ESTADÍSTICAS EN SALUD II

ESTADÍSTICA INFERENCIAL

AUTORA:

Mg. Silvana Torres

-AÑO 2009-

PRÓLOGO

Es de fundamental importancia que los profesionales de salud conozcan la ciencia estadística, asumiendo que esta es una herramienta imprescindible para poder cuantificar y analizar todos los aspectos relacionados con el proceso salud-enfermedad. Es conocido que la información confiable y correctamente analizada es la base de las decisiones coherentes.

También como herramienta básica de la investigación, el conocimiento de la metodología estadística es indispensable, ya que la investigación trabaja con fenómenos eminentemente variables. El concepto de variabilidad como la sistematización y organización racional de una metodología para el análisis y conocimiento de fenómenos cambiantes, es un componente de conocimiento relativamente nuevo.

Decimos que la estadística es un conjunto de procedimientos que tienen por finalidad recolectar, elaborar, caracterizar y analizar un conjunto de datos. Cuando hablamos de análisis estadístico es necesario puntualizar que este comprende dos grandes áreas de la metodología estadística: el análisis descriptivo y el inferencial. El primero, más ampliamente conocido, consiste en describir un conjunto de datos (recolectarlos, presentarlos en forma tabular y gráfica, calcular las medidas de resumen, etc.), para interpretar el comportamiento de las variables. El análisis inferencial, bastante más complejo que el anterior, consiste en aplicar determinadas técnicas estadísticas, para tratar de generalizar o inferir resultados en la población a partir del análisis de una parte de ella (muestra).

Es obvio que para poder comprender las técnicas de estadística inferencial es imprescindible manejar con fluidez la metodología de estadística descriptiva, conocimiento que el alumno de esta materia posee previamente. En este curso estudiaremos un concepto más amplio de la metodología estadística, que no sólo contempla la descripción de datos, sino básicamente el conocimiento de herramientas que le permitan al alumno aplicar e interpretar algunas técnicas inferenciales básicas (muestreo, estimación de parámetros, pruebas de hipótesis, etc).

INDICE

CAPÍTULO I-CAMPOS DE LA METODOLOGÍA ESTADÍSTICA

Áreas de la Estadística

La metodología estadística se refiere al grupo de técnicas o métodos que se han desarrollado para la recopilación, presentación, análisis de datos y para el uso adecuado de los mismos, con el objeto de fundamentar científicamente las conclusiones y decisiones que se asuman a partir de los mismos.

Las dos grandes áreas de la metodología estadísticas son:

- ***Estadística descriptiva:*** mediante estas técnicas, ante un conjunto de datos, se los describe y caracteriza; esto significa, una vez recolectados los datos, se los presenta en forma tabular y gráfica, y se calculan las medidas de resumen (posición y variabilidad); para luego ser analizados mediante la descripción de las características que se observan en los mismos.
- ***Estadística inferencial:*** consiste en aquellos métodos mediante los cuales se pueden realizar inferencias o generalizaciones acerca de una población; mediante procedimientos estadísticos basados en la teoría de las probabilidades, a partir de una muestra o parte de una población. Es decir que, a partir del análisis de una parte de la población (muestra), se puede conocer el comportamiento de los fenómenos en la población, midiendo y controlando, a través de la teoría de las probabilidades, el grado de error que se estaría cometiendo por el hecho de estudiar las características en sólo una parte de la población y no en la totalidad de la misma.

La inferencia estadística comprende básicamente tres grandes metodologías:

- Técnicas de muestreo
- Estimación de parámetros
- Pruebas de hipótesis

Para poder entender y aplicar las técnicas de estadística inferencial es indispensable el conocimiento de la metodología básica de estadística descriptiva por lo que es fundamental repasar los siguientes temas:

- Presentación de datos: tabular y gráfica
- Distribución de frecuencias: absolutas y relativas

- ❖ Medidas descriptivas o de resumen: de variables cuantitativas (posición: media, mediana, modo, cuartos, etc; variabilidad: rango, desvío estándar, etc.), en variables cualitativas (proporciones, razones, etc.)

Una vez que hayamos recordado los elementos básicos de estadística descriptiva, debemos aceptar que todo el campo de conocimientos que abordaremos, se desenvuelve en torno al manejo de datos cuantitativos inferidos (y a las técnicas que ello implica), es decir obtenidos a través de operaciones matemáticas y lógicas que nos permitan inferir juicios con márgenes de error cuantificados. Es decir, podremos conocer lo que ocurre en una población determinada con sólo el análisis de los datos de una muestra extraída de ella: decidiremos si los resultados de una investigación son demostrativos o no de las hipótesis que nos hayamos planteado; emitiremos juicios acerca del comportamiento futuro de un fenómeno (predicción), todo esto en el terreno de las probabilidades, es decir, sin certeza absoluta, sino con un mayor o menor grado de seguridad.

En síntesis, logrado un lote de datos como producto de una investigación, el tratamiento que le daremos será diverso según se haya trabajado con:

1) Una muestra y no el total de unidades de la población: se podrá inferir que es lo que ocurre en la población a la que pertenece la muestra estudiada, usando para ello la metodología de ***estimación de parámetros***.
2) Una hipótesis acerca del comportamiento de un fenómeno: se necesitará comprobar la hipótesis enunciada durante la investigación, para lo que necesitaremos la metodología de ***pruebas de hipótesis.***
3) El estudio de un problema en un momento cronológicamente bien definido, pero necesitamos inferir el comportamiento futuro de las variables: para esto necesitaremos conocer los conceptos de ***predicción***.

Variables

Antes de avanzar en el desarrollo de las técnicas básicas de estadística inferencial es necesario recordar el concepto de variables y su clasificación.

Definición: es una característica que puede tomar diferentes valores (no necesariamente numéricos) en las distintas unidades de observación. Por ejemplo, edad, sexo, tiempo de evolución de la diabetes, tipo de tratamiento administrado, etc. Las propiedades o características susceptibles de tomar distintos valores o intensidades es lo que se conoce con el nombre de ***variables***.

Clasificación: Si la variable presenta un atributo o cualidad se denomina ***cualitativa***. Si la variable presenta valores numéricos es ***cuantitativa***.

Las variables cualitativas a su vez se clasifican en ***nominales***, cuando sus categorías no presentan ningún orden preestablecido (por ejemplo sexo: varón o mujer); y ***ordinales*** cuando las categorías de la variable tiene un orden preestablecido (por ejemplo nivel de instrucción: primario, secundario, universitario)

Dentro de estas variables cuantitativas podemos distinguir dos tipos: las que pueden variar sólo en números enteros o en fracciones bien definidas sin valores intermedios, como por ejemplo: *Número de hijos de una mujer:* puede tener 1, 2, 3,…hijos; nunca 1,5 o 3,8 hijos. Son por ello llamadas ***variables cuantitativas discontinuas o discretas.***

Las que pueden variar en forma continua, como el contenido de hemoglobina en la sangre, la presión arterial, estatura, edad, etc. son llamadas ***variables cuantitativas continuas***, estas pueden tomar cualquier valor entre un máximo y un mínimo.

Para determinar los grupos o categorías en la ***escala cualitativa*** basta con enunciar las posibilidades que se presentan. Veamos algunos ejemplos:

Variable	**Categorías**
➢ Sexo	• Varón • Mujer
➢ Alfabetismo	• Alfabeto • Analfabeto
➢ Rendimiento	• Bueno • Regular • Malo

Para el caso de las variables cuantitativas discretas se deben definir los intervalos de clase. Por ejemplo número de cigarrillos fumados presenta los siguientes intervalos de clase:

0- 9: los que fuman menos de 10 cigarrillos

10-20: los que fuman desde 10 a 20 cigarrillos inclusive

+ 20: los que fuman más de 20 cigarrillos

Mayor problema se presenta cuando se trata de ***escalas cuantitativas continuas*** pues para determinar los intervalos de clases debemos tomar en cuenta varios aspectos:

- Siempre se pierde algo de información por el hecho de agrupar los datos.
- Es necesario definir con claridad los límites de estos grupos o intervalos de clase, de modo que sepamos bien a qué intervalo pertenece una observación individual. Estos intervalos deben ser exhaustivos (tener en cuenta todos los posibles valores) y mutuamente excluyentes. Por ejemplo la variable edad, los intervalos podrían ser: 0-4, 5-9, 10-14, 15-19, 20-24 años, etc.

En resumen las variables se clasifican en:

- Por su escala de medición.
  - *Cualitativas*
    - *Nominales*
    - *Ordinales*
  - *Cuantitativas*
    - *Discretas*
    - *Continuas*
- Por su lugar en la investigación:
  - *Independiente:* que precede a la aparición del fenómeno en estudio, frecuentemente llamado causa o factores relacionados con el fenómeno en estudio.
  - *Dependiente:* frecuentemente llamada efecto, los valores que asume dependen de otras variables. Es la variable principal que mide el fenómeno que se quiere estudiar.

Medidas Descriptivas

Según el tipo de variable que se estudie y los objetivos de la investigación se calcularán las medidas descriptivas correspondientes:

- En el caso de variables cualitativas:
  - Tasa
  - Razón
  - Proporción
- En el caso de variables cuantitativas
  - Medidas de Posición central
    - Media aritmética

        - Mediana
        - Modo
    - Medidas de Posición no central:
        - Mínimo y máximo
        - Percentiles
        - Cuartiles
    - Medidas de Variabilidad o Dispersión
        - Desvío estándar
        - Rango
        - Rango inter-cuartos
- En el caso de relación entre dos variables
    - Coeficiente de correlación
    - Riesgo relativo
    - Odds Ratio

Parámetros y estadísticos

Si estas medidas, también designadas características son calculadas con los datos de la población se denominan ***parámetros***, si estas son calculadas a partir de datos muestrales son llamadas ***estadísticos.***

Cuando no es posible calcular directamente los parámetros de la población, estos pueden ser estimados a través del cálculo de los estadísticos de las muestras.

Veamos algunos ejemplos de parámetros y estadísticos y los símbolos con los que usualmente se representan:

Características	Símbolo del **Parámetro**	Símbolo del **Estadístico**
Media	μ	$\bar{x}$
Desvío estándar	σ	s
Varianza	σ^2	s^2
Correlación	ρ	r
Proporción	π	p
Riesgo Relativo	RR	$R\hat{R}$
Odds Ratio	OR	$O\hat{R}$

CAPÍTULO II- NOCIONES BÁSICAS DE MUESTREO

Población

Generalidades

En cualquier investigación, el interés central es la obtención de conclusiones que sean aplicables no solamente a quienes fueron escogidos para participar en ella; sino al conjunto de la población de la cual se obtuvo la información.

Uno de los primeros pasos en una investigación, consiste en delimitar con exactitud los elementos (personas, objetos, etc.) sobre los cuales se desea realizar el estudio. Los elementos de una población son las unidades de las cuales se busca información; son los individuos, elementos, unidades elementales que forman la población; éstas son las unidades de análisis, y su naturaleza se determina mediante los objetivos de la investigación.

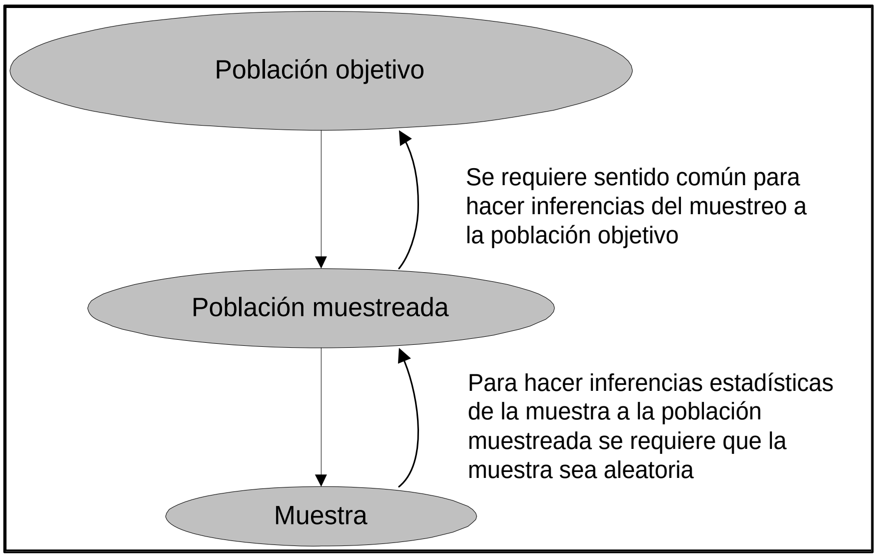
Definición

- La población es el agregado de los elementos, siendo éstos las unidades básicas que forman y definen la población.
- Población: conjunto de unidades (individuos, objetos, instituciones, etc.) en las que existe una característica común susceptible a ser medida y a partir de las cuales se obtendrán los datos.
- Es el agregado de los elementos que cumplen con un conjunto predeterminado de criterios.

