

Clase 11: Dualidad entre TdH e IdC

Matías Carrasco

5 de octubre de 2019

Índice

1. Estimación por intervalos de confianza	1
2. El z -intervalo (exacto) para la media de datos normales	6
3. El z -intervalo (asintótico) para muestras grandes	10
4. El t -intervalo (exacto) para la media de datos normales	11

1. Estimación por intervalos de confianza

Supongamos que tenemos un modelo (distribución de probabilidad) para datos observados con un parámetro desconocido. Hemos visto cómo un TdH usa los datos para probar la hipótesis de que el parámetro desconocido tiene un valor particular.

También hemos visto cómo las estimaciones puntuales, como el estimador de máxima verosimilitud, utilizan los datos para proporcionar una estimación del parámetro desconocido. Por sí sola, una estimación puntual como $\bar{x} = 2.2$ no contiene información sobre su precisión; es solo un número, independientemente de si se basa en diez datos o en un millón de datos.

Por esta razón, los estadísticos engordan las estimaciones puntuales con *intervalos de confianza*. Por ejemplo, para estimar una media desconocida μ podríamos decir que nuestra mejor estimación de la media es $\bar{x} = 2.2$ con un intervalo de confianza del 95% igual a $[1.2, 3.2]$. Otra forma de describir el intervalo es: $\bar{x} \pm 1 = 2.2 \pm 1$.

Dejaremos para más adelante la explicación de qué significa exactamente el nivel de confianza del 95%. Por ahora, notaremos que, tomados en conjunto el ancho del intervalo y el nivel de confianza, proporcionan una medida de la fuerza de la evidencia que apoya la hipótesis de que μ está cerca de nuestra estimación \bar{x} . Se puede interpretar al nivel de confianza de un intervalo como análogo al nivel de significación de un TdH. No es casualidad que se use $\alpha = 0.05$ como nivel de significación y $1 - \alpha = 0.95$ como nivel de confianza.

Primero exploraremos los intervalos de confianza en situaciones en las que podemos calcularlos fácilmente. Más adelante veremos cómo se relacionan con los TdH.

Precisión en la estimación de una proporción

Comencemos con un ejemplo. Una caja contiene una gran cantidad de bolitas de varios colores. Estamos interesados en la proporción θ de bolitas rojas en la caja.

En un primer intento sacamos 10 bolitas de la caja, de las cuáles 9 son rojas. Podríamos estimar θ con $\hat{\theta} = 9/10 = 0.9$. Pero surgen dos preguntas claves:

1. Por casualidad, podríamos haber sacado (en proporción) muchas más bolitas verdes de las que hay en la caja. O incluso, aunque parezca bastante improbable, podríamos haber sacado la única bola de otro color que hay en la caja. ¿Qué *confianza* tenemos en nuestra estimación?
2. ¿Cuán precisa es nuestra estimación? Será que la verdadera proporción de bolitas rojas es 0.91? ¿0.85? ¿Cuán lejos de 0.9 puede estar?

Para responder a la primera pregunta podemos repetir el procedimiento varias veces. Así que devolvemos las bolitas a la caja, entreveramos bien, y sacamos otra vez 10 bolitas. Resulta que hay 4 bolitas verdes entre las 10.

Esto complica un poco las cosas, ¿con cuál nos quedamos?. Mejor repetir el experimento para estar seguros. Así que repetimos el procedimiento 8 veces más, obteniendo 7, 8, 8, 9, 6, 5, 7, y 8. Pongamos los resultados en una tabla

Resultados de sacar 10 bolitas										
Intento	1	2	3	4	5	6	7	8	9	10
Estimación $\hat{\theta}$	0.9	0.4	0.7	0.8	0.8	0.9	0.6	0.5	0.7	0.8

¿Con qué estimación nos quedamos? Aquí surge una respuesta natural: con el promedio de las 10, que es 0.71. Pero esto es un poco tramposo, porque el promedio de las 10 corresponde a sacar 100 de una vez y estimar la proporción entre esas 100.

En la Figura 1 se muestran los resultados de diferentes estimaciones. Cada línea punteada contiene una extracción de bolitas de la caja representada por un punto. La coordenada horizontal de ese punto es la proporción de bolitas rojas en esa extracción. Las primeras 10 líneas corresponden a las 10 extracciones que están en la tabla anterior. Las siguientes 10 líneas corresponde a extracciones de 20 bolitas. Cada 10 líneas aumenta la cantidad de bolitas extraídas de la caja, en la progresión

$$10, 20, 40, 80, 100, 150, 200, 250, 500, 750, 1000.$$

La figura muestra que aunque parezca difícil decirlo correctamente, hay un cierto sentido de precisión y confianza en la estimación, que aumenta a medida que la cantidad de bolitas extraídas de la caja aumenta. Confiamos más en un $\hat{\theta}$ de las últimas 10 líneas que en uno de las primeras. Nuestro objetivo ahora es cuantificar este fenómeno.

