4

DISTRIBUCIONES MUESTRALES

4.1 INTRODUCCIÓN

Hemos explicado antes que la inferencia estadística tiene como problema general el establecimiento de las propiedades de un fenómeno aleatorio estudiando una parte del mismo. Igualmente se ha dicho que para lograr esto es necesario conocer la distribución de probabilidades de la variable aleatoria a través de la cual se expresa el fenómeno. Por tales razones, en los capítulos anteriores se ha hecho énfasis en conocer algunas ideas básicas de la teoría de probabilidad y deducido las propiedades de algunos modelos que describen la distribución probabilística de cierto tipo de variables aleatorias. En la Figura 4.1 se presenta un modelo gráfico de las relaciones que existen entre la Probabilidad y la Estadística Inferencial.

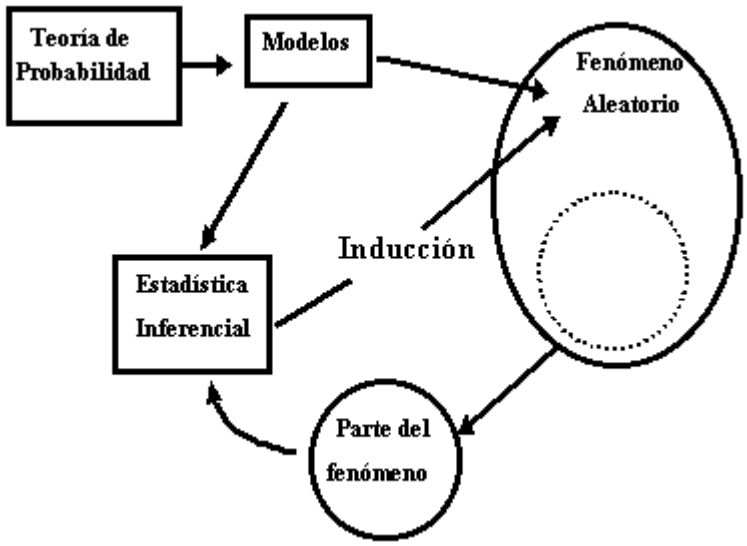


Figura 4.1: Modelo de relaciones entre la probabilidad y la estadística inferencial

El esquema anterior nos muestra que la teoría de probabilidad genera los modelos que describen la distribución de probabilidades de los resultados de un experimento aleatorio, mientras que los métodos de inferencia estadística evalúan las características de una parte del fenómeno y utilizando esos mismos modelos de probabilidad producen por inducción, conclusiones sobre la totalidad del fenómeno.

En la estadística inferencial existe toda una terminología que identifica las diferentes partes y procesos del modelo presentado en la Figura 4.1. Con el propósito de manejar adecuadamente ésta terminología será necesario definir algunos conceptos básicos, para luego estudiar algunas propiedades de la porción estudiada del fenómeno, así como la relación funcional que existe entre ella con el colectivo.

4.2 ALGUNOS CONCEPTOS IMPORTANTES

4.2.1 Universo, población y muestra

Un fenómeno aleatorio sería toda manifestación material susceptible de observarse o medirse mediante los sentidos o instrumentos en individuos, cosas o elementos similares que forman parte de un colectivo denominado Universo. Este colectivo puede estar formado por un número finito o infinito de tales unidades. Una Observación es un dato o valor numérico que se obtiene al calificar o cuantificar una característica en las diferentes unidades. El conjunto de observaciones origina una Población**,** la cual puede estar formada por un número finito o infinito de datos o valores numéricos. Una Muestra es un conjunto formado por n observaciones extraídas de la población. El número n de observaciones define el tamaño de la muestra. En la Figura 4.2 se esquematizan las relaciones que existen entre estos conceptos.

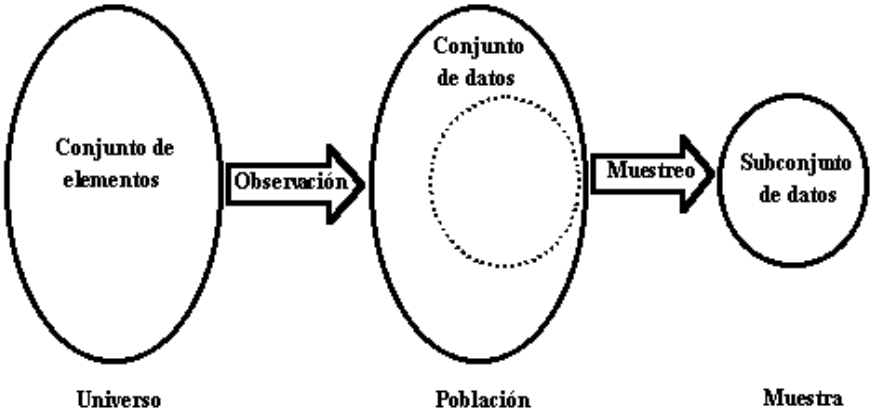


Figura 4.2**:** Esquema de relaciones entre Universo, Población y Muestra.

A continuación se ejemplificarán los conceptos de Universo, Población y Muestra:

Ejemplo 4.1

Un productor agrícola quiere conocer algunas características de las mazorcas producidas en una parcela sembrada con plantas de maíz. Para tal fin selecciona 50 mazorcas y cuenta el número de granos en cada mazorca.

Universo: Conjunto formado por todas las mazorcas de maíz que produjo la parcela. Este universo es finito porque lo forma el total de mazorcas producidas.
Característica observada: El número de granos de cada mazorca.
Población: Conjunto de todos los valores de la característica número de granos de cada mazorca. Esta población es finita porque lo forma el total de valores del número de granos obtenidos del universo.
Muestra: Conjunto finito de 50 valores de la característica número de granos.

Ejemplo 4.2

El mismo productor seleccionó del ejemplo anterior, seleccionó 20 mazorcas y determinó el peso de cada una.

Universo: El mismo del ejemplo anterior.
Característica observada: El peso de cada mazorca.
Población: Conjunto finito de todos los valores de peso de cada mazorca.
Muestra: Conjunto finito de 20 valores de la característica peso de cada mazorca.

Ejemplo 4.3
Un biólogo quiere conocer algunas características de los rabipelados *Didelphus marsupialis*. Para lograr esto, seleccionó 100 individuos de la especie en cuestión y le determinó a cada uno el número de glándulas sebáceas en los miembros anteriores.

Universo: Conjunto de rabipelados de la especie *Didelphus marsupialis*. Este universo es infinito porque está formado por todos los ejemplares que viven actualmente, los que vivieron alguna vez y los que van a existir en el futuro.
Característica observada: El número de glándulas sebáceas.
Población: Conjunto de todos los valores de la característica número de glándulas sebáceas. Esta población es infinita por las mismas razones que hacen infinito al universo de rabipelados.
Muestra: Conjunto finito de 100 valores de la característica número de glándulas sebáceas.

Ejemplo 4.4
El biólogo del ejemplo anterior midió el contenido de Hemoglobina en la sangre de 500 rabipelados.

