Contrastes paramétricos con una muestra

Contrastes sobre la media

- Sea $X_1, X_2, \ldots, X_n$ una m.a.s. extraída de una población normal X con media desconocida μ. Se desea contrastar:

$$H_0 : \mu = \mu_0$$

- Si H_0 es cierta, $X \in N(\mu_0, \sigma)$, de donde

$$D_1 = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n} \quad \in \quad N(0,1)$$
$$D_2 = \frac{\overline{X} - \mu_0}{\widehat{S}}\sqrt{n} \quad \in \quad t_{n-1}$$

siendo $\widehat{S}$ la desviación típica muestral corregida.

Ejemplo

Se cree que en la actualidad la longitud media del antebrazo de los varones adultos de una determinada población supera los $45'5$ centímetros. Seleccionada al azar una muestra de 15 varones adultos de la citada población, se obtuvieron las siguientes longitudes (en centímetros) de sus antebrazos: $43'9$, $46'7$, $53'1$, $42'7$, $47'5$, $52'1$, $45'5$, $51'8$, $46'5$, $52'1$, $48'3$, $44'5$, 46, $43'4$ y $47'8$.
Interesa contrastar la hipótesis $H_0: \mu = 45'5$ frente a $H_1: \mu > 45'5$, donde μ denota la longitud media del antebrazo.

Ejemplo

Supuesta la normalidad de la variable X = *longitud del antebrazo* y la independencia de las observaciones muestrales, el p-valor para el contraste anterior es:

$$p = P\left(t_{14} > \frac{\overline{X} - 45'5}{\widehat{S}}\sqrt{15}\right)$$

De las 15 observaciones muestrales se obtiene $\overline{X} = 47'447$ y $\widehat{S} = 3'403$, de modo que $p = P\left(t_{14} > 2'2159\right) = 0'0219$. Obsérvese que el contraste es significativo a un 5% pero no a un 1%. Quizá sería apropiado aumentar el tamaño de la muestra.

Estadística I

Mario Francisco

Introducción

Hipótesis estadística

Planteamiento y método

Tipos de error

Criterios de decisión

Etapas en la resolución de un contraste

Nivel crítico o p-valor

Potencia de un contraste

Contrastes paramétricos con una

Contrastes sobre la media

- En ausencia de la normalidad de X, el T.C.L. garantiza que para un tamaño muestral suficientemente grande ($n > 30$), la distribución de $\overline{X}$ también puede aproximarse por una $N\left(\mu, \sigma/\sqrt{n}\right)$.
- Si σ es desconocida, debemos acudir al estadístico de contraste D_2, que, en este nuevo contexto se considera la aproximación

$$\overline{X} \in N\left(\mu_0, \widehat{S}/\sqrt{n}\right)$$

- si $n \leq 30$ habrá que acudir entonces a contrastes no paramétricos, como el contraste de los signos o el de los rangos con signo de Wilcoxon.

Contrastes sobre la media. Determinación del tamaño muestral

- Supongamos que, a efectos prácticos, sea importante que el contraste, realizado a un nivel α, resulte significativo cuando el valor real de μ sea $\mu_0 + \varepsilon \in H_1$, $\varepsilon \in \mathbb{R}$. ¿Qué tamaño deberá tener una muestra para que ello ocurra con una potencia prefijada de antemano $1 - \beta(\mu_0 + \varepsilon)$?
- Consideraremos en primer lugar un contraste unilateral de la forma $H_1 : \mu > \mu_0$ y supondremos que se conoce σ.
- Si H_0 es cierta $D_1 \in N(0, 1)$, pero si es H_1 la correcta, y en concreto $\mu = \mu_0 + \varepsilon$, $\varepsilon > 0$, entonces $D_1 \in N\left(\frac{\varepsilon}{\sigma}\sqrt{n}, 1\right)$.

Contrastes sobre la media. Determinación del tamaño muestral

- Fijado α, la probabilidad de error de tipo II es

$$\begin{aligned}
\beta(\mu_0+\varepsilon) &= P\left(\text{“no rechazar } H_0\text{” } / \text{ “}H_1 \text{ es cierta”}\right) \\
&= P\left(D_1 \leq z_\alpha \Big/ D_1 \in N\left(\frac{\varepsilon}{\sigma}\sqrt{n}, 1\right)\right) \\
&= P\left(Z \leq -\frac{\varepsilon}{\sigma}\sqrt{n} + z_\alpha \Big/ Z \in N(0,1)\right)
\end{aligned}$$

- Equivalentemente,

$$-\frac{\varepsilon}{\sigma}\sqrt{n} + z_\alpha = z_{1-\beta(\mu_0+\varepsilon)} = -z_{\beta(\mu_0+\varepsilon)}$$

Contrastes sobre la media. Determinación del tamaño muestral

- Denotando $\beta = \beta(\mu_0 + \varepsilon)$, se deduce de la expresión anterior

$$n = \left(\frac{z_\alpha + z_\beta}{\varepsilon} \right)^2 \sigma^2$$

que especifica el tamaño muestral como función del nivel de significación y del error de tipo II cuando $\mu = \mu_0 + \varepsilon$.

