Temas de Estadística Práctica

Antonio Roldán Martínez

Proyecto http://www.hojamat.es/

Muestreo aleatorio simple

Resumen teórico de los principales conceptos estadísticos

Muestreo aleatorio simple

Definiciones

Cuando el colectivo que se pretende estudiar es muy extenso o inaccesible, se recurre a un subconjunto del mismo llamado ***muestra***, y al conjunto de técnicas usadas se le denomina ***muestreo***.

Población

Es el conjunto de referencia que pretendemos estudiar, formado por elementos que comparten una misma propiedad: Españoles adultos, alumnos de la Enseñanza Privada de Madrid, fresnos existentes en la Sierra de Guadarrama.

Censo

Si es posible estudiar toda la población, por ejemplo, los alumnos de un colegio, a este estudio le llamaremos ***censo***. Un censo no siempre es posible, especialmente por motivos económicos.

Muestra

Una ***muestra*** es un subconjunto de la población, y es el que verdaderamente se estudia en la inmensa mayoría de los experimentos y estudios. Se debe acudir a muestras cuando la población es demasiado numerosa (*población infinita*), o bien resulta muy caro un estudio exhaustivo. Otro motivo suele ser que el experimento requiera pruebas destructivas, y no es caso de destruir la población.

Una muestra es ***representativa*** cuando tiene una estructura y unos parámetros muy parecidos a la población. Desgraciadamente, esta definición no es útil, pues generalmente no se conoce con seguridad la población, o existe la sospecha de que sus características hayan cambiado. Llamaremos ***muestreo*** al conjunto de técnicas que nos ayudan a elegir una muestra representativa.

Muestreo

La operación de elegir una muestra puede ser tan compleja que llena libros enteros. Aquí sólo repasaremos las técnicas de muestreo más frecuentes;

Aleatorio: Una muestra es aleatoria cuando su elección se hace depender del azar. En concreto, si todos los elementos de la muestra han tenido las mismas oportunidades de ser elegidos, diremos que constituye una ***muestra aleatoria simple (m.a.s.)***. Esta es la muestra que consideraremos aquí.

Intencional: Se llama así cualquier técnica que dependa de la libre voluntad del experimentador, sin recurso al azar.

Errática: Una muestra errática es la que nos encontramos ya formada, sin intervención nuestra, como puede ser el conjunto de alumnos asignados al principio de curso.

Distribuciones en el muestreo

Es fácil confundir las distintas distribuciones estadísticas que concurren en el muestreo. Fundamentalmente son tres:

Distribución en la población: Es el conjunto de frecuencias y medidas que se dan en la población. Salvo mediante un censo, esta distribución sólo se conoce aproximadamente. Las medidas tomadas en la población se llaman ***parámetros***. Los más importantes son

* la media μ

* la desviación típica σ

* cualquier proporción P

* su tamaño N

Distribución en la muestra

Es el conjunto de características de la **muestra concreta que hemos elegido**. Su parecido a la de la población depende totalmente del azar: podemos elegir una muestra representativa sin saberlo, o elegir una muestra sesgada por pura mala suerte. Sus medidas se llaman ***estadísticos***. Los más importantes son

* la media $\bar{X}$

* la desviación típica S

* cualquier proporción p

* su tamaño n

Distribución muestral

Es la resultante de considerar, de forma teórica, **todas las posibles muestras que se puedan elegir**. Es una distribución teórica, construida sobre variables aleatorias, y sus elementos se obtienen mediante técnicas matemáticas. A la media de cualquier estadístico teórico D la representaremos por m_D y a su desviación típica $\mathbf{s_D}$.

También usaremos el lenguaje de las variables aleatorias: E(D) representa la media, VAR(D) a la varianza y DESV(D) a la desviación típica.

Principales distribuciones muestrales

La teoría que sigue no contiene justificaciones matemáticas de las propiedades que figuran en ella. Todas se pueden demostrar, algunas con técnicas elementales y otras mediante teoremas del límite. Remitimos a textos especializados en Estadística Inferencial.

Distribución muestral de la media

Media: La media de todas las medias muestrales cocincide con la de la población. Es decir, si elegimos muchas muestras distintas, no todas tendrán la misma media que la población; incluso muchas de ellas la tendrán muy alejada. No obstante, si pudiéramos considerar **todas las muestras**, el promedio de todas las medias coincidiría con la media de la población:

$$E(\overline{X}) = \mu$$

por tener esta propiedad, diremos que la media es un estimador insesgado.

