Probabilidad y Estadística

Clase 8

Gonzalo Blanco

2023

Estimación puntual

1. Introducción

En una fábrica se producen articulos, nos interesa la producción de la misma durante un día de trabajo completo. En particular de todos los artículos producidos tenemos especial interés en aquellos que están defectuosos.

En este caso, la **población** estaría conformada por todos los articulos fabricados en un día. Y llamaremos **p** a la proporción de artículos defectuosos de ese día de producción.

Si tomamos una **muestra** de 25 artículos, entonces puedo definir la variable aleatoria *X="n° de artículos defectuosos de la muestra"* y podemos asumir que esta v.a X sigue una distribución binominal *B(25,p)*.

En **probabilidades** se conocen todos los datos sobre X, es decir, conocemos su función de distribución de probabilidad y los parámetros de la misma; entonces podriamos calcular por ejemplo la probabilidad de que 5 de los 25 productos esten defectuosos, $\mathbf{P(X=5)}$.

En **estadística** desconocemos total o parcialmente las características de X y a partir de la **muestra** vamos a tratar de **inferir** información sobre la distribución de **X**, o lo que es lo mismo, trataremos de *obtener información de la población.*

Sigamos con el ejemplo de la fábrica. Supongamos que no conocemos **p**, entonces la probabilidad de que en una muestra M_1 un artículo sea defectuoso la llamo $\hat{p}$ y la podemos calcular como $\hat{p} = \frac{\#art.defectuosos}{\#muestra}$. Ahora bien, si tomo otra muestra M_2 con 25 artículos y calculo $\hat{p}$, seguramente el valor de $\hat{p}$ encontrado a partir de la muestra M_2 sea distinto al estimado a partir de la muestra M_1, entonces para cada muestra M_i que yo tome, el valor de $\hat{p}$ seguramente difiera, enonces podemos decir que $\hat{p}$ es una variable aleatoria, y se la conoce como **estadístico.**

Un **estadístico** es entonces una función que depende de la muestra aleatoria. En este caso $\hat{p} = h(x_1, x_2, ..., x_{25})$.

2. Muestreo Aleatorio

Para que las inferencias sean válidas, la muestra debe ser representativa de la población. A este tipo de muestras se las conoce como muestras aleatorias.

Diremos que una muestra es aleatoria cuando la misma sea seleccionada a partir de un mecanismo aleatorio. Un ejemplo a modo ilustrativo de selección de una muestra aleatoria sería "meter todos mis datos en una bolsa, mezclar, y sacar n datos al azar". En consecuencia, la selección de una muestra aleatoria es un experimento aleatorio, por que cada muestra que saque va a ser distina, y cada observación/dato x_i de la muestra es el valor observado de una variable aleatoria.

Los datos de toda mi **población** determinan la distribución de probabilidad que seguirán todas esas variables aleatorias.

Entonces para definir una muestra aleatoria, sea X la v.a que representa el resultado de tomar un dato al azar de la población y sea f(x) la fdp de X, entonces nuestra muestra aleatoria estará conformada por n observaciones/datos que son independientes, obtenidos bajo las mismas condiciones y que ademas siguen la misma distribución de probabilidad f(x). **Cuando todo esto se cumpla diremos que nuestra muestra es una muestra aleatoria y la podemos considerar representativa de la población, es decir que va a ser útil para inferir carácteristicas de la población**.

2.1. Estadísticos usuales

Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de una v.a X donde $E[X] = \mu$ y $Var(X) = \sigma^2$.

Si desconocemos μ y σ^2 de X, es decir, de la población y queremos estimarlos a partir de la muestra, entonces los estadísticos usuales que se uilizan para esto son

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}$$

Siendo $\bar{x}$ un estimador para la media conocido como **media muestral** y S^2 un estimador para la varianza conocido como **varianza muestral**. Por supuesto el **desvío estándar muestral** lo calculamos como $S = \sqrt{S^2}$.

Uno de los objetivos principales de los estadísticos es estimar parámetros desconocidos de la población, cuando lo estadísticos se usan para esto se los llaman **estimadores puntuales**. Por ejemplo si X sigue una $N(\mu, \sigma^2)$ o una $P(\lambda)$, podemos estimar un valor para μ, σ^2 o para λ a partir de una muestra aleatoria.

Si el parámetro que queremos determinar es, por ejemplo, λ entonces es habitual denotar a su estimador como $\hat{\lambda}$ y el mismo será de la forma $\hat{\lambda} = h(x_1, x_2, ..., x_n)$.

2.2. Criterios para evaluar estimadores puntuales

¿Cómo puedo saber si un estimador aproxima bien al parámetro de interés? ¿Cómo puedo determinar si es un buen estimador puntual?.

Para determinar si un estimador es bueno o de entre varios estimadores puntuales cual es el mejor, vamos a analizar algunas características que estos deben cumplir. Supongamos que tengo un parámetro θ y un estimador del mismo $\hat{\theta}$, diremos que el mismo es un buen estimador si cumple las siguientes propiedades:

1. **Sesgo**: diremos que el estimador es insesgado cuando su valor esperado sea el parámetro de interés, es decir cuando su sesgo "b"sea nulo.

$$b = E[\hat{\theta}] - \theta$$

2. **Eficiente**: la eficiencia de un estimador está asociado a la varianza del mismo, mientras menor varianza tenga diremos que es más eficiente. Entonces si tenemos varios estimadores puntuales para el mismo parámetro nos quedaremos con el más eficiente, es decir aquel que tenga menor varianza.