Universo: Igual al anterior
Característica observada: La concentración de hemoglobina.
Población: Conjunto de todos los valores de la concentración de hemoglobina Esta población es infinita por las mismas razones que hacen infinito al universo de rabipelados.
Muestra: Conjunto finito de 500 valores de la característica concentración de hemoglobina

Ejemplo 4.5
El mismo biólogo desea conocer el tamaño de los sapos del género *Atelopus* que viven actualmente en la selva nublada de Monte Zerpa. Con éste propósito capturó 35 individuos y midió la longitud del cuerpo de cada ejemplar.

Universo: Conjunto formado por todos los sapos del género *Atelopus* que viven actualmente en Monte Zerpa. Este universo es finito porque el biólogo sólo esta interesado en el conjunto de individuos que existe actualmente y no en los que vivieron o vivirán.
Característica observada: La longitud.
Población: Conjunto de todos los valores del tamaño. Esta población es finita por las mismas razones que hacen infinito al universo de sapos.
Muestra: Conjunto finito de 35 valores de longitud o tamaño.

Ejemplo 4.6
Se quiere conocer el valor promedio de la temperatura del agua del río Albarregas a la altura de los 2000 m.s.n.m. y para el año 1997. Para lograr esto se midió la temperatura del agua en el mismo lugar y hora todos los días de ese año.

Universo: Este universo se puede considerar infinito puesto que en el lugar escogido para realizar el experimento existen infinitos puntos sobre la superficie de la corriente de agua en los cuales se puede introducir el termómetro.
Característica observada: La temperatura del agua.
Población: Conjunto de todos los valores de la temperatura. Esta población es infinita por las mismas razones que hacen infinito al universo de puntos de medición de la temperatura.
Muestra: Conjunto de 365 valores de temperatura del agua.

De los ejemplos anteriores se pueden obtener dos conclusiones importantes: la primera es que los conceptos de universo y población son relativos y es el investigador quien determina, según su interés, la extensión del universo y consecuentemente la de la población a estudiar. Así vemos como en los ejemplos 3 y 4 el biólogo al decidir estudiar los rabipelados al nivel taxonómico de especie, estaba también decidiendo estudiar un universo infinito. Por el contrario, en el ejemplo 5 limitó su estudio a los sapos del género *Atelopus* que viven en un sitio determinado, es decir que decidió trabajar con un universo finito. La segunda conclusión que puede obtenerse es que de un universo se pueden generar varias poblaciones. Así vimos que del mismo universo de mazorcas se generó una población de números de granos y otra de peso de los granos.
De acuerdo a la nueva terminología se puede rediseñar el esquema de la Figura 4.1. En la Figura 4.2 se muestran estas nuevas relaciones entre la Estadística Inferencial y la Teoría de Probabilidad.

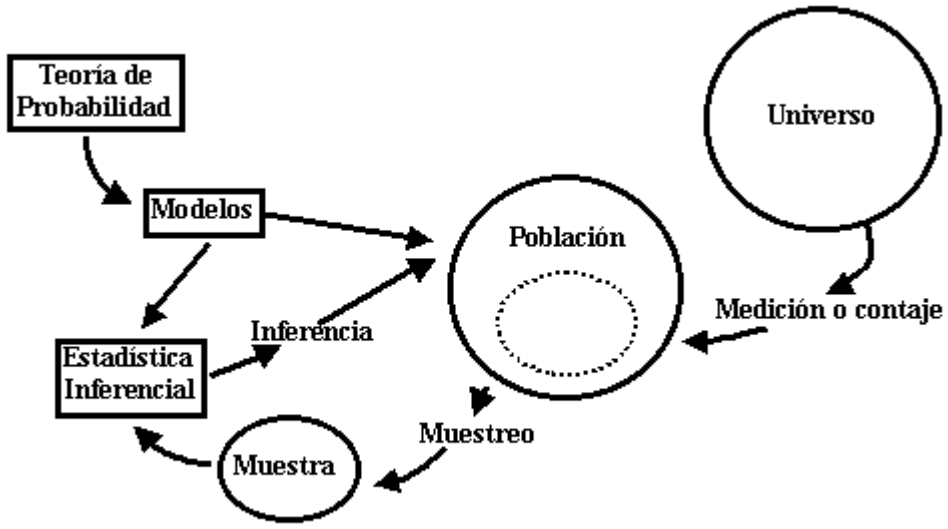


Figura 4.2: Relaciones entre la probabilidad y la estadística inferencial

4.2.2 Parámetros y estadísticos

Cuando estudiamos un fenómeno aleatorio, realmente lo que estamos haciendo es analizar las propiedades de las diferentes poblaciones de las variables que lo caracterizan. Muchas de las propiedades poblacionales son descritas por valores que reciben el nombre genérico de Parámetros (Figura 4.3). Por lo general los parámetros se identifican mediante una letra griega y son valores únicos que no cambian entre tanto no cambie la composición de la población.

Algunos de los parámetros poblacionales más importantes son: el promedio (μ), la varianza (σ^2) y la desviación (σ).

Las muestras también tienen características propias y relacionadas funcionalmente con las propiedades de la población. Estas características muestrales reciben el nombre de Estadísticos (Figura 4.3), y a diferencia de los parámetros son variables y cambian de muestra a muestra. Los estadísticos se identifican con letras del alfabeto romano y entre los más importantes se pueden señalar la media aritmética ($\overline{X}$); la varianza (S^2) y la desviación estándar (S).

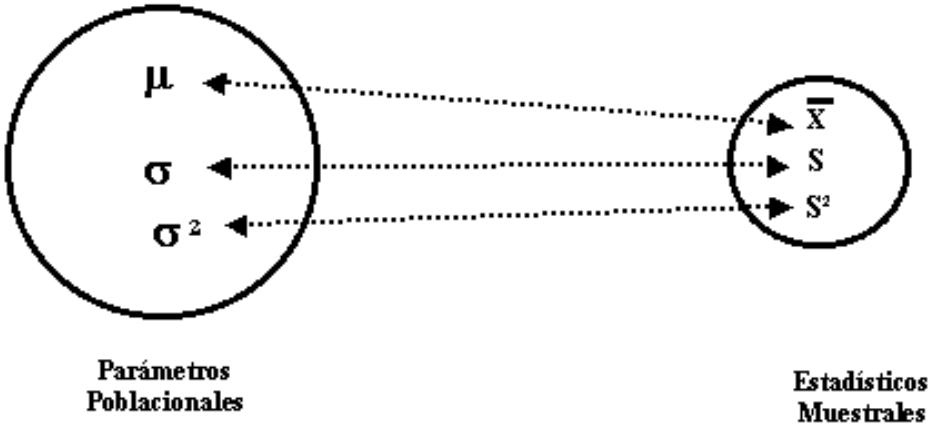


Figura 4.3

4.3 MUESTRAS Y MUESTREO

4.3.1 Muestra representativa

Las muestras deben proporcionar la información necesaria (estadísticos), a partir de la cual se infieren las propiedades (parámetros) de la población (Figura 4.3). En una buena muestra debe estar representada toda o al menos una gran parte de la información presente en la población. Para que una muestra sea representativa debe incluir los valores de la variable en la misma proporción como ellos se encuentran repartidos en la población. Los resultados que se exponen en la Tabla 4.1, representan la producción porcentual de cuatro diferentes variedades de soya obtenida en una determinada región y los valores de ésta misma producción de acuerdo a lo estimado por dos muestras de la producción.

