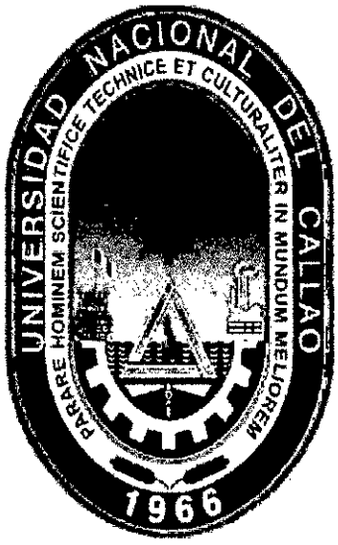
IF

NOV 2018

UNIVERSIDAD NACIONAL DEL CALLAO

FACULTAD DE CIENCIAS DE LA SALUD

UNIDAD DE INVESTIGACIÓN DE LA CIENCIAS DE LA SALUD



RECIBIDO
UNIVERSIDAD NACIONAL DEL CALLAO
VICE RECTORADO DE INVESTIGACIÓN
651
12 SEP 2018
HORA: 15:50
FIRMA:

RECIBIDO
UNIVERSIDAD NACIONAL DEL CALLAO
VICE RECTORADO DE INVESTIGACIÓN
704
02 OCT 2018
HORA: 15:50
FIRMA:

INFORME FINAL DEL TEXTO

"TEXTO: Estadística Aplicada a la Investigación en Ciencias Sociales e Ingeniería Basado en Competencias"

AUTOR: **Dra. ANA LUCY SICCHA MACASSI**

PERIODO DE EJECUCION:

1 de Setiembre del 2016 al 31 de Agosto del 2018

Resolución de aprobación N° 781-2016, Rectificada N° 978-2016-R

Callao, 2018

PERÚ

RECIBIDO
UNIVERSIDAD NACIONAL DEL CALLAO
VICE RECTORADO DE INVESTIGACIÓN
744
30 OCT 2018
HORA: 12:40
FIRMA:

I. INDICE

ÍNDICE DE TABLAS

ÍNDICE DE GRÁFICOS

II. PROLOGO

Existen textos y material de consulta sobre Estadística para la investigación, así como bioestadística aplicada en ciencias de la salud, estadística para ciencia epidemiológica, y similares como psicometría, Econometría y Quimiometria entre otros como se detalla en los referenciales y que en muchos casos están orientados al manejo de los datos.

La metodología que se utilizará para la elaboración del "Texto: **Estadística Aplicada a la Investigación en Ciencias Sociales e Ingeniería Basado en Competencias"** se sustentarán en la revisión bibliográfica como en los ejemplos aplicados como resultado dela actividad en la docencia del dictado de los cursos de: Bioestadística, Metodología de la Investigación y Tesis de los programas de Posgrado y Pregrado.

Los docentes e investigadores orientan sus resultados y culminación de sus investigaciones con la aplicación de la Estadística a la obtención del p-valor significativo ($p<0.05$), sin orientar a una adecuada aplicación de la técnica estadística y adicionalmente al cumplimiento de los supuestos condicionantes de la prueba. Así también con el uso indiscriminado de programas estadísticos, los investigadores orientan los objetivos de su estudio a la realización de asociación, relación, comparación, y otros ya sea de forma univariada, o bivariada. El presente texto es un aporte para vencer estos obstáculos en la culminación del trabajo de tesis

III. INTRODUCCIÓN

El presente libro es de aplicación técnica para el desarrollo de tesis de pregrado y posgrado como resultado de la enseñanza en los cursos dictados de seminario de tesis, estadística, bioestadística en las distintas facultades de la Universidad Nacional del Callao.

El estudiante en el último ciclo de estudios con el objetivo de obtener su licenciatura desarrolla su tesis el cual que no logra alcanzar con la culminación de su desarrollo, encontrando un desconocimiento en el uso de herramientas para la presentación, discusión y análisis de los resultados.

Los ejemplos utilizados en el desarrollo de este texto están basados en el enfoque de las diferencias disciplinas existentes en la Universidad Nacional del Callao, Ciencia de Sociales (Enfermería, Economía, Contabilidad, Administración) e Ingeniería (Química, Ambiental, Industrial).

Este texto será de mayor utilidad para aquellos que tengan una base fundamental de matemática y orientará para que el lector tenga confianza para adquirir destreza en su aplicación y análisis en su trabajo de tesis.

El presente texto: "**Estadística Aplicada a la Investigación en Ciencias Sociales e Ingeniería Basado en Competencias**" tiene por objetivo; proponer un texto universitario que facilite los conocimientos teóricos y prácticos de Estadística para el desarrollo, presentación y análisis de datos.

IV. CUERPO DEL TEXTO O CONTENIDO

CAPÍTULO 1

REVISIÓN DE CONCEPTOS

La estadística es una ciencia necesaria y útil en toda carrera profesional, ya que las técnicas y procedimientos estadísticos son aplicables a características de diferente naturaleza, como, por ejemplo: la ocurrencia de fallas en un dispositivo, las ventas diarias de una empresa, entre otras. Los datos estadísticos se caracterizan por ser aleatorios, ya que el dato es inesperado y casual; inciertos, es decir, no se tiene conocimiento del valor que puede tener; y variables, no constantes. Para la comprensión de los datos estadísticos se debe partir por la organización, presentación y resumen de los mencionados datos.

Capacidades adquiridas:

- ✓ Comprender los conceptos básicos de la estadística.
- ✓ Clasificar los tipos de variables.
- ✓ Organizar y representar los datos en forma tabular y gráfica.
- ✓ Calcular las medidas resumen.
- ✓ Determinar la forma de distribución de los datos.

Escala ordinal

Consigue distinguir entre valores y además establece un orden entre ellos. Consideremos que a un individuo que manifiesta su satisfacción se le pide que seleccione entre 4 centros de salud (A, B, C y D) en función de aceptar más o menos. Su respuesta es [A = 1, D = 2, C = 3 y B 4]. Es obvio que el individuo no sólo no prefiere igual el modelo A que el B, sino que, además, prefiere el modelo A más que el B.

Escala de intervalo

Una escala de intervalo posee las características de una nominal (diferentes valores representan diferentes características de los objetos) y de la ordinal (mayor valor representa mayor presencia de la característica, por ejemplo, la preferencia). Sin embargo, la escala de intervalo añade una nueva propiedad: las diferencias también tienen sentido.

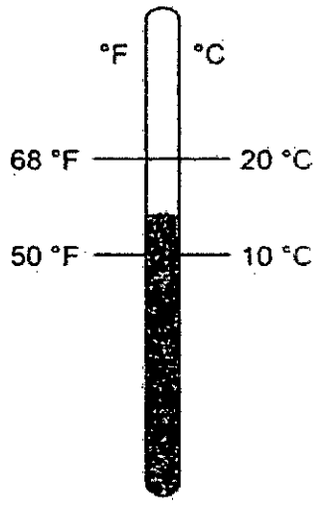


Gráfico 1.1. Ilustración de una escala de intervalo

Escala de razón

Esta escala tiene las mismas propiedades que el de intervalo, pero, además, las razones tienen sentido. Estas escalas tienen un valor base 0 natural: la edad, los ingresos, etc.

En dos procesos (nominal) distintos de obtención de un producto el tiempo fue de 40 minutos y el otro 20 minuto no solo son distintos tiempos sino el primero es mayor que el otro (ordinal).

Si un individuo tiene 20 años y otro tiene 10, no sólo tienen distintas edades (nominal), el primero es mayor que el segundo (ordinal), hay la misma diferencia de edad entre el primero y el segundo que entre el primero y un sujeto de 30 años (intervalo), sino que podemos afirmar sin problemas que el primero tiene el doble de edad que el segundo.

Variables *discretas* y *continuas*

Una variable discreta sólo puede tomar un número determinado de valores en un intervalo determinado: admisiones en un hospital. Número de productos. etc. Una variable continua, por el contrario, puede tomar potencialmente cualquier valor numérico en un intervalo dado.

El peso de un individuo como 80,0 kg como de 80,1223 kg.

Proceso Estadístico

En las diversas etapas de los procesos estadísticos es importante tener presente los términos para aplicar, por ejemplo, términos como 'población', parámetros, estimadores, factores, etc.

Los términos más usados en la aplicación de la estadística se presentan:

Variable

Es todo factor o característica que se encuentra en estudio en una muestra o población.

Clases de variables

a) Variables cualitativas

Son aquellas cuyos resultados posibles no pueden ser expresados en forma numérica. Las variables cualitativas pueden ser:

a.1) Variables cualitativas nominales

Son aquellas cuyas categorías posibles no tienen por qué ser presentadas en un orden definido. Por ejemplo:

- Color, textura olor de preferencia de las personas.
- La acción de mayor cotización en la bolsa de valores
- Tipo de proceso

a.2) Variables cualitativas jerárquicas

Son aquellas cuyas categorías posibles deben ser presentadas en un orden definido. Por ejemplo:

- Calidad de los artículos producidos por una empresa.
- Grado de instrucción de los empleados
- Calidad de gestión de una institución de salud

b) Variables cuantitativas

Son aquellas cuyos resultados posibles pueden ser expresados en forma numérica. Las variables cuantitativas pueden ser:

b.1) Variables cuantitativas discretas

Son aquellas que tienen un número finito o infinito numerable de valores posibles, se les asocia a procesos de conteo, donde el valor es un número entero. Por ejemplo:

- Cantidad de productos semanales en una empresa.
- Cantidad de empresas con problemas financieros.

b.2) Variables cuantitativas continuas

Son aquellas que tienen un número finito no numerable de valores posibles. Por ejemplo:

- Volumen mensual de producción de harina de pescado.
- Peso de un paciente
- Talla de un paciente.

Observación

Es el dato o registro que resulta de la apreciación de una característica en un individuo o unidad elemental. Una observación puede ser cualitativa o cuantitativa. Por ejemplo:

- "Rojo" es la observación del color preferido por una determinada persona.
- "240 toneladas" es la observación del volumen de producción de harina de pescado de una empresa.
- "Bueno" es la observación de la calidad de un producto después de ser revisada por un técnico en control de calidad.

Parámetro

Es una función de todas las observaciones de una población. Un parámetro es un indicador que resume la información contenida en las observaciones proporcionadas por todas las unidades elementales comprendidas en una población, por lo cual su valor es único y constituye usualmente un valor desconocido que todo investigador desea conocer. Los parámetros se definen también como valores constantes que caracterizan a una población. Algunos de los parámetros a los cuales se hará referencia son:

- Media poblacional, cuya notación es μ (se lee "mu").
- Variancia o varianza poblacional, cuya notación es σ^2 (se lee "sigma al cuadrado").
- Moda poblacional, cuya notación es Mo.

Estadístico o estimador

Es una función de las observaciones muestrales y que no depende de parámetros algunos. Un estadístico o estimador permite resumir la información contenida en las observaciones que comprende una muestra. Se caracterizan porque pueden tomar valores diferentes de muestra a muestra, debido a que las observaciones captadas en nuestras diferentes no son necesariamente a iguales. Los estadísticos o estimadores son útiles porque permiten obtener estimaciones (aproximaciones) del valor de los parámetros respectivos. Algunos de los estimadores a los cuales se hará referencia son los siguientes:

- Promedio o media muestral, cuya notación es $\overline{X}$.
- Variancia muestral, cuya notación es S^2.
- Moda muestral, cuya notación es m_o.

Los valores que se obtienen al aplicar los estimadores o estadísticos a una muestra particular son llamados "estimados" de los parámetros. Es decir, si para una muestra se obtiene:

$\bar{X} = 128.45$, $S^2 = 9.16$, $m_o = 125.42$,

estos valores son los estimados de los parámetros definidos como: media poblacional (μ), variancia poblacional (σ^2) y moda poblacional (Mo), respectivamente.

Hipótesis

Clasificación de hipótesis

Las hipótesis se han dividido en tres, basados principalmente en los tipos de investigación:

a) Hipótesis Descriptiva: Son hipótesis que reflejan alguna situación en un momento determinado, buscando la probable dirección de una variable. Es propia de estudios descriptivos y no siempre se encuentran redactadas de forma explícita.

Ejemplo:

El nivel de depresión que presentan los pacientes del Instituto Nacional de Ciencias Neurológicas, luego del evento cerebro vascular es moderado-alto.

b) Hipótesis Correlacional o Asociativa: Buscan una probable asociación entre variables que existen o varían al mismo tiempo, pero para las que no se propone una relación causa-efecto. Podemos incluir por ejemplo a los estudios de casos y controles y estudios de cohortes.

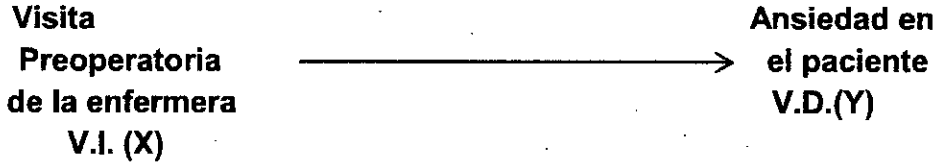
Ejemplo:

"El uso de ayudas mecánicas en la movilización de los pacientes reduce los tiempos de trabajo de enfermería y las lesiones de espalda en estos profesionales".

c) Hipótesis Causal: Describen la existencia de una relación en la que la variable independiente causa un efecto en la variable dependiente y modifica su comportamiento. Estas hipótesis pueden estar constituidas por dos, tres o más variables. Es propia de estudios de corte experimental y se busca comprobar una relación de causa efecto entre las variables. Aquí se pueden mencionar como ejemplos los estudios clínicos o estudios de intervención.

Ejemplo:

"La visita preoperatoria de la enfermera quirúrgica disminuye la ansiedad en el paciente sometido a cirugía de hernia de núcleo pulposo".



CAPÍTULO 2

PROTOCOLOS DE INVESTIGACIÓN

La evidencia empírica indica que un documento de investigación bien redactado requiere de cinco a diez, revisiones previas a la versión final. Esto puede ser tedioso pero vale el esfuerzo por el rigor y calidad que han de caracterizar a este tipo de documentos relacionados con la investigación científica.

Es necesario guardar coherencia en las normas ortográficas; utilizar correctamente las numeraciones, negrillas, cursivas, mayúsculas, así como citas y referencias bibliográficas, datos, gráficas, tablas, etc.

Capacidades adquiridas:

- ✓ Sabe identificar temas de interés para realizar investigación científica en el campo de su disciplina.
- ✓ Conoce los criterios para determinar la pertinencia e campo de su disciplina.
- ✓ Sabe plantear el problema y los objetivos de una investigación.
- ✓ Sabe justificar, delimitar y definir el tipo de estudio por realizar.
- ✓ Sabe elaborar el marco teórico y plantear la hipótesis de investigación.
- ✓ Sabe presentar la bibliografía consultada en la elaboración de la propuesta o anteproyecto de investigación

Consideraciones para elaborar un buen título

El título de la investigación debe reflejar una coherencia metodológica con el problema de investigación planteado y el objetivo general del estudio, a fin que motive al lector revisar la investigación y generar expectativa en lo que se espera encontrar en la publicación.

Contenido de un proyecto o protocolo de investigación

La estructura o elementos que debe contener un proyecto de investigación están generalmente determinados por el formato que exige la institución superior de origen o la institución pública o privada de donde proviene el investigador. Por lo tanto, luego de una revisión de los diversos componentes de los proyectos de investigación, se presentan a continuación cuatro modelos a manera de propuesta.

Estructura del informe de investigación

Elaborar un informe representa un reto para las personas que han decidido incursionar en el ámbito de la investigación. Se afirma que es un reto porque4 es necesario realizar un esfuerzo de discernimiento para identificar, seleccionar, evaluar y decidir respecto a la información relevante que debe contener un documento, el cual es resultado final de un proceso exploratorio y sobre todo, la forma de integrarlo. Este documento que puede ser un libro, una tesis, un artículo especializado, entre otros, debe describirse tomando en cuenta el perfil del usuario, quien toma sus decisiones con base en los resultados de la investigación.

Como todo proceso en los distintos campos del conocimiento, éste se basa en normas específicas, establecidas por instituciones académicas y científicas de cada país, que definen los protocolos y sistemas que deben ser considerados para la presentación de los informes.

Generalmente el Diseño Metodológico debe comprender los siguientes elementos:

- Tipo de estudio
- Población de estudio
- Muestra
- Procedimientos para la recolección de datos.
- Procesamiento y análisis de datos
- Consideraciones éticas.

Las fases de la investigación y la econometría tradicional

La investigación científica en economía está constituida por diversas fases interconectadas de una manera lógica, secuencial, pero también iterativa; es decir, que pueden atravesarse las distintas fases de la investigación una y otra vez.

Un resumen de todo el procedimiento hipotético deductivo anterior nos lo proporciona Figueroa (2009).