La población puede estar constituida por personas, animales, registros médicos, nacimientos, muestras de laboratorio, accidentes viales, instituciones, etc..

El tamaño de la Población (N) es el total de unidades que la componen. Si la población tiene un tamaño limitado recibe el nombre de ***población finita***, en estos casos es posible obtener un listado o inventario de los elementos de la población (por ej. Alumnos de la Escuela de Enfermería). Por su parte una ***población infinita*** o hipotética es aquella que contiene una cantidad ilimitada o muy grande de elementos, por lo que resulta imposible en la práctica producir un listado o inventario de ellos (por ej. portadores de HIV).

Con fines didácticos se acostumbra a diferenciar dos tipos de poblaciones: ***población bajo estudio*** y ***población objetivo***. La primera representa la población a partir de la cual se obtendrán los datos, o se extraerá la muestra, denominándose también población muestreada; en tanto que la población objetivo es aquella a la que se desea generalizar los resultados del estudio. Generalmente ambas poblaciones son coincidentes.



Si todos los elementos que integran una población poseen similares características, es decir baja variabilidad, se dice que esta es ***homogénea***; en caso de no cumplir esta condición, la población es ***heterogénea,*** o sea que presenta mucha variabilidad. En sentido estricto, las poblaciones humanas son siempre heterogéneas, puesto que las unidades que la componen presentan variaciones individuales respecto a una o más características personales (edad, sexo, raza, actividad laboral, nivel de instrucción, tipo de alimentación, etc.).

Si en un estudio de investigación se obtiene la información de la totalidad de la población se denomina **censo**. Cabe destacar que los datos censales están también sujetos a error: errores de medición.

Muestreo

Con frecuencia resulta muy difícil, y a veces imposible, desarrollar una investigación que comprenda el total de la población, ya sea porque el número de elementos es demasiado grande o infinito, porque están distribuidos en forma muy heterogénea o porque la inclusión de todos los elementos de la población generaría complicaciones o encarecería excesivamente el estudio. De allí la utilidad del diseño de muestras, con las que habiéndose operado adecuadamente, pueden obtenerse resultados similares a los que se alcanzan incluyendo a todos los elementos de la población.

Definiciones

Muestra: es el subconjunto de unidades provenientes de la población (parte de la población), que con algún criterio o sin él, son seleccionadas a los efectos de ser estudiadas en una o más características.

Es el subconjunto o parte de la población en la que se llevará a cabo la investigación con el fin posterior de generalizar los hallazgos a la población.

La cantidad de elementos que integran la muestra constituyen el tamaño muestral (n). El muestreo puede definirse como la selección de una parte de la población con el fin de hacer inferencias acerca de la totalidad de la misma.

Unidad de observación: son los elementos de la población en los cuales se medirán o estudiarán las variables de interés.

Unidad de muestreo: es el elemento utilizado para seleccionar la muestra. En muchos casos las unidades de observación y muestreo coinciden (Ej. Alumnos de la Facultad), pero hay casos en que no (Ej. Si se desea estudiar las infecciones respiratorias agudas en menores de 2 años, no se tendría un listado de niños, sino que se tendría que tomar muestras de casas para llegar a los niños)

Ventajas del muestreo: Razones para realizar un muestreo

1. *Costo*: las muestras al examinar parte de la población es mucho menos costosa que el censo. El costo es un argumento a favor del muestreo debido a que frecuentemente una muestra puede proporcionar datos con la suficiente precisión y a un costo mucho más bajo que el censo.

2. *Precisión*: (calidad de información) se puede ejercer mucho mejor control sobre los errores que no son de muestreo (fallas en las encuestas, respuestas incompletas, información imprecisa, errores de medición, errores de proceso, etc..) usando el muestro en lugar del censo. Por ejemplo en una encuesta por muestreo generalmente obtenemos mejor supervisión y entrenamiento de encuestadores; mayor control de respuestas y procesos.

3. *Tiempo necesario*: las muestras producen información más rápidamente por dos razones principales: 1°- tomar una muestra requiere menos tiempo que levantar un censo y 2°- el procesamiento de los datos toma menor tiempo.

4. *Cantidad de información*: puede obtenerse información más detallada debido a que la muestra toma menos tiempo, menor costo y permite poner más cuidado en la ejecución. (Ej. en el Censo 2001 se llevó a cabo una muestra para conocer ciertas características de la población).

5. *Pruebas destructivas*: cuando la medición implica la destrucción del elemento poblacional la única manera de realizar la investigación es a través de muestras. Ej. Control de estados de conservación de las vacunas (cadena de frío), efectividad.

Diseño de muestreo

El diseño de muestras tiene dos aspectos fundamentales: un *proceso de selección*, que consiste en las reglas y operaciones mediante las cuales se incluyen en la muestra algunos elementos de la población; y un *proceso de estimación* (inferencias) para calcular los estadísticos de la muestra que son estimaciones muestrales de valores de la población (proporciones, medias, desvío estándar, etc.).

Otros aspectos importantes que se deben considerar en el diseño de muestras son:

- *Definición de variables*: especificar la naturaleza de las características, categorías de clasificación y unidades para expresarlas.
- *Métodos de observación*: (mediciones) que incluye tanto la recolección como el procesamiento de los datos.
- *Métodos de análisis estadísticos*: reducen los datos de la encuesta a resultados que pueden comprenderse y utilizarse.
- *Utilización de los resultados*: que estos sirvan para tomar decisiones concretas que se basen en los resultados.
- *La precisión que se desea*: fijar la probabilidad de error por estar trabajando con una muestra en lugar de la población; es decir el riesgo que se corre de que el resultado obtenido a partir de la muestra no estime con exactitud al de la población, se simboliza con *p (p value)* y generalmente varía del 0,01 al 0,05 (1% al 5%).

Sesgos por muestras inadecuadas o información incompleta

Una muestra constituye una parte de la población, y si el método de selección de las unidades se realiza a través de un procedimiento adecuado se espera que ésta represente a la población en su conjunto, es decir que todas las unidades que la constituyan posean idénticas características a las que tiene la población. En cambio si la muestra es integrada por algunas de las unidades y por lo tanto corresponden sólo a una parte de la población de origen se dice que la ***muestra es sesgada.***

Podemos hablar de ***error sistemático*** *o* ***sesgo*** cuando existe una tendencia a obtener resultados que difieren en forma sistemática de los valores verdaderos.

Las principales fuentes de error sistemático son:

- Sesgo de selección
- Sesgo de medición (o clasificación) cuando las mediciones y/o clasificaciones de la variable son inexactas.

Si el error no es sistemático no lo denominamos "sesgo".

Sesgo de selección**:** el sesgo de selección se produce cuando existe una diferencia sistemática entre las características de la población seleccionada para el estudio y las características de la población no seleccionada. El marco de referencia no está constituido por la totalidad de la población.

El ***error de muestreo*** mide la discrepancia que se presenta a partir de una enumeración incompleta de la población. Estos errores pueden presentarse debido a que la población no ha sido definida debidamente o no corresponde a la población bajo estudio. La población meta (a ser estudiada) difiere de la población muestreada. Ejemplo: analizar la cobertura de vacunación de los niños de Tucumán, tomando una muestra de los que concurren a los Centros Asistenciales oficiales.

Los errores de muestreo (si no son debido a problemas del diseño) pueden reducirse aumentando el tamaño y/o complejidad de la muestra.

Métodos de muestreo

Los métodos de muestreo se clasifican en ***no probabilísticos*** y ***probabilísticos.***

Muestreo no probabilístico

Se caracterizan porque *se desconoce la probabilidad que tienen los elementos de la población de ser escogidos para constituir la muestra*, y la selección tiene lugar siguiendo criterios para los fines del estudio, sin recurrir al azar. En estos métodos *no se puede medir o controlar el error probable de muestreo*. Algunas de las técnicas de muestreo no probabilístico más usadas son:

- *Muestras casuales o fortuitas*: los elementos van siendo incorporados a medida que acuden al sitio donde se efectúan las mediciones, hasta alcanzar un tamaño muestral previamente establecido.
- *Selección experta*: muestra de juicio utilizadas por expertos para seleccionar elementos representativos o típicos de la población.
- *Por cuotas*: el investigador propone estratos de acuerdo a las variables que considere relevantes, y se incorporan elementos hasta completar la cantidad o cuota prefijada.

Muestreo probabilístico

En el muestreo probabilístico *cada elemento de la población tiene una probabilidad conocida y no nula de ser seleccionado*. Esta probabilidad se obtiene a través de una operación mecánica de aleatorización (fracción de muestreo: f_h). En éste tipo de muestreo *se puede cuantificar la probabilidad de cometer error de muestreo (p).*

***El muestreo irrestricto aleatorio** (mia)* es el proceso de selección y todos los demás procedimientos pueden verse como modificaciones de él, introducidos para proveernos de diseños más prácticos económicos y precisos. Las técnicas de muestreo probabilísticas más usadas son:

- *Aleatorio Simple:* En este tipo de muestreo cada unidad tiene la misma probabilidad de ser seleccionada para formar la muestra. Este se basa en la aleatoriedad del proceso de selección (selección al azar) esto puede ser llevado a cabo a través de diferentes maneras: sorteo, empleo de tablas de números aleatorios, números aleatorios de computadoras o de calculadoras. Para poder realizar el sorteo de las unidades que conformarán la muestra es necesario previamente enumerar el listado de elementos de la población (marco muestral), y luego seleccionar las unidades de la muestra.
- *Sistemático:* Consiste en seleccionar al azar la primera unidad y a partir de allí tomar las siguientes unidades a intervalos constantes, establecidos de acuerdo al tamaño de la muestra que se desea obtener.
- *Estratificado:* De acuerdo a un criterio predeterminado la población se divide en subgrupos (bloques o estratos) mutuamente excluyentes, con igual o diferente tamaño, en cada uno de los cuales están contenidas unidades que poseen características semejantes, es decir que cada estrato es homogéneo dentro del mismo, y son heterogéneos entre los estratos. De cada estrato se obtiene una muestra aleatoria. El

muestreo estratificado se aplica cuando la población bajo estudio es heterogénea y las características de los estratos difieren entre sí, obteniéndose así resultados con menor nivel de error.

- *Por conglomerados:* Consiste en dividir la población bajo estudio en subgrupos, llamados conglomerados, que tienen existencia en la realidad (tales como barrios, manzanas, escuelas, etc.), y que contienen unidades de distinta clase, aplicando luego en cada uno de ellos el método de muestreo aleatorio simple. A diferencia del muestreo estratificado, donde los subgrupos son homogéneos, en el muestreo por conglomerados resultan conjuntos heterogéneos, los cuales no fueron divididos por el investigador sino que se encuentran divididos en la realidad, además se desconoce de antemano cómo están distribuidas las unidades.
- *Polietápico:* consiste en realizar el muestreo en varias etapas sucesivas, utilizando iguales o distintos procedimientos en las diferentes fases. Se comienza con la división del universo en unidades de primer grado, a partir de las cuales se obtiene la muestra inicial, a continuación esta vuelve a dividirse para formar las unidades de segundo grado, de donde se selecciona una nueva muestra, así en forma continuada hasta llegar a las unidades finales del muestreo, unidades de observación de las cuales se obtendrán los datos.

Tamaño de la muestra (n)

La cantidad de unidades incluidas en la muestra (tamaño muestral) debe ser la adecuada para demostrar con una probabilidad razonable, prefijada por el investigador, la existencia de una diferencia estadísticamente significativa (no atribuible al azar), cuando la misma realmente existe.

El tamaño muestral depende entre otros factores de:

- El diseño de la investigación
- La frecuencia con que se presenta en la población la característica, factor o fenómeno que se desea estudiar
- El nivel de confianza con que se pretende efectuar las estimaciones
- La variabilidad de las mediciones (dispersión)
- La precisión que se desea (*p value)*

En general el tamaño muestral es menor cuanto mayor sea la frecuencia con que ocurre la característica, factor o fenómeno que se estudia en la población, más homogénea sea su distribución en la misma (menor variabilidad) y menor sea la precisión que se desea obtener. El criterio empleado, para la elección del tamaño muestral es realizar el estudio con el menor número posible de elementos que sean suficientes para obtener los resultados con la precisión fijada.

