UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE MEDICINA

DEPARTAMENTO DE MEDICINA PREVENTIVA
Y SALUD PÚBLICA

ESTADÍSTICA MÉDICA

Profesor Responsable
Dra. JULIA PISCOYA SARA

Colaboradores:

Dr. JORGE ALARCÓN VILLAVERDE
Ing. EDITH ALARCÓN MATUTI
Ing. LUZ BULLÓN CAMARENA
Dra. ELSY CUELLAR FRETEL
Lic. ESPERANZA GARCÍA CRIBILLEROS
Dr. CÉSAR GUTIERREZ VILLAFUERTE
Mg. MARTHA MARTINA CHÁVEZ
Dra. MARÍA TERESA PERALES DÍAZ

2 0 0 5

ESTADÍSTICA

Dra. Julia Piscoya S.

Antiguamente la estadística solo era considerada para hacer los consolidados numéricos de hechos ocurridos, hoy en día la estadística juega un papel muy importante tanto en nuestra vida cotidiana como en la investigación y situaciones especiales para la toma de decisiones, que sería muy extenso de mencionar. Las diferentes técnicas que ha desarrollado hacen que su uso sea aplicable en las diferentes áreas del conocimiento científico: física, química, antropología, biología, por mencionar algunas. En lo que respecta a nuestro campo, la Bioestadística cada vez es más utilizada tanto para describir, extrapolar resultados, tomar decisiones, establecer diseños de investigación, etc, lo cual hace que sea imposible concebir un especialista de la salud que no tenga conocimientos y un mínimo de habilidades de esta disciplina.

La estadística es una rama de la matemática referida a un sistema o método científico usado en la recolección, organización, análisis e interpretación numérica de la información. El método estadístico nos ayuda a elaborar inferencias inductivas a partir de la constatación de hechos particulares, independientemente del método de investigación con que se haga esta constatación.

Desde el punto de vista descriptivo-analítico la estadística se define como u conjunto sistemático de procedimientos para observar y describir numéricamente los fenómenos y descubrir las leyes que regulan la aparición, transformación y desaparición de los mismos.

En el campo de la Estadística se diferencian dos partes:

- ESTADÍSTICA DESCRIPTIVA O DEDUCTIVA es la que, como su nombre lo indica, se limita a la descripción de un conjunto de datos sin llegar a generalizar con respecto a un grupo mayor.
- ESTADÍSTICA INFERENCIAL O INDUCTIVA es la que se dedica al análisis y trata de llegar a conclusiones o generalizaciones acerca de un grupo mayor, basado en un grupo menor o “muestra”.

EL METODO ESTADÍSTICO

El método estadístico comprende las siguientes etapas:

1. Planificación de estudio
2. Recolección de la información
3. Elaboración o tabulación de los datos recogidos
4. Análisis e interpretación

1.- PLANIFICACIÓN DEL ESTUDIO

En la planificación des estudio tenemos que tener en cuenta :

1. Planteamiento del problema

2. Naturaleza e importancia del problema que se estudia
3. Determinación de los objetivos
4. Búsqueda y evaluación de la información existente.
5. Diseño del estudio
- Problema
- Hipótesis
- Variables: tipo, definición operacional de variables, control de variables extrañas, instrumentos de medición
- Sujetos de estudio (Población, muestra)
- Tipo de estudio (Descriptivo, analítico)
- Fuentes de información (Primaria, secundaria)
- Plan de Análisis

6. Cronograma de Actividades
7. Presupuesto

Es importante notar que en la planificación del estudio se debe desarrollar el PLAN DE ANÁLISIS. Uno de los errores frecuentes es obviar este punto y solo después de que se ha recolectado la información se piensa en cómo debe presentarse la información; esto trae consigo una recolección inadecuada de la información, ya sea por exceso o por defecto.

2.- RECOLECCIÓN DE INFORMACIÓN

Preparada la investigación comienza la recolección de datos. La recolección de la información puede hacerse de muchas manera. El método seleccionado dependerá de:

- Los objetivos y diseño de estudio
- Disponibilidad de los recursos humanos
- Recursos financieros

Esta etapa es muy importante, deberá hacerse con mucho cuidado, porque en muchos casos esta no puede repetirse para una corrección. De ella depende todo el resultado posterior; si esta mal realizada se hará una elaboración y tabulación inadecuada de los datos, dando origen a un análisis erróneo e interpretaciones equivocadas.

Otro punto a tenerse en cuenta es que tipo de resultados se quieren producir, si se pretende producir resultados cuantitativos con cierto grado de precisión o bien datos cualitativos que proporcionen información de tipo descriptivo. Es frecuente que los objetivos del estudio requieran de información cuantitativa y cualitativa, lo que implica que se debe emplear más de un método de recolección de información.

2.1. MÉTODOS DE RECOLECCIÓN CUALITATIVA

Entrevista no estructurada
Grupos focales
Observación directa y otros

2.2. MÉTODOS DE RECOLECCIÓN CUANTITATIVA

Entrevistas estructuradas: Encuestas
Censos
Sistemas de registros
Entrevista indirecta

2.2.1. ENCUESTA

Es una técnica o procedimiento de recolección de datos en muestras poblacionales. El instrumento que se utiliza en una encuesta es el cuestionario. Este consiste en un conjunto de preguntas formuladas y escritas que sirven para recoger datos orientados a un fin específico; puede ser desarrollado a través de una entrevista o auto administrado.

La entrevista es una conversación guiada por preguntas que el entrevistador (llamado también encuestador) realiza a la persona entrevistada. En este caso, las preguntas del cuestionario son leídas por el entrevistador, quien a su vez consigna las respuestas del entrevistado en el cuestionario correspondiente.
Cuando es auto administrado, el entrevistado recibe el cuestionario directamente o por correo para que él mismo consigne sus respuestas.

ETAPAS DE UNA ENCUESTA:

1. Definición de los objetivos de la encuesta.

El objetivo de la encuesta es recoger información (datos) para resolver un problema científico determinado. Esta información estará en relación a la/las hipótesis que formule el investigador y al conjunto de variables que le permitan describir o explicar el fenómeno en estudio.

Por ejemplo: si el problema es saber ¿Cuál es la relación entre hipertensión y consumo de sal? y la hipótesis es que "El excesivo consumo de sal en la dieta está asociado a la hipertensión" el objetivo de la encuesta será recoger datos acerca de presión arterial y la historia de consumo de sal en la dieta. Pero además de estas dos variables principales sabemos que hay otros factores que pueden tener importancia en la hipertensión arterial como: edad, raza, sexo, ocupación, antecedentes familiares, etc. Por lo tanto los objetivos de esta encuesta serán:

- Recoger datos sobre presión arterial
- Recoger datos sobre historia de consumo de sal en la dieta
- Recoger datos respecto a la edad, sexo, raza, ocupación, historia familiar, etc.

2. Delimitación de la población a estudiar.

Es importante delimitar la población que va ser estudiada, por lo que será necesario definir criterios precisos que permitan establecer qué sujetos pertenecen o no a la población objeto de estudio.

3. Hacer un estudio exploratorio.

El estudio exploratorio consiste en reconocer las características sociales, culturales, ambientales y la distribución geográfica de la población en estudio. Tiene por finalidad establecer la factibilidad del estudio y los instrumentos más adecuados para recoger información. Por ejemplo, si un alto porcentaje de la población es analfabeta no se podrá aplicar un cuestionario auto administrado. En esta etapa es útil el empleo de técnicas como el estudio de grupos focales y la observación.
También nos permitirá evaluar el grado de aceptabilidad que tendrá el estudio en la población seleccionada

3. Preparación del instrumento.

El cuestionario es un instrumento con objetivos definidos que servirá para obtener información de las variables que se han seleccionado en el estudio, hay que tener en cuenta lo siguiente:

- Tipo de pregunta.

El cuestionario es un conjunto de preguntas o ítems, en donde cada pregunta puede corresponder a una variable, una clasificación de la variable o a un indicador de la variable. Las preguntas pueden ser de dos tipos: cerradas y abiertas.
Las preguntas cerradas son aquellas que ya tienen escritas las opciones de respuesta.
Las preguntas abiertas son aquellas que no tienen ninguna opción de respuesta por lo que el entrevistador tendrá que escribir la respuesta que le dé el entrevistado.

- Orden de las preguntas.

Es importante tener en cuenta el orden de las preguntas. Uno de los criterios importantes es la ubicación de las preguntas llamadas sensitivas, debido a la reacción que producen en el entrevistado. Por este motivo, se colocarán primero las preguntas menos sensitivas.
Otro criterio importante es mantener el orden lógico de las preguntas. Por ejemplo, no se puede preguntar qué resultados tuvo en la alimentación de su niño con la leche materna, si previamente no se sabe si tiene hijos.

- Claridad de las preguntas.

Las preguntas deben expresar claramente el contenido de la variable, deben ser comprensibles para la persona que va ser entrevistada, no debe haber dos preguntas en una sola, no deben ser ambiguas, no deben sugerir ninguna respuesta.

- Diseño del cuestionario.

El diseño del cuestionario es importante para que la persona que entrevista no se equivoque en el llenado. En el diseño hay que tener en cuenta que para una mejor disposición de las preguntas es conveniente reunir todas las de una determinada área en bloques. Debe ser ágil, es decir que cada pregunta tenga las indicaciones pertinentes para ser respondidas, así como la forma de pasar a la siguiente en el caso de que ella se derive otro grupo de preguntas (se debe indicar el salto de los ítems).

5. Prueba piloto.

Cuando se tiene diseñado el cuestionario se procede al pretest o prueba piloto. La prueba piloto consiste en aplicar el cuestionario en una pequeña muestra de la población o en una población con características similares. El número adecuado para aplicarla es entre 5-30, según sea el número de personas del grupo a quien va dirigido; si el grupo poblacional es pequeño no se puede tomar muchos individuos para la prueba piloto, pues ellos ya no serán incluidos en la aplicación del cuestionario final.
Esta prueba piloto se podrá repetir las veces que sean necesarias, pero como hemos mencionado dependerá del número de individuos a quien va dirigida.

Es importante que el investigador y los encuestadores participen de esta prueba piloto, mediante esta prueba se puede obtener mucha información. Por ejemplo, de la entrevista: qué hora es la más adecuada para hacer la entrevista, el tiempo que demora; del cuestionario: es necesaria la pregunta; las alternativas para las preguntas

son suficientes, han sido demasiadas, han sido pocas; en las preguntas abiertas hay suficiente espacio para el llenado; orden de las preguntas; reacción del entrevistado frente al cuestionario, alguna de las preguntas despierta una reacción inadecuada al cuestionario; han sido claras las preguntas para el entrevistado; el diseño del cuestionario permite un llenado fácil o tiene elementos que dificultan su manejo.
Para el investigador, la prueba piloto puede servir para aprender algo nuevo del problema, introducir nuevas preguntas e incluso reformular su hipótesis. También le sirve para decidir cuáles de los encuestadores son aptos para participar en el estudio, ver el tiempo que demoran en cada encuesta, evaluar el tiempo que demoran para desplazarse en la zona, lo que le permitirá ajustar mejor los tiempos y hacer un cronograma de actividades más exacto. Para los encuestadores les sirve para familiarizarse con el cuestionario.

6. Aplicación del cuestionario.

Concluida la prueba piloto se tiene el cuestionario final para su aplicación, cabe señalar que para esta etapa los encuestadores ya deben estar capacitados y con el manual de encuestadores aprendido. Es importante que durante el desarrollo de la encuesta halla una o más personas encargadas de la supervisión del llenado completo del cuestionario; así, si alguno de los encuestadores omitió alguna pregunta puede regresar a completar la información.

3. ELABORACIÓN O TABULACIÓN DE LOS DATOS RECOGIDOS

Revisión y corrección de la información recogida
Procesamiento de los datos
Preparación y selección de tablas y gráficos más adecuados

- **REVISIÓN Y CORRECCIÓN DE LA INFORMACIÓN RECOGIDA**

Una vez recogida toda la información es necesario someterla a un examen crítico con la finalidad de comprobar que cumple con las condiciones indispensables. El objeto de la crítica es clasificar el material en tres grupos: material bueno, material incorrecto pero corregible y material incorregible o desechable; la clase e importancia del error cometido determinan la admisión o no de los datos recogidos.

- **PROCESAMIENTO DE LA INFORMACIÓN**

Terminada la revisión y corrección se inicia la etapa del procesamiento de los datos.
Si es que se ha aplicado una encuesta o si es que se han recolectado los datos en un formato determinado, la primera etapa del procesamiento es la codificación. Esta consiste en el traslado de las respuestas a un lenguaje sencillo (números) con el objeto de facilitar el análisis; antes de iniciar la codificación es necesario tener el "libro de códigos", que no es sino un listado de valores para cada una de las respuestas que existen en el cuestionario. Es importante que una vez terminada la codificación se haga un control de calidad, seleccionando un pequeño grupo de encuestas y revisando si ha sido correcta la codificación; si hay muchos errores habrá que revisar nuevamente este paso, pues esta es una fuente de error en los resultados. Si el cuestionario o el formato utilizado ha sido precodificado, no se tendrá que hacer este paso.

Terminada la codificación se proceder a la tabulación de los datos, ésta puede realizarse en forma manual o mediante el uso de máquinas (computadoras).
En el caso de usar computadoras, concluida la codificación se procede a la digitación que no es sino la introducción de los datos a una "base de datos" de algún programa

determinado. Una vez terminada la digitación es conveniente que se realice un control de calidad de este ingreso, para hacer la corrección respectiva, sino se puede convertir en otra fuente de error. Posteriormente, se podrá usar para el análisis algún programa estadístico (SPSS, EPI INFO, MINITAB, etc) que facilitará la obtención de los resultados.

- **PREPARACIÓN Y SELECCIÓN DE TABLAS Y GRÁFICOS MÁS ADECUADOS**

Realizada la tabulación inicial, es importante que se seleccionen algunas tablas y gráficos para que describan de una manera sencilla y adecuada el tipo de datos que se ha recolectado.

4. ANÁLISIS E INTERPRETACIÓN

El análisis puede ser de tipo descriptivo o inferencial, de acuerdo a lo que el investigador propuso en el plan de análisis, esta etapa no es sino la consolidación de lo que ya estuvo planificado anteriormente.
Los resultados serán interpretados por el investigador quien se encargará de la descripción de los hallazgos en relación a su/sus hipótesis planteadas.
Es muy importante que una vez finalizado un estudio se den a conocer los resultados obtenidos, mejor si estos son publicados; si es así, se deberá incluir como anexo el cuestionario utilizado, con el objeto de facilitar la interpretación a otros investigadores, quienes podrán efectuar réplicas si lo creen necesario.

USOS DE LA ESTADÍSTICA

1. En el diseño de investigaciones.

- Construcción de escalas de medición.
- Control de variables intervinientes.
- Selección de sujetos de estudios.

2. En el análisis de resultados.

- Descripción de variables.
- Describir asociaciones.

3. En la toma de decisiones (inferencia).

- Respecto a un valor obtenido.
- Respecto a una asociación observada.

VARIABLES

Dra. Julia Piscoya Sara *Dra. María Teresa Perales Díaz*

Variable es toda característica o atributo susceptible de tomar un valor y ser medido. Esta característica puede ser de las personas, objetos, lugares o cosas. Como su nombre lo dice, varía de acuerdo a cada sujeto de estudio; por lo tanto, para convertirse en variable la característica debe tener mas de dos valores.

Ejemplos:
Sexo, numero de hijos por familia, peso, numero de intervenciones quirúrgicas por paciente, edad, episodios de crisis asmática por paciente, estatura, nivel de educación, etc.

CLASIFICACIÓN DE VARIABLES

Las variables pueden ser de dos tipos:

1.- CUALITATIVAS O CATEGÓRICAS

Son variables que determinan una cualidad o atributo, solo se pueden clasificar o categorizar mediante el conteo. Pueden ser:

- Dicotómicas, si solo tienen dos categorías. Por ejemplo, la variable estado de salud tiene dos categorías: Sano y Enfermo.

- Politómicas, si tienen más de dos categorías. Por ejemplo, la variable estado civil tiene más de dos categorías: Soltero, Casado, Divorciado, Viudo.

2.- CUANTITATIVAS O NUMÉRICAS

Son variables que se expresan numéricamente, se pueden medir. Estas a su vez pueden ser discretas o continuas.

- Variables cuantitativas discretas o discontinuas, toman valores enteros y no pueden tomar un valor entre dos consecutivos. Por ejemplo: número de camas hospitalarias, número de médicos por país.

- Variables cuantitativas continuas, toman valores que pueden ser cualquiera de los números reales, encontrando infinitos valores entre dos distintos. Por ejemplo: edad, peso.

ESCALAS DE MEDICION

La escala de medición es el grado de precisión con que se va expresar la medida de una variable. Esta va determinar la forma de presentación de la información y resumen, así como los métodos estadísticos que se usarán para analizar los datos.
Existen cuatro escalas de medición: nominal, ordinal, intervalo y razón

1. ESCALA NOMINAL

Como su nombre lo indica, sólo nomina o nombra, es la más simple de las escalas de medición, clasifica los valores de los datos sin indicar orden o jerarquía. Por ejemplo, en datos dicotómicos, las categorías, valores o clases de las variables serán: si y no, presencia y ausencia, sano y enfermo. En otros datos, como departamentos del Perú, los valores de la escala serán: Ica, Lima, Moquegua, Tumbes, etc, dependiendo de los departamentos que se estudie.

2. ESCALA ORDINAL

Esta escala no sólo clasifica sino que existe un orden o jerarquía inherente entre las categorías, las observaciones se clasifican como en la escala nominal pero algunas tienen "mas" o son "mas grandes" que otras". Por ejemplo, en la variable desnutrición, las categorías o clases serán: leve, moderada y severa o también o también: tipo I, tipo II y tipo III

3. ESCALA INTERVALO

Esta escala ya no solo nomina y ordena sino que establece distancias es decir que permite medir. El cero de la escala de intervalo es arbitrario o convencional, este no indica la ausencia del fenómeno estudiado. Por ejemplo, en la variable temperatura, el valor "0" de las escalas Centígrada y Fahrenheit no indican la ausencia del fenómeno, sino que se han tomado como punto de partida con relación a determinados fenómenos físicos; a esto es lo que se llama cero convencional.

4. ESCALA DE RAZÓN

Al igual que la anterior, esta escala también nomina, ordena y establece distancias, permite hacer mediciones. El cero de la escala de razón es real; esto quiere decir que el valor "0" indica la ausencia del fenómeno estudiado. Por ejemplo, en la variable temperatura, el valor "0" de la escala Kelvin indica la ausencia del fenómeno. La escala de razón permite todas las operaciones matemáticas.

PROCEDIMIENTOS PARA HACER UNA ESCALA DE MEDICION

1. Determinar el tipo de variable para la que se quiere construir la escala de medición.
2. Ver el instrumento de medición que se va utilizar (será un resultado numérico o solo dará un resultado como mayor, igual o menor que etc.)
3. Dar nombres a las categorías o clases, algunas veces se pueden usar números.
4. Cuidar que las categorías sean:

- Exhaustivas: es decir que en las categorías o clases deben estar contenidas todos los valores de la variables estudiada
- Mutuamente excluyentes: las categorías o clases deben estar claramente delimitadas, de manera que cuando se tenga que clasificar un dato no haya duda en dónde debe ser ubicado.

5. Tener en cuenta que si la variable es cualitativa, los números que se utilicen para designar las categorías no se pueden emplear para realizar operaciones aritméticas. Si voy a usar una escala ordinal con valores 1, 2, 3, estos no servirán para realizar operaciones aritméticas.

DISTRIBUCIÓN DE FRECUENCIAS

Dra. Julia Piscoya S.

Antes de desarrollar el procedimiento para construir una distribución de frecuencia es necesario que se definan algunos conceptos importantes que se utilizan en este procedimiento.

DATOS

Conjunto de valores que representan los diversos estados que pueden tomar una o más características de uno o más individuos.

FRECUENCIA

Es el número de veces que una característica o valor se repite en un conjunto de datos (población o muestra). A esta frecuencia es la que se le conoce como FRECUENCIA ABSOLUTA. La suma de esta frecuencia nos dará el tamaño de la población o muestra estudiada.

FRECUENCIA RELATIVA

Es la relación que existe entre las frecuencias absolutas y el tamaño de la población o muestra estudiadas. Siempre es menor que la unidad.

FRECUENCIA ACUMULADA

Es el número de observaciones menores o iguales a un determinado valor de la variable.

ORGANIZACION DE UNA DISTRIBUCIÓN DE FRECUENCIAS

Las frecuencias pueden organizarse en serie simple y en serie agrupada.
Cuando se organiza en serie simple los valores de cada clasificación (clases) están representados por un solo valor. En cambio, cuando se organiza en serie agrupada los valores están representados por un intervalo (intervalo de clase)

EJEMPLOS DE DISTRIBUCION DE FRECUENCIA

Edad (Clases)	Fc
1	2
2	3
3	2
4	1
5	2
TOTAL	10

Distribución de frecuencia
Serie simple

Edad (Intervalo de clase)	Fc
1-2	5
3-4	7
5-6	8
7-8	3
9-10	7
TOTAL	30

Distribución de frecuencia
Serie agrupada

- **Serie Simple**

a.- Para datos cualitativos

Ejemplo: estado civil de los trabajadores de una empresa
soltero-conviviente-divorciado-casado-casado-soltero-casado-conviviente-viudo-soltero-casado-soltero-viudo-soltero-conviviente-casado-soltero-soltero-soltero-soltero-conviviente- divorciado-casado-conviviente-conviviente

Tabla N° 1. Estado civil de los trabajadores de una Empresa

ESTADO CIVIL	Fc	%
Soltero	9	36
Casado	6	24
Conviviente	6	24
Divorciado	2	8
Viudo	2	8
TOTAL	25	100

b.- Para datos cuantitativos

Para organizar una serie simple solo se ordenarán los valores y se contará las veces que se repite cada uno de ellos obteniéndose la frecuencia, así como sigue:

Edad	Conteo	Fc
10	IIIII	5
11	IIIII IIIII IIIII IIIII	20
12	IIIII IIIII IIIII IIIII IIIII III	28
13	IIIII IIIII II	12
14	IIII	4
15	I	1
TOTAL		70

Ejemplo: edad de 30 pacientes:

28-28-28-28-28-30-30-30-30-30-35-35-35-35-45-45-45-45-56-56- 56-56-68-68-68-68-70-70-70-70-

Tabla N° 2. Edad de 30 pacientes

	Fc	%
28	5	32.5
30	5	32.5
35	4	13
45	4	13
56	4	13
68	4	13
70	4	13
TOTAL	30	100

-

- **Serie Agrupada**

Para organizar una serie agrupada hay que seguir algunos pasos previos, antes de obtener la frecuencia. Veamos el siguiente ejemplo:

Peso en onzas de tumores malignos extraídos a 57 pacientes

68-65-12-23-63-43-32-43-42-25-49-27-27-74-38-49-30-51-42-28-36-36-27-23-28-42-31-19-32-28-50-46-79-31-38-30-27-28-21-43-22-25-16-49-23-45-24-12-24-12-69-25-57-47-44-51-23

¿Cuántos intervalos debo formar con estos datos?

Según Daniel, lo más importante es el **conocimiento de los datos**.

Si se usan pocos, se pierde información.
Si son muchos, se pierde el objetivo de resumir la información

El recomienda una regla empírica es que no sea menor de 5, ni mayor de 15.
Si se quiere algo más exacto se puede usar el siguiente procedimiento.
REGLA DE STURGES: $k= 1+3.322 \text{ Log}_{10} n$

"k" es el № de intervalos y "n" el № de observaciones

¿Cuánto sería la amplitud de cada intervalo?

W= Rango / № de intervalos

W es el ancho del intervalo, Rango= V. Max - V. min

En el ejemplo sería

$k= 1+3.322 \text{ Log}_{10} 57 \cong 7$

$w=79 - 12 / 7 = 9.6 \cong 10$

Veamos la distribución de serie agrupada con los otros cálculos de frecuencias:

Int de clase	Fc	Fc. Acum.	Fc. Relat.	%	Fc. Relat. Acum.	% Ac.	Punto medio
10-19	5	5	.0887	8.87	.0887	8.87	15
20-29	19	24	.3333	33.33	.4210	42.10	25
30-39	10	34	.1754	17.54	.5964	59.64	35
40-49	13	47	.2281	22.81	.8245	82.45	45
50-59	4	51	.0702	7.02	.8947	89.47	55
60-69	4	55	.0702	7.02	.9649	96.49	65
70-79	2	57	.0351	3.51	1.0000	100.00	75
TOTAL	57		1.00	100.00			

Otro valor a tener en cuenta en la serie agrupada es el Punto Medio, este es importante porque representa al conjunto de valores del intervalo del cual es calculado. Se calcula:

Punto medio = Límite real inferior + Amplitud del Intervalo / 2

En el ejemplo, para el primer intervalo será:

Punto Medio = 10 + 10/2= 10 + 5= 15

Cabe señalar que en la organización de frecuencias no es necesario que se muestren todos estos cálculos (frecuencia relativa, frecuencia acumulada, frecuencia relativa acumulada, etc.) el investigador seleccionará los que sean necesarios para demostrar su hipótesis. *En el ejemplo, hemos puesto todos estos cálculos por fines didácticos*.

Observaciones a tener en cuenta:

- Algunas veces, con este procedimiento no se obtiene una amplitud del intervalo muy conveniente, se debe usar el sentido común para elegir la amplitud.
- Algunas reglas empíricas recomiendan, que si los datos lo permiten, la amplitud del intervalo sea de 5 ó 10 unidades, ya que estas hacen el resumen más comprensible.
- El límite inferior del primer intervalo debe contener a la medición más pequeña y el límite superior del último a la medición más grande.

PRESENTACIÓN DE DATOS

Dra. Julia Piscoya S.

Efectuada la recolección de datos estos deben ser sometidos a tratamiento estadístico y deben seguir los siguientes pasos: descripción, análisis y generalización.

Para la descripción se utiliza tres formas de presentación:

- Tabular
- Gráfica
- Medidas resumen numérico

1.- PRESENTACION TABULAR

La presentación tabular es básica, insustituible y fundamental; es donde se reflejan los conceptos e hipótesis que plantea el investigador. Se utiliza tanto para las variables cualitativas como para las cuantitativas

PARTES DE UNA TABLA

- Título
- Talón
- Cuerpo
- Notas explicativas

T í t u l o

Talón ⇓	Cuerpo ⇓

*Notas explicativas

TIPOS DE TABLAS

- Tablas específicas

N° orden	Edad
1	3
2	4
3	2
4	1
5	4

Edad	Fc	%
1	2	20
2	3	30
3	2	20
4	1	10
5	2	20
TOTAL	10	100

Edad	Fc	%
1-2	5	10
3-4	7	14
5-6	18	36
7-8	13	26
9-10	7	14
TOTAL	50	100

- Tablas de contingencia (En el ejemplo una tabla de 2X2)

RESULTADO	VACUNADO	NO VACUNADO	TOTAL
Sano			
Enfermo			
TOTAL			

REQUISITOS QUE DEBE TENER UNA TABLA:

1. Ser lo más simple posible, es mejor 2 ó 3 tablas sencillas a una muy compleja.
2. Debe explicarse por si misma, por eso:

- Si se usan abreviaturas o símbolos deben aclararse en las notas explicativas.
- Cada fila y columna deben estar tituladas clara y concisamente
- El título debe ser claro, conciso y exacto, debe responder a las preguntas ¿Qué? ¿Dónde? ¿Cuándo? Y en algunos casos ¿Cómo?. Asimismo, debe consignarse el número de tabla.
- Deberán colocarse los totales, se dispondrán en la última fila inferior y en la última columna de la derecha.

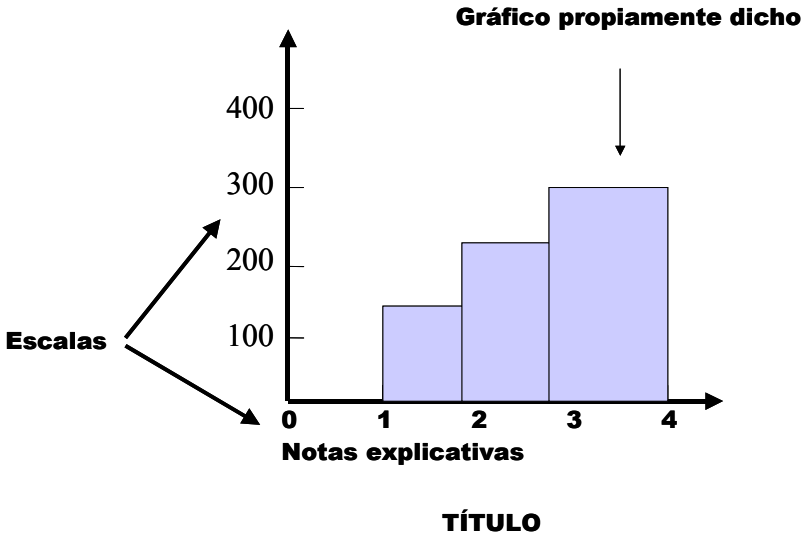
3. Si los datos no son originales debe mencionarse la fuente en las notas explicativas

2.- PRESENTACIÓN GRÁFICA

Es la forma de exponer los datos de manera que permita su comprensión global y de una manera rápida permite una impresión panorámica del material presentado.
El gráfico depende del tipo de variable y de la escala de medición que se ha utilizado.

PARTES DE UN GRÁFICO

- Título
- Escalas
- Gráfico propiamente dicho
- Notas explicativas



REQUISITOS QUE DEBE TENER UN GRÁFICO

1. El tipo de gráfico que alcance su objetivo con la mayor sencillez será el más efectivo. No debe contener más líneas o símbolos que los que el ojo pueda seguir cómodamente.
2. Todo gráfico debe explicarse por sí mismo; por eso debe indicarse claramente título, origen, escalas y leyendas.
3. No deben indicarse más ejes coordenados que los necesarios.
4. Las líneas del gráfico deben ser más gruesas que los ejes.
5. Por lo general, la frecuencia se presenta en el eje vertical y el método de clasificación en el eje horizontal. La escala de las frecuencias debe comenzar en 0 (Excepción del gráfico semilogarítmico que empieza en 1).

Además de estos requisitos, cada tipo de gráfico tiene sus particularidades que se deben de tener en cuenta en el momento de construirlos.

Antes de seleccionar el gráfico debemos tener en cuenta el tipo de variable, qué escala de medición se ha utilizado, cuál es el propósito que se persigue con la construcción; es decir queremos mostrar las frecuencias, queremos mostrar la proporción de determinados datos, queremos mostrar cómo evoluciona la variable en relación al tiempo. El siguiente cuadro nos dará algunas ideas para la selección, hay que señalar que hay otros tipos de gráficos (de caja o boxplot, de hojas, de correlación, etc), pero los que aquí se mencionan son los que se usan con mayor frecuencia.

