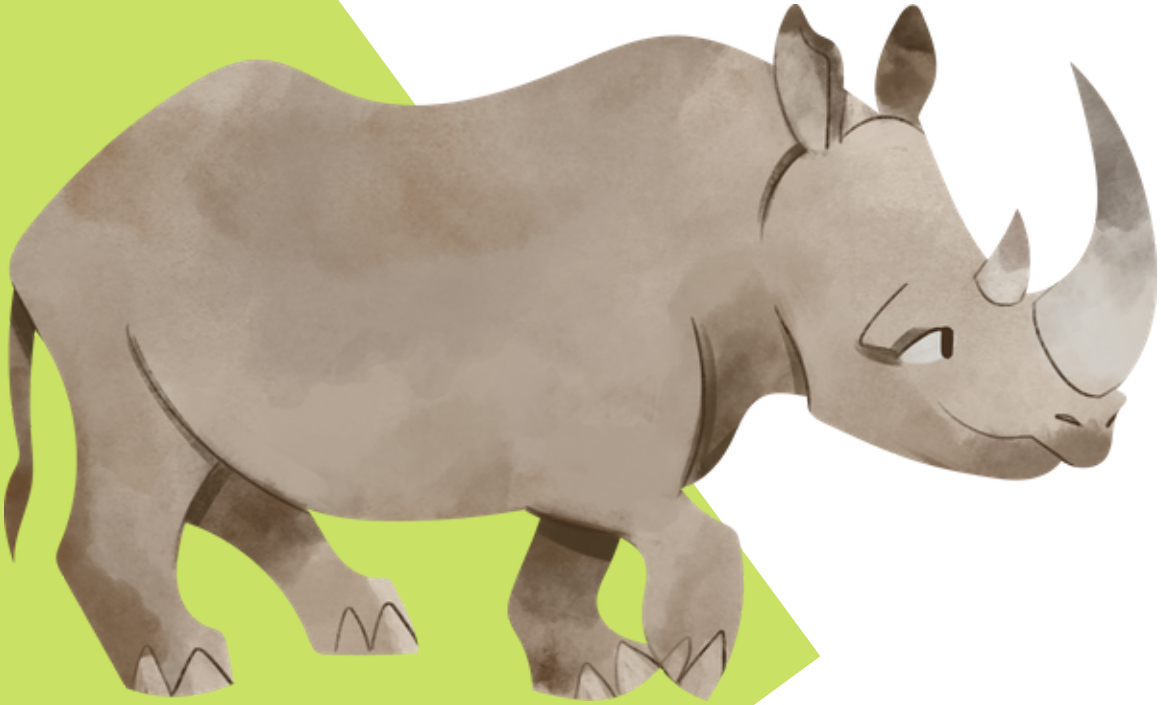
Fundamentos



Teoría Estadística

Índice general

Lista de Ecuaciones

Este es un resumen y mis apuntes de estudio para el curso XS0100-Fundamentos de teoría estadística. Siendo esta su última versión compilada [29 de junio de 2021].

Clase 1: Estimación Puntual

Fecha: 17/05/2021

Profe: Alexander Franck *Por: Ricardo Huapaya*

Partimos de lo que conocimos y probamos sobre estimadores de los parámetros poblacionales, los cuales nos dan una estimación puntual, es decir, no hemos incorporado una noción probabilística a la hora de aproximarnos al verdadero valor poblacional.

Queremos encontrar un intervalo que tenga dos condiciones:

1. Que contenga al parámetro θ
2. Su amplitud sea relativamente pequeña.

Los límites de ese intervalo los llamaremos límites de confianza.

Un intervalo de confianza contiene en sus límites (variables aleatorias) el parámetro fijo (θ) con una probabilidad que se denota con $1-\alpha$ Esta probabilidad, $1-\alpha$ indicará la proporción de veces que la estimación de una muestra aleatoria caerá en ese intervalo.

$$P[\theta_L \leq \theta \leq \theta_U] = 1-\alpha \tag{1.1}$$

El intervalo $[\theta_L, \theta_U]$ es el intervalo de confianza bilateral.

