Escuela Nacional de la Salud Pública

Análisis Estadístico

Dra. María Elena Astraín Rodríguez MSc

AÑO 2001

LA ESTADÍSTICA COMO MÉTODO

El **método estadístico** ofrece una doble vertiente. Es por un lado es un proceso de pensamiento asistido por todo el rigor de la ciencia matemática, con un enfoque peculiar de la mentalidad científica ante la investigación y por otro lado es la herramienta que nos permitirá obtener conclusiones y presentar resultados con verdadera conocimiento de la confianza de nuestras afirmaciones.
El desconocimiento de las posibilidades del Método Estadístico hace que se recurra a la estadística al terminar la investigación, cuando se tiene una gran cantidad de datos y se pretende analizar las consecuencias que de todo ello se pueda deducir. El método estadístico debe estar presente desde que se inicia la etapa de planificación de la investigación. De esta forma se evitaran graves errores, costos y esfuerzos innecesarios.
Dentro de la **etapa de planificación del método científico** el método estadístico está presente en:

• Definición de los objetivos
•Definición del universo y la muestra
• Definición de las unidades de observación
•Determinación de la información necesaria y de la fuente de obtención. En el diseño de instrumentos de recolección
•Definición de unidades de medidas de escalas de medición y de clasificación
•Elaboración del olan de tabulación y análisis
•Organización de la investigación

En la **etapa de ejecución** estará presente en**:**

•Recolección de la información
•Revisión y clasificación
•Recuento y presentación de datos
•Cálculo de medidas de resumen y análisis estadístico de los resultados
•Inferencia estadística de conclusiones respecto a la hipótesis

El método estadístico nos permite valorar la influencia de la variabilidad biológica y social, ya que no basta con el “sentido común” para obtener las mejores conclusiones de los trabajos científicos

ESTADÍSTICA DESCRIPTIVA

Conceptos básicos

Debemos distinguir dos acepciones diferentes de la estadística. Estadísticas en plural se refiere a datos numéricos, por ejemplo, cuando hablamos de estadísticas de mortalidad estamos haciendo alusión a datos de mortalidad, estadísticas de fecundidad, datos de fecundidad, sin embargo, estadística en singular es para algunos la ciencia ,para otros el método que nos brinda las

herramientas necesarias para la recolección**,** procesamiento y análisis de la información.

La Estadística se divide en dos grandes ramas: la estadística descriptiva y la estadística inferencial.

La estadística descriptiva nos brinda las herramientas necesarias para caracterizar a un grupo de elementos o unidades de análisis pero sin pretender generalizar esos resultados a un grupo mayor, objetivo que sí persigue, como veremos más adelante la estadística inferencial.

Otro aspecto que debemos tener presente es que vamos a entender por **población** o **universo.**

Población es un conjunto de elementos, objetos o unidades de análisis que al menos comparten una característica que nos interesa conocer o estudiar. Por ejemplo, los niños residentes en Santa Cruz que asistieron a la escuela durante el año 1998. Esta población tiene 3 características que la define, todos los elementos que la integran son niños, todos residen en la misma localidad, Santa Cruz, y todos tienen que haber asistido a la escuela durante el año 1997. Otro ejemplo sería las historias clínicas de un hospital para un periodo de tiempo dado.

Es muy importante en una investigación, definir como parte del método, la población objeto de estudio de manera que quede claro que elementos la integran.

Muestra no es mas que una parte cualquiera de la población, un conjunto cualesquiera no vacío, o sea, que teóricamente un solo elemento de la población puede ser considerado como una muestra, aunque en la práctica esto ocurra con poca frecuencia.

Los conceptos de muestra y población son conceptos relativos, lo que en un estudio puede ser una población en otro puede ser una muestra. Los niños de Santa Cruz que asistieron a la escuela en el año 1997 pueden ser parte de una muestra que se halla diseñado para un estudio nacional.

La selección de la muestra puede realizarse aplicando técnicas estadísticas y obtener una muestra probabilística, o simplemente definir la muestra por criterios de expertos.

Un concepto trascendental para la estadística, pues del se parte para definir métodos de representación a utilizar, medidas de resumen a calcular y que técnicas de análisis se pueden aplicar es el de **variable**.

Consideraremos como variable cualquier característica de la población que puede asumir diferentes comportamientos, valores, o grados de intensidad entre

los diferentes elementos individuos o unidades de análisis que la conforman. Por ejemplo la edad es una característica que asume diferentes valores de un individuo a otro, el sexo, el estado civil, los servicios de un hospital son todos variables.

Las variables pueden ser **cualitativas** o **cuantitativas**. Las variables cualitativas son aquellas en las que las diferencias entre un elemento y otro son atributos, cualidades, no medibles en términos numéricos. Ej. el sexo, la nacionalidad, las especialidades médicas con que cuenta un centro asistencial, las causas de defunción entre otras.

Las variables cuantitativas son aquellas en las que las diferencias existentes entre los diferentes elementos de la población son medibles numéricamente. Ej. la edad, la talla, el peso, el número de camas de un hospital.

Las variables cualitativas a su vez se clasifican en **nominales** u **ordinales** o **cuasicuantitativas.**

Una variable se considera cualitativa nominal cundo las diferencias entre los elementos son cualidades, atributos que no sólo no son medibles numéricamente, sino, que tampoco traducen diferencias de magnitudes o de intensidad. Ej. el estado civil. Las variables cualitativas ordinales son aquellas en que las diferencias si bien no son cuantificables, si tienen implícito diferencias de magnitud o de intensidad y dan una idea de ordenamiento. Ej. el estado de un paciente: grave, de cuidado, mejorado, el estado técnico de una vivienda: buena, regular, mala.

A su vez las variables cuantitativas también se subdividen en dos subgrupos: **discretas** o **discontinuas** y **continuas**.

Las variables discretas son aquellas que sólo asumen valores enteros, Ej. el número de hijos. Una pareja podrá tener 5 hijos, 10 hijos o ningún hijo, pero lo que no podrá tener nunca es 1.6 hijos.

Sin embargo, las variables cuantitativas continuas sí pueden asumir valores fraccionarios. Ej. el peso, la talla, la edad.

Una vez identificadas las variables que formaran parte del estudio, el siguiente paso que se impone es definir las **escalas de clasificación** para esas variables.

Una escala de clasificación no es mas que el conjunto de **clases o categorías** que se definen para clasificar la información en función de determinada variable. En el caso de las variables cualitativas la propia naturaleza de la variable establece las posibles categorías de la escala. Por Ej. la variable estado civil puede asumir 5 categorías en su escala: soltera, viuda, divorciada, acompañada y casada. Puedo utilizar esas 5 o hacer reagrupaciones y definir solo dos categorías en función de los intereses que persiga con mi estudio: uniones

estables (casada y acompañada) y uniones inestables(solteras divorciadas y viudas).Sin embargo, en el caso de las variables cuantitativas pueden existir diferentes y diversas escalas.

Sea cual sea la escala que se construya para la clasificación de la información debe cumplir con dos requisitos o condiciones para que sea una buena escala.

En primer lugar que sea **exhaustiva**, que el número de clases o categorías que la constituyen garanticen que todos los elementos que integran la población puedan ser clasificados.

La segunda condición que deben cumplir es que las categorías sean **mutuamente excluyentes**, de forma tal que no quede duda de en que clase debe ser incluido cada uno de los elementos que conforman la muestra o la población.

A modo de ejemplo:

Supongamos que se quiere hacer un estudio de fecundidad y una de las variables a estudiar es la edad a la que se produjo el nacimiento.

Edad.
(años)

20 -25

25 - 29

30 - 34

35 - 39

40 - 44

Si se producen nacimientos en menores de 20 años, algo bastante probable, no tendrían una categoría donde ser ubicados por tanto no se cumpliría el principio de la exhaustividad, por otra parte si nace un niño que la madre tiene 25 años no sabríamos en que categoría incluirlo, si en la primera o en la segunda por tanto las clases de la escala no son mutuamente excluyentes.

Las escalas, al igual que las variables, pueden ser cualitativas nominales u ordinales y cuantitativas discretas o continuas.

Para el caso particular de las variables cuantitativas es necesario identificar algunos elementos particulares.

Las clases o categorías que conforman las escalas cuantitativas se denominan intervalos de clase, en el ejemplo anterior de la edad de las madres, esa escala está formada por 5 intervalos de clase.

El menor valor y el mayor valor que delimitan los intervalos de clase se denominan **límite de clase inferior** y **límite de clase superior.**

Así, en el mismo ejemplo 20,25,30,35 y 40, serian los límites inferiores, y 24 (subsanando el error), 29 34 39 y 44, serían los límites superiores.

Los **límites reales** se obtienen de la semisuma de los límites de clase o de notación.
L.R.I. = L.C.S + L.C.I. \ 2.

Retomando el ejemplo:	Límites Reales.
24 + 25 / 2 = 24.5	---------- 24.5
29 + 30 / 2 = 29.5	24.5 ----------29.5
34 + 35 / 2 = 34.5	29.5 ---------- 34.5
39 + 40 / 2 = 39.5	34.5 ----------39.5
	39.5 ----------

Nos faltaría por calcular el límite real inferior del primer intervalo de clase y el límite real superior del último. Para ello suponemos, según el nivel de aproximación conque se esté trabajando la escala, cual sería el límite superior del intervalo de clase que antecediera al primero, en este caso concreto 19 años, y cual sería el límite inferior del intervalo de clase que seguiría después del último, para este ejemplo, 45 años, y con estos valores calculamos los límites que faltan:

19 + 20 / 2 = 19.5
44 + 45 / 2 = 44.5

Fíjese, que con los límites reales no se puede clasificar la información, pues no se cumple con el requisito de que los intervalos de clase sean mutuamente excluyente. Estos se usan con otros fines, uno de ellos lo veremos de inmediato.

Amplitud de intervalo de clase. Es el alcance, cuantas unidades abarca el intervalo de clase. Se obtiene por la diferencia de los límites reales superior e inferior del intervalo de clase.

AMP = L.R.S. - L.R.I.

Ejemplo.

Para el primer intervalo de clase:

19.5 -24.5 = 5

La amplitud es de 5, y compruebe que en esta escala todos los intervalos tiene la misma amplitud.
Marca de clase. Es el punto medio del intervalo de clase, el valor central, se obtiene por la semisuma de los límites de clase superior e inferior del intervalo.

M.C. = L.S + L.I / 2.

Ejemplo.

Para el primer intervalo de clase:

20 + 24 / 2 = 22

La marca de clase del primer intervalo es 22 años. Le sugerimos que calcule las restantes.
En ocasiones para cumplir con el criterio de exhaustividad se hace necesario agregar un intervalo de clase al principio o al final de la escala, porque no tenemos claro que límite inferior asumir en el caso del primer intervalo o que límite superior en el caso del último. Siguiendo con el ejemplo de la edad de las madres, se puede asumir que deben ocurrir nacimientos por debajo de los 20 años, pero cuán joven puede ser la madre. En este caso que no tenemos bien definido que limite inferior asumir, simplemente lo omitimos y creamos un intervalo de clase **abierto**: 19 años o menos, o menor de 20 años. Lo mismo pudiera ocurrir con el último intervalo, y sería, 45 años o más.

Cuando una escala de clasificación tiene el primero, el último o ambos intervalos de clase abierto se dice que la **escala es abierta**.

Una vez definida la variable y su escala de clasificación, el siguiente paso sería clasificar la información, obtener el número de elementos que corresponden a cada una de las clases de la escala, la **frecuencia absoluta** de cada clase.

Cuando tenemos una escala de clasificación y la frecuencia absoluta calculada estamos en presencia de una **distribución de frecuencia** para esa variable.

Edad (años)	Frecuencia absoluta	Frecuencia relativa Proporción	%	Frecuencias acumuladas Absoluta	Relativa
20–24	20	0.29	29	20	29
25–29	18	0.27	27	38	56
30–34	15	0.22	22	53	78
35–39	10	0.15	15	63	93
40–44	5	0.07	7	68	100
Total	68	1.00	100		

Además de la frecuencia absoluta también se puede obtener la **frecuencia relativa**. Esta se obtiene dividiendo la frecuencia absoluta de cada clase entre el total de observaciones, el valor que se obtiene se conoce por el nombre de **proporción** y si este resultado se multiplica por cien (x 100) estaremos en presencia del **por ciento o porcentaje**, que es una medida de la contribución que hace cada clase, la importancia relativa con relación al total de observaciones, al total de la población o al tamaño de la muestra.

Observe que la suma de todas las frecuencias absolutas es igual al total de observaciones, y la suma de todas las frecuencias relativas es igual a 1o a 100 en dependencia de si se trabaja con proporciones o porcentajes.

También pueden calcularse las **frecuencias acumuladas** absolutas o relativas. Para ello se suman las frecuencias de todas las clases que anteceden a la clase para la cual se esta realizando el cálculo. La frecuencia acumulada del primer intervalo de clase coincide con la absoluta o relativa del mismo, pues como es obvio nadie la antecede y la del último intervalo es el total de observaciones 1 o 100.

Veamos un ejemplo. La frecuencia absoluta acumulada del tercer intervalo de clase sería: 20 + 18 + 15 = 53 y la relativa en forma de por cientos: 29 + 27 + 22 = 78 %. Compruebe el resto de los cálculos con el ejemplo de la tabla anterior.

Las frecuencias absolutas y relativas pueden calcularse tanto para variables cualitativas como para las cuantitativas, sin embargo como es fácil de apreciar las frecuencias acumuladas sólo pueden calcularse para las variables cuantitativas.

Ejercicios resueltos

1.- Los datos que aparecen a continuación corresponden a las variables que se estudiaron en un grupo de trabajadores de la Industria Azucarera para determinar la frecuencia de accidentes del trabajo y enfermedades profesionales.

Variables y datos que se pidieron:

a. Edad.
b. Sexo.
c. Nivel de Escolaridad.
d. Talla.
e. Estado Civil.
f. Número de hijos.
g. Años de trabajo en la ocupación actual.
h. Número de accidentes sufridos.
i. Existencia de enfermedad profesional.
Clasifique cada una de las variables.

Respuestas:
Cualitativa nominal b,e,i.
Cualitativa ordinal c.
cuantitativa discreta f,h.
cuantitativa continua a,d,g.

2- En un estudio realizado sobre accidentes de trabajo en una industria sideromecánica, se dispone de datos para 300 obreros de los 5000 con que cuenta la industria.
El interés de los autores era observar las diferencias en el número de accidentes en trabajadores de diferentes ocupaciones dentro de la industria, según tipo de adiestramiento previo y años de trabajo en esa ocupación.

a. Identifique la población en estudio y cuál es la muestra en caso de existir.
b. identifique las variables en estudio y clasifíquelas.

Respuesta.
a. población, los 5000 obreros de la industria.
muestra, los 300 obreros estudiados.
b. Número de accidentes, cuantitativa discreta.
Tipos de ocupaciones, cualitativa nominal.
Tipo de adiestramiento, cualitativa nominal.
Años de trabajo, cuantitativa continua.

3. En la siguiente tabla aparecen los casos vistos por enfermedades diarreicas agudas (EDAD) por grupos de edad para un centro de salud en un mes.

Edad	No
1–4	154
5–9	81
10–14	66
15–24	116
25–59	310
60–64	30
Total	757

a. Cuál es la amplitud de cada intervalo de clase?
b. Determine la marca de clase.
c. Construya una distribución de frecuencias relativas y de frecuencias acumuladas absolutas y relativas.

Respuesta.

a. Del primer intervalo la amplitud es de 4, los dos siguientes y el último tienen una amplitud es de 5
El intervalo, de 15-24, tiene una amplitud de 10.
El de 25-59 ,tiene una amplitud de 35.
b. Las marcas de clase son: 2.5, 7, 12, 19.5, 42 y 62.
c.

Edad	No	%	Acumuladas	
			No	%
1–4	154	20.3	154	20.3
5–9	81	10.7	235	31.0
10–14	66	8.7	301	39.7
15–24	116	15.3	417	55.0
25–29	310	40.9	727	95.9
60–64	30	4.0	757	100.0

PRESENTACIÓN ESTADÍSTICA

Existen dos tipos de presentación la presentación tabular y la gráfica.
Presentación tabular.
La finalidad de los cuadros es presentar de forma resumida e inteligible determinado material numérico.
Las tablas cumplen generalmente uno de los siguiente propósitos:

- presentar una o más distribuciones de frecuencias tanto de variables cualitativas como cuantitativas.
- Presentar índices asociados a determinadas variables
- Presentar una serie cronológica

Las tablas pueden perseguir más de uno de estos propósitos ya que estos no son excluyentes.
Partes de un cuadro:

- Identificación de la tabla. Consiste en darle un número consecutivo según el orden en que aparezca en el texto del trabajo: "cuadro1, cuadro2..." y a la hora de hacer referencia al mismo se hace por el número asignado, " como puede apreciarse en el cuadro1....
- Titulo,. Es la parte más importante y difícil de diseñar correctamente. El cuadro o tabla estadística debe ser auto explicativo, debe entenderse sin necesidad de acudir la texto, si esto no se cumple quedo mal diseñado. La mejor manera de saber si una tabla tiene errores en su confección es dársela a leer a un colega, si comienza a hacernos preguntas por que no es capaz de entenderla algo no ha quedado claro o bien diseñado.

Es por eso que el titulo juega un papel tan importante, el nos brinda la mayor cantidad de información sobre el material numérico que aparece en el cuerpo de la tabla.

Para que un titulo este bien elaborado debe cumplir con dos requisitos básicos:

Que sea **completo.** Para que cumpla con este primer requisito debe responder las siguientes preguntas:

¿Qué?, se refiere a los elementos que conforman la muestra o la población. ¿ Qué se estudió?, defunciones o nacimientos o niños o camas de un hospital...

¿Cómo?, se estudio esa información. Se refiere a las variables por las que se clasificó la misma. ¿Cómo se estudio...? por sexo, por sexo y edad, por peso al nacer....

¿Dónde? se realizó el estudio, a que "espacio" pertenecen esos datos, Ciudad de la Habana, Cuba, Hospital x, Municipio...

¿Cuándo?, se refiere a la fecha o periodo de tiempo : año 2000, primer semestre del año, mes de enero, de 1990 a 2000...

Si nos hacemos esas preguntas en ese mismo orden y las vamos contestando tendremos elaborado un titulo completo.

El segundo requisito que debe cumplir el titulo es que sea **conciso,** que sea **breve**, pero nunca debe dejar de ser completo en aras de lograr una mayor concisión.

. Ejemplo

Titulo 1: Distribución de las defunciones de Cuba ocurridas durante el año 1999 clasificadas según sexo y edad de los fallecidos.

Veamos primero si el titulo es completo:

¿Qué se estudia?, defunciones

¿Cómo?, por edad y sexo

¿Dónde?, Cuba

¿Cuándo?, 1999

Obviamente el titulo es completo, responde las 4 preguntas pero es innecesariamente largo, si nos limitamos a dar respuesta a las preguntas quedará además de completo conciso que era la segunda característica de un buen titulo:

Defunciones por edad y sexo. Cuba 1999.(correcto)

Cuerpo de la tabla. Está formado por la unión de filas y columnas, las filas en sentido horizontal y las columnas en sentido vertical .

La primera columna se denomina columna matriz y en ella se coloca la variable con su escala. En la primera fila, denominada de los encabezamientos se rotulan los aspectos que denominan los números que aparecen el resto del cuerpo, nombre de los índices, número, % etc.

La última fila y la última columna son las de los totales

Nombre de la variable	FILA DE LOS ENCABEZAMIENTOS		TOTAL
E			
S			
C			
A			
L			
A			
TOTAL			

Notas aclaratorias

Fuente

Notas aclaratorias. Se acostumbra a colacarla al pie de la tablas y tiene la finalidad de esclarecer algunos aspectos importantes para la comprensión de la misma. Puede ser las unidades de medidas de algunas variables o el 10^n con que se trabajo alguna tasa, por citar algunos ejemplos.

