Estadística Inferencial

Dr. Alejandro Rodríguez

23 de marzo de 2022

Estadística inferencial

Estadística inferencial

Se llama estadística inferencial o inferencia estadística a la rama de la Estadística encargada de hacer **deducciones**, es decir, **inferir propiedades**, **conclusiones** y **tendencias**, a partir de una muestra del conjunto. **Su papel es interpretar, hacer proyecciones y comparaciones.**

Estadística inferencial

Estadística inferencial

- Estimación estadística
- Prueba de Hipótesis
- Regresión Lineal y Correlación
- Diseño de experimentos

Estimación estadística

Estimación estadística

La estimación es la determinación de un elemento o factor. Esto, usualmente tomando como referencia una base o conjunto de datos.

Estimación estadística

Estimación estadística

La estimación es la determinación de un elemento o factor. Esto, usualmente tomando como referencia una base o conjunto de datos.

La estimación es un cálculo que se realiza a partir de la evaluación estadística. Dicho estudio suele efectuarse sobre una muestra y no sobre toda la población objetivo. Sea $X_1, X_2, X_3, \ldots, X_n$ una muestra aleatoria de una distribución, un estimador es un estadístico $\hat{\theta} = T(X_1, X_2, X_3, \ldots, X_n)$ que sirve para estimar el valor θ.

Estimación estadística

Estimación estadística

La estimación es la determinación de un elemento o factor. Esto, usualmente tomando como referencia una base o conjunto de datos.

La estimación es un cálculo que se realiza a partir de la evaluación estadística. Dicho estudio suele efectuarse sobre una muestra y no sobre toda la población objetivo. Sea $X_1, X_2, X_3, \ldots, X_n$ una muestra aleatoria de una distribución, un estimador es un estadístico $\hat{\theta} = T(X_1, X_2, X_3, \ldots, X_n)$ que sirve para estimar el valor θ.
Nos sirve para calcular indicadores estadísticos como la **media**, la **mediana**, **moda** y **desviación estándar**.

Estimación estadística. Puntual

Estimación estadística Puntual

La estimación puntual consiste en encontrar un valor para θ, denotado por $\hat{\theta}$. Es decir seleccionar aquel estadístico que mejor nos permita describir la muestra.

Estimación estadística. Puntual

Estimación estadística Puntual

La estimación puntual consiste en encontrar un valor para θ, denotado por $\hat{\theta}$. Es decir seleccionar aquel estadístico que mejor nos permita describir la muestra.

Ejemplo: Por ejemplo, si se pretende estimar la talla media de un determinado grupo de individuos, puede extraerse una muestra y ofrecer como estimación puntual la talla media de los individuos.

Estimación estadística. Puntual

Podría ser que ni el estimador eficaz estime con exactitud el parámetro de la población. Sabiendo que la exactitud de la estimación aumenta cuando las muestras son grandes; pero incluso así no tenemos razones para esperar que una estimación puntual de una muestra dada sea exactamente igual al parámetro de la población que se supone debe estimar.

Hay muchas situaciones en que es preferible determinar un intervalo dentro del cual esperaríamos encontrar el valor del parámetro. Tal intervalo se conoce como **estimación por intervalo**.

Estimación estadística. Estimación por intervalo

En la estimación por intervalo calculamos el intervalo de confianza. Este intervalo intervalo $E_+ < \bar{x} < E_-$, que se calcula a partir de la muestra seleccionada, se llama entonces intervalo de confianza y los extremos, E_+ y E_-, se denominan límites de confianza inferior y superior.

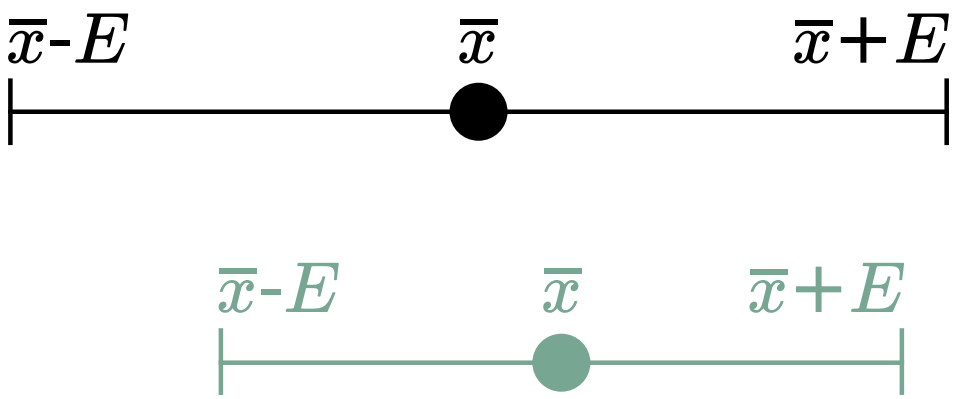

Figura: Ejemplo de dos posibles valores y sus intervalos de confianza.

