Contrastes de hipótesis para la comparación de dos poblaciones

Recordemos algunos de los contrastes más habituales para la comparación de dos poblaciones.

1. Comparación de dos poblaciones normales: contraste de la *t* de Student.

El test de la *t* se utiliza para contrastar hipótesis sobre dos medias poblacionales cuando se asume que la distribución de estas poblaciones es normal. Este test tiene varias versiones, dependiendo de si las observaciones recogidas de ambas poblaciones son independientes o no, y en el caso de independencia dependiendo de si la varianza de ambas poblaciones se supone igual o no.

El test *t* para muestras independientes se usa para contrastar la hipótesis nula $H_0 : \mu_1 = \mu_2$ de igualdad de medias de dos poblaciones normales Y_1 e Y_2 cuando se tiene una muestra de observaciones $(y_{11},\ldots,y_{1n_1})$ y $(y_{21},\ldots,y_{2n_2})$ de cada población. Para poder aplicar este contraste, los elementos de una población no deben estar emparejados de algún modo con los de la otra población, y los elementos dentro de cada grupo no deben estar relacionados con (esto es, son independientes de) otros elementos de ese grupo. Se asume además que ambas poblaciones siguen una distribución normal y que ambas tienen la misma varianza o desviación típica, aunque esta es desconocida.

El estadístico de contraste de este test es esencialmente una diferencia estandarizada de las dos medias muestrales, esto es,

$$t_0 = \frac{\overline{y}_1 - \overline{y}_2}{s\sqrt{1/n_1 + 1/n_2}}$$

donde $\overline{y}_1$ e $\overline{y}_2$ son las medias muestrales de cada grupo. La desviación típica común se estima combinando la desviación típica de ambas muestras mediante la expresión

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

Bajo la hipótesis nula H_0 el estadístico t_0 sigue una distribución *t* de Student con n_1+n_2-2 grados de libertad. Un intervalo de confianza $100(1-\alpha)\%$ para la diferencia de medias, útil para dar un rango plausible de esa diferencia, viene dado por

$$\overline{y}_1 - \overline{y}_2 \pm t_{\alpha/2,n_1+n_2-2}\, s\sqrt{1/n_1 + 1/n_2}$$

donde $t_{\alpha/2,n_1+n_2-2}$ denota el percentil $100(1-\alpha/2)$ de la distribución *t* con n_1+n_2-2 grados de libertad, esto es, el valor de la distribución tal que $P(t_{n_1+n_2-2} > t_{\alpha/2,n_1+n_2-2}) = \alpha/2$.

Cuando se sospecha que las dos poblaciones tienen varianzas diferentes, se emplea una versión modificada del contraste anterior, que se conoce como contraste de Welch y se basa en el estadístico de contraste

$$t_0 = \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

En este caso, el estadístico t_0 sigue una distribución *t* con *v* grados de libertad, donde

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_2^4}{n_2^2(n_2 - 1)}}$$

Cuando cada individuo de una muestra está relacionado con otro individuo de la otra muestra, o cuando cada individuo de la muestra se observa dos veces (por ejemplo, antes y después de llevar a cabo un tratamiento), se aplica el llamado contraste *t* para muestras pareadas. Este contraste se basa en el cálculo de las diferencias $d_i = y_{1i} - y_{2i}$ entre las observaciones de los *i*-ésimos elementos emparejados de cada muestra, *i* = 1,…,*n*. Estas diferencias d_i siguen una distribución normal con media μ, y en este caso la hipótesis nula es $H_0 : \mu = 0$, esto es, que la media de las diferencias es 0. El estadístico de contraste que se emplea viene dado por

$$t_0 = \frac{\overline{d}}{s/\sqrt{n}}$$

donde $\overline{d}$ es la media muestral de las diferencias entre observaciones emparejadas y *s* es su desviación típica. Bajo la hipótesis nula anterior, el estadístico t_0 sigue una distribución *t* con *n* – 1 grados de libertad. Un intervalo de confianza 100(1 – α)% para la media poblacional μ viene dado por

$$\overline{d} \pm t_{\alpha/2,n-1}\, s/\sqrt{n}$$

2. Contrastes no paramétricos para la comparación de dos poblaciones.

Uno de los supuestos de los contrastes *t* anteriores es que los datos se distribuyen normalmente, esto es, son unimodales y simétricos. Cuando estos supuestos se incumplen de manera pronunciada, puede ser aconsejable usar los análogos no paramétricos del test *t*, conocidos como el contraste de rangos de Wilcoxon Mann-Whitney y el contraste de rangos con signo de Wilcoxon. Básicamente, estos contrastes descartan los valores concretos observados y solo retienen el ranking u ordenación de las observaciones.

