

FACULTAD DE TECNOLOGÍA DE LA SALUD

ESTADÍSTICA

INFERENCIAL

Autores:

**HÉCTOR BAYARRE
RUBEN HERSFORD**

Tema 1- Teoría Elemental de Probabilidades.

Introducción

Si tu único propósito como investigador, es describir los resultados de un experimento concreto, los métodos analizados en el curso anterior, sobre Estadística Descriptiva, pueden considerarse suficientes. No obstante, si lo que pretendes es utilizar la información obtenida para extraer conclusiones generales, sobre la población de la cual fue seleccionada la muestra estudiada, entonces estos métodos constituyen sólo el principio del análisis, y debes recurrir a métodos de inferencia estadística, los cuales implican el uso inteligente de la teoría de la probabilidad.

Comenzaremos este tema interpretando la noción de probabilidad y la terminología subyacente a esta área de las matemáticas, ya que la probabilidad constituye por sí misma un concepto básico que refleja su relación con la faceta del mundo exterior que pretende estudiar: los fenómenos aleatorios, los cuales obedecen unas ciertas reglas de comportamiento. De alguna manera, el concepto de probabilidad, se relaciona o nos recuerda las propiedades de la frecuencia relativa.

Nos centraremos posteriormente, en el eslabón que une la teoría de la probabilidad y la estadística aplicada: la noción de variable aleatoria, mostrando de esta manera, como puede emplearse la teoría de la probabilidad para sacar conclusiones precisas acerca de una población sobre la base de una muestra extraída de ella, y que muchos de los estudios estadísticos son de hecho, estudio de las propiedades de una o más variables aleatorias.

Tal como hemos citado anteriormente, en las aplicaciones prácticas es importante poder describir los rasgos principales de una distribución, es decir, caracterizar los resultados del experimento aleatorio mediante unos parámetros. Llegamos así al estudio de las características asociadas a una variable aleatoria introduciendo los conceptos de valor esperado y varianza, relacionándolos con los conceptos de media y varianza de una variable estadística.

El cálculo de probabilidades nos suministra las reglas para el estudio de los experimentos aleatorios o de azar, constituyendo la base para la estadística inductiva o inferencial.

No pretendemos atiborrarte de fórmulas de cálculo, tampoco pasarnos todo el tema realizando ejercicios sobre urnas con bolas rojas y blancas, nuestro interés es proveerte de intuición suficiente

sobre la materia, de manera que el aspecto probabilístico de la inferencia estadística pueda ser fácilmente comprendido.

1.1 Probabilidad. Conceptos básicos.

La teoría de probabilidades tiene su origen en la solución de problemas relacionados con los juegos de azar, los cuales tienen sus precedentes en épocas muy remotas.

Desde mediados del siglo XVII y hasta nuestros días, el cálculo de probabilidades ha sufrido un proceso de transformaciones sucesivas, hasta alcanzar métodos aplicables a la solución de problemas complejos, que el desarrollo científico técnico plantea al hombre.

Muchas son las ciencias que emplean los resultados de esta disciplina, entre ellas se destacan la Física, la Química, la Biología, todas las ramas de la ingeniería, etc. Los progresos teóricos de esta rama de la matemática, han abierto nuevos campos de aplicación en Estadística. Así vemos que, en el campo de la salud pública es muy frecuente el uso de las probabilidades, no solo en el ámbito de la investigación, sino también en la práctica profesional cotidiana.

Cuando decimos que la oportunidad que tiene un paciente de sobrevivir a una intervención quirúrgica es de 50%. O bien, que un médico está 90 % seguro de que un paciente tiene tal enfermedad. Estos ejemplos, entre otros tantos que escuchamos a diario, nos muestran el empleo del concepto de probabilidad en el marco de la salud pública.

Seguro te preguntarás ¿qué entendemos por probabilidad?. Pues bien, te diré que existen varias definiciones, las que pudieran resultar algo complicadas dado el carácter matemático de las mismas. Apelamos a tu intuición y capacidad de razonamiento lógico.

Ante todo debes saber, que las probabilidades constituyen la medida numérica de la verosimilitud o el grado de certidumbre de ocurrencia real de un fenómeno.

Actualmente, existen dos puntos de vista mediante los cuales pueden ser abordadas las probabilidades: objetivo y subjetivo. El primero se basa en que la teoría de probabilidad permite descubrir las leyes que rigen una serie de fenómenos que se presentan en la práctica, cuyo análisis y estudio conducen al conocimiento de la realidad objetiva; mientras que el segundo, sostiene que la probabilidad mide la confianza que un individuo tiene en la certeza de una proposición particular, es decir, que las conclusiones

basadas en esta teoría no reflejan la realidad objetiva, sino el conocimiento del observador sobre el fenómeno que estudia.

Aunque este último punto de vista ha gozado de gran popularidad, los estadísticos con orientación tradicional aun no la aceptan del todo.

El concepto de probabilidad objetiva puede ser visto desde dos enfoques principales: el empírico y el teórico.

El enfoque teórico se corresponde con la llamada **probabilidad clásica** o **a priori** y puede ser definida como sigue:



Sea **E** un experimento aleatorio con **N** resultados igualmente posibles y mutuamente excluyentes, y si **m** de estos resultados poseen una característica **A**, la **probabilidad** de ocurrencia de **A** se define como la relación entre el número de estos resultados favorables a A (**m**) y el número total de resultados posibles del experimento (**N**). Se denota:

$$P(A) = \frac{m}{N}$$

¿Por qué a priori?, Ya que no necesita ser realizado el experimento para conocer la probabilidad de ocurrencia de los resultados posibles.

Ahora bien, para trabajar con el cálculo de probabilidades es necesario fijar previamente cierta terminología. Vamos a introducir parte de ella en las próximas líneas.

1.1.1 - Experimentos y sucesos aleatorios

Cuando en un experimento no se puede predecir el resultado final, hablamos de experimento aleatorio. Éste debe cumplir los siguientes requisitos:

1. Se puede repetir indefinidamente, siempre en las mismas condiciones;
2. Antes de realizarlo, no se puede predecir el resultado que se va a obtener;
3. En cada repetición se obtiene un resultado que pertenece a un conjunto conocido previamente de resultados posibles. A este conjunto, de resultados posibles, lo denominaremos **espacio muestral** o **conjunto fundamental** y lo denotaremos normalmente mediante la letra **E**. Los elementos del espacio muestral, cada uno

de los resultados posibles, se denominan **sucesos elementales** (e_1, e_2, \dots, e_n).

El desconocimiento de todas las circunstancias que describen el experimento aleatorio, causa la incertidumbre del resultado. Esto motiva la aparición de interrogantes sobre dicho resultado, a su vez estas preguntas determinan un subconjunto de sucesos elementales. A estos subconjuntos suele llamárseles **eventos** o **sucesos aleatorios**, existiendo una probabilidad de ocurrencia asociada a éstos.

Observa que los sucesos elementales son sucesos aleatorios compuestos por un sólo elemento, es decir, solo un resultado posible. Por supuesto los sucesos aleatorios son más generales que los elementales, ya que son conjuntos que pueden contener no a uno sólo, sino a una infinidad de sucesos elementales, e incluso no contener ninguno.

Dos o más sucesos son llamados **mutuamente excluyentes**, si la ocurrencia de cualquiera de ellos excluye la ocurrencia del(os) otro(s), es decir, no pueden ocurrir simultáneamente. Así tenemos que los sucesos elementales son mutuamente excluyentes, mientras que los sucesos aleatorios pueden en ocasiones no serlo, lo que significa que algunos de sus elementos coinciden. Ilustremos estos conceptos con un ejemplo simple.

Ejemplo 1.1.1

Si realizamos el experimento aleatorio de lanzar un dado al aire, tenemos los siguientes resultados posibles: $E = \{1, 2, 3, 4, 5, 6\}$

El conjunto E es el espacio muestral de este experimento y cada uno de sus elementos son sucesos elementales. Ahora, las preguntas ¿el resultado será un número par?, ¿Un múltiplo de tres? o ¿un número mayor que dos?, definen, respectivamente, los siguientes subconjuntos (sucesos aleatorios) del espacio muestral

E : $A = \{2, 4, 6\}$ $B = \{3, 6\}$ $C = \{3, 4, 5, 6\}$

Observa que todos los elementos de E son mutuamente excluyentes, pues si sale el 1 no puede simultáneamente salir otro valor, mientras que los sucesos aleatorios mencionados no son mutuamente excluyentes, si por ejemplo sale el número 6, habrán ocurrido los tres sucesos simultáneamente: A , B y C . Esto no es así siempre, por ejemplo, un suceso D es definido por la pregunta ¿el resultado será un múltiplo de 5?, tendrá los siguientes elementos:

$D = \{5\}$

Como verás, A y D son mutuamente excluyentes, al igual que B y D, no sucediendo así con C y D.

A continuación veremos como se realiza el cálculo de probabilidades en este ejemplo, según la definición clásica.

En este caso $N = 6$ (las 6 caras del dado), y si el dado está perfectamente balanceado, es decir, no está trucado, la probabilidad de ocurrencia de cada uno de los resultados posibles será la misma para todos e igual a: $\frac{1}{6}$

Esta es la probabilidad de ocurrencia de los sucesos elementales, veamos que sucede con los sucesos aleatorios. Pues muy simple, la probabilidad de ocurrencia de cada uno de ellos está determinada por las probabilidades individuales de sus elementos:

$$P(A) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$$

Para los demás es similar:

$$P(B) = \frac{2}{6} = \frac{1}{3} \quad P(C) = \frac{4}{6} = \frac{2}{3} \quad P(D) = \frac{1}{6}$$

Otro concepto importante es el de **suceso contrario o complemento**. Hasta ahora hemos hablado de probabilidad de ocurrencia de un suceso cualquiera **A**, denotada por $P(A)$, no te has preguntado ¿cuál será la probabilidad de que **A** no ocurra?, si es así, te diré que la no ocurrencia de un suceso (**A**) es a lo que suele llamársele su complemento o el suceso contrario de **A** y se denota como **no A**, o más comúnmente como \bar{A} y su probabilidad está dada por:

$$P(\bar{A}) = 1 - P(A)$$

Veamos en el ejemplo 2.1.1. El complemento de los sucesos A, B, C y D será:

$$\begin{aligned} \bar{A} &= \{1, 3, 5\} & \bar{B} &= \{1, 2, 4, 5\} \\ \bar{C} &= \{1, 2\} & \bar{D} &= \{1, 2, 3, 4, 6\} \end{aligned}$$

Las respectivas probabilidades puedes calcularlas fácilmente. Hazlo.

La definición clásica de probabilidad tiene dos desventajas: las palabras "igualmente posibles" pueden ser sinónimas de "igualmente probables", por lo que estamos definiendo exactamente la probabilidad con sus propios términos, además, este concepto supone que el número de resultados posibles es finito, lo que no ocurre frecuentemente en la práctica.

Por esta razón ha sido propuesta una definición estadística de probabilidad, también llamada **frecuencia relativa de probabilidad** o **probabilidad a posteriori**, que se corresponde con el **método empírico**.

Pues sí, nos referimos a la misma frecuencia relativa que viste en estadística descriptiva. A continuación te aclaro esto.

Cuando repetimos un experimento aleatorio, un número grande de veces, y registramos la frecuencia de ocurrencia de un resultado e , se observa que en la medida que el número de experimentos aumenta, las frecuencias relativas con las que ocurre ese suceso e , $Fr(e)$, tiende a converger hacia cierta cantidad que denominamos probabilidad de e . Veamos un ejemplo.

Ejemplo 1.1.2

En el siguiente gráfico, se presenta la evolución de la frecuencia relativa del número de caras obtenido en el lanzamiento de una moneda en 100 ocasiones (simulado por una computadora, no pensarás que lanzamos la moneda 100 veces). En principio la evolución de las frecuencias relativas es errática, pero a medida que el número de tiradas aumenta, tiende a lo que entendemos por probabilidad de cara, es decir, tiende a 0.5 (fíjate que esta moneda está perfectamente balanceada).

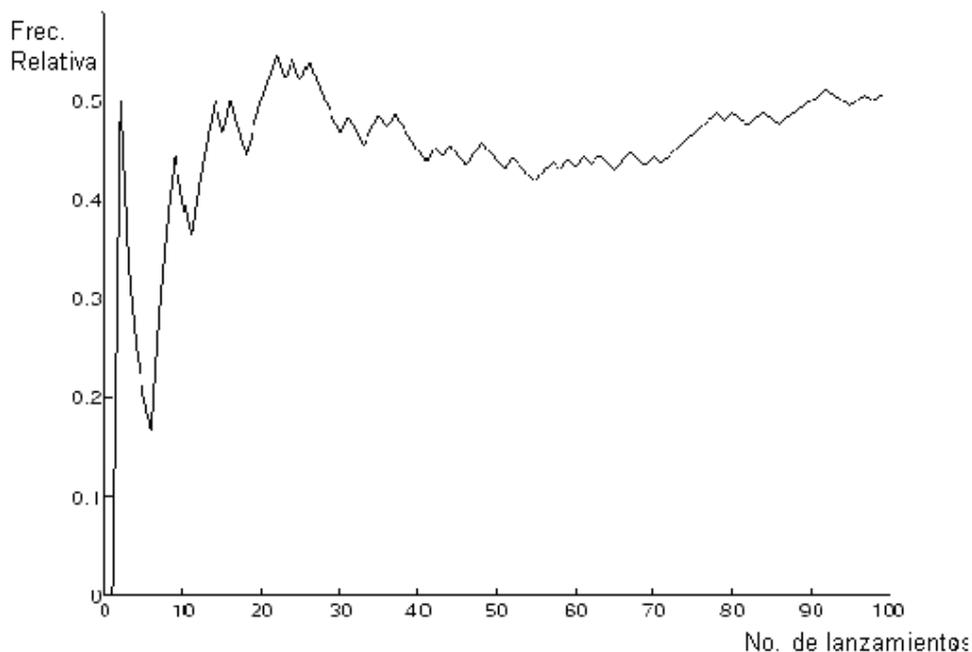


Figura 1.1.1- Tendencia de la frecuencia relativa del número de caras obtenido en lanzamientos sucesivos de una moneda.

De acuerdo a lo anterior:



La probabilidad estimada o **probabilidad empírica** de un suceso, se toma como la frecuencia relativa de la ocurrencia del suceso, cuando el número de observaciones es muy grande.

¿Por qué a posteriori?, Pues porque a diferencia de la definición clásica, en este caso para calcular la probabilidad de ocurrencia de un suceso es necesario realizar el experimento un número grande de veces, después de lo cual es que se conocerá dicha probabilidad.

Esta definición también tiene dificultades en la práctica, pues se requiere realizar un número infinito de veces un experimento para calcular una probabilidad. Por ejemplo, lanzar infinitas veces un dado para ver que las frecuencias relativas de la aparición de cada cara convergen a $1/6$. Esto puede suplirse en la práctica realizando el experimento un número suficientemente elevado de veces, hasta que tengamos la precisión que requieran nuestros cálculos.

Sin embargo, los experimentos aleatorios a veces no pueden ser realizados, como es el caso de calcular la probabilidad de morir jugando a la ruleta rusa con un revólver: No es posible (o no se debe) calcular esta probabilidad repitiendo el experimento un número indefinidamente alto de veces para aproximarla mediante la frecuencia relativa. Para ello existen métodos mucho más seguros, como es la definición axiomática de probabilidad, lo cual escapa del alcance de este texto, si estás interesado en profundizar en el tema puedes consultar la bibliografía citada al final del capítulo.

Por lo pronto, de acuerdo con los objetivos del texto, nos quedamos con la definición estadística de probabilidad. De hecho, prácticamente todo lo concerniente a probabilidades en el campo de la salud se obtiene de forma empírica, basándonos en el comportamiento pasado de los fenómenos. Por ejemplo, la probabilidad de supervivencia de un enfermo de cáncer, se basa en el conocimiento de los porcentajes de supervivencia de pacientes con características similares, con el mismo nivel de la enfermedad y sometidos al mismo tratamiento. Como puedes ver, estos porcentajes no son más que la expresión porcentual de la frecuencia relativa de este suceso.

El método empírico es también la base del antiguo método de diagnóstico, según el cual es más verosímil (probable) que un paciente tenga una enfermedad común que una rara.

Es importante que te aclare, que esta definición supone estabilidad en las circunstancias o factores que determinan el suceso en sí. En el ejemplo anterior, si se modifica alguno de esos factores como incluir pacientes con diferentes niveles de la enfermedad o sometidos a diferentes tratamientos, estas probabilidades empíricas ya no son válidas.

La clave está en que la probabilidad, basada en antiguos resultados, es cierta ahora y en el futuro, solamente en circunstancias semejantes.

Habrás notado, que las probabilidades pueden ser expresadas tanto en forma fraccional como porcentual. Ambas formas son válidas y equivalentes, así por ejemplo, una probabilidad del 20 % es lo mismo que de 0.2. Pero, ¿cómo interpretar su valor?, Para interpretar el valor de una probabilidad es necesario recurrir al concepto frecuentista de la misma, por ejemplo, $P(A) = 0.2$ significa que de cada 100 repeticiones del experimento se espera como promedio que ocurra 20 veces el evento A.

Ahora bien, independientemente de cómo la expreses, existen propiedades básicas que cumple la probabilidad de cualquier suceso e, $P(e)$. Estas son:

- La $P(e)$ es un número no negativo y menor o igual que 1 (entre 0 y 100%), esto no te es ajeno puesto que las probabilidades son proporciones y conoces las propiedades de una proporción.

$$0 \geq P(e) \leq 1$$

- La suma de las probabilidades de ocurrencia, de todos los resultados posibles de un experimento aleatorio, es igual a 1 (100%).

$$P(e_1) + P(e_2) + \dots + P(e_n) = 1$$

Esta es la propiedad de exhaustividad, se refiere al hecho de que el observador de un proceso probabilístico, debe contemplar todos los resultados posibles, y cuando se toman todos, la suma de sus probabilidades siempre es igual a 1. Ahora comprendes por qué la probabilidad del complemento de un suceso es 1 menos la probabilidad de ese suceso, pues según esta propiedad:

$$P(A) + P(\bar{A}) = 1$$

- Si dos eventos E_1 y E_2 , son mutuamente excluyentes. La probabilidad de ocurrencia de uno o el otro es igual a la suma de sus probabilidades individuales.

$$P(E_1 \text{ o } E_2) = P(E_1) + P(E_2)$$

Si no son mutuamente excluyentes el cálculo se torna algo complejo.

1.2 Variables aleatorias.

Hasta aquí, al analizar los resultados de un experimento o fenómeno aleatorio, hemos establecido la correspondencia entre los posibles resultados y la noción de probabilidad de ocurrencia de un determinado resultado o conjunto de resultados (sucesos aleatorios), de todos los posibles. Como consecuencia, a cada resultado o suceso aleatorio, está asociado un número real, que es su probabilidad.

La noción de variable aleatoria está asociada al concepto de probabilidad de ocurrencia de un suceso, y es también una correspondencia que se establece entre cada resultado de un fenómeno o experimento aleatorio y un número real determinado.

En el proceso de investigación científica y en particular en la investigación biomédica, es frecuente expresar el resultado de un experimento aleatorio en forma numérica, pues resulta más fácil utilizar valores numéricos en lugar de trabajar directamente con los elementos de un espacio muestral. De este modo aparece el concepto de variable aleatoria como sigue:



Una **variable aleatoria** (v.a) es toda función que atribuye un único número real a cada suceso elemental del espacio muestral de un experimento aleatorio.

Es decir, una variable aleatoria (v.a) es simplemente, la formalización matemática de alguna de las variables reales que estamos habituados a tratar. Cuando se trata de un carácter cuantitativo, la variable aleatoria correspondiente se define mediante su aplicación idéntica, mientras que si es un carácter cualitativo se emplea una convención numérica conveniente.

Esto puede aplicarse tanto si el resultado del experimento es una cantidad que varía como si se trata de un resultado categórico.

Por lo tanto, una variable aleatoria será definida como una cantidad variable que expresa el resultado de un experimento aleatorio.

Dicho de otro modo, se define una variable aleatoria haciendo corresponder a cada uno de los sucesos aleatorios, un número real

cualquiera, de modo que esta transformación nos permita definir diferentes operaciones aritméticas con sus elementos.

Estas variables se denotan con las últimas letras del alfabeto, en mayúsculas (X, Y, Z).

Ejemplo 1.2.1

Si el experimento consiste en observar el sexo de un recién nacido, es posible que nos interese determinar el número de varones que se obtienen de un total de 3 nacimientos.

El espacio muestral de este experimento estará formado por todas las formas posibles en que pueden ocurrir 3 nacimientos en cuanto a sexo. Te explico mejor, cada uno de los 3 nacimientos puede ser de 2 formas: hembra (H) o varón (V), por lo tanto si queremos saber de cuantas formas pueden ocurrir los 3 nacimientos tendríamos 8 posibilidades, 2 del primero, multiplicado por 2 del segundo, multiplicado a su vez por 2 del tercer nacimiento: $2 \times 2 \times 2 = 8^1$.

Veamos claramente; los resultados pueden ser:

$$\mathbf{E = \{VVV, HHH, VHH, VVH, HVV, HHV, HVH, VHV\}}$$

Como puedes ver, resulta bastante complejo trabajar con este espacio muestral tal como es, imagínate si aumentamos el número de nacimientos, sería el caos. Por lo tanto será mucho más conveniente que el suceso "hembra" sea representado por el 0 y el suceso "varón" por el 1, de esta forma se define una variable aleatoria **X** como el número de varones, de la siguiente forma:

$$\mathbf{X = \{0, 1, 2, 3\}}$$

Así pues hagamos corresponder el espacio muestral anterior con los valores de **X**

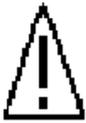
$$X = \begin{cases} 0 \Rightarrow & \mathbf{HHH} \\ 1 \Rightarrow & \mathbf{VHH, HVH \text{ y } HHV} \\ 2 \Rightarrow & \mathbf{VVH, VHV \text{ y } HVV} \\ 3 \Rightarrow & \mathbf{VVV} \end{cases}$$

Más adelante retomaremos este ejemplo.

Ahora bien, el calificativo de aleatoria no se debe al hecho de que atribuyes de modo imprevisible un valor cualquiera a un elemento del espacio muestral, ya que este valor está determinado de forma precisa. Lo que es aleatorio en realidad, es que al hacer el experimento, no sabemos que resultado puede ocurrir.

¹ Esto es de teoría combinatoria, que no creo necesario abundar mucho, si lo deseas puedes consultar la bibliografía citada al final del tema.

En función de los valores que tome la variable, ésta puede clasificarse en discreta o continua del siguiente modo:



Variable aleatoria discreta: Es aquella que sólo puede tomar un número finito o numerable de valores. Por ejemplo, el sexo de un recién nacido.

Variable aleatoria continua: Es la que puede tomar un número infinito no numerable de valores. Por ejemplo, la talla, y en general variables que expresan tiempo, medidas de longitud, peso, etc.

Si sobre los elementos de E existe una distribución de probabilidad, esta se transmite a los valores que toma la variable X . Es decir, toda v.a. conserva la estructura probabilística del experimento aleatorio que describe.

Retomando el ejemplo 1.2.1, supongamos que la probabilidad de que nazca un niño varón es similar a la de que nazca una hembra, e igual a 0.5 (no es así realmente). ¿Cómo serán las probabilidades de los valores de X ? Bueno en este caso como las probabilidades de cada sexo son iguales, entonces los elementos de E son igualmente probables, por lo que se puede aplicar el concepto de probabilidad clásico. Veamos: E tiene 8 resultados posibles, así que la probabilidad correspondiente a cada uno será igual a $1/8$. Por ejemplo.

El resultado VVV, su probabilidad se calcula así:

$$P(VVV) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$$

De forma similar para el resto de los resultados. Ahora calculemos las probabilidades de X :

$$P(X=0) = \frac{1}{8} \quad \text{recuerda que } X = 0 \text{ solo cuando } E = HHH$$

$$P(X=1) = \frac{3}{8} \quad X = 1 \text{ cuando } E = HVH, VHH \text{ y HHV}$$

$$P(X=2) = \frac{3}{8} \quad X = 2 \text{ cuando } E = VVH, VHV \text{ y HVV}$$

$$P(X=3) = \frac{3}{8} \quad X = 3 \text{ solo cuando } E = VVV$$

Vamos a estudiar a continuación los conceptos más importantes relacionados con la distribución de probabilidad de una v.a., diferenciando entre los casos de v.a. discreta y v.a. continua.

1.3 Distribución de probabilidad de una variable aleatoria.

El concepto de variable aleatoria nos permite, como vimos en el ejemplo 2.2.1, definir y denotar sucesos, y aún más, calcular la probabilidad de éstos, a través de la probabilidad de ocurrencia de los valores originales. Así pues, la distribución de probabilidad de una variable aleatoria puede ser definida de acuerdo al tipo de v.a de la siguiente forma:



Distribución de probabilidad de una v.a discreta: se define como el conjunto de todos los valores de la variable, junto con sus correspondientes probabilidades. Puede ser representada tanto tabular como gráficamente.

Veamos la distribución de probabilidades de la variable **X**: número de varones en tres nacimientos, vista anteriormente en el ejemplo 1.2.1.

Tabla 1.3.1 – Distribución de probabilidad del número de varones en 3 nacimientos.

Valores de X (x)	P (X = x)
0	0.125
1	0.375
2	0.375
3	0.125
	1.000

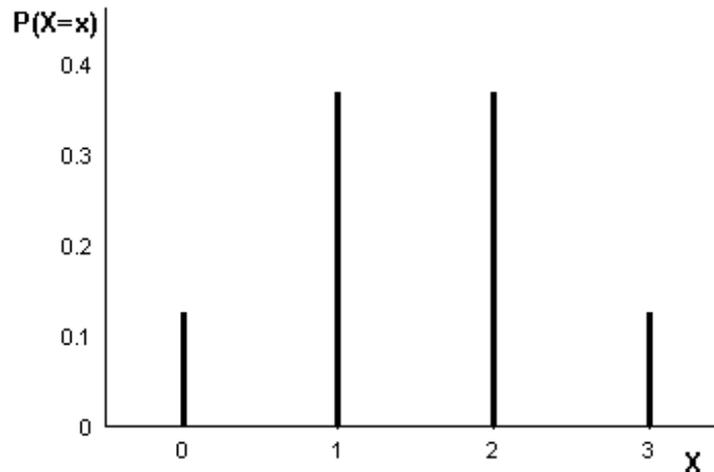


Figura 1.3.1 – Representación gráfica de la distribución de probabilidad del número de varones en 3 nacimientos.

Como puedes observar, la representación gráfica de la función de probabilidad se realiza mediante un diagrama de barras análogo al de la distribución de frecuencias relativas para variables discretas. Donde la longitud de cada barra equivale al valor de probabilidad del correspondiente valor de x . Veamos un ejemplo con más categorías.

Ejemplo 1.3.1

Supongamos que como parte del diagnóstico de salud del área, necesitas conocer el número de personas por familia, para ello tomas una muestra de 50 familias y en cada una de ellas registras la variable de interés. Construyamos la distribución de probabilidad para esta variable en dicha población.

Tabla 1.3.2 – Distribución de probabilidad del número de personas por familia en una población de 50 familias.

Número de personas	Frecuencia absoluta	$P(X=x)$
1	1	0.02
2	4	0.08
3	6	0.12
4	5	0.10
5	9	0.18
6	10	0.20
7	7	0.14

8	4	0.08
9	2	0.04
10	2	0.04
	50	1.00

Te habrás dado cuenta que esto es análogo a la distribución de frecuencias relativas, pero sustituyendo las frecuencias relativas por probabilidades. Por lo tanto, podemos pensar en las distribuciones de probabilidad como las formas teóricas o ideales de las distribuciones de frecuencia relativa, cuando el número de observaciones es grande (poblaciones).

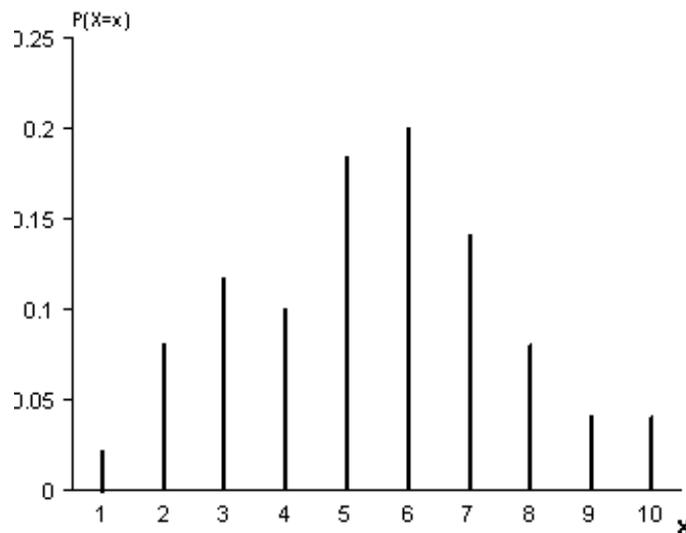


Figura 13.2 – representación gráfica de la distribución de probabilidad del número de personas por familia en una población de 50 familias.