Determinación del tamaño muestral. Comentarios

1. El valor de n anterior debe interpretarse como el tamaño muestral mínimo para que un $100(1-\beta)\%$ de las veces que realicemos el contraste a un nivel de significación α, detectemos que la verdadera media dista ε de μ_0.
2. Si el contraste propuesto es unilateral a la izquierda ($H_1: \mu < \mu_0$), la expresión anterior es igualmente válida para detectar $\mu = \mu_0 - \varepsilon$, $\varepsilon > 0$.
3. Si el contraste es bilateral la distancia ε puede manifestarse en las dos direcciones: $\mu_0 + \varepsilon$ y $\mu_0 - \varepsilon$, obteniéndose:

$$n \simeq \left(\frac{z_{\alpha/2} + z_\beta}{\varepsilon}\right)^2 \sigma^2$$

Ejemplo

En el ejemplo de los antebrazos, supuesto que se va a resolver con un nivel de significación del 5%, ¿qué número de varones adultos habría que examinar para detectar con probabilidad $0'9$ una longitud media real del antebrazo igual a $45'5 + 0'5\sigma$, siendo σ la desviación típica de X?

Se pretende que la potencia del contraste en $45'5 + 0'5\sigma$ sea igual a $0'9$ y, equivalentemente, la probabilidad de cometer un error de tipo II en $45'5 + 0'5\sigma$ sea $0'1$. De acuerdo con las fórmulas anteriores el tamaño muestral será

$$n \geq \left(\frac{z_{0'05} + z_{0'1}}{0'5} \right)^2 = \left(\frac{1'6449 + 1'2816}{0'5} \right)^2 = 34'26$$

El número mínimo requerido sería de 35 varones.

Contrastes sobre la varianza en poblaciones normales

Si se ha tomado una muestra de tamaño n de una población normal X y se está interesado en contrastar la hipótesis nula $H_0: \sigma^2 = \sigma_0^2$, bajo H_0,

- Si la media es desconocida:

$$\frac{(n-1)\widehat{S}^2}{\sigma_0^2} \in \chi^2_{n-1}$$

- Si la media es conocida:

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma_0} \right)^2 \in \chi^2_n$$

Contrastes para una proporción

- El objetivo es contrastar un valor postulado para la proporción de individuos de una población que verifican determinada característica A, es decir, $H_0 : p = p_0$.
- Se toma una muestra de tamaño n y se evalua sobre cada una de las n unidades muestrales el cumplimiento o no de A. La proporción muestral $\widehat{p}$, si $n > 30$, bajo $H_0 : p = p_0$, el estadístico

$$\frac{\widehat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}}$$

se distribuye, aproximadamente, según una $N(0, 1)$.

Ejemplo

Informes del Ministerio de Educación apuntan a que la proporción (p) de estudiantes de Informática que llegan a la Facultad manejando con soltura algún lenguaje de programación es de un 40%. El Decanato de la Facultad no da excesivo crédito a tal afirmación y opta por preguntar a diez estudiantes de primer curso acerca de este tema, resultando que en el momento de matricularse sólo 3 dominaban bien un lenguaje de programación. ¿Tenía razón el Decanato al dudar de la afirmación inicial?

Ejemplo

Utilizando la aproximación normal resulta:

$$\frac{0'3 - 0'4}{\sqrt{\frac{0'4 \cdot 0'6}{10}}} = -0'6456$$

de modo que un valor aproximado para el p-valor es:

$$p = 2P\left(Z \geq 0'6456\right) = 2 \cdot 0'2593 = 0'5186$$

Contraste de igualdad de varianzas para dos poblaciones normales independientes

- Sean $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ dos m.a.s. obtenidas de dos poblaciones normales e independientes
- Se desea contrastar la hipótesis nula $H_0 : \sigma_X^2 = \sigma_Y^2$
- La normalidad e independencia de las poblaciones permite afirmar que si H_0 es cierta, entonces

$$\left.\begin{array}{l} \dfrac{(n-1)\widehat{S}_X^2}{\sigma_X^2} \in \chi^2_{n-1} \\ \dfrac{(m-1)\widehat{S}_Y^2}{\sigma_Y^2} \in \chi^2_{m-1} \end{array}\right\} \Longrightarrow \frac{\widehat{S}_X^2}{\widehat{S}_Y^2} \in F_{n-1,m-1}$$

Contrastes para comparar medias. Muestras independientes y varianzas conocidas

- Sean $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ dos m.a.s. de X e Y que supondremos normales. Interesa contrastar la hipótesis nula de $H_0 : \mu_X = \mu_Y$ o, equivalentemente, $H_0 : \mu_X - \mu_Y = 0$.
- Si ambas muestras han sido recogidas con independencia, bajo H_0,

$$\frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \in N(0,1)$$

Contrastes para comparar medias. Muestras independientes y varianzas desconocidas pero iguales

Las hipótesis de independencia y normalidad garantizan que, si H_0 es cierta

$$\frac{\overline{X}-\overline{Y}}{\widehat{S}_T\sqrt{\dfrac{1}{n}+\dfrac{1}{m}}} \in t_{n+m-2}$$

con $\widehat{S}_T^2 = \dfrac{(n-1)\widehat{S}_X^2+(m-1)\widehat{S}_Y^2}{n+m-2}$ el estimador insesgado más eficiente de la varianza poblacional utilizando ambas muestras.