Tabla 4.1. Producción (%) de diferentes variedades de soya

Variedad	Producción real	Muestra 1	Muestra 2
A	52 %	25%	49%
B	24%	35%	26%
C	18%	22%	17%
D	6%	18%	8%

De los resultados presentados en la tabla anterior se deduce que la distribución de la producción de soya evidenciada por la muestra 2 y la distribución de la producción real son muy parecidos, por lo tanto, se puede decir que la muestra es representativa de la producción de la población. Por el contrario la muestra 1 proporciona una distribución de la producción que no se corresponde con la de la región y obviamente no es representativa.

Lograr que una muestra sea representativa es una tarea difícil, especialmente si se trata de poblaciones infinitas. Una manera de hacerlo es tomando muestras grandes. En la medida que se aumenta el tamaño de la muestra se incrementa la posibilidad de que todos los grupos de valores de la variable que caracteriza la población estén representados, llegando al extremo de afirmar que la mejor muestra es aquella que tiene el mismo tamaño de la población. Sin embargo, este procedimiento además de desvirtuar el fundamento de la estadística inferencial, trae consigo una serie de problemas de orden técnico y económico. Un aumento en el tamaño de la muestra puede significar incrementos importantes en los costos, en el tiempo o en la dificultad para manejar una mayor cantidad de información.

4.3.2 Muestreo aleatorio

Otra manera de lograr que una muestra sea representativa es eligiendo aleatoriamente los valores que van a formar parte de la muestra. Mediante el muestreo aleatorio todos los valores de la población tienen la misma posibilidad de ser elegidos, de modo que si se toma una muestra de un tamaño adecuado y se eligen aleatoriamente los elementos que la conforman se está asegurando la representatividad de la muestra. El muestreo aleatorio puede ser simple o restringido.

El ejemplo siguiente puede aclarar el funcionamiento del muestreo aleatorio simple. Suponga que se quieren seleccionar 24 ratones de un grupo de 100 con el propósito de determinar la concentración promedio de una hormona en el grupo de animales. En primer lugar es necesario advertir que un universo de este tipo puede ser bastante heterogéneo, puesto que puede estar formado por individuos con diferentes progenitores, sexo, tamaño, peso, edad, etc. Consecuentemente la población de valores de la hormona también es heterogénea. Para que la muestra sea representativa es necesario que en ella estén presentes valores provenientes de todas las categorías y en la misma proporción como están repartidas en la población. Si elegimos aleatoriamente los ratones, cada uno de ellos tiene la misma posibilidad de ser seleccionado y la probabilidad de que cada característica sea escogida es proporcional a su tamaño. Estas dos cualidades del proceso de elección deben hacer que la composición de la muestra se aproxime a la de la población.

En el ejemplo anterior la selección aleatoria de los 24 ratones se puede hacer de diferentes maneras. Una manera, sería marcando cada uno de los ratones con un número entre 1 y 100. Igualmente se pueden numerar un lote de tarjetas desde el 1 hasta el 100, se colocan en un envase, se mezclan y se escogen 24 tarjetas. Del grupo de 100 ratones se escogerán aquellos cuyos números coincidan con los marcados en las tarjetas seleccionadas. Otra forma es utilizando una tabla de números aleatorios. Estas tablas están formadas por columnas de números de tres, cuatro, cinco o más dígitos, que han sido generados aleatoriamente, mediante algún procedimiento matemático. Para usar la tabla hay que entrar de una forma aleatoria. Una manera es cerrando los ojos y colocar la punta de un lápiz sobre el contenido de la tabla. Los dos últimos dígitos del número marcado por la punta del lápiz formarán el primer número seleccionado. Luego se puede avanzar hacia el número siguiente y de aquí al siguiente y así sucesivamente hasta haber seleccionado el total de números requeridos. En el ejemplo se necesitan escoger 24 números. El avance al número siguiente puede ser hacia abajo, o hacia

arriba, hacia la derecha o hacia la izquierda, en todo caso es indiferente. En cada avance se anota las dos últimas cifras de cada número. Si se llega al final de una columna o una fila sin haber completado el total de números que se desea, se pasa a la columna o fila siguiente y se comienza a avanzar en cualquiera de los sentidos que se prefiera. No se deben tomar en cuenta los números repetidos. Si el total de número a elegir es de tres cifras, se sigue el mismo procedimiento pero en lugar de anotar los dos últimos dígitos, se registran los tres últimos.

4.3.3 Muestreo estratificado

En muchas ocasiones el tamaño de la muestra no es lo suficientemente grande para asegurar que las distintas categorías de valores de una población estén representadas proporcionalmente. Si no es posible aumentar el tamaño de la muestra, se puede recurrir a un muestreo aleatorio restringido, el cual aumenta la posibilidad de obtener muestras representativas. Entre los varios tipos de muestreo restringido que existen se pueden mencionar los siguientes: el muestreo estratificado, el muestreo por agrupamiento, el muestreo sistemático, el muestreo secuencial, etc. Aquí sólo nos referiremos al muestreo estratificado por ser el más utilizado en las ciencias biológicas.

En el muestreo estratificado se divide la población en estratos o subpoblaciones dentro de las cuales se procede a realizar un muestreo aleatorio simple. El tamaño de las muestras pueden ser proporcional al tamaño de los estratos o todas pueden ser del mismo tamaño independientemente del tamaño de los estratos. Volvamos al ejemplo de los ratones a fin de entender este procedimiento. Como indicamos anteriormente el universo de ratones puede ser muy heterogéneo con relación a diferentes características como el sexo, el tamaño, la edad, etc. Estas mismas características nos pueden servir para estratificar la población. Por ejemplo, podemos clasificar los ratones de acuerdo al estado de desarrollo del proceso reproductivo en tres categorías: inmaduros, maduros y post-reproductivos. De acuerdo al ejemplo, la muestra de 24 valores de la hormona que se está estudiando se puede medir seleccionando aleatoriamente el mismo número de ratones dentro de cada una de estas categorías, o seleccionando dentro de cada categoría un número de ratones que sea equivalente a su proporción en la población. Supóngase que la distribución de los individuos dentro de cada categoría es la presentada en la Tabla 4.2.

Tabla 4.2. Distribución del número de individuos en las diferentes etapas de desarrollo de una raza de ratones

Etapa de desarrollo	Población	Muestras	
		Estratificación uniforme	Estratificación Proporcional
Inmaduro	60	8	14
Maduro	30	8	8
Post-reproductivo	10	8	3
Total	100	24	24

De acuerdo al muestreo estratificado uniforme para cada categoría se eligieron aleatoriamente ocho ratones. En cambio de acuerdo al muestreo estratificado proporcional, los 24 ratones de

la muestra se repartieron en las tres categorías manteniendo la misma proporción de la población, aproximadamente el 60%, 30% y 10% respectivamente.

4.4 DISTRIBUCIONES MUESTRALES

Como ya sabemos un estadístico es una propiedad muestral cuyo valor cambia de muestra a muestra, por lo cual se comporta como una variable aleatoria. En consecuencia debe existir un modelo o función de probabilidad que describa su distribución de probabilidades, la cual se denomina distribución muestral**.** Estas relaciones se muestran en la Figura 4.4.

