es grande

$$\hat{\lambda} \sim N\left(\lambda, \frac{\lambda}{n}\right),$$

donde

$$\hat{\lambda} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

En una variable normal, no hace falta usar el teorema central del límite para justificar (6.1), pues la normalidad de la media muestral es exacta al ser la media muestral una combinación lineal de variables normales. Por tanto, en poblaciones normales, para cualquier tamaño muestral

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right). \tag{6.2}$$

6.2.2. Intervalos de confianza

En esta sección vamos a proponer un procedimiento para añadir la información de la incertidumbre que tenemos sobre la estimación realizada con la muestra. Dicha incertidumbre se describirá mediante la utilización de un intervalo de valores dentro de los cuales estará el valor poblacional μ con cierta probabilidad. Un ejemplo de este tipo de resultados sería decir que la estimación de la media es $\hat{\mu} = 100$ y que con una probabilidad del 95 % el valor verdadero μ estará en el intervalo $\mu \in (95, 105)$. Cuanto más estrecho sea dicho intervalo, menos incertidumbre existirá sobre el verdadero valor del parámetro. El intervalo sobre el valor del parámetro, que se construye utilizando las propiedades del estimador, se denomina **intervalo de confianza**. A la probabilidad de que con la información de la muestra, el parámetro esté dentro del intervalo se le denomina **nivel de confianza**. Lo habitual es realizar intervalos de confianza con nivel de confianza del 95 %.

Al valor numérico que se obtiene en la estimación se le denomina también **estimación puntual**, mientras que al uso de un intervalo de confianza se le denomina **estimación por intervalo**. Veamos a continuación cómo se construyen intervalos de confianza para μ, basado en el uso de **muestras grandes**.

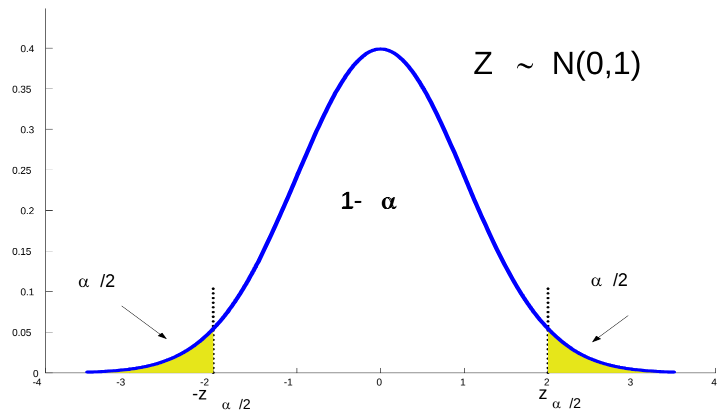
De (6.1) tenemos que, si n es grande, podemos estandarizar obteniendo una normal estándar.

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1). \tag{6.3}$$

Llamemos $z_{\alpha/2}$ al valor de la $N(0, 1)$ que deja un área a la derecha de valor $\alpha/2$. Entonces, por la simetría de la normal, a la izquierda de $-z_{\alpha/2}$ quedará un área igual a $\alpha/2$. Por tanto

$$P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = (1 - \alpha). \tag{6.4}$$

La siguiente figura ilustra la localización de estos valores $z_{\alpha/2}$ y $-z_{\alpha/2}$.



De (6.3) y (6.4) obtenemos por tanto

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

Operando tenemos que

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = (1 - \alpha). \tag{6.5}$$

La expresión (6.5) quiere decir que tenemos una probabilidad $1-\alpha$ de seleccionar una muestra tal que μ esté entre los valores $\bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$. Por tanto dada una muestra, podemos construir el siguiente **intervalo de confianza de nivel** $100 \times (1-\alpha)\,\%$ **para** μ

$$IC(1-\alpha) : \mu \in \left\{ \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\}. \tag{6.6}$$

El intervalo de confianza (6.6) se interpreta de la siguiente manera:

> **Si tuviésemos un número infinito de muestras de la población, y construyésemos con cada una un intervalo como en (6.6), entonces el 100(1-α) % de dichos intervalos contendría al verdadero valor del parámetro μ.**

En la práctica, sólo tenemos una muestra, y por eso sólo podemos construir un intervalo. No tiene entonces sentido interpretar el intervalo como la región en la que estará μ con probabilidad (1-α), puesto que en el intervalo calculado, la media μ estará o no estará. Por eso, para expresar nuestra incertidumbre sobre si el intervalo calculado con nuestra muestra contiene o no al parámetro μ emplearemos la palabra **nivel de confianza**.

Nótese que si la muestra es suficientemente grande (por ejemplo, más de 30 datos), el intervalo (6.6) es válido para cualquier variable aleatoria X, sea o no normal. **Por intervalo válido queremos decir que su nivel confianza es realmente** $100 \times (1-\alpha)\,\%$. Si la muestra es pequeña, entonces $\bar{X}$ tendrá una distribución en el muestreo que no será normal, y por tanto no tendremos ninguna garantía en que (6.6) tenga el nivel de confianza deseado. Es importante darse cuenta de que **si X es normal, $\bar{X}$ es siempre normal**, para cualquier tamaño muestral. Por tanto, si hay normalidad (6.6) es siempre un intervalo exacto.

Ejemplo 1 *Una muestra aleatoria extraída de una población con $\sigma^2 = 100$ de $n = 144$ observaciones tiene una media muestral $\bar{X} = 160$. se pide:*
(a) Calcular un intervalo de confianza del 95 % para la media poblacional μ.
(b) Calcular un intervalo de confianza del 90 % para la media poblacional μ.
SOLUCIÓN:
Como se conoce la varianza de la población, $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ será una variable $N(0,1)$, luego

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

Por tanto el intervalo es:

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{6.7}$$

(a) Un nivel de confianza del 95 % equivale a usar $\alpha = 0{,}05$ y por tanto, $z_{\alpha/2} = 1{,}96$. El intervalo es entonces

$$IC(95\,\%) : \mu \in [158{,}36, 161{,}63]$$

(b) Un nivel de confianza del 90 %, equivale a usar $\alpha = 0{,}10$ y por tanto $z_{\alpha/2} = 1{,}65$. El nuevo intervalo es

$$IC(90\,\%) : \mu \in [158{,}625, 161{,}375]$$

que al ser de menor confianza, es más estrecho que el anterior.

Si σ^2 es desconocido se sustituye por una estimación $\hat{\sigma}^2$. Si el tamaño muestral es suficientemente grande tendremos que la estimación proporcionará un valor muy cercano al verdadero, y por tanto seguiremos utilizando el intervalo (6.6) reemplazando σ por su estimación $\hat{\sigma}$, es decir, el intervalo de confianza para tamaño muestral grande será

$$IC(1-\alpha) : \mu \in \left\{ \bar{x} \pm z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \right\} \tag{6.8}$$

Ejemplo 2 *En una encuesta se pregunta a 10.000 estudiantes de Bachillerato sobre su consumo de refrescos semanal, encontrándose una media de 5 botes, con una desviación típica estimada de 2. Hallar un intervalo de confianza para el consumo medio de toda la población de estudiantes de Bachillerato, al 95 %.*

SOLUCIÓN:

Como el tamaño muestral es muy grande podemos emplear el intervalo

$$\left(\bar{x} - z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \right),$$

donde $\alpha = 0{,}05 \quad z_{\alpha/2} = 1{,}96$, *luego tenemos que*

$$\left(5 - 1{,}96 \frac{2}{100}, 5 + 1{,}96 \frac{2}{100} \right) = (4{,}96, 5{,}04)$$

6.2.3. La cuasivarianza

Una opción para estimar σ^2 es, de acuerdo con el método de los momentos, usar la varianza muestral, es decir

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

Se puede demostrar, sin embargo, que este estimador es sesgado. Se puede demostrar que

$$E(S^2) = \sigma^2 \frac{(n-1)}{n}, \tag{6.9a}$$

y por tanto el sesgo de S^2 como estimador de σ^2 es

$$\begin{aligned} \text{sesgo}(S^2) &= E(S^2) - \sigma^2 \\ &= \sigma^2 \frac{(n-1)}{n} - \sigma^2 = -\frac{\sigma^2}{n}. \end{aligned}$$

Vemos que el sesgo de S^2 es negativo; es decir, que S^2 subestima la verdadera varianza. Usaremos entonces el siguiente estimador alternativo

$$\hat{\sigma}^2 = \hat{S}^2 = \frac{\sum_{i=1}^n \left(X_i - \bar{X} \right)^2}{n-1},$$

que a partir de (6.9a) es fácil comprobar que es insesgado. Es decir

$$E(\hat{S}^2) = \sigma^2.$$

Este estimador $\hat{S}^2$ recibe el nombre de **cuasivarianza, pseudovarianza o varianza corregida**. Si n es grande tendremos que $\hat{S}^2 \approx \sigma^2$, y por tanto podemos hacer la estandarización (6.3) usando $\hat{S}^2$ en lugar de σ^2. Al estadístico que resulta de la estandarización de $\bar{X}$ usando $\hat{S}$ le llamaremos T en lugar del estadístico Z usado en (6.3). Por tanto, **para muestras grandes** tendremos que

$$T = \frac{\bar{X} - \mu}{\hat{S}/\sqrt{n}} \sim N(0,1),$$

pero si n no es elevado, la distribución del estadístico T ya no será normal. Será, en general, desconocida y dependerá de la distribución de los datos X.