En ambos ejemplos (1.2.1 y 1.3.1), observarás que la longitud de cada barra vertical indica la probabilidad para el valor correspondiente de x . Puedes usar alternativamente la representación gráfica o tabular, no necesariamente las dos. Observarás también que todos los valores de $P(X=x)$ son positivos, menores que 1, y la suma de los mismos es igual a 1. Estas características no son particulares de estos ejemplos, sino que se cumplen para todas las distribuciones de probabilidad de variables discretas.

Esta distribución te sirve para hacer afirmaciones (conclusiones) sobre la variable aleatoria en la población, hacer pronósticos y planificar medidas de prevención, entre otras cosas.

Las ideas vistas anteriormente, pueden extenderse al caso en que la variable pueda asumir un conjunto continuo de valores. En

este caso el concepto de distribución de probabilidad es más complejo.

Cuando la variable es continua, no tiene sentido hacer una suma de las probabilidades de cada uno de los términos, ya que el conjunto de valores que puede tomar la variable es **no numerable**. En este caso, lo que generaliza de modo natural el concepto de suma es el de integral. Por otro lado, para variables continuas no tiene interés hablar de la probabilidad de un valor específico de la variable, $P(X = x)$, ya que esta debe de valer siempre 0, para que la **suma infinita no numerable** de las probabilidades de todos los valores de la variable no sea infinita.

De este modo es necesario introducir un nuevo concepto que sustituya en v.a. continuas, al de función de probabilidad de una v.a. discreta. Este concepto es el de **función de densidad de una v.a. continua**, que se define como sigue:



A una función no negativa, $f(x)$, se le llama **distribución de probabilidad (función de densidad de probabilidad)** para la variable aleatoria continua X , si el área total delimitada por su curva y el eje de las x , es igual a **1** y si la subárea delimitada por la curva, el eje de las x y por las líneas perpendiculares levantadas sobre dos puntos cualesquiera **a** y **b** da la probabilidad de que X esté entre los puntos **a** y **b**.

Para que comprendas mejor, considera los datos presentados en la tabla 1.3.3 y la figura 1.3.3

Ejemplo 1.3.2

La talla de un recién nacido acostumbra a tener valores entre 47 y 53 cm, pero no todas estas tallas son igualmente probables, como tú bien sabes, lo habitual son tallas próximas a los 50 cm. Lo cual se puede observar tanto en la distribución de frecuencias de la tabla 1.3.3 como en el histograma de la figura 1.3.3, alrededor del 60 % de los niños se sitúan entre 49 y 51 cm de talla.

Tabla 1.3.3 – Distribución de frecuencias de la talla de 100 recién nacidos.

Talla (cm)	Frecuencia absoluta	Frecuencia relativa	Frecuencia acumulada
46	1	0.01	0.01
47	6	0.06	0.07
48	10	0.10	0.17
49	18	0.18	0.35
50	25	0.25	0.60
51	23	0.23	0.83
52	10	0.10	0.93
53	7	0.07	1.00
Total	100	1.00	

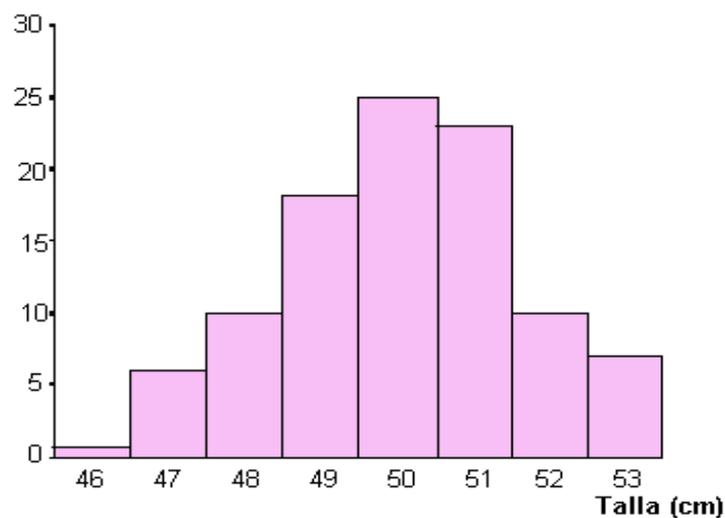


Figura 1.3.3 – Histograma de la talla (cm) de 100 recién nacidos.

Como recordarás del tema 1 del curso anterior, las subáreas del histograma corresponden a las frecuencias de ocurrencia de los valores de la variable entre los límites de la escala horizontal de subáreas, por lo tanto el área total se corresponde con el total de observaciones. En caso de representar frecuencias relativas en vez de absolutas, las subáreas se corresponderán con los valores de probabilidad de cada clase representada y por supuesto el área total será igual a 1.

Imagínate ahora una situación donde el número de valores de la variable aleatoria es muy grande y la dimensión de los intervalos de clase es muy pequeña, digamos que contamos con la talla de todos los RN del país durante un año (aproximadamente 150000), en este caso el histograma resultante tendrá mayor cantidad de intervalos y si se conectan los puntos medios de las celdas del mismo se obtendrá una curva suave como la distribución continua representada en la figura 1.3.4.

Estas curvas se utilizan para representar gráficamente las distribuciones de v.a continuas, lo cual tiene implicaciones importantes cuando se trabaja con distribuciones de probabilidad. En primer lugar, el área bajo la curva es igual a 1, como lo es para el histograma, y en segundo lugar, la frecuencia relativa de ocurrencia para los valores entre dos puntos específicos cualesquiera (a y b), sobre el eje de las X , es igual al área total delimitada por la curva, el eje de las X y las rectas perpendiculares levantadas sobre ambos puntos, tal como lo muestra la figura 1.3.5. Siendo la probabilidad de que la variable tome valores entre a y b , y se denota por: **$P(a < X < b)$**

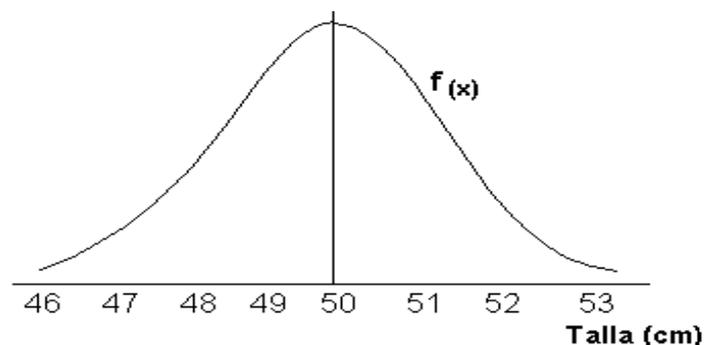


Figura 1.3.4 – Distribución de probabilidad de las tallas al nacer.

A esta función $f(x)$, se le llama función de densidad de probabilidad de X .

Por ser f una función integrable, la probabilidad de un punto es nula, es decir, la probabilidad de un valor específico de la v.a es cero, puesto que un valor específico se representa por un punto en el eje de las abscisas y el área encima de un punto es igual a cero. Por ello al calcular la probabilidad de un intervalo no afectará nada el que éste sea abierto o cerrado por cualquiera de sus extremos, pues estos son puntos y por tanto de probabilidad nula.

Así pues: **$P(a < X < b) = P(a \leq X \leq b)$**

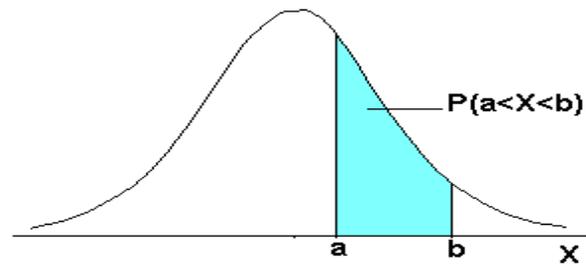


Figura 1.3.5 – Gráfico de una distribución de probabilidad continua que muestra el área entre a y b.

Para calcular las áreas bajo la curva y obtener de ese modo las probabilidades de presentarse algún intervalo específico de valores de la variable, se utiliza el proceso de integración, lo cual escapa al alcance de este texto por su grado de complejidad.

En la práctica existe una serie de distribuciones teóricas, tanto para variables discretas como para variables continuas, que sirven de modelo para representar las distribuciones empíricas más frecuentemente encontradas; y con el uso de éstas no necesitas hallar ninguna integral, pues al alcance de todos existen los valores de probabilidad tabulados para cada distribución teórica, solamente debes aprender a usarlas convenientemente. Además, puedes apelar a los software estadísticos.

A continuación te presentaremos, de forma muy resumida, algunas de las distribuciones de probabilidad teóricas que a nuestro juicio son las más usadas en las investigaciones que se realizan en la atención primaria de salud. Así que presta mucha atención porque esta es la base para comprender la inferencia estadística.

Pero antes debes conocer algunas “medidas de resumen” para las variables aleatorias. Así, la distribución de probabilidad de una v.a puede resumirse con la media, llamada en este caso esperanza matemática o valor esperado, y la varianza, las que se denotan por $E(X)$ y $V(X)$, respectivamente. Su cálculo, análogo al de una distribución de datos, se efectúa con las probabilidades de cada valor, en lugar del número de veces (frecuencia) con que el valor aparece en la muestra. Estas medidas constituyen los parámetros de la distribución.

En aras de simplificar, veremos sus formas de cálculo en cada distribución en particular que estudiaremos seguidamente.

1.4 La distribución Binomial.

Entre las distribuciones de probabilidad discretas, una de las más empleadas es la Binomial, por estar asociada a problemas de muestreo. Para realizar su estudio comenzaremos por una distribución más simple, cuyos parámetros serán utilizados para determinar la ley de probabilidad Binomial, esta es la llamada distribución Bernoulli.

En la práctica diaria, se presentan muchos problemas cuyo resultado, solo puede ser uno de dos categorías posibles, mutuamente excluyentes. Por ejemplo, vida o muerte, enfermo o no, varón o hembra, fuma o no, etc. Esta clase de experimentos, recibe el nombre de experimentos Bernoulli y sus resultados suelen denominarse como éxito y fracaso. La asignación del éxito o fracaso a las categorías originales, depende de los objetivos y de las características del experimento en particular, no existen reglas precisas al respecto. Pero si es muy importante definirlo adecuadamente y de forma clara, pues de esto dependerá la interpretación de los resultados, como verás más adelante.

En este caso las categorías de la variable original serán denotadas por 0 y 1, usando el valor 1 para el éxito.

Ejemplo 1.4.1

El sexo de un recién nacido puede considerarse como un experimento Bernoulli, quedando conformada la siguiente variable aleatoria X:

$$X = \begin{cases} 1 & = \text{varón} \\ 0 & = \text{hembra} \end{cases}$$

La probabilidad de ocurrencia de cada resultado, será precisamente, la probabilidad de éxito y la probabilidad de fracaso, denotadas por las letras **p** y **q** (minúsculas), respectivamente. De modo que:

$$p = P(X = 1)$$

$$p + q = 1 \quad \text{por lo tanto: } q = 1 - p$$

Pues **p** y **q** son las probabilidades de sucesos mutuamente excluyentes y **q** es el complemento de **p**.

El valor esperado y la varianza de esta distribución Bernoulli son:

$$E(X) = p \quad \text{y} \quad V(X) = pq$$

Ahora bien, cuando repetimos n veces un experimento Bernoulli, de forma independiente, en cada uno de ellos obtendremos uno de los dos resultados posibles: éxito o fracaso. Así, podemos estar interesados en conocer la probabilidad de ocurrencia de un número k de éxitos, en n repeticiones del

experimento. Para esto se construye una variable Binomial, que puede ser definida como sigue:



Variable aleatoria Binomial: es el número de éxitos obtenidos en n repeticiones independientes de un experimento Bernoulli. La que sigue una ley de distribución Binomial con parámetros n y p . Y se denota como:

$$X \sim B(n, p)$$

Donde: n = número de repeticiones.

p = probabilidad de éxito, la cual se mantiene constante a través de las n repeticiones.

Esta variable tomará valores entre 0 y n , es decir el número de éxitos puede ser ninguno (cero), uno, dos, y así sucesivamente hasta n (todos). Su valor esperado y varianza están dados por:

$$E(X) = np \text{ y } V(X) = npq$$

El número de éxitos esperado, $E(X)$, será el producto de la cantidad de repeticiones por la probabilidad de ocurrencia de un éxito, esto es algo bastante intuitivo.

Retomemos el ejemplo 2.2.1, la variable X definida como el número de varones en tres nacimientos, es una variable Binomial, si consideramos el nacimiento de un varón como el éxito (lo cual es así en algunas regiones). En este caso $n = 3$, el número de nacimientos, y $p = 0.5$, recuerda que asumimos que la probabilidad de nacer varón era similar a la de nacer hembra, e igual a 0.5, por tanto esta variable se distribuye Binomial con parámetros 3 y 0.5.

$$X \sim B(3, \frac{1}{2})$$

A su vez $X = \{0, 1, 2, 3\}$ y su valor esperado y varianza serán:

$$E(X) = 3 * \frac{1}{2} = \frac{3}{2} = 1.5 \quad V(X) = 3 * \frac{1}{2} * \frac{1}{2} = \frac{3}{4} = 0.75$$

Como habrás notado esto es absurdo, pues es imposible que nazca un varón y medio, pero debes tener en cuenta que esta es una operación matemática con un número fraccionario (probabilidad), por lo que resultados como este no son nada raros, entonces utiliza tu sentido común y realiza la aproximación correspondiente.

Este ejemplo es poco ilustrativo dado el pequeño número de repeticiones, aumentemos los nacimientos a 100, ahora el valor esperado será aproximadamente de 50 varones por cada 100 nacimientos.

También pueden calcularse las probabilidades de que ocurran k nacimientos varones, $P(X=k)$, lo cual nos puede servir en la

práctica, para planificar algún servicio en particular que se relacione con el sexo.

Estos cálculos se realizan mediante una fórmula matemática no muy compleja, pero escapa del alcance de este texto. Por otro lado, existen valores de probabilidad tabulados para diferentes valores de n y p , que podrás encontrar en la bibliografía citada al final del tema, con explicaciones sencillas de cómo usarlas.

1.5 La distribución de Poisson.

Esta distribución teórica ha sido ampliamente usada en Biología y Medicina. Constituye, junto a la Ley Binomial, las distribuciones discretas más comúnmente empleadas.

La distribución de Poisson es el resultado de un conjunto de suposiciones acerca de un proceso implícito para formar un conjunto de observaciones numéricas. Este proceso surge al contar el número de eventos o sucesos, distribuidos al azar, que ocurren en un espacio o intervalo de tiempo dado. Cumpliendo los siguientes requisitos:

- Las ocurrencias de los eventos son independientes, es decir, la ocurrencia de un evento en un intervalo de tiempo o espacio, no tiene efecto en la probabilidad de una segunda ocurrencia del evento en el mismo o en algún otro intervalo.
- Es posible la ocurrencia del evento, teóricamente, un número infinito de veces dentro del intervalo.
- La probabilidad de una sola ocurrencia del evento en un intervalo dado es proporcional a la dimensión del intervalo.
- En cualquier fracción infinitesimal (muy, pero que muy pequeña) del intervalo, la probabilidad de más de una ocurrencia del evento es insignificante.

De esta forma:



Una variable que se distribuye **Poisson** se define como el número de veces que ocurre un evento en un intervalo de tiempo o en un espacio dado. Se denota como:

$$X \sim P(\lambda)$$

Se lee: la variable² X se distribuye Poisson con parámetro lambda. Donde lambda (λ), es el número promedio de ocurrencias

² Te habrás fijado que usamos casi siempre la letra X para denotar a las variables, esto lo hacemos por comodidad, cuando se trabaja con más de una variable en un mismo ejercicio entonces es necesario diferenciar.

por unidad de tiempo. A la vez que constituye el valor esperado y la varianza de la distribución.

$$E(X) = \lambda$$

$$V(X) = \lambda$$

Esta distribución es una aproximación a la Binomial para el caso de n muy grande y p muy pequeña de forma tal que $\lambda = np$ no sea una magnitud grande. Se plantea que la aproximación es buena cuando $np \leq 5$.

Existen fórmulas para calcular la probabilidad de cualquier número de eventos, pero no creo necesario que las manejes, en la práctica puedes usar los valores tabulados para diferentes valores de λ , que podrás encontrar en la bibliografía citada al final del tema. Ilustremos con un ejemplo.

Ejemplo 1.5.1

El número de personas que acuden al cuerpo de guardia de un policlínico, con crisis aguda de asma bronquial, entre las 12 de la noche y las 6 de la mañana, constituye una variable aleatoria que sigue aproximadamente un proceso de Poisson. El parámetro λ , en este caso, está determinado por el número promedio de casos por intervalo, el cual se obtiene de la observación anterior de ese fenómeno. Esta información es útil para la dirección de dicho centro, pues conociendo las probabilidades de ocurrencia de k número de casos por intervalo de tiempo pueden planificar el abastecimiento de medicamentos destinados al tratamiento de esta enfermedad, además de hacer pronósticos y predicciones.

1.6 La distribución normal.

Hasta el momento hemos visto distribuciones de probabilidad discretas, veamos a continuación algunas distribuciones de variables continuas, entre las que la distribución normal constituye la más empleada en estadística.

La mayoría de las variables continuas que se presentan en la práctica diaria (peso, talla, colesterol sanguíneo, presión arterial, etc.), pueden ser modeladas mediante esta distribución³, a pesar de que ellas no siguen exactamente una distribución normal. Además, la ley normal tiene gran interés en estadística inferencial, pues sirve como modelo para describir la distribución muestral de las medias, como veremos en los temas siguientes. También, muchos de los estadígrafos empleados en los contrastes de hipótesis, que trataremos en temas posteriores, siguen esta distribución, a la vez

³ Streiner y Norman (1998) plantean que las distribuciones estrictamente normales son tan raras en la naturaleza como los dientes de gallina.

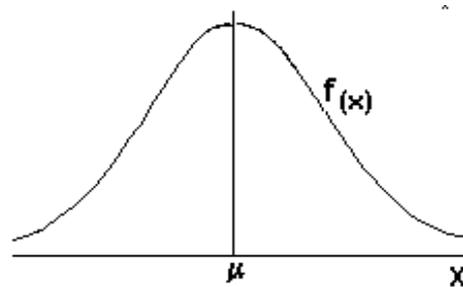
que muchos de los tests estadísticos que veremos, dan por supuesto que los datos provienen de una distribución normal.

La distribución normal está definida como la función de distribución de probabilidad de una variable X , cuya representación gráfica se muestra en la figura 2.6.1, y se denota por:

$$X \sim N(\mu, \sigma^2)$$

Donde:

μ es la media o valor esperado



σ^2 es la varianza

Figura 1.6.1 – Representación gráfica de la función de densidad de la distribución Normal.

Como puedes ver en la figura 1.6.1, la curva normal tiene forma de campana y como toda función de densidad de probabilidad el área bajo la curva es igual a 1. De aquí podemos deducir las características fundamentales de esta distribución:

1. Es simétrica respecto a la media
2. Tiene su punto máximo en la media
3. La media, la mediana y la moda son iguales entre sí
4. El área total bajo la curva es igual a 1 y como es simétrica, el 50% del área está a la derecha e izquierda de la media.
5. Siempre es positiva
6. Las colas de la curva están cada vez más próximas al eje x a medida que nos alejamos de la media, pero nunca lo alcanzan.
7. Si se levantan dos perpendiculares a una distancia de una desviación estándar desde la media hacia ambos lados, el área delimitada por la curva, el eje x y las perpendiculares

será aproximadamente del 68 %. Si hacemos lo mismo pero con dos desviaciones estándar, el área será aproximadamente del 95 %.

8. Los parámetros μ (mu) y σ (sigma), determinan completamente la forma de la distribución. Para cada valor de μ y σ habrá una distribución normal diferente. Así vemos en la figura 1.6.2 que diferentes valores de μ , trasladan la curva horizontalmente sobre el eje X, mientras que σ , determina el grado de apuntamiento de la curva (a menor desviación estándar, menos dispersión alrededor de la media y por lo tanto más alta la curva) figura 1.6.3.

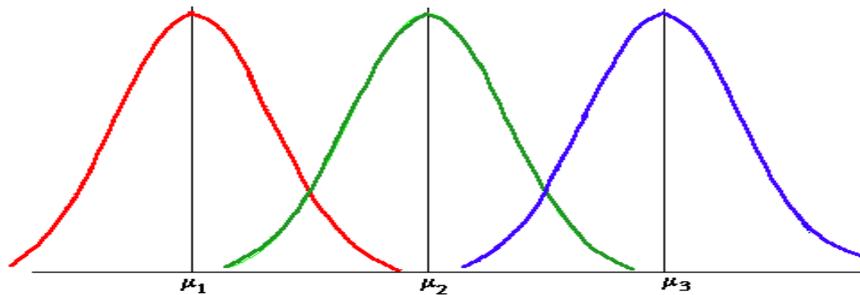


Figura 1.6.2 – Tres distribuciones normales con diferentes medias pero igual variabilidad.

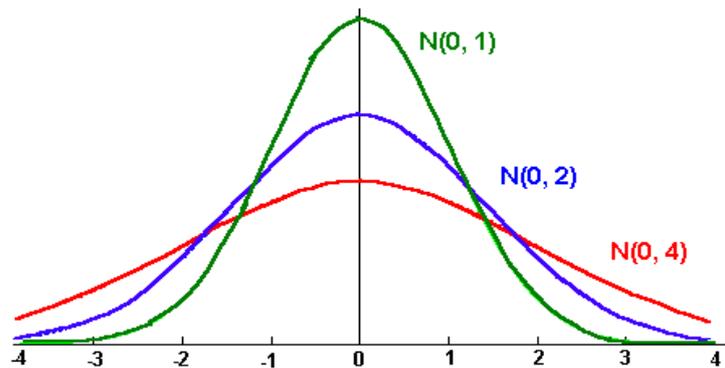


Figura 1.6.3 - Tres distribuciones normales con diferentes desviaciones estándar y con la misma media.

Para calcular áreas bajo la curva normal y así calcular las probabilidades de que la variable X tome valores en un intervalo dado, se hace necesario el uso de integrales. Pero no te asustes,

que en la práctica esto se hace fácilmente utilizando tablas estadísticas.

Como podrás suponer, no existen tablas para cada una de las posibles distribuciones que se pueden presentar en la práctica, la solución a esta problemática es la existencia de una distribución normal especial, llamada Normal estándar o típica, que se caracteriza por tener media cero y varianza uno, figura 1.6.4. Ésta se denota por:

$$X \sim N(0, 1)$$

Los valores de probabilidad para esta distribución se encuentran tabulados, y a partir de ella se pueden obtener los valores correspondientes a cualquier otra distribución normal, mediante la estandarización. La tabla A del Apéndice, es una muestra de estas tablas.

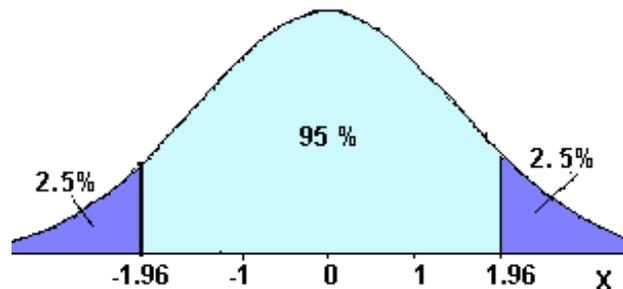


Figura 1.6.4- Áreas bajo la curva normal estándar.

Así, cualquier variable aleatoria que se distribuya aproximadamente normal, puede ser estandarizada y transformada en una normal estándar, mediante la siguiente fórmula:

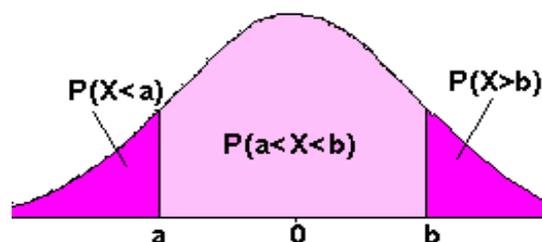
$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

Entonces, a cada valor de X se le hace corresponder un valor de la distribución normal estándar cuyas probabilidades son equivalentes, ya que:

$$X \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

A continuación, veremos varios ejemplos de cálculo de probabilidades utilizando la tabla de valores de la normal estándar, pero antes debes recordar, que en la distribución normal, como en toda distribución continua, la probabilidad para un valor específico de X es cero, lo que se halla es la probabilidad para intervalos.

Así tenemos, que se pueden calcular las probabilidades de que la variable X tome valores en un intervalo (a, b) , $P(a \leq X \leq b)$, o



para cualquier otro intervalo, por ejemplo $P(X \leq a)$, $P(X \leq b)$, $P(X \geq a)$, etc. Algunas de las cuales se encuentran representadas en la figura 1.6.5.

Figura 1.6.5 – Subdivisiones del área bajo la curva normal.

Ilustremos el uso de la tabla A del Apéndice. En esta tabla se muestra el área a la derecha de cualquier valor que tome la variable Z (normal estándar), como se representa en la figura adjunta a la tabla. Vemos que la columna de la izquierda, se corresponde con los valores de z , mientras que en la derecha, encontramos los valores correspondientes al área bajo la curva, a la derecha de z . Fíjate que te damos los valores de z solo con un decimal, pero debes saber que existen tablas mucho más explícitas.

Veamos, si queremos conocer el área a la derecha de $z = 0.1$, miramos en la columna derecha que le corresponde un valor de 0.4602, este se interpreta como la probabilidad de que la variable Z tome valores mayores o iguales que 0.1, $P(Z \geq 0.1)$.

Ahora, si quisieras saber ¿cuál es la probabilidad de valores menores o iguales a 0.1, $P(Z \leq 0.1)$?, que problema, la tabla no tiene esa información. Ah! Sabes como resolvemos esto, muy simple, recuerda que esta es una distribución simétrica, además el área bajo la curva es igual a 1, entonces:

$$P(Z \leq 0.1) = 1 - P(Z \geq 0.1)$$

$$P(Z \leq 0.1) = 1 - 0.4602 = 0.5398$$

Otro problema, la tabla solo me da los valores para $Z \geq 0$, ¿cómo calcular el área a la derecha de un valor de Z negativo?. Nada complicado, ya sabes que se trata de una distribución simétrica y que a ambos lados de la media se encuentra el 50 % del área, por lo tanto:

$$P(Z \leq -z) = P(Z \geq z)$$

Entonces $P(Z \leq -0.1) = 0.4602$

Mira que curioso, la probabilidad de Z menor que -0.1 es igual a la probabilidad de Z mayor que 0.1. esto lo verás mejor en la figura 1.6.6

Ahora, la probabilidad de Z mayor que -0.1 es igual a la probabilidad de Z menor que 0.1. Parece un trabalenguas, pero si te fijas en las figuras podrás entender perfectamente.

$$P(Z \geq -0.1) = P(Z \leq 0.1)$$

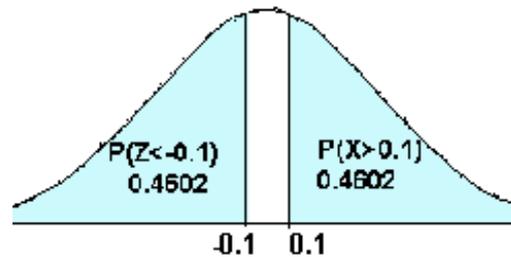


Figura 1.6.6 Distribución normal estándar

A modo de resumen te doy algunas reglas básicas para esta tabla⁴.

$$\begin{aligned}
 P(Z \geq z) &= \text{valor tabulado} \\
 P(Z \leq z) &= 1 - P(Z \geq z) \\
 P(Z \geq -z) &= P(Z \leq z) \\
 P(Z \leq -z) &= P(Z \geq z)
 \end{aligned}$$

Estas solo son algunas reglas que te brindamos, pero existen muchas otras formas de obtener los mismos resultados, debes usar tu sentido común e improvisar. Juguemos un poco con la tabla para que te entrenes en su uso:

- Probabilidad de que Z sea menor que 1: Es el área a la izquierda de 1, se denota como:

$$P(Z \leq 1) = 1 - P(Z \geq 1) = 1 - 0.1587 = 0.8413.$$

Se representa en la figura 2.6.7

- Probabilidad de Z mayor que 2 pero menor que 3:

$$P(2 \leq Z \leq 3) = P(Z \geq 2) - P(Z \geq 3)$$

Es decir, necesitamos el área entre la curva, el eje x y dos perpendiculares en los valores 2 y 3. La tabla me da el área total a la derecha de $z = 2$, entonces le resto el "pedacito" correspondiente al área a la derecha de $z = 3$ y obtengo el área entre 2 y 3. Figura 1.6.8

$$P(2 \leq Z \leq 3) = 0.02275 - 0.00135 = 0.0214$$

- Probabilidad de Z mayor que -1 y menor que 1.5:

$$\begin{aligned}
 P(-1 \leq Z \leq 1.5) &= P(Z \geq -1) - P(Z \geq 1.5) \\
 &= 1 - P(Z \geq 1) - P(Z \geq 1.5) \\
 &= 1 - 0.1587 - 0.0668 = 0.7745
 \end{aligned}$$

Ver figura 1.6.9

⁴ Si utilizas otra tabla puede ser diferente, esto depende del área que esté tabulada, si la izquierda o la derecha. A la vez que dominas una, entiendes cualquiera.

