Tema 5.- Técnicas de muestreo

Curso 2023-24

April 25, 2024

Muestreo

Muestreo

- Me interesa un conjunto de datos. A este conjunto lo llamaremos población
- Tengo acceso a un subconjunto de esos datos llamados muestra
- Me interesa una característica, numérica o no, de estos datos (media, una proporción, varianza, saber si unos datos se parecen a una determinada variable teórica, etc)
- Intentamos calcular una aproximación (estimación) de ese valor basándonos en la muestra

Niveles del proceso de muestreo

Nos centramos en el caso en que nos interesa estimar un parámetro (inferencia paramétrica)

- Primer nivel: intentamos calcular un estimador del parámetro en que estamos interesados
- Segundo nivel: desarrollamos un test para intentar comprobar si las conclusiones (hipótesis) de los pasos anteriores están equivocadas
- Tercer nivel: calculamos un intervalo (llamado intervalo de confianza) en el que estemos bastante seguros de que va a estar el valor a estimar

Características (deseables) de un estimador

- Los estimadores pueden ser vistos como variables aleatorias (varían con la muestra)
- Los estimadores más usados cumplen
  - Se conoce qué tipo de variable aleatoria son (binoimial, normal, etc)
  - Su media coincide con el parámetro que interesa estudiar (estimadores no sesgados)
  - Su varianza es pequeña cuando el número de datos es grande (estimadores robustos)

Estimadores usuales

- Tenemos una muestra $\{x_1, \cdots, x_n\}$
- Para estimar:
  - La media m: usaremos el estimador $\hat{\mathrm{m}} = \overline{X} = \frac{x_1 + \cdots + x_n}{n}$ (media muestral)
  - Una proporción p: usaremos el estimador $\hat{\mathrm{p}}$ = proporción de elementos de la muestra que cumplen la propiedad indicada (proporción muestral)
  - La varianza v: usaremos el estimador

$$\hat{v} = S^2_{n-1} = \frac{(x_1 - m)^2 + \cdots + (x - n - m)^2}{n - 1}$$

    (cuasi-varianza muestral)
- Es conocido que estos tres estimadores son no sesgados, robustos y se conoce qué tipo de variable aleatoria son

Estimador para la media

Teorema Central del Límite

- Teorema central del Límite
  - Tenemos una variable aleatoria X con media m y desviación típica s. Cogemos $\{X_1, \cdots, X_n\}$ una muestra aleatoria simple (independientes)
  - Si el tamaño de la muestra es suficientemente grande ($n \geq 30$) entonces
    - $X_1 + \cdots X_n \simeq N(n \cdot m, s \cdot \sqrt{n})$
    - $\hat{m} = \overline{X} = \frac{X_1 + \cdots X_n}{n} \simeq N(m, \frac{s}{\sqrt{n}})$
- Consecuencias
  - $E[\hat{m}] = m$ (es no sesgado)
  - $var(\hat{m}) = \frac{v}{n}$ que tiende hacia 0 si $n \to \infty$ (es robusto)
  - Tipificando obtenemos que $\frac{\hat{m} - m}{\frac{s}{\sqrt{(n)}}} \simeq N(0, 1)$
  - Dado que $N(0, 1)$ es simétrica respecto al eje Y, también podemos escribir $\frac{m - \hat{m}}{\frac{s}{\sqrt{n}}} \simeq N(0, 1)$
  - En esta fórmula estamos suponiendo que conocemos s

Estimador para la proporción

Estimador para la proporción

- Supongamos que nos interesa la proporción p de elementos de una población que cumplen una determinada propiedad P
- Queremos usar el Teorema Central del Límite
- Tenemos una muestra aleatoria simple $X_1, \cdots, X_n$ y definimos las siguientes variables

$$Y_i = \begin{cases} 1 \, si \, X_i \, cumple \, P \\ 0 \, si \, X_i \, no \, cumple \, P \end{cases}$$

- Estas variables cumplen
  - Son variables de Bernouilli con parámetro p
  - Son independientes
  - Para nuestra muestra $\hat{p} = \frac{Y_1 + \cdots + Y_n}{n}$
- Por el Teorema Central del Límite tendremos que

$$\hat{p} \simeq N(p, \sqrt{\frac{p(1-p)}{n}})$$

Estimador para la proporción

- Por tanto $\hat{p} \simeq p + \sqrt{\frac{p(1-p)}{n}} N(0,1)$
- Despejando tenemos $\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \simeq N(0,1)$
- Si la muestra es suficientemente grande, podemos suponer que p es bastante semejante a $\hat{p}$. Usando, además, que $N(0,1)$ es simétrica respecto al eje Y obtenemos

$$\frac{p-\hat{p}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \simeq N(0,1)$$

