Tema 6. Muestreo

Eva Romero Ramos
evarom03@ucm.es

Universidad Complutense de Madrid

Outline

Outline

Conceptos básicos

Definición

La inferencia es la parte de la estadística que se centra en obtener información relevante para la población de interés a partir de la información muestral.

La inferencia engloba por tanto todos los métodos y técnicas que permiten obtener información poblacional a partir de la información muestral.

Algunos de los más importantes son la estimación, que puede ser puntual o por medio de intervalos y los contrastes de hipótesis.

Población

Definición

En estadística se denomina Población al conjunto que incluye todos los elementos que son de interés para el estudio o para los que se desea obtener información.

- La población está por tanto formada por elementos, que pueden ser individuos o personas, empresas, países, productos etc ...
- Los elementos de la población están caracterizados mediante variables aleatorias que se definen por sus distribuciones de probabilidad.
- Llamaremos N al tamaño población y mediante la inferencia buscaremos obtener información válida para todos los elementos de la población.

Muestra

Si disponemos de información sobre los datos de una variable aleatoria para todos los elementos de la población decimos que tenemos un censo.

Pero en la mayoría de las casos será imposible acceder a toda esa información y tendremos que conformarnos con extraer una muestra.

Definición

Una muestra es el subconjunto de elementos de la población a cuyos datos tendremos acceso.

- Una muestra debe ser un buen reflejo de la población a la que pertenece, reflejando sus principales característica de forma rigurosa.
- Si es así diremos que la muestra es representativa.
- Denominaremos n al tamaño muestral y con la información para los n individuos de la muestra, aplicando técnicas de inferencia seremos capaces de inferir conclusiones aplicables a toda la población.

Distribución de la población y de la muestra

Ejemplo

Simularemos 500 valores aleatorios de una distribución normal $N(480; 180^2)$ y asumiremos que dichos valores constituyen la población de interés en referencia a la variable: ***'Gasto anual de las familias en telefonía e internet'****.*

Calcularemos para nuestra población la esperanza y la varianza.

Para esa población obtendremos una muestra aleatoria simple y compararemos la media muestral con la media poblacional.

Debemos observar:

- *¿Son iguales?*
- *¿Son parecidas?*
- *¿Cuál de las dos es fija?*
- *¿Cuál de las dos es una variable aleatoria?*

- En general en una muestra concreta, sus características (momentos, etc.) no tienen por qué coincidir exactamente con las correspondientes de la población.
- Si la muestra ha sido tomada con la máxima objetividad, es de esperar que los valores de las características muestrales no se alejen demasiado de los poblacionales, es por esto que podremos, a partir de la muestra hacer inferencia sobre la población.
- Desde el punto de vista de la estadística frecuentista, los momentos poblacionales son valores desconocidos pero fijos y los momentos muestrales son variables aleatorias, ya que cambiarán si obtenemos una muestra distinta, por lo que tomarán un valor u otro con una determinada probabilidad, siempre que haya aleatoriedad en la extracción de la muestra.

Outline

Muestreo

Definición (Muestreo)

El muestreo es la técnica empleada para seleccionar los elementos de una muestra. Esta técnica debe garantizar que la muestra seleccionada sea lo más representativa posible de la población a la que pertenece.

Hay que tener en cuenta que si un muestreo no se hace de forma correcta, se puede sesgar la muestra.

Existen numerosas técnicas de muestreo que se clasifican en dos grandes bloques:

- Los muestreos probabilísticos.
- Los muestreos no probabilísticos.

Muestreo

Definición (Muestreo probabilístico)

Los muestreos probabilísticos son aquellos que garantizan que todos los individuos tienen la misma probabilidad de estar incluidos en la muestra o desde otro punto de vista, todas las posibles muestras que se pueden extraer de una población la misma probabilidad.

Definición (Muestreo no probabilístico)

Los muestreos no probabilísticos no garantizan la equiprobabilidad de todos los individuos a la hora de entrar en la muestra.

De hecho, en ocasiones podría ocurrir que parte de los individuos que resultan de interés no tengan ninguna probabilidad de estar incluidos en la muestra.

