t-Student y F-Snedecor

Introducción

La prueba t-Student se utiliza para contrastar hipótesis sobre medias en poblaciones con distribución normal. También proporciona resultados aproximados para los contrastes de medias en muestras suficientemente grandes cuando estas poblaciones no se distribuyen normalmente (aunque en este último caso es preferible realizar una prueba no paramétrica). Para conocer si se puede suponer que los datos siguen una distribución normal, se pueden realizar diversos contrastes llamados de bondad de ajuste, de los cuales el más usado es la prueba de Kolmogorov. A menudo, la prueba de Kolmogorov es referida erróneamente como prueba de Kolmogorov-Smirnov, ya que en realidad esta última, sirve para contrastar si dos poblaciones tienen la misma distribución. Otros tests empleados para la prueba de normalidad son debidos a Saphiro y Wilks.

Existen dos versiones de la prueba t-Student: una que supone que las varianzas poblacionales son iguales y otra versión que no asume esto último. Para decidir si se puede suponer o no la igualdad de varianza en las dos poblaciones, se debe realizar previamente la prueba F-Snedecor de comparación de dos varianzas.

La prueba t-Student fue desarrollada en 1899 por el químico inglés William Sealey Gosset (1876-1937), mientras trabajaba en técnicas de control de calidad para las destilerías Guiness en Dublín . Debido a que en la destilería, su puesto de trabajo no era inicialmente de estadístico y su dedicación debía estar exclusivamente encaminada a mejorar los costes de producción, publicó sus hallazgos anónimamente firmando sus artículos con el nombre de "Student".

William Sealey Gosset (“Student”)   George W. Snedecor

La distribución F es conocida con este nombre gracias al matemático americano George W. Snedecor (1882-1974) quien la bautizó de este modo en honor de R. A. Fisher (1890-1962) que ya la había estudiado anteriormente en 1924.

Las pruebas de bondad de ajuste mencionadas son debidas a Nikolai Vasil’yevich Smirnov (1890-1966), Andrei Nikolaevich Kolmogorov (1903-1987) gran teórico probabilista que fundó las bases de la teoría de la medida en 1929 y finalmente Samuel S. Shapiro (actualmente profesor de matemáticas en los EE.UU) y Martin .B. Wilk (matemático canadiense) que publicaron sus hallazgos en la revista “Biometrika” en 1965.

Fórmulas básicas

En el caso de que se estén estudiando dos variables donde una de ellas es cuantitativa normal considerada como variable respuesta Rta y la otra variable es dicotómica considerada como variable explicativa Exp, se pueden aplicar técnicas de estimación por IC para diferencia de medias, la prueba t-Student para contrastar la diferencias de medias, técnicas de estimación por IC para el cociente de varianzas y la prueba F-Snedecor para igualdad de varianzas. Los IC para diferencia de medias y la prueba t-Student para diferencia de medias tienen expresiones distintas dependiendo si se puede asumir o no la igualdad de varianzas poblaciones (para esto último está la prueba F-Snedecor de igualdad de varianzas). La igualdad de varianzas se conoce como homocedasticidad y la no igualdad de varianzas como heterocedasticidad.

Intervalo de confianza para la diferencia de medias y prueba t-Student para dos medias

Cálculo de los estadísticos descriptivos básicos

Si se denota por n_1 y n_2 a los tamaños muestrales del primer y del segundo grupos, las medias y las desviaciones típicas para los dos grupos son:

$$\bar{x}_1 = \frac{\sum x_{1i}}{n_1}$$

$$\bar{x}_2 = \frac{\sum x_{2i}}{n_2}$$

$$s_1 = \sqrt{\frac{1}{n_1 - 1}\sum\left(x_{1i} - \bar{x}_1\right)^2}$$

$$s_2 = \sqrt{\frac{1}{n_2 - 1}\sum\left(x_{2i} - \bar{x}_2\right)^2}$$

donde x_{1i} indica los valores de la variable Rta para el grupo 1 y x_{2i} indica los valores de la variable Rta para el grupo 2.

