10

Teoría de muestras

En este capítulo se resume la Teoría de Muestras estadística, la cual trata el concepto de estudiar una población desconocida tomándole muestras, y a través del estudio de las mismas poder hacer inferencias acerca de toda la población. Primero se analiza el caso del muestreo aleatorio simple y estratificado, mostrando el manejo de una tabla de números aleatorios. Luego se ven los tipos no-aleatorios en el muestreo y se discute acerca de las ventajas y desventajas de cada uno de los métodos. Se explican los métodos usados en Bioquímica para lograr que las muestras extraídas a los pacientes cumplan los requisitos de aleatoriedad, aunque sea aproximadamente. Lo mismo para el caso de la industria Farmacéutica y para las simulaciones en los muestreos de mercadeo, usadas en el comercio en general. De manera tal de poder aplicar luego los modelos estadísticos que exigen tal requisito. En la Tabla 3, del fascículo de las tablas, se presenta la Tabla Aleatoria, más conocida como: "Random Numbers".

10.1 Introducción

Las razones para efectuar un muestreo a una población, en lugar de estudiarla directamente, pueden ser varias como se puntualiza a continuación:

. *El tamaño de la población es infinito.*
. *El muestreo es de tipo destructivo.*
. *La población es finita, pero demasiado grande.*
. *Sería muy caro estudiar a toda la población y basta con deducciones aproximadas.*
. *Tomaría demasiado tiempo analizar la población total.*

Cuando las muestras son lo suficientemente grandes, se pueden hacer inferencias analíticas bastante extensas, con pocos y simples recursos, en comparación con técnicas más refinadas de la Estadística. Esto es conveniente desde un punto de vista didáctico. La Teoría del muestreo es el estudio de las relaciones entre una población y las muestras que se extraen de ella. Del análisis de las muestras se pueden *estimar o inferir* datos de la población como su media (μ), varianza (σ^2), etc., llamados *parámetros poblacionales*, denotados usualmente con letras griegas, a partir de los valores obtenidos de la muestra, tales como la media muestral $\overline{x}$, la varianza muestral DS^2, etc. Por ejemplo, en Bioquímica el verdadero valor de glucosa de un paciente μ es siempre desconocido y además variable con el tiempo, por ello se le extrae una muestra de sangre para poder estimarlo a través de mediciones. Así, el valor $\overline{x}$ obtenido (ya sea con una o más mediciones) es el que se pone en el informe de la determinación clínica. El problema consiste en extraer muestras lo más representativas posibles, de la población desconocida, para que tengan sentido las estimaciones realizadas a través de ellas. Cuando en Farmacia las mediciones se hacen a través de encuestas; se miden los porcentajes de las respuestas obtenidas. Esto se emplea usualmente en investigaciones de mercado, técnicas de propaganda, estudios poblacionales, etc. Por ejemplo, en una encuesta sobre el uso de determinado producto cada respuesta favorable se puede considerar un éxito, y la proporción de éxitos p en el total de las encuestas realizadas se puede usar para estimar la verdadera proporción π en la población tomada como marco de referencia del estudio.

Cuando la población sea finita y de un tamaño manejable en tiempo y costo, los valores poblacionales se calculan directamente, sin necesidad del muestreo. Por ejemplo, si se trata de revisar diamantes, a nadie se le ocurriría tomar muestras sino que se controlarían uno por uno.

Ahora bien, cuando se efectúan mediciones de magnitudes clínicas de tipo cuantitativo, la idea teórica es que se pueden efectuar infinitas mediciones y, en tal caso, el tamaño de la población será infinito. Es el caso de mediciones repetidas de un mismo objeto como pesar un cuerpo, o medirle su longitud, etc., se puede efectuar todas las mediciones que se deseen. Esto no ocurre cuando el ensayo es destructivo para la muestra. Por ejemplo, en mediciones clínicas de calcio, colesterol, hierro, etc., porque al suero se le adicionan los reactivos químicos y sirve para una sola vez. O bien, cuando se controla la cantidad de los componentes activos en un remedio, no hay otra forma que destruirlo para poder revisarlo. Entonces, el número total de determinaciones posibles dependerá de la cantidad de material disponible, que es finito. En resumen, en estos casos se acostumbra a considerar al valor verdadero de la magnitud medida como desconocido. El único medio para estimarlo es la toma de muestras.

10.2 Muestras aleatorias

Las muestras extraídas deben ser representativas de la población en estudio para que las conclusiones obtenidas sean válidas. El estudio de los diferentes métodos de muestreo y las distintas variantes ocasionadas se llama *diseño de experimentos*. Para extraer una muestra representativa se usa el método del *muestreo al azar* o *aleatorio*, que *asigna a todo integrante de la población la misma probabilidad de ser elegido*. Cuando por lo menos uno de los elementos de la población tiene una probabilidad mayor (o menor) de ser seleccionado que el resto de los integrantes de la misma, entonces se trata de un método de *muestreo no aleatorio*.

Un sorteo es el único método seguro de hacer equiprobables a todos los integrantes de la población en el proceso de selección. Por ejemplo, se le asigna un número identificatorio a cada uno, luego se colocan estos números en una urna, se mezclan bien, para luego ir sacando los números de la misma hasta completar el tamaño de muestra deseado. O bien, se procede a sortearlos con un bolillero como en la Lotería Nacional. En el caso de poblaciones humanas se puede usar el número de documento de identidad, o el de la libreta universitaria en el caso de alumnos de una Facultad. Se debe insistir en que el sorteo sea al azar. Muchas veces, equivocadamente se cree que es lo mismo elegir "a dedo" a los integrantes de la muestra porque el investigador no tiene preferencias por alguno de ellos, y por lo tanto elegirá a cualquiera. No es así. Se trata de una creencia subjetiva del mismo, indemostrable para los demás. Por ejemplo, hay que sacar una rata de una jaula que contiene muchas de ellas; la mala costumbre es abrir la tapa y tomar una sin más. Por ser ratas, el investigador cree no tener preferencia por ninguna y, por lo tanto, todas tendrían la misma probabilidad de salir. Sin embargo, en realidad toma a la que se deja atrapar, ya sea por no ser tan veloz para escapar como sus compañeras, ya sea por estar enferma, o por ser muy gorda como para reaccionar ágilmente. La *única forma de asegurar el azar* en la selección es numerar las muestras de alguna manera y hacer un sorteo para elegirla. Es seguro que así nadie podrá ponerle objeciones a la mecánica de selección empleada.

