Probabilidad y estadística para la Inteligencia Artificial, clase 6

Camilo Enrique Argoty Pulido

Especialización en Inteligencia Artificial

15 de abril de 2025

Origen de la inferencia estadística

Determinar la verdad: El ejemplo del sistema judicial

A lo largo de la historia, siempre ha existido el problema de determinar de los hechos a nivel judicial

Determinar la verdad: El ejemplo del sistema judicial

A lo largo de la historia, siempre ha existido el problema de determinar de los hechos a nivel judicial

Por ello, en derecho se habla de la **verdad verdadera** y la **verdad judicial**

Determinar la verdad: El ejemplo del sistema judicial

A lo largo de la historia, siempre ha existido el problema de determinar de los hechos a nivel judicial

Por ello, en derecho se habla de la **verdad verdadera** y la **verdad judicial**

- La **verdad verdadera** es la realidad de los hechos. Suele ser imposible de conocer con total seguridad.

Determinar la verdad: El ejemplo del sistema judicial

A lo largo de la historia, siempre ha existido el problema de determinar de los hechos a nivel judicial

Por ello, en derecho se habla de la **verdad verdadera** y la **verdad judicial**

- La **verdad verdadera** es la realidad de los hechos. Suele ser imposible de conocer con total seguridad.
- La **verdad judicial** es lo que se puede establecer de los hechos a partir a partir de las evidencias. Puede ser distinta a la verdad verdadera.

La presunción de inocencia

Como todos sabemos, puede haber dos tipos de injusticias que se pueden cometer en un juicio:

La presunción de inocencia

Como todos sabemos, puede haber dos tipos de injusticias que se pueden cometer en un juicio:

Injusticia I: Condenar un inocente.

La presunción de inocencia

Como todos sabemos, puede haber dos tipos de injusticias que se pueden cometer en un juicio:

Injusticia I: Condenar un inocente.

Injusticia II: Absolver un culpable

La presunción de inocencia

Como todos sabemos, puede haber dos tipos de injusticias que se pueden cometer en un juicio:

Injusticia I: Condenar un inocente.

Injusticia II: Absolver un culpable

En la sociedad occidental moderna se considera peor cometer la Injusticia I que la Injusticia II.

La presunción de inocencia

Como todos sabemos, puede haber dos tipos de injusticias que se pueden cometer en un juicio:

Injusticia I: Condenar un inocente.

Injusticia II: Absolver un culpable

En la sociedad occidental moderna se considera peor cometer la Injusticia I que la Injusticia II.

Por lo anterior, el proceso para establecer la verdad judicial está determinado por el denominado **Principio de Presunción de Inocencia:**

La presunción de inocencia

Como todos sabemos, puede haber dos tipos de injusticias que se pueden cometer en un juicio:

Injusticia I: Condenar un inocente.

Injusticia II: Absolver un culpable

En la sociedad occidental moderna se considera peor cometer la Injusticia I que la Injusticia II.

Por lo anterior, el proceso para establecer la verdad judicial está determinado por el denominado **Principio de Presunción de Inocencia:**

Toda persona es inocente hasta que se demuestre lo contrario

¿Qué es **demostrar culpabilidad**?

En un tribunal, ante el juez se presentan las evidencias y contra-evidencias.

¿Qué es **demostrar culpabilidad**?

En un tribunal, ante el juez se presentan las evidencias y contra-evidencias.

Pero, ¿qué proceso se sigue para determinar si se ha demostrado la culpabilidad o no?

¿Qué es **demostrar culpabilidad**?

En un tribunal, ante el juez se presentan las evidencias y contra-evidencias.

Pero, ¿qué proceso se sigue para determinar si se ha demostrado la culpabilidad o no?

Fundamentalmente, lo que el juez determina es si, asumiendo la inocencia, es o no creíble que existan las evidencias que se presentan.

¿Qué es **demostrar culpabilidad**?

En un tribunal, ante el juez se presentan las evidencias y contra-evidencias.

Pero, ¿qué proceso se sigue para determinar si se ha demostrado la culpabilidad o no?