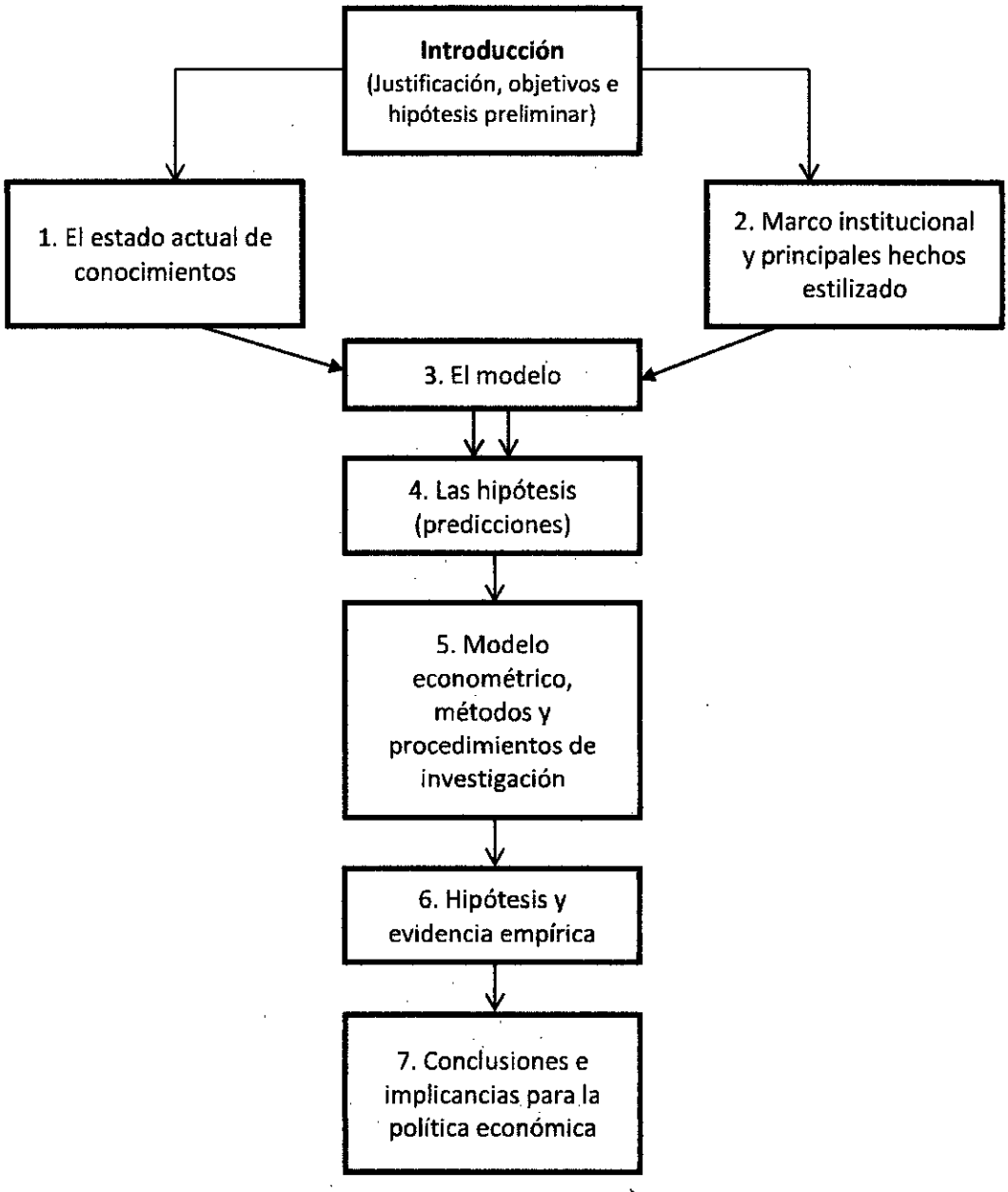


Gráfico 2.1: Fases de la investigación y la econometría tradicional

Consideraciones éticas en los proyectos de investigación

El aumento cada vez mayor de investigaciones con seres humanos ha generado preocupaciones y controversias en materia ética.

Según Polit (2000), la siguiente información se debe considerar esencial para la obtención de que los datos que proporcionen serán utilizados con fines científicos:

1. Condiciones del participante.- Los participantes potenciales deben ser informados de que los datos que proporcionen serán utilizados con fines científicos.
2. Objetivo del estudio.- Se explica el objetivo general de la investigación, de preferencia en términos coloquiales y no técnicos.
3. Tipos de datos.- Debe especificarse el tipo de datos que se solicitarán durante el estudio.
4. Naturaleza del compromiso.- Debe proporcionarse información relativa a la duración del estudio y el tiempo de participación estimado.
5. Patrocinio.- Se menciona quién financia o patrocina el estudio, o si bien la investigación es para optar un título profesional.
6. Selección de los participantes. - Se debe explicar cómo se seleccionó a los participantes.
7. Procedimientos.- Loa participantes deben conocer los procedimientos que se utilizaran para recabar los datos, así como cualquier tratamiento especial o experimental.
8. Riesgos o costos ocultos.- Se debe informar a los participantes de cualquier riesgo potencial previsible (físico, psicológico o económico) y sobre los costos en que se podría incurrir por participar.

9. Beneficios potenciales.- Los beneficios particulares para los participantes, deben especificarse, así como la información sobre posibles beneficios para terceros. Si habrá algún pago a los sujetos, éstos deben estar previamente enterados.
10. Garantía de confidencialidad.- Debe asegurarse a los participantes que su privacidad estará protegida en todo momento.
11. Consentimiento informado.- Se debe indicar con toda claridad que la participación es estrictamente voluntaria y que no habrá sanciones ni pérdidas de beneficios de ningún tipo si no se cumple.
12. Derecho a retirarse.- Se debe informar a los participantes que, aun cuando acepten colaborar en la investigación, tendrán derecho a retirarse del estudio y rehusarse a proporcionar información específica en cualquier momento.
13. Alternativas.- Si resulta apropiado, el investigador debe ofrecer información sobre procedimientos a tratamientos alternativos que pueden resultar beneficiosos para el paciente.
14. Información para el establecimiento de contactos.- El investigador debe informare con quién pueden ponerse en contacto en caso de que los participantes tengan dudas, comentarios o quejas relacionadas con la investigación, registrando telefonía móvil o fija.

2.1 Diseño de investigación

La investigación tiene sus bases en la ciencia y la ciencia es investigación; la investigación adquiere valor científico a través de las aportaciones de la metodología. Como medio de trabajo intelectual, la investigación se define como un procedimiento reflexivo, sistemático, controlado, metódico y crítico que conduce hacia el descubrimiento en cualquier campo del conocimiento. La investigación aplicada a las disciplinas sociales se define como proceso de obtención sistemática de respuestas a las preguntas significativas utilizando el método científico de captación e interpretación e interpretación de información.

Las características de cualquier investigación científica deben ser de rigor científico basado en cimientos teóricos y metodológicos; debe ser comprobable, objetiva, real, generalizada y precisa.

Existen dos tipos de investigación científica: según los propósitos que persigue, ésta puede ser investigación básica o investigación aplicada. Según las fuentes de datos puede ser documental, de campo o mixta.

La investigación científica es un proceso lógico y riguroso que aplica el método científico. Asimismo, la investigación social, económico-administrativo, etc., es un proceso organizado, asistemático, crítico y científico de captación de información sobre aspectos de las disciplinas para dar respuestas a problemas de investigación.

Las etapas del proceso de investigación científica aplicada son: tema, problema, antecedentes, variables y marco teórico; hipótesis y sus concretizaciones, metodología, procesamiento y medición, interpretación y presentación.

La metodología describe y analiza los métodos; la metodología científica es el uso de aquellos métodos que obtengan la información requerida para el problema y la hipótesis planteada con el mínimo de tiempo, recursos humanos y gastos.

Para la investigación social se aplican preferentemente cuatro métodos específicos: El método de encuesta que utiliza un cuestionario para recabar información de fuentes primarias. El cuestionario se define como una hoja de cuestiones o preguntas ordenadas y lógicas que sirven para obtener información fidedigna y confiable de la muestra del universo para poder enjuiciar las hipótesis. El método de entrevista es un intercambio

conversacional entre dos personas con la finalidad de obtener información para reforzar el problema e hipótesis. La entrevista puede ser dirigida o informal. El método de observación es la acción de mirar detenidamente una cosa para asimilar en detalle la naturaleza investigada, un conjunto de datos, hechos y fenómenos. El método de observación en la investigación social es aquel en el que el mismo objeto de estudio sirve de fuente de información al investigador; el cual recoge directamente los datos de las conductas observadas. El método de observación indirecta es el que mayor uso tiene en las investigaciones sociales aplicadas; consiste en tomar nota de un hecho que sucede ante los ojos de un observador entrenado, midiendo el comportamiento externo del individuo en la interacción con su medio o dentro de la propia organización.

El método experimental es aquel que se emplea para comprobar y medir variaciones o efectos que sufre una situación cuando en ella se introduce una nueva causa, dejando las demás causas en igual estado. Para realizar un diseño experimental, en la ciencia se requiere por lo menos de tres tipos de variables: 1) una variable independiente cuyos efectos han de ser medidos; 2) una variable independiente cuyos efectos han de ser controlados, y 3) las variables dependientes que son observadas para determinar las consecuencias del experimento.

Entre los muchos diseños de investigación que se pueden aplicar, los diseños antes-después sin grupo de control y antes-después con grupo de control son los más útiles y sencillos de aplicar. Otros diseños experimentales como el cuadro latino y diseño factorial son más complejos. Cuanto más grupo de control se establezcan en un diseño experimental, mayor confiabilidad se tendrá en el resultado del diseño experimental de investigación.

A) Tipos de estudio

Se clasifican en:

A.1) Según el tiempo de ocurrencia de los hechos y registros de la información

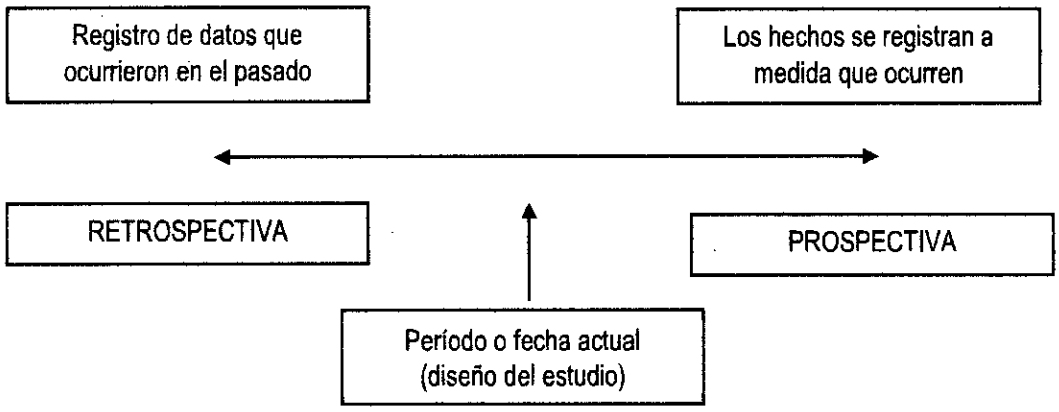


Gráfico 2.2: Según el tiempo de estudio

A.2) Según el período y la secuencia del estudio

Los estudios se clasifican en transversales y longitudinales. Una investigación es *Transversal* cuando se estudian las variables simultáneamente en un determinado momento, haciendo un corte en el tiempo. En este, el tiempo no es importante en relación con la forma en que se dan los fenómenos.

A.3) Según la manipulación o introducción deliberada de una variable

a.3.1) Diseño No Experimental u Observacional (Estudio Descriptivo)

Diseño Analíticos / Correlacionales

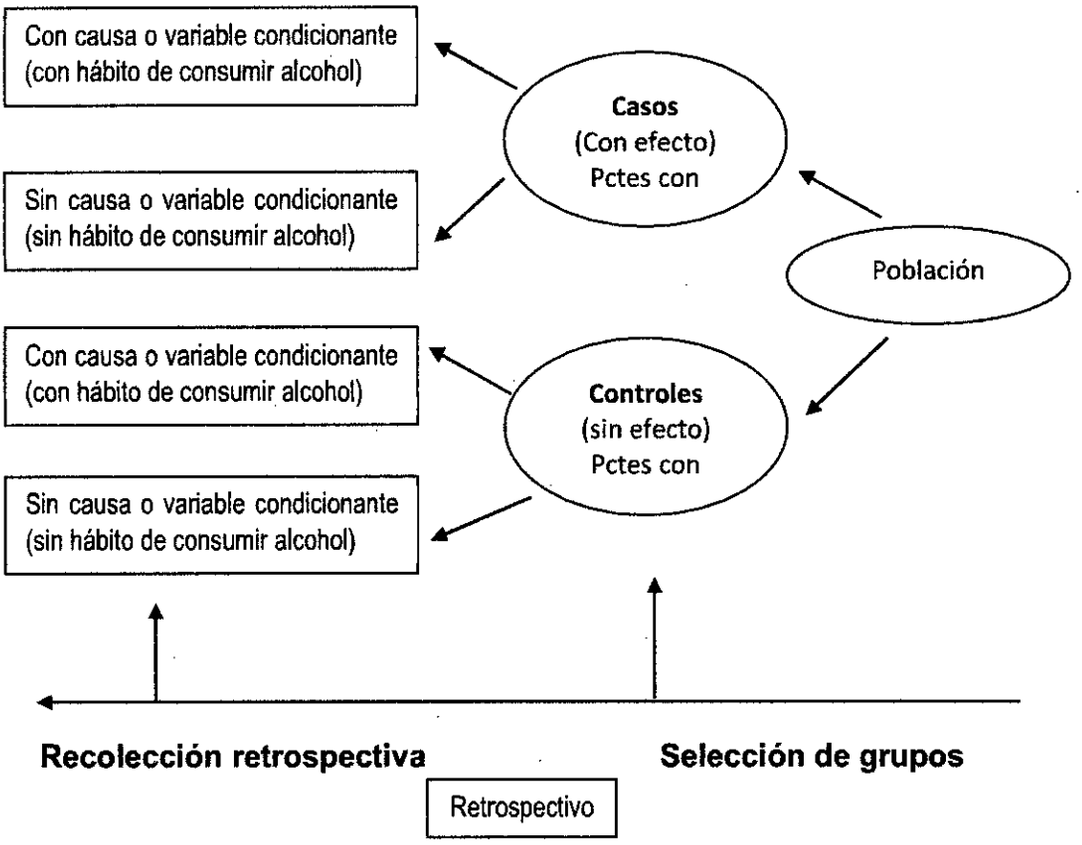


Gráfico 2.3: Estudios de Casos y Controles

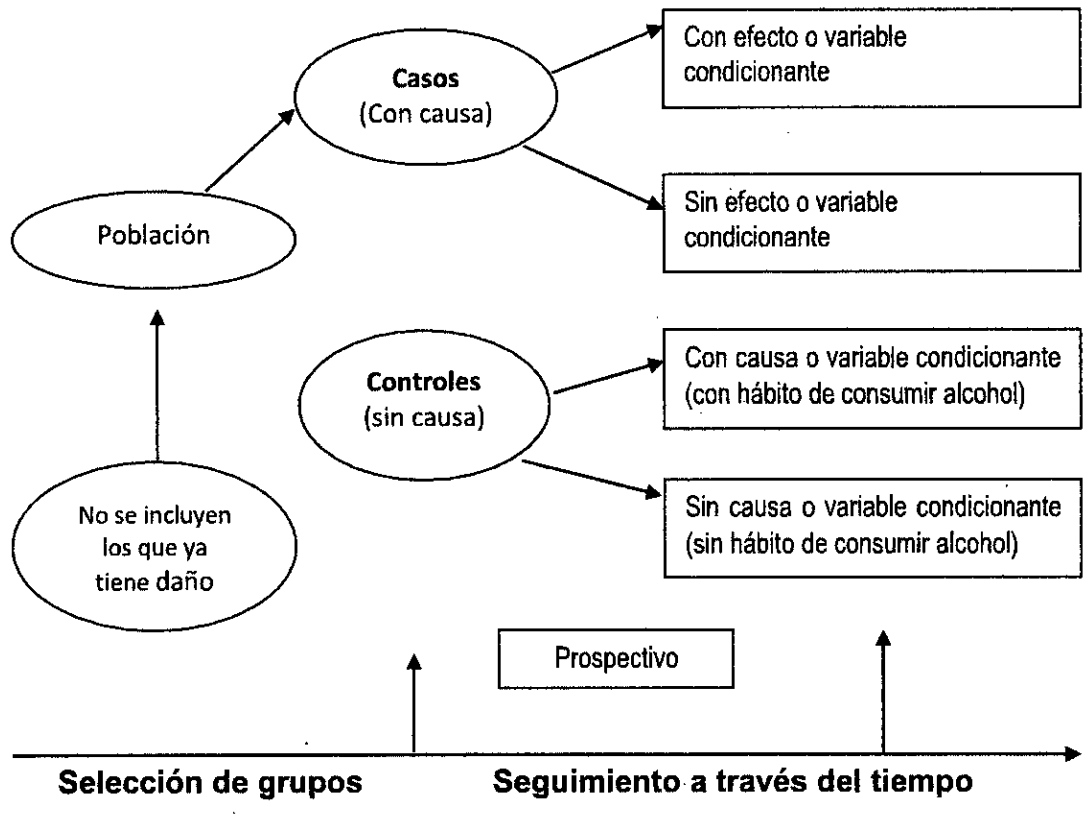


Gráfico 2.4: Estudios analítico de Cohorte

a.3.2) Diseño Experimental

- Estudio Preexperimental

El esquema que presenta un estudio experimental es el siguiente:

Grupo Experimental	ANA O_1 X O_2

Donde:

ANA = Asignación no aleatoria

O_1, O_2 = Observaciones o mediciones

X = Aplicación o manipulación de la variable independiente (V.I.)