Hay casos especiales donde no es tan simple numerar a los integrantes de la población para poder efectuar el sorteo, como sacar una muestra de granos de arroz de una bolsa de 50 Kg, o fideos de su contenedor, o yerba mate de una ponchada, etc. En tales casos, la solución es la homogeinización de la población previo a la extracción de la muestra. Por ejemplo, si se colocan los granos de arroz en una mezcladora y se agitan el tiempo suficiente, luego se toma la muestra en cualquier sector cuando esté bien homogeneizada. Lo mismo al tomar muestras de agua de río. Se supone que la correntada mezcla bien e impide la formación de grupos de coloides como los formados por el sulfato de aluminio cuando se trata de hacerla potable. La misma idea se aplica en el caso de extracción de sangre a pacientes para efectuar análisis clínicos. El torrente sanguíneo la homogeiniza tan bien que al puncionar en la vena se tiene una muestra representativa del paciente. Ese motivo hace desaconsejable la extracción en otra parte diferente de cuerpo del paciente, como pinchar un dedo o una oreja, excepto por fuerza mayor. Hay casos más difíciles o imposibles de homogeneizar como tomar una muestra representativa de tierra de un campo, o de arena en una playa. Aquí hay que subdividir la superficie en sectores, y efectuar un sorteo para seleccionar los lugares de extracción.

Para realizar un sorteo al azar hay diversos procedimientos. La manera clásica es colocar papeletas numeradas o con los nombres de los participantes en una bolsa, revolver y sacar una por una las mismas sin mirar. Una mejor manera es numerar a cada integrante de la población y usar un bolillero. Nadie, salvo la Lotería Oficial, puede manejar un bolillero con miles de números. Es más práctico usar un bolillero pequeño con diez bolillas numeradas de 0 a 9. Se procede a efectuar un sorteo para el número que va en la primera unidad, se repone la bolilla extraída y se hace un segundo sorteo para el número correspondiente a la decena, un tercero para la centena y así sucesivamente. De esta manera, no importa el tamaño de la población. Por ejemplo, si se tienen 87.000 integrantes se usan las diez bolillas para los tres últimos números; en cambio para los dos primeros se desechan los que no corresponden; si sale un 9 en el caso del primero, se realiza otra extracción para que salga alguno de los posibles. En Estadística se usa la Tabla de Números al Azar (Tabla 3) en lugar de bolilleros.

Dicha tabla consiste de una serie de números obtenidos con sorteos, o con un *software* de computadora generador de los *random numbers*. En la Tabla 3 se muestra una de estas tablas. El procedimiento es como sigue.

. Se parte de cualquier lugar de la tabla elegido a dedo. O lo que es mejor, se pone el dedo y se eligen dos números para identificar la fila de inicio, y los dos siguientes para identificar la columna de inicio. Como solo hay 45 filas y 50 columnas, si las dos cifras resultan mayores se sigue tomando de dos en dos hasta encontrar un valor adecuado.

. Con el número de fila y columna se localiza el primer número, desde donde se comienza.

. Se toman los números secuencialmente, empezando con el primero a la derecha del punto de partida y se lee la tabla de derecha a izquierda. Al terminar el renglón se sigue con el renglón de abajo de derecha a izquierda y así sucesivamente.

. También se puede usar otro sentido, como de izquierda a derecha, de arriba hacia abajo o viceversa. Pero *no se puede avanzar en diagonal*.

. Los números leídos se toman en grupos correspondientes al tamaño de la población. De a uno si hay menos de 10 integrantes, de a dos si hay menos de 100, de a tres si son menos de 1000, etc. Cuando la secuencia elegida no corresponde a ninguno de los integrantes, se desecha. Por ejemplo, si hay 85 integrantes y el número que sale es mayor, no se lo toma en cuenta y se sigue con el siguiente grupo de dos números.

La división de a cinco números en la tabla para formar columnas no tiene significado alguno, como el dejar un renglón en blanco cada cinco. Se ponen así para facilitar la lectura de la misma. Algunos autores objetan la elección a dedo del primer número. Su argumento es que al usar mucho la tabla de estas se memorizan ciertos grupos y el inconsciente puede jugar una mala pasada. Para evitar este inconveniente, recomiendan lo siguiente: a dedo se elige el lugar de inicio, pero como se explicó antes los dos primeros números se usan para elegir la columna, y los dos siguientes para elegir la fila, del número inicial como punto de partida. Así, si los cuatro primeros números son 2305, entonces el número de comienzo no es donde se colocó el dedo sino el dígito ubicado en la columna 23 y en la fila 5. En síntesis, con una simple hoja de papel se evita cualquier bolillero.

El proceso para fabricar estas tablas debe cumplir con las condiciones de equiprobabilidad e independencia, esto es:

- Todo número de la tabla tiene la misma probabilidad de salir que los demás. Esto es, en cualquier lugar de la tabla la cifra escogida tiene una probabilidad $p = 0,1$ de aparecer.

- La aparición de cierta cifra en la tabla no condiciona para nada la cifra siguiente ni es condicionada por la anterior.

Además de usarse para escoger *muestras aleatorias simples*, estas tablas tienen otros usos. Por ejemplo, cuando se va a testear a dos grupos de pacientes, uno para control y el otro que recibirá el tratamiento. Del total de pacientes posibles se usa la tabla para seleccionar a los integrantes de cada grupo, pero si hay que tomar en cuenta el sexo entonces se debe hacer un sorteo dentro de los pacientes varones, y otro dentro de las mujeres. Así, se garantiza igual probabilidad para cada *estrato*. Esto es, se divide primero a la población en estratos, como sexo, edad, raza, religión, nivel de educación, nivel de ingresos, etc., y luego se procede a seleccionar con un muestreo al azar dentro de cada estrato. A este tipo de procedimiento se lo denomina *muestreo aleatorio estratificado*. La teoría demuestra que la información obtenida en un muestreo estratificado es mucho más rica que en uno simple.

Casi siempre los pacientes entran a las pruebas clínicas, o a las farmacias, de un modo temporal, o sea, a medida que van llegando. Otra vez se puede usar la tabla para garantizar la aleatoriedad. Para el ejemplo anterior se puede hacer un sorteo para determinar cuál paciente irá a control y cuál recibirá el tratamiento, usando una regla de reparto. Por ejemplo: los que obtengan de 0 a 4 serán de control y de 5 a 9 para el tratamiento. Si de la tabla se extraen los números: 3, 8, 5, 0, 4 ... significa que al primer paciente que llegue le toca el número 3, o sea, irá para control el segundo un 8 y recibe el tratamiento; lo mismo el tercero al que le tocó un 5 los dos siguientes a control y así sucesivamente. Cuando se completa uno de los grupos, se sigue el orden del sorteo sólo con los del otro hasta terminar. Como se ve, siempre se puede encontrar la manera de aleatoreizar una muestra de una manera práctica. En el caso de arribo de clientes a una farmacia se puede seguir un procedimiento similar para, por ejemplo, otorgar un premio a modo de propaganda. En una industria, generalmente se seleccionan en forma aleatoria los productos finales o materias primas que irán al laboratorio de control de calidad.