Fundamentalmente, lo que el juez determina es si, asumiendo la inocencia, es o no creíble que existan las evidencias que se presentan.

Si, asumiendo la inocencia del acusado, es creíble que existan las evidencias que se presentan.

¿Qué es **demostrar culpabilidad**?

En un tribunal, ante el juez se presentan las evidencias y contra-evidencias.

Pero, ¿qué proceso se sigue para determinar si se ha demostrado la culpabilidad o no?

Fundamentalmente, lo que el juez determina es si, asumiendo la inocencia, es o no creíble que existan las evidencias que se presentan.

Si, asumiendo la inocencia del acusado, es creíble que existan las evidencias que se presentan.

Si, por el contrario y asumiendo la inocencia del acusado, las evidencias son de difícil ocurrencia, entonces se considera que la culpabilidad ha sido demostrada.

Traducción a lenguaje estadístico

Si se traduce esto a lenguaje estadístico, se resumiría de la siguiente manera:

Traducción a lenguaje estadístico

Si se traduce esto a lenguaje estadístico, se resumiría de la siguiente manera:

Si $P(\text{Evidencias} \mid \text{El acusado es inocente})$ en menor que un cierto umbral, entonces se considera que la culpabilidad se ha demostrado. Si no, se declara inocente al acusado.

Traducción a lenguaje estadístico

Si se traduce esto a lenguaje estadístico, se resumiría de la siguiente manera:

Si $P(\text{Evidencias} \mid \text{El acusado es inocente})$ en menor que un cierto umbral, entonces se considera que la culpabilidad se ha demostrado. Si no, se declara inocente al acusado.

Este umbral se denomina **significancia** o **significación estadística**.

Traducción a lenguaje estadístico

Si se traduce esto a lenguaje estadístico, se resumiría de la siguiente manera:

Si $P(\text{Evidencias} \mid \text{El acusado es inocente})$ en menor que un cierto umbral, entonces se considera que la culpabilidad se ha demostrado. Si no, se declara inocente al acusado.

Este umbral se denomina **significancia** o **significación estadística**.

Por otro lado, en estadística las evidencias se convierten en funciones calculadas a partir de los datos experimentales. En otras palabras, **estadísticos**. Estos estadísticos se denominan **estadísticos de prueba**.

Traducción a lenguaje estadístico

Si se traduce esto a lenguaje estadístico, se resumiría de la siguiente manera:

Si $P(\text{Evidencias} \mid \text{El acusado es inocente})$ en menor que un cierto umbral, entonces se considera que la culpabilidad se ha demostrado. Si no, se declara inocente al acusado.

Este umbral se denomina **significancia** o **significación estadística**.

Por otro lado, en estadística las evidencias se convierten en funciones calculadas a partir de los datos experimentales. En otras palabras, **estadísticos**. Estos estadísticos se denominan **estadísticos de prueba**.

El valor $P(\text{Evidencias} \mid \text{El acusado es inocente})$ se denomina **p-valor**

La hipótesis nula y la hipótesis alternativa

En estadística, una **hipótesis** es una afirmación sobre una o varias variables aleatorias, en general sobre parámetros como la media o la varianza.

La hipótesis nula y la hipótesis alternativa

En estadística, una **hipótesis** es una afirmación sobre una o varias variables aleatorias, en general sobre parámetros como la media o la varianza.

Así, en inferencia estadística siempre se confrontan 2 hipótesis:

La hipótesis nula y la hipótesis alternativa

En estadística, una **hipótesis** es una afirmación sobre una o varias variables aleatorias, en general sobre parámetros como la media o la varianza.

Así, en inferencia estadística siempre se confrontan 2 hipótesis:

H_0: **Hipótesis nula**. Se presume, es decir, se considera verdadera y no debe demostrarse.

La hipótesis nula y la hipótesis alternativa

En estadística, una **hipótesis** es una afirmación sobre una o varias variables aleatorias, en general sobre parámetros como la media o la varianza.