- Diseño Cuasiexperimental:

Grupo Experimental	ANA O_1 X O_2
Grupo Control	ANA O_3 O_4

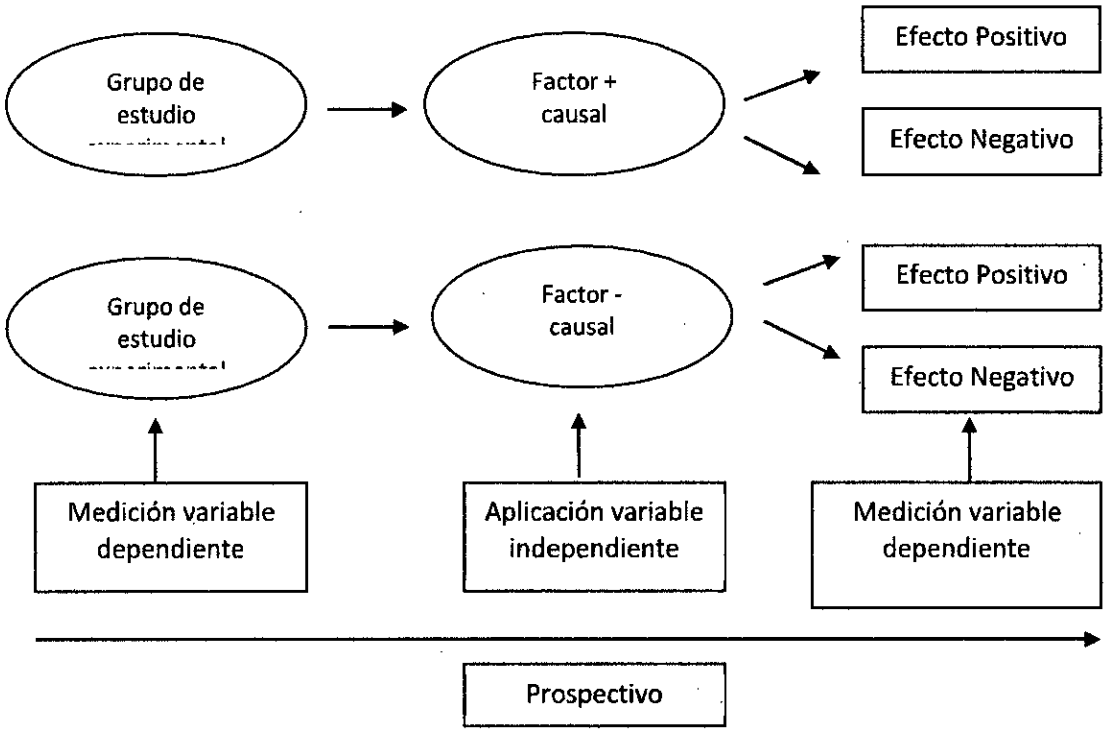


Gráfico 2.5: Diseño de estudio experimental

Requisitos de un instrumento de investigación

Existen dos requisitos fundamentales que deben poseer los instrumentos de medición como parte de la exigencia científica, la validez y la confiabilidad.

a) Confiabilidad

La evaluación de la confiabilidad de un instrumento puede evaluarse de diversas formas, tenemos tres aspectos importantes que deben ser

considerados para hallar la confiabilidad de los instrumentos; estabilidad, equivalencia y homogeneidad; dependerá de la naturaleza del instrumento para ser elegido uno de ellos.

En la actualidad, la técnica de homogeneidad es la que más recurren los investigadores para estimar la confiabilidad de un instrumento y se recomienda el uso del coeficiente alfa de Crombach, aunque debemos mencionar que existe también el coeficiente de Kudder Richardson (KR-21) como alternativa ante determinados instrumentos, dependiendo del tipo de ítem elegido para el instrumento. La confiabilidad la medimos por intervalos de valores, el intervalo normal de valores fluctúa entre 0.00 y +1.00. el coeficiente de 1.00 una confiabilidad perfecta y 0.00 que no hay confiabilidad. Una confiabilidad de 0.80 se considera el coeficiente mínimo aceptable para una herramienta de medición. Para Burns (2004), la confiabilidad que supera el 0.70 es aceptable.

b) Validez

La validez se calcula a través de varias pruebas y métodos, pero la más utilizada es a través de la correlación de Pearson.

Plan de tabulación

Los pasos que se siguen según Pineda y Alvarado (2008), para la construcción de un plan de tabulación se resumen a continuación:

1. Detallar las variables identificadas y que serán objeto de estudio.
2. Determinar las variables que ameritan sean analizadas individualmente o presentadas en cuadros simples de una variable, según los objetivos e hipótesis.
3. Determinar las variables que deben cruzarse, según los objetivos y las hipótesis.
4. Esquematizar, en los casos que lo ameriten, el cuadro para determinar relación de variables y las escalas de cruce de variables.
5. Hacer el listado de los cuadros que deberán presentarse.

Procesamiento y análisis de la información

El fin del procesamiento de la información es resumir, organizar y comunicar la información. Consiste en procesar los datos (dispersos, desordenados, individuales) obtenidos de la población objeto de estudio durante el trabajo de campo y tiene como finalidad generar resultados.

Los datos:

El proceso tiene varios momentos:

a) Ordenamiento de la información

Es recomendable conocer el número total de encuestas que tiene asignados un número o código.

Estos números se utilizan para controlar que, siempre que se tabule algo, los datos deben coincidir con estos totales.

Por ejemplo, en un estudio multicéntrico en cuatro institutos, sería importante saber el número de encuestas por sede de estudio (Instituto Nacional de Ciencias Neurológicas = 334 participantes, Instituto de Salud del Niño = 232 participantes; Instituto Materno Perinatal = 189 participantes y Instituto Nacional de Enfermedades Neoplásicas = 150 participantes) y el total de los cuatro (total de personas en la muestra = 905).

b) Revisión y depuración de la información – control de calidad de los datos

Se deben revisar los datos originales, a fin de corregir o subsanar la información incorrecta, incompleta o ilegible.

c) Captura de la información

La codificación de la información se hace para las variables cualitativas. Las variables son numéricas, no se necesita codificar.

En caso de que la información esté precodificada en el instrumento de recolección de datos, se respeta esa codificación.

En el caso de preguntas que se pueda marcar más respuestas, cada opción debe tomar una variable:

¿Qué hace en su tiempo libre?

1. Leer ( )
2. Escuchar música ( )
3. Practicar deportes ( )
4. Otros

"Leer" será una variable y tendrá dos opciones: "sí" para los que marcaron el paréntesis respectivo y "no" para los que lo dejaron en blanco y así sucesivamente con las opciones de respuesta a los incisos 2, 3 y 4

En el caso de preguntas abiertas que requieran respuestas que no han sido precodificadas, pues no se sabe el tipo de contestación que pueden dar las personas encuestadas, lo primero que se debe hacer es leer y hacer una pretabulación de una parte de las respuestas y elaborar los códigos.

d) Análisis e interpretación de datos

Determinará sí se da respuesta al problema, a la hipótesis o a las preguntas de la investigación.

Campos (1982), citado por Pineda y Alvarado (2008), afirma que significa determinar y exponer el plan que se deberá seguir para el tratamiento estadístico de los datos; en general, consiste en describir cómo será tratada la información.

Mencionaremos a continuación las principales pruebas estadísticas que se pueden aplicar para el análisis de los datos según el nivel de medición de las variables: moda, mediana, frecuencia, promedio, desviación estándar, chi cuadrado, prueba F, rho Spearman, Kruskall Wallis, Mann Whitney, análisis de Varianza, correlación de Pearson, prueba t, correlaciones lineales y múltiples.

Lo más importante en el uso de la estadística no es saber calcular un valor a través de alguna técnica, sino qué técnica usar y cómo interpretar el resultado final.

Para el análisis de los datos será necesario seguir los siguientes pasos:

1. Descripción de la muestra: obtener una frecuencia de todas las variables que describen la muestra. Ejemplo: edad, sexo, nivel socioeconómico, procedencia, religión.
2. Análisis descriptivo: Describe y sistematiza los datos hallados.
3. Análisis inferencial: Se puede estimar parámetros, prueba de hipótesis, prueba de diferencias de medias o proporciones entre grupos.

CAPÍTULO 3

RECOLECCIÓN DE DATOS

El procedimiento de datos debe realizarse mediante la asesoría de expertos en estadística y el uso adecuado de herramienta; por ejemplo, algún programa de estadística que se pueda acceder y esté disponible para trabajar en el computador.

Cada proyecto tiene sus particularidades, lo que conduce a que el procesamiento de los datos se haga de manera particular; por tanto, las herramientas estadísticas deben ser las adecuadas en función de los objetivos y las hipótesis de la investigación, si las hubo. Cabe recordar que, para estos efectos se puede utilizar el aplicativo de Excel.

Capacidades adquiridas:

- ✓ Conocer el procedimiento para recolectar la información necesaria en el desarrollo de la investigación.
- ✓ Conoce y utiliza diferentes herramientas estadísticas para procesar la información obtenida en el trabajo de campo.
- ✓ Conoce los criterios que se requieren para analizar los resultados del procesamiento de datos obtenidos en el desarrollo de la investigación.
- ✓ Sabe redactar las conclusiones de los resultados de la investigación.
- ✓ Sabe redactar el marco teórico definitivo del estudio.
- ✓ Sabe redactar las referencias bibliográficas para el informe final .

3.1 Medidas de tendencia central

Las medias de tendencia central o dispersión nos indican donde se sitúa un dato dentro de una distribución de datos. Las medidas de dispersión, variabilidad o variación nos indican si esos datos están próximos entre sí o sí están dispersos, es decir, nos indican cuán esparcidos se encuentran los datos. Estas medidas de dispersión nos permiten apreciar la distancia que existe entre los datos a un cierto valor central e identificar la concentración de los mismos en un cierto sector de la distribución, es decir, permiten estimar cuán dispersas están dos o más distribuciones de datos. Estas

medidas permiten evaluar la confiabilidad del valor del dato central de un conjunto de datos, siendo la media aritmética el dato central más utilizado. Cuando existe una dispersión pequeña se dice que los datos están dispersos o acumulados cercanamente respecto a un valor central, en este caso el dato central es un valor muy representativo. En el caso que la dispersión sea grande el valor central no es muy confiable. Cuando una distribución de datos tiene poca dispersión toma el nombre de distribución homogénea y si su dispersión es alta se llama heterogénea.

Las medidas de tendencia central se utilizan con bastante frecuencia para resumir un conjunto de cantidades o datos numéricos a fin de describir los datos cuantitativos que los forman. Ejemplos de ello, pueden ser: la edad promedio o la estatura promedio de los estudiantes de la universidad o el peso promedio de las bolsas de cereal que son llenadas por una determinada máquina en un proceso de producción o las ventas de un negocio.

Las medidas de tendencia central que se van a tratar en esta unidad son:

Media Aritmética o promedio. La media es un concepto estadístico básico que representa en un valor las características que presenta una variable de un conjunto de datos, y sólo puede usarse con variables cuantitativas. La media puede considerarse un concepto base para la comprensión de variable aleatoria y sus distribuciones, ya que la distribución se caracteriza principalmente por las medidas de tendencia central y de dispersión, siendo frecuentemente la media uno de los parámetros de las distribuciones. (Estrella 2016)

La media aritmética, o promedio aritmético, es la suma de los valores del grupo de datos dividida entre la cantidad de valores. Su fórmula se puede describir de la siguiente manera:

Mediana. Es el valor del elemento central del conjunto. Para encontrar la mediana, primero arreglar los valores del conjunto de acuerdo a su magnitud; es decir, arreglar los valores del más pequeño al más grande o del más grande al más pequeño y después localizar el valor central, es decir, el número de valores sobre la mediana es el mismo que el número de valores debajo de la mediana. Si el número de valores en un conjunto de datos no agrupados es par, no hay mediana verdadera.

Moda. También llamada modo o promedio típico de un conjunto de valores; la moda es el valor el cual ocurre más frecuentemente en el conjunto. Si un valor es seleccionado al azar del conjunto dado, un valor modal es el valor más probable a ser seleccionado. Así, la moda es generalmente considerada como el valor más típico en una serie de datos la cual es llamada, por esa razón, **Unimodal.** Un conjunto pequeño de datos en el que no se repiten valores medidos carece de moda. Cuando dos valores no adyacentes son casi iguales en cuanto a frecuencias máximas asociadas con ellos, la distribución se llama **Bimodal**, aquéllas con varias modas se llaman multimodales.

3.2 Medidas de dispersión

Los estudios estadísticos permiten hacer inferencias de una característica de una población a partir de la información contenida en una muestra. Los métodos numéricos que describen a los conjuntos de observaciones tienen como objetivo dar una imagen mental de la distribución de frecuencias.

Una vez localizado el centro de la distribución de un conjunto de datos, lo que procede es buscar una medida de dispersión de los datos.

La dispersión o variación es una característica importante de un conjunto de datos porque intenta dar una idea de cuán esparcidos se encuentran éstos.

Existen diversas medidas de dispersión, algunas de ellas son:

Rango. El rango de un conjunto de números es la diferencia entre el mayor y el menor de todos ellos. Hay 2 maneras de expresar ésta medida:

- La diferencia entre los valores mayores y menor.
- Los valores mayor y menor del grupo.

Desviación estándar. La desviación típica o desviación estándar (denotada con el símbolo σ o s, dependiendo de la procedencia del conjunto de datos) es una medida de dispersión para variables de razón (variables cuantitativas o cantidades racionales) y de intervalo. Se define como la raíz cuadrada de la varianza de la variable.

Varianza. Encontramos varianza, que es como la mayor parte de los textos científicos en castellano se refieren a la media aritmética de los cuadrados de las desviaciones de cada valor respecto de la media aritmética de los datos (por lo que a veces también se denomina desviación cuadrática media). La desviación estándar es la raíz cuadrada de la varianza. En algunos textos en castellano se ve variancia en vez de varianza, pero esta grafía se usa muy poco, pese a ser la recomendada por la Real Academia. La varianza es la media aritmética de los cuadrados de las desviaciones respecto a la media aritmética, es decir, es el promedio de las desviaciones de la media elevadas al cuadrado.

3.3 Medidas de forma: Asimetría y Curtosis

Asimetría

Las distribuciones pueden ser simétricas o asimétricas. Se dice que son simétricas cuando las dos colas de su histograma (derecha e izquierda) tienen la misma longitud. Esto es más fácil de visualizar que de explicar". Los tres histogramas que recoge la figura 2.18 corresponden a tres posibles

situaciones en cuanto a la asimetría; en cada situación es posible calcular un *coeficiente de asimetría,* que puede tomar valores negativos o positivos. La expresión matemática del coeficiente de asimetría es complicada y habitualmente se recurrirá al ordenador para calcularla. Cuando hay asimetría positiva, la cola de la derecha es más prolongada y su coeficiente de asimetría será positivo. En caso de asimetría negativa, la cola de la izquierda será más larga y el coeficiente, negativo. *Lo ideal para muchos procedimientos estadísticos es que la asimetría no sea grande y el coeficiente de asimetría esté lo más próximo posible a 0.*

En una variable que no puede tomar valores negativos, solo con conocer la media y la desviación estándar, ya podría decirse que tendrá siempre asimetría positiva cuando su desviación estándar sea superior al 50% de la media (es decir, si su coeficiente de variación es superior al 50%).

