Inferencia estadística

1. Muestras

En ocasiones, resulta imposible efectuar medidas sobre todos los objetos de una población, bien debido a que su tamaño lo hace imposible, bien porque el proceso de medida es destructivo (por ejemplo cuando se mide la duración de una población de bombillas) o por otras razones. En este caso se toma una muestra y a partir de los resultados de las medidas efectuadas sobre la muestra se trata de sacar conclusiones sobre la población.

Estas conclusiones tienen necesariamente un carácter probabilístico y por tanto, junto al dato poblacional que se deduzca habrá que especificar cuál es la probabilidad de que sea erróneo.

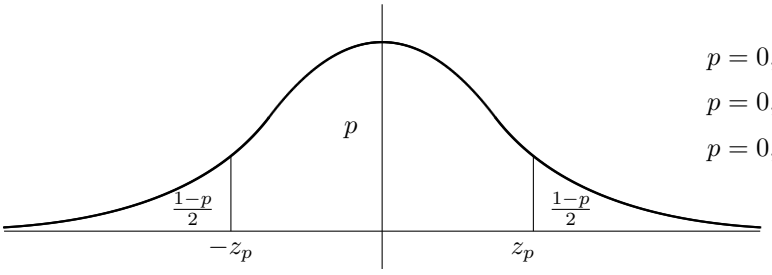
Para poder sacar conclusiones es esencial que el **muestreo** es decir, el proceso de obtención de la muestra sea **aleatorio**. El muestreo es aleatorio si todos los elementos de la población tienen la misma probabilidad de ser elegidos.

El muestreo puede ser **simple** o **estratificado**:

- El muestreo es simple si todas las muestras son posibles. Esto se puede hacer, por ejemplo, numerando los elementos, introduciendo los números en una urna y extrayendo números al azar.
- A veces, los elementos de la población pueden dividirse en varias categorías (por ejemplo por edades). Puede ser conveniente considerar solamente las muestras que respeten la proporción de cada categoría. Cuando las muestras se obtienen de esta forma, el muestreo se llama estratificado.

2. Intervalos característicos

Consideremos una distribución normal de media cero y desviación típica 1, $N(0,1)$. Se llaman **intervalos característicos** a intervalos centrados en la media $(-z_p, z_p)$ a los que corresponde una probabilidad dada p.



$p = 0,9 \quad z_p = 1,645$

$p = 0,95 \quad z_p = 1,96$

$p = 0,99 \quad z_p = 2,575$

Junto a la figura se han puesto los valores de los z_p para los valores de la probabilidad que se presentan más frecuentemente en la práctica.

El cálculo de intervalos característicos en una distribución normal cualquiera $N(\mu, \sigma)$ puede hacerse sin dificultad recordando las fórmulas de cambio a puntuaciones típicas:

$$z = \frac{x-\mu}{\sigma} \Longleftrightarrow x = \mu + z\sigma$$

por lo que el intervalo característico correspondienta a una probabilidad p será:

$$(\mu - z_p\sigma, \mu + z_p\sigma)$$

3. Distribución de medias muestrales

Supongamos que en una población se ha medido una magnitud y se ha obtenido una media μ y una desviación típica σ. Tomamos una muestra de tamaño N y nos preguntamos por los valores de la media de los valores de la muestra $\bar{x}$, es decir, nos preguntamos cuál es la probabilidad de que la media muestral $\bar{x}$ se encuentre en un cierto intervalo (a, b).

El **teorema central del límite** establece que si los valores de la variable en la población se distribuyen normalmente segun la distribución $N(\mu, \sigma)$, las medias muestrales se distribuyen normalmente con la misma media μ y una desviación típica $\sigma/\sqrt{N}$. Además, para muestras grandes ($N > 30$) se puede aplicar el teorema aunque la población de partida no sea normal.

En resumen, si la población es normal o si no siéndolo $N > 30$ podemos obtener probabilidades para los valores de la media muestral a partir de la distribución:

$$N(\mu, \frac{\sigma}{\sqrt{N}})$$

En la práctica, el valor de la desviación típica de la población σ es desconocido por lo que se toma como valor aproximado, la desviación típica de la muestra s.

4. Distribución de proporciones muestrales

Consideremos una población en la que una proporción de elementos p cumplen una determinada condición. Tomemos en esta población una muestra de tamaño N. La cantidad de elementos de la muestra que cumplen la condición puede variar desde 0 a N y la probabilidad de obtener un determinado valor está dada por $B(N, p)$ y si la muestra es grande, las probabilidades de esta distribución binomial pueden obtenerse pol la distribución normal:

$$N\left(Np, \sqrt{Npq}\right)$$

donde $q = 1 - p$. Si en lugar de preguntarnos por el número de elementos de la muestra que cumplen la condición nos preguntamos por la proporción de ellos que la cumplen, tendremos que dividir por el número de elementos N. Sabemos que en este caso, la media y la desviación típica quedan también divididas por el mismo número de forma que las proporciones muestrales se distribuyen según:

$$N\left(p, \sqrt{\frac{pq}{N}}\right)$$

La condición para poder aplicar esta distribución es que se pueda aplicar la aproximación normal de la distribución binomial que se vió en un tema anterior.

Como en el caso anterior, en muchos problemas la proporción poblacional p, a partir de la cual se calcula la desviación tipica es desconocida y es preciso aproximarla por la proporción muestral p_r.