Estimador para la varianza

Estimador para la varianza

- Partimos de una muestra aleatoria simple $\{X_1, \ldots, X_n\}$ proveniente de una población con distribución normal donde su media y varianza son desconocidas
- Para estimar la varianza de esta población se utiliza como estimador la cuasi-varianza muestral $\hat{S}^2$
- Es un estimador no sesgado y robusto
- Se sabe que

$$\frac{(n-1)\hat{S}^2}{s^2} \simeq \chi^2_{n-1}$$

Modificaciones del estimador para la media

Estimador para la media con s desconocida y población normal

- Muchas veces la desviación típica de la población no es conocida
- En ese caso se sustituye s por su estimador $\hat{s}$
- No sigue siendo cierto que $\hat{m} \simeq m + \frac{\hat{s}}{\sqrt{n}} N(0,1)$
- Si la población sigue una distribución normal, se tiene una fórmula semejante en la que aparece la variable aleatoria t de Student

$$\hat{m} \simeq m + \frac{\hat{s}}{\sqrt{n}} t_{n-1}$$

Siendo n el tamaño de la muestra

Coeficiente corrector de poblaciones finitas

- Las muestras se toman sin reemplazamiento (es decir, los elementos de la muestra son distintos)
- Esto hace que las variables correspondientes no sean independientes
- Este efecto es especialmente relevante si el tamaño de la muestra es comparado con el de la población
- Para prevenir este problema, se debe incluir una modificación en la varianza de la variable $\hat{m}$
- Esta modificación de conoce como 'coeficiente corrector de poblaciones finitas'
- En este caso la fórmula queda como

$$\hat{m} \simeq m + \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} N(0,1)$$

donde N es el tamaño de la población

Tests de hipótesis

Test de hipótesis

- Partimos de una hipótesis inicial sobre un parámetro x (normalmente m o p)
- Esta hipótesis (llamada hipótesis nula H_0,puede ser
  - De igualdad: $x = x_0$
  - Desigualdad: $x \leq x_0$
- Para testear esta hipótesis
  - Cogemos una muestra y calculamos su estimador $\hat{x}$
  - Si el valor de $\hat{x}$ es coherente con nuestra hipótesis H_0 la muestra no nos da ninguna razón adicional para dudar de nuestra hipótesis
  - Si el valor obtenido es extraño (dentro de nuestra hipótesis) supondremos que la hipótesis seguramente será falsa
  - Un valor extraño de $\hat{x}$ se dará cuando sea demasiado grande o demasiado pequeño si la hipótesis inicial es cierta
  - La probabilidad si la hipótesis es cierta de que se de un valor tan extremo (o más) como el obtenido se llama el p-valor del test

Test de hipótesis

- Partimos de una hipótesis inicial sobre un parámetro x (normalmente m o p)
- Esta hipótesis (llamada hipótesis nula H_0,puede ser
  - De igualdad: $x = x_0$
  - Desigualdad: $x \leq x_0$
- Para testear esta hipótesis
  - Cogemos una muestra y calculamos su estimador $\hat{x}$
  - Si el valor de $\hat{x}$ es coherente con nuestra hipótesis H_0 la muestra no nos da ninguna razón adicional para dudar de nuestra hipótesis
  - Si el valor obtenido es extraño (dentro de nuestra hipótesis) supondremos que la hipótesis seguramente será falsa
  - Un valor extraño de $\hat{x}$ se dará cuando sea demasiado grande o demasiado pequeño si la hipótesis inicial es cierta
  - La probabilidad si la hipótesis es cierta de que se de un valor tan extremo (o más) como el obtenido se llama el p-valor del test

Test de hipótesis

Pasos del test de hipótesis

- Partimos de una hipótesis inicial H_0 sobre un parámetro x (normalmente m o p)
- Para testear esta hipótesis
  - Cogemos una muestra y calculamos su estimador $\hat{x}$
  - Fijamos un nivel de significación e (este nivel significa que consideramos raros aquellos sucesos que tengan una probabilidad $\leq e$)
  - Calcularemos la probabilidad de haber obtenido un valor tan extremo (o más) que el obtenido si la hipótesis H_0 es cierta (su p-valor)
  - Si este p-valor es menor que e rechazaremos nuestra hipótesis