Fuente. Se refiere al lugar de donde se obtuvo la información, casi siempre hace alusión a una fuente secundaria, un sistema de información, una publicación, un documento oficial etc. El declarar la fuente de donde se obtuvieron los datos obedece a varios motivos, en primer lugar el lector debe saber cuan confiable es esa información, además puede acceder a la misma si así lo quisiera, y también es un reconocimiento a las personas o instituciones que obtuvieron esa información.

En ocasiones la fuente de donde se obtuvo la información es una fuente primaria, digamos una encuesta, si toda las tablas tienen como única fuente esa encuesta, fruto del trabajo del propio investigador, y quedo bien declarada en el método, es innecesario poner en todas las tablas "fuete: encuesta" como suele ocurrir con mucha frecuencia, es redundante, monótono, y le resta elegancia al trabajo.

Errores más frecuentes al diseñar una tabla

- Título o encabezamiento incorrecto o inadecuado.
- Disposición incorrecta de los datos.
- Errores de cálculo.
- Tablas que muestran solo %
- Cuadros sobrecargados
- No poner la fuente en casos necesarios, o ponerla cuando los datos son originales

Algoritmo para el análisis de una tabla

Al escribir el informe final de una investigación, un articulo científico o simplemente un informe de trabajo donde aparezca determinado material numérico presentado en un cuadro, este debe ser analizado por el autor y escribir en el texto sólo los aspectos más relevantes, los juicios de valor que se desprenden de ese análisis, las conclusiones a las que se arriba después de realizado el mismo.

Sin embargo, en un número no despreciable de casos nos encontramos que se hace una descripción detallada de la tabla, se describe celda por celda todos los valores que aparecen en las mismas, como nos gusta decir a modo de ilustración, " se hace un retrato escrito de la tabla" y no se da ni un solo juicio de valor.

Para evitar este mal habito le sugerimos que siga estos pasos para el análisis de la tabla y una vez concluido el mismo escriba en el texto las conclusiones a las que arribó, eso es lo que se espera de usted como autor, los datos numéricos los tiene el cuerpo de la tabla, no hay que repetirlos en forma de texto.

Algoritmo para el análisis de una tabla:

- Leer cuidadosamente el título.

- Leer las notas aclaratorias
- Precisar unidades de medidas
- Fijarse en los totales
- Relacionar los totales con los valores particulares de las categorías de las variables
- Relacionar entre sí los valores de las variables estudiadas

Presentación gráfica.

El gráfico es la representación de los datos numéricos en el plano, con el fin de obtener una impresión visual de conjunto del material presentado, que facilite su rápida comprensión y que permita evidenciar de forma más objetiva que la tabla, la evolución de un fenómeno o hacer más evidentes ciertas relaciones poco aparente en el material tabulado.
El gráfico es un auxiliar de la tabla estadística, no la sustituye, si no, la complementa, además no es necesario tener tantos gráficos como tablas aparecen en el informe final de un trabajo, por el contrario, le restan lucidez al trabajo lejos de proporcionar calidad u rigor científico.

Características generales de los gráficos

- Generalmente se inscriben en los llamados ejes de coordenadas cartesianos o rectangulares.
- En la ordenada(y) se representan las frecuencias o índices
- En la abcisa (x) se representan las clases de la escala de clasificación
- Los ejes deben ser de igual magnitud, solo las abcisas pueden ser 1.5 o 2 veces mayor que las ordenadas. Este especto es muy importante porque de no tenerse en cuenta pueden ocurrir las llamas falacias o distorsión visual del fenómeno que se analiza. Si el tamaño del eje de las y es más del doble dará la impresión de cambios marcados en el comportamiento del fenómeno, si por el contrario es menor que el eje de las x parecerá que apenas hay modificaciones en el comportamiento de las variables estudiadas
- De ser posibles las escalas deben comenzar en cero. En caso de valores muy altos y con pocas oscilaciones puede “partirse la escala”
- Los ejes de las ordenadas y de las abcisas deben estar rotulados
- Las escalas no deben tener muchas subdivisiones, deben usarse siempre que sea posible números redondos. El detalle se encuentra en la tabla, rl gráfico da la idea general.

Partes de un gráfico

- Identificación. Al igual que en la tabla su objetivo es numerarlos, darles un número consecutivo si es más de uno el que aparece en el trabajo.
- Título. Debe cumplir los mismos requisitos que el de las tablas.
- Gráfico propiamente dicho. Su forma y características dependen del tipo de variables como ya veremos mas adelante.
- Leyenda. Se utiliza cuando por razones de espacio o estética se desea identificar colateralmente alguna parte del gráfico.
- Notas aclaratorias. Tienen el mismo objetivo que en la tabla.
- Fuente. La fuente de todo gráfico es la tabla estadística que le dio origen. Una tabla en ocasiones puede ser fuente de varios gráficos

Principales tipos de gráficos

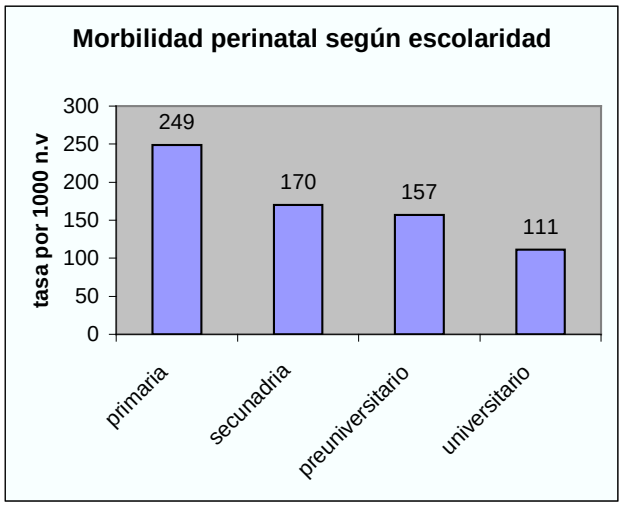
1- Diagramas de barras. Los diagramas de barras son los gráficos de elección para representar distribuciones de frecuencias de **cualitativas y cuantitativas discretas.** El tipo gráfico de barras que se utilice dependerá del número de variables a graficar.

a) Barras simples: Se utilizan para representar **una sola variable cualitativa o cuantitativa discreta.**

Como en todo gráfico en el eje de las y si inscribe la frecuencia absoluta o relativa o el índice utilizado para el análisis del información y en el eje de las x la variable con su escala.

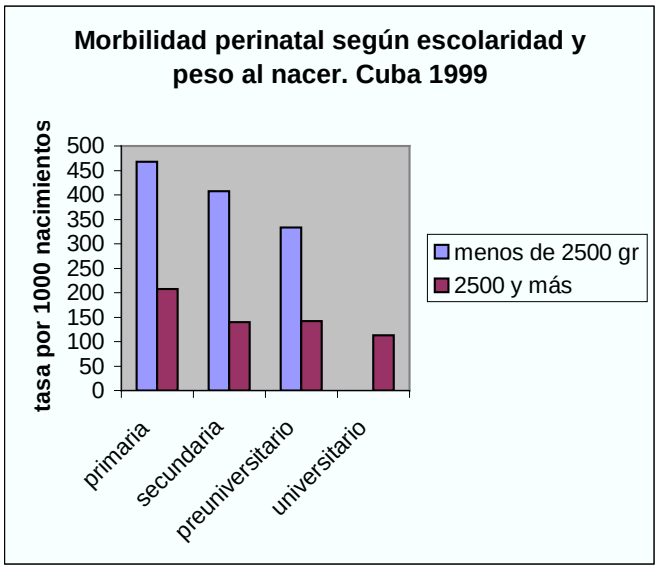
La frecuencia es representada por la altura de la barra, todas la barras tienen el mismo ancho, a libre elección, y la separación entre barras es también arbitraria pero la misma para todas; con frecuencia se toma el mismo ancho de la barra o la mitad del ancho como medida de separación.

Cada barra se rotula con el nombre de la clase que representa. Cuando el



número de categorías de la escala es muy numerosa se hace difícil el rotulado de las barras y en estos casos se acostumbra a “rotar los ejes”, se colocan las barras en posición vertical para facilitar la identificación de las barras.
Si los datos que se representan no están en una secuencia que deba conservarse, es aconsejable ordenar las barras de mayor o menor o viceversa, con lo que se mejora la comprensión y el diseño.

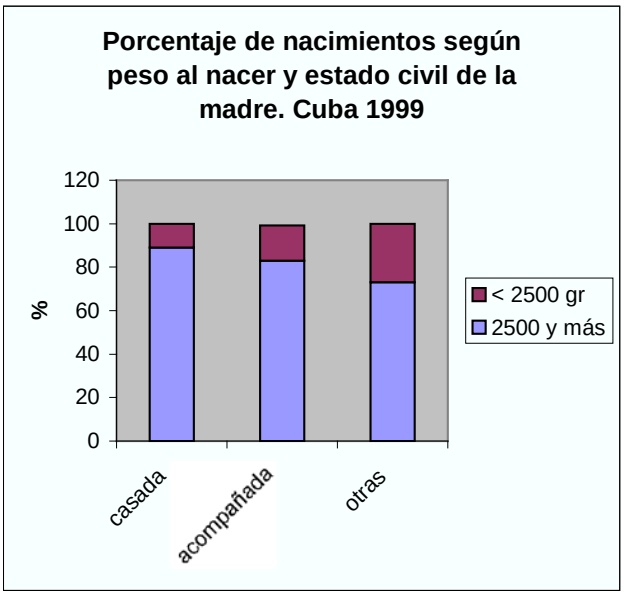
b) Barras múltiples: Se usan para representar **dos variables cualitativas o cuantitativas discretas.** Las barras pueden ser dobles, triples etc, depende del número de clases o categorías de la escala de una de las variables y el número de grupos de barras estará dado a su vez por el número de clases de la escala de la otra variable. Se sugiere que la variables que más interés se tenga en resaltar será la que formará los



grupos de barras.

Observe en el ejemplo, que se trata de dos variables, una cualitativa ordinal, la escolaridad y la otra aunque por su naturaleza es continua se trabajó con una escala discreta, < 2500 y 2500 y más. Como la variable que más se quería destacar era el peso de los recién nacidos es la que forma los grupos de barras y al ser la escala dicotómica da origen a un gráfico de barras dobles. Al tener la escolaridad 4 clases en la escala se formaron 4 grupos de barras dobles. Si nuestro interés hubiese sido poner como variables principal la escolaridad tendríamos entonces dos grupos de barras cuádruples.

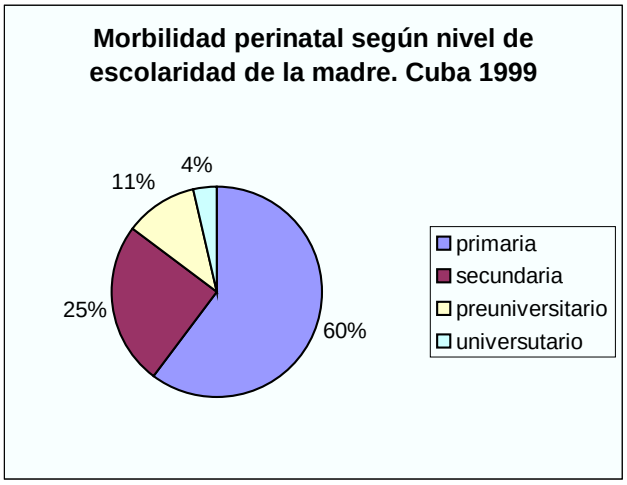
c) Barras compuestas: Al igual que las barras múltiples se usan para variables cualitativas y cuantitativas discretas, por lo que deviene en una alternativa para variar el diseño.



Sólo tiene que cumplirse un requisito adicional y es que las distribuciones de

frecuencias tienen que ser relativa expresadas en %. La barra en su totalidad representa el 100% y cada parte en que se divide la barra el aporte que hace la otra variable.

2- Gráfico circular o de pastel. Se usa para representar la distribución de **una sola variable cualitativa o discreta**, por lo que constituye una alternativa de presentación de las barras simple, pero al igual que las barras compuestas la distribución tiene que estar en %



Cuando la variable tiene un número considerable de categorías en la escala no es

aconsejable utilizar este tipo de gráfico porque tendría tantas subdivisiones que obstaculizarían su comprensión, en estos casos el grafico de barras simples es el de elección. Está misma situación se daría en las barras compuestas y se prefiere en este caso las barras múltiples.

3- Histograma. Hasta ahora los gráficos que hemos visto son para variables cualitativas y cuantitativas el caso que sean discretas. Las variables **cuantitativas continuas** también tienen sus gráficos para ser representadas, y es el caso del histograma que se utiliza cuando queremos graficar **una sola variable continua.**
No es precisamente este gráfico el que más se usa, sino, el polígono de frecuencias como veremos más adelante. Si nos detenemos a dar una explicación exhaustiva del mismo, es porque nos ayudará mucho a entender las distribuciones teóricas de frecuencias en la parte inferencial.
En el eje de las **x** se inscriben los límites inferiores de los intervalos de clase y en el de las **y** la frecuencia absoluta o relativa.
La figura que se usa en la presentación es también el rectángulo, como en los gráficos de barras. Cuando todos los intervalos de clase tienen la misma amplitud, simplemente la altura de la barra dependerá de la frecuencia del intervalo, el ancho a criterio del diseñador y todas la barras tendrán el mismo ancho, pero para dar el sentido de continuidad que tiene la variable las barras no se separan, sino que van unidas.

Sin embargo, basta con que un solo intervalo da clase tenga una amplitud diferente al resto para que no podamos asumir el diseño del histograma como lo vimos

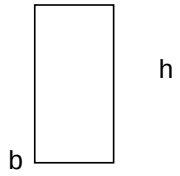
anteriormente, hay que “promediar por cada unidad de escala” y veamos que quiere decir esto con un ejemplo.

Defunciones por accidentes en menores de 40 años según edad. Matanzas 2000

Edad en años	Número
<1	10
1-4	80
5-9	125
10-19	240
20-29	50
30-39	10

Si realizáramos el gráfico dándole la altura al rectángulo en dependencia del valor de la frecuencia, el grupo de edad de 20 a 29 años sería el mayor y por ende, diríamos que los accidentes son más frecuentes en ese grupo de edad.
Pero ¿ cuántos años contempla ese grupo de edad? ¿ Tienen todos los grupos la misma amplitud y por consiguiente son esos grupos comparables?
Evidentemente no, las amplitudes cambian, los tres últimos intervalos tiene 10 de amplitud, abarcan 10 edades, mientras que los primeros intervalos tienen 1, 4 y 5 respectivamente. Es lógico que los intervalos de mayor amplitud, que contemplan un mayor número de edades, que abarcan una mayor parte de la población tengan más fallecidos.
Para hacer esos grupos comparables se divide la frecuencia da cada intervalo entre la amplitud correspondiente y así obtenemos las defunciones promedio por año(promedio por unidad de escala) y este resultado es lo que nos da la altura de la barra.
Otro aspecto a tener en cuenta que las barras no pueden tener todas el mismo ancho, el ancho de la barra dependerá de la amplitud de la clase.
Es por ello que se dice que el histograma es un diagrama de superficie, o sea que no es la altura de la barra la que representa la frecuencia, sino, que la frecuencia de la clase está representada por toda el área de la barra.
Si recordamos de la enseñanza media la formula de cálculo del área de un rectángulo está es:

A= bxh
Donde:
A= área del rectángulo
b= base del rectángulo
h= altura del rectángulo



Tenemos el área, que es la frecuencia del intervalo, tenemos la base que es la amplitud, lo que no conocemos es que altura darle a cada rectángulo para que se cumpla que toda el área represente la frecuencia.
Si despejamos de la fórmula tenemos:
h= A/b
A= frecuencia
b= amplitud

Sustituyendo:
h= frecuencia / amplitud

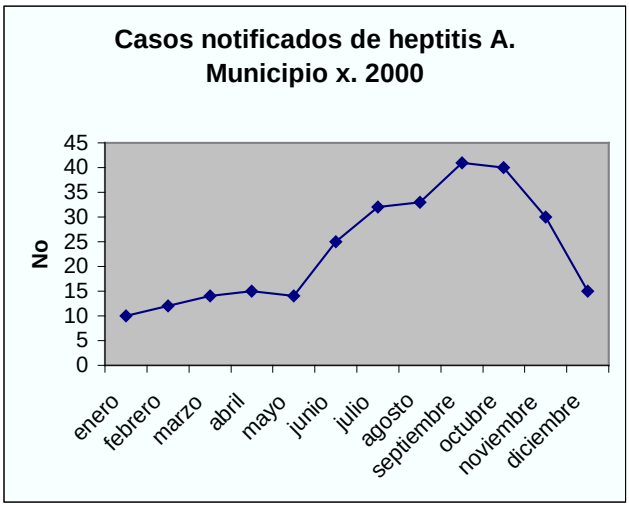
Podríamos pensar, que en buena ley, siempre se debe dividir la frecuencia entre la amplitud para tener una altura que asegure que toda el área sea la frecuencia. Teóricamente debe ser así , .lo que ocurre es que al tener todos los intervalos igual amplitud en la práctica este es un paso innecesario, porque al dividir por el mismo valor todas las frecuencias la diferencia entre la altura de las barras es la misma.
Si se sumaran todas las áreas de los rectángulos, esta suma es igual al total de casos, si se trata de la frecuencia absoluta. Si se trabajó con la frecuencia relativa, la suma de todas las áreas de los rectángulos tiene que ser igual a la unidad, si se trata de proporciones o a 100 si se trabajó con porcentajes.

4-Polígono de frecuencia. Se utiliza más que el histograma y no sólo me permite representar la distribución de una variable continua, sino, que puedo tener más de una distribución en el mismo gráfico, o sea, puedo tener dos variables, la cuantitativa continua que es la variables principal y otra cualitativa casi siempre aunque puede ser discreta y valorar comparativamente los cambios en la distribución de la variable principal.
Por ejemplo: el peso por sexo, para hombres y mujeres en un mismo gráfico porque en este caso la forma de representación es una curva, o como se suele decir es un gráfico de tipo lineal.

El polígono se construye exactamente igual que el histograma, lo único que cambia es la forma de representación, en lugar de dibujar un rectángulo a la misma altura que tendría la barra se ubica un “punto” en el centro del intervalo de clase, a nivel de la marca de clase, en el punto medio del intervalo, luego esos puntos se unen y queda una línea curva que sustituye a los rectángulos.
Como el polígono es también un diagrama de superficie, el área bajo la curva es igual al tamaño e la muestra o de la población, al total de observaciones, si se trabajó con frecuencias absolutas, y si fue con frecuencia relativas, el área bajo esa curva es equivalente a 1 si se usó la proporción o a 100 si fue el porcentaje

5- Gráfico Aritmético Simple (GAS). Es el diagrama de elección cuando se quiere representar series de tiempo donde su uso rebasa los límites de la representación para convertirse el algunos casos es técnicas de análisis como ocurre con los canales endémicos.

En el eje de las x aparece el tiempo expresado en años, meses, semanas, días u horas y en el eje de las y la frecuencia o el índice calculado para el análisis de la información. El valor de la frecuencia para cada unidad de tiempo queda representada por un punto, que posteriormente se unen para dar lugar a una curva que representa el comportamiento del fenómeno estudiado.



Ejercicios resueltos:

1. En una investigación sobre riesgos profesionales en los trabajadores de la industria del cromo en 1995, se encontró en una muestra de 897 trabajadores, que 509 de ellos presentaban perforación del tabique nasal.

Del total de trabajadores de la muestra, 41 habían trabajado menos de 6 meses en la industria del cromo; 117 lo habían hecho por un lapso de 6 meses a 3 años; 370 se habían mantenido en ese trabajo por un período de 4 a 10 años; 369 llevaban más de 10 años trabajando en la industria.

Entre los 509 trabajadores portadores de perforación del tabique nasal. Sólo 1 había permanecido en la industria durante menos de 6 meses; 46 lo habían hecho durante un período de 6 meses a 3 años; 205 habían trabajado durante 4 a 10 años en ambiente saturado de derivados crómicos; y 257 estaban hacía más de 10 años en la industria del cromo.

a- Presente estos datos en una tabla estadística completa.

b- Describa la información presentada.

c- Qué gráfico utilizaría y por qué?

