



Madrid Internet
of Things Institute
Make-build-learn

8 de abril de 2020

Estadística para Data Science

Sesión 6: Estadística Inferencial

Jesús Hernando Corrochano



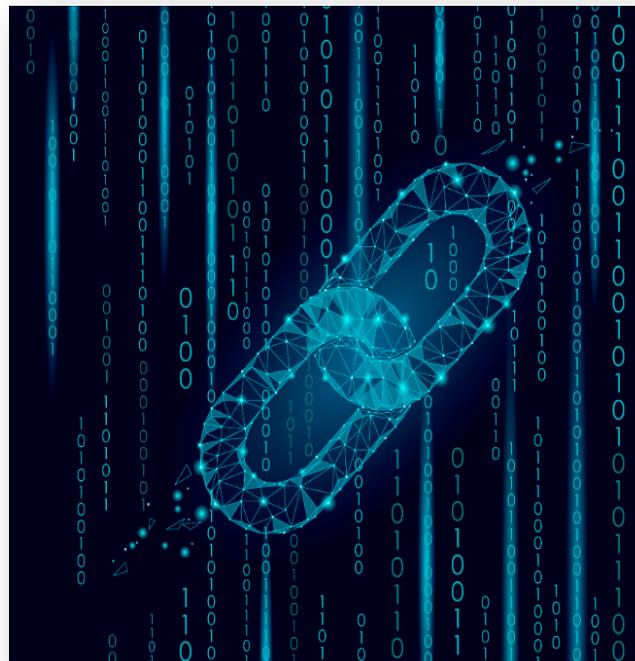
Estadística para Data Science

● Programa



	Sesión 1 13/2	Sesión 2 20/2	Sesión 3 27/2	Sesión 4 5/3	Sesión 5 26/3	Sesión 6 8/4	Sesión 7	Sesión 8
Introducción a la estadística	■							
Introducción a la combinatoria y la probabilidad		■						
Estadística descriptiva			■	■				
Regresión y correlación					■			
Estadística inferencial								■
Probabilidad Total. Teorema de Bayes. Test A/B								■

- 1. Teorema Central del Límite**
- 2. T-Student**
- 3. Contraste de Hipótesis**
- 4. Verosimilitud**
- 5. Anova**



Teorema Central del Límite

● Teorema Central del Límite....

No importa como se distribuya una población. Las medias de las muestras obtenidas de una población siempre se distribuyen según una distribución Normal.

Si la media de la población es μ y su desviación estándar es σ , entonces las medias de sus muestras se distribuirán con una normal:

$$\bar{x} \rightarrow N(\mu ; \frac{\sigma}{\sqrt{n}})$$

n es el tamaño de la muestra o número de elementos que forman la muestra

Aunque lo anterior es cierto, solamente podemos utilizar la tabla de la Distribución Normal si:

- El tamaño de la muestra es grande (mayor o igual a 30) y
- Conocemos la desviación típica de la población

Si esto no es así hay que utilizar la tabla de la distribución t de Student





T – Student: grados de libertad

La Distribución t de Student, tiene por función de densidad

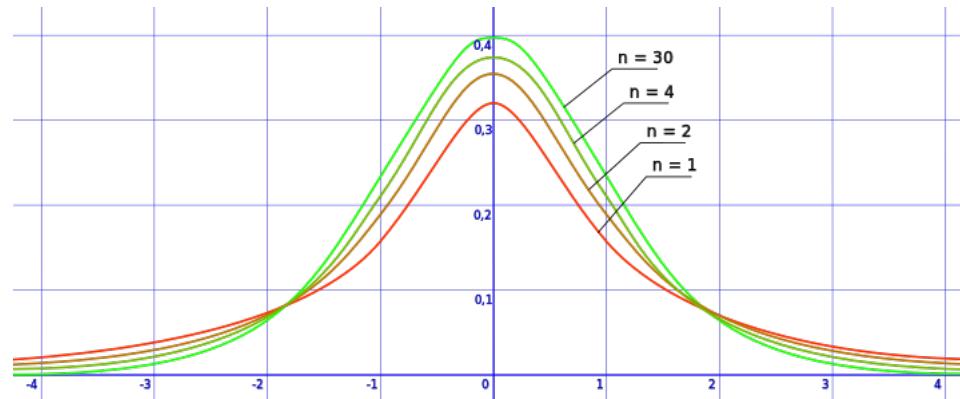
$$t_n(x) = \frac{1}{\sqrt{n\pi}} \cdot \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \cdot \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

Donde el parámetro n de t_n se denomina grados de libertad de la distribución.

La distribución t de Student existe para todos los valores de x reales, y es simétrica respecto al eje y.