Por tanto, si σ^2 es desconocida y la estimamos con la cuasivarianza, un intervalo de confianza de nivel $(1-\alpha)$ valido para cualquier tipo de variable aleatoria, con tal que n sea elevado ($n > 30$), será:

$$IC(1-\alpha) : \mu \in \left\{ \bar{x} \pm z_{\alpha/2} \frac{\hat{s}}{\sqrt{n}} \right\} \tag{6.10}$$

6.3. Determinación del tamaño de la muestra

Del intervalo (6.6) puede verse que **si aumentamos n reducimos la amplitud el intervalo y por tanto nuestra incertidumbre sobre μ para un mismo nivel de confianza**. El intervalo de confianza para muestras grandes puede escribirse como

$$\mu \in \left\{ \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\} = \bar{x} \pm L.$$

La longitud del intervalo será entonces $2L$. Si queremos un valor de L determinado ¿cuál debe ser el tamaño muestral n?

$$L = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \Rightarrow n = \left(\frac{z_{\alpha/2}\sigma}{L} \right)^2 \tag{6.11}$$

Si σ es desconocido lo estimamos en una muestra piloto, obteniendo $\hat{\sigma}_0$. Entonces

$$n = \left(\frac{z_{\alpha/2}\hat{\sigma}_0}{L} \right)^2$$

Ejemplo 3 *Sea X el consumo unitario de cierto material en un proceso productivo (miligramos por unidad de producto obtenido). Se sabe que X tiene desviación típica $\sigma = 20$ mg. Se toma una muestra aleatoria de 200 observaciones obteniéndose una media muestral del consumo de $\bar{X} = 120$ mg.*

(a) A partir de esta información muestral, estimar mediante un intervalo con un 95 % de confianza el consumo medio de este producto.

(b) ¿Qué tamaño muestral sería necesario tomar para que un intervalo del 95 % de confianza tuviese una longitud de 2mg? (longitud del intervalo=diferencia entre sus extremos).

SOLUCIÓN:

(a) Nos piden el intervalo de confianza para la media de una distribución normal con $\sigma = 20mg$. conocida. Sabemos que, al ser un tamaño muestral grande, la media muestral verifica que

$$\bar{X} \sim N\left(\mu, \sigma^2/n\right)$$

y para un nivel de confianza del 95 % coeficiente de confianza;

$$1-\alpha = 0{,}95 \Longrightarrow \alpha = 0{,}05 \Longrightarrow \alpha/2 = 0{,}025$$

El intervalo para el consumo medio del material es;

$$\mu \in \left[\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = \left[120 \pm 1{,}96 \times \frac{20}{\sqrt{200}}\right] = [120 \pm 2{,}77] = [117{,}23{,}122{,}77]$$

(b) Se quiere garantizar una amplitud del intervalo como máximo de 2 mg. con el mismo nivel de confianza. Es decir, que el intervalo del 95 % sea $\bar{X} \pm 1$. De (6.11) tenemos que $L = 1$, y entonces,

$$n = \left(\frac{z_{\alpha/2}\sigma}{L}\right)^2 = \left(\frac{1{,}96 \times 20}{1}\right)^2 \approx 1537 \text{ observaciones}$$

6.4. Introducción al contraste de hipótesis

6.4.1. Introducción

En la sección anterior aprendimos a estimar los parámetros de una población a partir de una muestra no sólo asignando una estimación puntual sino asignando un intervalo de confianza. En muchos problemas de ingeniería, sin embargo, no se pretende estimar el valor de un parámetro sino comprobar que dicho parámetro cumple alguna restricción o suposición. Por ejemplo, podemos estar interesados en comprobar mediante una muestra que el valor del parámetro no ha cambiado respecto del valor que teníamos estimado en el pasado. En este caso, no queremos utilizar la muestra para asignar un nuevo valor al parámetro sino sólo comprobar si la información de la muestra es consistente con nuestra restricción o suposición, o por el contrario nuestra suposición debe rechazarse a la luz de la nueva evidencia mostrada por la muestra. Otras veces, sólo queremos saber si un procedimiento es más rápido que otro. En ese caso se comparan ambos procedimientos con algún protocolo de pruebas. El objetivo de estas pruebas no es saber cuánto se tarda en completarlas, sino en saber cuál de los dos procedimientos es más rápido; o más concretamente, cuál tiene un tiempo medio de ejecución menor.

A estas restricciónes o suposiciones sobre los valores de los parámetros de una población les llamaremos **hipótesis** y a la prueba estadística consistente en comprobar si la muestra apoya o no dichas hipótesis le llamaremos **contraste de hipótesis**. Nuestras hipótesis serán siempre **restricciones sobre los parámetros de una población**. El planteamiento de una hipótesis implica una división de los posibles valores del parámetro. Por un lado tendremos los valores que cumplen la hipótesis y por otro a los valores que no la cumplen. Veamos un ejemplo.

Ejemplo 4 *Un fabricante de transistores del tipo BC547B (datos:* ***transistor.sf3****) sabe por su información histórica que cuando su producción se mantiene en los niveles de calidad deseables, el valor de la llamada* ganancia en corriente *de los transistores (conocida por β, adimensional) sigue una distribución de media 290 y varianza 760. Dicho fabricante desea saber si su producción sigue manteniendo el mismo nivel de calidad; es decir, si se sigue* ***manteniendo la media*** *y que* ***no haya aumentado la varianza****. Con ese fin, el fabricante toma una muestra de transistores. Esa muestra no la quiere para obtener una nueva estimación de la media o de la varianza, sino para contrastar las hipótesis de que la media no ha cambiado y la varianza no ha aumentado. Más concretamente, desea saber si:*

$$\mu = 290,$$

o, por el contrario,

$$\mu \neq 290;$$

y por otra parte desea saber si

$$\sigma^2 \leq 760$$

o, por el contrario,

$$\sigma^2 > 760.$$

6.4.2. Contraste de una hipótesis estadística

En esta sección se dará una panorámica general de cómo se hace un contraste de hipótesis como los planteados en el ejemplo de arrriba. A la hipótesis que se quiere comprobar se le denomina **hipótesis nula**, y la denotaremos por H_0. Por tanto, utilizando el ejemplo de los transistores BC547B tenemos

$$\mathrm{H}_0 : \mu = 290,$$

Al rango de valores que toma el parámetro cuando la hipótesis nula sea falsa le denominaremos **hipótesis alternativa**, y la denotaremos por H_1,es decir

$$\mathrm{H}_1 : \mu \neq 290.$$

Análogamente, si queremos contrastar que la varianza no ha aumentado, plantearemos el siguiente contraste de hipótesis:

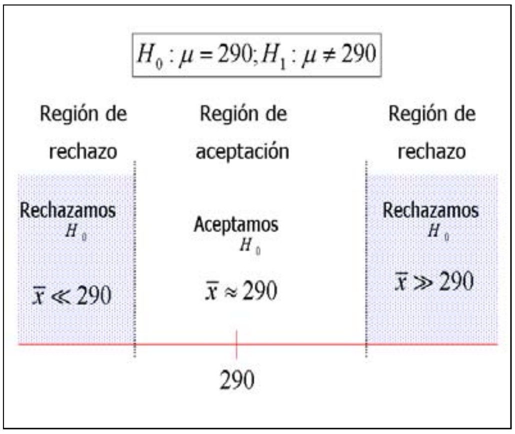
$$\begin{aligned} \mathrm{H}_0 &: \sigma^2 \leq 760, \\ \mathrm{H}_1 &: \sigma^2 > 760. \end{aligned}$$

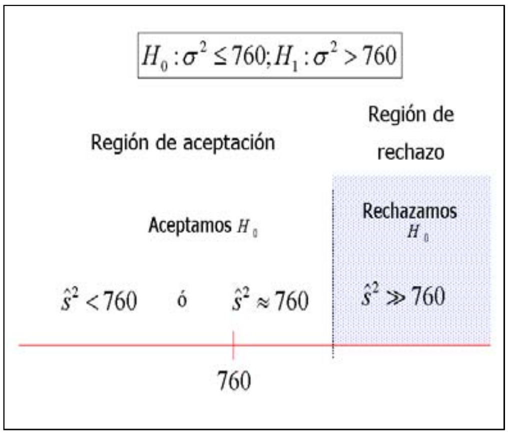
Como vemos, **entre H_0 y H_1 está todo el rango de valores de los parámetros implicados**. El método estadístico que vamos a emplear (la metodología de Neyman-Pearson) introduce algunas restricciones en la forma en que podemos plantear las hipótesis del contraste.