CAPÍTULO III- MEDIDAS DE ASOCIACIÓN ENTRE VARIABLES

Las medidas descriptivas consideradas hasta ahora son apropiadas sólo para resumir observaciones sobre una característica o variable (como por ejemplo: edad, nivel de colesterol en sangre, etc.). Sin embargo, gran parte de la investigación en salud se interesa en la relación entre dos o más características, tratando de establecer el grado de asociación entre ellas.

Cuando se quiere describir el grado de asociación entre dos características hay que tener en cuenta no sólo si ellas son cuantitativa o cualitativas sino también algunas características de su distribución, como ser si ellas son simétricas o asimétricas, los objetivos y el diseño de la investigación.

Para cumplir con el objetivo de evaluar la asociación entre dos variables, se debe tratar de cuantificar esta relación, establecer la posible relación causa-efecto; para esto es necesario definir medidas de asociación. Los objetivos de asociación dependen del tipo de variables involucradas, estos pueden ser:

1. Evaluar si la presencia de un factor (causa) produce cambios en la frecuencia de presentación de una patología (evento de interés). *Relación entre dos variables cualitativas dicotómicas.*
2. Evaluar si la distribución de una variable efecto (evento de interés) cambia según los niveles de un factor. *Relación entre una variable cualitativa y la otra cuantitativa.*
3. Evaluar si la magnitud del cambio observado en los valores de la variable efecto se puede explicar por un cambio en los valores del factor. *Relación entre dos variables cuantitativas.*

Relación entre dos características cualitativas dicotómicas

Uno de los más difíciles, pero también más importante, de los problemas en investigaciones médicas concierne a la existencia de asociación entre el llamado f*actor de riesgo* y la *incidencia de una enfermedad.* La investigación de tales tópicos es parte de lo que se conoce formalmente como epidemiología.

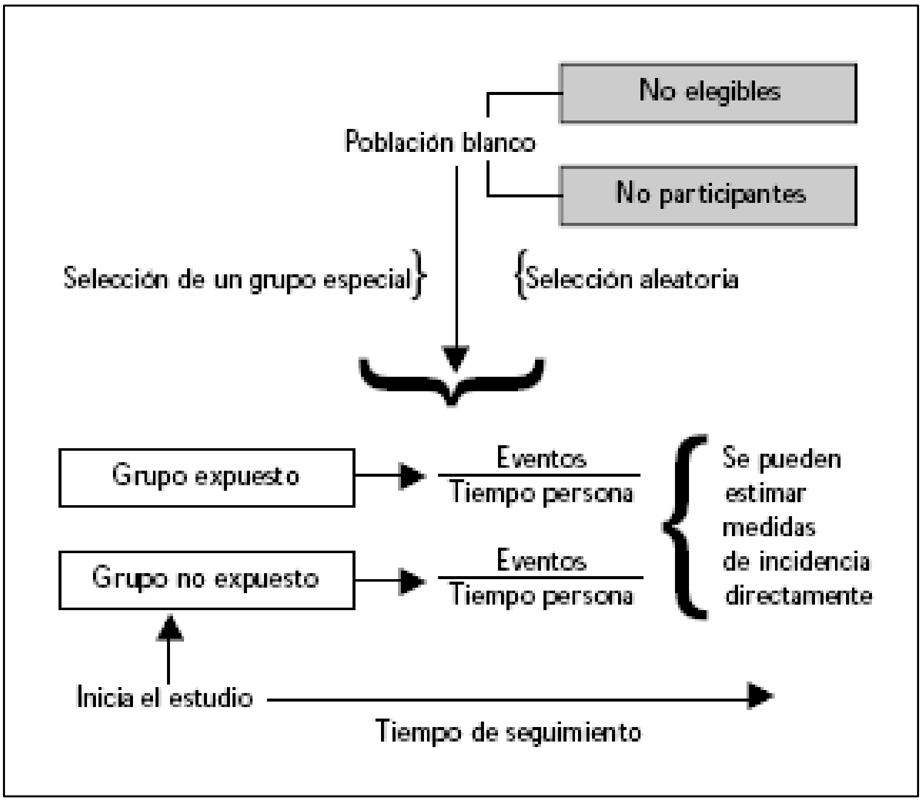
El principal interés está en dar medidas e interpretación de la asociación entre el posible factor de riesgo y la particular enfermedad. Tres diseños epidemiológicos principales son los que se tratarán en esta sección: el ***estudio prospectivo***, ***estudio retrospectivo*** y el de *corte transversal.*

Ejemplo: supóngase que existe una sospecha clínica de que existe una relación entre la obesidad y el diagnóstico de cardiopatía coronaria. Hay muchas formas para investigar esta suposición, se mostrarán las más importantes.

Estudios Prospectivos o de cohorte

Un estudio prospectivo comienza con dos grupos de personas, el primero contiene personas expuestas al factor de riesgo de interés, y el segundo contiene personas sin tal exposición. Al comienzo de la investigación se supone que los dos grupos están sanos, es decir libres del evento (patología) que se está estudiando. Las personas dentro de cada grupo son seguidas durante un período de tiempo especificado, y al final del mismo se comparan, de alguna manera, las estimaciones de la incidencia de enfermos en cada grupo.

Diseño de Estudio de Cohorte Prospectivo



Tanto en el grupo de expuestos como en el grupo de de no expuestos al factor se pueden calcular las incidencias de la enfermedad, por lo tanto es posible conocer el *riesgo de enfermar* en cada grupo.

Los resultados de esta investigación se pueden resumir en una tabla de contingencia de 2x2.

Tabla 1: Datos genéricos de un estudio prospectivo

Factor de riesgo	**Enfermos**	**No enfermos**	**Total**
Presente Ausente	a c	b d	a+b c+d
Total	a+b	c+d	a+b+c+d

En esta tabla sólo los totales a+b, c+d son conocidos ya que ellos están fijos por el diseño del estudio, incluyen al total de personas expuestas (factor presente) y total de personas no expuesta (factor ausente) respectivamente.

Determinar Incidencia o riesgo de enfermar en cada grupo:

- Incidencia en Expuestos (Factor presente): $I_F = a/(a+b)$
- Incidencia en no expuestos (Factor ausente): $I_{NF} = c/(c+d)$

Conocida la incidencia en cada grupo (Riesgo Absoluto) debemos construir medidas para poder comparar las mismas, entre las más usadas tenemos: Riesgo Relativo y Riesgo Atribuible.

Riesgo Relativo

El ***riesgo relativo*** se calcula como

$$\boxed{RR = \frac{I_F}{I_{NF}} = \frac{a/(a+b)}{c/(c+d)}}$$

En el ejemplo presentado anteriormente donde se pretende conocer si la obesidad podría ser considerado un factor de riesgo para enfermedad coronaria, la investigación podría comenzar con un grupo de obesos y no obesos (ambos sin enfermedad coronaria) y hacer un seguimiento en el tiempo para determinar la incidencia de enfermedad coronaria entre los obesos y entre los no obesos. Un conjunto hipotético de datos obtenidos después de un número de años de seguimiento se muestra en la siguiente tabla.

Tabla 2: Enfermedad Coronaria según obesidad

Obesidad	Enfermos	No enfermos	Total
Presente	65	500	**565**
Ausente	25	650	**675**
Total	90	1150	**1240**

A partir de estos datos se encuentra que la proporción de personas con enfermedad coronaria entre los obesos es 65/565 (11,5%), mientras que la del grupo de no obesos es 25/675 (3,7%). Estas dos proporciones son las estimaciones de la incidencia (también llamado **riesgo**) de la enfermedad entre los que presentan el factor y los que no presentan el factor considerado de riesgo respectivamente.

Un procedimiento comúnmente usado para resumir los resultados de un estudio prospectivo es comparar estos riesgos estimados mediante lo que se conoce como el ***riesgo relativo.*** Esto es simplemente el cociente del riesgo estimado en cada grupo, dando aquí el valor 11,5/3,7=3,11.

El riesgo relativo le indica cuánto se incrementa el riesgo de enfermarse en un paciente expuesto al factor de riesgo, comparado con un paciente no expuesto al factor; y cuantifica el beneficio que puede tener el paciente si el factor de riesgo fue removido.

Así en el ejemplo el riesgo de sufrir enfermedad coronaria entre aquellas personas con obesidad se estima que es aproximadamente tres veces el riesgo de las personas que no presentan obesidad.

Si tanto obesos como no obesos tuvieran el mismo riesgo de enfermedad coronaria, el riesgo relativo en la población debería ser igual a 1. De esta manera, *un valor del RR igual a uno o cercano a uno indicaría que el factor considerado no es un factor de riesgo para la enfermedad.*

- *Si* $I_F \cong I_{NF} \Rightarrow RR \cong 1 \Rightarrow$ *NO ES FACTOR DE RIESGO*
- *Si* $I_F > I_{NF} \Rightarrow RR > 1 \Rightarrow$ *ES FACTOR DE RIESGO*
- *Si* $I_F < I_{NF} \Rightarrow RR < 1 \Rightarrow$ *ES FACTOR PROTECTOR*

Riesgo Atribuible

También se denomina diferencia de riesgos, se calcula mediante una resta entre la incidencia de expuestos y la incidencia entre no expuestos. También se puede calcular el riesgo atribuible porcentual.

Riesgo Atribuible

$$RA_{exp} = I_F - I_{NF}$$

Riesgo atribuible porcentual

$$\%RA_{exp} = RA_{exp} / I_F \times 100$$

El Riesgo Atribuible mide el exceso de riesgo a enfermar de los expuestos al factor en comparación a los que no presentan el factor.

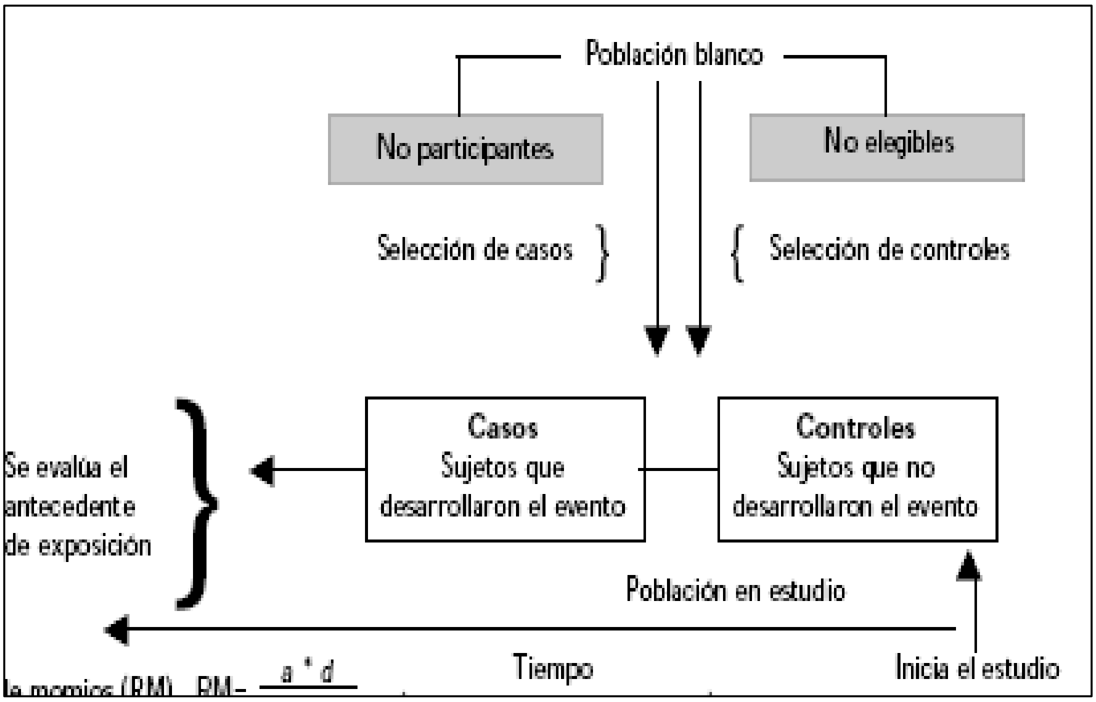
- *Si* $I_F \cong I_{NF} \Rightarrow RA \cong O \Rightarrow$ *NO ES FACTOR DE RIESGO*
- *Si* $I_F > I_{NF} \Rightarrow RR > 0 \Rightarrow$ *ES FACTOR DE RIESGO*
- *Si* $I_F < I_{NF} \Rightarrow RR < 0 \Rightarrow$ *ES FACTOR PROTECTOR*

Esto es simplemente una resta del riesgo estimado en cada grupo, en el ejemplo planteado este valor sería 11,5% - 3,7% = 7,8%. Lo que se interpretaría que entre los obesos la incidencia de enfermedad coronaria es un 7,8% mayor.