3. **Consistencia:** Diremos que un estimador es consistente cuando cumpla que:

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| > \epsilon) = 0$$

   Es decir, cuando la muestra es muy grande el valor del estimador puntual toma valor del parámetro de interés.

   Para comprobar la consistencia de un estimador puntual nos basta con chequiar lo siguiente:

   a) $\lim_{n\to\infty} E[\hat{\theta}] = \theta$.

   b) $\lim_{n\to\infty} Var(\hat{\theta}) = 0$

   Si el estimador cumple a) y b) entonces podemos decir que es consistente.

Ejemplo: Queremos encontrar la altura media μ de los estudiantes de una facultad pero no estamos seguros de como calcularla, tenemos los siguientes estimadores:

a) La media muestral $\hat{x_1} = \frac{1}{n}\sum_{i=1}^{n} x_i$

b) Primer y ultimo dato $\hat{x_2} = \frac{x_1+x_n}{2}$

c) tomo un dato al azar $\hat{x_3} = x_i$

Recordemos que cada estudiante tomado al azar x_i es una variable aleatoria cuya media y varianza son las de la población μ y σ^2.

Veamos cual es el mejor estimador.

Primero analicemos el **sesgo** de acada uno:

$E[\hat{x_1}] = E[\frac{1}{n}\sum_{i=1}^{n} x_i] = \frac{1}{n}E[\sum_{i=1}^{n} x_i] = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{n\mu}{n} = \mu$

$b_1 = E[\hat{x}_1] - \mu = 0$, entonces es insesgado.

$E[\hat{x}_2] = E[\frac{x_1+x_n}{2}] = \frac{1}{2}(E[x_1] + E[x_2]) = \frac{1}{2}(\mu + \mu) = \frac{1}{2}2\mu = \mu$

Por lo tanto x_2 también es insesgado.

$E[\hat{x_3}] = E[x_i] = \mu$, también es insesgado.

Los tres estimadores son insesgados, veamos ahora cual es el más **eficiente**, es decir, cual de ellos presenta menor varianza.

$Var[\hat{x_1}] = Var[\frac{1}{n}\sum_{i=1}^{n} x_i] = \frac{1}{n^2}Var[\sum_{i=1}^{n} x_i] = \frac{1}{n^2}\sum_{i=1}^{n} Var[x_i] = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$

Recordar que $Var[\sum_{i=1}^{n} x_i] = \sum_{i=1}^{n} Var[x_i]$ por que los x_i son independientes, entonces los términos que incluyen a la coravrianza son nulos.

$Var[\hat{x}_2] = Var[\frac{x_1+x_n}{2}] = \frac{1}{2^2}Var(x_1 + x_2) = \frac{1}{2^2}(\sigma^2 + \sigma^2) = \frac{1}{2^2}2\sigma^2 = \frac{\sigma^2}{2}$

$Var(\hat{x}_3) = Var(x_i) = \sigma^2$

De los 3 estimadores, el más eficientes es $\hat{x}_1$ ya que es el de menor varianza (suponiend que n>2).

Veamos ahora si este estimador es **consistente**, entonces probemos que

a) $\lim_{n\to\infty} E[\hat{x}_1] = \mu$.

Vimos que era insesgado por lo tanto cuando n tienda a infinito el valor esperado de este estimador también tenderá a μ.

b) $\lim_{n\to\infty} Var(\hat{x}_1) = 0$

Vimos que $Var(\hat{x}_1) = \frac{\sigma^2}{n}$, es decir que cuando n tienda a infinito la varianza tenderá a 0.

Como cumple a) y b) entonces es consistente.

Intervalos de confianza

3. Introducción

Las clases anteriores vimos como construir a partir de una muestra aleatoria un estimador puntual de un parámetro desconocido de interés. En ese caso, siempre necesitábamos dar alguna característica del estimador, como por ejemplo su varianza o decir si tiene sesgo. A veces es más simple y útil dar un intervalo de valores posibles del parámetro desconocido, de manera tal que este intervalo contenga al parámetro con una determinada probabilidad.

Que quiero decir con esto, supongamos que tengo una muestra aleatoria y me interesa estudiar un parámetro μ, entonces a partir de la muestra aleatoria se va a construir un intervalo $[\hat{\mu_1}, \hat{\mu_2}]$, donde los extremos de este intervalo son dos estadísticos $\hat{\mu_1}$ y $\hat{\mu_2}$ tal que $P(\hat{\mu_1} < \mu < \hat{\mu_2}) = 1-\alpha$, siendo α un valor entre 0 y 1. Al valor $1-\alpha$ se lo conoce como **nivel de confianza** y representa la certeza que tenemos de que nuestro parámetro de interés caiga en ese intervalo. Entonces $P(\hat{\mu_1} < \mu < \hat{\mu_2}) = 1-\alpha$ la podemos traducir como *la probabilidad de que mi parámetro de interés caiga dentro del intervalo* $[\hat{\mu_1}, \hat{\mu_2}]$ *es de* $1-\alpha$.

Un valor usual para este intervalo de confianza es $1-\alpha = 0{,}95$, es decir busco aquel intervalo tal que la probabilidad de que mi parámetro caiga en él es del 95 %, o lo que es lo mismo, tengo un 95 % de confianza en ese intervalo.

