Tema 8. Muestreo

Indice

1. Población y muestra.

Es muy común encontrar medios de comunicación que nos presenten resultados de estudios estadísticos. Muchas veces estos estudios se realizan desde una forma partidista, es decir se enfocan para obtener los resultados que el medio desea. La forma más sencilla de "trucar" los resultados es la elección de una muestra favorable a los intereses (por ejemplo si deseamos calcular la media de altura en España no parece lógico hacer el estudio entre las personas que juegan en algún equipo de baloncesto).

Para definir los parámetros más importantes de las muestras veamos 2 ejemplos:

- Ejemplo1: estimar porcentaje de votos de un partido en unas elecciones.
- Ejemplo 2: conocer la duración media de bombillas que se fabrican.

En todo estudio estadístico se denotan:

- ***Población:*** al conjunto o colectivo de elementos sometidos al estudio estadístico (N tamaño población). En nuestros ejemplos:
  - Ejemplo 1: el conjunto total de votantes a dichas elecciones
  - Ejemplo 2:el conjunto de todas las bombillas fabricadas
- ***Muestra:*** muchas veces no es posible realizar el estudio para la población total : a) población muy numerosa, b) proceso destructivo. Se hace entonces necesario realizar el estudio a una muestra, un subconjunto de la población sobre las cuales se realiza el estudio y a partir de este se generaliza a toda la población (n tamaño de la muestra). En nuestros ejemplos:
  - Ejemplo 1: si la población es muy grande no podremos hacer el estudio para toda la población por motivos económicos y de temporalidad.
  - Ejemplo 2: no podemos medir la duración de todas las bombillas pues entonces no podremos venderlas (proceso destructivo).
- ***Muestreo:*** es el proceso que utilizamos para elaborar la muestra a partir de la población total. La forma de elegirla es básica para que la estadística sea válida. En nuestros ejemplos:
  - Ejemplo 1: si elegimos un barrio obrero o rico cambiará la estadística.
  - Ejemplo 2: no sería lógico elegir las bombillas fabricadas seguidas.

2. Tipos de muestreos.

Como hemos comentado en el apartado anterior la muestra elegida debe de ser representativa de la población estudiada, ya que si no, las conclusiones obtenidas en el estudio no son fiables.

Para que una muestra sea fiable este tiene que cumplir dos condiciones fundamentales:

- Tener un tamaño adecuado
- Que los elementos sean elegidos de forma totalmente aleatoria.

Si cumple las dos condiciones diremos que la muestra es ***representativa***, en caso contrario la muestra se dice que es ***sesgada***.

Los muestreos más importantes y utilizados son los siguientes:

- ***Muestreo aleatorio simple:*** todos los elementos de la muestra tienen la misma probabilidad de ser elegidos. Veamos cómo realizarlo en nuestros ejemplos:
  - Ejemplo 1: numeramos los elementos de la población que puedan votar de 1 a N, y elegimos al azar n de ellos.
  - Ejemplo 2: elegimos n bombillas al azar, ordenándolas según el orden de fabricación.
- ***Muestreo aleatorio sistemático:*** Se ordenan los elementos de la población se selecciona uno de los elementos y se eligen los n-1 restantes de forma que estén igualmente espaciados. Ejemplo: si tenemos una población de N=10.000 y queremos una muestra de n=100 (uno de cada 10.000/100=100), elegimos un al azar 3.211, serán entonces (3.211, 3.311, 3411,…, 9.911, 11, 111, …, 3.111).
- ***Muestreo aleatorio estratificado:*** En ocasiones, la característica que se estudia en la población varía según diferentes grupos delimitados por ejemplo por edades, sexos o barrios. En este caso la población se divide en grupos homogéneos denominados estratos y posteriormente se extrae una muestra aleatoria de cada estrato de forma que el tamaño de la muestra mantenga la misma proporción que los elementos de la población de dicho estrato. Por ejemplo en la población que va a votar la dividimos en tres