SELECCIÓN DEL GRÁFICO DE ACUERDO AL TIPO DE VARIABLE

TIPOS DE DATOS	VARIABLE	TIPO DE GRÁFICO
DISTRIBUCIONES DE FRECUENCIA	Cualitativa	Barras: simples y todas sus variedades Gráficos circulares Pictogramas
	Cuantitativa discreta	
	Cuantitativa Continua	Histogramas Polígonos de frecuencia
TENDENCIAS	Cuantitativa	Curvas Gráficos lineales Gráficos Semilogarítmicos

Gráfico de Barras Horizontales

Figura N° 1. Hogares visitados según distrito. Campaña de Educación para la Salud. Lima, 2004

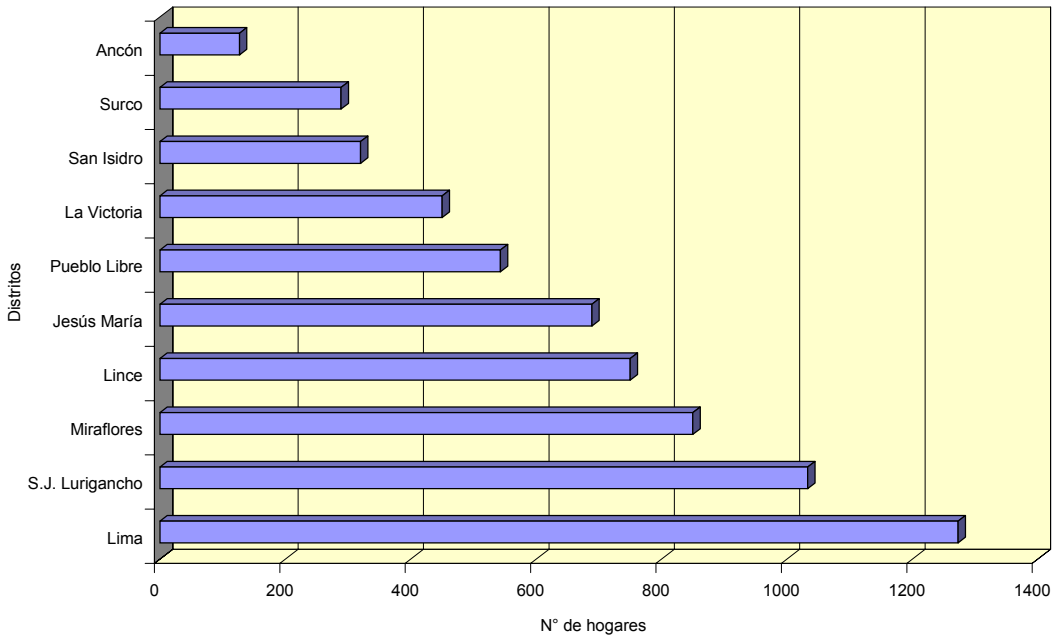
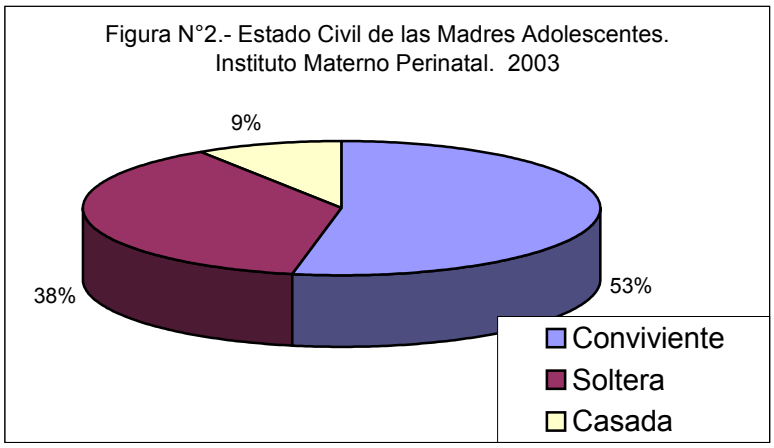
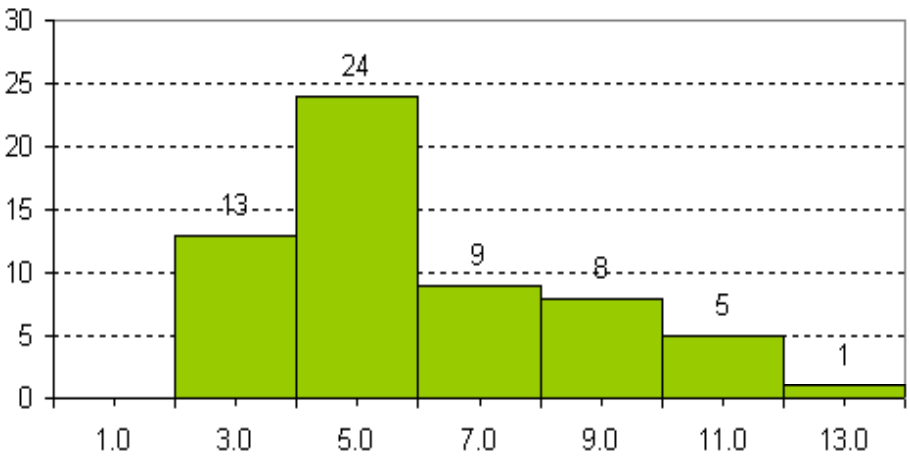


Gráfico Circular



Histograma



Polígono de frecuencia

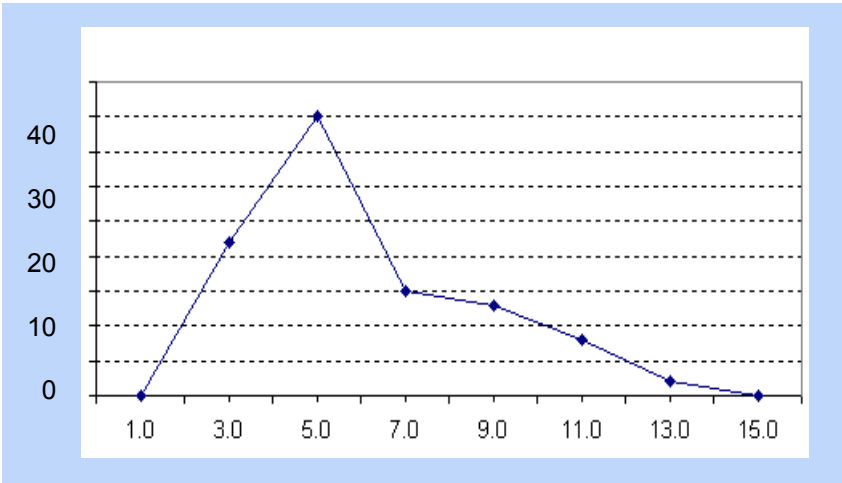
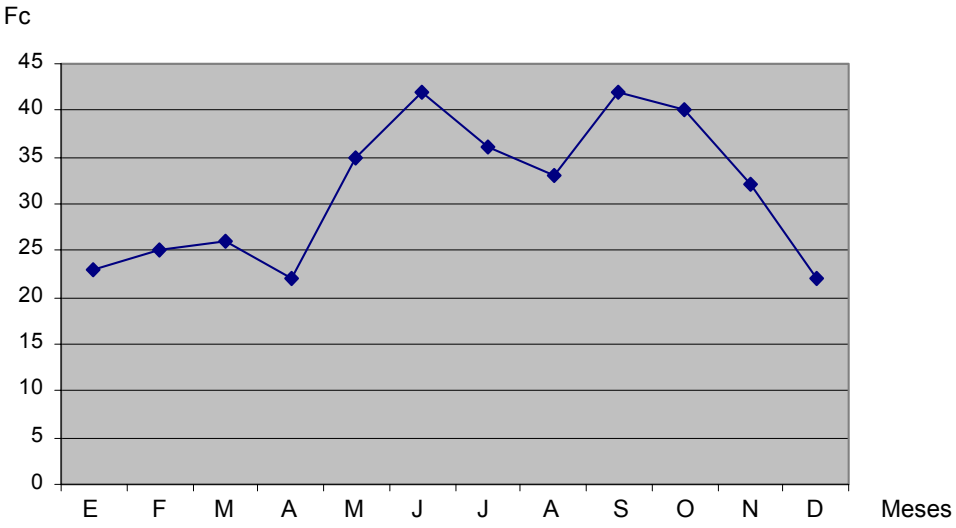


Gráfico lineal



Existen otros gráficos como el gráfico de tallo y hojas (stemplot), el gráfico de caja (boxplot) usados en el análisis exploratorio de datos.

GRÁFICO DE TALLO Y HOJAS

- Se utiliza en el análisis exploratorio de datos
- Muestra la distribución de datos cuantitativos.
- Tiene gran similitud con el histograma
- No se pierden los datos individuales
- Es fácil notar la mayor concentración de los datos
- En su construcción se usan los datos originales
- Se observa con facilidad los valores máximo y mínimo.
- Son más eficientes en conjuntos relativamente pequeños de datos

Ejemplo:

Edad de pacientes en un estudio sobre diabetes:

54-59-35-41-46-25-47-60-54-46-49-46-41-34-22

Ordenamos los datos:

22, 25, 34, 35, 41, 41, 46, 46, 46, 47, 49, 54, 54, 59, 60

2	**2 5**
3	**4 5**
4	**1 1 6 6 6 7 9**
5	**4 4 9**
6	**0**

Tallo **Hojas**

MEDIDAS DE RESUMEN NUMÉRICO

Dra. Elsy Cuellar Fretel *Dra. María Teresa Perales Díaz*

El médico en el desempeño de su profesión actuará a nivel individual y a nivel de comunidad, en el primer caso llegará a un diagnóstico y a un plan de tratamiento para un solo paciente mediante la historia clínica, el examen físico y pruebas de laboratorio; pero desempeñar esta misma función en el segundo caso, es decir en el campo de la salud pública, exige utilizar herramientas y técnicas estadísticas adecuadas para evaluar el estado de salud de una comunidad. Tradicionalmente, estos medios consisten en los datos demográficos que incluyen el número de nacimientos, muertes, enfermedades y diversas mediciones que pueden calcularse a partir de ellos.

Cuando tenemos un conjunto de datos y ya los hemos ordenado y clasificado (distribución de frecuencias) es importante que con uno o dos valores podamos tener una idea del conjunto de los datos.

La tarea de resumir consiste en presentar, en lugar de toda la distribución, solamente unas pocas características que indiquen los aspectos fundamentales de la distribución de frecuencias.

Estas medidas de resumen varían de acuerdo al tipo de variable y los datos que éstas generen. En el caso de los datos cualitativos, cuál es la relación, cuál es la mayor proporción de un determinado valor. En el caso de los datos cuantitativos, cuál es la regularidad (medidas de tendencia central) y cuál es la variabilidad (medidas de dispersión) de los datos estudiados.

MEDIDAS DE RESUMEN NUMÉRICO PARA VARIABLES CUALITATIVAS

Las medidas de resumen numérico empleadas para variables cualitativas son:

- RAZON
- PROPORCIÓN
- TASA

RAZÓN

Es la comparación por cociente entre dos cifras de diferente o similar naturaleza, en donde el numerador y el denominador son excluyentes.

Por ejemplo, si tengo 380 camas hospitalarias y 95 enfermeras y quiero encontrar la razón entre ellas, tengo que dividir:

380 camas hospitalarias / 95 enfermeras = 4

Este número constituye un valor que refleja una relación. En este caso, el número 4 se interpreta como que por cada cuatro camas hospitalarias hay una enfermera.

Otro ejemplo, en la enfermedad del SIDA en el Perú, en 1987 por cada caso notificado en una mujer se habían registrado 25 casos en varones a diferencia de 1998 donde la razón disminuye a 1 de 4 (Una mujer infectada por cada 4 varones infectados)

La manera correcta de expresar el resultado del cálculo de una razón es señalar el número de elementos del numerador que existen por cada elemento del denominador.

PROPORCIÓN

Es la comparación por cociente entre el número de elementos de un subconjunto y el número de elementos de un conjunto al que pertenece dicho subconjunto. En este caso el numerador esta incluido en el denominador, por este motivo los valores siempre van a ser menores que la unidad.

Por ejemplo, si en la población existen 175 casos de cáncer pulmonar de un total de 1925 casos de todos los tipos de cáncer, la proporción se calculará

175 / 1925 = 0.09

Este valor indica la magnitud o importancia del subconjunto de casos de cáncer pulmonar entre todos los tipos de cáncer.

Si en otra población los casos de cáncer de pulmón fueran 194 y los casos totales de cáncer fueran 13 486

194/13 496 = 0,01

Este nuevo cálculo de la proporción en la segunda población, no permite comparar la magnitud o importancia del cáncer pulmonar entre las dos poblaciones, no podemos afirmar que la primera población tiene mayor riego de cáncer pulmonar que la segunda en base a las proporciones obtenidas, a pesar de que 0,01 es menor que 0,09.

Como hemos expresado, los valores que se obtienen en el cálculo de las proporciones son menores que la unidad, lo cual no es muy práctico, por esta razón estas proporciones son multiplicadas por 100 obteniéndose los porcentajes, facilitando la comprensión y comunicación.

TASA

Cuando los clínicos dicen que una enfermedad es frecuente y otra es rara presuponen una diferencia en las tasas.
Todos los clínicos saben que la enfermedad coronaria es mucho más frecuente en un hombre de mediana edad que en una adolescente. Saben que el cáncer del páncreas es mucho más común en las personas de edad avanzada que en los jóvenes. Saben que la anemia de células falciformes es mucho más probable en una persona de raza negra que en una de raza blanca. El médico puede apreciar el significado de las tasas sobre la base de su experiencia clínica personal y la valoración científica y objetiva de artículos de investigación.
¿Qué es una TASA?

Es la comparación por cociente entre un número de eventos ocurridos en un tiempo y lugar determinados y la población que estuvo expuesta al riesgo de que le ocurrieran dichos eventos en el mismo tiempo y lugar.

Otra definición dice que: la tasa es una proporción en la que el denominador representa a la población expuesta al riesgo de sufrir un daño en un lugar y tiempo determinados. En la composición de la tasa tiene mucha importancia el denominador, del cual se supone que salen los casos que conforman el numerador.

La Tasa es una probabilidad o frecuencia relativa o proporción, en la cual el numerador es el número de veces que ocurre un suceso y el denominador es el número de veces que podría haber ocurrido. Como en todas las proporciones el numerador está incluido en el denominador. Las tasas realmente son un tipo especial de medida en la que el denominador también incluye una unidad de tiempo.

En epidemiología, las mediciones más empleadas están referidas a la magnitud del daño, la velocidad de propagación y las asociaciones. Las dos primeras se expresan por tasas.

La tasa está constituida por tres elementos:

1) El numerador del cociente, que consiste en el número de veces que ocurrió el evento en un lugar y tiempo determinado.

2) El denominador del cociente que es la población expuesta al riesgo de que le ocurra el fenómeno en el mismo lugar y tiempo.

3) Una constante por la cual se multiplica el resultado del cociente. Debido a que usualmente la división del numerador entre el denominador resulta en una cifra inferior a la unidad el resultado suele multiplicarse por 100, 1000, 10000 ó 100000 para una mejor comprensión y fácil lectura.

La fórmula matemática corresponde a:

$$\text{TASA} = \frac{\text{Nº de veces de ocurrencia de un evento en un lugar y tiempo determinados}}{\text{Pob. expuesta al riesgo de sufrir el evento, en el mismo lugar y tiempo}} \times 10^{n}$$

TASAS DE USO FRECUENTE

En medicina, una función importante de las tasas y de las proporciones es la de caracterizar la historia natural de la enfermedad. Con frecuencia se usan tres tipos de medidas:

1) Tasa de prevalencia: toma en cuenta los casos antiguos y nuevos, mide la probabilidad de tener una enfermedad en un momento dado. La prevalencia sólo proporciona una idea de magnitud del problema.

2) Tasa de incidencia: sólo toma en cuenta los casos nuevos, los cuales provienen de la población expuesta, delimitada al inicio del período de observación. Esta característica hace que la Incidencia tenga un poder predictivo mayor que el de la tasa de prevalencia. Una incidencia alta nos

informa que el problema se extenderá en poco tiempo a toda la población susceptible.

3) Tasa de Letalidad: probabilidad de morir por una enfermedad durante un espacio de tiempo a partir de su diagnóstico.

(Ver construcción de fórmulas en Cuadro de Indicadores de Morbilidad y Mortalidad)

AJUSTE DE TASAS

Una tasa permite expresar de manera cuantitativa y sintética la relación entre un evento y la población en que dicho evento puede ocurrir. En este sentido, una tasa refleja la concurrencia de toda una constelación de factores que influyen para que el resultado de la tasa sea mayor o menor.

Así por ejemplo, al encontrar que la tasa de intoxicación crónica por plomo es más alta en una población que en otra, de inmediato se evoca una imagen mental en la que la población con tasa alta tiene condiciones que favorecen el contacto a la exposición con dicho metal, tal vez más repetida o intensamente que la población con tasa baja. Algunas tasas suelen ser consideradas como indicadores que reflejan condiciones ambientales y se emplean para comparar el riesgo que una población tiene con respecto a otra de padecer problemas colectivos de salud derivados de la contaminación.

Sin embargo, la comparación de tasas puede perder casi totalmente su validez si no se efectúan procedimientos que corrijan el importante efecto que suele tener la diferente estructura, respecto a una característica (grupos de edad, sexo, etc.), de las poblaciones a comparar; a ese procedimiento que permite una buena comparación entre dos poblaciones diferentes se llama AJUSTE DE TASAS.

Veamos por ejemplo:

1. ANTECEDENTES

MORTALIDAD POR EDADES EN DOS CIUDADES

Edad (años)	Población Ciudad A	Defuncio-nes en A	**Tasa de Mortalidad en A por 1000**	Población Ciudad B	Defuncio-nes en B	**Tasa Mortalidad en B por 1000**
0-14	500	2	**4,0**	400	1	**2,5**
15-20	2000	8	**4,0**	300	1	**3,3**
30-44	2000	12	**6,0**	1000	5	**5,0**
45-59	1000	10	**10,0**	2000	18	**9,0**
60-74	500	10	**40,0**	2000	70	**35,0**
75 y más	100	15	**150,0**	400	50	**125,0**
Total	6100	67	**11,0**	6100	145	**23,77**

Según esta tabla vemos que la mortalidad en la ciudad B es casi el doble que en la ciudad A, así también vemos que la composición de las poblaciones es diferente, por lo que es necesario hacer el ajuste de tasas, en este caso se hará el ajuste por edad.

AJUSTE DE TASAS

1.- Construcción de "Población tipo

La población tipo puede ser cualquiera de las dos poblaciones, la A, la B, o la suma de ambas, generalmente se usa la suma de ambas, como lo vamos a ver en el ejemplo.

Edad	Población A	Población B	Población tipo AB
0-14	500	400	900
15-29	2000	300	2300
30-44	2000	1000	3000
45-59	1000	2000	3000
60-74	500	2000	2500
75 y más	100	400	500

A la población tipo AB se le aplica las tasas de mortalidad específica de A y de B y se tiene el Nº de muertes esperadas que habría ocurrido en la población tipo si ésta hubiese estado en las condiciones de A o de B. Con este número de muertes se calcula la tasa de mortalidad general ajustada.

2.- Las defunciones teóricas se calculan por un despeje de la fórmula de la tasa de mortalidad:

$$\text{Tasa de mortalidad} = \frac{\text{N° de defunciones}}{\text{Población}} \times 1000$$

$$\text{N° de defunciones} = \frac{\text{Tasa de mortalidad} \times \text{Población}}{1000}$$

Edades (años)	N° de individuos	Tasa Población A	N° de muertes Esperadas
0-14	900	4,0	3,6
15-29	2300	4,0	9,2
30-44	3000	6,0	18,0
45-59	3000	10,0	30,0
60-74	2500	40,0	100,0
75 y más	500	150,0	75,0
TOTAL	12200	19,32	235,8

Edades (años)	N° de individuos	Tasa Población B	N° de muertes Esperadas
0-14	900	2,5	2,25
15-29	2300	3,3	7,59
30-44	3000	5,0	15,0
45-59	3000	9,0	27,0
60-74	2500	35,0	87,0
75 y más	500	125,0	62,5
TOTAL	12200	16,54	201,84

3.- Con estas muertes esperadas se calcula la tasa para cada una de las ciudades.

La tasa ajustada para la población A es de 19,32 por 1000 y para la población B es de 16,54 por 1000. Como podemos apreciar, estas cifras son completamente distintas de las primeras que vimos en la tabla, porque en estas últimas se controló el factor edad.

INDICADORES DE SALUD

MORTALIDAD

Medida	Numerador	Denominador	Unidad Poblacional (10^n)
Tasa de mortalidad	Nº muertes durante un período de tiempo	Población entre la que ocurrieron las muertes.	1.000 ó 100.000
Tasa cruda o bruta de mortalidad.	Nº total de muertes durante un período de tiempo	Población a mitad del período.	1.000 ó 100.000
Tasa de mortalidad por causas.	Nº muertes asignadas a una causa durante un período.	Población a mitad del período.	100.000
Tasa de Mortalidad proporcional	Nº de muertes asignadas a una causa específica durante un período.	Nº total de muertes por causas durante el mismo período.	100 ó 1.000
Tasa de mortalidad neonatal	Nº total de muertes por debajo de 28 días de edad durante un período.	Nº de nacidos vivos durante el mismo período.	1.000
Tasa de mortalidad infantil.	Nº de muertes por debajo de 1 año de edad durante un período.	Nº de nacidos vivos durante el mismo período.	1.000
Tasa de mortalidad materna	Nº de muertes asignadas a causas relacionadas con el embarazo, parto y puerperio.	Nº de nacidos vivos durante el mismo período.	10.000 ó 100.000
Tasa de Letalidad	Nº de muertes por una enfermedad durante un período	Nº de casos diagnosticados con la misma enfermedad al inicio del período	100

MORBILIDAD

Medida	Numerador	Denominador	Unidad poblacional (10^n)
Tasa de incidencia	Nº de casos nuevos de enfermedad en un período determinado de tiempo.	Población sujeta a riesgo en el mismo período.	1.000 ó 100.000
Tasa de prevalencia	Nº de casos antiguos y nuevos de una enfermedad en un período determinado.	Población sujeta a riesgo en el mismo período.	1.000 ó 100.000
Tasa de ataque	Nº de casos nuevos de enfermedad en un período epidémico.	Población expuesta al inicio del período de estudio.	1.000
Tasa de ataque secundario	Nº de casos nuevos en contactos de casos conocidos.	Población de contactos a riesgo	1.000

MEDIDAS DE RESUMEN PARA VARIABLES CUANTITATIVAS

Lic. Esperanza García C.

Antes de aplicar cualquier técnica estadística para resumir datos, se debe realizar un análisis exploratorio de los mismos, con la finalidad de:

1. Evaluar la calidad. Es este momento se descubre por ejemplo datos no registrados o valores discordantes, a los cuales, por separado, se les dará una solución o explicación según sea el caso.

2. Determinar si los datos siguen una distribución normal o por lo menos con tendencia a ésta. La distribución normal concentra la mayoría de valores al centro (campana); en la distribución que no es normal, la concentración de valores se dan a la derecha o a la izquierda de la distribución. La normalidad se puede apreciar construyendo un gráfico de tallo y hojas, de caja o un histograma.

La normalidad o no de las distribución permitirá seleccionar las medidas de resumen más convenientes para la descripción respectiva.

Las medidas para resumir datos correspondientes a variables cuantitativas son: medidas de tendencia central, medidas de dispersión, medidas de posición o localización.

MEDIDAS DE TENDENCIA CENTRAL

Son valores que indican el centro de la distribución de las observaciones referentes a variables cuantitativas continuas o discontinuas.

En el área biomédica, la experiencia indica que para las variables medidas en escala de razón los datos tienden a concentrarse alrededor de un sector de la variable. Se trata entonces, de aceptar determinados criterios para representar con un valor de la distribución esa tendencia de las observaciones, que se llama tendencia central.

Estas medidas se pueden calcular a partir de los datos de una muestra o de una población:

- Una medida descriptiva calculada a partir de una muestra se llama *estadístico.*
- Una medida calculada a partir de los datos de una población se llama *parámetro*.

Las tres medidas de tendencia central usadas con más frecuencia en el área biomédica son: la media, la mediana y el modo.

MEDIA ARITMÉTICA

Llamada también promedio, resulta de sumar los valores de todas las observaciones y dividir la sumatoria entre el total de ellas. Se caracteriza por ser única, fácil de calcular y porque es afectada por todos y cada uno de los valores del conjunto, de tal manera

que los muy grandes o muy pequeños que salen del rango esperado pueden distorcionarla, en tal caso, el valor discordante se analizará por separado. Ejemplo; si se analiza un conjunto de datos de la variable talla (cm) de un grupo de varones adultos, donde uno de ellos mide 230 centímetros, el valor discordante será 230, éste distorcionará la media, luego, para evitar esa inconveniencia será mejor analizarlo por separado o de lo contrario, se debe calcular una mediana. La media se calcula con las siguientes fórmulas:

a.- A partir de una muestra (estadístico)

$$\overline{X} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

b.- A partir de una población (parámetro):

$$\mu = \frac{\sum x_i}{N}$$

Donde:
N es la población
n es la muestra
$x_1, x_2, x_3, \ldots x_n$ son los valores de la variable
$\overline{X}$ es la media.

Ejemplo 1

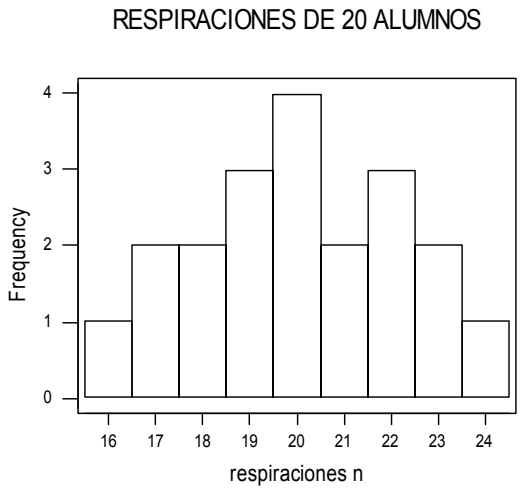
En la práctica del curso de Bioestadística, del II Semestre Académico de 2004, los estudiantes del tercer año, entre otros datos, contabilizaron el número de respiraciones por minuto en situación de reposo. Se desea saber la media de las respiraciones de los estudiantes. Los datos se presentan a continuación:

19	19	17	20	16	14	20	21	16	15	20	20	20	16
24	19	19	20	18	20	15	18	18	16	22	24	19	17
16	20	20	17	20	20	22	20	18	20	18	16	21	22
24	16	24	15	19	20	20	15	21	23	21	24	21	24
20	24	18	17	18	18	20	17	22	17	16	19	20	21
14	22	21	22	19	21	18	26	18	16	17	21	22	17
20	23	28	22	23	18	16	24	22	20	18	22	13	20

Población de alumnos: 98, la media se calculará de la siguiente manera:

PASOS PARA CALCULAR LA MEDIA

1. Se verifica la normalidad de la distribución de los datos, en este caso usamos un histograma, se observa que los datos tienen una distribución aproximadamente normal, entonces podemos calcular la media.



$$\mu = \frac{\sum x_i}{N}$$

$$\mu = \frac{19 + 19 + ... + 20}{98} = 19.5$$

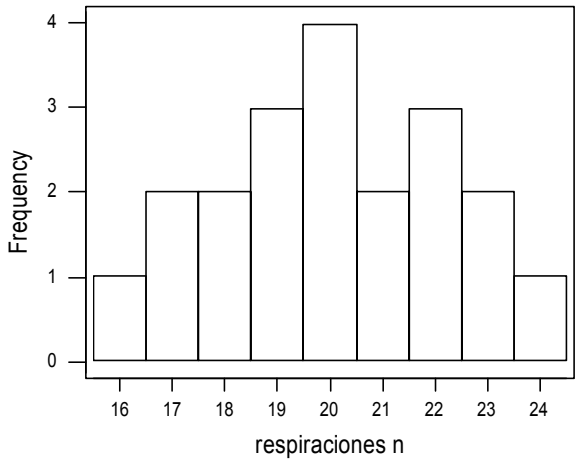
INTERPRETACIÓN: Los alumnos del tercer año que llevaron la asignatura de Bioestadística el año 2004, tuvieron en promedio 20 respiraciones por minuto.

Ejemplo 2:

De la misma población se obtuvo una muestra de 20 alumnos para calcular el promedio de las respiraciones en las mismas condiciones. Los datos se presentan a continuación.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
19	16	20	22	21	24	23	19	22	17	20	20	20	21	18	22	18	17	19	23

RESPIRACIONES DE 20 ALUMNOS



En el gráfico observamos la tendencia a simetrìa de la distribución, por lo tanto la media es la medida de resumen adecuada.

$$\overline{X} = \frac{\sum x_i}{n} = \frac{19+16+...+18}{15} = \frac{302}{15} = 20.05$$

INTERPRETACIÓN:
Los alumnos tuvieron en promedio 20 respiraciones por minuto

MEDIANA

Es el valor que divide al conjunto ordenado de datos en dos grupos de igual tamaño en cuanto al número de observaciones se refiere. El primero será igual o menor que la mediana y el otro igual o mayor. Se usa con datos ordinales o con numéricos de distribución normal preferentemente. La mediana de un conjunto de datos se

caracteriza por ser única, su cálculo es muy fácil y a diferencia de la media los valores extremos no afectan su valor.

Pasos:

1. Los datos se ordenan en forma creciente: $x_1 + x_2 + \ldots + x_n$
2. Calcular la posición de la mediana teniendo en cuenta la fórmula:

$$Me = \frac{n+1}{2}$$

Donde:

Me es la mediana
n es el tamaño de la muestra

3.- Se establece el valor de acuerdo a la posición calculado, teniendo en cuenta si "n" es par o impar.

Ejemplo 3:

Con los datos usados para obtener la media, ahora calculamos la mediana.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
19	16	20	22	21	24	23	19	22	17	20	20	20	21	18	22	18	17	19	23

1.- Se ordenan los datos de menor a mayor,

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
16	17	17	18	18	19	19	19	20	20	20	20	21	21	22	22	22	23	23	24

2.- Calcular la posición de la mediana teniendo en cuenta la fórmula:

$$Me = \frac{n+1}{2} \qquad Me = \frac{20+1}{2} = 10.5$$

3.- Como "n" es par, la posición de la mediana es 10.5, en este caso el valor de la mediana se localiza entre los dos valores centrales de la distribución.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
16	17	17	18	18	19	19	19	20	20	20	20	21	21	22	22	22	23	23	24
									Me										

Valor de la mediana: promedio de los valores que se encuentran en las posiciones diez y once, es decir

Valor de la $Me = \frac{20+20}{2} = 20$ respiraciones por minuto

INTERPRETACIÓN: El 50% de los alumnos tuvieron 20 respiraciones o menos y el otro 50% , 20 respiraciones o más.

3.1.-Si *n* es impar:

$$Me = \frac{n+1}{2} \qquad Me = \frac{21+1}{2} = 11$$

La posición de la mediana se encuentra en el onceavo lugar

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
16	17	17	18	18	19	19	19	20	20	20	20	21	21	22	22	22	23	23	24	25

Me ;

El valor de la mediana será el dato que se encuentre en el centro de la distribución, en este caso es 20.

INTERPRETACIÓN: El 50% de los estudiantes, tuvieron 20 respiraciones o menos y el otro 50% 20 respiraciones o más.

MODA

Valor que se presenta con mayor frecuencia en un conjunto de datos. Se usa solamente cuando se tiene interés en resaltar el o los valores más frecuentes. Un conjunto de datos puede tener más de una moda o ninguna.

Ejemplo 4:

Una muestra de 17 alumnos, ingresantes a la universidad, fueron sometidos a un examen bucodental para determinar la presencia de alguna enfermedad oral. Entre otros datos se registró la edad de cada uno de ellos, los cuales se presentan a continuación, ¿Cuál es el valor modal?

Alumno	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Edad	16	15	17	18	18	16	18	15	18	19	18	17	17	16	19	20	16

Para una mejor visualización del valor más frecuente se ordenan los datos:

Alumno	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Edad	15	15	16	16	16	16	17	17	17	18	18	18	18	18	19	19	20

Mo

El valor modal de la edad fue 18, pues, el dato se repite 5 veces.

CUANTILES

Se conocen también como medidas de localización. Se usan con datos numéricos sesgados o cualitativos medidos en escala ordinal.

1.- PERCENTILES (P)

Son 99 valores que dividen a un conjunto de datos en 100 partes iguales. Un percentil indica el porcentaje de los valores de un conjunto de datos que es menor o igual al valor de un determinado percentil. Su importancia radica en su uso para comparar un valor individual con una norma. Se usa intensamente en la interpretación y desarrollo de tablas de crecimiento físico, y de mediciones de destreza de inteligencia, así como también para determinar rangos normales de valores de laboratorio. Los límites

normales para la mayoría de los análisis de laboratorio se establecen en los percentiles 2.5 y 97.5, de modo que estos límites normales contienen el 95% central de la distribución. Los percentiles se emplean cuando se usa la mediana, también se emplea cuando se usa la media, pero el interés es comparar un valor individual de la variable con un conjunto de normas. Por ejemplo, comparar el peso de un niño de 24 meses con lo establecido para esa edad en una tabla de control del niño sano.

La fórmula para calcular percentiles es:

$$P_k = \frac{k(n+1)}{100}$$

Donde:

k es el número del percentil
n es la muestra
P_k Es el percentil que se desea calcular.

Con esta fórmula se calcula la posición que tiene el percentil *k* en el arreglo ordenado, luego se procede a ubicar el valor de la variable en la posición que le corresponde.

Ejemplo:

Calcular el percentil 90 en la distribución de los niveles de glucosa de 100 niños.

1. Ordenar los datos de menor a mayor:

50	55	55	55	56	56	56	57	57	57	57	58	58	59	59	59	60	60	60	61
61	61	61	62	62	62	62	62	63	63	63	63	64	64	64	64	65	65	65	65
65	65	65	65	65	65	65	65	66	66	66	66	66	67	67	67	67	67	67	68
68	68	68	68	68	68	69	69	69	69	71	71	72	72	72	72	73	73	73	73
73	73	73	74	74	75	75	75	75	75	75	76	76	77	**79**	**80**	80	80	81	81

P_{95}

2. Calcular la posición del percentil

$$P_{95} = \frac{95(n+1)}{100} = 95.95 \ posición$$

El percentil 95 (P_{95}) es un valor que está ubicado en la posición 95.95 del segmento de datos, entonces hay que calcular el valor de la variable en esa posición haciendo exrtapolación.
A la posición 95 le corresponde el valor 79 y a la 96 el valor 80, a partir de estos valores se obtendrá el valor del percentil 95, finalmente:

P_{95} = 79+0.95(80-79) = 79.95

INTERPRETACIÓN: El 95% de los niños tuvieron un nivel de glucosa igual o menor que 79.95

USOS:

Se usa para comparar un valor individual con un conjunto de normas. Ampliamente se utiliza para desarrollar e interpretar tablas de crecimiento físico, mediciones de destreza e inteligencia y especialmente para determinar rangos normales de valores de laboratorio. Para muchos de los análisis, los límites normales están entre el

percentil 2.5 y 97.5, de modo que el 95% central de los valores se encuentran entre estos dos percentiles.

2.- CUARTILES

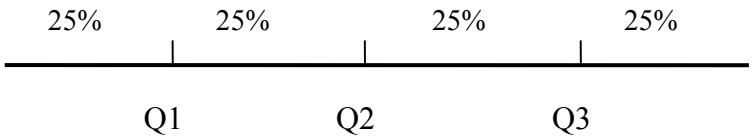
Son tres medidas de posición que dividen a un conjunto de datos cuyos valores están ordenados generalmente de menor a mayor en cuatro partes iguales. Se les nombra como Q_1 , Q_2 , Q_3. Las fórmulas para su cálculo son tres, a saber:

$$Q_1 = \frac{n+1}{4} \qquad Q_2 = \frac{2(n+1)}{4} \qquad Q_3 = \frac{3(n+1)}{4}$$

Donde:

Q_i es el cuartil que se desea calcular$_i$

n tamaño de muestra



Intervalo Intercuartilar (IQ)

Es la medida que describe el 50 % central de una distribución, sin importar su forma, no es afectada por las fluctuaciones extremas de la serie. Mide la dispersión de los valores de la variable alrededor de la mediana. Mientras más próximos estén sus límites, mayor será la concentración alrededor de ésta. Comprende entre el percentil 25 y 75, entre Q_1 y Q_3, tiene como centro el percentil 50, el cuartil 2 o la mediana.

$$IQ = Q_3 - Q_1$$

Desviación cuartilar (Q)

Es la mitad del intervalo cuartilar. Si la serie es perfectamente simétrica, la mediana es el punto que divide a la serie en dos partes iguales. Se calcula con la fórmula:

$$Q = \frac{Q_3 - Q_2}{2}$$

CÁLCULO DEL INTERVALO CUARTILAR Y DE LOS CUARTILES

Con los datos ordenados de mayor a menor se calcula los cuartiles 1 y 3. Usaremos los datos:

50	55	55	55	56	56	56	57	57	57	57	58	58	59	59	59	60	60	60	61
61	61	61	62	**62**	62	62	62	63	63	63	63	64	64	64	64	65	65	65	65
65	65	65	65	65	65	65	65	66	**66**	**66**	66	66	67	67	67	67	67	67	68
68	68	68	68	68	68	69	69	69	69	71	71	72	72	**72**	72	73	73	73	73
73	73	73	74	74	75	75	75	75	75	75	76	76	77	79	80	80	80	81	82
				Q_1						**Q_2**				**Q_3**					

$$Q_1 = \frac{100 + 1}{4} = 25.25 \Rightarrow \quad El \ \ valor \ \ Q_1 = 62$$

Interpretación: El 25% de los niños tienen un nivel de glucosa igual o menor que 62

El Q_2 es la mediana

$$Q_3 = \frac{3(n+1)}{4} = 75.75 \quad posición \quad \Rightarrow \quad El \quad valor \quad Q_3 = 72$$

Interpretación: El 75% de los niños tienen un nivel de glucosa igual o menor que 72.