- La hipótesis nula debe tener siempre el signo $=$. Puede ser $\geq$, $\leq$ ó $=$. La H_1 será entonces del tipo $\neq$, $<$ ó $>$. Cuando H_1 es del tipo $\theta \neq \theta_0$ diremos que **la alternativa es bilateral,** pues los valores que implica estarán a ambos lados de H_0; mientras que cuando es del tipo $\theta > \theta_0$ ó $\theta < \theta_0$ diremos que **la alternativa es unilateral**, pues ésta se encuentra sólo a un lado de H_0.
- El método favorece a la hipótesis nula, en el sentido de que **aceptaremos H_0 salvo que la muestra revele mucha evidencia en contra**. La situación es parecida a la de un jurado que debe decidir si condena o no al acusado. El jurado querrá ser justo, y tomar la decisión correcta. Sin embargo, como la decisión se basa en un conjunto limitado de pruebas, se enfrenta a dos posibles tipos de error. Por una parte puede condenar al acusado siendo éste inocente. Por otra parte puede dejar libre al acusado siendo éste culpable. Es evidente que es más grave el error de encarcelar a un inocente que el error de liberar a un culpable. Por esa razón se dice que el acusado es inocente salvo que se demuestre feacientemente lo contrario. En nuestro caso, la hipótesis nula hace el papel de la declaración de inocencia del jurado: se considerará que H_0 es cierta salvo que se pruebe feacientemente que es falsa. Por tanto, **el método tenderá a aceptar H_0 salvo que los datos hagan lo que dice H_1 de forma muy evidente**. Es muy importante tener este aspecto en cuenta a la hora de obtener conclusiones prácticas del resultado de un contraste de hipótesis.

Por ejemplo, si queremos a partir de unos datos contrastar $\mathrm{H}_0 : \mu = 290$ frente a la alternativa $\mathrm{H}_1 : \mu \neq 290$, obtendremos el valor de la media muestral, que denotaremos por $\bar{x}$, y sólo rechazaremos H_0 si los datos muestran una fuerte evidencia en contra, es decir, si $\bar{x} >> 290$ ó $\bar{x} << 290$. El método estadístico que vamos a emplear nos ayudará a delimitar cuándo consideramos que $\bar{x} >> 290$ ó $\bar{x} << 290$; es decir, nos ayudará a delimitar la frontera a partir de la cual el alejamiento respecto a H_0 es ya demasiado grande. Análogamente, si queremos contrastar a partir de una muestra si $\mathrm{H}_0 : \sigma^2 \leq 760$ frente a la alternativa $\mathrm{H}_1 : \sigma^2 > 760$, obtendremos el valor de $\hat{S}^2$ en la muestra, que denotaremos por $\hat{s}^2$, y sólo rechazaremos H_0 si obtenemos $\hat{s}^2 >> 760$.

El método de Neyman-Pearson para la realización de contrastes consistirá, entonces, en establecer unas **regiones** para los valores del estimador $\hat{\theta}$ del parámetro θ implicado, en las cuales podamos decir si aceptamos o rechazamos la hipótesis nula. La región próxima a H_0 se le llama **región de aceptación o de no rechazo**, mientras que a las regiones alejadas de H_0 se les denomina **región de rechazo**. Estas regiones se ilustran en las siguientes figuras.





Si los datos son tales que nos colocamos en la región de aceptación (también llamada de no rechazo) diremos que la diferencia encontrada entre la estimación y lo que dice H_0 **no es significativa**, y puede perfectamente explicarse por el azar de la muestra. Por el contrario, si los datos son tales que nos colocamos en la región de rechazo, diremos entonces que la diferencia entre lo que dice la muestra y nuestra hipótesis nula **es significativa**, siendo muy improbable que se deba al azar de la muestra.

El método estadístico para la realización del contraste de hipótesis consiste, por tanto, en **valorar el alejamiento entre los datos y nuestra hipótesis nula**. Si lo que dicen los datos está cerca de lo que dice H_0 (o mejor dicho, no se aleja mucho) nuestra decisión será aceptar H_0. Por el contrario, si los datos se alejan mucho de lo que dice H_0 rechazaremos H_0 a favor de H_1.

La simple comparación de la estimación con el parámetro ($\bar{x}$ con $\mu_0 = 290$ ó $\hat{s}^2$ con $\sigma_0^2 = 760$) no es una buena indicación de si debemos aceptar o no H_0. Por ejemplo, si en el caso de los transistores del Ejemplo 4 obtenemos que la media muestral es $\bar{x} = 295$: ¿es 295 una cantidad cercana a 290 o no? La respuesta depende de varios factores, entre otros del tamaño muestral con que se ha obtenido el valor de $\bar{X}$. No es lo mismo un alejamiento de 295-290=5 unidades a partir de 20 datos que a partir de 2000 datos. En este segundo caso, la distancia de 5 unidades debe interpretarse como más relevante que con sólo 20 datos, donde la variabilidad muestral es grande. Usaremos entonces una medida de discrepancia entre el valor del parámetro estimado con la muestra, $\hat{\theta}$, y el valor que implica H_0, θ_0, cuya magnitud sea más fácil de interpretar que la simple comparación de $\hat{\theta}$ con θ_0. Esta medida de discrepancia resumirá toda la información contenida en los datos que sea relevante para hacer el contraste. A esta medida de discrepacia le llamaremos **estadístico de contraste** y será diferente según sea el parámetro que estemos considerando y de las características de la población

En resumen, hemos visto hasta ahora que un contraste de hipótesis tiene los siguientes pasos:

1. Determinar H_0 y H_1 teniendo en cuenta que H_0 debe tener el signo $=$ y que el método favorecerá dicha hipótesis
2. Buscar el **estadístico de contraste,** que será la medida de discrepancia entre el valor del parámetro estimado con la muestra y el valor que implica H_0. Más adelante veremos qué **estadístico de contraste** usar en cada caso.
3. A partir de las propiedades del estadístico de contraste, delimitamos las **regiones de aceptación y rechazo.**
4. Calcular a partir de la muestra el valor que toma el **estadístico de contraste** y comprobar si cae en la región de aceptación o en la de rechazo.

Es importante ver que nunca sabremos con certeza si las hipótesis son ciertas o falsas. La decisión correcta sobre qué hipótesis es la verdadera sólo puede hacerse sin error si conociésemos el verdadero valor del parámetro. Lo que hacemos en un contrate de hipótesis es simplemente una apuesta, a la luz de la información limitada de la muestra, sobre si H_0 es o no verosímil. Por tanto, al igual que en un juicio a un acusado, podemos tomar la decisión equivocada. Por ejemplo, si queremos contrastar $H_0 : \theta = \theta_0$ pudiera ser que realmente se cumpla $\theta = \theta_0$ pero que nuestra muestra resulte $\hat{\theta} >> \theta_0$. Por ejemplo, si θ es una longitud media de una población, podríamos haber seleccionado a elementos que casualmente tengan una longitud superior a la media, con lo que $\hat{\theta}$ sería $>> \theta_0$. En este caso rechazaremos erróneamente H_0.Este tipo de situaciones serán muy improbables; pero pueden darse, lo cual ha de tenerse en cuenta a la hora de interpretar el resultado y tomar decisiones en la vida real. A este tipo de error: rechazar H_0 siendo cierta le llamaremos **error Tipo I**. En el caso del simil del juicio a un acusado, en el que la hipótesis nula es la inocencia, el **error Tipo I** sería condenar a un inocente.

Por otra parte, puede ser que la población no cumpla la restricción que estamos contrastanto, es decir, que H_0 sea falsa, pero que casualmente la muestra seleccionada no lo refleje así. Por ejemplo, si queremos contastar $H_0 : \theta \geq \theta_0$ pudiera ser que esta hipótesis sea falsa y el valor poblacional fuese $\theta < \theta_0$ pero tengamos la mala suerte de que en la muestra recogida tengamos $\hat{\theta} > \theta_0$. Entonces concluiremos erróneamente que H_0 es cierta. A este tipo de error, aceptar H_0 cuando es falsa, le llamaremos **error Tipo II**. En el cado del simil del juicio, el **error Tipo II** sería declarar inocente a un acusado que es culpable. El siguiente cuadro resume estas posibles equivocaciones.