Generalmente se usan cuando no es posible utilizar métodos probabilísticos, habitualmente porque estos resultan extremadamente costosos.

Muestreo aleatorio simple

El **muestreo aleatorio simple** es un muestreo probabilístico, por lo que garantiza que:

- Todos los elementos de la población tienen la misma probabilidad de formar parte de la muestra.
- Todas la muestras de igual tamaño tienen la misma probabilidad.

Desde el punto de vista teórico se asume que el tamaño de la población es infinito, así que en la práctica se debe aplicar a poblaciones suficientemente grandes.

Se puede realizar asignando un valor numérico a cada uno de los elementos de la población y eligiendo tantos números aleatorios como elementos se quieren incluir en la muestra, de forma que sus elementos asociados serán los componentes de la muestra.

Los valores numéricos se deben elegir con reemplazamiento de manera que un elemento podría estar 2 veces en la muestra, pero la probabilidad de que esto suceda es muy pequeña debido al tamaño infinito de la población.

Muestra aleatorio simple

Una muestra aleatoria simple (m.a.s.) de tamańo n de una variable aleatoria X se denota por:

$$(X_1, X_2, \ldots, X_n)$$

Para una m.a.s. se garantiza que todos los elementos X_i son independientes y están idénticamente distribuidos (i.i.d.).

Una m.a.s. de tamaño n de una v.a. X es en la práctica el conjunto de valores obtenidos al observar la variable X en n individuos.

Outline

Estadístico

Definición (Estadístico)

Un estadístico es cualquier función de los elementos muestrales, que no contenga parámetros desconocidos.

- Trataremos con estadísticos muy concretos: los momentos muestrales (media, varianza, covarianza), el valor máximo o mínimo de la muestra, etc...
- Como los elementos que integran la muestra son v.a., cualquier función de estos elementos (es decir, un estadístico), también será variable aleatoria.
- Por eso, el estadístico tendrá su propio campo de variación y su distribución de probabilidad, determinados, a su vez, unívocamente por el campo de variación y la distribución de la población.

Distribución en el muestreo

- El campo de variación del estadístico es el conjunto de valores que toma para cada una de las posibles muestras que se puede obtener.
- Dado que un estadístico se genera en el proceso de muestreo, su distribución de probabilidad recibe el nombre de distribución de probabilidad en el muestreo

Ejemplo

Una v.a. X *presenta los valores* $1, 2, 3$ *con probabilidades* $0.1, 0.2$ *y* 0.7. *Tomamos muestras aleatorias simples de tamaño* 3 *y consideramos como estadístico la media muestral. Obtenga la distribución de la media muestral.*

Solución

Elementos	*Muestras distintas*	*Media Muestral*	*Probabilidad*
$(1, 1, 1)$	1	1	$0.1 \cdot 0.1 \cdot 0.1 = 0.001$
$(1, 1, 2)$	3	1.33	$0.1 \cdot 0.1 \cdot 0.2 = 0.002$
$(1, 1, 3)$	3	1.67	$0.1 \cdot 0.1 \cdot 0.7 = 0.007$
$(1, 2, 3)$	6	2	$0.1 \cdot 0.2 \cdot 0.7 = 0.014$
$(2, 2, 1)$	3	1.67	$0.1 \cdot 0.2 \cdot 0.2 = 0.004$
$(2, 2, 2)$	1	2	$0.2 \cdot 0.2 \cdot 0.2 = 0.008$
$(2, 2, 3)$	3	2.33	$0.2 \cdot 0.2 \cdot 0.7 = 0.028$
$(3, 3, 1)$	3	2.33	$0.7 \cdot 0.7 \cdot 0.1 = 0.049$
$(3, 3, 2)$	3	2.67	$0.7 \cdot 0.7 \cdot 0.2 = 0.098$
$(3, 3, 3)$	1	3	$0.7 \cdot 0.7 \cdot 0.7 = 0.343$

Solución (cont.)

