Bioestadística: Inferencia Estadística. Análisis de Una Muestra

M. González

Departamento de Matemáticas. Universidad de Extremadura

Estimación Puntual e Intervalos de Confianza

Planteamiento del Problema de Inferencia

- X variable cuantitativa que define una población. La distribución de probabilidad de X es desconocida en todo o en parte.
- θ parámetro desconocido asociado a la distribución de probabilidad de X (por ejemplo, la media poblacional, μ, la varianza poblacional, σ^2, la proporción asociada a una distribución binomial, p etc.)
- $X_1, \ldots, X_n$ Muestra Aleatoria Simple de la variable X. Nuestra principal herramienta para obtener información sobre la distribución de probabilidad de X ó sobre cualquier parámetro θ.

Estimación puntual

Un estimador puntual de θ es una función $\hat{\theta}$ de $X_1, \ldots, X_n$ que aproxima el valor de θ.

Cuando se da el valor del estimador $\hat{\theta}$, hay que dar también una estimación del error que se comete al aproximar el valor del parámetro θ mediante $\hat{\theta}$.

Intervalos de Confianza

Si T_1 y T_2 son dos valores, obtenidos a partir de la muestra y tales que

$$P(T_1 \leq \theta \leq T_2) = 1 - \alpha$$

entonces $[T_1, T_2]$ es un intervalo de confianza para θ con un nivel de confianza de $1 - \alpha$.

- Intuitivamente, T_1, T_2 son dos valores tales que el $100(1-\alpha)\%$ de las veces que repitamos el experimento en esa población, el valor desconocido de θ estará entre estos dos valores.
- El nivel de confianza $1 - \alpha$ es un valor entre 0 y 1 que debe estar próximo a 1(0.90, 0.95, 0.99, . . .). De ello resulta que el valor de α es próximo a 0 (0.1, 0.05, 0.01, . . .).

Estimación de la Media, μ (RESUMEN II)

Sea $X_1, \ldots, X_n$ una muestra aleatoria simple de una población $N(\mu, \sigma^2)$.
Sean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ y $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Estimador Puntual de μ:

$$\hat{\mu} = \bar{X} \qquad \text{Error estándar } = \frac{S}{\sqrt{n}}$$

Intervalo de Confianza para μ al Nivel de Confianza $1-\alpha$:

σ^2 CONOCIDA	σ^2 DESCONOCIDA
$\mu \in \bar{x} \pm z_\alpha \frac{\sigma}{\sqrt{n}}$	$\mu \in \bar{x} \pm t_\alpha(n-1)\frac{s}{\sqrt{n}}$

z_α en TABLA III-1 para la distribución $N(0,1)$.
$t_\alpha(n-1)$ TABLA IV para distribuciones t-Student.

Si $n \geq 60$ los intervalos anteriores son válidos aunque no se verifique la hipótesis de normalidad.

5. Un dermatólogo investiga cierto tipo de afección de piel induciéndolo en una muestra aleatoria de 25 ratas y tratándolas luego con un nuevo fármaco. Se cuenta el número de horas hasta que desaparece dicha afección, con los resultados siguientes: $\bar{x} = 132$ horas, $s = 40$ horas. Supondremos que el número de horas hasta que desaparece la afección se distribuye normalmente.

(a) Estima el número medio de horas que tarda en desaparecer la afección dermatológica con el nuevo fármaco. ¿Cuál es el error máximo de esta estimación? Utiliza un nivel de confianza del 95%.

$X =$número de horas hasta que aparece la afección $\sim N(\mu, \sigma^2)$

$n = 25 \qquad X_1, \ldots, X_{25}$ v.a. i.i.d. $\qquad x_1, \ldots, x_{25} \qquad \bar{x} = 132\text{h}, s = 40\text{h}$

RESUMEN II $\qquad 1 - \alpha \qquad \mu \in \bar{x} \pm t_\alpha(n-1)\dfrac{s}{\sqrt{n}}$

$1 - \alpha = 0.95 \Longleftrightarrow \alpha = 0.05$ $\qquad$ TABLA IV $\quad t_{0.05}(24) = 2.064$

$$\mu \in 132 \pm 2.064\frac{40}{\sqrt{25}} = 132 \pm 16.512 = [115.49, 148.51]$$

con una confianza de un 95%.

