Capítulo 4

INFERENCIA ESTADÍSTICA

4.1. Introducción

Inferir: Sacar una consecuencia de una cosa. Sacar consecuencia o deducir una cosa de otra.

La estadística, ciencia o rama de las Matemáticas que se ocupa de recoger datos, analizarlos y organizarlos, y de realizar las predicciones que sobre esos datos puedan deducirse, tiene dos vertientes básicas:

a) Estadística descriptiva: Básicamente se ocupa de la 1ª parte, es decir, a partir de ciertos datos, analizarlos y organizarlos. Es aquí donde tiene sentido calcular la media, mediana, moda, desviación media, desviación típica, etc.

b) Estadística inferencial: Se ocupa de predecir, sacar conclusiones, para una población tomando como base una muestra (es decir , una parte) de dicha población. Como todas las predicciones, siempre han de hacerse bajo un cierto grado de fiabilidad o confianza.

Será esta última vertiente de la estadística la que estudiemos en este tema.

4.2. Muestreos

Ya sabemos que una población es el conjunto de individuos sobre los que hacemos cierto estudio, y que una muestra es un subconjunto de la población. Es evidente que los resultados de una determinada encuesta tendrán un mayor grado de fiabilidad si dicha encuesta se realiza sobre la población completa. Sin embargo, en la mayoría de las ocasiones esto no es posible, debido a múltiples razones:

* Imposibilidad material (Hacer una encuesta a los casi 41 millones de españoles es imposible,hacer un estudio sobre la fecha de caducidad de un producto. Si lo hacemos con todos los productos ¿qué vendemos luego?)

* Imposibilidad temporal (Hacer un estudio sobre la duración de una bombilla. ¿Cuánto debemos esperar para saberlo?).

Por tanto, es habitual que tengamos que manejarnos con muestras, de modo que es importante saber elegir bien una muestra de la población, una muestra que represente bien a dicha pbolación.

Hay muchas maneras de elegir una muestra de una población.

Antes de pasar a analizar dichas formas de extracción de muestras, lo que si hemos de dejar claro es que todas las muestras han de cumplir varias condiciones indispensables.

Es evidente que para que el estudio a realizar sea fiable, hay que cuidar mucho la elección de la muestra, para que represente en la medida de lo posible a la población de la que se extrae. Si la muestra está mal elegida, diremos que *no es representativa*.

En este caso, se pueden producir errores imprevistos e incontrolados. Dichos errores se denominan *sesgos* y diremos que *la muestra está sesgada*.

Una de las condiciones para que una muestra sea representativa es que el *muestreo* (o sistema para elegir una muestra de una población) que se haga sea *aleatorio*, es decir, todas las personas de

la población tengan las mismas posibilidades de ser elegidas, mientras que si la elección de la muestra es subjetiva, es probable que resulte sesgada.

Las distintas maneras de elegir una muestra de una población se denominan muestreos. Básicamente hay dos tipos de muestreos:

1. *Muestreo no probabilístico:* El investigador no elige la muestra al azar, sino mediante determinados criterios subjetivos.

2. *Muestreo probabilístico:* Cuando la muestra se elige al azar. En este caso podemos distinguir varios tipos:

   a) *Muestreo aleatorio simple:* Aquel en el que cada individuo de la población tiene las mismas posibilidades de salir en la muestra.

   b) *Muestreo sistemático:* En el que se elige un individuo al azar y a partir de él, a intervalos constantes, se eligen los demás hasta completar la muestra.

   c) *Muestreo estratificado:* En este muestreo se divide la población en clases o estratos y se escoge, aleatoriamente, un número de individuos de cada estrato proporcional al número de componentes de cada estrato.

   d) *Muestreo por conglomerados:*Si no disponemos de la relación de los elementos de la población, o de los posibles estratos, no podemos aplicar los muestreos anteriores.

   Aquí entra el llamado muestreo por conglomerados, donde en lugar de elegir individuos directamente, se eligen unidades más amplias donde se clasifican los elementos de la población, llamados conglomerados. En cada etapa del muestreo en lugar de seleccionar elementos al azar seleccionamos conglomerados.

   Los conglomerados deben ser tan heterogéneos como la población a estudiar, para que la represente bien. Luego se elegirían algunos de los conglomerados al azar, y dentro de éstos, analizar todos sus elementos o tomar una muestra aleatoria simple.

   No debemos confundir estrato y conglomerado. Un estrato es homogéneo (sus elementos tienen las mismas características), mientras que un conglomeardo es heterogéneo (debe representar bien a la población).