Consideremos nuevamente el caso en que sacamos 10 bolitas de la caja. Hemos aprendido en las últimas dos clases a poner a prueba las siguientes hipótesis

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_A : \theta \neq \theta_0 \end{cases}$$

Supongamos que nuestra estimación es $\hat{\theta} = 0.8$, es decir, 8 de las 10 bolitas extraídas son rojas. Basados en esta estimación, podemos definir un intervalo de valores creíbles de θ_0 , por ejemplo $\theta_0 = 0.85$ es claramente creíble, pero $\theta_0 = 0.1$ no lo es tanto.

Proporciones de bolitas rojas

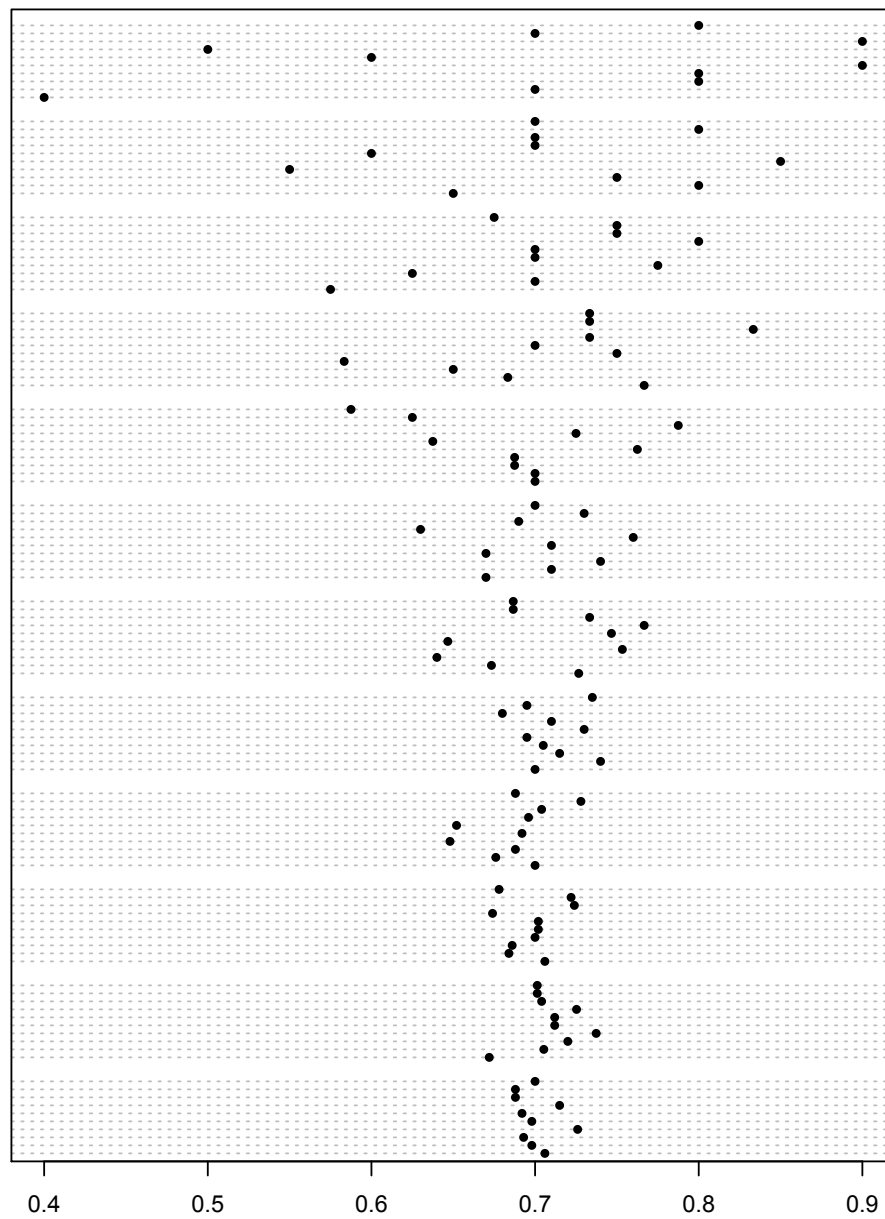
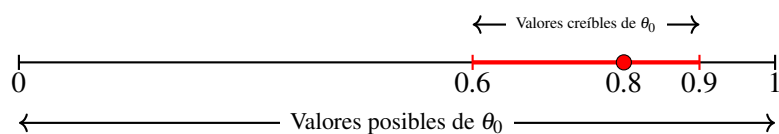


Figura 1: Variación en las proporciones de bolitas rojas.