Encontrar el intervalo cuartilar $IQ = Q_3 - Q_1 = 72 - 62 = 10$

Interpretación: **el 50% central de los niños tuvieron un nivel de glucosa entre 62 y 72.**

3. Dividir el valor del intervalo cuartilar entre 2 para obtener la desviación cuartilar

$$Q = \frac{10}{2} = 5$$

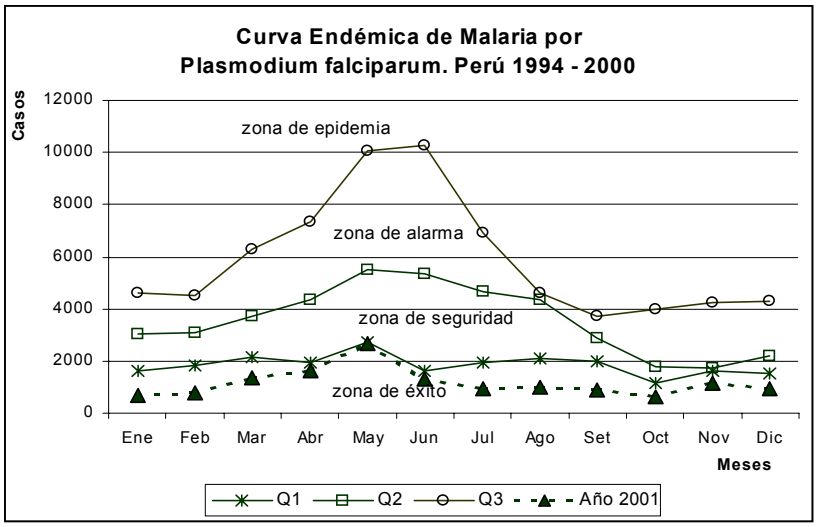
USO DE LOS CUARTILES:

- Para describir el 50% central de una distribución
- En epidemiología, para construir la curva endémica
- Para construir el gráfico de caja, útil en el análisis exploratorio de datos y para comparar poblaciones.

Ejemplo:

El siguiente gráfico es la curva endémica de malaria por Plasmodium falciparum del Perú en el año 1994 al 2 000).
La curva endémica es un patrón de comportamiento de la enfermedad a partir de la incidencia mensual de la enfermedad en varios años.

En el siguiente gráfico, Q_1 está representado por la línea inferior (*), Q_2 por la línea intermedia (), Q_3 por la superior (○) y la línea con triángulos representa el comportamiento de la enfermedad en el año 2001 (año que interesa estudiar).



MEDIDAS DE DISPERSIÓN

Son aquellas que miden la variabilidad de un conjunto de datos. La magnitud de la variabilidad es pequeña cuando los valores son diferentes pero están cercanos entre sí; si éstos son muy diferentes la dispersión es grande. Los sinónimos de dispersión son variabilidad y expansión. Ejemplo: si 10 estudiantes son pesados en una misma balanza bajo las mismas condiciones, y se encuentra que cada uno pesa 50 kilos, entonces no hay variación en los pesos., el peso es constante. Otro grupo de 10 estudiantes. fueron pesados en iguales condiciones, los pesos fueron: 55, 60, 53, 56, 48, 50, 51, 58, 62, 59, en este caso, se observa que son diferentes, entonces hay dispersión. Las medidas de dispersión que estudiaremos son: rango, varianza, desviación estándar y coeficiente de variación.

RANGO

Es la diferencia entre el valor máximo y el mïnimo de un conjunto de datos. Los demás valores se encuentran entre estos. Es una medida apropiada cuando se quiere enfatizar los valores extremos. Su uso es limitado por que toma en consideración solo la diferencia de dos valores. La fórmula para calcularlo es:

$$R = x_{máximo} - x_{mínimo}$$

Donde:

$x_{máximo}$ es el valor más grande lde la variable

$x_{mínimo}$ es el valor más pequeño dela variable

Ejemplo:
La siguiente serie corresponde a las edades años de una muestra de 11 niños.

2	5	6	8	11	14	15	17	21	24	26

R= 26 – 2 = 24

El rango, es decir la diferencia entre el valor máximo y el mínimo es 24, obsérvese que los demás valores están entre los valores extremos.

VARIANZA

Es la medida que cuantifica la variabilidad de los datos respecto al valor de la media. Si los valores de las distancias son iguales, el valor de la varianza es cero. Si los datos son diferentes pero cercanos entre sí, la varianza es pequeña. Si los datos están muy distantes, la varianza es grande. Se puede definir también como la sumatoria de las diferencias de cada uno de los datos con respecto a la media dividida entre ***n*-1.** Se calcula restando de cada observación el valor de la media; las diferencias se elevan al cuadrado, luego la sumatoria se divide entre ***n*-1** si los datos corresponden a una muestra, o, entre ***N*** si pertenecen a una población.
Las diferencias se elevan al cuadrado para desaparecer los signos negativas que se generan al restar la media a cada uno de los valores x_i, de esta manera se evita que la suma algebráica de éstas diferencias den como resultado cero. La varianza tiene las propiedades matemáticas necesarias para analizar mejor los datos en comparación a la desviación media, medida que se obtiene de sumar las diferencia de los valores $\mathbf{x_i}$ con su media, sin tomar en cuenta el signo y dividiendo la sumatoria entre el número de observaciones.

Cuando se trata de una muestra el símbolo de la varianza es $\mathbf{s^2}$ y cuando corresponde a una población σ^2 .

La fórmula para obtener la varianza cuando los datos no están agrupados es la siguiente:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n-1} = \frac{\sum (x - \bar{x})^2}{n-1}$$

Donde:
x_i representa los valores de la variable, x_1 , x_2 , …., etc.
n número de observaciones de la muetra
$\bar{x}$ es la media aritmética

La fórmula alternativa para un gran número de datos es:

$$s^2 = \frac{\left(x_1^2 + x_2^2 + ... + x_n^2\right) - n(\bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}{n-1}$$

USOS:

1. Se usa como elemento importante para realizar diferentes pruebas de inferencia estadística.
2. Sirve para calcular la desviación estándar, medida muy utilizada en las ciencias de la salud para analizar la variabilidad de los datos cuantitativos.
3. Sirve para calcular el tamaño de muestras cuando se requiere estudiar una variable cuantitativa.

PASOS PARA CALCULAR LA VARIANZA

Ejemplo:

Los datos de la siguiente tabla son los mismos del ejemplo 2 que se usaron para el cálculo de las media y mediana

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
19	16	20	22	21	24	23	19	22	17	20	20	20	21	18	22	18	17	19	23

1. Antes de aplicar cualquier técnica de resumen, es necesario un análisis previo de los datos que se dispone para evaluar las bondades de los mismos y solucionar problemas en el diseño de la investigación y en la recogida de los datos(ausentes y atípicos) . Las tareas que suelen realizarse en un análisis previo son: Análisis exploratorio y tratamiento de los datos ausentes y atípicos (outliers). En el ejemplo, se construye un gráfico de caja en el cual observamos que la distribución de los datos tiende a ser simétrica por lo tanto la medida de resumen más adecuada en este caso es la media y la desviación estándar.

También se puede apreciar que la mediana se ubica aproximadamente a la misma distancia del cuartil 1 y 3. No hay ningún dato que sea discordante (outlier) en el conjunto.

2.- Calcular la media aritmética:

$$\overline{X} = \frac{\sum x_i}{n} = \frac{19+16+...+23}{20} = \frac{401}{20} = 20.05$$ respiraciones por minuto

3.- Calcular la varianza, para lo cual se usará la fórmula que corresponde a una muestra, dado que es la medida con suficientes propiedades para usarla en inferencia estadística.

$$s^2 = \frac{(19-20.05)^2 + (16-20.05)^2 + ... + (23-20.05)^2}{20-1} = 4.89\ respiraciones^2$$

La varianza es 4.89 respiraciones^2. Se puede apreciar que la variabilidad de los datos es pequeña, debido a que estos son valores cercanos entre sí. La medida se expresa en unidades al cuadrado, y por lo tanto no se usa para su interpretación; sin embargo, a partir de ella podemos calcular la desviación estándar, medida muy usada en el análisis de datos en salud.

DESVIACIÓN ESTÁNDAR

Es la raíz cuadrada positiva de la varianza. Mide la variabilidad de los datos en las unidades en que se midieron originalmente. Los símbolos son: ***s*** si se trata de una muestra y σ^2 **;** si es una población. La fórmula es:

$$S = \sqrt{s^2}$$

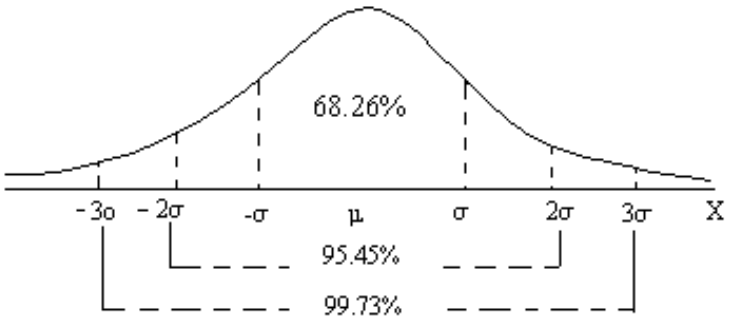
La desviación estándar se calcula cuando también es apropiado el cálculo de la media, es decir, cuando la variable es cuantitativa y además su distribución es o tiende a ser simétrica, la media se ubica al centro de la distribución o muy cercana a ella.

Características de la desviación estándar:

1. Siempre es un valor positivo
2. Está influenciada por todos los valores de la muestra o población. Mayor influencia ejercen los valores extremos que los que están cerca al promedio, debido a que son elevados al cuadrado en el cálculo.
3. Sirve para definir la dispersión de los datos alrededor de la media.

Si la distribución de la población sigue una distribución normal, en forma de campana (campana de Gauss), las observaciones se concentrarán en la parte central e incluirán, aproximadamente:

$\mu \pm 1\sigma$	**68%**
$\mu \pm 2\sigma$	**95%**
$\mu \pm 3\sigma$	**99%**



Estos valores son importantes a tener en cuenta cuando tenemos que hacer la interpretación de la desviación estándar.

CÁLCULO DE LA DESVIACIÓN ESTÁNDAR:

El cálculo es sumamente fácil, consiste en extraer la raíz cuadrada de la varianza. En el ejemplo se tiene que:

$$S = \sqrt{s^2} = \sqrt{4.89} = 2.21 \text{ respiraciones por minuto}$$

La descripción de las variables numéricas se hace se hace con los valores de la media y la desviación estándar, porque con estos dos valores tenemos una idea del conjunto de los datos ($\mu \pm 3\sigma$ incluirá el 99%). También nos dará la regularidad y la variabilidad de los datos.

INTERPRETACION: El 68% de los estudiantes tienen entre 17.84 y 22.26 (20.05 ± 2.21) respiraciones por minuto, o mejor aún, entre 18 y 22 respiraciones por minuto, por ser una variable cuantitativa discreta.

DESVIACIÓN MEDIA

Es una medida que expresa la forma en que las observaciones se dispersan alrededor de la media. Consiste en sumar las desviaciones de las observaciones respecto a su media y dividir la sumatoria entre ***n***. Es el promedio simple de las desviaciones, la fórmula es la siguiente.

$$DM = \frac{\sum(x_i - \bar{x})}{n}$$

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
19	16	20	22	21	24	23	19	22	17	20	20	20	21	18	22	18	17	19	23

El cálculo se hace considerando los valores absolutos de las desviaciones, se obvia los signos. La sumatoria de éstas diferencias se divide entre el número de observaciones.

$$DM = \frac{(19 - 20.05) + (16 - 20.05) + .. + 23 - 2.05)}{20} = 1.76$$

El promedio de las desviaciones de las observaciones respecto a la media es 1.76.

COEFICIENTE DE VARIACION

Es una medida de relativa, útil para comparar la dispersión en **dos o más conjuntos de datos,** los que pueden ser medidos en las mismas unidades o no. Expresa en porcentaje la relación de la desviación estándar y la media, la fórnula es:

$$CV = \frac{s}{\bar{x}} \times 100$$

La media y la desviación estándar se expresan en la misma unidad de medida, las que se anulan cuando se hace el cálculo, obteniéndose una medida independiente a la unidad de medición. El coeficiente de variación es útil también para comparar los resultados obtenidos por diferentes personas que efectúan investigaciones en las que se estudian la misma variable. Ejemplo: comparar la dispersión de los pesos de una muestra de sujetos obtenidos en libras con el peso de otra muestra expresada en kilogramos.

Si el coeficiente es:

< 10 % **se dice que hay poca dispersión**
10 – 33% **la dispersión es aceptable**
34 – 50% **dispersión es alta**
> 50% **la dispersión es muy alta**

CALCULO DEL COEFICIENTE DE VARIACION

$$\mathrm{CV} = \frac{2.212}{20.05} \times 100 = 11.03\%$$

La variación relativa de las respiraciones en los estudiantes fue 11.03%.

Ejercicios de repaso

En los siguientes ejercicios: Identifique la naturaleza de la variable y la escala de medición. Calcule: la media, mediana, desviación estándar y el coeficiente de variación. Interprete los resultados

1. En un programa para la detección de hipertensión, en una muestra de 30 hombres en edades entre 30 y 40 años, la distribución de la presión diastólica (mínima) en mm Hg fue la siguiente:

95	**90**	**70**	**100**	**65**	**80**	**90**	**95**	**90**	**95**	**110**	**100**	**85**	**80**	**75**
70	**85**	**85**	**75**	**65**	**90**	**110**	**95**	**90**	**70**	**60**	**75**	**80**	**120**	**85**

2. Niños atendidos diariamente, en una clínica pediátrica, durante el último trimestre del año 2004.

7	**10**	**12**	**4**	**8**	**7**	**3**	**8**	**5**	**7**	**12**	**11**	**3**	**8**	**1**
1	**13**	**10**	**4**	**6**	**4**	**5**	**5**	**8**	**7**	**7**	**3**	**2**	**3**	**5**
8	**13**	**1**	**7**	**17**	**3**	**4**	**5**	**5**	**4**	**3**	**1**	**17**	**10**	**4**
7	**7**	**12**	**8**	**3**										

ANÁLISIS EXPLORATORIO DE DATOS

Dr. Jorge Alarcón V.

El análisis de los datos puede realizarse de dos maneras, de acuerdo al uso de la estadística será:

- Estadístico
- No estadístico (cualitativo)

La estadística cumple con algunas funciones como:

- Producir datos
- Interpretar datos:
    - Existentes
    - Producidos de acuerdo a un plan o diseño

Los datos contienen información acerca de las características de un conjunto de individuos, expresadas como VARIABLES. Estos datos provienen de diversas fuentes como:

- Registros continuos
- Muestreo (encuestas)
- Experimentos (diseños controlados)
- Censos, etc

ANÁLISIS ESTADÍSTICO

El análisis estadístico sigue una secuencia ordenada de procedimientos; primero evalúa el valor de los datos, para ello analiza las fuentes y examina la validez, exactitud, consistencia; asimismo examina sus características, construye modelos, así como extrae el conocimiento que brindan los mismos generando "información"

ENFOQUES DEL ANÁLISIS ESTADÍSTICO

El análisis estadístico tiene dos enfoques: el confirmatorio o clásico y el exploratorio (Tukey, 1977). Ambos enfoques tienen sus particularidades, según Bertrand, podriamos hacer la siguiente comparación.

EXPLORATORIO	CONFIRMATORIO
Enfoque descriptivo	Enfoque inferencial
Indica las hipótesis a probar	Prueba hipótesis
Usa estadísticos resistentes	Usa estadísticos sensibles
Plan de investigación flexible y poco definido	Plan de investigación riguroso y bien definido.
Usa los datos disponibles	Usa datos sin error (ideal)
Privilegia la representación gráfica.	Poca importancia a la representación gráfica.
Tiene visión intuitiva de los datos.	Tiene una visión precisa de los datos.
Semeja una investigación policial.	Semeja un juicio

■

ANÁLISIS EXPLORATORIO

Es un conjunto de conceptos y herramientas (técnicas) que permite examinar los datos para describir sus principales características, privilegiando la representación visual de los mismos.

Los objetivos del análisis exploratorio son:

- Examinar las características del conjunto de datos.
- Comprobar si cumplen ciertas condiciones, como la condición de normalidad.
- Comprobar detectar y corregir los datos anómalos.
- Generar modelos óptimos.

Las estrategias que desarrolla el análisis exploratorio son:

- Examinar cada variable por separado
- Examinar las relaciones entre variables
- En el uso de las técnicas:
    - Iniciar con gráficos (de acuerdo al tipo de variable)
    - Luego pasar a resúmenes numéricos de aspectos específicos de los datos.

REPRESENTACIÓN GRÁFICA

En este análisis se privilegia la representación gráfica, los gráficos que utiliza son:

1.- Gráfico de barras y sectores

- Muestra la distribución de variables cualitativas.
- En su construcción se puede usar la frecuencia absoluta o relativa de las categorías.

2.- Histograma

- Muestra la distribución de datos cuantitativos.
- Es un gráfico de áreas, el área es proporcional a la frecuencia.
- Se construye con la frecuencia absoluta o relativa de los datos.
- No siempre es fácil de construir
- Se pierde información individual

3.- Grafico de tallo y hojas (Stemplots)

- Muestra la distribución de datos cuantitativos.
- Es muy parecido al histograma
- No se pierde la información individual
- Muestra con facilidad la concentración de los datos

4.- GRÁFICO DE CAJAS (BOXPLOTS)

- Muestra la distribución de datos cuantitativos.
- Permite examinar mejor la simetría de una distribución.
- Usa la mediana (más estable).
- Muestra el núcleo central de los datos (rango intercuartil) y sus colas.
- Se construye con los datos originales.
- Detecta datos anómalos, es decir aquellos datos que escapan al patrón general de la distribución.

¿Qué es lo que tenemos que observar en estos datos?

- Debemos observar la forma, el centro y la dispersión para determinar: simetría, número de modas.
- Buscar los casos anómalos e investigar su origen (¿análisis cualitativo?)
- Buscar la tendencia temporal cuando las observaciones se han hecho a través del tiempo.
- Estas observaciones nos permitirán hacer correcciones

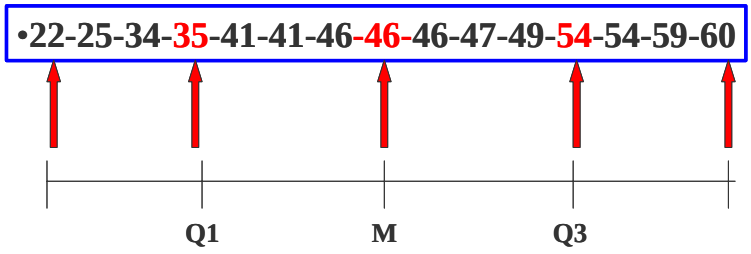
Los cinco datos básicos para la construcción del gráfico de caja son:

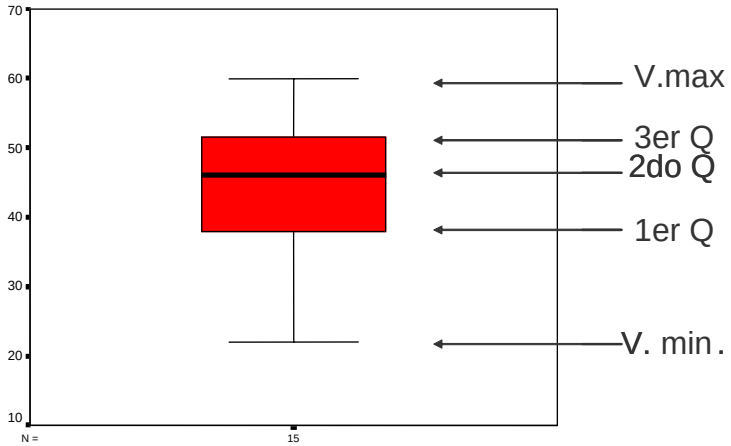
- Valor mínimo
- Cuartil 1 (Q1)
- Mediana (M)
- Cuartil 3 (Q3)
- Valor Máximo

Ejemplo: Edad de pacientes en un estudio sobre hipertensión:

54-59-35-41-46-25-47-60-54-46-49-46-41-34-22

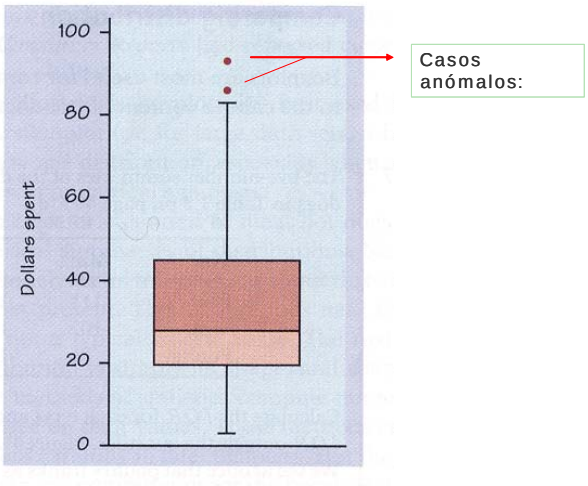
a.- Ordenamos los datos y ubicamos los datos básicos para su construcción





El grafico de cajas muestra:

- Una caja central que une los cuartiles y representa el 50% de las observaciones.
- La mediana, representada por la línea media de la caja.
- La simetría de la distribución
- Casos posiblemente anómalos, aquellos que se encuentran más allá de 1.5 del rango intercuartil, sobre Q3 o debajo de Q1
- Línea que muestra observaciones que corresponden al patrón de distribución



OTRAS TÉCNICAS PARA EL ANÁLISIS EXPLORATORIO

- Análisis de residuos
- Transformación de los datos para encontrar la escala que mejor simplifique o clarifique el análisis.

PROBABILIDADES:

Ing. Luz Bullón Camarena

CONCEPTOS BÁSICOS

TEORIA DE CONJUNTOS: Fundamentos y desarrollo: George Cantor (1845-1918)

CONJUNTO

Colección de objetos bien definidos

A, B, C	designan conjuntos
a, b, c,	designan los elementos del conjunto
	“a” pertenece a “A”
	“a” no pertenece a “B” (no es elemento de B)

EXPERIMENTO ALEATORIO (ε)

Cualquier experimento o ensayo real o hipotético cuyo resultado no puede predecirse con certeza y del cual es posible describir todos los resultados posibles

ESPACIO MUESTRAL (Ω) (S)

Es el conjunto de todos los resultados posibles de un experimento aleatorio

EVENTO

Cualquier subconjunto del espacio muestral

EVENTOS MUTUAMENTE EXCLUYENTES

Dos eventos que no pueden ocurrir juntos $A \cap B = \Phi$

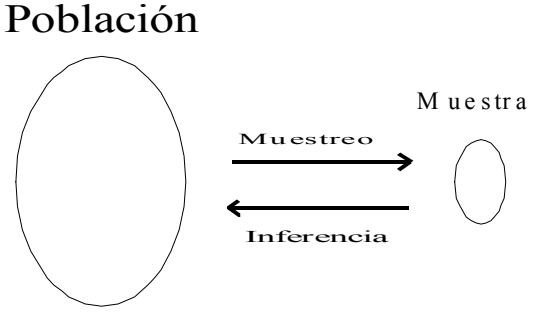
TEORÍA ELEMENTAL DE PROBABILIDADES

Ing. Luz Bullón Camarena

Debido a que las ciencias naturales y las ciencias sociales entre otras, no son ciencias exactas, raras veces se puede predecir un **evento** con absoluta certeza. Con frecuencia podemos encontrar afirmaciones como las siguientes:

- "La probabilidad del nacimiento de un individuo albino es 1/4, si ambos padres, normales, son portadores de un gen de albinismo"
- "La ocurrencia de determinado tipo sanguíneo es tan probable en varones como en mujeres"
- "Hay una gran probabilidad de supervivencia prolongada y una vida normal de un paciente con anemia aplásica grave sometido a trasplante"

Una comprensión de la teoría de probabilidades es necesaria para la toma de decisiones y para **formular conclusiones acerca de una población, basadas en el conocimiento e información de una muestra de esa población**, es decir, para hacer **Inferencia Estadística**.



DEFINICIÓN

Se va a definir la probabilidad en términos de una **frecuencia relativa** (es decir, una proporción). Así, la probabilidad P de que un evento *E* ocurra, es estimada por

$$P(E) = \frac{Número\ de\ veces\ en\ que\ E\ ocurre}{Número\ de\ veces\ en\ que\ E\ \ puede\ ocurrir}$$

De esta forma, la teoría de probabilidades, también, ayuda a la comprensión o interpretación de los datos presentados en tablas y gráficos.

PROPIEDADES

1. La probabilidad es un número entre 0 y 1
   a. Un valor de 0 significa que el evento es imposible, no puede ocurrir.
   b. Un valor de 1 significa que el evento es seguro, definitivamente ocurrirá.
   c. Un valor de 0.5 significa que el evento es igualmente probable que ocurra como que no ocurra.

2. La suma de las probabilidades (o frecuencias relativas) de todos los eventos que pueden ocurrir en la muestra debe ser 1 (o 100%)

EJEMPLO 1:

En un estudio de seguimiento en el primer año de vida de 122 niños nacidos con bajo peso (menor que 2500 g) y puntuaciones de APGAR a los 10 minutos muy bajas, se encontraron los siguientes resultados:

TABLA 1. Decesos en Niños nacidos con bajo peso y Puntuaciones APGAR muy bajas (0 a 3)

RESULTADO	NÚMERO DE CASOS	FRECUENCIA RELATIVA
Deceso (D)	42	0.3443
No Deceso (D^c)	80	0.6557
TOTAL	122	1.0000

La probabilidad de que el evento D ocurra, es igual a:

$$P(D) = \frac{42}{122} = 0.3443$$

Pero hay un grupo de niños en los cuales el evento D no ocurre, esto es, el evento complemento de D, representado por D^c, cuya probabilidad es igual a:

$$P(D^c) = 1 - \frac{42}{122} = \frac{80}{122} = 0.6557$$

EJEMPLO 2 :

En un estudio se quiere encontrar la relación entre puntuaciones de APGAR a los 10 minutos y riesgo de muerte en el primer año de vida en niños con bajo peso al nacer (menor que 2500g). Se investigó un total de 467 niños encontrándose los resultados siguientes:

TABLA 1. Decesos en Niños nacidos con bajo peso y Puntuaciones APGAR muy bajas (0 a 3)

APGAR a los 10 minutos	EVENTO RESULTADO		Total
	Deceso (D)	No Deceso (D^c)	
Muy baja 0 a 3 (E)	42	80	122
Intermedia 4 a 6 (E^c)	43	302	345
TOTAL	85	382	467

Ahora utilizaremos la definición y las propiedades para responder:

1. ¿Cuál es la probabilidad del evento deceso ó P(D) ?

$$P(D) = \frac{85}{467} = 0.182$$

2. ¿Cuál es la probabilidad de que un niño con bajo peso al nacer tenga una puntuación APGAR muy baja (de 0 a 3) ? ó ¿ P(E) ?

$$P(E) = \frac{122}{467} = 0261$$

PROBABILIDAD CONJUNTA DE DOS EVENTOS

La probabilidad conjunta de dos o más eventos, es la probabilidad de que **dichos eventos ocurran simultáneamente**. Es la probabilidad de la **intersección** dos eventos.

EJEMPLO 3. ¿Cuál es la probabilidad de una puntuación APGAR muy baja a los 10 minutos **y** morir en el primer año de vida? En símbolos, $P(E \cap D)$

TABLA 2. Puntuaciones APGAR a los 10 minutos y Muerte en el primer año de vida en nacidos con bajo peso

APGAR a los 10 minutos (Exposición)	EVENTO RESULTADO		Total
	Deceso (D)	No Deceso (D^c)	
Muy baja 0 a 3 (E)	42	80	122
Intermedia 4 a 6 (E^c)	43	302	345
TOTAL	85	382	467

Numero de niños con bajo peso al nacer, con puntuaciones APGAR muy bajas y que murieron en el primer año de vida 42
Número total de casos en los que el evento puede ocurrir 467

$$P(E \cap D) = \frac{42}{467} = 0.0899$$

PROBABILIDAD DE DOS EVENTOS CUALESQUIERA

La probabilidad de la **unión** de dos eventos cualesquiera, mutuamente excluyentes o no, es la probabilidad de que cualquiera de ellos ocurra, o que dichos eventos ocurran simultáneamente.

EJEMPLO 4. ¿Cuál es la probabilidad de una puntuación APGAR muy baja **o** deceso en el primer año de vida?. En símbolos, $P(E \cup D)$

TABLA 2 Puntuaciones APGAR a los 10 minutos y Muerte en el primer año de vida en nacidos con bajo peso

APGAR a los 10 minutos (Exposición)	EVENTO RESULTADO		Total
	Deceso (D)	No Deceso (D^c)	
Muy baja 0 a 3 (E)	42	80	122
Intermedia 4 a 6 (E^c)	43	302	345
TOTAL	85	382	467

Numero de niños con puntuaciones APGAR muy bajas (E) = 122
Numero de muertes en el primer año de vida (D) = 85
Numero de niños con puntuaciones APGAR muy bajas y que fallecieron en el primer año de vida ($E \cap D$) = 42

$$P(E \cup D) = P(E) + P(D) - P(E \cap D)$$
$$= \frac{122 + 85 - 42}{467} = 0.3533$$

PROBABILIDAD CONDICIONAL

La probabilidad condicional es la probabilidad de que un evento ocurra “dado que” o “sabiendo que” otro evento ya ha ocurrido

PREGUNTA**.** ¿Cual es la probabilidad de muerte en el primer año de vida, si se sabe que la puntuación APGAR a los 10 minutos resultó muy baja?

La probabilidad solicitada se refiere sólo al grupo con APGAR muy baja, esto es, para poder calcular la probabilidad del evento deceso en el primer año de vida, (D), ***primero debe haber ocurrido*** la puntuación APGAR muy baja (E). En símbolos,:

D/E : Ocurrencia del evento D ***dado que*** ocurrió el evento E

FÓRMULA DE LA PROBABILIDAD CONDICIONAL:

La probabilidad condicional, P(D/E), se puede definir en términos de la probabilidad conjunta P(D∩E), usando la formula:

$$P(D/E)=\frac{P(D\cap E)}{P(E)}$$

P(E) debe ser diferente de cero.

EJEMPLO 5. Calculemos la probabilidad condicional de deceso en el primer año de vida, dado que (o condicionado a que) la puntuación APGAR a los 10 minutos fue muy baja.

Número de casos en los **que ocurre** el evento deceso y Puntuación APGAR muy baja (D∩E) = 42

Número de casos en los **que ocurre** el evento Puntuación APGAR muy baja (E) = 122

Número total de casos en los **que puede** ocurrir el evento Deceso y el evento APGAR muy baja = 467

Calculando las probabilidades:

$$P(D\cap E)=\frac{42}{467} \qquad\qquad P(E)=\frac{122}{467}$$

Luego:

$$P(D/E)=\frac{\frac{42}{467}}{\frac{122}{467}}=\frac{42}{122}=0.3443$$

APLICACIONES DE PROBABILIDADES
MEDIDAS EPIDEMIOLÓGICAS

Ing. Luz bullón Camarena

1. RIESGO RELATIVO (RR)

El concepto de riesgo relativo resulta de utilidad cuando se quiere comparar las probabilidades de cierto resultado E, por ejemplo enfermedad, en dos situaciones o grupos diferentes.

$$RR = \frac{P\ (E \mid expuesto\ )}{P\ (E \mid no\ expuesto\ )}$$

Es una medida natural, directa e intuitiva del efecto de exposición.

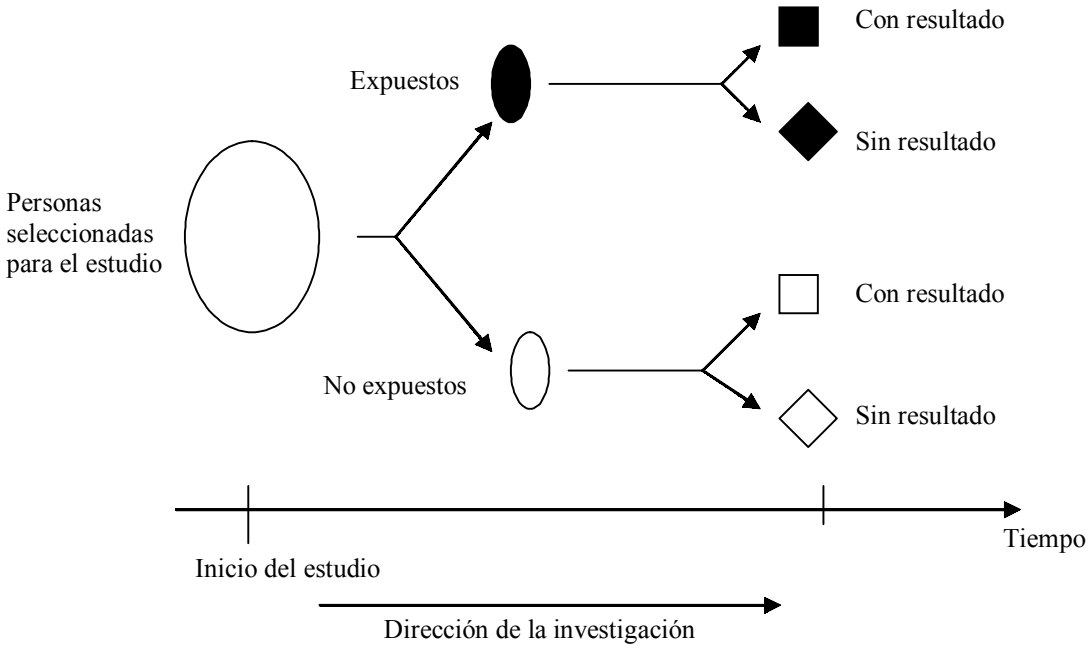
RR = 1 Implica que la probabilidad de desarrollar la enfermedad en los grupos *expuesto* y *no expuesto* son idénticas, por lo tanto no hay asociación entre exposición y enfermedad

RR > 1 significa aumento del riesgo de enfermedad, entre los expuestos

RR < 1 significa disminución del riesgo de enfermedad, entre los expuestos

El **riesgo relativo** puede calcularse sólo en un **estudio de cohortes** en donde se identifica primero, un grupo de personas expuestas a un factor de riesgo y otro no expuesto. Luego de un seguimiento se determina la ocurrencia de un evento como enfermedad, deceso o recuperación.