Estudio retrospectivo o de caso control

Un estudio retrospectivo comienza, como el prospectivo, con dos grupos de personas. Sin embargo, ahora, uno de los grupos está compuesto por individuos que ya tienen la enfermedad y el otro por personas que no tienen la enfermedad.

Diagrama de un estudio retrospectivo o de caso control



Nuevamente los resultados de este estudio se pueden resumir en una tabla de contingencia de 2x2 en la cual inicialmente se conocen los totales de casos (enfermos) y controles (no enfermos), y a partir de ellos se examina en cada grupo, por ejemplo por medio de una encuesta, cuantas personas han estado expuestas al factor de riesgo en el pasado obteniéndose una tabla como la siguiente:

Tabla 3: Datos genéricos de un estudio retrospectivo

Factor de riesgo	Casos (Enfermos)	Controles (No enfermos)	Total
Presente	a	b	a+b
Ausente	c	d	C+d
Total	a+c	b+d	a+b+c+d

En el ejemplo, para analizar la posible asociación entre relación entre la obesidad y el diagnóstico de cardiopatía coronaria, usando un diseño retrospectivo, se debería comenzar con un grupo de pacientes que haya sido diagnosticados con cardiopatía coronaria (casos) y un grupo que no presenta dicha patología (control). Una vez elegidos los grupos se busca la información acerca de la presencia de obesidad.

Odds Ratio de Exposición al Factor

Un conjunto hipotético de datos se muestra en la tabla siguiente:

Tabla 4: Enfermedad Coronaria según obesidad

Obesidad	Casos (Enfermos)	Controles (No enfermos)	Total
Presente	70	40	110
Ausente	30	60	90
Total	**100**	**100**	**200**

A partir de esta tabla podemos calcular las proporciones de expuestos al factor (obesidad) en cada grupo. Así, para el grupo de enfermos la proporción es 70/100 (70%), mientras que en el control (no enfermos) es 40/100 (40%).

Aquí *no podemos estimar el riesgo* en cada grupo como lo hicimos en el estudio prospectivo, ya que el número de expuestos al factor y no expuestos al factor no están bajo el control del investigador y las proporciones calculadas (70% y 40%) no son la incidencia de la enfermedad., lo que se puede calcular son las chances u odds de exposición en casos y controles, lo que permite contar con otra medida que indica el ***grado de asociación*** entre las variables consideradas conocida como ***odds ratio*** **(OR)**:

Chance u "odds" de exposición

Odds exposición en Enfermos (casos)= $odds_E$= a/c

Odds exposición en No Enfermos (controles)= $odds_{NE}$= b/d

Odds Ratio de Exposición al Factor.

$$OR^{(F)} = \frac{odds_E}{odds_{NE}} = \frac{a/c}{b/d} = \frac{a \times d}{c \times b}$$

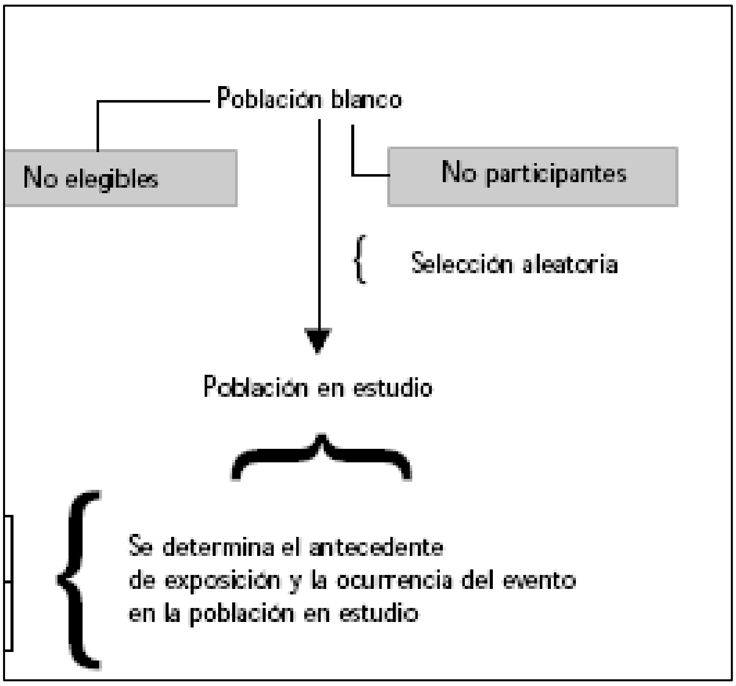
Un odd ratio igual a 1 indica que no hay asociación entre las variables estudiadas. Así si en un caso particular se obtiene valores de OR cercanos a uno se podría inferir que la asociación no es significativa. El odd ratio estimado para los datos del ejemplo es OR=70x60/30x40=3,5 lo que indicaría que dentro del grupo de los enfermos (casos) la proporción de expuestos al factor es mayor que la de los no enfermos (control).

- *Si* $odds_E \cong odds_{NE}$ *OR≅1⇒ NO HAY ASOCIACIÓN ENTRE FACTOR Y ENFERMEDAD*
- *Si* $odds_E > odds_{NE}$ *OR>1⇒ FACTOR POSITIVAMENTE ASOCIADO A ENFERMEDAD*
- *Si* $odds_E < odds_{NE}$ *OR<1⇒ FACTOR NEGATIVAMENTE ASOCIADO A ENFERMEDAD*

Debe notarse que OR no es una medida de riesgo, sino solo de asociación

Estudios de Corte Transversal

En este tipo de estudio donde se parte de un grupo de individuos sin predeterminar con anterioridad ninguna característica. De la población se selecciona un grupo de personas y se observa si está enfermo o no y si estuvo expuesto o no al factor de riesgo. Es decir que inicialmente solo conocemos en total de individuos incluidos en el estudio.



Obtendríamos finalmente la siguiente información:

Tabla 5: Datos genéricos de un estudio de corte transversal

Factor de riesgo	Enfermos	No enfermos	Total
Presente	a	b	a+b
Ausente	c	d	C+d
Total	a+c	b+d	a+b+c+d

Razón de Prevalencias

En este tipo de estudios es posible calcular la prevalencia de la enfermedad en los grupos de exposición (presente y ausente), por lo que la medida de asociación adecuada a calcular es la Razón de Prevalencias (RP). Esta se calcula de la misma forma que el Riesgo Relativo, pero su interpretación es distinta ya que no es posible conocer la incidencia de la enfermedad sino solo la prevalencia.

Determinar Prevalencia en:

Expuestos al Factor de Riesgo: $P_F = a/(a+b)$

No expuestos al Factor de Riesgo: $P_{NF} = c/(c+d)$

$$\boxed{RP = \frac{P_F}{P_{NF}} = \frac{a/(a+b)}{c/(c+d)}}$$

Esta medida se interpreta como una razón que nos indica cuántas veces es más frecuente la prevalencia de la enfermedad entre los expuestos al factor en relación a los no expuestos.

- *Si $P_F \cong P_{NF}$ $RP \cong 1 \Rightarrow$ NO HAY ASOCIACIÓN ENTRE EL FACTOR Y ENFERMEDAD (La prevalencia de la enfermedad es igual en expuestos y no expuestos)*
- *Si $P_F > P_{NF}$ $RP > 1 \Rightarrow$ FACTOR POSITIVAMENTE ASOCIADO A LA ENFERMEDAD (La prevalencia de la enfermedad es mayor en el grupo de expuestos al factor)*
- *Si $P_F < P_{NF}$ $RP < 1 \Rightarrow$ FACTOR NEGATIVAMENTE ASOCIADO A LA ENFERMEDAD (La prevalencia de la enfermedad es menor en el grupo de expuestos al factor)*

Odds Ratio de Enfermedad

En los estudios de corte transversal también es posible calcular las chances u odds de la Enfermedad

Chance u "odds" de la enfermedad

Odds de Enfermedad entre expuestos al factor: $odds_F$= a/b

Odds de Enfermedad entre no expuestos al factor: $odds_{NF}$= c/d

Odds Ratio de la Enfermedad:

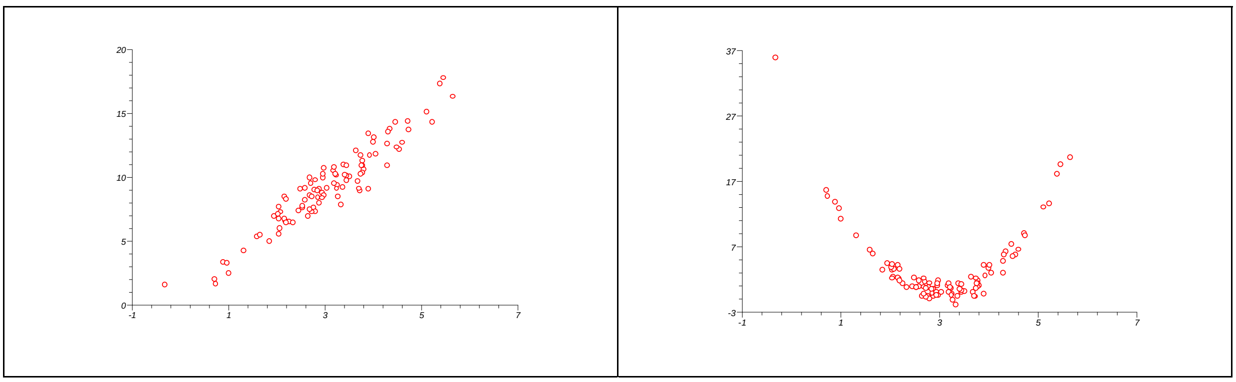
$$\boxed{OR^{(E)} = \frac{odds_F}{odds_{NF}} = \frac{a/b}{c/d} = \frac{a \times d}{c \times b}}$$

- *Si* $odds_F \cong odds_{NF}$ $\quad OR \cong 1 \Rightarrow$ *NO HAY ASOCIACIÓN ENTRE FACTOR Y ENFERMEDAD*
- *Si* $odds_F > odds_{NF}$ $\quad OR > 1 \Rightarrow$ *FACTOR POSITIVAMENTE ASOCIADO A ENFERMEDAD*
- *Si* $odds_F < odds_{NF}$ $\quad OR < 1 \Rightarrow$ *FACTOR NEGATIVAMENTE ASOCIADO A ENFERMEDAD*

Relación entre variables cuantitativas

Suponga, por ejemplo, que se desea estimar la relación entre la concentración de colesterol y el cambio en el diámetro vascular medio en pacientes con angina de pecho estable (variables cuantitativas), en el sentido de que *cuánto se asocia un cambio en el diámetro vascular medio a un cambio en los niveles de colesterol*. Como ya se vio en el Módulo de Estadística Descriptiva, una manera de graficar esta relación es a través de un diagrama de dispersión o correlación. ¿Qué debemos observar de ese diagrama para que se pueda visualizar algún tipo de asociación? La respuesta es ver si los puntos de ese diagrama son tales que se puede pensar que están agrupados alrededor de alguna línea, como se puede observar en los gráficos siguientes, donde el primero muestra una agrupación alrededor de una línea recta, y el segundo alrededor de una curva.

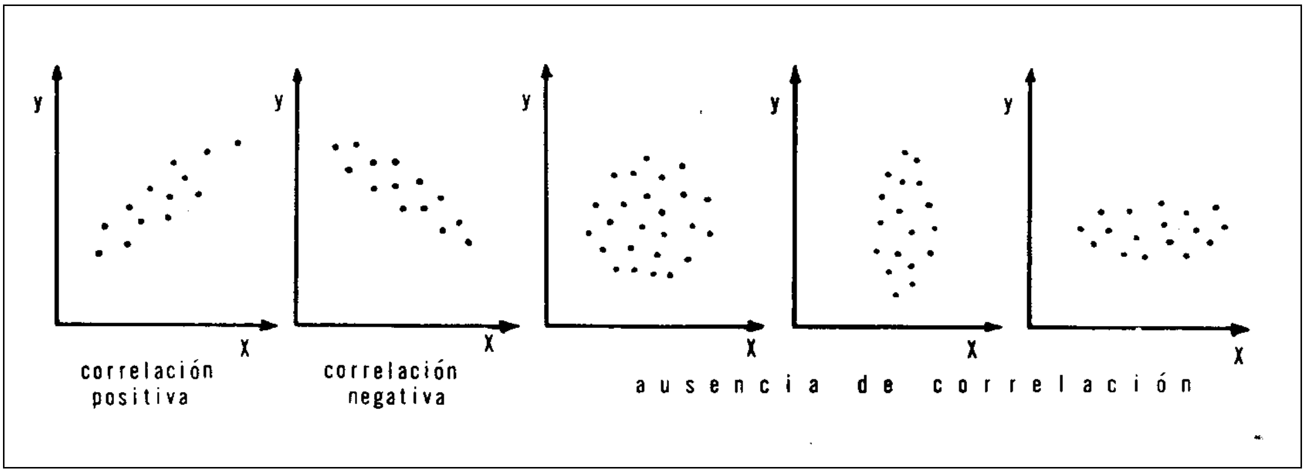
Gráfico N°1: Gráfico de dispersión o correlación



Coeficiente de Correlación de Pearson

Si se consideran los tipos de agrupamientos alrededor de una línea, una medida de este tipo de asociación o correlación (lineal) es el ***Coeficiente de Correlación de Pearson***. En la figura 1 del gráfico 2 se observa que cuando aumenta el valor de X, también aumenta el valor de Y, mientras que en la figura 2 cuando aumenta el valor de X, disminuye el de Y (X,Y son las variables que se están analizando). En el primer caso se dice que hay una correlación positiva y en el segundo una negativa. Las tres últimas figuras muestran casos de ausencia de correlación, donde una aumento o disminución de X no implica un aumento o disminución de Y.