Respuesta.

a- Distribución de trabajadores según tiempo de trabajo y lesión de tabique nasal. Industria de cromo. 1995.

Tiempo	*Con perforación*		*Sin perforación*		*Total*
	No	*%*	*No*	*%*	
< de 6 meses	1	0.2	40	10.3	41
6 meses a 3 años	46	9.0	71	18.3	117
4 a 10 años	205	40.3	165	42.5	370
Más de 10 años	257	50.5	112	28.9	369
Total	509	100	388	100	897

b- Más de la mitad de los trabajadores presentaron perforación del tabique nasal, aproximadamente el 50% tenían 4 años o más de trabajo en la industria, a medida que se aumenta el tiempo de exposición la frecuencia de lesionados aumento.

c- Barras múltiples. Tenemos dos variables, la perforación del tabique es una variable cualitativa dicotómica, pues solo admite dos respuestas perforó o no perforo. El tiempo de trabajo, aunque por su naturaleza es cuantitativa continua, se ha trabajado con una escala cuantitativa discreta, hecho este que ocurre con relativa frecuencia en las variables cuantitativas donde la naturaleza de la variable y el tipo de escala no siempre coinciden.

Observe que en este caso la escala tiene solo 4 intervalos de clase, con amplitudes muy diferentes, medidos en distintas unidades de tiempo, y dos de ellos son abiertos por lo que no se ha tenido mucha precisión en la medición.

2. Según datos reflejados en el anuario estadístico del MINSAP del año 1999.Las tasas de fecundidad por edades de la madre por 1000 mujeres para los años 90, 93 y 94 se comportaron de la siguiente forma.

Para el año 1990, para el grupo de 15 - 19 años la tasa fue de 77,5 de 20 - 24 años, 113.0 de 97.4 para el grupo de 25 - 29 años, 56,1 para el de 30 -34 años,

para el grupo de 35 - 39 años, 17,5 para el de 40 - 44 años, 3,3 y para el de 45 - 49 años de 0,3. La tasa global total fue de 62,1.
Para los años 95-99 con el mismo ordenamiento de los grupos etáreos las tasas fueron:

1995	1999
60.2	53.9
91.4	86.7
78.8	87.4
46.8	53.1
17.1	23.7
2.5	2.8
0.5	0.3
48.3	49.2

Además en el anuario se aclara que los datos del año 99 son provisionales.

a - Construya una tabla estadística para presentar estos datos.
b - Comente brevemente el comportamiento de la fecundidad.
c - Qué gráfico Usted utilizaría y por qué?

Respuesta.

a- Fecundidad según edad de la madre. Cuba 1990,1995 y 1999

Edad de la madre	1990	1995	1999
15-19	77.5	60.2	53.9
20-24	113.0	91.4	86.7
25-29	97.4	78.8	87.4
30-34	56.1	46.8	53.1
35-39	17.5	17.1	23.7
40-44	3.3	2.5	2.8
45-49	0.3	0.5	0.3
Total	62.1	48.3	49.2

Fuente: Anuario estadístico del MINSAP.1999

b- Se produjo un descenso de la fecundidad muy marcado del año 90 al 95 , producto de la crisis económica por la que atraviesa el país, ya en el año 99 se

observa una ligera tendencia el incremento que pensamos que se deba por una parte a una cierta mejoría de las condiciones socio económicas y por otro lado las mujeres que postergaron en un inicio sus embarazos ya en el año 99 se deciden a tener sus hijos como se refleja en el incremento de las tasas en el grupo de 30-34 años que casi iguala sus niveles al grupo de 15-19, lo que no ocurre con el resto de los años analizados donde el grupo de las adolescentes ocupa la tercera posición y que es en el año 99 en el único año donde el grupo de 25- 29 supera al de 20 a 24

c- Polígono de frecuencias. Se trata de dos variables, la edad de la madre cuantitativa continua y la otra variable es el tiempo con tres años solamente. Tendríamos en el mismo gráfico representado el comportamiento de las tasas para esos tres momentos lo que además facilita la comparación.

3. Qué gráfico Ud. utilizaría para presentar las siguientes distribuciones y por qué?

a - Distribución de pacientes con TB. según edad y sexo
b - Distribución de camas de un hospital por servicio.
c - Distribución de recién nacidos según peso al nacer.
d - Distribución de mujeres según número de hijos y escolaridad.
e - Porcentaje de remisiones según especialidad.

Respuestas.

a - Polígono de frecuencias por tratarse de 2 variables, una cuantitativa continua (edad) y la otra cualitativa nominal(sexo) .
b - Barras simples o pastel, por tener una variable cualitativa nominal(servicio).
c - Histograma o polígono. Una sola variable cuantitativa continua, peso al nacer.
d - Barras múltiples, por ser dos las variables, una cuantitativa discreta (número de hijos) y la otra cualitativa ordinal.
e - barras simples o pastel, una sola variable cualitativa nominal.

Medidas de resumen para variables cuantitativas.

Las distribuciones de frecuencias son de gran utilidad en los análisis. El número de consultas médicas realizadas por especialidades por un centro de salud puede ser una información muy importante para la planificación de recursos.

Sin embargo, en ocasiones manejar un gran número de datos no es lo más aconsejable para determinados análisis, y si es muy beneficioso tener toda la distribución de una variable cuantitativa resumida en un solo valor o en muy pocos valores que representan con suficiente aproximación todo la distribución.

Las medidas de resumen para las variables cuantitativas se dividen en dos grandes
grupos: las medidas de tendencia central y las denominadas medidas de posición.

Medidas de tendencia central.

Este grupo de medidas van a resumir toda una distribución, generalmente en un solo valor, que tiende a ocupar una posición central entre el menor valor y el mayor valor de la serie de datos y al rededor del cual se agrupan los valores que asume la variable. Ellas son la media aritmética, la mediana y la moda.

Media aritmética. También denominada promedio o simplemente media, es la más importante de todas las medidas de resumen. La media para su cálculo, tiene en cuenta todos los valores de la variable, este elemento hace que ella represente muy bien el comportamiento de la variable.

En una serie simple de datos, la media se obtiene sumando todos los datos y dividiéndolos entre el total de observaciones.

Ej.
Peso en kg. de 5 niños.
3,4,6,8,9,

La media se obtendría sumando los 5 pesos y dividiéndolos entre 5.

3+4+6+8+9/5=30/5
$\overline{X}$ =6

La media o el peso promedio de los 5 niños es 6kg.Su expresión de cálculo es:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Para una serie simple pero con las frecuencias absolutas previamente calculadas, la media se obtendrá multiplicando la frecuencia por el valor que asume la variable, sumar esos productos y dividirlos entre el total de observaciones.
Peso en kg. de 30 niños:

Peso (Kg)	No.
5	3
7	8
6	10
9	6
11	3
Total	30

$\overline{X}$ = (5x3)+ (7x8)+ (6x10)+ (9x6)+ (11x3) /30
$\overline{X}$ = 15+56+60+54+33/30
$\overline{X}$ = 218/30
$\overline{X}$ = 7.3kg.

La media de los pesos de los 30 niños es 7.3kg.

La expresión de cálculo es:

$$\overline{X} = \frac{\sum X.f_i}{n}$$

Veamos por último el cálculo de la media cuando se tienen los datos agrupados, cuando se tiene una escala con intervalos de clase.

Peso en kg. de 25 niños de una escuela primaria.

Peso en kg	**No.**
20–24	**4**
25–29	**10**
30–34	**8**
35–39	**3**
Total	**25**

Para el cálculo de la media tengo ahora la dificultad de no saber el peso exacto de los niños, los 4 niños que clasificaron en el primer intervalo de clase pueden haber tenido cualquier peso entre 20 y 24 kg. Para resolver este inconveniente

asumimos que el peso de los niños es el valor central del intervalo de clase, la marca de clase.

Se procede de igual forma que para el cálculo de una serie simple con frecuencia. Multiplicamos la marca de clase por la frecuencia del intervalo, sumamos esos resultados y lo dividimos entre el total de observaciones.

Marca de clase	No	Mc. fr
22	4	88
27	10	270
32	8	256
37	3	111
Total	25	725

$\overline{X}$ = 725/25
$\overline{X}$ = 29kg

EL peso medio de los 25 escolares es de 29kg.
La expresión de cálculo es:

$$\overline{X} = \frac{\sum mc.f_i}{n}$$

Medias o promedios ponderados.

En ocasiones puede ocurrir que tengamos la información resumida por más de una media y después nos percatamos de la necesidad de calcular una media general.

Supongamos que le maestría en gerencia tiene una matricula de 60 alumnos, 40 son varones y 20 son hembras. La edad promedio de los hombres es de 34.8 años y la de las mujeres es de 29.7 años. Si quisiéramos conocer la edad promedio general del grupo es lógico pensar que esas medias parciales que ya están calculadas por sexo me pueden servir para abreviar los cálculos.

Una tendencia generalizada es asumir que el promedio de las medias parciales sería la forma de calculo de esa media general. Sin embargo, esa no es la forma correcta de operar, es necesario tener en cuenta el tamaño de las muestras de donde se obtuvieron esas medias, y para ello se multiplica el valor de la media por su **n** correspondiente. Este proceso, que se usa mucho en estadística se conoce con el nombre de **ponderación.**

La expresión de calculo de la media ponderada sería entonces:

$$\overline{X} = \frac{n_1\overline{x_1} + n_2\overline{x_2} + n_3\overline{x_3}....n_i\overline{x_i}}{n_1 + n_2 + n_3....n_i}$$

La media, a pesar de ser la mejor medida de tendencia central y precisamente su precisión en los cálculos está dado en buena medida por el hecho de tener en cuenta todos los valores que asume la variable, esto puede convertirse en determinadas circunstancias en una limitación para su uso.

Supongamos que las edades de los estudiantes de la maestría se mueven en un rango de 25 a 35 años y al calcular la edad promedio fue de 30.5 años. Un profesional de cierto prestigio y que a dedicado buena parte de su vida a trabajar dentro del Sector Salud, matricula el curso a pesar de tener 60 años.

Al calcular la edad promedio de los cursistas ya no tendrá un valor de 30.5 años, valor que reflejaba muy bien el comportamiento del grupo, sino un valor mucho mayor que se aleja mucho del comportamiento real. En este caso, cuando existen valores extremos o aberrantes la media se ve afectada en su cálculo por los mismos y deja de ser una buena medida de resumen. Entonces preferimos para resumir la información la **mediana**, que como veremos a continuación, por su forma de cálculo no se ve afectada ante esta situación.

Mediana. Es la observación que en una serie **ordenada**, ocupa la posición central, por tanto divide a la serie en dos parte iguales. Por encima de ella se encuentra el 50% de las observaciones y a su vez su valor supera al 50% restante.

Con la expresión: (n+1)/2 se obtiene cual posición en la serie ordenada es la mediana.

Por Ejemplo.

Peso en kg. de 5 niños.

5, 6, 7, 9, 11,

Ya se encuentra ordenada de menor a mayor, aunque a simple vista se aprecia que la mediana es 7kg,apliquemos la expresión para encontrarla:

5+1/2=3

En efecto la tercera posición corresponde al niño que peso 7kg.

Cuando la serie es par no hay un único valor central, la mediana sería en este caso la semisuma de los dos valores centrales.

3, 5, 6 ,7, 9, 11

El 6 y el 7 ocupan la posición central, si aplicamos la expresión matemática tendremos:
6+1/2=3.5

La mediana está entre la tercera y la cuarta posición.

Me = 6+7/2
Me = 6.5kg

La mediana por su propia forma de cálculo, no se verá afectada por valores extremos pues al realizar el ordenamiento de los datos estos serian la primera o la última observación.

Es una medida muy utilizada en epidemiología, sobre todo el los estudios de series cronológicas como veremos en el tema 1.5,precisamente por poder se aplicada a series irregulares.

Puede ser calculada para datos agrupados, pero se usa con poca frecuencia. En el caso de tener escalas abierta, la media no puede ser calculada, porque los intervalos abiertos no tienen marca de clase, elemento presente en la expresión de cálculo. En este caso se prefiere utilizar la mediana que no se ve afectada por este hecho.

La moda es la menos utilizada de las medidas de tendencia central, y como su nombre lo indica es el valor más frecuente.

Ej. Edad en años de 10 pacientes con SIDA.

20, 27, 23, 35, 23, 40,32, 23, 19, 17

La moda en este caso es 23 años

En este caso:

20, 24, 15, 34, 20, 38, 24, 34, 24, 20

Hay dos modas, 20 y 24 años, se dice que esa distribución es bimodal.

Por último, puede ocurrir que la moda no exista:

20, 35, 25, 18, 36, 24, 19, 33, 27, 15,

Puede ser calculada para datos agrupados, en cualquier libro de Estadística Ud. puede encontrar la formula.

Medidas de dispersión.

Una distribución de frecuencias para una variable cuantitativa no puede ser resumida con solo utilizar una medida de tendencia central. El hecho estriba en que si bien el valor que esa medida asuma nos indica que el resto de las observaciones están ubicadas a su alrededor, no sabemos cuan lejos o cerca pueden estar situadas y es fácil entender que mientras más próximas estén las observaciones de esa medida, mejor reflejará esta el comportamiento de la distribución.

Veamos un ejemplo.

Días de hospitalización de 5 niños ingresados en un servicio de E.D.A.

3, 5, 7, 9, 11
$\overline{X}$ = 7 días
Me= 7 días

Días de hospitalización de 5 niños en un servicio de I.R.A.

1, 2, 7, 12, 13

$\overline{X}$ = 7 días
Me= 7 días

Aunque en los dos ejemplos coinciden los valores de las medias y las medianas, los valores de las observaciones no se distribuyen alrededor de las mismas de igual manera.

En el caso de las E.D.A. la dispersión es menor, los valores se concentran más alrededor de la media y la mediana, lo que nos hace pensar que en este caso las medidas de tendencia central reflejan mejor el comportamiento de la variable.

Por consiguiente una medida de la dispersión, de la variabilidad, debe ser añadida a la medida de resumen para completar el análisis.

.

El **rango, recorrido o amplitud** se obtiene por la diferencia entre el mayor y el menor valor observado.

En el ejemplo de la E.D.A. sería: 11-3=8

Para las I.R.A. tendríamos: 13- 1=12

Aunque logra cuantificar la variabilidad de los datos, no es una buena medida porque no tiene en cuenta el resto de las observaciones, solo los dos valores extremos.

Por consiguiente, una buena medida, debe tener en cuenta todos los valore que asume la variable, e ir midiendo cuan cerca o lejos están de esa medida de tendencia central. Además para obtener un valor único, lo lógico es hacer un promedio de esas diferencias. Apliquemos estas ideas al ejemplo de las E.D.A.

$\overline{X}$ =7
(3-7)+ (5-7)+ (7-7)+ (9-7)+ (11-7)/5
(-5)+(-2)+(0)+(2)+(5)/5

Como puede apreciarse, al realizar la suma algebraica de las diferencias de los valores con respecto a la media, estos se anulan y es algo que va a ocurrir siempre, invariablemente.

Esta situación se da por que los valores se alejan de la media en dos sentidos, por defecto y por exceso, dicho de otra manera, unos serán menores que la media y otros la superarán.

Sin embargo, nuestro interés es tener una medida de la dispersión, pero no tiene que precisar en que sentido se alejan de la media, lo que se traduce en que no nos interesa el signo que se obtenga al encontrar las diferencias.

Así surge la **desviación media**, y para ello lo que se calcula es la diferencia modular.

D.M.= $\frac{\sum \left\| x_i - \overline{x} \right\|}{n}$

Con el mismo ejemplo de las E.D.A. tendremos:

D.M.= (|3-7|)+ (| 5-7|)+ (|7-7|)+(|9-7|)+(|11-7|)/5
D.M.= 4+ 2+ 0+ 2+ 4/5
D.M.=2.4 días.

Los niños permanecieron ingresados en el servicio de E.D.A. en promedio 7 días con una dispersión o variabilidad promedio de 2.4 días.

Varianza. Otra forma de lograr que la suma de las diferencias de los valores de la variable con respecto a la media no se anulen, es elevando esta diferencia al cuadrado. De esta forma el promedio de las desviaciones al cuadrado es la medida que se conoce bajo el nombre de **varianza**.

Su expresión de cálculo para una serie simple de datos es:

$$S^2 = \frac{\sum (x - \bar{x})^2}{n}$$

Con el mismo ejemplo anterior tendremos:

S^2 =(3-7) ^{2} + (5-7) ^{2} +(7-7) ^{2} +(9-7) ^{2} +(11-7) ^{2} /5
S^2 =16+4+0+4+16/5
S^2 =8

Como el resultado sería 8 días al cuadrado, algo realmente difícil de entender y analizar, se le saca la raíz cuadrada a ese resultado y obtendremos un valor de fácil interpretación:2.8 días.

Esta medida que hemos obtenido al calcular la raíz cuadrada a la varianza es la denominada **Desviación estándar**, la más conocida y utilizada de las medidas de dispersión.
Su expresión de cálculo para una serie simple es:

$$S = \sqrt{(x_1 - \bar{x})^2} \text{ /n}$$

Para el cálculo con datos agrupados utilizaremos el mismo ejemplo que se trabajo para la media, los pesos de 25 niños en edad escolar, que fue de 29 kg.

Al igual que en el caso de la media, se asume que la marca de clase es el valor de la variable para el número de observaciones que clasificaron en el intervalo.

M.C	Fi	(M.C.– $\bar{x}$)^2	(M.C.– $\bar{x}$)^2 fi
22	4	7= 49	196
27	10	2= 4	40
32	8	3= 9	72
37	3	8= 64	192
Total	25		Σ500

S^2 =500/25=20 kg.
S= $\sqrt{\ }$ 20

S=4.5 kg.

Su expresión de calculo es:

$$S^2 = \frac{\sum (mc_i - \bar{x})^2 f_i}{n}$$

La media de los pesos de los niños fue de 29 kg., con una desviación estándar de 4.5kg.Una forma muy intuitiva de interpretar esta medida es la siguiente, si a la media se le resta y se le suma la desviación estándar, se forma un intervalo donde se encuentra una buena parte de las observaciones. Para este ej. entre 15.5 y 24.5kg se encuentra los pesos de una parte importante de los niños estudiados.

Todas las medidas de dispersión que hemos visto hasta ahora se consideran medidas de **dispersión absoluta**.

Cuando queremos saber entre 2 o más variables cuales presentan una mayor variabilidad o dispersión, no lo podemos conocer comparando algunas de las medidas de dispersión analizadas, pues los valores que alcancen dependerá de la unidad de medida propia de cada variable.

Para realizar estas comparaciones se utilizan las denominadas **medidas de dispersión relativas** concretamente el **coeficiente de variación.**

El coeficiente de variación indica el tanto por ciento de la media que representa la desviación estándar.

Cv = $\bar{X}$ /Sx100

Supongamos que el peso promedio de un grupo de estudiantes fue de 62kg,con una DE. de 6kg, y la media de la talla de 162cm con una desviación estándar de 8 cm. Si comparamos las DE. parece ser la talla la más variable de las dos . Veamos que ocurre cuando calculamos el CV.

CV_{peso} = 6/62 x 100=9.6%

CV_{talla} = 8/162 x 100=4.9%

El peso tiene una mayor dispersión que la talla.

Ejercicios resueltos

1. A continuación aparecen la ganancia de peso durante el embarazo de 10 mujeres a termino.

peso en kg.

10, 9, 8, 11, 10, 8, 7, 11, 10, 9

Calcule:

a- la media aritmética.

b- la mediana.

c- la moda

d- la desviación estándar.

Respuesta.

a-9.3

b-9.5

c-10

d-1.3

2. El cuadro que aparece a continuación muestra el peso en kg. de 100 estudiantes universitarios.

Peso en kg	No
60–62	5
63–65	18
66–68	42
69–71	27
72–74	8
Total	100

Calcule.

a- La media.

b- La desviación estándar.

Comentario [FV1]:

Respuestas:

1- 67.45 kg.

2- 2.92 kg.