10.2.1 Aplicaciones en Medicina

Los métodos más comunes se conocen con el nombre de *ciegos* o *doble-ciegos*. Esto se refiere a que hay casos donde el director del proyecto de investigación es la única persona que conoce la identidad de las muestras individuales. Por ejemplo, cuales son las muestras de control y cuales son las de los casos a investigar, o bien cuando en un control de calidad se introducen las muestras de control sin que los operarios del sistema lo sepan. De esta forma, el personal que trabaja efectuando mediciones no sabe (*está ciego*) con cuales muestras su jefe los está controlando. Así se puede verificar el comportamiento del personal de un laboratorio, sin que estos lo sepan. Cuando se hace un trabajo de comprobación de la efectividad de un medicamento, los enfermos son separados en dos grupos, uno al cual se le administra el medicamento (casos) y el otro al cual se le suministra un placebo (control). El paciente nunca sabe si se trata de un caso o de un control. Si el director del proyecto decide hacer un trabajo del tipo *ciego*, entonces evita que su personal sepa que pacientes son los controles y quienes los casos. De esta forma se asegura que ambos grupos sean tratados por igual: mismas condiciones ambientales y alimenticias, misma atención personalizada, etc. Se asegura que las dosis y los placebos tengan las mismas características físicas, para evitar que alguien pueda distinguirlos, y mediante un muestreo aleatorio designa quienes van a formar parte de cada grupo. Por sorteo, también asigna al personal y a los pacientes que van a conformar su grupo de trabajo.

En ciertas ocasiones, tampoco el director conoce quienes son los controles y quienes los casos. Es decir, una tercera persona ajena al proyecto de investigación, es el único que los puede identificar, y solo al terminar el trabajo procede a hacerlo. Generalmente, los nombres de los pacientes y su condición se guardan en un sobre cerrado, para ser abierto al final. A este caso, se lo conoce como un *doble-ciego*. Es otra forma de evitar influencias externas que puedan alterar los resultados como consecuencia de la tendencia humana a tener intereses en uno u otro sentido.

Cuando se trata de evaluar la calidad de un test o un diagnóstico clínico, el mejor procedimiento para evitar influencias ajenas al proyecto es la siguiente:

1) El investigador decide el tamaño muestral total N que va a usar en su experimento, de entre todos los casos disponibles que tenga.
2) Por casos disponibles se entiende aquellos sujetos que pueden ser clasificados con certeza en sanos o enfermos (los casos dudosos se descartan). Así escoge con un muestreo aleatorio a los casos que va a usar con una proporción del 50%. O sea, TE = TS = N/2.

A este método se lo conoce como muestreo basado en la enfermedad. De esta forma los cuatro casos mutuamente excluyentes de la tabla de verdad son naturales, así como los totales de casos diagnosticados como positivos y negativos. Si en el estudio, se usan estos valores, entonces se puede tener la tranquilidad de que los resultados no tengan ningún tipo de influencia. En cambio, si se usan los índices que contengan algunos los valores elegidos por el investigador, como Sensibilidad, Especificidad, etc., ya no se puede estar tan seguro. Notar que si el investigador se decide por un muestreo basado en los diagnósticos, la prevalencia de la enfermedad en la población de referencia podría afectar los datos. Es decir, si el investigador decide adoptar N con un 50% de positivos y el otro de negativos (TP = TN = N/2), el total de enfermos que obtenga no sería el mismo en un hospital general, que en un servicio especializado adonde se derivan los pacientes con esa enfermedad. Por ejemplo, la cantidad de enfermos que concurren a un centro médico con un diagnóstico presunto de infarto de miocardio, normalmente es mucho menor porcentualmente, que la cantidad que llega a una unidad coronaria.

En Epidemiología se usan ambas maneras de efectuar los muestreos. En estos casos se trata de armar una tabla de 2 x 2 para analizar la relación entre un factor de riesgo (o uno de protección como la inmunización de una vacuna) con la enfermedad estudiada. Un *muestreo basado en la exposición* es el que se usa cuando el investigador decide adoptar el número de sujetos expuestos al factor de riesgo y el de no-expuestos. Aquí las costumbres no se basan en un 50% para cada caso sino en criterios médicos derivados del conocimiento que se tiene de la situación, o en casos similares ya estudiados y publicados, o sencillamente en el número de casos disponibles cuando la enfermedad es rara como una meningitis, etc. En cambio, *un muestreo basado en la enfermedad* consiste en que el investigador elige el número total de enfermos y de sanos que va a emplear en su experimento, y generalmente la relación no es del 50% sino en alguna otra derivada de experiencias anteriores. Por ejemplo, en el caso de la meningitis no es raro que la proporción usada sea un caso de enfermedad y cuatro casos como control. En este punto, conviene hacer una aclaración importante:

1) El primer paso consiste en seleccionar todos aquellos sujetos que se adapten a los requerimientos específicos de la investigación y descartar a los demás.
2) Una vez que se tiene el número de total de sujetos disponibles, el investigador debe decidir el tamaño muestral N generalmente menor que el total.
3) De acuerdo al tipo de muestreo que vaya a efectuar (por enfermedad o por exposición) el siguiente paso es adoptar el tamaño muestral de cada total marginal de la tabla.
4) Recién entonces, empleando un muestreo aleatorio selecciona a los sujetos.

Esta recomendación se hace, porque a veces se hace la selección al azar primero, y luego se eligen quienes formaran parte de la muestra y quienes no de acuerdo a ciertos criterios médicos. Pero esto invalida la representatividad de la muestra en la población, porque la eliminación de ciertos sujetos debe hacerse antes del sorteo y no después.

10.3 Muestras no aleatorias

Cuando el método de extracción de las muestras no asegure a cada individuo de la población o del estrato, igual probabilidad de ser elegido, entonces la muestra obtenida no es aleatoria. A veces, esto se hace por razones de practicidad en el sentido del costo o del tiempo. Si se desea tomar una muestra probabilística de la población argentina no parece razonable usar a cada individuo como unidad de muestreo. Lo mismo cuando se desea hacer un muestreo a los escolares de una provincia, es muy difícil empadronar a todos primero para luego sortear, y se tardaría demasiado para ubicarlos uno por uno hasta terminar el trabajo.

En el *muestreo de etapas múltiples* se utiliza para el caso de grandes poblaciones humanas. Acá, la unidad de muestreo en la primera etapa son los departamentos de cada provincia. Se los lista y se hace un primer sorteo para la selección. En una segunda etapa se distingue la población rural de la urbana, subdividiendo en fracciones (diferentes superficies con densidad de población semejante). Otra vez se sortea para elegir, y se continúa con otra división en radios dentro de las fracciones, segmentos dentro de radios, y así sucesivamente. La razón es repartir equitativamente el trabajo del encuestador.

En el *muestreo por conglomerados* se eligen conjuntos donde naturalmente se agrupan los individuos. Es, por ejemplo, el caso de las escuelas para hacer un muestreo alumnos en el sistema educativo, o las facultades para los universitarios. Si se trata de estudiar las condiciones laborales de los empleados de comercio que trabajan en supermercados, primero se empadronan a los lugares naturales de trabajo (supermercados), y luego se sortea entre estos conglomerados para elegir a uno. Luego se entrevista a todos los empleados del supermercado elegido, y se acepta esto como una muestra representativa del sector.

