Clase 17 de Bioestadística

Intervalos de confianza

Ernesto Mordecki

CMAT, Facultad de Ciencias, Universidad de la República.

Uruguay

Contenidos de la clase

Intervalos de confianza

- Tenemos una muestra aleatoria simple

$$X_1, \ldots, X_n$$

de v.a.i.i.d. con esperanza μ y varianza σ^2 finita y positiva.

- Sabemos estimar la esperanza μ mediante el promedio:

$$\widehat{\mu} = \overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

- Aplicando la ley fuerte de los grandes números obtenemos la consistencia:

$$\widehat{\mu} \to \mu \quad \text{cuando } n \to \infty$$

- Sin emabargo, no tenemos información sobre la precisión de la estimación
- Es decir: ¿cuál es el error de la estimación?
- ¿Cuan lejos está el verdadero valor de la estimación $\widehat{\mu}$?
- Podemos dar un intervalo cerca de $\widehat{\mu}$ que contenga el verdadero valor con probabilidad alta?
- Un tal intervalo se llama intervalo de confianza

Intervalo de confianza con varianza conocida

- Para construir el intervalo de confianza utilizamos el Teorema Central del Límite (TCL)
- Suponemos entonces que conocemos σ^2 y queremos estimar μ (desconocido)
- El TCL nos dice que

$$\frac{\sqrt{n}}{\sigma}\left(\overline{X}_n - \mu\right) \approx Z$$

  donde $Z \sim \mathcal{N}(0,1)$
- Involucra la diferencia $\overline{X}_n - \mu$ que es la que nos interesa

- Más precisamente nos dice que

$$\overline{X}_n - \mu \approx \frac{\sigma}{\sqrt{n}} Z$$

- ¿Cómo construímos nuestro intervalo para μ?
- Primero decidimos cuanto estamos dispuestos a equivocarnos
- Es decir, cual es la probabilidad que toleramos de
- Esa tolerancia, se llama nivel de significancia y se designa α

- El valor usual para $\alpha = 0,05$ (es decir, un 5 %) pero puede variar dependiendo del problema
- Otro valor de α usual es $\alpha = 0{,}01$ (si queremos estar mas seguros de no equivocarnos)

Procedemos de la siguiente forma:

- Asumimos que n es suficientemente grande [1]
- Entonces utilizamos que

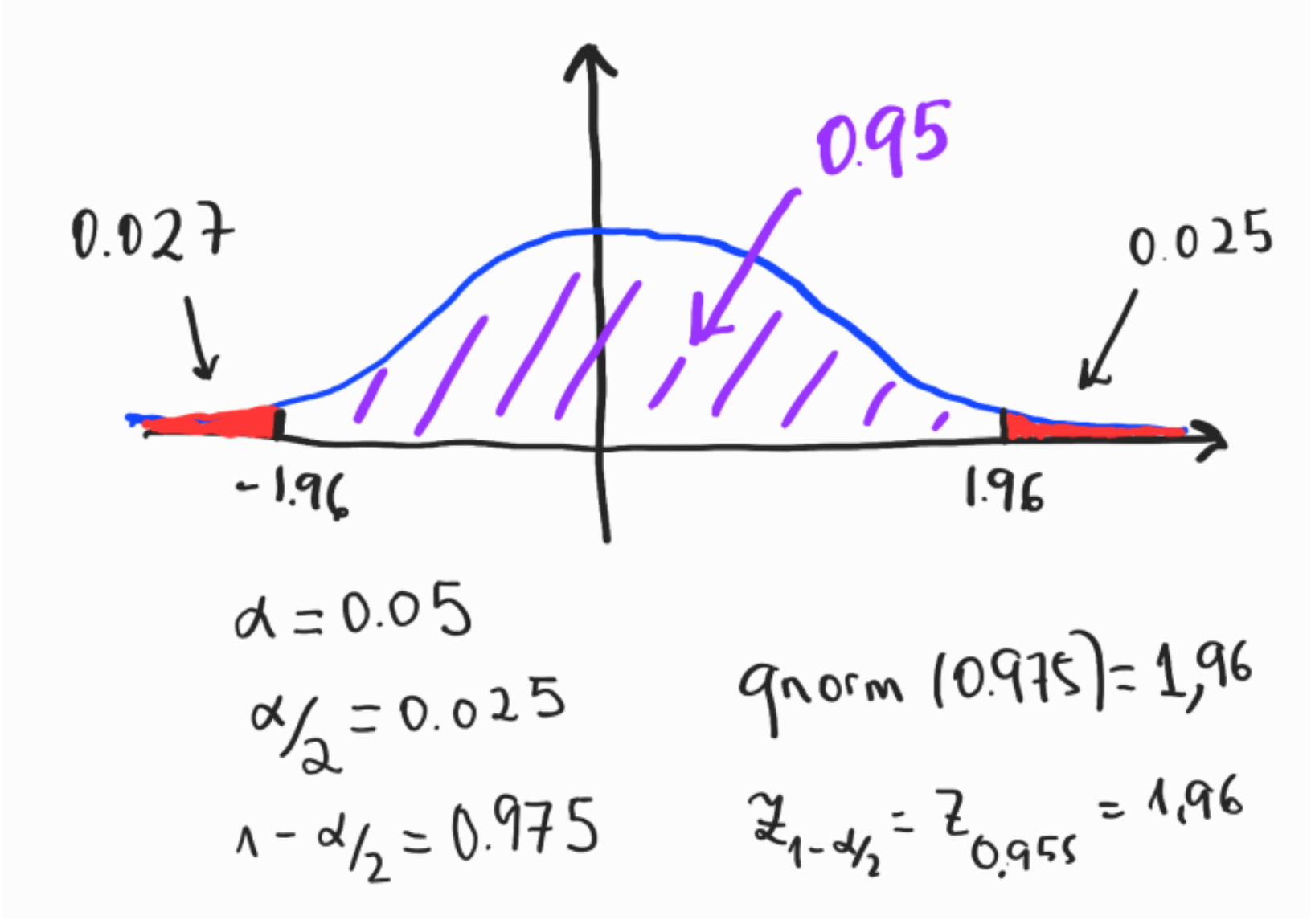
$$\overline{X}_n - \mu \approx \frac{\sigma}{\sqrt{n}} Z$$

- Construímos primero un intervalo simétrico para Z, que es normal estándar:

[1] En la práctica se asume $n = 30$ o mayor

Eso significa encontrar el cuantil $z_{1-\alpha/2}$ tal que

$$P(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1 - \alpha$$



- Asumimos una significancia $\alpha = 0{,}05$
- Es significa que tenemos un intervalo de probabilidad 0.95 para Z:

$$P(-1{,}96 \leq Z \leq 1{,}96) = 0{,}95$$

- Nuestro objetivo es encontrar un intervalo para μ
- Multiplicamos

$$P\left(-1{,}96\frac{\sigma}{\sqrt{n}} \leq \frac{\sigma}{\sqrt{n}}Z \leq 1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

- Ahora aplicamos el TCL y cambiamos el término azul

$$P\left(-1{,}96\frac{\sigma}{\sqrt{n}} \leq \overline{X}_n - \mu \leq 1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

- Despejando, es equivalente a

$$P\left(\overline{X}_n - 1{,}96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}_n + 1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

- Estamos diciendo que con probabilidad 0,95 el parámetro desconocido μ se encuentra en el intervalo

$$\left[\overline{X}_n - 1{,}96\frac{\sigma}{\sqrt{n}}\ ,\ \overline{X}_n + 1{,}96\frac{\sigma}{\sqrt{n}}\right]$$