Distribución de probabilidad en el muestreo de la media muestral:

$\overline{x}$	1	1.33	1.67	2	2.33	2.67	3
$f(\overline{x})$	0.001	0.006	0.033	0.092	0.231	0.294	0.334

Esperanza de la media muestral:

$$\mathbb{E}(\overline{x}) = 1 \cdot 0.001 + \cdots + 3 \cdot 0.343 = 2.6$$

Varianza de la media muestral:

$$\mathbb{V}\text{ar}(\overline{x}) = \mathbb{E}\left(\overline{x}^2\right) - [\mathbb{E}(\overline{x})]^2 = 6.9067 - 2.6^2 = 0.1467$$

Esperanza de la población:

$$\mu = 1 \cdot 0.1 + 2 \cdot 0.2 + 3 \cdot 0.7 = 2.6$$

Varianza de la población:

$$\sigma^2 = \mathbb{E}\left(x^2\right) - \mu^2 = 7.2 - 2.6^2 = 0.44$$

Outline

Momentos muestrales con respecto al origen

- Sea cual sea la distribución de la variable aleatoria poblacional, la esperanza matemática de cada momento muestral (α_r) respecto al origen es igual al correspondiente momento poblacional (M_r).

$$\mathbb{E}\left(\alpha_r\right) = M_r$$

- Esto sucede aunque la muestra no sea aleatoria simple, es decir, aunque sus elementos no sean independientes.
- Para la varianza tenemos:

$$\mathbb{V}\mathrm{ar}\left(\alpha_r\right) = \frac{1}{n}\left(\alpha_{2r} - \alpha_r^2\right)$$

- El resultado obtenido exige que las variables aleatorias sean independientes, es decir, la muestra tiene que ser aleatoria simple

Momentos muestrales con respecto al origen

- Media muestral

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \alpha_1$$

- Su esperanza es

$$\mathbb{E}\left(\overline{x}\right) = \mathbb{E}\left(\alpha_1\right) = M_1 = \mu$$

- Su varianza es

$$\mathbb{V}\mathrm{ar}\left(\overline{x}\right) = \mathbb{V}\mathrm{ar}\left(\alpha_1\right) = \frac{1}{n}\left(\alpha_2 - \alpha_1^2\right) = \frac{\sigma^2}{n}$$

Momentos muestrales con respecto a la media

- El valor esperado de los momentos muestrales respecto a la media no coincide con sus respectivos momentos poblacionales

$$\mathbb{E}\left(m_r\right) = \mu_r + O\left(\frac{1}{n}\right)$$

- La diferencia será menor cuanto mayor sea el tamaño de la muestra

Momentos muestrales con respecto a la media

- Varianza muestral

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- Esperanza matemática de la varianza muestral:

$$\mathbb{E}\left(S^2\right) = \frac{n-1}{n}\sigma^2$$

- La esperanza de la varianza muestral es casi la varianza poblacional
- Varianza de la varianza muestral:

$$\mathbb{V}\text{ar}\left(S^2\right) = \frac{\mu_4 - \mu_2^2}{n} - \frac{2\left(\mu_4 - 2\mu_2^2\right)}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}$$

Momentos muestrales con respecto a la media

- Cuasivarianza muestral:

$$\boxed{S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} \Longrightarrow S_1^2 = \frac{n}{n-1}S^2$$

- Estadístico muy importante en inferencia
- Esperanza matemática de la cuasivarianza muestral:

$$\mathbb{E}\left(S_1^2\right) = \sigma^2$$

- Varianza de la cuasivarianza muestral:

$$\mathbb{V}\mathrm{ar}\left(S_1^2\right) = \frac{1}{\left(n-1\right)^2}\left[\left(n-2\right)\mu_4 - \left(n-4\right)\mu_2^2 + \frac{\mu_4 - 3\mu_2^2}{n}\right]$$