Pienso que ya estás listo para calcular probabilidades de otras distribuciones normales, es decir, estandarizando las variables.

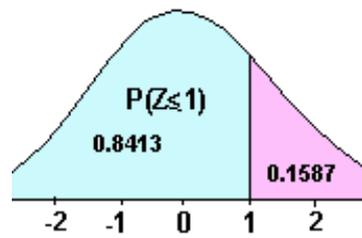


Figura 1.6.7 Probabilidades de $(Z \geq 1)$ y $(Z \leq 1)$

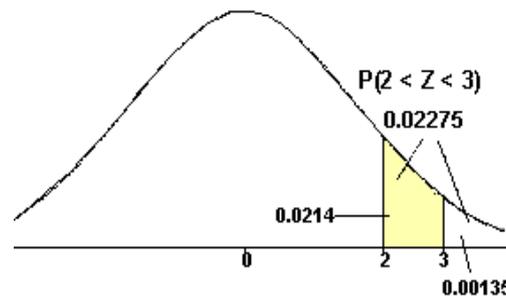


Figura 1.6.8 – Curva normal estándar. Probabilidad de $(2 \leq Z \leq 3)$

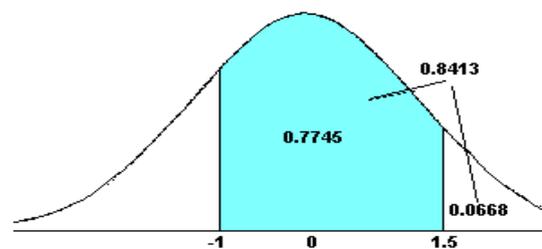


Figura 2.6.9 – Curva normal estándar. $P(-1 \leq Z \leq 1.5)$

Ejemplo 1.6.1

Se conoce que las cifras de glucosa en sangre, tienen una distribución aproximadamente normal, con media 100 (mg/100 mL) y desviación estándar 40. ¿Cuál es la probabilidad de que al seleccionar una persona de dicha población tenga su nivel de glucosa por encima de 120?

Primero, la variable X se define como el nivel de glucosa en sangre de las personas. Ésta se distribuye normal:

$$X \sim N(100, 1600)$$

Nos piden la probabilidad de X mayor que 120: $P(X \geq 120)$.
Tenemos que estandarizar la variable:

$$Z = (120 - 100) / 40 = 20/40 = 0.5$$

Entonces la $P(X \geq 120) = P(Z \geq 0.5)$ y por lo tanto:

$$P(X \geq 120) = 0.3085$$

R/ La probabilidad de que al seleccionar una persona de dicha población su nivel de glicemia sea mayor que 120 es del 30.85%. Figura 1.6.10

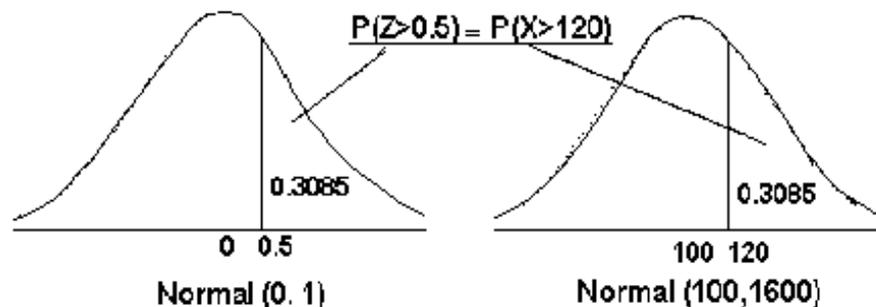


Figura 2.6.10 – Equivalencia entre el área de la curva normal (0, 1) y la normal (100,1600)

Aunque la importancia de esta distribución en el campo de la estadística es indiscutible, podrás darte cuenta que no es una ley inherente a todas las variables continuas que conoces. A pesar de que muchas de ellas tienen una distribución aproximadamente normal, no es posible encontrar en la naturaleza una distribución exactamente normal, no obstante al utilizar la ley normal para modelar estas variables, es posible establecer afirmaciones de probabilidad más útiles y convenientes que si empleáramos una distribución más compleja.

1.6.1 Relación entre la distribución Normal, Binomial y de Poisson.

Cuando en una distribución Binomial n es grande y p y q no están muy cercanas a cero, ésta puede aproximarse a la distribución normal, lo que facilitaría enormemente los cálculos de

probabilidades asociadas. Si miras la figura 2.3.2, verás que si unimos los extremos superiores de cada barra vertical, obtenemos una curva parecida a la normal.

Entonces, en la medida que n aumenta y p se mantiene constante, la aproximación será cada vez mejor. De forma general se plantea que hay buena aproximación cuando n mayor o igual que 30. Si p está cercana a 0.5, se considera bueno un tamaño muestral incluso mayor que 10, en la medida que p se aleja de 0.5, se necesitarán mayores tamaños muestrales.

Ahora bien, sabemos que la distribución de Poisson es una aproximación a la Binomial cuando n es grande y p es pequeña, por lo tanto también puede ser aproximada a la normal.

1.7 Otras distribuciones de probabilidad.

Además de las distribuciones ya estudiadas, existen otras de amplio uso en estadística inferencial, entre las que se destacan las famosas t de Student y la Ji cuadrado (χ^2), muy usadas ambas en la realización de pruebas de hipótesis y en la construcción de intervalos de confianza, como veremos más adelante.

A continuación abordaremos, brevemente, las características principales de cada una de ellas.

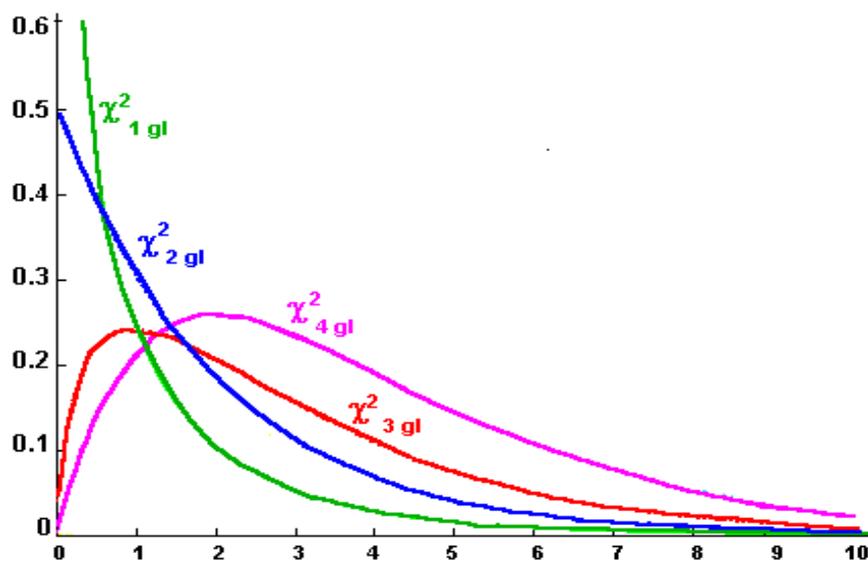
1.7.1 La distribución Ji cuadrado (χ^2)

Esta distribución deriva de la distribución normal. Surge cuando tenemos n variables independientes, que siguen aproximadamente una distribución normal estándar (X_1, X_2, \dots, X_n). La variable resultante de sumar los cuadrados de éstas, sigue una distribución Ji cuadrado con n grados de libertad. Se denota como.

$$X \sim \chi^2 (n \text{ gl})$$

Su representación gráfica puedes verla en la figura 2.7.1, para diferentes grados de libertad.

No necesariamente deben ser distribuciones normal estándar,



también pueden ser variables estandarizadas.

Figura 1.7.1 Representación gráfica de la función de densidad de la distribución Ji cuadrado para diferentes grados de libertad.

Se trata de una función positiva, no puede tomar valores negativos, pues como podrás suponer, se deriva de la suma de valores elevados al cuadrado. Su forma depende de los grados de libertad, con tendencia a aproximarse a la curva normal en la medida que aumentan los grados de libertad. Al igual que la normal y cualquier distribución continua, el área bajo la curva es igual a 1, pero no es simétrica.

Su valor esperado y varianza son:

$$E(\chi^2) = n \quad \text{y} \quad V(\chi^2) = 2n$$

Los grados de libertad se refieren, al número de términos independientes que es necesario, para obtener el valor de la variable χ^2 .

Una propiedad importante, es que la Ji cuadrado con un grado de libertad, es igual al cuadrado de la normal estándar.

Esta distribución es muy empleada para probar hipótesis, cuando los datos están en forma de frecuencias. Estos procedimientos se estudian bajo el título de pruebas de bondad de ajuste, como verás en cursos posteriores. Siendo más adecuado su uso, cuando estamos en presencia de variables categóricas.

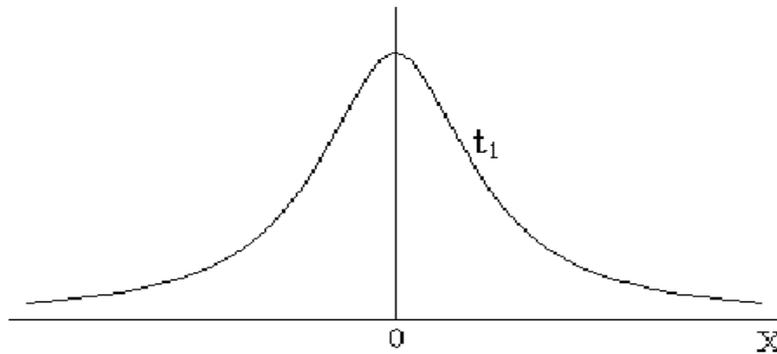
La cantidad χ^2 , es una medida del grado de congruencia entre las frecuencias observadas y las esperadas, bajo el cumplimiento de una hipótesis dada, esto lo entenderás mejor cuando veamos la aplicación concreta de dicha distribución.

Por ahora, baste que conozcas que existen tablas con los valores de los percentiles de esta distribución, para diferentes magnitudes de n . La tabla C del Apéndice es un ejemplo de éstas. Como puedes ver están representados los valores de algunos percentiles⁵ de χ^2 para diferentes grados de libertad. Por ejemplo, el percentil 95 de la χ^2 con 12 gl = 21.0

1.7.2 La distribución t de Student.

⁵ Los más frecuentemente usados para pruebas de hipótesis.

Este modelo teórico, al igual que la Ji cuadrado, se deriva del modelo normal, pero es un proceso algo complejo. Se denota por la



letra t minúscula, y su forma gráfica se representa en la figura 2.7.2.

Figura 1.7.2 – Función de densidad de la distribución t de Student con un grado de libertad.

Como puedes ver es muy similar a la curva normal estándar, pero más dispersa, su media también es cero. Es simétrica respecto a la media, su forma depende de los grados de libertad, así en la medida que aumentan los grados de libertad mayor es la aproximación a la normal. Los grados de libertad tienen el mismo significado que en la Ji cuadrado.

Puede tomar valores tanto positivos como negativos, y al igual que toda distribución continua, el área bajo la curva es igual a uno.

Su valor esperado y varianza:

$$E(t) = 0 \quad \text{y} \quad V(t) = \frac{n}{n-2}$$

Al igual que la Ji cuadrado existen tablas con los percentiles de ella, los que son empleados en las pruebas de hipótesis y en la construcción de intervalos de confianza. La tabla B del Apéndice es un ejemplo de estas tablas, en ella se muestran los valores de la distribución t para diferentes grados de libertad, correspondientes a los percentiles más empleados en inferencia estadística, veremos su uso en la medida que sea necesario y así lo entenderás mejor.

Resumen

En este tema estudiaste que:

1. La probabilidad es una medida numérica, del grado de certidumbre real de ocurrencia de un fenómeno. Puede ser estimada mediante la frecuencia relativa de ese evento o fenómeno.

2. El conocimiento acerca de las probabilidades de los procesos y fenómenos de la vida práctica, nos permite hacer inferencias y predicciones futuras, sobre el comportamiento de dichos fenómenos, lo que a su vez, es útil para tomar medidas de prevención y protección específica.
3. Una variable aleatoria es la correspondencia que se establece, entre cada uno de los resultados posibles de un experimento aleatorio y un número real, lo que nos permite realizar operaciones aritméticas con dichos resultados, facilitando así el manejo de los datos.
4. El conjunto de todos los valores de una variable aleatoria discreta, junto a sus correspondientes valores de probabilidad, constituye la distribución de probabilidades de dicha variable. Puede ser representada tabular y (o) gráficamente y la suma de todas sus probabilidades es igual a 1.
5. Las distribuciones de probabilidad de las variables continuas, suelen ser complejas, se les llama función de densidad de probabilidades y su representación gráfica se asemeja al polígono de frecuencias relativas de las variables continuas. El área bajo la curva es igual a 1, siendo el valor de probabilidad para un valor específico de la variable igual a cero.
6. La distribución discreta más empleada en estadística, es la Binomial. Esta se caracteriza, por ser un proceso equivalente a la repetición de n experimentos Bernoulli, independientes, cada uno con solo dos resultados posibles: éxito y fracaso, donde la probabilidad del éxito se mantiene constante de una repetición a otra. Mediante su uso podemos calcular el número de éxitos esperados en n repeticiones, así como las probabilidades de obtener cualquier número k de éxitos.
7. La distribución normal constituye el modelo teórico más ampliamente utilizado en estadística. Es una función continua, mediante la cual pueden ser modeladas la mayoría de las variables continuas usadas en la práctica diaria. También es muy usada para realizar pruebas de hipótesis y construir intervalos de confianza.
8. Existen valores tabulados de la distribución normal estándar (media cero y varianza uno), que nos permiten realizar adecuadamente el cálculo aproximado de las probabilidades de cualquier otra distribución normal, mediante la estandarización.
9. Hay otras distribuciones teóricas de uso en estadística, entre las que se destacan las famosas Ji cuadrado y t de Student. Ambas pueden deducirse de la normal y son empleadas en inferencia

estadística en la realización de pruebas de hipótesis y la construcción de intervalos de confianza.

Ejercitación

1- Responde Verdadero (V) o falso (F), las siguientes afirmaciones:

- La probabilidad clásica puede ser calculada solamente después de observar un número elevado de repeticiones del experimento.
- Un experimento que realizado bajo las mismas circunstancias, se obtiene en cada repetición un resultado diferente, se le denomina experimento aleatorio.
- La definición estadística de probabilidad, considera a la frecuencia relativa de un evento, como una estimación de la probabilidad de ocurrencia de dicho evento.
- Masculino y femenino, constituyen los sucesos elementales del experimento aleatorio consistente, en observar el sexo de un recién nacido.

2- De las variables aleatorias listadas a continuación, identifique que distribución de probabilidad siguen aproximadamente.

- a) Número de pacientes diabéticos de un área de salud.
- b) Talla de los niños preescolares de una escuela primaria.
- c) Número de personas que sufren un accidente automovilístico, diariamente en una ciudad.
- d) Valores de colesterol de los adultos hipertensos, de un área de salud.

3- Cierta enfermedad tiene una probabilidad muy baja de ocurrir, $p=1/100.000$, en una población de 500 000 habitantes.

- a) Diga, ¿Qué distribución de probabilidad sigue esta enfermedad en dicha población?
- b) ¿Con qué otra distribución puede ser aproximada? Justifique.
- c) Calcule el valor esperado de la variable para ambas distribuciones.

4- El peso en libras de los niños recién nacidos, hijos de madres diabéticas, se distribuye aproximadamente normal, con media 9.5 y desviación estándar 1.

- a) Calcule la probabilidad de que el peso de uno de esos RN sea mayor que 12.
- b) ¿Qué probabilidad existe de que nazca un niño con peso inferior a 8 libras?

- c) ¿Cuál será el peso aproximado del 95 % de los niños?
- d) ¿Cuántos de cada 100 niños RN de madres diabéticas, tendrán un peso menor que 12 libras?

Autoevaluación

- 1- Explica con tus palabras, ¿Qué entiendes por variable aleatoria?. Da tres ejemplos de aplicación en el contexto de la Atención Primaria de Salud. Clasifícalas.
- 2- El número de fumadores en tú área de salud es una variable aleatoria. Diga, qué tipo de v.a y qué distribución sigue aproximadamente?.
- 3- Ofrece un ejemplo de variable que se distribuya aproximadamente Poisson.
- 4- El coeficiente intelectual (CI) de los individuos admitidos en la escuela para retrasados mentales, tiene una distribución aproximadamente normal con media 60 y desviación estándar 10.
 - a) Calcula la cantidad de individuos con CI mayor que 75, si sabes que la matrícula es de 250.
 - b) ¿Cuál es la probabilidad de que un individuo, elegido al azar, tenga un CI entre 55 y 75?
 - c) Calcula $P(50 \leq X \leq 70)$.
- 5- Mencione dos aplicaciones de las distribuciones Ji cuadrado y t de Student.

Bibliografía

1. Armitage P, Berry G. **Statistical Methods in Medical Research**. 3rd ed. Oxford: Blackwell Scientific Publications, 1994
2. Altman DG. **Practical statistics for medical research**. London: Chapman and Hall. 1992
3. Daniel WW. **Bioestadística. Base para el análisis de las ciencias de la salud**. 3ra edición. México. D. F: Limusa; 1997
4. Norman GR, Streiner DL. **Bioestadística**. España: Hartcourt Brace; 1998
5. Spiegel MR. **Teoría y problemas de Estadística**. La Habana: Pueblo y educación; 1977
6. Doménech Massons, JM. **Métodos estadísticos en ciencias de la salud**. Unidad didáctica 3. Barcelona: Signo; 1995

7. Swinscow, TVD. **Statistics at square one**. 9th edition. BMJ publishing group; 1997
8. Dawson-Saunders B, Trapp R G. **Bioestadística**. 2da ed. El manual moderno. México; 1999.
9. Mtnez canalejo H. **Principios Básicos de la Teoría de Probabilidades**. Ed. Pueblo y Educación. La Habana; 1989
10. Armitage P, Berry G. **Estadística para la Investigación Biomédica**. Doyma, Barcelona, 1992.
11. Hamilton LC. **Modern Data Analysis**. Brooks/Cole Publishing Company, Pacific Grove, 1990.
12. Martín Andrés A, Luna Del Castillo JD. **Bioestadística para las Ciencias de la salud**. Norma, Granada, 1994.
13. Marascuilo LA, Serlin RC. **Statistical Methods for the Social and Behavioral Sciences**. W.H. Freeman and Company, Nueva York, 1988.
14. Versión electrónica del manual de la Universidad de Málaga. **Bioestadística: métodos y aplicaciones**. Málaga; 1998.

Tema 2. Nociones de muestreo.

Introducción

Uno de los problemas que debes enfrentar como médico, es la necesidad de conocer el comportamiento de una enfermedad, de un factor de riesgo o de cualquier otra variable, en los elementos de un determinado colectivo. Para ello, tendrías que examinar cada uno de estos elementos y medir el conjunto de variables que sean de tú interés.

Dicho así parece simple, pero a menudo sucede que el número de sujetos a observar es muy grande y por tanto no es posible examinar a cada uno de ellos, ya sea por cuestiones de tiempo y (o) de recursos. En estos casos lo que se hace comúnmente es estudiar una parte de dicho colectivo y extender luego los resultados de ese examen al conjunto de donde proceden los elementos ya observados y analizados.

De aquí surgen las siguientes interrogantes: ¿Qué parte de ese colectivo debe analizarse?, ¿Cuántos elementos de ese colectivo deben ser examinados?, ¿De qué forma serán seleccionados los elementos a observar?, ¿Qué aspectos debemos tener en cuenta, al realizar la selección de los sujetos, para reducir al máximo los errores al realizar las conclusiones?.

Las respuestas a estas preguntas contienen dentro de sí el origen de la llamada teoría del muestreo y sus diferentes métodos.

En este tema te presentamos una serie de definiciones, procedimientos y técnicas sobre muestreo, que podrás aplicar a problemas de investigación que se te presenten en tú trabajo diario.

2.1 Muestreo. Conceptos básicos.

El muestreo constituye una herramienta de la investigación científica. Su función básica es determinar qué parte de una realidad en estudio debe ser examinada con el objetivo de hacer inferencias (conclusiones) sobre el todo de la cual procede, siendo esa parte de la realidad la **muestra** y el todo de donde procede la **población** o **universo**.

Debes distinguir entre la población que deseas estudiar, llamada **población objeto u objetivo** y la población realmente estudiada: la **población muestreada**. Lo deseable es que ambas coincidan, pero esto no es posible en todas las circunstancias, ya que no siempre se tiene acceso a la población que se desea estudiar. Por consiguiente, la población objetivo debes definirla de forma clara,

sin ambigüedades, de manera que no existan dudas al decidir si un individuo pertenece o no al universo bajo consideración, lo que facilitará en gran medida la realización de la inferencia estadística.

En muchas ocasiones te encontrarás en una situación tal que no tienes acceso a una población de la cual extraer una muestra, sino que deberás trabajar con los datos que has podido obtener, en estos casos las inferencias recaerán sobre aquella población de la que se supone que la muestra es representativa.

Veamos a continuación algunos conceptos básicos que te ayudarán a comprender el contenido del tema.

Unidades de análisis: se llama así a los elementos de la población objeto de estudio. Pueden ser de variada naturaleza en dependencia de los objetivos y del fenómeno estudiado. Por ejemplo, personas, animales de laboratorio, muestras de organismos (tejidos, secreciones, fluidos corporales, etc.) e incluso áreas o comunidades.

Unidades de muestreo: son las partes en que se puede dividir la población objeto de estudio antes de seleccionar la muestra, las cuales deben abarcar toda la población sin intersectarse, es decir, cada unidad de análisis pertenece a una y sólo una unidad de muestreo. A su vez, una unidad de muestreo puede contener un conjunto de unidades de análisis o, incluso, un conjunto de unidades de muestreo correspondientes a una etapa posterior de selección. Por ejemplo, podemos dividir la población según áreas de salud y a la vez cada área estará dividida en grupos básicos de trabajo o directamente en consultorios.

Marco muestral: no es más que la lista de todas las unidades de muestreo. En el ejemplo anterior, el marco muestral es el listado de todas las áreas de salud que existan en la población objeto de estudio.

Problema de muestreo: se llama así a la situación donde se quiere conocer una característica general o parámetro de una población, por ejemplo, el porcentaje de discapacitados en una comunidad, y en lugar de examinar a todas y cada una de las unidades de análisis se decide medir solo a una parte de ellas y estimar el número desconocido a partir de dicha información.

Según Silva (1993), resolver este problema implica:

1. Delimitar el número de unidades de análisis a seleccionar (tamaño muestral)
2. Establecer la forma en que se efectuará la selección (método de muestreo a emplear)

3. Determinar el modo en que se procesarán los datos para realizar la estimación. (análisis)
4. Dar el procedimiento de cálculo del error que se comete en el proceso de estimación.

Estas cuatro etapas se condicionan mutuamente, como verás más adelante cuando abordemos cada tipo de muestreo por separado.

Error de muestreo o aleatorio: es el error que se comete debido al hecho de sacar conclusiones sobre una población a partir del estudio de una muestra de ella. Podrás suponer, que la magnitud del error aleatorio será mayor en la medida que el tamaño de la muestra estudiada sea menor. Como investigador, es obvio, que desearás que ese error sea pequeño y a la vez poder cuantificar su magnitud.

Probabilidad de inclusión: como su nombre lo indica, es la probabilidad que tiene un elemento de la población objeto de estudio de ser incluido en la muestra a observar. Se puede calcular de diferentes formas, en dependencia del tipo particular de muestreo empleado.

Método probabilístico de muestreo: es aquel que otorga una probabilidad conocida, no nula⁶, de integrar la muestra a cada una de las unidades de análisis de la población objeto de estudio. Cuando la probabilidad de inclusión es la misma para todos los elementos de la población objeto, se dice que el método es **equiprobabilístico**.

Muestra probabilística: es aquella generada a partir del uso de un método probabilístico de muestreo. De aquí que, la muestra será **equiprobabilística** si el método empleado también lo es.

El empleo de procedimientos probabilísticos de muestreo nos permite cuantificar la magnitud del error aleatorio, es decir, medir la precisión con que se realizan las estimaciones.

Muestra representativa: no existe una definición formal que nos permita afirmar que una muestra es o no representativa de la población objeto de estudio, a pesar de que la mayoría de los investigadores se refieren a este término en sus investigaciones. Según Silva (1993), para conseguir representatividad lo que debe procurarse es que la muestra exhiba internamente el mismo grado de variabilidad que la población, así, una muestra puede considerarse representativa de **ciertos aspectos específicos** de la población, cuando el error en que se incurre al sacar conclusiones sobre esos aspectos no excede ciertos límites prefijados.

⁶ Es decir, una probabilidad diferente de cero.

En esto juega un papel esencial el azar, el cual no necesariamente provee de representatividad a la muestra obtenida, pero asegura la imparcialidad en la conducta del investigador. Por esta razón, la confianza que pueda depositarse en una muestra depende vitalmente de la que merezca el procedimiento que la produjo.

No obstante, debo aclarar que en muchas ocasiones no es posible emplear un procedimiento probabilístico de muestreo, en estos casos se hace necesario recurrir a otros métodos no probabilísticos. Los cuales siempre y cuando se utilicen de manera adecuada (excepto la selección sin método alguno), de acuerdo a los objetivos del estudio y teniendo en cuenta este inconveniente a la hora de interpretar los resultados, pueden resolver problemas de investigación y nos permite acercarnos al cumplimiento final del objetivo de interés.

A continuación, vamos a desarrollar someramente los tipos principales de muestreo, que podrás aplicar en los problemas de investigación que se te presenten en la práctica diaria.

2.2 – Muestreo simple aleatorio (MSA)

El muestreo simple aleatorio constituye el más sencillo y conocido procedimiento probabilístico de muestreo. Es ampliamente utilizado en el diseño experimental, además de ser un procedimiento básico componente de otros métodos más complejos que abordaremos más adelante.



Puede definirse al **Muestreo Simple Aleatorio** como un procedimiento mediante el cual las unidades de análisis a integrar la muestra son seleccionadas de manera equiprobabilística, además, todos los subconjuntos de tamaño n susceptibles de ser formados a partir de la población objeto tendrán la misma probabilidad de selección.

Aclaremos un poco esta definición. Supongamos que se quiere seleccionar una muestra de tamaño n , de una población de N unidades. Es simple imaginar que existe un número determinado de muestras posibles, es decir, que existen diferentes subconjuntos de tamaño n de esa población que pudieran ser seleccionados. Con este proceder, todos los subconjuntos posibles de tamaño n de esa población tendrán la misma probabilidad de ser seleccionados. Esta constituye la característica distintiva de este método.

Ejemplo 2.2.1

Veamos este ejemplo, tomado de Silva (1993), ello viabilizará la comprensión de lo planteado con anterioridad.

Supongamos que se tienen $N = 10$ niños ordenados alfabéticamente en una lista y que se desea seleccionar una muestra de ellos de tamaño $n = 5$. Consideremos los siguientes procedimientos de selección.

- a) Tomar los 5 primeros de la lista
- b) Dividir la lista en dos grupos: los 5 primeros y los 5 restantes. Luego lanzar una moneda y si sale cara se toma el primer grupo y a los del segundo en caso contrario.
- c) Numerar los 10 elementos. Poner en una caja 10 papелitos numerados sucesivamente. Seleccionar ciegamente 5 papeles y por último admitir en la muestra a los niños cuyos números aparezcan en los papeles seleccionados.