Dado que los extremos del intervalo $[\hat{\mu_1}, \hat{\mu_2}]$ son estadísticos, es decir que los calculamos a partir de los datos de mi muestra aleatoria, entonces este intervalo varía con cada muestra aleatoria que tome (ya que cada muestra aleatoria no es idéntica a la anterior o a las sucesivas). Entonces la probabilidad $P(\hat{\mu_1} < \mu < \hat{\mu_2}) = 0,95$ también se puede traducir como "si tomo 100 muestras distintas, tendremos 100 intervalos distintos $[\hat{\mu_1}, \hat{\mu_2}]$ para mi parámetro μ de los cuales aproximadamente 5 de ellos no contendrán a mi parámetro de interés".

4. Intervalos de confianza para la media de una población

4.1. Población con distribución normal y varianza conocida

Analicemos el caso más simple de todos.

Sea $(x_1, x_2, ..., x_n)$ una muestra aleatoria de tamaño n de una v.a. X (o de una población) donde X $\backsim N(\mu, \sigma^2)$, con σ^2 conocido y μ desconocido. Se quiere construir un intervalo de confianza para μ de nivel $1-\alpha$.

Supongamos $\alpha = 0{,}05$. Veamos cómo hacemos esto:

1. Tomamos un estimador puntual de μ, sabemos que un buen estimador de la media es la media muestral, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

2. A partir de $\bar{x}$ construimos el estadístico $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$, al cual llamaremos **pivote**.

   Notar que Z contiene al parámetro de interés μ y bajo las condiciones dadas $Z \backsim N(0,1)$. Además este pivote solo contiene un único parámetro desconocido, en este caso μ. Esta es una condición importante que debe cumplir el estadístico para que podamos utilizarlo como pivote y determinar el intervalo de confianza de μ.

3. Como conocemos la distribución de Z entonces podemos buscar un intervalo tal que $P(-z < Z < z) = 0{,}95$. Siendo $1-\alpha = 0.95$.

A partir de la tabla de la distribución normal estándar obtengo $z = 1,96$ que cumple esto:

$$P(-1,96 < Z < 1,96) = 0{,}95$$

$$P(-1,96 < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < 1,96) = 0,95$$

$$P(\bar{x} - 1,96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1,96\frac{\sigma}{\sqrt{n}}) = 0,95$$

Es decir, el intervalo $[\bar{x} - 1,96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1,96\frac{\sigma}{\sqrt{n}}]$ tiene un 95 % de contener al parámetro μ. Esto lo puedo hacer para cualquier nivel de confianza, es decir para cualquier $1-\alpha$.

En resumen, defino el nivel de confianza con el que quiero estimar el intervalo para μ y hago los pasos 1, 2 y 3.

Tener en cuenta que esto lo hacemos suponiendo que X sigue una distribución gaussiana con varianza conocida.

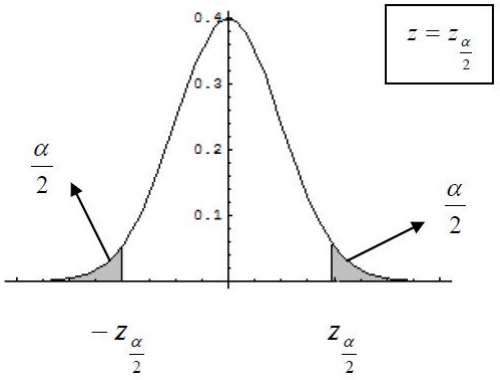

Figura 1: Distribución Normal estándar.

Al valor de z que verifica $P(-z < Z < z) = 1 - \alpha$ lo denotamos como $z_{\frac{\alpha}{2}}$, ya que al ser la gaussiana una función simétrica, vale que

$$P(-z < Z < z) = P(Z < z) - P(Z < -z) = 1 - 2P(Z < -z) = 1 - 2P(Z < -z_{\frac{\alpha}{2}})$$

Es decir, $P(Z < -z_{\frac{\alpha}{2}}) = P(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$.
En el ejemplo anterior $z_{\frac{\alpha}{2}} = 1{,}96$.

Veamos algunas cosas importantes cuando estemos trabajando con intervalos de confianza :

1. El largo del intervalo L viene dado por la diferencia entre sus extremos, la diferencia entre $\bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ y $\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$. Por lo tanto $L = 2z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$.
2. Si aumentamos el nivel de confianza del intervalo, es decir si agrandamos el $1 - \alpha$, o lo que es lo mismo achicamos α, el valor de $z_{\frac{\alpha}{2}}$ aumenta, por lo tanto el largo del intervalo $[\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$ crece.

Notar que si α y σ están fijos, si el n aumenta $\Longrightarrow$ el largo del intervalo L disminuye, es decir mientras más datos use, más preciso se hace mi estimador ya que el parámetro puede tomar una cantidad más acotada de valores.

Entonces podríamos calcular el n necesario para que mi intervalo no supere cierto largo.

En general si queremos hallar n tal que $L = 2z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < l$, donde l es un largo dado, entonces el tamaño mínimo de la muestra para que esto se cumple es de

$$n > \left(\frac{2z_{\frac{\alpha}{2}} \sigma}{l} \right)^2$$

Si estimamos puntualmente a la media μ con el promedio simple $\bar{x}$ estamos cometiendo un error en la estimación menor o igual a $\frac{L}{2} = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$, y se lo conoce como **precisión del estimador**.