Distribución normal

- La varianza de la varianza muestral y de la cuasivarianza muestral se simplifican mucho cuando la población sigue una distribución normal:

$$\mathbb{V}\text{ar}\left(S^2\right) = \frac{2\left(n-1\right)\sigma^4}{n^2}$$

$$\mathbb{V}\text{ar}\left(S_1^2\right) = \frac{2\sigma^4}{n-1}$$

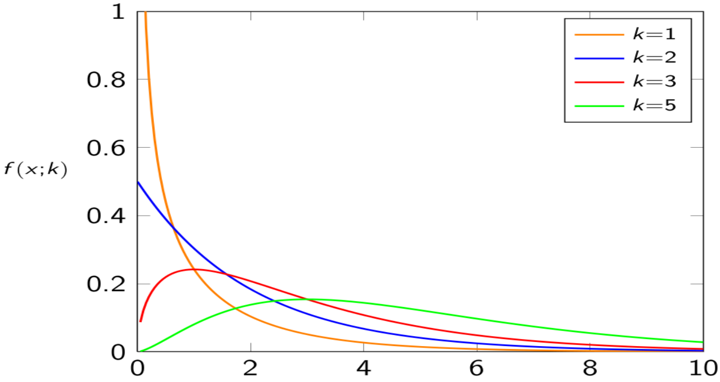
Outline

Motivación

- Sean n variables aleatorias independientes $Z_1, Z_2, \ldots, Z_n$ distribuidas $\mathcal{N}(0; 1)$.
- Se define la variable $X = Z_1^2 + Z_2^2 + \cdots + Z_n^2$ que recibe el nombre de $\chi^2(n)$.
- El número de variables aleatorias que la integran, n, se denomina grados de libertad.
- Como el campo de variación de las variables normales es el intervalo $(-\infty; \infty)$ y la variable $\chi^2(n)$ es suma de sus cuadrados, su campo es el intervalo: $[0; \infty)$.

Función de densidad

$$f(x;k)=\frac{1}{2^{k/2}\Gamma(k/2)}x^{(k/2)-1}e^{-x/2}, \quad x>0$$



Propriedades

- La función de densidad queda completamente definida con sus grados de libertad n.
- La forma de la función cambia con el valor que toma n, haciendose más simétrica cuando n aumenta.
- Para valores intermedios, la función de densidad es asimétrica positiva.
- Sólo tienen densidad los valores positivos y los valores de la "cola derecha" normalmente se consideran como anómalos.
- Esperanza: $\boxed{\mathbb{E}\left(X\right) = n}$
- Varianza: $\boxed{\mathbb{V}\text{ar}\left(X\right) = 2n}$

Propriedades

- Propiedad aditiva o reproductiva: sean k v.a. independientes $X_j \sim \chi^2(n_j)$ para $j = 1, \ldots, k$. Entonces

$$Y = X_1 + \cdots + X_k \sim \chi^2(n_1 + \cdots n_k)$$

- Cuando el número de grados de libertad es suficientemente grande, la χ^2 converge a una $N(0,1)$:

$$\sqrt{2\chi^2(n)} - \sqrt{2n-1} \xrightarrow[n\to\infty]{N} (0,1)$$

Manejo de tablas

- La tabla nos da la probabilidad del suceso $\{\chi^2(n) \geq a\}$
- La mecánica del cálculo es similar al utilizado para la $\mathcal{N}(0;1)$, con las siguientes excepciones:
    - No es simétrica.
    - No admite valores negativos.

Ejemplo

Hállese $\Pr\left[3.94 \leq \chi^2(10) \leq 15.987\right]$.

Ejemplo

Hállese $\Pr\left[3.94 \leq \chi^2(10) \leq 15.987\right]$.

Solución

$$\begin{aligned}
\Pr\left[3.94 \leq \chi^2(10) \leq 15.987\right] &= \Pr\left[\chi^2(10) \geq 3.94\right] - \\
&- \Pr\left[\chi^2(10) \geq 15.987\right] = \\
&= 0.95 - 0.1 = 0.85
\end{aligned}$$

Ejemplo

Si $\Pr\left[8.907 \leq \chi^2(19) \leq a\right] = 0.725$ *hállese* a.

Ejemplo

Si $\Pr\left[8.907 \leq \chi^2(19) \leq a\right] = 0.725$ *hállese* a.