Veamos la diferencia de estos muestreos mediante un ejemplo:

Imaginemos que hemos de recoger una muestra de 20 alumnos de entre los de un instituto de 600.

-Muestreo aleatorio simple: Elegiríamos un alumno al azar (probabilidad de elegirlo $\frac{1}{600}$. Lo devolvemos a la población y se elige otro (probabilidad de elegirlo $\frac{1}{600}$), y así hasta 20. Notemos que si no devolviésemos al alumno, entonces, la probabilidad de escoger al 2º alumno sería $\frac{1}{599}$, y ya no todos tendrían la misma probabilidad de ser elegidos. El problema es que entonces permitimos que se puedan repetir individuos.

-Muestreo sistemático: Como hemos de elegir 20 alumnos de 600, es decir, 1 de cada 30, se procede así: Se ordenan los alumnos y se numeran, se elige uno al azar, por ejemplo el alumno 27, y luego los demás se eligen a partir de este a intervalos de 30 alumnos. Escogeríamos por tanto a los alumnos:

$$27,57,87,117,147,177,207,237,267,297,327,357,387,417,447,477,507,537,567,597$$

y el alumno 627 ya es otra vez el 27.

-Muestreo estratificado: Si queremos que la muestra sea representativa, lo mejor será conocer cuántos alumnos de cada curso hay, es decir, si hay 200 alumnos de 3º ESO, 150 de 4º ESO, 150 de 1º Bachillerato y 100 de 2º Bachillerato, procederíamos:

Como de 600 en total hemos de elegir a 20, de 200 de 3º de ESO hemos de elegir x:

$$\frac{20}{600} = \frac{x}{200} \longrightarrow x = \frac{4000}{600} = 6'6 \approx 7 \text{ alumnos de } 3º$$

(Utilizando la regla de tres)

De igual manera podemos calcular los alumnos correspondientes a los demás cursos:

$$\frac{20}{600} = \frac{y}{150} \longrightarrow y = \frac{3000}{600} = 5 \text{ alumnos de } 4^{\underline{o}}$$

$$\frac{20}{600} = \frac{z}{150} \longrightarrow z = \frac{3000}{600} = 5 \text{ alumnos de } 1^{\underline{o}}$$

$$\frac{20}{600} = \frac{t}{100} \longrightarrow t = \frac{2000}{600} = 3'3 \text{ alumnos de } 2^{\underline{o}}$$

De modo que en nuestra muestra de 20, 7 alumnos son de $3^{\underline{o}}$, 5 de $4^{\underline{o}}$, 5 de $1^{\underline{o}}$ y 3 de $2^{\underline{o}}$. Para la elección de cada alumno dentro de cada curso, utilizamos el muestreo aleatorio simple.

-Muestreo por conglomerados: Para ver este muestreo, hemos de cambiar el ejemplo.

Supongamos que queremos extraer una muestra aleatoria de los estudiantes universitarios del país. Necesitariamos una lista con todos ellos para poder realizar algún muestreo del tipo de los 3 anteriores, lo cuál es muy difícil de conseguir. Sin embargo, los estudiantes estan clasificados por Universidades, Facultades y Clases.

Podemos seleccionar en una primera etapa algunas Universidades, después algunas facultades al azar, dentro de las facultades algunas clases y dentro de las clases, algunos estudiantes por muestreo aleatorio simple. Los conglomerados en cada etapa serían las diferentes Universidades, las diferentes facultades y los diferentes clases.

Como vemos los conglomerados son unidades amplias y heterogéneas.

Ejercicio:

En una población de 1500 jóvenes,7500 adultos y 1000 ancianos, se hace una encuesta a 200 personas para conocer sus actividades de ocio preferidas. Si se utiliza un muestreo estratificado, ¿qué tamaño muestral corresponde a cada estrato?.

4.3. Estimación por puntos

Como el objetivo principal de la estadística inferencial es el estudio de la población y realizar predicciones a cerca de ella pero a partir de una muestra de ella , no de la población entera, en principio, tendremos que estimar los índices de la población a partir de los índices correspondientes para la muestra.

En una primera aproximación, parece lógico pensar que si queremos determinar la media de una cierta población, si hemos cogido una muestra representativa de dicha población, la media de la muestra (que es fácilmente calculable porque tenemos los datos) será muy parecida a la de la población y por tanto sirva para estimarla.

Distinguiremos, por tanto, entre:

1. *Parámetros poblacionales:* Que son los índices centrales y de dispersión que definen a una población.

   Representaremos la media poblacional μ y la desviación típica poblacional σ.