Una forma de decidir si un valor de θ_0 es creíble es haciendo el test: si el test rechaza H_0 entonces θ_0 no es creíble, si no rechaza entonces sí lo es.

En la siguiente tabla se muestra la región de rechazo para varios valores posibles de θ_0 . Hemos usado el nivel de significación $\alpha = 0.1$ (para variar un poco). La naturaleza discreta del problema hace que no se pueda elegir la región de rechazo de forma que el nivel de significación sea exactamente α .

En la tabla, las regiones de rechazo están indicadas por las celdas pintadas de rojo. Cada fila corresponde a un valor particular de θ_0 , y las columnas a los diferentes valores de $\hat{\theta}$. Si $\hat{\theta}$ cae en una celda roja, se rechaza H_0 , y si cae en una celda verde no se rechaza. En cada fila, la suma de las probabilidades de las celdas rojas es aproximadamente 0.1.

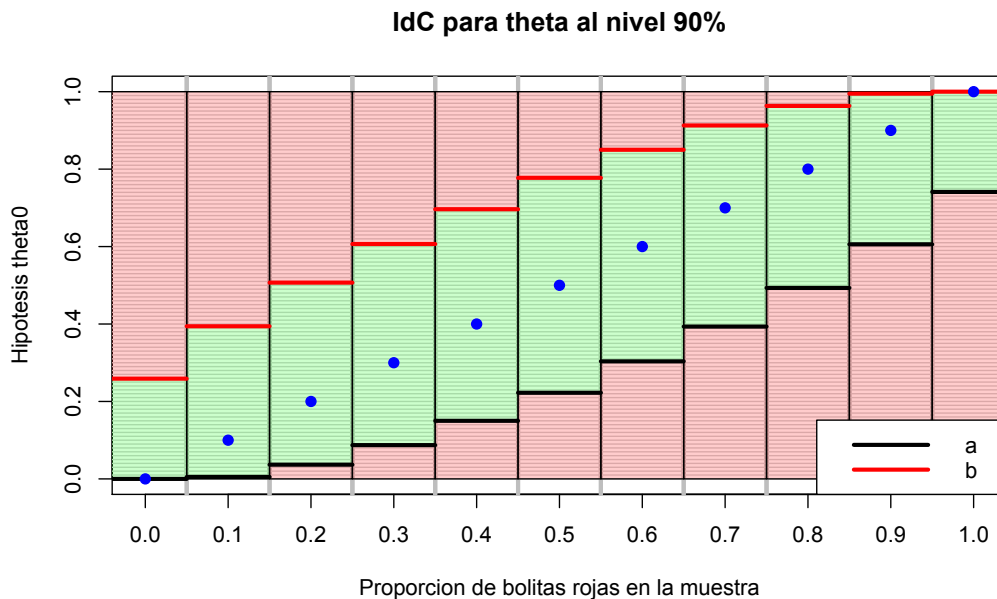
$\hat{\theta} =$	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$\theta_0 = 1.0$.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.00
$\theta_0 = 0.9$.000	.000	.000	.000	.000	.001	.011	.057	.194	.387	.349
$\theta_0 = 0.8$.000	.000	.000	.001	.006	.026	.088	.201	.302	.268	.107
$\theta_0 = 0.7$.000	.000	.001	.009	.037	.103	.200	.267	.233	.121	.028
$\theta_0 = 0.6$.000	.002	.011	.042	.111	.201	.251	.215	.121	.040	.006
$\theta_0 = 0.5$.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001
$\theta_0 = 0.4$.006	.040	.121	.215	.251	.201	.111	.042	.011	.002	.000
$\theta_0 = 0.3$.028	.121	.233	.267	.200	.103	.037	.009	.001	.000	.000
$\theta_0 = 0.2$.107	.268	.302	.201	.088	.026	.006	.001	.000	.000	.000
$\theta_0 = 0.1$.349	.387	.194	.057	.011	.001	.000	.000	.000	.000	.000
$\theta_0 = 0.0$	1.00	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