Gráfico N°2: Gáficos de correlación



Volviendo al ejemplo, si con X designamos a la concentración de colesterol, y con Y al cambio del diámetro vascular medio, y si se denota con r al ***Coeficiente de Correlación de Pearson***, su cálculo puede hacerse usando la expresión:

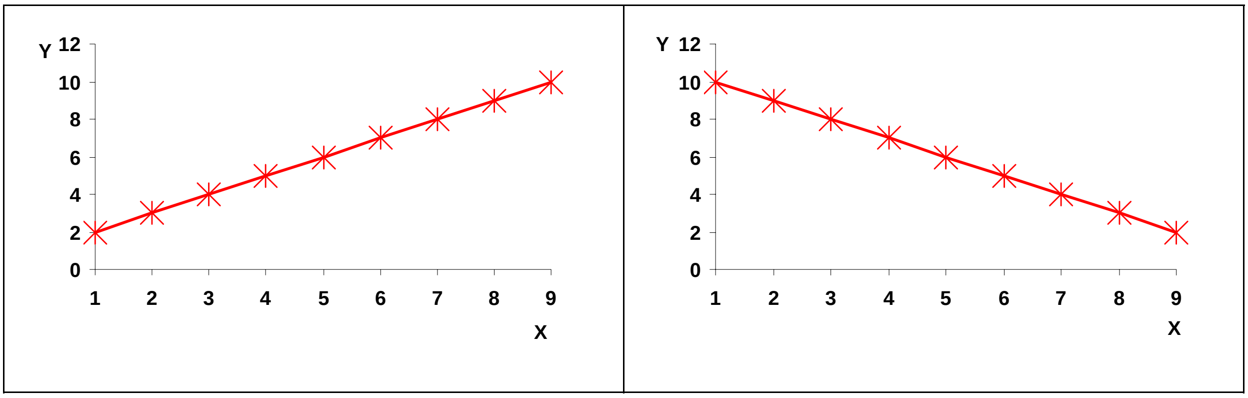
$$r = \frac{\sum\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)}{\sqrt{\sum\left(X-\overline{X}\right)^2}\sqrt{\sum\left(Y-\overline{Y}\right)^2}}$$

donde $\overline{X}$ e $\overline{Y}$ son las medias de la concentración de colesterol y del diámetro vascular medio respectivamente.

Con los datos de los valores típicos de colesterol total/LAD, y cambio del diámetro vascular medio en 39 pacientes con angina de pecho, cuya lesión no creció; se calculó el coeficiente de correlación y se obtuvo r = 0.5. ¿Qué significa una correlación de 0.5 entre colesterol y cambio en el diámetro vascular?

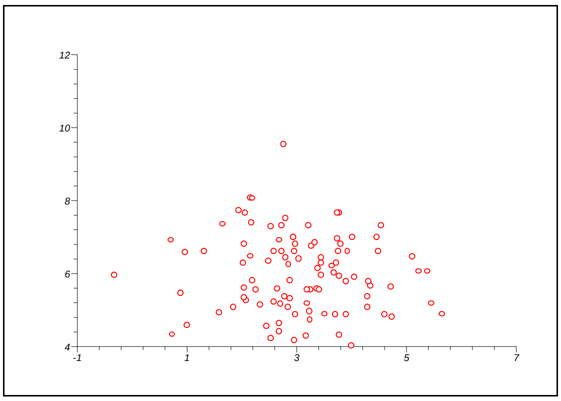
Los posibles valores de r oscilan entre –1 y +1, en donde +1 describe una línea recta perfecta con pendiente positiva, como se observa en la primera figura del Gráfico Nº3 y -1 describe una línea recta perfecta con pendiente negativa, como se muestra en la segunda figura del mismo gráfico

Gráfico Nº3: Tipos de correlación lineal



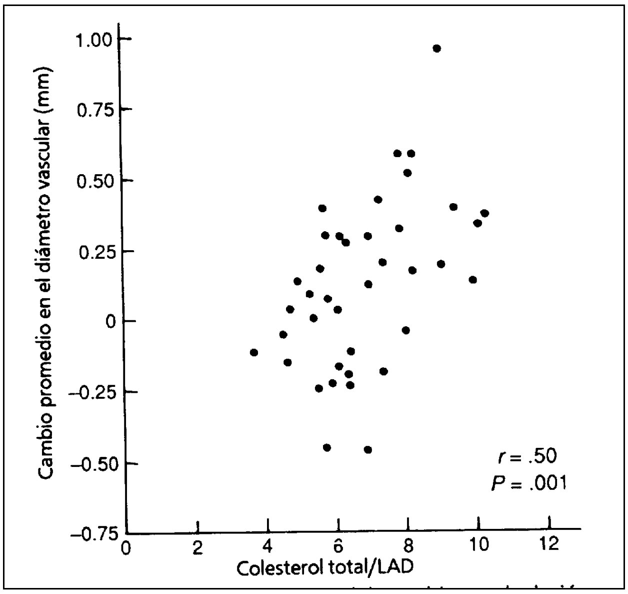
Una correlación de 0 (o sea r=0) significa que no hay relación lineal entre las dos variables. Existe una correspondencia entre la cifra del coeficiente de correlación y cuán dispersas están o no las observaciones alrededor de una línea recta. Cuando la correlación se aproxima a 0, la forma del gráfico de dispersión es más o menos circular, como se observa en el siguiente gráfico.

Gráfico Nº4: Correlación entre dos variables



Conforme el valor de la relación se aproxima a +1 o –1, la forma se vuelve más elíptica, hasta que, en +1 o –1, las observaciones quedan directamente sobre la línea recta. Con una correlación de 0.5, cabe esperar una dispersión de datos en forma más o menos oval, como en el siguiente gráfico.

Gráfico N°5: Correlación entre colesterol total y cambio en el diámetro vascular.



En ocasiones, la correlación se eleva al cuadrado (r^2) para formar un valor estadístico importante llamado ***coeficiente de determinación***. Para los datos de colesterol y diámetro vascular, el coeficiente de determinación es $(0.5)^2$ ó 0.25, que significa que el colesterol total explica el 25 % de los cambios del diámetro vascular medio. De otra manera si se conociera el valor del colesterol de los pacientes y se tomara en consideración al examinar el cambio del diámetro del vascular medio, la variabilidad de esta última medición podría reducirse un 25 %.

El coeficiente de correlación posee varias características.

- *Es independiente de cualquier unidad usada para medir las variables,* es decir es adimensional
- *El valor del coeficiente de correlación se altera en forma importante por la presencia de un valor alejado o distante.* Por tanto la correlación no proporciona una descripción adecuada entre dos variables cuando la distribución de una u otra variable está sesgada o incluye valores distantes.
- *El coeficiente de correlación de Pearson mide sólo el grado de asociación lineal*; de hecho, dos factores pueden guardar una relación no lineal fuerte, aún cuando la correlación lineal es bastante pequeña. Por tanto, cuando se analizan las relaciones entre dos características, los datos se deberán graficar antes de calcular el coeficiente de correlación.

- *Correlación no implica causalidad*. El enunciado de que una característica causa otra, debe justificarse basándose en observaciones experimentales o argumentos lógicos, no con fundamento en el tamaño de un coeficiente de correlación.
- El coeficiente de correlación de Pearson *no es adecuado cuando las variables provienen de distribuciones asimétricas.*

Coeficiente de Correlación de Spearman.

En el caso de falta de conocimiento a cerca de la distribución, o de la linealidad de la relación entre las variables, se puede usar métodos de rango para evaluar una relación más general entre los valores de las variables mencionadas.

El coeficiente de correlación que se usa con frecuencia para describir la relación entre dos características que no cumplen con las condiciones antes mencionadas ya sea porque las distribuciones de las variables son asimétricas o están medidos en escala ordinal (o una ordinal y una numérica) o hay observaciones alejadas, o el tamaño de muestra es pequeño, es el ***coeficiente de correlación de Spearman.***

Este coeficiente puede variar de –1 a +1 igual que el coeficiente de Pearson, pero +1 o –1 indican concordancia perfecta entre posiciones de los valores o categorías de la variable ordinal en lugar de entre los valores mismos.

Para calcular el coeficiente de correlación de rangos de Spearman ($\mathbf{r_s}$), se ordena separadamente de menor a mayor los valores de ambas variables y se les asigna rangos (indicando el orden que ocupan), luego se sustituye los valores de los rangos en la fórmula del coeficiente de correlación de Pearson.

Aunque la fórmula de Spearman se deriva de la de Pearson, esto no implica que los valores de ambos coeficientes vayan a coincidir siempre.

Relación entre dos características ordinales

Otras medidas de correlación de datos ordenados son ***Tau de Kendall y W de Kendall*** basadas también en rangos. La Tau generalmente da valores menores de la correlación si se la compara con el coeficiente de correlación de Pearson cuando se puede usar este último.

CAPÍTULO IV- ESTIMACIÓN DE PARÁMETROS

Antes de desarrollar los conceptos básicos sobre las técnicas estadísticas que nos permitirán estimar parámetros a partir de datos muestrales es importante dar algunos conceptos sobre ***distribución muestral.***

Un ***estadístico*** (media aritmética, mediana, proporción, RR, OR, etc.) es una forma de combinar los datos (función de los datos muestrales), por ejemplo la media es la suma de los datos divida en el número de datos.

Para un conjunto particular de datos (valores observados en una muestra) el valor obtenido de esa función nos da una ***estimación*** del ***parámetro poblacional***. Obviamente con cada posible muestra tendríamos diferentes valores del estadístico, es decir obtendríamos diferentes estimaciones. Por lo tanto, **un estadístico es una variable** que toma diferentes valores, y estos valores o estimaciones depende de la particular muestra con que estemos trabajando.

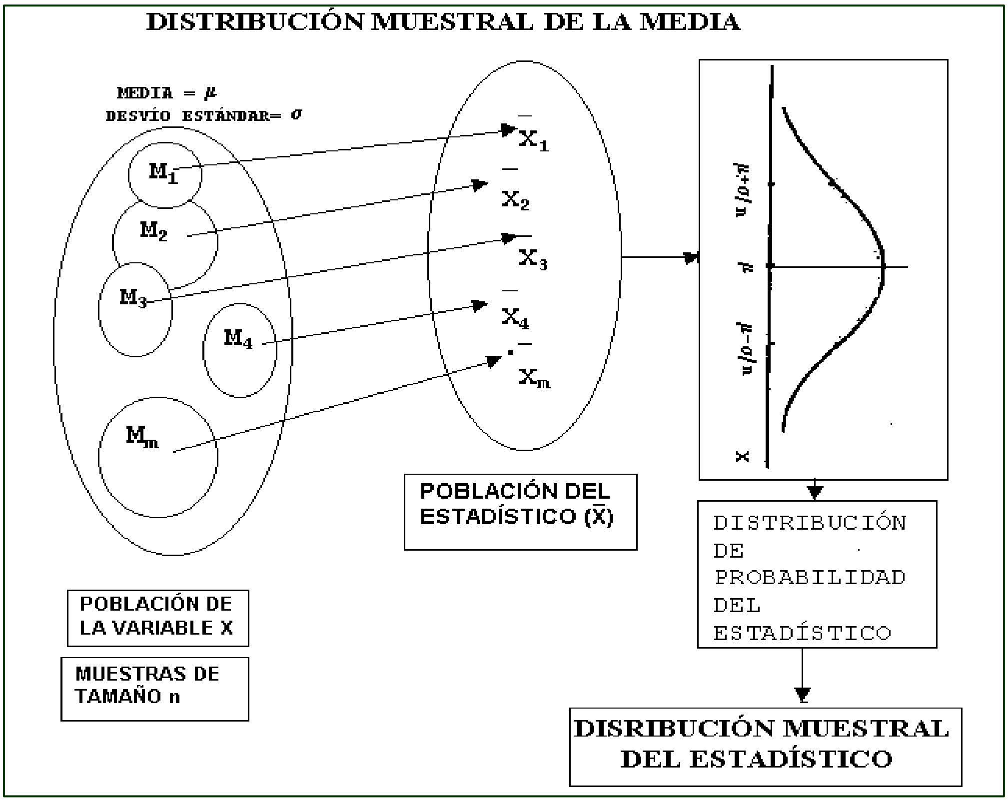
Si pudiéramos extraer todas las posibles muestras de un dado tamaño de la población de interés y con cada una de ellas calculamos el estadístico correspondiente obtendríamos todas las estimaciones posibles, a partir de ella podríamos construir la distribución del estadístico que recibe el nombre de **distribución muestral del estadístico.**

Cabe preguntarse por qué es tan importante el concepto de distribución muestral, la respuesta es simple. Cuando se quiere estimar un parámetro poblacional (característica de la población) a partir de una muestra surgen interrogantes como: "¿qué tan buena es la estimación obtenida?, ¿se puede llegar a la conclusión de que el parámetro de la población es idéntico al estadístico de la muestra o es probable que exista algún error?. Si es así, ¿qué tan grande es dicho error?" Para responder a estas preguntas se debe comparar los resultados obtenidos a partir de las muestras con los resultados "esperados". Los resultados esperados son justamente dados por la distribución muestral del estadístico y de allí la importancia de ella.

