# MUESTREO. INFERENCIA ESTADÍSTICA
## Resumen

## 1) Muestreo

La Estadística estudia grandes masas de datos. La totalidad de ellos constituye la *población* objeto de estudio. No suele ser posible o rentable, ni en tiempo ni en dinero, estudiar la población completa. Por ello, se recurre a extraer algunos datos representativos de la población, a los que se llamará *muestra*.

A partir de una muestra, elegida de forma que sea representativa de la población, se intentan deducir datos sobre ésta. La técnica de elegir la muestra de manera que sea, en efecto, representativa y, así, las conclusiones sobre la población sean fiables, se denomina *muestreo*. Los tipos más importantes de muestreo son:

- Muestreo aleatorio simple. La elección de los elementos de la población que compondrán la muestra es aleatoria. Cada dato de la población tiene las mismas probabilidades de ser elegido. La elección se realiza con reemplazamiento, de manera que un mismo dato puede aparecer más de una vez en la misma muestra. Una muestra se diferencia de otra por el orden en que se enumeran sus componentes, aunque estos coincidan.
- Muestreo sistemático. Si $n$ es el tamaño que queremos para la muestra y la población total consta de $N$ elementos, hallamos $k = \frac{N}{n}$ (se redondea el resultado). Ordenamos la población y elegimos un elemento de ella al azar. A partir de él, elegimos los restantes datos de la muestra de $k$ en $k$. Por ejemplo, si tenemos una población de 25 individuos y queremos extraer una muestra de tamaño 4, $k = 25/4 = 6.25$, que se redondea a $k = 6$. Elegimos, al azar, un elemento de la población, que se supone ordenada. Supongamos que es el 15. Entonces, el primer elemento de la muestra es el que ocupa el lugar 15; el segundo: 15 + 6 = 21; el tercero: 21 + 6 = 27 pero como hemos rebasado el total de la población, será: 27 – 25 = 2. Y el cuarto y último, el 2 + 6 = 8.
- Muestreo aleatorio estratificado con afijación proporcional. La población está dividida en estratos de diferente tamaño. Elegimos la muestra de manera que tenga la misma proporción de datos de cada estrato que en la población. Si conocemos dicha proporción, en forma de fracción o en tantos por ciento, el número de datos de cada estrato en la muestra será dicha proporción o tanto por ciento del tamaño muestral $n$. Si lo que conocemos es cuántos elementos hay en cada estrato, hacemos una regla de tres en la que la proporción de datos de la muestra con respecto a los de la población en ese estrato se compara con el tamaño de la muestra respecto al de la población. Dentro de cada estrato, la elección de los elementos es por muestreo aleatorio simple.

Notación:

| | Población | Muestra |
|---|---|---|
| Tamaño | $N$ | $n$ |
| Media | $\mu$ | $\bar{x}$ |
| Desv.Típica | $\sigma$ | $s$ |
| Varianza | $\sigma^2$ | $s^2$ |

## 2) Teorema Central del Límite (Lindeberg y Lévy)

Sea una población en la que se realiza un estudio estadístico cuyos resultados son los de la variable aleatoria $X$. Esta población seguirá una distribución. Llamemos $\mu$ a la media poblacional y $\sigma$ a la desviación típica de la población.

En dicha población, se extrae una muestra aleatoria de tamaño $n$. Si calculamos su media (muestral), su resultado es, por tanto, aleatorio. El experimento aleatorio es, ahora, la extracción de una muestra y calcular su media. Por tanto, una nueva variable aleatoria puede recoger los resultados de las medias aritméticas de las muestras. Llamemos $\bar{x}$ a la nueva variable aleatoria, que seguirá su propia distribución.

Es decir, tenemos dos variables aleatorias: la principal $X$, con media $\mu$ y desviación típica $\sigma$, que recoge los resultados de realizar el experimento aleatorio en la población principal, y una variable aleatoria secundaria $\bar{x}$, con media $\mu_{\bar{x}}$ y desviación típica $\sigma_{\bar{x}}$, que da el resultado de extraer una muestra de $n$ resultados de $X$ y calcular su media.