Así, en inferencia estadística siempre se confrontan 2 hipótesis:

H_0: **Hipótesis nula**. Se presume, es decir, se considera verdadera y no debe demostrarse.

H_1: **Hipótesis alternativa**. No se presume. Debe demostrarse.

La hipótesis nula y la hipótesis alternativa

En estadística, una **hipótesis** es una afirmación sobre una o varias variables aleatorias, en general sobre parámetros como la media o la varianza.

Así, en inferencia estadística siempre se confrontan 2 hipótesis:

H_0: **Hipótesis nula**. Se presume, es decir, se considera verdadera y no debe demostrarse.

H_1: **Hipótesis alternativa**. No se presume. Debe demostrarse.

Con respecto a estas hipótesis, se pueden cometer los siguientes errores:

La hipótesis nula y la hipótesis alternativa

En estadística, una **hipótesis** es una afirmación sobre una o varias variables aleatorias, en general sobre parámetros como la media o la varianza.

Así, en inferencia estadística siempre se confrontan 2 hipótesis:

- H_0: **Hipótesis nula**. Se presume, es decir, se considera verdadera y no debe demostrarse.
- H_1: **Hipótesis alternativa**. No se presume. Debe demostrarse.

Con respecto a estas hipótesis, se pueden cometer los siguientes errores:

- Error tipo I: Rechazar H_0 siendo esta verdadera.

La hipótesis nula y la hipótesis alternativa

En estadística, una **hipótesis** es una afirmación sobre una o varias variables aleatorias, en general sobre parámetros como la media o la varianza.

Así, en inferencia estadística siempre se confrontan 2 hipótesis:

H_0: **Hipótesis nula**. Se presume, es decir, se considera verdadera y no debe demostrarse.

H_1: **Hipótesis alternativa**. No se presume. Debe demostrarse.

Con respecto a estas hipótesis, se pueden cometer los siguientes errores:

Error tipo I: Rechazar H_0 siendo esta verdadera.

Error tipo II: No rechazar H_0 siendo esta falsa.

Funcionamiento general de una prueba de hipótesis I

Por lo anterior, funcionamiento general de una prueba de hipótesis es de la siguiente manera. O bien,

Funcionamiento general de una prueba de hipótesis I

Por lo anterior, funcionamiento general de una prueba de hipótesis es de la siguiente manera. O bien,

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0$$

Funcionamiento general de una prueba de hipótesis I

Por lo anterior, funcionamiento general de una prueba de hipótesis es de la siguiente manera. O bien,

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0$$

o bien,

Funcionamiento general de una prueba de hipótesis I

Por lo anterior, funcionamiento general de una prueba de hipótesis es de la siguiente manera. O bien,

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0$$

o bien,

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta > \theta_0$$

Funcionamiento general de una prueba de hipótesis I

Por lo anterior, funcionamiento general de una prueba de hipótesis es de la siguiente manera. O bien,

$$
\begin{aligned}
H_0 &: \theta = \theta_0 \\
H_1 &: \theta \neq \theta_0
\end{aligned}
$$

o bien,

$$
\begin{aligned}
H_0 &: \theta = \theta_0 \\
H_1 &: \theta > \theta_0
\end{aligned}
$$

o bien,

$$
\begin{aligned}
H_0 &: \theta = \theta_0 \\
H_1 &: \theta < \theta_0
\end{aligned}
$$

Funcionamiento general de una prueba de hipótesis I

Por lo anterior, funcionamiento general de una prueba de hipótesis es de la siguiente manera. O bien,

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0$$

o bien,

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta > \theta_0$$

o bien,

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta < \theta_0$$

La primera se conoce como **prueba bilateral o de dos colas,** mientras que las otras se conocen como **pruebas unilaterales o de una cola.**

Funcionamiento general de una prueba de hipótesis II

En ambos casos se calcula un **estadístico de prueba** $T_0 = T(X_1, X_2, \ldots, x_n)$ con base en los datos, a partir del cual se calcula:

Funcionamiento general de una prueba de hipótesis II

En ambos casos se calcula un **estadístico de prueba** $T_0 = T(X_1, X_2, \ldots, x_n)$ con base en los datos, a partir del cual se calcula:

$$p = P(T > T_0 | H_0)$$

Funcionamiento general de una prueba de hipótesis II

En ambos casos se calcula un **estadístico de prueba** $T_0 = T(X_1, X_2, \ldots, x_n)$ con base en los datos, a partir del cual se calcula:

$$p = P(T > T_0 | H_0)$$

Si $p < \alpha$, se rechaza H_0.