Cálculo del IC(1 - a)% para la diferencia de medias suponiendo igualdad de varianzas

Para calcular el IC(1 - α)% para la diferencia de medias se necesita calcular el error estándar de la diferencia de medias que, en el supuesto de igualdad de varianzas, tiene la expresión:

$$EE\left(\bar{x}_1 - \bar{x}_2\right) = \sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

donde s^2 recibe el nombre de varianza conjunta ("pooled variance"), que tiene por expresión:

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

En segundo lugar para calcular el IC deseado se necesita el valor de la t-Student $t_{1-\alpha/2;gl}$ con grados de libertad gl = (n_1 - 1) + (n_2 – 1) = (n_1 + n_2 – 2) , con lo que:

$$IC(1-\alpha)\%\left(\bar{x}_1 - \bar{x}_2\right) = \left[\left(\bar{x}_1 - \bar{x}_2\right) \pm t_{1-\alpha/2,gl} EE\left(\bar{x}_1 - \bar{x}_2\right)\right]$$

proporciona el IC buscado.

Cálculo del IC(1 - a)% para la diferencia de medias suponiendo no igualdad de varianzas

Para calcular el IC(1 - α)% para la diferencia de medias se necesita calcular el error estándar de la diferencia de medias que, en el supuesto de no igualdad de varianzas, tiene la expresión:

$$EE\left(\bar{x}_1 - \bar{x}_2\right) = \sqrt{EE\left(\bar{x}_1\right)^2 + EE\left(\bar{x}_2\right)^2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

En segundo lugar, para calcular el IC deseado se necesita el valor de la t-Student $t_{1-\alpha/2;gl}$ con grados de libertad gl dados por la siguiente expresión, llamada de Sattherwaite:

$$gl = \frac{\left[EE\left(\bar{x}_1 - \bar{x}_2\right)\right]^4}{\frac{1}{n_1 - 1}\left[EE\left(\bar{x}_1\right)\right]^4 + \frac{1}{n_2 - 1}\left[EE\left(\bar{x}_2\right)\right]^4}$$

con lo que:

$$IC(1-\alpha)\%\left(\bar{x}_1 - \bar{x}_2\right) = \left[\left(\bar{x}_1 - \bar{x}_2\right) \pm t_{1-\alpha/2,gl} EE\left(\bar{x}_1 - \bar{x}_2\right)\right]$$

proporciona el IC buscado.

Cálculo de la prueba t-Student para la diferencia de medias suponiendo igualdad de varianzas

Para llevar a cabo el contraste:

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 \neq 0$$

suponiendo igualdad de varianzas poblacionales, se construye el estadístico de contraste experimental t dado por:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{EE(\bar{x}_1 - \bar{x}_2)} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

que bajo la hipótesis nula sigue una distribución t-Student con grados de libertad gl = $(n_1 - 1) + (n_2 - 1) = (n_1 + n_2 - 2)$.

Cálculo de la prueba t-Student para la diferencia de medias suponiendo no igualdad de varianzas

Para llevar a cabo el contraste:

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 \neq 0$$

suponiendo no igualdad de varianzas poblacionales, se construye el estadístico de contraste experimental t dado por:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{EE(\bar{x}_1 - \bar{x}_2)} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

que bajo la hipótesis nula sigue una distribución t-Student con grados de libertad gl dados por:

$$gl = \frac{\left[EE\left(\bar{x}_1 - \bar{x}_2\right)\right]^4}{\frac{1}{n_1 - 1}\left[EE\left(\bar{x}_1\right)\right]^4 + \frac{1}{n_2 - 1}\left[EE\left(\bar{x}_2\right)\right]^4}$$

que recibe el nombre de grados de libertad de Sattherthwaite.

Intervalo de confianza para el cociente de varianzas y prueba F-Snedecor para dos varianzas

Cálculo del IC(1 - a)% para el cociente de varianzas

La expresión para calcular el IC(1 - α)% para el cociente de varianzas es:

$$IC95\%\left(\frac{\sigma_1^2}{\sigma_2^2}\right)=\left(\frac{s_1^2 / s_2^2}{F_{1-\alpha/2;gln;gld}}\,;\,\frac{s_1^2}{s_2^2}F_{1-\alpha/2;gld;gln}\right)$$

donde: $F_{1-\alpha/2;\ gln;\ gld}$ se calcula a partir de una F-Snedecor siendo gln los grados de libertad del numerador, que se calculan como el tamaño muestral del grupo con mayor varianza muestral menos uno, y gld los grados de libertad del denominador que se calculan como el tamaño muestral del grupo con menor varianza muestral menos uno.

Cálculo de la prueba F-Snedecor para la igualdad de varianzas

Para llevar a cabo el contraste:

$$H_0: \sigma_1 - \sigma_2 = 0$$

$$H_1: \sigma_1 - \sigma_2 \neq 0$$

mediante la prueba F-Snedecor de comparación de varianzas se construye el estadístico de contraste experimental F dado por:

$$F=\frac{\text{máx}\left\{s_1^2\,;\,s_2^2\right\}}{\text{mín}\left\{s_1^2\,;\,s_2^2\right\}}$$

que bajo la hipótesis nula sigue una distribución F-Snedecor siendo gln los grados de libertad del numerador y gld los grados de libertad del denominador. En el caso de no poder rechazar la hipótesis nula (p-valor > 0.05) se considera que las dos varianzas son iguales (homogéneas).