3- En un grupo de 150 estudiantes entre 15 y 19 años se encontró que el peso promedio era de 62 kg. y la desviación estándar de 6 kg.

La estatura promedio fue de 162 cm y la desviación estándar de 8 cm.

Cuál de las dos variables tiene una mayor dispersión relativa?.

Respuesta.

CV = 9.6% (peso)

CV = 4.9% (talla)

EL peso tiene una mayor variabilidad que la talla.

Medidas de resumen para variables cualitativas.

Las medidas de resumen para variables cualitativas también se conocen como frecuencias relativas y medidas por cocientes. Este último nombre deriva del hecho de que todas ellas establecen una relación por cocientes, por lo que siempre habrá determinados aspectos a analizar incluidos el numerador y otros elementos que constituirán el denominador. Estas relaciones que se establecen, permiten hacer más evidentes ciertos comportamientos de las variables analizadas que por el simple análisis de las frecuencias absolutas no seria posible evidenciar.

Supongamos que tenemos una población hipotética de 300 personas,200 son mujeres y 100 son hombres. Aunque es evidente la superioridad numérica femenina, se puede hacer más elocuente si establecemos una relación entre esas dos categorías.

R: h/m= 100/200= 1/2=0.50=50x100.

En la población estudiada hay 1 hombre por cada 2 mujeres,0.50 son hombres y multiplicado ese resultado por 100 para facilitar el análisis podemos decir que por cada 100 mujeres hay 50 hombres.

Esta medida que hemos calculado es la **razón**, por tanto podemos definirla diciendo: **es la relación por cociente que se establece entre dos categorías cualesquiera de una variable, por lo tanto el numerador nunca estará contenido en el denominador.**

Su expresión de cálculo es:

R= a/b

Otra medida que puede calcularse es **la proporción o por ciento** tal como vimos en el tema 1.1. En este caso la relación por cociente se establece entre cada categoría de la escala de la variable y el total general(tamaño de la muestra o de la población).Por tanto, y a diferencia de la razón, el numerador siempre estará incluido en el denominador.

Su expresión de cálculo es:

p= a/a+b

Veamos el ejemplo.

p/h = 100/300 = 1/3=0.33=33%
p/m = 200/300 =2/3=0.67=67%

Hay 1 hombre por cada 3 individuos que conforman la población, la proporción de hombres es de 0.33,el 33% son hombres, por cada 100 personas de esa población 33 son hombres.

Mientras que las mujeres se dan en una relación de 2 mujeres por cada 3 personas,0.67 son mujeres, lo que equivale a decir que de cada 100 individuos en esa población 67 son mujeres.

Observe como con el cálculo de la proporción o el por ciento lo que estamos midiendo es la importancia relativa, lo que aporta, el peso que tiene cada categoría de la escala de la variable con relación al total.

El por ciento como medida tiene un uso muy difundido y es de gran utilidad, tiene la ventaja de poder comparar poblaciones con totales diferentes porque todo se lleva a una base común que es 100.

Sin embargo, en ocasiones se le da un uso al por ciento que desborda sus posibilidades de análisis, y se cometen graves errores al interpretar los resultados de un estudio.
Veamos un ejemplo.

Defunciones perinatales según edad gestacional de la madre. Ciudad de la Habana.1998.

Edad gestacional (Semanas)	**Defunciones**	
	No	**%**
20-27	31	6.2
28-32	108	21.2
33-36	124	25.6
37-41	197	39.7
42 y más	36	7.3
Total	**496**	**100**

Pudiera resultar llamativo, para quien no conozca bien estas medidas, que el mayor por ciento de defunciones se concentra en las edades óptimas para parir, 39.7% de defunciones entre las 37y 41 semanas de gestación.

Este resultado es lógico, pues es precisamente en estas edades gestacionales donde ocurre un mayor número de nacimientos.

Lo que ocurre es que se está pensando en función del riesgo, por todos es conocido, que desde el punto de vista clínico, el riesgo que tiene un niño de fallecer es menor cuando se produce el parto a termino que cuando ocurre con pocas semanas de gestación.

No es el por ciento quien como medida mide **riesgo**, el riesgo lo mide la **tasa**.

Una tasa no es mas que la relación por cociente que se establece entre el número de veces que se produce el efecto que se esta midiendo, muchas veces no deseado como morir o enfermar, y en el denominador la población expuesta al riesgo que se está analizando en el numerador. Como ese resultado será siempre menor que la unidad, para facilitar su análisis se multiplica por una constante que es 10^n

Veamos el ejemplo de la mortalidad perinatal.

Edad gestacional	Defunciones		Nacimientos	
	No	%	No	tasa x 1000
20–27	31	6.2	81	382.7
28–32	108	21.2	817	132.2
33–36	124	25.0	1912	64.8
37–41	197	39.7	29713	6.6
42 y más	36	7.3	2848	12.6
Total	496	100	35371	14.0

Observe como los valores más altos de las tasas, los mayores riesgos, se encuentran en los menores tiempos de gestación, y a medida que las semanas de gestación aumentan las tasas van disminuyendo para alcanzar su valor más bajo de las 37a las41 semanas, y finalmente ascender ligeramente en los partos post-término con 42 semanas y más.

En las tasa debe existir concordancia entre numerador y denominador en tres aspectos fundamentales:

La naturaleza del hecho que se está analizando. La población tiene que estar expuesta al riesgo que se esta midiendo en el numerador, por ejemplo si se va a calcular la tasa de mortalidad por cáncer de próstata, el denominador de la tasa no puede ser la población total, pues la población femenina por razones obvias nunca estará sometida a ese riesgo y debe ser excluida del denominador el cual se limitara a la población masculina.

La zona geográfica. Tanto el numerador como el denominador de la tasa deben referirse al mismo territorio.

El período de tiempo. Las tasas pueden calcularse para diferentes períodos de tiempo, las más frecuentes son anuales. Las poblaciones entre censos se estiman y para el cálculo de las tasas anuales la poblaciones deben ser estimadas a mitad de año, al 30 de junio o al 1 de julio.

Las tasas pueden ser **brutas** o **especificas**. Una tasa es **bruta, cruda, general o global** cuando en el denominador de la tasa aparece la población total, y es

especifica cuando la población esta restringida a una parte, por ejemplo un grupo de edad.

Distinguir cuando una tasa es bruta o especifica es un elemento muy importante para el análisis, porque como veremos más adelante, las tasas brutas de diferentes territorios o de un mismo territorio en diferentes periodos de tiempo, no se pueden comparar si no se les somete previamente a un procedimiento denominado tipificación o estandarización.

Veamos algunos ejemplos de tasas de uso mas frecuente en los análisis de la situación de salud.

Tasa de natalidad = total de nacidos vivos /total de la población x 1000 hab.

Tasa bruta de mortalidad = total de defunciones/ total de la población x 1000 hab.

Tasa específica de mortalidad por edad = total de defunciones para un grupo de edad/ total -de la población para ese grupo de edad x 1000 hab.

Ej. Tasa de mortalidad de 60 años y más = defunciones de 60 años y más/total de la población de 60 años y más x 1000 hab.

Tasa de mortalidad especifica por sexos = defunciones por un sexo x/total de la población para ese sexo x 1000 hab.

Tasa de mortalidad por causa = total de defunciones por una causa x / total de la población x 100 000 hab.

Nótese, que la tasa de mortalidad por causa es una tasa bruta, aunque sea específica para una causa de defunción, porque en el denominador se encuentra toda la población. No obstante puede calcularse específica por grupos de edad, por sexo, y por grupos de edad y sexo.

Tasas especiales de mortalidad.

Tasa de mortalidad infantil = defunciones menores de 1 año/ total de nacidos vivos x 1000 nacidos vivo (n.v.)

Componentes de la mortalidad infantil:

Tasa de mortalidad neonatal precoz = defunciones menores de 7 días /total de nacidos vivos x 1000 n.v.

Tasa de mortalidad neonatal tardía = defunciones de 7 a 27 días /total de nacidos vivos x 1000 n.v.

Tasa de mortalidad posneonatal = defunciones de 28 días a 11 meses/ total de nacidos vivos x 1000 n.v.

Tasa de mortalidad perinatal = defunciones fetales tardías + def. menores de 7 días/ total de nacidos vivos + def .fetales tardías x 1000 n.v.+ def. fetales tardías.

Nota: Defunción fetal tardía es aquella defunción fetal que pesa 1000gr o más, también se puede considerar las semanas de gestación, de 28 semanas o más, pero el peso es más preciso como criterio de clasificación.

La tasa puede calcularse sólo considerando en el denominador los nacidos vivos, pero se es menos preciso en el cálculo.

Tasa de mortalidad materna = defunciones maternas directas/total de nacidos vivos x 100 000 n.v.

Defunción materna directa, es aquella defunción que ocurre por complicaciones del embarazo, parto y puerperio.

Tasa de letalidad = defunciones por una causa x / total de enfermos de esa causa x 100 enfermos.

Tasa de mortalidad proporcional por causa = total de defunciones por una causa x /total de defunciones x 100.

Nota: La tasa de letalidad y de mortalidad proporcional, no son realmente tasa .son proporciones, por consiguiente no miden riesgo, sino importancia relativa.

Tasas de morbilidad:

Tasa de incidencia = total de casos nuevos diagnosticados de una enfermedad x en un período de tiempo dado / total de la población x 100 000 hab.

Tasa de prevalencia = total de casos(nuevos y viejos) diagnosticados para una enfermedad x / total de la población x 100 000 hab.

Estos son algunos de los indicadores más utilizados en Salud Pública, otros de gran importancia serán abordados en el módulo 5 de estadística se salud.

Usted debe memorizar las expresiones de cálculo y saber interpretar los resultados que brindan.

Ejercicios resueltos

1- A continuación se dan cifras sobre población, nacimientos, defunciones y casos de algunas enfermedades para Cuba en 1999

Población estimada a Julio 1, 1999	11142691
Defunciones	79486
Nacidos vivos (estimados)	150590
Defunciones en menores de 1 año	966
Defunciones maternas directas	44
Defunciones por enf. del corazón	21517
Defunciones por accidentes	5188
Defunciones por influenza y neumonía	5388
Defunciones por tumores malignos	16376
Defunciones por enfermedades cerebro vasculares.	8402
Casos nuevos de tuberculosis	1111
Casos de meningitis meningocóccica	70

A partir de los datos anteriores obtenga los siguientes indicadores.

a) - Tasa de Natalidad.
b) - Tasa Mortalidad general.
c) - Tasa Mortalidad infantil.
d) - Tasa Mortalidad por enfermedades del corazón
e) - Tasa Mortalidad por accidentes.
f) - Tasa Mortalidad por influenza y neumonía.
g) - Tasa Mortalidad materna.
h) - Incidencia de meningoencefalitis.
i) - Incidencia de tuberculosis.
j) - Mortalidad proporcional por causas.

Respuesta.

a- 13.5x 1000 hab.
b- 7.0 x 1000 hab
c- 6.4 x1000 nv
d- 193.1x 100 000 hab.
e- 46.6x100 000 hab.
f- 48.4x 100 000 hab.
g-29.2x100 000 n.v.
h- 0.6x 100 000 hab.
i- 10.0x100 000 hab.

j- Causa	%
Enf. del corazón	27.1
Accidentes	6.5
Influenza y neumonía	6.8
Tumores malignos	20.6
Cerebro vasculares	10.6

2 - ¿ Qué indicador usted calcularía y que datos necesitaría para medir los siguientes aspectos en una población?.

1. Riesgo de morir por TB.
2. Riesgo de morir por cualquier causa en la población X.
3. Riesgo de estar enfermo de tifoidea en un momento dado.
4. Riesgo de morir de los enfermos de sarampión.
5. Riesgo de morir en el segundo decenio de la vida.
6. Importancia de la tuberculosis como causa de muerte.
7. Importancia de la tuberculosis como causa de muerte en los mayores de 15 años.

Respuesta.

1- Total de defunciones por TB./total de la población x 100 000 hab. Tasa de mortalidad por TB.

2- Total de defunciones / total de la población x 1000 hab. Tasa de mortalidad general.

3- Total de casos de fiebre tifoidea/ total de la población x 100 000 hab. Tasa de incidencia.

4- Total de defunciones por sarampión/ total de enfermos por sarampión x 100 hab. Tasa de letalidad

5- Defunciones de 10 a 19 años / población de 10 a 19 años x 1000 hab. Tasa especifica de mortalidad por edad.

6- Total de defunciones por TB. / total de defunciones x 100 defunciones. Mortalidad proporcional.

7- Total de defunciones por TB. en los mayores de 15 años / total de defunciones en los mayores de 15 años x 100 def. Mortalidad proporcional.

ESTADÍSTICA INFERENCIAL

Introducción a la estadística inferencial.

Es bueno recordar que la estadística inferencial trata de las bases lógicas y aporta las técnicas necesarias para extraer conclusiones sobre la población a partir de una muestra.

En general la inferencia estadística nos permite resolver dos grandes problemas:

- La estimación de los parámetros de la población de referencia, a partir de la muestra.

- La verificación de hipótesis estadísticas, mediante las denominadas pruebas o contrastes de hipótesis, que permiten determinar el grado de compatibilidad de los valores muestrales con los parámetros de la población.

Este proceso de generalización o de inferencia de los resultados encontrados en una muestra a la población de origen, lleva siempre implícito un grado de **incertidumbre** sobre la certeza de nuestras estimaciones o decisiones .

Es por eso que la **teoría de probabilidades** es el fundamento básico de la inferencia estadística, pues nos va o facilitar los procedimientos necesarios para cuantificar la incertidumbre.
Sin embargo, ello no implica que se deba ser une experto en Teoría de probabilidades, para utilizar los métodos estadísticos de forma efectiva, basta con saber los supuestos teóricos básicos para poder utilizar estas técnicas y lo que es más importante aún saber interpretar adecuadamente los resultados No existe una sola definición de probablidades**.**
El concepto estadístico de probabilidades surge de su aplicación en los juegos de azar y actualmente se utiliza en cualquier ciencia cuando lo que queremos es **cuantificar el grado de incertidumbre** que tenemos la tomar decisiones o al hacer estimaciones de los parámetros poblacionales a partir de una muestra.
Vamos a trabajar con la denominada **definición frecuentista de probabilidades:**
La probabilidad de ocurrencia de un determinado suceso indica la ***frecuencia relativa*** *con que dicho suceso tendrá lugar* ***a largo plazo en pruebas sucesivas y en condiciones similares.***
Si un evento ocurre **X** veces de **n** veces su frecuencia relativa es **x/n** , proporción o porcentaje, y el valor al cual se aproxima **x/n** cuando **n** tiende a ser infinito, se le llama **límite de la frecuencia relativa**. El concepto de límite es un concepto matemático complejo, que implica toda una serie de demostraciones y cálculos, por eso vamos a hacer **límite** sinónimo de **a largo plazo** que es más intuitivo y fácil de entender.

Veamos un ejemplo clásico. Una moneda balanceada, que tiene por una de sus caras una estrella y por la otra un escudo; cuando decimos que la probabilidad de que salga cara al lanzar al aire la moneda es de ½, 0.50 o 50% no quiere decir que necesariamente si yo lanzo 10 veces la moneda al aire 5 veces va a salir cara y 5 veces escudo, o si lanzo la moneda 100 veces no tiene necesariamente que salir 50 veces estrella y 50 veces escudo, pero si podemos esperar que si una moneda balanceada (condiciones similares) se lanza un gran número de veces, infinitas veces (pruebas repetidas a largo plazo) podemos esperar obtener muy cerca del 50% de estrellas y el 50% de escudos.
Puesto que la definición de probabilidad que hemos dado es en termino de frecuencia relativa, el valor numérico que toma la probabilidad siempre será un valor positivo, y oscilara entre 0 y 1 si la expreso en forma de proporción o de 0 y 100 si es en forma porcentual.
Tan bien de aquí se desprende que la **probabilidad de no ocurrencia** de un evento será la substracción de la unidad o sea **1-la probabilidad de ocurrencia.**
Si la probabilidad que tiene un alumno de la maestría de graduarse con resultados satisfactorios es de un 90%, la de no graduarse será 100 – 90= 10%

Veamos ahora un concepto muy importante, el **de *distribución teórica de frecuencias o de probabilidad***.
Las distribuciones de frecuencias absolutas y relativas, describen el comportamiento de una variable en una muestra o en una población, en caso de disponer de información para esta última.

En inferencia siempre trabajamos con muestras y desconocemos la distribución real de las variables en la población. Por eso los matemáticos crearon unos modelos teóricos que nos permiten acercarnos a esa realidad.

La distribución de frecuencia relativas de una variable X, en una muestra, es objetiva, es real, nosotros mismos la calculamos. Esos modelos denominados, distribución teórica de frecuencias, parten de los mismos principios de las muestrales, pero con la diferencia que no son reales, no se trabaja con datos empíricos, sino, con elementos o modelaciones teóricas y el nombre de distribuciones teóricas de probabilidad, con que también se le conoce, deriva precisamente del concepto frecuentista de probabilidad.

En estadística descriptiva, al tener calculada la distribución de frecuencias absolutas y relativas para una variable cuantitativa continua la representábamos gráficamente con un histograma o un polígono de frecuencias.

El polígono queda representado por una curva que su forma varia en función de la variable y su distribución. Si ese polígono se cierra, el área que queda delimitada debajo de esa curva, si estamos trabajando con frecuencias absolutas, es igual al total de observaciones, y si estamos trabajando con

frecuencias relativas esa área será igual a 1 si son proporciones o 100 si se trata de por cientos.

Ahora ya UD. está en condiciones de entender la definición, o una de las propiedades más importante de las distribuciones teóricas de probabilidad.

Una distribución teórica de frecuencia o de probabilidad, f(x), representa un porcentaje teórico de la distribución de x. De ahí que los porcentajes deben totalizar un valor de 100%, o sea, las probabilidades sumaran en total 100 o la unidad. Matemáticamente:

$$\sum f(x) = 1$$

La distribución normal.

La distribución normal, curva normal o distribución o curva de Gauss tiene sus orígenes a principios del siglo XVIII cuando De Moivre encontró la función matemática de define este tipo de curva.
A finales del propio siglo Gauss y La Place al hacer mediciones astronómicas cometían errores en las mediciones que realizaban. A los valores numéricos de estos errores de medición le calcularon las distribuciones de frecuencias absulotas y relativas y le realizaron los polígonos de frecuencias correspondientes y observaron que al igual que las distribuciones de otras variables cuantitativas continuas, siempre la curva que describían tenía forma de campana.
Por lo que decidieron utilizar la función que describe este tipo de curva en sus trabajos como veremos más adelante.
Es bueno recordar, de la enseñanza media, que una recta también tiene una función matemática que la define: y = a+bx ; donde **a** es el intercepto, el lugar donde la recta corta al eje de las y u ordenadas y **b** es la pendiente de la recta. Como a y b pueden tomar infinitos valores habrán también infinitas rectas.
La función que define la curva normal es:

$$F(x)= \frac{1}{\sqrt{2\pi\sigma}} e^{-1/2\frac{(x-\mu)^2}{\sigma^2}}$$

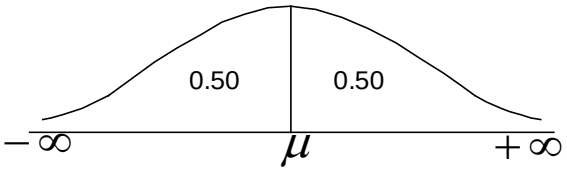
En esta expresión **e** y π son constantes; **x** es el valor que asume la variable por consiguiente los parámetros que definen esa función son: μ y σ, la media y la desviación estándar poblacional.
Variando el valor de μ se desplaza la curva sobre el eje de las x o abcisas a la derecha y a la izquierda, mientras que al variar el valor de σ, varía la dispersión

por lo que la curva será mas aplanada (más dispersión) o más apuntada (menos dispersión), por lo que existen infinitas curvas normales.

Características de la curva normal

-La distribución normal es simétrica con respecto a una perpendicular trazada desde el punto más alto de la curva al eje de las x. Queda así dividida en dos mitades exactamente iguales, que cada una tiene un área de 0.5 o el 50% pues si recordamos, todo el área bajo la curva es igual a 1 o 100.