La columna correspondiente a $\hat{\theta} = 0.8$ está marcada por una doble línea. ¿Cuáles son los valores creíbles de θ_0 ? Las celdas verdes de la columna $\hat{\theta} = 0.8$ corresponden a aquellos valores de θ_0 para los cuales el test “aceptaría” H_0 . Según este criterio entonces, los valores creíbles de θ_0 son

$$\{0.6, 0.7, 0.8, 0.9\}.$$

Hemos obtenido así una especie de intervalo de valores creíbles, en el cual confiamos que aproximadamente el 90% de las veces contiene el verdadero valor de la proporción de bolitas verdes en la urna.

Es importante entender bien la información contenida en esta tabla. Las columnas contienen la distribución de probabilidad para cada hipótesis nula posible θ_0 . Es decir, la función $p(x; \theta_0)$ con x como variable. Las columnas contienen la función de verosimilitud $p(\hat{\theta}, \theta_0)$ con θ_0 como variable. Por ejemplo, cuando $\hat{\theta} = 0.8$, vemos que el estimador de máxima verosimilitud de θ es justamente 0.8, pues es el valor que maximiza la verosimilitud.



En realidad, nuestro intervalo está bastante incompleto, porque nuestra tabla está bastante incompleta. Tendríamos que incluir una fila para cada valor posible de θ_0 en el intervalo $[0, 1]$. Como esto es obviamente imposible, lo que podemos hacer es cambiar la tabla por un gráfico.

El gráfico anterior muestra el resultado. La coordenada vertical representa cada hipótesis alternativa θ_0 posible. La coordenada horizontal es discreta, y representa los 11 valores posibles de proporción de bolitas rojas entre las 10 extraídas, desde 0/10 a 10/10. En cada columna, la zona verde representa los valores creíbles de θ_0 . Los puntos en azul representan el estimador de máxima verosimilitud en cada caso.

Usando el gráfico, y razonando de la misma forma que con la tabla, el intervalo creíble de θ_0 es $[0.49, 0.96]$. Esto responde a las dos preguntas que hicimos más arriba. En lugar de decir que la estimación es 0.8, decimos que la estimación está entre 0.49 y 0.96 con una confianza del 90%. Esto nos da la precisión (ancho del intervalo) y la confianza al mismo tiempo.

Definición informal de intervalo de confianza

Cuando queremos estimar un parámetro desconocido θ , la estimación por un intervalo de confianza consiste en dar un rango de valores creíbles de θ , en el cual confiemos al nivel $(1 - \alpha)100\%$. La precisión dada por el intervalo es su ancho.

La estimación mediante intervalos de confianza no tiene porqué contener al verdadero valor del parámetro, pero la probabilidad de que sí lo contenga es $1 - \alpha$.

Definición de intervalo de confianza

Sea X_1, \dots, X_n un muestreo aleatorio de X que tiene distribución $p(x; \theta)$ que depende de un parámetro θ .

Dado $\alpha \in (0, 1)$, un *intervalo de confianza* al nivel de confianza $1 - \alpha$ para θ es un intervalo de la forma

$$I_\alpha = I_\alpha(X_1, \dots, X_n) = [a(X_1, \dots, X_n), b(X_1, \dots, X_n)],$$

en donde a y b son tales que

$$\mathbf{P}(\theta \in I_\alpha(X_1, \dots, X_n)) = 1 - \alpha.$$

Es decir, los extremos del intervalo son variables aleatorias que dependen de la muestra X_1, \dots, X_n , y son tales que la probabilidad de que entre ellos esté el parámetro θ es $1 - \alpha$.

Técnicamente, un intervalo de confianza no es más que un par de estadísticos puntuales que proporcionan los límites inferior y superior del intervalo. Es importante enfatizar que el intervalo es un estadístico, por la siguiente razón: el intervalo es aleatorio, nuevos datos producirán un nuevo intervalo.

Una advertencia pedante para prevenir malas interpretaciones: el nivel de confianza no es la probabilidad de que el parámetro desconocido se encuentre en el intervalo de confianza, sino la probabilidad de que el intervalo de confianza contenga al parámetro desconocido. El parámetro desconocido es fijo, y no tiene sentido hablar de la probabilidad de que pertenezca a un intervalo dado.