Varianza: La varianza de la media tiene, en principio, una distribución más complicada;

$$VAR(\bar{X}) = \frac{\sigma^2}{n}\sqrt{\frac{N-n}{N-1}}$$

La expresión se simplifica si la población es infinita, pues en ese caso la raíz cuadrada tiende a 1, y nos queda una expresión más simple.

$$VAR(\bar{X}) = \frac{\sigma^2}{n}$$

Este resultado es muy interesante: **Cuanto mayor sea el tamaño de la muestra, más pequeña será la varianza de la media, lo que minimizará los errores.**

Podemos deducir de la fórmula anterior la expresión de la desviación típica del estimador media, y obtendríamos

$$e = \sqrt{\frac{\sum (x-\bar{x})^2}{N \cdot (N-1)}}$$

también llamado ***error muestral*** o ***error de estimación.***

Distribución muestral: Para saber cómo se distribuye la media deberemos distinguir varios casos:

* Si la distribución de la población es **Normal**, y se conoce la **s** de la población, la de la media muestral también será **normal**.
* Si la muestra es **de tamaño mayor o igual que 30**, y se

conoce la **s** de la población, aunque la población no sea normal, la media de la muestra sí se comportará como **normal**. Este hecho fundamental se conoce por el nombre de **Teorema Central del Límite**.

* Si la población es aproximadamente normal, y **no se conoce la s** de la población, en muestras grandes (n>120) puede usarse la distribución normal, de forma aproximada, pero en muestras más pequeñas hay que acudir a la **Distribución T de Student**.

Distribución muestral de la proporción

La proporciones **p** en las muestras forman una distribución binomial. Si llamamos **P** a la proporción equivalente en la población, la distribución muestral, para poblaciones infinitas, queda:

E(p) = P

por tener esta propiedad, diremos que la proporción es un estimador insesgado.

Es decir, la media de la proporción de las muestras coincide con la proporción en la población.

VAR(p) = PQ/n , llamando **Q** a **1-P**

Como en la media, el aumento del tamaño disminuye los

errores.

Si n<30, la proporción sigue la distribución binomial.

Si n>=30, se puede aproximar a la normal.

Si P no se conoce, en la fórmula de la varianza PQ/n podemos sustituir P y Q por p y q, con un pequeño error. Más aún, en la práctica se puede tomar como p y q el valor 1/2, que se puede demostrar daría el error máximo. Así, la varianza quedaría como **VAR(p) < 1/(4n)**. Esta fórmula es muy útil en la práctica.

Distribución muestral de la varianza

La varianza de las muestras sigue un proceso distinto a los de la media y proporción. La causa es que el promedio de todas las varianzas de las muestras no coincide con la varianza de la población □**2**. Se queda un poco por debajo. En concreto, se verifica que

$$E(S_n^2) = \frac{n-1}{n}\sigma^2$$

Hemos usado el subíndice n para recordar que en la varianza se divide entre n.

Si deseamos que la media de la varianza coincida con la varianza de la población, tenemos que acudir a la

cuasivarianza o varianza insesgada, que es similar a la varianza, pero dividiendo las sumas de cuadrados entre n-1.

$$S_{n-1}^2 = \frac{\sum (x - \bar{x})^2}{n-1}$$

Su raíz cuadrada es la cuasidesviación típica o desviación estándar.

Si se usa esta varianza, si coinciden su media y la varianza de la población

$$E(S_{n-1}^2) = \sigma^2$$

lo que nos indica que la cuasivarianza es un estimador insesgado, y la varianza lo es sesgado.

Distribución muestral de la varianza

La suma de cuadrados de la varianza, dividida entre la varianza de la población

$$\frac{\sum (x - \bar{x})^2}{\sigma^2} = \frac{nS_n^2}{\sigma^2} = \frac{(n-1)S_{n-1}^2}{\sigma^2}$$

se distribuye según una **chi-cuadrado χ^2 con n-1 grados de libertad**

Estimación

Es la operación mediante la cual identificamos el valor de un **parámetro** de la población con el valor de un **estadístico** de la muestra. Es como un acto de confianza: suponemos que la estructura de la muestra permite que sus medidas sean también las de la población. Puede ser una operación arriesgada.

Estimación puntual

La estimación se llama **puntual** cuando identificamos, sin más, el parámetro con el estadístico. En ese caso añadiremos un acento circunflejo al parámetro para representar que estamos estimando.

Un estimador es **insesgado** cuando su media muestral coincide con el parámetro. Así, son insesgadas (y recomendables) estas estimaciones:

$$\hat{\mu} = \overline{X}$$

El estimador insesgado de la media de la población es la media de la muestra

$$\hat{P} = p$$

El estimador insesgado de la proporción es la proporción de la muestra

$$\hat{\sigma}^2 = S_{n-1}^2$$

El estimador insesgado de la varianza no es la varianza de la población, sino la cuasivarianza.

Estimación por intervalos

Al ser la estimación una operación arriesgada (¿cuándo aciertan totalmente las encuestas políticas?), en lugar de apostar por una estimación puntual, se rodea esta de un intervalo de seguridad, lo que la prensa llama "la horquilla", que técnicamente es el **Intervalo de confianza**.

Para construir un intervalo de confianza, además de la elección del estimador, debemos fijar el **nivel de confianza**, que para no correr riesgos, se suele tomar como una probabilidad grande: 95%, 96%, 99%…

A este nivel de confianza lo representaremos por **1 -a**.