   En el caso de proporciones, la proporción de población que tiene una determinada característica la denotaremos por p y la proporción que no la cumple por $q = 1 - p$. (Como en la Distribución binomial)

2. *Estadísticos poblacionales:* Son los índices centrales y de dis persión que definen a una muestra.

   Representaremos la media muestral por $\overline{x}$ y la desviación típica muestral por s.

   En el caso de proporciones, la proporción de muestra que tiene una determinada característica la denotaremos por $\hat{p}$ y la proporción que no la cumple por $\hat{q} = 1 - \hat{p}$.

¿Cuál es el problema de la estimación entonces?. Como vamos a disponer de una muestra, lo que podemos calcular es $\overline{x}$ y s (o bien $\hat{p}$ y $\hat{q}$), y a partir de estos intentar estimar quienes tienen que ser μ y σ (o bien p y q), los reales para la población.

En la estimación por puntos, el conocimiento de un estadístico muestral nos permitirá decidir cuál es el correspondiente parámetro de la pobla ción. Para ello hemos de conocer cuál es la relación entre un estadístico y el corresp ondiente parámetro.

4.4. Distribución muestral de medias

Si tenemos una población de parámetros desconocidos μ y σ, y tomamos una muestra, podemos calcular la media muestral,$\overline{x_1}$, que tendrá cierta relación con μ.

Podríamos tomar otra muestra, de igual tamaño, y calcular de nuevo su media muestral $\overline{x_2}$, que también estará relacionada con μ.

Así sucesivamente, considerando varias muestras y haciendo las medias muestrales respectivas, tenemos una serie de medias, relacionadas de alguna manera con μ ¿cómo?. De la siguiente forma:

Propiedad: Si la población sigue una distribución normal $N(\mu, \sigma)$, donde μ y σ son desconocidos, si elegimos todas las muestras de cierto tamaño (n) , de forma que sean representativas, entonces:

a) La media de las medias muestrales de todas las muestras posibles, es igual a la media poblacional, es decir:

$$\overline{x} = \frac{\overline{x_1} + \overline{x_2} + \ldots + \overline{x_k}}{k} = \mu$$

b) La desviación típica de las medias muestrales posibles es:

$$s_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

donde σ es la desviación típica poblacional y n es el tamaño de las muestras.

Conclusión: Las medias de las muestras de tamaño n extraídas de una población de parámetros μ y σ , siguen una distribución:

$$\overline{X} \longrightarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

siempre que dichas muestras tengan un tamaño $n \geq 30$.

Notas importantes:

a) Este resultado es consecuencia del *Teorema Central del límite.*

b) *Si la población es normal*, el resultado se cumple para muestras de CUALQUIER tamaño (incluso menor que 30).

c) Si σ es desconocida, el mismo resultado sigue siendo cierto sustituyendo en la fórmula σ por s.

Ejemplo: La altura de los estudiantes de una población se distribuye según una normal de media 167 y desviación típica 3'2.

a) Calcula la probabilidad de que *un* estudiante mida menos de 165 cm.

b) Se toma una muestra de 10 estudiantes. Calcula la probabilidad de que *la media muestral* sea menor que 165 cm.

En el apartado a) , manejamos la variable

$$X \longrightarrow N(165; 3'2)$$

siendo X= "altura de un estudiante".

La probabilidad pedida será:

$$p(X < 165) = p\left(\frac{X - 167}{3'2} < \frac{165 - 167}{3'2}\right) = p(Z < -0'63) = 0'2676$$

En el apartado b), la variable que manejamos ya no es X, sino que tenemos una muestra de 10 estudiantes. Como la población inicial es normal, podemos aplicar el resultado anterior aunque la muestra sea de tamaño menor que 30. Así, la variable a estudiar es

$\overline{X}$="media de las alturas de 10 estudiantes", que según lo dicho, sigue una distribución

$$\overline{X} \longrightarrow N\left(165; \frac{3'2}{\sqrt{10}}\right) = N(165; 1'012)$$

Nos piden:

$$p(\overline{X} < 165) = p\left(\frac{\overline{X} - 167}{1'012} < \frac{165 - 167}{1'012}\right) = p(Z < -1'97) = 0'0244$$

Ejemplo: Los pesos de los tornillos que fabrica cierta máquina se distribuyen según una $N(142'32; 8'5)$ (medidas en gr.). Se toman muestras de 25 tornillos. Calcular:

a) Distribución que siguen las medias de esas muestras.

b) Probabilidad de que una muestra elegida al azar de 25 tornillos tenga un peso medio superior a 144'6 gr.

c) La misma pregunta si la muestra es de 100 tornillos.