Para ser más precisos, la definición que dimos arriba es la de un IdC exacto. Pero frecuentemente uno se encuentra con otros tipos de IdC:

1. IdC *exacto*: es cuando $\mathbf{P}(\theta \in I_\alpha(X_1, \dots, X_n)) = 1 - \alpha$.
2. IdC *asintótico*: es cuando $\lim_{n \rightarrow \infty} \mathbf{P}(\theta \in I_\alpha(X_1, \dots, X_n)) = 1 - \alpha$.
3. IdC *conservador*: es cuando $\mathbf{P}(\theta \in I_\alpha(X_1, \dots, X_n)) \geq 1 - \alpha$.

2. El z -intervalo (exacto) para la media de datos normales

Deducción a partir de la definición

Supongamos que X_1, \dots, X_n es un muestreo aleatorio de X . Supongamos además que X tiene distribución normal $N(\mu, \sigma^2)$, en donde σ^2 es conocido, pero μ es desconocido. Ya nos hemos encontrado con ejemplos de esta situación, sobre todo en el control de calidad.

Sabemos que un estimador consistente para μ es

$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n}.$$

Recordar que

$$\mathbf{E}(\bar{X}_n) = \mu \text{ y } \mathbf{var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Además, \bar{X}_n tiene distribución normal $N(\mu, \sigma^2/n)$.

Queremos construir un intervalo de confianza exacto para la media μ . Basámonos en la simetría de la distribución normal, busquemos un intervalo simétrico y centrado en \bar{X}_n .

Nos damos:

- un error $\varepsilon > 0$,
- y una probabilidad de error $\alpha \in (0, 1)$.

Como es costumbre, típicamente $\alpha = 0.1, 0.05, 0.01$.

Nos gustaría que encontrar la relación entre α y ε a partir de

$$\mathbf{P}(|\bar{X}_n - \mu| > \varepsilon) = \alpha.$$

Calculando:

$$\begin{aligned} \mathbf{P}(|\bar{X}_n - \mu| > \varepsilon) &= \mathbf{P}\left(\left|\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}\right| > \frac{\varepsilon\sqrt{n}}{\sigma}\right) \\ &= \mathbf{P}\left(|Z| > \frac{\varepsilon\sqrt{n}}{\sigma}\right) \\ &= \mathbf{P}\left(Z > \frac{\varepsilon\sqrt{n}}{\sigma}\right) + \mathbf{P}\left(Z < -\frac{\varepsilon\sqrt{n}}{\sigma}\right) \end{aligned}$$

Aquí Z es una variable con distribución $N(0, 1)$.

Como la función de distribución Φ es simétrica:

$$\begin{aligned} \mathbf{P}(|\bar{X}_n - \mu| > \varepsilon) &= 1 - \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) + \Phi\left(-\frac{\varepsilon\sqrt{n}}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) + 1 - \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) \\ &= 2 \left[1 - \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)\right]. \end{aligned}$$

Es decir que

$$\mathbf{P}(|\bar{X}_n - \mu| > \varepsilon) = \alpha \Leftrightarrow \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) = 1 - \frac{\alpha}{2}.$$

Sea $z_{\alpha/2} \in \mathbb{R}_+$ el valor crítico de la normal estándar asociado a la probabilidad $\alpha/2$, que cumple $\Phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2}$. Entonces

$$\mathbf{P}(|\bar{X}_n - \mu| > \varepsilon) = \alpha \Leftrightarrow \varepsilon = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

Notar que

$$|\bar{X}_n - \mu| \leq \varepsilon \Leftrightarrow \mu \in [\bar{X}_n - \varepsilon, \bar{X}_n + \varepsilon].$$

Si tomamos

$$\varepsilon = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

entonces

$$\mathbf{P}\left(\mu \in \left[\bar{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right]\right) = 1 - \alpha.$$

El intervalo

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

es un intervalo de confianza (exacto) al nivel de confianza $1 - \alpha$ para μ .

Es importante observar que los límites del intervalo

$$a(X_1, \dots, X_n) = \bar{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \text{ y } b(X_1, \dots, X_n) = \bar{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

son funciones de α y de la muestra X_1, \dots, X_n . Pero no dependen de ningún parámetro desconocido. En este caso no hay problema pues hemos asumido que σ^2 es conocido, pero si no lo fuera, la fórmula de arriba no sería un IdC pues no lo podríamos calcular a partir de los datos.

La interpretación es que si realizamos un número muy grande de veces el mismo experimento, de forma independiente, y procedemos a calcular el intervalo correspondiente a los datos obtenidos, entonces aproximadamente el $(1 - \alpha)100\%$ de las veces el parámetro desconocido μ estará en dicho intervalo. Pero a no olvidarse, que también el $\alpha 100\%$ el parámetro μ no estará en intervalo.