La distribución de probabilidad de esta función para valores menores de un x dado, que representamos por:

$$P(t_n < x) = \int_{-\infty}^x t_n(u) du$$

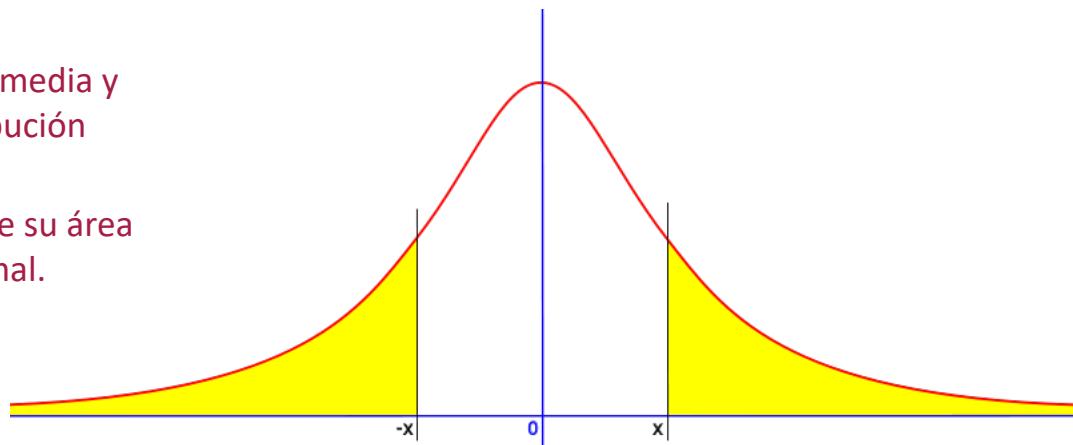


● T - Student

La T-Student es una distribución de tipo Normal, es decir, los valores estimados se van a representar como una distribución de Gauss (conocida como campana de Gauss).

En probabilidad y estadística la **distribución t de student** es una distribución de probabilidad que surge del problema de estimar la media de una población normalmente distribuida cuando el tamaño de la muestra es pequeña.

- La distribución t student es menor en la media y mas alta en los extremos que una distribución normal.
- Tiene proporcionalmente mayor parte de su área en los extremos que la distribución normal.



T – Student: grados de libertad

- Existe una distribución t para cada tamaño de la muestra, por lo que “*Existe una distribución para cada uno de los grados de libertad*”.
- Los grados de libertad son el numero de valores elegidos libremente

TABLA DISTRIBUCIÓN t DE STUDENT

Ejemplos:
Para $n-1 = 10$ grados de libertad
 $P(t > 1.812) = 0.05$
 $P(t < -1.812) = 0.05$

α	0.25	0.2	0.15	0.1	0.05	0.025	0.01	0.005	0.0005
1	1.000	1.3764	1.9626	3.0777	6.3138	12.7062	31.8205	63.6567	636.6192
2	0.8165	1.0607	1.3862	1.8856	2.9200	4.3027	6.9646	9.9248	31.5991
3	0.7649	0.9785	1.2498	1.6377	2.3534	3.1824	4.5407	5.8409	12.9240
4	0.7407	0.9410	1.1896	1.5332	2.3181	2.7764	3.7469	4.6041	8.6103
5	0.7267	0.9195	1.1558	1.4759	2.0150	2.5706	3.3649	4.0321	6.8688
6	0.7176	0.9057	1.1342	1.4394	1.9432	2.4469	3.1427	3.7074	5.9588
7	0.7111	0.8960	1.1192	1.4149	1.8946	2.3646	2.9980	3.4995	5.4079
8	0.7064	0.8889	1.1081	1.3968	1.8955	2.3060	2.8965	3.3554	5.0413
9	0.7027	0.8834	1.0997	1.3830	1.8331	2.2622	2.8214	3.2498	4.7809
10	0.6998	0.8791	1.0931	1.3722	1.8125	2.2281	2.7638	3.1693	4.5869
11	0.6974	0.8755	1.0877	1.3634	1.7959	2.2010	2.7181	3.1058	4.4370

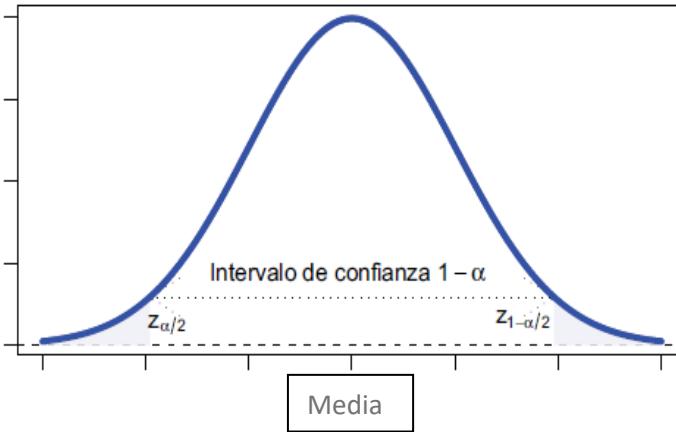


TABLA 2: DISTRIBUCIÓN t DE STUDENT

Puntos de porcentaje de la distribución t

Ejemplo:
Para $\alpha = 0.10$ grados de libertad:
 $P(t > 1.812) = 0.05$
 $P(t < -1.812) = 0.05$