Ejemplo:

Dada la producción de sogas en una fábrica se analiza la carga que puede soportar un tipo en particular. Se sabe que la carga que soportan las mismas se distribuye normalmente con una varianza de 400 kg^2. Al tomar al azar 12 sogas se determinó que el peso promedio que soportan es 1000 kg.

a) Construya un intervalo de confianza del 95 % para la resistencia a la carga promedio que soporta la soga.

b) Construya un intervalo de confianza del 99 % para la resistencia a la carga promedio que soporta la soga. Compare el ancho de este intervalo de confianza con el ancho encontrado en el inciso a).

c) ¿Qué tamaño n de muestra se necesita para que el intervalo tenga un nivel de confianza del 95 % y la longitud sea la mitad del intervalo hallado en a)?

a) Queremos el intervalo de confianza para de nivel 95 %. Por lo tanto $1 - \alpha = 0{,}05 \longrightarrow \frac{\alpha}{2} = 0{,}025$. Busco en la tabla de N(0,1) el $z_{\frac{\alpha}{2}} = z_{0{,}025}$, es decir aquel z tal que $P(Z > z) = 0,025$.

De la tabla obtengo $z_{0{,}025} = 1,96$.

Entonces mi intervalo con un 95 % de confianza para le media vendrá dado por

$$[1000 - 1,96\frac{20}{\sqrt{12}}, 1000 + 1,96\frac{20}{\sqrt{12}}] = [988kg, 1011kg]$$

El largo de mi intervalo viene dado por $L_1 = 1011 - 988 = 23kg$

b) Ahora buscamos el intervalo con un nivel de confianza del 99 %. Por lo tanto $1 - \alpha = 0,01 \longrightarrow \frac{\alpha}{2} = 0{,}005$.

Busco en la tabla de N(0,1) el $z_{\frac{\alpha}{2}} = z_{0{,}005}$ y obtengo $z_{0{,}005} = 2{,}575$.

Entonces mi intervalo vendrá dado por

$$[1000 - 2{,}575\frac{20}{\sqrt{12}}, 1000 + 2{,}575\frac{20}{\sqrt{12}}] = [985kg, 1015kg]$$

El largo de mi intervalo es $L_2 = 1015 - 985 = 30kg$

Como era esperable el largo del segundo intervalo es mayor que el del primero ya que el aumento del nivel de confianza se ve reflejado en un crecimiento en el largo del intervalo. O lo que es lo mismo, al aumentar el nivel de confianza, crece el intervalo y perdemos precisión en la estimación (el parámetro de interés puede tomar una cantidad mayor de valores posibles).

c) Si ahora queremos disminuir a la mitad el intervalo de confianza de del inciso a) manteniendo el nivel en 95 %, entonces busco el n tal que

$$n > \left(\frac{2z_{\frac{\alpha}{2}}\sigma}{l}\right)^2 = \left(\frac{2 * 1{,}96\sigma}{11{,}5}\right)^2 = 48$$

Entonces si tomara el promedio de 48 sogas en vez de 12, obtendríamos un intervalo de confianza del 95 % para μ de la mitad del largo que para 12 sogas, es decir la estimación es mucho más precisa.

Para muestras tomadas de una población normal , o para muestras de tamaño $n > 30$ de una población cualquiera, siempre con varianza conocida, el intervalo de confianza dado anteriormente proporciona buenos resultados. En el caso de que la población de la que se extrae la muestra no sea normal, pero $n > 30$, el nivel de confianza del intervalo dado previamente es aproximadamente $1 - \alpha$, es decir podemos aproximar a la distribución del pivote con una normal estándar. Pero para muestras pequeñas tomadas de poblaciones que no son normales no se puede garantizar que el nivel de confianza sea realmente de $1 - \alpha$ si se lo utiliza.

4.2. Población con distribución normal y varianza desconocida

Nuevamente como queremos encontrar un intervalo de confianza para μ nos basamos en la media muestral $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ que sabemos que es un buen estimador de μ. En el caso anterior, donde la varianza era conocida, habíamos partido del estadístico $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$, pero ahora desconocemos σ^2, entonces Z ya no nos sirve como pivote. Recordemos que era necesario que Z contenga solo al parámetro de interés como única incógnita para poder determinar su intervalo de confianza. Entonces vamos a aproximar a σ^2 con el estimador de la varianza muestral:

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

Sabemos que S^2 es un buen estimador de σ^2, entonces defino el pivote

$$T = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

Para poder usar a T como pivote, debemos conocer su distribución densidad de probabilidad. En el caso anterior como conocíamos σ sabíamos que Z seguía una distribución normal estándar, pero ahora no es tan simple ya que S no es un parámetro fijo conocido, sino que es una variable aleatoria que depende de la muestra aleatoria a partir de la cual la calcule. Se puede probar que T sigue una distribución llamada **distribución de Student con n-1 grados de libertad**.

Nota: Decimos que una variable aleatoria x sigue una distribución de Student con k grados de libertad si su f.d.p es de la forma:

$$f(x) = \frac{\Gamma\left[\frac{(k+1)}{2}\right]}{\sqrt{\pi k}\,\Gamma\left(\frac{k}{2}\right)} \frac{1}{\left[\left(\frac{x^2}{k}\right)+1\right]^{\frac{k+1}{2}}} \qquad -\infty < x < \infty$$

Figura 2: Distribución de Student con k grados de libertad.

Notar que esta f(x) depende de una función $\Gamma(k)$, esta función es conocida como función gamma y también es una función de densidad de probabilidad que se aplica en ciertos casos. A nosotros no nos va a importar el valor de esta función ya que la distribución de Student está tabulada, entonces seguiremos una proceso similar al aplicado anteriormente para la N(0,1).