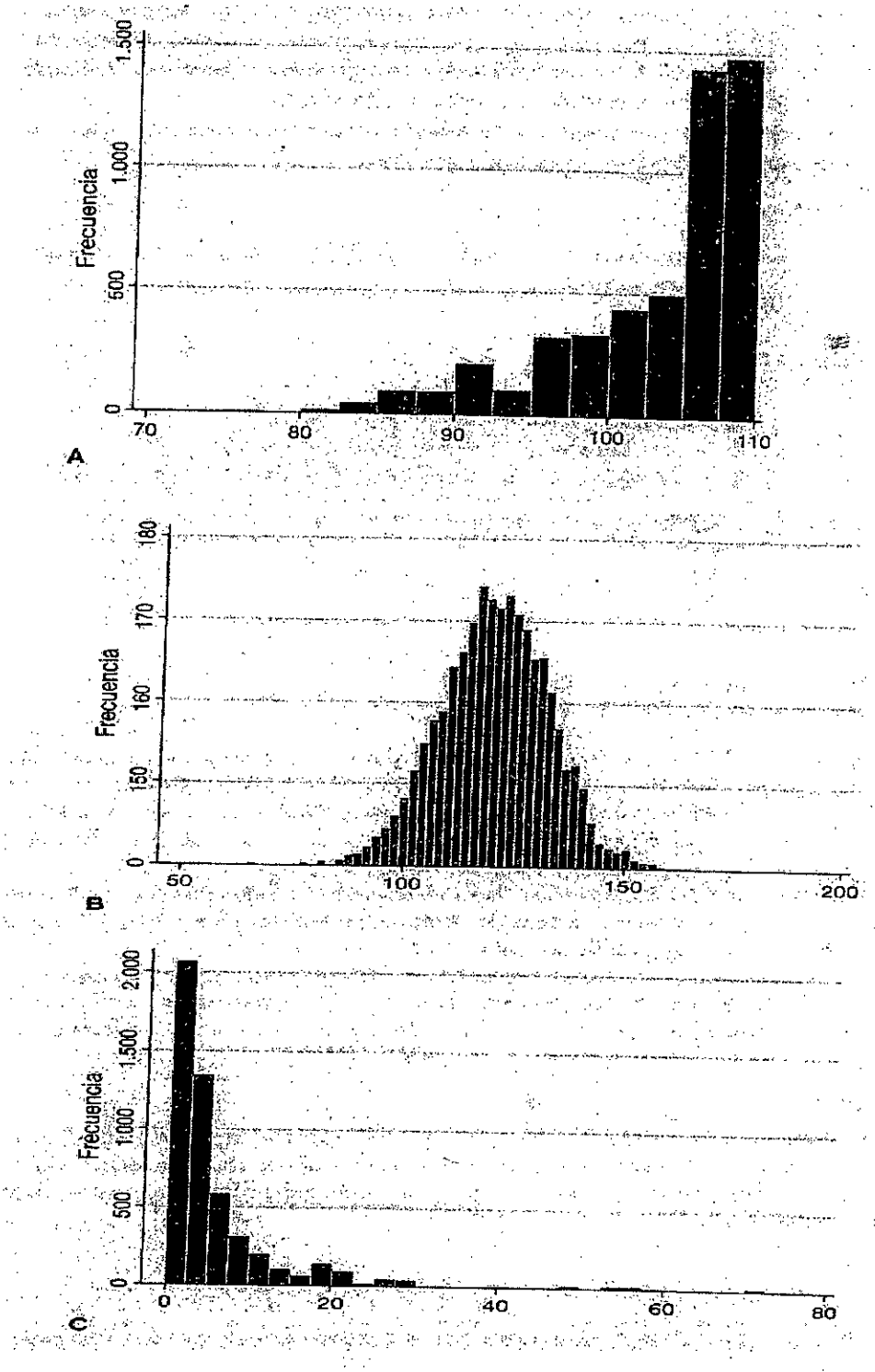


Grafico 3.1: Medida de forma: coeficiente de asimetría. A. Asimetría negativa < 0. B. Simetría perfecta = 0. C. Asimetría positiva > 0.

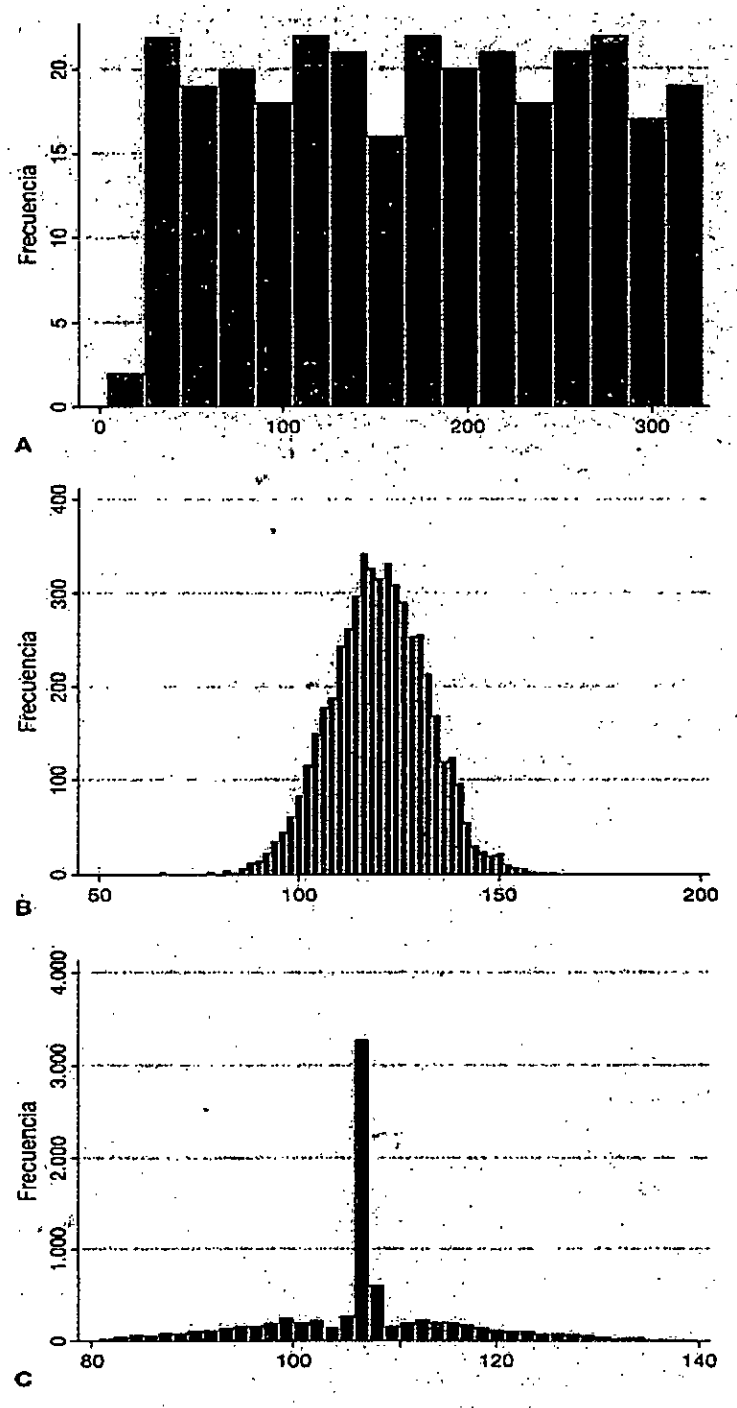


Grafico 3.2: **Medida de forma: coeficiente de curtosis. A. Curtosis negativa, <3 (en STATA), <0 (en otros), platicúrtica. B. Mesocúrtica (normocúrtica): curtosis = 3 (STATA), curtosis = 0 (otros). C. Curtosis positiva, > 3 (STATA), > 0 (otros), leptocúrtica.**

Curtosis o apuntamiento

El apuntamiento o curtosis mide el grado en el que un histograma resulta picudo o aplastado (fig. 2.19). Lo ideal es que el valor de la curtosis sea intermedio (próximo al valor nulo, mesocúrtico o normocúrtico). En casi todos los programas de estadística, el valor nulo de la curtosis es 0. Sin embargo, STATA suma tres unidades al calcular el coeficiente de curtosis y entonces el valor nulo es 3. Cuando se cumple esta condición y la asimetría es casi inexistente, se podrá considerar la distribución de los datos como normal. Como se verá, este tipo de distribución facilita enormemente el trabajo.

3.4 Medidas de posición: Cuantiles, percentiles

Los cuantiles son medidas de posición. Indican qué puesto ocupa un determinado valor de una variable en el conjunto ordenado de los datos de esa variable. Este puesto o posición se expresa como la proporción o porcentaje de los datos que queda por debajo de ese valor. A esta cantidad se le llama percentil. Así, que un niño esté en el percentil 80 del peso para su edad quiere decir que el 80% de los niños de su edad pesan menos que él. Si un alumno está en el percentil 100 de las notas de la clase, es que es el que mejor nota tiene de toda la clase.

Para calcular los percentiles se ordenan todas las observaciones de la distribución de menor a mayor y se busca aquel valor que deja un determinado porcentaje de las observaciones por debajo de él. Ya se ha visto que la mediana es el percentil 50 (P_{50}) porque deja por debajo al 50% de los sujetos. El percentil 5 es el que deja al 5% debajo de él, el percentil 90, al 90% de los individuos de la muestra, y así sucesivamente.

Al hablar de los diagramas de caja ya se habían mencionado los percentiles 25 (P_{25}) y 75 (P_{75}). La mediana y estos dos percentiles (P_{25} y P_{75}) son tres puntos de corte que dividen la muestra ordenada en cuatro partes iguales. Estos tres puntos de corte se llaman cuartiles. El rango intercuartílico (RIC) es la distancia entre el primer y el tercer cuartil ($RIC = P_{75} - P_{25}$).

También se habla de *terciles,* que son aquellos dos valores que dividen la muestra en tres grupos de igual tamaño. El primer tercil (o tercil 1) sería equivalente al percentil 33,33 y el segundo tercil, al percentil 66,67. Hay cuatro quintiles correspondientes a dar puntos de corte en los percentiles 20, 40, 60 y 80. También podría hablarse de *deciles*. Existen nueve puntos de corte (del percentil 10 al percentil 90) para definir 10 deciles.

No obstante, son términos equívocos y en la literatura científica es muy común el uso de, por ejemplo, quintil para hacer referencia tanto a los cuatro puntos de corte (P_{20}, P_{40}, P_{60} y P_{80}) como a los cinco grupos de observaciones que quedan delimitados por estos cuatro cortes. De esta manera, el grupo de observaciones que queda por debajo del P_{20} se denominaría el primer quintil, entre P_{20} y P_{40} el segundo quintil, etc. A su vez, al grupo situado por encima de P_{80} se le llamará el quinto quintil. Conviene prestar atención para identificar en qué caso nos encontramos.

Para explicar cómo calcular un percentil se usará un ejemplo sencillo. Se dispone de las edades ordenadas de menor a mayor de ocho sujetos:

28 31 33 33 34 38 40 42

Se aplica una interpolación. Si se desea calcular, por ejemplo, el *percentil* 25, se debe calcular la siguiente expresión, donde *i* es el percentil expresado en tanto por uno:

$$\text{Puesto} = i(n + 1)$$

$$\text{Puesto} = 0{,}25 \times (8 + 1) = 2{,}25.^{\circ}$$

El puesto que le correspondería al percentil 25 es el número de orden 2,25.° Para hallar el percentil 25 (P_{25}) se buscará, por tanto, el valor que ocupa el puesto 2,25. ° en el conjunto ordenado de datos. El puesto 2. ° está ocupado por el valor 31. El siguiente valor (el 3.er puesto) es 33. Interpolando resulta:

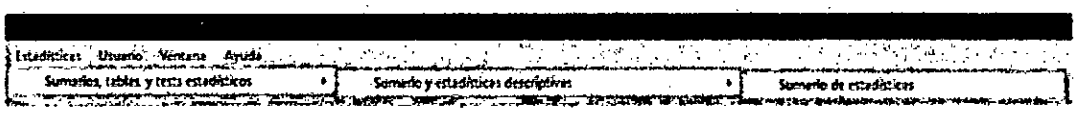
$$P_{25} = 31 + [0{,}25 \times (33 - 31)] = 31 + (0{,}25 \times 2) = 31{,}5$$

El percentil 25 valdrá por tanto *31,5*. Puede comprobarse que P_{75} = 39,5. El fundamento de este procedimiento es el siguiente: el decimal del número de puesto sirve de «factor de peso» para interpolar una fracción de la diferencia entre el puesto previo y el posterior. De este modo, el valor del percentil será más cercano a aquel de los dos valores que lo flanquean que se acerque más a su posición. El resultado del puesto o número de orden (2,25.° para el percentil 25) indica que el percentil 25 está a un 25% de la distancia que hay entre el puesto 2.° (valor = 31) y el 3.° (valor = 33). Se calcula cuál es el 25% de la distancia entre 31 y 33, y se suma esa distancia a 31. Por eso se dice que el cálculo se basa en la interpolación. No es el único modo de calcular percentiles. Hay otras aproximaciones. Por ejemplo, cuando se usa STATA para hacer gráficos de caja, a veces se obtiene otro resultado, porque STATA buscará los valores que se hayan observado realmente y estén más próximos al percentil teórico cuando se dibuja el gráfico de caja. No hay que preocuparse por esto. Habitualmente se hará con ordenador y se debe aceptar el gráfico resultante. Cuando el tamaño de muestra es grande, estas diferencias no se suelen notar.

3.5 Representación. Procedimientos descriptivos con programas

Análisis exploratorio de datos Uno de los objetivos de la Estadística es el de describir en unas pocas medidas resumen las principales características de un amplio conjunto de datos, de forma que estas medidas reflejen lo más fielmente las principales peculiaridades de dicho conjunto. El análisis exploratorio de datos tiene por objetivo resumir y visualizar los datos de manera que se facilite la identificación de tendencias o patrones que los subyacen y que son relevantes para responder a la pregunta o preguntas de interés. Dicho análisis se basa en el uso de gráficos, tablas y estadísticos descriptivos que permiten explorar la distribución de los datos identificando características tales como: valores atípicos u "outliers", sesgo, saltos o discontinuidades, concentraciones de valores, forma de la distribución, etc.

Estadísticas descriptivas para variables cuantitativa



Comando ***summarize***[9]

```
. sum peso
```

Variable	Obs	Mean	Std. Dev.	Min	Max
peso	75	3149.933	627.01	907	4394

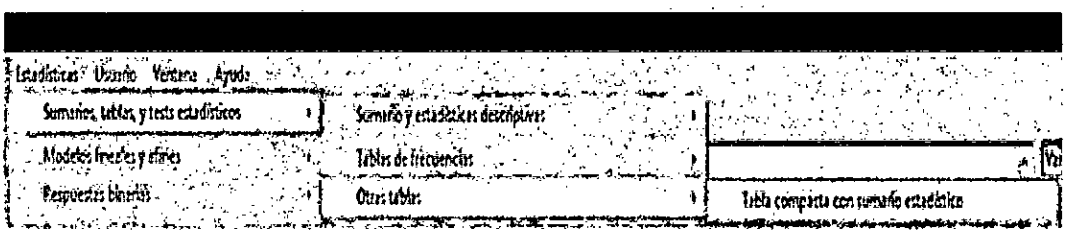
Si deseamos información de las estadísticas descriptivas de forma más detallada, agregamos la opción ***detail***

```
. sum peso, detail
```

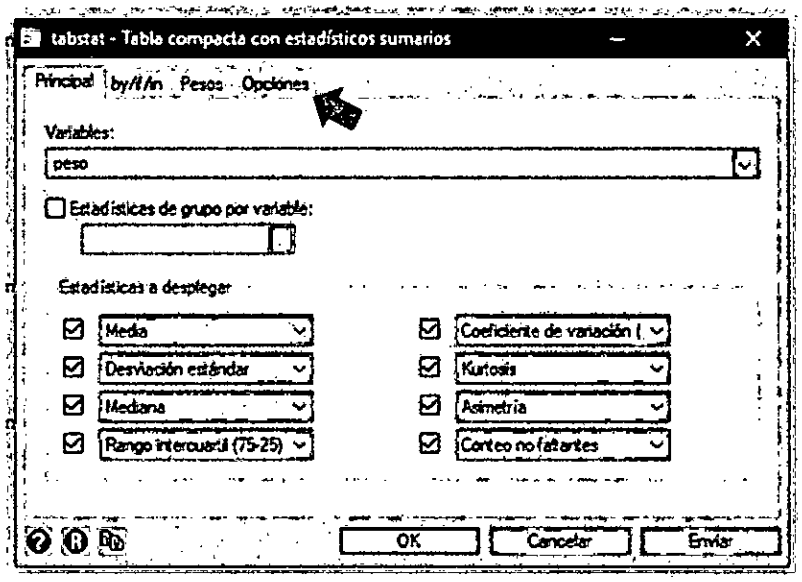
peso

	Percentiles	Smallest		
1%	907	907		
5%	1928	1644		
10%	2495	1899	Obs	75
25%	2778	1928	Sum of Wgt.	75
50%	3175		Mean	3149.933
		Largest	Std. Dev.	627.01
75%	3629	4050		
90%	3912	4082	Variance	393141.5
95%	4050	4139	Skewness	-.6813775
99%	4394	4394	Kurtosis	4.066858