α	0.25	0.2	0.15	0.1	0.05	0.025	0.01	0.005	0.0005
1	1.000	1.376	1.962	3.077	6.313	12.706	31.821	63.656	636.619
2	0.816	1.060	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	0.765	0.978	1.249	1.637	2.353	3.182	4.541	5.841	12.924
4	0.740	0.941	1.189	1.533	2.015	2.570	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	1.943	2.571	3.365	4.032	6.869
6	0.718	0.896	1.134	1.440	1.843	2.447	3.143	3.707	5.407
7	0.711	0.886	1.119	1.415	1.895	2.365	2.998	3.499	5.041
8	0.709	0.889	1.108	1.397	1.860	2.306	2.898	3.355	5.041
9	0.709	0.883	1.100	1.386	1.833	2.282	2.821	3.250	4.781
10	0.705	0.879	1.092	1.374	1.812	2.240	2.744	3.169	4.387
11	0.697	0.876	1.088	1.368	1.796	2.200	2.718	3.106	4.437
12	0.693	0.873	1.083	1.360	1.779	2.180	2.697	3.076	4.391
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.693	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	0.866	1.074	1.343	1.753	2.131	2.602	2.947	4.073
16	0.689	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	0.863	1.069	1.330	1.740	2.110	2.567	2.898	3.965
18	0.688	0.861	1.067	1.327	1.734	2.100	2.552	2.861	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.851	3.893
20	0.687	0.860	1.064	1.328	1.725	2.086	2.523	2.845	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.768
24	0.684	0.857	1.059	1.317	1.708	2.060	2.492	2.787	3.745
25	0.684	0.856	1.058	1.316	1.708	2.050	2.485	2.787	3.725
26	0.684	0.856	1.058	1.315	1.706	2.046	2.479	2.779	3.707
27	0.684	0.855	1.057	1.314	1.703	2.032	2.473	2.771	3.689
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.660
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
31	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.591
32	0.680	0.850	1.049	1.302	1.683	2.019	2.419	2.697	3.580
33	0.677	0.845	1.041	1.299	1.658	1.980	2.358	2.617	3.373
34	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.290

T – Student: grados de libertad

La siguiente tabla representa los ANS alcanzados por un proveedor de desarrollo en 13 meses consecutivos:

92.59
87.39
80.21
92.66
96.24
92.05
98.77
97.81
93.52
97.37
97.63
95.09
99.24

<i>N (número de valores)</i>	13	\rightarrow	<i>N-1</i>	12
<i>Media</i>	93.89			
<hr/>				
<i>Porcentaje (acordado)</i>	$\alpha = 95 \rightarrow$	<i>Grado de libertad</i>		
<i>1 - α</i>	0,05	\rightarrow	<i>1 - $\alpha/2$ = 0.025</i>	
<hr/>				
<i>Máxima</i>	99.24			
<i>Mínima</i>	80.21			
<hr/>				
<i>t $\alpha, n-1$ (tabla T-Student 0,025)</i>	2.1788			

● T – Student: grados de libertad

Una vez calculados estos parámetros, y basándonos en fórmulas y valores de tabla correspondientes a la Distribución T-Student, se calcula el valor de la Desviación estándar, y a partir de esos cálculos, se obtiene el Intervalo de Confianza, objetivo final de este análisis:

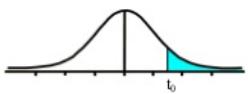
$$\mu = 3.19$$

$$\text{Media} = 93.89$$

Por último, a partir de este dato, se obtiene el Intervalo de Confianza:

$$\text{Intervalo Confianza} = \text{Media} \pm \mu \rightarrow (90.70, 97.08)$$

Tabla t-Student



Grados de libertad	0.25	0.1	0.05	0.025	0.01	0.005
1	1.000	3.0777	6.3137	12.7062	31.8210	63.6559
2	0.8165	1.8856	2.9200	4.3027	6.9645	9.9250
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8408
4	0.7407	1.5332	2.1318	2.7765	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0321
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	2.3646	2.9979	3.4995
8	0.7064	1.3968	1.8595	2.3080	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.6955	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.6938	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.6924	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.6912	1.3400	1.7531	2.1315	2.6025	2.9467
16	0.6901	1.3368	1.7459	2.1199	2.5835	2.9208
17	0.6892	1.3334	1.7396	2.1098	2.5669	2.8982
18	0.6884	1.3304	1.7341	2.1009	2.5524	2.8784
19	0.6876	1.3277	1.7291	2.0930	2.5395	2.8609
20	0.6870	1.3253	1.7247	2.0860	2.5280	2.8453
21	0.6864	1.3232	1.7207	2.0796	2.5176	2.8314
22	0.6858	1.3212	1.7171	2.0739	2.5083	2.8188
23	0.6853	1.3195	1.7139	2.0687	2.4999	2.8073
24	0.6848	1.3178	1.7109	2.0639	2.4922	2.7970
25	0.6844	1.3163	1.7081	2.0595	2.4851	2.7874
26	0.6840	1.3150	1.7056	2.0555	2.4786	2.7787
27	0.6837	1.3137	1.7033	2.0518	2.4727	2.7707
28	0.6834	1.3125	1.7011	2.0484	2.4671	2.7633
29	0.6830	1.3114	1.6991	2.0452	2.4620	2.7564
30	0.6828	1.3104	1.6973	2.0423	2.4573	2.7500
31	0.6825	1.3095	1.6955	2.0395	2.4528	2.7440
32	0.6822	1.3086	1.6939	2.0369	2.4487	2.7385
33	0.6820	1.3077	1.6924	2.0345	2.4448	2.7333
34	0.6818	1.3070	1.6909	2.0322	2.4411	2.7284
35	0.6816	1.3062	1.6896	2.0301	2.4377	2.7238
36	0.6814	1.3055	1.6883	2.0281	2.4345	2.7195
37	0.6812	1.3049	1.6871	2.0262	2.4314	2.7154
38	0.6810	1.3042	1.6860	2.0244	2.4286	2.7116
39	0.6808	1.3036	1.6849	2.0227	2.4258	2.7079
40	0.6807	1.3031	1.6839	2.0211	2.4233	2.7045
41	0.6805	1.3025	1.6829	2.0195	2.4208	2.7012
42	0.6804	1.3020	1.6820	2.0181	2.4185	2.6981
43	0.6802	1.3016	1.6811	2.0167	2.4163	2.6951
44	0.6801	1.3011	1.6802	2.0154	2.4141	2.6923
45	0.6800	1.3007	1.6794	2.0141	2.4121	2.6896
46	0.6799	1.3002	1.6787	2.0129	2.4102	2.6870
47	0.6797	1.2998	1.6779	2.0117	2.4083	2.6846
48	0.6796	1.2994	1.6772	2.0106	2.4066	2.6822
49	0.6795	1.2991	1.6766	2.0096	2.4049	2.6800