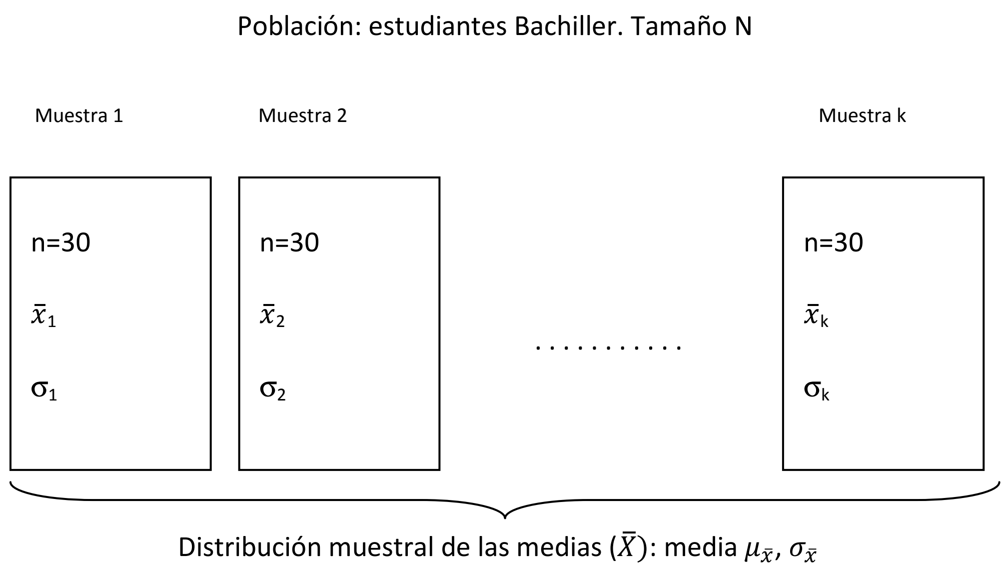
estratos de 18 a 40 años (estrato 1), de 40 a 60 años (estrato 2) y más de 60 años (estrato 3). Si la población de 30 millones distribuida de la siguiente forma: 15 millones del estrato 1, 10 millones del estrato 2 y 5 millones del estrato 3. Si queremos elegir una muestra de 600 personas, tendremos que elegir:

- $n_1=\frac{30.000.000}{15.000.000}\cdot 600=300$
- $n_2=\frac{30.000.000}{10.000.000}\cdot 600=200$
- $n_3=\frac{30.000.000}{5.000.000}\cdot 600=100$

3. Distribución muestral de las medias.

Consiste en el estudio estadístico de las medias de una población fijada el tamaño de la muestra. Es decir ahora en vez de individuos con un valor de x, tenemos una muestra con su correspondiente valor medio $\bar{x}$.

Para entender el significado veamos un ejemplo: queremos estudiar el peso medio de los estudiantes de bachillerato de España. Para el estudio elegimos diferentes muestras de 30 alumnos (n=30) y calculamos los distintos valores de las medias de los pesos ($\bar{x}$):



La variable de estudio estadístico no es x, peso de los alumnos, sino $\bar{x}$ que es la media de los pesos de las muestras de 30 alumnos. La media de esta variable estadística es la media de las medias, $\mu_{\bar{x}}$, y la desviación es la desviación de las medias, $\sigma_{\bar{x}}$.

La pregunta es la siguiente, ¿qué distribución sigue $\bar{x}$?, ¿depende de la distribución que siga x?.

Si podemos obtener todas las muestras posibles de n=30 de la población total por muestreo aleatorio simple entonces se cumple:

- La media de $\bar{x}$, es igual a la media de la población total: $\mu_{\bar{x}} = \frac{\mu_{\bar{x}1}+\cdots+\mu_{\bar{x}_k}}{k} = \mu$
- La desviación es menor cuanto mayor es el tamaño de la muestra: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.
- La distribución de las medias se comporta de forma aproximada como una distribución normal (independientemente si x sigue esta distribución u otra): $\bar{X} \to N(\mu_{\bar{x}} = \mu, \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}})$. Esta aproximación se da por buena si n>30 y N>2n.