El *muestreo sistemático* se usa para el caso de sucesiones de elementos. Por ejemplo, el caso de las historias clínicas de pacientes, certificados de nacimiento, tarjetas de catálogo en una biblioteca, etc. Son los casos donde la información está en archivos y hay que trabajar con estos para obtenerlas. Se elige una cifra entera, razonable, tomando en cuenta el tamaño de la muestra y el de la población. Por ejemplo, hay que tomar una muestra de tamaño 25 de un archivo que contiene 488 fichas; luego, el cociente entre población y muestra es 488 /25, aproximadamente 19. Notar que si se elige 20 el tamaño muestral no llega a 25. Entonces, se cuentan las fichas y a llegar a la décimo novena se la extrae, se sigue hasta la número 38 que será la segunda escogida, y así sucesivamente hasta tener las 25 fichas necesarias. Es también el caso de los soldados que se numeran de 1 en adelante y cada 5 (u otro número cualquiera) dan un paso al frente. Es un método sencillo y rápido de selección.

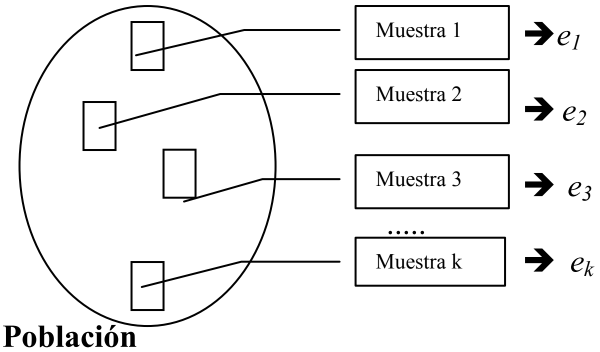
Hay otros casos de muestreo no aleatorios de uso común. Como el de tomar una lista de nombres, cerrar los ojos y con la punta de un lápiz marcar a uno de ellos, para escogerlo. Son los casos de los programas de TV donde se toma la guía telefónica, se la abre en una página cualquiera y se escoge a uno de los números que figuran, para luego hacer un llamado con premio. En los juegos infantiles se hace una ronda con los participantes, se vendan los ojos del que va a elegir, se le hace dar varias vueltas con los ojos vendados para que marque a alguno. Todos los casos vistos tienen una característica común: *los individuos de la población no son equiprobables*.

10.4 Distribuciones de probabilidad en el muestreo

También llamadas *distribuciones muestrales* de un estadígrafo cualquiera obtenido a través de la muestra. La idea es la siguiente: si se toman k muestras, *todas las posibles* de tamaño n (con o sin reemplazamiento) de una población de tamaño N_P, y a cada muestra se le calcula un estadígrafo e (media, mediana, varianza, etc.), se obtienen una serie de k valores: $e_1, e_2, e_3, \dots, e_k$

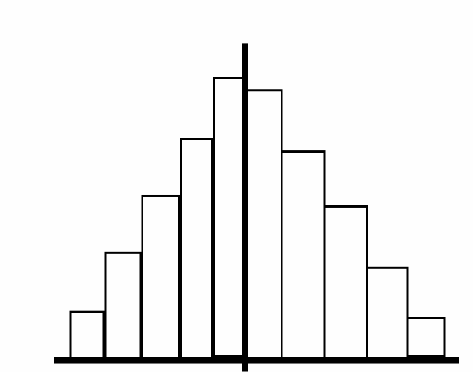
Estos valores pueden agruparse mediante un histograma de frecuencias para poder apreciar la forma de la distribución de los mismos. En la Figura 10.1 se esquematiza esta situación:

Figura 10.1 Distribuciones muestrales.



Población

Histograma con los valores ei



De una población cualquiera se extraen k muestras; cada una permite calcular k estadígrafos con los cuales se puede hacer un histograma como el de la derecha de la Figura 10.1. Se aprecia que este histograma adquiere forma de campana si se suavizan los escalones, al achicar los intervalos. Esta curva obtenida a partir de datos muestrales, observados a través del muestreo, tiende asintóticamente a otra curva teórica a medida que k aumenta, y los intervalos se hacen infinitesimales. Dicha curva teórica es la función de Gauss de acuerdo con el Teorema Central del Límite, el principal de la Estadística.

El Teorema Central del Límite permite establecer que, en condiciones muy generales, si la muestra es lo *suficientemente grande*, la distribución teórica de los k valores obtenidos es aproximadamente la función de Gauss. Esta es la base de la *Teoría de las Grandes Muestras*. Las principales *Distribuciones Muestrales* son funciones de Gauss identificadas en forma unívoca con sus dos parámetros μ y SE. En la Tabla 10.1 se presentan estos dos valores para cada uno de los estadígrafos más usuales. En la primer columna de la tabla se muestra cada estadígrafo, en la segunda columna se da la fórmula para el cálculo del error típico de estimación SE. Finalmente en la tercera columna se muestra la estimación puntual para obtener el valor esperado del estadígrafo μe, con las aclaraciones respecto al tamaño muestral requerido para que tal estimación sea considerada aceptable.

10.4.1 Distribución muestral de la media

Si el estadígrafo elegido es la media, se tendrán $\overline{X}_1$, $\overline{X}_2$, $\overline{X}_3$,..., $\overline{X}_k$ medias muestrales; estas se distribuyen normalmente si k es muy grande. En la práctica, 30 o más valores son suficientes. En la teoría, cuando $k \to \infty$ entonces la distribución muestral de la media es *asintóticamente normal* y coincidirá con la función de Gauss. Esta distribución tendrá un valor esperado y una varianza que permitirán estimar los respectivos valores poblacionales. O sea,

$$\mu_{\overline{x}} = \mu \qquad \sigma^2_{\overline{x}} = \sigma^2 / n = SE^2(\overline{x}) = VAR(\overline{x})$$

Tabla 10.1. Errores típicos para algunas distribuciones muestrales.