Parámetro μ (Intervalo de Confianza)

$$\mu \in \left[\bar{x} - t_{\alpha, n-1} \frac{s}{\sqrt{n}}; \bar{x} + t_{\alpha, n-1} \frac{s}{\sqrt{n}} \right]_{1-\alpha}$$

Cuasivarianza muestral

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

Contraste de Hipótesis

● Contrate de hipótesis

Contraste de hipótesis:

busca aceptar o no una hipótesis considerando una cierta probabilidad expresada mediante un intervalo de confianza y basándose en evidencias obtenidas de una muestra real.

También puede realizarse a partir de una muestra simulada obtenida a partir de números generados aleatoriamente.

La altura media de los españoles era 170 cm. en el año 1970.

Un fabricante de colchones desea contrastar al 95% si esa altura es aún correcta.

Para ello ha realizado una muestra de 10 individuos con el siguiente resultado:

160	168	171	174	175
180	165	195	184	168

La media de esta muestra es 174. También se sabe que esta variable sigue una distribución normal de desviación típica 20.

● Contraste de hipótesis

En la realidad nos encontramos situaciones en las que existe una **idea preconcebida** sobre una característica de la población que estamos estudiando.

Por ejemplo, cuando nos planteamos si los niños de las distintas comunidades españolas tienen la misma altura. Este tipo de circunstancias son las que nos llevan al estudio de la parcela de la Estadística Inferencial conocida como **Contraste de Hipótesis** (también denominado *Test de Hipótesis o Contraste de significación*).

Así diremos que una **hipótesis** no es más que una creencia sobre la población, principalmente sobre alguno de sus parámetros (*media, proporción...*).

Si queremos contrastarla, se debe establecer antes del análisis. El test será la herramienta que nos permitirá extraer

El contraste de hipótesis implica, en cualquier investigación, la existencia de dos teorías o hipótesis implícitas, complementarias, que denominaremos **hipótesis nula (H_0)** e **hipótesis alternativa (H_1)**, que reflejarán esa idea que tenemos a priori y que pretendemos contrastar con la “realidad”.

Los contrastes de hipótesis se realizan:

- Suponiendo a priori que la distribución de la población es conocida.
- Extrayendo una muestra aleatoria de dicha población.
- Si la distribución de la muestra es “diferente” de la distribución de probabilidad que hemos asignado a priori a la población, concluimos que probablemente sea errónea la suposición inicial.

● Contrate de hipótesis

La hipótesis nula H_0 , es la hipótesis de partida (la que contrastamos).

Debe recoger el hecho que queremos someter a prueba.

La hipótesis alternativa H_1 es la que, como su nombre indica, ofrecemos como alternativa a la nula.

Esta hipótesis, representa que se ha producido un cambio con respecto a la situación descrita por la hipótesis nula.

Nos encontraremos con dos situaciones posibles:

- a) La hipótesis nula es cierta
- b) La hipótesis nula es falsa

Esto hace que podamos cometer dos clases de errores diferentes:

1. Aceptar la hipótesis nula cuando ésta es falsa
2. Rechazar la hipótesis nula cuando es cierta

Situaciones	Decisiones	
	Aceptar H_0	Rechazar H_0
H_0 cierta	Decisión correcta	Error de tipo I
H_0 falsa	Error de tipo II	Decisión correcta

Una buena forma de hacer pequeños los dos errores y, por tanto, de mejorar nuestros resultados, es aumentar el tamaño de las muestras que utilizamos.

● Nivel de significación α o nivel de confianza $1-\alpha$.