Ejemplos de pruebas de hipótesis

Media con varianza conocida

Procedimiento de prueba para una sola media (varianza conocida)

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha/2} \quad \text{o} \quad z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_{\alpha/2}$$

Si $-z_{\alpha/2} < z < z_{\alpha/2}$, no se rechaza H_0. El rechazo de H_0, desde luego, implica la aceptación de la hipótesis alternativa $\mu \neq \mu_0$. Con esta definición de la región crítica debería quedar claro que habrá α probabilidades de rechazar H_0 (al caer en la región crítica) cuando, en realidad, $\mu = \mu_0$.

Media con varianza conocida

Procedimiento de prueba para una sola media (varianza conocida)	$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha/2} \quad \text{o} \quad z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_{\alpha/2}$$ Si $-z_{\alpha/2} < z < z_{\alpha/2}$, no se rechaza H_0. El rechazo de H_0, desde luego, implica la aceptación de la hipótesis alternativa $\mu \neq \mu_0$. Con esta definición de la región crítica debería quedar claro que habrá α probabilidades de rechazar H_0 (al caer en la región crítica) cuando, en realidad, $\mu = \mu_0$.

Un fabricante de equipo deportivo desarrolló un nuevo sedal para pesca sintético que, según afirma, tiene una resistencia media a la rotura de 8 kilogramos con una desviación estándar de 0.5 kilogramos. Pruebe la hipótesis de que $\mu = 8$ kilogramos contra la alternativa de que $\mu \neq 8$ kilogramos si se prueba una muestra aleatoria de 50 sedales y se encuentra que tienen una resistencia media a la rotura de 7.8 kilogramos. Utilice un nivel de significancia de 0.01.

Media con varianza conocida

1. H_0: $\mu = 8$ kilogramos.
2. H_1: $\mu \neq 8$ kilogramos.
3. $\alpha = 0.01$.
4. Región crítica: $z < -2.575$ y $z > 2.575$, donde $z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$.
5. Cálculos: $\bar{x} = 7.8$ kilogramos, $n = 50$, en consecuencia, $z = \frac{7.8 - 8}{0.5 / \sqrt{50}} = -2.83$.
6. Decisión: rechazar H_0 y concluir que la resistencia promedio a la rotura no es igual a 8 sino que, de hecho, es menor que 8 kilogramos.

Como la prueba en este ejemplo es de dos colas, el valor de P que se desea es el doble del área de la región sombreada en la figura 10.11 a la izquierda de $z = -2.83$. Por lo tanto, si usamos la tabla A.3, tenemos

$$P = P(|Z| > 2.83) = 2P(Z < -2.83) = 0.0046,$$

que nos permite rechazar la hipótesis nula de que $\mu = 8$ kilogramos a un nivel de significancia menor que 0.01.

Media con varianza desconocida

El estadístico t para una prueba sobre una sola media (varianza desconocida)

Para la hipótesis bilateral

$$H_0\colon \mu = \mu_0,$$
$$H_1\colon \mu \neq \mu_0,$$

rechazamos H_0 a un nivel de significancia α cuando el estadístico t calculado

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

excede a $t_{\alpha/2,n-1}$ o es menor que $-t_{\alpha/2,n-1}$.

Media con varianza desconocida

El estadístico t para una prueba sobre una sola media (varianza desconocida)

Para la hipótesis bilateral

$$H_0\colon \mu = \mu_0,$$
$$H_1\colon \mu \neq \mu_0,$$

rechazamos H_0 a un nivel de significancia α cuando el estadístico t calculado

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

excede a $t_{\alpha/2,n-1}$ o es menor que $-t_{\alpha/2,n-1}$.