Distribuciones usadas en tests de hipótesis

- Para calcular dónde debería estar la aproximación del parámetro en que estemos interesados usaremos:
  - Para m si conozco s:

$$\frac{\hat{m} - m}{\frac{s}{\sqrt{n}}} \simeq N(0,1)$$

  - Para m si no conozco s:

$$\frac{\hat{m} - m}{\frac{\hat{s}}{\sqrt{n}}} \simeq t_{n-1}$$

  - Para una proporción p:

$$\frac{\hat{p} - p}{\sqrt{\frac{p \cdot (1-p)}{n}}} \simeq N(0,1)$$

Test de hipótesis de igualdad para la media (si s es conocida)

Vamos a a empezar a estudiar el caso en que tenemos una hipótesis de igualdad

- Creemos que $m = m_0$
- Tenemos una muestra con n datos de la que calculamos $\hat{m}$
- Un valor de $\hat{m}$ será extraño si se aleja mucho de m_0
- Fijamos un nivel de significación e
- Sabemos que $\hat{m} \simeq m + \frac{s}{\sqrt{n}} N(0,1)$
- Si nuestra hipótesis es cierta, $\hat{m} \simeq m_0 + \frac{s}{\sqrt{n}} N(0,1)$
- Calculamos el p-valor como

$$P(|\frac{s}{\sqrt{n}} N(0,1)| > |\hat{m} - m_0|)$$

- Si nuestro p-valor es menor que e rechazamos nuestra hipótesis

Test de hipótesis de igualdad para la media (si s no es conocida)

- Creemos que $m = m_0$
- Tenemos una muestra con n datos de la que calculamos $\hat{m}$
- Un valor de $\hat{m}$ será extraño si se aleja mucho de m_0
- Fijamos un nivel de significación e
- Sabemos que $\hat{m} \simeq m + \frac{s}{\sqrt{n}} t_{n-1}$
- Si nuestra hipótesis es cierta, $\hat{m} \simeq m_0 + \frac{s}{\sqrt{n}} t_{n-1}$
- Calculamos el p-valor como

$$P(|\frac{s}{\sqrt{n}} t_{n-1}| > |\hat{m} - m_0|)$$

- Si nuestro p-valor es menor que e rechazamos nuestra hipótesis

Test de hipótesis para desigualdades de la media (si s es conocida)

- Creemos que $m \leq m_0$
- Tenemos una muestra con n datos de la que calculamos $\hat{m}$
- Un valor de $\hat{m}$ será extraño si es muy grande
- Fijamos un nivel de significación e
- Sabemos que $\hat{m} \simeq m + \frac{s}{\sqrt{n}} N(0,1)$
- Si nuestra hipótesis es cierta, $\hat{m} \simeq m_0 + \frac{s}{\sqrt{n}} N(0,1)$
- Calculamos el p-valor como

$$P(m_0 + \frac{s}{\sqrt{n}} N(0,1) > \hat{m})$$

- Si nuestro p-valor es menor que e rechazamos nuestra hipótesis

Test de hipótesis para desigualdades de la media (si s no es conocida)

- Creemos que $m \leq m_0$
- Tenemos una muestra con n datos de la que calculamos $\hat{m}$
- Un valor de $\hat{m}$ será extraño si es muy grande
- Fijamos un nivel de significación e
- Sabemos que $\hat{m} \simeq m + \frac{s}{\sqrt{n}} \simeq t_{n-1}$
- Si nuestra hipótesis es cierta, $\hat{m} \simeq m_0 + \frac{s}{\sqrt{n}} t_{n-1}$
- Calculamos el p-valor como

$$P(m_0 + \frac{s}{\sqrt{n}} t_{n-1} > \hat{m})$$

- Si nuestro p-valor es menor que e rechazamos nuestra hipótesis

Test de hipótesis de igualdad para una proporción p

- Creemos que $p = p_0$
- Tenemos una muestra con n datos de la que calculamos $\hat{p}$
- Un valor de $\hat{p}$ será extraño si se aleja mucho de p_0
- Fijamos un nivel de significación e
- Sabemos que $\hat{p} \simeq p + \sqrt{\frac{p(1-p)}{n}} N(0,1)$
- Si nuestra hipótesis es cierta, $\hat{p} \simeq p_0 + \sqrt{\frac{p_0(1-p_0)}{n}} N(0,1)$
- Calculamos el p-valor como