Solución

$$\begin{aligned}
\Pr\left[8.907 \leq \chi^2(19) \leq a\right] &= \Pr\left[\chi^2(19) \geq 8.907\right] - \\
&- \Pr\left[\chi^2(19) \geq a\right] = \\
&= 0.975 - \Pr\left[\chi^2(19) \geq a\right] = 0.725
\end{aligned}$$

$$\Longrightarrow \Pr\left[\chi^2(19) \geq a\right] = 0.25 \Longrightarrow a = 22.7$$

Outline

Motivación

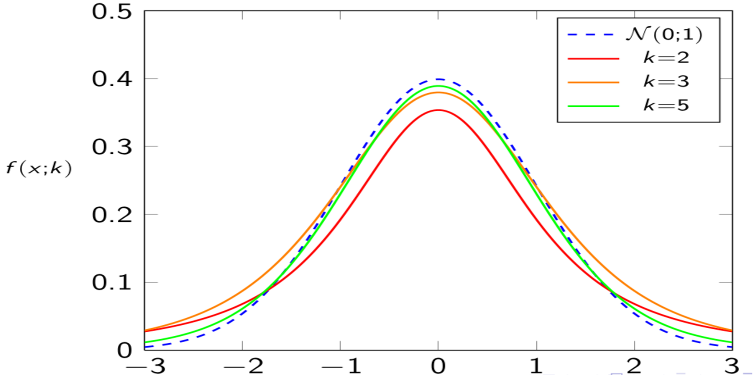
- Sean $n+1$ v.a. $\mathcal{N}(0;\sigma)$ e independientes $(X; X_1, X_2, \ldots, X_n)$.
- Definimos la variable t de Student con n grados de libertad, llamada $t(n)$, como:

$$\boxed{t(n) = \frac{X}{\sqrt{\dfrac{X_1^2 + X_2^2 + \cdots + X_n^2}{n}}}}$$

- El número de grados de libertad es igual al número de variables que figuran en el denominador de $t(n)$.
- El campo de variación es el eje real.
- La función es simétrica respecto al eje de ordenadas y tiene dos asíntotas en $-\infty$ y ∞.
- Se consideran valores anómalos los que se alejan de cero (positivos o negativos).

Función de densidad

$$f(x;k) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)}\left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}}$$



Importancia y propriedades

- La importancia de esta v.a. reside en el hecho de que su función de densidad no depende de la varianza de las variables que la integran, y su utilidad se verá plenamente cuando se estudie la Inferencia Estadística en los casos donde la varianza se desconoce.
- Esperanza: $\boxed{\mathbb{E}\left(X\right) = 0}$
- Varianza: $\boxed{\mathbb{V}\mathrm{ar}\left(X\right) = \dfrac{n}{n-2}}$ para $n > 2$

Manejo de tablas

- Las tablas que utilizaremos de la distribución $t(n)$ proporcionan la probabilidad del suceso $t(n) \geq a$.
- Para el manejo de las tablas es importante recordar la simetría de la función de densidad.
- Por otra parte, y teniendo en cuenta las diferencias con la tabla de distribución $\mathcal{N}(0;1)$ la mecánica es similar, recurriendo cuando se necesite a la interpolación lineal.
- Cuando el número de grados de libertad es elevado (por ejemplo, mayor que 30), la distribución $t(n)$ es aproximadamente $\mathcal{N}\left(0; \sqrt{\frac{n}{n-2}}\right)$.

Ejemplo

Hállese $\Pr\left[t\left(7\right) \leq 1.1192\right]$.

Ejemplo

Hállese $\Pr\left[t\left(7\right) \leq 1.1192\right]$.

Solución

$$\begin{aligned} \Pr\left[t\left(7\right) \leq 1.1192\right] &= 1 - \Pr\left[t\left(7\right) > 1.1192\right] = \\ &= 1 - 0.15 = 0.85 \end{aligned}$$

Ejemplo

Si $\Pr\left[a \leq t\left(16\right) \leq 0.865\right] = 0.7$ *hállese* a.