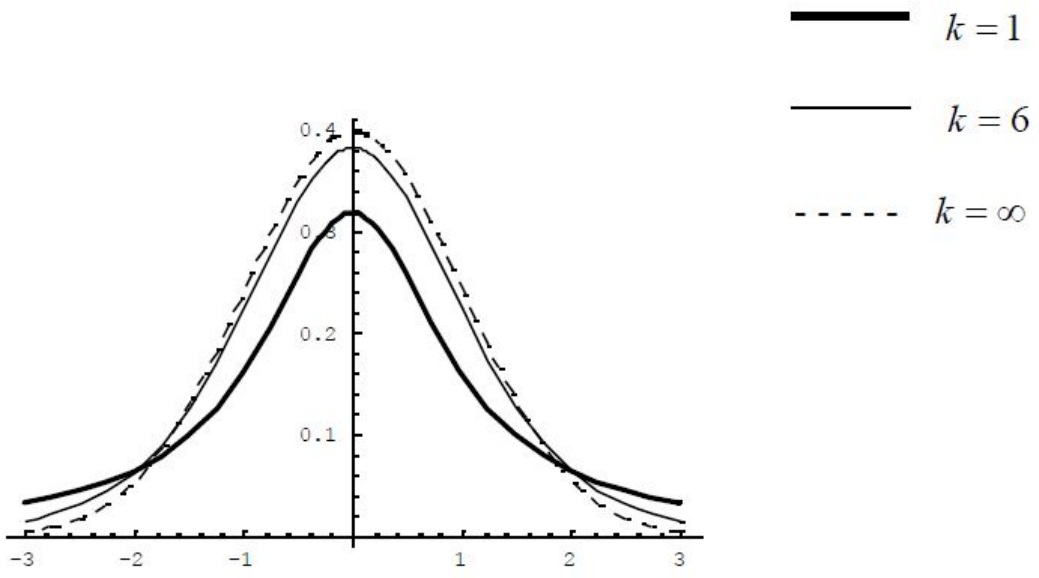

Figura 3: Distribución de Student con k grados de libertad.

La distribución de Student (que se muestra en la figura 3) Es una función simétrica con forma de campana muy similar a la Gaussiana. Cuando k tiende a infinito la distribución de Student tiende a la N(0,1).

Notación: $T \backsim t_k$ significa que la v.a T sigue una dist. de Student con k grados de libertad.

Ya tenemos definido el pivote T y conocemos su fdp, entonces podemos buscar el intervalo de confianza de μ con un nivel de confianza de $1-\alpha$ como hicimos antes.

Comenzamos planteando la ecuación:

$$P(-t < T < t) = 1 - \alpha$$

Entonces buscamos el valor de t que cumple esta probabilidad.

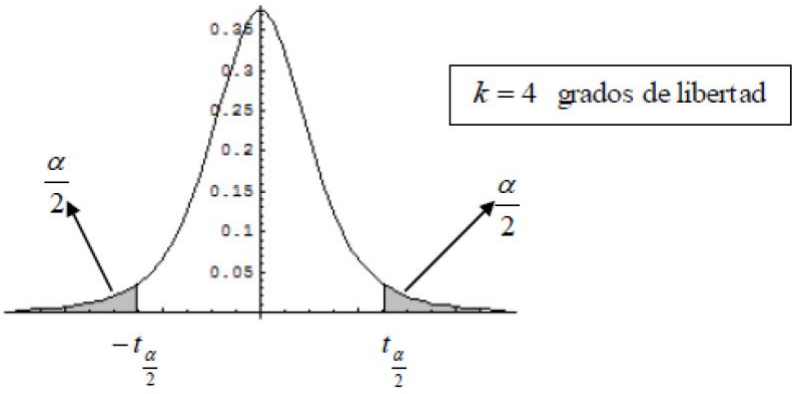

Figura 4: Distribución de Student con 4 grados de libertad.

Al ser simétrica tiene las mismas propiedades que la gaussiana, entonces igual que en caso anterior busco t al que $P(T > t) = \frac{\alpha}{2}$, y a este t lo llamo $t_{\frac{\alpha}{2},n-1}$. El α referido al nivel de confianza y el n-1 a los grados de libertad. En nuestros casos de aplicación n viene dado por la cantidad de datos de la muestra aleatoria.

Si reemplazo T para poder despejar μ queda:

$$P(-t < \frac{\bar{x} - \mu}{S/\sqrt{n}} < t) = 1 - \alpha$$

$$P(\bar{x} - t\frac{S}{\sqrt{n}} < \mu < \bar{x} + t\frac{S}{\sqrt{n}}) = 1 - \alpha$$

Cosas a tener en cuenta sobre la tabla de la distribución de Student, veamos la figura de abajo.

Tabla IV. Valores críticos de la distribución t de Student:
Abcisas $t_{\alpha;\nu}$ que dejan a su derecha un área α en una t con ν grados de libertad.

ν	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.326	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.213	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869

Figura 5: Tabla de la distribución de Student.

En la primera columna v hace referencia a los grados de libertad. Y la primera fila corresponde a los distintos valores de α.