$$P(|\sqrt{\frac{p_0(1-p_0)}{n}} N(0,1)| > |\hat{p} - p_0|)$$

- Si nuestro p-valor es menor que e rechazamos nuestra hipótesis

Test de hipótesis para desigualdades de una proporción

- Creemos que $p \leq p_0$
- Tenemos una muestra con n datos de la que calculamos $\hat{p}$
- Un valor de $\hat{p}$ será extraño si es muy grande
- Fijamos un nivel de significación e
- Sabemos que $\hat{p} \simeq p + \sqrt{\frac{p(1-p)}{n}} N(0,1)$
- Si nuestra hipótesis es cierta, $\hat{p} \simeq p_0 + \sqrt{\frac{p_0(1-p_0)}{n}} N(0,1)$
- Calculamos el p-valor como

$$P(p_0 + \sqrt{\frac{p_0(1-p_0)}{n}} N(0,1) > \hat{p})$$

- Si nuestro p-valor es menor que e rechazamos nuestra hipótesis

Intervalos de confianza

Intervalo de confianza para un parámetro

- Un intervalo de confianza para un parámetro x con confianza p es un intervalo (a, b) de modo que la probabilidad de que x esté en ese intervalo sea p
- Debe interpretarse en el sentido de que si repetimos esta construcción varias veces, el porcentaje de veces en que el parámetro esté en este intervalo será p

Intervalo de confianza para la media (si s es conocida)

- En el caso de la media, construiremos el intervalo de confianza en dos pasos
  - Paso 1.- Construimos un intervalo (c, d) de modo que la probabilidad de que una variable $N(0, 1)$ esté en él sea p. Para ello
    - Sea $e = \frac{1-p}{2}$
    - El intervalo viene dado por los percentile e y $1 - e$ de la variable $N(0, 1)$. $c = qnorm(e, 0, 1)$ y $d = qnorm(1 - e, 0, 1)$
  - Paso 2.- Despejamos m
    - $m = \hat{m} + \frac{s}{\sqrt{n}} N(0, 1)$
    - Como $N(0, 1)$ está entre c y d obtenemos que m debe estar entre $\hat{m} + \frac{s}{\sqrt{n}} c$ y $\hat{m} + \frac{s}{\sqrt{n}} d$ (con seguridad p)

Intervalo de confianza para la media (si s no es conocida)

- Si s no es conocida, puede aproximarse su valor por la desviación típica de la muestra $\hat{s}$
- En este caso, debemos modificar la fórmula obteniendo $\frac{m-\hat{m}}{\frac{\hat{s}}{\sqrt{n}}} \simeq t_{n-1}$
- Igual que antes, construiremos el intervalo de confianza en dos pasos
  - Paso 1.- Construimos un intervalo (c, d) de modo que la probabilidad de que una variable t_{n-1} esté en él sea p. Para ello
    - Sea $e = \frac{1-p}{2}$
    - El intervalo viene dado por los percentile e y $1-e$ de la variable t_{n-1}. $c = qt(e, n-1)$ y $d = qt(1-e, n-1)$
  - Paso 2.- Despejamos m
    - $m = \hat{m} + \frac{\hat{s}}{\sqrt{n}} t_{n-1}$
    - Como t_{n-1} está entre c y d obtenemos que m debe estar entre $\hat{m} + \frac{\hat{s}}{\sqrt{n}} c$ y $\hat{m} + \frac{\hat{s}}{\sqrt{n}} d$ (con seguridad p)

Intervalo de confianza para la proporción

- Ahora, construiremos el intervalo de confianza en dos pasos
  - Paso 1.- Construimos un intervalo (c, d) de modo que la probabilidad de que una variable $N(0,1)$ esté en él sea p. Para ello
    - Sea $e = \frac{1-p}{2}$
    - El intervalo viene dado por los percentile e y $1-e$ de la variable $N(0,1)$. $c = qnorm(e, 0, 1)$ y $d = qnorm(1-e, 0, 1)$
  - Paso 2.- Despejamos p
    - $p = \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} N(0,1)$
    - Como $N(0,1)$ está entre c y d obtenemos que p debe estar entre $\hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} c$ y $\hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} d$ (con seguridad p)