Ejemplo

Si $\Pr\left[a \leq t\left(16\right) \leq 0.865\right] = 0.7$ *hállese* a.

Solución

$$\begin{aligned}\Pr\left[a \leq t\left(16\right) \leq 0.865\right] &= \Pr\left[t\left(16\right) > a\right] - \Pr\left[t\left(16\right) > 0.865\right] = \\ &= \Pr\left[t\left(16\right) > a\right] - 0.2 = 0.7\end{aligned}$$

$$\Longrightarrow \Pr\left[t\left(16\right) > a\right] = 0.9 \Longrightarrow a \textit{ es negativo}$$

$$\Longrightarrow \Pr\left[t\left(16\right) > -a\right] = 0.1 \Longrightarrow a = -1.337$$

Ejemplo

Si $\Pr\left[-0.686 \leq t\,(22) \leq a\right] = 0.6$ *hállese* a.

Ejemplo

Si $\Pr\left[-0.686 \leq t\left(22\right) \leq a\right] = 0.6$ *hállese* a.

Solución

$$\begin{aligned} \Pr\left[-0.686 \leq t\left(22\right) \leq a\right] &= \Pr\left[t\left(22\right) > -0.686\right] - \Pr\left[t\left(22\right) > a\right] = \\ &= 1 - \Pr\left[t\left(22\right) > 0.686\right] - \Pr\left[t\left(22\right) > a\right] = \\ &= 1 - 0.25 - \Pr\left[t\left(22\right) > a\right] = 0.6 \end{aligned}$$

$$\Longrightarrow \Pr\left[t\left(22\right) > a\right] = 0.15 \Longrightarrow a = 1.061$$

Ejemplo

Hállese $\Pr\left[t\left(507\right) \geq 0.4\right]$.

Ejemplo

Hállese $\Pr\left[t\left(507\right) \geq 0.4\right]$.

Solución

$$\Pr\left[t\left(507\right) \geq 0.4\right] \approx \Pr\left[\mathcal{N}\left(0;1\right) \geq 0.4\right] = 0.3446$$

Outline

Motivación

- Consideremos $m+n$ v.a. $\mathcal{N}(\mu;\sigma)$ e independientes $Y_1, Y_2, \ldots, Y_m$ y $T_1, T_2, \ldots, T_n$, la variable:

$$\boxed{F(m;n) = \frac{\dfrac{Y_1^2 + Y_2^2 + \cdots + Y_m^2}{m}}{\dfrac{T_1^2 + T_2^2 + \cdots + T_n^2}{n}}}$$

sigue la distribución F de Fisher-Snedecor con m y n grados de libertad

Motivación

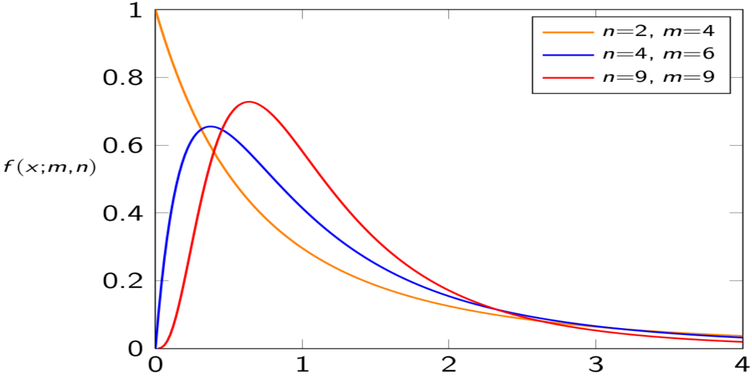
- Consideremos $m+n$ v.a. $\mathcal{N}(\mu;\sigma)$ e independientes $Y_1, Y_2, \ldots, Y_m$ y $T_1, T_2, \ldots, T_n$, la variable:

$$\boxed{F(m;n) = \frac{\dfrac{Y_1^2+Y_2^2+\cdots+Y_m^2}{m}}{\dfrac{T_1^2+T_2^2+\cdots+T_n^2}{n}}}$$

  sigue la distribución F de Fisher-Snedecor con m y n grados de libertad
- Campo de variación: $X \geq 0$