El nivel de **significación α** o **nivel de confianza $1-\alpha$** , de un contraste es el error máximo de tipo I (rechazar H_0 cuando es cierta) que estamos dispuestos a asumir.

La probabilidad de cometer el error de tipo I es el nivel de significación α , mientras que la probabilidad de cometer el error de tipo II la denotamos por la letra β .

$$\alpha = P[\text{rechazar } H_0 \mid H_0 \text{ es cierta}]$$

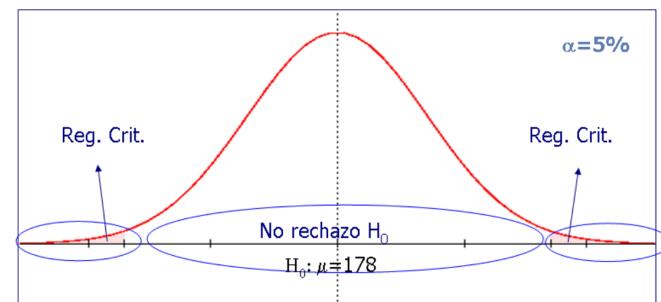
$$\beta = P[\text{no rechazar } H_0 \mid H_0 \text{ es falsa}]$$

El nivel de significación hay que fijarlo cuando se comienza el estudio.

Se suele utilizar el valor estándar de $\alpha=0,05$ (otros niveles utilizados son del orden de 0,1, de 0,01,.....).

Un nivel $\alpha = 0,05$ significa que, aunque la hipótesis nula sea cierta, los datos de cinco de cada cien muestras nos harán rechazarla.

Es decir, aceptamos que podemos rechazar la hipótesis nula de forma equivocada cinco de cada cien veces.



● Entendamos que

Cuanto **menor** sea el valor de α que fijemos, más tendencia tendremos a aceptar la hipótesis nula. El caso extremo sería fijar un nivel de significación 0, de manera que aceptaríamos siempre la hipótesis nula y nunca se daría el error de tipo I, pero en esta situación, el estudio que hacemos no aportaría nada nuevo.

Al tomar un α muy pequeño tendremos que β se puede aproximar a uno.

- Lo ideal a la hora de definir un test es encontrar un compromiso satisfactorio entre α y β (aunque siempre a favor de H_0).
- La potencia de un contraste $(1-\beta)$ es la capacidad de una prueba para detectar una diferencia cuando ésta realmente existe, es decir:

$$\text{Potencia del contraste} = 1-\beta = P[\text{rechazar } H_0 \mid H_0 \text{ es falsa}]$$

Los errores de tipo I y II están relacionados de manera que cuando α decrece β crece. Por tanto no es posible encontrar tests que hagan tan pequeños como queramos ambos errores simultáneamente.

De este modo es siempre necesario privilegiar a una de las hipótesis, de manera que no será rechazada a menos que su falsedad se haga muy evidente.

● ¿Aceptar, rechazar?

Debemos tener claro que:

- Cuando aceptamos la hipótesis nula, no estamos seguros de que sea realmente cierta, ya que no controlamos el error de tipo II (el error que cometemos cuando aceptamos la hipótesis nula y ésta es falsa).
- Cuando rechazamos la hipótesis nula, estamos seguros de que tenemos que rechazarla porque tenemos acotado el error de tipo I (el error que cometemos cuando rechazamos la hipótesis nula y ésta es cierta).

¿“Aceptamos” o “no rechazamos”?

Puesto que cuando aceptamos la hipótesis nula no estamos demasiado seguros, normalmente, en lugar de decir “Aceptamos la hipótesis nula”, decimos “No rechazamos la hipótesis nula”.

● Estadístico de Contraste

Una vez fijadas las hipótesis, así como el error de tipo I que estamos dispuestos a asumir, para decidir si rechazamos la hipótesis nula o no, utilizaremos el llamado **estadístico de contraste**.

Consiste en definir un estadístico T relacionado con la hipótesis que deseamos contrastar.

A continuación, suponiendo que H_0 es verdadera se calcula un intervalo, denominado **intervalo de aceptación de la hipótesis nula** (T_i, T_s), de manera que al calcular sobre la muestra $T = T_c$ el criterio a seguir sea:

Si T_c está en (T_i, T_s) no rechazamos H_0

Si T_c NO está en (T_i, T_s) rechazamos H_0 y aceptamos H_1

● La SIGNIFICACIÓN (*p valor*)

El **p-valor** asociado a una observación del estadístico de contraste es el menor nivel de significación que nos permite rechazar la hipótesis nula.

Cuando el p-valor sea pequeño, indicará que el valor del estadístico de contraste que hemos observado tenía una probabilidad pequeña de salir bajo la hipótesis nula y, por tanto, deberemos rechazar la hipótesis nula.

En cambio, cuando sea grande, indicará que era un valor bastante probable bajo la hipótesis nula y, por tanto, es lógico que aceptemos H_0 .

- Si el p-valor es inferior al nivel de significación α , rechazaremos la hipótesis nula.
- Si el p-valor es superior o igual al nivel de significación α , aceptaremos la hipótesis nula.

El p-valor siempre es conocido después de realizar el contraste de hipótesis.