	H_0 es cierta	H_0 es falsa
Se rechaza H_0	Error Tipo I	**Decisión correcta**
Se acepta H_0	**Decisión correcta**	Error Tipo II

6.5. Contraste de hipótesis de la media con muestras grandes

Se desea contrastar alguna de las siguientes hipótesis:

1. $H_0 : \mu = \mu_0$; frente a $H_1 : \mu \neq \mu_0$,

2. $\mathrm{H}_0 : \mu \geq \mu_0$; frente a $\mathrm{H}_1 : \mu < \mu_0$,

3. $\mathrm{H}_0 : \mu \leq \mu_0$; frente a $\mathrm{H}_1 : \mu > \mu_0$.

El **estadístico de contraste** es la operación que realizaremos con los datos y que contendrá toda la información relevante para el contraste. El **estadístico de contraste** se basa siempre en las propiedades del estimador del parámetro que estemos contrastando. En el caso de contrastes sobre μ, el contraste se basará en las siguientes propiedades de la media muestral $\bar{X}$ obtenida con una muestra aleatoria simple de tamaño n. **Para cualquier variable X de media μ y varianza σ^2, si el tamaño muestral es suficientemente grande** se cumple que

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \tag{6.12}$$

y si σ^2 es desconocido, y utilizamos la cuasivarianza $\hat{S}^2$ como estimador de σ^2 tenemos, si el tamaño muestral es suficientemente grande,

$$T = \frac{\bar{X} - \mu}{\hat{S}/\sqrt{n}} \sim N(0,1). \tag{6.13}$$

Como **estadístico de contraste** utilizaremos (6.12) y (6.13), pero imponiendo el valor μ_0, es decir,

$$d \equiv Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}},$$

para el caso de σ conocida, y

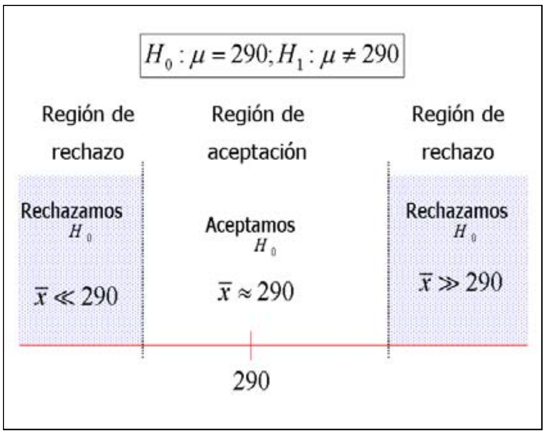
$$d \equiv T_0 = \frac{\bar{X} - \mu_0}{\hat{S}/\sqrt{n}}, \tag{6.14}$$

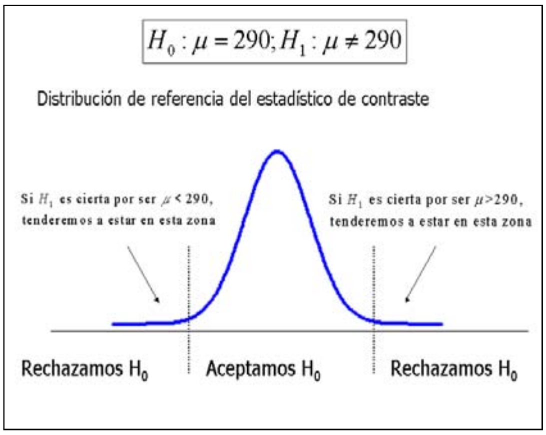
cuando estimemos σ con la cuasidesviación típica. Estos estadísticos d miden la distancia entre el valor de la hipótesis nula μ_0 y el obtenido en la muestra $\bar{x}$.

Para interpretar el valor de d necesitamos una distribución de referencia que nos ayude a determinar cuando d toma un valor alto o bajo, y así poder valorar la distancia entre $\bar{x}$ y μ_0. La elección de la distribución de referencia, que nos ayudará a evaluar d, es similar para cualquier parámetro θ y se elige de la siguiente manera:

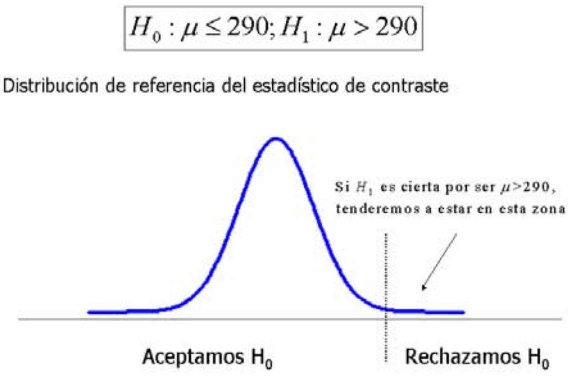
La distribución de referencia en el contraste de un parámetro θ es la distribución que sigue el estadístico de contraste d cuando $\theta = \theta_0$.

Por tanto, en nuestro caso, de acuerdo con (6.12) y (6.13) la distribución de referencia es la $N(0,1)$. Cuanto más lejos esté $\bar{x}$ de lo que dice H_0 más lejos estará d de la $N(0,1)$. Por tanto, y siguiendo el ejemplo anterior sobre los transistores, si queremos hacer el contraste bilateral de $\mathrm{H}_0 : \mu = 290$ frente a la alternativa $\mathrm{H}_1 : \mu \neq 290$, calcularemos d usando los valores muestrales $\bar{x}$ y $\hat{s}$ y tomando que $\mu_0 = 290$. Vemos que d depende no sólo de $\bar{X}$ sino de n y de σ^2 (a través de su estimador $\hat{S}^2$). Si obtenemos que d está en la zona de la cola derecha de la $N(0,1)$ (ver figura siguiente), tendremos mucha seguridad para concluir que $\bar{x} >> 290$, y que por tanto es muy *plausible* que realmente $\mu > 290$ y deberemos rechazar H_0. Análogamente, si d está en la zona de la cola de la izquierda, concluiremos que $\bar{x} << 290$ y por tanto es muy *plausible* que también $\mu < 290$ y deberemos rechazar H_0, como se ilustra en la siguiente figura.

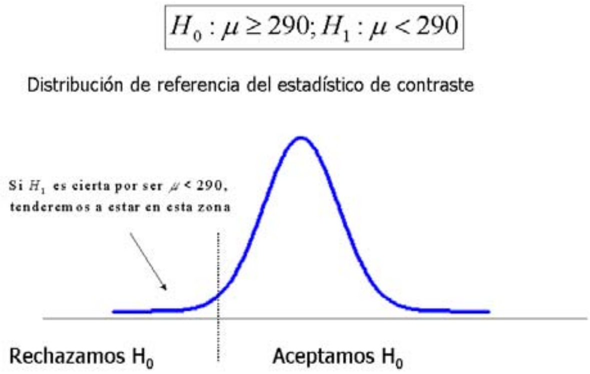




Es importante ver que si el contraste fuese unilateral sólo nos interesaría una de las dos colas de la distribución. Por ejemplo, si el contraste a resolver fuese $H_0 : \mu \leq 290$ frente a la alternativa $H_1 : \mu > 290$ rechazaríamos H_0 sólo si $\bar{x} >> 290$, lo que equivaldría a obtener el estadístico de contraste con un valor en la cola de la derecha, como se muestra en la siguiente figura

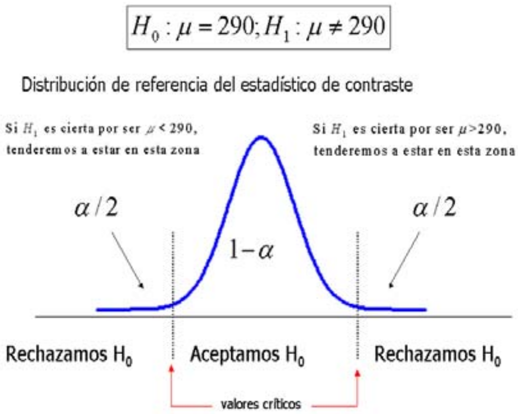


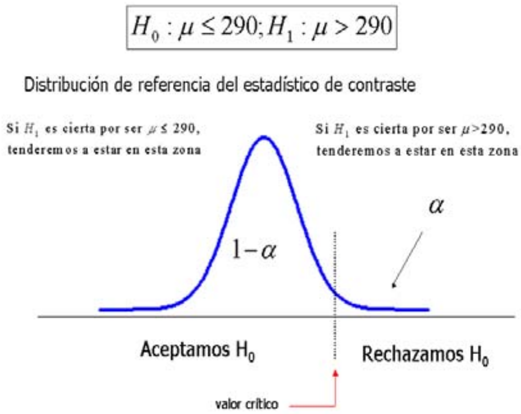
Usando el mismo argumento, si el contraste fuese $H_0 : \mu \geq 290$ frente a la alternativa $H_1 : \mu < 290$ rechazaríamos H_0 sólo cuando d estuviese en la zona de la izquierda de la distribución de referencia, como se muestra a continuación.