También podemos definir el intervalo de confianza unilateral de la forma

$$P[\theta_L \leq \theta] = 1-\alpha \tag{1.2}$$

En este caso el intervalo está formado por un solo límite, pero tendría la forma: $[\theta_L, \infty[$ Análogamente podríamos obtener:

$$P[\theta_U \geq \theta] = 1-\alpha$$

Con intervalo de confianza $[-\infty, \theta_u[$

1.1. Métodos

1.1.1. Método Pivotal

El método que funciona conceptualmente de base para todas las estimaciones por intervalos es el método pivotal. Este consiste en *encontrar* una variable pivote que cumpla con las siguientes dos condiciones

- Que sea función de los valores muéstrales y el parámetro desconocido (θ), y este es el único parámetro desconocido.

- Que su distribución de probabilidad no dependa del parámetro (θ).

Luego partimos de las siguientes dos propiedades algebráicas:

Hecho 1.1.1. *Sea c una constante arbitraria desconocida, $c > 0$, y $P(a \leq Y \leq b) = 1 - \alpha$ un intervalo de confianza para Y; entonces note que:*

1. $P(ca \leq cY \leq cb) = 1 - \alpha$
2. $P(c + a \leq c + Y \leq c + b) = 1 - \alpha$

Ejemplo 1.1.2 (8.4 del Mendelhall)**.** Suponga que obtenemos una sola observación Y de una distribución exponencial con media θ. Use Y para construir un intervalo de confianza para θ con un coeficiente de confianza de .90.

La función de densidad de probabilidad esta dada por:

$$f(y) = \begin{cases} \left(\frac{1}{\theta}\right) e^{\frac{-y}{\theta}}, & y \geq 0 \\ 0, & y < 0 \end{cases}$$

Para ello tome $U = \frac{Y}{\theta}$, entonces:

$$f_U(u) = \begin{cases} e^{-u}, & u \geq 0 \\ 0, & u < 0 \end{cases}$$

La distribución de U no depende de θ. Entonces, podemos emplear $U = \frac{Y}{\theta}$ como cantidad pivote. Como buscamos un estimador de intervalo con coeficiente de confianza igual a .90, encontramos dos números a y b tales que:

$$P(a \leq U \leq b) = 0{,}90$$

Para ello debe elegir un $a \wedge b$ de la forma:

$$P(U < a) = \int_{-\infty}^{a} e^{-u} du = 0{,}05 \qquad P(U > b) = \int_{b}^{\infty} e^{-u} du = 0{,}05$$

$$1 - e^{-a} = 0{,}05 \qquad e^{-b} = 0{,}05$$

$$a = 0{,}051 \qquad b = 2{,}996$$

$$\therefore$$

$$P(0{,}051 \leq U \leq 2{,}996) = 0{,}90 = P\left(0{,}051 \leq \frac{Y}{\theta} \leq 2{,}996\right)$$

$$P\left(\frac{0{,}051}{Y} \leq \frac{1}{\theta} \leq \frac{2{,}996}{Y}\right) = 0{,}90$$

$$P\left(\frac{Y}{0{,}051} \geq \theta \geq \frac{Y}{2{,}996}\right) = 0{,}90$$

$$P\left(\frac{Y}{2{,}996} \leq \theta \leq \frac{Y}{0{,}051}\right) = 0{,}90$$

Entonces, vemos que $Y/2{,}996$ y $Y/{,}051$ forman los límites de confianza inferior y superior, respectivamente, que estábamos buscando. Para obtener los valores numéricos de estos límites debemos observar un valor real para Y y sustituirlo en las fórmulas dadas para los límites de confianza. Sabemos que límites de la forma $Y/2{,}996, Y/{,}051)$ incluirán los valores (desconocidos) verdaderos de θ para 90 % de los valores de Y que obtendríamos por muestreo repetido a partir de esta distribución exponencial■

Clase 2: Intervalos para medias y varianzas poblacionales

Fecha: 20/05/2021

Profe: Alexander Franck *Por: Ricardo Huapaya*

2.1. Métodos

2.1.1. Métodos para la Media

Considere a X es una normal con σ^2 conocida, usamos $\overline{x}$ para estimar μ.