Distribución muestral	Error típico	Notas especiales
Medias	$SE(\bar{x}) = \sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$	Se cumple para muestras grandes o pequeñas. La distribución muestral de medias se ajusta mucho a una normal para $n \geq 30$ incluso para poblaciones no normales. $\mu_{\bar{x}} = \mu$ en todos los casos.
Proporciones	$SE(\pi) = \sigma_{\pi} = \sqrt{\dfrac{\pi(1-\pi)}{n}}$	Las notas anteriores para medias son igualmente aplicables aquí con $n > 25$. $\mu_P = \pi$ en todos los casos.
Desviaciones típicas	(*1*) $SE(DS) = \sigma_{DS} = \dfrac{\sigma}{\sqrt{2n}}$ (*2*) $SE(DS) = \sigma_{DS} = \sqrt{\dfrac{\mu_4 - \mu_2^2}{4n\mu_2}}$	Para $n \geq 100$, la distribución muestral de *s* es muy próxima a una normal. σ_s esta dada por (*1*) solamente cuando la población es normal (o aproximadamente normal). Si la población no es normal, puede utilizarse (*2*). Nótese que (*2*) pasa a ser (*1*) cuando $\mu_2 = \sigma^2$ y $\mu_4 = 3\sigma^4$, lo que se cumple para poblaciones normales. Para $n \geq 100$, $\mu_S = \sigma$ con gran aproximación.
Medianas	$SE(Med.) = \sigma_{med} = \dfrac{1{,}2533\ \sigma}{\sqrt{n}}$	Para $n \geq$ *30, la* distribución muestral de la mediana es muy próxima a una normal. Los resultados dados son válidos solamente si la población es normal (o aproximadamente normal) $\mu_{med} = \mu$
Cuartiles primero y tercero	$SE(Q2) = \sigma_{Q2} = \sigma_{Q3} = \dfrac{1{,}3626\ \sigma}{\sqrt{n}}$	Las notas anteriores para medianas son igualmente aplicables aquí. μ_{Q1} y μ_{Q3} son casi iguales al primero y tercer cuartil de la población. Nótese que $\sigma_{Q2} = \sigma_{med}$
Varianzas	(*1*) $SE(Var) = \sigma_{Var} = \sigma^2\sqrt{\dfrac{2}{N}}$ (*2*) $SE(Var) = \sigma_{Var} = \sqrt{\dfrac{\mu_4 - \mu_2^2}{n}}$	Las notas para desviaciones típicas son igualmente aplicables aquí. Nótese que (2) pasa a ser (1) en caso de que la población sea normal. $\mu_s^2 = \sigma^2 (n-1) / n$ que es casi igual a σ^2 para valores grandes de n.
Coeficiente de variación	$SE(CV) = \sigma_{CV} = \dfrac{CV}{\sqrt{2n}}\sqrt{1 + 2(CV)^2}$	Aquí $CV = \sigma / \mu$ es el coeficiente de variación de la población. Los resultados dados son válidos para poblaciones normales (o aproximadamente normales) y $n \geq 100$.

Esto es: la media aritmética de las k medias muestrales obtenidas es aproximadamente igual a la media poblacional (o valor verdadero). Sin embargo, esta aproximación tiene un error de estimación denominado *error típico* o *error estándar* de estimación que en el caso de la media es: $\sigma_{\bar{x}}$. En la bibliografía clínica la nomenclatura más empleada es $SE(\bar{x})$. En la Tabla 10.1 se muestran los valores anteriores para el caso de la media aritmética.

Las relaciones anteriores son válidas solo si la población es infinita, o si es finita, pero el muestreo es con reemplazamiento. Caso contrario, cuando la población es finita y se realizan muestreos sin reposición, entonces se deben ajustar dichas relaciones con:

$$\mu_{\bar{x}} = \mu \qquad \sigma^2_{\bar{x}} = (\sigma^2 / n)\,[(N_P - n) / (N_P - 1)] = SE^2(\bar{x}) = VAR(\bar{x})$$

En el Cuadro 10.1 siguiente se presenta un problema de aplicación para un caso donde se conoce a toda la población, los parámetros poblacionales se calculan directamente aplicando las fórmulas vistas en el Tema 4 resultando: $\mu = 4,5$ y $\sigma^2 = 1,25$. Se pueden verificar las relaciones anteriores de dos maneras. En la primera se toman las seis muestras posibles de tamaño 2, para un muestreo sin reemplazamiento. A cada muestra se le calcula su media respectiva, luego con estos 6 promedios se pueden calcular: el promedio y la varianza de esas seis muestras. Ahora, el promedio de todas las medias muestrales da exactamente igual al valor medio poblacional y la varianza de las medias muestrales verifica la relación anterior, si se aplica el factor de corrección para muestras de tamaño finito. La segunda manera (*Bootstrap procedure*) es tomando muestras con reposición, primero se toman todas las 16 muestras posibles con reemplazamiento de tamaño 2. Luego, con las dieciséis muestras se calculan las 16 medias respectivas. Por último, se calcula el promedio y la varianza de estos 16 valores, verificando de nuevo las relaciones vistas más arriba, para el caso de muestras con reposición. Para el otro problema, se suponen conocidos los valores poblacionales, y tomando 50 muestras de tamaño 3 hay que determinar la cantidad de casos donde el resultado esté comprendido en un intervalo (6 ; 7,796). La forma de proceder es calculando primero las probabilidades de obtener esos resultados límites, luego por diferencia calcular la probabilidad gaussiana asociada al intervalo y entonces, multiplicando dicha probabilidad por el tamaño muestral, se puede contestar la pregunta efectuada.

CUADRO 10.1: Distribución Muestral de Medias.

Caso 1) Un jefe de laboratorio quiere determinar cuántos Calcios realiza su personal en una hora de trabajo. Tiene a cuatro bioquímicos trabajando y los resultados obtenidos son:

Bioquímico	*Cantidad*	
A	4	Si considera a estos datos como su población
B	6	esta se representa con el conjunto (3 , 4 , 5 , 6)
C	5	Donde los parámetros poblacionales serán
D	3	$\mu = 4,5$ y $\sigma^2 = 1,25$ ($\sigma = 1,118$)

Si se toman todas la muestras posibles, sin reemplazamiento, de tamaño 2, se tendrán 6 pares en total, que son las combinaciones de 4 elementos tomados de a 2 (C = 4! / 2! . 2!) O sea,

Combinaciones	*Resultados*	*Media muestral*
(A , B)	(4 , 6)	5
(A , C)	(4 , 5)	4,5
(A , D)	(4 , 3)	3,5
(B , C)	(6 , 5)	5,5
(B , D)	(6 , 3)	4,5
(C , D)	(5 , 3)	4

El promedio de estas medias es

$$\frac{5 + 4,5 + 3,5 + 5,5 + 4,5 + 4}{6} = 4,5$$

valor igual al poblacional

Esto es $\mu_{\bar{x}} = 4,5 = \mu$

Y la varianza de estos valores es $\sigma^2_{\bar{x}} = 0,417 = (1,25 / 2) [(4-2) / (4-1)]$ verificando la relación vista más arriba. Si se toman todas las muestras con reemplazamiento, hay 16 casos:

(AA) (AB) (AC) (AD) (BA) (BB) (BC) (BD) (CA) (CB) (CC) (CD) (DA) (DB) (DC) (DD)

O sea, las muestras son:

(4;4) (4;6) (4;5) (4;3) (6;4) (6;6) (6;5) (6;3) (5;4) (5;6) (5;5) (5;3) (3;4) (3;6) (3;5) (3;3)

Y las medias de las muestras respectivas son:

4 ; 5 ; 4,5 ; 3,5 ; 5 ; 6 ; 6,5 ; 4,5 ; 4,5 ; 5,5 ; 5 ; 4 ; 3,5 ; 4,5 ; 4 ; 3

Luego el promedio de este conjunto de 16 medias es $\mu_{\bar{x}} = 4,5 = \mu$. Y la varianza de estos valores es $\sigma^2_{\bar{x}} = 0,625 = 1,25 / 2 = \sigma^2 / n$. Las relaciones se cumplen para ambos casos de muestreo, tanto si es con o sin reposición.
El método del *jacknife* consiste en sacar secuencialmente cada dato, calcular la media del grupo y luego reinsertarlo a los restantes para formar cada muestra. Para la población (4, 6, 5, 3) sería:

Muestras	*Medias*	*Promedio general*
(6, 5, 3)	4,6666	
(4, 5, 3)	4	4,5
(4, 6, 3)	4,3333	
(4, 6, 5)	5	

Notar que se puede "mirar" la estabilidad de la media que acá varía entre 4 y 5.