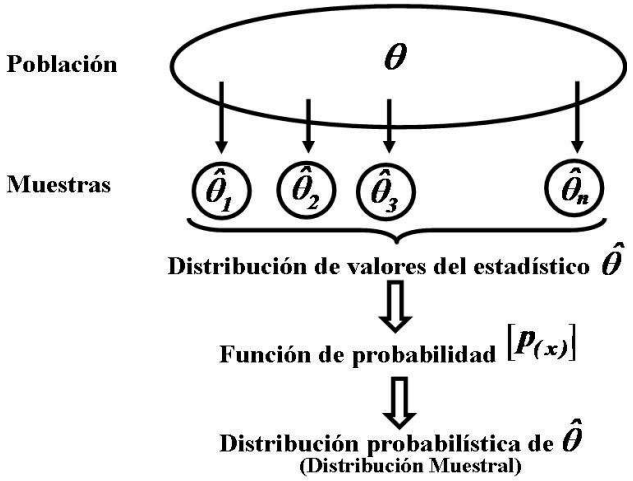


Figura 4.4. Relación funcional entre un parámetro poblacional (θ) y su respectivo estadístico muestral (θ). El símbolo θ representa un parámetro poblacional como la media (μ), la varianza (σ^2), la desviación (σ), etc. El símbolo θ representa el estadístico muestral respectivo como la media $\left(\overline{X}\right)$, la varianza (S^2), la desviación estándar (S), etc.

La importancia de conocer y comprender las distribuciones muestrales resulta del valor que ellas tienen para la inferencia estadística, tema que se tratará en los próximos capítulos. Por los momentos lo primordial es conocer las principales distribuciones y familiarizarse con sus propiedades.

4.4.1 Distribución de la media muestral

Si de una población de valores de una variable aleatoria *X* que se distribuye normalmente con media μ_x y varianza σ_x^2 se extrae una muestra de tamaño *n*, entonces se puede calcular la media $(\overline{x})$ de la muestra. Esta media representa una de las muchas medias muestrales que se pueden extraer de la población de valores de la variable *X*. Por lo tanto, la media muestral a su vez es una nueva variable aleatoria $\overline{X}$ que conforma una nueva población cuyos parámetros $\mu_{\overline{x}}$ y $\sigma_{\overline{x}}^2$ se pueden deducir mediante la aplicación de la denominada Propiedad Reproductiva de la distribución normal.

4.4.4.1 Propiedad Reproductiva de la distribución normal.

Sean X_1, X_2, X_3, ..., X_n, variables que se distribuyen normalmente, con la misma media: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_n$ y la misma varianza: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots = \sigma_n^2$. La variable que resulta de la suma de las n variables individuales: $X = X_1 + X_2 + X_3 + \ldots + X_n$, también se distribuye normalmente con una media: $\mu_x = \mu_1 + \mu_2 + \mu_3 + \ldots + \mu_n = n\mu$ y una varianza: $\sigma_x = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \ldots + \sigma_n^2 = n\sigma^2$

Puesto que es posible demostrar que cada uno de los valores $(x_1, x_2, x_3, \ldots . x_n)$ que forman parte de una muestra es una variable aleatoria que proviene de una misma población, se puede concluir que la media muestral es una variable que resulta de la suma de varias variables que tienen la misma μ y la misma varianza σ^2.

$$\overline{X} = \frac{\sum_{x=1}^{x=n} x_i}{n} = \frac{1}{n}\left(x_1 + x_2 + x_3 + \ldots\ldots + x_n\right) = \frac{x_1}{n} + \frac{x_2}{n} + \frac{x_3}{n} + \cdots + \frac{x_n}{n}$$

Por lo tanto, la media y la varianza de la distribución de medias muestrales serán:

$$\mu_{\overline{x}} = n\frac{\mu}{n} = \mu \qquad \text{y} \qquad \sigma_{\overline{x}}^2 = n\frac{\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

En resumen, si se toman muestras aleatorias de la población de una variable *X* que se distribuye normalmente, la distribución de las medias muestrales también es normal con una media igual a la media de la población de la variable *X* y una varianza igual a la de la población dividida entre el tamaño de la muestra (Figura 4.5). La desviación de la distribución de medias muestrales se le denomina error estándar y se obtiene como el cociente de dividir la desviación de la población de la variable X entre la raíz cuadrada del tamaño de la muestra $\sigma_{\overline{x}} = \sigma / \sqrt{n}$

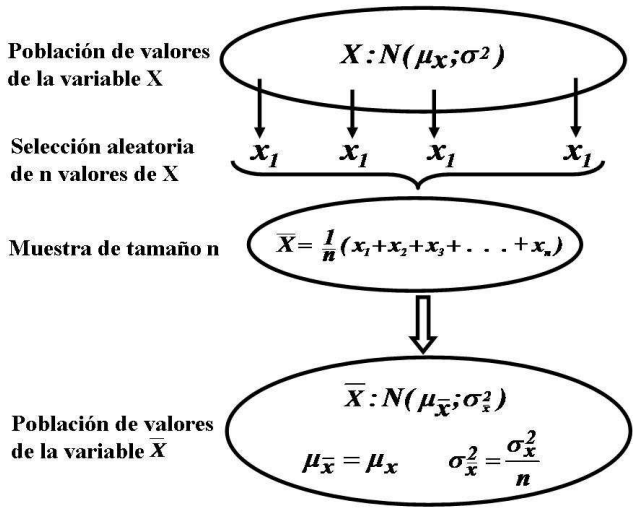


Figura 4.5. Relación entre las distribuciones de las variables X y $\overline{X}$.

La aplicación inmediata que le podemos dar a la distribución de la media muestral es calcular la probabilidad de obtener una media de un valor determinado.

Ejemplo 4.7
Sea una población de una variable que se encuentra distribuida normalmente con una media y una varianza igual a 800 y 1600 respectivamente, de la cual se seleccionan aleatoriamente 16 valores ¿Cuál es la probabilidad de que la muestra tenga un valor menor a 775?

Por la propiedad reproductiva sabemos que la media de una muestra obtenida de una población de valores distribuidos normalmente, también se distribuye normalmente con una media y una varianza igual a:

$$\mu_{\overline{x}} = \mu_x = 800 \qquad y \qquad \sigma^2_{\overline{x}} = \frac{\sigma^2_x}{n} = \frac{1600}{16} = 100$$

Además sabemos que para poder encontrar la probabilidad de ocurrencia de la variable aleatoria $\overline{X}$ que sigue una distribución normal es necesario hallar su valor equivalente en términos de la variable ***Z***, para lo cual se recurrimos al estadístico,

$$\frac{\overline{x} - \mu_x}{\sigma_x / \sqrt{n}}$$

Por lo tanto, la probabilidad deseada será:

$$P\left(\overline{X} \le 775\right) = P\left(Z \le \frac{\overline{x} - \mu_{\overline{x}}}{\sigma_{\overline{x}}}\right) = P\left(Z \le \frac{\overline{x} - \mu_x}{\sigma_x / \sqrt{n}}\right) = P\left(Z \le \frac{775 - 800}{40 / \sqrt{16}}\right) = P\left(Z \le -2{,}5\right) = 0{,}0062$$

Ejemplo 4.8
La concentración de fructuosa en el semen de toro se sabe que se distribuye normalmente con una media de 80 mg/100 ml. y una varianza igual a 900 (mg/100 ml)². Si en 36 ocasiones se midió el contenido de fructuosa en 100 ml de semen ¿Cuál es la probabilidad de que la media muestral se encuentre dentro del intervalo definido entre 70 y 90 mg/100 ml de semen?