a) Como las muestras son de tamaño n=25 y la población es normal $N(142'32; 8'5)$, las medias muestrales siguen una distribución:

$$\overline{X} \longrightarrow N\left(142'32; \frac{8'5}{\sqrt{25}}\right) = N(142'32; 1'7)$$

b) Nos piden:

$$p(\overline{X} \geq 144'6) = p\left(Z \geq \frac{144'6 - 144'32}{1'7}\right) = p(Z \geq 1'34) = 1 - p(Z \leq 1'34) = 0'0901$$

c) Si las muestras son de tamaño n=100, las medias muestrales siguen una distribución:

$$\overline{X} \longrightarrow N\left(142'32; \frac{8'5}{\sqrt{100}}\right) = N(142'32; 0'85)$$

y por tanto:

$$p(\overline{X} \geq 144'6) = p\left(Z \geq \frac{144'6 - 144'32}{0'85}\right) = p(Z \geq 2'68) = 1 - p(Z \leq 2'68) = 0'0037$$

Ejercicio: Una máquina ha fabricado piezas de precisión con un peso medio de 150 gr. y una desviación típica de 20 gr. Calcular la probabilidad de que una muestra de 80 piezas tenga un peso medio de más de 155 gr. (Solución: 0'0129)

4.5. Distribución muestral de proporciones

Nos planteamos ahora determinar qué proporción de una población posee un cierto atributo, por ejemplo si es fumador o no fumador, si tiene ordenador o no, si tiene alergia o no,etc... El estudio de este tipo de proporciones es equiparable al de una distribución binomial (donde sólo hay dos posibilidades). Si la proporción éxito es p y la de fracaso q, y se toma una muestra de la población de tamaño n, al

igual que en el caso anterior, para cada muestra tendremos una proporción muestral que denotaremos por $\hat{p}$ y una desviación típica muestral que denotaremos por $s_{\hat{p}}$.

Entonces,utilizando razonamientos similares a los del apartado anterior, se verifica que $\hat{p} = p$, y $s_{\hat{p}} = \sqrt{\frac{p \cdot q}{n}}$ por tanto:

Conclusión: Las proporciones muestrales de tamaño $n \geq 30$, extraídas de una población en la que la probabilidad de éxito es p, se ajustan a una normal

$$N\left(p; \sqrt{\frac{p \cdot q}{n}}\right)$$

Ejemplo: Una fábrica de pasteles fabrica, en su producción habitual, un 3 % de pasteles defectuosos. Un cliente recibe un pedido de 500 pasteles de la fábrica.

a) Probabilidad de que encuentre más del 4 % de pasteles defectuosos.

b) Probabilidad de que encuentre menos de un 1 % de pasteles defectuosos.

a) En este caso éxito= "pastel defectuoso", y la proporción poblacional de éxito es de $p = \frac{3}{100}$ tanto, $q = \frac{97}{100}$. La muestra que recibe el cliente es de tamaño n=500.

Por tanto, las proporciones muestrales siguen una distribución:

$$\hat{p} \longrightarrow N\left(\frac{3}{100}; \sqrt{\frac{\frac{3}{100} \cdot \frac{97}{100}}{500}}\right) = N(0'03; 0'076)$$

puesto que las muestras tienen tamaño mayor que 30.

La probabilidad pedida es que la proporción de pasteles defectuosos en la muestra sea mayor del 4 %, es decir:

$$p(\hat{p} \geq 0'04) = p\left(Z \geq \frac{0'04 - 0'03}{0'0076}\right) = p(Z \geq 1'32) = 0'0934$$

b) En este caso es

$$p(\hat{p} \leq 0'01) = p\left(Z \leq \frac{0'01 - 0'03}{0'0076}\right) = p(Z \leq -2'63) = 0'0043$$

Ejercicios:

1. De una población de 120 alumnos , hay 48 que tienen 2 o más hermanos. Si de dicha población se toman muestras de tamaño 40.

   a) ¿Qué distribución siguen las proporciones muestrales?.

   b) ¿Cuál es la probabilidad de que se encuentre en dicha muestra una proporción de más del 55 % de alumnos con 2 o más hermanos?.

2. Las notas de cierto examen se distribuyen según una normal de media μ=5'3 y desviación típica σ=2'4. Hallar la probabilidad de que un estudiante tomado al azar tenga una nota:

   a) Superior a 6'5

   b) Inferior a 5'2

   c) Comprendida entre 5 y 6'5

   Halla las mismas probabilidades para de la media de las notas de 16 estudiantes elegidos al azar.