El método a) no es probabilístico: los niños del 6 al 10 de la lista nunca serán admitidos en la muestra, es decir, su probabilidad de inclusión es cero.

El método b), en cambio, es probabilístico. En efecto la probabilidad de inclusión de cada niño es la misma que tiene el grupo al cual pertenece; ésta es obviamente igual a 0.5, la probabilidad de que salga cara o escudo. Sin embargo, no se trata de MSA ya que existen subconjuntos de 5 elementos que no pueden resultar elegidos como consecuencia de la aplicación de este procedimiento: el subconjunto formado por los niños que ocupan los lugares impares en la lista o los niños que ocupan los números 1, 2, 3, 4 y 6, estos y otros muchos tienen probabilidad cero de ser seleccionados.

¿Comprendes ahora a qué me refiero?. Espero que sí.

Por último, el método c) es un MSA ya que cualquiera de los 252 subconjuntos diferentes de tamaño 5 que pueden ser formados a partir de 10 niños, puede resultar seleccionado y, dado el modo de selección, ninguno de dichos subconjuntos tiene más probabilidad que otro de ser el que a la postre resulte elegido. Además todos los niños tienen la misma probabilidad de ser incluidos en la muestra y es igual a:

$$f = \frac{n}{N} = \frac{5}{10} = 0.5$$

Te estarás preguntando de dónde salió ese 252, pues te diré: de la teoría combinatoria elemental, se sabe que el número de

subconjuntos diferentes de n elementos susceptibles de ser formados de un conjunto de tamaño N es igual a:

$$k = \binom{N}{n} = \frac{N!}{n! (N-n)!} \quad 7$$

Donde k es el número de muestras posibles y la probabilidad de selección de cada una de estas muestras en el MSA es igual a:

$$\frac{1}{k}$$

En el ejemplo anterior sería así:

$$k = \binom{10}{5} = \frac{10!}{5! (10-5)!} = \frac{10*9*8*7*6*5*4*3*2*1}{(6*4*3*2*1)(6*4*3*2*1)} = 252$$

Te parece muy engorroso, pero con el uso de una calculadora científica se limita a oprimir un par de teclas. Aclarado este punto podemos continuar.

La forma de seleccionar los sujetos que formarán la muestra puede ser muy variada, pero es necesario poseer una lista de todas las unidades de análisis de la población objeto, lo cual en la práctica no siempre es posible y de contarse con esta información sería realmente trabajoso aplicar dicho mecanismo. Para facilitar este proceso se crearon las tablas de números aleatorios, que en la actualidad están en desuso, ya que con una calculadora o mejor con el uso de una computadora, pueden generarse números aleatorios muy fácilmente. De todas formas, por la imposibilidad frecuente de contar con listas de las poblaciones a estudiar, este método de muestreo es muy poco usado de manera pura.

Como se ha dicho, el objetivo del muestreo es estimar parámetros poblacionales, entre los que se encuentran la media y desviación estándar, para variables continuas, y la proporción o porcentaje en caso de variables categóricas. Veamos a continuación cómo se estiman estos parámetros después de haber empleado el MSA para seleccionar la muestra.

2.2.1 Estimación de la media.

⁷ Por si no lo sabes o no recuerdas, el factorial de un número ($N!$), es igual a la productoria de todos los números desde cero hasta N , considerándose convencionalmente el factorial de cero = 1, sino todos fueran = 0. Ejemplo: $3! = 3*2*1 = 6$

El promedio poblacional de cierta variable X , puede calcularse si medimos esa variable en cada uno de los N elementos de dicha población y aplicamos la fórmula ya estudiada en el curso anterior.

$$\bar{X} = \frac{\sum_{i=1}^N X_i}{N}$$

Donde X_i es la observación de X para el i ésimo elemento de la población.

La media poblacional suele denotarse por la letra griega (mu) μ . Si en lugar de estudiar todos los elementos de la población objeto, seleccionamos una muestra de tamaño n de la misma y medimos la variable X para cada elemento de la muestra, entonces podemos calcular la media muestral mediante el empleo de la misma fórmula:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

La media muestral, así obtenida, puede considerarse una variable aleatoria, puesto que cada muestra posible de tamaño n que se seleccione de esa población, tendrá un valor para la media. Esto puede resultarte extraño y nuevo, aclaremos un poco.

Retomando el ejemplo 3.2.1, supongamos que seleccionamos las 252 muestras posibles de tamaño 5 y en cada caso medimos las tallas de los niños y calculamos la media; al final de ese proceder tendremos 252 valores para la talla promedio de los 10 niños, muchos de estos valores coincidirán exactamente con la media del grupo pero otras estarán más o menos alejadas. ¿Cómo saber cuál de estos valores es la verdadera media del grupo?. En este caso es fácil, basta con medir a los 10 niños y calcular la media, pero esto es solo un ejemplo hipotético, en otros casos en que no es posible examinar al total de la población ¿Qué hacer?.

Está demostrado que el promedio de todas esas medias coincide exactamente con la media poblacional. Por lo tanto, la media muestral puede ser considerada como un estimador insesgado⁸ de la media poblacional, siempre que la muestra sea equiprobabilística. Así que, cuando queremos estimar la media poblacional podemos utilizar a la media obtenida en la muestra estudiada, como un buen estimador de ella. Por supuesto, este proceder está sujeto a error, pues ¿Cómo saber si la media muestral obtenida es de las más o menos alejada de la verdadera media poblacional?. A ciencia cierta,

⁸ Un estimador es insesgado (carente de error sistemático), cuando el valor esperado de su distribución, bajo el supuesto de muestreo sucesivo de tamaño n de una población de N elementos, coincide exactamente con el parámetro.

esto nunca lo sabremos, a menos que examinemos a todos los elementos de la población, lo cual no tiene ningún sentido práctico. Entonces ¿Qué hacer?. Simple, podemos medir la magnitud del error con que nuestra muestra estima al parámetro, y así conoceremos cuán precisa es nuestra estimación. Esto lo hacemos a través del cálculo de la varianza del estimador, es decir, la varianza de la media muestral respecto a la media poblacional, $V(\bar{x})$.

$$V(\bar{x}) = \frac{s^2}{n} \quad \text{donde} \quad s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \text{ es la varianza}$$

muestral.

Con tamaño muestral fijo, en la medida que la varianza muestral sea menor, mayor será nuestra precisión (menor error). Pero lo más importante, en la medida que el tamaño muestral aumenta, independientemente de la dispersión de las observaciones, menor será el error de estimación.

Ahora bien, ya vimos que bajo el supuesto de muestreo sucesivo de tamaño n de una población de N elementos, la media muestral se convierte en una variable aleatoria. Toda variable aleatoria, como vimos en probabilidades, tiene una determinada distribución de probabilidades, y la variable aleatoria media muestral se distribuye normal. Esto es así en el caso en que la variable original se distribuya normal en la población, aunque, aún cuando esto no se cumple, si el tamaño de muestra es grande, digamos mayor o igual que 30, entonces la media muestral seguirá aproximadamente una distribución normal con media igual a la media poblacional y varianza s^2/n :

$$\bar{X} \sim \left(\mu, \frac{s^2}{n} \right)$$

Como recordarás del tema anterior, entre las propiedades de esta distribución, está el hecho de que aproximadamente el 95% de las observaciones distan de la media no más que dos veces la desviación estándar. Siendo exactamente así: en el intervalo formado por los valores

$\left(\bar{x} \pm 1.96 \sqrt{\frac{s^2}{n}} \right)$ se encuentra aproximadamente el 95% de los valores de la variable (media muestral).

Esta propiedad es muy importante ya que nos permite calcular el error de muestreo, e , y a la vez construir intervalos de confianza para la media poblacional.

Esto lo veremos en detalle en el tema siguiente. Por ahora te adelanto que el error de muestreo puede calcularse mediante la siguiente fórmula:

$$e = \mathbf{C} \sqrt{V(\bar{x})}$$

Donde **C** es el coeficiente de confiabilidad y se corresponde con el percentil $100(1 - \alpha/2)$ de la distribución normal estándar (en este caso concreto de la media). Sería así: $Z_{(1 - \alpha/2)}$. A su vez α es el nivel de significación, como verás en el próximo tema, de manera que $100(1 - \alpha) \%$ es el nivel de confiabilidad.

Entonces para un nivel confianza del 95%, $\alpha = 0.05$ (5%) y **C** = 1.96, que es el percentil 97.5 de la distribución normal estándar, es decir, $Z_{(1 - \alpha/2)} = Z_{0.975} = 1.96$. En el caso del 99%, **C** = 2.58. Esto lo entenderás mucho mejor cuando abordemos el tema siguiente, espero esto te sirva de introducción.

Ahora bien, con el error de muestreo se puede construir un intervalo de confianza (IC) para la media poblacional. Restando y sumando el error de muestreo a la media muestral obtendremos los límites inferior y superior de dicho intervalo, respectivamente. Para un IC al 95% será así:

$$\bar{x} \pm e = \bar{x} \pm \mathbf{1.96} \frac{s}{\sqrt{n}}$$

Notarás de la fórmula anterior, que el error es la mitad de la longitud del intervalo, por lo tanto, si efectivamente ocurriese que la media poblacional está contenida en el intervalo, entonces la distancia máxima entre la estimación \bar{x} y μ es, a lo sumo, igual al error, de ahí que se considere ese valor **e** como el **error asociado a la estimación**.

Quizás te preguntes ¿cómo saber que el error cometido es pequeño?. El error expresado en términos absolutos no nos dice mucho, a menos que conozcamos muy bien el comportamiento del fenómeno estudiado, en cambio si lo expresamos, como se hace frecuentemente, en términos relativos, podremos entonces comparar y determinar si es grande o no. El error relativo de la estimación se calcula como sigue:

$$e_x = \frac{e}{\bar{x}}$$

De esta forma el error pierde sus unidades de medida y puede expresarse tanto en términos de proporción como de porcentaje. Esta fórmula es válida para el error relativo de la estimación de cualquier otro parámetro, consiste básicamente, en medir qué parte representa el error de la estimación.

En general se considera que un error relativo no superior al 10% resulta aceptable. Esto es importante a la hora de calcular el tamaño de muestra necesario, como veremos muy pronto.

2.2.2 - Estimación de la proporción poblacional π .

Imagínate que deseas estimar la proporción de sujetos en la población que presentan cierta característica A: discapacidad física por ejemplo. Para resolver este problema resulta conveniente codificar esta variable cualitativa de la siguiente manera: se define una variable aleatoria X que solo toma valores 0 y 1, en dependencia de la presencia o ausencia de la característica en cuestión. Por ejemplo, la característica estudiada es la presencia de discapacidad física:

$$X = \begin{cases} 0 & \Rightarrow \text{No discapacidad física} \\ 1 & \Rightarrow \text{Discapacidad física} \end{cases}$$

De esta forma, la suma de todas las observaciones de la variable X en la muestra será igual al número de personas con discapacidad física en dicha muestra, a .

$$a = \sum_{i=1}^n x_i$$

Por lo tanto, la proporción muestral de personas con discapacidad física será.

$$p = \frac{\sum_{i=1}^n x_i}{n} = \frac{a}{n}$$

Esto quiere decir, que la proporción es un caso particular de la media aritmética, cuando se define una variable aleatoria de la forma anterior. Por lo tanto, cabe aplicar lo aprendido sobre la estimación de la media. En el caso de muestreo equiprobabilístico, la proporción muestral p , es un estimador insesgado de la proporción poblacional π , siendo la varianza de dicho estimador:

$$V_{(p)} = \frac{pq}{n} \quad \text{donde } q = 1 - p$$

De aquí que el error asociado a la estimación es igual a:

$$e = \mathbf{C} \sqrt{\frac{pq}{n}}$$

Además, puedes construir IC para π , de manera similar al caso de la media.

$$p \pm e = p \pm \mathbf{C} \sqrt{\frac{pq}{n}}$$

El término **C** tiene el mismo significado que en el caso de la media. Esto lo retomaremos en el siguiente tema.

2.2.3 Tamaño de la muestra.

Seguramente te has estado preguntando ¿cuál debe ser el tamaño muestral necesario para realizar determinado estudio?. La respuesta a esta interrogante la hallarás en las líneas siguientes.

Para calcular el tamaño de muestra necesario debemos, en primer lugar, establecer el valor máximo de error que estamos dispuestos a admitir, que será denotado E_0 . Definir este aspecto de vital importancia es responsabilidad plena del investigador conocedor del problema a estudiar y no del muestrista. Esto debe llevarse a cabo de acuerdo al problema en particular y teniendo en cuenta los objetivos del estudio. Generalmente suele emplearse entre el 5 y el 10 % de la magnitud supuesta del parámetro a estimar.

La fórmula empleada para calcular el tamaño muestral n , es la que sigue:

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \quad \text{donde} \quad n_0 = \frac{(C)^2 \sigma^2}{E_0^2}$$

Como puedes ver, el tamaño muestral depende de los siguientes elementos: el tamaño poblacional N , la varianza poblacional σ^2 , el error máximo admisible E_0 y el coeficiente de confiabilidad **C**; éste último con el mismo significado que en el caso del cálculo del error aleatorio, **e**, todos estos elementos deberán ser especificados.

El tamaño poblacional no siempre se conoce, no obstante, el efecto de N sobre el tamaño muestral puede ser considerado, a los efectos prácticos, como despreciable, lo que realmente influye en la precisión que se obtenga no es la magnitud relativa de la muestra respecto a la población, sino el tamaño absoluto de n . Por lo tanto, en estos casos n_0 puede ser considerada como una aproximación aceptable del valor de n .

En cuanto al error máximo admisible, quedamos en que es responsabilidad del investigador que conoce bien el problema de estudio y que suele tomarse entre el 5 y el 10 % de la magnitud que presumiblemente tenga el parámetro a estimar.

La varianza poblacional, por su parte, es un parámetro desconocido y no podemos emplear la varianza muestral puesto que aún no hemos seleccionado la muestra, en este caso se puede realizar un estudio piloto, a pequeña escala, y con esos resultados se estima la

varianza desconocida, o puede obtenerse un valor aceptable a partir de trabajos anteriores.

El coeficiente de confiabilidad, dijimos que tiene el mismo significado que al calcular el error de muestreo.

Esta forma de determinar el tamaño muestral es bastante esquemática. En la práctica este proceso es mucho más complejo. Estimar el error máximo admisible, exige del investigador un nivel de conocimientos del problema a investigar que no siempre es posible; y un problema aún mayor, se trata de una decisión eminentemente subjetiva. De hecho, para calcular el error máximo admisible, es necesario que el investigador se pronuncie hacia cual será la magnitud esperada del parámetro a estimar, lo cual es paradójico, si el parámetro se desea estimar es porque es desconocido, ¿Entonces?, se suele emplear las referencias de estudios similares, los criterios de expertos o un estudio piloto.

De forma similar ocurre con la varianza poblacional. Por otro lado, tomar un nivel de confiabilidad u otro es algo completamente subjetivo. De forma casi generalizada se emplea el 95 % de confiabilidad, ¿por qué no 96 o 94?, como ves esto también es subjetivo.

De esta forma, puede darse la situación en que dos investigadores siguiendo la misma estrategia general para estimar el mismo parámetro, pueden llegar a tamaños de muestra abismalmente diferentes.

Otro problema que surge al calcular el tamaño muestral, es que habitualmente se estima más de un parámetro en un estudio, por lo tanto, al aplicar la fórmula para el tamaño muestral obtendremos diferentes tamaños para cada parámetro. La fórmula vista es para el caso de la media, si lo que queremos estimar es una proporción la fórmula a emplear es la siguiente:

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}} \quad \text{donde } n_0 = \frac{C^2 pq}{E_0^2}$$

En este caso p es la proporción de personas con determinada característica, que supuestamente existen en esa población y $q = 1 - p$.

Entonces en un estudio que se desea estimar un promedio y una proporción ¿qué hacer?. Se plantea por algunos que el investigador ante esta situación debe “escoger” el parámetro más importante y operar con él, cayendo nuevamente en un margen de subjetividad importante.

Mi interés con lo dicho no es desanimarte ni predisponerte con el uso de las fórmulas, solo te he querido advertir sobre algunos de

los problemas que se presentan al determinar el tamaño muestral necesario para un estudio. En fin, el objetivo es tener una idea aproximada de la magnitud necesaria de la muestra, la decisión final corre por parte del investigador, teniendo en cuenta los recursos disponibles, el tiempo, los antecedentes de trabajos similares, etc. Y lo más importante, cualquiera que sea el tamaño muestral, es que el error asociado a la estimación puede ser calculado a posteriori.

Una exposición detallada sobre estos y otros aspectos relacionados con el tamaño muestral puedes encontrar en Silva(1993 y 1997).

2.2.4 Efecto de diseño (DEFF)

Como se ha dicho, en la práctica el MSA puro es muy poco usado y los cálculos de tamaño muestral vistos con anterioridad, se realizan bajo el supuesto de que será empleado el MSA. Para compensar el hecho de que se empleará otro tipo de método de selección, el tamaño muestral n calculado debe ser aumentado.

En este punto, debo presentarte un concepto de singular importancia: el llamado efecto de diseño, DEFF (design effect de los ingleses). Puede decirse que consiste en una “multa” que debe pagarse cuando se emplean diseños complejos.

Un diseño muestral complejo, se refiere a aquellos métodos de selección que involucran varias etapas, siendo en general, al menos teóricamente, menos eficientes que el MSA, en lo concerniente a precisión de los estimadores. El DEFF se expresa en términos de los errores de muestreo, es decir, en términos de la varianza de los estimadores. Por lo tanto, el tamaño de muestra necesario para estimar un determinado parámetro, será mayor si se emplea un diseño complejo que uno simple, a los efectos de tener igual precisión.

De esta forma, el DEFF se define como la razón entre la varianza del estimador calculada según el método complejo y según el MSA. En el caso de la estimación de la media será así:

$$DEFF = \frac{V_C(\bar{x})}{V_S(\bar{x})}$$

Donde el numerador es la varianza de la media muestral usando la fórmula adecuada al diseño complejo y el denominador es la varianza de la media muestral en caso de MSA.

Lo habitual es que este indicador sea mayor que 1, pues se supone que con el MSA hay mayor precisión (menor varianza del estimador) que con un diseño complejo, aunque en ocasiones se gana en precisión al usar otro método. El grado en que éste excede

la unidad expresa cuanta eficiencia se ha perdido como consecuencia de hacer un diseño menos preciso pero más económico que el MSA.

Lo que se hace en la práctica, es calcular el tamaño muestral empleando las fórmulas vistas y después se multiplica este resultado por el DEFF correspondiente al diseño empleado, siempre mayor que 1, generalmente alrededor de 2. De esta forma el Tamaño muestral resultante será mayor.

En síntesis, todo lo visto hasta este momento, se puede aplicar a otros métodos de muestreo, pero con algunas particularidades, las que veremos a continuación para los diseños más usados.

2.3 Muestreo sistemático (MS)

Aplicar el MSA presupone la existencia de una lista de todos los elementos de la población objeto, del que será seleccionado un número de ellos mediante el empleo de números aleatorios.

Esto desde el punto de vista práctico, además de engorroso, puede resultar imposible, pues con frecuencia no se dispone de listas de la población e incluso en algunos casos no se puede siquiera definir de forma precisa la población objeto. Supongamos que se desea estudiar la efectividad de un nuevo tratamiento para la úlcera péptica, para ello se necesita incluir en el estudio a pacientes ulcerosos que cumplan ciertas condiciones, la población objeto de estudio, teóricamente, será la totalidad de pacientes portadores de úlcera péptica susceptibles de ser tratados con el nuevo tratamiento. Te imaginarás que es prácticamente imposible disponer de una lista de esta población.

Todo esto hace necesario el empleo de otros métodos de muestreo, entre los que se encuentra el **muestreo sistemático**.

Ilustremos con un ejemplo en que consiste este método. En el mismo estudio del tratamiento para la úlcera péptica (UP), supongamos que por estudios anteriores se conoce que el número de pacientes con UP que acude a una consulta de gastroenterología es aproximadamente 1200 en el año, se decide basado en algunas consideraciones tomar una muestra de 200 de esos pacientes. Como no se dispone de una lista, la forma más cómoda de hacer la selección es elegir 1 sujeto de cada 6 que acudan a consulta ($1200/200$), para ello se toma un número aleatorio, r , entre 1 y 6, eligiéndose la persona número r que asista a dicha consulta, a partir de aquí se van incluyendo en el estudio sucesivamente las 199 personas restantes de la siguiente forma:

$r, r + 6, r + 12, r + 18 \dots$ hasta el término del período de inclusión (1 año), quedando exactamente 200 personas seleccionadas si realmente asisten 1200 pacientes a la consulta ese año. Por ejemplo, si $r = 4$, se tomarán el número 4, 10, 16, 22,, 1198.

A este número r seleccionado aleatoriamente se le llama **arranque aleatorio**, el cual puede tomar valores entre uno y $\frac{N}{n}$.

Para aplicar este método basta tener un marco muestral con cierta ordenación explícita, por ejemplo, visitas sucesivas de pacientes a un centro, tarjetas ordenadas en un tarjetero, historias clínicas en un estante, e incluso, una lista de las unidades de análisis. Cuando el criterio de ordenación de los elementos en la lista es tal que los elementos más parecidos tienden a estar más cercanos, el muestreo sistemático suele ser más preciso que el aleatorio simple, ya que recorre la población de un modo más uniforme. Ahora bien, el ordenamiento de la población determinará las posibles muestras a seleccionar y por ende la varianza de los estimadores, mientras que para el MSA el ordenamiento es irrelevante. No obstante, esto no resta precisión al método como piensan muchos.

Ya vimos que para seleccionar r es necesario calcular la siguiente relación:

$$k = \frac{N}{n}$$

Al número k se le llama **intervalo de selección** y al procedimiento empleado, **método regular de selección**.

Supondrás que la magnitud k en ocasiones no es un número entero, en este caso pueden usarse varias alternativas, las que no abordaremos por considerarlas fuera del alcance de este texto. En estos casos te sugerimos tomes la parte entera de K y operes de la forma antes expuesta.

Debes saber, que al aplicar el muestreo sistemático, no es posible computar los errores asociados a las estimaciones a partir de la muestra utilizada. Esto no invalida el uso del método, lo que se hace en la práctica es utilizar los resultados como si se hubiera realizado un MSA.



Vale decir, que cuando se emplea el **muestreo sistemático** en el acto de selección, tanto el **tamaño de la muestra**, como las **estimaciones** de los parámetros y la estimación de los errores se calculan como si se fuese a utilizar o se hubiese empleado un MSA.

2.4 Muestreo aleatorio estratificado (MAE)

El MSA es un método de selección que descansa exclusivamente en el azar, sin embargo, como hemos visto, el azar no es una garantía de representatividad. Hay ocasiones en que éste no resuelve el problema, pues se sabe que en la población objeto existen diferentes grupos de elementos cuya representación en la muestra se quiere asegurar. Para ello se realizan listas separadas para dichos grupos y se procede a seleccionar submuestras de cada uno de ellos.

A estos grupos previamente formados se les llama **estratos** y deben abarcar a toda la población de estudio (exhaustividad) sin intersectarse (mutuamente excluyentes), es decir, cada miembro de la población pertenece a uno y sólo un estrato⁹. La idea es que los grupos sean internamente homogéneos en algún sentido y consecuentemente diferentes entre sí.



Quando la selección de la muestra implica la división de la población en estratos y dentro de los estratos se emplea un procedimiento de selección en que interviene el azar, sea MSA o MS, se dice que se aplica un **Muestreo Aleatorio Estratificado, (MAE)**.

Por ejemplo, se desea realizar un estudio en la población de estudiantes de una Universidad, en el que a través de una muestra de ellos queremos obtener información sobre el uso del condón.

Primeramente pudiéramos pensar en hacer un muestreo aleatorio simple, pero en su lugar podemos reflexionar sobre el hecho de que el comportamiento de la población con respecto a este carácter no es homogéneo, y atendiendo a él, podemos dividir a la población en dos estratos de acuerdo al sexo:

- Estudiantes femeninos (60% del total);
- Estudiantes masculinos (40% restante).

De modo que se garantice la representación de ambos grupos en la muestra, ya que de hacerlo al azar tendremos seguramente más

⁹ Esto te recordará las características de las escalas de clasificación.

hembras que varones por lo que los resultados no serán exactamente los reales.

2.4.1 Procedimientos de estimación.



Cuando se emplea el **MAE** no es posible, en todos los casos, producir muestras equiprobabilísticas, por lo tanto, la media y la proporción muestrales no son estimadores insesgados de los respectivos parámetros poblacionales.

Veamos, la población objeto de N elementos es dividida en k estratos de tamaños N_1, N_2, \dots, N_k , de manera que:

$$N_1 + N_2 + \dots + N_k = N$$

Dentro de cada estrato se selecciona una submuestra de tamaño n_i , de manera que:

$$n_1 + n_2 + \dots + n_k = n$$

En este caso, la probabilidad de inclusión de cada sujeto en la muestra depende del tamaño del estrato al cual pertenece y del tamaño de la submuestra seleccionada en dicho estrato, siendo igual a: $\frac{n_i}{N_i}$

Entonces, para que el procedimiento sea equiprobabilístico, es necesario que las probabilidades de inclusión de cada estrato sean iguales entre sí, esto se logra mediante la **asignación proporcional**, de esta forma el tamaño de la muestra de cada estrato es proporcional al tamaño del estrato correspondiente con respecto a la población total:

$$n_i = n \frac{N_i}{N}$$

Por lo tanto $\frac{n_i}{N_i} = \frac{n}{N}$ para todo i .

Solamente en este caso, el MAE produce muestras equiprobabilísticas y por tanto la media y la proporción muestral serán estimadores insesgados de los correspondientes parámetros poblacionales, de lo contrario es necesario introducir ponderaciones al realizar el proceso de estimación. Entonces un estimador insesgado de la media poblacional es:

$$\bar{x}_w = \sum_{i=1}^k W_i \bar{x}_i \quad \text{donde } W_i = \frac{N_i}{N} \text{ y } \bar{x}_i \text{ es la media del estrato } i.$$

De manera similar, un estimador insesgado de la proporción poblacional es:

$$P_w = \sum_{i=1}^k W_i P_i$$
 donde W_i igual al caso de la media y P_i es la proporción muestral del estrato i .

Para calcular los errores de muestreo es necesario tener en cuenta tanto la variabilidad interna dentro de cada estrato como la externa entre estratos. Es lógico que en la medida que los valores de n_i sean mayores aumentará la precisión y a su vez ésta es más alta en la medida que los estratos sean más homogéneos internamente. Así, las varianzas de los estimadores se calculan empleando fórmulas bastante complejas que escapan del alcance matemático de este texto. Al lector interesado lo remito a la bibliografía citada al final del tema.

En general, podemos decir que cuanto más homogéneos sean los estratos, más precisas resultarán nuestras estimaciones, pudiendo ser incluso en algunos casos más precisas que de haberse empleado el MSA.

El enfoque a utilizar para determinar el **tamaño muestral**, será esencialmente el mismo empleado en el MSA, ya que en la práctica estas fórmulas siempre producirán tamaños mayores que si se emplearan las propias del MAE.

2.5 Muestreo por conglomerados.

Si intentamos hacer un estudio sobre los habitantes de una gran ciudad, el muestreo aleatorio simple puede resultar muy costoso, ya que estudiar una muestra de tamaño n implica enviar a los encuestadores a n puntos distintos de la misma, de modo que en cada uno de ellos sólo se realiza una entrevista.

En esta situación es más económico realizar el denominado **muestreo por conglomerados**, que consiste, en este caso, en elegir aleatoriamente ciertos barrios dentro de la ciudad, para después elegir calles y edificios. Una vez elegido el edificio, se entrevista a todos los vecinos, o a una parte de ellos. De forma general puede definirse este método de selección de la siguiente forma:



El **muestreo por conglomerados** consiste en dividir la población objeto de estudio en cierto número de partes o conglomerados, llamadas unidades de primera etapa (UPE),

éstas a su vez pueden ser divididas también en partes, llamadas unidades de segunda etapa (USE) y así sucesivamente, hasta llegar a cierto nivel de subdivisión en que sean seleccionadas las unidades de análisis que conformarán la muestra.

Como puedes ver, la selección de la muestra puede ser realizada en una o varias etapas, existiendo así el muestreo por conglomerados monoetápico, bietápico y polietápico. De esta forma la necesidad de listas de la población se limita a aquellas unidades de muestreo seleccionadas en la etapa anterior.

Naturalmente, en cada etapa pueden aplicarse diversos métodos de selección de los ya vistos, pudiendo tornarse bastante compleja la estimación de los parámetros. A continuación abordaremos el muestreo por conglomerados monoetápico, por considerarlo como uno de los que más podrás usar, para el empleo de otros diseños más complejos te sugerimos consultes a un bioestadístico.