FIG 1. Esquema de un estudio de cohortes



* Reproducido de Greenberg, R.S.: Prospective studies, Encyclopedia of Statistical Sciences, Vol 7, Wiley, 1988

Las áreas sombreadas representan sujetos expuestos al factor antecedente, las áreas claras a no expuestos. Los cuadrados representan sujetos con la enfermedad o consecuencia que se estudia, los rombos son los sujetos sin ella.
Los datos proporcionados por un estudio de cohortes, pueden ser presentados en una tabla como sigue:

Exposición	Enfermedad		Total
	Sí	No	
Sí	*a*	*b*	n_1
No	*c*	*d*	n_0

Los tamaños de los grupos comparados expuesto y no expuesto son cantidades fijadas de antemano, mientras que las cantidades *a, b, c* y *d* son aleatorias y conocidas al final del estudio.

EJEMPLO 1 Relación entre puntuaciones de APGAR a los 10 minutos y riesgo de muerte en el primer año de vida en niños con bajo peso al nacer (menor que 2 500g)

APGAR	Evento resultado		Total
	Deceso	No deceso	
Muy baja (0 a 3)	42	80	122
Intermedia (4 a 6)	43	302	345
Total	85	382	467

Factor de Riesgo: Puntuación APGAR muy baja

Decesos entre nacidos con APGAR muy baja: 42/122 = 0.3443
Decesos entre nacidos con APGAR intermedia: 43/345 = 0.1246

Cálculo del RR de muerte con APGAR muy baja:

$$RR = \frac{42/122}{43/345} = \frac{0.3443}{0.1246} = 2.762$$

INTERPRETACIÓN:

Un RR de 2,76 significa que recién nacidos con bajo peso al nacer y puntuaciones APGAR muy bajas a los 10 minutos, tienen una probabilidad casi tres veces mayor de fallecer en su primer año de vida que los recién nacidos con APGAR intermedio a los 10 minutos.

Las magnitudes de las probabilidades no tienen importancia (aún cuando los eventos sean raros o poco probables), sólo es importante el cociente de estas probabilidades. De esta forma podemos comparar los decesos por cáncer pulmonar, de baja probabilidad en ambos grupos: fumadores y no fumadores pero de elevado riesgo para el primero. Se conoce que la probabilidad de muerte de un hombre mayor de 35 años por cáncer pulmonar perteneciendo al grupo de fumadores es .002679, mientras que esta probabilidad entre los no fumadores es .000154. Calculando el riesgo relativo, RR = .002679 / .000154 = 17.4, éste resulta elevado.

2. RAZÓN DE CHANCES (Odds Ratio, OR)

Otra medida frecuentemente empleada en la comparación de grupos mediante probabilidades es la razón de chances. *La chance* en favor de un evento E, que ocurre con probabilidad *p,* se define como $\frac{p}{1 - p}$.

Por ejemplo, si la probabilidad de E es $p = \frac{1}{2}$, la chance de E es $\frac{1/2}{1/2}$ = 1 a 1. En otro caso, si $p = \frac{2}{3}$, esta chance es de $\frac{2/3}{1/3}$ = 2 a 1, es decir la probabilidad de que E ocurra es dos veces mayor que la probabilidad que no ocurra.

Con esta definición previa, la *Razón de Chances*, es definida como la chance a favor de la enfermedad entre individuos expuestos dividida por la chance de enfermedad entre los no expuestos:

$$OR = \frac{P(E \mid expuesto) / [1 - P(E \mid expuesto)]}{P(E \mid no\ expuesto) / [1 - P(E \mid no\ expuesto)]}$$

OR = 1 Indica que la exposición no tiene un efecto en la probabilidad de la enfermedad

RR y *OR* son dos medidas que intentan explicar el mismo fenómeno

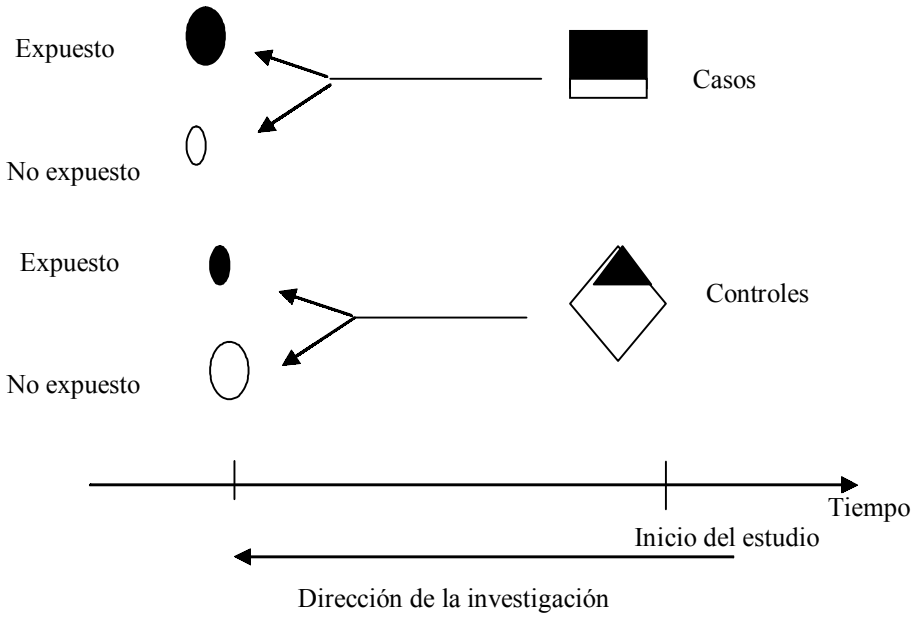
OR tiene mejores propiedades estadísticas

La **razón de chances** puede calcularse en un **estudio de casos y controles.** En ese tipo de estudio se identifica primero, un grupo de personas enfermas (los casos), se busca un segundo grupo comparable con el primero en las demás características pero sin la enfermedad en estudio y se averigua retrospectivamente la exposición a un factor de riesgo.

Los datos de un estudio de casos y controles pueden ser presentados en una tabla como la siguiente. Los tamaños de las muestras de casos y controles son fijados inicialmente y las cantidades *a, b, c* y *d* resultan conocidas después.

Enfermedad	Exposición		Total
	Sí	No	
Sí (Casos)	*a*	*b*	n_1
No (Controles)	*c*	*d*	n_0

FIG 2. Esquema de un estudio de casos y controles



* Reproducido de Greenberg, R.S.: Retrospective studies, Encyclopedia of Statistical Sciences, Vol 8, Wiley, 1988

EJEMPLO 2. Asociación entre la toma de anticonceptivos orales y el riesgo de trombosis venosa en un grupo de mujeres hospitalizadas

Trombosis	Toma de Anticonceptivos		Total
	Si	No	
Casos	12	30	42
Controles	53	347	400
Total	65	367	442

$$OR = \frac{\frac{12 / 65}{53 / 65}}{\frac{30 / 377}{347 / 377}} = \frac{12 \times 347}{30 \times 53} = 2.619$$

La interpretación del valor del OR es similar a la del RR.

Cuando el valor está alrededor de 1, no hay asociación entre enfermedad y factor de riesgo. La asociación si existe se expresa en dos sentidos: un valor del OR menor de 1 indica una asociación en sentido inverso mientras que un valor bastante mayor de 1 indica una mayor asociación directa entre factor y enfermedad.

VARIABLE ALEATORIA

Ing. Luz Bullón Camarena

Una variable aleatoria es aquella variable que asume diferentes valores a consecuencia de la aleatoriedad. Estas variables pueden ser discretas o continuas.

V. A. DISCRETA. Asume sólo un número limitado de valores, Los valores respuesta son números enteros.

EJEMPLO: Un centro de detección del cáncer mamario no puede conocer con exactitud cuántas mujeres solicitarán ser examinadas en un día cualquiera. Por lo tanto, el número de mujeres que serán atendidas mañana es una variable aleatoria. Los valores de esta variable son los números correspondientes a cada resultado posible: 0, 1, 2, ...

V. A. CONTINUA. Asume cualquier valor dentro de los límites de un intervalo continuo. Teóricamente, la variable aleatoria X puede asumir un número infinito de posibles valores

EJEMPLOS: Cuando se mide el peso o la estatura de un individuo, las respuestas pueden ser 60.40 kg, 175.50 cm. Claramente, los valores de las respuestas varían en un rango permisible, pero siempre será posible encontrar un tercer individuo entre dos cualesquiera.

DISTRIBUCIÓN DE PROBABILIDADES

D.P. DISCRETA: Es un listado (una tabla, un gráfico) de las probabilidades de todos los resultados posibles de una variable aleatoria discreta que pueden presentarse.

D.P.CONTINUA: Es una función, un modelo matemático que da lugar a curvas y las probabilidades van a ser áreas bajo las curvas.

DISTRIBUCION BINOMIAL

Ing. Luz Bullón Camarena

PROBLEMA: Suponga que la tasa de mortalidad para cierta enfermedad es de 0.20 y que tres personas de una comunidad contraen la enfermedad. ¿Cuál es la probabilidad de que dos enfermos mueran?

Como el ejemplo, hay muchos en los que se dan las siguientes condiciones:

- Una situación que conduce a dos resultados posibles: estado nutricional normal o no, opinión favorable o no favorable, tener la enfermedad o no, sobrevivir a una enfermedad o fallecer.
  Uno de los resultados o evento es llamado "éxito" y el otro "fracaso"
  {éxito, fracaso} = {1, 0}

- La probabilidad de la ocurrencia del evento éxito es: P(éxito) = π , por lo tanto P(fracaso) = 1- π.
  π y (1-π) no necesariamente son iguales

- El evento puede repetirse un número *n* de veces, en forma independiente, es decir, la ocurrencia del evento en un individuo, no depende de la ocurrencia del mismo evento en otro individuo

CALCULO DE PROBABILIDADES

La variable aleatoria se define como

X: Número de éxitos que ocurren en las *n* repeticiones

X puede tomar los valores $x = 0, 1, 2, \ldots, n$

Se dice: X tiene distribución Binomial,

$$P(X = x) = \binom{n}{x} \pi^x (1-\pi)^{n-x}$$

La notación $\binom{n}{x}$ representa al número combinatorio, cuenta el número de secuencias diferentes que contienen *x* éxitos y por tanto *n-x* fracasos, tiene desarrollo $\frac{n!}{x!(n-x)!}$.

Calculando la probabilidad en el ejemplo,

Tres personas de una comunidad contraen la enfermedad, $n = 3$,

La tasa de mortalidad para la enfermedad es de 0.20 $\pi = 0.2$

¿Cuál es la probabilidad de que dos enfermos mueran? $x = 2$

$$P(X=2)=\binom{3}{2}0.2^2\,(1-0.2)^{3-2}$$

$$=\frac{3!}{2!1!}0.2^2\times 0.8=0.096$$

Explicando la fórmula:

Los dos resultados posibles de un individuo son Muerte = M o Sobrevivencia = S, los tres individuos se combinan obteniéndose el número de muertes:

Resultados:		SSS	SSM	SMS	MSS	SMM	MSM	MMS	MMM
Número de muertes	*x*	0	1	1	1	2	2	2	3

$\binom{3}{2}=\frac{3!}{2!1!}=3$ es el número de formas en que ocurre que dos de los tres mueran

El resultado *x* = 2 está formado por SMM MSM MMS cada uno con igual probabilidad $=0.2^2\times 0.8=0.032$

Luego, la probabilidad de que dos de los tres mueran es $=\binom{3}{2}\times 0.032=0.096$

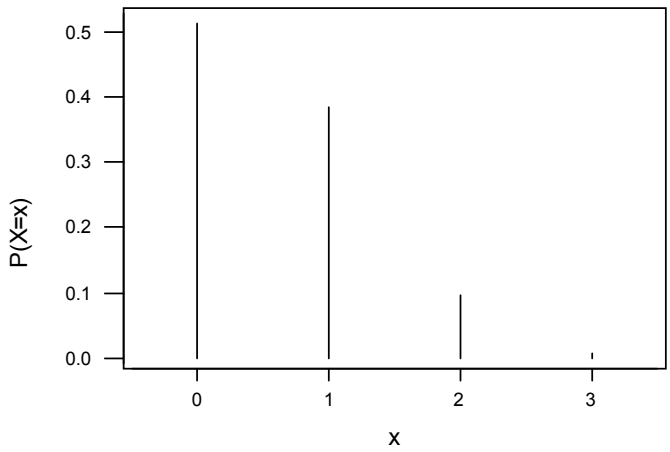
DISTRIBUCIÓN DE PROBABILIDADES

Es una tabla o un gráfico. Presentada en una tabla la distribución de probabilidades de la variable X: Número de muertes en las tres personas que contraen la enfermedad es:

TABLA 1. Distribución de probabilidades del número de muertes

x	P (X = *x*)
0	0.512
1	0.384
2	0.096
3	0.008
Total	1.000

GRÁFICO 1. Distribución de probabilidades del número de muertes



Es posible calcular probabilidades acumuladas de la forma:

$P(X \leq 2) = P(X=0) + P(X=1) + P(X=2) = 0.992$

El mismo resultado puede ser calculado por el complemento:

$P(X \leq 2) = 1 - P(X>2) = 1 - P(X=3) = 1 - 0.008 = 0.992$

DISTRIBUCION DE POISSON

Ing. Luz Bullón Camarena

PROBLEMA: Durante el estudio de cierto organismo acuático, un gran número de muestras fue tomado de una laguna y se contó el número de organismos en cada muestra. El número promedio de organismos encontrados por muestra fue de dos. ¿Cuál es la probabilidad de que la siguiente muestra tenga tres organismos?

Como en el ejemplo, existen procesos tales como: accidentes automovilísticos en un cruce, demanda (necesidades) de servicios en una institución asistencial, clientes que llegan a una farmacia. Estos procesos tienen en común que pueden ser descritos por una variable aleatoria discreta que asume valores enteros. Así, el número de pacientes que llegan a un consultorio médico en cierto intervalo de tiempo será 0, 1, 2, 3, 4 o algún otro número.

Características:

- La media o promedio de ocurrencias por intervalo de tiempo, (por área o espacio determinado) se conoce o puede ser estimado
- El número de ocurrencias en un intervalo determinado de tiempo, no depende del número de ocurrencias en cualquier otro intervalo de igual magnitud.

CALCULO DE PROBABILIDADES

La variable aleatoria se define como

X: número de ocurrencias en un intervalo determinado de tiempo

X puede tomar los valores x = 0, 1, 2, . . .
Se dice: X tiene distribución de Poisson,

$$P(X = x) = \frac{\mu^x \, e^{-\mu}}{x!}$$

donde:

μ es la media o promedio de ocurrencias por intervalo de tiempo, de la variable en estudio

e es el número base del sistema de logaritmos naturales (=2.718281 . . .)

Calculando la probabilidad del problema:

Datos: μ = 2, x=3

$$P(X = 3) = \frac{2^3 \, e^{-2}}{3!} = 0.1804$$

Si se quiere saber:

¿Cuál es la probabilidad de que una muestra tenga un organismo o menos?

$$P(X \leq 1) = P(X = 0) + P(X = 1)$$
$$= \frac{2^0 e^{-2}}{0!} + \frac{2^1 e^{-2}}{1!} = 0.1353 + 0.2707 = 0.406$$

¿Cuál es la probabilidad de que una muestra tenga por lo menos dos organismos?

$$P(X \geq 2) = \sum_{x=2}^{\infty} P(X = x) = 1 - P(X < 2)$$
$$= 1 - 0.406 = 0.594$$

DISTRIBUCIÓN DE PROBABILIDADES

TABLA 2. Distribución de probabilidades del número de organismos por volumen de agua

x	P (X=x)
0	0.1353
1	0.2707
2	0.2707
3	0.1804
4	0.0902
5	0.0361
6	0.0120
7	0.0034
8	0.0009
9	0.0002
10	0.0000
...	
Total	1.0000

DISTRIBUCIÓN NORMAL DE PROBABILIDADES

Ing. Luz Bullón Camarena

La función de densidad de probabilidad de una variable aleatoria continua se dice que presenta una distribución normal de probabilidades si su función de densidad es,

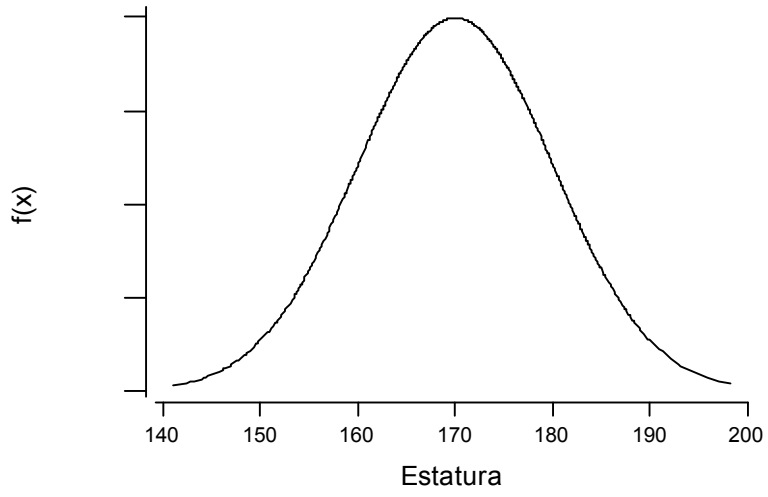
$$f(x)=\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty<x<+\infty, \quad -\infty<\mu<+\infty, \quad \sigma>0$$

μ, y σ, son los parámetros de la distribución y corresponden a la media y desviación estándar, e es la base de los logaritmos naturales (2.718282) y π es la constante geométrica (3.141593).

La gráfica de tal función de densidad es lo que se conoce como curva normal, la cual es utilizada para describir el comportamiento de muchas variables en la naturaleza, tal como la estatura y peso de niños menores de cinco años, estatura y peso de hombres adultos, entre muchas otras características antropométricas. También puede describir el comportamiento de la variable temperatura (medida a intervalos fijos de tiempo) en el día de hoy o medida en la misma hora sucesivos días, contenido real con el que se envasa una marca de yogurt, longitud o diámetro de una pieza para ensamblaje de automóviles, etc.

Consideremos la estatura del hombre adulto, peruano, y supongamos que la media o promedio de la misma sea 170 cm. y además la desviación estándar sea 10 cm. El gráfico siguiente es de la distribución de probabilidades de dicha estatura.

FIG 1. Distribución de probabilidades de la estatura del hombre adulto peruano ($\mu = 170$, $\sigma^2 = 10^2$)



La curva normal presenta las siguientes características:

1. Distribución en forma de campana, la curva es simétrica alrededor del valor central μ
2. Media, mediana y moda coinciden.
3. El área total entre la curva y el eje horizontal es 1
4. $E(X) = \mu$ y $V(X) = \sigma^2$
5. Los extremos de la distribución son extienden asintóticamente al eje horizontal
6. La posición de la curva está determinada por el valor central μ, y el grado de apuntamiento alrededor del valor central está determinado por σ^2. A mayor variancia, más achatada será la curva.
7. La probabilidad que la v.a. X se encuentre dentro de un intervalo es dada por el área bajo la curva normal para ese intervalo:
    a. $P(X = b) = 0$
    b. $P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$
8. 68.27% del área de la distribución cae entre $\mu - 1\sigma$ y $\mu + 1\sigma$, o dentro de una desviación estándar de la media.
9. 95.45% del área de la distribución cae entre $\mu - 2\sigma$ y $\mu + 2\sigma$, o dentro de dos desviaciones estándares de la media.
10. 99.73% del área de la distribución cae entre $\mu - 3\sigma$ y $\mu + 3\sigma$, o dentro de tres desviaciones estándares de la media.

FIG 2. Distribuciones normales, con medias diferentes pero variancias iguales

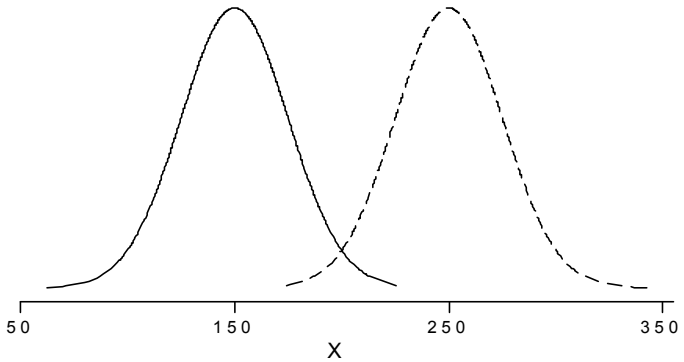
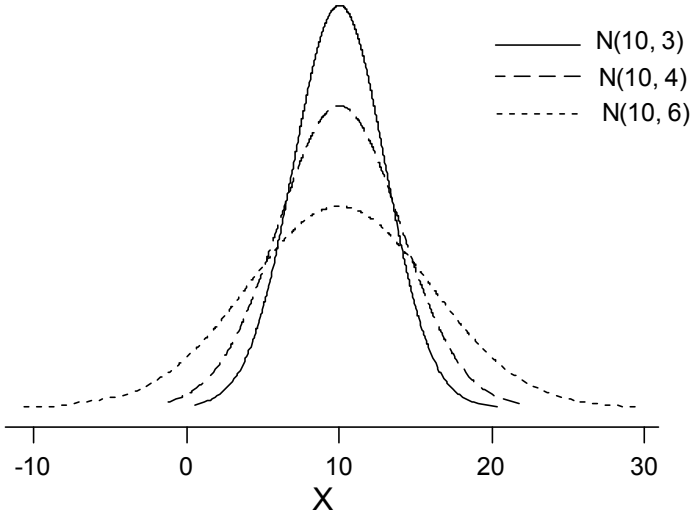


FIG 3. Distribuciones normales, con igual media pero variancias diferentes



DISTRIBUCIÓN NORMAL ESTÁNDAR

Una variable aleatoria normal estandarizada, representada por Z, corresponde a una distribución normal con $\mu = 0$, y $\sigma^2 = 1$. La función de densidad de probabilidad de la variable normal estandarizada **Z** está dada por

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad -\infty < z < +\infty, \quad \mu = 0, \quad \sigma^2 = 1$$

1. Distribución en forma de campana, la curva es simétrica alrededor de 0 (μ)
2. 68.27% del área de la distribución cae entre –1 y +1
3. 95.45% del área de la distribución cae entre –2 y +2
4. 99.73% del área de la distribución cae entre –3 y +3
5. Cualquier distribución normal, donde X tiene media μ y variancia σ^2 puede ser transformada en la distribución normal estándar, lo cual nos permite utilizar tablas de áreas bajo la curva de la distribución normal estándar para responder a preguntas sobre probabilidades de ocurrencia de un valor **x** de la variable aleatoria. X

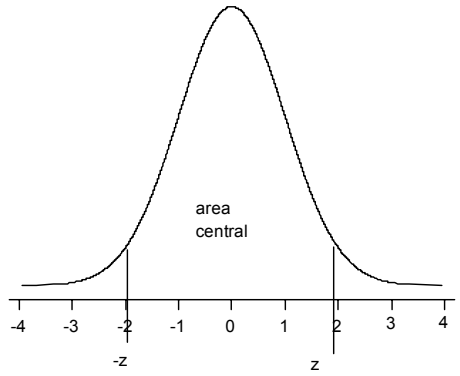
Si **X** es una variable aleatoria normal, con media μ y variancia σ^2, entonces

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

es una variable aleatoria normal estandarizada con media 0 y variancia 1,

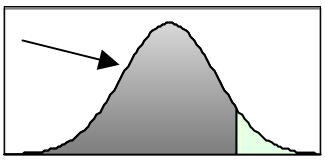
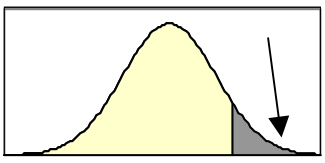
$$P(a \le X \le b) = P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma} \right)$$

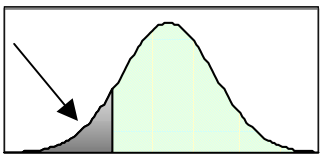
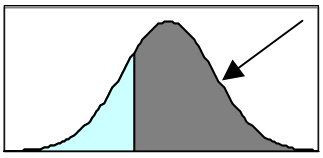
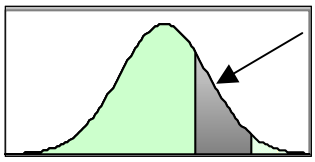
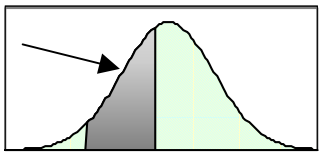
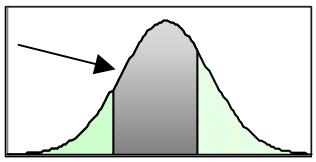
Distribución Normal Estándar : Z



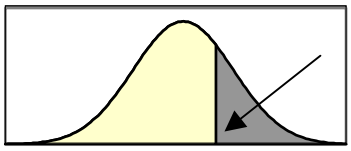
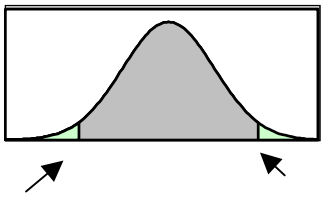
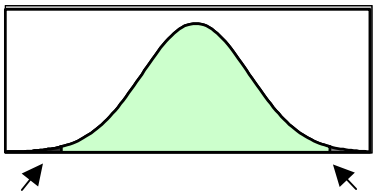
Valor de Z	1.0	1.282	1.645	1.96	2.0	2.326	2.576	2.807	3.0	3.09
Área en extremos α		0.20	0.10	0.05		0.02	0.01	0.005		0.002
Área central (1- α)		0.80	0.90	0.95		0.98	0.99	0.995		0.998

ÁREAS BAJO LA CURVA NORMAL

Uso de la tabla de probabilidades normal I	
1.- Hallar $P(Z \le z_0)$ 	o Hallar $P(Z \le 1.45)$ o De la tabla área a la izquierda de 1.45 = 0.9265 o $P(Z \le 1.45) = 0.9265$
2.- Hallar $P(Z \ge z_0)$ 	o Hallar $P(Z \ge 1.75)$ o Área bajo la curva es 1.00 o De la tabla $P(Z \le 1.75) = 0.9599$ o $P(Z \ge 1.75) = 1.00 - P(Z \le 1.75)$ $= 1.00 - 0.9599 = 0.0401$

Uso de la tabla de probabilidades normal I	
3. Hallar $P(Z \leq z_0)$, $z_0 < 0$ 	o Hallar $P(Z \leq -1.20)$ o $P(Z \leq -1.20) = 0.1151$ o Área a la derecha de -1.20 = 1- 0.1151 o $P(Z \geq -1.20) = -0.8849$
4. Hallar $P(Z \geq z_1)$, $z_1 < 0$ 	o Hallar $P(Z \geq -0.75)$ o Área a la izquierda de - 0.75 = 0.2266 o $P(Z \geq -0.75) = 1- 0.2266 = 0.7734$
5. Hallar $P(z_1 \leq Z \leq z_2)$ 	o Hallar $P(0.70 \leq Z \leq 1.96)$ o Área a la izquierda de 1.96 = 0.9750 o Área a la izquierda de 0.70 = 0.7580 o $P(0.70 \leq Z \leq 1.96) = P(Z \leq 1.96) - P(Z \leq 0.70)$ $= 0.9750 – 0.7580 = 0.2170$
6. Hallar $P(z_1 \leq Z \leq z_2)$, donde $z_1 < 0$ y $z_2 < 0$ 	o Hallar $P(-1.75 \leq Z \leq -0.30)$ o Área a la izquierda de - 0.30 = 0.3821 o Área a la izquierda de - 1.75 = 0.0401 o $P(-1.75 \leq Z \leq -0.30) = 0.3821 – 0.0401 = 0.342$
7. Hallar $P(z_1 \leq Z \leq z_2)$, donde $z_1 < 0$ y $z_2 > 0$ 	o Hallar $P(-1.20 \leq Z \leq 0.70)$ o Área a la izquierda de 0.70 = 0.7580 o Área a la izquierda de - 1.20 = 0.1151 o 0.7580 – 0.1151 = 0.6429 o $P(-1.20 \leq Z \leq 0.70) = 0.6429$

CÁLCULO DE VALORES DE LA DISTRIBUCIÓN NORMAL

Uso de la tabla de probabilidades normal II	
Hallar el valor de z_0, tal que $P(Z < z_0) = 0.75$ (z_0 es el tercer cuartil) 	– $P(Z < z_0) = 0.75$ – En tabla, $P(Z < 0.68) = 0.7517$ – $Z_0 = 0.68$ (aproximadamente)
Hallar z_0, tal que $P(- z_0 < Z < z_0) = 0.90$ 	– Área a la izquierda de $-z_0 = 0.01 / 2 = 0.05$ – Área a la derecha de $+z_0 = 0.05$ – Área a la izquierda de $+z_0 = 0.95$ – De la tabla, $z_0 = 1.64$
Hallar z_0, tal que $P(-z_0 < Z < z_0) = 0.99$ 	– $P(-z_0 < Z < z_0) = 0.99$ – Área a la izquierda de $- z_0 = 0.005$ – Área a la izquierda de $+ z_0 = 0.995$ – De la tabla, $z_0 = 2.57$ (Aprox.)

EJERCICIOS

1. Sea X una v.a. con media 10 y variancia 4. Halle:

   a. Probabilidad de una observación elegida al azar sea menor de 7

   Tenemos que $Z = \frac{X-10}{\sqrt{4}}$

   Para X = 7, corresponde z = $\frac{7-10}{2} = -1.5$

   P(X < 7) = P(Z < -1.5) = 0.0668

   b. Probabilidad que el valor de X se encuentre entre 7 y 13

   Para X = 13, corresponde z = $\frac{13-10}{2} = 1.5$

   P(7 < X < 13) = P(-1.5 < Z < 1.5) = 2 $\times$ 0.0668 = 0.1336

   c. Probabilidad que el valor de X se encuentre entre 8.9 y 11.4

   Tenemos que $\frac{8.9-10}{2} = -0.55$ y $\frac{11.4-10}{2} = 0.7$

   Entonces, P(8.9 < X < 11.4) = P(-0.55 < Z < 0.7)

   P(Z < 0.7) = 0.7580

   P(Z < -0.55) = 0.2912

   P(8.9 < X < 11.4) = 0.7580 - 0.2912 = 0.4668

2. Sea X una v.a. con media 68 y desviación estándar 5. Halle:

   a. El valor x_0 tal que se cumple P(X $\geq$ x_0) = 0.15

   Tenemos que $Z = \frac{X-68}{5}$, luego para X = x_0, corresponde z_0 = $\frac{x_o-68}{5}$

   P(X $\geq$ x_o) = 0.15 = P(Z $\geq$ z_o), entonces, P(Z $\leq$ z_o) = 1 – 0.15 = 0.85

   De la tabla se tiene z_o = 1.04 (Aprox.)

   Luego, x_o = 68 + 1.04 $\times$ 5 = 73.2

   P(X $\geq$ 73.2) = 0.15

   b. Los valores x_1 y x_2 tal que se cumple P($x_1 \leq X \leq x_2$) = 0.94,

   P(X $\leq$ x_1) = 0.03, y P(X $\geq$ x_2) = 0.03

   Para X = x_1, corresponde z_1 = $\frac{x_1-68}{5}$

   Para X = x_2, corresponde z_2 = $\frac{x_2-68}{5}$

$P(z_1 \leq Z \leq z_2) = 0.94$, $P(Z \leq z_1) = 0.03$, y $P(Z \geq z_2) = 0.03$. Notar $z_1 = -z_2$

$P(Z \leq z_2) = 0.94 + 0.03 = 0.97$

De la tabla, $z_2 = 1.88$ (aprox.) ,

$P(Z \leq -1.88) = 0.03$, $P(Z \geq 1.88) = 0.03$, $P(-1.88 \leq Z \leq 1.88) = 0.94$

$x_1 = 68 - 1.88 \times 5 = 58.6$

$x_2 = 68 + 1.88 \times 5 = 77.4$

EJEMPLO DE APLICACIÓN

Una marca de yogurt, afirma que fabrica su producto con un contenido medio de grasa (en mg. / unidad) de 4.5 y una desviación estándar de 0.3. Ud. adquiere una unidad,

a. ¿cuál es la probabilidad de que esté consumiendo un producto con más de 5 mg. de grasa?

$P(X > 5) = P(Z > \frac{5 - 4.5}{0.3}) = P(Z>1.67) = 0.0475$

b. ¿cuál es la probabilidad de que su compra tenga un contenido de grasa, diferente de la media en 1 unidad?

Es la probabilidad de que ocurra $X - \mu = -1$ ó $X - \mu = 1$, o expresado de otra forma,
$P(X = \mu - 1)$ ó $P(X = \mu + 1)$, esas probabilidades son puntuales e iguales a cero

c. El dueño, afirma que se devolverá el dinero si el contenido de grasa supera 7 mg, ¿cuál es la probabilidad que le retornen su dinero?

$P(X > 7) = P(Z > \frac{7 - 4.5}{0.3}) = P(Z>8.33) = 0.0000...$

Prácticamente no le devolverán su dinero desde que es casi improbable que adquiera una unidad con más de 7 mg de grasa.

MUESTREO

Ing Edith Alarcón M. *Mg. Martha Martina Ch*
.

INTRODUCCIÓN

Un investigador está interesado en determinar el nivel de conocimientos y percepciones de las madres de familia de una comunidad urbana marginal acerca del calendario de vacunaciones; otro investigador, le interesa determinar los hábitos de estudios y su relación con el nivel de aprendizaje de los estudiantes en una Universidad Pública; y, probablemente otro investigador, está motivado por demostrar la eficiencia de un nuevo procedimiento en el tratamiento de las úlceras por decúbito en pacientes adultos mayores, mediante un ensayo clínico controlado.
Todos estos ejemplos, sugieren las siguientes interrogantes:

En el primer caso:

¿Es necesario estudiar a todas las madres de familia para estudiar cuáles son sus conocimientos acerca del calendario de vacunaciones?