3. En un saco mezclamos judías blancas y pintas en la relación de 14 blancas por cada pinta. Extraemos un puñado de 100 judías. Calcula la probabilidad de que la proporción de judías pintas esté comprendida entre 0'05 y 0'1.

4. El cociente intelectual, CI, de unos universitarios se distribuye normalmente con media 100 y desviación típica 11.

   a) Se elige al azar una persona. Hallar la probabilidad de que su CI esté entre 100 y 103.

   b) Se elige al azar una muestra de 25 personas. Encontrar la probabilidad de que la media de sus cocientes intelectuales esté entre 100 y 103.

4.6. Intervalos de probabilidad

En una variable normal cualquiera $N(\mu, \sigma)$, se verifica que:
1. En el intervalo $(\mu - \sigma, \mu + \sigma)$ está el 68'26 % de la población.
2. En el intervalo $(\mu - 2 \cdot \sigma, \mu + 2 \cdot \sigma)$ está el 95'44 % de la población.
3. En el intervalo $(\mu - 3 \cdot \sigma, \mu + 3 \cdot \sigma)$ está el 99'74 % de la población.

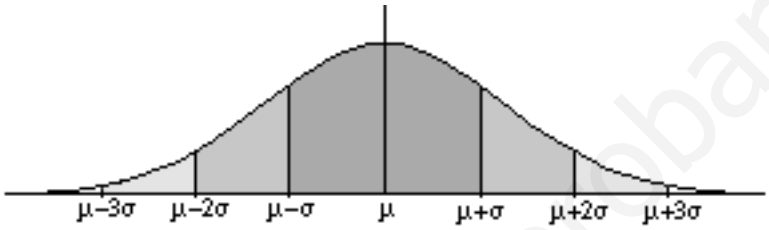

Figura 4.1: Porcentajes de población en los diferentes intervalos simétricos de una normal $N(\mu, \sigma)$.

Es evidente que a medida que el intervalo se amplía, hay mayor porcentaje de la población en él.

En general, dado un porcentaje del N %, siempre es posible encontrar un intervalo simétrico respecto de la media de forma que dicho intervalo contenga a dicho porcentaje de población.

Más explicitamente, se denomina *intervalo de probabilidad* a aquel intervalo para el cuál se sabe que hay una seguridad del N % de que los parámetros muestrales ($\overline{x}$ o $\hat{p}$) se encuentren en dicho intervalo. La seguridad N viene fijada previamente.

Si queremos que el N % de la población esté en el intervalo, denominaremos *nivel de confianza* al número:

$$1 - \alpha = \frac{N}{100}$$

y unido a este, se encuentra el llamado *nivel de significación*, que viene dado por α. Este nivel en general vendrá explicitado en las condiciones del problema, si bien los valores más comunes suelen ser del 90 %,95 % y 99 %.

Ejemplo: Si queremos que el 88 % de la población esté en el intervalo, el nivel de confianza será $1-\alpha = \frac{88}{100} = 0'88$, mientras que el nivel de significación será $\alpha = 1 - 0'88 = 0'12$.

4.6.1. Intervalo de probabilidad para la media muestral $\overline{x}$

Si la población sigue una distribución de parámetros μ y σ, y las muestras son de tamaño $n \geq 30$ (o bien la población ya es normal y las muestras son de cualquier tamaño), sabemos que la media muestral $\overline{x}$ sigue una distribución:

$$\overline{X} \longrightarrow N\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$$

Se trata de encontrar el valor de k como en la figura:

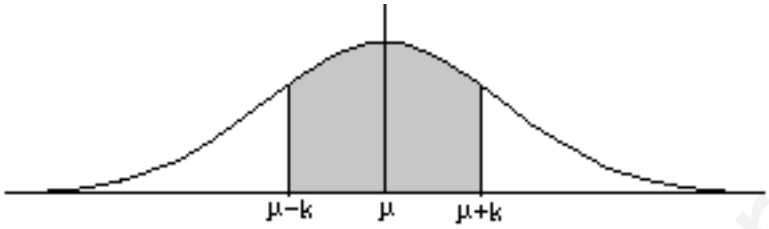

Figura 4.2: Buscamos el valor de k que deje en el intervalo $(\mu - k, \mu + k)$ al $(1 - \alpha) \cdot 100\,\%$ de la población.

Razonemos ahora sobre la normal Z $\longrightarrow$ N(0;1) que es la que se encuentra tabulada Si queremos que el intervalo buscado contenga a la media muestral con una confianza de $1 - \alpha$, entonces fuera del intervalo el área tiene que ser de α, y como la curva es simétrica, en cada una de las ramas fuera de la región rayada, tenemos un área de $\dfrac{\alpha}{2}$. Llamaremos $z_{\frac{\alpha}{2}}$ al punto situado en el eje x que separa la región rayada de la otra.