El Edison Electric Institute publica cifras del número de kilowatts-hora que gastan anualmente varios aparatos electrodomésticos. Se afirma que una aspiradora gasta un promedio de 46 kilowatts-hora al año. Si una muestra aleatoria de 12 hogares, que se incluye en un estudio planeado, indica que las aspiradoras gastan un promedio de 42 kilowatts-hora al año con una desviación estándar de 11.9 kilowatts-hora, ¿esto sugiere que las aspiradoras gastan, en promedio, menos de 46 kilowatts-hora al año a un nivel de significancia de 0.05? Suponga que la población de kilowatts-hora es normal.

Media con varianza desconocida

1. H_0: $\mu = 46$ kilowatts-hora.
2. H_1: $\mu < 46$ kilowatts-hora.
3. $\alpha = 0.05$.
4. Región crítica: $t < -1.796$, donde $t = \dfrac{\bar{x} - \mu_0}{s \,/\, \sqrt{n}}$ con 11 grados de libertad.
5. Cálculos: $\bar{x} = 42$ kilowatts-hora, $s = 11.9$ kilowatts-hora y $n = 12$.
   En consecuencia,

$$t = \frac{42 - 46}{11.9/\sqrt{12}} = -1.16, \qquad P = P(T < -1.16) \approx 0.135.$$

6. Decisión: no rechazar H_0 y concluir que el número promedio de kilowatts-hora que gastan al año las aspiradoras domésticas no es significativamente menor que 46. ▟

Prueba t dos muestras con varianzas desconocidas

Prueba t agrupada de dos muestras

Para la hipótesis bilateral

$$H_0: \mu_1 = \mu_2,$$
$$H_1: \mu_1 \neq \mu_2,$$

rechazamos H_0 al nivel de significancia α cuando el estadístico t calculado

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p \sqrt{1/n_1 + 1/n_2}},$$

donde

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

excede a $t_{\alpha/2,\, n_1+n_2-2}$ o es menor que $-t_{\alpha/2,\, n_1+n_2-2}$.

Prueba t dos muestras con varianzas desconocidas

Se llevó a cabo un experimento para comparar el desgaste por abrasivos de dos diferentes materiales laminados. Se probaron 12 piezas del material 1 exponiendo cada pieza a una máquina para medir el desgaste. Se probaron 10 piezas del material 2 de manera similar. En cada caso se observó la profundidad del desgaste. Las muestras del material 1 revelaron un desgaste promedio (codificado) de 85 unidades con una desviación estándar muestral de 4; en tanto que las muestras del material 2 revelaron un promedio de 81 y una desviación estándar muestral de 5. ¿Podríamos concluir, a un nivel de significancia de 0.05, que el desgaste abrasivo del material 1 excede al del material 2 en más de 2 unidades? Suponga que las poblaciones son aproximadamente normales con varianzas iguales.

Prueba t dos muestras con varianzas desconocidas

Representemos con μ_1 y μ_2 las medias de la población del desgaste abrasivo para el material 1 y el material 2, respectivamente.

1. $H_0: \mu_1 - \mu_2 = 2$.

2. $H_1: \mu_1 - \mu_2 > 2$.

3. $\alpha = 0.05$.

4. Región crítica: $t > 1.725$, donde $t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}}$ con $v = 20$ grados de libertad.

5. Cálculos:

$$\bar{x}_1 = 85, \quad s_1 = 4, \quad n_1 = 12,$$
$$\bar{x}_2 = 81, \quad s_2 = 5, \quad n_2 = 10.$$

En consecuencia,

$$s_p = \sqrt{\frac{(11)(16) + (9)(25)}{12 + 10 - 2}} = 4.478,$$
$$t = \frac{(85 - 81) - 2}{4.478\sqrt{1/12 + 1/10}} = 1.04,$$
$$P = P(T > 1.04) \approx 0.16. \qquad \text{(Véase la tabla A.4).}$$