Ejemplos

Intervalo de confianza para la diferencia de medias y prueba t-Student para dos medias

Se tienen los siguientes datos experimentales correspondientes a 17 individuos de los que se ha recogido el valor que presentan en dos variables, una de ellas cuantitativa con distribución normal considerada como variable respuesta (Rta), y la otra variable dicotómica considerada como variable explicativa (Exp). Los datos se presentan de forma que en las filas hay varios individuos para facilitar la lectura:

Rta	Exp	Rta	Exp
15	1	16	2
15	1	25	2
25	1	28	2
25	1	28	2
25	1	28	2
33	1	28	2
43	1	35	2
15	2	43	2
16	2		

Calcular un intervalo de confianza al 90% para la diferencia de medias asumiendo igualdad de varianzas y no asumiendo la igualdad de éstas y realizar el siguiente contraste:

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 \neq 0$$

mediante la prueba t-Student para dos medias en los dos supuestos de igualdad y no igualdad de varianzas.

Cálculo de los estadísticos descriptivos básicos

Para los datos del ejemplo se tiene que los tamaños muestrales son: $n_1 = 7$ y $n_2 = 10$. Las medias y las desviaciones típicas para los dos grupos son:

$$\bar{x}_1 = \frac{\sum x_{1i}}{n_1} = \frac{181}{7} = 25.8571$$

$$\bar{x}_2 = \frac{\sum x_{2i}}{n_2} = \frac{262}{10} = 26.2000$$

$$s_1 = \sqrt{\frac{1}{n_1 - 1}\sum \left(x_{1i} - \bar{x}_1\right)^2} = 9.8561$$

$$s_2 = \sqrt{\frac{1}{n_2 - 1}\sum \left(x_{2i} - \bar{x}_2\right)^2} = 8.8669$$

donde x_{1i} indica los valores de la variable Rta para el grupo 1 y x_{2i} indica los valores de la variable Rta para el grupo 2.

Cálculo del IC90% para la diferencia de medias suponiendo igualdad de varianzas

Para calcular el IC90% para la diferencia de medias se necesita calcular el error estándar de la diferencia de medias que, en el supuesto de igualdad de varianzas, tiene la expresión:

$$EE\left(\bar{x}_1-\bar{x}_2\right)=\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

donde s^2 recibe el nombre de varianza conjunta ("pooled variance"), que tiene por expresión:

$$s^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}=86.0305$$

con lo que:

$$EE\left(\bar{x}_1-\bar{x}_2\right)=\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}=4.5709$$

En segundo lugar para calcular el IC deseado se necesita el valor de la t-Student $t_{1-\alpha/2;gl}$ para $\alpha = 0.10$ (confianza del 90%) y con grados de libertad gl = $(n_1 - 1) + (n_2 - 1) = (n_1 + n_2 - 2) = 15$, que resulta ser $t_{1-\alpha/2;gl} = 1.7531$, con lo que el intervalo de confianza buscado es:

$$IC90\%\left(\bar{x}_1-\bar{x}_2\right)=\left[\left(\bar{x}_1-\bar{x}_2\right)\pm t_{1-\alpha/2,gl}EE\left(\bar{x}_1-\bar{x}_2\right)\right]$$
$$=\left[-0.3429\pm 1.7531\cdot 4.5709\right]=\left[-8.3559\,;7.6702\right]$$

que cubre al valor de cero para la diferencia de medias poblacionales de los dos grupos.

Cálculo del IC90% para la diferencia de medias suponiendo no igualdad de varianzas

Para calcular el IC90% para la diferencia de medias se necesita calcular el error estándar de la diferencia de medias que, en el supuesto de no igualdad de varianzas, tiene la expresión:

$$EE\left(\bar{x}_1-\bar{x}_2\right)=\sqrt{EE\left(\bar{x}_1\right)^2+EE\left(\bar{x}_2\right)^2}=\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}=\sqrt{3.7253^2+2.8040^2}=4.6626$$