Estimación estadística. Estimación por intervalo

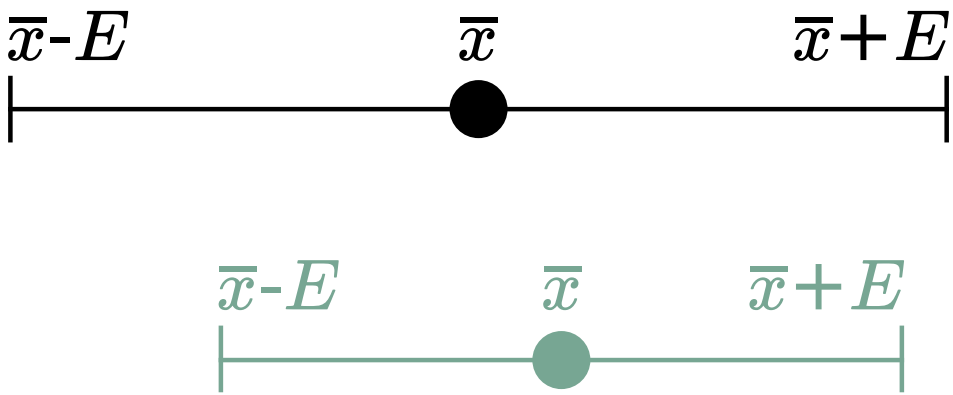

Figura: Ejemplo de dos posibles valores y sus intervalos de confianza.

Para hacer la estimación de los límites de confianza emplearemos la siguiente expresión:

$$E_{\pm} = \frac{1{,}96 \times \sigma}{\sqrt{n}}$$

Siendo: σ desviación estándar $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$ y n el numero de elementos.

Prueba de significación o Hipótesis

Prueba de Hipótesis

La **verificación de hipótesis** es el proceso que lleva a juzgar la credibilidad de declaraciones tentativas (hipótesis) relativas a las poblaciones de las que fueron extraídas las muestras.

Ejemplo

Analizamos el contenido de sodio en unas muestras de un lote de unas galletas y obtenemos un valor medio de $352mg$. Afirmamos entonces que el contenido medio de sodio en ese lote es de $352mg$.

Prueba de Hipótesis

La **verificación de hipótesis** es el proceso que lleva a juzgar la credibilidad de declaraciones tentativas (hipótesis) relativas a las poblaciones de las que fueron extraídas las muestras.

Ejemplo

Analizamos el contenido de sodio en unas muestras de un lote de unas galletas y obtenemos un valor medio de $352mg$. Afirmamos entonces que el contenido medio de sodio en ese lote es de $352mg$.
Entonces otra persona, que puede ser, nuestro cliente, obtiene de unas muestras de ese mismo lote un valor medio del contenido de sodio igual a $375mg$.

Prueba de Hipótesis

La **verificación de hipótesis** es el proceso que lleva a juzgar la credibilidad de declaraciones tentativas (hipótesis) relativas a las poblaciones de las que fueron extraídas las muestras.

Ejemplo

Analizamos el contenido de sodio en unas muestras de un lote de unas galletas y obtenemos un valor medio de $352mg$. Afirmamos entonces que el contenido medio de sodio en ese lote es de $352mg$.
Entonces otra persona, que puede ser, nuestro cliente, obtiene de unas muestras de ese mismo lote un valor medio del contenido de sodio igual a $375mg$.
¿Contradice este resultado o no mi afirmación de que el contenido medio es de 352 mg? ¿Aceptará o rechazará el lote mi cliente?

Prueba de Hipótesis

La **verificación de hipótesis** es el proceso que lleva a juzgar la credibilidad de declaraciones tentativas (hipótesis) relativas a las poblaciones de las que fueron extraídas las muestras.