Las tablas que veníamos usando hasta ahora eran tablas de funciones de densidad de probabilidad ACUMULADA, es decir nos daban valores para $P(Z < z)$. Ahora bien, la tabla de la distribución de Student nos da valores de probabilidad para $P(T > t)$, veamos los recuadros naranjas. Para una variable aleatoria T con 4 (v) grados de libertad el t tal que $P(T > t) = 0{,}025 = \frac{\alpha}{2}$ es $t = 2{,}776$.

Entonces si $(x_1, x_2, ..., x_n)$ es una muestra aleatoria de tamaño n de una variable aleatoria X que sigue una $N(\mu, \sigma^2)$ con σ^2 desconocida, un intervalo para μ con nivel de confianza $1 - \alpha$ viene dado por

$$[\bar{x} - t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}]$$

Si la muestra es grande entonces podemos probar que el estadístico T sigue aproximadamente una distribución N(0,1) y puedo construir el intervalo de μ como

$$[\bar{x} - z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}]$$

Ejemplo:

Se hicieron 10 mediciones sobre la resistencia de cierto tipo de alambre que dieron valores tal que $\bar{x} = 10{,}48ohm$ y $S = 1,36ohm$. Supongamos que la resitencia del alambre sigue una $N(\mu, \sigma^2)$. Obtener un intervalo de confianza del 90 % para μ. Defino X: "*resistencia del alambre*".

Tenemos que $1 - \alpha = 0{,}9$, entonces $\frac{\alpha}{2} = 0,05$.

Se que X sigue una $N(\mu, \sigma^2)$, pero le desconozco tanto la media como la varianza, entonces para calcular el intervalo de confianza para μ voy a utilizar el pivote

$$T = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

Sabemos que T sigue una distribución de Student, y para este caso con 9 grados de libertad (n-1).

Entonces en la tabla de Studen busco el $t_{0{,}05;9}$, es decir el t tal que $P(T > t) = 0{,}05$ con 9 grados de libertad.

De la tabla se obtiene que $t_{0{,}05;9} = 1{,}833$

Entonces el intervalo de confianza vendrá dado por

$$[10{,}48 - 1{,}833\frac{1{,}36}{\sqrt{10}}, 10{,}48 + 1{,}833\frac{1{,}36}{\sqrt{10}}] = [9,69; 11{,}27]$$

5. Intervalo de confianza para la Varianza de una población

5.1. Población con distribución normal

Supongamos que se quiere hallar un intervalo de confianza para la varianza σ^2 de una v.a. que sigue una distribución normal $N(\mu, \sigma^2)$. Sea $(y_1, y_2, ..., y_n)$ una muestra aleatoria de esa población. Podemos tomar como estimador puntual de σ^2 a la varianza muestral S^2

$$S^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

Luego a partir de este estimador podemos definir el estadístico

$$X = \frac{(n-1)S^2}{\sigma^2}$$

Este estadístico X contiene como única incognita a nuestro parámetro de interés σ^2 y se puede probar que sigue una **distribución chi-cuadrado con n-1 grados de libertad**.

Nota: sea X una v.a continua, se dice que X sigue una distribución chi-cuadrado con k grados de libertad, si su fdp es de la forma:

$$f(x) = \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} x^{(k/2)-1} e^{-x/2} \qquad x > 0$$

Figura 6: Distribución chi-cuadrado con k grados de libertad.

Esta distribución es asimétrica. A continuación, se la grafica para distintos grados de libertad.

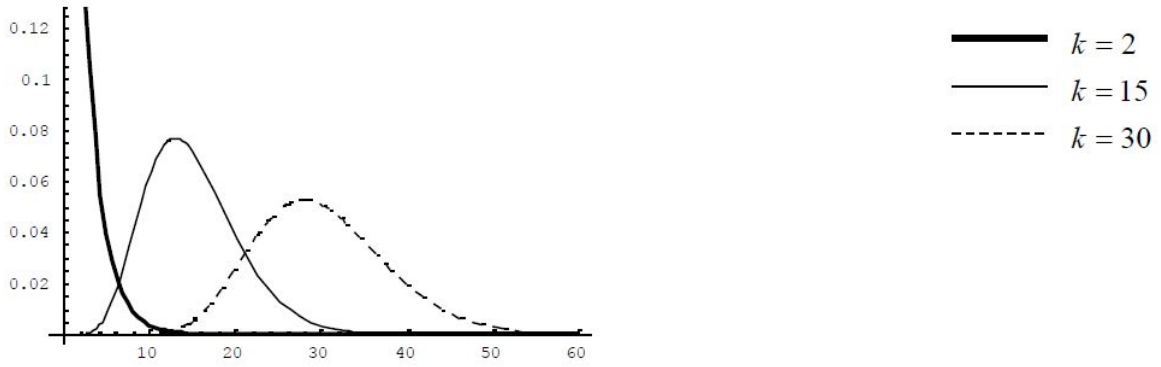

Figura 7: Distribución chi-cuadrado con k grados de libertad.

Notación: $Y \backsim X_k^2$ entonces Y sigue una distribución chi-cuadrado con k grados de libertad.

Denotaremos $X_{\alpha,k}^2$ al valor de Y, tal que $P(Y > X_{\alpha,k}^2) = \alpha$. Es lo mismo que planteamos arriba con la distribución de Student, con la diferencia que la distribución chi-cuadrado no es simétrica.

Propiedades

1. Vale que si $Y_1, Y_2, ..., Y_n$ son variables aleatorias independientes con fdp N(0,1) entonces la v.a $Z = Y_1 + Y_2 + ... + Y_n$ sigue una distribución chi-cuadrado con n grados de libertad.