Aplicando las propiedades de las distribuciones de las variables $\overline{X}$ y Z, se tiene que:

$$P\left(70 \le \overline{X} \le 90\right) = 0.9544\left(z_1 \le Z \le z_2\right) = P\left(\frac{70 - 80}{30 / \sqrt{36}} \le Z \le \frac{90 - 80}{30 / \sqrt{36}}\right) =$$

$$= P\left(-2 \le Z \le +2\right) = P\left(Z \le 2\right) - P\left(Z \le -2\right) = 0{,}9772 - 0{,}0228 = 0{,}9544$$

4.4.1.2 Teorema del límite central.

Supóngase ahora que se tiene una variable de la cual se conoce la media μ_x y la varianza σ_x^2 pero no la forma de su distribución. Esto impide la aplicación de la propiedad reproductiva y consecuentemente la deducción de los parámetros que caracterizan la distribución de las medias muestrales. Sin embargo, se puede recurrir a otra propiedad de la distribución normal conocida como el Teorema del Límite Central, que establece lo siguiente:

Sean $X_1, X_2, X_3, \ldots, X_n$ variables independientes con una misma función de probabilidad y por tanto con una misma distribución e igual $\mu_1=\mu_2=\mu_3=\ldots=\mu_n$ e igual varianza $\sigma_1^2=\sigma_2^2=\sigma_3^2=\ldots=\sigma_n^2$. La variable que resulta de la suma de las n variables independientes $X = X_1 + X_2 + \ldots + X_n$ también se distribuye normalmente con una media: $\mu_x=\mu_1+\mu_2+\mu_3+\ldots+\mu_n=n\mu$ y una varianza: $\sigma_x^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \ldots + \sigma_n^2 = n\sigma^2$ cuando n es grande.

En términos menos formales, el teorema anterior establece que las medias provenientes de muestras grandes tomadas de poblaciones con una distribución desconocida, se distribuyen normalmente con media y varianza igual a:

$$\mu_{\bar{x}} = \mu_x \qquad \text{y} \qquad \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

Por lo tanto, si se desconoce la distribución de una variable se puede suponer que aumentando el tamaño de la muestra, la distribución de la media muestral se aproximará progresivamente a una normal. En general y para efectos prácticos se considera que una muestra de tamaño $n \geq 30$ es lo suficientemente grande para que se cumpla el teorema.

Ejemplo 4.9

Se sabe que la concentración de albúmina en una sustancia tiene un valor promedio igual a 20,9 g/l y una desviación estándar de 3,45 g/l. Cuál será la probabilidad de encontrar una media superior a los 22 g/l en una muestra formada por 36 valores.

En éste caso no es posible aplicar la propiedad reproductiva puesto que se desconoce la forma de la distribución de la variable aleatoria. Sin embargo como el tamaño de la muestra es grande $(n \geq 30)$, se puede recurrir al Teorema del Límite Central y suponer que la media muestral tiene una distribución normal. Bajo esta situación es posible conocer los valores de $\mu_{\bar{x}}$ y $\sigma_{\bar{x}}$.

$$\mu_{\bar{x}} = \mu_x = 20,9 \qquad \text{y} \qquad \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{3,45}{\sqrt{36}} = 0,575$$

El valor de Z es $z = \dfrac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \dfrac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}} = \dfrac{22 - 20.9}{0,575} = 1,91$ y la probabilidad buscada será

$$P\left(\bar{X} \geq 22\right) = P\left(Z \geq 1,91\right) = 1 - P\left(Z \leq 1,91\right) = 1 - 0,9719 = 0,0281$$

4.4.1.3 El teorema de límite central y los fenómenos biológicos.

La tendencia que tiene la distribución de la media muestral de ser normal en la medida que aumenta el tamaño de la muestra independientemente de su distribución original, posiblemente explique porque tantos fenómenos biológicos siguen una distribución normal. Por ejemplo, considérese la variable altura en los humanos. Esta variable que sigue una distribución normal, es una característica fenotípica que resulta de la acción conjunta de factores genéticos y ambientales. Son varios los genes que afectan el crecimiento, igualmente algunos factores ambientales como la alimentación, las enfermedades, el ejercicio, etc., actúan sobre la talla de las personas. Unos de estos factores tienden a aumentar el tamaño y otros a disminuirlo. De manera que la altura (variable aleatoria), debe cumplir con el Teorema del Límite Central, si se le considera como el promedio (media muestral) que resulta de la suma de numerosos factores (variables independientes) sobre cada individuo (muestra aleatoria). Es importante puntualizar que dada la complejidad de los sistemas vivos, las variables biológicas no siempre tienen una distribución normal aunque sean resultado de la integración de muchos factores internos y/o externos. Posiblemente los postulados del teorema también se cumplan en otros fenómenos naturales que dependen de la interacción de muchos factores. En parte esto podría explicar, la gran frecuencia con que la distribución normal aparece en las mediciones de los fenómenos naturales.

4.4.2 Distribución de la diferencia de medias muestrales

Muchas veces es necesario estudiar dos poblaciones de una misma variable. Supongamos que la variable se distribuye normalmente en ambas poblaciones y que de cada una se extrae independientemente una muestra aleatoria con tamaños n_1 y n_2 respectivamente, y que además se calcula la media de las dos muestras. A partir de éstas dos medias muestrales es posible generar nuevas variables que relacionen las dos poblaciones Por ejemplo, se pueden sumar, restar, multiplicar o dividir los valores de las dos medias muestrales y originar otras variables cuyos valores estarían representadas por el resultado de las operaciones realizadas. De estas nuevas variables la más conveniente para la inferencia estadística es la diferencia de medias muestrales debido que se conocen las propiedades de su distribución de frecuencia. Cuando el muestreo de una variable se hace a partir de poblaciones que se distribuyen normalmente, la diferencia de medias muestrales es una nueva variable (Figura 4.6) que de acuerdo a la propiedad reproductiva también se distribuye normalmente con media y varianza igual a:

$$\mu_{(\bar{x}_2-\bar{x}_1)} = \mu_{(\bar{x}_2)} - \mu_{(\bar{x}_1)} = \mu_{(x_2)} - \mu_{(x_1)}$$

$$\sigma^2_{(\bar{x}_2-\bar{x}_1)} = \sigma^2_{\bar{x}_2} + \sigma^2_{\bar{x}_1} = \frac{\sigma^2_{x_2}}{n_2} + \frac{\sigma^2_{x_1}}{n_1}$$

Conocido el modelo de probabilidad que describe la distribución de la diferencia de medias muestrales, se puede calcular la probabilidad de ocurrencia de un determinado valor de la diferencia de medias muestrales, utilizando la transformación de Z.