Ejercicios resueltos

Ejercicio 1: La altura de los alumnos de bachillerato tiene una media de 176 cm y desviación de 10 cm. Se elije al azar un muestra de 45 alumnos de bachillerato. Calcular las probabilidades de a) que la media sea mayor que 180cm, b) que la medie esté entre 170 y 178cm.

Solución: Aunque no nos dicen nada de la distribución de x (altura de los alumnos), las medias de la población de las alturas ($\bar{x}$) sigue una distribución normal $\bar{x} \to N(\mu_{\bar{x}} = \mu, \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}})$; $\bar{x} \to N(176, 10/\sqrt{45})$; $\bar{x} \to N(176, \frac{10}{\sqrt{45}} = 1.5)$;

a) $P(\bar{x}\geq 180)=P(Z\geq\frac{180-176}{1.5})=P(Z\geq 2,67)=1-P(Z\leq 2,67)=1-0.9962=0.0038$

b) $P(170\leq\bar{x}\leq 178)=P(\frac{170-176}{1.5}\leq Z\leq\frac{178-176}{1.5})=P(-4\leq Z\leq 1,33)=P(Z\leq 1,33)-P(Z\leq -4)=$
$=0,9082-(1-1)=0,9082$

Ejercicio 2: Se dispone de una muestra aleatoria de 36 alumnos de 1º Bachillerato. Se sabe de estudios anteriores que la altura de estos alumnos se distribuye de tal forma que la media de la altura es de 170 cm y la desviación típica de 12cm. Calcular a) la probabilidad de que la media muestra comprendida entre 166 y 172 cm, b) la probabilidad que la media sea superior a 174cm.

Solución: Aunque no nos dicen que distribución sigue las alturas (x) de los alumnos, al tratarse del estudio de muestras y por tanto de medias muestrales ($\bar{x}$) estos siguen una distribución normal: $\bar{x} \rightarrow N(\mu,\sigma/\sqrt{n})=N(170,12/\sqrt{36})=N(170,2)$

a) $p(166\leq\bar{x}\leq172)=p(\frac{166-170}{2}\leq Z\leq\frac{172-170}{2})=p(-2\leq z\leq1)=p(z\leq1)-(1-p(z\leq2))=$
$=0,8413-(1-0,9772)=0,8185$

b) $p(\bar{x}\geq174)=p(Z\geq\frac{174-170}{2})=p(z\geq2)=1-p(z\leq2)=0.0228.$

Ejercicio 3: La media de edad de los alumnos que se presentan a Selectividad es de 18,1 años con una desviación típica de 0,6 años. Se elige una muestra al azar de 100 alumnos que se presentan a selectividad, ¿cuál es la probabilidad de que la media de edad de esta muestra esté entre 17,9 años y 18,2?

Solución: Aunque no nos dicen que distribución sigue las edades (x) de los alumnos, al tratarse del estudio de muestras y por tanto de medias muestrales ($\bar{x}$) estos siguen una distribución normal: $\bar{x} \rightarrow N(\mu,\sigma/\sqrt{n})=N(18.1,0,6/\sqrt{100})=N(18.1, 0,06)$

$p(17,9\leq\bar{x}\leq18.2)=p(\frac{17.9-18.1}{0.06}\leq Z\leq\frac{18.2-18.1}{0.06}) =p(-3.33\leq z\leq1.67)=p(z\leq1.67)-(1-p(z\leq3.3)=$
$=0.9525-(1-0.9996)=0.9521.$