Contrastes para comparar medias. Muestras independientes y varianzas desconocidas y no iguales

Cuando H_0 es cierta, el estadístico:

$$\frac{\overline{X}-\overline{Y}}{\sqrt{\frac{\widehat{S}_X^2}{n}+\frac{\widehat{S}_Y^2}{m}}} \in t_{n+m-2-\delta},$$

siendo δ el entero más próximo a

$$\psi = \frac{[(m-1)\psi_1 - (n-1)\psi_2]^2}{(m-1)\psi_1^2 + (n-1)\psi_2^2}$$

con $\psi_1 = \frac{\hat{S}_X^2}{n}$ y $\psi_2 = \frac{\hat{S}_Y^2}{m}$.

Contrastes para comparar medias. Muestras apareadas

- Se considera la nueva variable aleatoria $D = X - Y$. Si D se supone normal con media μ_D y varianza σ_D^2.
- Contrastar H_0: $\mu_X = \mu_Y$ equivale a contrastar H_0': $\mu_D = 0$ y para ello disponemos de una m.a.s. sin más que considerar los valores $D_i = X_i - Y_i$ obtenidos a partir de la muestra apareada (X_i, Y_i).
- Bajo H_0',

$$\frac{\overline{D}}{\widehat{S}_D}\sqrt{n} \in t_{n-1}$$

y este puede ser utilizado como estadístico de contraste transformando el problema original en el contraste de una media bajo condiciones de normalidad.

Contrastes para comparar medias. Ausencia de normalidad

- Si las muestras son grandes el T.C.L. garantiza que la aproximación normal para las distribuciones de $\overline{X}$ e $\overline{Y}$ (caso de muestras independientes), o de $\overline{D}$ (caso de muestras apareadas) es adecuada y, por consiguiente, la violación de la hipótesis de normalidad no es especialmente preocupante. Más aún, se ha demostrado que el test t de Student es muy *robusto* (insensible) a las desviaciones de la normalidad y por todo ello los estadísticos utilizados proporcionan regiones críticas razonablemente válidas para tamaños muestrales grandes (pongamos $n > 30$).
- Si las muestras son pequeñas se dispone de los contrastes no paramétricos.

Contrastes para comparar dos proporciones

- Contrastar la hipótesis nula $H_0 : p_X = p_Y$ donde p_X y p_Y denotan, respectivamente, las proporciones de elementos de las poblaciones X e Y que verifican una cierta característica que denotaremos por A.
- Si $n > 30$, $m > 30$, $\widehat{p}_X$ y $\widehat{p}_Y$ son independientes y con distribuciones aproximadas $N\left(p_X, \sqrt{p_X(1-p_X)/n}\right)$ y $N\left(p_Y, \sqrt{p_Y(1-p_Y)/m}\right)$ respectivamente.

Contrastes para comparar dos proporciones

- Bajo $H_0 : p_X = p_Y = p$, se deduce que

$$\frac{\widehat{p}_X - \widehat{p}_Y}{\sqrt{p(1-p)\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

es aproximadamente $N(0,1)$.

- Como no se conoce p, se sustituye por su estimador

$$\widehat{p} = \frac{n\widehat{p}_X + m\widehat{p}_Y}{n+m}$$

Relación entre intervalos de confianza y contrastes de hipótesis

- Si θ_0 pertenece al intervalo de confianza construido para θ a un nivel de confianza $1-\alpha$, entonces $H_0: \theta = \theta_0$ será aceptada frente a $H_1: \theta \neq \theta_0$ cuando el contraste se realiza con un nivel de significación α y con la misma información muestral.
- Por ejemplo, a un nivel α, $H_0: \mu = \mu_0$, con μ la media de una población normal X con varianza desconocida, será aceptada siempre y cuando

$$\overline{X} \in \left(\mu_0 - t_{n-1,\alpha/2}\frac{\widehat{S}}{\sqrt{n}}, \mu_0 + t_{n-1,\alpha/2}\frac{\widehat{S}}{\sqrt{n}} \right)$$

o, equivalentemente, cuando

$$\mu_0 \in \left(\overline{X} - t_{n-1,\alpha/2}\frac{\widehat{S}}{\sqrt{n}}, \overline{X} + t_{n-1,\alpha/2}\frac{\widehat{S}}{\sqrt{n}} \right)$$