$$Z = \frac{(\bar{x}_2 - \bar{x}_1) - \mu_{(\bar{x}_2 - \bar{x}_1)}}{\sigma^2_{(\bar{x}_2 - \bar{x}_1)}} = \frac{(\bar{x}_2 - \bar{x}_1) - (\mu_{\bar{x}_2} - \mu_{\bar{x}_1})}{\sqrt{\sigma^2_{\bar{x}_2} + \sigma^2_{\bar{x}_1}}} = \frac{(\bar{x}_2 - \bar{x}_1) - (\mu_{x_2} - \mu_{x_1})}{\sqrt{\frac{\sigma^2_{x_2}}{n_2} + \frac{\sigma^2_{x_1}}{n_1}}}$$

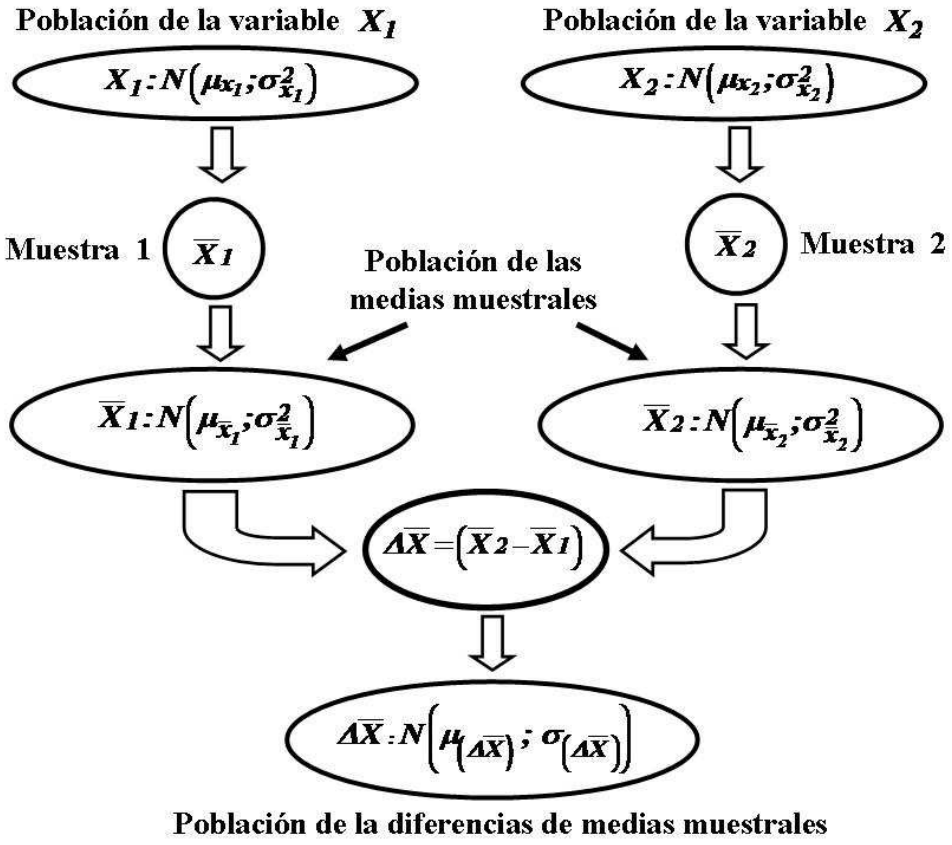


Figura 4.6. Relación funcional entre las variables X_1 y X_2 con las medias de muestras y la diferencia de medias muestrales.

Ejemplo 4.10

Una muestra de tamaño 5 se obtiene aleatoriamente en una población de una variable normalmente distribuida con media igual a 50 y varianza igual a 9 y se registra la media muestral. Otra muestra aleatoria de tamaño 4 se selecciona en una segunda población de la misma variable cuya media es igual a 40 y su varianza igual a 4. Encuentre la probabilidad de que el valor de la diferencia de las medias muestrales sea menor a 8,2.

Por la propiedad reproductiva de la distribución normal sabemos que $\left(\overline{X}_2 - \overline{X}_1\right)$ se distribuye normalmente con una media y una varianza igual a:

$$\mu_{(\bar{x}_2 - \bar{x}_1)} = \mu_{(\bar{x}_2)} - \mu_{(\bar{x}_1)} = \mu_{(x_2)} - \mu_{(x_1)} = 50 - 40 = 10$$

$$\sigma^2_{(\bar{x}_2-\bar{x}_1)} = \sigma^2_{\bar{x}_2} + \sigma^2_{\bar{x}_1} = \frac{\sigma^2_{x_2}}{n_2} + \frac{\sigma^2_{x_1}}{n_1} = \frac{9}{5} + \frac{4}{4} = \frac{14}{5} = 2{,}8$$

Además sabemos que para poder encontrar la probabilidad de ocurrencia de una variable que sigue una distribución normal es necesario transformarla en la variable $Z : N(0;1)$. De modo que para el caso presente:

$$Z = \frac{(\bar{x}_2 - \bar{x}_1) - \mu_{(\bar{x}_2-\bar{x}_1)}}{\sigma^2_{(\bar{x}_2-\bar{x}_1)}} = \frac{(\bar{x}_2 - \bar{x}_1) - (\mu_{x_2} - \mu_{x_1})}{\sqrt{\frac{\sigma^2_{x_2}}{n_2} + \frac{\sigma^2_{x_1}}{n_1}}} = \frac{8{,}2 - 10}{\sqrt{2{,}8}} = \frac{-1{,}8}{1{,}6733} = -\,1{,}08$$

Por lo tanto, la probabilidad deseada será:

$$P\left(\bar{X}_2 - \bar{X}_1 \le 8.2\right) = P\left(Z \le -1{,}08\right) = \ 0{,}1401$$

4.4.2.1 La diferencia de medias muestrales y el Teorema del Límite Central.

Cuando se desconoce la distribución de la variable, se pueden deducir las propiedades de la distribución de la diferencia de medias muestrales a partir del Teorema del Límite Central. Por lo tanto, si el muestreo se realiza a partir de poblaciones con distribución desconocida y el tamaño de las muestras es grande (n_1 y $n_2 \ge 30$), se aplica el teorema y la distribución de la diferencia de medias muestrales tendrá una media y una varianza igual a:

$$\mu_{(\bar{x}_2-\bar{x}_1)} = \mu_{(\bar{x}_2)} - \mu_{(\bar{x}_1)} = \ \mu_{(x_2)} - \mu_{(x_1)}$$

$$\sigma^2_{(\bar{x}_2-\bar{x}_1)} = \sigma^2_{\bar{x}_2} + \sigma^2_{\bar{x}_1} = \frac{\sigma^2_{x_2}}{n_2} + \frac{\sigma^2_{x_1}}{n_1}$$

Ejemplo 4.11

Dadas dos poblaciones A y B no distribuidas normalmente con los parámetros poblacionales siguientes: $\mu_A = 1400$; $\sigma^2_A = 40000$; $\mu_B = 1200$ y $\sigma^2_B = 10000$. Se extrae una muestra aleatoria n_A = 125 de la población A y una muestra n_B = 100 de la población B; determine la probabilidad de que la media muestral de A sea mayor a la media muestral de B en más de 160 unidades.