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

$$P\left(\mu - 1{,}96\frac{\sigma}{\sqrt{n}} \leq \overline{x} \leq \mu + 1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

$$P\left(\overline{x} - 1{,}96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + 1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

Considere el caso de una σ^2 desconocida.

$$P\left(\overline{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right) = 1 - \alpha \tag{2.1}$$

$$\text{con } t_{n-1} = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \tag{2.2}$$

2.1.2. Métodos para la Varianza

Defina: $Q = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$

$$P\left(q_1 \leq \frac{(n-1)s^2}{\sigma^2} \leq q_2\right) = 1 - \alpha \tag{2.3}$$

$$P\left(\frac{(n-1)s^2}{q_2} \leq \sigma^2 \leq \frac{(n-1)s^2}{q_1}\right) = 1 - \alpha \tag{2.4}$$

Note que para los caso unilaterales es de la forma:

$$P(Q \geq q_1) = \frac{\alpha}{2} \wedge P(Q \leq q_2) = \frac{\alpha}{2}$$

Ejemplo 2.1.1. Un fabricante afirma que tiene un artefacto que produce mediciones con una exactitud de 1.5 micras. Con base en esta afirmación, el fabricante otorga la garantía. Se obtuvo una muestra de cinco lecturas de un mismo objeto y cuyas mediciones son las siguientes: 135, 137, 138, 137 y 139. Determine un intervalo de confianza de $90\,\%$ para la varianza poblacional.

Solución:(Excel) Note que tenemos una muestra con un tamaño de 5 observaciones, para ello podemos calcular la varianza tabulando los datos en excel para ello la formula `VAR()` y encontramos que s^2 de la muestra es $2{,}2$.

Con ello nada más queda encontrar los quintiles con la fámula de excel `INV.CHICUAD(p, k_g.l.)` que devuelve el inverso de la probabilidad de cola izquierda de la distribución chi cuadrado. Asi $q_1 = 0{,}71 \wedge q_2 = 9{,}48$.

Finalizamos sustituyendo:

$$P\left(\frac{(5-1)2{,}2}{9{,}48} \leq \sigma^2 \leq \frac{(5-1)2{,}2}{0{,}71}\right) = 0{,}90$$
$$P\left(0{,}93 \leq \sigma^2 \leq 12{,}38\right) = 0{,}90$$

Por lo tanto para cada 100 muestras que se efectúen el 90 de estas su varianza muestral se encuentran dentro de [0.93, 12.38]. ■

2.1.3. Métodos para comparación de medias

Sea $X_1, X_2, \ldots, X_m$ muestra aleatoria de tamaño m con distribución normal con media μ_x varianza σ^2, y sea $Y_1, Y_2, \ldots, Y_n$ muestra aleatoria de tamaño n con distribución normal con media μ_y y varianza σ^2.

$$Q = \frac{\overline{x} - \overline{y}}{s_p\sqrt{\frac{1}{m} + \frac{1}{m}}} = \frac{\overline{x} - \overline{y}}{S_{\overline{x}-\overline{y}}}$$

Con: $s_p^2 = \frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2}$

Por ello tenemos entonces:

$$P(-t_{\alpha/2}m + n - 2 \leq Q \leq t_{\alpha/2}m + n - 2) = 1 - \alpha \tag{2.5}$$
$$P((\overline{x} - \overline{y}) - t_{\alpha/2,m+n-2}S_{\overline{x}-\overline{y}} \leq \overline{x} - \overline{y} \leq (\overline{x} - \overline{y}) + t_{\alpha/2,m+n-2}S_{\overline{x}-\overline{y}}) = 1 - \alpha \tag{2.6}$$