-La distribución normal es unimodal



-La media la mediana y la moda son iguales

-En el punto de inflección de la curva, donde cambia de concavidad (cóncava para convexa) se encuentra a una distancia igual a σ, y se ha demostrado matemáticamente que:

$\mu \pm \sigma$ abarca el 68% del área

$\mu \pm 2\sigma$ abarca el 95% del área

$\mu \pm 3\sigma$ abarca el 99% del área

Con fines prácticos a mi me interesaría saber que proporción del área bajo la curva quede entre dos puntos de las abcisas o eje de las x, lo que equivale a decir entre que dos valores de la variable; o conocer para un determinado valor qué área está por encima y qué área queda por debajo. Esto implicaría sustituir los valores de μ y σ en la función que define esta curva, y realizar múltiples cálculos.

Para darle solución a este problema se creó la **normal estándar**, que no es más que una distribución normal pero con $\mu = 0$ y $\sigma = 1$. Para esta distribución normal se calcularon las áreas bajo la curva que aparecen en Tablas.

Aparecen como apéndice A-1y apéndice A-2 dos tablas de la distribución normal estándar.
En la primera columna aparecen los valores de Z con aproximación a la décima, la aproximación a la centésima se obtiene en la primera fila. En el cuerpo de la tabla
aparecen los valores de las áreas que delimitan los valores de Z, las probabilidades.

En la tabla A-1, aparecen las áreas tabuladas para la mitad de la curva, tabla de una cola, por la propiedad de ser las dos mitades idénticas, no se hace imprescindible trabajar con toda la curva, solo es necesario tener presente que los valores de Z numéricamente son los mismos pero que a la derecha de μ, para la mitad derecha de la curva, tienen signo positivo, y a la izquierda los valores son negativos, algo muy fácil de entender si recordamos que $\mu = 0$.
En la práctica, es mucho más probable, que la distribuciones tengan medias diferentes de cero y varianza distintas a 1 y esto me limitaría el uso de las tablas. Para resolver este problema se transforman los valores de una distribución cualquiera a normal estándar mediante la expresión:

$$Z = \frac{x - \mu}{\sigma}$$

Es bueno aclarar que existen otras formas de representar el valor del área que delimita Z y UD. puede encontrarse en otras publicaciones otras tablas de la normal estándar, digamos que con otro diseño, pero por supuesto en esencia son las mismas.
Veamos algunos ejemplos que nos ayudaran a entender el manejo e interpretación de las áreas bajo la curva.

1- Cuál es el área bajo la curva normal estándar:
   a) por encima de 1.00
   b) por debajo de –1.00
   c) entre 1y –1

Repuestas:

a) Localizamos Z = 1 en la tabla de una cola y el área que delimita es de 0.3413, pero me están pidiendo el área por encima de Z =1.
   Como la mitad del área es 0.50, le resto la que delimita z = 1 y me queda el resto que es precisamente el área por encima de 1 que es la que me están pidiendo.

0.5-0.3413 = 0.1587
La probabilidad que la normal estándar tome valores por encima de 1 es de un 15,87

b) El valor es el mismo porque la curva es simétrica.
c) Si Z = 1 delimita un área de 0.3414
Y Z = -1 delimita la misma área pero al izquierda de la curva , entre –1 y 1 se obtiene por la suma de las dos áreas:
0.3413 +0.3413= 0.6824
Por consiguiente la probabilidad de que la normal estándar tome valores entre –1 y 1 es de un 68%. Y hemos corroborado una de las propiedades de la normal cuando decíamos que la media mas y menos una desviación estándar delimitaban un área del 68 %

$\mu \pm \sigma$ = 68% del área.
En este caso como μ = 0 y σ = 1
0 $\mp$ 1= -1y 1 que es el área que acabamos de encontrar

2-En la curva normal estándar que valor de Z intercepta o delimita:
a) un 10% a la derecha
b) un 5 % a la izquierda

Respuestas:

Ahora buscamos en el cuerpo de la tabla el área que nos piden y vemos que valor de Z delimita esa área:
a) 1.28
b) –1.65

3. Supongamos que los niveles de glucosa en sangre de diabéticos en ayuna se distribuye de forma normal con μ = 105 y σ = 9
    a) Qué proporción de diabéticos presentan niveles de glicemia entre 90 y 125
    b) Qué niveles de glucosa abarcan el 95% de los diabéticos

Respuestas:
a) Lo primero que tenemos que hacer es transformar los valores de la variable, glicemia en sangre, en normal estándar.
Vamos a denominar por **z** los valores de la estándar y por **x** los de la variable en estudio.

Z = $\frac{x-\mu}{\sigma}$
Para x_1 = 90
Z = $\frac{90-105}{9}$ = - 1.67

Z_1 = - 1.67

Para x_2 = 125

Z = $\frac{125-105}{9}$ = 2.22

Z_2 = 2.22

Ahora buscamos en la tabla:

Z = -1.67= 0.4525

Z = 2.22 = 0.4865

0.4525+0.4865= 0.9390

El 94% de los diabéticos presentan niveles de glicemia en ayunas entre 90 y 125.

b) Ahora buscamos el área que nos piden, es de 0.95, como la tabla viene tabulada para la mitad de la curva y la curva es simétrica, dividimos 0.95/2

0.95/2 = 0.475

Buscamos en la tabla que valor de Z delimita un área de 0.475 y es 1.96

Con signo positivo y negativo, a la derecha y a la izquierda, $\pm$ 1.96

Como lo que quiero saber es valor de la variable **x** tengo que estandarizar y despejar en la fórmula:

Z = $\frac{x-\mu}{\sigma}$

Sustituyendo:

Para Z =- 1.96

-1.96 = $\frac{x-105}{9}$

despejando x

X_1= (-1.96 x 9) +105

X_1= 87.36

Para Z = 1.96

X_2 = (1.96x9) +105

X_2 = 122.6

El 95 % de los diabéticos tiene niveles de glucosa entre 87.4 y 122.6 mg/l

En este último ejemplo donde ya se ve una aplicación concreta de la normal hemos asumido que conocemos la media y la desviación estándar poblacional, algo que no ocurre en la realidad.

Por eso aunque ya tenemos modelos teóricos, las distribuciones teóricas de frecuencia o de probabilidad, que nos van a permitir cuantificar la incertidumbre al hacer inferencias a la población, todavía no estamos preparados para dar este paso, falta un elemento vital, un concepto trascendental que es el concepto de:

***Distribución muestral de un estimador**.*

Antes de definir este concepto debemos aclarar que se entiende por un ***estimador**.*

Estimador**. Función particular de las observaciones en una muestra escogida para estimar el parámetro de una población. La media de la muestra es utilizada para estimar la media de la población(parámetro poblacional), el valor numérico obtenido se llama **estimación.

El poder utilizar esas medidas de tendencia central, como la media muestral, u otras como la proporción, la varianza y la desviación estándar, por sólo mencionar las más utilizadas, como estimadores de los parámetros poblacionales tiene su sustento teórico en el concepto de ***Distribución muestral de un estimador.***

Distribución muestral de un estimador. *Dada una muestra aleatoria de una población, se llama distribución muestral de un estimador a la distribución de frecuencias de los valores que ese estimador toma en un número infinito de muestras del mismo tipo y tamaño.*
Aclaremos un poco este concepto con un ejemplo; supongamos que tenemos una población de niños menores de un año y comenzamos a sacar muestras de tamaño 10 y calculamos el peso promedio de los niños para cada muestra.
Debemos aclarar que se trata de un muestreo con reemplazo, o sea, cada vez que saco al azar una muestra de 10 niños y le calculo su peso promedio, esos niños los devuelvo a la población antes de obtener las otras muestras.
De tal suerte que tendremos un listado de pesos pro medios, tantos como muestras haya obtenido.
Esas medias yo puedo tratarlas como si fueran valores de una variable, puedo ordenarlas, calcular su distribución de frecuencias absolutas y relativas, calcular la media de esas medias y una medida de dispersión, digamos la desviación estándar.
Se ha comprobado que esa distribución de las medias va a tener tres propiedades.

Propiedades de la distribución de las medias en el muestreo.

1- La media de la distribución de las medias en el muestreo es igual a la media poblacional μ
2- La forma de la distribución de las medias en el muestreo es aproximadamente el de las distribución normal, independiente de la forma de distribución de la población, suponiendo que **n** sea suficientemente grande (Teorema central del límite).

De ahí la importancia tan grande que tiene la distribución normal en la estadística inferencial, porque puede que no sepa cual es la distribución de mi población o que incluso la población no se distribuya normal, que si el tamaño de la muestra es suficientemente grande, por esta segunda propiedad que se conoce como teorema central del límite, la muestra si sigue esta distribución.

3- La desviación estándar de la distribución de las medias en el muestreo es $\frac{\sigma}{\sqrt{n}}$ y se conoce como error estándar.

Ya tenemos todos los elementos teóricos necesarios para trabajar en los dos campos de acción de la inferencia estadística, en la **estimación** de parámetros de una población a partir de una muestra y en las **pruebas de hipótesis**.

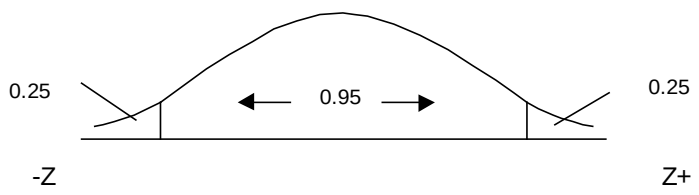
ESTIMACIÓN

Estimación por intervalo para la media poblacional.

Introducción:

En muchos estudios en el campo de la Salud Pública ocurre con cierta frecuencia qu Z = $\frac{x-\mu}{\sigma}$ e se quiere conocer a partir de una muestra el valor de determinado parámetro poblacional, por ejemplo conocer a partir de una muestra de mujeres que residen en un área de salud la proporción que usan diafragma como método anticonceptivo, o saber partir de una muestra de mujeres, la edad promedia de las primera relaciones sexuales en un municipio.
En este ultimo ejemplo se quiere"conocer", se quiere **estimar** la media poblacional(μ) a partir del calculo de la media muestral ($\bar{x}$). Esta sería una estimación puntual de la media poblacional, estoy dando un único valor, pero no tenemos la certeza de que ese valor de la media de la muestra sea el de la población.
Si recordamos la primera propiedad de la distribución de la media en el muestreo, planteaba que la media de la distribución de las medias en el muestreo era igual a la media poblacional, pero en la práctica no se trabaja con todas las posibles muestras de la población, si no , con una sola muestra y por lo tanto no es aconsejable dar una estimación puntual.
Una estimación mas precisa y segura sería dar entre que par de valores se encuentra la media de la población (u otro parámetro que se quiera estimar) con una probabilidad dada.
Si recordamos la distribución normal estándar se puede determinar entre que dos valores de Z se encuentra el 95% del área, y esos valores son –1.96 y 1.96

.



La expresión matemática es:

Pr = $\{-1.96 \leq Z \leq 1.96\}$= 0.95
Que equivale a decir que el 95% de la curva estándar está comprendida entre 1.96y -1.96 sobre la abscisa.
Como en la práctica es poco probable trabajar con poblaciones que la media sea 1 y la varianza 1 vamos a sustituir la Z la expresión matemática para la tipificación: z = $\frac{X-\mu}{\sigma}$, pero como vamos a trabajar con una muestra la expresión queda como:

Z = $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$
Recordemos que ahora en el denominador no es σ, sino el error estándar y en el numerador no es X_i si no la media de la muestra $\bar{x}$.
Sustituyendo tenemos:
Pr = $\left\{-1.96 \leqq \frac{\bar{x}-\mu}{\sigma/n} \leq 1.96\right\}$= 0.95
Como mi parámetro desconocido es μ, lo despejamos de la desigualdad con simples manipulaciones matemáticas y tendremos la siguiente expresión:
Pr = $\left\{\bar{x}-1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leqq \bar{x}+1.96\frac{\sigma}{\sqrt{n}}\right\}$= 0.95
Por tanto la media poblacional se va a encontrar entre los dos valores que forman las expresiones $\bar{x}$ -1.96$\frac{\sigma}{\sqrt{n}}$ y $\bar{x}$ +1.96$\frac{\sigma}{\sqrt{n}}$ con un 95 % de confianza, y de esta forma hemos obtenido un estimación por intervalo para la media poblacional o como también se le conoce un **intervalo de confianza para la media poblacional μ.**

Si se desea tener un intervalo con una certeza mayor, digamos al 99% de confianza trabajaría con los percentiles de la normal que me delimitan un área del 99%, en este caso –2.58 y2.58 y los intervalos serían:

Límite inferior: $\bar{x} - 2.58\frac{\sigma}{\sqrt{n}}$

Límite superior: $\bar{x}$ +2.58$\frac{\sigma}{\sqrt{n}}$

Veamos un ejemplo:

Supongamos que en una muestra de 25 mujeres del municipio Plaza la media de las primeras relaciones sexuales fue de 18.5 años con una varianza poblacional de 2.7años. La estimación por intervalo para la media de la población del municipio al 95% seria:

Límite superior: 18.5+(1.96) (2.7/√25) = 19.55

Límite inferior: 18.5-(1.96) (2.7/√25)= 17.44

Por tanto con una confiabilidad de un 95% la edad media de las primeras relaciones sexuales en el municipio se encuentra entre 17.44 y 19.55 años. Existen otras formas de interpretar estos resultados, ahora bien, lo que debe quedar bien claro que esos "límites de confianza" no abarcan el 95% de la población, error que se comete con frecuencia y constituye un hecho grave, por ejemplo sería lamentable decir que el 95% de las mujeres del municipio tuvieron sus primeras relaciones sexuales entre los 19.55y017.44 años. **El intervalo sólo se relaciona con la media poblacional,** o con el parámetro que se esté estimando.

Estimación por intervalo para la media poblacional con varianza poblacional desconocida.

Si observamos detenidamente la expresión con que realizamos las estimación de los intervalos de confianza para la media, vemos que aparece σ que es la desviación estándar poblacional, supuestamente conocida. ¿ Qué hacer entonces en el caso más frecuente en la práctica cuando σ es desconocida?

La varianza poblacional se estima con la varianza muestral, s^2, y se sustituye en la expresión.

Pero ocurre que σ es un parámetro poblacional y tiene también su distribución muestral y por consiguiente tiene su variación en el muestreo, su error estándar y esto hace que la distribución teórica de probabilidades que sustenta esta estimación ya no la " normal" sino otra distribución teórica de frecuencias conocida como **t de student** o simplemente la **distribución t,** con n-1 grados de libertad.

La distribución t, al igual que la normal es una distribución teórica de probabilidades que también tiene una función matemática que la define y aparecen tablas con las tabulaciones del área bajo la curva.

La curva es simétrica, acampanada, muy parecida a la curva de la distribución normal, pero más dispersa. en ella tenemos que tener en cuenta un nuevo

concepto, el de **grados de libertad** (gl) que también aparece en otras distribuciones teóricas de frecuencia o de probabilidades.
Para diferentes grados de libertad también existen diferentes distribuciones t, por eso las tablas vienen con las áreas bajo la curva para valores muy concretos, los más utilizados, 95%, 99% , etc. Y hasta 30 gl casi simple.
¿Por qué hasta 30 gl?. Cuando el tamaño de la muestra es suficientemente grande (teorema central del límite) la estimación de σ a partir de la desviación estándar muestral (S) es "tan buena" , es lo suficientemente fiable y se puede trabajar con la aproximación a la normal, por tanto se sustituye en la expresión matemática de estimación de los intervalos σ por S y se sigue distribuyendo normal si **n** es grande
Si **n,** el tamaño de la muestra, es pequeña la estimación de σ no es tan fiable y no se puede trabajar con la distribución normal sino con la distribución t
¿Cuán grande tiene que ser entonces el tamaño de la muestra? En la práctica, cuando el tamaño de muestra garantiza que los gl sean $\geq$ a a 30 (tamaño de muestra > de 30) los valores de la t y la normal se aproximan mucho y entonces se puede obviar el uso de la t de student y utilizar la normal.
Resumiendo:

Situación	Distribución teórica de probabilidades	Límites de confianza
Varianza poblacional conocida	NORMAL	$\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ *
Varianza poblacional desconocida y n $\leq$ 30	T DE STUDENT	$\bar{x} \pm t_{\frac{\alpha}{2}}$ n-1gl $\frac{s}{\sqrt{n}}$ **
Varianza poblacional desconocida y n > 30	NORMAL	$\bar{x} \pm z_{\frac{\alpha}{2}} s/\sqrt{n}$ *

*Percentil de la normal
** percentil de la t

Veamos un ejemplo con σ desconocida y **n** pequeña.
Supongamos que en el mismo estudio sobre la edad media de las primeras relaciones sexuales, la varianza poblacional fue desconocida y con el tamaño de muestra de 25 no se puede utilizar la aproximación a la normal.
Supongamos que la estimación de σ fue S= 3.2

Datos:
n=25
gl= 25-1=24

α = 0.05

$t\alpha_{24gl}$ = 2.064

Estadígrafo:

$$\bar{X} \pm t\alpha_{n-1gl} \frac{s}{\sqrt{n}}$$

Sustituyendo:

$$18.5 \pm (2.064)\left(\frac{3.2}{\sqrt{25}}\right)$$

18.5±1.32

Límite inferior = 18.5-1.32= **17.88**

Límite superior = 18.5+ 1.312 = **19.82**

La edad promedio de las primeras relaciones sexuales de las mujeres del municipio Plaza se encuentra entre 17.9 y 19.9 años con una confiabilidad del 95% .

Intervalo de confianza para las proporciones.

La distribución **binomial**, otra distribución teórica de frecuencia o de probabilidades, es el sustrato teórico sobre el que se desarrolla la inferencia sobre proporciones, pero el igual que pasaba con la media poblacional, cuando el tamaño de muestra es “suficientemente grande” se puede utilizar la aproximación a la normal. Nos limitaremos al estudio de las “muestras grandes” porque el estudio de las “ muestras pequeñas” rebasa los objetivos de este curso.

Colton plantea que un tamaño de muestra que nos permite trabajar con la aproximación a la normal es el que cumpla que:

$np \geq 5$ y $n(1-p) \geq 5$

Donde:

n= tamaño de la muestra

p= proporción

La expresión de cálculo sería:

$$P \pm z\left(\sqrt{pq/n}\right)$$

P = proporción muestral

q = 1-p

A modo de ejemplo:

Supongamos que se estudio una muestra de 60 mujeres en edad fértil de un área de salud y que 10 refirieron utilizar el diafragma como método anticonceptivo. ¿ Cuál será la proporción de mujeres del área que hacen uso de este método anticonceptivo?

Datos:
n = 60
p = 10/60 = 0.17

Vamos a probar si n es lo suficientemente grande para usar la normal
np = 60 x 0.17 = 10.2
nq = 60x 0.83 = 49.8

Se cumple que es mayor de 5
Sustituyendo en la expresión:

0.17$\pm$ 1.96 ($\sqrt{0.17x0.83/60}$)
0.17 $\pm$ 1.96 $\sqrt{0.1411/60}$
0.17$\pm$ 1.96x 0.048
0.17$\pm$0.09

Límite inferior: 0.08
Límite superior: 0.26
El porcentaje de mujeres del área que usan el diafragma como método anticonceptivo está entre el 8 y el 26% con una confiabilidad del 95%.

Ejercicios resueltos

1. En un estudio sobre conocimiento y uso de métodos anticonceptivos realizado en un área de salud, se estudio una muestra de 25 mujeres en edad fértil. Se encontró que la edad media de las primeras relaciones sexuales fue 18.5 años con una D.E. de 3.2. Estime con una confiabilidad del 95 % el intervalo de confianza para la media.