Surgen ahora otros interrogantes, ¿cómo es en realidad una distribución muestral y, por lo tanto, cuáles son los resultados esperados? La distribución muestral del estadístico depende de:

- la distribución de la población,
- el estadístico que se elija para estimar el parámetro,
- la forma de selección aleatoria de la muestra,
- el tamaño de la misma.

Distribución muestral de la media aritmética (estadístico).



La relación existente entre ambas distribuciones (distribución de la población de individuos y distribución muestral del estimador) es la que nos permite hacer afirmaciones sobre el parámetro poblacional y cuantificar el error de dichas afirmaciones.

Debemos ser conscientes que en general es imposible obtener todas las posibles muestras de una población, pero la estadística inferencial nos provee herramientas que nos permiten conocer la distribución muestral teórica del estadístico, y a partir de ella hacer afirmaciones sobre la precisión de la estimación y cuantificar el error de las afirmaciones que se hagan sobre ella.

A fin de clarificar estos conceptos, se considera por el momento el caso en que el parámetro poblacional es la media **μ** y el estadístico para estimarla es la media aritmética $\bar{x}$, obtenida a partir de una muestra de tamaño ***n*** de la población.

Es claro que si se quiere obtener la distribución muestral de $\bar{x}$, extrayendo todas las muestras de tamaño *n*, esto consumiría más tiempo que el requerido para tomar la información de toda la población y, en consecuencia, sería poco práctico. En su lugar, es posible usar la teoría estadística para determinar la distribución muestral de la media aritmética en cualquier situación particular. Por ejemplo, supongamos una población de personas adultas donde la presión arterial sistólica (PAS) tiene media poblacional μ=120 mm Hg con un desvío estándar σ=10 mm Hg. Supongamos ahora, que se desconoce esta información y se decide estimar la media poblacional tomando una muestra aleatoria de tamaño n=100 de la población. La media aritmética calculada a partir de la muestra dio un valor $\bar{x}$=121 mm Hg. Para hacer afirmaciones sobre la precisión de la estimación que nos dé algún grado de confianza en el valor encontrado a través de la muestra, necesitamos conocer la distribución muestral de $\bar{x}$. Las propiedades de la distribución muestral de $\bar{x}$ son la base para uno de los teoremas más importantes de la teoría estadística, llamado Teorema del Límite Central.

Dada una población con media **μ** y desviación estándar **σ**, la distribución muestral de la media basada en muestras aleatorias repetidas de tamaño ***n*** tiene las siguientes propiedades:

1. *La media de una **distribución muestral** o media de medias, es igual a la media de la población μ.*

2. *La desviación estándar en la distribución muestral de $\bar{x}$ es igual a $\frac{\sigma}{\sqrt{n}}$. Esta cantidad denominada **error estándar de la media (SEM),** tiene una función importante en numerosos procedimientos estadísticos.*

3. *Con muestras de tamaño grande la distribución muestral de $\bar{x}$ sigue un modelo teórico denominado modelo de **distribución normal,** sin importar la forma de la distribución de la población original.*

Esto nos da la base para toda la inferencia estadística sobre la media. Así, en el ejemplo si se supone σ conocido, es decir conocemos que σ=10 mm Hg. se puede afirmar que la distribución muestral *de* $\bar{x}$ en este caso tiene un error estándar

$$\text{SEM}= \frac{\sigma}{\sqrt{n}} = 10/10 = 1$$

ya que la raíz cuadrada de n=100 es 10.

La importancia del SEM radica en que a partir de él se puede hacer la siguiente afirmación: ***si el tamaño de muestra es grande aproximadamente el 95% de las muestras darían valores de*** $\bar{x}$ ***en un intervalo que va desde*** μ***-2SEM a*** μ***+2SEM.***

Obviamente, en la práctica siempre se desconoce el valor de μ y casi siempre el valor de σ, de manera que esto es sólo el respaldo teórico de toda la inferencia estadística, como se verá en las secciones subsiguientes.

En el ejemplo de PAS, si μ y σ fueran conocidos entonces se puede afirmar que aproximadamente el 95% de las muestras de tamaño n=100 darían valores *de* $\bar{x}$ entre 120-2 y 120+2.

Si el estadístico elegido para estimar la media poblacional no es la media aritmética sino cualquier otro estimador, por ejemplo la mediana, entonces el cálculo de su error estándar y su distribución muestral no sigue exactamente lo enunciado para el caso *de* $\bar{x}$, sino que ello deberá determinarse en cada caso, basándose en herramientas de la teoría estadística.

ESTIMACIÓN

Nuestro objetivo es estimar, de alguna forma, el o los parámetros que caracterizan a nuestra población. El estadístico a utilizar, dependerá del parámetro a estimar, y para un dado parámetro en general hay más de un estadístico que nos podría resultar de utilidad.

Estimación Puntual

Se denomina Estimación Puntual al valor obtenido del estadístico o estimador a partir de una muestra. A partir de los datos muestrales se calculan los estadísticos (media, desvío estándar, proporciones, etc.) que son los estimadores puntuales de los parámetros poblacionales.

Debemos preguntarnos cuáles son las propiedades de que debería tener un estimador para que sea considerado como bueno y que nos facilite la elección entre un estadístico y otro. Al obtener una estimación puntual debemos ser conscientes que él depende de la muestra que hayamos seleccionado y que el valor encontrado puede cambiar de muestra en muestra. Así, las propiedades deseables serían que con cada estimación no estemos muy alejados del verdadero valor del parámetro y que no haya demasiada diferencia entre los valores del estadístico obtenidos de muestra en muestra. Esto se puede formalizar diciendo que:

Las propiedades a tener en cuenta de los estadísticos son las siguientes:

1. **Insesgado**: Ausencia de error sistemático, el estadístico tiene como valor promedio el parámetro que se busca estimar.
2. **Varianza Mínima**: Las estimaciones obtenidas para distintas muestras varían poco entre ellas.

Si la variabilidad de las estimaciones se mide a través del desvío estándar, este desvío estándar recibe el nombre de ***error estándar del estimador* (SE).**

No debe confundirse entre "desvío estándar" de la distribución de la población (variabilidad entre los individuos) y "error estándar" del estimador que es el desvío estándar de la distribución muestral (variabilidad entre las estimaciones de las muestras).

Si se consideran todos los estadísticos (estimadores) insesgados posibles de algún parámetro, aquél con la varianza más pequeña recibe el nombre de **estadístico insesgado más eficiente** del parámetro. Se puede demostrar que la media aritmética es un estimador insesgado y eficiente de la media poblacional.

Es muy probable que el estadístico insesgado más eficiente no estime el parámetro poblacional con "exactitud", esto se debe a que en realidad cuando realizamos la estimación sólo tomamos una muestra, y obtenemos uno de los posibles valores del estadístico que en general no tiene porque coincidir con el valor del parámetro que se quiere estimar. Si bien la precisión se incrementa con muestras grandes no hay razón para esperar que la estimación puntual de una muestra dada deba ser exactamente igual al parámetro poblacional que se supone estima. Entonces, existen muchas situaciones en las cuales es preferible determinar un intervalo dentro del cual se esperaría encontrar el valor del parámetro, tal metodología se conoce como estimación por intervalos de confianza.

Estimación por intervalos

Consiste en un conjunto de procedimientos mediante los cuales, según el nivel de significación deseado, se calcula el intervalo en el cual se podría encontrar el parámetro estimado a un determinado nivel de confianza, que generalmente varía entre el 95% y 99%.

Intervalos de Confianza: Son intervalos aleatorios obtenidos a partir de los datos muestrales y en los cuales hay un *grado de confianza* prefijado (medido en %) de que dicho intervalo contenga al verdadero valor del parámetro (valor poblacional) que se quiere estimar.

El grado de confianza se denomina ***nivel de confianza*** y lo denotaremos como $100(1-\alpha)\%$. Usualmente este valor corresponde a un 95%. Donde:

$(1-\alpha)$ = es el grado de confianza o coeficiente de confianza

α = es la probabilidad de error de que el parámetro poblacional no se encuentre dentro del intervalo de confianza. Este valor es fijado por el investigador y generalmente varía entre el 1% y 5% (α=0,01 o α =0,05)

Para encontrar estos intervalos debemos conocer la distribución muestral de cada estimador (tipo de distribución teórica y error estándar del estimador), que como ya se vio esto depende del parámetro de interés y del estadístico que se elija para estimar dicho parámetro. Sin embargo es posible dar la forma general que adopta un intervalo de confianza en cualquier caso.

Supongamos que se quiere estimar un parámetro Q de la población a través del estadístico q, si el error estándar de la distribución de q lo denotamos con SE(q), entonces un **intervalo de confianza para Q con una confianza del 95% ($IC_{95\%}$)** viene dado por la expresión:

$$\boxed{\mathbf{IC_{95\%} = [q - k_1 SE(q)\ ,\ q + k_2 SE(q)]}}$$

donde k_1 y k_2 dependen de la forma de la distribución de q. Los límites inferior y superior del intervalo están dados por q – k_1SE(q) y q + k_2SE(q) respectivamente

En el ejemplo de la PAS el intervalo de confianza para la media poblacional con una confianza del 95% está dado por:

$$IC_{95\%} = [121 - 1.96x1 ; 121 + 1.96x1] = [119.04 ; 122.96]$$

Interpretación y usos del Intervalo de Confianza: Hay dos aspectos importantes en la interpretación de un intervalo de confianza a saber:

- La amplitud del IC mide el grado de precisión de la estimación puntual, es decir cuando menor es la amplitud mayor es la precisión de la estimación. Esto viene del hecho de que la amplitud del IC depende del SE el cual, como ya se vio, mide el grado de variabilidad de las estimaciones de muestra en muestra.
- Dado el $IC_{100(1-\alpha)\%}$ para Q, existe una probabilidad del 100(1-α)% que el intervalo estimado contenga al verdadero valor del parámetro poblacional Q. En el ejemplo podríamos decir que existe una probabilidad del 95% que el $IC_{95\%}$ =[119.04, 122.96] contenga al verdadero valor de μ.
- El IC nos permite hacer comparaciones entre poblaciones o diferentes estimaciones de una misma población. Por ejemplo, supongamos que la PAS media en otra población B es de μ_B=115 mm Hg , luego a partir del IC se puede concluir que la PAS media en nuestra población es significativamente mayor que la correspondiente a la población B, ya que μ_B es menor que el extremo menor del intervalo.

A continuación vamos a ver algunos ejemplos de estimaciones de parámetros.

Estimación de la media aritmética

Sabemos que $\bar{x}$ (media muestral) estima a μ (media poblacional), y se calcula con la siguiente fórmula:

$$\bar{x} = (\sum x_{obs})/n$$

El error estándar de la media (SEM)es igual a $\frac{\sigma}{\sqrt{n}}$ en el caso que se conozca σ, pero si σ es desconocido lo reemplazamos por SD (desvío estándar) donde:

$$SD = \sqrt{\frac{suma\,de\,los\,desvios\,al\,cuadrado}{n-1}} = \sqrt{\frac{suma\,(x_{obs} - \bar{x})^2}{n-1}}$$

y el SEM es igual a $\dfrac{SD}{\sqrt{n}}$

Bajo el supuesto de normalidad; es decir que suponemos que la distribución muestral de la media es normal; podemos decir que $\bar{x}$ es una estimación precisa de μ cuando el tamaño muestral (n) es grande.