2.5.1 Muestreo por conglomerados monoetápico (MCM).



El **muestreo por conglomerados monoetápico (MCM)**, consiste en dividir la población objeto de estudio en **M** partes o conglomerados, se procede a seleccionar de forma aleatoria un número **m** de éstos, quedando conformada la muestra por el total de unidades de análisis contenidas en ellos.

En efecto la selección de la muestra se realizó en una sola etapa. Ahora, la forma de dividir la población puede basarse en aspectos geográficos, administrativos o de otra índole, pudiendo incluso referirse a instituciones. Por ejemplo, municipios, áreas de salud, hospitales, escuelas, provincias.

Este método produce **muestras equiprobabilísticas**, la probabilidad de inclusión de cada miembro de la población es igual a la del conglomerado al cual pertenece, puesto que cada unidad de análisis queda automáticamente incluida en la muestra una vez que ha sido seleccionada la UPE que la contiene, es evidente que esta probabilidad es igual a:

$$\frac{m}{M}$$

En este diseño el **tamaño muestral** es completamente **aleatorio**, éste depende del número de unidades de análisis contenidas en las UPE seleccionadas, por lo tanto no se podrá planificar adecuadamente los recursos necesarios para el estudio hasta que no se haya seleccionado la muestra, constituyendo la desventaja principal de este procedimiento de selección. Pero, si los tamaños de los conglomerados no son tan diferentes entre sí, entonces esta desventaja queda, a los efectos prácticos, anulada.

En cuanto al proceso de estimación, podemos decir que se realiza de forma similar a como se hace en caso del MSA, es decir, la media y la proporción muestrales son estimadores insesgados de los respectivos parámetros poblacionales. Pero la estimación del error asociado suele ser complejo lo cual escapa del alcance de este texto.

Supondrás que este método puede extenderse a varias etapas e incluso emplear diferentes procedimientos de selección en cada etapa. En caso de necesitar el empleo de estos diseños complejos, te sugerimos consultes un bioestadístico.

Resumen

En este tema estudiaste que:

1. El muestreo constituye un recurso de la investigación científica, cuya función básica consiste en determinar que parte de una realidad objeto de estudio deberá ser analizada, con el objetivo de hacer inferencias al todo de la cual procede. A esa parte analizada se le llama **muestra** y al todo de la cual procede **población** o **universo**. Debes distinguir entre **población objeto**, la que realmente se desea estudiar, y la **población muestreada**, la que realmente se estudió, no siempre ambas coinciden como es lo deseado, por lo que se debe ser cuidadoso a la hora de emitir conclusiones, éstas deben recaer sobre la población muestreada.
2. A partir de los resultados obtenidos al estudiar los elementos de una muestra, es posible realizar estimaciones de los parámetros poblacionales, siendo ésta la razón de ser de la muestra. Este procedimiento de estimación se realiza de manera diferente en dependencia del método de selección empleado¹⁰ e inevitablemente estará sujeto a un margen de error, el cual es llamado **error de muestreo** que puede ser cuantificado siempre y cuando se emplee un método probabilístico de selección. A su vez, la magnitud del

¹⁰ Esto es algo muy importante que frecuentemente es obviado por los investigadores, lo que trae consigo graves errores al interpretar los resultados.

error depende del tamaño muestral y de la variabilidad de los datos, siendo menor en la medida que aumenta el tamaño de la muestra.

3. Se dice que un **método** de selección de muestras es **probabilístico**, cuando otorga una probabilidad conocida, no nula, a cada uno de los elementos de la población de ser incluidos en la muestra. A su vez, éste será **equiprobabilístico**, si la probabilidad de inclusión es la misma para todos los elementos de dicha población. Siempre que se emplean métodos equiprobabilísticos, la media y la proporción muestrales son estimadores insesgados de los respectivos parámetros poblacionales.

4. El muestreo simple aleatorio es un método que descansa exclusivamente en el azar. Además de ser equiprobabilístico, se distingue por otorgar igual probabilidad de ser seleccionada a todas y cada una de las muestras de tamaño n susceptibles de ser formadas de esa población de N elementos. Es el método que se supone ha sido empleado para realizar los procedimientos de inferencia estadística que veremos en el tema siguiente.

5. El muestreo sistemático constituye una buena opción cuando existe algún tipo de ordenación de la población objeto, aunque no se disponga de una lista de la misma. Cuando el criterio de ordenación de los elementos es tal que los elementos más parecidos tienden a estar más cercanos, el muestreo sistemático suele ser más preciso que el aleatorio simple, ya que recorre la población de un modo más uniforme. Por otro lado, es a menudo más fácil no cometer errores con un muestreo sistemático que con este último.

6. El muestreo aleatorio estratificado implica la división de la población en k subconjuntos o estratos mutuamente excluyentes y exhaustivos, dentro de los cuales se seleccionan las unidades de análisis que conformarán la muestra mediante un procedimiento que involucra el azar. Éste produce muestras equiprobabilísticas solamente cuando se realiza la **asignación proporcional**, es decir, el tamaño de muestra de cada estrato es proporcional al tamaño del estrato correspondiente con respecto a la población total, de manera que:

$$\frac{n_i}{N_i} = \frac{n}{N} \text{ para todo } i$$

7. El muestreo por conglomerados suele ser más económico que el MSA cuando se desea estudiar una población grande, pues permite controlar la dispersión geográfica de la muestra al seleccionar las unidades de análisis solamente de algunas partes o conglomerados de la población. La selección puede ser realizada en una o varias

etapas, lo que le da nombre a los tipos específicos de este método, siendo el muestreo por conglomerado monoetápico el más empleado, éste produce muestras equiprobabilísticas pero el tamaño muestral es aleatorio, ya que depende del tamaño de los conglomerados seleccionados. No obstante, este sesgo en la práctica puede ser despreciado y emplear la media y proporción muestrales como estimadores de los parámetros correspondientes.

8. Determinar el tamaño de muestra mínimo necesario para realizar una investigación, constituye uno de los problemas más complejos de la teoría muestral. Siendo de vital importancia para el proceso inferencial, de él dependerá la precisión de nuestras estimaciones, además de ser un factor clave para el contraste de hipótesis, como veremos en el tema siguiente. Este proceder exige que el investigador piense en términos para los cuales generalmente no está preparado, por su parte, el uso de las fórmulas requiere de una serie de elementos con un gran margen de subjetividad. Pero, lo más importante, ninguna fórmula sustituye el buen juicio y el sentido común de un investigador bien preparado, además, al final lo que cuenta son los recursos disponibles.

Ejercitación

1. Responde verdadero(V) o falso(F) las siguientes afirmaciones, justifique brevemente las F.

- El MSA se distingue de otros métodos de muestreo porque siempre genera muestras equiprobabilísticas.
- El muestreo sistemático a diferencia del MSA es influenciado por la ordenación de las unidades de análisis en el marco muestral.
- Un muestreo estratificado con un solo estrato es equivalente a un muestreo simple aleatorio.
- Se dice que una muestra es representativa de una población cuando la media muestral de los estimadores coincide con los parámetros respectivos.
- Un mismo diseño muestral puede ser equiprobabilístico a los efectos de estimar un parámetro y no serlo si se trata de estimar otro parámetro diferente.

2. Identifique en las siguientes situaciones el método muestral empleado.

a - Se desea estudiar la incidencia de bajo peso al nacer en una comunidad, para ello se tomó 1 de cada 10 nacimientos ocurridos

en el área en un periodo de un año, obteniéndose una muestra de 300 recién nacidos.

b - Con el objetivo de estimar la prevalencia de anemia durante el embarazo en un área de salud, se estudió una muestra de 250 embarazadas tomadas aleatoriamente de un total de 600.

c - En un estudio realizado con el fin de evaluar el grado de madurez escolar en niños de primer grado, fueron seleccionados 300 niños: 150 provenientes de círculo infantil y los 150 restantes cursaron el preescolar en escuelas.

d - En un área de salud se desea evaluar el desempeño de los médicos de familia, para ello deciden seleccionar 2 de los 4 grupos básicos de trabajo existentes y dentro de estos tomar aleatoriamente 10 consultorios, para obtener una muestra de 20 médicos.

3. Se planifica realizar un estudio de fecundidad en una ciudad donde hay aproximadamente 3489 mujeres en edad fértil. Supongamos que se quiere conocer la talla promedio de dichas mujeres con error no mayor de 2 cm. Se conoce por estudios anteriores que la desviación estándar poblacional es alrededor de 25 cm.

a - Calcule el tamaño de muestra mínimo necesario para realizar dicho estudio, si se quiere emplear el MSA.

b - ¿Cuál será el tamaño muestral resultante si se desea aplicar un diseño complejo que involucra varias etapas?

c - Si la media obtenida en dicho estudio, después de aplicar MSA, es de 168 cm, ¿Cuál es la media poblacional?

d - Calcula el error asociado a dicha estimación. Con una confiabilidad del 95% y sabiendo que $\sigma = 22$.

4. Un investigador está interesado en estimar la proporción de muertes debidas a cáncer de estómago en relación con el número de defunciones por cualquier tipo de neoplasia. Su experiencia le indica que sería sorprendente que tal proporción supere el valor de $1/3$.

a - ¿Qué tamaño de muestra debe tomar para estimar la anterior proporción, con una confianza del 99%, para que el valor estimado no difiera del valor real en más de 0,03?.

b - Si queremos estimar dicha proporción con un error máximo del 4%, para una confianza del 95%, ¿Qué tamaño de muestra debemos tomar?

Autoevaluación

1. ¿En qué consiste el MSA?, ¿Qué desventajas posee respecto al muestreo sistemático?
2. ¿Qué entiende usted por una muestra representativa?
3. ¿A qué llamamos muestra probabilística? Y ¿Por qué es importante obtener este tipo de muestra?
4. Se desea estimar el tiempo medio de sangría en fumadores de más de 20 cigarrillos diarios, con edades comprendidas entre 35 y 40 años, con una precisión de 5 segundos. Ante la ausencia de cualquier información acerca de la variabilidad del tiempo de sangría en este tipo de individuos, se tomó una muestra preliminar de 5 individuos, en los que se obtuvieron los siguientes tiempos (en segundos):
97, 80, 67, 91, 73.
 - a - Determinar el tamaño mínimo de muestra, al 95%, para cumplir el objetivo anterior.
5. Se quiere estimar la incidencia de la hipertensión arterial en el embarazo. ¿Cuántas embarazadas tenemos que observar para, con una confianza del 95%, estimar dicha incidencia con un error del 2% en los siguientes casos:
 - a - Sabiendo que un sondeo previo se ha observado un 9% de hipertensas.
 - b - Sin ninguna información previa.

Bibliografía

1. Silva LC. **Muestreo para la investigación en salud**. Madrid: Díaz de Santos; 1993
2. Silva LC. **Cultura Estadística e investigación científica en el campo de la salud: una mirada crítica**. Madrid: Díaz de Santos. 1997
3. Daniel WW. **Bioestadística. Base para el análisis de las ciencias de la salud**. 3ra ed. México. D.f: Limusa; 1997
4. Armitage P, Berry G. **Statistical methods in medical research**. 3ra ed. Oxford. Blackwell science. UK; 1994
5. Altman DG. **Practical statistics for medical research**. London: Chapman and Hall; 1992.
6. Norman GR, Streiner DL. **Bioestadística**. España: Hartcourt Brace; 1998

7. Armitage P, Berry G. **Estadística para la Investigación Biomédica**. Doyma, Barcelona, 1992.
8. Hamilton LC. **Modern Data Analysis**. Brooks/Cole Publishing Company, Pacific Grove, 1990.
9. Martín Andrés A, Luna Del Castillo JD. **Bioestadística para las Ciencias de la salud**. Norma, Granada, 1994.
10. Marascuilo LA, Serlin RC. **Statistical Methods for the Social and Behavioral Sciences**. W.H. Freeman and Company, Nueva York, 1988.
11. Peña Sánchez De Rivera D. **Estadística: Modelos y Métodos, 1**. Alianza Universidad Textos, Madrid, 1994.
12. Rivas Moya T, Mateo MA, Ríos Díaz F, Ruiz M. **Estadística Aplicada a las Ciencias Sociales: Teoría y Ejercicios (EAC)**. Secretariado de Publicaciones de la Universidad de Málaga, Málaga, 1991.
13. Versión electrónica del manual de la Universidad de Málaga. **Bioestadística: métodos y aplicaciones**. España; 1998.

Tema 3 – Introducción a la Inferencia Estadística.

Introducción

Generalmente, al realizar un estudio se desea arribar a conclusiones acerca de la naturaleza de una población. Al ser la población grande y no poder ser estudiada en su integridad en la mayoría de los casos, las conclusiones obtenidas deben basarse en el examen de solamente una parte de ésta, lo que nos lleva, en primer lugar a la justificación, necesidad y definición de las diferentes técnicas de muestreo, estudiadas en el tema anterior. Toca el turno al proceso de generalización, la inferencia estadística.

La **inferencia estadística** es el procedimiento por medio del cual se llega a conclusiones acerca de una población mediante los resultados que se obtienen a partir de una muestra extraída de ella.

Pero, existe un inconveniente a la hora de dar estas conclusiones: **el error aleatorio**. Cualquier medición basada en una muestra de individuos presentará alguna diferencia respecto del verdadero valor como consecuencia de la aleatoriedad del proceso.

El investigador solo estudia una de las posibles muestras que pueden obtenerse de la población objeto de estudio. Y en cada una de éstas, la variable de interés puede presentar diferentes valores simplemente por azar. Las técnicas estadísticas se basan en el hecho de que esta variabilidad propia del muestreo, sigue unas leyes conocidas, por lo que puede ser cuantificada.

La **Estadística inferencial** se usa justamente, para determinar la probabilidad de que una conclusión sacada a partir del análisis de los datos de una muestra sea cierta.

En este tema te introducimos los elementos de estadística inferencial necesarios para realizar este proceso de generalización de la muestra a la población, que se aplicarán a dos indicadores particularmente sencillos y muy empleados: la media y la proporción.

3.1 - Distribuciones muestrales.

El primer término al que debemos hacer referencia es el de estimador.

Dentro de este contexto, será necesario asumir un **estimador** como una variable aleatoria con una determinada distribución, que se expresa en función de la muestra estudiada, con el objetivo de aproximar el valor de un parámetro poblacional desconocido; y que

será la pieza clave en las dos amplias categorías de la inferencia estadística: la estimación y el contraste de hipótesis.

El concepto de estimador, como herramienta fundamental, lo caracterizamos mediante una serie de propiedades que nos servirán para elegir el "mejor" para un determinado parámetro de una población, así como algunos métodos para la obtención de ellos, tanto en la estimación puntual como por intervalos. Esto lo veremos más adelante.

Es de vital importancia que entiendas claramente en qué consisten las distribuciones muestrales, ya que este concepto es la clave para comprender la inferencia estadística. Así pues, la distribución muestral de un estimador se ha definido como sigue.



La distribución de frecuencias de todos los valores posibles que puede asumir un estimador, calculados a partir de muestras del mismo tamaño, extraídas aleatoriamente de la misma población, se llama **distribución muestral** de ese estimador.

Podemos referirnos a una distribución de medias, de proporciones, de varianzas, etc.

Las distribuciones muestrales pueden construirse empíricamente a partir de poblaciones finitas y discretas. Mediante el siguiente procedimiento:

1. De una población finita de tamaño **N**, se extraen de manera aleatoria todas las muestras posibles de tamaño **n**.
2. Se calcula el estimador de interés para cada muestra.
3. Se lista en una columna los diferentes valores observados del estimador y en otra columna las frecuencias correspondientes de cada valor observado.

Como supondrás, esto en la práctica es una tarea colosal y en el caso de poblaciones infinitas es imposible. Por tanto este problema debe ser resuelto por medio del uso de la Matemática. Pero no te asustes que los procedimientos aplicados para ello no son compatibles con el nivel matemático de este texto. Baste saber que las distribuciones muestrales pueden deducirse matemáticamente.

Ahora bien, como estudiaste en probabilidad, para cualquier distribución de probabilidad se tiene interés en conocer tres cosas

básicas: su valor esperado (media), la varianza y su apariencia gráfica.

En este punto cabe señalar que todos los procedimientos de estadística inferencial se basan en el supuesto de muestreo simple aleatorio, sin importar que se haya empleado otro método de selección, lo que se considera por algunos como una deficiencia de esta rama de la estadística, y en estos momentos se encuentra en proceso de discusión y análisis por los grandes estadísticos. Dejemos que ellos se ocupen de ese problema y sigamos nosotros con nuestro tema.

Pues bien, dado que se trata de muestras elegidas al azar, lo más frecuente es que los valores promedio observados (proporción \mathbf{p} y media \bar{x}) sean valores cercanos a los verdaderos parámetros de la población (proporción π y media μ). No obstante, el error aleatorio propio del muestreo produce algunas muestras con valores de \mathbf{p} y \bar{x} alejadas de π y μ .

Habrás notado que a los valores muestrales de la variable se les denota por letras romanas mientras que a los parámetros poblacionales con letras griegas, esto evita confusiones y favorece la comprensión de los conceptos que se definirán más adelante.

En el siguiente cuadro te brindo una lista de la nomenclatura utilizada.

Término estadístico	Muestra	Población
Media	\bar{x}	μ (mu)
Proporción	\mathbf{p}	π (pi)
Desviación estándar	\mathbf{s}	σ (sigma)
Diferencia	\mathbf{D}	δ (delta)

Espero que te sirva de guía y no te confundas a lo largo del tema, a partir de ahora emplearemos esta nomenclatura.

3.1.1 - Distribución muestral de la media.

Apliquemos el concepto de distribución muestral, introducido anteriormente, al caso concreto de la media muestral.



La distribución de frecuencias de las medias \bar{x} observadas en todas las muestras posibles de tamaño n , extraídas al azar de una población de media μ y varianza σ^2 conocidas, recibe el nombre de **distribución muestral de medias**.

Como vimos en el tema anterior, la media muestral bajo el supuesto de muestreo sucesivo de tamaño n de una población con N elementos, se convierte en una variable aleatoria que sigue una distribución normal.

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Esta distribución tiene las siguientes características:

1. Su valor esperado es igual a la media poblacional

$$E(\bar{x}) = \mu$$

2. Su varianza es igual a:

$$\text{Var}(\bar{x}) = \frac{\sigma^2}{n}$$

3. El error estándar de la media muestral viene dado por:

$$EE_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

4. Su representación gráfica se muestra a continuación en la figura 4.1.1.

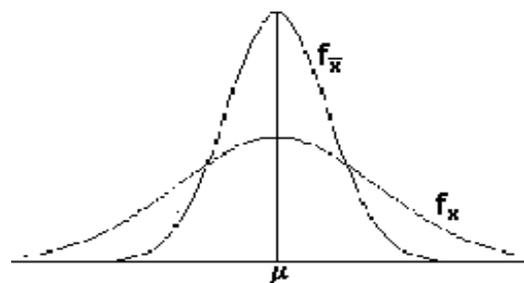


Figura 3.1.1 – Distribución de la media muestral y la respectiva distribución normal.

En la medida que aumenta el tamaño de muestra la variabilidad de la media tiende a disminuir y por lo tanto la curva tiende a ser más aguda.

Es importante señalar que la media muestral sigue una distribución normal cuando la distribución de la variable en la población sigue una ley normal, y cuando el tamaño de muestra n es grande aunque la distribución de la variable en la población no siga una ley normal. Esto se debe al **teorema central del límite**, que explica las leyes de los errores causales, según la cual, cuando el error asociado a un determinado fenómeno es el resultado de muchos errores independientes, se observa que la distribución de ese error es parecida a la ley normal. Esta ley fue descubierta por De Moivre.

El significado de la distribución de De Moivre reside en que la suma de una gran cantidad de variables aleatorias independientes, arbitrariamente distribuidas, siempre tiene una distribución aproximadamente normal. La distribución de dicha suma será más cercana a la normal cuanto mayor sea el número de variables aleatorias.

Este teorema constituye la razón por la cual muchas distribuciones muestrales pueden aproximarse por la ley normal a partir de un cierto tamaño de muestra.

Al llegar a este punto surge una pregunta lógica. ¿Cuán grande debe ser la muestra para que el teorema central del límite sea aplicable?. Una regla empírica establece que, en la mayoría de las situaciones prácticas, una muestra de tamaño 30 es suficiente.

El conocimiento de la distribución muestral de la media y de sus características, es muy importante cuando se desea realizar inferencias a partir de medias, ya sea comparar las medias de dos poblaciones o comparar una media con una magnitud específica, como veremos en las próximas secciones.

3.1.2 Distribución muestral de la proporción

Ya vimos el caso en que la variable era continua y por tanto el estadístico de interés era la media, veamos ahora cuando estamos interesados, como sucede frecuentemente, en una característica categórica, siendo la proporción muestral el estadístico para estas situaciones.

La distribución muestral de la proporción se puede obtener experimentalmente de la misma forma que vimos para la media. Lo cual prácticamente no se realiza, claro está.

Así tenemos que:



El conjunto de los valores de una proporción p , observados en todas las posibles muestras de tamaño n , extraídas aleatoriamente, de una población de N elementos con proporción π , recibe el nombre de **distribución muestral de la proporción**.

Esta distribución sigue una ley Binomial con las siguientes características:

1. Su valor esperado es igual a la proporción poblacional

$$E(p) = \pi$$

2. Su varianza es igual a:

$$\text{Var}(p) = \frac{\pi(1-\pi)}{n}$$

3. El error estándar de la proporción p está dado por:

$$EEp = \sqrt{\frac{\pi(1-\pi)}{n}}$$

Pero, cuando la muestra es grande, la distribución de las proporciones de la muestra es aproximadamente normal de acuerdo al teorema central del límite. Considerándose en la práctica que la distribución de las proporciones muestrales observadas se aproxima satisfactoriamente a una ley normal de media π y varianza $\frac{\pi(1-\pi)}{n}$, si el tamaño muestral cumple que $n\pi \geq 5$ y $n(1-\pi) \geq 5$.

De aquí se deduce que en la medida que π se acerque a 0.5 el tamaño de muestra necesario será menor que cuando está cerca de 0 o de 1. De forma general se considera adecuado un tamaño muestral de 50 o más.

Al igual que en el caso de la media, el conocimiento de todos estos elementos son muy útiles al hacer inferencias sobre proporciones, como veremos en las próximas líneas.

3.1.3 – Distribución muestral de la diferencia entre medias y proporciones.

Vimos las distribuciones de la media y la proporción, pero resulta de interés conocer que sucede con las diferencias entre medias y entre proporciones. Es muy simple, veamos.

A menudo estamos interesados en comparar dos poblaciones en cuanto a una variable en particular. Cuando ésta es continua

estamos entonces interesados en la diferencia entre las medias de dicha variable para ambas poblaciones, si la variable es categórica, entonces nos interesa la diferencia entre las proporciones correspondientes, es decir.

$$(\bar{x}_1 - \bar{x}_2) \text{ y } (p_1 - p_2) \text{ respectivamente.}$$

Ahora bien, estas diferencias bajo el supuesto de muestreo sucesivo del mismo tamaño¹¹, de cada población, se comportarán como variables aleatorias que siguen una distribución de probabilidades conocida.

En el caso de las medias, sigue una distribución normal con media igual a la diferencia entre ellas a escala poblacional ($\mu_1 - \mu_2$) y varianza igual a la suma de las varianzas de cada media:

$$(\bar{x}_1 - \bar{x}_2) \sim N\left(\mu_1 - \mu_2, \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)\right)$$

De forma análoga, en el caso de las proporciones tenemos que sigue una distribución aproximadamente normal si los tamaños muestrales n_1 y n_2 son grandes.

$$(p_1 - p_2) \sim N\left(\mu_1 - \mu_2, \left(\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right)\right)$$

3.2 - Estimación de parámetros.

En esta sección analizaremos la **estimación**, que es la primera de las dos áreas generales de la Estadística inferencial, la segunda, pruebas de hipótesis, la estudiaremos en la próxima sección.

El proceso de estimación implica calcular, a partir de los datos de una muestra, algún estadístico que se ofrece como una aproximación del parámetro correspondiente de la población de la cual fue extraída la muestra (media, proporción, etc.).

Aunque no lo habíamos definido anteriormente, habrás notado que en general, un parámetro es cualquier indicador descriptivo de una población.

Como vimos anteriormente, se denomina estimador de un parámetro, a cualquier v.a. que se exprese en función de la muestra aleatoria y que tenga por objetivo aproximar el valor de un parámetro poblacional desconocido.

Observa que el estimador no es un valor concreto sino una variable aleatoria, ya que aunque depende unívocamente de los valores de la muestra observados, la elección de la muestra es un

¹¹ Se refiere al mismo tamaño de muestra para cada población, pudiendo ser diferentes de una población a la otra.

proceso aleatorio. Una vez que la muestra ha sido elegida, se denomina estimación al valor numérico que toma el estimador sobre esa muestra.

Entre las características deseables para esta nueva variable aleatoria (que usaremos para estimar el parámetro desconocido) tenemos las siguientes:

- **Consistencia:** Cuando el tamaño de la muestra crece arbitrariamente, el valor estimado se aproxima al parámetro desconocido.
- **Carencia de sesgo:** El valor medio que se obtiene de la estimación para diferentes muestras debe ser el valor del parámetro.
- **Eficiencia:** Al estimador, al ser v.a., no puede exigírsele que para una muestra cualquiera se obtenga como estimación el valor exacto del parámetro. Sin embargo podemos pedirle que su dispersión con respecto al valor central (varianza) sea tan pequeña como sea posible.
- **Suficiencia:** El estimador debería aprovechar toda la información existente en la muestra.

Veremos que es posible calcular dos tipos de estimaciones para cada uno de estos parámetros: una estimación puntual y una estimación por intervalos.



Una **estimación puntual** es un solo valor numérico empleado para estimar el parámetro correspondiente de la población.

Una **estimación por intervalos o confidencial** consta de dos valores numéricos que definen un intervalo que, con un grado de confianza específico, se considera que incluye al parámetro por estimar.

Este parámetro será habitualmente una proporción en el caso de variables dicotómicas, y la media o la varianza para variables continuas.

3.2.1 Estimación puntual

La estimación puntual es muy simple, podrás suponer que consiste en escoger el mejor estimador para el parámetro en cuestión.

Ya vimos en secciones anteriores, que la media de la muestra, la proporción de la muestra, la diferencia entre las medias de dos muestras y la diferencia entre las proporciones de dos muestras, son cada una estimadores insesgados de sus parámetros correspondientes. Esta propiedad está implícita cuando se dijo que los parámetros eran las medias de las distribuciones muestrales correspondientes. Por lo que se les considera como la estimación puntual de cada uno de estos parámetros.

Ahora bien, es muy importante considerar aquí algo que vimos en el tema anterior, y es la diferencia que existe entre **población objeto** y **población muestreada**. La población objeto es la que realmente se desea estudiar, mientras que la muestreada es de la que se extrajo la muestra, no siempre ambas coinciden, desafortunadamente. Los procedimientos de inferencia estadística permiten inferir respecto a la población muestreada (siempre y cuando se hayan empleado técnicas de muestreo correctos). Por lo que esto debes tenerlo muy en cuenta a la hora de hacer tus estimaciones.

3.2.2 Estimación confidencial (por intervalo).

La técnica de la estimación confidencial consiste en asociar a cada muestra un intervalo que se sospecha que debe contener al parámetro. A éste se le denomina intervalo de confianza

La estimación confidencial ofrece una información más amplia que la simple estimación puntual, lo cual suele ser muy ventajoso. Por ejemplo, dos investigadores estiman la proporción de fumadores de la población donde reside cada uno, construyen los siguientes intervalos de confianza (IC) al 95%:

Población A (56,4% ; 63.6%)
 Población B (23.5% ; 96.5%)

Como puedes ver en ambos casos la estimación puntual es del 60% (punto medio del intervalo), sin embargo en la población A la estimación es mucho más precisa que en la B donde la proporción de fumadores puede oscilar desde un 23.5% hasta un 96%. Esto nos demuestra que si nos hubiéramos referido solamente a la estimación puntual podíamos haber pensado que en ambas poblaciones la prevalencia de fumar es similar. Es decir, El IC nos

ofrece además de la estimación puntual, información acerca de la precisión con que hicimos nuestra estimación.

Evidentemente esta técnica no tiene porqué dar siempre un resultado correcto. A la probabilidad de que hayamos acertado al decir que el parámetro estaba contenido en dicho intervalo se la denomina nivel de confianza. También se denomina nivel de significación a la probabilidad de equivocarnos.