Ejemplo: Suponiendo que las proteínas totales se distribuyen normalmente entre los 2.500 varones sanos de una cierta población, con edades de 14 a 50 años, con un valor promedio de 7 g/dl y un desvío de 0,9 g/dl. Un investigador decide tomar 50 muestras aleatorias de tamaño 3, con reemplazamiento. Desea saber en cuántas muestras cabe esperar una media entre 6 y 7,796 g/dl.

El primer paso para resolver este problema es encontrar los parámetros de la distribución muestral de medias, usando los valores poblacionales conocidos. Entonces:

$$\mu_{\bar{x}} \approx 7 \text{ g/dl} = \mu \quad ; \quad \sigma^2_{\bar{x}} \cong \sigma^2 / n = (0,9)^2 / 3 = 0,27 \quad ; \quad \sigma_{\bar{x}} = SE(\bar{x}) = 0,52$$

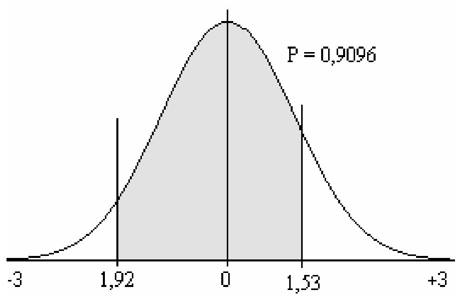
Cabe esperar que la distribución muestral de medias sea una curva de Gauss con tales parámetros. Para calcular la probabilidad de que una muestra tenga una media entre 6 y 7,796 g/dl, se tipifican estos valores con:

$$z_6 = (x - \mu_{\bar{x}}) / \sigma_{\bar{x}} = (6 - 7) / 0,52 = -1,92$$

$$z_{7,6} = (x - \mu_{\bar{x}}) / \sigma_{\bar{x}} = (7,796 - 7) / 0,52 = 1,53$$

La probabilidad bajo la curva de Gauss tipificada dentro del intervalo - 1,92 ; + 1,53 se obtiene calculando las probabilidades:

$$P(-1,92 ; 0) = 0,4726$$
$$P(0 ; 1,53) = \underline{0,4370}$$
$$P(-1,92 ; 1,53) = 0,9096$$



Entonces, el número esperado de muestras buscado será la proporción anterior, multiplicada por el número total de muestras extraídas. Esto es, (0,9096) 50 = 45,48 muestras. Es de esperar encontrar 45 muestras con valores de proteínas totales entre 6 y 7,6 g/dl.

10.4.2 Distribución muestral de proporciones

Cuando el estadígrafo obtenido de cada muestra de una población sea la proporción de ocurrencia de un suceso dado, un "éxito", se trata de una *distribución muestral de proporciones*. Por ejemplo, el caso del lanzamiento de una moneda es uno de población infinita porque se puede realizar la prueba cuantas veces se desee. La probabilidad de sacar un éxito, una cara es p = 0,5, la misma para todos los casos posibles, mientras que la probabilidad de "fracaso" al lanzar la moneda y salir seca es q = 1 - p = 0,5. Si el experimento consiste en realizar **n** lanzamientos de la moneda, una cantidad lo suficientemente grande, luego la frecuencia relativa del suceso cara, tendrá una distribución aproximadamente normal de parámetros:

$$\mu_p = \pi \qquad \text{y} \qquad \sigma^2 \pi = [\pi .(1-\pi)] / n = SE^2(\pi)$$

O bien $\mu_r = r = n . \pi$ y $\sigma^2{}_r = n . [\pi .(1-\pi)] = SE^2 (r)$

CUADRO 10.2: Distribución muestral de proporciones.

Ejemplo) En un laboratorio farmacéutico, históricamente se ha encontrado que el 3% de las ampolletas fabricadas de cierto medicamento salen falladas. Se desea saber cuánto vale la probabilidad de encontrar r = 20 defectuosas o menos, en una partida de n = 500 producidas.

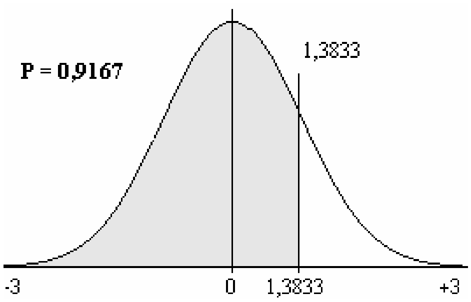
(a) La probabilidad exacta se puede calcular con la fórmula de la Binomial acumulada:

$$P_{Bi} (r \leq 20) = P_{Bi} (r =0) + P_{Bi} (r = 1) + \ldots + P_{Bi} (r = 20) = 0{,}9202$$

(b) El segundo método es usando la aproximación normal a la binomial con la relaciones siguientes: $\mu_r = n . \pi = 15$; $\sigma^2{}_r = n [\pi .(1-\pi)] = (500 .0{,}03. 0{,}97) = 14{,}55$; SE(r) = σ_r = 3,8144

Como se trata de una variable discreta y la función de Gauss es continua, hay que efectuar una corrección por continuidad de Yates o de Molenaar:

$z_{20} = [(4 . 20 +3)(1-0{,}03)]^{1/2} – [(4 . 500 – 4 . 20 - 1)(0{,}03)]^{1/2} = 1{,}3833$ O sea P = 0, 9167



La probabilidad pedida es el área a la izquierda de Z_{20} = 1,3883 en la curva Gauss tipificada.

Con la corrección de Yates sería: z = [(20 – 15) –1/2] / 3,8144 = 1,1797 y P = 0,881 O sea una aproximación peor al valor exacto dado en (a). Notar que si no se hubiese hecho la corrección por continuidad, la estimación de la probabilidad pedida sería peor aún: P = 0,8168.