Puede verse que **la región de rechazo esta siempre en la zona de la distribución de referencia que señale $\mathbf{H_1}$**. Esta localización de la región de rechazo en el lugar que señale H_1 es también aplicable a todos los contrates de hipótesis que veremos en los siguientes temas.

La pregunta ahora es ¿dónde ponemos la frontera entre la región de aceptación y de rechazo? La frontera la delimitaremos reservando un área de la distribución de referencia para la región de rechazo. El tamaño de la región de rechazo lo decide el analista. Ese área suele ser una cantidad pequeña: 1 %, 5 % ó 10 % colocada en la cola de la distribución de referencia. Este área se denomina **nivel de significación** y se denota con la letra $\boldsymbol{\alpha}$. En el caso de contrastes bilaterales, el área α que reservamos para la región de rechazo se divide en dos partes iguales y se deja la mitad de dicho área en cada cola de la distribución. de referencia. Para el caso de nuestro ejemplo de los transistores, las regiones de rechazo usando un nivel de significación α serán las que se muestran en las siguientes figuras.





A los valores del eje de abcisas que delimitan las regiones de aceptación y rechazo se les denomina **valores críticos**. La siguiente tabla resume las características de este contraste.

Contrastes	Estadísticos de contraste	Distribución de referencia	Región de rechazo
(1)-$\mathrm{H}_0: \mu = \mu_0$; $\mathrm{H}_1: \mu \neq \mu_0$ (2)-$\mathrm{H}_0: \mu \geq \mu_0$; $\mathrm{H}_1: \mu < \mu_0$ (3)-$\mathrm{H}_0: \mu \leq \mu_0$; $\mathrm{H}_1: \mu > \mu_0$	$Z_0 = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$ $T_0 = \dfrac{\bar{X} - \mu_0}{\hat{s}/\sqrt{n}}$	$N(0,1)$	(1) $\lvert z_0\rvert, \lvert t_0\rvert > z_{\alpha/2}$ (2) $z_0, t_0 < -z_\alpha$ (3) $z_0, t_0 > z_\alpha$

Ejemplo 5 *Utilizando la información del Ejemplo 4 sobre los transistores BC547B (fichero* ***transistor.sf3****) mencionados anteriormente deseamos hacer el contraste de si se mantiene el valor nominal* $\mu = 290$ *como media de la distribucion de valores* β*, es decir,*

$$\begin{aligned} H_0 &: \mu = 290 \\ H_1 &: \mu \neq 290 \end{aligned}$$

Los 100 datos muestran que

$$\begin{aligned} \bar{x} &= 282{,}3; \hat{s} = 27{,}69; \\ t_0 &= \frac{\bar{x} - \mu_0}{\hat{s}/\sqrt{n}} = \frac{282{,}3 - 290}{27{,}69/10} = -2{,}78. \end{aligned}$$

Como es un contaste bilateral necesitamos dos valores críticos. Como la distribución de referencia es la N(0,1), que es simétrica de media cero, ambos valores críticos serán iguales pero de signo contrario. Uusando un nivel de significación $\alpha = 0{,}05$*, se tiene que los valores críticos son* $z_{0{,}025} = 1{,}96$ *y* $-z_{0{,}025} = -1{,}96$*. Por tanto, como* $|t_0| = 2{,}78 > 1{,}96$ *y rechazamos* H_0*. Por tanto, la diferencia entre la media muestral* $\bar{x} = 282{,}3$ *y la hipótesis nula* $\mu_0 = 290$ *es suficientemente grande como para considerarla sólo debida al azar del muestreo. Se dice entonces que, con un nivel de significación de* $\alpha = 0{,}05$*, la diferencia detectada es significativa. Hay dos opciones para interpretar este resultado, o bien el proceso ha cambiado la media o el aparato de medida con que hemos obtenido estas observaciones está descalibrado.*

Ejemplo 6 *El fichero* ***estaturas.sf3*** *contiene las estaturas de un grupo de 50 mujeres y 50 hombres entre 18 y 25 años. El sistema de recogida de datos consistió en preguntar a los encuestados por su estatura, seleccionando a los encuestados al azar entre estudiantes tanto del campus de la Carlos III como de otros campus universitarios. Es, por tanto, estatura declarada, no estatura medida.*

Según los estudios antropométricos, los jóvenes de este rango de edad tienen una estatura media de 164 cm ellas y 177 cm ellos. La pregunta que queremos contestar mediante un contraste de hipótesis es ¿son más altos, por término medio, los jóvenes universitarios madrileños que la media de los jóvenes españoles?, es decir ¿$\mu_{Hombres} > 177$*? ¿*$\mu_{Mujeres} > 164$*?. Como* H_0 *debe tener el signo = y nosotros queremos saber si nuestros datos son el reflejo de una población que superan dichas medias, nuestra hipótesis debe asignarse a* H_1*. Las hipótesis serán:*

$$\begin{aligned} H_0^{(H)} &: \mu_H \leq 177; H_1^{(H)} \mu_H > 177, \\ H_0^{(M)} &: \mu_M \leq 164; H_1^{(M)} \mu_M > 164, \end{aligned}$$

y sólo nos decantaremos por H_1 *si los datos lo apoyan de forma muy clara, es decir, si* $\bar{x}_H >> 177$

y si $\bar{x}_M >> 164$*. Los datos del fichero* ***estaturas.sf3*** *proporcionan los siguientes valores:*

$$\begin{aligned} t_0^{(H)} &= \frac{\bar{x}-\mu_0}{\hat{s}/\sqrt{n}} = \frac{175{,}9-177}{5{,}93/\sqrt{50}} = -1{,}31, \\ t_0^{(M)} &= \frac{166{,}2-164}{4{,}65/\sqrt{50}} = 3{,}35, \\ z_{0,05} &= 1{,}65. \end{aligned}$$

En el caso de los hombres, no haría falta hacer el contraste pues al ser $\bar{x}_H = 175{,}9 < \mu_0$ *está muy claro que no rechazaremos* H_0*, pues ese rechazo sólo ocurriría cuando haya mucha evidencia a favor de* H_1*, es decir cuando* $\bar{x} >> \mu_0$*. En ambos casos la región de rechazo, utilizando* $\alpha = 0{,}05$*, viene delimitada por el valor crítico* $z_{0,05} = 1{,}65$*, y la región de rechazo serán los valores* $t_0 > 1{,}65$*. Por tanto, rechazamos* H_0 *para las mujeres pero no para los hombres. Si la muestra recogida es realmente representativa de las jóvenes madrileñas y respondieron de forma sincera a su estatura, podemos decir que, con un nivel de significación del 5 %, las jóvenes universitarias madrileñas son más altas que las jóvenes españolas (por término medio).*

Se puede demostrar que la probabilidad de incurrir en un error Tipo I (rechazar H_0 cuando es cierta) mencionado anteriormente es precisamente el nivel de significación α. Para verlo, recordemos que α es el área de la distribución de referencia que reservamos para posicionar la región de rechazo. Recordemos también que la distribución de referencia es precisamente la distribución del estadístico de contraste cuando $\theta = \theta_0$. Es decir, si $\theta = \theta_0$ el estadístico de contraste que obtengamos con nuestros datos será un valor que proceda de la distribución de referencia. Puede entonces estar tanto en la zona central como en los extremos. Será más probable que dicho valor sea de la zona de más probabilidad de la distribución de referencia, y será poco probable que sea de las zonas de las colas. Por lo tanto, si H_0 es cierta podemos todavía obtener valores tanto en la región de aceptación, lo que llevaría a una decisión correcta (aceptar), como en la región de rechazo, lo que llevaría a una decisión incorrecta (rechazar). Vemos así que la probabilidad de que siendo $\theta = \theta_0$ estemos en la región de rechazo es entonces α. Por tanto, si el estadístico de contraste nos da un valor que cae en la región de rechazo, habría dos interpretaciones posibles:

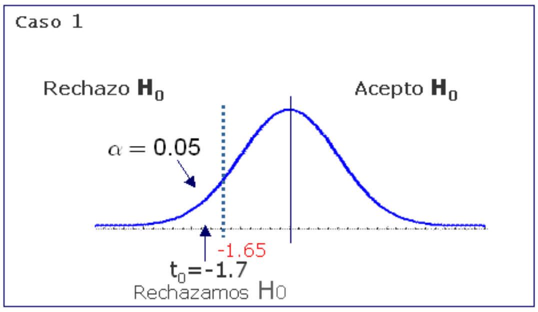
1. Que H_0 es cierta, pero por azar de la muestra el estadístico ha tomado un valor muy extremo que cae en la región de rechazo. Esto ocurrirá en un $\alpha \times 100\,\%$ de las muestras y el azar ha querido que la nuestra sea una de ellas. Si α es pequeño, será un suceso muy improbable. La probabilidad de tomar una decisión errónea será entonces igual a α. O bien,