En la sección siguiente abordaremos el procedimiento para la construcción de IC para los parámetros más usados. De esta forma podrás comprender mejor los conceptos antes expuestos.

3.3 Intervalos de confianza

Ya vimos en la sección anterior que la estimación de parámetros poblacionales puede hacerse tanto de forma puntual como por intervalos, siendo esta segunda forma más informativa y preferida por muchos investigadores. A continuación veremos cómo se pueden construir estos IC para algunos parámetros.

Pero antes debes saber, que los IC se construyen con la información obtenida en la muestra estudiada, utilizando la estimación puntual de dicho parámetro, teniendo en cuenta el error aleatorio o de muestreo y por supuesto la distribución muestral del estimador.

3.3.1 Intervalo de confianza para la media

Si conocemos la media \bar{x} , de una determinada variable continua, observada en una muestra, sabemos que este valor es la mejor estimación puntual del parámetro μ . No obstante, es deseable conocer su precisión, es decir, lo próximo o alejado que \bar{x} puede estar del verdadero valor μ . Como sabemos, la precisión depende fundamentalmente del tamaño muestral: cuanto mayor sea el tamaño de la muestra, más precisa será esta estimación, y en segundo lugar de la variabilidad de los datos, σ^2 .

Así pues, la estimación puntual de μ , y de cualquier parámetro, debe ir acompañada de su precisión e , indicadora de las posibles diferencias entre el valor estimado y el verdadero valor del parámetro.

De esta forma, el intervalo de confianza $1 - \alpha$ del parámetro μ , es el intervalo $\bar{x} \pm e$ alrededor de la estimación puntual \bar{x} , que tiene una probabilidad grande ($1 - \alpha$) de contener en algún punto del intervalo la desconocida μ .

Al valor $1 - \alpha$, se le llama **confiabilidad o nivel de confianza**, suele ser considerada como 0.95 (95%) o mayor, pues en estadística por convenio, se considera que una probabilidad es grande cuando su valor es igual o superior a 0.95. Por su parte, el valor α es un indicador del riesgo de que el intervalo de confianza no contenga el verdadero valor del parámetro μ , fijándose habitualmente en 0.05 (5%), suele llamársele nivel de significación.

Recordemos que el error estándar de la media se define como:

$$EE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Y que el error de muestreo e se define, en este caso, como:

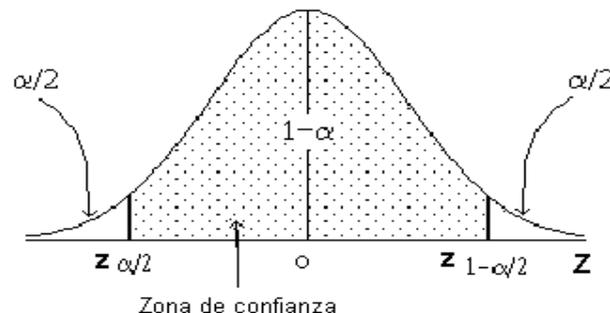
$$e = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

De aquí que el IC para μ está dado por:

$$\bar{x} \pm e \Rightarrow \bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Donde $Z_{\alpha/2}$ es el valor de la distribución normal estándar que se corresponde con un área a su izquierda igual a $\alpha/2$, es decir, el valor de Z para el que se cumple que:

$P(Z < z) = \alpha/2$. Este valor de Z recibe el nombre de



percentil¹² $\alpha/2$.

Veamos en la figura siguiente su representación gráfica.

Figura 3.3.1 Percentiles $\alpha/2$ y $1 - \alpha/2$ de la distribución normal estándar.

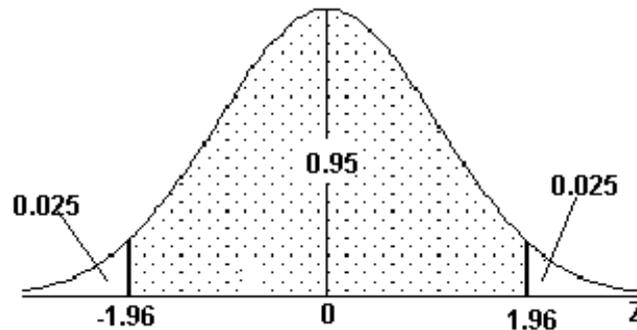
Recordarás que la distribución normal es simétrica respecto a la media, y en el caso de la normal estándar el percentil $Z_{\alpha/2}$ es igual al percentil $Z_{1 - \alpha/2}$ solo con signos diferentes.

$$-Z_{\alpha/2} = Z_{1 - \alpha/2}$$

¹² Recordarás que los percentiles son aquellos valores de la variable que dividen a una serie ordenada de datos (en este caso la distribución) en 100 partes iguales, existiendo entonces 99 percentiles.

Por ejemplo, para un valor de α igual 0.05 (5%), $\alpha/2 = 0.025$, significa que el área bajo la curva normal a la izquierda de $Z_{\alpha/2}$ y a la derecha de $Z_{1-\alpha/2}$ es igual a 0.025. Si buscas en la tabla A del Apéndice podrás ver que a este valor de probabilidad le corresponde un valor de $Z = -1.96$ y 1.96 respectivamente, es decir los percentiles 2.5 y 97.5 de la distribución normal estándar. Veamos el gráfico de la figura 3.3.2.

Figura 3.3.2 – Áreas bajo la curva de la distribución normal



estándar.

Al área comprendida entre estos dos percentiles se le llama región de confianza, se corresponde con un IC al 95% para la normal estándar.

Es decir, cuando construimos un IC con una confiabilidad de $100(1 - \alpha) \%$, significa que encontramos dos valores de esa distribución de manera tal, que entre ellos se encuentra el $100(1 - \alpha) \%$ de los valores que toma la variable.

De forma similar puedes calcular los percentiles para otros valores de alfa, te adelanto que para $\alpha = 0.01$, el valor correspondiente de Z es 2.58 (percentil 99.5).

El procedimiento anterior para construir IC para la media, se basa en que la variable original sigue aproximadamente una distribución normal y que se conoce la varianza poblacional σ^2 . Esto en la práctica es casi imposible, es muy raro conocer la varianza poblacional y no a la media, pues ésta es necesaria para obtener la varianza. ¿qué hacer?, gracias al teorema central del límite si la muestra es grande (≥ 30) garantizamos que la distribución de las medias muestrales sea aproximadamente normal, además el estadístico:

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Utilizado para estandarizar la distribución de las medias muestrales, sigue una distribución normal, incluso en el caso que no se conozca la varianza poblacional σ^2 , siempre que el tamaño

muestral sea grande, en estos casos se sustituye el valor de σ^2 por el de la varianza muestral s^2 .

En aquellos casos en que la muestra sea pequeña, no se conozca la varianza poblacional σ^2 , y (o) la variable no sigue una distribución normal es necesario emplear otras fórmulas para construir el IC para la media, que involucran a la distribución t de Student. No te asustes que son muy similares a las vistas anteriormente, sería así:

$$\bar{x} \pm e \Rightarrow \bar{x} \pm t_{(1-\alpha/2)} \frac{s}{\sqrt{n}}$$

Como ves, la diferencia radica en el uso de s sustituyendo a σ , y en lugar del percentil $1 - \alpha/2$ de la normal estándar usamos el mismo percentil pero de la distribución t de Student con $n-1$ grados de libertad. Veamos un ejemplo.

Ejemplo 3.3.1

Se ha medido el volumen diario de bilis, expresado en litros, en 10 individuos sanos, obteniéndose los siguientes resultados:

0,98; 0,85; 0,77; 0,92; 1,12; 1,06; 0,89; 1,01; 1,21; 0,77.

Estimemos la producción diaria media de bilis en individuos sanos suponiendo que la muestra ha sido obtenida por muestreo aleatorio simple sobre una población normal.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{9.58}{10} = 0.958$$

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n} = 0.02131$$

Por lo tanto el error estándar de la media es:

$$EE_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{0.146}{\sqrt{10}} = 0.0462$$

Calculemos el error de muestreo, con una confiabilidad del 95%, teniendo en cuenta que se trata de una muestra pequeña y no conocemos la varianza poblacional:

$$e = t_{(1-\alpha/2)} \times EE_{\bar{x}} = 2.2622 \times 0.0462 = 0.1045$$

El valor del percentil $t_{1-\alpha/2}$, surge de buscar en la tabla B del Apéndice el valor correspondiente de la distribución t para 9 ($n - 1$) grados de libertad, que el área bajo la curva a su derecha es de $(1 - \alpha/2)$, es decir, el percentil 97.5 para α igual a 5% y por tanto una confiabilidad del 95%. Como podrás observar en dicha tabla se muestran los valores de algunos percentiles de la distribución t para diferentes grados de libertad.

Entonces un IC al 95% para la producción media de bilis poblacional será:

$$\bar{x} \pm e = 0.958 \pm 0.1045 = (0.8535 ; 1.0625)$$

Habrás notado que el IC es bastante estrecho lo cual habla a favor de su precisión, a menor amplitud mayor precisión. Mientras que el error relativo es aproximadamente el 11%:

$$\frac{0.1045}{0.958} * 100 = 10.91\%$$

Se obtuvo esta precisión relativamente buena con tan bajo tamaño muestral, porque la variabilidad de las observaciones es baja. Veamos otro ejemplo.

Ejemplo 3.3.2

Un investigador está interesado en estimar el nivel promedio de una enzima, en una población humana determinada. Para ello tomó una muestra de 25 sujetos y obtuvo una media muestral $\bar{x}=22$. Se sabe que los niveles de esta enzima siguen una distribución normal con varianza $\sigma^2 = 45$.

Con esta información podemos construir un IC al 95 % para la media poblacional de dicha enzima. Podemos emplear la distribución normal pues se conoce la varianza poblacional y la variable se distribuye normal. Entonces:

$$EE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{45}{25}} = 1.3416$$

$$e = Z_{1 - \alpha/2} \times EE_{\bar{x}} = 1.96 \times 1.3416 = 2.63$$

El IC al 95 % para μ será:

$$\bar{x} \pm e = 22 \pm 2.63 = (19.37 ; 24.63)$$

Siendo el error relativo del 12 % aproximadamente:

$$\frac{2.63}{22} * 100 = 11.95 \approx 12 \%$$

¿Cómo se interpretan estos intervalos de confianza?. En estos ejemplos en que la confiabilidad fue del 95 %, se dice que, al repetir el muestreo, aproximadamente el 95% de los IC construidos mediante las fórmulas vistas, incluyen la media poblacional, esta es la interpretación probabilística. La interpretación práctica plantea que se tiene un $100(1 - \alpha)$ % de confianza que el intervalo único calculado, contenga la media poblacional. En el ejemplo 4.3.2 diremos que se tiene un 95 % de confianza de que el nivel medio poblacional de dicha enzima se encuentre entre **19.37 y 24.63**.

Es importante tener en cuenta el error relativo, ya que nos indica el grado de precisión de nuestras estimaciones. Como

comprenderás, cuando se construye un IC, se hace con una confiabilidad determinada de forma arbitraria por el investigador, por lo tanto, si comparamos las estimaciones realizadas por dos investigadores sobre un mismo parámetro, utilizando el mismo nivel de confianza, no es lo mismo un error relativo del 10 % que del 15%, el primer caso ofrecerá mayor crédito que el segundo a pesar de que en ambos casos se tiene igual % de confianza que los IC contengan al parámetro.

Ya estás en condiciones de construir IC para la media.

3.3.2 IC para la diferencia entre las medias de dos poblaciones.

También se pueden construir IC para la diferencia entre las medias de dos poblaciones, para ello utilizamos el mismo procedimiento anterior pero adaptado a la distribución muestral de dicha diferencia.

Ya vimos que el estimador $(\bar{X}_1 - \bar{X}_2)$, proporciona una estimación insesgada de $(\mu_1 - \mu_2)$, que es la diferencia entre las medias de las poblaciones. La varianza del estimador es $(\sigma_1^2 / n_1) + (\sigma_2^2 / n_2)$. También aprendimos que este estimador sigue aproximadamente una distribución normal, de modo que en muchas ocasiones se emplea la teoría adecuada a las distribuciones normales para calcular IC para la diferencia entre las medias de dos poblaciones.

Cuando se conoce la varianza de las dos poblaciones, el IC al $100(1 - \alpha)\%$ para $(\mu_1 - \mu_2)$ está dado por:

$$(\bar{X}_1 - \bar{X}_2) \pm Z_{1-\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Esta fórmula se aplica si la variable se distribuye normal y aún cuando no es así pero las muestras son grandes.

Si muestras pequeñas y (o) no se conocen las varianzas poblacionales entonces hay dos maneras de construir los IC:

$$(\bar{X}_1 - \bar{X}_2) \pm t_{1-\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad \text{caso que se supone } \mathbf{varianzas}$$

diferentes.

$$(\bar{X}_1 - \bar{X}_2) \pm t_{1-\alpha/2} \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} \quad \text{caso que se suponen } \mathbf{varianzas}$$

iguales

En ambos casos se utiliza el percentil de la distribución t con $(n_1 + n_2 - 2)$ grados de libertad. En el caso en que se suponen varianzas iguales se calcula la varianza muestral de manera conjunta para ambas muestras y se denota por s_p^2 .

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

3.3.3 – Intervalos de confianza para una proporción.

De manera análoga al caso de la media, también se pueden construir intervalos de confianza para una proporción poblacional π .

Conoces además, que de forma general los IC se construyen mediante la siguiente fórmula:

Estimador \pm error de muestreo

Sabemos que la proporción muestral p , es un estimador insesgado de la proporción poblacional π . Vimos también que cuando el tamaño muestral es grande, es decir, np y $n(1-p)$ son mayores que 5, la distribución muestral de p es aproximadamente normal. Por lo tanto, se puede calcular un IC al $100(1-\alpha)\%$ de confiabilidad para π , empleando la siguiente fórmula:

$$p \pm Z_{(1-\alpha/2)} \sqrt{\frac{p(1-p)}{n}}$$

Este intervalo se interpreta de forma similar al caso de la media, tanto de forma probabilística como práctica.

Ejemplo 3.3.4

En una determinada región se tomó una muestra aleatoria de 125 individuos, de los cuales 12 padecían afecciones pulmonares.

Estimemos la proporción de afecciones pulmonares en dicha región. Primero calculamos la proporción muestral p , pues sabemos que será nuestra estimación puntual de π .

$$p = \frac{12}{125} = 0.096$$

Calculemos ahora un IC al 95% para la proporción poblacional de afecciones pulmonares.

$$p \pm e = 0.096 \pm 1.96 * \sqrt{\frac{0.096(1-0.096)}{125}} = 0.096 \pm 0.0516$$

(0.0444 ; 0.1476)

Veamos cuánto es el error relativo de nuestra estimación.

$$\frac{0.0516}{0.096} * 100 = 53.75\%$$

Como puedes, ver la amplitud del intervalo parece pequeña, sin embargo el error relativo es mayor que el 50 % de la estimación. En la figura 4.3.3 se representa este IC de forma gráfica.

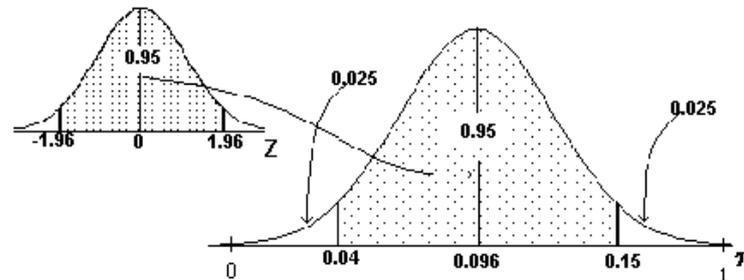


Figura 3.3.2 IC al 95% para la proporción poblacional de afecciones pulmonares.

Observa la equivalencia entre la distribución normal estándar y la distribución muestral de la proporción. Dentro de esos límites se encuentra el 95% de los valores estimados de π bajo el supuesto de muestreo sucesivo.

3.3.4 - IC para la diferencia entre dos proporciones.

En ocasiones se tiene interés en conocer la diferencia entre las proporciones de dos poblaciones ($\pi_1 - \pi_2$). Para ello el estimador puntual viene dado por ($p_1 - p_2$).

Imaginarás que para calcular un IC al $100(1-\alpha)$ % emplearemos la fórmula general antes vista:

Estimador \pm error de muestreo

Entonces en este caso será:

$$(p_1 - p_2) \pm Z_{1-\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

La interpretación de este intervalo se realiza de forma similar a los ya vistos.

Por ejemplo, en un estudio sobre prevalencia de hábitos tóxicos se obtuvo que las proporciones de fumadores de dos áreas de salud (P_1 y P_2) fueron de 0.23 y 0.15 respectivamente, para una diferencia de 0.08. A continuación construiremos un IC al 95% para dicha diferencia, si sabemos que los tamaños de muestra (n_1 y n_2) fueron de 350 y 475 sujetos, respectivamente.

$$0.08 \pm 1.96 \sqrt{\frac{0.23(-0.23)}{350} + \frac{0.15(-0.15)}{475}}$$

$$0.08 \pm 0.055 \Rightarrow 0.025 \text{ a } 0.135$$

Entonces podemos concluir, con una confiabilidad del 95%, que la diferencia entre las proporciones de fumadores de dichas áreas se encuentra entre un 2.5 y 13.5 %.

3.4 Pruebas de hipótesis.

Al igual que en el caso de la estimación, el propósito de las pruebas de hipótesis es ayudar al médico, investigador o administrador a tomar una decisión en torno a una población, al examinar una muestra de ella.

Pueden presentarse en la práctica, situaciones en las que exista una teoría preconcebida relativa a una característica de la población sometida a estudio. Tal sería el caso, por ejemplo si pensamos que un tratamiento nuevo puede tener un porcentaje de mejoría mayor que otro estándar, o cuando nos planteamos si los niños de distintas comunidades tienen la misma talla. Este tipo de circunstancias son las que nos llevan al estudio de la parte de la Estadística Inferencial que se recoge bajo el título genérico de **Contraste de Hipótesis**.

El contraste de hipótesis no es tan distinto a la estimación, de hecho es posible emplear los IC para llegar a las mismas conclusiones que se alcanzan con una prueba de hipótesis, como veremos más adelante.

Una hipótesis, se define simplemente como una proposición acerca de una o más poblaciones. En general las hipótesis se refieren a los parámetros de las poblaciones para las cuales se hace la proposición. Por medio de las pruebas de hipótesis se determina si tales proposiciones son compatibles o no con los datos disponibles.

Los investigadores se interesan en dos tipos de hipótesis: de investigación y estadísticas. Las primeras¹³ son aquellas suposiciones o conjeturas que motivan la investigación, éstas conducen directamente a las segundas, las que se establecen de forma tal que pueden ser evaluadas por medio de técnicas estadísticas adecuadas.

¹³ La mayor parte de las investigaciones que realizas, por el nivel de conocimientos que generan, sólo incluyen hipótesis de trabajo.

Implica, en cualquier investigación, la existencia de dos teorías o hipótesis estadísticas, que denominaremos hipótesis nula e hipótesis alternativa, que de alguna manera reflejarán esa idea a priori que tenemos (hipótesis de investigación) y que pretendemos contrastar con la "realidad".

De la misma manera aparecen, implícitamente, diferentes tipos de errores que podemos cometer durante el procedimiento. No podemos olvidar que, habitualmente, el estudio y las conclusiones que obtengamos para una población cualquiera, se habrán apoyado exclusivamente en el análisis de sólo una parte de ésta. De la probabilidad con la que estemos dispuestos a asumir estos errores, dependerá, por ejemplo, el tamaño de la muestra requerida.

Desarrollamos en esta sección los contrastes de hipótesis para los parámetros más usuales que venimos estudiando en las secciones anteriores: medias y proporciones, para una o dos poblaciones. Pero antes, veremos un algoritmo general para realizar pruebas de hipótesis que contiene 9 pasos¹⁴:

1. Datos¹⁵: Es necesario definir el tipo de datos que se analizará, pues de ello depende el tipo de prueba a aplicar. Se debe precisar si son variables cuantitativas o no, etc.
2. Suposiciones: De manera similar al proceso de estimación, se necesita evaluar el cumplimiento de una serie de suposiciones o supuestos que son importantes para decidir el tipo de prueba a realizar. Por ejemplo, respecto a la normalidad de la distribución de la población, igualdad de varianzas, independencia de las muestras, etc.
3. Hipótesis: Se trabaja con dos hipótesis estadísticas, como dijimos anteriormente, que deben enunciarse de manera explícita. La primera es la llamada **hipótesis nula**, es la que debe probarse y se designa como **H₀**. Suele denominarse como hipótesis de no diferencia, ya que es una proposición de no diferencia respecto a condiciones que se suponen ciertas en la población. En general se establece con el propósito de ser rechazada. En consecuencia, el complemento de la hipótesis del investigador se convierte en el enunciado de la hipótesis nula.

La otra hipótesis planteada es la llamada **hipótesis alternativa**, **H₁**, constituye el complemento de la nula, se plantea en correspondencia con la hipótesis del investigador. Esta

¹⁴ Modificado de Daniel 1997

¹⁵ Se refiere al tipo de variables.

hipótesis no es directamente sometida a prueba pero resulta conveniente plantearla puesto que es la más verosímil cuando la prueba estadística nos conduce a rechazar la H_0 . Si la H_0 no se rechaza, se dirá que los datos no proporcionan evidencia suficiente que cause el rechazo. De rechazarse, se dice que los datos disponibles no son compatibles con H_0 .

4. Estadígrafo de prueba: Es un estadístico que se puede calcular con los datos de la muestra. Como regla, existen muchos valores posibles que puede asumir el estadístico de prueba y el valor particular observado depende de la única muestra estudiada. Dependiendo de la magnitud del estadígrafo se procederá o no al rechazo de H_0 , como veremos más adelante. Una fórmula general para un estadígrafo de prueba, empleada en muchas pruebas de hipótesis (PH), es como sigue:

$$\text{Estadígrafo de prueba} = \frac{\text{estimador} - \text{parámetro supuesto}}{\text{error estándar del estimador}}$$

5. Distribución del estadígrafo de prueba: Hemos señalado, que la clave para la estadística inferencial es la distribución muestral. Veremos en cada caso particular qué distribución sigue el estadígrafo de prueba.
6. Regla de decisión: Todos los valores posibles que el estadígrafo de prueba puede asumir, constituyen puntos sobre el eje horizontal del gráfico de la distribución de dicho estadígrafo, y se dividen en dos grupos: los valores “compatibles” con la H_0 (región de aceptación) y los “no compatibles” con H_0 (región crítica o de rechazo). La regla de decisión señala que se debe rechazar la H_0 si el valor del estadígrafo de prueba, calculado a partir de los datos obtenidos en la muestra, es uno de los valores de la zona de rechazo, y que no se debe rechazar en caso contrario.

El intervalo de aceptación o más precisamente, de no rechazo de la hipótesis nula, se establece fijando una cantidad α , suficientemente pequeña denominada **nivel de significación**, de modo que la probabilidad de que el estadístico del contraste tome un valor fuera del mismo (**región crítica**) cuando la hipótesis nula es cierta sea igual a α . Esto se ha de entender como sigue:

Si H_0 es correcta el criterio de rechazo sólo se equivoca con probabilidad α , que es la probabilidad de que una muestra dé un valor del estadístico del contraste extraño (fuera del intervalo de aceptación).

La decisión de rechazar o no la hipótesis nula está al fin y al cabo basado en la elección de una muestra tomada al azar, y por tanto es posible cometer decisiones erróneas. Los errores que se pueden cometer se clasifican como sigue:

Error de tipo I: Es el error que consiste en rechazar H_0 cuando es cierta. La probabilidad de cometer este error es lo que anteriormente hemos denominado nivel de significación. Es una costumbre establecida el denotarlo siempre con la letra α . Suele tomarse un valor pequeño de α para hacer que la probabilidad de rechazar la H_0 sea pequeña; los valores que frecuentemente se utilizan son 0.01, 0.05 y 0.1.

Error de tipo II: Es el error que consiste en no rechazar H_0 cuando es falsa. La probabilidad de cometer este error la denotamos con la letra β .

En general, aunque se tome un valor pequeño para α , no se tiene control sobre β , a pesar de saber que en la mayoría de las situaciones prácticas es mayor que α . No obstante, nunca se sabe si cometimos uno de estos errores, cuando se rechaza o no la H_0 , ya que se desconoce el verdadero estado de las cosas a nivel poblacional.

Los errores de tipo I y II están relacionados del siguiente modo: Cuando α decrece β crece. Por tanto no es posible encontrar tests que hagan tan pequeños como queramos ambos errores simultáneamente. De este modo es siempre necesario privilegiar a una de las hipótesis, de manera que no será rechazada, a menos que su falsedad se haga muy evidente. En los contrastes, la hipótesis privilegiada es H_0 , que sólo será rechazada cuando la evidencia de su falsedad supere el umbral del $100(1 - \alpha)\%$.

Al tomar α muy pequeño tendremos que β se puede aproximar a uno. Lo ideal a la hora de definir un test es encontrar un compromiso satisfactorio entre α y β (aunque siempre a favor de H_0). La manera que tiene el investigador de reducir el valor de β y mantener α constante es aumentar el tamaño de la muestra.

Denominamos **potencia de un contraste** a la cantidad $1 - \beta$.

A continuación te muestro las posibles acciones que como investigador puedes emprender para varias condiciones de una prueba de hipótesis y las condiciones en que se produce cada tipo de error.

Condición de la hipótesis nula	Acción posible	
	No rechazar H_0	Rechazar H_0
Verdadera	Correcto Probabilidad $1 - \alpha$	Error tipo I Probabilidad α
Falsa	Error tipo II Probabilidad β	Correcto Probabilidad $1 - \beta$

En el momento de elegir una hipótesis privilegiada podemos en principio dudar entre si elegir una dada o bien su contraria. Criterios a tener en cuenta en estos casos son los siguientes:

- **Simplicidad científica:** A la hora de elegir entre dos hipótesis científicamente razonables, tomaremos como H_0 aquella que sea más simple.
- **Las consecuencias de equivocarnos:** Por ejemplo al juzgar el efecto que puede causar cierto tratamiento médico que está en fase de experimentación, en principio se ha de tomar como hipótesis nula aquella cuyas consecuencias por no rechazarla siendo falsa son menos graves.

7. Cálculo del estadígrafo de prueba: A partir de los datos contenidos en la muestra, se calcula un valor del estadígrafo de prueba y se compara con las regiones de aceptación y rechazo que ya fueron especificadas.
8. Decisión estadística: La decisión consiste en el rechazo o no de la H_0 , se rechaza H_0 si el valor del estadígrafo cae en la zona de rechazo.
9. Conclusión: Las conclusiones de una prueba de hipótesis debemos expresarlas en términos de rechazo o no rechazo de H_0 , o en los términos equivalentes de valor significativo¹⁶ y no significativo. Un resultado no significativo sólo indica que es compatible con la H_0 porque la discrepancia observada puede ser explicada por el error aleatorio, esto no demuestra que la H_0 sea cierta, simplemente que los datos no aportan suficiente

¹⁶ Un resultado significativo se refiere a diferencias no atribuibles al azar.

evidencia para dudar de la validez de H_0 . Por lo tanto ante este resultado debe mantenerse la H_0 hasta tanto se obtengan otras pruebas.

Por el contrario, un resultado estadísticamente significativo, indica que no es compatible con la H_0 porque es muy poco probable que la discrepancia observada haya sido producida por el error aleatorio del muestreo. En este caso se rechaza la H_0 por poco verosímil. Fíjate que se habla de rechazo y no rechazo, no siendo este último sinónimo de acepto.

Lo importante es ser muy cuidadoso a la hora de dar conclusiones, el que se rechace la H_0 no implica necesariamente que la alternativa sea verdadera, además tampoco se puede llegar a un juicio de causalidad por este motivo.

Además, la decisión estadística no debe interpretarse como definitiva, ella solo es una parte de la evidencia que influye sobre la decisión administrativa o clínica, como veremos más adelante.

A continuación estudiaremos pruebas de hipótesis específicas basadas en estos comentarios generales.

3.4.1 Pruebas de hipótesis a partir de medias.

Existen dos condiciones básicas en que realizamos PH a partir de medias: para **una sola población** y para **dos poblaciones**. Veremos cada caso por separado, a la vez que nos detendremos en las particularidades de cada una. Pero antes, debes conocer que las pruebas de hipótesis se pueden realizar de forma unilateral y bilateral, en dependencia de la forma en que son enunciadas las hipótesis nula y alternativa. Así, una PH bilateral es aquella en que sólo interesa conocer la existencia de diferencias, sin definir el sentido de éstas, como ocurre en el caso unilateral.

La media de una sola población.