5. (b) Si repetimos este experimento exactamente en las mismas condiciones, la longitud del intervalo que obtendríamos, ¿sería la misma?. Razona la respuesta.
(c) Supongamos ahora que $\sigma = 32$ horas. Calcula un intervalo de confianza al 90% para el número medio de horas que tarda en desaparecer la afección dermatológica. En estas condiciones, ¿qué tamaño de muestra se necesitaría para tener el 90% de confianza de que la media se estima dentro de ± 5 horas?

$X =$número de horas hasta que aparece la afección $\sim N(\mu, \sigma^2)$

$$n = 25 \qquad X_1, \ldots, X_{25} \text{ v.a. i.i.d.} \qquad x_1, \ldots, x_{25} \qquad \bar{x} = 132\text{h}, \, s = 40\text{h}$$

$$\sigma = 32\text{h} \qquad \mu \in 132 \pm 10.496 \text{ con una confianza del } 90\%$$

$$n? \qquad z_\alpha \frac{\sigma}{\sqrt{n}} \leq \text{PRECISIÓN}$$

$$1 - \alpha = 0.9 \iff \alpha = 0.1 \qquad \text{TABLA III-1} \qquad z_{0.1} = 1.645$$

$$n? \qquad 1.645 \frac{32}{\sqrt{n}} \leq 5 \iff n \geq \left(1.645 \frac{32}{5}\right)^2 = 110.166 \Longrightarrow n = 111$$

16. La ingestión media de calorías por persona y por día en una determinada región es de 2900 calorías. En una región vecina, se efectuó un muestreo para estudiar el consumo medio de calorías. Se eligieron aleatoriamente 50 personas y los resultados fueron de un consumo medio de 3000 calorías por persona y por día, con una desviación típica muestral de 100 calorías. Suponiendo que la distribución del consumo de calorías en esa región es normal, contesta las siguientes preguntas:
(a) ¿Podemos admitir, con un nivel de significación del 5%, que las dos regiones tienen diferente consumo medio de calorías por persona y por día?

$$X = \text{consumo de calorías por persona y por día} \sim N(\mu, \sigma^2)$$

$$n = 50 \qquad X_1, \ldots, X_{50} \text{ v.a. i.i.d.} \qquad x_1, \ldots, x_{50} \qquad \bar{x} = 3000\text{cal.}, \quad s = 100\text{cal.}$$

$$\mu = 2900 \quad \text{vs.} \quad \mu \neq 2900$$

$$H_0 : \mu = 2900$$
$$H_1 : \mu \neq 2900$$

RESUMEN III: Sea $X_1, \ldots, X_n$ una muestra aleatoria simple de una población $N(\mu, \sigma^2)$ y μ_0 un valor conocido. Sean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ y $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

	H_0	H_1	RECHAZAMOS H_0 A UN NIVEL α SI	
			σ^2 CONOCIDA	σ^2 DESCONOCIDA
I.	$\mu = \mu_0$	$\mu \neq \mu_0$	$\dfrac{\lvert\bar{x} - \mu_0\rvert}{\frac{\sigma}{\sqrt{n}}} > z_\alpha$	$\dfrac{\lvert\bar{x} - \mu_0\rvert}{\frac{s}{\sqrt{n}}} > t_\alpha(n-1)$
II.	$\mu \leq \mu_0$	$\mu > \mu_0$	$\dfrac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_{2\alpha}$	$\dfrac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} > t_{2\alpha}(n-1)$
III.	$\mu \geq \mu_0$	$\mu < \mu_0$	$\dfrac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} < -z_{2\alpha}$	$\dfrac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} < -t_{2\alpha}(n-1)$

X =consumo de calorías por persona y por día $\sim N(\mu, \sigma^2)$

$$n = 50 \qquad \bar{x} = 3000\text{cal.}, s = 100\text{cal.}$$

$$H_0 : \mu = 2900$$
$$H_1 : \mu \neq 2900$$

Rechazamos H_0 al nivel de significación α si $\dfrac{|\bar{x} - \mu_0|}{s/\sqrt{n}} > t_\alpha(n-1)$

$$t_{exp} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{3000 - 2900}{100/\sqrt{50}} = 7.07$$
$$\alpha = 0.05 \Rightarrow t_{0.05}(49) \simeq 2.009 \Rightarrow |t_{exp}| > t_{0.05}(49)$$