- Se dice que el contraste es **estadísticamente significativo** si $p\text{-valor} < \alpha$.
- En caso contrario, si $p\text{-valor} > \alpha$ el contraste **no es significativo**

● LA SIGNIFICACIÓN (p-valor)

Podemos describir el procedimiento para plantear y resolver un contraste de hipótesis en seis pasos:

- Paso 1: Fijar las hipótesis nula y alternativa.
- Paso 2: Fijar un nivel de significación.
- Paso 3: Elegir el estadístico del contraste y determinar su distribución.
- Paso 4: Construcción de la región de aceptación
- Paso 5: Calcular el valor que toma el estadístico del contraste para la muestra, así como el p-valor asociado a nuestro estadístico de contraste calculado.
- Paso 6: Aceptación o rechazo de la hipótesis nula, e interpretación de la decisión en el contexto del enunciado del problema. Comparar el p-valor con el nivel de significación y tomar una decisión.

Ejemplo

Un estudio afirma que la media de las alturas de los chicos talaveranos de 18 años es de 178 cm. Visitamos un instituto y escogemos a diez chicos al azar; su altura media es de 171 cm.

¿Podemos pensar que los muchachos de este instituto tienen una altura diferente de la del conjunto de chicos de Talavera?

Sabemos, por ejemplo, que la variable “altura” sigue una distribución normal. Por tanto, una altura es una observación de una variable $N(\mu, \sigma^2)$.

Supondremos también que σ es conocida y es igual a 3. Entonces, cuando decimos que los chicos de 17 años tienen una altura media de 178 cm, en realidad proponemos que la hipótesis nula expresada en términos del parámetro μ es ésta:

$$\begin{aligned} H_0: \mu &= 178 \\ H_1: \mu &\neq 178 \end{aligned}$$

sabemos que si tenemos una muestra de alturas de n chicos escogidos al azar, bajo la hipótesis nula ($\mu=178$) podemos definir la variable:

$$z = \frac{\bar{x} - 178}{\frac{3}{\sqrt{n}}} \quad \text{Que sigue una } N(0,1).$$

Ejemplo

Si la hipótesis nula es cierta, el valor observado z debería estar en la zona en la que la distribución normal estándar concentra una mayor probabilidad, es decir, alrededor del cero.

Si nos sale un valor muy alejado del cero, este valor será poco probable bajo la hipótesis nula, y nos llevará a decidir rechazarla, ya que pensaremos que su aparición no puede ser debida al azar, sino al hecho de que la hipótesis nula debe de ser falsa.

De este modo, utilizaremos la **regla de decisión** siguiente:

Aceptaremos H_0 si $|z| \leq z_{\alpha/2}$

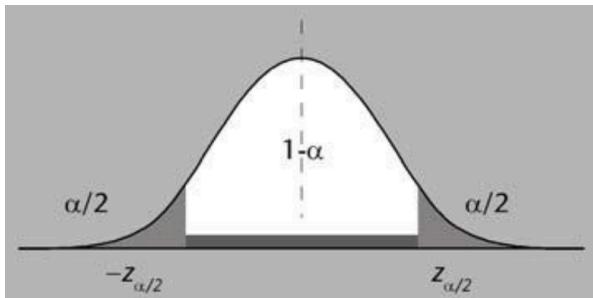
Rechazaremos H_0 si $|z| > z_{\alpha/2}$

Donde $Z_{\alpha/2}$ es el llamado valor crítico

En este caso, bajo la hipótesis nula, sigue una distribución normal estándar.

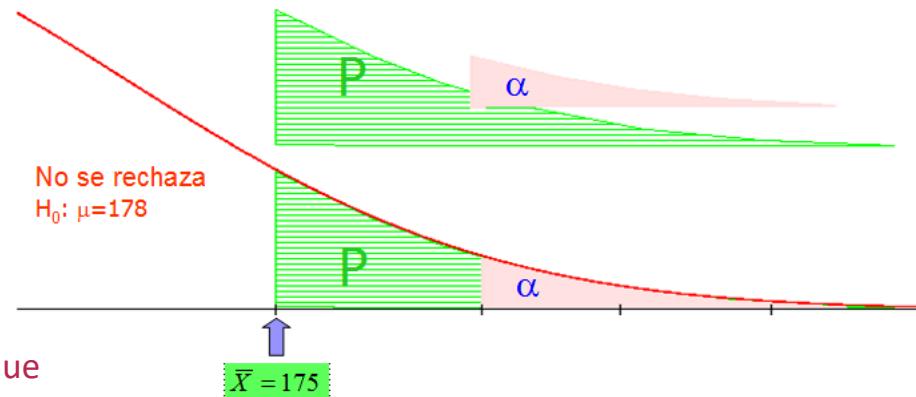
Por ejemplo, para $\alpha=0,05$ encontramos (recordad las tablas de la normal) que $z_{\alpha/2}=1,96$.