Su significado intuitivo es que si repitiéramos muchas veces un experimento con un nivel de confianza, pongamos el 95%, sólo corremos el riesgo de equivocarnos en la estimación un 5% de las veces, mientras acertaríamos un 95%. Así, el símbolo **a** representa el riesgo de que la estimación sea errónea.

Una vez elegido el nivel, sabiendo las distribuciones muestrales, se puede rodear al estimador de todo un intervalo en el que existe una probabilidad 1 - a de que se encuentre en su interior el parámetro estimado.

Los intervalos más populares son (para muestras con n>=30)

Intervalo para la media

$$\left(\bar{x} + z_{\alpha/2} \cdot \sigma/\sqrt{n}, \bar{x} + z_{1-\alpha/2} \cdot \sigma/\sqrt{n}\right)$$

Los valores de **z** son uno negativo y otro positivo, por lo que rodean la media. Corresponden a la distribución normal.

σ es la desviación típica de la población, supuesta conocida y **n** el número de elementos de la muestra.

Si no es conocida, recurriríamos a la **t de Student** o a la normal si la muestra es mayor que 120.

Estos casos los puedes consultar en los manuales.

Intervalo para la proporción

$$\left(p + z_{\alpha/2} \cdot \sqrt{\frac{pq}{n}}, p + z_{1-\alpha/2} \cdot \sqrt{\frac{pq}{n}} \right)$$

Los significados de **z, p, q** y **n** ya están explicados con anterioridad.

Intervalo para la varianza

$$\left(\frac{nS_n^2}{\chi^2_{1-\alpha/2}}, \frac{nS_n^2}{\chi^2_{\alpha/2}} \right)$$

donde la **chi-cuadrado** se toma con n-1 grados de libertad

Distribuciones en la Regresión y Correlación

En las estimaciones correspondientes a la Regresión lineal se admite como hipótesis el siguiente modelo teórico:

Se supone que en la población se han medido dos variables X e Y, que están relacionadas siguiendo estas hipótesis:

(1) - $\mathbf{Y_i = a + bX_i + e_i}$ **,** donde **a** y **b** son parámetros de la población (ordenada en el origen y pendiente) y $\mathbf{e_i}$ es el error de cada observación respecto al modelo lineal

(2) La media de los errores $\mathbf{e_i}$ es cero. La varianza de los errores $\mathbf{e_i}$ coincide con la de la población.

(3) Los errores de las observaciones son independientes entre sí.

Designaremos por **r** al coeficiente real de correlación entre X e Y que presenta la población estudiada.

Estimadores

Llamaremos A al estimador de **a** , B al de **b** , y R al del coeficiente de correlación **r**

Estimador B de la pendiente b

La fórmula del estimador B de la pendiente presenta es:

$$B = \frac{N\sum X_i Y_i - \sum X_i \sum Y_i}{N\sum X_i^2 - \sum X_i \sum X_i}$$

que en realidad es un desarrollo de la que se estudió en el Tema 5 y equivale al cociente entre la covarianza y la varianza de X

$$B = \frac{S_{xy}}{S_x^2}$$

Estimador de la ordenada en el origen a

La fórmula del estimador A de la ordenada en el origen es, como en el Tema 5:

$$A = \bar{Y} - B\bar{X}$$

Estimador de la varianza

La varianza se estima mediante

$$S^2 = \frac{\sum (Y - Y')^2}{N - 2}$$

N-2 son los grados de libertad y el numerador equivale a la suma de los cuadrados de las diferencias entre los valores de Y y sus pronósticos.

Estimador del coeficiente de correlación r

También nos vale la clásica fórmula de Pearson.

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

que equivale al cociente de la covarianza entre las dos desviaciones típicas (X e Y).

Distribuciones de los estimadores

Estimador B

La varianza del estimador de la pendiente B viene dada por la expresión

$$VAR(B) = \frac{\sigma^2}{\sum (X_i - \overline{X})^2}$$

Si suponemos que la población es normal y su varianza conocida, el estimador B también seguirá una distribución normal. Si la varianza es desconocida, su distribución será la T de Student, y se deberá sustituir la varianza por su estimador S^2.

Estimador A

El estimador A posee una varianza algo más complicada de calcular

$$VAR(A) = \frac{\sigma^2 . \sum X^2}{N \sum (X_i - \overline{X})^2}$$

También A se distribuye normalmente o mediante la T de Student, según sea conocida o no la varianza de la población. En este último caso se deberá sustituir la varianza por su estimador S^2.

Estimador S^2

El cociente

$$\chi^2 = \frac{S^2(N-2)}{\sigma^2}$$

se distribuye según una χ2 con N-2 grados de libertad

Estimador r

El cociente $\chi^2 = \frac{S^2(N-2)}{\sigma^2}$

sigue una T de Student con N-2 grados de libertad. El valor de T puede dar una idea de si **r** es significativamente distinto de cero.

Si se aplica al coeficiente **r** la transformación de Fisher

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r}$$

el estadístico resultante se distribuye de forma aproximadamente normal con una varianza igual a 1/(N-3)

Se puede usar esta transformación para construir un intervalo de confianza para el coeficiente de correlación.