5. Intervalos de confianza

En una población, una variable estadística tiene una media μ y una desviación típica σ. Se toma una muestra de tamaño N y se mide la media muestral de la misma magnitud $\bar{x}$. Si las medias muestrales se

distribuyen normalmente, podemos decir que con una probabilidad c, llamada **nivel de confianza**, $\bar{x}$ se encontrará en el intervalo $(\mu - z_c\sigma/\sqrt{N}\ ,\ \mu + z_c\sigma/\sqrt{N})$.

Con la misma probabilidad podemos decir que la media poblacional se encontrará en el siguiente intervalo:

$$\left(\bar{x} - z_c\frac{\sigma}{\sqrt{N}}\ ,\ \bar{x} + z_c\frac{\sigma}{\sqrt{N}}\right)$$

llamado **intervalo de confianza** para la media correspondiente a un nivel de confianza c.

De forma similar, podemos obtener intervalos de confianza para proporciones. Sea p_r la proporción muestral. Con un nivel de confianza c se puede decir que la proporción en la población p se encuentra en el intervalo:

$$\left(p_r - z_c\sqrt{\frac{pq}{N}}, p_r + z_c\sqrt{\frac{pq}{N}}\right)$$

6. Contraste de hipótesis

Decisiones estadísticas

En muchas ocasiones es preciso tomar decisiones respecto a parámetros poblacionales a partir de los datos obtenidos a partir de muestras extraídas de la población. Para ello se formula una hipótesis respecto a la media poblacional (**hipótesis nula**) y se establecen intervalos de aceptación y de rechazo de la hipótesis para la media muestral. El intervalo de rechazo de la hipótesis se llama **región crítica**.

Nivel de significación

La probabilidad α de que, siendo cierta la hipótesis, la media de una muestra aleatoria se encuentre en la región crítica se llama **nivel de significación** del ensayo. Los valores más habituales de α son 10 %, 5 % y 1 %.

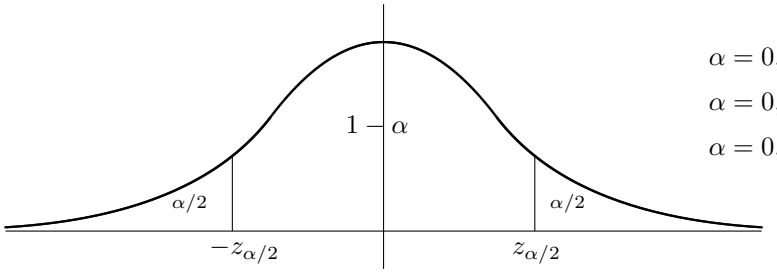
Ensayos bilaterales y unilaterales

Si la hipótesis es de la forma $\mu = \mu_0$, se podrá rechazar bien porque $\overline{x}$ sea mayor o menor que μ_0. Por consiguiente, la región crítica tendrá dos partes.La región de aceptación de la hipótesis con un nivel de significación α es:

$$\left(\mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\ ,\ \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right)$$

y la región crítica está formada por los dos intervalos que quedan a ambos lados de éste:

$$\text{Región crítica } = \left(-\infty\ ,\ \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right) \cup \left(\mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\ ,\ +\infty\right)$$



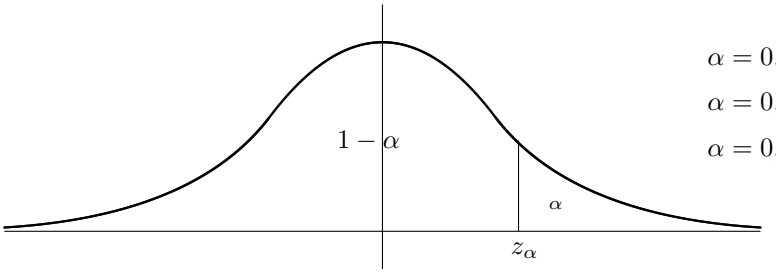
$\alpha = 0,1 \quad z_{\alpha/2} = 1,645$

$\alpha = 0,05 \quad z_{\alpha/2} = 1,96$

$\alpha = 0,01 \quad z_{\alpha/2} = 2,575$

Una hipótesis del tipo $\mu \leq \mu_0$ no puede rechazarse mediante valores de $\bar{x}$ menores que μ_0. Por consiguiente, la región crítica está formada por un solo intervalo. La región de aceptación de la hipótesis es:

$$\left(-\infty\ ,\ \mu_0 + z_\alpha\frac{\sigma}{\sqrt{N}}\right) \quad \text{y la región crítica será:} \quad \left(\mu_0 + z_\alpha\frac{\sigma}{\sqrt{N}}\ ,\ +\infty\right)$$



$\alpha = 0,1 \quad z_\alpha = 1,28$

$\alpha = 0,05 \quad z_\alpha = 1,645$

$\alpha = 0,01 \quad z_\alpha = 2,33$

Errores de tipo I y de tipo II

El rechazar una hipótesis que deberia ser aceptada se conoce como error de tipo I. La probabilidad de cometer ese error es el nivel de significación α. El aceptar una hipótesis falsa es el error de tipo II. La probabilidad de cometer este error es más dificil de evaluar y depende también del tamaño de la muestra.