Respuesta.
Datos
α = 0.05
$\overline{x}$ = 18.5
s = 3.2
n = 25

Recordando la expresión:

$\overline{x} \pm$ t. s / $\sqrt{n}$
Buscamos en la tabla de la distribución t, con 24 gl y α = 0.025 (el error se reparte en las dos colas) y el valor es de 2.064
Sustituyendo:

18.5 $\pm$ 2.064 (3.2 / $\sqrt{25}$)
18.5 $\pm$ 1.32
Límite inferior 17.18
Límite superior 19.82

La edad promedio de las primeras relaciones sexuales se encuentra entre 17.2 y 19.8 años con una confiabilidad del 95 %

2. En un pesquisaje de estomatitis subprótesis, en un grupo de portadores de prótesis estomatológicas se seleccionó una muestra de 50 portadores y se encontró que 25 tenían la enfermedad.
Estime el intervalo de confianza al 99 % para la proporción de enfermos entre los portadores de prótesis.

Respuesta.

Datos.
α = 0.01
n = 50
p = 25 / 50 = 0.5
q = 1- 0.5
q = 0.5
z = 2.58 (el error se reparte en las dos colas y el percentil de z que limita un área de 0.495 es 2.58)
Recordemos la expresión:

$p \pm z. \sqrt{pq/n}$

Sustituyendo:

$0.50 \pm 2.58 \times \sqrt{0.5x0.5/50}$

$0.50 \pm 2.58 \times 0.07$
0.50 ± 0.18
Límite inferior 0.32
Límite superior 0.68

La proporción de enfermos se encuentra entre el 32 % y el 68 % con una confiabilidad del 99%.

PRUEBA DE HIPÓTESIS.

En el capitulo anterior se abordó uno de los grandes campos de trabajo de la inferencia estadística, la estimación de parámetros poblacionales, el otro campo de acción de la estadística inferencial es la **verificación de hipótesis estadística,** también conocida como **prueba, dócima o contraste de hipótesis.**
Vamos a analizar algunos conceptos básicos comunes todas las pruebas y esenciales para su comprensión, y posteriormente veremos pruebas particulares para problemas concretos.

Definiciones Básicas

-Hipótesis.

Supongamos que se conoce por estudios anteriores que la prevalencia del habito de fumar en la población de un área de salud es del 45% y se realiza una intervención para tratar de reducir este nocivo habito. Después de realizada la misma se estudia una muestra de esa población y constató que la prevalencia de fumadores era del 30%.

¿ Puedo por estos resultados asegurar que la reducción del porcentaje de fumadores se debió al impacto de la intervención?. ¿Será esa diferencia entre esos dos porcentajes producto del azar, de la casualidad, del propio hecho de trabajar con una muestra?. ¿ Son realmente iguales o diferentes esas proporciones?. ¿ Es el por ciento de fumadores menor después que se realizó la intervención?.

Esas interrogantes, esas preguntas las vamos a formular en forma de **hipótesis estadística.** Una hipótesis no es mas que una afirmación sobre el comportamiento poblacional que se asume para esa variable.

La primera hipótesis que se plantea es la llamada **hipótesis nula**, y se designa con $\mathbf{H}_0$, algunos autores la designan con el nombre de hipótesis de no diferencia. En esta hipótesis se expresa la no existencia de diferencia entre los resultados obtenidos en la práctica y los resultados teóricos. Dicho de otra forma, en la hipótesis nula se declara lo opuesto a lo que estamos intentando probar, lo opuesto a lo que el investigador quiere probar. En general esta hipótesis se establece con el propósito de ser **rechazada.**

El por qué de esta condición la explicaremos mas adelante. Volviendo al ejemplo la hipótesis nula sería

$H_0: p_1 = p_2$ (La prevalencia de fumadores es la misma antes y después de la intervención)

La hipótesis nula da lugar a una segunda hipótesis, la **hipótesis alternativa**, que se designa como $\mathbf{H}_1$ y en ella se expresa lo contrario a lo que se planteo en la hipótesis nula, lo que el investigador quiere probar.

En dependencia de cómo se plantee la hipótesis alternativa se derivan de ella tipos de pruebas de hipótesis.

Siguiendo con el ejemplo.

- ✓ Que simplemente difieren los porcentajes de fumadores, sin precisar el sentido de esa diferencia.

  $H_1 = p_1 \neq p_2$

  En este caso se dice que la prueba es de **dos colas o contraste bilateral**, pues la diferencia puede darse en cualquier sentido, que difieran por exceso o por defecto, $p_1 > p_2$,o, $p_1 < p_2$

- ✓ Darle sentido a la desigualdad o diferencia.
  $H_1 = p_1 > p_2$ (Sería el caso del ejemplo)
  $H_1 = p_1 < p_2$
  En este caso se diece que la prueba es de **una cola o contraste unilateral.**

-Errores

Al tomar la decisión de rechazar o no las hipótesis podemos equivocarnos en nuestra decisión, podemos cometer un **error.** La decisión siempre se toma en función de la hipótesis nula, sobre esta hipótesis y se pueden cometer dos tipos de errores:
1er Error. **Error de tipo I o α**
Rechazar la hipótesis nula, siendo cierta

2do Error. **Error de tipo II o β**
Aceptar la hipótesis nula, cuando esta es falsa

-***Nivel de significación***

Siempre al tomar una decisión estoy expuesto a cometer un error, por eso es muy beneficioso tener una **cuantificación de cuán buena o no ha sido mi decisión**, medir en términos de probabilidades si mi decisión ha sido o no acertada y tener alguna medida de la confianza de mis decisiones.
El error que se mide con más frecuencia es el de tipo I o α, donde se fija una probabilidad pequeña, como es lógico, de cometer este tipo de error, de equivocarme en mi decisión. Por convención se fija una probabilidad de 0.05 o un 5%, o de 0.01 o un 1% yb se acostumbra a denotar esta probabilidad por **α,** así tendríamos α= 0.05 ,o, α= 0.01
La especificidad de una probabilidad pequeña designada por **α** de cometer el error de tipo I, es lo que se conoce como **nivel de significación de la prueba.**
Ahora resulta fácil entender porque la hipótesis nula se expresa en términos de lo que ``esperamos rechazar´´, lo contrario a lo esperado por el investigador, pues el error que fijamos con una probabilidad pequeña de equivocarnos es el de rechazarla siendo cierta, y desde el momento mismo en que la formulamos se hace en función de lo que **no se espera que ocurra.**
En el ejemplo de la intervención para reducir la prevalencia del habito de fumar, la hipótesis nula se formuló en términos de igualdad en los proporciones de fumadores, cuando lo que el investigador espera es que sea menor la prevalencia de los que fuman después de la intervención.

El rechazo de la Hipótesis Nula equivale a la aceptación de la Hipótesis Alternativa, si en el ejemplo que nos ocupa rechazamos la igualdad de prevalencia de fumadores, aceptamos la alternativa de que la prevalencia disminuyó después de aplicado el plan de intervención.
El error de tipo II o β también puede ser `` medido´´, pero en la práctica su uso se limita a casos muy especiales. Es por eso que al no tener en valor de probabilidad fijado de cometer este tipo de error- aceptar la hipótesis nula cuando es falsa- trato de no cometer este error al realizar la prueba y por eso al no poder rechazar la hipótesis nula nunca digo que **la acepto**, si no que no puedo rechazarla. Esta forma de expresar el no rechazo de H_0 es denominada por algunos autores como **reservar el juicio** y simplemente lo que se trata es de no cometer el error de tipo II al utilizar la palabra **acepto**.

-Estadígrafo o estadístico de prueba.
Para realizar una prueba de hipótesis hay que tener en cuenta algunos aspectos de diseño de la investigación como es la naturaleza de las variables, en que escala están medidas, las características de la muestra, y el cumplimiento de algunos supuestos pre establecidos para decidir que tipo de prueba se va a utilizar.
Siempre existe para cualquier tipo de prueba un estadístico o estadígrafo (expresión o formula matemática) que se calcula con los datos de la muestra. Este estadígrafa bajo el supuesto de que H_0 sea cierta sigue una determinada distribución teórica de frecuencia, distribución que puede variar según el tipo de prueba, y que en ocasiones le da el nombre a la prueba estadística.

-**Regla de decisión.**
La distribución teórica de frecuencia o de probabilidad que caracteriza a cada estadígrafo y el nivel de significación que se fije para realizar la prueba, son los elementos esenciales que van a influir en la decisión que se tome en cuento al rechazo o no de la hipótesis nula.
Generalmente cuando se produce el rechazo de H_0 y por ende la aceptación de la alternativa se dice que la prueba fue **significativa** a un 5% o un 1 % en dependencia del nivel de significación con que se halla trabajado la prueba**.** Este aspecto se explicará con mas detalle, en aras de facilitar su comprensión, cuando desarrollemos algunas pruebas de hipótesis especificas.

Antes creemos pertinente realizar algunas observaciones sobre el termino **estadísticamente significativo,** que con frecuencia se confunde con el significado corriente de la palabra **significativo,** y se hace sinónimo el resultado de una prueba, al de relevante, importante desde el punto de vista del marco teórico de la ciencia en la cual se está investigando. Por tanto se recomienda usar en la literatura científica la palabra significativo solo en caso de referirse al resultado de una prueba estadística y no al discurso en general.
Otro aspecto que vale la pena aclarar es que el resultado de una prueba de hipótesis no puede analizarse al margen del marco teórico de la ciencia

particular en que esta ha sido usada. El resultado de la prueba estadística sólo es una parte de la evidencia que influye en la decisión del investigador. La decisión estadística no debe considerarse como definitiva, si no que es un elemento mas a considerar junto con el análisis de toda la información científica que existe sobre el problema que se investiga.

Prueba de hipótesis para la comparación de dos medias. Muestras independientes.

En muchas ocasiones nos encontramos ante la situación de decidir si la diferencia entre dos medias muestrales son debida a la causalidad, al hecho de trabajar con muestras y no con las poblaciones o si realmente ese par de medias difieren.
Vamos a partir de algunos supuestos básicos:

- Que las dos muestras son independientes, incluso pueden ser de diferentes tamaños.
- Que las varianzas de las dos muestras son iguales (Homocedasticidad de la varianza)

Pueden darse tres situaciones diferentes a partir del cumplimiento de estos upuestos.

1. Que las poblaciones de donde se obtuvieron las muestras este normalmente distribuidas t además que las varianzas sean conocidas. En este caso se trabaja con la dostribución normal y el estadigrafo de prueba sería:

   $$Z= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}}$$

   Si H_0 es cierta Z~ (0,1), con $\mu_1 - \mu_2 = 0$ y $\sigma_1 - \sigma_2 = 1$

Esta situación es poco frecuente en la práctica. El cumplimiento del cumplimiento de normalidad para las poblaciones de donde se obtuvieron las muestras es posible conocerlo o probarlo con otras pruebas estadísticas (pruebas de bondad y ajuste) pero lo que en la práctica resulta muy poco frecuente es que las varianzas poblacionales sean conocidas e iguales por lo tanto no nos detendremos detallar este caso

2. Que las poblaciones de donde se obtuvieron las muestras estén normalmente distribuidas, pero las varianzas sean desconocidas.

En este caso sería necesario estimar las varianzas poblacionales a partir de las estimaciones de las varianzas muestrales, para a partir de ellas obtener la estimación del error estándar poblacional.

EE= $\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}$

Como los tamaños de muestras pueden diferir lo lógico sería ponderar las estimaciones de las varianzas poblacionales. Por tanto la estimación de la desviación estándar mancomunada para las dos muestras y ponderada sería:

D:E $_p$ = $\sqrt{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 / n_1 + n_2 - 2}$

Pero como hay que trabajar con el error estándar hay que dividir la DE entre la √ de los tamaños muestrales. Al trabajar con las estimaciones muestrales de las varianzas esto hace que el estadígrafo ya no se distribuya Normal, si no t de student con $n_1 + n_2 - 2$ g.l

Pero como las poblaciones de donde se obtuvieron las muestras se distribuyen normal, se asume que la estimación de las varianzas es una buena estimación y se puede trabajar con la aproximación a la normal estándar.

3. Que las poblaciones de donde se obtuvieron las muestras no se tenga la certeza que se distribuyen normal y las varianzas poblacionales son desconocidas.

Este es el caso que se da con más frecuencia en la práctica. Se utiliza el mismo estadígrafo de prueba que en el caso anterior.

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(n_1 - 1)s_1^2 + (n_2 - 2)s_2^2 / n_1 + n_2 - 2}.\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1 + n_2 - 2\ gl}$$

Ahora bien, en función del tamaño de la muestra y por ende de los grados de libertad pueden darse dos situaciones:

a) Por las propiedades de la distribución muestral de las medias, si **n** era suficientemente grande, se podía trabajar con la aproximación a la distribución normal (teorema central del límite). De ahí, que si el tamaño de la muestra es lo suficientemente grande como para garantizar que los gl, sean $\geq$**30 se puede** trabajar con la **aproximación a la normal** para obtener el valor crítico y aplicar la regla de decisión.

b) Si el tamaño de la muestra hace que los gl sean**< 30**, entonces se utiliza la **distribución t de student** que sigue el estadígrafo de prueba con $\mathbf{n_1 + n_2 - 2}$ **gl.** Veamos un ejemplo:

Para probar el efecto que tiene el entrenamiento físico sobre la capacidad física de un individuo, uno de los parámetros fisiológicos que se miden es la frecuencia cardiaca en reposo (F.C),Se sabe por estudios realizados que un individuo bien entrenado tiene una F.C es menor que la de una persona no, o mal entrenada. Para comp.`robar al efectividad de un nuevo tipo de entrenamiento se selecciono una muestra de 60 hombres con edades comprendidas entre 20 y 40 años, y se sometieron al entrenamiento por espacio de dos meses, al finalizar se midio la F.C. en reposo de cada uno de los integrantes del grupo y se calculo la media: 68 latidos por minuto y la D.E. de ± 6
Se seleccionó un grupo control de 70 hombres comprendido entre las mismas edades, lo que nunca había recibido entrenamiento. La frecuencia cardiaca promedio en reposo fue de 84 latidos por minuto y la D.E. de ± 8

Datos:

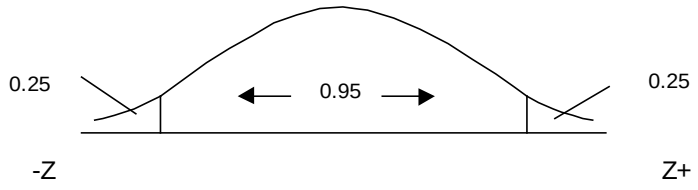
$n_1 = 60$	$n_2 = 70$
$\bar{x}_1 = 68$	$\bar{x}_2 = 84$
DE= 6	DE= 8

Hipótesis:

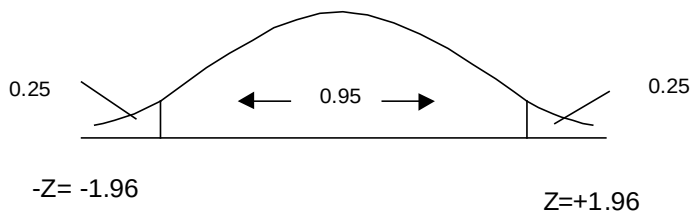
$H_0 = \bar{x}_1 = \bar{x}_2$
Aunque en la práctica solo se trabaja con una alternativa, vamos a analizar dos alternativas, que dan origen a pruebas de una y dos colas para analizar como se procede en cada caso.
1-Prueba de dos colas o contraste bilateral.

$H_1 = \bar{x}_1 \neq \bar{x}_2$
Esta alternativa responde a una prueba de dos colas o contraste bilateral.
Los grados de libertad serian: (60+70)-2= 128 , es mayor de 30 se puede trabajar con la distribución normal estándar.
Fijamos el nivel de significación o el error de tipo I que estoy dispuesta a cometer: α = 0.05
En esta alternativa solo estamos planteando que las medias difieren, no en el sentido que se da esa diferencia, por consiguiente puede que esa diferencia sea por exceso o por defecto, por consiguiente el contraste es bilateral y el α tengo que “repartirlo” en las dos colas de la distribución.

Por lo tanto $\frac{\alpha}{2} = \frac{0.05}{2} = 0.25$



El área que queda delimitada en el centro de la curva es de 0.95 y el percentil de la normal que delimita esa área es Z = 1.96



Regla de decisión: Por consiguiente si H_0 es cierta, el estadígrafo de prueba calculado tendrá que caer dentro de los límites de – 1.96 a 1.96, zona de aceptación o no rechazo de la hipótesis nula. Si el valor calculado del estadígrafo de prueba cae en las "colas", fuera del intervalo –1.96 a 1.96, estamos en la región de rechazo de la hipótesis nula, pues se corresponde con el valor del error de tipo I que estoy dispuesto a cometer.
De forma más pragmática también podemos decir, que si el valor modular del estadígrafo de prueba es mayor que el valor de la tabla de la distribución normal estándar, se rechaza H_0

$|Z|_{calaculada} > |Z|_{tabulada}$ = Rechazo de H_0

Cálculo del estadígrafo de prueba:

$$Z=\frac{68-64}{\sqrt{\frac{(59x36)+(69x64)}{60+70-2}}\sqrt{\frac{1}{60}+\frac{1}{70}}} = -12.72$$

Como –12.72 es menor que –1.96 cae en el gráfico anterior en la cola de la izquierda que es zona de rechazo.

También podemos apreciar que valor modular de Z calculado es 12,72 y es mayor que el valor modular del valor de la Z tabulada, 1.96

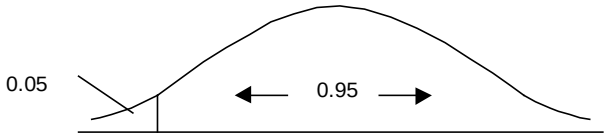
12.72> 1.96

Por consiguiente se produce el rechazo de la hipótesis nula y podemos plantear que las medias difieren. Los promedios de FC en reposo no son iguales para los pacientes con y sin entrenamiento.

2- Prueba de una cola o contraste unilateral.
La hipótesis alternativa sería:

$H_1 = \bar{x}_1 < \bar{x}_2$

Ahora el error de tipoI, el α, el nivel de significación de la prueba está concentrado en una sola cola, como se plantea que la media de la primera muestra es menor que la de la segunda, el error se concentra a la izquierda (si el sentido ce la desigualdad hubiese sido > que ... , el error se concentraba a la derecha).



-Z= -1.675

Busco en la tabla qué percentil de la normal delimita a la izquierda un área de 0.95, y es Z= - 1.675
Ya el estadígrafo de prueba lo tengo calculado, lo único que cambia es la regla de decisión:
Como –11.67 es menor que –1.675 cae en la cola de la izquierda que es la zona de rechazo.
Además también es fácilmente apreciable que:
11.76> 1.675
Por tanto se rechaza la hipótesis nula de igualdad de media y se acepta la alternativa y podemos decir que la frecuencia cardiaca en reposo de los pacientes entrenados es menor que la de los pacientes que no recibieron

entrenamiento, con una probabilidad de equivocarme en esta afirmación de un 5 %), por lo que podemos pensar que el entrenamiento fue efectivo.
Antes de concluir con esta prueba queremos hacer dos aclaraciones que nos parecen oportunas y que son validas para cualquier prueba de hipótesis
La primera es reiterar que en la práctica se trabaja con una sola alternativa y la segunda, que los paquetes de programas para trabajar con microprocesadores, calculan la probabilidad exacta de cometer el error de tipo I al rechazar la hipótesis nula, por que trabajan con la función de densidad de las distribuciones teóricas de frecuencias que siguen los estadígrafos de prueba.
Por consiguiente la regla de decisión se limitara a analizar si el valor de probabilidad que aparece en las salidas de los programas es menor que el valor crítico admisible se 0.05, rechazamos H_0 y decimos que la prueba fue significativa con una $p < 0.05$.

ANÁLISIS DE VARIANZA. (ANOVA)

Introducción.