2.1.4. Método para la Comparación de Varianzas

$$F = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} \sim F_{n_1-1,n_2-1} \tag{2.7}$$
$$P\left(\frac{s_1^2/s_2^2}{F_{1-\alpha/2}} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2/s_2^2}{F_{\alpha/2}}\right) \tag{2.8}$$

2.1.5. Métodos para una proporción y diferencia de proporciones

$$P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \tag{2.9}$$

$$P\left((\hat{p}_x - \hat{p}_y) - z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{m} + \frac{\hat{p}_y(1-\hat{p}_y}{n}} \leq \hat{p}_x - \hat{p}_y \leq (\hat{p}_x - \hat{p}_y) + z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{m} + \frac{\hat{p}_y(1-\hat{p}_y}{n}}\right) \tag{2.10}$$

2.2. Practica

Ejercicio 2.2.1. (8.50 Mendenhall) Consulte el Ejemplo 8.8. En este ejemplo, p1 y p2 se usaron para denotar las proporciones de refrigeradores de las marcas A y B, respectivamente, que fallaron durante los períodos de garantía

1. En el nivel aproximado de $98\,\%$ de confianza, ¿cuál es el mayor "valor creíble" para la diferencia en las proporciones de fallas de refrigeradores de las marcas A y B?
2. En el nivel aproximado de $98\,\%$ de confianza, ¿cuál es el menor "valor creíble" para la diferencia en las proporciones de fallas de refrigeradores de las marcas A y B?
3. Si p1 - p2 es realmente igual a 0.2251, ¿cuál marca tiene la mayor proporción de fallas durante el período de garantía? ¿Qué tanto más grande?
4. Si p1 - p2 es realmente igual a –0.1451, ¿cuál marca tiene la mayor proporción de fallas durante el período de garantía? ¿Qué tanto más grande?
5. Como se observó en el Ejemplo 8.8, cero es un valor creíble de la diferencia. ¿Concluiría usted que hay evidencia de una diferencia en las proporciones de fallas (dentro del período de garantía) para las dos marcas de refrigeradores? ¿Por qué?

Solución

Podemos reescribir el Intervalo de Confianza para las proporciones de la forma,

$$(\hat{p}_x - \hat{p}_y) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{m} + \frac{\hat{p}_y(1-\hat{p}_y}{n}}$$

Usando excel podemos hacer el calculo de $Z_{\alpha/2}$ `INV.NORM.ESTAND(0,99)`, en este caso nos da 2,33 con ello podemos calcular el coeficiente de la forma.

$$(0{,}24 - 0{,}20) \pm 2{,}33\sqrt{\frac{0{,}24(1-0{,}24)}{50} + \frac{0{,}20(1-0{,}20)}{60}}$$

Por lo que el intervalo de confianza nos da $[-0{,}145, 0{,}225]$.

(1) En el nivel $98\,\%$ de confianza el mayor valor creíble es 0,225. (2) De forma análoga el menor valor creíble de diferencia entre las proporciones es de -0.145.

(3) En este caso la marca A sería la que presente más fallos y sería casi el doble de fallos (4) dado caso la marca B sería la que más presente fallos y seria cuatro veces la cantidad de fallo. (5) No hay evidencia pues 0 es un valor dentro del intervalo.

Ejercicio 2.2.2. 33

Ejercicio 2.2.3. (8.58 Mendenhall) Los administradores de un hospital deseaban estimar el número promedio de días necesarios para el tratamiento de enfermos internados entre las edades de 25 y 34 años. Una muestra aleatoria de 500 pacientes entre estas edades produjo una media y una desviación estándar igual a 5.4 y 3.1 días, respectivamente. Construya un intervalo de confianza del 95 % para la duración media de permanencia de la población de pacientes de la cual se extrajo la muestra.