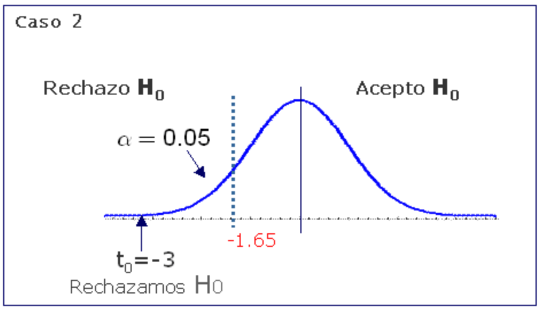
2. Que H_0 es falsa y por eso era de esperar un valor del estadístico de contraste tan extremo que nos colocase en la región de rechazo, y debemos rechazar H_0

Lo que hacemos en la práctica es que como α es suficientemente pequeño, es más sensato decantarnos por la interpretación 2 y rechacemos H_0. No debemos pensar que usando un α excesivamente pequeño, por ejemplo $\alpha = 10^{-7}$ mejoraremos las propiedades del contraste, pues si diminuimos α aumentaremos la región de aceptación, y con ello la probabilidad de aceptar H_0 siendo falsa (error Tipo II). Por tanto, la simple manipulación de α no nos ayuda a evitar errores pues lo que hace es disminuir un tipo de error (error Tipo I) a costa de aumentar el otro (error Tipo II).

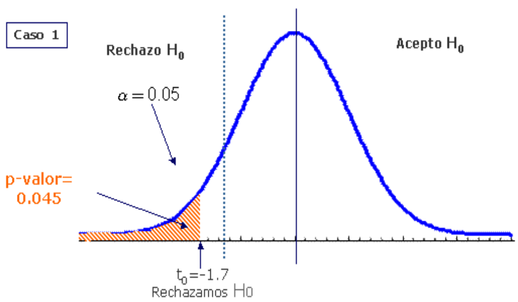
6.6. Interpretación de un contraste usando el p-valor

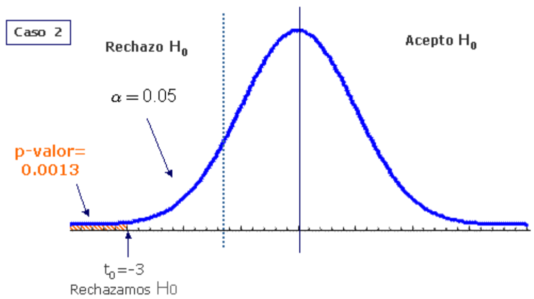
En la sección anterior hemos visto cómo resolver un contraste comparando el valor del estadístico de contraste con el valor crítico que corresponde con el α utilizado. Con este procedimiento, el resultado del contraste es sólo si rechazamos o no rechazamos H_0. Si por ejemplo concluimos que se rechaza H_0 no sabremos si el valor del estadístico estaba cerca del valor crítico, y por lo tanto ha sido rechazado 'por los pelos', o por el contrario el valor del estadístico de contraste estaba muy dentro de la región de rechazo, y el rechazo ha sido con mucha seguridad. La figura siguiente ilustra estas dos posibles situaciones. En la figura de la izquierda (Caso 1), el valor del estadístico t_0 (-1.7) está muy próximo al valor crítico (-1.65), rechazándose H_0 por escaso margen. En la figura de la derecha (Caso 2) el estadístico de contraste ($t_0 = -3$) está mucho más dentro de la región de rechazo, rechazándose la hipótesis nula con más seguridad que en el Caso 1. En este segundo caso, la muestra refleja más claramente que la población no verifica dicha hipótesis.





Existe otra forma de dar el resultado de un contraste, que nos diga si la decisión tomada (rechazar o aceptar) se hace con más o menos incertidumbre. Consiste en utilizar el llamado **p-valor del contraste**. **El p-valor es el nivel de significación que deberíamos usar para dejar al valor del estadístico de contraste justo en la frontera de la región de rechazo**. O bien, es el mínimo nivel de significación que nos llevaría a rechazar la hipótesis nula. Las figuras siguientes nos muestran el p-valor correspondiente con los dos casos mostrados en la figura anterior.



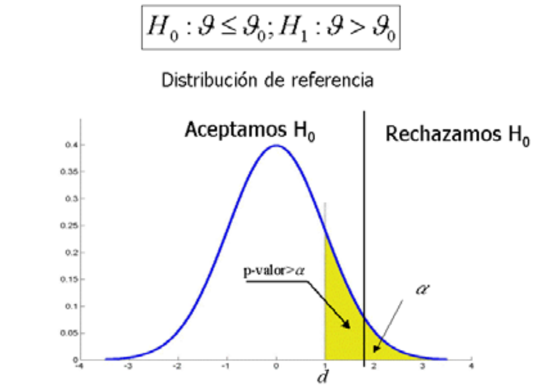


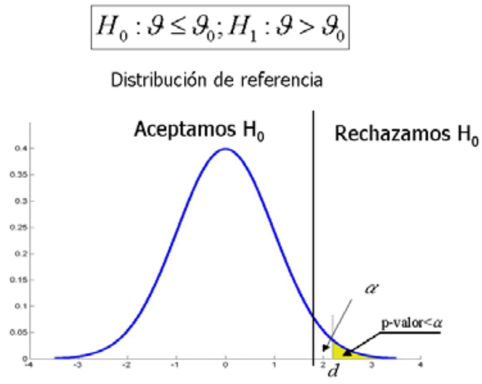
La regla de decisión basada en el p-valor será entonces:

- Si el p-valor es mayor que α entonces estamos dentro de la región de aceptación.
- Si el p-valor es menor que α entonces estamos en la región de rechazo.

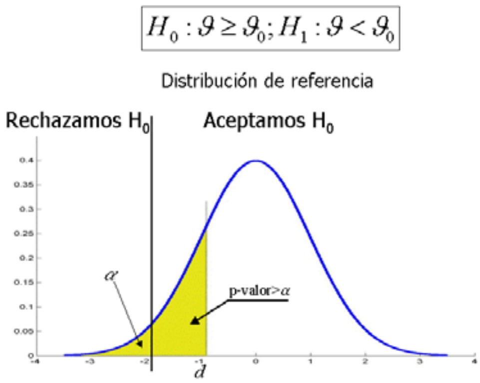
Cuanto menor sea el p-valor más seguridad tendremos en rechazar $\mathbf{H}_0$, mientras que un elevado p-valor nos dará más seguridad a la hora de aceptar $\mathbf{H}_0$. Si denotamos por τ a la distribución de referencia usada en el contraste (por ejemplo, en el caso de la media $\tau \sim N(0,1)$) y d al valor concreto del estadístico de contraste (z_0 ó t_0 en el caso de la media) obtenido con unos datos, el p-valor se calculará de la siguiente forma:

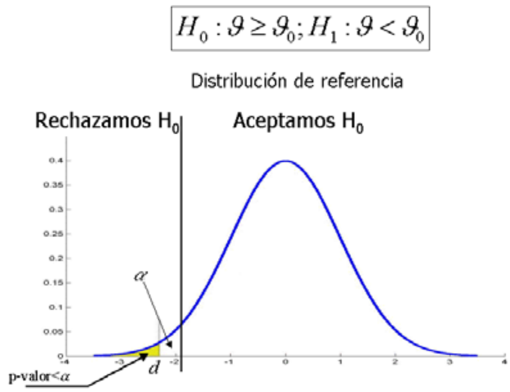
- Si el contraste es unilateral con $\mathrm{H}_1 : \theta > \theta_0$ el p-valor será $P(\tau > d)$. La siguiente figura ilustra la posición de este p-valor