Las que se obtienen de la relación vista más arriba con reemplazamiento de $\pi = r / n$. Por muestra lo suficientemente grande, se entiende en la práctica $n > 25$. Por otra parte, cuando la población es finita y el muestreo se realiza sin reposición, entonces las ecuaciones anteriores se modifican con:

$$\mu_p = \pi \quad \text{y} \quad \sigma^2 \pi = [\pi .(1- \pi) / n] [(N_P - n) / (N_P - 1)] = SE^2 (\pi)$$

O bien $\mu_r = r = n . \pi$ y $\sigma^2_r = n [\pi .(1- \pi)] [(N_P - n) / (N_P - 1)] = SE^2 (r)$

Usando la distribución de Gauss, se pueden calcular las probabilidades asociadas en ejemplos como los presentados en el Cuadro 10.2.

Tanto la distribución de medias como la de proporciones se usan muy frecuentemente en la práctica. El problema básico es conocer los valores poblacionales porque es muy raro saberlos de antemano, salvo el caso de la moneda y otros juegos de azar; lo común es ignorar estos parámetros. Entonces, usando los valores obtenidos del muestreo aleatorio se pueden aproximar los valores desconocidos poblacionales, como se verá en el tema *inferencia estadística*.

10.4.3 Diferencias de medias y de proporciones

Muchas veces es necesario comparar dos poblaciones, como cuando se testea un tratamiento en pacientes contra un control o blanco; o bien, cuando se comparan dos sistemas de medición entre sí (estos sistemas pueden ser: técnicas de laboratorio, marcas comerciales, instrumentos de diferente marca, etc.). El problema es ver si hay diferencias entre ambos casos. El método es encontrar un estadístico para la diferencia (o suma) de ambos estadísticos muestrales. Se tienen dos poblaciones, se efectúa un muestreo aleatorio de tamaño n_1 en la primera de ellas, y se obtienen los valores de un estadígrafo cualquiera: e_1. La población gaussiana de la distribución muestral tendrá los parámetros μ_{e1} y σ_{e1}. Análogamente, se procede con la segunda población y se obtienen los parámetros μ_{e2} y σ_{e2}. Se pueden combinar ambas poblaciones a través de un estadígrafo que sea la diferencia de ambos: $e = e_1 - e_2$. Este nuevo índice también tendrá una distribución gaussiana, pues la diferencia de dos funciones de Gauss es otra función gaussiana, siempre y cuando las muestras no dependan unas de otras; esto es, sean *independientes*. Sus parámetros son:

$$\mu_{e1-e2} = \mu_{e1} - \mu_{e2} \quad \text{y} \quad \sigma^2_{e1-e2} = \sigma^2_{e1} + \sigma^2_{e2} = SE^2 (e_1 - e_2)$$

Si el estadígrafo es la media muestral $e = \overline{x}$ entonces, la *distribución muestral de la diferencia de medias* para poblaciones infinitas, de parámetros ($\mu_1 ; \sigma_1$) y ($\mu_2 ; \sigma_2$). Entonces:

$\mu_{\overline{x}_1 - \overline{x}_2} = \mu_1 - \mu_2$ es el valor esperado de la diferencia de medias y su error de estimación es:

$$SE^2 (\overline{x}_1 - \overline{x}_2) = \sigma^2_{\overline{x}_1 - \overline{x}_2} = (\sigma^2_{\overline{x}_1} + \sigma^2_{\overline{x}_2}) = [(\sigma^2_1 / n_1) + (\sigma^2_2 / n_2)]$$

Si el estadígrafo es una proporción $e = \pi$ entonces, la *distribución muestral de la diferencia de proporciones* para poblaciones infinitas, de parámetros ($\mu_1 ; \sigma_1$) y ($\mu_2 ; \sigma_2$). Luego:
$\mu_{P1 - P2} = \mu_1 - \mu_2 = \pi_1 - \pi_2$ es el valor esperado de la diferencia de proporciones poblacionales.

$SE^2 (\pi_1 - \pi_2) = \sigma^2_{(\pi 1 - \pi 2)} = [\pi_1 (1-\pi_1) / n_1] + [\pi_2 (1-\pi_2) / n_2]$

Es el error estándar al cuadrado de tal estimación para la diferencia de proporciones.

En el Cuadro 10.3 se presentan dos ejemplos de aplicación de lo visto. En el primer caso se trata de una diferencia de medias usando la vida útil de dos medicamentos, mientras que en el segundo caso se trata de una diferencia de Sensibilidades de dos métodos clínicos (tomando como proporciones a las Sensibilidades) para poder compararlos. En ambos casos las probabilidades encontradas son muy chicas lo que indica diferencias entre las medias muestrales. Para poder decidir si tales diferencias son válidas se emplea la Teoría de la decisión estadística (Tema 12).

CUADRO 10.3: **Diferencia de 2 muestras.**

Caso 1) La vida útil de un medicamento fabricado por el Laboratorio A es de 1.400 días, con una desviación estándar de 200 días. Por su parte, el mismo medicamento fabricado por el Laboratorio B de la competencia tiene una duración de 1.200 días con un desvío de 100 días. Se eligen 125 medicamentos de cada Laboratorio con un muestreo al azar. Calcular la probabilidad que los del Laboratorio A duren 250 días más que los del B.

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

$\sigma^2_{\bar{x}_A - \bar{x}_B} = (\sigma^2_{\bar{x}_A} + \sigma^2_{\bar{x}_B}) = [(\sigma^2_A / n_A) + (\sigma^2_B / n_B)]$ Como $n_A = n_B = 125$

$\sigma^2_{\bar{x}_A - \bar{x}_B} = [(200)^2 + (100)^2] / 125 = 400$ O sea: $SE(\bar{x}_A - \bar{x}_B) = \sigma_{\bar{x}_A - \bar{x}_B} = 20$ días

La tipificación de esta diferencia se hace con: $Z_{A-B} = (250 - 200) / 20 = 2,5$

Entonces, la probabilidad pedida es igual al área a la derecha de este valor en la curva de Gauss. $P = (0,5 - 0,4938) = 0,0062$. Lo que significa que hay un 0,6% de probabilidad porcentual de que la vida útil del medicamento fabricado por A dure 250 días más que el de su competencia.

Caso 2) La Sensibilidad de una prueba clínica es del 85% si se usa la marca A, y del 80% si es la marca B. Se toman 100 muestras al azar para cada caso, y de desea saber la probabilidad que la diferencia de sensibilidades entre ambas sea del 10 % o más. Se estima $\pi_1 \approx 0,85$ y $\pi_2 = 0,8$

(a) $\mu_{A-B} = \mu_A - \mu_B = 0,85 - 0,80 = 0,05$ y $SE(_{A-B}) = \sigma_{A-B} = 0,02681$ pues:

$\sigma^2_{A-B} = [\pi_1 (1-\pi_1) / n_1] + [\pi_2 (1-\pi_2) / n_2] = (1/100) [(0,85 . 0,15) + (0,8 . 0,2)] = 0,002875$

Si se hace la Ho de que no hay diferencia entre ambas la variable tipificada es:

$Z_{A-B} = [(0,1 - 0) - (0,85 - 0,80)] / 0,05362 = 0,09325$ O sea $P(x < 10) = 0,5371$ y la probabilidad pedida se calcula con $P (X \geq 10) = (1 - 0,5371) = 0,4629$

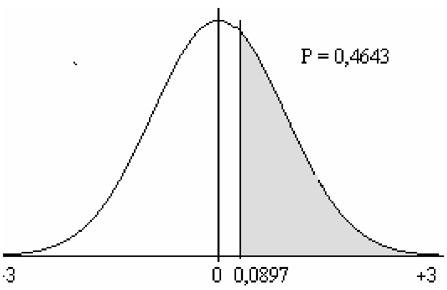
(b) Otra forma de resolver este problema es hacer el supuesto de que las muestras provienen de dos poblaciones diferentes, la primera con una probabilidad de 0,85 y la otra con 0,80.