Solución: (Excel)

Para ello ello aplicamos

$$P\left(\overline{x} - 1{,}96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + 1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

Sustituimos valores:

$$P\left(5{,}4 - 1{,}96\frac{3{,}1}{\sqrt{500}} \leq \mu \leq 5{,}4 + 1{,}96\frac{3{,}1}{\sqrt{500}}\right) = 0{,}95$$

$$P\left(5{,}13 \leq \mu \leq 5{,}67\right) = 0{,}95$$

En este caso el valor de $1{,}96$ es el $z_{\alpha/2}$ que se obtiene en excel aplicando la formula:

```
INV.NORM.ESTAND(I.de Confianza + (1-I.de Confianza)/2)
```

Ejercicio 2.2.4. (8.62 Mendenhall) Los siguientes estadísticos son el resultado de un experimento realizado por P. I. Ward para investigar una teoría relativa al comportamiento de cambio de piel del macho Gammarus pulex, un pequeño crustáceo. Si el macho cambia de piel mientras se aparea con una hembra, éste debe liberarla y perderla. La teoría es que el macho Gammarux pulex es capaz de posponer dicho cambio, con lo cual reduce la posibilidad de perder su pareja. Ward asignó aleatoriamente 100 parejas de machos y hembras a dos grupos de 50 cada uno. Las parejas del primer grupo se mantuvieron juntas (normal); las del segundo grupo fueron separadas. Se registró el tiempo de muda para machos y hembras, y las medias, desviaciones estándar y tamaños muestrales se ilustran en la tabla siguiente. (El número de crustáceos en cada una de las cuatro muestras es menor que 50 porque algunos en cada grupo no sobrevivieron hasta el tiempo de muda.)

Clase 3: Contraste Hipótesis estadisticas

Fecha: 20/05/2021

Profe: Alexander Franck *Por: Ricardo Huapaya*

3.1. Preliminares

3.1.1. Hipótesis estadísticas

Por ejemplo, si se quiere probar que $\mu = \mu_0$ son base en una muestra aleatoria $X_1, X_2, \ldots, X_m$ la hipótesis nula es $H_0 : \mu = \mu_0$

Esta hipótesis se contrasta con la **hipótesis alternativa** H_a que especifica un valor alternativo para μ puede ser $H_a : \mu = \mu_0$ o $H_a : \mu \; neq \mu_0$; también $H_a : \mu > \mu_0$ o $H_a : \mu < \mu_0$

Hipótesis simple: se especifica completamente la distribución de la cual se tomó la muestra, $H_0 : \mu = \mu_0$

Hipótesis compuesta: no se especifica completamente la distribución de la cual se tomó la muestra, por ejemplo, $H_a : \mu > \mu_0$

3.1.2. Tipos de Error

Cuando contrastamos H_0 vs H_1, podemos cometer dos tipos de error:

1. Rechazar H_0 cuando esta es verdadera. Error tipo I $P(I) = \alpha$. Se llama nivel de significancia del contraste.
2. No rechazar H_0 cuando esta es falsa. Error tipo II. $P(II) = \beta$

3.2. Procedimientos de contraste de hipótesis

Necesitamos un estimador de θ,$\hat{\theta} = h(x_1, x_2, \ldots, x_n)$ y determinar su distribución muestral $g(\hat{\theta}; \theta_0)$.

Dividir la región de todos los valores posibles $\mathcal{R}$ en dos regiones: una región de rechazo RR, en la que θ_0 es poco probable, tal que $P(\hat{\theta} \in RR | \theta = \theta_0) = P(I) = \alpha$.

Calcular el valor de $\hat{\theta}$ con la muestra aleatoria y rechazar H_0 si cae en la región de rechazo, y, si no, no rechazar H_0.