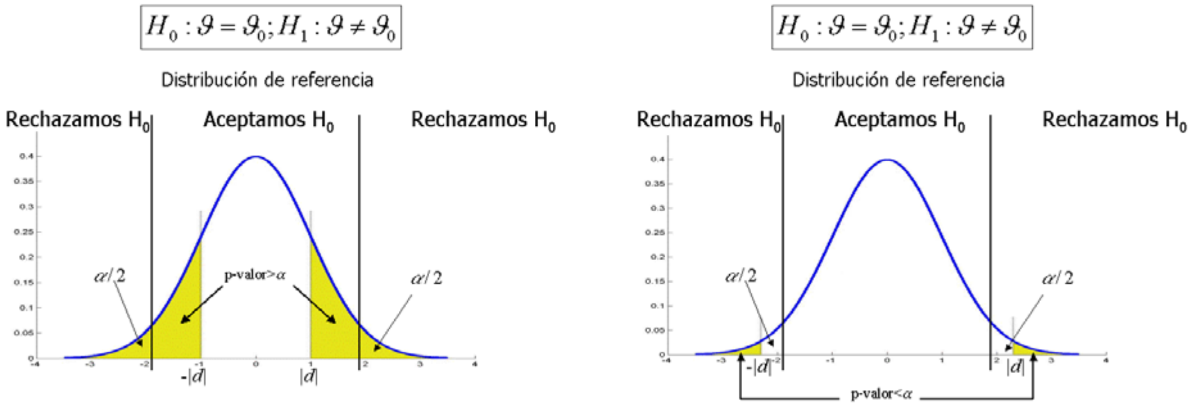


- Si el contraste es unilateral con $\mathrm{H}_1 : \theta < \theta_0$ el p-valor será $P(\tau < d)$. La siguiente figura ilustra la posición de este p-valor





- Si el contraste es bilateral, con $\mathrm{H}_1 : \theta \neq \theta_0$ el p-valor será $P(\tau > |d|) + P(\tau < -|d|)$. La localización de este p-valor a ambos lados de la distribución de referencia se ilustra en la siguiente figura



6.7. Relación entre contrastes de hipótesis e intervalos de confianza

Tanto los intervalos de confianza como los contrastes de hipótesis se basan en las mismas propiedades del estimador que esté implicado en el contraste. En ambos casos se acude a la distribución del estimador en el muestreo. Por esta razón, ambos puden interpretarse comno formas diferentes de utilizar la misma información: la variabilidad de la estimación de un parámetro debido a la selección de la muestra.

Se puede demostrar que la realización de un contraste de hipótesis bilateral con nivel de significación α es equivalente a realizar un intervalo de confianza de nivel $(1-\alpha)$ y comprobar si μ_0 está dentro o fuera de dicho intervalo.

Utilizando los datos del Ejemplo 5, podemos construir un intervalo de confianza para la media poblacional del coeficiente β de los transistores a partir de la muestra observada. El intervalo de confianza de nivel $(1-\alpha)$ es

$$\mu \in \left\{ \bar{x} \pm z_{\alpha/2} \frac{\hat{s}}{\sqrt{n}} \right\} = 282{,}3 \pm 1{,}96 \frac{27{,}69}{10} = 282{,}3 \pm 5{,}4 = [276{,}9; 287{,}7] \,,$$

y que puede verse que con un nivel de confianza $(1-\alpha)$ no contiene a $\mu_0 = 290$. Por tanto, rechazamos la hipótesis nula de $\mu = 290$ con un nivel de significación de $\alpha = 0{,}05$.

6.8. Inferencia sobre una proporción en muestras grandes

6.8.1. Estimación

Sea p una proporción poblacional. Es decir, la probabilidad de que un elemento de la población elegido al azar tenga alguna característica, o bien la proporción de individuos de la población que tiene dicha característica. Por ejemplo, la probabilidad de que un artículo sea defectuoso, la propoción de individuos que votarán a cierto candidato, la probabilidad de que cierta variable supere cierto valor, etc. Esta probabilidad p, es también el parámetro de un proceso de Bernoulli. Sea X_i una variable aleatoria de Bernoulli que toma el valor 0 si el elemento no tiene el atributo de interés, y el valor 1 cuando el elemento sí tiene el atributo. Entonces p es la probabilidad de que una variable X_i tome valor 1. Recordemos que $E(X_i) = p$, $\text{Var}(X_i) = p(1-p)$.

La proporción o probabilidad poblacional p la estimaremos con la proporción muestral, es decir

$$\hat{p} = \frac{\text{número de elementos de la muestra con el atributo de interés}}{\text{número total de elementos en la muestra}}$$

Como ya vimos en capítulos anteriores, esta proporción puede escribirse como el promedio de las variables de Bernoulli asociadas a cada elemento, es decir:

$$\hat{p} = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

donde $X_i = 1$ si el elemento i-ésimo tiene el atributo de interés, y $X_i = 0$ en caso contrario. Por tanto estamos ante una media muestral y se puede aplicar lo visto en la sección anterior, con la única diferencia de la notación. En lugar de $\bar{X}$ usamos $\hat{p}$, en lugar de μ usamos p y en lugar de σ^2 usamos $p(1-p)$. Se tiene entonces que, usando (6.1), si n es elevado ($n > 30$) las diferentes proporciones muetrales $\hat{p}$ que podemos obtener al cambiar la muestra tienen la siguiente distribución

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right). \tag{6.15}$$

6.8.2. Intervalo de confianza de p

A partir de (6.15) es ya fácil construir un intervalo de confianza para p. Para construir un intervalo de confianza sobre el verdadero valor de p a partir de una muestra, podemos aplicar el resultado (6.7) basado en las propiedades de la media muestral con tamaño muestrales grandes. Un intervalo de confianza de nivel de confianza $(1-\alpha)$ será, si n es grande,

$$p \in \left\{\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right\} \tag{6.16}$$

6.8.3. Determinación del tamaño muestral

Al igual que antes, el intervalo de confianza se puede expresar como

$$p \in \left\{\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right\} = \hat{p} \pm L \tag{6.17}$$

El ancho del intervalo será entonces $2L$. Si queremos un valor de L máximo ¿Cuál debe ser el n mínimo para un valor de p dado? Operando en (6.17) tenemos que

$$L = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \Rightarrow n = \left(\frac{z_{\alpha/2}\sqrt{p(1-p)}}{L}\right)^2. \tag{6.18}$$

El problema para aplicar (6.18) es que p es desconocido. Una opción es hacer una muestra piloto y dar una primera estimación de p, y usar ese valor para obtener un n. Otra opción es usar el caso más desfavorable del valor $p(1-p)$ que proporcionaría un valor de n más alto del que realmente necesitásemos, pero que nos garantizaría un ancho L deseado. El valor más alto de $p(1-p)$ se observa cuando $p = 0{,}5$. Por tanto, un valor de n será

$$n = \left(\frac{z_{\alpha/2}}{2L}\right)^2$$

Como normalmente usamos $\alpha = 0{,}05$, y como para ese valor $z_{0{,}025} = 1{,}96 \approx 2$ es muy frecuente usar como tamaño muestral

$$n = \frac{1}{L^2}.$$

Ejemplo 7 *Con el objeto de determinar la proporción de personas que poseen coche en una provincia determinada se realizó un muestreo aleatorio simple a 100 perdonas, de tal forma que de los 100 encuestados, 30 de ellos tenían coche.*

1. *Determinar un intervalo de confianza para la proporción de la población que poseen coche* $(\alpha = 0{,}05)$.
2. *Si se desease estimar dicha proporción con una precisión de 0.02* $(L = 0{,}02)$ *y un 95 % de confianza, ¿a cuántas personas se debería encuestar?*

SOLUCIÓN:

1. *El intervalo es de la forma* $\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$, *con* $\widehat{p} = 0{,}3 \quad z_{\alpha/2} = 1{,}96$. *Entonces*

$$IC(1-\alpha) : p \in 0{,}3 \pm 1{,}96\sqrt{\frac{0{,}3 \cdot 0{,}7}{100}} = (0{,}21, 0{,}39)$$

2. *De acuerdo con (6.18)*

$$L = 0{,}02 = z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = 1{,}96\sqrt{\frac{0{,}3 \cdot 0{,}7}{n}} \Longleftrightarrow n = \left(\frac{1{,}96}{0{,}02}\right)^2 0{,}3 \cdot 0{,}7 \approx 2017 \text{ \textit{encuestados.}}$$

6.8.4. Contraste de hipótesis de una proporción

Se quiere contrastar alguna de las siguientes hipótesis:

1. $\mathrm{H}_0 : p = p_0$; $\mathrm{H}_1 : p \neq p_0$
2. $\mathrm{H}_0 : p \geq p_0$; $\mathrm{H}_1 : p < p_0$
3. $\mathrm{H}_0 : p \leq p_0$; $\mathrm{H}_1 : p > p_0$

El contraste se basa en las propiedades de la proporción muestral p, que si n es grande verifica que

$$\hat{p} \sim N\left(p, \frac{pq}{n}\right)$$

y por tanto, estandarizando,

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

El estadístico de contraste y la región de rechazo serán:

Contrastes	Estadísticos de contraste	Distribución de referencia	Región de rechazo
(1)-$H_0 : p = p_0$; $H_1 : p \neq p_0$ (2)-$H_0 : p \geq p_0$; $H_1 : p < p_0$ (3)-$H_0 : p \leq p_0$; $H_1 : p > p_0$	$Z_0 = \dfrac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}$	$N(0,1)$	(1) $\lvert z_0 \rvert > z_{\alpha/2}$ (2) $z_0 < -z_\alpha$ (3) $z_0 > z_\alpha$