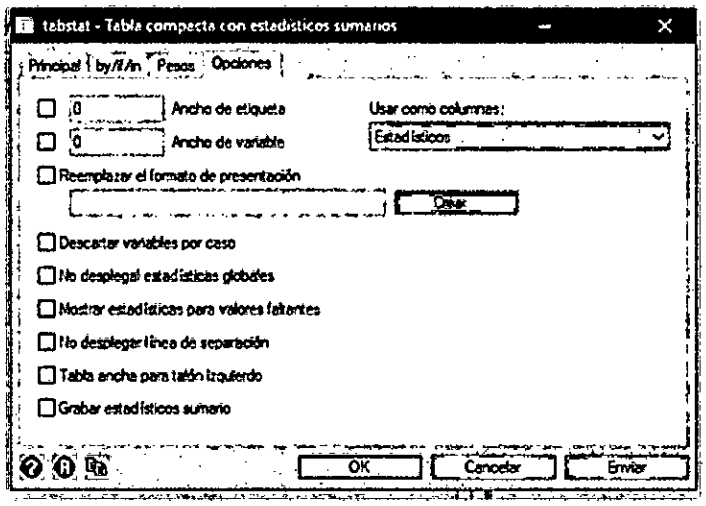
Otra manera es según el siguiente recorrido: **(Statistics/Summaries, tables, and tests/Other tables/Compact table of summary statistics)**



Obtendremos la siguiente ventana, donde colocamos la variable cuantitativa. Además podemos indicar los estadísticos. Antes de terminar el proceso nos vamos a Options



Obtendremos la siguiente ventana, indicaremos que use como columnas estadísticas. Finalmente presionamos OK

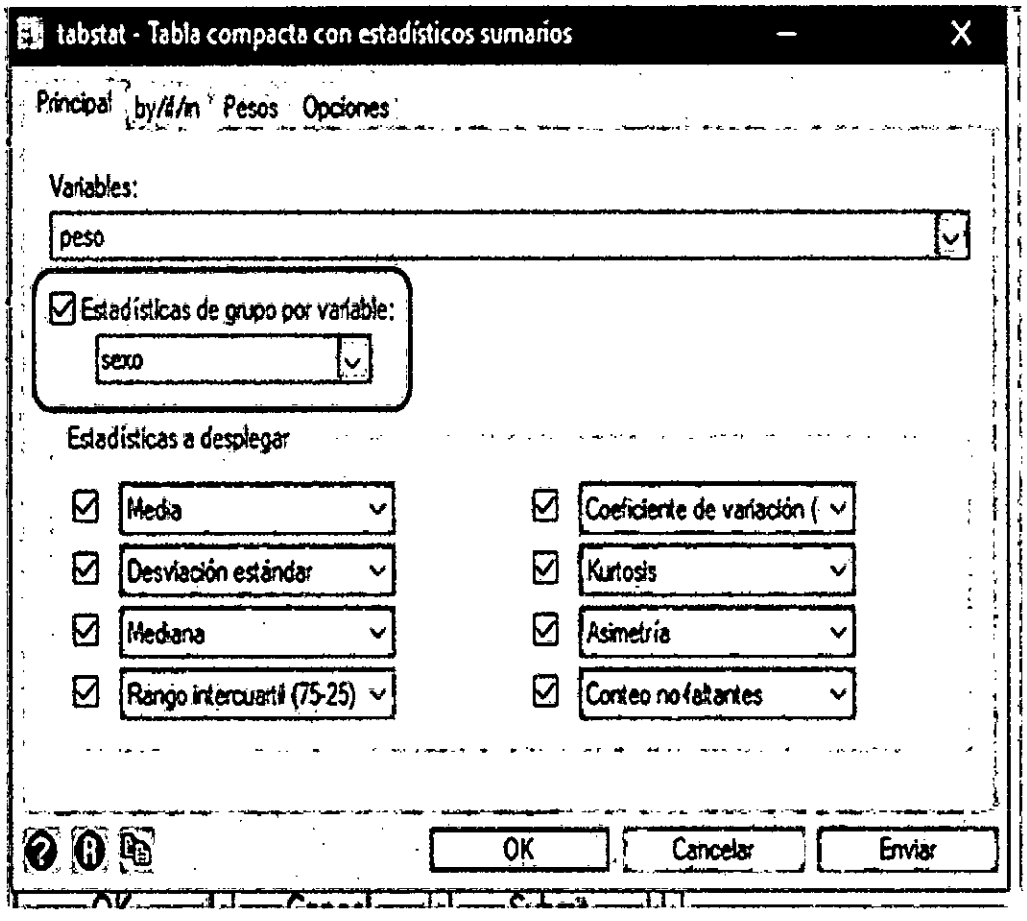


Obtenemos la siguiente salida

```
. tabstat peso, statistics( mean sd median iqr cv kurtosis skewness count ) columns(statistics)
```

variable	mean	sd	p50	iqr	cv	kurtosis	skewness	N
peso	3149.933	627.01	3175	851	.199055	4.066858	-.6813775	75

Si deseo obtener el peso de recién nacido según sexo. Agregamos la siguiente opción en la ventana



Luego obtenemos la siguiente salida

```
. tabstat peso, statistics( mean sd median iqr cv kurtosis skewness count ) by(sexo) columns(statistics)

Summary for variables: peso
     by categories of: sexo (sexo)
```

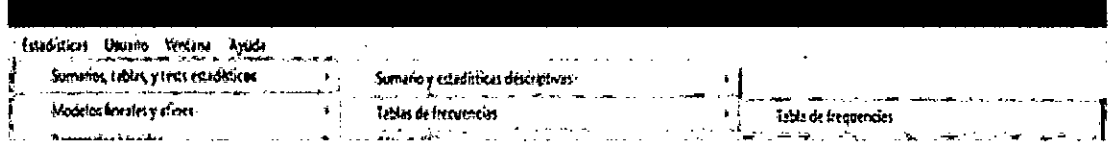
sexo	mean	sd	p50	iqr	cv	kurtosis	skewness	N
femenino	3256.667	509.0509	3345	964	.1560034	1.694817	-.1223441	33
masculino	3066.071	701.2437	3062	794	.2287108	3.928625	-.6749534	42
Total	3149.933	627.01	3175	851	.199055	4.066858	-.6813775	75

Observemos que para obtener las estadísticas descriptivas de la variable cuantitativa (peso) según otra variable en el comando se agregó ***by*** *(variable de agrupación)*

Estadísticas descriptivas para variable cualitativa

Comando ***tabulate***[b]

Otra manera es según el siguiente recorrido



Procedimientos descriptivos con Excel

En el programa Excel, si se selecciona:

Insertar → Función ...

Aparecerá un menú (figura) que ofrece múltiples posibilidades de solicitar índices estadísticos.

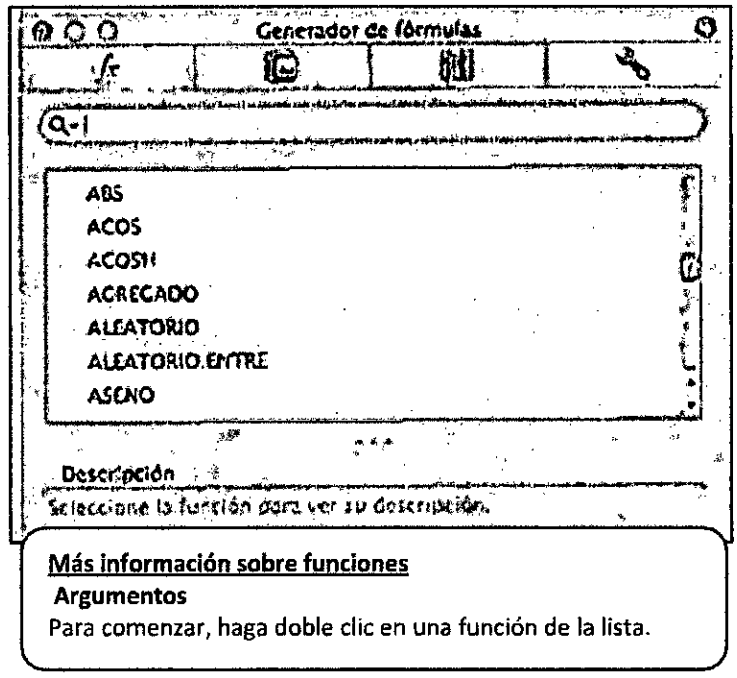


Gráfico 3.3: Menú para seleccionar funciones en Excel. Aparecerá cuando se selecciona: Insertar → Función ...

Cada una de estas funciones viene adecuadamente explicada en las múltiples ayudas y ventanas que ofrece este programa. Para que una función se ejecute se debe escribir su nombre en una silla, pero siempre debe precederse del signo igual (=). Luego, se debe dar una indicación entre paréntesis de cuáles son las casillas en que están situados los datos. Por ejemplo, = **PROMEDIO (11:A9)** significa que se pide la media aritmética de los nuevos datos que ocupan las casillas A1, A2, A3, A4, A5, A6, A7, A7, A8 y A9.

F unciones descriptivas en SPSS

Casi todas las medidas de tendencia central en SPSS están situadas en:

Analizar → Estadísticos Descriptivos

La opción más usada es:

Analizar → Estadísticos Descriptivos → Frecuencias...

Esta opción ofrece un menú, donde se selecciona la variable de interés: por ejemplo, ***edad***. Si luego se pulsa el botón:

Estadísticos...

Aparecerá la siguiente figura en el momento en que se habían seleccionado (cuando se hizo la captura de pantalla) las tres opciones de medidas de tendencia central (media, mediana y moda).

Después se seleccionaron otras (cuartiles, asimetría y curtosis, etc.). finalmente se pulsa:

Continuar → Aceptar

O bien:

Continuar → Pegar

(Esta opción "Pegar" es la adecuada si lo que se desea es seguir trabajando con sintaxis.)

Resumen de las instrucciones es STATA y SPSS

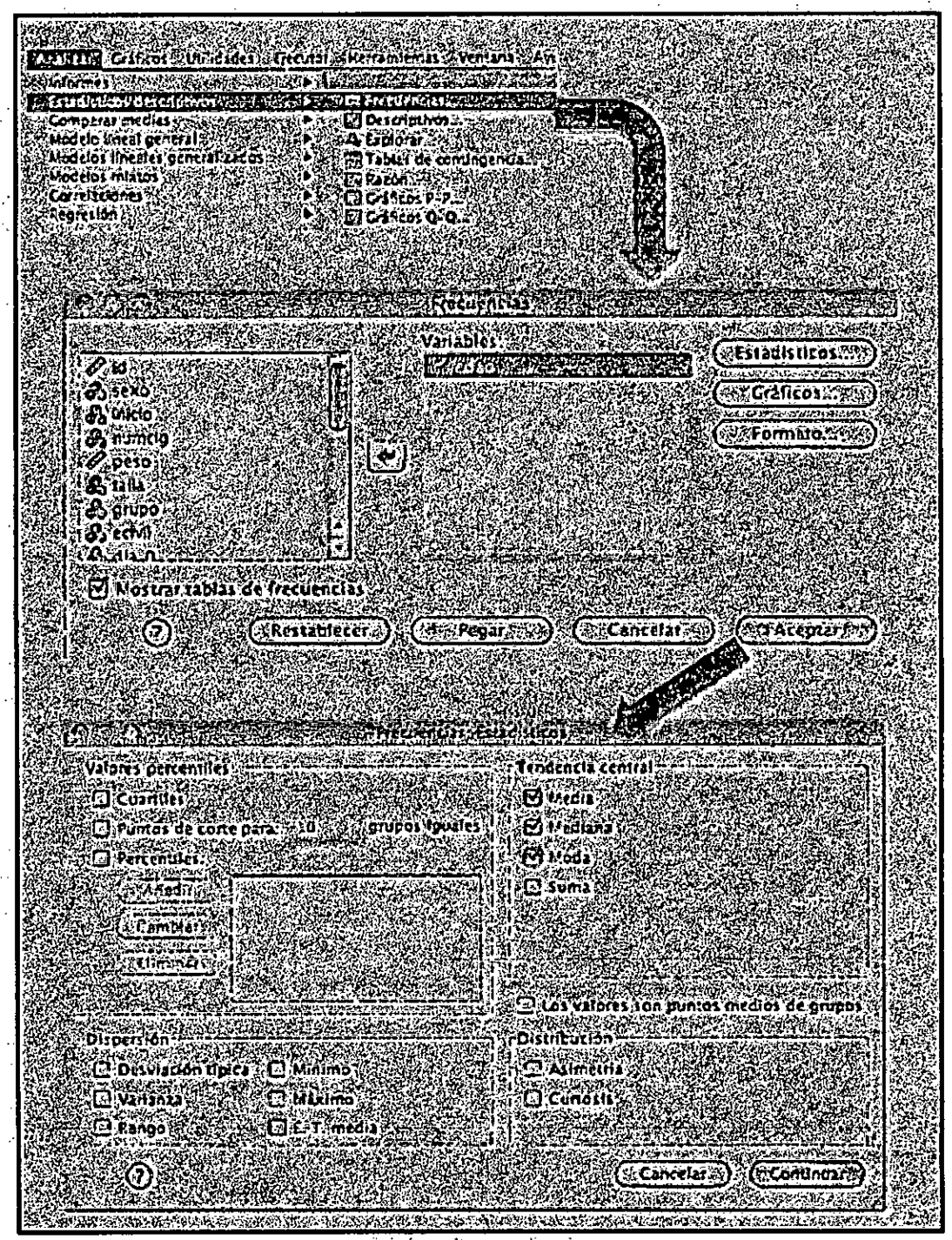


Gráfico 3.4: Estadístico descriptivo con SPSS

CAPÍTULO 4
PRUEBA DE HIPÓTESIS

Una hipótesis es una suposición respecto al problema de investigación, y lo que se ha hace en la prueba de hipótesis es determinar si la proposición es consistente con los datos obtenidos una vez realizada la investigación. Si la hipótesis o proposición no es consistente con los datos obtenidos, se rechaza la hipótesis.

Se usan los datos para intentar rechazar la hipótesis nula y optar por la hipótesis alternativa. Se decidirá entre una y otra. Cuando se rechaza H_0 se dirá que la comparación resultó estadísticamente *significativa* y se concluirá que los datos apoyaban la hipótesis *alternativa*. Las hipótesis (nula o alternativa) se plantean para la población, no para la muestra.

Sin embargo, los datos que se usan en el contraste se obtienen en la muestra. Lamentablemente, el contrate de hipótesis mal usado puede llevar al *automatismo* y acabar por convertirse en un libro de recetas prefabricadas como sucedáneo del raciocinio. Es imprescindible entenderlo bien para que esto no suceda.

Capacidades adquiridas:

- ✓ Conocer los criterios que se requieren para analizar las hipótesis y los resultados del procesamiento de datos obtenidos en la investigación.
- ✓ Saber redactar el marco teórico definitivo del estudio.

4.1 Introducción al contraste de Hipótesis

El contraste de hipótesis se define como el procedimiento estadístico que permite determinar la verdad o falsedad de una afirmación acerca de uno o más parámetros.

Por ejemplo, cuando nos planteamos si existen diferencias entre las medias de dos variables, tenemos que proponer una hipótesis de partida y, por medio del método estadístico más apropiado, concluir en términos probabilísticos, si la hipótesis inicial se rechaza o no se rechaza. En este capítulo vamos a dar una serie de nociones básicas sobre el contraste de

hipótesis necesarias para entender los planteamientos de problemas basados en el contraste

Errores de tipo I y tipo II

Ninguna prueba de hipótesis es 100% cierta. Puesto que la prueba se basa en probabilidades, siempre existe la posibilidad de llegar a una conclusión incorrecta. Cuando usted realiza una prueba de hipótesis, puede cometer dos tipos de error: tipo I y tipo II. Los riesgos de estos dos errores están inversamente relacionados y se determinan según el nivel de significancia y la potencia de la prueba. Por lo tanto, usted debe determinar qué error tiene consecuencias más graves para su situación antes de definir los riesgos.

Error de tipo I

Si usted rechaza la hipótesis nula cuando es verdadera, comete un error de tipo I. La probabilidad de cometer un error de tipo I es α, que es el nivel de significancia que usted establece para su prueba de hipótesis. Un α de 0.05 indica que usted está dispuesto a aceptar una probabilidad de 5% de estar equivocado al rechazar la hipótesis nula. Para reducir este riesgo, debe utilizar un valor menor para α. Sin embargo, usar un valor menor para alfa significa que usted tendrá menos probabilidad de detectar una diferencia si está realmente existe.

Error de tipo II

Cuando la hipótesis nula es falsa y usted no la rechaza, comete un error de tipo II. La probabilidad de cometer un error de tipo II es β, que depende de la potencia de la prueba. Puede reducir el riesgo de cometer un error de tipo II al asegurarse de que la prueba tenga suficiente potencia. Para ello, asegúrese de que el tamaño de la muestra sea lo suficientemente grande como para detectar una diferencia práctica cuando está realmente exista.

La probabilidad de rechazar la hipótesis nula cuando es falsa es igual a 1–β. Este valor es la potencia de la prueba.