Remarquemos algunas observaciones que se deducen de la fórmula que dedujimos recién:

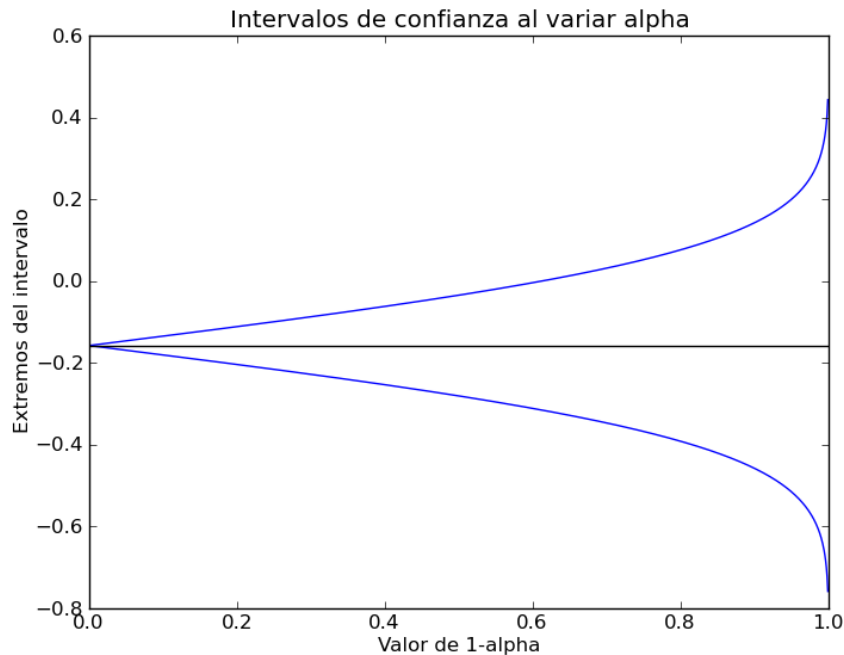
- Notar que el ancho del intervalo es $2z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$.
- Cuanto más grande el número de datos n , más chico es el intervalo. Es decir, mejor es la estimación desde el punto de vista de la precisión.
- Cuanto más grande la varianza σ^2 también más grande es el intervalo, por lo que más difícil es estimar μ .
- Y al variar $\alpha \in (0, 1)$:

$$\lim_{\alpha \rightarrow 0} 2z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = +\infty \text{ y } \lim_{\alpha \rightarrow 1} 2z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 0.$$

Para ilustrar la dependencia del intervalo con α , consideremos la siguiente tabla que contiene datos simulados por computadora de una distribución $N(0, 1)$.

-0.612	-2.819	-0.336	-0.371	-0.811	0.033	0.658	-1.426	0.184	0.609
0.291	-0.951	-1.131	0.332	-0.449	0.693	0.568	0.309	0.589	-0.632
-0.100	0.484	-1.044	0.196	-0.057	1.084	1.038	0.825	0.180	-2.061

Un intervalo de confianza al nivel 0.05 es $[-0.515, 0.200]$. Esto también se puede escribir como -0.1575 ± 0.3575 . La siguiente figura muestra cómo varían los extremos del intervalo al variar α en $(0, 1)$.



Deducción como valores creíbles

El enfoque que seguimos para obtener el IdC para μ fue bastante diferente a lo que hicimos con la proporción de bolitas verdes en la sección anterior. Pero podemos proceder de la misma manera.

La media μ puede en principio ser cualquier número real. Supongamos que tenemos el muestreo X_1, \dots, X_n , podemos entonces considerar el siguiente TdH:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_A : \mu \neq \mu_0 \end{cases}$$

La región de “aceptación” al nivel de significación, es como vimos en las clases anteriores