Ejemplo

Analizamos el contenido de sodio en unas muestras de un lote de unas galletas y obtenemos un valor medio de $352mg$. Afirmamos entonces que el contenido medio de sodio en ese lote es de $352mg$.
Entonces otra persona, que puede ser, nuestro cliente, obtiene de unas muestras de ese mismo lote un valor medio del contenido de sodio igual a $375mg$.
¿Contradice este resultado o no mi afirmación de que el contenido medio es de 352 mg? ¿Aceptará o rechazará el lote mi cliente?

Para esclarecer esto tenemos que seguir los procedimientos de las pruebas de hipótesis.

Prueba de Hipótesis

Prueba de Hipótesis

Lo que se puede hacer es afirmar que tiene tal o cual probabilidad de ser verdadera o falsa. Si la probabilidad de ser verdadera es muy alta (95 % o 99 %) por ejemplo, se concluye que la hipótesis es altamente creíble y se califica provisionalmente como cierta.
Si no se consigue probar que es verdadera (se rechaza), se acepta provisionalmente como falsa.

Prueba de Hipótesis

Casi siempre se plantean dos hipótesis: **la hipótesis nula** y **su contraria, la hipótesis alternativa**, planteándose generalmente como: H_0 y H_1.
Pasos a seguir para la prueba de hipótesis:

- Definir la hipótesis a contrastar.
- Definir la prueba estadística a emplear.
- Establecer el nivel de significación.
- Hacer la experimentación (muestrear y evaluar el o los estadísticos.
- Tomar la decisión sobre la hipótesis.

Prueba de Hipótesis F-Fisher

Distribuciones F-Fisher

Prueba F para igualdad de varianzas
Hipótesis nula: $H_0 : \sigma_1^2 = \sigma_2^2$
Valor estadístico de prueba: $f = s_1^2/s_2^2$ Siendo α el grado insignificancia, por lo tanto $1 - \alpha$ cuantificaría el nivel de confianza.

Hipótesis alternativa	Región de rechazo para una prueba de nivel α
$H_a : \sigma_1^2 > \sigma_2^2$	$f \geq F_{\alpha,m-1,n-1}$
$H_a : \sigma_1^2 < \sigma_2^2$	$f \leq F_{1-\alpha,m-1,n-1}$
$H_a : \sigma_1^2 \neq \sigma_2^2$	o $f \geq F_{\alpha/2,m-1,n-1}$ o $f \leq F_{1-\alpha/2,m-1,n-1}$

Como los valores críticos se tabulan sólo para $\alpha = 0.10$, 0.05, 0.01 y 0.001. Con software estadístico se obtienen otros valores críticos F.

Distribuciones F-Fisher. Ejemplo

Ejemplo

La variabilidad en la cantidad de grasa presente en un lote de un complemento dietético, utilizada para un proceso de fabricación de un alimento, depende del origen del complemento. Un fabricante que recibe el complemento de dos proveedores **1** y **2**, hizo una comparación analizando muestras de ambos proveedores. Muestras de $n_1 = 10$ y $n_2 = 16$ mediciones de dos lotes produjeron las varianzas:
$S_1^2 = 1{,}25$ y $S_2^2 = 0{,}5$
¿Presentan los datos evidencia suficiente para indicar que la variabilidad en el contenido de grasa es menor para el producto que se recibe del proveedor 2? Realice una prueba con un a o $\alpha = 0.05$.

Distribuciones F-Fisher. Ejemplo

Solución

Tenemos nuestras hipótesis nula, la cual aceptamos a priori:

$H_a : \sigma_1^2 > \sigma_2^2 \quad : \quad f \geq F_{\alpha,m-1,n-1}$

Y rechazamos si:

$f \geq F_{\alpha,m-1,n-1}$

Luego, debemos buscar en tablas con una exactitud de 0.05 el valor de F_{tabla}.