2. Si $Y_1, Y_2, ..., Y_n$ son variables aleatorias independientes tal que cada Y_i sigue una distribución chi-cuadrado con k_i grados de libertad, entonces la v.a $Z = Y_1 + Y_2 + ... + Y_n$ sigue una distribución chi-cuadrado con $k = k_1 + k_2 + ... + k_n$ grados de libertad.

Retomemos, queremos entonces encontrar un intervalo de confianza para la varianza σ^2 utilizando el estadístico $X = \frac{(n-1)S^2}{\sigma^2}$ que sigue una distribución chi-cuadrado con n-1 grados de libertad.

Notar que al ser esta distribución asimétrica, entonces puedo armar distintos intervalos de confianza de valor $1 - \alpha$.

Para desarrollar el intervalo de confianza, planteamos hallar dos números a y b tal que:

$$P(a < X < b) = 1 - \alpha$$

$$P(a < \frac{(n-1)S^2}{\sigma^2} < b) = 1 - \alpha$$

Se puede probar que la mejor elección de a y b es:

$$a = X_{1-\frac{\alpha}{2},n-1}^2 \qquad b = X_{\frac{\alpha}{2}^2,n-1}$$

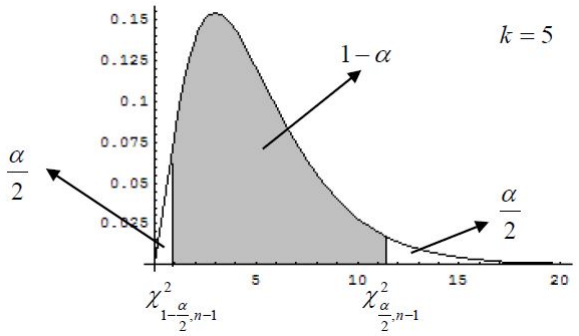

Figura 8: Distribución chi-cuadrado con 5 grados de libertad. En gris el área que representa una probabilidad de $1 - \alpha$.

Por lo tanto,

$$P(a < \frac{(n-1)S^2}{\sigma^2} < b) = 1 - \alpha$$

$$P(X^2_{1-\frac{\alpha}{2},n-1} < \frac{(n-1)S^2}{\sigma^2} < X_{\frac{\alpha}{2}^2,n-1}) = 1 - \alpha$$

Despejando σ^2 se llega a

$$P\left(\frac{(n-s)S^2}{X^2_{\frac{\alpha}{2},n-1}} < \sigma^2 < \frac{(n-1)S^2}{X^2_{1-\frac{\alpha}{2},n-1}}\right) = 1 - \alpha$$

Entonces si $y_1, y_2, ..., y_n$ son una muestra aleatoria de una variable aleatoria Y, con $Y \backsim N(\mu, \sigma^2)$. Un intervalo con un nivel de confianza de $1 - \alpha$ para σ^2 viene dado por

$$\left[\frac{(n-1)S^2}{X^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)S^2}{X_{1-\frac{\alpha}{2}^2,n-1}}\right]$$

Un intervalo de confianza para el desvío estandar σ vendrá dado por

$$\left[\sqrt{\frac{(n-1)S^2}{X^2_{\frac{\alpha}{2},n-1}}}, \sqrt{\frac{(n-1)S^2}{X_{1-\frac{\alpha}{2}^2,n-1}}}\right]$$

Ejemplo:

Un fabricante de detergente líquido está interesado en la uniformidad de la máquina utilizada para llenar las botellas. De manera específica, es deseable que la desviación estándar σ del proceso de llenado sea menor que 0.15 onzas de líquido; de otro modo, existe un porcentaje mayor del deseable de botellas con un contenido menor de detergente. Supongamos que la distribución del volumen de llenado es aproximadamente normal. Al tomar una muestra aleatoria de 20 botellas, se obtiene una varianza muestral $S^2 = 0{,}0153\, onzas^2$. Hallar un intervalo de confianza de nivel 0.95 para determinar si efectivamente existen botellas con un contenido menor de detergente del deseado.

Sea X= volumen de llenado de las botellas.

$X \backsim N(\mu.\sigma^2)$ con ambos parámetros desconocidos.

Defino el pivote $X = \frac{(n-1)S^2}{\sigma^2}$ y se que sigue una distribución chi-cuadrado con 19 grados de libertad.

Tenemos que $1 - \alpha = 0{,}95 \Longrightarrow \alpha = 0,05$

$X^2_{\frac{\alpha}{2},n-1} = X^2_{0{,}025,19} = 32{,}85$

$X^2_{1-\frac{\alpha}{2},n-1} = X^2_{0{,}975,19} = 8{,}91$

Ambos valores obtenidos de la tabla de distribución chi-cuadrada para 19 grados de libertad.

Entonces el intervalo de confianza para σ^2 viene dado por:

$$\left[\frac{(n-1)S^2}{X^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)S^2}{X_{1-\frac{\alpha}{2}^2,n-1}}\right] = \left[\frac{19 * 0{,}0153}{32{,}85}, \frac{19 * 0{,}0153}{8{,}91}\right] = [0{,}00884; 0{,}0326]$$

Y el intervalo de confianza para el desvío estandar σ viene dado por $[\sqrt{0{,}00884}; \sqrt{0{,}0326}] = [0{,}09; 0{,}18]$

Según el intervalo de confianza obtenido para σ, este puede tomar valores de hasta 0.18, es decir que pueden exister botellas con un contenido menor que el deseado.