6. Decisión: no rechazar H_0. No podemos concluir que el desgaste abrasivo del material 1 excede al del material 2 en más de 2 unidades. ▟

Tabla pruebas con medias varianza conocida

H_0	Valor del estadístico de prueba	H_1	Región crítica
$\mu = \mu_0$	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$; σ conocida	$\mu < \mu_0$ $\mu > \mu_0$ $\mu \neq \mu_0$	$z < -z_\alpha$ $z > z_\alpha$ $z < -z_{\alpha/2}$ o $z > z_{\alpha/2}$
$\mu = \mu_0$	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$; $v = n - 1$, σ desconocida	$\mu < \mu_0$ $\mu > \mu_0$ $\mu \neq \mu_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ o $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$z = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$; σ_1 y σ_2 conocidas	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$z < -z_\alpha$ $z > z_\alpha$ $z < -z_{\alpha/2}$ o $z > z_{\alpha/2}$

Tabla pruebas con medias varianza desconocida

$\mu_1 - \mu_2 = d_0$	$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}}$; $v = n_1 + n_2 - 2$, $\sigma_1 = \sigma_2$ pero desconocidas $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ o $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t' = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$; $v = \dfrac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$, $\sigma_1 \neq \sigma_2$ y desconocidas	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$t' < -t_\alpha$ $t' > t_\alpha$ $t' < -t_{\alpha/2}$ o $t' > t_{\alpha/2}$
$\mu_D = d_0$ observaciones pareadas	$t = \dfrac{\bar{d} - d_0}{s_d/\sqrt{n}}$; $v = n - 1$	$\mu_D < d_0$ $\mu_D > d_0$ $\mu_D \neq d_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ o $t > t_{\alpha/2}$

ANOVA

$$\sum_i \sum_j (y_{ij} - \bar{y})^2 = n \sum_i (y_i - \bar{y})^2 + \sum_i \sum_j (y_{ij} - y_i)^2$$

ANOVA

$$\sum_i \sum_j (y_{ij} - \bar{y})^2 = n \sum_i (y_i - \bar{y})^2 + \sum_i \sum_j (y_{ij} - y_i)^2$$

$$SS_{total} = SS_{fact} + SS_{error}$$

ANOVA

$$\sum_i \sum_j (y_{ij} - \bar{y})^2 = n \sum_i (y_i - \bar{y})^2 + \sum_i \sum_j (y_{ij} - y_i)^2$$

$$SS_{total} = SS_{fact} + SS_{error}$$

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F
Intergrupo	SS_{Factores}	t - 1	$T = \frac{SS_{\mathrm{Factores}}}{t-1}$	$F = \frac{T}{E}$
Intragrupo o Error	SS_{Error}	N - t	$E = \frac{SS_{\mathrm{Error}}}{N-t}$	
Total	SS_{Total}	N - 1		

Prueba de independencia χ^2

Una prueba muy útil e interesante es la prueba χ^2 de independencia para variables categóricas.

Prueba de independencia χ^2

Una prueba muy útil e interesante es la prueba χ^2 de independencia para variables categóricas.

Example

En una encuesta electoral realizada a 500 personas, se obtjvo la siguiente distribución en función de sus edades e intención de voto:

Prueba de independencia χ^2

Una prueba muy útil e interesante es la prueba χ^2 de independencia para variables categóricas.

Example

En una encuesta electoral realizada a 500 personas, se obtjvo la siguiente distribución en función de sus edades e intención de voto:

Partido \ Edad	18-35	35-50	50 o más
A	10	40	60
B	15	70	90
C	45	60	35
D	30	30	15

Prueba de independencia χ^2

Una prueba muy útil e interesante es la prueba χ^2 de independencia para variables categóricas.

Example

En una encuesta electoral realizada a 500 personas, se obtjvo la siguiente distribución en función de sus edades e intención de voto:

Partido \ Edad	18-35	35-50	50 o más
A	10	40	60
B	15	70	90
C	45	60	35
D	30	30	15

Determinar si las variables **partido** y **edad** son independientes a una significancia de 5 %.