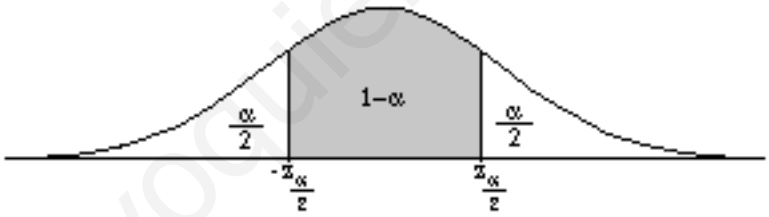

Figura 4.3: Buscamos el valor de $z_{\frac{\alpha}{2}}$ que deje en el intervalo $(-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}})$ al $(1 - \alpha)$ de la población en la N(0;1)

Es evidente que se cumple:

$$p\left(Z \geq z_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$$

o dicho de otro modo:

$$p\left(Z \leq z_{\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}$$

probabilidad que se busca *dentro de la tabla* como hemos visto anteriormente en el tema de la normal.

Ahora bien, este valor sólo sirve para la normal estándar N(0;1). Nosotros manejamos la normal $N\left(\mu; \dfrac{\sigma}{\sqrt{n}}\right)$ y para pasar a la normal estándar deberemos tipificar:

$$\frac{k - \mu}{\frac{\sigma}{\sqrt{n}}} = z_{\frac{\alpha}{2}}$$

de donde despejando, encontramos k, el valor buscado:

$$k = \mu + \frac{\sigma}{\sqrt{n}} \cdot z_{\frac{\alpha}{2}}$$

Así, dado el nivel de significación α o el de confianza $1-\alpha$, podemos determinar el intervalo de probabilidad para la media muestral, que será:

$$\left(\mu - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \mu + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

Ejemplo Determinar, en la distribución N(0;1), el valor que concentra el 75 % de la población en un intervalo simétrico respecto a la media

Ahora $1-\alpha = 0'75$, es decir $\alpha = 0'25$ y por tanto $\frac{\alpha}{2} = 0'125$, es decir, buscamos el valor $z_{0'125}$, de modo que, como en la figura, dejemos el 75 % de la población en el centro.

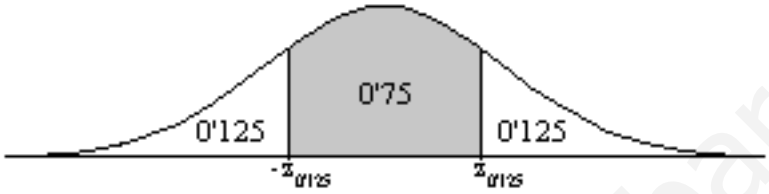

Figura 4.4: Buscamos el valor de $z_{0'125}$ que deje en el intervalo $(-z_{0'125}, z_{0'125})$ al 0'75 de la población en la N(0;1)

Se cumple que $p(Z \geq z_{0'125}) = 0'125$, es decir $p(Z \leq z_{0'125}) = 0'875$,y si buscamos en la tabla, resulta que el valor es:

$$z_{0'125} = 1'15$$

Ejercicio: Encuentra el valor correspondiente que concentre el 88 % de la población.

Ejemplo: Calcular el intervalo de probabilidad con un nivel de confianza del 95 % para la media de una muestra de 100 recién nacidos, sabiendo que la población de recién nacidos sigue una normal de media μ=3100 gr. y desviación tipica σ=150 gr.

Como el nivel de confianza es 0'95, entonces $1-\alpha = 0'95$ y por tanto $\alpha = 0'05$ y en cada rama fuera de la región queda $\frac{\alpha}{2} = 0'025$.

Buscamos entonces $z_{0'025}$, que es el valor que deja a su derecha un área de 0'025, es decir:

$$p(Z \geq z_{0'025}) = 0'025 \Longrightarrow p(Z \leq z_{0'025}) = 0'975$$

Buscando este valor dentro de la tabla se obtiene que el valor de $z_{0'025} = 1'96$, y por tanto el intervalo para la media muestral es:

$$\left(3100 - 1'96 \cdot \frac{150}{\sqrt{100}}, 3100 + 1'96 \cdot \frac{150}{\sqrt{100}}\right) = (3100 - 1'96 \cdot 15, 3100 + 1'96 \cdot 15) = (3070'6, 3129'4)$$

Esto significa que el 95 % de las muestras de tamaño 100 tendrá su media comprendida entre estos 2 valores: (3070'6,3129'4)

Ejercicio: Calcular el mismo intervalo con una confianza del 99 %.