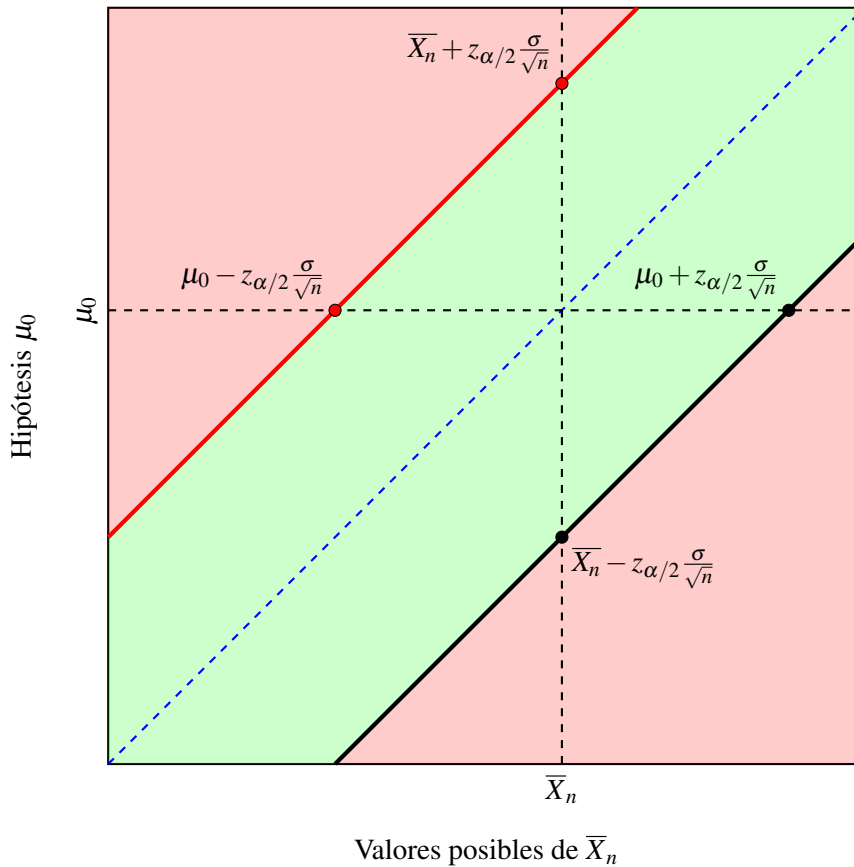
$$R_\alpha = \left[\mu_0 - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \mu_0 + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right].$$

Esto quiere decir que si $\bar{X}_n \in R_\alpha$, entonces no rechazamos H_0 y consideramos a μ_0 como un valor creíble.

Pero observar que

$$\bar{X}_n \in R \Leftrightarrow |\bar{X}_n - \mu_0| \leq z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \Leftrightarrow \mu_0 \in I_\alpha.$$

Es decir que los valores creíbles de μ_0 son exactamente aquellos que pertenecen al intervalo I_α . Dicho de otro modo, el intervalo de confianza I_α coincide con el intervalo de valores creíbles de μ .



La figura de arriba es la equivalente para este caso de la figura que hicimos para la estimación de bolitas verdes en la caja. La coordenada vertical corresponde a los posibles valores de la hipótesis nula μ_0 , la horizontal a los posibles valores de \bar{X}_n . En este caso ambas cantidades son continuas.

La línea azul representa el estimador de máxima verosimilitud para μ , que en este caso es el promedio \bar{X}_n . Por eso divide a la figura en dos partes iguales.

Esta figura hace evidente la dualidad existente entre los IdC y los TdH:

1. Para cada hipótesis nula μ_0 , si trazamos una línea horizontal por μ_0 , la intersección de ésta con la zona verde indica la región de aceptación R_α .
2. Para cada valor observado del promedio \bar{X}_n , si trazamos una línea vertical por \bar{X}_n , la intersección de ésta con la zona verde representa el intervalo de confianza I_α , constituido por los valores creíbles de μ_0 .

3. El z -intervalo (asintótico) para muestras grandes

Al igual que en los TdH, si el tamaño de la muestra es grande (en este curso grande significa $n \geq 30$) podemos usar el TCL para aproximar la distribución del promedio \bar{X}_n .

De hecho, como consecuencia del Teorema de Slutsky

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \Rightarrow N(0, 1).$$

en donde

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

es la varianza muestral.

Si procedemos de la misma forma que en la sección anterior, podemos probar que

$$I_\alpha = \left[\bar{X}_n \pm z_{\alpha/2} \frac{S_n}{\sqrt{n}} \right]$$

es un intervalo de confianza asintótico al nivel de confianza $1 - \alpha$.

Veamos un ejemplo de aplicación de este método.

Supongamos que en una encuesta electoral se toma una muestra de 2500 votantes, de los cuales 31 % afirman votar por el candidato A.

¿Cómo estimamos la cantidad de votantes de A en todo el país?