¿Cómo llegará a las madres de familia? De puerta en puerta? A través de los comedores populares?

En el segundo caso:

¿Estudiará a todos los estudiantes de la Universidad?

¿Seleccionará a los estudiantes por sexo?, por Facultad de procedencia? Por nivel de rendimiento?

En el tercer caso:

¿Cuántos pacientes requiere para probar su hipótesis: el nuevo procedimiento es eficiente para el tratamiento de las úlceras por decúbito.

Esta es parte de la preocupación de los investigadores, y está relacionado con el tema que abordaremos en las próximas líneas: Población y Muestra.

ALGUNOS CONCEPTOS BÁSICOS

- **POBLACIÓN:**

Es todo conjunto de objetos, situaciones o sujetos con un rasgo común. Es un conjunto de casos que satisface una serie predeterminada de criterios. No siempre se refiere a personas ya que pudiera referirse al total de expedientes clínicos archivados en un determinado hospital; al total de anotaciones de enfermería; al total de punciones lumbares; etc. Sea cual fuere la unidad fundamental, la población siempre abarca el total de elementos que interesan al investigador y se debe partir de los criterios específicos que se desean incluir. **A la población se le denota por: N**

Puede diferenciarse en dos niveles:

- la población objetivo que es el gran conjunto de unidades a los que se generalizarán los resultados del estudio, y están definidas por las condiciones clínicas y demográficas; y,
- la población accesible que es el subconjunto de la población que se encuentra disponible para el estudio y está determinada por las características geográficas y temporales. En la cual supuestamente se podrán localizar a todas las unidades que integrarán la muestra. También se conoce como marco muestral.

- **ELEMENTOS O UNIDADES MUESTRALES:**

Es la unidad básica alrededor de la cual se recaba la información. Es el elemento que da origen al valor de las variables (un expediente, una radiografía, un paciente, una enfermera, un estudiante, un animal de laboratorio, etc.). Las unidades de muestreo cubren toda la población. Dichas unidades deben estar claramente definidas, identificables y observables.

- **MUESTRA:**

Es el subconjunto de la población integrado por las unidades muestrales seleccionadas. **A la muestra se le denota por: n**

- **MARCO MUESTRAL:**

Es una lista detallada de las unidades de muestreo de donde se obtiene la muestra. Dependiendo de la complejidad de la investigación a veces es imposible disponer de un marco muestral. Se le define también como la población operativamente factible o la que puede ser muestreada realmente. Son ejemplos de marcos muestrales: el directorio telefónico, el listado de alumnos de una universidad, el listado de Centros de Salud, el listado de manzanas de una comunidad, etc.

RECUERDE SIEMPRE QUE EL INVESTIGADOR:

Estudia la muestra, la que debe ser representativa (calidad) y significativa(cantidad) y partir de este estudio infiere (deduce) lo que sucede en la población de la cual fue extraída dicha muestra
LA META DEL INVESTIGADOR: Es obtener una muestra que represente realmente todas las características de la población de la cual es extraída y que sólo difiera en el tamaño. El investigador no sólo debe preocuparse del tamaño de la muestra sinó también de seleccionar cuidadosamente las unidades que formarán parte de la

¿Cuáles son las razones para realizar un Muestreo?

Al efectuar una investigación existen varias razones para realizar muestreo:

- Rapidez
- Costo
- Factibilidad
- Exactitud.

En cuanto a las tres primeras razones, es obvio que existe mayor rapidez y menor costo en estudiar cien personas que mil o más y es más posible hacerlo por situaciones de recursos humanos, físicos y apoyos logísticos. En cuanto a exactitud, se refiere al hecho de que a menor volumen de trabajo, es posible emplear personal mejor capacitado que garantice una medición del fenómeno de interés con mayor precisión y poder supervisar mejor para producir resultados más exactos.

¿Cuáles son las preguntas habituales que hace un investigador respecto a población y muestreo?

¿Cuál es la población en estudio? ¿El investigador determina su propia población?

El investigador determina la población en estudio de acuerdo con el problema que quiere investigar. Influye también el tiempo y los recursos económicos que dispone.

¿Cuántas personas se requieren en la muestra?

Para responder a esta pregunta, el investigador debe recordar que **la muestra debe reunir dos condiciones:**

1.- Representativa:

Las características importantes de la población (sexo,edad, etc) deben estar presentes en la muestra, en proporciones similares. De esta manera, el investigador podrá hacer inferencias válidas respecto a la población de donde obtuvo su muestra. Es decir, si en la población una de las características relevantes es el sexo femenino y éste se encuentra en un 60%, en la muestra también estará representado el sexo femenino en un 60%.

2.- Adecuada:

Está relacionado con el tamaño de la muestra. Se calcula con diversas fórmulas establecidas de acuerdo a si el estudio busca una proporción existente en una población (por ejemplo un estudio de prevalencia), diferencias entre las medias o las proporciones de dos poblaciones, correlación entre dos o más factores, factores de riesgo (estudios de riesgos relativos o razones de momios), pruebas diagnósticas (estudios de sensibilidad, especificidad y valores predictivos), etc.
No existe una fórmula única para la determinación del tamaño de una Muestra.

¿Cómo seleccionar una muestra?

Para responder a esta tercera pregunta, el investigador debe conocer que existen diferentes métodos de muestreo, los cuales están relacionados con el diseño de la investigación..

¿Cuáles son los tipos de Muestreo?

Se divide en dos grandes grupos: No probabilísticos y Probabilísticos

MUESTREO NO PROBABILÍSTICO

Si la muestra es escogida por medio de un proceso subjetivo o arbitrario de modo que la probabilidad de selección de cada unidad de la población no es conocida (se utiliza con frecuencia cuando no se conoce el marco muestral). Es decir la selección de la muestra depende del juicio personal del investigador. Éste tipo de muestreo es usado con frecuencia en la investigación de mercados y en investigaciones cualitativas.

MUESTREO PROBABILÍSTICO

Cuando el método de selección de la muestra permite que todos los elementos de la población tengan la misma probabilidad de ser seleccionados en la muestra. Utiliza procedimientos de selección aleatoria para asegurar que cada unidad de la muestra se seleccione por probabilidad (es factible si se conoce el marco muestral, es decir, se cuenta con un listado completo de todas las unidades que componen la población).

NO PROBABILÍSTICO

1. Por conveniencia (a criterio)
2. Por casos consecutivos
3. Por cuota
4. Por Bola de Nieve

PROBABILÍSTICO

1. Aleatorio simple
2. Sistemático
3. Estratificado
4. Por conglomerados
5. Multietápico

<u>Tipos de Muestreo no probabilístico</u>: *Es aquel muestreo en el que la probabilidad de selección de cada unidad muestral no es igual ni conocida.*

TIPO DE MUESTREO NO PROBABILÍSTICO	VENTAJAS	DESVENTAJAS
<u>Por conveniencia:</u> Se seleccionan a las unidades de estudio que se encuentren disponibles al momento de la recolección de datos. Una variación de éste es el llamado muestreo a criterio o juicio donde además de encontrarse disponibles, se elige a los que se suponen más apropiados para participar en el estudio, generalmente es el investigador que en base a su experiencia realiza la elección.	Es más fácil, económico y accesible y puede dar una visión inicial buena. Se usa en estudios exploratorios	Puede ser poco representativo, algunas unidades estarán subrepresentadas y otras sobrerepresentadas

Por casos consecutivos: Consiste en elegir a cada paciente que cumpla con los criterios de selección dentro de un intervalo de tiempo específico o hasta alcanzar un número definido de pacientes.	Es el mejor y el más fácil de los muestreos no probabilísticos ya que su limitante solamente es la duración del estudio.	Su problema es precisamente cuando la duración es demasiado corta para representar adecuadamente todos los factores estacionales o cambios que puedan producirse con el tiempo y que sean importantes para la pregunta que se investiga (por ejemplo, prevalencia de infecciones respiratorias en un estudio que abarque dos meses e inicie en junio).
Por cuotas: Se seleccionan unidades de estudio de cada uno de los subgrupos que componen la población en una cuota predeterminada. Ej, si hablamos de edades, seleccionar un porcentaje de cada uno de los grupos de edad. Asegura que un determinado número de unidades de muestreo de diferentes categorías aparezcan en la muestra de modo que todos queden representados. Útil para balancear las unidades de estudio pero no se obtiene la representatividad de la población	Bajos costos y la mayor conveniencia para los entrevistadores al seleccionar los elementos para cada cuota	No permite la evaluación del error de muestreo
Por Bola de Nieve: Se selecciona un grupo inicial de entrevistados por lo general en forma aleatoria, después de la entrevista se pide a los participantes que identifiquen a otros que pertenecen a la población objetivo, por lo tanto los entrevistados subsecuentes se eligen en base a la referencias de los primeros. El proceso se continúa generando un efecto de bola de nieve	Permite estimar las características raras en la población. Bajos costos	No permite la evaluación del error de muestreo

La decisión de seleccionar uno u otro tipo de muestra dependerá :

- Del tipo de fenómeno a estudiar
- La oportunidad de acercamiento hacia los sujetos de estudio
- Los objetivos e hipótesis del estudio

Tipos de Muestreo Probabilístico: Es aquel muestreo donde el método de selección de la muestra permite que todos los elementos de la población tengan la ***misma probabilidad de ser seleccionados en la muestra.***

MUESTREO ALEATORIO SIMPLE

Cada individuo tiene la misma probabilidad de ser seleccionado para el estudio. Generalmente la selección se hace "sin reemplazo" esto es, que el individuo seleccionado no vuelve a ser tomado en cuenta para el sorteo.

Procedimiento:

1. Elaboración o construcción del Marco muestral en forma de lista. Cada unidad es identificada con un número
2. Aplicación de la tabla de números aleatorios. Seleccionar al primer elemento entre 1 y N
3. Selección del segundo elemento entre 1 y N. Si se repite se desecha
4. El proceso continua hasta completar los elementos de la muestra

Tabla de números aleatorios. Esta tabla es un conjunto de números enteros generado de modo que, comúnmente, la tabla contendrá todos los diez enteros (0,1,.....9), en proporciones aproximadamente iguales, sin tendencias en el patrón en que se generaron los dígitos. Si un número aleatorio ocurre dos veces, se omite la segunda ocurrencia y se selecciona otro número como su reemplazo.

927415	**956121**	168117	756409	536712	590261	196843
926937	515**107**	014658	436902	523498	490256	387130
867169	**388342**	670947	326078	638712	532780	683064
512500	542747	198302	251938	036528	280029	736209
729053	843**384**	105463	271167	129645	338639	393877
290366	**488**369	527892	190364	389462	462388	456297
337854	773025	837659	014517	639701	286593	649302
739285	536829	284561	746202	859274	183620	196387
483761	**479**401	026847	539028	274904	910477	690254
610537	993062	209385	598728	493672	290365	458926

Ejemplo: Si tenemos que seleccionar 70 estudiantes(muestra) de un listado general 500 estudiantes(marco muestral), entonces, se eligen tres dígitos y se empieza a seleccionar las unidades que conformarán la muestra. Se elige un punto de comienzo al azar, En este caso, empezamos por el 956(el cual no interviene), continuamos y es seleccionado, el 388, el 488, 479, 121, 107, 342, 384, 369, 025, 401, 062, 168, 014,..... y así sucesivamente, hasta completar los 70 estudiantes.

Con una calculadora científica también es posible obtener números aleatorios.

MUESTREO SISTEMÁTICO

Todas las unidades tienen la misma probabilidad de ser elegidos. Se incorpora un criterio importante al muestreo anterior, que es, el ORDEN, ello en función de un criterio que determina el investigador. Con este ordenamiento se gana en representatividad.

Todos los individuos se seleccionan a intervalos regulares, cada K elementos. Se selecciona dividiendo el total de población entre el número de elementos deseados lo que nos dará el intervalo de cada cuántos se eligen (por ejemplo, en una población de 300 elementos y un tamaño de muestra requerido de 60, 300/60 = 5, se escogerá cada quinto elemento). Puede tomarse el elemento inicial de cada grupo o el centro, aunque esto se comporta erráticamente, por lo que es preferible tomar el primer elemento de manera aleatoria los demás de acuerdo con la sistematización que se haya determinado
Por ejemplo, si el primer elemento elegido aleatoriamente fue el Nº 4, el siguiente será 4 + 5 = 9, el que le sigue será el 14, etc. ***No debe utilizarse cuando existe repetición cíclica inherente al marco de muestreo*** (por ejemplo los días de la semana).
Una ventaja sobre el aleatorio simple: Es más fácil sacar una muestra sin errores y ahorra tiempo. Desventajas sobre el aleatorio simple: El riesgo de sesgo es mayor.

CARACTERÍSTICAS:

- Asigna probabilidades iguales de selección
- No requiere Tabla de Números aleatorios
- Eficiente solo en poblaciones homogéneas
- La muestra se distribuye uniformemente en toda la población, siempre que exista una “buena” ordenación en el marco de muestreo
- Aplicable en encuestas de pequeña escala y la selección de campo
- Forma parte de diseños de muestra más complejos

PROCEDIMIENTO DE SELECCIÓN:

1. Ordenar los elementos de la población y trasladarlos al marco muestral mediante algún criterio de ordenamiento relacionada con la investigación.
2. Calcular el intervalo de selección

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

3. Seleccionar el arranque aleatorio entre 1 y k
4. Seleccionar las unidades a partir del arranque aleatorio, hasta completar el tamaño de la muestra

<u>Ejemplo:</u> Se tiene una población de 150 médicos y el tamaño de la muestra es de 30. Se decide el muestreo sistemático. Se desea investigar opinión de los médicos acerca del liderazgo que ejercen sus jefes, entonces, se elabora un marco muestral(150) ordenadas por el tiempo de servicios en la institución.

.

$$k = \frac{150}{30} = 5$$

Los 150 médicos están ordenados y numerados por años de servicio en la institución. Se elige la primera unidad muestral al azar entre 1 y 5, a partir de éste, se cuenta de cinco en cinco, hasta completar las 30 unidades. Se elige el nº 3 + 5 = 8; y así sucesivamente

Marco Muestral: 150 médicos numerados

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120
121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140
141	142	143	144	145	146	147	148	149	150										

MUESTREO ALEATORIO ESTRATIFICADO

Se divide primero a la población en estratos pertinentes (subgrupos) y luego de cada estrato se selecciona la muestra aleatoria, es decir, las extracciones de la muestra deben hacerse independientemente en los diferentes estratos (es una muestra aleatoria simple en cada estrato). Es posible sólo cuando se conoce la proporción de la población en estudio que pertenece a cada grupo de interés. Las subpoblaciones deben ser mutuamente excluyentes y en su conjunto corresponden a toda la población

El muestreo estratificado se utiliza en algunas situaciones como:

a) Cuando se requiere tener una precisión conocida en algunas subdivisiones de la población;
b) Por conveniencia administrativa;
c) Por dificultades específicas en algunas partes de la población, y
d) Para favorecer el análisis de grupos más homogéneos dentro de la heterogeneidad de la población.

En el muestreo estratificado puede mejorarse la precisión de la medición sobre el aleatorio simple si se cumplen tres requisitos que son:

a) La población consta de subconjuntos que varían mucho en tamaño;
b) Las principales variables a medir están íntimamente relacionadas con los tamaños de los subconjuntos y
c) Si se cuenta con una buena medida del tamaño para establecer los estratos.

El problema que se presenta es que ***la mejor asignación para una característica no necesariamente es la mejor para otra***, por lo que se sugiere reducir las características consideradas en la asignación a un número relativamente pequeño (es decir, estratificar de acuerdo con el menor número de variables en estudio posible), y calcular la asignación óptima para cada característica por separado y verificar hasta que punto existe desacuerdo.

La diferencia del muestreo aleatorio estratificado con el sistemático es que el sistemático estratifica la población en *n* estratos que consisten en las primeras k unidades, las segundas k unidades, etc. y las unidades ocurren en la misma posición relativa del estrato, mientras que en el aleatorio estratificado, la posición dentro del estrato se determina separadamente por aleatorización dentro de cada estrato.

La distribución de la muestra en función de los diferentes estratos se denomina afijación y puede ser de diferentes tipos:

- Afijación simple: A cada estrato le corresponde igual número de elementos muestrales
- Afijación Proporcional: A cada estrato le corresponde un número proporcional al tamaño del peso de la población en cada estrato
- Afijación Optima: En cada estrato se toma en cuenta además de la proporcionalidad, la dispersión de la variable.

CARACTERÍSTICAS

- Requiere información auxiliar de una o más variables para estratificar la población
- Eficiente en poblaciones heterogéneas
- Reduce costos
- Forma parte del diseño de muestras complejas
- Las estratificaciones se realizan a partir de la población, no de la muestra
- Es importante definir el número de estratos, lo que es decidido por el propio investigador
- Las muestras extraídas de cada estrato, son muestras independientes.
- El número y formación de los estratos depende del criterio del investigador

Ejemplo:

MUESTREO ESTRATIFICADO, CON AFIJACIÖN PROPORCIONAL Y SISTEMÁTICO

En la investigación sobre "Ambiente laboral y condiciones de salud de las enfermeras en los Hospitales de las Fuerzas Armadas, IPSS y Clínicas Privadas de Lima Metropolitana"

Estrato	Número	% respecto a "N"	Nº enfermeras de "n"
FFAA	1333	40.9	110.83 = 111
IPSS	1502	46.08	124.89 = 125
Clínicas	424	13.01	35.25 = 35
TOTAL	3259 enfermeras	99.99%	271 enfermeras

POBLACIÓN: 3259 enfermeras

MUESTRA: 271 enfermeras

(después de efectuar los cálculos respectivos)

Una vez determinado que son 271 enfermeras, tomando en cuenta el peso de cada estrato, se seleccionan 111 de las FFAA; 135 del IPSS y 35 de las Clínicas

HOSPITAL DE LAS FFAA	Tamaño de la Población	Tamaño de la muestra
Centro Médico Naval	316	26
Hospital Militar Central	371	31
Hosp. de la Sanidad de la	420	35
Hosp.Central de Aeronaútica	226	19
TOTAL	1333	111

Las 111 enfermeras deben ser seleccionadas de las 1333 que conforman todo este estrato. Para ello, se toma en cuenta nuevamente, el peso (proporción) que representa cada hospital respecto al estrato de las FFAA.

Dentro de cada Hospital, la investigadora decidió, el Muestreo sistemático
Ej. Centro Médico Naval

- Lista ordenada de las 316 enfermeras
- Se determinó el intervalo dividiendo 316/26= 12
- Se eligió al azar un número entre 1 y 12.
- Cada 12 enfermeras se eligieron las enfermeras(unidades muestrales) hasta completar las 26 enfermeras que participarán en la investigación.

MUESTREO POR CONGLOMERADO

Es la selección de grupos de unidades de estudio, en lugar de unidades de estudio individuales Generalmente son unidades geográficas u organizacionales. Ejemplo: Servicio de Medicina; Facultades de Ciencias de la Salud; Un conjunto de sectores de la comunidad de Villa El Salvador.
Su principal ventaja es que no se necesita el marco muestral de las unidades de estudio individuales. Su desventaja es que si no se incluyen en el estudio a todos los individuos de cada conglomerado se puede generar sesgo. Es un método menos preciso y requiere muestras de mayor tamaño. Su principal uso es en estudios epidemiológicos.

CARACTERÍSTICAS

- Las unidades de muestreo suelen ser un grupo de elementos que comúnmente es llamado conglomerado de elementos. El muestreo de estas unidades es llamado muestreo por conglomerado
- El marco muestral es una lista de conglomerados
- La medición se realiza a todos los elementos del conglomerado seleccionado
- Se utiliza en las investigaciones por muestreo a gran escala
- Reduce el costo del muestreo, al no utilizar una lista de elementos de la población
- Se pueden utilizar mapas de áreas territoriales como marco de muestreo. Es decir, se puede aprovechar, la organización existente: manzanas, los centros de salud, hogares, etc.

DESVENTAJAS

- Pérdida de información en las estimaciones si los conglomerados están “mal formados”
- Exceso de información al entrevistar a todos los elementos del conglomerado (ejem: si sale elegido una manzana con 50 casas(conglomerado) se tendría que entrevistar a las 50 viviendas; una alternativa, sería entrevistar 5 de las manzanas vecinas.
- La eficiencia de este tipo de muestreo disminuye al aumentar el tamaño del conglomerado

Ejemplo: En una investigación titulada “Factores de Riesgo reproductivo de la población femenina en edad fértil de la comunidad José Carlos Mariátegui del distrito de Villa María del Triunfo en 1991”realizada por una enfermera, se tomaron las siguiente acciones: La Comunidad José Carlos Mariátegui del distrito de Villa María del Triunfo está constituida VII Sectores; 515 manzanas, 12500 mujeres de 15 a 49 años:
MARCO MUESTRAL: Plano con viviendas
CONGLOMERADO: Cada sector de la comunidad JCM

UNIDAD MUESTRAL: Manzanas elegidas mediante muestreo de cada sector
UNIDAD DE OBSERVACIÓN: Madre de familia

LA ENCUESTA PILOTO en 10 manzanas, dio como resultado que:
Cada manzana habían 25 lotes y un aproximado de 2 mujeres en edad fértil en cada lote.
Por lo tanto: 8 mz x 25lotes= 200 lotes = 400 mujeres
En consecuencia debían elegirse 8 manzanas de toda la Comunidad José Carlos MariáteguiFinalmente la muestra quedó seleccionada de la siguiente manera:

SECTORES	Manz.	Mz.
Sector I: Gabriel Bajo	171	2
Sector II: 30 Agosto	56	1
Sector III: Vallecito Bajo	68	1
Sector IV: Vallecito Alto	104	2
Sector V: Gabriel Alto y Limatambo	116	2
TOTAL	**515**	**8**

MUESTREO MULTIETÁPICO

Se efectúa en pasos o fases (etapas) y habitualmente involucra más de un método de muestreo. Sus principales ventajas son que no se requiere un listado de las unidades de estudio, inicialmente el listado de los conglomerados es suficiente y luego sólo se requiere la lista de los conglomerados seleccionados y de la muestra de las unidades. Además, la muestra es más fácil de seleccionar ya que las unidades están físicamente unidas en grupos en vez de diseminadas en toda la población de estudio.

Su desventaja es que hay más probabilidad que la muestra final no sea representativa de la población y depende del número de conglomerados seleccionados en la primera etapa; a más conglomerados seleccionados existe mayor representatividad.

¿Cómo se calcula el tamaño de la muestra?

Existen diversas formas para calcular el tamaño de una muestra:

- emplear fórmulas preestablecidas
- emplear tablas precalculadas; o,
- emplear algún paquete estadístico.

El tamaño dependerá de:

1. Variabilidad de la característica de interés en la población: a mayor variabilidad , mayor tamaño de muestra. Está relacionado directamente con la varianza de la característica en estudio. Para el caso de las variables cuantitativas el valor está dado por la varianza; mientras que para el caso de las variables cualitativas, está dado por la proporción en que esté presente la variable en la población, multiplicado por su complemento.
2. Margen de error permisible (lo que está dispuesto a tolerar el investigador), se refiere al nivel de precisión o de aproximación, que el investigador desea tener respecto al valor real de la población. Esto lo decide el investigador. Significa en otras palabras, a cuantas unidades de la media poblacional o, en su defecto, a

cuantas unidades porcentuales de una característica determinada, desea el investigador, aproximarse con los resultados de su estudio.

3. Nivel de confianza: Está referido a la determinación probabilística de que una asociación o presencia de un fenómeno se observe por una asociación o presencia real de(l) (los) fenómeno(s) y que por lo tanto no obedezca al azar. Por convención, los más utilizados son 95% y 99%. En la fórmula está dado por los valores "z". Para 95%=1,96 y 99%=2,57

Algunas precisiones respecto a las fórmulas, que el investigador debe tomar en cuenta:

- A mayor variabilidad , mayor tamaño de la muestra.
- A mayor nivel de confianza, mayor tamaño de la muestra
- Cuánto más se aleje del verdadero valor de la población, menor será el tamaño de la muestra

TAMAÑO DE UNA MUESTRA PARA VARIABLES CUALITATIVAS

Fórmula del tamaño de muestra para una proporción

$$n = \frac{Z^2 pq}{d^2}$$

Descripción	
η	Tamaño de muestra
Z	1,96 = 95% confianza 2,57 = 99% confianza Nivel de confianza: Está referido a la determinación probabilística de que una asociación o presencia de un fenómeno se observe por una asociación o presencia real de(l) (los) fenómeno(s) y que por lo tanto no obedezca al azar. Por convención, los más utilizados son 95% y 99%. En la fórmula está dado por los valores "z". Para 95%=1,96 y 99%=2,57 Este valor indica el grado de confianza que se tendrá de que el verdadero valor del parámetro en la población caiga dentro del intervalo obtenido. Cuanta más confianza se desee, menor será el valor de α, mayor el valor de Zα y más elevado el número de sujetos necesarios.
p	proporción de casos de la población que tiene la característica que se desea estudiar Cuando se desconoce la proporción buscada, se utiliza p = 0.50 en la fórmula, que es la que proporciona el máximo valor de n.
q	1-p ó 100-p proporción de individuos de la población que no tiene la característica de interés y por tanto representa la probabilidad de obtener al azar un individuo sin esa característica.
d^2	margen de error permisible, establecido por el investigador. Cuanta más precisión se desee, más estrecho deberá ser este intervalo y más sujetos deberán ser estudiados.

Cuando el tamaño total de la población es menor de 5,000 (población finita), se requiere efectuar un ajuste en la fórmula:

$$nf = \frac{n}{1 + n/N}$$

n_f = corrección por tamaño de la muestra
N = Tamaño de la poblacional

Ejemplo: Se desea realizar un estudio sobre factores epidemiológicos y clínicos sobre el Asma Bronquial. Se conoce que la población consultante en un determinado año, corresponde a 2312 pacientes de los cuales el 85% corresponde a niños. Se quiere determinar cuál es el tamaño de muestra con un margen de error del 5% y con un nivel de confianza del 95%.
Solución:

$$n = \frac{1{,}96^{2}\, pq}{d^{2}} \qquad n = \frac{3{,}8416(85)(15)}{5^{2}} = 195{,}9$$

$$nf = \frac{n}{1 + n/N} \qquad nf = \frac{196}{1 + (196/2312)} = 181 pacientes$$

Respuesta: Se requiere 181 pacientes para realizar el estudio sobre factores epidemiológicos y clínicos sobre el Asma Bronquial.

TAMAÑO DE UNA MUESTRA PARA VARIABLES CUANTITATIVAS

Fórmula del tamaño de muestra para una media.

$$n = \frac{z^{2}\sigma^{2}}{d^{2}}$$

	Descripción
η	Tamaño de muestra
Z	1,96 = 95% confianza 2,57 = 99% confianza Nivel de confianza: Está referido a la determinación probabilística de que una asociación o presencia de un fenómeno se observe por una asociación o presencia real de(l) (los) fenómeno(s) y que por lo tanto no obedezca al azar. Por convención, los más utilizados son 95% y 99%. En la fórmula está dado por los valores "z". Para 95%=1,96 y 99%=2,57 Este valor indica el grado de confianza que se tendrá de que el verdadero valor del parámetro en la población caiga dentro del intervalo obtenido. Cuanta más confianza se desee, menor será el valor de α, mayor el valor de Zα y más elevado el número de sujetos necesarios.
σ^{2}	varianza de la población, (si no se conoce su valor, se estimará mediante una muestra piloto)
d^{2}	margen de error permisible, establecido por el investigador. Cuanta más precisión se desee, más estrecho deberá ser este intervalo y más sujetos deberán ser estudiados.

Cuando el tamaño total de la población es menor de 5,000 (población finita), se requiere efectuar un ajuste en la fórmula:

$$nf = \frac{n}{1 + n/N}$$

n_f = corrección por tamaño de la muestra
N = Tamaño de la poblacional

Ejemplo: Un grupo de investigadores desea estudiar edad de los niños al comenzar a caminar y su relación con habilidades psicomotrices. Revisan las historias clínicas de un hospital y encuentran 890 registrados. Desean obtener una muestra con un 95% de confianza y que el verdadero valor no exceda de 0.5 mes

Solución: Como se desconoce la varianza poblacional se realiza un estudio piloto y se obtiene que la varianza es igual a: 3,705

El promedio de edad de los niños es igual a: 12,08 meses

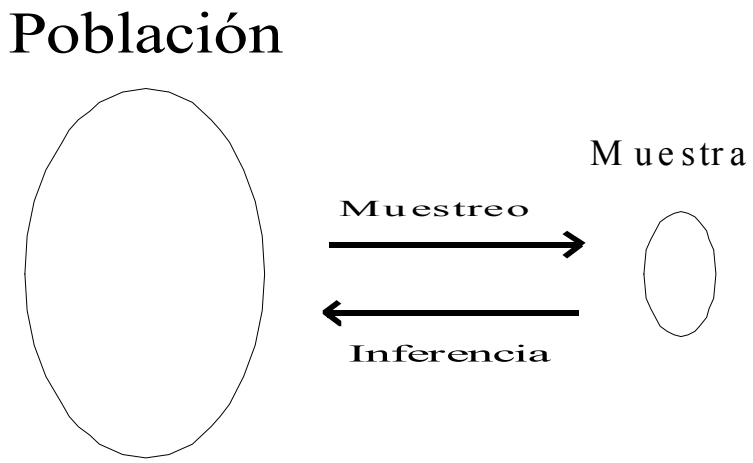
$$n = \frac{1.96^2 \sigma^2}{d^2} \qquad n = \frac{(3,8416)(3,705)}{0,25} = 56.93 = 57$$

$$nf = \frac{n}{1 + n/N} \qquad nf = \frac{57}{1 + 57/890} = 53,57$$

Respuesta: El tamaño de la muestra adecuado a fin de obtener un 95% de confianza y que el verdadero valor no exceda de 0,5 meses, corresponde a al intervalo entre 54 y 57 niños.

INFERENCIA ESTADÍSTICA

Ing. Luz Bullón Camarena



Rama de la estadística que, basada en conceptos de probabilidad, toma decisiones acerca de una población usando los resultados de una muestra extraída de esa población.
Existen dos procedimientos para la inferencia: estimación de parámetros y prueba de hipótesis

ESTIMACIÓN DE PARÁMETROS

Parámetro*,* es alguna característica descriptiva de los elementos de la población. Es un valor que queremos "estimar" con alguna exactitud razonable. Por ejemplo, la media de alguna variable cuantitativa, la proporción de algún atributo. Se puede hacer dos tipos de estimaciones: estimación puntual y estimación por intervalo

ESTIMACIÓN PUNTUAL

Es un número que estima el valor verdadero del parámetro desconocido de la población.
$\bar{x}$ la media de la muestra estima la media poblacional μ
p la proporción en la muestra, estima la proporción poblacional π

La estimación puntual es a menudo insuficiente, puesto que o acierta o se equivoca. Si está equivocada, se ignora el grado de error y no se puede estar seguro de la confiabilidad de la estimación. Por tanto, la estimación puntual es mucho más útil si se acompaña de una estimación del error que puede haber.

ESTIMACIÓN POR INTERVALO

Es un conjunto de valores que sirven para estimar el valor del parámetro de una

población. Indica el error en dos formas: por el tamaño del intervalo y por la probabilidad de que el verdadero valor del parámetro de la población se encuentre dentro de él.
En general, se expresa:

estimador ± coeficiente de confiabilidad × error estándar

• INTERVALO DE CONFIANZA PARA UNA PROPORCIÓN

Para estimar una proporción π de la población, se extrae una muestra de la población de interés y se calcula la proporción *p* en la muestra, luego, *el intervalo de confianza del* $100(1 - \alpha)\%$ se obtiene por medio de:

$$p \pm z \sqrt{\frac{p(1-p)}{n}}$$

La fórmula es usada cuando la muestra es de gran tamaño y la proporción de la población no está muy cerca de 0 ó de 1. Un criterio para usar de forma válida esta aproximación es que *n p* y *n (*1- *p*) deben ser mayores que 5.

Interpretación: Se tiene el $100(1 - \alpha)\%$ de confianza de que el intervalo calculado, contenga la proporción poblacional del atributo de interés

EJEMPLO

En una muestra de 120 pacientes de una población de pacientes infartados, se encontraron, entre otros, los siguientes resultados.

	N° pacientes	*P* (%)
Obesidad	36	0.30 (30%)
Diabetes	18	0.15 (15%)

Estime la proporción poblacional de obesos de esa población y encuentre un intervalo de 95% de confianza.

Solución:
La proporción de obesos en la población: π se estima por *P* = 0.30 (30%)
Calculando el intervalo de 95% de confianza, (α= 0.05) para π :

$0.30 \pm 1.96 \sqrt{\frac{0.30 \times 0.70}{120}}$ se obtiene [0.218 , 0.382]

Interpretación:
Con 95% de confianza, el intervalo encontrado de 21.8% a 38.2%, incluirá la proporción de obesos en la población de pacientes infartados.

INTERVALO DE CONFIANZA PARA LA MEDIA

Cuando el muestreo se realiza a partir de una población con distribución normal con variancia conocida el intervalo para μ se expresa como

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

Interpretación: Se tiene el 100(1- α)% de confianza de que el intervalo calculado, contenga la media de la población

Cuando la variancia poblacional es desconocida, lo que generalmente sucede si se desconoce la media, si el tamaño de la muestra es grande, se puede confiar en "*s"* como una aproximación de σ y e justifica la utilización de la teoría de la distribución normal. Cuando no es posible suponer que la población de interés tiene distribución normal, por el teorema central del límite, el intervalo anterior sirve si se puede observar una muestra suficientemente grande (n>30), pues la media de la muestra presenta una distribución aproximadamente normal sin importar cómo está distribuida la población original.

EJEMPLO

La muestra simple aleatoria de 120 pacientes de una población de pacientes infartados, proporciona entre otros, los siguientes resultados:

	Media $\bar{x}$	Desv. Estándar *s*
Edad	66.3	10.49
Colesterol Total	222.7	57.1

Estime la edad promedio en la población y encuentre un intervalo de confianza del 95%.

Solución:

La edad promedio en la población se estima con el promedio muestral. La estimación puntual de la media poblacional es 66 años.