Ejercicio: Las notas de una población de 150 alumnos siguen una distribución de media 5'5 y varianza 4'1616. Extaremos muestras de tamaño 36. Calcula el intervalo de probabilidad para un nivel de confianza del: a)75 % b) 86'64 %, e interpreta los resultados.

(NOTA: Recordemos que la varianza y la desviación típica de una distribución están relacionadas porque la varianza es el cuadrado de la desviación típica y se representa por σ^2).

4.6.2. Intervalo de probabilidad para la proporción muestral $\hat{p}$

Razonando de igual manera se puede llegar a que para el nivel de significación α el intervalo para la proporción muestral $\hat{p}$ es

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

donde p y q son las proporciones poblacionales y $n \geq 30$.

Ejercicio: Sabiendo que la proporción de alumnos con vídeo de una población de 120 alumnos es de p=0'7, halla el intervalo de probabilidad para la proporción de:

a) las muestras de tamaño 30 con una confianza del 75 %.

b) las muestras de tamaño 49 con una confianza del 90 %.

c) las muestras de tamaño 49 con una confianza del 99 %.

4.7. Estimación por intervalos

La estimación anterior, la puntual, se utiliza poco, pues no tenemos datos suficientes que nos indiquen el grado de fiabilidad del dato muestral que hemos tomado. Lo que tiene más sentido plantearse es cuál es la probabilidad de que la media o proporción poblacional pertenezcan a un intervalo determinado.

4.7.1. Estimación de la media de una población μ

La media μ de una población es desconocida y deseamos conocerla. Para ello, basándonos en los intervalos de probabilidad, sabemos que si la población tiene parámetros μ y σ, la media muestral $\overline{x}$ sigue una distribución:

$$\overline{X} \longrightarrow N\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$$

, siendo n el tamaño de la muestra, y sabemos que el intervalo de probabilidad a nivel de confianza $1 - \alpha$ para $\overline{x}$ es:

$$\left(\mu - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \mu + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

es decir, que:

$$\mu - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \overline{x} \leq \mu + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

De la primera desigualdad se sigue que:

$$\mu - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \overline{x} \Longrightarrow \mu \leq \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Y de la segunda:

$$\overline{x} \leq \mu + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \Longrightarrow \mu \geq \overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Luego se deduce que:

$$\overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Es decir, que el intervalo de confianza con nivel de confianza $1 - \alpha$ para la media poblacional μ desconocida es:

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

NOTA:

a) Hay que añadir que para aplicar este resultado, o bien las muestras tienen tamaño $n \geq 30$, o bien la población sigue una distribución normal.

b) Si la desviación típica de la población σ, es desconocida, se utilizará, la desviación típica muestral s en su lugar,y el intervalo sería:

$$\left(\overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}\right)$$

Al valor $\dfrac{\sigma}{\sqrt{n}}$ se le denomina *error típico o estándar.*

Ejemplo: Para estimar la media de los resultados que obtendrían al resolver un cierto test los alumnos de 4º de E.S.O. de la Comunidad de Castilla-León, se les pasa el test a 400 alumnos escogidos al azar, con los resultados de la tabla:

Puntuación	Número de alumnos
1	24
2	80
3	132
4	101
5	63

A partir de ellos, estima con un nivel de confianza del 95 % el valor de la media poblacional.

Aprovechando repasaremos el cálculo de algunos parámetros estadísticos.

Como sólo disponemos de la muestra, no conocemos la media ni la desviación típica poblacional, hemos de calcular la media y la desviación típica muestral.

Para ello, calculamos la tabla siguiente:

X	Frec.absoluta f_i	$X \cdot f_i$	X^2	$X^2 \cdot f_i$
1	24	24	1	24
2	80	160	4	320
3	132	396	9	1188
4	101	404	16	1616
5	63	315	25	1575
Total	400	1299		4723

Resulta:

$$\overline{x} = \frac{1299}{400} = 3'25$$

Varianza=$s^2 = \dfrac{4723}{400} - (3'25)^2 = 11'81 - 10'56 = 1'25$

$$s = \sqrt{s^2} = \sqrt{1'25} = 1'12$$

Ya tenemos los parámetros muestrales. Hemos de determinar el intervalo de confianza para μ. Como $1 - \alpha$= 0'95, resulta que $\alpha = 0'05$ y queda $\dfrac{\alpha}{2} = 0'025$.