4. Distribución muestral de las proporciones.

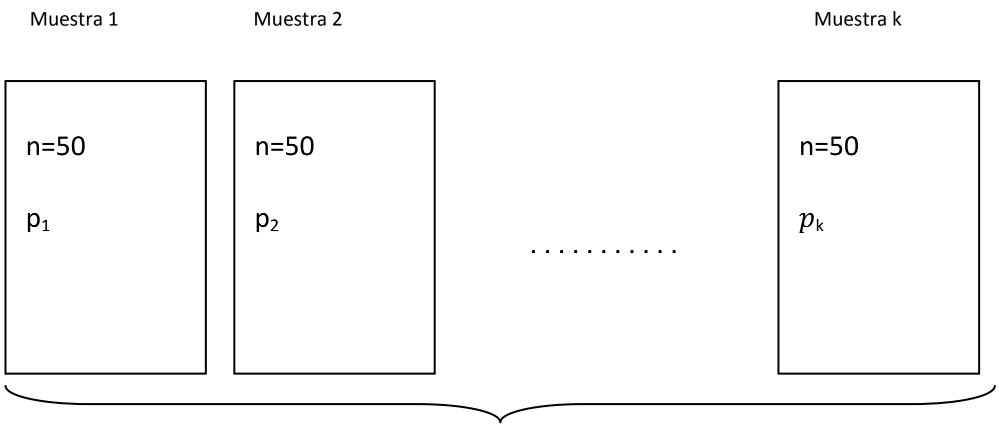
Supongamos una población grande (N) , como por ejemplo todos los alumnos de una universidad; y sea el suceso A="sea aficionado al cine" (cuya respuesta es sí o no, es decir binomial). Si n_a son los que afirman gustarles el cine, entonces $p=n_a/N$ será la proporción de alumnos aficionados, siendo q=1-p la proporción de no aficionados.

Tomemos ahora una muestra aleatoria de tamaño n=50 (M_1), en esta muestra la proporción $p_1=n_{a1}/50$ nos da la proporción de aficionados al cine en la muestra M_1.

Podremos extraer otras muchas muestras (M_2, M_3,…) con el mismo tamaño, y calcular las proporciones relativas a las mismas (p_2, p_3, …) que variarán de una muestra a otra.

Todas las proporciones constituyen otra distribución estadística: la llamada ***distribución muestra de la proporción***.

Población: estudiantes universidad. Tamaño N



Distribución muestral de las proporciones (P): media μ_p, σ_p

La variable aleatoria que estudia esta nueva distribución se denota como ***P, la proporción muestral,*** de la que se cumple que su media aritmética y su desviación son (si tomamos 100):

$$\mu_p = \frac{p_{a1}+p_{a2}+\cdots+p_{a100}}{100} \qquad \text{y} \qquad \sigma_p = \sqrt{\frac{\sum(p_{ai}-\mu_p)}{100}}$$

Si en vez las 100 muestras se forman todas las posibles de tamaño n mediante un muestreo aleatorio simple, la media y la desviación típica cumplen lo siguiente:

$$\boldsymbol{\mu_p = p} \qquad \textbf{y} \qquad \boldsymbol{\sigma_p = \sqrt{\frac{p\cdot q}{n}}}$$

Teorema del límite para proporciones: si la distribución de una población grande (N>60) tiene una proporción p de una cierta característica A, entonces la variable aleatoria P, de las proporciones muéstrales extraídas de esta población (distribuida binomialmente) de tamaño grande (n>30) se aproxima a una distribución normal de las siguientes características:

$$P \rightarrow N\left(\mu_p = p,\ \sigma_p = \sqrt{\frac{p\cdot q}{n}}\right)$$

<u>Ejercicios resueltos</u>

Ejercicio 1: Una máquina fabrica piezas de las se sabe que el 3% son defectuosas. Un cliente recibe una remesa de 500 piezas. a) ¿cuál es la probabilidad de que encuentre menos del 1% de piezas defectuosas?, b) ¿cuál es la probabilidad de que encuentre más de un 5% de ellas defectuosas?, c) ¿Cuál es la probabilidad de que estén defectuosas entre 15 y 30 piezas?

Solución: la proporción de defectuosas p=0,003. Se puede considerar que la muestra de tamaño n=500 proviene de una producción N muy grande, con lo que podemos aplicar el teorema del límite, por tanto la proporción de piezas defectuosas P, sigue una distribución normal: $P \rightarrow N(\mu_p = p = 0{,}03,\ \sigma_p = \sqrt{\frac{p\cdot q}{n}} = 0{,}0076)$

a) $p(P \leq 0{,}01) = P\left(Z \leq \frac{0{,}01-0{,}03}{0{,}0076}\right) = p(Z \leq -2.6316) = 1 - p(Z \leq 2.6316) = 0{,}0043$
b) $p(P \geq 0{,}05) = P\left(Z \geq \frac{0{,}05-0{,}03}{0{,}0076}\right) = p(Z \geq 2.6316) = 1 - p(Z \leq 2.6316) = 0{,}0043$
c) $p\left(\frac{15}{500} \leq P \leq \frac{30}{500}\right) = P\left(\frac{0{,}03-0{,}03}{0{,}0076} \leq Z \leq \frac{0{,}06-0{,}03}{0{,}0076}\right) = p(0 \leq z \leq 3.94) = 0{,}5.$