Si fijamos **α** = 0,05 , recordemos que **α** es el error, y 1-**α** es el nivel de confianza.

(1-α**)** 100% = (1-0,05)100%= 95%; entonces el intervalo del 95% de confianza de μ es:

$IC_{95\%}$ [$\bar{x}$-$Z_{\alpha/2}$ SEM ; $\bar{x}$ + $Z_{\alpha/2}$ SEM] (1)

Como la distribución de $\bar{x}$ es normal $Z_{\alpha/2}$ =$Z_{0,025}$ = 1,96 (este valor se obtiene de la tabla de valores de la distribución normal)

Ejemplo del cálculo e interpretación de un intervalo de confianza: Supongamos que en una muestra de tamaño n=36, el peso promedio de recién nacidos prematuros es de 1930 grs. con un desvío estándar de 325 grs. Se quiere conocer cual es el intervalo del 95% de confianza para el peso promedio de los prematuros.

n= 36 ; $\bar{x}$= 1930 grs. ; SD= 325 grs. ; SEM= $\dfrac{325}{\sqrt{36}}$ = 54,17

α= 0,05 ; α/2= 0,025 ; $Z_{0.025}$= 1,96

Usando la formula (1) el intervalo del 95% de confianza para la media es

$IC_{95\%}$ = [1930 – 1,96 (54,17) ; 1930 +1,96 (54,17)]

$IC_{95\%}$ = [1930 – 106,17 ; 1930 +106,17]

$IC_{95\%}$ = [1823,83 ; 2036,17]

El intervalo del 95% de confianza para el peso promedio de los prematuros es de 1824grs. a 2036grs.

Estimación de una proporción

Proporción (p): es un cociente que relaciona una parte con el total. Es un cociente que se obtiene dividiendo el número de individuos que poseen determinada característica con el total de individuos. Esta se calcula de la siguiente forma:

$$p = \frac{x}{n}$$

Donde ***x*** representa al número de individuos que pertenecen a una determinada categoría y ***n*** es el total de individuos.

Como ***p*** estima a la proporción en la población, el error estándar de la proporción $SE(p)$ es igual a:

$$SE(p) = \sqrt{\frac{p \times q}{n}} \qquad \text{donde } q = 1\text{-}p$$

Bajo el supuesto de normalidad; es decir que suponemos que la distribución muestral de la proporción es normal; podemos decir que *p* es una estimación precisa de la proporción de la población cuando el tamaño muestral (n) es grande.

Si fijamos $\alpha = 0{,}05$, recordemos que α es el error, y $1-\alpha$ es el nivel de confianza.

$(1-\alpha)\,100\% = (1-0{,}05)100\% = 95\%$; entonces el intervalo del 95% de confianza de *p* es:

$IC_{95\%}$ $[p - Z_{\alpha/2}\, SE(p)\ ;\ p + Z_{\alpha/2}\, SE(p)]$ (2)

Como la distribución de *p* es normal $Z_{\alpha/2} = Z_{0,025} = 1{,}96$ (este valor se obtiene de la tabla de valores de la distribución normal)

Ejemplo del cálculo e interpretación de un intervalo de confianza para una proporción: Supongamos que en una muestra de n=750 niños se encontraron en estado de desnutrición 34 de ellos; se quiere conocer cual es el intervalo del 95% de confianza para la proporción de desnutridos.

n= 750 ; $p = \frac{34}{750} = 0{,}045$; q= 1-0,045= 0,955 ; $SE(p) = \sqrt{\frac{0{,}045 \times 0{,}955}{750}} = 0{,}0076$

$\alpha = 0{,}05$; $\alpha/2 = 0{,}025$; $Z_{0.025} = 1{,}96$

Usando la formula (2) el intervalo del 95% de confianza para la proporción es

$IC_{95\%}$ = [0,045 – 1,96 (0,0076) ; 0,045 +1,96 (0,0076)]

$IC_{95\%}$ = [0,045 – 0,015 ; 0,045+0,015]

$IC_{95\%}$ = [0,030 ; 0,060]

$IC_{95\%}$ = [3% ; 6%]

El intervalo del 95% de confianza para la proporción de desnutridos es de 0,03 a 0,06.Las proporciones también pueden expresarse en porcentajes, en ese caso podríamos decir que la proporción de desnutridos es de 4,5% , y el intervalos del 95% de confianza es del 3% al 6%.

Estimación de medidas de asociación

En el capítulo anterior se definieron las principales medidas para analizar la asociación entre variables, a estas estimaciones puntuales se deben acompañar los respectivos intervalos de confianza. Para cuantificar la precisión de la asociación se realiza el cálculo de los intervalos de confianza, normalmente estimados para un nivel de confianza del 95%; esto es, si se repitiera el mismo estudio n veces, bajo las mismas suposiciones estadísticas, en 95% de los casos el estimador puntual (r, RR, RP, OR, etc) estará contenido dentro de los límites estimados. Para cada medida de asociación calculada se debe tener en cuenta cuál es el valor nulo o punto de corte en el cual se puede concluir que no existe asociación entre las variables, y se podrá inferir que existe asociación significativa sólo cuando el intervalo de confianza no contenga dicho valor.

El coeficiente de correlación de Pearson (r) mide la relación lineal entre dos variables cuantitativas, este puede tomar valores en el rango [-1 ; 1], siendo el valor nulo o punto de corte r=0; por lo tanto cuando el intervalo de confianza de este coeficiente contenga al cero no podremos concluir que exista asociación o correlación lineal significativa entre dos variables cuantitativas.

En el caso de variables cualitativas las medidas de asociación y riesgo, dependiendo del diseño del estudio, son RR (Riesgo Relativo), RP (Razón de Prevalencias) y OR (Odds Ratio), siendo en todos estos casos el valor nulo o punto de corte el uno (1); por lo tanto cuando el intervalo de confianza contenga este valor no podremos concluir que exista asociación significativa entre las variables en estudio.

En el caso de RR se podrá concluir que el factor estudiado se considera de riesgo cuando el intervalo se encuentre en valores mayores que uno, y será considerado factor protector a intervalos menores que uno; siempre teniendo en cuenta que el intervalo no contenga al valor nulo (uno). La interpretación del intervalo de confianza de RP es similar, pero se podrá concluir que la prevalencia es mayor o menor, o que existe asociación si el intervalo de confianza no contiene al uno.

También se puede calcular el impacto de la exposición mediante el riesgo atribuible (RA), llamado también diferencias de riesgos, el cual mide la proporción de la enfermedad que se evitaría si se lograra erradicar la exposición. El valor nulo del RA es el cero (0), por lo tanto cuando este valor se encuentre en el intervalo de confianza, la diferencias de riesgos no será significativa.

La interpretación de los resultados en los estudios retrospectivos es la siguiente: si OR igual a uno, la exposición no está asociada con el evento o enfermedad; si OR es menor de uno, la exposición está asociada de manera inversa con el evento, esto es, la exposición disminuye la posibilidad de desarrollar el evento; si OR es mayor de uno, la exposición se encuentra asociada positivamente con el evento, lo que quiere decir que la exposición aumenta la posibilidad de desarrollar el evento y si OR es igual a uno no existe asociación entre las variables; por lo tanto para poder inferir los resultados a la población es necesario considerar si el intervalo de confianza contiene el valor nulo del OR, es decir el uno. En caso de que el intervalo contenga este valor podremos concluir que la asociación entre las variables involucradas en el estudio no es significativa.

Características de las Medidas de Asociación, Correlación y Riesgo

Tipo Variables	*Tipo Estudio*	*Medida*	*Valor Nulo*	*Interpretación*
Cuantitativas	Corte Transversal	r	0	Mide correlación lineal
Cualitativas	Prospectivo	RR	1	Mide Riesgo
Cualitativas	Prospectivo	RA	0	Mide Riesgo
Cualitativas	Corte Transversal	RP	1	Mide asociación
Cualitativas	Retrospectivo	OR	1	Mide asociación

CAPÍTULO V- TEST DE HIPÓTESIS ESTADÍSTICA

En las secciones anteriores tratamos con estimación y precisión de las estimaciones que es una forma de inferencia estadística. En esta sección se introduce una forma diferente de inferencia, los test de hipótesis estadísticas.

Un test de hipótesis es una metodología o procedimiento que permite cuantificar la probabilidad del error que se cometería cuando se hace una afirmación sobre la población bajo estudio, es decir, nos permite medir la fuerza de la evidencia que tienen los datos a favor o en contra de alguna hipótesis de interés sobre la población. Es un procedimiento de decisión basado en datos muestrales.

Hipótesis estadística se define como una afirmación acerca de una o más poblaciones.

Para ilustrar los conceptos de los tests de hipótesis supongamos que estamos interesados en conocer cuanto influye el nivel de escolaridad de la madre sobre el estado nutricional del niño, es decir, nuestro interés es saber si en nuestra población el estado nutricional del niño está asociado al nivel de escolaridad de la madre. En principio, este interrogante se plantea porque tenemos la sospecha que realmente existe tal asociación. Para investigar sobre este punto, se toma una muestra de niños y se analiza en cada uno de ellos el estado nutricional y el nivel de escolaridad de sus madres. Una vez obtenido los datos como estamos conscientes que los hallazgos en la muestra pueden ser aleatorios necesitamos de algún procedimiento que estime la verosimilitud de los resultados obtenidos en la muestra y éste es precisamente un test de hipótesis estadística.

Los tests de hipótesis consisten en confrontar dos hipótesis, una llamada ***hipótesis nula*** que denotamos con $\mathbf{H_0}$ y otra llamada ***hipótesis alternativa*** denotada con $\mathbf{H_1}$. En el ejemplo las hipótesis que se plantean son:

1. "el estado nutricional de los niños está asociado al nivel de escolaridad de las madres" (hipótesis de trabajo)
2. No existe tal asociación

Cabe preguntarse ahora, cuál de ellas se debe considerar como hipótesis nula. En la mayoría de los tests que se usan en investigación médica la hipótesis nula se elige a aquella que se quiere rechazar, es decir, en este caso H_o: "No hay asociación".

La hipótesis nula generalmente es una hipótesis de igualdad, por lo que admite sólo una posibilidad; mientras la hipótesis alternativa admite varias posibilidades. Para aclarar esto veamos un ejemplo: Supongamos que queremos probar si la proporción (p) de desnutridos en una población infantil es igual o no al 20%. La $\mathbf{H_0}$ es que la proporción de desnutridos es igual al 20% (p=0,20); y la $\mathbf{H_1}$ admite tres posibilidades, de acuerdo a los datos muestrales, que la proporción de desnutridos sea menor al 20% (p<0,20), que sea mayor al 20% (p>0,20), o que la proporción sea distinta al 20% (p#0,20), se debe determinar como $\mathbf{H_1}$ una de estas tres posibilidades. En símbolos:

$$\mathbf{H_0}: p=0{,}20$$

$$\left.\begin{array}{l} \mathbf{H_1}: p<0{,}20 \\ \mathbf{H_1}: p>0{,}20 \\ \mathbf{H_1}: p\#0{,}20 \end{array}\right\} \quad \text{Seleccionar una de estas alternativas}$$

Bajo este planteo un test de hipótesis estadística no es otra cosa que ***un procedimiento para tomar una decisión, bajo incertidumbre****,* sobre la validez de la hipótesis nula usando la evidencia de los datos. Puesto que se trabaja bajo incertidumbre es claro que cualquiera sea la decisión que se tome siempre existe una probabilidad de cometer error. A fin de clarificar esto podemos presentar el siguiente esquema:

Situaciones posibles al probar una hipótesis estadística

Decisión	**Realidad sobre H_o**	
	Falsa	**Cierta**
Rechazar H_o	Decisión correcta	**Error Tipo I**
No rechazar H_o	**Error Tipo II**	Decisión correcta

Como se pueda ver en el esquema, con cada tipo de decisión que se tome hay asociado una posibilidad de cometer un error. Un procedimiento de este tipo sería óptimo cuando las probabilidades de cometer un error, cualquiera sea la decisión que se adopte, sean pequeñas. Lamentablemente, en la mayoría de los test de hipótesis sólo es posible controlar una de ellas, con el agravante de que estos errores son competitivos, es decir, cuando se disminuye mucho la probabilidad de uno aumenta la probabilidad del otro.