La anterior es lo que se conoce como **tabla de contingencia**

Prueba de independencia χ^2

Example

Siguiendo con el ejemplo, sumamos por fila y por columna en la tabla de contingencia

Prueba de independencia χ^2

Example

Siguiendo con el ejemplo, sumamos por fila y por columna en la tabla de contingencia

Partido \ Edad	18-35	35-50	50 o más	Total
A	10	40	60	110
B	15	70	90	175
C	45	60	35	140
D	30	30	15	75
Total	100	200	200	500

Prueba de independencia χ^2

Example

Ahora recalculamos los valores esperados asumiendo que las variables son independientes:

Prueba de independencia χ^2

Example

Ahora recalculamos los valores esperados asumiendo que las variables son independientes:

Partido \ Edad	18-35	35-50	50 o más	Total
A	22	44	44	110
B	35	70	70	175
C	28	56	56	140
D	15	30	30	75
Total	100	200	200	500

Prueba de independencia χ^2

Example

Ahora calculamos el estadístico de prueba

Prueba de independencia χ^2

Example

Ahora calculamos el estadístico de prueba

$$J = \sum \frac{(O - E)^2}{E}$$

Prueba de independencia χ^2

Example

Ahora calculamos el estadístico de prueba

$$J = \sum \frac{(O-E)^2}{E}$$

Este estadístico se distribuye χ^2 con $(m-1)(n-1)$ grados de libertad, donde m y n es el número de valores que toman las variables partido y edad respectivamente.

Prueba de independencia χ^2

Example

Ahora calculamos el estadístico de prueba

$$J = \sum \frac{(O-E)^2}{E}$$

Este estadístico se distribuye χ^2 con $(m-1)(n-1)$ grados de libertad, donde m y n es el número de valores que toman las variables partido y edad respectivamente. En este caso, el número de grados de libertad es $3 \cdot 2 = 6$.

Prueba de independencia χ^2

Example

Ahora calculamos el estadístico de prueba

$$J = \sum \frac{(O-E)^2}{E}$$

Este estadístico se distribuye χ^2 con $(m-1)(n-1)$ grados de libertad, donde m y n es el número de valores que toman las variables partido y edad respectivamente. En este caso, el número de grados de libertad es $3 \cdot 2 = 6$.
De acuerdo a esto, podemos plantear una prueba de hipótesis:

Prueba de independencia χ^2

Example

Ahora calculamos el estadístico de prueba

$$J = \sum \frac{(O - E)^2}{E}$$

Este estadístico se distribuye χ^2 con $(m-1)(n-1)$ grados de libertad, donde m y n es el número de valores que toman las variables partido y edad respectivamente. En este caso, el número de grados de libertad es $3 \cdot 2 = 6$.
De acuerdo a esto, podemos plantear una prueba de hipótesis:
H_0: Las variables son independientes

Prueba de independencia χ^2

Example

Ahora calculamos el estadístico de prueba

$$J = \sum \frac{(O-E)^2}{E}$$

Este estadístico se distribuye χ^2 con $(m-1)(n-1)$ grados de libertad, donde m y n es el número de valores que toman las variables partido y edad respectivamente. En este caso, el número de grados de libertad es $3 \cdot 2 = 6$.
De acuerdo a esto, podemos plantear una prueba de hipótesis:
H_0: Las variables son independientes
H_1: Las variables no son independientes

Prueba de independencia χ^2

Example

Ahora calculamos el estadístico de prueba

$$J = \sum \frac{(O-E)^2}{E}$$

Este estadístico se distribuye χ^2 con $(m-1)(n-1)$ grados de libertad, donde m y n es el número de valores que toman las variables partido y edad respectivamente.
En este caso, el número de grados de libertad es $3 \cdot 2 = 6$.
De acuerdo a esto, podemos plantear una prueba de hipótesis:
H_0: Las variables son independientes
H_1: Las variables no son independientes
En este caso, $J = 70{,}86$, lo que lleva a un p-valor mucho menor que $0{,}05$, lo que lleva a rechazar la hipótesis nula, esto es, que el partido y la edad son idependientes.