- Encontramos entonces nuestro intervalo de confianza

Comentarios

- Tenemos un intervalo centrado en $\overline{X}_n$
- El error del intervalo es

$$\epsilon_n = 1{,}96\frac{\sigma}{\sqrt{n}}$$

- La amplitud del intervalo es

$$2\epsilon_n = 2 \times 1{,}96\frac{\sigma}{\sqrt{n}}$$

- Cuanto menor es la amplitud es mas precisa la estimación:
  - La amplitud aumenta con σ^2
  - La amplitud disminuye con $\sqrt{n}$
- Muestras mas grandes dan mayor precisión

Intervalo de confianza con varianza estimada

- Para calcular el intervalo de confianza precisamos saber:
- $\overline{X}_n$ que se calcula a partir de los datos
- $\sqrt{n}$ que es la raíz del tamaño de la muestra
- σ que es el desvío estándar de los datos de la muestra.
- En general σ es desconocido

Estimación de σ

- Habíamos visto que

$$\widehat{\sigma_n^2} = \frac{1}{n}\sum_{k=1}^{n}\left(X_k - \overline{X}_n\right)^2$$

era un estimador de σ^2

- Se utiliza sin embargo el estimador alternativo

$$s_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \overline{X}_n\right)^2$$

que tiene mejores propiedades (es insesgado y consistente[2])

[2] La esperanza del estimador coincide con el parámetro estimado

- Si n es relativamente grande (como estamos suponiendo) no hay gran diferencia práctica

- El estimador de σ es entonces

$$s_n = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \overline{X}_n\right)^2}$$

- Como es usual sustituímos el parámetro desconocido por su estimador[3]

- Nuestro intervalo de confianza con nivel de confianza 0.95 (es decir $\alpha = 0{,}05$) es entonces

$$\left[\overline{X}_n - 1{,}96\frac{s_n}{\sqrt{n}}\ ,\ \overline{X}_n + 1{,}96\frac{s_n}{\sqrt{n}}\right]$$

[3] Aquí podemos hacer dos cosas: (1) cruzamos los dedos para que el estimador esté cerca del parámetro; (2) estudiar la teoría donde se muestra que hay convergencia del estimador al parámetro

Otros niveles de confianza

- El intervalo de confianza con un nivel α es

$$\left[\overline{X}_n - z_{1-\alpha/2}\frac{s_n}{\sqrt{n}}\ ,\ \overline{X}_n + z_{1-\alpha/2}\frac{s_n}{\sqrt{n}}\right]$$

- Por ejemplo si $\alpha = 0{,}01$ el intervalo tiene nivel de confianza 0,99.

- En ese caso calculamos

$$1 - \frac{\alpha}{2} = 0{,}995 \qquad \texttt{qnorm(0.995)=2.576}$$

- El intervalo es

$$\left[\overline{X}_n - 2{,}576\frac{s_n}{\sqrt{n}}\ ,\ \overline{X}_n + 2{,}576\frac{s_n}{\sqrt{n}}\right]$$

- Si queremos mayor confianza, el intervalo es mas grande.

Ejemplo

- Queremos estimar la utilidad de un medicamento que declara controlar la presión arterial.
- Elegimos 30 personas con hipertensión y medimos la diferencia de presión sistólica (alta) al inicio de un tratamiento y al final, luego de 60 días.
- Valores típicos de presión arterial (en mmHg) son 120/80 (a veces se dice ”doce-ocho”)
- Obtenemos 30 datos, cuyo promedio es (en mmHg)

$$\overline{X}_{30} = 5, \qquad s_n = 4{,}5$$

- Nuestro intervalo con 0,95 de confianza es

$$5 \pm 1{,}96 \frac{4{,}5}{\sqrt{30}} = [3{,}39\ ,\ 6{,}61]$$

Estimación del tamaño de muestra

- Al diseñar un experimento, es necesario conocer el tamaño de la muestra a utilizar
- Si aceptamos un nivel de confianza $1 - \alpha$ y un error ϵ a priori, podemos determinar n
- Para eso utilizamos la ecuación