Función de densidad

$$f(x;m,n)=\left(\frac{m}{n}\right)^{\frac{m}{2}}\frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\frac{x^{\frac{m}{2}-1}}{\left(1+\frac{m}{n}x\right)^{\frac{m+n}{2}}}$$



Características I

- No depende de la varianza de las variables integrantes.
- Puede expresarse como cociente de dos χ^2 que no dependen de σ^2:

$$F(m;n) = \frac{\dfrac{Y_1^2 + Y_2^2 + \cdots + Y_m^2}{m}}{\dfrac{T_1^2 + T_2^2 + \cdots + T_n^2}{n}} = \frac{n}{m}\frac{\chi^2(m)}{\chi^2(n)}$$

- Esperanza: $\boxed{\mathbb{E}(X) = \dfrac{n}{n-2}}$ para $n > 2$.
- Varianza: $\boxed{\mathbb{V}\mathrm{ar}(X) = \dfrac{2n^2(m+n-2)}{m(n-2)^2(n-4)}}$ para $n > 4$.

Características II

- La distribución no es simétrica.
- Su campo de variación, al proceder de una suma de cuadrados, es el intervalo $[0; \infty)$ y tiene una asíntota en ∞.
- Se consideran valores anómalos los de la cola de la derecha.
- Si una v.a. Y sigue la distribución $F(m; n)$, su inversa, $1/Y$ es $F(n; m)$, es decir:

$$\boxed{F(n; m; \alpha) = \frac{1}{F(m; n; 1-\alpha)}}$$

  donde α es la probabilidad que deja Y a la derecha
- Si $m = 1$ entonces $F(1; n) = t^2(n)$

Manejo de tablas

- Las tablas proporcionan $\Pr\left[F\left(m;n\right)\geq a\right]$.
- La presentación de las tablas de esta distribución es distinta de las anteriores, debido a la existencia de dos parámetros en la función de densidad.
- En general, las tablas se utilizan en el caso inverso, es decir, conocida la probabilidad del suceso $F\left(m;n\right)\geq a$ hallar el valor de a.

Ejemplo

Si $\Pr\left[F\left(1;10\right)\geq a\right]=0.05$ *hállese* a. *Hállese* $F\left(5;30;0.01\right)$. *Hállese* $F\left(9;40;0.05\right)$. *Si* $\Pr\left[F\left(5;10\right)\geq a\right]=0.99$ *hállese* a.

Ejemplo

Si $\Pr\left[F\left(1;10\right) \geq a\right] = 0.05$ *hállese* a. *Hállese* $F\left(5;30;0.01\right)$. *Hállese* $F\left(9;40;0.05\right)$. *Si* $\Pr\left[F\left(5;10\right) \geq a\right] = 0.99$ *hállese* a.

Solución

$$\Pr\left[F\left(1;10\right) \geq a\right] = 0.05 \Longrightarrow a = 4.96$$

$$F\left(5;30;0.01\right) = 3.70$$

$$F\left(9;40;0.05\right) = 2.12$$

$$F\left(5;10;0.99\right) = \frac{1}{F\left(10;5;0.01\right)} = \frac{1}{10.05} = 0.0995$$

Outline

Introducción

- Anteriormente hemos obtenido ciertos momentos muestrales sin indicar nada respecto a su distribución concreta en el muestreo, pues se ignoraba cuál era la de la población de la que procedía la muestra.
- Ahora estudiaremos qué ocurre en el caso particular de que la población sea normal.
- Sea una variable aleatoria X con distribución $\mathcal{N}\left(\mu;\sigma^2\right)$.
- Se toma una muestra aleatoria simple de tamaño n $(X_1, X_2, \ldots, X_n)$.
- En una m.a.s., cada elemento de la muestra es una variable aleatoria con la misma distribución que X (en este caso son, por tanto, v.a. normales $\mathcal{N}\left(\mu;\sigma^2\right)$).