Calculando el intervalo de confianza para μ

$66.3 \pm 1.96 \frac{10.49}{\sqrt{120}}$ y el resultado es: [64.42 , 68.18].

Interpretación:

Con 95% de confianza la verdadera edad promedio de la población de pacientes infartados, se encontrará en el intervalo de 64 a 68 años.

LA DISTRIBUCIÓN T

Cuando se tienen muestras pequeñas (30 o menos) la alternativa es el intervalo

$$\bar{x} \pm t \frac{s}{\sqrt{n}}$$

Se resalta que la muestra debe ser obtenida de una población normal, tolerándose desviaciones moderadas de este requisito.

PRUEBA DE HIPÓTESIS

Ing. Luz Bullón Camarena

En muchas situaciones el investigador tiene alguna idea, o conjetura, sobre el comportamiento de una variable, o de una posible asociación entre variables. En estos casos el diseño y planeamiento de la investigación debe ser de tal forma que permita con los datos muestrales, probar la veracidad de sus ideas sobre la población en estudio. La idea o conjetura es una *hipótesis* y se hará una *prueba de hipótesis*.
La finalidad de la prueba de hipótesis, no es poner en tela de juicio el valor calculado del estadístico muestral, sino emitir un juicio sobre la *diferencia* existente entre él y un supuesto parámetro de la población.

EJEMPLO 1:
Suponga que el responsable de Salud Pública de una población infantil afirma que la proporción de desnutridos es a lo más 0.08 (8%). Por su parte, estudiantes de la Facultad de Medicina sospechan de la veracidad de esta afirmación y deciden llevar a cabo un estudio de prevalencia.
Eligen aleatoriamente una muestra de 150 niños de dicha población, en la muestra encuentran 12% de desnutridos.¿Qué se puede concluir respecto a la afirmación inicial?

CONCEPTOS BÁSICOS DEL PROCEDIMIENTO

HIPÓTESIS NULA - Suposición o conjetura que se hace sobre el valor del parámetro de la población *antes* de empezar el muestreo, generalmente una suposición del "status quo" (situación actual). Se representa con el símbolo H_0

En el ejemplo: $H_0 : \pi \leq 0.08$
La proporción de desnutridos en la población es a lo más 0.08 (8%).

HIPÓTESIS ALTERNATIVA - Conclusión que se acepta cuando los datos no apoyan la hipótesis nula. Se representa simbólicamente H_1. Generalmente es la hipótesis del investigador.

En el ejemplo: $H_1 : \pi > 0.08$

DECISIONES CORRECTAS Y ERRORES EN LA PRUEBA DE HIPÓTESIS

Al probar una hipótesis realmente se está tomando una decisión entre dos acciones, una decisión entre H_0 y H_1.
La veracidad o falsedad de una hipótesis en particular nunca puede conocerse con certidumbre, a menos que pueda examinarse a toda la población. Por tanto, el procedimiento tiene en cuenta la probabilidad de llegar a una conclusión equivocada.

		Condición de la hipótesis nula	
		Verdadera	Falsa
Acción posible	No rechazar H_0	Acción correcta	Error tipo II
	Rechazar H_0	Error tipo I	Acción correcta

ERROR DE TIPO I - Rechazo de una hipótesis nula cuando es verdadera. La probabilidad de cometer este error al tomar una decisión se denomina **NIVEL DE SIGNIFICACIÓN** y se denota con la letra griega α (alfa). Valores típicos, fijados de antemano para α son 0.05 , 0.01 ó 0.10

ERROR DE TIPO II - Aceptación de una hipótesis nula cuando es falsa. La probabilidad de un error de tipo II se denota con la letra griega β (beta)

PASOS DEL PROCEDIMIENTO DE LA PRUEBA DE HIPÓTESIS

1. Identificar la variable aleatoria y los parámetros de interés
2. Formular las hipótesis
3. Fijar el nivel de significación
4. Seleccionar la prueba estadística
5. Formular la regla de decisión
6. Calcular la estadística de prueba
7. Formular la decisión estadística
8. Expresar la conclusión en términos del problema de investigación.

1.- PRUEBA DE HIPÓTESIS REFERIDA A UNA PROPORCIÓN

Para el EJEMPLO 1,

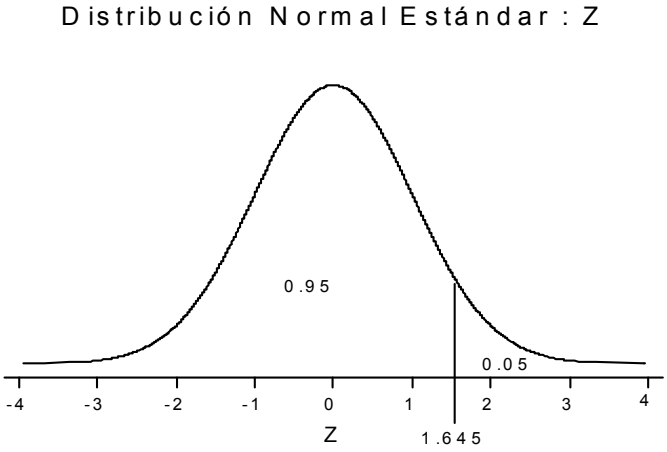
Paso 1. La variable en estudio es cualitativa, el parámetro de interés: la proporción poblacional π de desnutridos

Paso 2. Las hipótesis: $H_0 : \pi \leq 0.08$ versus $H_1 : \pi > 0.08$

Paso 3. Nivel de significación α=0.05

Paso 4. La prueba estadística es Z, la muestra es grande

Paso 5. Regla de decisión. La prueba es unilateral, hay una región de rechazo. La decisión es: rechazar la hipótesis nula sí el valor calculado de la estadística de prueba resulta mayor que el valor Z de la tabla de distribución normal estándar. Es decir, Rechazar H_0 sí $Z_{calc} > Z_{0.95} = 1.64$



Paso 6. Cálculo de la estadística de prueba:

$$Z_{calc} = \frac{p - \pi_0}{\sqrt{\frac{\pi_0 (1 - \pi_0)}{n}}} = \frac{0.12 - 0.08}{\sqrt{\frac{0.08(1-0.08)}{150}}} = 1.8058$$

Paso 7. Se rechaza H_0. La prueba resultó significativa.

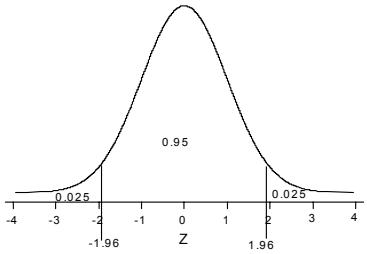
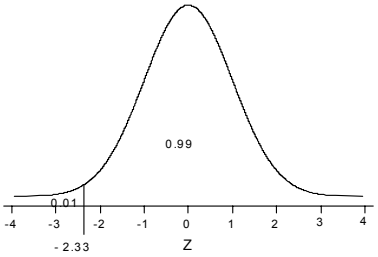
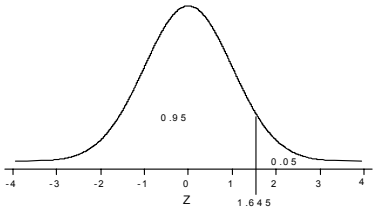
Paso 8. Es posible concluir que en la población bajo estudio, la proporción de individuos desnutridos supera el 8% (α=0.05), por lo tanto la afirmación del responsable es incorrecta.

FÓRMULA Y REGLA DE DECISIÓN

- Requisito: Muestra grande, n>50
- La estadística de prueba (para el cálculo del paso 6), en todos los casos es:

$$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0 (1 - \pi_0)}{n}}}$$

- Hay tres tipos de prueba, por la manera en que se formulan H_0 y H_1:

Tipos de prueba	Regla de decisión (paso 4)
• Prueba bilateral o de doble cola. $H_0 : \pi = \pi_0$ $H_1 : \pi \neq \pi_0$	Distribución Normal Estándar : Z  Valores tabulares para la prueba: $Z_{\alpha/2}$ y $Z_{1-\alpha/2}$
• Prueba unilateral de cola inferior. $H_0 : \pi \geq \pi_0$ $H_1 : \pi < \pi_0$	Distribución Normal Estándar : Z  Valor tabular para la prueba: Z_{α}
• Prueba unilateral de cola superior. $H_0 : \pi \leq \pi_0$ $H_1 : \pi > \pi_0$	Distribución Normal Estándar : Z  Valor tabular para la prueba: $Z_{1-\alpha}$

2.- PRUEBA DE HIPÓTESIS REFERIDA A LA DIFERENCIA ENTRE LAS PROPORCIONES DE DOS POBLACIONES

EJEMPLO 2:

En un estudio comparativo de obesidad se sospecha que la proporción de obesos es mayor en la población femenina. Se obtuvieron los siguientes resultados a partir de muestras de hombres y mujeres entre las edades de 20 y 75 años:

	Tamaño de la muestra	*N° de individuos con sobrepeso*
Varones	150	21
Mujeres	200	48

Paso 1. La variable en estudio es cualitativa, el parámetro de interés es: la proporción poblacional π de obesos en cada población

Paso 2. Las hipótesis: $H_0 : \pi_V = \pi_M$ o equivalentemente, $H_0 : \pi_V - \pi_M = 0$

$H_1 : \pi_V < \pi_M$

Paso3. Nivel de significación α=0.01

Paso 4. Prueba estadística Z (ver hoja de fórmulas), las muestras son grandes.

Paso 5. Regla de decisión. La prueba es unilateral, hay una región de rechazo. La decisión es: Rechazar la hipótesis nula sí el valor calculado de la estadística de prueba resulta menor que el valor de la tabla de distribución normal estándar Z. Es decir, Rechazar H_0 sí $Z_{calc} < Z_{tabla}$. El valor de tabla que corresponde al percentil inferior 0.01 (puesto que α=0.01) es $Z_{tab} = -2.33$

Paso 6. Cálculo de la estadística de prueba: (hoja de fórmulas)

Cálculos previos. $p_V = 21/150 = 0.14$ $p_M = 48/200 = 0.24$

$$\bar{p} = \frac{21+48}{150+200} = \frac{69}{350} = 0.1971$$

$$Z_{calc} = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)_0}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}} = \frac{(0.14-0.24)-0}{\sqrt{0.1971(0.8029)\left(\frac{1}{150}+\frac{1}{200}\right)}} = -2.327$$

Paso 7. Se rechaza H_0. La prueba resultó significativa.

Paso 8. Es posible concluir, a un nivel de significación de 1%, que la proporción de individuos obesos es menor en la población masculina.

$H_0 : \pi_1 = \pi_2$ o equivalentemente, $H_0 : \pi_1 - \pi_2 = 0$

Requisitos (Suposiciones)	*Estadística de prueba*	*Valores tabulares para la prueba*	
		Unilateral	*Bilateral*
Muestras grandes e independientes	$Z = \dfrac{(p_1 - p_2) - (\pi_1 - \pi_2)_0}{\sqrt{\dfrac{\bar{p}(1-\bar{p})}{n_1} + \dfrac{\bar{p}(1-\bar{p})}{n_2}}}$ donde $\bar{p} = \dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$	$H_1 : \pi_1 < \pi_2$ z_α	$H_1 : \pi_1 \neq \pi_2$ $z_{\alpha/2}$ $z_{1-\alpha/2}$
		$H_1 : \pi_1 > \pi_2$ $z_{1-\alpha}$	

3.- PRUEBA DE HIPÓTESIS REFERIDA A UNA MEDIA

EJEMPLO 3.

Muchos pacientes con cifoescoliosis desarrollan incapacidad pulmonar que puede conducir a insuficiencia respiratoria. Lisboa y otros (1985) deseaban valorar la función de músculos inspiratorios en pacientes adultos con cifoescoliosis grave. Estudiaron nueve adultos con cifoescoliosis en un estudio transversal. La capacidad pulmonar total (CPT) y la capacidad vital forzada se encontraron muy disminuidos en los pacientes cuando se compararon con un grupo normal. La presión inspiratoria máxima (Pimax) es una medición que refleja la fuerza combinada de todos los músculos respiratorios. La media en adultos normales es de 100 cm H_2O y se puede suponer que la desviación estándar es 20.

En el cuadro 1 se encuentran los datos obtenidos.

CUADRO 1 – Presión inspiratoria por la boca (Pimax) en pacientes con cifoescoliosis

Número de paciente	Pimax (cm H_2O)
1	44.8
2	62.0
3	63.3
4	84.2
5	80.3
6	66.3
7	69.3
8	94.6
9	76.6
Media	71.27
Desviación estándar	14.58

Pruebe la hipótesis correspondiente siguiendo los pasos del procedimiento de la prueba.

Paso 1. La variable en estudio es numérica, presión inspiratoria máxima (Pimax).
El parámetro de interés: μ, la media poblacional en pacientes con cifoescoliosis
La variancia poblacional se conoce, se puede suponer $\sigma = 20$

Paso 2. Las hipótesis:

$$H_0 : \mu = 100$$
$$H_1 : \mu < 100$$

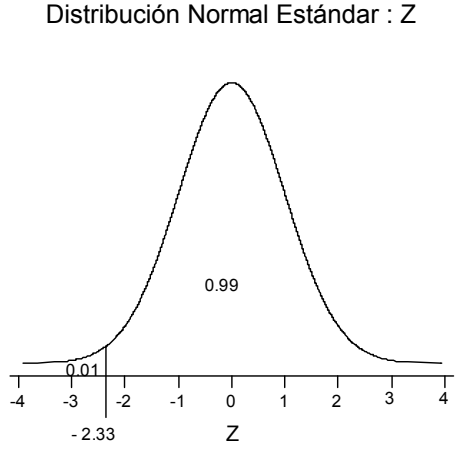
Paso 3. Nivel de significación α=0.01

Paso 4. La prueba estadística es Z. El problema se refiere a la media de una población, la

desviación estándar de la población es conocida (supuesta σ = 20) y es razonable afirmar que la distribución de la variable es normal.

Paso 5. Regla de decisión

La prueba es unilateral, hay una región de rechazo.



La decisión es: rechazar la hipótesis nula sí el valor calculado de la estadística de prueba resulta menor que el percentil 1% (puesto que □=0.01) de la tabla de distribución normal estándar Z. Es decir, rechazar H_0 sí $Z_{calc} < Z_{0.01}$= -2.33

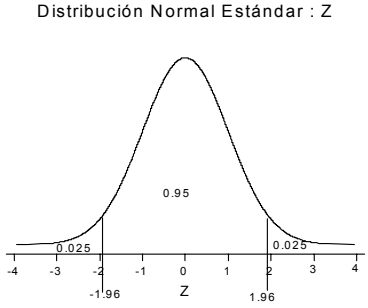
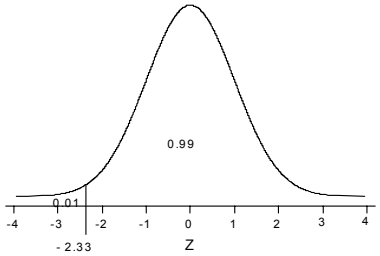
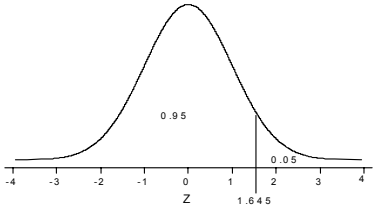
Paso 6. Cálculos. La media de los datos es $\bar{x} = 71.27$, luego por la fórmula, la estadística de prueba es

$$Z_{calc} = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{71.27 - 100}{20 / \sqrt{9}} = -4.31$$

Paso 7. La decisión es: rechazar H_0.

Paso 8. La prueba resultó significativa. Es posible concluir (al nivel de significación de 0.01) que la presión inspiratoria máxima (Pimax) en pacientes con cifoescoliosis es menor que 100 cm H_2O, correspondiente a individuos normales.

HIPÓTESIS Y REGLAS DE DECISIÓN

Tipos de prueba	Regla de decisión (paso 5)
• Prueba bilateral o de doble cola. $H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	 Valores tabulares para la prueba: $Z_{\alpha/2}$ y $Z_{1-\alpha/2}$
• Prueba unilateral de cola inferior. $H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$	 Valor tabular para la prueba: Z_{α}
• Prueba unilateral de cola superior. $H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$	 Valor tabular para la prueba: $Z_{1-\alpha}$

Requisitos (Suposiciones)	*Prueba Estadística*	*Cálculo de la estadística de prueba*
1. σ^2 conocida Población normal	*Z*	$Z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$
2. Muestra grande ($s \approx \sigma$)		
3. Muestra pequeña Población normal	*t*	$t = \dfrac{\bar{x} - \mu_0}{s / \sqrt{n}}$

EJEMPLO 4.

Una compañía farmacéutica afirma que una cierta cápsula contiene en promedio 2.50 miligramos de un determinado medicamento. Una oficina de protección al consumidor obtuvo una muestra aleatoria de 20 cápsulas y midió la cantidad del medicamento en cada cápsula. Los resultados son los siguientes:

2.68	2.57	2.48	2.51	1.62	2.70	2.34	2.11	2.71	3.02
2.78	2.56	2.80	3.50	2.75	2.60	3.19	2.82	3.48	3.09

Si se sabe que la variable contenido por cápsula se distribuye normalmente, realice una prueba de hipótesis para probar si lo que afirma la compañía farmacéutica es aceptable, a un nivel de significación del 5%.

Paso 1. La variable es cuantitativa: contenido por cápsula del medicamento, en mg.

El parámetro de interés: μ, el contenido medio poblacional (de la producción)

Paso 2. Las hipótesis:

H_o: El verdadero contenido promedio por cápsula es igual a 2.50 mg

H_1: El verdadero contenido promedio por cápsula es diferente de 2.50 mg

En símbolos, $H_o: \mu = 2.50$

$H_1: \mu \neq 2.50$

Paso 3. Nivel de significación α=0.05

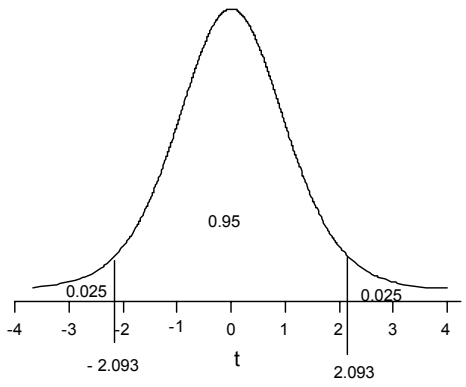
Paso 4. Prueba estadística

Prueba t, el problema se refiere a la media de una población, la muestra es pequeña (n = 20), la variancia de la población se desconoce. Se sabe que la variable tiene distribución normal.

Paso 5. Regla de decisión

La prueba es bilateral, las regiones de rechazo:

Distribución t (19 g.l.)



La hipótesis alternativa H_1 es de dos lados:

$\mu < 2.50$ ó $\mu > 2.50$ (μ diferente al valor en H_o)

La decisión es: rechazar la hipótesis nula sí el valor calculado de la estadística de prueba resulta menor que el valor del percentil 0.025 o mayor que el valor del percentil 0.975 de la distribución t de student con 19 grados de libertad.
Es decir, rechazar H_0 sí $t_{calc} < t_{(19)\ 0.025} = -2.093$ ó $t_{calc} > t_{(19)\ 0.975} = 2.093$

Paso 6. Cálculos

$n = 20$ $\bar{x} = 2.7155$ $s^2 = 0.4321^2$

La estadística de prueba es

$$t_{calc} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{2.7155 - 2.5}{0.4321 / \sqrt{20}} = 2.23$$

Paso 7. La decisión es rechazar H_0, 2.2304 > 2.093. La prueba resultó significativa.

Paso 8. Hay suficiente evidencia estadística para concluir que el contenido promedio de las cápsulas es diferente a 2.50 mg. El contenido promedio de las cápsula en la producción parece ser mayor que el contenido promedio especificado de 2.50 mg, siendo la diferencia significativa estadísticamente a un nivel de significación del 5%.

SOLUCIÓN UTILIZANDO MINITAB

One-Sample T: Contenido

Test of mu = 2.5 vs mu not = 2.5

Variable	N	Mean	StDev	SE Mean
Contenido	20	2.7155	0.4321	0.0966

Variable	95.0% CI	T	P
Contenido	(2.5133, 2.9177)	2.23	0.038

EJEMPLO 5. La infección por *E. Canis* es una enfermedad canina trasmitida por la garrapata, que algunas veces contraen los seres humanos. En la población general, el recuento medio de glóbulos blancos es 7250/mm^3. Se cree que las personas infectadas con *E. canis* deben tener en promedio un recuentos de glóbulos blancos más bajos.
Para una muestra de 11 personas infectadas, el recuento medio de glóbulos blancos, mm^3 fue el siguiente:

477 6501 689 6044 7242 2558 3149 1878 3215 4848 2093

¿Qué concluye Ud. a un nivel de significación de 0.05?

Siguiendo los pasos del procedimiento de la prueba de hipótesis:

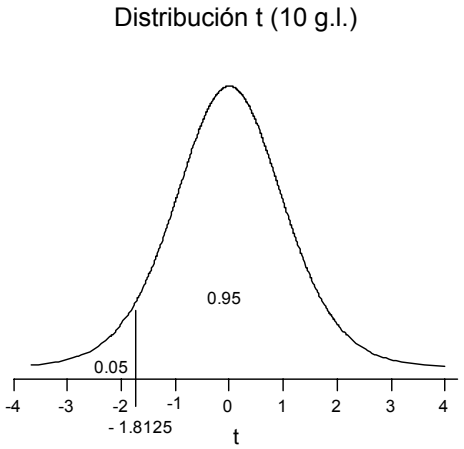
Paso 1. La variable en estudio es numérica: recuento de glóbulos blancos
El parámetro de interés: μ, la media poblacional de individuos infectados por *E. Canis*

Paso 2. Las hipótesis: $H_0 : \mu \geq 7250$
$H_1 : \mu < 7250$

Paso 3. Nivel de significación α=0.05

Paso 4. Prueba estadística (ver hoja de fórmulas)
Prueba t, el problema se refiere a la media de una población, la muestra es pequeña, la desviación estándar de la población se desconoce, y es razonable afirmar que la variable tenga distribución normal.

Paso 5. Regla de decisión
La prueba es unilateral, hay una región de rechazo.



La decisión es: rechazar la hipótesis nula sí el valor calculado de la estadística de prueba resulta mayor que el valor del percentil 0.95 de la distribución t de student con 10 grados de libertad.
Es decir, rechazar H_0 sí $t_{calc} < t_{(10)\ 0.05} = -1.8125$

Paso 6. Cálculos

La media de los 11 datos es $\bar{x} = 3518$, la desviación estándar de los datos es s = 2325. Luego, la estadística de prueba es

$$t_{calc} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{3518 - 7250}{2325 / \sqrt{11}} = -5.33$$

Paso 7. La decisión es rechazar H_0. La prueba resultó significativa.

Paso 8. Es posible concluir que el recuento medio de glóbulos blancos en la población de individuos infectados por *E. Canis* sea menor que en la población general de 7250 mm^3

Solución utilizando MINITAB

T-Test of the Mean

Test of mu = 7250 vs mu < 7250

Variable	N	Mean	StDev	SE Mean	T	P
GLOBB	11	3518	2325	701	-5.33	0.0002

4.- DIFERENCIA DE DOS MEDIAS (MUESTRAS INDEPENDIENTES)

EJEMPLO 6.
En un estudio de factores que se consideran responsables de los efectos adversos del tabaquismo sobre la reproducción humana, se midieron los niveles de cadmio (nanogramos por gramo) en el tejido de la placenta de una muestra de 14 mujeres embarazadas que fumaban y una muestra aleatoria independiente de 18 mujeres no fumadoras. Los resultados se detallan a seguir. Se quiere saber si es posible afirmar que el nivel medio de cadmio registrado es mayor entre las fumadoras que entre las no fumadoras?

Fumadoras:	30.0	30.1	15.0	24.1	30.5	17.8	16.8	14.8	13.4	28
	17.5	14.4	12.5	20.4						
No fumadoras:	10.0	8.4	12.8	25.0	11.7	9.8	12.5	15.4	23.5	9.4
	25.1	19.5	25.5	9.8	7.5	11.8	12.2	15.0		

Paso 1. La variable en estudio es numérica, nivel de cadmio registrado en la placenta
El parámetro de interés: la diferencia de las medias de dos poblaciones: μ_1 media en mujeres fumadoras y μ_2 la media en mujeres no fumadoras.

Paso 2. $H_0 : \mu_F = \mu_{NF}$ o equivalentemente, $H_0 : \mu_F - \mu_{NF} = 0$
$H_1 : \mu_F > \mu_{NF}$

Paso 3. Nivel de significación α=0.05

Paso 4. Prueba estadística (ver fórmulas)
Prueba t, el problema se refiere a la diferencia de dos medias, las muestras son pequeñas, no se conoce la variancia poblacional

Paso 5. Regla de decisión
La prueba es unilateral, hay una región de rechazo del lado superior de la distribución. La decisión es: Rechazar la hipótesis nula sí el valor calculado de la estadística de prueba resulta mayor que el valor de la tabla de distribución t de student con $(n_1 - 1) + (n_2 - 1) = 14 + 18 - 2 = 30$ grados de libertad. Es decir, Rechazar la H_0 sí $t_{calc} > t_{tabla}$.
El valor de tabla que corresponde al percentil superior 0.95 (puesto que α=0.05) es $t_{0.95}$=1.6973

Paso 6. Cálculo de la estadística de prueba (ver fórmulas)

	n	$\bar{x}$	s^2	*s*
Fumadoras	14	20.41	46.37	6.81
No fumadoras	18	14.72	38.44	6.20

Variancia ponderada

$$s_p^2 = \frac{(n_1 - 1)\ s_1^2 + (n_2 - 1)\ s_2^2}{n_1 + n_2 - 2} = \frac{(14 - 1)\,46.37 + (18 - 1)\ 38.44}{14 + 18 - 2} = 41.861$$

El valor de la estadística de prueba

$$t_{calc} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{(20.41 - 14.72) - 0}{\sqrt{41.861\left(\frac{1}{14} + \frac{1}{18}\right)}} = 2.47$$

Paso 7. La decisión es rechazar H_0, pues $t_{calc} > t_{0.95}$

Paso 8. Es posible concluir, a un nivel de significación de 5%, que el nivel medio de cadmio en la placenta en una población de mujeres embarazadas y fumadoras es mayor que en una población de comparable con la anterior pero que no fuman

$H_0 : \mu_1 = \mu_2$, o equivalentemente, $H_0 : \mu_1 - \mu_2 = 0$

Requisitos (Suposiciones)	*Estadística de prueba*	*Valores tabulares*	
		unilateral	*bilateral*
σ_1^2 , σ_2^2 son desconocidas pero se supone que son iguales* Muestras pequeñas Poblaciones normales	$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$ donde $s_p^2 = \frac{(n_1 - 1)\ s_1^2 + (n_2 - 1)\ s_2^2}{n_1 + n_2 - 2}$	$H_1 : \mu_1 < \mu_2$ $t_{(\alpha,\ n_1 + n_2 - 2)}$ $H_1 : \mu_1 > \mu_2$ $t_{(1-\alpha, n_1 + n_2 - 2)}$	$H_1 : \mu_1 \neq \mu_2$ $t_{(\alpha/2, n_1 + n_2 - 2)}$ $t_{(1-\frac{\alpha}{2}, n_1 + n_2 - 2)}$
σ_1^2 , σ_2^2 desconocidas pero diferentes** Muestras pequeñas Poblaciones normales	$t' = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ donde $t'_{(tabla)} = \frac{\frac{s_1^2}{n_1} t_1 + \frac{s_2^2}{n_2} t_2}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	con $t_1 = t_{(1-\alpha, n_1 - 1)}$ $t_2 = t_{(1-\alpha, n_2 - 1)}$ luego, $H_1 : \mu_1 < \mu_2$ $-t'_{(tabla)}$ $H_1 : \mu_1 > \mu_2$ $+t'_{(tabla)}$	$t_1 = t_{(1-\frac{\alpha}{2}, n_1 - 1)}$ $t_2 = t_{(1-\frac{\alpha}{2}, n_2 - 1)}$ $H_1 : \mu_1 \neq \mu_2$ $-t'_{(tabla)}$ $+t'_{(tabla)}$

* ** En la práctica esta suposición hay que probarla usando la prueba F de homogeneidad de variancias.

5.- DIFERENCIA DE DOS MEDIAS (MUESTRAS DEPENDIENTES, PAREADAS O EN PAREJAS)

Dato pareado es un dato bivariado (x,y) que corresponde a:

- Dos variables obtenidas para un mismo elemento de la población.
- Una variable obtenida en un mismo elemento de la población, en dos momentos distintos o por dos observadores.

La prueba compara los valores del par observado. Se toma la pareja de datos de la i-ésima observación y se obtiene la diferencia $x_i - y_i$, la cual puede ser cero, mayor que 0 ó menor que 0, es decir, tiene signo + ó -.

EJEMPLO 7. Doce individuos participaron en un experimento para estudiar la efectividad de cierta dieta, combinada con un programa de ejercicios, para la reducción de los niveles de colesterol en suero. ¿Existe la evidencia suficiente para concluir que el programa de ejercicios y dieta resultaron efectivos para la reducción de los niveles de colesterol en el suero?
La tabla siguiente muestra los niveles de colesterol en suero para los doce individuos al principio del programa (antes) y al final del mismo (después). La tabla también contiene las diferencias entre las dos mediciones

Individuo	***Colesterol en suero***		*Diferencia* $d_i = X_2 - X_1$
	Antes X_1	*Después* X_2	
1	201	200	-1
2	231	236	+5
3	221	216	-5
4	260	233	-27
5	228	224	-4
6	237	216	-21
7	326	296	-30
8	235	195	-40
9	240	207	-33
10	267	247	-20
11	284	210	-74
12	201	209	+8

Paso 1. La variable en estudio es numérica, es la diferencia en los niveles de colesterol antes y después de un programa experimental para reducción: $d_i = X_2 - X_1$
El parámetro de interés es la media de las diferencias individuales: μ_d
Paso 2. Planteamiento de las hipótesis
$H_0 : \mu_d = 0$ o equivalentemente, $H_0 : \mu_1 - \mu_2 = 0$
$H_1 : \mu_d < 0$

La hipótesis alternativa depende del sentido de la diferencia, en el ejemplo, se ha definido $d_i = X_2 - X_1$. Luego el programa será efectivo si la media de las diferencias es negativa.
Paso 3. Nivel de significación α=0.05
Paso 4. Prueba estadística t (ver fórmulas), la muestra es pequeña, la desviación

estándar poblacional es desconocida

Paso 5. Regla de decisión. La prueba es unilateral, la región de rechazo está del lado inferior de la distribución t de student con *n*-1 grados de libertad.
Se rechazará la hipótesis nula sí el valor calculado de la estadística de prueba resulta menor que el valor tabular, es decir, sí $t_{calc} < t_{0.05}$ = -1.7959

Paso 6. Cálculo de la estadística de prueba (ver fórmulas)

Cálculos previos, el promedio de las diferencias

$$\bar{d} = \frac{\sum d_i}{n} = \frac{-1+5+(-5)+\ldots+8}{12} = -20.17$$

Variancia de las diferencias

$$s_d^2 = \frac{\sum d_i^2 - n\,\bar{d}^{\,2}}{n-1} = 534.9969$$

El valor calculado de la estadística de prueba

$$t_{calc} = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} = \frac{-20.17}{23.13/\sqrt{12}} = -3.02$$

Paso 7. La decisión es rechazar H_0, pues $t_{calc} < t_{0.05}$.

Paso 8. Se puede concluir a un nivel de significación de 5%, que el programa de dieta y ejercicios resultó efectivo en la reducción de los niveles de colesterol

$H_0 : \mu_1 = \mu_2$ o equivalentemente, $H_0 : \mu_d = 0$

Requisitos (Suposiciones)	*Estadística de prueba*	*Valores tabulares*	
		Unilateral	*bilateral*
Distribución normal Existe correlación entre muestras	$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$ $s_d^2 = \frac{\sum d_i^2 - n\,\bar{d}^{\,2}}{n-1}$	$H_1 : \mu_d < 0$ $t_{(\alpha,\ n-1)}$ $H_1 : \mu_d > 0$ $t_{(1-\alpha,n-1)}$	$H_1 : \mu_d \neq 0$ $t_{(\alpha/2,n-1)}$ $t_{(1-\alpha/2,n-1)}$

COMPARACIÓN DE TRES Ó MÁS MEDIAS

Ing. Luz Bullón Camarena

Para la comparación de más de dos grupos, las pruebas Z o t no pueden aplicarse. Si se efectúan múltiples pruebas entre los pares diferentes de medias, el nivel de significación α empleado en cada comparación, se altera (incrementándose) respecto al nivel de significación de una prueba global o del experimento como un todo.
Los datos deben ser analizados mediante el **Análisis de Variancia - ANVA**. Este procedimiento evita esta alteración de α

Los datos muestrales:

	Grupos (tratamientos)				Total
	1	2	...	k	
	Y_{11} Y_{12} . . . Y_{1n1}	Y_{21} Y_{22} Y_{2n2}	Y_{ij}	Y_{k1} Y_{k2} Y_{knk}	
Nº de unidades Totales Medias Variancias	n_1 $\sum_j Y_{1j} = Y_1$ $\overline{Y}_1$ s_1^2	n_2 $\sum_j Y_{2j} = Y_2$ $\overline{Y}_2$ s_2^2		n_k $\sum_j Y_{kj} = Y_k$ $\overline{Y}_k$ s_k^2	$n = \sum n_i$ $\sum_i \sum_j Y_{ij} = Y$ $\overline{Y}$

Y_{ij} observación, *j*-ésima perteneciente al grupo o tratamiento *i*
k número de grupos comparados o tratamientos
n_i número de observaciones del *i*-ésimo tratamiento
n número total de observaciones del estudio
$\overline{Y}$ promedio general de todas las observaciones

El ANVA responde en un principio, si la media de alguno de los grupos es diferente de las demás o si hay una diferencia cualquiera entre los grupos. Si el ANVA resulta significativo, es decir si se ha encontrado alguna diferencia, se pueden hacer comparaciones entre pares o combinaciones de grupos.