Distribuciones F-Fisher. Ejemplo

Solución

Tenemos nuestras hipótesis nula, la cual aceptamos a priori:

$H_a : \sigma_1^2 > \sigma_2^2 \quad : \quad f \geq F_{\alpha,m-1,n-1}$

Y rechazamos si:

$f \geq F_{\alpha,m-1,n-1}$

Luego, debemos buscar en tablas con una exactitud de 0.05 el valor de F_{tabla}. $f \geq 2{,}5876$

$$f_{calculado} = \frac{1{,}25}{0{,}5} = 2{,}5$$

Distribuciones F-Fisher. Ejemplo

Solución

Tenemos nuestras hipótesis nula, la cual aceptamos a priori:

$H_a : \sigma_1^2 > \sigma_2^2 \quad : \quad f \geq F_{\alpha,m-1,n-1}$

Y rechazamos si:

$f \geq F_{\alpha,m-1,n-1}$

Luego, debemos buscar en tablas con una exactitud de 0.05 el valor de F_{tabla}. $f \geq 2{,}5876$

$$f_{calculado} = \frac{1{,}25}{0{,}5} = 2{,}5$$

Como $2.5 < 2.5876$ no se rechaza H_0, y se concluye con un $\alpha = 0{,}05$ que no existe suficiente evidencia para decir que la variabilidad del contenido de grasa del complemento del proveedor 2 es menor que la del complemento suministrado por el proveedor 1.

Distribuciones F-Fisher. La Tabla

Tabla VALORES F DE LA DISTRIBUCIÓN F DE FISHER

$1 - \alpha = 0.9$ $\quad \nu_1$ = grados de libertad del numerador

$1 - \alpha = P\,(F \leq f_{\alpha,\nu_1,\nu_2})$ $\quad \nu_2$ = grados de libertad del denominador

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	11	12
1	39.864	49.500	53.593	55.833	57.240	58.204	58.906	59.439	59.857	60.195	60.473	60.705
2	8.526	9.000	9.162	9.243	9.293	9.326	9.349	9.367	9.381	9.392	9.401	9.408
3	5.538	5.462	5.391	5.343	5.309	5.285	5.266	5.252	5.240	5.230	5.222	5.216
4	4.545	4.325	4.191	4.107	4.051	4.010	3.979	3.955	3.936	3.920	3.907	3.896
5	4.060	3.780	3.619	3.520	3.453	3.405	3.368	3.339	3.316	3.297	3.282	3.268
6	3.776	3.463	3.289	3.181	3.108	3.055	3.014	2.983	2.958	2.937	2.920	2.905
7	3.589	3.257	3.074	2.961	2.883	2.827	2.785	2.752	2.725	2.703	2.684	2.668
8	3.458	3.113	2.924	2.806	2.726	2.668	2.624	2.589	2.561	2.538	2.519	2.502
9	3.360	3.006	2.813	2.693	2.611	2.551	2.505	2.469	2.440	2.416	2.396	2.379
10	3.285	2.924	2.728	2.605	2.522	2.461	2.414	2.377	2.347	2.323	2.302	2.284
11	3.225	2.860	2.660	2.536	2.451	2.389	2.342	2.304	2.274	2.248	2.227	2.209
12	3.177	2.807	2.606	2.480	2.394	2.331	2.283	2.245	2.214	2.188	2.166	2.147
13	3.136	2.763	2.560	2.434	2.347	2.283	2.234	2.195	2.164	2.138	2.116	2.097

La columna indica el grado de libertad del numerador y la fila el grado de libertad del denominador. La Figura17 muestra una sección de la tabla de la distribución F para el caso de un nivel de significancia de 10 %, es decir $\alpha = 0{,}1$. Aparece resaltado el valor de **F** cuando $\nu_1 = 3$ y $\nu_2 = 6$ con nivel de confianza $1 - \alpha = 0{,}9$ es decir 90 %.

Regresión lineal simple y correlación

Regresión lineal simple

En la práctica a menudo se requiere resolver problemas que implican conjuntos de variables de las cuales se sabe que tienen alguna relación inherente entre sí.

Regresión lineal simple

En la práctica a menudo se requiere resolver problemas que implican conjuntos de variables de las cuales se sabe que tienen alguna relación inherente entre sí.

Sin embargo:

- No todas las casas ubicadas en la misma zona del país, con la misma superficie de construcción, se venden al mismo precio.
- Varios automóviles con un motor del mismo volumen; no tienen que tener el mismo rendimiento de combustible

Regresión lineal simple

Una forma razonable de relación entre la respuesta Y y el **regresor** x es la relación lineal.

$$Y = \beta_0 + \beta_1 x$$

en la que β_0 es la intersección y β_1 es la pendiente.

Regresión lineal simple

Una forma razonable de relación entre la respuesta Y y el **regresor** x es la relación lineal.

$$Y = \beta_0 + \beta_1 x$$

en la que β_0 es la intersección y β_1 es la pendiente.