Aunque no sabemos como se distribuye la variable estudiada, de acuerdo al Teorema del Límite Central la diferencia de medias muestrales se distribuye normalmente con media y desviación igual a:

$$\mu_{(\bar{x}_A-\bar{x}_B)} = \ \mu_A - \ \mu_B = 1400 - 1200 = 200$$

$$\sigma_{(\bar{x}_A-\bar{x}_B)}=\sqrt{\frac{\sigma_A^2}{n_A}+\frac{\sigma_B^2}{n_B}}=\sqrt{\frac{40000}{125}+\frac{10000}{100}}=20,49$$

La probabilidad buscada será:

$$P\ (\bar{x}_A\text{-}\bar{x}_B \geq 160)=P\ (Z \geq z)=P\left(Z \geq \frac{(\bar{x}_A-\ \bar{x}_B)-(\mu_A-\ \mu_B)}{\sqrt{\left(\sigma_A^2/n_A\right)+\left(\sigma_B^2/n_B\right)}}\right)=P\left(Z \geq \frac{160-200}{20,49}\right)=$$

$$=P\ \left(Z \geq -1,95\right)=1-P\ (Z \leq -1,95)=1-0,0256=0,9774$$

4.5 EJERCICIOS

1. Se extrae aleatoriamente una muestra de tamaño 9 de una población de valores de una variable que se distribuye normalmente con media igual a 1200 y desviación igual a 400
   a. ¿Cuál es la media de la media muestral?
   b. ¿Cuál es la varianza de la media muestral?
   c. ¿Cuál es el error estándar?
   d. ¿Cuál es la probabilidad de que la media muestral tenga un valor menor a 1050?

2. El número de huevos puestos por una trucha es una variable aleatoria que se distribuye normalmente con una media igual a 115000 huevos y una desviación típica de 25000 huevos. Si se registra la cantidad de huevos producidos por 100 truchas responda lo siguiente:
   a. ¿Cuál es la probabilidad de que la media muestral de oviposición sea menor a 110000 huevos?
   b. ¿Cuál es la probabilidad de que la media muestral de oviposición esté entre 113000 huevos y 117000 huevos?
   c. ¿Cuál es la probabilidad de que la media muestral de oviposición esté entre 114000 y 116000 huevos?
   d. Sin hacer cálculos, razonar en cuál de los siguientes rangos es más probable que se encuentre la media muestral de oviposición:
      d1. Entre 113000 y 115000 huevos
      d2. Entre 114000 y 116000 huevos,
      d3. Entre 115000 y 117000 huevos,
      d4. Entre 116000 y 118000 huevos.
   e. Supongamos que, después de haber realizado los cálculos anteriores, alguien afirma que la distribución poblacional del número de huevos puestos es casi con toda seguridad no normal. ¿Se deben desechar los cálculos y plantearse nuevamente el problema?

3. En 16 ocasiones se midió el tiempo que tardan en reaccionar las sustancias A y B. Si se sabe que la distribución de dichos tiempos en la población sigue una distribución normal con media de 87 minutos y desviación típica de 22 minutos
   a. ¿Cuál es el error estándar de la media muestral de los tiempos de reacción?
   b. ¿Cuál es la probabilidad de que la media muestral sea menor a 100 minutos?
   c. ¿Cuál es la probabilidad de que la media muestral sea mayor a 80 minutos?
   d. ¿Cuál es la probabilidad de que la media muestral tome un valor que esté entre 85 y 95 minutos?
   e. Si se repite la reacción 15 veces en forma independiente diga, sin hacer los cálculos, si las probabilidades calculadas en los apartados (b), (c) y (d) serán mayores, menores o iguales para esta segunda muestra. Utilizar gráficos para ilustrar las respuestas.

4. Se sabe que la desviación típica de la cantidad de insectos consumidos diariamente por los individuos de una especie de anfibio es de 40 insectos. Si se examina el contenido estomacal de 100 de estos anfibios con el fin de estimar el consumo medio diario de insectos para el total de la población de anfibios.
   a. ¿Cuál será error estándar de la media muestral del consumo diario?
   b. ¿Cuál es la probabilidad de que la media muestral exceda a la media poblacional en más de 5 insectos?
   c. ¿Cuál es la probabilidad de que la media muestral esté más de 4 insectos por debajo de la media poblacional?
   d. ¿Cuál es la probabilidad de que la media muestral difiera de la media poblacional en más de tres insectos?

5. Una compañía fabrica células fotoeléctricas cuya duración promedio es de 800 horas con una desviación de 40 horas. Si la variable duración se distribuye normalmente, encuentre la probabilidad de que una muestra aleatoria de 16 fotocélulas tenga una vida promedio menor a 775 horas.

6. La concentración de fructuosa en muestras de semen de toro se distribuye normalmente con media igual a 80 mg/100 ml y varianza igual a 400 $(mg/100ml)^2$. Si se toman 36 alícuotas de 100 ml de semen ¿Cuál es la probabilidad de que la media muestral se encuentre entre 70 y 90 mg/100 ml?

7. Si en una comunidad la longitud craneal se distribuye normalmente con μ =185,6 mm y σ =12,7 mm, ¿Cuál es la probabilidad de que una muestra de tamaño 10 proporcione una media mayor a 190 mm?

8. Si la media y la desviación de la concentración de hierro en el suero de hombres sanos es de 120 y 15 mg/100ml respectivamente, ¿Cuál es la probabilidad de que una muestra aleatoria de 50 hombres sanos proporcione una media cuyo valor se encuentra entre 115 y 125 mg/100 ml?

9. En una población de mamíferos se sabe que el consumo de calorías por individuos es una variable que se distribuye normalmente con μ = 4000 cal/día y $\sigma^2 = (1200)^2$. Si se

selecciona una muestra constituida por 36 animales, calcule la probabilidad de obtener un valor promedio:

a. Inferior a 4200 cal/día.
b. Superior a 3500 cal/día.
c. Inferior a 3800 cal/día.
d. Inferior a 3000 cal/día
e. Inferior a 4000 cal/día
f. Que difiera de μ en 100 cal/día o más.

10. Usando los datos del Ejercicio 9 calcule la probabilidad de seleccionar una muestra formada por 100 individuos, cuyo consumo promedio:
    a. Sea de 3820 cal/día o más.
    b. Sea de 3820 cal/día o menos.
    c. Esté comprendido entre 3820 y 4000 cal/día.
    d. Esté comprendido entre 4000 y 4100 cal/día.
    e. Sea superior a μ en 100 cal/día o menos.
    f. Sea inferior a μ en 100 cal/día o más.
    g. Difiera de μ en 100 cal/día o menos.
    h. Difiera de μ en 100 cal/día o más.

11. Usando los datos del Ejercicio 9, encuentre el tamaño que debe tener una muestra para que con una probabilidad del 95,44% su media difiera de μ en 100 cal/día o menos.

12. En una región productora de caña de azúcar el 17,62% de las medias de todas las muestras de tamaño n = 61 tienen un valor superior a 10 toneladas de azúcar por hectárea. Si σ = 4 toneladas/hectárea: ¿Cuál es el rendimiento promedio en la región?