De hecho ya sabemos cómo hacerlo, pues es igual al caso de las bolitas verdes en la caja. Podemos modelar el experimento mediante una variable Bernoulli de parámetro $\theta \in (0, 1)$ que representa la proporción real de votantes de A en la población. Es decir,

$$X = \begin{cases} 1 & \text{si la persona vota a A} \\ 0 & \text{sino.} \end{cases}$$

Tenemos una muestra aleatoria de X con $n = 2500$ y $\bar{X}_n = 0.31$. Notar que el 31 % de 2500 es 775.

Fijemos un nivel de confianza de 95 %. Si aplicamos el método que desarrollamos en esa sección (para lo cual precisamos una computadora) obtenemos el intervalo de confianza

$$I_1 = [0.29190, 0.32854].$$

Este es un intervalo conservativo para θ .

Pero como la muestra es grande, podemos aproximar la distribución de \bar{X}_n por una normal.

Como $X_i^2 = X_i$ en este caso, tenemos que

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

De la tabla de la normal sabemos que el valor crítico $z_{0.05/2} = 1.96$. Entonces, el z -intervalo de confianza asintótico para θ es

$$I_2 = \left[\bar{X}_n \pm 1.96 \sqrt{\frac{\bar{X}_n (1 - \bar{X}_n)}{n-1}} \right] = [0.29187, 0.32813]$$

que es muy cercano al anterior.

4. El t -intervalo (exacto) para la media de datos normales

¿Qué pasa cuando la varianza σ^2 es desconocida? Podemos usar la distribución t de Student en ese caso.

Supongamos que X_1, \dots, X_n es un muestreo aleatorio de X , que tiene distribución normal $N(\mu, \sigma^2)$, en donde tanto μ como σ^2 son desconocidos.

Queremos construir un intervalo de confianza exacto para la media μ . Al igual que en el caso anterior, buscaremos un intervalo simétrico y centrado en \bar{X}_n .

Haremos el mismo cálculo que en el caso anterior, salvo que en lugar de dividir entre σ dividiremos entre S_n el desvío muestral. Calculando:

$$\begin{aligned} \mathbf{P}\left(\left|\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n}\right| > t\right) &= \mathbf{P}(|T| > t) \\ &= \mathbf{P}(T > t) + \mathbf{P}(T < -t) \\ &= 1 - F_{t,n-1}(t) + F_{t,n-1}(-t). \end{aligned}$$

Aquí T es una variable con distribución t de Student con $n - 1$ grados de libertad, y $F_{t,n-1}$ es su f.d.a..

Como la función de distribución $F_{t,n-1}$ es simétrica:

$$\begin{aligned} \mathbf{P}\left(|\bar{X}_n - \mu| > t \frac{S_n}{\sqrt{n}}\right) &= 1 - F_{t,n-1}(t) + F_{t,n-1}(-t) \\ &= 1 - F_{t,n-1}(t) + 1 - F_{t,n-1}(t) \\ &= 2[1 - F_{t,n-1}(t)]. \end{aligned}$$

Es decir que

$$\mathbf{P}\left(|\bar{X}_n - \mu| > t \frac{S_n}{\sqrt{n}}\right) = \alpha \Leftrightarrow F_{t,n-1}(t) = 1 - \frac{\alpha}{2}.$$

Sea $t_{n-1}(\alpha/2) \in \mathbb{R}_+$ el valor crítico de la distribución de Student asociado a la probabilidad $\alpha/2$, que cumple $F_{t,n-1}(t_{n-1}(\alpha/2)) = 1 - \frac{\alpha}{2}$. Entonces

$$\mathbf{P}\left(|\bar{X}_n - \mu| > t \frac{S_n}{\sqrt{n}}\right) = \alpha \Leftrightarrow t = t_{n-1}(\alpha/2).$$

Notar que

$$|\bar{X}_n - \mu| \leq t \frac{S_n}{\sqrt{n}} \Leftrightarrow \mu \in \left[\bar{X}_n - t \frac{S_n}{\sqrt{n}}, \bar{X}_n + t \frac{S_n}{\sqrt{n}}\right].$$

Si tomamos $t = t_{n-1}(\alpha/2)$, concluimos que

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

El intervalo

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

es un intervalo de confianza (exacto) al nivel de confianza $1 - \alpha$ para μ .

Notar que el valor crítico $t_{n-1}(\alpha/2)$ es más grande que el correspondiente $z_{\alpha/2}$ del caso en que σ^2 es conocido. Esto es razonable: cuando no conocemos la varianza resulta un poco más difícil estimar μ .