Ejemplo



Para determinar el valor crítico $z\alpha/2$, sólo hay que imponer que el error de tipo I (probabilidad de rechazar H_0 cuando es cierta) sea menor o igual que el nivel de significación α , es decir:

$$P(|Z| > z_{\frac{\alpha}{2}}) = P(Z > z_{\frac{\alpha}{2}}) + P(Z < -z_{\frac{\alpha}{2}}) \leq \alpha$$



si el valor obtenido para la altura media extraída de una muestra es de 175 cm, no rechazaríamos nuestra hipótesis nula de partida, ya que la probabilidad de la región crítica que comienza exactamente en ese valor (en color verde) es mayor que la probabilidad α .

En este caso el contraste no es significativo ($p>\alpha$).

Ejemplo

Si nuestro estadístico de contraste observado fuese 1,61 y denotamos por Z una variable aleatoria que tiene una distribución normal estándar, que es la ley del estadístico de contraste bajo la hipótesis nula.

Supongamos, además, que hubiésemos fijado un nivel de significación $\alpha=0,1$.

Así:

Si hacemos el contraste $H_0: \mu = 178$ contra $H_1: \mu \neq 178$, entonces el p-valor es (probabilidad de las dos colas):

$$P(Z > |1,61|) = P(Z > 1,61) + P(Z < -1,61) = 2 \cdot 0,0537 = 0,1074 > \alpha$$

Luego **no rechazamos H_0**

Verosimilitud



● Verosimilitud

Pregunta:

Si tiro una moneda al aire 5 veces y 3 veces me sale cara.....

ENTONCES: ¿Debo pensar que si tiro la moneda una vez, la probabilidad de que salga cara será 0,6?

¿Es correcto este razonamiento?

Razónese la respuesta

$$P(x = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

● Función Verosimilitud

Es la función que proporciona la probabilidad de que ocurra lo que ha ocurrido.

Como en el ejemplo es una binomial, la probabilidad de que salgan 3 caras al tirar la moneda 5 veces será:

$$\mathcal{L}(p) = \binom{5}{3} p^3 (1-p)^{5-3}$$

La \mathcal{L} proviene del nombre en inglés (*likelihood*)

Podemos optimizar esta función calculando logaritmos y derivadas (esto se llama **calculo del estimador de máxima verosimilitud**)
O podemos tantear como hemos hecho antes y el resultado será aproximadamente el mismo.

Si ocurre el suceso 1 y el 2 ... y así hasta el suceso n, la función de verosimilitud conjunta será:

$$\mathcal{L}(p) = \mathcal{L}_1(p) \cdot \mathcal{L}_2(p) \cdot \mathcal{L}_3(p) \cdot \dots \cdot \mathcal{L}_n(p)$$

Habrá que calcular el valor de p. Porque ese valor debe ser único, el mismo para todos los sucesos que han ocurrido

Pensemos....

- 1) Se han realizado tests en tres poblaciones que tienen el mismo número de habitantes y la misma significación como muestra de la población total de la zona afectada.
 - a. En la primera se realizaron 5 tests de los que 2 fueron positivos
 - b. En la segunda se realizaron tests hasta obtener el segundo positivo, lo que ocurrió en el test numero 6
 - c. En la tercera se realizaron 3 tests de los que solamente uno fue positivo.

Supuestos igualmente significativos los tres resultados se desea conocer la prevalencia más verosímil actualmente con una precisión de un decimal (un 10% de diferencia respecto al valor real) (3 puntos)

Anova



El ANOVA es una técnica estadística que nos permite comparar las diferencias de medias de una variable de resultado (dependiente) en dos o más grupos (niveles) de una variable independiente (factor).

Si solo hay dos niveles (por ejemplo, hombre / mujer) de la variable independiente (predictor), los resultados son análogos a la prueba t de Student.

También es cierto que ANOVA es un caso de los modelos de regresión, por lo que a medida que aumente el número de niveles, podría tener más sentido probar uno de esos enfoques.

ANOVA también permite comparaciones de diferencias de medias entre múltiples factores (Factorial o Nway ANOVA) que no abordaremos aquí.

Surge como una generalización del contraste para dos medias de la t de Student, cuando el número de muestras a contrastar es mayor que dos.

Necesitamos poder comparar simultáneamente todas las medias. El test que lo permite es el test **ANOVA (de ANalysis Of VAriance)**. Como su nombre indica, compara varianzas aunque lo que contrastamos sean medias.

Para ello parte de 3 requisitos previos:

Variables cuantitativas
Mas de dos grupos
Compara medias

Independencia: las k muestras son independientes,

Normalidad: $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \dots, k$, y

Homocedasticidad: $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2$.