Clase 4: Lemma Neyman Pearson

Fecha: 20/05/2021

Profe: Alexander Franck *Por: Ricardo Huapaya*

4.1. Concepto de Espacio paramétrico y potencia

Sea una población descrita por $f_{x;\theta}$ y Ω el conjunto de valores que puede tomar θ, llamado espacio paramétrico. Sea ω el subconjunto de Ω que contiene todos los valores de θ que establece la hipótesis nula.

$$H_0 : \theta = \theta_0$$
$$H_a : \theta \in \Omega - \{\theta_0\}$$

Sea $\hat{\theta}$ un estimador de θ que se utiliza para contrastar H_0 con H_a y RR la zona de rechazo. La función potencia $\Pi(\theta)$ definida para $\theta \in \Omega$ por:

$$\Pi(\theta) = \begin{cases} P(\hat{\theta} \in RR) = 1 - \beta & \text{si } \theta \neq \theta_0 \\ \alpha & \text{si } \theta = \theta_0 \end{cases} \tag{4.1}$$

4.2. Neyman-Pearson

Con un nivel de α fijado, queremos maximizar la potencia de la prueba. Para eso se usa el lema de Neyman-Pearson (solo para contrastar hipótesis simples)

Lemma 4.2.1. *(Neyman-Pearson) Para contrasta $H_0 : \theta = \theta_0$ con $H_a : \theta = \theta_a$, con base en una muestra aleatoria $Y_1, Y_2, \ldots, Y_n$ de una distribución con parámetro θ. Sea $L(\theta)$ la función de verosimilitud de la muestra cuando el valor del parámetro es θ. Entonces, para un α dado, la prueba que maximiza la potencia en θ_a tiene una región de rechazo RR determinada por:*

$$\frac{L(\theta_0)}{L(\theta_a)} < k$$

El valor de k se escoge para que tenga el valor deseado para α

Clase 5: Estadistica Bayesiana

Fecha: 28/6/2021

Profe: Alexander Franck *Por: Ricardo Huapaya*

5.1. Contraste con el Enfoque frecuentista

Coincide con los enfoques de la probabilidad, incluido los enfoques **a priori** racionalista de inferencias y estocástico determinado por la cantidad de resultados; **el enfoque clásico** sea frecuentista que de manera escéptica el enfoque de inferencia y estima a partir del estudio empírico usando las muestras.

Enfoque subjetivo: basado en creencias dado a que incorpora información subjetiva en el calculo matemático, considera que θ debe ser una distribución de probabilidad. Pues así antes de ver la probabilidad de los datos asignamos una distribución a priori que llamamos $\Pi(\theta)$, $0 \leq \theta \leq 1$. Queremos encontrar una distribución a posteriori.

La distribución posteriori $\Pi(\theta|y)$ se obtiene por medio del teorema de Bayes.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{5.1}$$

$$\Pi(\theta|Y) = \frac{P(Y|\theta) \cdot P(\theta)}{P(Y)} \tag{5.2}$$

También puedo escribir la función de distribución posterior como:

$$\Pi(\theta|y) \propto f(y|\theta) \cdot \Pi(\theta)$$

Ejemplo 5.1.1. Considere la función de verosimilitud de un solo lanzamiento de dados se puede escribir como $P(Y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i}$, es decir $P(Y_i = 1) = \theta$ y $P(Y_i = 0) = 1 - \theta$.

Entonces,

$$\begin{aligned} P(Y_1|\theta, \ldots, Y_n|\theta) &= \theta^{y_1}(1-\theta)^{1-y_1} \cdot \cdots \cdot \theta^{y_n}(1-\theta)^{1-y_n} \\ &= \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \\ &= \theta^{\sum y_i}(1-\theta)^{n-\sum y_i} \end{aligned}$$

La especificación del modelo requiere una distribución a priori de $0 \leq \theta \leq 1$. ■