En ese caso se puede suponer que la proporción $\pi = 0{,}05$ es la diferencia real entre ambas poblaciones y que tiene un desvío $\sigma = 0{,}05362$. En n = 200 muestras el valor esperado se calcula con $\mu_r = 200 \,.\, 0{,}05 = 10$ y se pide hallar la probabilidad de que $r \geq 10$. O sea $[1 - P_{Bi}\,(r \leq 9)]$

Calculando el valor tipificado con la corrección de Molenaar resulta:

$$z = [(4 \,.\, 9 + 3)(1 - 0{,}05)]^{1/2} - [(4 \,.\, 200 - 4 \,.\, 9 - 1)(0{,}05)]^{1/2} = 0{,}0897$$

O sea P = 0, 5357 y la probabilidad pedida es (1 – 0,5357) = 0,4643



El área a la derecha del valor tipificado es p = 0,4643 O sea, existe una probabilidad del 46,43% que la marca A tenga una sensibilidad mayor del 10% a la de la marca B.

(c) Utilizando los mismos supuestos que en el punto anterior, pero trabajando con la probabilidad exacta, la respuesta se calcula con:

$P_{Bi}\,(x > 10) = 1 - P_{Bi}\,(x \leq 9) = 1 - [P_{Bi}\,(x = 0) + P_{Bi}\,(x = 1) + ... + P_{Bi}\,(x = 9) = 0{,}4547$

Notar que ambas aproximaciones normales son bastante buenas. La idea de usar la diferencia de dos poblaciones binomiales como si fuese una sola se explicará más adelante en más detalle.

10.5 Problemas propuestos

1) Marcar la respuesta correcta a cada una de las afirmaciones siguientes, o completar la frase:

1) Una muestra es aleatoria si todo componente tiene la misma probabilidad de ser elegido. **V F**
2) El muestreo sistemático es de tipo aleatorio. **V F**
3) Los muestreos no aleatorios son: .... **...............**
4) En una tabla de números al azar se puede ir en diagonales a través de ella. **V F**
5) La base de la Teoría de las Grandes Muestras es el Teorema ....
6) Una distribución muestral de un estadígrafo tiende asintóticamente a una normal **V F**
7) En una distribución de medias el valor esperado es .... y su varianza es ....
8) En el punto anterior: valor esperado y varianza no cambian si la población es finita. **V F**
9) Cuando se realizan muestreos sin reposición hay que aplicar un factor de corrección. **V F**
10) En una distribución muestral de proporciones se puede trabajar igual tamaño muestral **V F**
11) Estas distribuciones valen si se toman todas las muestras posibles de tamaño N **V F**
12) En una distribución de proporciones el valor esperado es .... y su varianza es ....

13) Solo si las muestras son independientes se puede usar la distribución de diferencias. **V** **F**
14) Los casos vistos en este capítulo suponen la posibilidad de acceder a toda la población **V** **F**
15) Si los parámetros poblacionales son desconocidos no se puede aplicar lo visto más arriba. **V** **F**
16) El valor esperado y varianza de una diferencia de medias se expresa como:....................... **................**

2) En una medición de pesos se efectuaron 64 determinaciones de un objeto, resultado una media de 65 g con un desvío estándar de 4 g. Luego se llevó el objeto a una balanza calibrada en el INTI el cual informa un valor de referencia (que puede ser tomado como patrón) de 62,5 g. Determinar cuanto vale la probabilidad del valor tipificado de la media encontrada.

3) Si la Sensibilidad de una prueba clínica A es del 85% para un determinado punto de corte y la de otra prueba B para hacer el mismo análisis es del 90%. Se pide calcular el valor promedio y el desvío de la diferencia entre ambas, junto con su valor tipificado para TE = TS = 100.

4) El fabricante de cierto kit para medir glucosa entrega un estándar de 100 mg/dl con un error del 8%. Si un nuevo envase viene con otro estándar de 105 mg/dl y el mismo error, ¿cuánto vale el valor tipificado de la diferencia entre ambos envases si la diferencia de medias fue de 6 mg/dl?

5) Las 64 determinaciones de Glucosa de un paciente muestran una media de 88 g/dl con un desvío estándar de 5 mg/dl. El Bioquímico a cargo decide realizar 36 determinaciones más de la misma sangre pero usando un segundo kit cuya media fue de 90 y un desvío de 4 mg/dl. Tipificar esta diferencia.

6) En una farmacia se realiza una encuesta entre 100 clientes que concurren a la misma. El 75% de los encuestados se mostró favorable al uso de la marca A de jarabe para la tos. Luego de efectuada una propaganda intensa a través de los medios de comunicación, se vuelve a realizar la encuesta a otros 100 clientes y se obtiene una respuesta favorable en el 88% de los casos. Averiguar la probabilidad asociada a la diferencia de proporciones antes y después de la campaña.

7) En una balanza se pesaron 49 veces una cierta cantidad de droga resultando una media de 58 g con un desvío de 2 g. La misma droga se pesó 36 veces en otra balanza y se obtuvo un valor promedio de 55 g con un desvío de 4 g. Calcular la probabilidad asociada a la diferencia de medias entre las dos balanzas empleadas.

8) Suponiendo que las mediciones de glucosa se distribuyen normalmente entre los 5.000 varones sanos de una cierta población, con edades de 4 a 60 años, con un valor promedio de 9,5 y un desvío de 1,8. Un investigador decide tomar 90 muestras aleatorias de tamaño 4, con reemplazamiento. Desea saber en cuántas muestras cabe esperar una media entre 7 y 1,1.

9) Explicar como puede evitarse la influencia del factor humano en un muestreo basado en enfermedad, para la estimación de los índices clínicos.

10) Explicar cuales son los pasos a seguir en un estudio epidemiológico para evitar la influencia del investigador en estudios basados en exposición, en protección y en enfermedad.

11) Cuando se trabaja con proporciones y se puede aproximar con la función normal: ¿ Cuál es la mejor corrección por continuidad que se puede usar, para que la aproximación sea lo mejor posible ? ¿ Cuanto debe valer el tamaño de muestra para poder usar aproximaciones normales ?