6. Intervalo de confianza para una proporción

Sea una población de tamaño N (eventualmente puede ser infinito) de cuyos individuos nos interesa cierta propiedad A (ej: de toda la producción de una fábrica los productos defectuosos). Supongamos que la probabilidad de que un individuo de la población verifique A es P(A) = p. El significado del parámetro p es, en consecuencia, el de proporción de individuos de la población que verifican la propiedad A. Podemos definir una variable aleatoria X_i que mide, para cada individuo de la población, si cumple o no la propiedad A.

La variable X_i aleatoria tendrá la siguiente distribución:

$$\text{p(x)=} \begin{cases} p(1) = P(X_i = 1) = p \\ \\ p(0) = P(X_i = 0) = 1 - p \end{cases}$$

Es decir, X_i es una variable aleatoria que puede tener solo dos valores posibles, $X_i = 1$ si X_i verifica la propiedad A con probabilidad p, o $X_i = 0$ si no cumple la propiedad A con probabilidad 1-p.

Supongamos que considero una muestra aleatoria de tamaño n, y defino el estadístico $X = X_1 + X_2 + ... + X_n$, es evidente que este estadístico mide el número de elementos de mi muestra que cumplen la propiedad A, por lo tanto X sigue una distribución Binomial con parámetros n y p. Entonces la variable aleatoria $\hat{P} = \frac{X}{n}$ representa la proporción de elementos de la muestra que cumplen la propiedad A.

Se puede probar que $\hat{P}$ es un estimador consistente.

$E[\hat{P}] = E[\frac{X}{n} = E[X]\frac{1}{n} = np\frac{1}{n} = p$
$Var(\hat{P}) = Var(\frac{X}{n}) = \frac{1}{n^2}Var(X) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$
En el límite cuando n tiende a infinito $E[\hat{P}] = p$ y $Var(\hat{P}) = 0$.

Para construir un intervalo de confianza para p definimos el estadístico que será nuestro pivote

$$Z = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Si n es lo suficientemente grande entonces $Z \backsim N(0, 1)$.

Dado que $\hat{P}$ es un estimador puntual consistente de p entonces podemos reemplazarlo en el denominador de Z

$$Z = \frac{\hat{P} - p}{\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}}$$

Este pivote Z también seguirá una N(0,1), siempre que n sea lo suficientemente grande.

Entonces ya podemos calcular el intervalo de p con un nivel de confianza 1 – α de la misma forma que veníamos haciendo, el intervalo nos queda:

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

Siendo $z_{\frac{\alpha}{2}}$ aquel valor tal que $P(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$.

Nota: Este procedimiento depende de la aproximación normal a la distribución binomial. Será válido siempre que se cumpla $n\hat{P} > 10$ y $n(1-\hat{P}) > 10$. Es decir, la muestra debe contener un mínimo de 10 éxitos y 10 fracasos.

Ejemplo:

Un fabricante de componentes electrónicos compra un lote de dispositivos de segunda mano y desea saber la proporción de dispositivos que están fallados. Con ese fin experimenta con 140 dispositivos elegidos al azar y encuentra que 35 de ellos están fallados.

a) Calcular un intervalo de confianza del 99 % para la proporción poblacional de productos fallados.

b) ¿De qué tamaño deberá extraerse la muestra a fin de que la proporción de productos defectuosos de la muestra no difiera de la proporción poblacional en más de 0.03 con un 95 % de confianza?

a) El tamaño de la muestra es n = 140 (muestra grande). La proporción de productos defectuosos de la muestra es $\hat{P} = \frac{35}{140} = 0{,}25$.

El nivel de confianza $1–\alpha = 0{,}99 \longrightarrow \alpha = 0{,}01 \longrightarrow \frac{\alpha}{2} = 0{,}005$

De la tabla de la distribución normal estándar vemos que $z_{0{,}005} = 2{,}58$. Entonces el intervalo buscado es:

$$\left[0{,}25 - 2{,}58\sqrt{\frac{0{,}25(1-0{,}25)}{140}}, 0{,}25 + 2{,}58\sqrt{\frac{0{,}25(1-0{,}25)}{140}}\right] = [0{,}155; 0{,}344]$$

b) Buscamos el tamaño n de la muestra tal que, con una confianza del 95, % $\hat{P}$ no difiera de p en mas de 0,03. Recordemos que la mitad del largo del intervalo representa la precisión de mi estimador, entonces buscamos el n tal que $\frac{L}{2} < 0{,}03$, por lo tanto como $1–\alpha = 0{,}95 \longrightarrow \alpha = 0{,}05 \longrightarrow \frac{\alpha}{2} = 0{,}025$. De la tabla de N(0,1) obtengo $z_{0{,}025} = 1{,}96$.

El largo del intervalo de $\hat{P}$ viene dado por

$$L = 2z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$$

buscamos n tal que $\frac{L}{2} \leq 0{,}03$, despejando llegamos a que

$$n \geq \left(\frac{2z_{\frac{\alpha}{2}}}{n}\right)^2 \hat{P}(1-\hat{P})$$

$$n \geq \left(\frac{2 * 1{,}96}{2 * 0{,}03}\right)^2 0{,}25(1 - 0{,}25) = 801$$

Es decir que si tomo una muestra de 801 elementos puedo decir con un 95 % de confianza que la proporción de productos defectuosos de la muestra, $\hat{P}$, diferirá de la poblacional p como máximo en 0.03.