Ejercicio 2: Un lanzador de tiro al blanco acierta el 85% de las veces a la diana. Determina la distribución en el muestreo de la proporción de veces que da a la diana para muestras de 100 lanzamientos. Calcular la probabilidad de que acierte más de 90 lanzamientos en la diana:

Solución: $P \rightarrow N(\mu_p = p = 0{,}85,\ \sigma_p = \sqrt{\frac{p\cdot q}{n}} = 0{,}0357)$

$p(P \geq 0{,}90) = p\left(Z \geq \frac{0{,}9-0{,}85}{0{,}0357}\right) = p(Z \geq 1{,}4) = 1 - p(Z \leq 1{,}4) = 0{,}0808.$

Ejercicio 3: Se sabe que de cada 5 personas accidentadas en una cierta carretera hay una mujer. Calcular la probabilidad de que en los próximos 200 accidentes automovilísticos en esa carretera a) sean hombres menos el 30% de los accidentados, b) sean hombres más del 80% de los accidentados, c) la proporción de hombres accidentados esté comprendida entre el 40% y el 60%.

Solución: Designemos por p=4/5=0,8 la proporción de hombres accidentados. El tamaño de la muestra es N=200, y que procede de una población muy grande.

Si P es la variable aleatoria que representa la proporción de hombres accidentados en una muestra de tamaño N=200. La media muestral de P es μ_p =p=0,8 de media y de desviación $\sigma_p = \sqrt{\frac{p\cdot q}{n}} = 0{,}0283$.

Puede considerarse que la proporción P sigue la ley normal N(p=0.8, σ_p=0,0283)

a) p(P≤0,3)=p$(z \le \frac{0{,}3-0.8}{0{,}0283}) = p(z \le -17{,}67) = 1 - p(z \le 17{,}67) \approx 1 - 1 = 0$
b) p(P≥0,8)=p(z≥$\frac{0{,}8-0.8}{0{,}0283}) = p(z \ge 0) = 0{,}5$
c) p(0,4≤P≤0,6)=p($\frac{0{,}4-0.8}{0{,}0283}$≤z≤$\frac{0{,}6-0.8}{0{,}0283}$)=p(-14,13≤z≤-7,07)≈0

Ejercicios propuestos:

Problema 1: un cierto equipo electrónico formado por 100 componentes. Si cada componente tiene la probabilidad p=0,02 de romperse hallar la probabilidad de que al conectar el equipo se rompan 3 o más componentes.

Problema 2: Supongamos que un tirador tiene 50 sobre 100 posibilidades de hacer blanco. Hallar la probabilidad de que después de haber hecho 40 disparos haya acertado entre 8 y 12.

Problema 3: si los alumnos de preescolar de León tiene una altura cuya media es de 95cm y su desviación típica es de 16cm y cogemos una muestra aleatoria de de 36 de estos alumnos, ¿cuál es la probabilidad de que la media de esta muestra este entre 90 y 100cm?

Problema 4: Se conoce que el número de días que permancen los nefermos en un hospital es una distribución aleatoria de media 8,1 días y desviación 9 días. Se elije al azar a 100 enfermos. Calcular la probabilidad de que estén de media a) más de 7 días, b) entre 7 y 8 días, c) y menos de 9 días.

Problema 5: Las notas de un examen sigue una ley normal de media 5,6 y varianza 9. Selecciona 16 estudiantes al azar y calculamos las medias de sus notas. Calcular las probabilidades de a) la media comprendida entre 4,7 y 6,5; b) si seleccionamos 25 alumnos ¿aumentarán o disminuirá la probabilidad calculada en el apartado anterior.