Rechazamos H_0 al nivel de significación 0.05

$$\alpha = 0.01 \Rightarrow t_{0.01}(49) \simeq 2.678 \Rightarrow |t_{exp}| > t_{0.01}(49) \Rightarrow H_1$$
$$\alpha = 0.001 \Rightarrow t_{0.001}(49) \simeq 3.497 \Rightarrow |t_{exp}| > t_{0.001}(49) \Rightarrow H_1$$

$$p < 0.001$$

17. Un laboratorio ha obtenido una sustancia que en cierto modo alivia determinadas afecciones cutáneas. Dicho laboratorio afirma, según sus experiencias, que dicha sustancia causa alivio, en los pacientes tratados, por espacio de 10 horas por término medio. Se ha elegido una muestra de 120 individuos afectados por esta dolencia, se les trató con dicha sustancia y se contabilizó el tiempo de alivio en los pacientes. Los resultados fueron un tiempo medio de 9.6 horas, con una desviación típica de 1.2 horas.
(a) ¿Podemos desmentir la afirmación del laboratorio con un nivel de significación del 5%? Calcula el valor de p.

X =número de horas durante las que causa alivio la sustancia del laboratorio

$$\mu, \quad \sigma^2$$

$n = 120 \qquad X_1, \ldots, X_{120}$ v.a. i.i.d. $\qquad x_1, \ldots, x_{120} \qquad \bar{x} = 9.6\text{h}, s = 1.2\text{h}$

$$H_0 : \mu = 10$$
$$H_1 : \mu < 10$$

Rechazamos H_0 al nivel de significación α si $\dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} < -z_{2\alpha}$

17.

X =número de horas durante las que causa alivio la sustancia del laboratorio

$$\mu, \quad \sigma^2$$

$$n = 120 \qquad \bar{x} = 9.6\text{h}, s = 1.2\text{h}$$

$$H_0 : \mu = 10$$
$$H_1 : \mu < 10$$

Rechazamos H_0 al nivel de significación α si $\dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} < -z_{2\alpha}$

$$t_{exp} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{9.6 - 10}{1.2/\sqrt{120}} = -3.65$$

$$\alpha = 0.05 \Rightarrow -z_{0.1} = -1.645 \Rightarrow t_{exp} < -z_{0.1} \Rightarrow H_1$$
$$\alpha = 0.01 \Rightarrow -z_{0.02} = -2.326 \Rightarrow t_{exp} < -z_{0.02} \Rightarrow H_1$$
$$\alpha = 0.005 \Rightarrow -z_{0.01} = -2.576 \Rightarrow t_{exp} < -z_{0.01} \Rightarrow H_1$$

$$p < 0.005$$

17. (b) Como máximo, ¿en cuánto podemos estimar el tiempo medio de alivio que produce esa sustancia?

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right) \qquad s^2 \approx \sigma^2$$

$$1-\alpha = P\left(\frac{\bar{X}-\mu}{s/\sqrt{n}} \geq -z_{2\alpha}\right) = P\left(\mu \leq \bar{X} + z_{2\alpha}\frac{s}{\sqrt{n}}\right)$$

$\bar{X} + z_{2\alpha}\frac{s}{\sqrt{n}}$ es un límite superior de confianza para μ al nivel $1-\alpha$

$$1-\alpha = 0.95 \Rightarrow 9.6 + z_{0.1}\frac{1.2}{\sqrt{120}} = 9.78$$

18. Se observaron 12 mujeres primíparas y se obtuvieron los tiempos en minutos de la duración del parto:

353 496 568 422 410 380 463 430 310 518 446 368

(a) ¿Es normal la variable?

$$\begin{array}{lcl} H_0 & : & X \text{ sigue distribución Normal} \\ H_1 & : & X \text{ no sigue distribución Normal} \end{array}$$

RESUMEN VI - TABLA VII: TEST DE D'AGOSTINO

310	353	368	380	410	422	430	446	463	496	518	568
$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$	$x_{(11)}$	$x_{(12)}$

$$D_{exp} = \frac{(1*310+2*353+\ldots+12*568) - 13*(310+\cdots+568)/2}{12*\sqrt{12*((310^2+\ldots+568^2)-(310+\ldots+568)^2/12)}} = 0.2832$$