Esta situación surge cuando al investigador le interesa probar que la media μ de una determinada variable en una población es igual o diferente a un valor determinado μ_0 . Estas pruebas pueden realizarse en tres condiciones diferentes que veremos a continuación:

1. La población¹⁷ se distribuye normal con varianza conocida.
2. La población se distribuye normal con varianza desconocida.
3. La población no se distribuye normal.

Aunque la teoría para las condiciones 1 y 2 se basa en que la población sigue una distribución normal, en la práctica es común aplicar este proceder aún cuando la población sólo está distribuida aproximadamente normal. Esto es satisfactorio siempre que la desviación de la normalidad sea moderada.

1- Población normal con varianza conocida.

Suponemos que $X \sim N(\mu, \sigma^2)$ donde σ^2 es conocida y queremos contrastar si es posible que μ (desconocida) sea en realidad cierto valor μ_0 fijado. Esto es un supuesto teórico que nunca se dará en la realidad pero servirá para introducir la teoría sobre contrastes.

Test de dos colas con varianza conocida :

El test se escribe entonces como:

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

Estadígrafo de prueba:

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Si H_0 es cierta, este estadígrafo tiene una distribución normal estándar. Le llamaremos al valor de Z obtenido en la muestra Z observado y lo denotaremos como Z_0

Fijemos ahora el nivel de significación α , de manera que queden definidas las zonas de aceptación y de rechazo (crítica) respectivamente. Así, tomaremos como región crítica C , a los valores que son muy extremos y con probabilidad α en total, o sea:

$$P(Z_0 \leq Z_{\alpha/2}) = \alpha / 2$$

$$P(Z_0 \geq Z_{1-\alpha/2}) = \alpha / 2$$

Recuerda que $Z_{\alpha/2} = -Z_{1-\alpha/2}$

Lo anterior implica que:

¹⁷ Suele decirse que la población se distribuye normal, pero en realidad nos referimos a la distribución de la variable en la población. Así aparece en todos los textos.

$P(-Z_{1-\alpha/2} \leq Z_0 \leq Z_{1-\alpha/2}) = 1 - \alpha$ siendo esta la zona de aceptación.

Entonces la región crítica **C** consiste en:

$$\mathbf{C = Z_0, \text{ tal que } |Z_0| > Z_{1-\alpha/2}}$$

Luego rechazaremos la hipótesis nula si:

$$\mathbf{|Z_0| > Z_{1-\alpha/2}}$$

Aceptando en consecuencia la hipótesis alternativa (figura 3.4.1).

Puedes observar en la figura 3.4.1, que la región de rechazo de la hipótesis nula es la sombreada. Se rechaza H_0 cuando el estadígrafo Z_0 toma un valor comprendido en la zona sombreada de la gráfica pequeña, $N(0, 1)$, o equivalentemente, cuando el estadígrafo \bar{x} toma un valor en la zona sombreada de la gráfica grande, $N(\mu_0, \sigma^2)$.

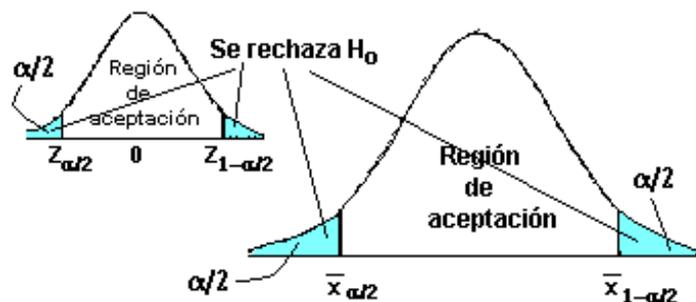


Figura 3.4.1 representación gráfica de las zonas de aceptación y rechazo de la H_0

Tests unilateral (de una cola) con varianza conocida

Consideremos un contraste de hipótesis donde ahora las hipótesis se enuncian:

$$\mathbf{H_0: \mu = \mu_0} \quad \text{A veces se escribe} \quad \mathbf{H_0: \mu \geq \mu_0}$$

$$\mathbf{H_1: \mu < \mu_0} \quad \mathbf{H_1: \mu < \mu_0}$$

El estadígrafo de prueba es el mismo del caso anterior, pues la H_0 es igual.

Como región crítica consideraremos aquella formada por los valores extremadamente bajos de Z_0 , con probabilidad α , es decir:

$$\mathbf{P(Z_0 \leq Z_\alpha) = \alpha \text{ implica } P(Z_0 \geq Z_\alpha) = 1 - \alpha}$$

Entonces la región de aceptación, o de modo más correcto, de no rechazo de la hipótesis nula es (figura 3.4.2): $\mathbf{Z_0 \geq Z_\alpha}$; y por consiguiente, se rechaza H_0 si: $\mathbf{Z_0 < Z_\alpha}$

Así pues, se rechaza la hipótesis nula, cuando uno de los estadísticos Z o \bar{x} toma un valor en la zona sombreada de sus gráficas respectivas.

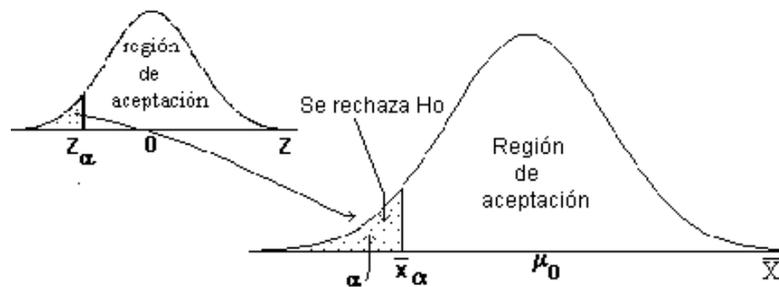


Figura 3.4.2 Zonas de aceptación y rechazo para test unilateral.

Es evidente que si en el contraste de significación, hubiésemos tomado como hipótesis alternativa su contraria, es decir:

$$\begin{array}{ll} \mathbf{H_0: \mu = \mu_0} & \text{O también} & \mathbf{H_0: \mu \leq \mu_0} \\ \mathbf{H_1: \mu > \mu_0} & & \mathbf{H_1: \mu > \mu_0} \end{array}$$

Por simetría con respecto al caso anterior, la región donde no se rechaza la hipótesis nula es (véase la figura 3.4.3 y contrástese con la 3.4.2): $Z_0 \leq Z_{1-\alpha}$, lo que implica que se rechaza H_0 si: $Z_0 > Z_{1-\alpha}$

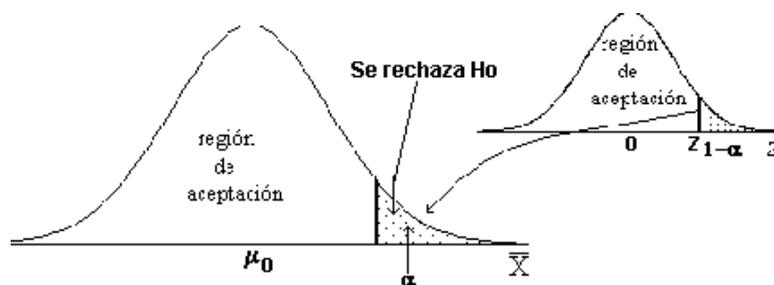


Figura 3.4.3 - Regiones de aceptación y rechazo para el test unilateral contrario al anterior.

2- La variable se distribuye normal con varianza desconocida.

Test de dos colas con varianza desconocida

Sea $X \sim N(\mu, \sigma^2)$ donde ni μ ni σ^2 son conocidos y queremos realizar el contraste

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

Al no conocer σ^2 va a ser necesario estimarlo a partir de su estimador insesgado: la varianza muestral, S^2 . Por ello la distribución del estimador del contraste será una t de Student, con $n-1$ grados de libertad.

Si H_0 es cierta implica:

$$T_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1} \text{ gl}$$

Consideramos como región crítica C , a las observaciones de T_0 extremas:

$$P(T_0 \leq t_{n-1, \alpha/2}) = \alpha/2$$

$$P(T_0 \geq t_{n-1, 1-\alpha/2}) = \alpha/2$$

Recuerda que la distribución t es simétrica respecto a la media y tiene media 0, por lo que:

$$t_{n-1, \alpha/2} = -t_{n-1, 1-\alpha/2}$$

Todo lo anterior implica que:

$$P(t_{n-1, \alpha/2} \leq T_0 \leq t_{n-1, 1-\alpha/2}) = 1 - \alpha$$

o sea

$$C = (T_0 < -t_{n-1, \alpha/2} \text{ ó } T_0 > t_{n-1, 1-\alpha/2})$$

De forma similar al caso anterior, para dar una forma homogénea a todos los contrastes de hipótesis es costumbre denominar al valor del estadístico del contraste calculado sobre la muestra como valor observado y a los extremos de la región crítica, como valores teóricos, en este caso T_t . Definiendo entonces:

$$T_t = t_{n-1, 1-\alpha/2}$$

El resultado del contraste es el siguiente (figura 3.4.4):

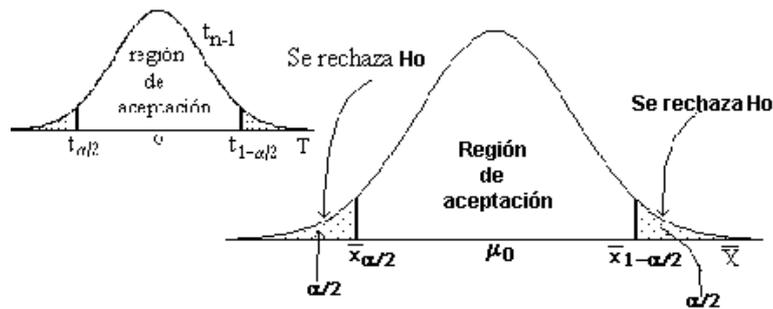


Figura 3.4.4 - Región crítica para el contraste bilateral de una media con varianza desconocida.

Entonces, si: $|T_o| \leq T_t \Rightarrow$ no rechazamos H_0
 $|T_o| > T_t \Rightarrow$ rechazamos H_0

Tests de una cola con varianza desconocida

Si realizamos el contraste

$$H_0: \mu = \mu_0$$

$$H_1: \mu < \mu_0$$

El estadígrafo de prueba es el mismo que en el caso bilateral, pero el valor teórico se modifica al igual que la zona crítica (figura 3.4.5).

$$T_t = t_{n-1, \alpha}$$

De forma tal que: $T_o \geq T_t \Rightarrow$ no rechazamos H_0
 $T_o < T_t \Rightarrow$ rechazamos H_0 y aceptamos H_1

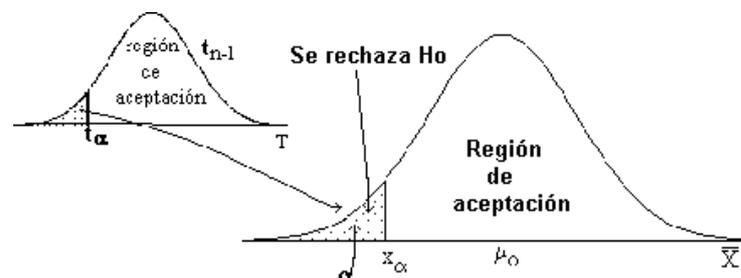


Figura 3.4.5 - Región crítica para uno de los contrastes unilaterales de una media con varianza desconocida.

Para el contraste contrario: $H_0: \mu = \mu_0$

$$H_1: \mu > \mu_0$$

Definimos T_o y T_t como anteriormente y el criterio a aplicar es

(véase la figura 3.4.6): $T_t = t_{n-1, 1-\alpha}$

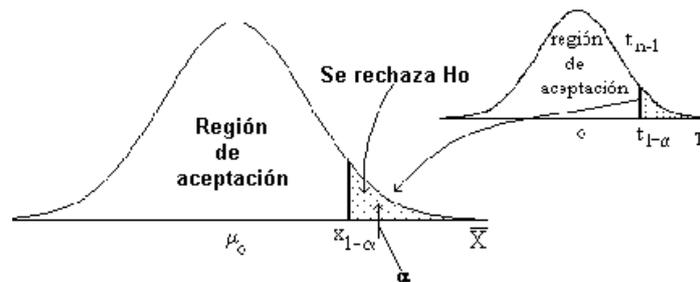


Figura 3.4.6 - Región crítica para el contraste unilateral de una media contrario al anterior. Varianza desconocida.

De aquí que: **Si $T_o \leq T_t \Rightarrow$ no rechazamos H_o**
Si $T_o > T_t \Rightarrow$ rechazamos H_o y aceptamos H_1

Ejemplo 3.4.1

Conocemos que la talla (X) de los individuos de una ciudad, se distribuye aproximadamente normal. Deseamos contrastar con un nivel de significación de $\alpha=0.05$ si la talla media es diferente de 174 cm. Para ello nos basamos en el estudio de una muestra de 25 personas. Se obtuvo los siguientes resultados:

$$\bar{x} = 170\text{cm y } s = 10\text{cm}$$

Solución:

El contraste que se plantea es:

$$H_o: \mu = 174 \text{ cm}$$

$$H_1: \mu \neq 174 \text{ cm}$$

Como puedes ver, se trata de un test con varianza desconocida, planteado de forma bilateral, entonces el estadígrafo de prueba será:

$$T_o = \frac{170 - 174}{10 / \sqrt{25}} = \frac{-4}{2} = -2 \sim t_{24} \text{ gl}$$

La regla de decisión estará dada por el valor teórico del estadígrafo:

$$T_t = t_{24, 1 - \alpha / 2} = t_{24, 0.975} = 2.06^{18}$$

Comparemos ahora el valor observado con el teórico:

$$|T_o| = 2 < T_t = 2.06 \Rightarrow \text{No rechazamos } H_o$$

Por lo tanto, aunque podamos pensar que ciertamente el verdadero valor de μ no es 174, no hay una evidencia suficiente para rechazar esta hipótesis al nivel de confianza del 95%. En la figura 3.4.7 vemos que el valor de T_o no está en la región crítica (aunque ha quedado muy cerca), por tanto al no ser la evidencia en

¹⁸ Revisa la tabla B del Apéndice.

contra de H_0 suficientemente significativa, esta hipótesis no se rechaza.

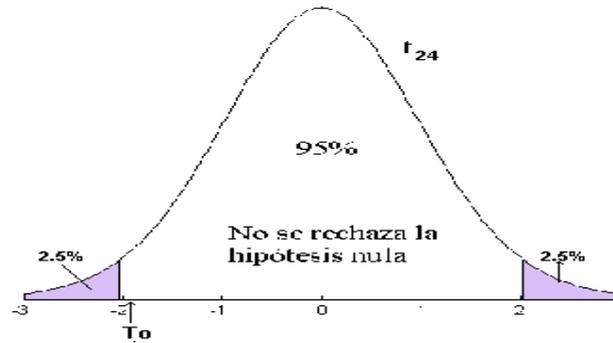


Figura 3.4.7 – Región de rechazo de la hipótesis nula: $\mu = 174$, del ejemplo 4.4.1

Ejemplo 3.4.2

Consideramos el mismo ejemplo anterior. Visto que no hemos podido rechazar el que la talla media de la población sea igual a 174 cm, deseamos realizar el contraste sobre si la talla media es menor de 174 cm.

Solución:

Ahora el contraste es **$H_0: \mu = 174 \text{ cm}$**

$H_1: \mu < 174 \text{ cm}$

De nuevo la técnica a utilizar consiste en suponer que H_0 es cierta y ver si el valor que toma el estadígrafo T es aceptable bajo esta hipótesis, con un nivel de confianza del 95%. Se aceptará la hipótesis alternativa (y en consecuencia se rechazará la hipótesis nula) si:

$$T_0 < t_{24, \alpha} = -t_{24, 1 - \alpha} = t_{24, 0.95} = -1.71^{19}$$

Recordamos que el valor de T_0 obtenido fue de:

$$T_0 = -2 < T_t = -1.71$$

Por ello hemos de rechazar la hipótesis nula y por tanto, aceptar la alternativa (véase la figura 3.4.8).

¹⁹ Revisar la tabla B del Apéndice.

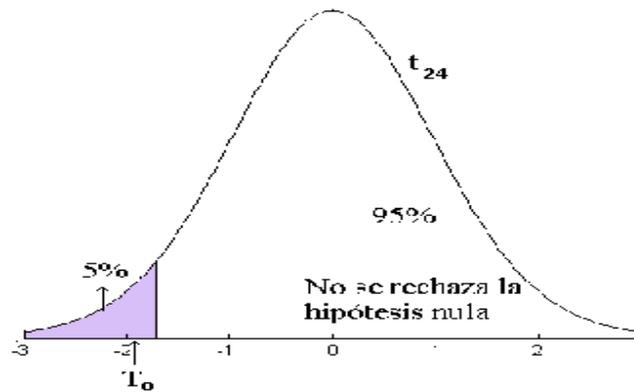


Figura 3.4.8 – Región crítica para el estadígrafo t del ejemplo 4.4.2.

Podemos observar en el gráfico, que el valor T_0 está en la región crítica, por tanto existe una evidencia significativa en contra de H_0 , y a favor de H_1 .

Es importante observar este hecho curioso: Mientras que en el ejemplo anterior no existía una evidencia significativa para decir que $\mu \neq 174$ cm, el "simple hecho" de plantearnos un contraste que parece el mismo pero en versión unilateral nos conduce a rechazar de modo significativo que $\mu = 174$ cm y aceptamos que $\mu < 174$ cm. Esto no es raro que suceda, de hecho se plantea que al usar pruebas unilaterales las diferencias encontradas suelen ser mucho más significativas que si se aplica el test bilateral. Por ello, es aceptable la actitud conservadora de muchos investigadores que sistemáticamente emplean contrastes bilaterales.

3- Población que no presenta una distribución normal.

Si, como ocurre con frecuencia, la muestra en la cual se basa la prueba de hipótesis acerca de la media de una población proviene de una distribución desconocida o diferente de la normal, si la muestra es grande, mayor o igual que 30, es posible aplicar el teorema del límite central y usar el mismo estadígrafo Z visto con anterioridad, incluso en el caso en que no conocemos la varianza se puede sustituir ésta por la varianza muestral.

$$Z = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \sim N(\mu_0, s / \sqrt{n})$$

El resto del contraste se realiza de forma similar a lo visto anteriormente.

Observaciones

Es necesario que aclaremos algunos aspectos antes de continuar, los cuales serán válidos para el resto de la sección.

1. **Test Bilateral Vs Unilateral:** Una prueba se denomina **bilateral o de dos colas**, cuando la hipótesis alternativa está planteada sin especificar el sentido de la diferencia. Mientras que en el caso **unilateral o de una cola** se especifica el sentido o la dirección de la diferencia esperada. En el primer caso la zona de rechazo está dividida en dos partes, a ambos extremos de la distribución siendo las probabilidades a cada lado igual a $\alpha/2$, para que entre sí sumen un total de α . Mientras que en el caso unilateral la zona crítica se encuentra hacia un extremo o el otro de dicha distribución, cuya área es igual a α . La decisión de cuál test emplear depende del objetivo de la investigación.
2. **Valor de p (probabilidad) asociado al estadígrafo:** Para llegar a una conclusión sobre el resultado de la prueba podemos utilizar, además del valor directo del estadígrafo de prueba, la llamada **p** asociada al estadígrafo, que seguro has oído hablar de ella. El valor **p** para la prueba, es la probabilidad de obtener, cuando H_0 es cierta, un valor del estadígrafo mayor o igual (según la dirección de la diferencia) que el observado a partir de la muestra. Suele usarse con mayor frecuencia que el valor del estadígrafo, incluso se exige en la publicación de artículos científicos de algunos editores. Por supuesto que las conclusiones son equivalentes. En este caso se compara el valor de **p** con el valor de α prefijado y si:

$$p < \alpha \Rightarrow \text{se rechazará } H_0$$

Este valor lo puedes obtener de las tablas de las distribuciones teóricas del estadígrafo, pero los programas estadísticos de computación suelen darlo con exactitud.

3. **Pruebas de hipótesis por medio de intervalos de confianza:** Se pueden realizar pruebas de hipótesis empleando para ello los intervalos de confianza vistos en secciones anteriores, cuyos resultados son equivalentes al test de hipótesis tradicional. Consisten básicamente, en calcular un IC para el parámetro que se desea contrastar, empleando el mismo nivel de significación

que en el test, entonces la regla de decisión será basada en si el Intervalo contiene o no al valor del parámetro hipotetizado, de ser así no se puede rechazar la H_0 , en caso de no contenerlo se rechazará la H_0 y se aceptará la hipótesis alternativa, siendo la probabilidad de cometer el error tipo I igual a α . Recuerda que el parámetro a contrastar puede ser un valor único o la diferencia entre dos valores. En general, cuando se prueba una hipótesis nula por medio de un intervalo de confianza bilateral, se rechaza H_0 en el nivel α de significación, si el parámetro supuesto no está contenido dentro del intervalo de confianza del $100(1-\alpha)\%$.

4. **Elección del test**²⁰: Como estudiamos, se pueden probar las mismas hipótesis empleando diferentes estadígrafos: prueba Z y prueba t, la escogencia entre uno de ellos dependerá del cumplimiento de las suposiciones²¹ necesarias para cada caso. No lo olvides.

Hemos tratado de ser explícitos y a la vez de no complicar mucho las cosas para facilitarte la comprensión del contenido, aún así te percatarás que no es un tema sencillo. Lo explicado hasta el momento podrás aplicarlo a otras pruebas de hipótesis que veremos a continuación.

Pruebas para las medias de dos poblaciones

La prueba de hipótesis que comprende la diferencia entre las medias de dos poblaciones, se utiliza con mayor frecuencia para determinar si es razonable o no concluir que las dos son distintas entre sí. Al igual que en el caso de una sola población se distinguen diferentes situaciones:

1. Muestras pareadas
2. Muestras independientes

Ambas situaciones serán abordadas por separado y teniendo en cuenta otras particularidades.

1- Muestras pareadas.

Se dice que dos muestras están pareadas si la probabilidad de que la variable en estudio tome un valor en el i ésimo sujeto de la

²⁰ El término test es un anglicismo comúnmente empleado en la literatura consultada, por lo que decidimos utilizarlo en este contexto.

²¹ El cumplimiento de los supuestos puede verificarse mediante la realización de pruebas estadísticas, o sencillamente darlos por sentado.

muestra 2, depende del valor del *i*ésimo sujeto de la muestra 1. Te aclaro, el objetivo de seleccionar muestras pareadas es eliminar al máximo las fuentes de variación por medio de la formación de parejas similares con respecto a tantas variables como sea posible, pero nunca respecto a la variable de estudio. Así se pueden formar muestras pareadas de diferentes formas, por ejemplo, todos los estudios donde se compara observaciones de una variable en dos o más momentos diferentes sobre los mismos individuos, cada medición constituye una muestra y el interés está en la diferencia antes – después. También, en los estudios de casos y testigos suelen formarse parejas del mismo sexo o grupos de edades, etc.

En estos casos, en lugar de llevarse a cabo el análisis con las observaciones individuales, se puede utilizar como variable de interés la diferencia entre los pares individuales de observaciones. Las hipótesis se enuncian generalmente de la siguiente forma:

$$H_0: \mu_1 = \mu_2 \quad \Rightarrow \quad \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 \neq \mu_2$$

Entonces, para cada pareja se registra la diferencia **d**, de esta forma es posible calcular la diferencia media \bar{d} , como la media aritmética de las diferencias observadas y la desviación estándar de dichas observaciones s_d . Así, el promedio poblacional de las diferencias se denota por μ_d y es igual a la diferencia entre $(\mu_1 - \mu_2)$, que en este caso es cero.

Estadígrafo de prueba:

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$$

Si H_0 es cierta el estadígrafo se distribuye *t* de Student con $n - 1$ grados de libertad.

La regla de decisión: se rechaza H_0 si: $|T_o| > T_{n-1, 1-\alpha/2}$
Este es el caso bilateral que es el recomendado.

Ejemplo 3.4.3²²

A continuación se muestran los valores correspondientes al puntaje de ansiedad para 10 pacientes recibiendo un medicamento (M) y un placebo (P) en orden aleatorio.

²² Este es un ejemplo didáctico, en la práctica puede ser abordado por una prueba no paramétrica

Pte.	Puntaje de ansiedad		Diferencia d = M - P
	M	P	
1	19	22	-3
2	11	18	-7
3	14	17	-3
4	17	19	-2
5	23	22	1
6	11	12	-1
7	15	14	1
8	19	11	8
9	11	19	-8
10	8	7	1

Calculemos la diferencia promedio y su desviación estándar:

$$\bar{d} = \frac{\sum d_i}{n} = \frac{-13}{10} = -1.3 \qquad s^2 = \frac{\sum (d_i - \bar{d})^2}{n-1} = \frac{186.1}{9} = 20.68$$

Calculemos el estadígrafo de prueba

$$T_o = \frac{\bar{d} - \mu_0}{s/\sqrt{n}} = \frac{-1.3 - 0}{4.55/\sqrt{10}} = \frac{-1.3}{1.439} = -0.9$$

Busquemos en la tabla B del Apéndice el valor teórico de T:

$$T_t = t_{n-1, 1-\alpha/2} = t_{9, 0.975} = 2.821$$

Como puedes ver $|T_o| = 0.9 < T_t = 2.821$

Por lo tanto no se puede rechazar H_o , podemos concluir que no existen evidencias que hagan pensar que el medicamento sea efectivo para disminuir la ansiedad a pesar de que las diferencias parecen favorecer al medicamento.

Observaciones:

Este caso es bastante frecuente en la práctica. Cuando el tamaño muestral es grande, $n > 30$, la distribución t puede aproximarse a la normal, no importa la distribución de la variable en la población ni tampoco es necesario conocer la varianza poblacional. De esta forma, siempre que el tamaño muestral lo

permita, se puede sustituir en la regla de decisión el percentil de la distribución t por el correspondiente de la normal estándar. A continuación te muestro una guía para la regla de decisión en caso bilateral y unilateral.

Tabla 3.4.1- regla de decisión PH para dos medias.

Hipótesis alternativa	Regla de decisión Rechazar Ho si:
$\mu_d > 0$ ($\mu_1 > \mu_2$)	$ T_o > Z (1 - \alpha)$
$\mu_d < 0$ ($\mu_1 < \mu_2$)	$ T_o < Z (1 - \alpha)$
$\mu_d \neq 0$	$ T_o > Z (1 - \alpha / 2)$

2- Muestras independientes

Suposiciones generales: las muestras son aleatorias e independientes, la variable es continua y se distribuye normal, además las observaciones son independientes.

Hipótesis: **H₀: $\mu_1 = \mu_2 \Rightarrow \mu_1 - \mu_2 = 0$**

H₁: $\mu_1 \neq \mu_2$

La hipótesis alternativa puede enunciarse también en forma unilateral:

H₁: $\mu_1 > \mu_2 \Rightarrow \mu_1 - \mu_2 > 0$ ó H₁: $\mu_1 < \mu_2 \Rightarrow \mu_1 - \mu_2 < 0$

Si las varianzas poblacionales son conocidas el estadígrafo de prueba es:

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \text{ se distribuye normal estándar}$$

En el numerador no se incluye μ_d porque bajo Ho $\mu_d = 0$.

En el caso de varianzas desconocidas hay dos posibilidades:

- Varianzas desconocidas pero se supone que son iguales
- Varianzas desconocidas pero se supone que son diferentes.

El estadígrafo de prueba en cada caso es:

$$a) T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad \text{donde } s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Este estadígrafo se distribuye t de Student con $(n_1 + n_2 - 2)$ grados de libertad

$$b) d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Se distribuye t de Student con ν (nu) grados de libertad, pero si el tamaño de las muestras es grande se distribuye normal estándar; Que suerte! porque la fórmula para calcular los grados de libertad se las trae.

En cada caso la regla de decisión puedes enunciarla a partir de la tabla 4.4.1, solo tienes que adecuarlo a la distribución específica de cada estadígrafo.

3.4.2 Pruebas de hipótesis a partir de proporciones.

Las pruebas de hipótesis a partir de proporciones se realizan casi en la misma forma utilizada cuando nos referimos a las medias, cuando se cumplen las suposiciones necesarias para cada caso. Pueden utilizarse pruebas unilaterales o bilaterales dependiendo de la situación particular.

La proporción de una población

Las hipótesis se enuncian de manera similar al caso de la media.

$$H_0: \pi = \pi_0$$

$$H_1: \pi \neq \pi_0$$

En caso de que la muestra sea grande $n > 30$, el estadígrafo de prueba es:

$$Z = \frac{p - \pi_0}{\sqrt{\frac{p(1-p)}{n}}} \quad \text{se distribuye normal estándar.}$$

Regla de decisión: se determina de acuerdo a la hipótesis alternativa (si es bilateral o unilateral), lo cual puedes fácilmente hacerlo auxiliándote de la tabla 3.4.1.