	Verdad acerca de la población	
Decisión basada en la muestra	H_0 es verdadera	H_0 es falsa
No rechazar H_0	Decisión correcta (probabilidad = 1 - α)	**Error tipo II** - no rechazar H_0 cuando es falsa (probabilidad = β)
Rechazar H_0	**Error tipo I** - rechazar H_0 cuando es verdadera (probabilidad = α)	Decisión correcta (probabilidad = 1 - β)

Tabla 4.1: Tipo de errores

4.2 Hipótesis estadísticas

Una prueba de hipótesis examina dos hipótesis opuestas sobre una población:

La hipótesis nula: H_0

La hipótesis alternativa: H_1

La hipótesis nula es el enunciado que se probará. Por lo general, la hipótesis nula es un enunciado de que "no efecto" o "no diferencia". La hipótesis alternativa es el enunciado que se desea poder concluir que es verdadero de acuerdo con la evidencia proporcionada por los datos de la muestra.

Con base en los datos de muestra, la prueba determina si se puede rechazar la hipótesis nula. Usted utiliza el valor p para tomar esa decisión. Si el valor p es menor que el nivel de significancia (denotado como α o alfa), entonces puede rechazar la hipótesis nula.

Al diseñar una prueba de hipótesis, esperamos rechazar la hipótesis nula. Antes establecemos el nivel de significancia para que sea pequeño antes del análisis (por lo general, un valor de 0.05), cuando rechazamos la hipótesis nula, tenemos evidencia suficiente de que la hipótesis alterna es verdadera. En cambio, si no tenemos evidencia suficiente no podemos rechazar la hipótesis nula, se concluye que la hipótesis nula es verdadera.

Pasos para realizar una prueba de hipótesis básica

1. Especificar las hipótesis

   Se plantean la hipótesis nula y la hipótesis alterna
2. Se establece un nivel de significancia

   También denominado alfa o α que generalmente se establece como α= 0.05
3. Se decide que estadístico de prueba se va utilizar.

   Las diferentes pruebas de hipótesis utilizan diferentes estadísticos de prueba según el modelo de probabilidad asumido en la hipótesis nula. Las pruebas comunes y sus respectivos estadísticos de prueba incluyen:

Prueba de hipótesis	**Estadístico de prueba**
Prueba Z	Estadístico Z
Prueba t	Estadístico t
ANOVA	Estadístico F
Prueba de chi-cuadrado	**Estadístico de chi-cuadrado**

4. Se establece la regla de decisión

   Si $p<0.05$ se rechaza la hipótesis nula

   Si $p>0.05$ no se rechaza la hipótesis nula
5. Se realizan los cálculos adecuados para determinar el valor de p
6. Decisión estadística, en base a los resultados se rechaza o no se rechaza la hipótesis nula.
7. Conclusión

En base a la evidencia se llega a una conclusión y ésta puede ser estadísticamente significativa

p-valor

Además del valor del estadístico de prueba, los programas estadísticos brindan información del p-valor para tomar la decisión de aceptar o rechazar la hipótesis nula. Cuando la prueba es unilateral con cola o la derecha, este valor se define como el área bajo la curva a la derecha del valor del estadístico de prueba, si la prueba es unilateral con cola a la izquierda, este valor se define como el área bajo la curva a la izquierda del valor del estadístico de prueba, pero si la prueba es bilateral, este valor se define como dos veces el área bajo la curva a la derecha o izquierda del valor del estadístico de prueba en caso de que el valor del estadístico de prueba sea positivo o negativo respectivamente; luego, el p-valor se compara con el nivel de significancia α y se rechaza la hipótesis nula, siempre y cuando, el p-valor sea menor que α.

CAPÍTULO 5

COMPARACIÓN DE MEDIAS ENTRE 2 GRUPOS

La distribución T es una nueva distribución teórica de probabilidad cuyos valores (que se llaman *t*) se interpretan del mismo modo que los valores *z* de la distribución normal. La peculiaridad de la distribución T es que, para cada error α, proporciona un valor de *t* que es distinto *para cada tamaño de muestra.* En cambio, la distribución normal da siempre el mismo valor *z* para cada error α, sea cual sea el tamaño muestral. Cuando el tamaño de muestra es muy grande resulta indiferente usar una u otra, ya que entonces se cumple que $t \approx z$.

Esta distribución fue descrita por W.S. Gosset a principios del siglo xx usando como seudónimo "estudiantes" (Student); este nombre ha perdurado. Al utilizar la *t* para calcular intervalos de confianza para una media, basta saber que los grados de libertad son n – 1 (gl = n -), siendo *n* el tamaño de la muestra.

Capacidades adquiridas:

- ✓ Conocer el procedimiento para recolectar la información necesaria en el desarrollo de la investigación.
- ✓ Conocer y utiliza diferentes herramientas estadísticas para procesar la información obtenida en el trabajo de campo.
- ✓ Organizar y representar los datos en forma tabular y gráfica.
- ✓ Calcular las medidas resumen.
- ✓ Determinar la forma de distribución de los datos.

5.1 Contrastación de hipótesis correspondiente a dos medias poblacionales, Muestras Independientes, Muestras relacionadas, Valor Z, Valor t.

Prueba de hipótesis para la media poblacional

La hipótesis para la media poblacional μ se presentan a continuación:

Hipótesis simple

$$H_o : \mu = \mu_o$$

$$H_1 : \mu =< \mu_o \qquad H_1 : \mu \neq \mu_o \qquad H_1 : \mu > \mu_o$$

Hipótesis compuesta

$$H_o : \mu \geq \mu_o \qquad H_o : \mu \leq \mu_o$$

$$H_1 : \mu < \mu_o \qquad H_1 : \mu > \mu_o$$

En este tipo de prueba se presentan los siguientes casos:

a) **CASO I: Prueba de hipótesis para la media poblacional (μ) cuando la varianza de la poblacional (σ^2) es conocida**

Sea $X_1, X_2, \ldots, X_0$ una muestra aleatoria tomada de una población normal con media μ desconocida y varianza σ^2 **conocida.**

El estadístico de prueba que corresponde es:

$$Z_{cal} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

La región de rechazo se establece a partir de la hipótesis alterna definida y el nivel de significancia dado.

b) **CASO II: Prueba de hipótesis para la media poblacional (μ) cuando la varianza de la población (σ^2) es desconocida y el tamaño de la muestra es menor o igual que 30**

Sea $X_1, X_2, \ldots, X_0$ ($n \leq 30$) una muestra aleatoria tomada de una población normal con media μ desconocida y varianza σ^2 **conocida.**

El estadístico de prueba que corresponde es:

$$T_{cal} = \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

La región de rechazo se establece a partir de la hipótesis alterna definida y el nivel de significancia dado.

c) **CASO III: Prueba de hipótesis para la media poblacional (μ) cuando la varianza de la población (σ^2) es desconocida y el tamaño de la muestra es menor o igual que 30**

Sea $X_1, X_2, \ldots, X_0$ (n >30) una muestra aleatoria tomada de una población normal con media μ desconocida y varianza σ^2 **conocida.**

El estadístico de prueba que corresponde es:

$$Z_{cal} = \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n}}} \sim N(0,1)$$

La región de rechazo se establece a partir de la hipótesis alterna definida y el nivel de significancia dado.

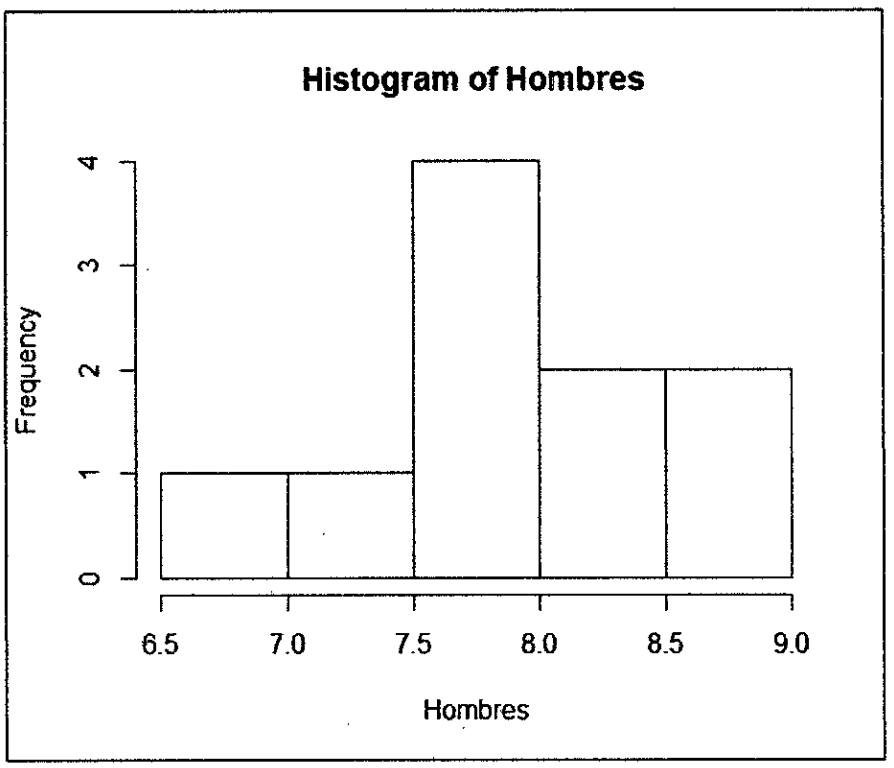
Pruebas paramétricas

t de Student para una muestra

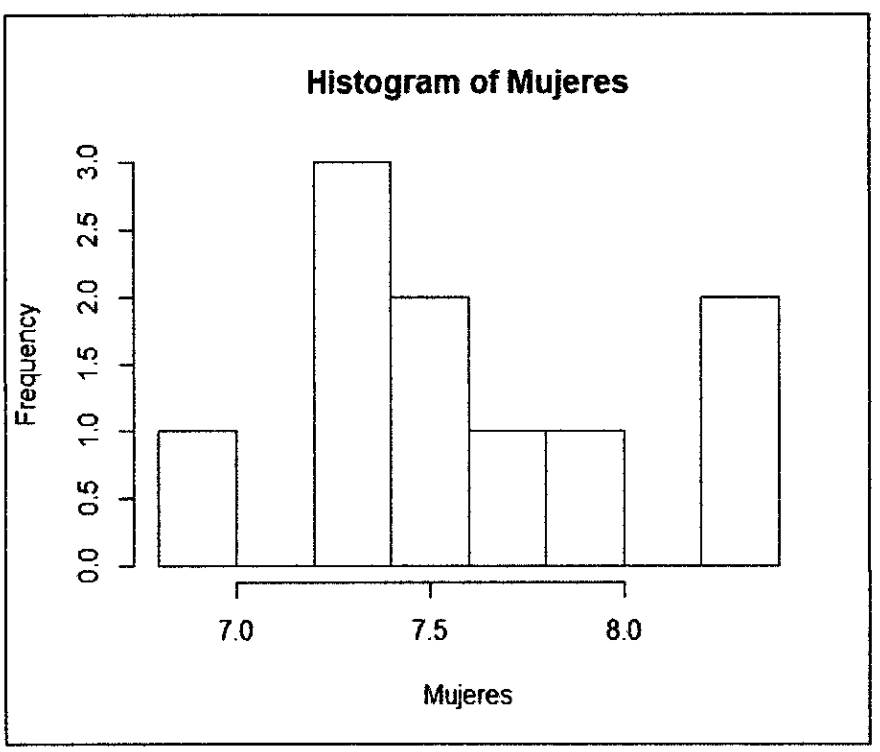
En un estudio se obtuvo 253 observaciones sobre patrones de sueño en estudiantes universitarios. Para ello se realizó un seguimiento de estos alumnos durante dos semanas. La variable *sueno_medio* contiene el número medio de horas de sueño de cada estudiante durante este periodo. En primer lugar, leemos los datos y presentamos un histograma de esta variable:

```
> Hombres=c(7.55,8.57,8.49,7.56,7.38,8.85,7.82,8.13,6.77,7.88)
> Mujeres=c(8.34,7.26,6.95,7.92,7.51,7.58,7.24,7.62,7.37,8.34)
> Hombres
 [1] 7.55 8.57 8.49 7.56 7.38 8.85 7.82 8.13 6.77 7.88
> Mujeres
 [1] 8.34 7.26 6.95 7.92 7.51 7.58 7.24 7.62 7.37 8.34
```

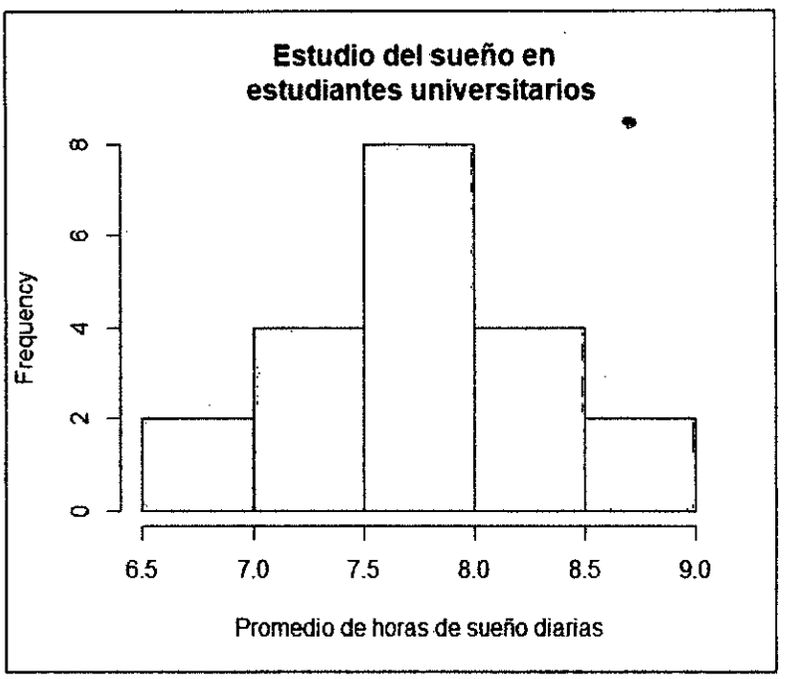
```
> hist(Hombres)
```




```
> hist(Mujeres)
```



```
> Total=c(7.55,8.57,8.49,7.56,7.38,8.85,7.82,8.13,6.77,7.88,8.34,
7.26,6.95,7.92,7.51,7.58,7.24,7.62,7.37,8.34)
> hist(Total, col="lightSalmon", xlab="Promedio de horas de sueño
diarias", main="Estudio del sueño en \n estudiantes universitario
s")
```



Tomemos la variable *Total* como si fuera una sola muestra. Queremos determinar si es admisible la hipótesis de que estos alumnos duermen por término medio 8 horas diarias. Para ello tomamos la variable Total y utilizamos el siguiente comando:

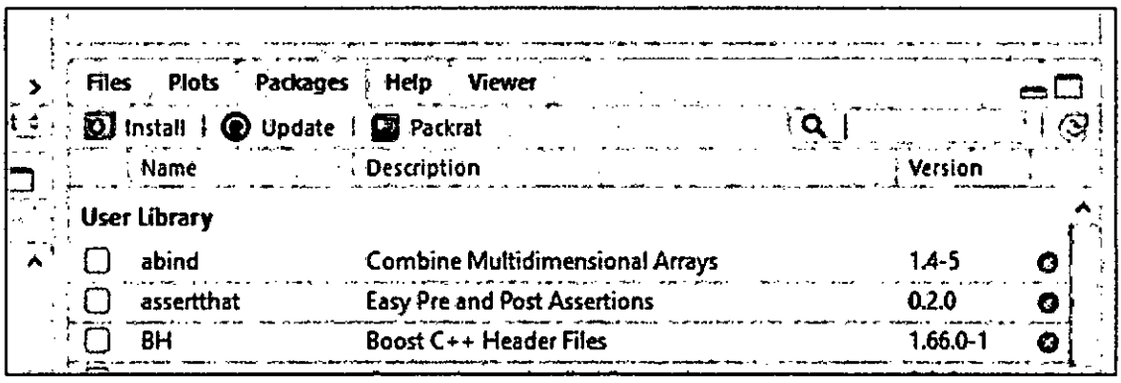
```
> t.test(Total,mu=8)

        One Sample t-test

data:  Total
t = -1.9572, df = 19, p-value = 0.06518
alternative hypothesis: true mean is not equal to 8
95 percent confidence interval:
 7.496101 8.016899
sample estimates:
mean of x
   7.7565
```

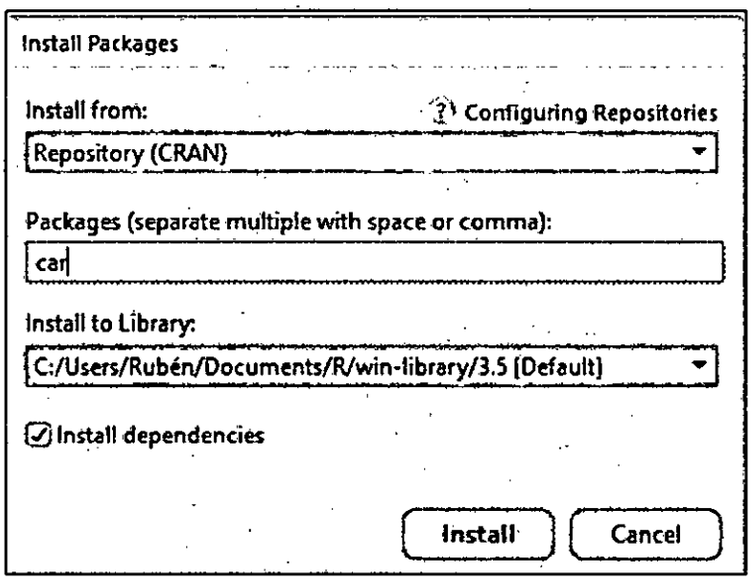
Obsérvese que para llevar a cabo el contraste basta con especificar la media que se desea poner a prueba mediante mu=8. Como resultado del

procedimiento se muestra el valor del estadístico t, sus grados de libertad (df) y el p-valor del contraste (0.06518), que indica que la hipótesis planteada es admisible. Además, obtenemos también la estimación del número medio de horas de sueño en la muestra (7.7565) y un intervalo de confianza al 95%.