$$\epsilon_n = 1{,}96\frac{\sigma}{\sqrt{n}}$$

- Debemos tener una idea a prior de σ, para despejar

$$\sqrt{n} = 1{,}96\frac{\sigma}{\epsilon_n}$$

- Es decir

$$n = \left(1{,}96\frac{\sigma}{\epsilon_n}\right)^2$$

Tamaño de muestra: ejemplo

- Queremos diseñar un estudio del fármaco que reduce la presión arterial
- Tenemos una estimación previa de $s = 4{,}5$
- Aceptamos un nivel de significancia de $\alpha = 0{,}05$ y un error $\epsilon = 1$ en mmHg
- Precisamos

$$n = \left(1{,}96\frac{\sigma}{\epsilon_n}\right)^2 = \left(1{,}96\frac{4{,}5}{1}\right)^2 = 77{,}79 \approx 78$$

pacientes.

- Si queremos una significancia de $\alpha = 0{,}01$, entonces

$$n = \left(2{,}58\frac{\sigma}{\epsilon_n}\right)^2 = \left(2{,}58\frac{4{,}5}{1}\right)^2 = 134{,}8 \approx 135$$

pacientes.

Intervalos de confianza para proporciones

- Un caso particular muy importante de la situación anterior es la estimación de proporciones
- Se trata de caso en donde las variables $X_1, \ldots, X_n$ toman dos valores[4]: 0 (fracaso) y 1 (éxito)
- El parámetro de interés es

$$p = P(X_1 = 1) = P(\text{éxito}) = E(X_1)$$

(corresponde al μ en el desarrollo anterior)

[4] A veces se dice que las variables son binarias

- En este caso recordamos que el estimador de p es

$$\widehat{p_n} = \overline{X}_n = \frac{\#\text{ éxitos}}{n}$$

- Recordamos también que la varianza en este caso es

$$\sigma^2 = var(X_1) = p(1-p)$$

- Tenemos entonces un estimador de la varianza

$$\widehat{\sigma_n^2} = \widehat{p_n}(1-\widehat{p_n})$$

- El intervalo de confianza queda entonces

$$\left[\overline{X}_n - z_{1-\alpha/2}\frac{\sqrt{\widehat{p_n}(1-\widehat{p_n})}}{\sqrt{n}}\ ,\ \overline{X}_n + z_{1-\alpha/2}\frac{\sqrt{\widehat{p_n}(1-\widehat{p_n})}}{\sqrt{n}}\right]$$

Ejemplo: rock versus cumbia

- Se plantea una acalorada discusión en un grupo de estudiantes de facultad:
- ¿Que es mas popular, el rock o la cumbia?
- Decididos a dirimir la cuestión científicamente, los estudiantes se proponen hacer una encuesta.
- Se proponen entonces consultar a n estudiantes de facultad sobre sus gustos musicales: ¿prefieren el rock o la cumbia?
- Aceptan una significancia del 5 %, y quieren un error máximo del 1 %
- ¿Qué tamaño de muestra precisan?

Solución

- Tomamos la fórmula del tamaño de muestra con esa significación:

$$n = \left(1{,}96\frac{\sigma}{\epsilon_n}\right)^2$$

- Como $\sigma^2 = p(1-p)$ es desconocido tomamos el peor caso
- El peor caso (máximo de la varianza) es con $p = 1/2$

Solución

- En ese caso $\sigma^2 = p(1-p) = \frac{1}{4} = 0{,}25$, entonces $\sigma = 0{,}5$, resultando

$$n = \left(1{,}96 \times \frac{0{,}5}{0{,}01}\right)^2 = 9604$$

estudiantes

- Como son demasiados estudiantes, deciden tolerar un error del 3 %:

$$n = \left(1{,}96 \times \frac{0{,}5}{0{,}03}\right)^2 = 1067$$

estudiantes

- El número 1067 (aprox. 1000) es el que se utiliza en las encuestas de opinión