En el caso de muestras pequeñas se utiliza la distribución Binomial. No lo abordaremos por ser complicado y poco frecuente su uso.

Diferencia entre las proporciones de dos poblaciones

La situación más frecuente es suponer que existen diferencias entre las proporciones de dos poblaciones, para ello suelen enunciarse las hipótesis de forma similar al caso de las medias:

$$H_0: \pi_1 = \pi_2 \Rightarrow \pi_1 - \pi_2 = 0$$

$$H_1: \pi_1 \neq \pi_2$$

Puede la hipótesis alternativa enunciarse unilateralmente.

El estadígrafo de prueba para el caso de **muestras independientes**:

$$Z = \frac{p_1 - p_2}{\sqrt{pq \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \text{ donde } p = \frac{a_1 + a_2}{n_1 + n_2}$$

Siendo a_1 y a_2 , el número de sujetos con la característica objeto de estudio en las muestras 1 y 2 respectivamente, es decir, en vez de calcular la varianza para cada muestra, se calcula una **p** conjunta para ambas muestras bajo el supuesto que no hay diferencias entre ambas proporciones y así se obtiene la varianza conjunta. Recuerda que **q = 1-p**.

Está de más que te diga que este estadígrafo se distribuye normal estándar.

La regla de decisión se determina de manera similar a los casos ya vistos anteriormente.

Si las **muestras son pareadas** nuestro interés no estará en las observaciones individuales sino en la pareja, es decir, tendremos n parejas de sujetos, para cada uno de los sujetos registramos la presencia o ausencia de cierta característica A que se desea estudiar, por ejemplo enfermo o no, así tendremos cuatro posibilidades:

Muestra		Número de pares
1	2	
A	A	a
A	No A	b
No A	A	c
No A	No A	d

Esta misma información podemos resumirla en una tabla de contingencia de dos filas y dos columnas, llamada tabla de 2x2 y muy empleada en investigación epidemiológica, como verás en la parte II del texto.

Tabla 2x2 para muestras pareadas

Muestra 1	Muestra 2		Total
	A	No A	
A	a	b	a+b
No A	c	d	c+d
Total	a+c	b+d	n

De esta forma podemos calcular las proporciones de sujetos con la característica en cada muestra:

$$p_1 = \frac{a+b}{n} \qquad p_2 = \frac{a+c}{n}$$

El objetivo de la prueba es comparar estas dos proporciones, como estimadores insesgados de las respectivas proporciones poblacionales π_1 y π_2 . En este caso, la diferencia entre estas dos proporciones depende de aquellos pares discrepantes, es decir, los pares donde solo uno de los miembros de la pareja presenta la característica.

$$p_1 - p_2 = \frac{a+b}{n} - \frac{a+c}{n} = \frac{b-c}{n}$$

Las hipótesis se enuncian de forma similar al caso de muestras independientes.

$$\mathbf{H_0: \pi_1 = \pi_2 \Rightarrow \pi_1 - \pi_2 = 0}$$

$$\mathbf{H_1: \pi_1 \neq \pi_2}$$

Recuerda que la H_1 también puede plantearse de forma unilateral. Si la H_0 es cierta entonces $b - c = 0$ y el estadígrafo de prueba es:

$$Z = \frac{b - 1/2(b+c)}{1/2\sqrt{b+c}}$$

Este estadígrafo se distribuye normal estándar si el tamaño de muestra es grande, en caso de muestras pequeñas se emplea la prueba no paramétrica²³ de Mc Nemar, cuyo estadígrafo de prueba es:

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} \text{ se distribuye } \chi^2 \text{ con 1 grado de libertad.}$$

La regla de decisión depende del planteamiento de la H_1 , sea bilateral o unilateral. La puedes construir basándote en lo que has aprendido.

Observaciones:

1. Aunque te parezca absurdo, debo aclararte que las pruebas de hipótesis se realizan sobre los parámetros poblacionales desconocidos, es decir, sólo tiene sentido realizarlas cuando se estudia una muestra de la población objeto y deseamos hacer inferencias hacia el total poblacional. Si estudiaste al total de los elementos de tú población objeto (definida de acuerdo a los objetivos de tú investigación), no tiene sentido realizar PH ni otro tipo de inferencia.
2. Antes de realizar una prueba de hipótesis, debes revisar cuidadosamente las características de los datos (naturaleza de las variables), la forma de selección de la muestra y su tamaño, en fin, valorar el cumplimiento de los supuestos necesarios para aplicar la prueba adecuada a cada caso. Fijando el nivel de significación antes de realizar la prueba y no después de obtener el resultado, al igual que debes valorar seriamente si debes enunciar el problema de forma bilateral o unilateral antes de realizar la prueba. Violar el cumplimiento de los supuestos implica que la prueba pierda potencia, pudiendo no encontrarse diferencias cuando realmente las hay o lo contrario.
3. Existen software que realizan estas y otras muchas pruebas de hipótesis, al alcance de cualquier persona, esto trae consigo el

²³ Una prueba de hipótesis no paramétrica, es aquella en que no es necesario el cumplimiento de las suposiciones propias de las pruebas paramétricas que hemos visto hasta el momento (distribución normal de la población, muestras aleatorias, independencia de las observaciones, etc.), por lo tanto para cada prueba que hemos visto existe su análoga no paramétrica que podrás estudiar en el próximo curso.

uso y “abuso” indiscriminado de las mismas sin un conocimiento sólido del basamento estadístico de cada prueba. Te recomiendo no te sumes a la lista de irresponsables que andan por ahí haciendo de las suyas con los números, si no estás seguro de cual es la prueba adecuada recurre a los servicios de un bioestadístico.

4. En el caso que contrastas la diferencia entre dos parámetros, sean medias o proporciones, debes tener en cuenta que el signo del estadígrafo te da la dirección de la diferencia. Si el valor del estadígrafo observado es menor que cero (negativo), quiere decir que el parámetro 1 es menor que el 2 (parámetro1 – parámetro2), esto lo tendrás en cuenta al dar las conclusiones.
5. Por último, ningún test estadístico supera el sentido común y la responsabilidad profesional, por lo tanto las conclusiones deben basarse no sólo en un resultado estadísticamente significativo logrado a toda costa de artificios matemáticos, sino que depende de todo un conjunto de análisis clínico, epidemiológico, económico, entre otros aspectos que deben ser adecuadamente balanceados teniendo en cuenta la razón perjuicio-beneficio.

Resumen

En este tema estudiaste que:

1. La estimación y las pruebas de hipótesis constituyen las dos partes esenciales de la estadística inferencial. La estimación se puede realizar tanto de forma puntual como por intervalos, llamados intervalos de confianza. Por su parte las pruebas de hipótesis pueden realizarse empleando un estadígrafo específico para cada caso o mediante la construcción de un intervalo de confianza.
2. El concepto de distribución muestral de un estadístico se ha definido como la distribución de frecuencias de los valores obtenidos al aplicar el estadístico a todas las muestras posibles de tamaño n extraídas aleatoriamente de una población con parámetros conocidos. Como se trata de muestras extraídas aleatoriamente, lo más frecuente es que los valores promedio observados (proporción p y media \bar{x}) sean valores cercanos a los verdaderos parámetros en la población (proporción π y media μ). No obstante el error aleatorio propio del muestreo

produce algunas muestras con valores de p y \bar{x} alejadas de π y μ .

3. La fuente principal de variabilidad de la distribución muestral no es la variabilidad de las observaciones en la población de sujetos, sino el error aleatorio del muestreo. Por esta razón, la raíz cuadrada de la varianza del estimador no se llama desviación estándar, sino error estándar del estimador. Se demuestra que el error estándar disminuye en la medida que aumentamos el tamaño muestral.
4. La construcción de intervalos de confianza para los parámetros poblacionales, parten del conocimiento de una muestra concreta de tamaño n , para la que se calculan los valores de interés p o \bar{x} , y resuelve un problema de estimación al inducir los límites del intervalo que tiene una probabilidad grande $(1-\alpha)$ de contener al verdadero parámetro poblacional μ o π .
5. La fórmula general para construir un intervalo de confianza para un parámetro cualquiera de la población es:
 Estimador \pm (coeficiente de confiabilidad) \times (error estándar del estimador)
6. Las pruebas estadísticas encaminadas a verificar la veracidad de una hipótesis estadística se denominan pruebas de hipótesis. La hipótesis a contrastar recibe el nombre de hipótesis nula y se simboliza por H_0 . Junto a ésta se plantea siempre su hipótesis complementaria, que recibe el nombre de hipótesis alternativa y se simboliza por H_1 . Esta última no es sometida directamente a prueba pero resulta conveniente plantearla puesto que es la más verosímil cuando la prueba nos conduce a rechazar la H_0 .
7. Las pruebas de hipótesis son construidas en función de la hipótesis nula. Por consiguiente, deberemos expresar sus conclusiones en términos de rechazo o de no rechazo de H_0 , o en los términos equivalentes de valor significativo o no significativo estadísticamente. No se deben emplear los términos acepto porque implican demostración y un resultado significativo no es una demostración de la veracidad de la H_0 .
8. Un resultado no significativo estadísticamente sólo indica que es compatible con la hipótesis nula porque la discrepancia observada puede ser explicada por el error aleatorio del muestreo. Pero la significación estadística depende tanto de la

magnitud de la diferencia observada como del tamaño de la muestra estudiada. Estudios con muestras pequeñas difícilmente obtendrán significación estadística aunque las diferencias sean importantes. De la misma manera, estudios con muestras grandes obtendrán significación estadística aunque las diferencias sean poco importantes a nivel clínico o epidemiológico. Por ello, afirmar simplemente que existe o no significación estadística aporta poca información en sí misma.

9. El intervalo de confianza para un determinado parámetro, constituye un conjunto aceptable de hipótesis, y proporciona información sobre la magnitud de dicho parámetro que ayuda al clínico o al epidemiólogo experto a valorar la importancia clínica de ese resultado.
10. Al hacer pruebas de hipótesis no estamos exentos de cometer errores. El error que cometemos al rechazar la H_0 cuando es verdadera se conoce como error tipo I y la probabilidad de cometerlo es igual al nivel de significación α . Mientras que el error que cometemos al no rechazar la H_0 cuando H_1 es verdadera se denomina error tipo II y la probabilidad de cometerlo es igual a β . Raras veces podemos cuantificar la magnitud de β , pero a mayor tamaño muestral garantizamos que éste será pequeño. Por su parte, α suele fijarse como un valor pequeño entre 1 y 10 %, siendo el valor más frecuente 5% (0.05).
11. Tabla resumen sobre regla de decisión en pruebas de hipótesis:

Hipótesis alternativa	Regla de decisión Se rechaza H_0 si:
Bilateral ($a \neq b$)	$ E_o > \text{Percentil } 1 - \alpha / 2$
Unilateral ($a > b$)	$E_o > \text{Percentil } 1 - \alpha$
Unilateral ($a < b$)	$E_o < \text{Percentil } 1 - \alpha$

Donde $|E_o|$ es el valor modular del estadígrafo observado y nos referimos al percentil correspondiente a la distribución del estadígrafo de prueba bajo el supuesto de H_0 verdadera

Ejercitación

NOTA: EN TODOS LOS PROBLEMAS QUE SIGUEN A CONTINUACIÓN, SE SUPONE QUE LAS MUESTRAS HAN SIDO ELEGIDAS DE MODO INDEPENDIENTE, Y QUE LAS CANTIDADES QUE SE MIDEN, SE DISTRIBUYEN APROXIMADAMENTE NORMAL.

1. En un estudio realizado sobre la altura (en metros) media de los habitantes de cierto pueblo. Se obtuvieron los siguientes resultados:

{1,50; 1,52; 1,48; 1,55; 1,60;
1,49; 1,55; 1,63}

- a) Realiza la estimación puntual y por intervalo para la altura promedio de los habitantes de esa población. Emplea el nivel de confiabilidad que desees.
2. La cantidad mínima requerida para que un anestésico surta efecto en una intervención quirúrgica fue por término medio de 50 mg, con una desviación típica de 10,2 mg, en una muestra de 60 pacientes. Construye un intervalo de confianza para la media al 99%.
 3. Un cardiólogo se encuentra interesado en encontrar límites de confianza al 90%, para el incremento de la presión sistólica tras un cierto ejercicio físico. Construye dicho intervalo si en 50 individuos se obtuvo $\bar{x} = 13$ mm Hg y $s = 3$ mm Hg.
 4. Se desea realizar una estimación confidencial del promedio de la estatura de los niños varones de 10 años de una ciudad con una confianza del 95%.
 - a) ¿Cuál será dicho intervalo si se toma una muestra de 101 niños al azar, entre todos los que reúnen las características deseadas, y medimos sus estaturas, y se obtienen las siguientes estimaciones puntuales: $\bar{x} = 138.6$ cm, $s^2 = 29.16$ cm²?
 5. En un determinado servicio de odontología se ha observado que el 22% de las visitas llevan consigo una extracción dentaria inmediata. Se tomó una muestra, $n=360$, del registro de visitas de cierto año, detectándose que 74 dieron lugar a una

extracción inmediata. ¿Entran en contradicción las cifras de ese año con el porcentaje establecido de siempre?

6. Sólo una parte de los pacientes que sufren un determinado síndrome neurológico consiguen una curación completa; Si de 64 pacientes observados se han curado 41. Da una estimación puntual y un intervalo al 99% de la proporción de los que sanan.
7. El calcio se presenta normalmente en la sangre de los mamíferos en concentraciones de alrededor de 6 mg por cada 100 ml del total de sangre. Una serie de nueve pruebas sobre un paciente revelaron una media muestral de 7,2 mg de calcio por 100 ml del volumen total de sangre, y una desviación típica muestral de 2 mg de calcio por cada 100 ml de sangre. ¿Hay alguna evidencia, para un nivel $\alpha = 0.05$, de que el nivel medio de calcio para este paciente sea más alto del normal?
8. El número de accidentes mortales en una ciudad es, en promedio, de 12 mensuales. Tras una campaña de señalización y mejoramiento de las vías urbanas se contabilizaron en 6 meses sucesivos: 8, 11, 9, 7, 10, 9 accidentes mortales. ¿Fue efectiva la campaña?
9. Una población infantil se dice que es susceptible de recibir una campaña de educación e higiene si su porcentaje de niños con dientes cariados es superior al 20%. En una población con 12.637 niños, ¿debe hacerse la campaña si de 387 de ellos 70 tenían algún diente cariado?
10. Para comprobar si un tratamiento con ácidos grasos es eficaz en pacientes con eczema atípico, se tomaron 10 pacientes con eczema de más de 9 meses y se les sometió durante 3 semanas a un tratamiento ficticio (placebo) y durante las tres siguientes a un tratamiento con ácidos grasos. Tras cada periodo, un médico ajeno al proyecto evaluó la importancia del eczema en una escala de 0 (no eczema) a 10 (tamaño máximo de eczema). Los datos fueron los siguientes:

Tratamiento	Observaciones									
Placebo	6	8	4	8	5	6	5	6	4	5
Medicamento	5	6	4	5	3	6	6	2	2	6

¿Es eficaz el medicamento?

Autoevaluación

1. En una muestra de 50 bebés varones de 12 semanas de vida, se obtuvo un peso medio de 5.900 gr y una desviación típica de 94 gr. Obtener un intervalo de confianza (al 95%) para el peso medio poblacional.
2. En una muestra de tabletas de aspirinas, de las cuales observamos su peso expresado en gramos, obtenemos:
1, 19; 1,23 ; 1,18 ; 1,21 ; 1,27; 1,17 ; 1,15 ;
1,14 ; 1,19 ; 1,2

Suponiendo la Normalidad para esta distribución de pesos, determinar un intervalo al 80% de confianza para la media.

3. Se desea estimar el peso medio de mujeres de 30 a 40 años. Un estudio realizado en 160 mujeres de tales edades da los siguientes resultados: $\bar{x} = 53$ kg y $s = 5$ kg. Construye un intervalo de confianza al 95 % para el peso promedio poblacional.
4. El director de cierto hospital está interesado en conocer el porcentaje de pacientes asmáticos de la población que atiende, para planificar una medida preventiva de las crisis. Se estudió una muestra de 250 personas de dicha población, encontrándose que de ellos 28 eran asmáticos. Construye un intervalo de confianza al 99% para la proporción de asmáticos de dicha población.
5. El promedio de las puntuaciones de un número elevado de alumnos de Bioestadística es de 96,50. Un determinado año se examinaron 50 alumnos con resultados promedio de 97,25 y desviación típica de 1. ¿Variaron las calificaciones?
6. Se conoce que un 20% de los individuos tratados crónicamente con digoxina sufren una reacción adversa por causa de ella. A 10 pacientes se les administró durante largo tiempo digoxina más otros medicamentos, y de ellos 5 desarrollaron la reacción adversa. ¿Puede afirmarse que la asociación entre la digoxina y los otros medicamentos hace variar el número de reacciones adversas?

7. Un 8% de los individuos que acuden a un servicio sanitario son hiperutilizadores del mismo (más de 11 visitas al año) y, de entre ellos, un 70% son mujeres. De entre los no hiperutilizadores, son mujeres el 51%. ¿Puede afirmarse que han variado los hábitos de éstas si, tras una campaña de información y control de visitas, de 90 mujeres elegidas al azar 6 resultaron hiperutilizadoras?
8. En un programa de Control de Enfermedades Crónicas, la hipertensión está incluida como la primera enfermedad a controlar. 15 pacientes hipertensos son sometidos al programa y controlados en su tensión sistólica antes y después de 6 meses de tratamiento. Los datos son los siguientes:

Antes	180	200	160	170	180	190	190	180	190	160
	170	190	200	210	220					
Después	140	170	160	140	130	150	140	150	190	170
	120	160	170	160	150					

¿Es efectivo el tratamiento?

9. La eliminación por orina de aldosterona está valorada en individuos normales en 12 mgs/24 h. por término medio. En 50 individuos con insuficiencia cardíaca se observó una eliminación media de aldosterona de 13 mgs/24 h., con una desviación típica de 2,5 mgs/24 h.
- ¿Son compatibles estos resultados con los de los individuos normales?
 - ¿La insuficiencia cardíaca aumenta la eliminación por orina de aldosterona?

Bibliografía

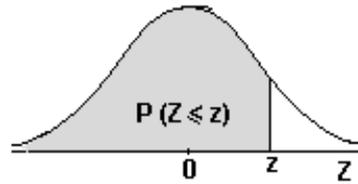
- Armitage P, Berry G. **Statistical Methods in Medical Research**. 3rd ed. Oxford: Blackwell Scientific Publications, 1994
- Altman DG. **Practical statistics for medical research**. London: Chapman and Hall. 1992
- Daniel WW. **Bioestadística**. Base para el análisis de las ciencias de la salud. 3ra edición. México. D. F: Limusa; 1997
- Norman GR, Streiner DL. **Bioestadística**. España: Hartcourt Brace; 1998

18. Doménech Massons, JM. **Métodos estadísticos en ciencias de la salud**. Unidades didácticas 5 y 6 . Barcelona: Signo; 1995
19. Swinscow, TVD. **Statistics at square one**. 9th edition. BMJ publishing group; 1997
20. Martínez Canalejo H, Santana Porben S. **Manual de procedimientos bioestadísticos**. Tomo I. ISCM de la Habana. 1989.
21. Silva LC. **Cultura estadística e investigación científica en el campo de la salud: una mirada crítica**. Díaz de Santos. 1997
9. Armitage P, Berry G. **Estadística para la Investigación Biomédica**. Doyma, Barcelona, 1992.
10. Hamilton LC. **Modern Data Analysis**. Brooks/Cole Publishing Company, Pacific Grove, 1990.
11. Martín Andrés A, Luna Del Castillo JD. **Bioestadística para las Ciencias de la salud**. Norma, Granada, 1994.
12. Peña Sánchez De Rivera D. **Estadística: Modelos y Métodos, 1**. Alianza Universidad Textos, Madrid, 1994.
13. Rivas Moya T, Mateo MA, Ríus Díaz F, Ruiz M. **Estadística Aplicada a las Ciencias Sociales: Teoría y Ejercicios (EAC)**. Secretariado de Publicaciones de la Universidad de Málaga, Málaga, 1991.
14. Rubio Calvo E, Martínez Terrer T Y Otros. **Bioestadística**. Colección Monografías Didácticas, Universidad de Zaragoza, Zaragoza, 1992.
15. Versión electrónica del manual de la Universidad de Málaga. **Bioestadística: métodos y aplicaciones**. Málaga; 1998.
16. Dawson-Saunders B, Trapp RG. **Bioestadística**. 2da Edición. El Manual Moderno. México; 1999.

Apéndices

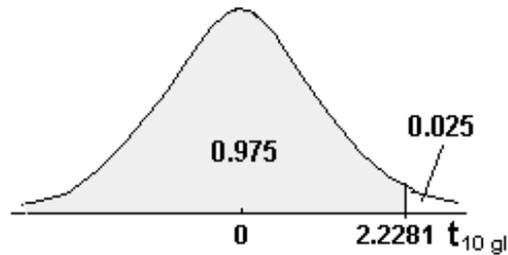
Apéndice

Tabla A – Áreas de la distribución normal estándar.



Z	Área	Z	Área
0.0	0.5000	2.0	0.02275
0.1	0.4602	2.1	0.01786
0.2	0.4207	2.2	0.01390
0.3	0.3821	2.3	0.01072
0.4	0.3446	2.4	0.00820
0.5	0.3085	2.5	0.00621
0.6	0.2743	2.6	0.00466
0.7	0.2420	2.7	0.00347
0.8	0.2119	2.8	0.00256
0.9	0.1841	2.9	0.00187
1.0	0.1587	3.0	0.00135
1.1	0.1657	3.1	0.00097
1.2	0.1151	3.2	0.00069
1.3	0.0968	3.3	0.00048
1.4	0.0808	3.4	0.00034
1.5	0.0668	3.5	0.00023
1.6	0.0548	3.6	0.00016
1.7	0.0446	3.7	0.00011
1.8	0.0359	3.8	0.00007
1.9	0.0287	3.9	0.00005
2.0	0.02275	4.0	0.00003

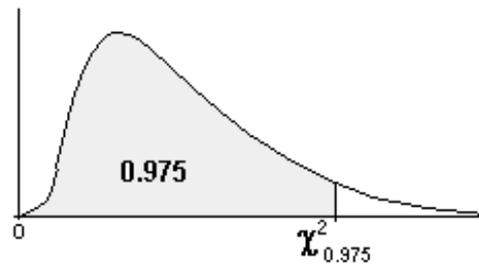
Tabla B - Percentiles de la distribución t de student.



g.l	t _{0.90}	t _{0.95}	t _{0.975}	t _{0.99}	t _{0.995}
1	3.078	6.3138	12.706	31.821	63.657
2	1.886	2.9200	4.3027	6.965	9.9248
3	1.638	2.3534	3.1825	4.541	5.8409
4	1.533	2.1318	2.7764	3.747	4.6041
5	1.476	2.0150	2.5706	3.365	4.0321
6	1.440	1.9432	2.4469	3.143	3.7074
7	1.415	1.8946	2.3646	2.998	3.4995
8	1.397	1.8595	2.3060	2.896	3.3554
9	1.383	1.8331	2.2622	2.821	3.2498
10	1.372	1.8125	2.2281	2.764	3.1693
11	1.363	1.7959	2.2010	2.718	3.1058
12	1.356	1.7823	2.1788	2.681	3.0545
13	1.350	1.7709	2.1604	2.650	3.0123
14	1.345	1.7613	2.1448	2.624	2.9768
15	1.341	1.7530	2.1315	2.602	2.9467
16	1.337	1.7459	2.1199	2.583	2.9208
17	1.333	1.7396	2.1098	2.567	2.8982
18	1.330	1.7341	2.1009	2.552	2.8784
19	1.328	1.7291	2.0930	2.539	2.8609
20	1.325	1.7247	2.0860	2.528	2.8453
21	1.323	1.7207	2.0796	2.518	2.8314
22	1.321	1.7171	2.0739	2.508	2.8188
23	1.319	1.7139	2.0687	2.500	2.8073
24	1.318	1.7109	2.0639	2.492	2.7969
25	1.316	1.7081	2.0595	2.485	2.7874
26	1.315	1.7056	2.0555	2.479	2.7787
27	1.314	1.7033	2.0518	2.473	2.7707
28	1.313	1.7011	2.0484	2.467	2.7633
29	1.311	1.6991	2.0452	2.462	2.7564
30	1.310	1.6973	2.0423	2.457	2.7500
35	1.3062	1.6896	2.0301	2.438	2.7239
40	1.3031	1.6839	2.0211	2.423	2.7045
45	1.3007	1.6794	2.0141	2.412	2.6896
50	1.2987	1.6759	2.0086	2.403	2.6778
60	1.2959	1.6707	2.0003	2.390	2.6603
70	1.2938	1.6669	1.9945	2.381	2.6480

80	1.2922	1.6641	1.9901	2.374	2.6388
90	1.2910	1.6620	1.9867	2.368	2.6316
100	1.2901	1.6602	1.9840	2.364	2.6260
120	1.2887	1.6577	1.9799	2.358	2.6175
140	1.2876	1.6558	1.9771	2.353	2.6114
160	1.2869	1.6545	1.9749	2.350	2.6070
180	1.2863	1.6534	1.9733	2.347	2.6035
200	1.2858	1.6525	1.9719	2.345	2.6006
∞	1.282	1.645	1.96	2.326	2.576

Tabla C – Percentiles de la distribución Ji cuadrado (χ^2).



GL	0.05	0.10	0.50	0.90	0.950	0.975	0.990	0.995
1	0.004	0.016	0.455	2.71	3.84	5.02	6.63	7.88
2	0.10	0.211	1.39	4.61	5.99	7.38	9.21	10.6
3	0.35	0.584	2.37	6.25	7.81	9.35	11.3	12.8
4	0.71	1.06	3.36	7.78	9.49	11.1	13.3	14.9
5	1.15	1.61	4.35	9.24	11.1	12.8	15.1	16.7
6	1.64	2.20	5.35	10.6	12.6	14.4	16.8	18.5
7	2.17	2.83	6.35	12.0	14.1	16.0	18.5	20.3
8	2.73	3.49	7.34	13.4	15.5	17.5	20.1	22.0
9	3.33	4.17	8.34	14.7	16.9	19.0	21.7	23.6
10	3.94	4.87	9.34	16.0	18.3	20.5	23.2	25.2
11	4.57	5.48	10.3	17.3	19.7	21.9	24.7	26.8
12	5.23	6.30	11.3	18.5	21.0	23.3	26.2	28.3
13	5.89	7.04	12.3	19.8	22.4	24.7	27.7	29.8
14	6.57	7.79	13.3	21.1	23.7	26.1	29.1	31.3
15	7.26	8.55	14.3	22.3	25.0	27.5	30.6	32.8
16	7.96	9.31	15.3	23.5	26.3	28.8	32.0	34.3
17	8.67	10.1	16.3	24.8	27.6	30.2	33.4	35.7
18	9.39	10.9	17.3	26.0	28.9	31.5	34.8	37.2
19	10.1	11.7	18.3	27.2	30.1	32.9	36.2	38.6
20	10.9	12.4	19.3	28.4	31.4	34.2	37.6	40.0
21	11.6	13.2	20.3	29.6	32.7	35.5	38.9	41.4
22	12.3	14.0	21.3	30.8	33.9	36.8	40.3	42.8
23	13.1	14.8	22.3	32.0	35.2	38.1	41.6	44.2
24	13.8	15.7	23.3	33.2	36.4	39.4	43.0	45.6

25	14.6	16.5	24.3	34.4	37.7	40.6	44.3	46.9
26	15.4	17.3	25.3	35.6	38.9	41.9	45.6	48.3
27	16.2	18.1	26.3	36.7	40.1	43.2	47.0	49.6
28	16.9	18.9	27.3	37.9	41.3	44.5	48.3	51.0
29	17.7	19.8	28.3	39.1	42.6	45.7	49.6	52.3
30	18.5	20.6	29.3	40.3	43.8	47.0	50.9	53.7
40	26.5	29.0	39.3	51.8	55.8	59.3	63.7	66.8
50	34.8	37.7	49.3	63.2	67.5	71.4	76.2	79.5
60	43.2	46.5	59.3	74.4	79.1	83.4	88.4	92.0

Nota: para grados de libertad mayores que 60 utilice la siguiente fórmula:

$$\chi_p^2 = \frac{1}{2} \left(Z_p + \sqrt{2k-1} \right)^2 \text{ donde } k = \text{grados de libertad.}$$

y Z_p = percentil correspondiente de la distribución normal estándar