En el capitulo anterior de prueba de hipótesis se analizó la comparación de dos medias, muestras independientes. Pero en muchas ocasiones nos enfrentamos al problema de querer saber si más de 2 medias difieren o son iguales.
Ante esta situación podemos pensar que la solución sería realizar varias veces, tantas como sean necesarias, dócimas de hipótesis de comparación de dos medias.
Por ejemplo, como parte de un estudio de calidad de la atención se quiere conocer sí el tiempo de hospitalización de los pacientes sometidos a uma intervención quirurgica menor y sin complicaciones, varía entre 5 cirujanos especialista de primer grado y con los mismos años de experiencia, de un servicio de Cirugía de hospital Clínico Quirúgico. De cada cirujano se tomaron datos de 8 historias clínicas seleccionado de pacientes que ellos habian intervenidos y se calcularon los tiempos promedios de estadía.
Si fuéramos al realizar todas las posibles comparaciones de pares de medias, el número de parejas posibles para este ejemplo sería:

$$\binom{n}{k}$$

Donde:
n= total de muestras o medias a comparar, en nuestro ejemplo 5
k = cuantos elementos forman las combinaciones, en nuestro ejemplo 2 porque son pares de medias lo que se quiere comparar
Sustituyendo tendríamos:

n!/ k! (n–k)! = $\frac{5x4x3x2x1}{(2x1)(3x2x1)}$ = 120/12 = 10 combinaciones de pares de media

No es difícil imaginarse, lo engorroso que sería realizar 10 veces lo prueba de hipótesis, para comparar los 10 pares de medias, pero esto no es lo más importante, la consecuencia más importante y negativa es que al realizar todas esas pruebas aumentaría considerablemente la probabilidad de cometer el error de tipo I, α , rechazar que el par de medias son iguales cuando en realidad es cierta.
Veamos rápidamente porque ocurre esto:
Si para prueba que realizamos fijamos un α = 0.05, la confiabilidad sería 1–α = 0.95. Por la regla de multiplicación de la probabilidad, si las pruebas son independientes la probabilidad de dejar de rechazar una hipótesis de no diferencia en los 10 casos sería: 0.95x0.95x0.95....10 veces, o sea $(0.95)^{10}$ = 0.5987
Por tanto α sería:

1– 0.5987 = 0.4013 y se podría cometer el error de tipo I el 40 % de las veces.

Resulta evidente que cuando queremos comparar más de dos medias esta prueba no soluciona el problema y irgue la necesidad de aplicar otra prueba, otra técnica estadística que solucione estos inconvenientes. Es precisamente el analice de varianza la prueba que nos permite la comparación simultanea de varias medias de manera bastante eficiente y teniendo en cuenta los problemas analizados.

¿ Qué es en síntesis el ANOVA?

El análisis de varianza puede definirse como **la técnica mediante la cual la variación total presente en un conjunto de datos se distribuye en varios componentes. Cada uno de estos componentes está asociado a una fuente específica de variación, de modo que en el análisis es posible averiguar la magnitud de la contribución de cada una de estas fuentes a la variación total.**

Esta técnica fue descollada por Fisher cuyas contribuciones a la estadística cubrieron los años 1912 hasta 1962.

Tipos de ANOVA

En el ANOVA la variable **dependiente o de respuesta** es **cuantitativa** y las variables **independientes, covcariables** o **variables explicativas** son **cualitativas.**

A las variables explicativas algunos autores le llaman **factor,** y a las categorías de la escala **niveles del factor,** otros autores denominan a las variables

independientes o explicativas, **tratamientos,** y a sus categorías**, tipos de tratamiento**. En el ejemplo que veníamos analizando la variable dependiente es el promedio de estadía y la variable explicativa, covariable o variable independiente, los cirujanos, que también se le puede denominar como factor con 5 niveles, 5 cirujanos, o tratamiento con 5 tipos de tratamiento, cada cirujano estudiado se corresponde con un tipo de tratamiento.

La variable de respuesta o dependiente, obviamente será siempre una sola, pero las variables explicativas pueden ser más de una, así en dependencia del número que sean tendremos distintos tipos de ANOVA.

Cuando tenemos una sola variable explicativa estaremos en presencia del **ANOVA de una vía o completamente aleatorizado.**

Si tenemos dos variables explicativas o factores estaríamos en presencia del **ANOVA de dos vías,** y si son mas de 2 el numeró de factores entonces se trata del **ANOVA de más de dos vías.**

En este capitulo nos limitaremos a abordar el ANOVA de una vía.

ANOVA de una vía.

Puede ser de efectos fijos o de efectos aleatorios. En el ANOVA de efectos fijos se trabaja con todos los posibles niveles del tratamiento, en el ejemplo de los cirujanos 5, serían todos los cirujanos con que cuenta el servicio que se quiere evaluar. Si fuera de efectos aleatorios, los 5 cirujanos constituyen una muestra de los cirujanos con que cuenta el servicio y los resultados que obtengo son validos para todos los cirujanos del servicio.

Las muestras pueden ser todas de igual tamaño o de diferentes tamaño. En el ejemplo que estamos desarrollando, vamos a trabajar para simplificar los calculo, con tres cirujanos y para que sea de efectos fijos estos tres especialitas son todos con los que cuenta el servicio. Vamos a trabajar con muestras de igual tamaño, 8 pacientes tomados al azar para cada cirujano.

La información que se obtuvo se muestra a continuación:

Cirujano 1 (días)	Cirujano 2 (días)	Cirujano 3 (días)
4	4	5
5	5	3
5	4	3
4	3	3
6	4	3
6	5	3
4	3	4

5	3	5
Total: 39	31	29
Media: 4.87	3.87	3.62

Modelo estructural

Veamos cual es el modelo matemático que sustenta esta técnica.
Una observación cualquiera del conjunto de datos es denominada por y_{ij}

Donde:

i = 1... k indica el grupo o nivel del factor, dicho de otra forma el número de muestra a la cual pertenece, en nuestro ejemplo i va desde 1 hasta 3
j = 1... n indica el número de orden del elemento dentro de l muestra, en el ejemplo que nos ocupa j va desde 1 hasta 8, porque hay 8 elementos en cada muestra.

Por tanto sería:

Tratamientos o Muestras

1	2	3	K
Y_{11}	Y_{21}	Y_{31}	Y_{k1}
Y_{12}	Y_{22}	Y_{32}	Y_{k2}
Y_{13}	Y_{23}	Y_{33}	Y_{k3}
Y_{1n}	Y_{2n}	Y_{3n}	Y_{kn}
$Y_{1.}$ (total)	$Y_{2.}$	$Y_{3.}$	$Y_{k.}$
$\bar{y}_{1.}$ (media)	$\bar{y}_{2.}$	$\bar{y}_{3.}$	$\bar{y}_{k.}$

Dentro de cada muestra o grupo existe una media μ_i , cda valor que aume la variable en ese grupo o muestra guarda relación con esa media, puede coincidir, ser mayor o menor.
Llamaremos **error** a la cantidad en que cualquier valor difiera de la media μ_i de su grupo, y lo vamos a representar por ε_{ij} . Este "error" se refiere a la **variación no controlada** que existe entré los elementos o unidades de cualquier población o muestra y es provocada por un sin número de factores.

Por tanto:

$Y_{ij} = \mu_{ij} + \varepsilon_{ij}$

Ahora bien, si tenemos en cuenta las k muestras (3 en el ejemplo) podemos tener también una gran media general, total, en conjunto para todas las observaciones, independientemente de que pertenezcan a una u otra muestra, que denominaremos por μ

$$\mu = \frac{\sum \mu_i}{k}$$

Al igual que una observación cuakquiera de una muestra difería de la medias μ_i de su grupo en una cantidad ε_{ij} , una media μ_i de cualquier grupo o muestra también va a diferir de la media general o total μ en una determinada cantidad. Esta cantidad en la cual difiere sé cono ce como **efecto del tratamiento**, y es lógico, mientras más difiera la media de la muestra de la media general, quiere decir que ese nivel del factor tiene un efecto sobre el comportamiento de la variable que hace que se aleje del comportamiento general.
El iesimo efecto del tratamiento o del nivel del factor se denota por:

$$\alpha_i = \mu_i - \mu$$

De donde, despejando μ_i

$$\mu_i = \mu + \alpha_i$$

Si recordamos habíamos dicho que cualquier observación de una muestra estaba definida por:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

Si sustituimos μ_i por su nueva expresión tenemos:

$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ υ modelo estructural o matemático

Para i = 1... k
j= 1... n

De donde una observación cualquiera está definida por la media general (μ), el efecto del tratamiento (α_i) y el error (ε_{ij}) que es el componente aleatorio, la variación no controlada.

Supuestos del modelo

- I**ndependencia** (representatividad). Los k conjuntos de datos constituyen k muestras aleatorias de poblaciones por tanto todas las variables y las observaciones son independientes entre sí.
- **Normalidad.** Cada una de las poblaciones de las cuales provienen las muestras deben estar normalmente distribuidas. Si recordamos el teorema central del límite, esto ocurre siempre que si *n,* el tamaño de la muestra de cada grupo, es mayor de 30 y también si la variable se distribuye normalmente en la población de referencia, aunque *n* sea < de 30. Únicamente no podemos garantizar que la media muestral se distribuya normalmente cuando las muestra sean menores de 30 y la variable dependiente no se distribuye normal en algún nivel del factor. Por eso si el tamaño de la muestra de algún grupo es menor de 30, debemos hacer una prueba para probar la normalidad.
- **Homocedasticidad.** Esta complicada palabra quiere decir que la varianza de las poblaciones de donde proceden las muestras, por consiguiente las varianzas de las muestras o grupos deben ser homogéneas, iguales.

Hipótesis

En la hipótesis nula plantearíamos la igualdad de las medias.

$H_o = \mu_1 = \mu_2 = \ldots \ \mu_k$

Si las medias de todos los grupos son iguales eso quiere decir que no hay efecto de los niveles de los tratamientos, por tanto $\alpha_i = 0$ y tambi´en se podría plantear la hipótesis en función de los α_i

$H_o = \alpha_i = 0$ para toda *i*

La hipótesis alternativa sería:

H_1= no todas las μ_i son iguales
En función de los α_i

H_1= no todos los α_i son iguales a cero

Estimadores

Ya vimos que el ANOVA se define como un proceso mediante el cual al variación total se distribuye en componentes atribuibles a diferentes fuentes.

En este contexto el termino "variación" se refiere a lo que conocemos como el numerador de la varianza, la suma del cuadrado de las desviaciones de las observaciones respecto a su media y se denomina **suma de cuadrados o variabilidad cuadrática.**

Suma de cuadrado total

Es la suma de las desviaciones al cuadrado de cada observación respecto a la media general.

$$SC_{total} = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} = \bar{y}_{..})^2$$

Ahora vamos a ver como se descompone la suma total de cuadrados:

Por medio de algunas operaciones algebraicas sencillas, que aparecen en el Daniels y puede ser consultado si se quiere ganar en precisión, se llega a los siguientes resultados:

$$SC_{total} = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} = \bar{y}_{i})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n} (\bar{y}_{i.} - \bar{y}_{i})^2$$

El primer termino es la **suma de cuadrado dentro de grupo**, la desviación al cuadrado de cada observación del grupo con relación a su media. También se conoce con el nombre **de cuadrado residual o error.**

El segundo termino es **suma de cuadrado entre grupos,** que no es más que la suma de las desviaciones al cuadrado de cada media particular de cada grupo con respecto a la media total.

Resumiendo:

Suma de cuadrado total = suma de cuadrado dentro + suma de cuadrado entre.

A partir de las sumas de cuadrados se pueden estimar las varianzas(σ^2) dividiendo las sumas de cuadrados entre sus respectivos grados de libertad y se denominan en este contexto como **cuadrado medio.**

Suma de cuadrado	Grados de libertad (gl)	Cuadrado Medio
Total	Nk–1 ó N–1	Total / N–1
Dentro de grupo	K(n–1) ó N –K	Dentro / N–K
Entre grupo	K–1	Entre / K–1

¿ Cuál sería entonces el estadígrafo para hacer la prueba de hipótesis?

Bajo la hipótesis nula de igualdad de las medias el cuadrado medio entre y dentro deben ser prácticamente iguales, de no ocurrir así, si hubiese efecto de algún tratamiento y por tanto, alguna de las medias difiere, el cuadrado medio entre grupos sería mayor que el cuadrado medio dentro.
Si hacemos una razón de varianzas tenemos:

R.V = $\frac{CMentre}{CMdentro}$

Bajo H_0, la razón varianza, R.V sería = 1

Si H_0 no se cumple, si las medias difieren y hay efecto de la variable que

constituye el factor o el tratamiento, la R.V sería mayor que 1. ¿ Pero cuanto mayor que 1?
Esta razón de varianzas sigue una distribución teórica de frecuencias o de probabilidades denominada **F** (Fisher)
Por tanto el estadígrafo de prueba sería esa razón de varianzas que bajo H_0 se distribuye:

$$\frac{CMentre}{CMdentro} \sim F_{(K-1)(N-K)gl}$$

Regla de decisión

Si la F calculada > F tabulada, rechazo H_0.

Tabla ANOVA

Para facilitar el análisis se acostumbra a resumir las estimaciones en la llamada **tabla ANOVA** que es la forma en aparecen los resultados en los paquetes de programas como el SPSS y el EPINFO.

Tabla ANOVA una vía, efectos fijos.

Fuente de variación	Suma de cuadrados	GL	Cuadrado medio	Razón de varianzas	Probabilidad
Entre grupos	$\sum_{i=1}^{k} \sum_{j=1}^{n} (\bar{y}_{i.} - \bar{y}_i)^2$	K-1	$\frac{SCE}{K-1}$		
Dentro de grupos	$\sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} = \bar{y}_i)^2$	N-K	$\frac{SCD}{N-K}$	$\frac{CME}{CMD}$	Según GL y nivel de significación de la prueba α

Total	$SC_{total}=$ $\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}=\bar{y}_{..})^2$	N-1	$\frac{SCT}{N-1}$		

A continuación aparece la salida del SPSS en la solución del ejemplo que veníamos desarrollando.

Source	D.F:	Sum. of Squares	Means Squares	F Ratio	Prob.
Between groups	2	7.00	3.5		
Within Groups	21	15.625	0.774	4.704	0.021
Total	23	22.625.			

Observe como el cuadrado medio entre grupo es mayor que el cuadrado medio dentro de grupo, precisamente por que las medias son diferentes y hay efecto de la variable " cirujano" lo que se corrobora cuando vemos que en la última columna aparece bajo el titulo de "Prob". la probabilidad exacta de cometer el error de tipo I o α de 0.021, menor que 0.05 que es el error máximo que habitualmente estamos dispuestos a cometer de rechazar la hipótesis nula siendo cierta.
Por tanto podemos decir, que el tiempo promedio que los pacientes estuvieron hospitalizados es significativamente diferente en dependencia del cirujano que realizó la operación.

Robustez de la F

Se dice que una prueba estadística es robusta cuando es poco sensible al no cumplimiento de los supuestos del modelo.
La prueba F para el ANOVA es robusta porque es poco sensible a los cambios del modelo, no siempre se logra en la práctica que se cumplan todos los supuestos del modelo y esto no afecta grandemente la sensibilidad de la prueba.
Veamos como se comporta la prueba ante el no cumplimiento de estos supuestos.
- **La Independencia**. Este es el supuesto que más afecta la sensibilidad de la prueba, pero por suerte depende del investigador, de que se haga un buen diseño del estudio, que cada muestra sea una muestra aleatoria y cada observación independiente del resto de las observaciones.
Si las observaciones no son independientes el valor del estadígrafo F puede ser seriamente afectado. El tipo de efecto depende de la naturaleza de la dependencia entre las observaciones.

El efecto que produce es una tendencia a hacer más pequeño el cuadrado medio dentro de grupo y al ser más pequeño el denominador del estadígrafo de prueba, CM entre/ CM dentro, me aumente el valor de la F y podemos rechazar

hipótesis nulas que no son falsas y aumentar el error de tipo I. Lo contrario también puede ocurrir.
La única solución a este problema, y por demás nada difícil, esa hacer un buen diseño del estudio para que no se incumpla con la independencia de las observaciones.
-**La normalidad.** El incumplimiento de este supuesto es el que menos afecta la prueba, y dentro de este supuesto el que la distribución sea asimétrica no tiene ningún impacto, si tiene alguna influencia negativa la curtosis.
Si la curtosis es pequeña, si la curva es leptocúrtica, hay poca variabilidad y por tanto el cuadrado medio dentro de grupo sería pequeño, y un denominador pequeño hace que la F se aumente y que se rechace lo que no debe ser echazado.
Si por el contrario la curva es platicurtica, con una curtosis alta, el CMD aumenta, por consiguiente aumenta el denominador y disminuye el valor de la F y puedo dejar de rechazar hipótesis realmente rechazables.

-**La homocedasticidad.** Si no se cumple al homocedasticidad se incrementa la variabilidad entre grupos, no solo por el posible efecto del factor, sino, por las varianzas diferentes entre los grupos, se produce un aumento del numerador del estadígrafo de prueba, se incrementa el valor de la F calculada y no se rechazarían hipótesis que son rechazables.
Si las **n** son grandes e iguales o lo más parecidas posibles, el no cumplimiento de igualdad de varianza afecta poco la sensibilidad del test.

Comparaciones post hoc o comparaciones a posteriori.
Cuando realizamos el ANOVA y rechazamos la hipótesisi nula de igualdad de las medias, sólo nos indica que existe al menos una media significativamente distinta de las demás, pero sin saber cual cuales son las medias que difieren, donde se encuentran esas diferencias. Para ello debemos realizar alguna de las pruebas de comparaciones múltiples a posteriori.

Existen varios test y los paquetes de programas traen esa opción, una de las mejores pruebas por su potencia y porque además admite trabajar con **n** diferentes es la prueba Scheffé.

Veamos la realización de las pruebas a posteriori con un ejemplo:

Se quiere comparar el nivel de la fecundidad para más de 100 países agrupados por regiones. El nivel de la fecundidad se midió a través de la tasa global de fecundidad: número promedio de hijos por mujer en edad fértil (15 a 44 años)
El factor o tratamiento tiene 6 niveles, las 6 regiones en que se agruparon los países: Comunidad Económica Europea, Europa Oriental. Medio Oriente, América latina, Pacífico y Asia y África.
La salida del programa SPSS se muestra a continuación
En la tabla ANOVA que se muestra a continuación se observa como la variación entre grupos es mucho mayor que la variación dentro de grupos y la prueba F

dio un error de tipo I de 0.000, por consiguiente podemos rechazar la hipótesis nula de igualdad de las medias, hijos promedios por mujer, y aceptar la alternativa y de diferencia entre el número promedio de hijos por mujer entre las regiones.

Este resultado era de esperar, pues se sabe que en los países de mayor desarrollo los niveles de fecundidad son más bajos que en los países menos desarrollados.

ANOVA

	Fertility: average number of kids				
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	250.679	5	50.136	38.080	.000
Within Groups	132.977	101	1.317		
Total	383.656	106			

Multiple Comparisons
Dependent Variable: Fertility: average number of kids
Scheffe

(I) Region or economic group	(J) Region or economic group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
OECD	East Europe	-.143	.405	1.000	-1.518	1.231
	Pacific/Asia	-1.637	.381	.004	-2.929	-.344
	Africa	-4.334	.363	.000	-5.568	-3.101
	Middle East	-2.978	.374	.000	-4.249	-1.707
	Latin America	-1.590	.354	.002	-2.792	-.388
East Europe	OECD	.143	.405	1.000	-1.231	1.518
	Pacific/Asia	-1.494	.428	.040	-2.948	-3.958E-02
	Africa	-4.191	.413	.000	-5.593	-2.789
	Middle East	-2.835	.423	.000	-4.270	-1.400
	Latín América	-1.446	.405	.032	-2.821	-7.197E-02
Pacific/Asia	OECD	1.637	.381	.004	.344	2.929
	East Europe	1.494	.428	.040	3.958E-02	2.948
	Africa	-2.697	.389	.000	-4.019	-1.376
	Middle East	-1.341	.400	.055	-2.698	1.565E-02
	Latín América	4.741E-02	.381	1.000	-1.245	1.340
Africa	OECD	4.334	.363	.000	3.101	5.568
	East Europe	4.191	.413	.000	2.789	5.593
	Pacific/Asia	2.697	.389	.000	1.376	4.019
	Middle East	1.356	.383	.035	5.612E-02	2.657
	Latin America	2.745	.363	.000	1.512	3.978
Middle East	OECD	2.978	.374	.000	1.707	4.249
	East Europe	2.835	.423	.000	1.400	4.270
	Pacific/Asia	1.341	.400	.055	-1.565E-02	2.698
	Africa	-1.356	.383	.035	-2.657	-5.612E-02
	Latín América	1.388	.374	.023	.118	2.659

Latín América	OECD	1.590	.354	.002	.388	2.792
	East Europe	1.446	.405	.032	7.197E-02	2.821
	Pacific/Asia	-4.741E-02	.381	1.000	-1.340	1.245
	Africa	-2.745	.363	.000	-3.978	-1.512
	Middle East	-1.388	.374	.023	-2.659	-.118

* The mean difference is significant at the .05 level.