Se obtiene que el valor es $z_{0'025} = 1'96$, por tanto el intervalo de confianza para μ, al 95 % es:

$$\left(3'25 - 1'96 \cdot \frac{1'12}{\sqrt{400}}, 3'25 + 1'96 \cdot \frac{1'12}{\sqrt{400}}\right) = (3'25 - 0'11, 3'25 + 0'11) = (3'14, 3'36)$$

Por tanto tenemos una confianza del 95 % de que la nota media de la población esté comprendida entre 3'14 y 3'36.

Ejercicio: De una variable estadística conocemos la desviación típica, 8, pero desconocemos la media. Para estimarla, extraemos una muestra de tamaño 60 cuya media es 37. Estimar la media poblacional con una confianza del 99 %.

Error máximo admisible:

Hemos visto que el intervalo de confianza para la media poblacional μ es:

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

Se cumple entonces que la diferencia, en valor absoluto, entre las medias poblacional y muestral es:

$$|\mu - \overline{x}| < z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Al valor

$$E = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

se le llama error máximo admisible. Dicho error tiene las siguientes propiedades:

a) El error es manor cuanto mayor sea el tamaño de la muestra (n), porque dividimos por un número cada vez mayor.

b) El error es mayor al aumentar el nivel de confianza, puesto que el valor $z_{\frac{\alpha}{2}}$ aumenta, como se observa en la tabla:

Confianza=$1-\alpha$	$z_{\frac{\alpha}{2}}$
0'9	1'645
0'95	1'96
0'99	2'575

Para reducir el error, por tanto, no hay que aumentarla confianza, sino el tamaño de la muestra elegida.

Si conocemos el error y el nivel de confianza, podemos calcular el tamaño de la muestra , usando la fórmula del error.

Ejercicio: Al medir un tiempo de reacción, un psicólogo sabe que la desviación típica del mismo es 0'5 segundos. ¿Cuál es el número de medidas que deberá realizar para que con una confianza del 99 %, el error de estimación no exceda de 0'1 segundos?.

4.7.2. Estimación de una proporción

Si para cierta población se desconoce la proporción p de individuos que poseen cierta característica, y deseamos dar un intervalo de confianza para el valor de p, como el intervalo de probabilidad para la proporción muestral,$\hat{p}$,para el nivel de confianza $1-\alpha$ en una muestra de tamaño n es:

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

Razonando igual que en el caso anterior, concluimos que:

El intervalo de confianza para p a un nivel de confianza de $1-\alpha$ es:

$$\left(\hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p \cdot q}{n}}, \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p \cdot q}{n}}\right)$$

Aunque como habitualmente no se conoce p en realidad se usa:

$$\left(\hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}, \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}\right)$$

NOTA: a) Es necesario que $n \geq 30$ para poder aplicar esta fórmula.

b) Habitualmente en las encuestas, se suele utilizar, en lugar de la última fórmula, el valor de p=q=0'5, que es la situación más desfavorable.

Ejercicio: Determina el intervalo de confianza, con una significación del 0'05 para la proporción poblacional de fumadores entre los jóvenes menores de 21 años, a partir de una muestra de tamaño 900, cuando no se conocen valores de p anteriores. Considera los dos casos anteriores (usando $\hat{p}$ y p=q=0'5). La proporción de fumadores en la encuesta ha sido de $\hat{p} = 0'3$.

El **error máximo admisible** en este caso es:

$$E = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p \cdot q}{n}}$$

o en caso de no conocer p:

$$E = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$$

Ejercicio: Para 96 familias españolas elegidas al azar se ha determinado que la TV permanece encendida en la casa una media de 217 minutos diarios, la desviación típica de la muestra fue de 40 minutos.

a) Para una fiabilidad del 95 % ¿qué error se asume cuando se da por bueno ese dato para el total de las familias españolas?.

b) ¿Qué tamaño muestral sería necesario para reducir ese error a la mitad?.

NOTA: Diferencia entre intervalos de probabilidad y de confianza

En un intervalo de probabilidad *lo que conocemos es la media y desviación típica poblacionales*, y damos el intervalo donde se encontrará (para un cierto nivel de confianza) la media muestral o la proporción muestral.

Sin embargo, en un intervalo de confianza entramos ya en el terreno de la estimación, es decir NO conocemos la media poblacional (y en ocasiones tampoco la desviación típica poblacional) ni la proporción poblacional , sino que sólo *conocemos, o podemos calcular, la media muestral o la proporción muestral*, y de lo que se trata es de dar un intervalo en el que se encuentre la media poblacional (o la proporción poblacional).