Validación de la hipótesis de normalidad

El paquete car proporciona la función qqPlot() que nos permite evaluar gráficamente si puede aceptarse la hipótesis de normalidad de una variable. Para obtener este grafico primero tenemos que cargar el paquete **car** en la ventana inferior derecho dando clic en **Install**



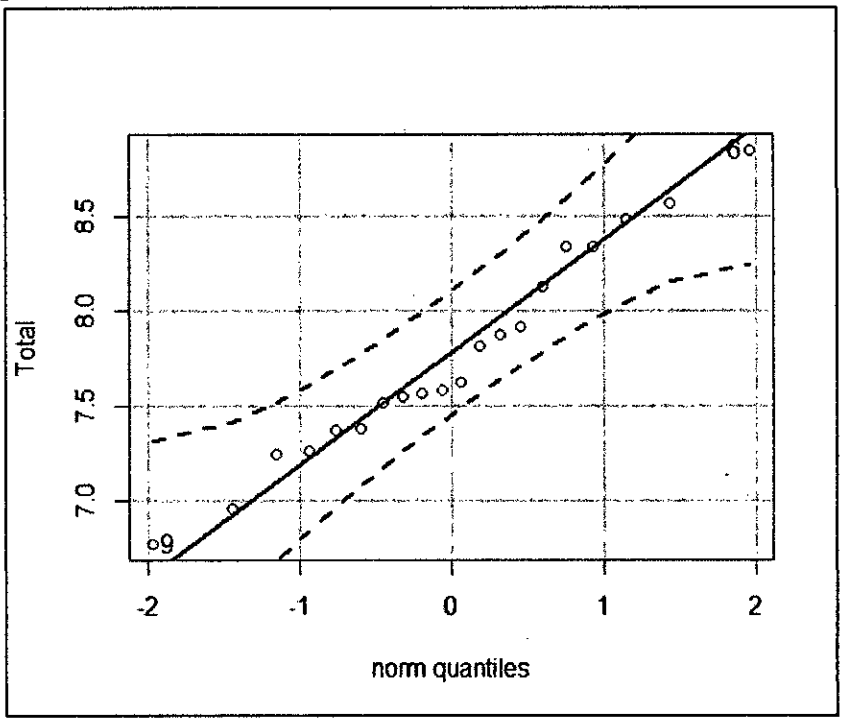


O en forma más sencilla escribiendo el comando

```
> install.packages("car")
```

Una vez descargado el paquete

```
> library(car)
> qqPlot(Total)
[1] 6 9
```



En este caso se aprecia una ligera asimetría en la cola inferior de la distribución. No obstante, el test de Shapiro-Wilk permite aceptar la normalidad de esta variable:

```
> shapiro.test(Total)

        Shapiro-Wilk normality test

data:  Total
W = 0.97182, p-value = 0.7927
```

t de Student para muestras independientes

Comparación de grupos con varianzas distintas. Para contrastar con los datos del estudio anterior si existen diferencias en el promedio de horas de

sueño diarias entre hombres y mujeres, asumiendo varianzas distintas, empleamos la siguiente sintaxis:

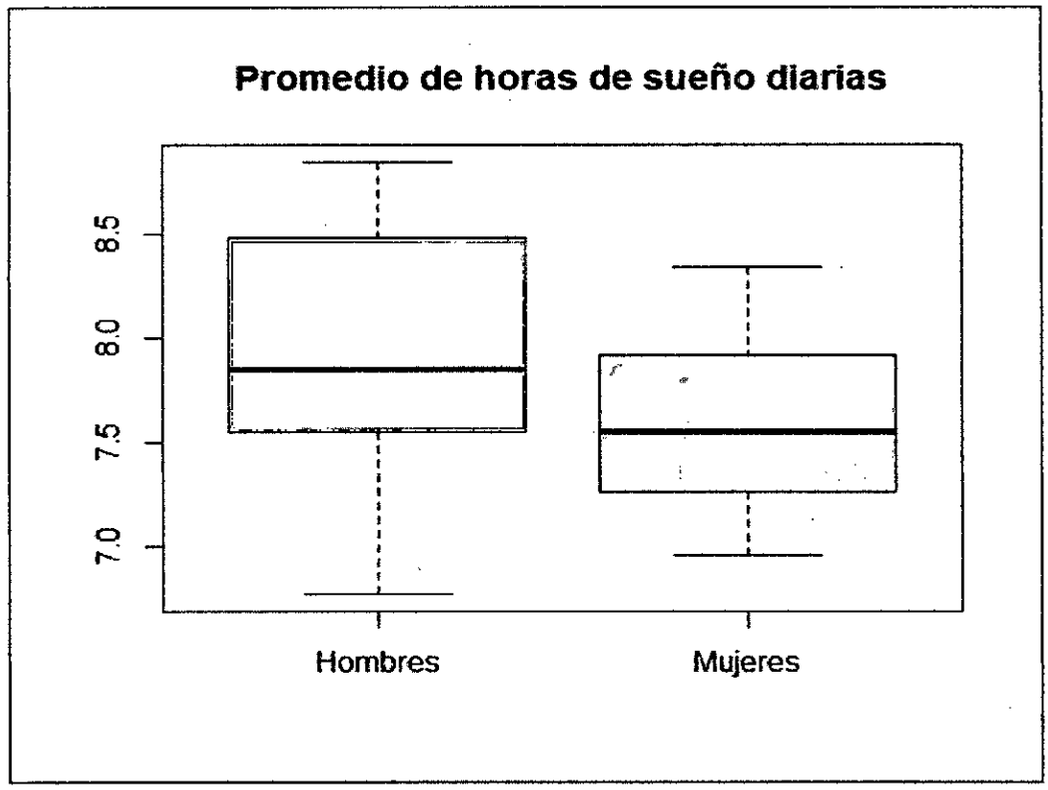
```
> t.test(Hombres, Mujeres)

        Welch Two Sample t-test

data:  Hombres and Mujeres
t = 1.1642, df = 16.55, p-value = 0.2609
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -0.2342163  0.8082163
sample estimates:
mean of x mean of y
    7.900     7.613
```

Como vemos, no existen diferencias significativas entre sexos (p-valor 0.2609).

Además el intervalo de confianza (-0.2342163 0.8082163) contiene al valor cero, una muestra de que no existen diferencias significativas

El boxplot que mostramos a continuación muestra que efectivamente ambos grupos son muy similares:

```
> boxplot(Hombres, Mujeres,names = c("Hombres","Mujeres"),main="P
romedio de horas de sueño diarias",col=c("cyan","lightpink"))
```



Para validar la aplicación del test, comprobamos la normalidad en cada grupo:

```
> shapiro.test(Hombres)

        Shapiro-Wilk normality test

data:  Hombres
W = 0.97464, p-value = 0.9302

> shapiro.test(Mujeres)

        Shapiro-Wilk normality test

data:  Mujeres
W = 0.92467, p-value = 0.3976
```

Comparación de grupos con varianzas iguales. En caso de que queramos especificar que las varianzas son iguales, utilizaríamos la opción *var.equal=TRUE:*

```
> t.test(Hombres, Mujeres, var.equal=TRUE)

        Two Sample t-test

data:  Hombres and Mujeres
t = 1.1642, df = 18, p-value = 0.2596
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -0.2309439  0.8049439
sample estimates:
mean of x mean of y
    7.900     7.613
```

t-Student muestras emparejadas o relacionadas

Para comparar las medias de dos poblaciones en un diseño emparejado podemos utilizar t.test() con la opción paired=TRUE. A modo de ejemplo, vamos a trabajar con un conjunto de datos sobre la influencia de dos tipos de droga sobre las horas de sueño. Este archivo contiene los datos de pulsaciones por minuto de un grupo de 10 estudiantes en dos situaciones: cuando se le administra la droga 1 y cuando se le administra la droga 2.

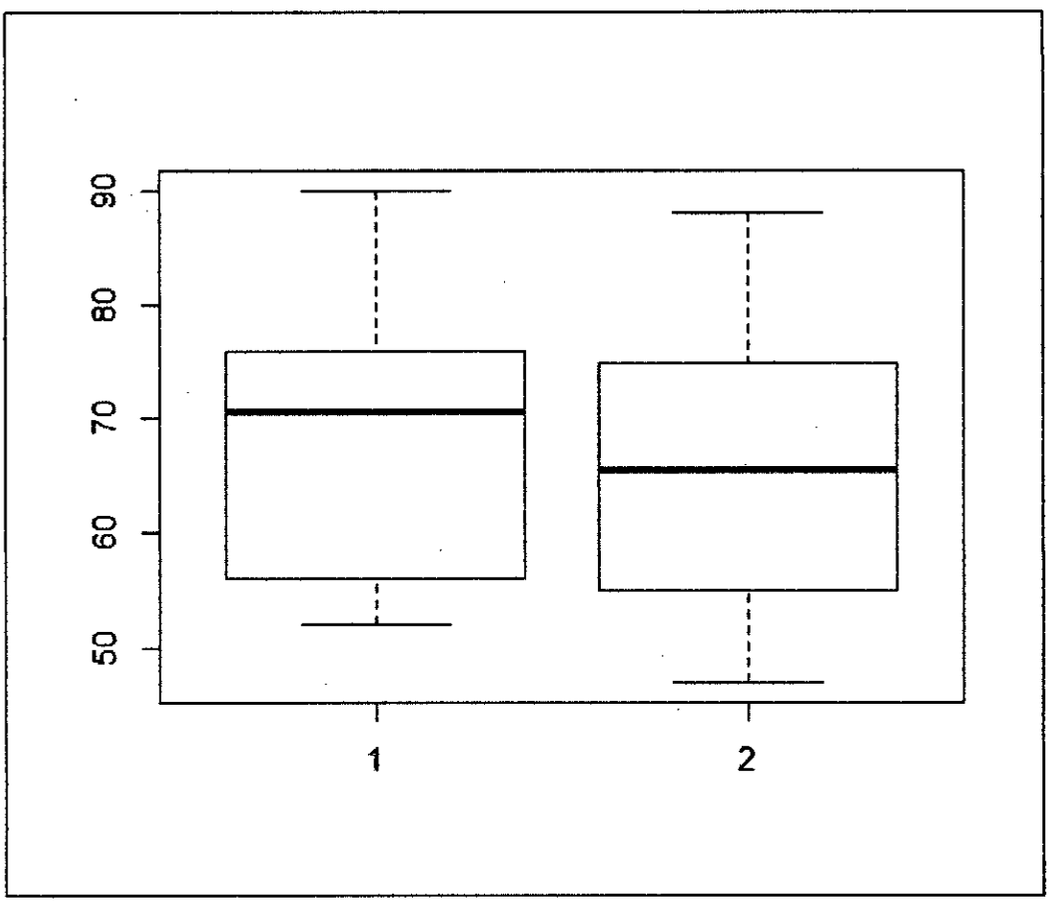
```
##   estudiante droga1 droga2
## 1          1    75     73
## 2          2    52     53
## 3          3    52     47
## 4          4    80     88
## 5          5    56     55
## 6          6    90     70
## 7          7    76     61
## 8          8    71     75
## 9          9    70     61
## 10        10    66     78
```

```
> DrogaA=c(75, 52, 52, 80,56, 90,76, 71, 70, 66)
> DrogaB=c(73,53, 47, 88, 55,70, 61, 75, 61, 78)
> t.test(DrogaA, DrogaB,paired=TRUE)

        Paired t-test

data:  DrogaA and DrogaB
t = 0.85952, df = 9, p-value = 0.4124
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
```

```
 -4.406119  9.806119
sample estimates:
mean of the differences
                    2.7
```

La media de las diferencias es de 2.7, pero esta diferencia no es estadísticamente significativa.

```
> boxplot(DrogaA,DrogaB)
```



Comprobamos la normalidad de los datos

```
> shapiro.test(DrogaA)

        Shapiro-Wilk normality test
data:  DrogaA
W = 0.94359, p-value = 0.5936

> shapiro.test(DrogaB)

        Shapiro-Wilk normality test

data:  DrogaB
W = 0.97225, p-value = 0.9109
```

ANOVA con R Studio

Ejemplo. Una compañía farmacéutica está evaluando tres pastillas para calmar el dolor de cabeza (migraña). Se seleccionaron 27 voluntarios para el experimento y se asignaron aleatoriamente 9 participantes para una de las drogas, formando tres grupos. Se instruyo a los participantes para que tomaran una pastilla cuando se les presentaba la migraña y puntuar el dolor sentido en una escala del 1 al 10 (10 el más doloroso, 0 sin dolor)

Droga A 4 5 4 3 2 4 3 4 4
Droga B 6 8 4 5 4 6 5 8 6
Droga C 6 7 6 6 7 5 6 5 5

Se introdujeron los datos de la manera adecuada, como una data frame

```
> dolor = c(4, 5, 4, 3, 2, 4, 3, 4, 4, 6, 8, 4, 5, 4, 6, 5, 8, 6,
6, 7, 6, 6, 7, 5, 6, 5, 5)
> droga = c(rep("A",9), rep("B",9), rep("C",9))
> migrana = data.frame(dolor,droga)
```

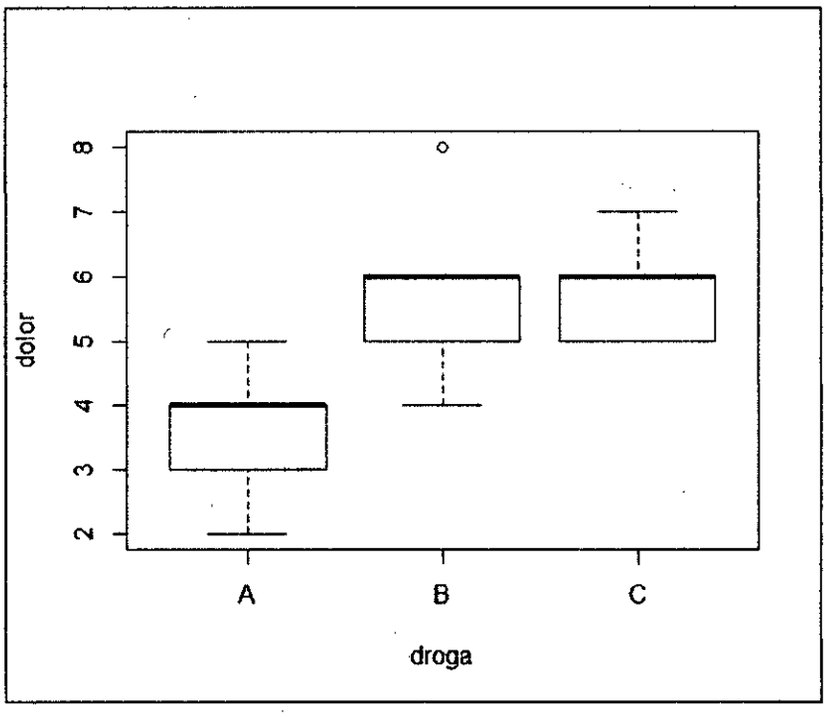
Nótese que el comando rep ("A",9) construye una lista de 9 As en una fila. La variable droga entonces se convierte en una lista de longitud 27 con 9 "A", 9"B" y 9"C". Colocar los datos en un data frame nos permite graficar un diagrama de cajas y realizar el análisis ANOVA

```
> migrana
   dolor droga
1      4     A
2      5     A
3      4     A
4      3     A
5      2     A
6      4     A
7      3     A
8      4     A
9      4     A
10     6     B
11     8     B
12     4     B
13     5     B
14     4     B
15     6     B
16     5     B
17     8     B
18     6     B
```

```
19      6       C
20      7       C
21      6       C
22      6       C
23      7       C
24      5       C
25      6       C
26      5       C
27      5       C
```

Ahora pedimos un gráfico de cajas

```
> plot(dolor ~ droga, data=migrana)
```



Observando el grafico parece ser que la media del dolor para la droga A es más baja que la media para la droga B y C. Ahora utilizaremos la función aov() en R Para ajustar el modelo ANOVA. La forma general es:

aov(response ~ factor, data=data_name)