Ejemplo 8 *Un proceso productivo que fabrica semiconductores produce un 2 % de artículos defectuosos cuando el proceso productivo funciona adecuadamente. Con el fin de mejorar la calidad de la producción se adquiere una nueva máquina basada en una tecnología más avanzada. Después de producir 200 artículos con la nueva máquina se encuentra que 2 son defectuosos. ¿Se puede afirmar que la nueva máquina ha mejorado la calidad de la producción?.*

SOLUCIÓN:

En este ejemplo tenemos que $p_0 = 0{,}02$ *y* $\hat{p} = 2/200 = 0{,}01$. *Si queremos saber si ha habido una disminución en* p, *nuestro contraste será*

$$\begin{aligned} H_0 &: p \geq 0{,}02 \\ H_1 &: p < 0{,}02 \end{aligned}$$

y consideraremos entonces que el proceso sigue igual salvo que demostremos que $\hat{p}$ *supone una disminución significativa, más allá de la variabilidad que se pueda asignar al azar de la muestra. Con los datos tenemos que, usando un nivel de significación del 5 %,*

$$\begin{aligned} z_0 &= \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} = \frac{0{,}01 - 0{,}02}{\sqrt{0{,}02 \times 0{,}98/200}} = -1{,}01 \\ -z_{0{,}025} &= -1{,}96 \end{aligned}$$

Como -1.01>-1.96 no podemos asegurar, con un nivel de significación del 5 %, que haya habido una disminución de la proporción de artículos defectuosos. Puede ser que la máquina no produzca dicha disminución (tal vez se esté usando incorrectamente) o tal vez la mejora sea pequeña y se necesite de una muestra mayor para apreciarla.

Si comparamos el intervalo de confianza (6.16) y el estadístico de contraste que se utiliza, vemos que en ambas fórmulas se utiliza una expersión diferente para la varianza de la proporción muestral $\sqrt{pq/n}$. En el intervalo de confianza se sustituye p par la estimación $\hat{p}$, mientas que en el estadístico de contraste se utiliza el valor p_0. Esta diferencia hace que el resultado de un contraste bilateral y un intervalo de confianza no sean ya exactamente iguales. Para que fuesen iguales, el estadístico de contraste debería ser

$$Z_0 = \frac{\hat{p} - p_0}{\sqrt{\hat{p}\hat{q}/n}},$$

en lugar de

$$Z_0 = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}.$$

Es frecuente no obstante seguir interpretando el intrvalo de confianza como si fuese un contraste, pero esa interpretación es menos precisa que si realizamos directamente el contraste de hipótesis.

6.9. Inferencia con estimadores de máxima verosimilitud (avanzado, I. Informática)

6.9.1. Intervalo de confianza

Sea $\hat{\theta}_{MV}$ es estimador de maxima verosimilitud del parámetro θ. Vimos en el tema anterior que este estimador es asintóticamente normal con las siguientes características

$$\hat{\theta}_{MV} \sim N\left(\theta, \mathrm{Var}(\hat{\theta}_{MV})\right), \tag{6.19}$$

donde

$$\mathrm{Var}(\hat{\theta}_{MV}) = -\left[\frac{d^2L(\theta)}{d\theta^2}\right]^{-1},$$

siendo $L(\theta)$ la función soporte. Al poder utilizar esta propiedad de normalidad asintótica, se pueden utilizar los resultados (6.3) y (6.6) para construir intervalos de confianza de θ en muestras grandes. Se tiene entonces que el intervalo de confianza de θ para muestras grandes, con nivel de confianza $(1-\alpha)$ será

$$\theta \in \left\{\hat{\theta}_{MV} \pm z_{\alpha/2}\sqrt{\mathrm{Var}(\hat{\theta}_{MV})}\right\}.$$

La varianza asintótica $\mathrm{Var}(\hat{\theta}_{MV})$ dependerá de θ, por lo que habrá que reemplazarla por un estimador $\widehat{\mathrm{Var}}(\hat{\theta}_{MV})$ que utilice $\hat{\theta}$ en lugr de θ. Un intervalo de confianza asintótico de θ basado en estimadores de máxima verosimilitud será entonces

$$\theta \in \left\{\hat{\theta}_{MV} \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{\theta}_{MV})}\right\}$$

Ejemplo 9 *La velocidad de una molécula, según el modelo de Maxwell, es una variable aleatoria con función de densidad,*

$$f(x) = \begin{cases} \frac{4}{\sqrt{\pi}}\frac{1}{\alpha^3}x^2e^{-(x/\alpha)^2}, & x \geq 0 \\ 0, & x < 0. \end{cases}$$

Donde $\alpha > 0$, es el parámetro de la distribución. Se desea estimar el parámetro α. Si calculásemos $E(X)$ veríamos que no guarda una relación sencilla con α, que permita estimar a patir de $\bar{X}$. Queremos construir un intervalo de confianza para α.

Este problema ya se planteó en el tema anterior obteniéndose que

$$\begin{aligned} L(\alpha) &= k - 3n\log\alpha - \frac{\sum_{i=1}^n x_i^2}{\alpha^2} \\ \hat{\alpha}_{MV} &= \sqrt{\frac{2\sum_{i=1}^n X_i^2}{3n}} \end{aligned}$$

Por tanto, podemos calcular la expresión de la varianza asintótica:

$$Var(\hat{\alpha}_{MV}) = -\left[\frac{d^2L}{d\alpha^2}\right]^{-1}.$$

Como

$$\frac{d^2L}{d\alpha^2} = \frac{3n\alpha^2 - 6\sum_{i=1}^n x_i^2}{\alpha^4},$$

entonces

$$Var(\hat{\alpha}_{MV}) = \frac{\alpha^4}{6\sum x_i^2 - 3n\alpha^2}.$$

Sustituyendo α por $\hat{\alpha}_{MV}$ y operando se obtiene un estimador de esta varianza asintótica:

$$\widehat{Var}(\hat{\alpha}_{MV}) = \frac{\hat{\alpha}_{MV}^4}{6\sum x_i^2 - 3n\hat{\alpha}_{MV}^2}$$

$$\widehat{Var}(\hat{\alpha}_{MV}) = \frac{\hat{\alpha}_{MV}^2}{6n}.$$

Por tanto

$$\alpha \in \left\{ \hat{\alpha}_{MV} \pm z_{\alpha/2}\sqrt{\frac{\hat{\alpha}_{MV}^4}{6\sum x_i^2 - 3n\hat{\alpha}_{MV}^2}} \right\}.$$

De esta forma, dada una muestra de tamaño n suficientemente grande, podemos calcular los valores de ese intervalo.

6.9.2. Contraste de hipótesis

Se quiere contrastar alguna de las siguientes hipótesis

1. $\mathrm{H}_0 : \theta = \theta_0$; $\mathrm{H}_1 : \theta \neq \theta_0$
2. $\mathrm{H}_0 : \theta \geq \theta_0$; $\mathrm{H}_1 : \theta < \theta_0$
3. $\mathrm{H}_0 : \theta \leq \theta_0$; $\mathrm{H}_1 : \theta > \theta_0$

Para la realización de cualquiera de estos contrastes utilizaremos la propiedad de normalidad asintótica (6.19). Entonces, el diseño del contraste es muy similar al que se hizo para el caso de la media μ en muestras grandes. Estandarizando en (6.19) tendremos que el estadístico de contraste será similar al estadístico Z_0 utilizado en (6.14):

$$Z_0 = \frac{\hat{\theta}_{MV} - \theta_0}{\sqrt{\widehat{\mathrm{Var}}(\hat{\theta}_{MV})}},$$

y la distribución de referencia será la $N(0, 1)$. La siguiente tabla resume las características de estos contrastes.:

Contrastes	Estadísticos de contraste	Distribución de referencia	Región de rechazo
(1)-$\mathrm{H}_0 : \theta = \theta_0$; $\mathrm{H}_1 : \theta \neq \theta_0$ (2)-$\mathrm{H}_0 : \theta \geq \theta_0$; $\mathrm{H}_1 : \theta < \theta_0$ (3)-$\mathrm{H}_0 : \theta \leq \theta_0$; $\mathrm{H}_1 : \theta > \theta_0$	$Z_0 = \frac{\hat{\theta}_{MV} - \theta_0}{\sqrt{\widehat{\mathrm{Var}}(\hat{\theta}_{MV})}}$	$Z_0 \sim N(0,1)$	(1) $\lvert z_0 \rvert > z_{\alpha/2}$ (2) $z_0 < -z_\alpha$ (3) $z_0 > z_\alpha$