Como se observa en la tabla anterior, se realizó una prueba a posteriori o de comparaciones múltiples, la prueba Sheffé, para identificar cuales eran las regiones que diferían significativamente en sus niveles de fecundidad.
Veamos algunos resultados. La primera región que se compara es la Comunidad Económica Europea con el resto de las regiones. Difieren signifivamente ($p< 0.05$) el resto de las regiones excepto Europa Oriental con similares niveles de fecundidad.

La última región que se compara es América latina, difiere significativamente de la Comunidad Económica Europea y Europa Oriental, regiones que tiene niveles mas bajos de fecundidad. También se aprecia diferencias significativas con África y Medio Oriente, en éste caso ambas regiones exhiben tasas mayores que las de América. La región del Pacífico y Asia presenta un comportamiento de la fecundidad similar al de latino América.

Inferencia sobre proporciones.

1- *Intervalo de confianza para proporciones.*

La distribución binomial, otra distribución teórica de frecuencia o probabilidades, es el modelo teórico sobre el que se desarrolla la inferencia sobre las proporciones.

Sin embargo, al igual que ocurría con las medias, cuando el tamaño de muestra es *suficientemente grande*, se puede utilizar la aproximación a la normal para la estimación por intervalos.

Nos limitaremos al caso de *muestras grandes,* ya que para muestras pequeñas no se puede utilizar la aproximación a la normal y rebasa los objetivos de este curso.

Colton plantes que para identificar si el tamaño de la muestra es *suficientemente grande*, debemos basarnos en la siguiente regla.

$np \geq 5$, y $nq \geq 5$

Donde:

n = tamaño de la muestra
p = proporción.
q = (1-p)

Si se cumple la aproximación a la normal la expresión de cálculo sería:

$$p \pm Z_{\alpha/2} (\sqrt{pq / n}$$

Donde:
p = proporción muestral.
q = (1-p)

Veamos un ejemplo.

En una muestra de 25 hombres con edades comprendidas entre los 40 y 60 años pertenecientes a un área de salud, se encontró que el 56% eran fumadores. Cuáles son los límites de confianza al 95% para la prevalencia de fumadores en el área estudiada?

$$0.56 \pm 1.96 \sqrt{(0.56)(0.44) / 25}$$

$$0.56 \pm o.195$$

Límite inferior = 0.365
Límite superior = 0.755

Podemos afirmar con una certeza del 95% que la prevalencia de fumadores en el área de salud para los hombres de ese grupo de edad se centra alrededor del rango de 36.5% al 75.5%.

- *Comparación de dos proporciones en muestras independientes*..

En ocasiones nos enfrentamos al problema de discernir si la diferencia observada entre dos proporciones de dos muestras independientes es realmente significativa o es debida al hecho de estar trabajando con muestras.
Por ejemplo, se está probando la efectividad de un nuevo tratamiento para la migraña. A una muestra de 400 pacientes que padecen esta enfermedad se le administró la droga A, el nuevo medicamento, y se observó una remisión de los síntomas en 350, el 87%. Como grupo de control se tomo un muestra de 300 pacientes y se le administró el tto. convencional, B y se encontró que 225 mejoraron sus síntomas (75%). El investigador quiere saber si esta diferencia entre la proporción de casos mejorados es porque la nueva droga es más

eficiente que la otra, o es producto de la casualidad, del hecho de trabajar con muestras.
La hipótesis nula sería:

$H_0 = p_1 = p_2$
Las alternativas podrían ser:

$H_1 = p_1 \neq p_2$
$H_1 = p_1 > p_2$

Estadígrafo de prueba:
La distribución teórica de frecuencia que subyace para esta prueba es la **binomial,** Sin embargo, cuando el tamaño de las muestras es " lo suficientemente grande" puede asumirse que la diferencia entre la proporciones tiende a la distribución normal con media $\mathbf{p}_i$ y un error estándar de :

EE= $\sqrt{p_i(1-p_1)/n_1 + p_2(1-p_2)/n_2}$

Se considera **n** suficientemente grande con fines prácticos y conservadores, cuando se cumple que np> 5 y n(1-p) > 5
El estadígrafo de prueba sería por consiguiente:

$$Z=\frac{p_1 - p_2}{\sqrt{P(1-P)(1/n_1 + 1/n_2}}$$

Donde: $p_1 = \frac{x_1}{n_1}$ y $p_2 = \frac{x_2}{n_2}$

Y donde P, es la probabilidad combinada de éxitos de ambas muestras.

$$P= \frac{x_1 + x_2}{n_1 + n_2}$$

Nota: a 1-p también se le denota por la letra Q
Veamos el ejemplo que nos ocupa:

Cálculo del estadígrafo de prueba:

$$Z=\frac{0.875 - 0.750}{\sqrt{(0.821x0.178)(1/400 + 1/300}} = 4.28$$

Nivel de significación de la prueba: $\alpha = 0.05$

Regla de decisión:
Para la alternativa de desigualdad de proporciones el valor de Z., como vimos en la comparación de dos medias, es de 1.96

Si trabajamos con la alternativa de que la proporción de casos en la primera muestra supera a los de la segunda, el percentil de la normal estándar sería de 1.645.
Como
4.28> 1.96 y
4.28> 1.645
Se produce el rechazo de la hipótesis nula y aceptamos la alternativa. La droga A es más efectiva que la B pues la proporción de casos mejorados con ella fue mayor.

Ejercicios resueltos

1. En un estudio realizado en un hospital dos investigadores revisaron los registros de pacientes admitidos con infarto del miocardio en la sala de cuidados intensivos coronarios, en dos años seleccionados, 1989 y 2000. El número promedio de días que los paciente clasificados como pacientes con un ataque leve, permanecieron ingresados en éste servicio se muestra a continuación.

1989	2000
n = 19	n = 26
$\bar{x}$ = 3.88	$\bar{x}$= 6.15
s = 1.37	s = 1.85

La diferencia entre el promedio de días en los dos años de estudio va más allá de lo que podría esperarse para una variación fortuita al azar?
Sea α = 0.01

Respuesta.

Ho = $\bar{x}$ - $\bar{x}$

Hi = $\bar{x}$ < $\bar{x}$

Estadígrafo de prueba:

$$t_{n1-n2-2} = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$t = \frac{3.89 - 6.15}{(18x1.37) + (25x1.85) / 43x\sqrt{0.05 + 0.03}}$$

t = 4.80

Como la suma del tamaño de las muestras es mayor de 30, se trabaja con la aproximación al normal.

Regla de decisión.

Z = 2.33 (para α 0.01, contraste de una cola).

Como 4.80 > 2.33

Rechazo la hipótesis nula.
El promedio de días de hospitalización de los pacientes fue menor en el año 85 en comparación con el año 95 con una probabilidad de error del 1%.
También se podría decir que existen evidencias estadísticamente significativas de que el promedio de días de hospitalización de los pacientes fue menor en el año 89 que en el 2000.

2. En un estudio realizado para evaluar la calidad de los servicios de un hospital, se aplicó a una muestra de pacientes egresados de diferentes servicios una encuesta para conocer la satisfacción con la atención recibida.
De los 45 pacientes egresados del servicio de Medicina, 38 estaban satisfechos con la calidad de la atención y de 50 encuestados del servicio de Cirugía General, 45 mostraron satisfacción con el servicio recibido.
Proporcionan estos datos evidencias suficientes que indique que existe una mayor satisfacción de los pacientes atendidos en el servicio de Cirugía

Datos.

Medicina	Cirugía
n_1= 45	n_2= 50
p_1= 38/45	p_2= 45/90
p_1= 0.84	p_2= 0.90

Estadígrafo de prueba.

$$Z = \frac{p_1 - p_2}{\sqrt{P.Q(1/n_1 + 1/n_2)}}$$

Donde:

$$P = \frac{p_1 + p_2}{n_1 + n_2}$$

Comentario [FV2]:

P = 38+45/45+50
P = 83/95
P = 0.87

Q= 1-P
Q = 1-0.87
Q = 0.13

Sustituyendo:

$$Z = \frac{0.84 - 0.90}{\sqrt{(0.87x0.13)(1/45 + 1/50)}}$$

Z = 0.06/0.22

Z = 0.27

Regla de decisión.

Sea α= 0.01

Z = 2.33 (prueba de una cola, contraste unilateral)

0.22< 2.33

No rechazo Ho, no existen evidencias con significación estadística para decir que las medias difieren. Con los resultados obtenidos en el estudio no se puede decir que en el servicio de cirugía la proporción de pacientes satisfechos sea mayor.

3. A continuación aparecen los resultados de la comparación de las esperanzas de vida promedio de mujeres de mas de 100 países, agrupados según regiones, similar al estudio de los niveles de fecundidad utilizado en el texto.

Esperanza de vida promedio de las mujeres de países agrupados segun regiones económicas

Fuente de variación	Suma de cuadrados	g.l.	Cuadrados medios	F	
Variación entre grupos	7568.810	6	1513.782	34.636	.000
Variación dentro de grupos	4501.539	103	43.704		
Variación total	12070.349				

Hipótesis

En la hipótesis nula plantearíamos la igualdad de las esperanzas de vida medias de las mujeres según regiones económicas.

$H_o = \mu_1 = \mu_2 = \ldots \mu_k$

Puede apreciarse en la tabla que la variabilidad entre grupos (regiones) es mucho mayor que la variabilidad dentro de los grupos , como evidencian los cuadrados medios.
La prueba F dio una probabilidad de error de tipo 1 de 0.000 por lo que se rechaza la hipotesis nula y se concluye que las esperanzas de vida promedio de las mujeres según regiones económicas difieren, resultado esperado pues la comparación se estableció entre regiones tan desarrolladas como la Comunidad Económica Europea y Africa.

ENFOQUE DE RIESGO

Con bastante frecuencia hablamos y oímos sobre un número no despreciable de *factores de riesgo*, para diferentes problemas de salud. Si embargo, lo que no siempre está claro para muchos de nosotros es qué es **riesgo**, que es un **factor de riesgo** y como se miden.

Riesgo. Es la probabilidad de ocurrencia de un evento, una enfermedad complicaciones de una enfermedad, la muerte. También se puede entender como la expresión de un peligro de sufrir un daño o perjuicio.
Por ejemplo el riesgo de tener un infarto, de morir por cáncer del pulmón, de nacer un niño con bajo peso, de que ocurra una mortalidad infantil.

El estudio de los riesgos comprende su determinación, los factores que lo condicionan, su cálculo en forma directa o su estimación.

Con frecuencia oímos decir que un grupo de alto riesgo para el infarto del miocardio lo constituyen sujetos obesos, sedentarios, fumadores, hipertensos, entre otros.
La probabilidad de tener un infarto del miocardio sería el riesgo, todas esas variables, factores que se asocian a la probabilidad de ocurrencia del infarto del miocardio son los factores de riesgo.

Factor de riesgo, es el conjunto de fenómenos de los cuales depende esa probabilidad.
Si el factor o los factores de riesgo son de tipo causal, la eliminación o control de los mismos disminuirán o evitaran el riesgo. De ahí la importancia del estudio de los factores de riego, pues permiten tomar acciones preventivas.

Cálculo del riesgo. Es la medición de la probabilidad de ocurrencia de un evento dado en forma global o específica. Las tasas me estiman esas probabilidades, por tanto el cálculo se hace en base a tasas de incidencia, prevalencia y mortalidad.

Cálculo del riesgo en estudios poblacionales.

Supongamos el ejemplo clásico del cáncer del pulmón y el hábito de fumar

Esquema

Factor de riesgo	Efecto presente	Efecto ausente	total
F+	a	b	a+b
F-	c	d	c+d
Total	a+c	b+d	n

Rt = a+c/n Tasa de la enfermedad en esa población (riesgo total)
Rf+ = a/a+b Tasa de enfermar los expuestos al factor de riesgo (riesgo específico)
Rf- = c/d+d Tasa de enfermar los no expuestos (riesgo específico)

No es difícil suponer que si el factor evaluado es un factor de riesgo, la tasa en los expuestos al factor debe ser mayor que la tasa en los no expuestos.

Por tanto si relacionamos ambas tasas, y la tasa de los expuestos ocupa el numerador de esa razón, y en el denominador, la tasa de los no expuestos, cabe esperar los siguientes resultados.

- Si el factor investigado es un factor de riesgo, ese cociente tiene que ser superior a 1, pues la tasa de los expuestos tiene que ser la mayor.
- Si el resultado es exactamente 1, los riesgos son iguales esté o no el factor presente.
- Si el resultado es menor que 1, el factor no es un factor de riesgo.

Esa relación por cociente, esa razón de tasas, es el **riesgo *relativo.***
RR = tasa en los expuestos al factor de riesgo/ tasa de los no expuestos.

El riesgo relativo es un medida de la intensidad de la asociación y es un elemento importante en el establecimiento de la causalidad.

Veamos un ejemplo concreto.

Bajo peso	Fallecido < 1 año	No fallecido	total
Si	19	720	739
No	17	1570	1587
total	36	2290	2326

Tasa M.I = 36/2326 x 1000 = 15.5
Tasa M.I = 19/739 x 1000 = 25.7
(exp.)
Tasa M.I = 17/1587 x 1000 = 10.7

(no exp.)

RR = 25.7/10.7 = 2.4

Los niños bajo pesos tienen un riesgo 2.4 veces mayor de fallecer que los niños normo pesos. También se puede interpretar que es 2.4 veces más probable o más frecuente que muera un niño bajo peso que un niño normo peso.

Riesgo atribuible.

Es una medida del **impacto** que provoca el factor de riesgo en el grupo sometido a su exposición.

RA = tasa de los expuestos - tasa de los no expuestos.

Si el factor no tuviese influencia las tasas serian prácticamente iguales y la diferencia sería 0, por tanto todo valor mayor que 0, es el impacto del factor.

El riesgo atribuible es una tasa y nos indica el beneficio ganado en tasa, en cuantas unidades yo reduciría la tasa en los expuestos si se eliminara el factor de riesgo. De ahí su gran utilidad para fines investigativos y administrativos pues permite priorizar los programas y las acciones en función de los beneficios de reducción de la tasa e incluso puedo calcular los beneficios desde el punto de vista económico.

Con el ejemplo de la mortalidad infantil:

RA = 25.7- 10.7
RA = 15 x 1000

En 15 unidades se reduciría la tasa de los bajo pesos si se eliminara el efecto del factor.

Por ciento de riesgo atribuible o fracción etiológica en los expuestos.

Es la expresión porcentual del riesgo atribuible, en que por ciento se reduciría la tasa de los expuestos si se eliminara el factor de riesgo.

%RA o FEE = tasa de los expuestos- tasa no expuestos/tasa expuestos x 100

Ejemplo.

% RA o FEE = 15/25.7 x 100
% RA o FEE = 58.3%

La tasa de los niños bajo peso se reduciría en un58% si se eliminaran los recién nacidos con esa condición.

Riesgo atribuible poblacional.

Es el exceso de mortalidad o morbilidad por la presencia del factor de riesgo pero en toda la población, lo que se reduciría la tasa general de no existir el factor de riesgo.

RAP = tasa total- tasa de los no expuestos.
Ejemplo.

RAP= 15.5 -10.7
RAP = 4.8 x 1000

La tasa de mortalidad se reduciría en esa comunidad en 4.8 x 1000 n.v.

Por ciento de riesgo atribuible poblacional o fracción etiológica.

Expresa el por ciento de reducción de la tasa total, en toda la comunidad si se elimina el factor de riesgo.

%RA o FE = tasa total - tasa no expuestos/ tasa total x 100

Ejemplo.

%RA o FE = 4.8/15.5
%RA o FE = 30.9
La tasa de mortalidad infantil se reduciría en esa comunidad en un 30.9% si se eliminara el bajo peso al nacer.

Creemos oportuno aclarar que aquel factor que tenga un mayor riesgo relativo no tiene que tener mayores riesgos atribuibles también. Eso depende de la frecuencia conque ocurra en la población el factor que se está investigando.

Por ejemplo si tenemos dos enfermedades con igual riesgo relativo, tendrá mayor riesgo atribuible aquella que tenga una incidencia o prevalencia mayor.

Estimación de los riesgos en estudios no poblacionales.

No siempre se dispone de la información necesaria para realizar estudios de riesgos en la población.

Los estudios que con más frecuencia se utilizan son los estudios de cohorte y los de casos y controles .

El diseño de estos estudios se abordara con mas detalle en el módulo de Metodología de la Investigación.

En los estudios de cohorte se obtienen estimaciones de las tasas y los cálculos se realizan igual que como ya vimos para los estudios poblacionales.

Como se trata de estimaciones es necesario realizar una prueba de hipótesis para determinar si existe asociación significativa entre el factor analizado y el efecto o riesgo estudiado. Para ello utilizamos la prueba Chi cuadrado de independencia .

Si hay significación se procede entonces a las estimaciones de las medidas de riesgo, pero no solo de manera puntual sino por intervalo.

No entraremos a profundizar en estos aspectos porque rebasaría los objetivos de éste curso.
En los estudios de casos y controles se estiman estas mismas medidas pero con expresiones de calculo diferentes, ya que en éste caso no se pueden realizar estimaciones de las tasas como ocurría en los estudios de cohorte.

Antes de iniciar las estimaciones es necesario realizar una prueba Chi cuadrado de independencia para probar si existe asociación significativa entre el factor y el efecto analizado.

Al igual que en los estudios de cohorte se realiza además de la estimación puntual la estimación por intervalo.

Vamos a abordar la estimación del riesgo relativo para los estudios de casos y controles, pues tienen u uso bastante difundido.

Una vez comprobado que existe asociación significativa se procede a la estimación de riesgo relativo mediante la denominada *razón de productos cruzados.*

Tomemos el esquema inicial de ejemplo.

Esquema

Factor de riesgo	Efecto presente	Efecto ausente	total
F+	A	b	a+b
F-	c	d	c+d
Total	a+c	b+d	n

a+c = grupo de casos (efecto presente)

b+d = grupo de controles (efecto ausente)

RR = ad / bc

La estimación por intervalo es mediante la siguiente expresión.

$$\widehat{RR} \quad 1 \pm \left(z / \sqrt{x^2} \right)$$

Donde:

Z = percentil de la normal estándar
X^2 = Chi cuadrado calculada

Para que el factor sea considerado como un factor de riesgo es necesario que el límite inferior del intervalo supere a la unidad.

Veamos un ejemplo.

Hipertensión	Con infarto	Sin infarto	total
Si	51	66	117
No	59	154	213
total	110	220	330

RR = 51x 154/ 59x66
RR = 2.02
X^2 = 8.57

$$2.02^{1 \pm (1.96 / \sqrt{8.57})}$$

Límite inferior = 1.26
Limite superior = 3.37

En primer lugar existe una asociación significativa entre la hipertensión y el tener un infarto del miocardio, la Ji cuadrda calculada de 8.57 es mayor que la tabulada de 3,76 con un α = 0.05 y 1gl, por consiguiente se rechaza la H_0 de no asociación. Como el límite inferior de la estimación por intervalo del riesgo

relativo es mayor que 1 se concluye que la hipertensión es un factor de riesgo para el infarto del miocardio por lo que es, en este caso, 2 veces más probable que tenga un infarto agudo del miocardio un paciente hipertenso que uno que no padezca de esta enfermedad.