Puesto que, el interés generalmente es "rechazar H_o" la probabilidad de error que se controla durante este procedimiento, es justamente el error asociado a esta decisión (Probabilidad del Error Tipo I), es decir, ***la probabilidad de rechazar H_o cuando es cierta***. La máxima probabilidad de error tipo I se denota con α y recibe el nombre de ***nivel de significación del test*** y él debe ser prefijado de antemano, generalmente varía entre el 1% y el 5% (α= 0,01 o α = 0,05). La probabilidad de Error Tipo II se denota con **β.**

La bondad de un test de nivel α se mide en términos de la cantidad 1-β denominada ***Poder del Test.***

El nivel de significación que se usa generalmente es α=0.05 lo que corresponde a un 5% en término de porcentaje.

Tipos de pruebas de hipótesis

El tipo de prueba de hipótesis depende de la alternativa que se seleccione.

Prueba de una cola: prueba de cualquier hipótesis estadística donde la alternativa es unilateral

$H_0 : p=p_0$

$H_1: p>p_0$ *(cola derecha)* ó $H_1: p<p_0$ (*cola izquierda*)

Prueba de dos colas: prueba de cualquier hipótesis estadística donde la alternativa es bilateral

$H_0 : p=p_0$

$H_1: p\#p_0$

Procedimiento general de un test de hipótesis

El procedimiento de un test de hipótesis lo podemos resumir en los siguientes pasos:

1. Establecer la hipótesis nula. Se supone que H_o es cierta.

2. Seleccionar la H_1 apropiada.

3. Seleccionar el nivel de significación α

4. Para confrontar esta suposición con la información (parcial) que proveen los datos sobre la realidad de H_o, se forma "una especie de indicador" de concordancia, denominado ***estadístico del test.***

5. Como el estadístico depende de la información de los datos, con cada muestra posible hay asociado un valor de este estadístico y en consecuencia se genera una nueva variable aleatoria. Asociada a esta variable hay una cierta distribución de probabilidad, a partir de la cual se determina la probabilidad de que la información de los datos concuerde con la hipótesis nula, denominado ***"P-value"***. De esta manera, el "P-value" representaría la probabilidad de cometer un error cuando se toma la decisión de rechazar H_o.

6. Decidir: es claro que, si de antemano se fijó que la máxima probabilidad de error al rechazar H_o debía ser igual a α, para tomar la decisión es necesario comparar el valor de P con α. Así

   Si $P \leq \alpha$ *entonces la decisión es* **Rechazamos H_o**

   Si $P > \alpha$ *la decisión es* ***No hay evidencia suficiente para rechazar H_o***

Prueba sobre una media

Para ilustrar el procedimiento a seguir para realizar un test de hipótesis sobre una media, presentaremos el siguiente ejemplo: Se conoce que el promedio días de estada de los pacientes de un hospital es de 8 días; si en una muestra aleatoria de 30 pacientes incluidos en un estudio indica que el promedio días de estada es de 6días con un desvío estándar de 2,31 días; se puede pensar a un nivel de significancia del 5% que el promedio días de estada es menor a 8 días.

Siguiendo el procedimiento se prueba la hipótesis nula

1. $H_0 : \mu$=8días
2. $H_1 : \mu$< 8días
3. α= 0,05
4. El estadístico de prueba en este caso es el test "t" para una media

5. Del cálculo surge que P= 0,00026
6. Se decide rechazar H_0 ya que P(0,00026)es menor que α (0,05). Se concluye que el promedio días de estada de los pacientes en el hospital es menor a 8 días.

Prueba sobre una proporción

También para ilustrar la aplicación del procedimiento para realizar un test de hipótesis sobre una proporción (p), presentaremos el siguiente ejemplo: Se cree que un medicamento es efectivo en un 60% (p=0,6) de los casos. De un nuevo medicamento administrado en una muestra aleatoria de 100 adultos, 70 mejoraron (p=0,7); al 5% de significancia se puede afirmar que el nuevo medicamento es mejor.

Siguiendo el procedimiento se prueba la hipótesis nula

1. H_0 : p=0,6
2. H_1 : p>0,6
3. α= 0,05
4. El estadístico de prueba en este caso es el test "Z" para una proporción
5. Del cálculo surge que P= 0,0248
6. Se decide rechazar H_0 ya que P(0,0248)es menor que α (0,05). Se concluye que el nuevo medicamento es mejor

Si en este ejemplo se hubiese fijado α= 0,01 (1%) al comparar P (0,0248) con α (0,01) la decisión sería: **no hay evidencia estadística suficiente para rechazar H_0,** por lo tanto no se puede afirmar que el nuevo medicamento sea mejor.

Prueba de Independencia o Test Chi Cuadrado (χ2)

Esta prueba se utiliza para probar la hipótesis de independencia entre dos variables. Se deben presentar las frecuencias observadas (datos de la muestra) en una tabla de contingencia (o tabla de clasificación cruzada) con sus respectivas frecuencias conjuntas (celdas) y frecuencias marginales (totales).

Ejemplo: Supongamos que queremos conocer si existe asociación entre las variables estado nutricional de los niños y el nivel de instrucción de la madre; para analizar tal asociación se toma una muestra aleatoria de 1000 niños y se obtienen los siguiente datos (frecuencias observadas) que presentamos en una tabla de contingencia:

Tabla N° 1: Estado Nutricional de los niños según nivel de instrucción de la madre.

Frecuencias observadas (o_i)

Estado Nutricional	Nivel de Instrucción de la Madre			Total
	Primario	Secundario	Terciario	
Eutrófico	182	213	203	598
Desnutrido	154	138	110	402
Total	336	351	313	1000

Las Hipótesis nula y alternativas serán:

H₀: *las variables son independientes*

H₁: *existe asociación entre las variables*

Se calculan las frecuencias esperadas (e_i) en caso de independencia:

$$e_i = \frac{(\text{Total de columna}) x (\text{Total de renglón})}{\text{Gran Total}}$$

Frecuencias esperadas (e_i)

Estado Nutricional	Nivel de Instrucción de la Madre			Total
	Primario	Secundario	Terciario	
Eutrófico	201	210	187	598
Desnutrido	135	141	126	402
Total	336	351	313	1000

Se debe calcular el estadístico de prueba, en este caso corresponde el χ^2 (*CHI CUADRADO)*. Este estadístico de prueba se calcula sumando las diferencias de las frecuencias observadas y esperadas al cuadrado divididas en las frecuencias esperadas,

$$\chi^2_{obs} = \sum (o_i - e_i)^2 / e_i$$

a este estadístico se le calculan los grados de libertad (*gl*) que son igual al número de categorías de la variable (renglones) menos 1 por el número de categorías de la otra variable (columnas) menos 1,

$gl = (r-1)(c-1)$ Se debe fijar $\alpha = 0{,}05$

Luego se compara el χ^2_{obs} con el χ^2_α con *(r-1)(c-1) grados de libertad* (este valor se lo busca en la tabla de distribución del χ^2_α). Si el χ^2_{obs} nos da un valor mayor que el $\chi^2_{\alpha(r-1)(c-1)}$ la decisión es **rechazar H_0** y concluir que existe asociación entre las variables; pero si χ^2_{obs} es menor que el χ^2_α **no se rechaza H_0** y se concluye que las variables son independientes, todo esto a un nivel de significación α.

Si calculamos $\chi^2_{obs} = 7{,}88$; y si α=0,05 el $\chi^2_{0,05;2} = 5{,}991$; al comparar estos dos valores tomamos la decisión de ***rechazar H_0*** y *se concluye que el estado nutricional del niño y el nivel de instrucción de la madre están asociados.* Al calcular la prueba a través de un programa estadístico este nos da un *p-value, P=0,0195;* al comparar este valor de *P con α=0,05 la decisión es rechazar H_0 ; pero si α= 0,01 la decisión es no hay evidencia estadística suficiente para rechazar H_0.*

ESTUDIOS ANALITICOS

RELACION ENTRE DOS VARIABLES MEDIDAS EN EL MISMO INDIVIDUO

Evaluar la asociación entre dos variables

VARIABLES	MEDIDAS SUGERIDAS (ANALISIS SUGERIDO)
DOS CONTINUAS (n grande)	*coef. de correlacion de Pearson (r)*
DOS CONTINUAS (n mediano o chico)	**coef. correlacion de Spearman (r_s)** **coef. correlacion de Kendall (τ)**
CONTINUA vs ORDINAL	**coef. correlacion de Kendall (τ)** (TEST DE TENDENCIA DE CUZICK)
CONTINUA vs **NOMINAL**	(ANOVA PARAMETRICO) (NO PARAMETRICO)
CONTINUA vs DICOTOMICA	(TEST T O TEST Z) (TEST MANN-WHITNEY)
DOS ORDINALES	**coef. correlacion de Kendall (τ)** (TEST DE TENDENCIA DE CUZICK) (TEST χ^2 PARA TABLA cxr)
ORDINAL Vs NOMINAL	(TEST χ^2 PARA TABLA cxr)
ORDINAL Vs DICOTOMICA	(χ^2 DE TENDENCIA LINEAL PARA TABLA 2xk)
DOS NOMINALES	(TEST χ^2 PARA TABLA cxr)
NONINAL vs DICOTOMICA	(χ^2 SIN TENDENCIA EN TABLA 2xk)
DOS DICOTOMICAS	RIESGO RELATIVO – ODDS RATIO (χ^2 PARA TABLA 2x2 TEST EXACTO DE FISHER)

COMPARACIÓN DE DOS GRUPOS INDEPENDIENTES

Evaluar influencia de condiciones (Factores o tratamientos) en problemas de salud

Ejemplo: Peso al nacer de niños de un grupo de madres con CPN comparado con aquellos de un grupo de madres sin CPN

TIPO DE DATOS	ANALISIS SUGERIDO
CONTINUOS	**TEST DE HIPOTESIS** **(para grupos independientes)**
ORDINAL	**χ^2 DE TENDENCIA LINEAL EN TABLA 2xk** **(más del 80% de las frec. esperadas deben ser > que 5)**
NOMINAL	**χ^2 SIN TENDENCIA PARA TABLA 2xk** **(más del 80% de las frec. esperadas deben ser > que 5)**
DICOTÓMICOS	**χ^2 PARA TABLA 2x2** **(más del 80% de las frec. esperadas deben ser > que 5)** **EXACTO DE FISHER**

COMPARACIÓN DE LA RESPUESTA DE "UN GRUPO" BAJO DIFERENTES CONDICIONES

Evaluar la respuesta a intervenciones (PLS o Tratamientos) sobre los problemas de salud
Ejemplo: Presión arterial antes y después del tratamiento

TIPO DE DATOS	ANALISIS SUGERIDO
CONTINUOS	**TEST DE HIPOTESIS** **(para grupos pareados)**
ORDINAL O NOMINAL	TEST DE SIGNO
DICOTÓMICOS	**TEST McNEMAR** TEST LIDDELL

BIBLIOGRAFÍA

1. Batellino, Luis y Susana Doronsoro. METODOLOGÍA DE LA INVESTIGACIÓN EN SALUD. Universidad Nacional de Córdoba, Facultad de Odontología. Córdoba. 1994.
2. Dawson-Sanders and Trapp, RG. BIOESTADÍSTICA MÉDICA. Ed. Manual Moderno. 1993
3. Hernández-Ávila, M; Garrido-Latore, F. y López-Moreno, S. DISEÑO DE ESTUDIOS EPIDEMIOLÓGICOS. Salud Pública de México. Vol. 42, Nº 2. Marzo-Abril 2000.
4. Norman y Striner. BIOESTADÍSTICA. Harcourt-Brace. España. 1998.
5. Pita-Fernández, S. TIPOS DE ESTUDIOS CLÍNICOS EPIDEMIOLÓGICOS. Madrid. 2001.
6. Kleinbaum, D.G; Kupper, L.L; Morgenstern H. EPIDEMIOLOGIC RESEARCH. PRINCIPLES AND CUANTITATIVE METHODS. Van Nostrand Reinhold Company. 1982
7. Kish, Leslie. MUESTREO DE ENCUESTAS. Editorial Trillas. México. 1979
8. Santana, Mirta. APUNTES DE ESTADÍSTICA INFERENCIAL. Cátedra Bioestadística. Facultad de Medicina. U.N.T. 2000
9. Santana, M; D'Urso,M y Lencina, V. BIOESTADÍSTICA I. Facultad Medicina. UNT. Tucumán. 2004.
10. Walpole y Myers. PROBABILIDAD Y ESTADÍSTICA .4° Edición. McGraw-Hill/Interamericana de México. México. 1991.