13. La producción promedio de una siembra de mangos es de 600 frutos/planta con una varianza de $(100)^2$. Si la variable en cuestión se distribuye normalmente:

    a. ¿De qué tamaño debe ser una muestra para que con un 99,74% de probabilidad su media difiere de μ en 50 frutos /planta o menos?

14. Usando los datos del ejercicio 4.5.13, calcule la probabilidad de seleccionar una muestra de 100 plantas cuyo promedio sea:
    a. Inferior a 627 frutos/planta.
    b. Superior a 586 frutos/planta.
    c. Entre 627 y 586 frutos/planta.
    d. Entre 610 y 627 frutos/planta.
    e. Entre 586 y 596 frutos/planta.
    f. Inferior a 590 frutos/planta.

15. Los niveles de vitamina A en dos poblaciones humanas se distribuye normalmente con el mismo valor promedio $(\mu_1 = \mu_2)$. ¿Cuál es la probabilidad de que una muestra de

tamaño 49 y otra segunda muestra de tamaño 36 proporcione un valor igual o mayor a 50, si las varianzas poblacionales son de 19600 y 8100 unidades respectivamente?

16. Se calculó la media de una muestra de 100 valores extraídos aleatoriamente de una población que está normalmente distribuida con media igual a 10 y varianza igual a 2. Igualmente se determinó la media de una segunda muestra aleatoria del mismo tamaño, seleccionada aleatoriamente e independientemente de otra población de la misma variable que también esta distribuida normalmente con media igual a 8 y la misma varianza anterior. Con esta información haga lo siguiente:

    a. Calcular la probabilidad de obtener un valor para la diferencia de los promedios muestrales que sea:
    - *a1.* Inferior a 2,4
    - *a2.* Superior a 1,90
    - *a3.* Se encuentre entre 0,70 y 1,90
    - *a4.* Se encuentre entre 0,60 y 2,4
    - *a5.* Se encuentre valor entre 0,70 y 2,00

    b. Calcule la probabilidad de obtener un valor de $\overline{X}_1$ mayor al valor de $\overline{X}_2$:
    - *b1.* En 2,6 o menos unidades.
    - *b2.* Igual a 1,6 o más unidades.
    - *b3.* En 2,3 o más unidades.
    - *b4.* En 1,5 o menos unidades.

    c. Calcule la probabilidad de obtener una diferencia de los promedios muestrales mayor a 1,6 unidades si se sabe que ésta diferencia es un valor que se encuentra entre 1,3 y 2,4 unidades.

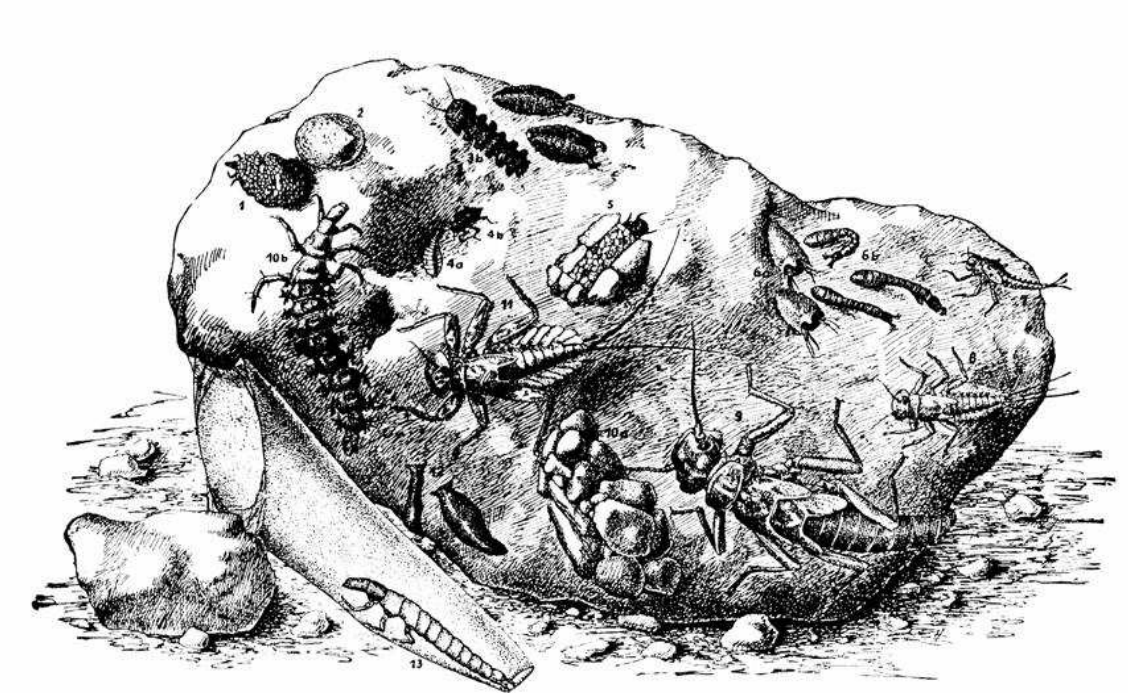
17. Se ha determinado que los individuos de una población de aves de cierta especie tiene una vida promedio de 6,5 años con una desviación de 0,9 años, mientras que los individuos de otra población de la misma especie de ave tienen una vida promedio de 6,0 años con una desviación de 0,8 años ¿Cuál es la probabilidad que la media de una muestra aleatoria de 36 individuos de la primera población no difiera en más de un año de la duración promedio de 49 individuos de la segunda población?

18. En cierto criadero el peso promedio de los adultos machos de los perros de la raza Pastor Alemán es de 35,6 kg con una desviación de 8,4 kg mientras que las hembras de la misma raza promedian 32,5 kg con una desviación de 5,5 kg. Si durante un año se vendieron 50 machos y 60 hembras ¿Cuál es la probabilidad que la diferencia entre los pesos promedios de los dos grupos se encuentre alejada por más de 1,5 kg de la media poblacional de las diferencias $(\mu_1 - \mu_2)$?

19. Se sabe que la grasa corporal total en los atletas se distribuye normalmente. En las nadadoras de alta competencia tiene un valor promedio de 12,2 kg con una desviación igual a 2,3 kg. En las corredoras de alta competencia el promedio de grasa total es de 7,8 kg con una desviación de 2,1 kg. Si el contenido promedio de grasa total se determinó en

10 nadadoras y 12 corredoras ¿Cuál es la probabilidad que la diferencia entre las medias muestrales se aleje de $\mu_1 - \mu_2$ en menos de 1,2 desviaciones estándar?

20. La concentración de cloruros en el suero sanguíneo de personas sanas es una variable que se distribuye con una media igual a 102 meq/l con una desviación de 10,2 meq/l. En las personas con problemas de tensión alta la concentración promedio de cloruro es de 121 meq/l con una desviación de 12,8 meq/l. Si de cada grupo se eligen aleatoriamente 30 y 36 individuos respectivamente ¿Cuál será la probabilidad que la concentración promedio de cloruros en la sangre de los hipertensos supere a la concentración promedio de las personas normales en más de 20 meq/l si ya se sabe que en éste último grupo el promedio de concentración de cloruros es menor al del primer grupo en un valor que se encuentra entre 17 y 24 meq/l?



Macroinvertebrados acuáticos