EL A N V A

Es una forma de dividir la **variación total** de las observaciones en dos partes. Si el valor observado en un individuo es Y_{ij}, se considera cuánto difiere éste de la media global de todos los individuos del estudio sin importar el grupo al que pertenecen, $\left(Y_{ij} - \overline{Y}\right)$.
Esta diferencia puede dividirse en dos partes; la diferencia entre el individuo y la media del grupo de este individuo y la diferencia entre la media del grupo y la media global o gran media. En símbolos,

$$\left(Y_{ij} \text{ - } \overline{Y}\right) = \left(Y_{ij} \text{ - } \overline{Y}_i\right) + \left(\overline{Y}_i \text{ - } \overline{Y}\right)$$

El ANVA considera la variación de los individuos de los k grupos y la divide en:
1. la variación de cada individuo y la media de su grupo
2. la variación entre la media de cada grupo y la media global.

Considerando la variación de todos los individuos del experimento

$$\sum_i^k \sum_j^{n_i} \left(Y_{ij} - \overline{Y}\right)^2 = \sum\sum \left[\left(\overline{Y}_i \text{ - } \overline{Y}\right) + \left(Y_{ij} \text{ - } \overline{Y}_i\right)\right]^2$$

$$= \sum\sum \left(\overline{Y}_i \text{ - } \overline{Y}\right)^2 + \sum_i \sum_j \left(Y_{ij} \text{ - } \overline{Y}_i\right)^2$$

Suma de cuadrados total: SCT

$$\sum_i^k \sum_j^{n_i} \left(Y_{ij} - \overline{Y}\right)^2 = \sum\sum Y_{ij}^{\ 2} - \frac{Y^2}{n}$$

Suma de cuadrados *entre* grupos (o entre tratamientos): SCTrat

$$\sum_i \sum_j \left(\overline{Y}_i \text{ - } \overline{Y}\right)^2 = \sum_i n_i \left(\overline{Y}_i - \overline{Y}\right)^2 = \sum_i \frac{Y_i^{\ 2}}{n_i} - \frac{Y^2}{n}$$

Suma de cuadrados *dentro* de grupos (debida al error aleatorio): SCError

$$\sum_i \sum_j \left(Y_{ij} \text{ - } \overline{Y}_i\right)^2 = \sum (n_i - 1)\, s_i^2$$

Si las medias de los grupos son bastante diferentes entre sí, habrá variación considerable entre éstas y la gran media, comparada con la variación dentro de cada grupo. Por el contrario, si las medias de los grupos no difieren mucho, la variación entre éstas y la media global no será mucho mayor que la variación entre individuos de cada grupo. Por lo tanto, puede usarse la **prueba F** para dos variancias para probar la razón de la variancia entre medias a la variancia de cada grupo.

La hipótesis nula para la prueba F es que las dos variancias son iguales; si lo son, la variación entre medias no es mucho mayor que la variación entre observaciones individuales dentro de un grupo dado. Por consiguiente, no hay evidencia suficiente para concluir que las medias son diferentes una de otra. De esta forma el ANVA es una prueba de igualdad de medias, aun cuando en el proceso se prueban las variancias.

La hipótesis nula es, $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$. Si la hipótesis nula se rechaza, se concluye que no todas las medias son iguales, o que alguna de ellas difiere de las demás; sin embargo, no se sabe cuáles no son iguales, por esta razón se hacen necesarios procedimientos de comparación posteriores.

CUADRO DEL ANVA

Fuente de Variación	*Grados de libertad*	*Suma de Cuadrados*	*Cuadrados medios*	*Razón F*
Factor (Entre Grupos)	k - 1	$SC_G = \sum \frac{Y_i^2}{n_i} - \frac{(\sum\sum Y_{ij})^2}{n}$	$CM_G = \frac{SC_G}{k - 1}$	$F_{calc} = \frac{CM_G}{CM_E}$
Error (Dentro de grupos)	$n - k$	$SC_E = SC_T - SC_G$	$CM_E = \frac{SC_E}{n - k}$	
Total	n - 1	$SC_T = \sum\sum Y_{ij}^2 - \frac{(\sum\sum Y_{ij})^2}{n}$		

Una fórmula semejante puede usarse para encontrar la variancia de las medias de grupos respecto a la gran media:

Estimación de la variancia de medias = $\frac{\sum n_i \ (\overline{Y}_i - \overline{Y})^2}{k - 1}$

donde n_i es el número de observaciones en cada grupo y k es el número de grupos. Esta estimación se denomina **cuadrado medio entre grupos** (CM_G) y tiene k -1 grados de libertad.

Para obtener la variancia de las observaciones respecto a su media del grupo, se emplea una variancia ponderada como en la prueba *t* para grupos independientes:

Estimación de variancias dentro de grupos = $\frac{\sum (n_i - 1) S_i^2}{\sum (n_i - 1)}$

Esta estimación se denomina **cuadrado medio dentro de grupos** o **cuadrado medio del error** (CM_E) y tiene k (n_i –1) grados de libertad o si el número total de observaciones es n, se tienen $n - k$ grados de libertad.

La razón F se forma dividiendo ambas estimaciones,

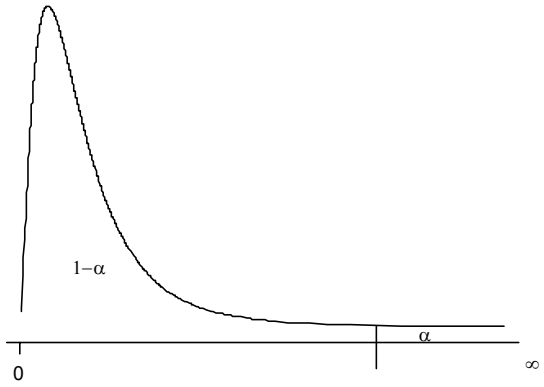
$$F = \frac{Cuadrado\ medio\ entre\ grupos}{Cuadrado\ medio\ dentro\ de\ grupos} = \frac{CM_G}{CM_E}$$

y tiene k -1 y n - k grados de libertad.

Si el valor observado de la razón *F* es mayor que el valor crítico de la tabla, se rechaza la hipótesis nula de igualdad de medias. El rechazo de la hipótesis nula no informa sobre los grupos que difieren, por lo tanto se debe estudiar la forma de determinar cuáles grupos específicos difieren.

Si $F_c \geq F_{1-\alpha}$ se rechaza H_0

Distribución F - Regiones de Decisión



EJEMPLO

Un estudio clínico realizado en Perú buscaba examinar la influencia de los suplementos de hierro y zinc en la absorción de estos minerales por los glóbulos rojos en mujeres embarazadas. Se seleccionaron 37 mujeres embarazadas (33 ± 1 semanas de embarazo) de características médicas y biológicas similares y se distribuyeron aleatoriamente en tres grupos: el grupo **A** de 10 mujeres, recibió un suplemento diario prenatal de 60 mg Fe y 250 μg folatos sin Zinc, el grupo **B** de 12 mujeres, recibió un suplemento diario prenatal de 60 mg Fe y 250 μg folatos con 15 mg de Zinc y el grupo **C**, "**Control**" en el que habían 15 mujeres, no recibió ningún suplemento férrico prenatal.
Los suplementos se administraron durante un período que se inició entre la semana 10 y la semana 24 hasta el parto. A continuación se presentan los niveles de ferritina sérica, μ**g/L,** de las pacientes,

Ferritina sérica (μg/L)			
A	**B**	**C**	
8.96	23.87	7.29	
18.98	30.23	5.45	
17.43	26.76	10.21	
13.51	34.45	10.73	
14.60	15.75	3.21	
26.12	16.17	3.44	
21.32	11.30	21.65	
15.96	12.51	14.01	
18.23	14.53	18.64	
27.39	23.00	18.28	
12.54		14.16	
5.85		14.15	
		8.87	
		6.55	
		10.13	
n_i	12	10	15	37
Total	200.89	208.57	166.77	576.23
Promedio	16.74	20.86	11.12	15.5738
Desv.Estánd	6.342	7.960	5.617	

Término de Corrección = ***TC*** = $\frac{(\Sigma\Sigma\, Y_{ij})^2}{n}$ = $\frac{576.23^2}{37} =$ **12359.10846**

- Suma de Cuadrados de Tratamientos:

$$\sum_i \sum_j \left(\overline{Y}_i - \overline{Y}\ \right)^2 = \sum_i \frac{Y_{i.}^2}{n_i} - TC = \frac{\Sigma Y_{1j}^2}{12} + \frac{\Sigma Y_{2j}^2}{10} + \frac{\Sigma Y_{3j}^2}{15} - TC$$

$$= \frac{200.89^2}{12} + \frac{208.57^2}{10} + \frac{166.77^2}{15} - 12359.10846 = \textbf{593.3}$$

- Suma de Cuadrados del Total:

$$\sum_i \sum_j \left(Y_{ij} - \overline{Y}_{..}\right)^2 = \sum_{ij} Y_{ij}^2 - \frac{Y_{..}^2}{3 \times 15} = Y_{11}^2 + Y_{12}^2 + \ldots + Y_{3,15}^2 - TC$$

= $8.96^2 + 18.98^2 + \ldots + 5.85^2 + \ldots + 6.55^2 + 10.13^2$ **–** 12359.10846 = **2047.7**

- Suma de Cuadrados del Error

S. C. Error = S. C. Total – S. C. Tratamientos

= 2047.7 - 593.3 **= 1454.5**

Cuadro del ANVA

Fuentes de Variación	G. L.	S. C.	C. M.	F_{calc}
Tratamientos	3 – 1 = 2	593.3	296.6	6.93
Error	37 - 3 = 34	1454.5	42.8	
Total	37 – 1 = 36	2047.7		

Hipótesis acerca de efectos de Tratamientos

H_o : $\mu_{1.} = \mu_{2.} = \mu_{3.} = \mu$

H_1 : al menos un $\mu_{i.} \neq \mu$, i = 1, 2, 3.

Nivel de significación $\alpha = 0.05$

Cálculo de la estadística de prueba (evidencia muestral)

$$F_{calc} = \frac{C.M.Tratamientos}{C.M.Error} = \frac{296.6}{42.8} = 6.93$$

Conclusión

Dado que $F_{cal} > F_{tab}$, existe suficiente evidencia estadística para rechazar la hipótesis nula. Podemos concluir que al menos uno de los grupos presenta un nivel medio de ferritina sérica distinto que el resto.

One-way ANOVA

```
Analysis of Variance for FERRITINA

Source     DF        SS        MS        F        P
Factor      2     593.3     296.6     6.93    0.003
Error      34    1454.5      42.8
Total      36    2047.7

Level       N      Mean     StDev
A          12    16.741     6.342
B          10    20.857     7.960
C          15    11.118     5.617
```

COMPARACIÓN DE TRES Ó MÁS MEDIAS

ANÁLISIS DE VARIANCIA

EJEMPLO. Se quiere determinar si las dietas A, B, C y D presentan diferencias en función de sus efectos sobre el incremento de peso en ratones. Se seleccionaron 20 ratones de cierta especie de la población general y luego los asignaron aleatoriamente a los tratamientos. Después de un periodo determinado, se midió el aumento de peso de cada ratón (en gramos) y se obtuvieron los datos que se muestran.

	Dieta A	Dieta B	Dieta C	Dieta D
	32	36	35	29
	37	38	30	30
	34	37	36	34
	33	30	29	31
	30	34	31	27
Medias	33.2	35.0	32.2	30.2
Desv.Est.	2.588	3.162	3.114	2.588

Las hipótesis:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ No hay diferencia en la respuesta media del **incremento de peso**, entre estas cuatro dietas

H_1 : Alguna dieta difiere de las demás

ANÁLISIS DE LA VARIANCIA

FUENTE	G.L	S.C	C.M	F_{CALC}
Dietas	3	60.15	20.05	2.42
Error	16	132.4	8.28	
Total	19	192.55		

Conclusión:

ESTADÍSTICA NO PARAMÉTRICA

Ing. Luz Bullón Camarena

Los procedimientos de inferencia presentados previamente trataron la estimación y prueba de hipótesis referidas a los parámetros de las poblaciones en estudio. Estas pruebas llamadas paramétricas utilizan los estadísticos calculados con los datos de las muestras provenientes de esas poblaciones. Sin embargo, la formulación de estas pruebas requiere suposiciones restrictivas tales como: muestras provenientes de poblaciones con distribución normal, variancias poblacionales homogéneas, conocidas, muestras grandes para garantizar normalidad por el Teorema central del límite, entre otras.

La estadística no paramétrica proporciona técnicas y pruebas alternativas las cuales no hacen suposiciones restrictivas sobre la forma de la distribución de la población. Estas pruebas alternativas son denominadas más comúnmente, pruebas de distribución libre. Las pruebas no paramétricas disponibles son muchas, de ellas abordaremos la Prueba de rangos con signo o de Wilcoxon para una muestra y para muestras pareadas, Prueba de rangos para dos muestras independientes denominada U de Mann- Withney.

Cuando se recurre a pruebas no paramétricas se hace un compromiso: perder eficiencia en la estimación de intervalos, pero adquirir la habilidad de utilizar menos información.

VENTAJAS DE LOS MÉTODOS NO PARAMÉTRICOS

1. No requieren hacer la suposición de que la población está distribuida normalmente o tiene otra forma específica.
2. En general, son más fáciles de comprender y aplicar
3. Requieren supuestos muy generales acerca de la población
4. La escala de medición puede ser de las inferiores

DESVENTAJAS DE LOS MÉTODOS NO PARAMÉTRICOS

1. Desperdician información al utilizar **signos o rangos** en lugar de los valores de las variables
2. No recomendables cuando una buena alternativa sea un método paramétrico, desde que a menudo no son tan eficientes o "exactas" como éstas.

RANGOS

Muchas pruebas no paramétricas usan los rangos en lugar de los datos. Un RANGO es un número asignado a una observación teniendo en consideración su importancia relativa (o jerarquía) respecto a los demás datos.

EJEMPLO 1:

Suponga los datos 14.5, 10.3, 11.0, 8.5 y 15.8.
Éstos pueden ordenarse de menor a mayor y tener rangos respectivamente:

Datos ordenados:	8.5	10.3	11.0	14.5	15.8
Rangos:	1	2	3	4	5

EMPATES EN LOS RANGOS.

En caso de empate o coincidencia de observaciones se asigna el promedio de los rangos

que ocupan las observaciones.

EJEMPLO 2:

Los números 9, 5, 11, 9, 12, 16 y 8 reciben los rangos de 1 a 7, pero hay un empate de 9 con los rangos 3 y 4. Se calcula media de las posiciones 3 y 4 (que es 3.5) y asignamos los rangos:

Datos ordenados:	5	8	9	9	11	12	16
Rangos:	1	2	3.5	3.5	5	6	7

De manera similar, si el empate es de las tres observaciones más pequeñas cuyos rangos son 1, 2 y 3, entonces a cada una se le asigna el rango medio (1+2+3) / 3 = 2

PRUEBA DEL SIGNO PARA UNA MUESTRA

Es una de las pruebas no paramétricas más sencillas de utilizar. Su nombre proviene del hecho en que se basa en la dirección (o signo) de los datos en lugar de su valor numérico.

La prueba se usa cuando:
- No es posible suponer normalidad de los datos
- Los datos disponibles están en escala ordinal, por lo menos

La prueba supone que la muestra se obtiene de una población simétrica en la cual la probabilidad de que un valor muestral sea menor que la mediana (media) es 1/2 e igual a la probabilidad de que sea mayor.

Para los cálculos de la prueba los datos son signos + y –, dependiendo si las observaciones están por arriba o por debajo de la mediana hipotética.

PROCEDIMIENTO

1. Las hipótesis

H_0: Me = μ	H_0: Me $\leq \mu$	H_0: Me $\geq \mu$
H_1: Me $\neq \mu$	H_1: Me > μ	H_1: Me < μ

   Si H_0 es cierta, se esperaría que el número de observaciones mayores que μ sea igual al número de observaciones menores, es decir, que la probabilidad de observar un signo + es igual a la probabilidad de observar un signo –, entonces también se puede plantear

   H_0: P(+) = P(–) = 1/2

2. Estadística de prueba: S = número de signos + ó – en la muestra

   H_1 determina la conveniencia de + ó –

   Sí H_1: P(+) < P(–), la estadística de prueba es el número de signos + y un número suficientemente pequeño de signos + causará el rechazo de H_0.

   Sí H_1: P(+) > P(–), la estadística de prueba es el número de signos – y un número

suficientemente pequeño de signos – causará el rechazo de H_0.

Sí H_1: $P(+) \neq P(-)$, un número suficientemente pequeño de + ó – causará rechazo. Se puede tomar como estadística de prueba el número de signos que ocurra con menos frecuencia

3. Distribución de la estadística de prueba

   Las observaciones constituyen ensayos de Bernoulli, luego S tiene distribución binomial con probabilidad de éxito igual a 1/2.

4. Decisión

   Cuando H_1: $P(+) < P(-)$, se rechaza H_0, sí bajo H_0 cierta, la probabilidad de observar s ó menos signos + es menor ó igual que α
   Cuando H_1: $P(+) > P(-)$, se rechaza H_0, sí bajo H_0 cierta, la probabilidad de observar s ó menos signos – es menor ó igual que α
   Cuando H_1: $P(+) \neq P(-)$, se rechaza H_0, sí bajo H_0 cierta, la probabilidad de obtener un valor de s tanto ó más extremo como el que se calculó, es menor ó igual que $\alpha/2$.

   En una prueba unilateral, el valor *p* = $P(S \leq s)$, si la prueba es bilateral, se rechazará H_0 si $P(S \leq s) < \alpha/2$

OBSERVACIONES

1. La estadística de prueba S tiene distribución Binomial (n, π =1/2)
2. Si el tamaño de muestra n, es pequeño, se usará la verdadera distribución.
   Si n es grande (n>30), la distribución de S se puede aproximar por la normal con parámetros $\mu = n\pi$ y $\Phi = n\pi(1-\pi)$, luego con la corrección de continuidad,

$$P(S \leq s) = P\left(Z \leq \frac{s + 1/2 - n\pi}{\sqrt{n\pi(1-\pi)}} \right)$$

EJEMPLO: El profesor de Estadística afirma que la nota promedio de la clase en el semestre anterior fue aprobatoria (mayor que 10.5). Para verificar su afirmación, se toma una muestra de 11 alumnos que cuyas notas obtenidas en el curso fueron las siguientes:

15.5 14.5 9.0 17.0 11.5 13.5 8.5 10.5 12.0 11.5 9.5

¿Qué puede concluir respecto a la afirmación del profesor? (Use □ □ 0.05?

PROCEDIMIENTO:

1. H_0: Me = 10.5 ó equivalentemente, H_0: $P(+) = P(-)$
   H_1: Me > 10.5

2. Signos de las diferencias respecto al valor planteado:

1	2	3	4	5	6	7	8	9	10	11
+	+	-	+	+	+	-	0	+	+	-

Si alguna diferencia resulta cero, se elimina la observación correspondiente, disminuyendo el tamaño de muestra.

3. Estadística de prueba: S = número de signos + (el que ocurre con menos frecuencia) en la muestra. Un número suficientemente pequeño de + causará rechazo
4. Distribución de la estadística de prueba. Las observaciones constituyen ensayos de Bernoulli, luego S tiene distribución binomial con probabilidad de éxito igual a 1/2 y tamaño de muestra reducida si hay ceros, n = 10

5. Decisión. La prueba es unilateral, se rechaza H_0 si el valor $p = P(S \leq s) < \alpha$

6. Cálculo de la probabilidad: $p = P(S \leq 3) = 0.0010 + 0.0098 + 0.0438 + 0.117 = 0.1717$

7. Como $p > \alpha$, no se rechaza la hipótesis nula.

MUESTRA GRANDE

Suponga una situación con una muestra grande (n = 40) donde se observaron 11 signos - y 29 signos +.

Se desea contrastar las hipótesis

H_0: P(+) = P(–) = 1/2
H_1: P(+) > P(–)

La estadística de prueba es el número de signos – y un número suficientemente pequeño de signos – causará el rechazo de H_0.

La distribución binomial de la estadística de prueba S se aproxima por la distribución normal de la forma siguiente:

$$P(S \leq 11) = P\left(Z \leq \frac{11 + 0.5 - 40(1/2)}{\sqrt{40(1/2)(1/2)}} \right)$$

$$P(Z \leq 2.69) < \alpha = 0.05$$

Luego, se rechaza H_0

PRUEBA DE WILCOXON (O DEL RANGO CON SIGNO)

Se usa cuando se desea probar una hipótesis con respecto a la media de una población, pero por alguna razón, ni *Z* ni *t* resultan adecuadas como estadística de prueba.
La prueba supone respecto a los datos:

- La muestra es aleatoria
- La variable es continua
- La población es simétrica
- La escala de medición es al menos de intervalo

La prueba utiliza las magnitudes de las diferencias entre las observaciones y el parámetro de interés ordenadas por rangos.
Las hipótesis que pueden probarse para alguna media de población no conocida:

H_0: $\mu = \mu_0$	H_0: $\mu \geq \mu_0$	H_0: $\mu \leq \mu_0$
H_1: $\mu \neq \mu_0$	H_1: $\mu < \mu_0$	H_1: $\mu > \mu_0$

PROCEDIMIENTO

1. Obtener las diferencias $d_i = x_i - \mu_0$. Si cualquier $d_i = 0$, eliminarla de los cálculos, reduciendo por lo tanto n
2. Ordenar las d_i de menor a mayor sin considerar el signo. Si dos ó más $|d_i|$ son iguales asignar a cada valor la media de la posición que ocupa en la lista. Por ejemplo, si las tres $|d_i|$ más pequeñas, sus posiciones son 1, 2 y 3, dentro del rango, luego a cada una se le asigna el rango (1+2+3)/3=2
3. A cada categoría se le asigna el signo de la diferencia correspondiente
4. Encontrar las estadísticas: T^+, la suma de las categorías con signo + y T^-, la suma de las categorías con signo -.

Si H_0 es verdadera, la probabilidad de una diferencia positiva de una magnitud dada, es igual a la probabilidad de una diferencia negativa de la misma magnitud, es decir, $P(d_i^+) = P(d_i^-)$. Luego, el valor esperado de T^+ es igual al valor esperado de T^-. A partir de una muestra no se espera una gran diferencia entre sus valores
La estadística de prueba es T^+ ó T^-, dependiendo de la hipótesis alternativa.
El valor calculado se compara con los valores críticos de la estadística de prueba de Wilcoxon que se encuentran en la tabla correspondiente. Los valores se presentan para todas las muestras de tamaño 4 hasta n = 50.

OBSERVACIONES

Si n>30, se define la estadística de prueba: $T = \dfrac{\sum R_i}{\sqrt{\sum R_i^{\,2}}}$

Sí no hay empates ésta se simplifica: $T = \dfrac{\sum R_i}{\sqrt{\dfrac{n(n+1)(2n+1)}{6}}}$

Luego se usa la aproximación de la distribución normal estándar.

EJEMPLO 1: En el estudio del **nivel de actividad física** (escala medida de 0 a 10) y su relación con el peso corporal en niños de 7 a 15 años, se ha determinado que el nivel adecuado sería de 5 o más. Para verificar esta afirmación en una población particular donde se sospecha que este nivel es menor, se toma una muestra de 11 niños de un centro educativo encontrándose los niveles de actividad física que se detallan. ¿Qué puede concluir respecto a la afirmación? (Use □□ 0.05?

4	4	7	3	2	4	8	5	4	4	7

PROCEDIMIENTO:

La hipótesis H_0: Me = 5
H_1: Me < 5

Los cálculos necesarios se muestran en la tabla siguiente:

Nivel de actividad física	Diferencia: $d_i = x_i - \mu_0$	Rango de $\lvert d_i \rvert$	Rango con signo de d_i
4	- 1	3	- 3
4	- 1	3	- 3
7	+ 2	7	+ 7
3	- 2	7	- 7
2	- 3	9.5	- 9.5
4	- 1	3	- 3
8	+ 3	9.5	+ 9.5
5	0		
4	- 1	3	- 3
4	- 1	3	- 3
7	+ 2	7	+ 7
			$T^+ = 23.5$
			$T^- = 31.5$

La segunda columna corresponde a los valores de la diferencia, de la observación menos el valor hipotético planteado

En la columna 3 se otorgan rangos a las diferencias sin tomar en cuenta el signo

Se suma los rangos con signo, por separado rangos negativos y positivos.

La estadística de prueba es el menor entre los valores T, en este caso $T^+ = 23.5$, la pregunta es ¿es suficientemente pequeño para rechazar H_0 ?

El valor crítico, (tabla de Wilcoxon) para una hipótesis unilateral es T = 11. Luego no se puede rechazar la hipótesis nula.

CONCLUSIÓN. Es posible afirmar a un nivel de significación de 5%, que el nivel de actividad física en la población estudiada, no es significativamente menor al recomendado.

EJEMPLO 2. Un estudio analizó la influencia de charlas educativas de nutrición en cambios de actitudes hacia la preparación de alimentos en familias de pocos ingresos. Se seleccionaron aleatoriamente quince familias de características similares, a las cuales se les instruyó en el valor nutritivo de distintos productos locales y en la importancia de preparar comidas balanceadas. A continuación se presenta los resultados de la evaluación de la preparación de alimentos antes y después de las charlas educativas. Los resultados se presentan en una escala de **1 = pobre** a **7 = alto valor nutritivo** de las comidas preparadas en casa. Realizar la prueba de los rangos signados de Wilcoxon. Usar α **= 0.05**.

Familia	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Antes	3	6	6	1	5	3	1	4	6	3	6	4	5	2	4
Después	4	5	4	6	5	5	7	7	6	5	3	3	7	5	6
$\lvert d_j \rvert$	1	1	2	5	0	2	6	3	0	2	-3	1	2	3	2
Rango	2	2	6	12	-	6	13	10	-	6	10	2	6	10	6
R_j con signo	-2	-2	-6	12	-	6	13	10	-	6	-10	-2	6	10	6

Hipótesis **H_o:** Charlas educativas nos tuvieron ningún impacto
H_1: Luego de las charlas mejoró la preparación de las comidas,

o equivalentemente, **H_o:** Me ≤ 0

H_1: Me >0

Estadístico de prueba y cálculo de la evidencia muestral

$\sum_{j=1}^{n} R_j$ = -2 – 2 – 6 + 12 + 6 + 13 + 10 + 6 – 10 - 2 + 6 + 10 + 6 = 47

$\sum_{j=1}^{n} R_j^2$ = $(-2)^2$ + $(–2)^2$ + $(–6)^2$ + 12^2 + 6^2 + 13^2 + 10^2 + 6^2 + $(–10)^2$ + $(-2)^2$ + 6^2 + 10^2 + 6^2 = 805

$$Z^*_{calc} = \frac{\sum_{i=1}^{13} R_i}{\sqrt{\sum_{i=1}^{13} R_i^2}} = \frac{47}{\sqrt{805}} = 1.6565$$

p-value = $P(Z \geq 1.6565) = 0.0488$

Conclusión

p-value = 0.0488 **<** 0.05 , por lo tanto rechazar la hipótesis nula y concluir que las charlas educativas sí tuvieron efecto positivo en la preparación de comidas de mayor nivel nutritivo.

PRUEBA DE MANN -WHITNEY
(PARA DOS MUESTRAS INDEPENDIENTES)

Alternativa a la prueba *t* para la diferencia de dos medias.

Las preguntas que se hacen y que la prueba va a responder son:

- ¿Hay tendencia de una población a producir valores más grandes que la otra población?
- ¿Son las medianas de las poblaciones iguales?

La prueba supone que las dos muestras, de tamaños n_1 y n_2 respectivamente, han sido extraídas independientemente y en forma aleatoria de sus poblaciones

- Si las poblaciones son diferentes, difieren sólo en lo que respecta a sus medianas
- La escala de medición es por lo menos ordinal

La prueba utiliza la información de los datos ordenados por rangos.

Las HIPÓTESIS se refieren a las medianas de las poblaciones:

H_0: $Me_x = Me_y$ H_0: $Me_x \leq Me_y$ H_0: $Me_x \geq Me_y$
H_1: $Me_x \neq Me_y$ H_1: $Me_x > Me_y$ H_1: $Me_x < Me_y$

PROCEDIMIENTO

Combinar los valores de ambas muestras aleatorias y luego asignar rangos (de menor a mayor) sin importar a que población pertenece cada valor. En caso de empate o coincidencia de observaciones se asigna el promedio de los rangos que ocupan las observaciones.

Si la mediana de la población X es, en efecto, más pequeña (o más grande) que la mediana de la población Y, es de esperar, (para muestras de igual tamaño) que la suma de los rangos asignados a las observaciones de X sea menor (o mayor) que la suma de los rangos asignados a las observaciones de la población Y

La prueba estadística se basa en

U = mín (U_1 , U_2), donde U_1 y U_2 son funciones de $\sum R_x$ $\sum R_y$ de la forma siguiente:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_x \qquad U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_y$$

Se rechaza H_0: Las medianas son iguales, las poblaciones son idénticas, sí $U < U_\alpha$

La tabla de valores U, presenta la probabilidad $P (U \geq U_\alpha) \leq \alpha$ para muestras pequeñas ($n \leq 20$)

EJEMPLO En un experimento diseñado para estimar los efectos de la inhalación prolongada de óxido de cadmio, 15 animales de laboratorio sirvieron de sujetos para el experimento, mientras que 10 animales similares sirvieron de controles. La variable de interés fue la concentración de hemoglobina (gramos) después del experimento. Los resultados se muestran a seguir.

Animales expuestos **X**	14.4	14.2	13.8	16.5	14.1	16.6	15.9	15.6	14.1	15.3
	15.7	16.7	13.7	15.3	14.0					
Animales no expuestos **Y**	17.4	16.2	17.1	17.5	15.0	16.0	16.9	15.0	16.3	16.8

Se desea saber si es posible concluir que la inhalación prolongada de óxido de cadmio disminuye el nivel de hemoglobina.

Las hipótesis: H_0: $Me_x \geq Me_y$
H_1: $Me_x < Me_y$

PROCEDIMIENTO:

Datos y rangos para el cálculo de la estadística de prueba:

X	*Rango*	Y	*Rango*
13.7	1		
13.8	2		
14.0	3		
14.1	4.5		
14.1	4.5		
14.2	6		
14.4	7		
		15.0	8.5
		15.0	8.5
15.3	10.5		
15.3	10.5		
15.6	12		
15.7	13		
15.9	14		
		16.0	15
		16.2	16
		16.3	17
16.5	18		
16.6	19		
16.7	20		
		16.8	21
		16.9	22
		17.1	23
		17.4	24
		17.5	25
	$\sum R_x = 145$		$\sum R_y = 180$

Cálculo de U = mín (U_1 , U_2), donde U_1 y U_2 son:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum R_x = 15\times 10 + \frac{15\times 16}{2} - 145 = 125$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum R_y = 15\times 10 + \frac{10\times 11}{2} - 180 = 25$$

U = mín (U_1 , U_2) = 25

El valor de la *Tabla de valores críticos de U de Mann-Whitney* es $U_\alpha = 44$

Luego, se rechaza H_0. La prueba resultó significativa. Es posible concluir que la inhalación prolongada de óxido de cadmio disminuye el nivel de hemoglobina (*p<0.05*)

OBSERVACIONES

Para muestras grandes, $n>20$, se usa la aproximación de la distribución normal:

- La prueba se puede basar en U_1 ó U_2 (pruebas equivalentes)
- Bajo H_0, las dos muestras provienen de poblaciones idénticas, se puede probar:

$\mu_U = \frac{n_1 n_2}{2}$ y $\Phi_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$, luego $Z = \frac{U_U - \mu_U}{\sigma_U}$ tiene distribución normal estándar.

DECISIÓN

Rechazar H_0: $Me_x = Me_y$, a favor de:

H_1: $Me_x \neq Me_y$ si $Z < -Z_{\alpha/2}$ ó $Z > Z_{\alpha/2}$
H_1: $Me_x < Me_y$ sí ó $Z > Z_{\alpha}$ (valores grandes de U corresponden a menores valores de R_x)
H_1: $Me_x > Me_y$ sí ó $Z < -Z_{\alpha}$

PRUEBA DE KRUSKAL -WALLIS
VARIAS MUESTRAS INDEPENDIENTES

ANÁLISIS UNILATERAL DE VARIANCIAS POR RANGOS

El ANVA, prueba de hipótesis de igualdad de las medias de varias poblaciones, supone normalidad, homogeneidad de variancias, aditividad …

- La prueba de Kruskal-Wallis es una alternativa no parámetrica al ANVA

- La prueba es una ampliación de la prueba de Mann-Whitney para más de dos muestras independencia

- Detecta diferencias entre los *k* grupos (tratamientos), respecto a ubicación, dipersión, forma.

La prueba supone:

- Las muestras han sido extraídas independientemente y en forma aleatoria de sus poblaciones
- Poblaciones con igual distribución o alguna tiende a producir valores más grandes que las otras poblaciones
- La escala de medición es por lo menos ordinal

Los datos muestrales:

muestra 1	muestra 2	. . .	muestra k
x_{11}	x_{21}		x_{k1}
x_{12}	x_{22}		x_{k2}
.	.		
.	.		
.	.		
x_{1n1}	.		
	x_{2n2}		x_{knk}

x_{ij} : observación, *j*-ésima perteneciente a la muestra *i*
k : número de grupos poblaciones, tratamientos)
n_i : número de observaciones del *i*-ésimo grupo
$n = \sum n_i$: número de observaciones en todos los grupos combinados

La HIPÓTESIS se refiere a un parámetro de localización (media, mediana) o a la forma de la distribución de las poblaciones

H_0: $\mu_1 = \mu_2 = \ . \ . \ . \ = \mu_k$

PROCEDIMIENTO

1. Combinar las $n_1, n_2, \ldots n_k$ observaciones de las *k* muestras aleatorias en una sola serie de tamaño *n* y luego asignar rangos (de menor a mayor) asignando el promedio de los rangos en caso de empate o coincidencia de observaciones.

2. Los rangos asignados a las observaciones en cada uno de los *k* grupos se suman por separado para dar *k* sumas de rangos.

3. La estadística de prueba se calcula como:

$$T = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

donde R_i = suma de los rangos en el *i*-ésimo grupo

4. Cuando hay 5 ó menos observaciones por grupo, el nivel de significación puede calcularse usando la distribución exacta de T
Cuando hay más de 5 observaciones en cada grupo, la estadística se compara con los valores tabulados de la distribución χ^2 con *k*-1 grados de libertad.