Distribución de la media muestral

- La media muestral es combinación lineal de variables aleatorias normales e independientes $(X_1, X_2, \ldots, X_n)$, por ser la muestra aleatoria simple, y por tanto, la combinación lineal sigue una distribución normal

$$\boxed{\overline{X} \sim \mathcal{N}\left(\mu; \frac{\sigma^2}{n}\right)}$$

Lema de Fisher

Distribución de la varianza muestral

- Las variables aleatorias media y varianza muestrales ($\overline{X}$ y S^2) son estadísticamente independientes.
- El estadístico:

$$\boxed{\frac{nS^2}{\sigma^2} \sim \chi^2\left(n-1\right)}$$

- Podemos interpretar este resultado de la siguiente manera: la variable $\frac{nS^2}{\sigma^2}$ es igual a la suma de los cuadrados de $n-1$ variables aleatorias $\mathcal{N}\left(0;1\right)$ e independientes.

Varianza poblacional desconocida

Distribución de la media muestral

- Para obtener la distribución muestral desconociendo σ^2, es preciso recurrir a una distribución independiente de la varianza σ^2, y esta distribución es la t de Student

$$\boxed{\frac{\sqrt{n}\left(\overline{X}-\mu\right)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}}=\frac{\sqrt{n}\left(\overline{X}-\mu\right)}{S_1}\sim t\left(n-1\right)}$$

Ejemplo

Los beneficios mensuales de un conjunto de empresas de desarrollo de aplicaciones informáticas se distribuyen según una distribución normal de media de 10000 *euros y varianza* 500^2. *Disponemos de una muestra aleatoria simple de* 16 *empresas de este sector, con la cual calculamos la media muestral y la varianza muestral.*

a) Calcula la probabilidad de que la media muestral sea superior a 9800 *euros.*
b) Calcula la misma probabilidad asumiendo que la varianza poblacional es desconocida, pero la cuasivarianza muestral es 745^2.
c) Calcula la probabilidad de que la varianza muestral sea superior a 625^2, *asumiendo una varianza poblacional conocida.*

Ejemplo

a) Calcula la probabilidad de que la media muestral sea superior a 9800 *euros.*

Solution

Sabemos que con varianza conocida $\overline{X} \sim \mathcal{N}\left(\mu; \frac{\sigma^2}{n}\right) = \mathcal{N}\left(10000; \frac{500^2}{16}\right)$

$$\begin{aligned}\Pr\left(\overline{X} > 9800\right) &= \Pr\left(Z > \frac{9800 - 10000}{500/4}\right) = \Pr\left(Z > -1.6\right) = \\ &= 1 - \Pr\left(Z > 1.6\right) = 1 - 0.0548 = 0.9452\end{aligned}$$

Ejemplo

b) Calcula la misma probabilidad asumiendo que la varianza poblacional es desconocida, pero la varianza muestral es 745^2.

Solution

Sabemos que con varianza desconocida $\frac{\sqrt{n}(\overline{X}-\mu)}{S_1} \sim t(n-1) \Longrightarrow \frac{\sqrt{16}(\overline{X}-10000)}{745} \sim t(15)$

$$\begin{aligned}
\Pr\left(\overline{X} > 9800\right) &= \Pr\left(t(15) > \frac{\sqrt{16}\,(9800-10000)}{745}\right) = \\
&= \Pr\left(t(15) > -1.074\right) = \\
&= 1-\Pr\left(t(15) > 1.074\right) = 1-0.15 = 0.85
\end{aligned}$$

Ejemplo

c) Calcula la probabilidad de que la varianza muestral sea superior a 625^2*, asumiendo una varianza poblacional conocida.*

Solution

Sabemos que con varianza conocida $\frac{nS^2}{\sigma^2} \sim \chi^2(n-1) \Longrightarrow \frac{16S^2}{500^2} \sim \chi^2(15)$

$$\begin{aligned} \Pr\left(S_1^2 > 625^2\right) &= \Pr\left(\chi^2(15) > \frac{16 \cdot 625^2}{500^2}\right) = \\ &= \Pr\left(\chi^2(15) > 25\right) = 0.05 \end{aligned}$$