Análisis de regresión

El concepto de **análisis de regresión** se refiere a encontrar la mejor relación entre Y y x cuantificando la fuerza de esa relación, y empleando métodos que permitan predecir los valores de la respuesta dados los valores del regresor x.

Correlación

Correlación

Tomemos a X como la antigüedad de un automóvil usado y Y representa su precio, se esperaría que los valores grandes de X correspondan a valores pequeños de Y y que los valores pequeños de X correspondan a valores grandes de Y.

Correlación

Tomemos a X como la antigüedad de un automóvil usado y Y representa su precio, se esperaría que los valores grandes de X correspondan a valores pequeños de Y y que los valores pequeños de X correspondan a valores grandes de Y.

Correlación

El **análisis de correlación** intenta medir la fuerza de tales relaciones entre dos variables por medio de un solo número denominado **coeficiente de correlación**.

Correlación

Tomemos a X como la antigüedad de un automóvil usado y Y representa su precio, se esperaría que los valores grandes de X correspondan a valores pequeños de Y y que los valores pequeños de X correspondan a valores grandes de Y.

Correlación

El **análisis de correlación** intenta medir la fuerza de tales relaciones entre dos variables por medio de un solo número denominado **coeficiente de correlación**.

r_{xy} representa el coeficiente de correlación de Pearson.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Correlación. Interpretación

El valor del índice de correlación varía en el intervalo $[-1, 1]$, indicando el signo el sentido de la relación:

- Si $r = 1$, existe una **correlación positiva perfecta**. El índice indica una **dependencia total** entre las dos variables denominada relación directa: *cuando una de ellas aumenta, la otra también lo hace en proporción constante*.

Correlación. Interpretación

El valor del índice de correlación varía en el intervalo $[-1, 1]$, indicando el signo el sentido de la relación:

- Si $r = 1$, existe una **correlación positiva perfecta**. El índice indica una **dependencia total** entre las dos variables denominada relación directa: *cuando una de ellas aumenta, la otra también lo hace en proporción constante*.
- Si $0 < r < 1$ entonces existe una **correlación positiva**.

Correlación. Interpretación

El valor del índice de correlación varía en el intervalo $[-1, 1]$, indicando el signo el sentido de la relación:

- Si $r = 1$, existe una **correlación positiva perfecta**. El índice indica una **dependencia total** entre las dos variables denominada relación directa: *cuando una de ellas aumenta, la otra también lo hace en proporción constante*.
- Si $0 < r < 1$ entonces existe una **correlación positiva**.
- Si $r = 0$ entonces **no existe relación lineal** pero esto *no necesariamente implica que las variables son independientes*: pueden existir todavía *relaciones no lineales* entre las dos variables.

Correlación. Interpretación

El valor del índice de correlación varía en el intervalo $[-1, 1]$, indicando el signo el sentido de la relación:

- Si $r = 1$, existe una **correlación positiva perfecta**. El índice indica una **dependencia total** entre las dos variables denominada relación directa: *cuando una de ellas aumenta, la otra también lo hace en proporción constante*.
- Si $0 < r < 1$ entonces existe una **correlación positiva**.
- Si $r = 0$ entonces **no existe relación lineal** pero esto *no necesariamente implica que las variables son independientes*: pueden existir todavía *relaciones no lineales* entre las dos variables.
- Si $-1 < r < 0$, existe una **correlación negativa**.

Correlación. Interpretación

El valor del índice de correlación varía en el intervalo $[-1, 1]$, indicando el signo el sentido de la relación:

- Si $r = 1$, existe una **correlación positiva perfecta**. El índice indica una **dependencia total** entre las dos variables denominada relación directa: *cuando una de ellas aumenta, la otra también lo hace en proporción constante*.
- Si $0 < r < 1$ entonces existe una **correlación positiva**.
- Si $r = 0$ entonces **no existe relación lineal** pero esto *no necesariamente implica que las variables son independientes*: pueden existir todavía *relaciones no lineales* entre las dos variables.
- Si $-1 < r < 0$, existe una **correlación negativa**.
- Si $r = -1$, existe una **correlación negativa perfecta**. El índice indica una **dependencia total** entre las dos variables llamada relación inversa: *cuando una de ellas aumenta, la otra disminuye en proporción constante*.