H_0 = las medias de los grupos son iguales
 H_A =no todas son iguales (alguna es diferente)

Existe distribución normal en cada uno de los grupos
Homogeneidad de varianza de grupos
Los grupos son independientes

Fundamentos del ANOVA (1)

	k grupos					
	1	2	...	i	...	k
	x_{11}	x_{21}	...	x_{i1}	...	x_{k1}
	x_{12}	x_{22}	...	x_{i2}	...	x_{k2}
	\vdots	\vdots	...	\vdots	...	\vdots
	x_{1n_1}	x_{2n_2}	...	x_{in_i}	...	x_{kn_k}
medias	\bar{x}_1	\bar{x}_2	...	\bar{x}_i	...	\bar{x}_k
varianzas	s_1^2	s_2^2	...	s_i^2	...	s_k^2

Fundamentos del ANOVA (2): El ANOVA se basa en la comparación de la variabilidad media que hay entre los grupos con la que hay dentro de los grupos, lo que nos permite dos estimaciones diferentes para 2 cuando disponemos de k muestras de una misma población:

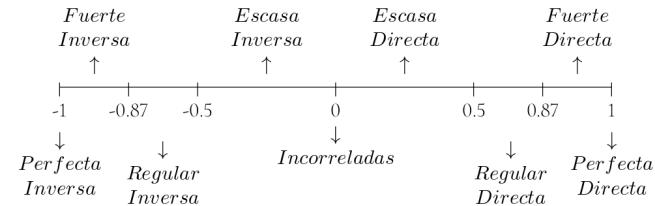
$$\hat{\sigma}^2 = \overline{s^2} = \frac{1}{k} \sum_{i=1}^k s_i^2 \quad (1) \quad y \quad \hat{\sigma}^2 = ns_{\bar{x}}^2 = \frac{n}{k-1} \sum_{i=1}^k (\bar{x}_i - \bar{x})^2 \quad (2)$$

Anova

MUESTRAS	TAMAÑO	VALORES	MEDIAS
1	n_1	$x_{11}, x_{12}, \dots, x_{1n_1}$	$\bar{x}_1 = (\sum_{j=1}^{n_1} x_{1j}) / n_1 = x_1 / n_1$
2	n_2	$x_{21}, x_{22}, \dots, x_{2n_2}$	$\bar{x}_2 = (\sum_{j=1}^{n_2} x_{2j}) / n_2 = x_2 / n_2$
...
k	n_k	$x_{k1}, x_{k2}, \dots, x_{kn_k}$	$\bar{x}_k = (\sum_{j=1}^{n_k} x_{kj}) / n_k = x_k / n_k$
	$N = \sum_{i=1}^k n_i$	$X = \sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij}$	$\bar{X} = X / N$

Realidad de la población		
	H_0	H_1
Conclusión del estudio (Decisión)	H_0	Acierto
	H_1	Error II
		Error I
		Acierto

TABLA DEL ANOVA			
Causa de var.	G.L.	Suma Cuadrática	Media Cuadrática
Entre Grupos	$k-1$	$SCE = \sum_{i=1}^k n_i (\bar{x}_i - \bar{X})^2$	$MCE = SCE / (k-1)$
Dentro Grupos	$N-k$	$SCD = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$	$MCD = SCD / (N-k)$
Total	$N-1$	$SCT = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{X})^2$	



Fundamentos del ANOVA (3): si las k muestras provienen de la misma población todas las medias son iguales, (H_0 es cierta) y tanto (1) como (2) son válidos.

¿Qué ocurre cuando las medias no son iguales?

Si suponemos que: $\mu_i = \mu + \alpha_i$

entonces:

$$ns_x^2 = \hat{\sigma}^2 + \frac{1}{k-1} \sum_{i=1}^k n_i \hat{\alpha}_{i|}^2 \quad (3)$$

Denotamos que (1) describe la variabilidad dentro de los grupos, mientras que (2) y (3) describen la variabilidad entre los grupos.

Test ANOVA (4):

Si la observación j -ésima del grupo i es de la forma

$$X_{ij} = \mu_i + \varepsilon_{ij}, \text{ con } \mu_i = \mu + \alpha_i,$$

las hipótesis

$$\mathbf{H}_0: \alpha_i = 0, \forall i \iff \mu_i = \mu, \forall i,$$

frente a

$$\mathbf{H}_1: \text{algún } \alpha_i \neq 0 \iff \text{las } \mu_i \text{ son distintas,}$$

se contrastan mediante el cociente de varianzas

$$F_0 = \frac{ns_x^2}{s^2} = \frac{\hat{\sigma}^2 + \frac{1}{k-1} \sum_{i=1}^k n_i \hat{\alpha}_i^2}{\hat{\sigma}^2} \quad (4).$$

Todo se reduce a obtener el valor del estadístico (4) que bajo las condiciones iniciales de *independencia, normalidad y homocedasticidad*, se distribuye como una $F_{k-1, n-k}$. La comparación con el valor teórico correspondiente nos dirá si debemos aceptar o rechazar \mathbf{H}_0 .

Comparación de Varianzas:

Si X_1, X_2, \dots, X_m y Y_1, Y_2, \dots, Y_n son muestras independientes de $N(\mu_1, \sigma_1^2)$ y $N(\mu_2, \sigma_2^2)$, respectivamente, un test para comparar la igualdad de varianzas se basa en que el cociente corregido de varianzas muestrales,

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{S_1^2}{S_2^2},$$

se distribuye como una $F_{m-1, n-1}$, una F de Fisher con $m - 1$ gl en el numerador y $n - 1$ gl en el denominador.



Describamos los pasos a realizar

Si rechazamos H_0 , tenemos la prueba o test de Tukey para determinar las medias diferentes

Ejercicio: descríbase el proceso



Calle Rufino González 25
28037 Madrid
+34810527241
www.mioti.es