Para dos muestras independientes, el contraste de rangos de Wilcoxon Mann-Whitney procede ordenando conjuntamente las observaciones de ambos grupos de mayor a menor, como si procedieran de la misma muestra. El estadístico de contraste se obtiene a partir de la suma de los rankings de una de las muestras, típicamente la muestra con una suma de rankings más baja. Así, si la suma de rankings de la primera muestra fuera menor que la de la segunda, el estadístico de contraste vendría dado por

$$U_1 = R_1 - \frac{n_1(n_1 + 1)}{2}$$

donde R_1 es la suma de los rankings de las observaciones de la primera muestra y n_1 es el número de observaciones en esa muestra. La hipótesis nula en este caso es que ambas poblaciones tienen una distribución idéntica, mientras que la hipótesis alternativa es que estas distribuciones difieren en términos de localización, esto es, tienen medianas diferentes. Bajo la hipótesis nula, el estadístico *U* se distribuye de manera aproximadamente normal, lo que permite establecer límites de significación estadística del modo habitual.

Para muestras pareadas se utiliza el contraste de rangos con signo de Wilcoxon, que procede ordenando las diferencias $d_i = y_{1i} - y_{2i}$, i = 1,…,n, por su valor absoluto $|d_i|$. El estadístico de contraste se obtiene entonces a partir de la suma de los rankings de las diferencias positivas $d_i > 0$.

3. Contraste de correlación de Pearson.

Al calcular el coeficiente de correlación de Pearson para dos variables de una muestra de tamaño n, de cara a evaluar si las correspondientes poblaciones se relacionan linealmente suele ser necesario evaluar la significatividad de la correlación obtenida, esto es, contrastar la hipótesis nula $H_0 : \rho = 0$, donde ρ denota el coeficiente de correlación poblacional entre ambas distribuciones. Este contraste se realiza mediante el estadístico

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

donde r denota el coeficiente de correlación muestral. Cuando ambas poblaciones son normales, bajo la hipótesis nula el estadístico t_0 se distribuye como una t de Student con n – 2 grados de libertad.

4. Contraste chi-cuadrado de independencia en tablas de contingencia.

Cuando la muestra está compuesta de n observaciones en dos variables categóricas, es posible representar esta información mediante una tabla de contingencia, que registra el número de observaciones que presentan cada posible combinación de una categoría o nivel de la primera variable y una categoría de la segunda variable. En este contexto, suele ser de interés contrastar la hipótesis nula de que ambas variables son independientes, lo que equivale a que todas las distribuciones condicionadas de una de las variables dado cualquier nivel de la otra han de ser idénticas, o similarmente que la distribución conjunta de ambas variables puede obtenerse como producto de sus distribuciones marginales.

Sea r el número de niveles de la primera variable y c el de la segunda. Bajo la citada hipótesis nula de independencia, la frecuencia esperada E_{jk} de la combinación de las categorías j y k (j = 1,…,r, k = 1,…,c) puede estimarse como $E_{jk} = n_{j\cdot} n_{\cdot k} / n$, donde $n_{j\cdot}$ es el total marginal de la categoría j de la primera variable y $n_{\cdot k}$ el total marginal de la categoría k de la segunda variable. Entonces, el estadístico de contraste para evaluar la hipótesis de independencia procede comparando estas frecuencias esperadas con las observadas mediante la expresión

$$X_0^2 = \sum_{j=1}^{r} \sum_{k=1}^{c} \frac{(n_{jk} - E_{jk})^2}{E_{jk}}$$

donde n_{jk} denota la frecuencia observada de la combinación de las categorías j y k de ambas variables. Bajo la hipótesis nula, el estadístico X_0^2 se distribuye asintóticamente como una $\chi^2_{(r-1)(c-1)}$.

5. Contraste de McNemar.

El anterior test de la chi-cuadrado se basa en el supuesto de que las observaciones de ambas variables se recogen de forma independiente. Sin embargo, es frecuente que datos categóricos de ese tipo se obtengan de manera pareada, como por ejemplo dos observaciones consecutivas

de la misma variable en los mismos individuos. Para este tipo de datos pareados categóricos, el procedimiento adecuado para contrastar la igualdad de ambas poblaciones es el contraste de McNemar. Aquí supondremos que la variable que se estudia es binaria, observándose la presencia o la ausencia de una determinada característica en los individuos muestreados. Sea *a* el número de individuos para el que la característica de interés estaba presente en la primera medición y ausente en la segunda, y sea *b* el número de observaciones con la característica ausente en la primera medición y presente en la segunda. Bajo la hipótesis nula de que ambas poblaciones tienen una misma probabilidad de presentar la característica de interés, el estadístico de contraste

$$X_0^2 = \frac{(a-b)^2}{a+b}$$

sigue una distribución chi-cuadrado con un único grado de libertad.