EJEMPLO: Se quiere comparar tres métodos de medición del nivel de contaminación de una planta industrial. Las medidas obtenidas por cada método se presentan a continuación:

Métodos de Medición		
A	*B*	*C*
94	85	89
87	82	67
91	79	72
74	84	76
86	61	69
97	72	
	80	

PROCEDIMIENTO:

Valores y rangos de los tres métodos					
A	Rango	B	Rango	C	Rango
94	17	85	12	89	15
87	14	82	10	67	2
91	16	79	8	72	4.5
74	6	84	11	76	7
86	13	61	1	69	3
97	18	72	4.5		
		80	9		
R_A = 84 n_A = 6		R_B = 55.5 n_B = 7		R_C = 31.5 n_C = 5	

LA HIPÓTESIS A PROBAR H_0: $\mu_1 = \mu_2 = \mu_3$

CÁLCULOS:

$$T = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) = \frac{12}{18\times 19}\left(\frac{84^2}{6} + \frac{55.5^2}{7} + \frac{31.5^2}{5}\right) - 3(19) = 6.67$$

El valor de la tabla de Distribución χ^2 con 2 grados de libertad es $\chi^2_{,95} = 5.991$. Luego, se rechaza H_0.

Se procede a comparaciones: H_0: $\mu_1 = \mu_2$, H_0: $\mu_1 = \mu_3$ y H_0: $\mu_2 = \mu_3$

ASOCIACIÓN DE VARIABLES

Lic. Esperanza García C.

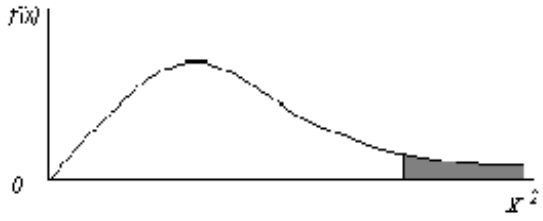
En muchos aspectos de nuestra vida diaria observamos situaciones que guardan relación, así por ejemplo, vemos que el nivel de nutrición de los niños está asociado con el nivel de aprendizaje, el peso con la talla de las personas, los contaminantes ambientales y la presencia de procesos respiratorios, el nivel de escolaridad de la madre y el número de hijos, etc.; vemos también, que muchos de estas observaciones son motivo del desarrollo de investigaciones, razón por el cual se hace necesario la demostración de esta asociación. En este contexto, es importante el tipo de variable (cualitativa o cuantitativa) que queremos relacionar:

- cuando las variables son cualitativas, la asociación podrá establecerse usando Ji-cuadrado.
- cuando las variables son cuantitativas, la naturaleza e intensidad de la relación se hará por medio del análisis de regresión y correlación.

JI CUADRADO (X^2)Y SUS APLICACIONES

La distribución de probabilidades χ^2, es sesgada a la derecha, sus valores empiezan en cero y por la derecha aumenta infinitamente**.** Supongamos una variable aleatoria Y, que tiene una distribución normal, con media μ y varianza σ^2, si se eligen muestras aleatorias e independientes de tamaño n=1, cada valor seleccionados puede transformarse en la variable normal estándar ($Z = \frac{X - \mu}{\sigma}$) Cada valor "z" puede elevarse al cuadrado, al estudiar la distribución muestral de z^2 se observa que sigue una distribución χ^2 con 1 grado de libertad.

CARACTERÍSTICAS:



1. La distribución de probabilidades χ^2 se lee con grados de libertad. Para cada grado de libertad hay una curva de probabilidades.
2. No tiene valores negativos. El área bajo la curva se inicia en cero y a la derecha se distribuye infinitamente.
3. Todas las curvas son asimétricas.
4. A medida que aumentan los grados de libertad las curvas son menos elevadas y más extendidas a la derecha.

APLICACIONES:

Karl Pearson, demostró que la distribución χ^2 puede emplearse como prueba de la congruencia entre la observación y la hipótesis, La estadística χ^2 es más adecuada para utilizarse con variables medidas en escala nominal u ordinal, Los datos utilizados para el cálculo de la estadística de prueba, son frecuencias asociadas con cada una de las categorías dos variables, de las cuales se desea saber si existe o no asociación. Estos se presentan en tablas de contingencia 2 x 2 ó r x c.
Las pruebas más usadas para probar hipótesis son: la prueba de bondad de ajuste, la prueba de independencia y la prueba de homogeneidad. Se usa la siguiente fórmula:

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Donde:

- O_i representa las frecuencias observadas en cada una de las celdas de la tabla
- E_i representa las frecuencias esperadas, dado que H_o es verdadera.

CONCEPTOS BÁSICOS:

1.- FRECUENCIA OBSERVADA: Es el número de objetos o individuos en la muestra que caen dentro de las categorías de la variable de interés. Por ejemplo, si en una muestra de 100 pacientes hospitalizados se puede observar que: 50 son casados, 30 solteros, 15 viudos y 5 son divorciados.

2.- FRECUENCIA ESPERADA*:* Es el número de individuos u objetos en la muestra que se esperaría observar si alguna hipótesis nula respecto a la variable es verdadera. Tomando el ejemplo, en la variable estado civil, la hipótesis nula pueden ser que las cuatro categorías de la variable tienen igual representación en la población de la que se extrajo. En este caso se puede esperar que la muestra contenga 25 casados, 25 solteros, 25 viudos y 25 divorciados. Su cálculo está basado en las probabilidades, específicamente en la intersección. Se usa la siguiente fórmula:

$$F.\,Esperada = \frac{Total\ de\ fila \times Total\ de\ Columna}{Total\ general}$$

La cantidad de χ^2 con (f - 1)(c - 1) grados de libertad, es una medida del grado con el que los pares de frecuencias observadas y esperadas concuerdan en una situación dada. Si la congruencia es estrecha entre O*i* - E*i*, el valor χ^2 será cero o próximo a cero para cada par de frecuencias en cada categoría, por lo que no es posible rechazar la hipótesis de nulidad. Por otro lado cuando la congruencia es pobre, dicho valor es muy grande. En consecuencia se necesita un χ^2 suficientemente grande para rechazar la hipótesis de nulidad.

3.- CÁLCULO DE LOS GRADOS DE LIBERTAD: Se obtiene de multiplicar el número de categoría de la primera variable menos uno por el número de las categorías de la segunda variable menos uno.

Grados de libertad = (f – 1)(c – 1)

PRUEBA DE INDEPENDENCIA

Se usa cuando el interés del investigador es probar que dos criterios de clasificación son independientes; es decir si la distribución de un criterio es la misma, sin importar cuál es la distribución del otro. Entonces, H_o expresará independencia (no relación o no asociación entre las variables). Por ejemplo, si el estado socioeconómico y área de residencia de los habitantes de cierta ciudad son independientes, se esperaría encontrar la misma proporción de familias en los grupos socioeconómicos alto, medio y bajo en todas las áreas de la ciudad. Se tiene interés en probar la hipótesis nula según la cual en la población, los dos criterios de clasificación son independientes. Si se rechaza la hipótesis de nulidad, se concluye que los dos criterios de clasificación no son independientes, y por lo tanto las variables están asociadas. Como intervienen dos variables, las frecuencias observadas se presentan en una tabla de contingencia 2 x 2, ó, r x f.

Los datos para esta prueba están medidos en escala nominal u ordinal. La característica principal es que *n* se extrae en forma aleatoria de una sola población, en consecuencia, las frecuencias que caen en las diferentes celdas suceden en forma aleatoria, por ende los totales marginales de las filas y columnas son también aleatorios.

RECOMENDACIONES DE COCHRAN SOBRE EL USO DE χ^2

A. Si las frecuencias están en tablas 2 x 2, el uso de χ^2 debe guiarse de las siguientes consideraciones:

- Si *n>40,* se usa χ^2 corregida por continuidad, con la fórmula:

$$\chi^2_{\text{corregido}} = \frac{n\left(|ad - bc| - 0.5n\right)^2}{(a+c)(b+d)(a+b)(c+d)}$$; con 1 grado de libertad

El uso de la corrección de Yates disminuye el riesgo de cometer error tipo I, se aconseja sobre todo cuando el χ^2 calculado está próximo al valor crítico. Algunos autores expresan su disconformidad al aplicar el ajuste cuando la muestra sobrepasa de 50.

- *n* está entre 20 y 40 se usará χ^2 en el caso que todas las frecuencias esperadas sean de 5 o más.

- Si *n* < 20, no se usa χ^2. En este caso se usará la prueba exacta de Fisher.

B. Si las frecuencias están en tablas con grados de libertad mayor que 1, se puede usar χ^2 si menos del 20% de las celdas tienen frecuencias esperadas menores que 5 y si no hay ninguna celda con una frecuencia esperada menor que 1 (si estos requisitos no se dan el investigador puede combinar categorías para aumentar las frecuencias en las diferentes celdas. El procedimiento de la prueba se ilustra con el siguiente ejemplo:

Ejemplo:

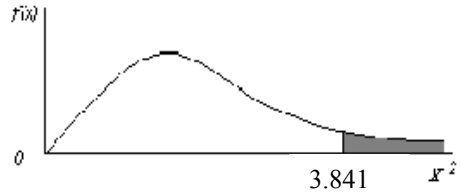
Una muestra de estudiantes universitarios participó en un estudio para evaluar el nivel de conocimientos respecto a determinado grupo de enfermedades comunes. La tabla siguiente presenta la clasificación de los estudiantes de acuerdo a su principal campo de estudio y el nivel de conocimientos sobre el grupo de enfermedades.

Conocimientos de enfermedades					
Campo de estudio	Buena		Deficiente		Total
	O	E	O	E	
Premédico	31	**12.20**	91	**109.80**	122
Otro	19	**37.80**	359	**340.20**	378
Total	50		450		500

¿Sugieren estos datos que existe relación entre el conocimiento del grupo de enfermedades y el principal campo de estudio de los estudiantes de nivel superior de los cuales se extrajo esta muestra. Sea α = 0.05.

Solución

1. Hipótesis
   - Ho: El conocimiento del grupo de enfermedades y el principal campo de estudio de los estudiantes del nivel superior no están asociados (son independientes)
   - H_1: El conocimiento del grupo de enfermedades y el principal campo de estudios de los estudiantes del nivel superior están asociadas (son dependientes).
2. Nivel de significación α= 0.05
3. Selección de la prueba:
   - Las variables son cualitativas
   - La muestra es aleatoria, obtenida de una población.
   - Estadística de prueba : $X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$
4. Criterios de decisión: Ho se rechazará si, $\chi^2_{(1\text{grado de libertad})}$ calculado es mayor o igual a 3.841.



5. Cálculo:

Conocimientos de enfermedades					
Campo de estudio	Buena		Deficiente		Total
	O	E	O	E	
Premédico	31	**12.20**	91	**109.80**	122
Otro	19	**37.80**	359	**340.20**	378
Total		50		450	500

$\chi^2 = (31\text{-}12.20)^2/12.20 + (91\text{-}109.80)^2/109.80 + \ldots + (359\text{-}340.20)^2/340.20 =$ **42.579**

6. Decisión y conclusión: χ^2 calculado mayor que 3.841, se rechaza la hipótesis nula y se concluye que a un nivel de significación del 0.05, es posible afirmar que el conocimiento del grupo de enfermedades y el principal campo de estudio de los estudiantes están asociados.

PRUEBA DE HOMOGENEIDAD

Se usa para probar hipótesis de nulidad que indica que **dos o más muestras** provienen de poblaciones homogéneas pero distintas con respecto a algún criterio de clasificación. Con frecuencia es usada en estudios donde hay intervención, es decir cuando se hacen estudios de tipo experimental. La prueba sirve para comparar dos o más muestras respecto a un determinado criterio y que han sido extraídas de poblaciones previamente seleccionadas. Los datos se presentan en tablas de contingencia en las que un conjunto de de totales marginales es *fijo* por que se conoce previamente, están bajo el control del investigador, mientras que el criterio de clasificación aplicado a las muestras, es aleatorio. El estadístico de prueba es el mismo que el usado en la prueba de independencia. La hipótesis nula y conclusiones se establecen en términos de homogeneidad (igualdad) de las poblaciones con respecto a la variable de interés, por ello a este procedimiento también se la conoce como prueba de las similitudes.

Ejemplo I:

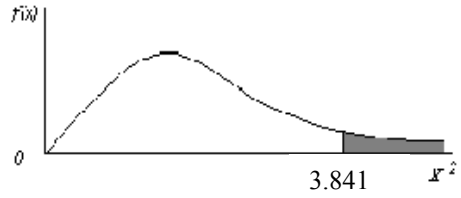
En la siguiente tabla se presentan los resultados de un estudio que analiza la efectividad de los cascos de seguridad para ciclistas, para prevenir lesiones en la cabeza en caso de accidentes. De los 147 de individuos que usaban casco al momento del accidente, 17 sufrieron lesiones en la cabeza que requirieron atención médica, mientras que 130 no lo necesitaron; entre los individuos que no empleaban cascos de seguridad 218 sufrieron lesiones en la cabeza serias y 428 no. Se desea saber si las poblaciones de usuarios del casco y los que no lo usan son similares respecto a las lesiones de cabeza sufridas en los accidentes.

	Uso del casco				
Lesión en la	Sí		No		
cabeza	O	E	O	E	Total
Sí	17	**43.56**	218	**191.44**	**235**
No	130	**103.44**	428	**454.56**	**558**
Total	147		646		793

Solución:

1. Hipótesis
   - H_o: Las dos poblaciones son homogéneas respecto a la lesión de cabeza sufrida en el accidente.
   - H_1: Las dos poblaciones no son homogéneas respecto a la lesión de cabeza sufrida en el accidente.
2. Nivel de significación α= 0.05
3. Selección de la prueba:
   - Las variables son cualitativas
   - Se tiene dos muestras de las poblaciones de usuarios y no usuarios de casco al momento del accidente de ciclismo y se las clasifica de acuerdo a la lesión de cabeza sufrida.
   - Estadística de prueba : $X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$

4. Criterios de decisión: Ho se rechazará si, χ^2 (1grado de libertad) calculado es mayor o igual a 3.841.



5. Cálculo:

	Uso del casco				
	Sí		No		Total
Lesión en la cabeza	O	E	O	E	
Sí	17	**43.56**	218	**191.44**	**235**
No	130	**103.44**	428	**454.56**	**558**
Total	147		646		793

χ^2 = (17- 43.56)^2/43.56 +(218 - 191.44)^2/191.44 + … + (428 - 454.56)^2/454.56 = **28.255**

6. Decisión y conclusión: χ^2 calculado mayor que 3.841, se rechaza Ho y se concluye que a un nivel de confianza de 0.05 es posible concluir que las poblaciones de usuarios y no usuarios del casco al momento del accidente fueron diferentes respecto a la lesión de cabeza al momento del accidente.

Ejemplo II:

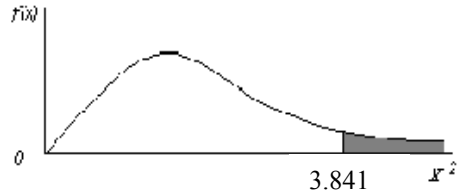
Un grupo de investigadores realizaron un estudio para comparar la curabilidad de dos tipos de enfermedad neoplásica respecto al tratamiento con quimioterapia. Los resultados en cuanto a la curabilidad fueron:

	Curabilidad			
Enfermedad	Sí	No	Total	
A	37	6	**43**	**n1**
B	27	18	**45**	**n2**
Total	64	24	88	

Solución:

1. Hipótesis
   - H_o: Las poblaciones son homogéneas respecto a la curabilidad.
   - H_1: Las poblaciones no son homogéneas respecto a la curabilidad.
2. Nivel de significación α= 0.05
3. Selección de la prueba:
   - Las variables son cualitativas, dicotómicas medidas en escala nominal
   - Se tiene dos muestras poblacionales
   - Estadística de prueba: $\chi^2 = \dfrac{n\left(\left[ad - bc\right] - n/2\right)^2}{(a+c)(b+d)(a+b)(c+d)}$

4. Criterios de decisión: Ho se rechazará si, χ^2(1grado de libertad) calculado es mayor o igual a 3.841.



5. Cálculo:

$$\chi^2 = \frac{\left(|37 \times 18 - 27 \times 6| - 44\right)^2}{64 \times 24 \times 43 \times 45} = 6.265$$

Remplazando términos se tiene que $\chi^2_{corregido} = 6.265$

6. Decisión y conclusión: χ^2 calculado mayor que 3.841, se rechaza Ho y se concluye que a un nivel de confianza de 0.05 es posible concluir que las poblaciones no son homogéneas respecto a la curabilidad.

EJERCICIO DE REPASO

Lea y responda en forma precisa o complete cuando sea necesario.

1. La prueba de independencia establece como hipótesis nula la no asociación de variables. Se usa cuando se analiza una muestra…..…….. procedente de una sola población.

2. Frecuencias observada es el número de elementos del estudio que cae en una celda de una tabla de contingencia y frecuencia esperada es ……
Es igual al producto de sus probabilidades…….....

3. La prueba de homogeneidad se caracteriza por que uno de los totales marginales es fijo, manipulado por el investigador, el otro sucede ....., Se usa para comparar.....o más....., respecto a un determinado criterio de clasificación.

4. En un estudio realizado en alumnos universitarios se les clasificó según su especialidad y su preferencia por un partido político. Se encuestaron a 310 estudiantes, 111 de Letras, 67 de Ingeniería, 68 de Agronomía y 74 de Educación y se obtuvieron los siguientes resultados:

Especialidad	Partido Político			
	UN	PP	APRA	Total
Letras	34	61	16	111
Ingeniería	31	19	17	67
Agronomía	19	23	16	68
Educación	23	39	12	74
Total	107	142	61	310

¿Cuántas muestras seleccionaron en el estudio? .....
¿Qué prueba se debe usar para probar la hipótesis? .....
Pruebe la hipótesis correspondiente según los pasos desarrollados en los ejemplos anteriores e interprete **(x^2 calculado = 16.161)**

6. Una muestra de 150 portadores crónicos de cierto antígeno y una muestra de 500 no portadores, revelaron la siguiente distribución de grupos sanguíneos. ¿Qué puede afirmar acerca de la distribución de grupos sanguíneos en los dos grupos de portadores y no portadores?. Use α = 0.05.

Antígeno	Grupo sanguíneo				Total
	O	A	B	AB	
Portadores	72	54	16	8	150
No portadores	230	192	63	15	500
Total	302	246	79	23	650

ANÁLISIS DE REGRESIÓN Y CORRELACIÓN LINEAL

Ing. Edith Alarcón Matutti

El objetivo de éste capítulo es analizar el grado de la relación existente entre variables cuantitativas, utilizando modelos matemáticos y representaciones gráficas. Así pues, para representar la relación entre dos o más variables desarrollaremos una ecuación que permitirá estimar una variable identificada como dependiente, en función de otra definida como Independiente.

Por ejemplo:

¿Será posible que un incremento en la calificación final del curso de estadística esta asociado con las horas destinadas para el estudio y la práctica de ejercicios?

¿Cree Ud. que la edad de la madre gestante, influye en el peso del recién nacido? ¿de manera positiva o negativa?

¿Podemos afirmar que el peso de un niño depende de la edad cronológica que dicho niño tenga al momento de la medición?

Para responder a las situaciones antes mencionadas, estudiaremos el grado de relación entre dos variables en lo que llamaremos ***análisis de correlación***. Para representar esta relación utilizaremos una representación gráfica llamada ***diagrama de dispersión*** y, finalmente, estudiaremos un modelo matemático para estimar el valor de una variable basándonos en el valor de otra, en lo que llamaremos ***análisis de regresión.***

ANÁLISIS DE CORRELACIÓN

Dadas dos variables aleatorias cuantitativas, nos interesa cuantificar la intensidad de la relación lineal entre las mismas. El parámetro estadístico que nos da tal cuantificación es el ***coeficiente de correlación lineal de Pearson***, denotado por el símbolo "**r**", este coeficiente en la población se denota por ***"ρ"*;** los valores que puede tomar éste parámetro están comprendidos dentro del siguiente intervalo del campo de los reales:

$$-1 \leq r \leq +1$$

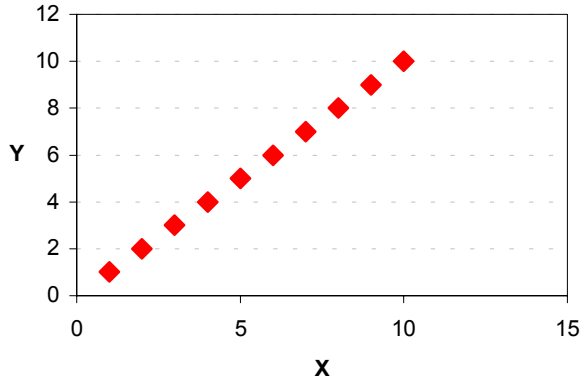
El cálculo del coeficiente de correlación lineal se realiza con la siguiente fórmula:

$$r = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}$$

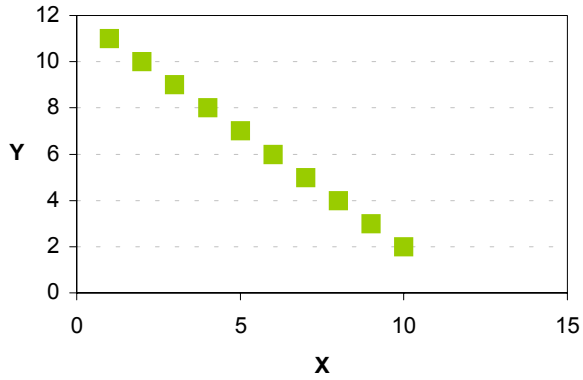
Donde los componentes, n es el tamaño de muestra conformado por los pares de datos correspondientes a las variables x e Y, las sumatorias simples de los datos, las sumas de los cuadrados de los datos y la suma del productos cruzados de las variables.

Gráficamente podemos visualizar las siguientes situaciones:

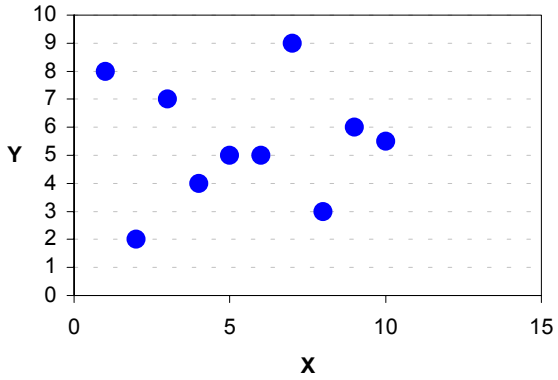
Variables con correlación positiva r > 0

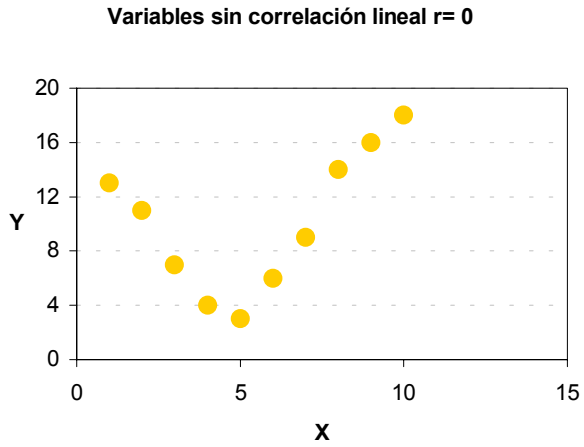


Variables con correlación negativa r< 0



Variables no correlacionadas r = 0





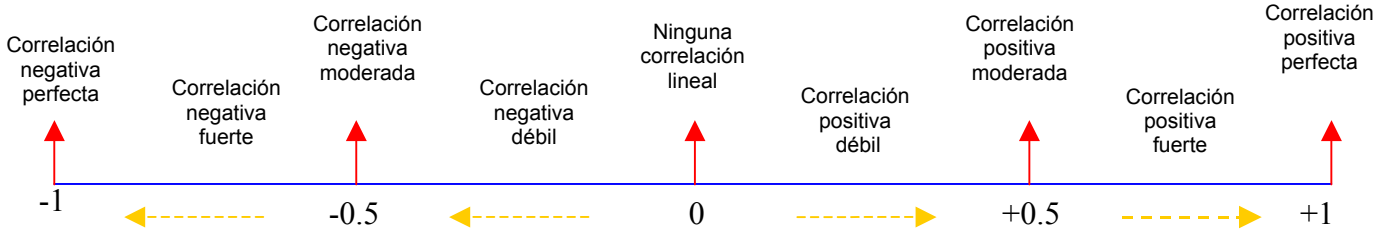
Como se observa en los diagramas anteriores, el valor de **r** se aproxima a +1 cuando la correlación tiende a ser ***lineal directa*** (mayores valores de X significan mayores valores de Y), y se aproxima a –1 cuando la correlación tiende a ser ***lineal inversa***.

Es importante notar que la existencia de correlación entre variables no implica causalidad.

¡Atención! si no hay correlación de ningún tipo entre dos variables alestorias, entonces tampoco habrá correlación lineal, por lo que r = 0.

Sin embargo, el que ocurra r = 0 sólo nos dice que no hay correlación lineal, pero puede que la haya de otro tipo.

El siguiente diagrama resume el análisis del coeficiente de correlación entre dos variables:

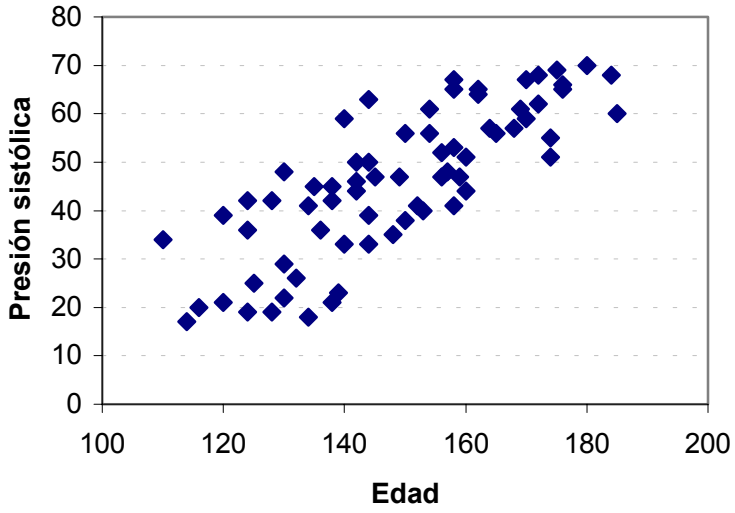


Ejemplo:

En la tabla siguiente se muestran los datos de 69 pacientes de los que se conoce su edad y una medición de su presión sistólica. Si estamos interesados en estudiar la variación en la presión sistólica en función de la edad del individuo, primero debemos verificar gráficamente con el diagrama de dispersión y luego calcular el coeficiente de correlación.

Nº	Tensión Sistólica	Edad	Nº	Tensión Sistólica	Edad	Nº	Tensión Sistólica	Edad	Nº	Tensión Sistólica	Edad
1	114	17	18	136	36	35	149	47	52	140	59
2	134	18	19	150	38	36	156	47	53	170	59
3	124	19	20	120	39	37	159	47	54	185	60
4	128	19	21	144	39	38	130	48	55	154	61
5	116	20	22	153	40	39	157	48	56	169	61
6	120	21	23	134	41	40	142	50	57	172	62
7	138	21	24	152	41	41	144	50	58	144	63
8	130	22	25	158	41	42	160	51	59	162	64
9	139	23	26	124	42	43	174	51	60	158	65
10	125	25	27	128	42	44	156	52	61	162	65
11	132	26	28	138	42	45	158	53	62	176	65
12	130	29	29	142	44	46	174	55	63	176	66
13	140	33	30	160	44	47	150	56	64	158	67
14	144	33	31	135	45	48	154	56	65	170	67
15	110	34	32	138	45	49	165	56	66	172	68
16	148	35	33	142	46	50	164	57	67	184	68
17	124	36	34	145	47	51	168	57	68	175	69
									69	180	70

Diagrama de dispersión



Observamos que existe una correlación positiva, el valor de r nos cuantificará la fuerza de dicha correlación.

Calculando r con la fórmula:

$$\sum X_i Y_i = (17 \times 114) + 18 \times 134) + ... + (70 \times 180) = 488606$$

$$\sum X_I^2 = 17^2 + 18^2 + ... + 70^2 = 162303$$

$$\sum Y^2 = 114^2 + 134^2 + \ldots + 180^2 = 1549424$$

$$\sum X_i = 17 + 18 + ... + 70 = 3183\,; \qquad \sum Y_i = 114 + 134 + \ldots + 180 = 10262$$

$$r = \frac{69 \times 488606 - 3183 \times 10262}{\sqrt{69 \times 162303 - 3183^2}\ \sqrt{69 \times 1549424 - 10262^2}} = 0.803$$

El coeficiente de correlación es 0.803, el grado de correlación es alto.

PRUEBA DE HIPÓTESIS ACERCA DE ρ

Por lo general, el interés radica en saber si es posible concluir que **X e Y** están correlacionadas. Luego, con los datos de la muestra se calcula *r*, el valor estimado de ρ y se prueba

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

La estadística de prueba adecuada es $t = r\sqrt{\frac{n-2}{1-r^2}}$

Cuando H_0 es verdadera y se cumplen las suposiciones, la estadística de prueba sigue una distribución t de Student con n-2 grados de libertad.

ANÁLISIS DE REGRESIÓN LINEAL SIMPLE

En los casos en que el coeficiente de correlación lineal sea "cercano" a **+1** ó **–1**, tiene sentido considerar la ecuación de la recta que "mejor se ajuste" a la nube de puntos (recta de mínimos cuadrados). Uno de los principales usos de dicha recta será el de predecir o estimar los valores de Y que obtendríamos para distintos valores de X.

La ecuación de la recta de mínimos cuadrados (en forma punto-pendiente) es la siguiente:

$$Y = \alpha + \beta\chi + \varepsilon$$

Donde:
α es el valor de la ordenada donde la línea se intercepta con el eje Y.
β es el coeficiente de regresión poblacional (pendiente de la recta).
ε es el error

SUPUESTOS QUE DEBEN CUMPLIR LOS DATOS:

1. Los valores de la variable independiente X son fijos es decir son manipulados por el investigador y por lo tanto son medidos sin error.

2. La variable Y es aleatoria

3. Para cada valor de X, existe una distribución normal de valores de Y (subpoblaciones de Y).

4. Las variancias de todas las subpoblaciones de Y son todas iguales.

5. Todas las medias de las subpoblaciones de Y están sobre la recta

6. Los valores de Y siguen una distribución normal y son estadísticamente independientes.

ESTIMACIÓN DE LA RECTA DE REGRESIÓN LINEAL SIMPLE

Para estimar la ecuación de la recta que mejor describe la relación entre dos variables, se usa el método de mínimos cuadrados y la recta resultante se conoce como la recta de Mínimos Cuadrados.

Luego, la ecuación de regresión estimada es:

$$Y' = a + bX$$

a es el estimador de α . Es el valor para un X = 0
Y´ es el valor estimado de la variable *Y*
b es el estimador de *β*. Es el coeficiente de regresión

b indica el número de unidades que varía Y cuando se produce un cambio en una unidad, en X (pendiente de la recta de regresión). Un valor negativo de b, se interpreta como la magnitud del decremento en Y por cada unidad de aumento en *X*.

Para calcular a y b utilizamos las siguientes fórmulas:

$$a = \overline{Y} - b\overline{X}$$

$$b = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \frac{\sum X_i Y_i - n\overline{X}\,\overline{Y}}{\sum X_i^2 - n\,\overline{X}^2}$$

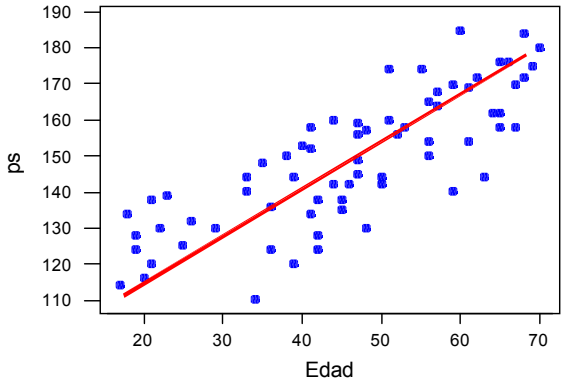
Para el ejemplo desarrollado anteriormente, estimaremos la ecuación de la recta de regresión que relaciona la presión sistólica en función de la edad:

$$b = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \frac{\sum X_i Y_i - n\overline{X}\,\overline{Y}}{\sum X_i^2 - n\,\overline{X}^2} = \frac{488606 - 69(46.13)(148.7299}{162303 - 69(46.13)^2} = 0.984$$

$$a = \overline{Y} - b\overline{X} = 148.72 - 0.984(46.13) = 103.353$$

Entonces la ecuación queda determinada por : Y= 103.353+0.984X o en términos de nuestras variables **Presión Sistólica = 103 + 0.984 Edad**

Gráficamente :



EVALUACIÓN DE LA ECUACIÓN DE REGRESIÓN

Para tener la suficiente garantía de que las estimaciones que se realicen son válidas se sugiere validar el modelo con pruebas de hipótesis referentes a la constante y a la pendiente de la ecuación hallada.

COEFICIENTE DE DETERMINACION:

Medida que permite evaluar el grado de dispersión de los puntos en torno a la recta de regresión con la dispersión en torno $\overline{Y}$ (promedio de los valores de Y). Nos cuantifica el efecto de la variable independiente sobre la respuesta, su valor está entre 0 y 1. En el ejemplo la evidencia gráfica es suficiente pero es el coeficiente de determinación una medida objetiva de la fuerza de la relación XY.

El cálculo lo haremos con la siguiente fórmula:

$$r^2 = \frac{\sum(Y'-\overline{Y})}{\sum(Y_i-\overline{Y})} = \frac{b^2\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right]}{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}$$

Para nuestro ejemplo

$$r^2 = \frac{103.353^2\left[162303 - \frac{3183^2}{69}\right]}{1549424 - \frac{10262^2}{69}} = 0.645 \approx 64.5\,\%$$

Se interpreta como que el 64.5% de la variación en la presión sistólica (Y) es explicada por la regresión de la presión sistólica en función de la edad(X).
Por lo tanto, para estimar la presión sistólica de un paciente que tiene 49 años, reemplazamos el valor de X por 49 y efectuamos las operaciones y obtenemos:

Presión Sistólica = 103 + 0.984 Edad = 103 + 0.984 (49) = 147