Pues bien:

- Si $X$ sigue una distribución normal: $X \in N(\mu; \sigma)$, entonces, $\boxed{\bar{x} \in N\left(\mu; \frac{\sigma}{\sqrt{n}}\right)}$.
- Aunque $X$ no siga una distribución normal, como consecuencia del *Teorema Central del Límite*, cuando $n$ es grande puede suponerse igualmente cierto lo anterior, por aproximación. En general, si $n \geq 30$ pueden considerarse válidos los cálculos obtenidos utilizando dicho resultado.
- De forma equivalente y por las mismas causas y en las mismas condiciones ($X$ es normal o $n \geq 30$), $\boxed{\sum_{i=1}^{n} x_i \in N\left(n\mu; \sigma\sqrt{n}\right)}$, donde $x_i$ son los valores de una muestra de tamaño $n$.
- Cuando las poblaciones son pequeñas ($N$ = tamaño de la población, finito) y es posible calcular todas las muestras de un determinado tamaño $n$, ya no estamos ante una aproximación, y la media de todas las medias de esas muestras ($\mu_{\bar{x}}$) valdrá: $\mu_{\bar{x}} = \mu$, y la desviación típica de todas las medias de las muestras: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.
- Si $\sigma$ es desconocida pero el tamaño muestral es grande ($n \geq 30$), podemos sustituirla por la *cuasidesviación típica muestral*: $s_{n-1} = \sqrt{s^2 \frac{n}{n-1}}$.
- Si tenemos dos poblaciones normales: $X_1 \in N(\mu_1; \sigma_1)$ y $X_2 \in N(\mu_2; \sigma_2)$, entonces la diferencia de medias muestrales seguirá una distribución: $\bar{x}_1 - \bar{x}_2 \in N\left(\mu_1 - \mu_2; \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$. Si las poblaciones no son normales, para muestras grandes ($n_1 \geq 30$ y $n_2 \geq 30$), se pueden aproximar los resultados reales a los que proporcione suponer que la diferencia de medias muestrales sigue la distribución que se ha mencionado.

## 3) Intervalo de confianza para la media poblacional

- En las condiciones anteriores ($X$ sigue una distribución normal de media $\mu$ y desviación típica $\sigma$, o bien, aplicando el *Teorema Central del Límite*, tomamos muestras de tamaño $n \geq 30$), el intervalo de confianza para la media poblacional $\mu$, con un *nivel de confianza* $1-\alpha$ es:

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

Lo que significa que la probabilidad de que $\mu$ esté dentro del intervalo es $1-\alpha$. Es decir, que el $(1-\alpha)\cdot 100\%$ de las muestras que se extraigan proporcionarán un intervalo tal que la verdadera media poblacional está dentro.

- $1-\alpha$ = Nivel de confianza
- $(1-\alpha)\cdot 100\%$ = *Nivel de confianza* en tantos por ciento
- $\alpha$ = Nivel de significación
- $z_{\alpha/2}$ es el número que proporcionan las tablas de la $N(0;1)$ tal que $P(Z \le z_{\alpha/2}) = 1 - \alpha/2$ (ver gráfico al final del documento). Se le llama *valor crítico* correspondiente a $p = 1-\alpha$.
- $E = z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ se llama *error máximo* admisible (*cota de error*).
- El centro del intervalo es $\bar{x}$, y el radio es el error $E$. La amplitud es $2E$.
- Si la desviación típica poblacional no es conocida pero $n \ge 30$, puede estimarse por la *cuasidesviación típica* de la muestra: $s_{n-1} = \sqrt{s^2\dfrac{n}{n-1}}$

## 4) Distribución de las proporciones muestrales

Sea una población en los que los individuos pueden poseer una determinada característica $C$, o no poseerla (es decir, elegir un individuo de dicha población es una prueba de Bernoulli, con resultado de *éxito* si posee la característica, o *fracaso* si no).
Sea $p$ la probabilidad de *éxito*, es decir, de que un individuo de la población posea la característica $C$, y $q = 1 - p$ la de *fracaso*.
Sea $x$ el número de *éxitos* en una muestra de tamaño $n$. Como cada individuo es un experimento de Bernoulli, la variable aleatoria que, elegida una muestra al azar, nos da $x$ = nº de éxitos, sigue una distribución *Binomial*: $x \in B(n; p)$
Sea $\hat{p}$ la proporción de *éxitos* en la muestra, es decir, $\hat{p} = x / n$.
Pues bien, si $n$ es suficientemente grande ($n \ge 30$) se puede considerar buena la aproximación de la Binomial por la Normal (consecuencia del *Teorema Central del Límite* que, particularizado en la Binomial es el *Teorema de Moivre-Laplace*), con lo que $x \in N(np;\ \sqrt{npq}\,)$. Por las características de las distribuciones normales, un teorema nos lleva a deducir de aquí que:

$$\hat{p} \in N\left(p;\ \sqrt{\frac{pq}{n}}\right)$$

## 5) Intervalo de confianza para una proporción

En las condiciones anteriores ($n \ge 30$), el intervalo de confianza para la probabilidad $p$ de éxitos en la población, o sea, de que un individuo de la población posea la característica $C$ (es decir, la proporción de individuos con dicha característica en la población, medida en tantos por uno), con un *nivel de confianza* de $1-\alpha$, es:

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

donde $\hat{p}$ es la proporción de éxitos en la muestra.

Esto significa que la probabilidad de que $p$ esté dentro de dicho intervalo es $1-\alpha$. Es decir, que el $(1-\alpha)\cdot 100\%$ de las muestras que se extraigan proporcionarán un intervalo tal que la verdadera proporción poblacional está dentro.

- $1-\alpha$ = *Nivel de confianza* (multiplicado por cien está en tantos por ciento)

- $\alpha$ = Nivel de significación
- $Z_{\alpha/2}$ es el número que proporcionan las tablas de la $N(0;1)$ tal que $P(Z \leq z_{\alpha/2}) = 1 - \alpha/2$.
- $E = z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ se llama *error máximo* admisible (*cota de error*).
- El centro del intervalo es $\hat{p}$, y el radio es el error $E$. La amplitud es $2E$.

## 6) Intervalo de confianza para la diferencia de medias poblacionales

- En las condiciones anteriores ($X_1$ y $X_2$ siguen distribuciones normales o, aplicando el *Teorema Central del Límite*, tomamos muestras de tamaño $n_1 \geq 30$ y $n_2 \geq 30$), el intervalo de confianza para la diferencia de medias poblacionales $\mu_1 - \mu_2$, con un *nivel de confianza* $1-\alpha$ es:

$$\left( \bar{x}_1 - \bar{x}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2}},\ \bar{x}_1 - \bar{x}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2}} \right)$$

  Lo que significa que la probabilidad de que $\mu_1 - \mu_2$ esté dentro del intervalo es $1-\alpha$. Es decir, que el $(1-\alpha)\cdot 100\%$ de las muestras que se extraigan proporcionarán un intervalo tal que la verdadera media poblacional está dentro.
- $1-\alpha$ = Nivel de confianza
- $(1-\alpha)\cdot 100\%$ = *Nivel de confianza* en tantos por ciento
- $\alpha$ = Nivel de significación
- $z_{\alpha/2}$ es el número que proporcionan las tablas de la $N(0;1)$ tal que $P(Z \leq z_{\alpha/2}) = 1 - \alpha/2$ (ver gráfico al final del documento). Se le llama *valor crítico* correspondiente a $p = 1 - \alpha$.
- $E = z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ se llama *error máximo* admisible (*cota de error*).
- El centro del intervalo es $\bar{x}_1 - \bar{x}_2$, y el radio es el error $E$. La amplitud es $2E$. Si la desviaciones típicas poblacionales no son conocidas pero $n_1 \geq 30$ y $n_2 \geq 30$, pueden estimarse por las *cuasidesviaciones típicas* de las muestras respectivas.

## 7) Valores más frecuentes de las tablas N(0; 1) para $Z_{\alpha/2}$

| 1–α | 1–α/2 | $z_{\alpha/2}$ |
|---|---|---|
| 0,90 | 0,95 | 1,645 |
| 0,91 | 0,955 | 1,7 |
| 0,92 | 0,96 | 1,75 |
| 0,93 | 0,965 | 1,81 |
| 0,94 | 0,97 | 1,88 |

| 1–α | 1–α/2 | $z_{\alpha/2}$ |
|---|---|---|
| 0,95 | 0,975 | 1,96 |
| 0,96 | 0,98 | 2,055 |
| 0,97 | 0,985 | 2,17 |
| 0,98 | 0,99 | 2,33 |
| 0,99 | 0,995 | 2,575 |