En segundo lugar para calcular el IC deseado se necesita el valor de la t-Student $t_{1-\alpha/2;gl}$ para $\alpha = 0.10$ (confianza del 90%) y con grados de libertad gl dados por la siguiente expresión:

$$gl=\frac{\left[EE\left(\bar{x}_1-\bar{x}_2\right)\right]^4}{\frac{1}{n_1-1}\left[EE\left(\bar{x}_1\right)\right]^4+\frac{1}{n_2-1}\left[EE\left(\bar{x}_2\right)\right]^4}=12.1290$$

que resulta ser $t_{1-\alpha/2;gl} = 1.7807$, con lo que el intervalo de confianza buscado es:

$$IC90\%\left(\bar{x}_1-\bar{x}_2\right)=\left[\left(\bar{x}_1-\bar{x}_2\right)\pm t_{1-\alpha/2,gl}EE\left(\bar{x}_1-\bar{x}_2\right)\right]$$
$$=\left[-0.3429\pm 1.7807\cdot 4.6626\right]=\left[-8.6456\,;7.9599\right]$$

que cubre al valor de cero para la diferencia de medias poblacionales de los dos grupos.

Cálculo de la prueba t-Student para la diferencia de medias suponiendo igualdad de varianzas

Para llevar a cabo el contraste requerido se construye el estadístico de contraste experimental t dado por:

$$t=\frac{\bar{x}_1-\bar{x}_2}{EE\left(\bar{x}_1-\bar{x}_2\right)}=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}=-0.0750$$

que bajo la hipótesis nula sigue una distribución t-Student con grados de libertad gl = $(n_1 - 1) + (n_2 - 1) = (n_1 + n_2 - 2) = 15$, que tiene asociado un p-valor de 0.9412.

Cálculo de la prueba t-Student para la diferencia de medias suponiendo no igualdad de varianzas

Para llevar a cabo el contraste requerido se construye el estadístico de contraste experimental t dado por:

$$t_{exp} = \frac{\overline{x}_1 - \overline{x}_2}{EE\left(\overline{x}_1 - \overline{x}_2\right)} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = -0.0735$$

que bajo la hipótesis nula sigue una distribución t-Student con grados de libertad gl dados por:

$$gl = \frac{\left[EE\left(\overline{x}_1 - \overline{x}_2\right)\right]^4}{\frac{1}{n_1 - 1}\left[EE\left(\overline{x}_1\right)\right]^4 + \frac{1}{n_2 - 1}\left[EE\left(\overline{x}_2\right)\right]^4} = 12.1290$$

que tiene asociado un p-valor de 0.9426.

Intervalo de confianza para el cociente de varianzas y prueba F-Snedecor para dos varianzas

Para los datos experimentales anteriores, calcular un intervalo de confianza al 95% para el cociente de varianzas y realizar el siguiente contraste:

$$H_0: \sigma_1 - \sigma_2 = 0$$
$$H_1: \sigma_1 - \sigma_2 \neq 0$$

mediante la prueba F-Snedecor de comparación de varianzas.

Cálculo del IC95% para el cociente de varianzas

La expresión para calcular el IC95% para el cociente de varianzas es:

$$IC95\%\left(\frac{\sigma_1^2}{\sigma_2^2}\right) = \left(\frac{s_1^2 / s_2^2}{F_{1-\alpha/2;gln;gld}}; \frac{s_1^2}{s_2^2} F_{1-\alpha/2;gld;gln}\right)$$

donde: gln son los grados de libertad del numerador que se calculan como el tamaño muestral del grupo con mayor varianza muestral menos uno, gld son los grados de libertad del denominador que se calculan como el tamaño muestral del grupo con menor varianza muestral menos uno, $F_{1-\alpha/2;\, gln;\, gld}$ es 4.3197 para $\alpha = 0.05$, gln = 6, gld = 9 y $F_{1-\alpha/2;\, gld;\, gln}$ es 5.5234 para $\alpha = 0.05$, gld = 9, gln = 6, con lo que el intervalo:

$$IC95\%\left(\frac{\sigma_1^2}{\sigma_2^2}\right) = [0.2860; 6.8245]$$

proporciona el intervalo de confianza buscado.

Cálculo de la prueba F-Snedecor para la igualdad de varianzas

Para llevar a cabo el contraste requerido se construye el estadístico de contraste experimental F dado por:

$$F = \frac{\text{máx}\left\{s_1^2; s_2^2\right\}}{\text{mín}\left\{s_1^2; s_2^2\right\}} = 1.2356$$

que bajo la hipótesis nula sigue una distribución F-Snedecor con gln los grados de libertad del numerador = 6 y gld los grados de libertad del denominador = 9, que tiene asociado un p-valor de 0.7440.