Donde *response* representa la variable respuesta y factor la variable que separa a los datos en grupos. Ambas variables deben estar contenidas en el data frame denominada data_name. Una vez que ajustamos el modelo

ANOVA, observamos los resultados utilizando la función summary(), el cual produce una tabla estándar de ANOVA

```
> results = aov(dolor ~ droga, data=migrana)
> summary(results)
            Df Sum Sq Mean Sq F value   Pr(>F)
droga        2  28.22  14.111   11.91 0.000256 ***
Residuals   24  28.44   1.185
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La nueva variable ***results*** contiene a partir de ahora nuestro Modelo1 de ANOVA. Observando los resultados de la tabla de ANOVA podemos ver que el valor F es 11.91con un p-value igual a 0.000256. Se rechaza la hipótesis nula de la igualdad de las medias para los tres grupos.

Ya sabemos que existen diferencias significativas entre los tres grupos, pero no sabemos si existen diferencias entre los grupos tomados dos a dos. Para saber eso realizamos un análisis post-hoc.

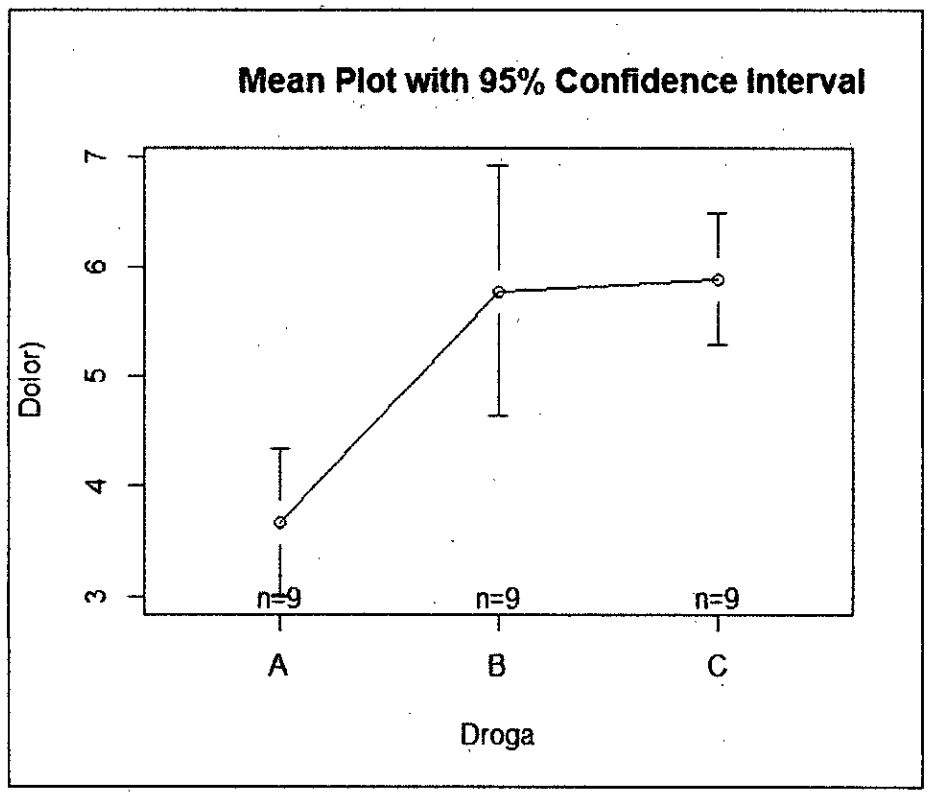
```
> TukeyHSD(results)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = dolor ~ droga, data = migrana)

$`droga`
         diff        lwr      upr     p adj
B-A 2.1111111  0.8295028 3.392719 0.0011107
C-A 2.2222222  0.9406139 3.503831 0.0006453
C-B 0.1111111 -1.1704972 1.392719 0.9745173
```

Se observa que la diferencia entre la droga B y A esa significativa, lo mismo que entre la droga C y A, pero la diferencia entre C y B no es significativa. Vamos a instalar el paquete **gplots** para visualizar mejor

```
> install.packages("gplots")
```

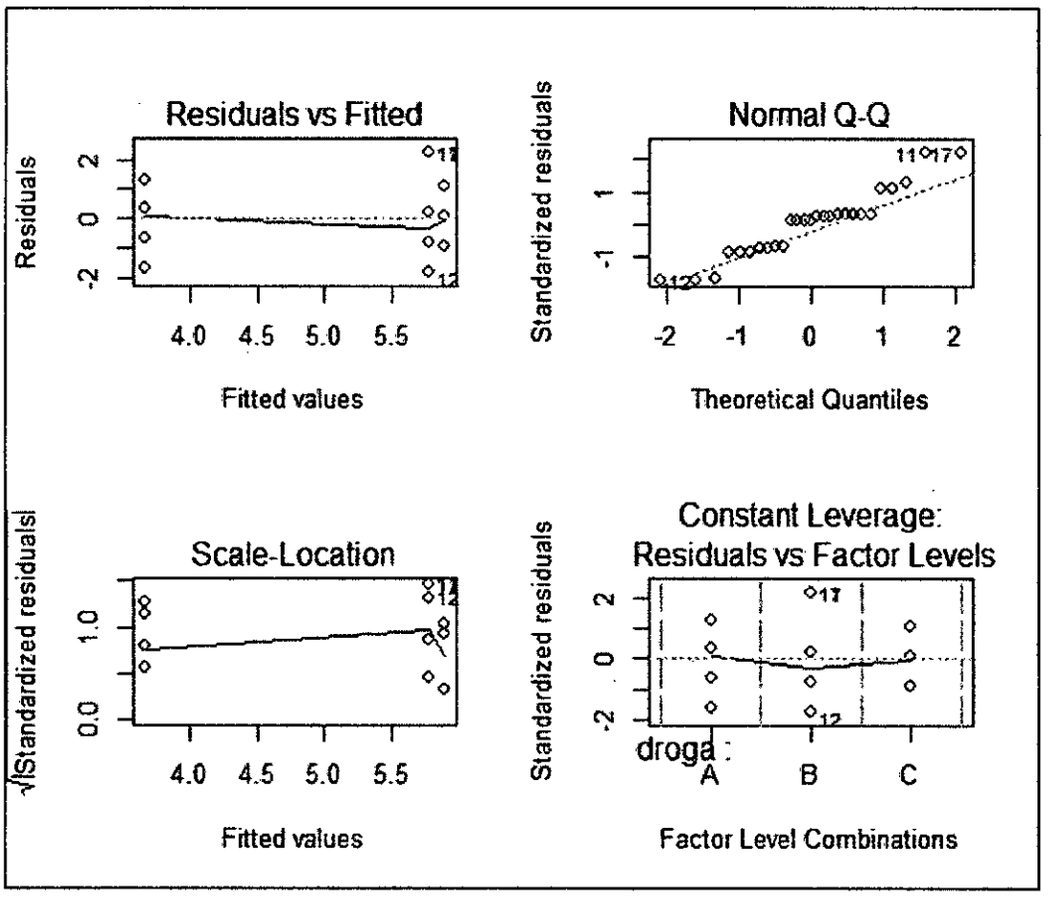
Después de instalarlo ya podemos correrlo

```
> library(gplots)
> plotmeans(dolor ~ droga, main="Fig.-3: Mean Plot with 95% Confi
dence Interval", ylab = "Dolor", xlab = "Droga")
```



Verificando las asunciones del modelo

```
> par(mfrow=c(2,2))
> plot(results)
```



El primer grafico muestra los residuales vs, los valores ajustados. Si los residuales tuvieran un patrón particular, como por ejemplo una línea diagonal, existirían otros predictores no considerados en el modelo. La línea se ajusta bien a un modelo con una sola variable explicadora. La curva Normal Q-Q muestra los cuantiles de los residuales estandarizados y los cuantiles esperados silos datos fueran normales, aparentemente existe normalidad. La grafica Scale-Location nos permite evaluar la homocestacidad. La grafica Residuals vs. Leverage nos indica que no valores extremos que pudieran distorsionar el modelo. Aparentemente el modelo de ANOVA se ajusta bien.

Vamos a constar los supuestos en forma analítica. Primero guardamos lo residuales

```
> uhat<-resid(results)
```

Aplicamos el Shapiro-Wilk para evaluar la normalidad de los residuales

Ho las observaciones de la muestra provienen de una población normal
Hi : las observaciones de la muestra fueron tomadas de una población no normal

```
> shapiro.test(uhat)

        Shapiro-Wilk normality test

data:  uhat
W = 0.93675, p-value = 0.1013
```

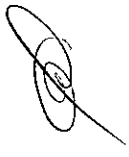
```
La muestra ha sido tomada de una población normal
```

Comprobamos la homogeneidad de la varianza entre los grupos mediante la prueba de Bartlett y Levene

```
> bartlett.test(dolor~droga)

        Bartlett test of homogeneity of variances

data:  dolor by droga
Bartlett's K-squared = 3.8192, df = 2, p-value = 0.1481
```

```
> library(car)
> leveneTest(dolor~droga)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2  1.6667   0.21
      24
Warning message:
In leveneTest.default(y = y, group = group, ...) : group coerced
to factor.
```

La muestra presenta homocestacidad.



5.2 Introducción. Prueba de Kruskall Wallis. Prueba de rangos de Friedman. Chi cuadrado: pruebas de bondad da ajuste, homogeneidad e independencias. Regla de Cochran. Prueba Exacta de Fisher. Muestreo Aleatorio Simple (M.A.S.)

Test de Kruskal-Wallis

El test de Kruskal-Wallis, también conocido como test H, es la alternativa no paramétrica al test ANOVA de una vía para datos no pareados. Se trata de una extensión del test de Mann-Whitney para más de dos grupos. Se trata por lo tanto de un test que emplea rangos para contrastar la hipótesis de que *k* muestras han sido obtenidas de una misma población.

Bajo ciertas simplificaciones puede considerarse que el test de Kruskal-Wallis compara las medianas.

- *H0*: todas las muestras provienen de la misma población (distribución).
- *HA*: Al menos una muestra proviene de una población con una distribución distinta.

Ejemplo

Un estudio compara el número de células cancerosas encontradas en una muestra de tejido bajo 3 condiciones distintas (Droga 1, Droga 2 y Droga 3. ¿Existen diferencias significativas en el número de células cancerosas dependiendo de las condiciones?

Ingresamos los datos de la forma adecuada

```
> datos <- data.frame(condicion = c(rep("Droga1", 18), rep("Droga
2", 18),rep("Droga3", 18)), n_celulas =c(1, 2, 3, 4, 5, 6, 7, 8,
9, 16, 27,28, 29, 30, 51, 52, 53, 66, 40, 41, 42, 43, 44, 45, 46,
47, 48, 67, 88,89, 90, 91, 92, 93, 94, 59, 19, 20, 21, 22, 23, 24
, 25, 26, 27, 28, 25,36, 37, 58, 59, 60, 71, 72))
```

Hemos creado un conjunto de datos en formato data. frame,

```
> datos <- data.frame
```

especificando primero las condiciones: Droga 1 con 18 datos, repetir...

```
condicion = c(rep("Droga1", 18), rep("Droga2", 18),rep("Droga3",
18))
```

Luego hemos establecido el número de células, de los cuales los 18 primeros pertenecen a la condición 1 (droga 1), los 18 siguientes a la condición 2 (Droga 29 y así sucesivamente

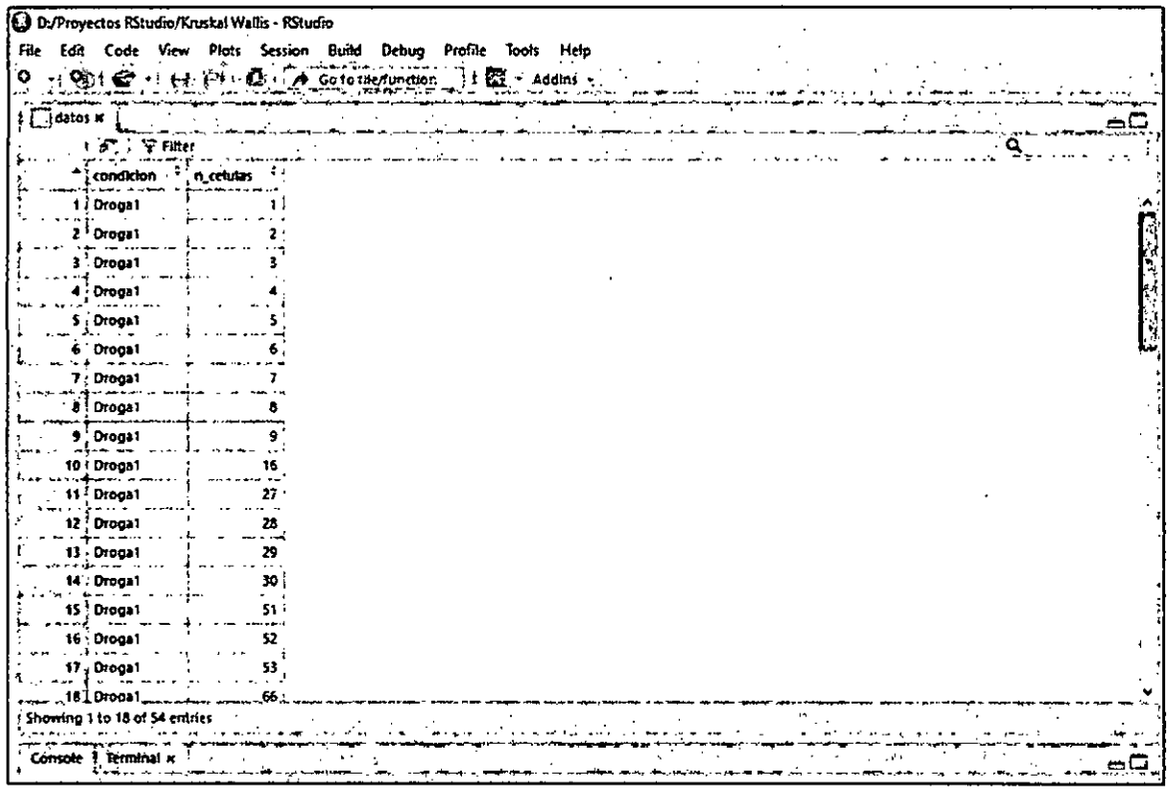
```
n_celulas =c(1, 2, 3, 4, 5, 6, 7, 8, 9, 16, 27,28, 29, 30, 51, 52
, 53, 66, 40, 41, 42, 43, 44, 45, 46, 47, 48, 67, 88,89, 90, 91,
92, 93, 94, 59, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 25,36, 37
, 58, 59, 60, 71, 72))
```

Observamos los primeros elementos de nuestros datos creados

```
> head(datos)
condicion n_celulas
1    Droga1         1
2    Droga1         2
3    Droga1         3
4    Droga1         4
5    Droga1         5
6    Droga1         6
```

O ponemos toda la base de datos en la ventana de Script

```
> View(datos)
```



Generamos datos para la estadística descriptiva

Mediana

```
> aggregate (n_celulas ~ condicion, data = datos, FUN = median)
  condicion n_celulas
1    Droga1      12.5
2    Droga2      53.5
3    Droga3      26.5
```

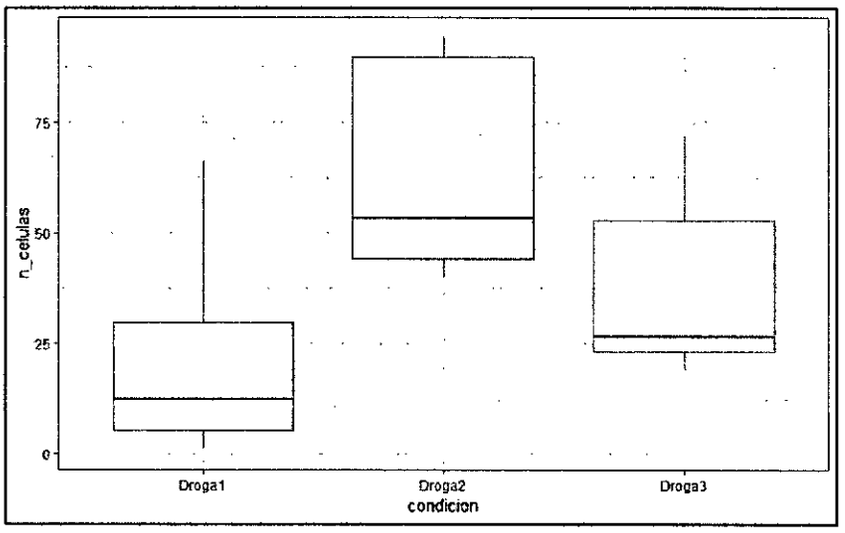
Desviación estándar

```
> aggregate(n_celulas ~ condicion, data = datos, FUN = sd)
  condicion n_celulas
1    Droga1  21.01672
2    Droga2  22.78064
3    Droga3  18.59097
```

Generamos un gráfico de caja y bigote

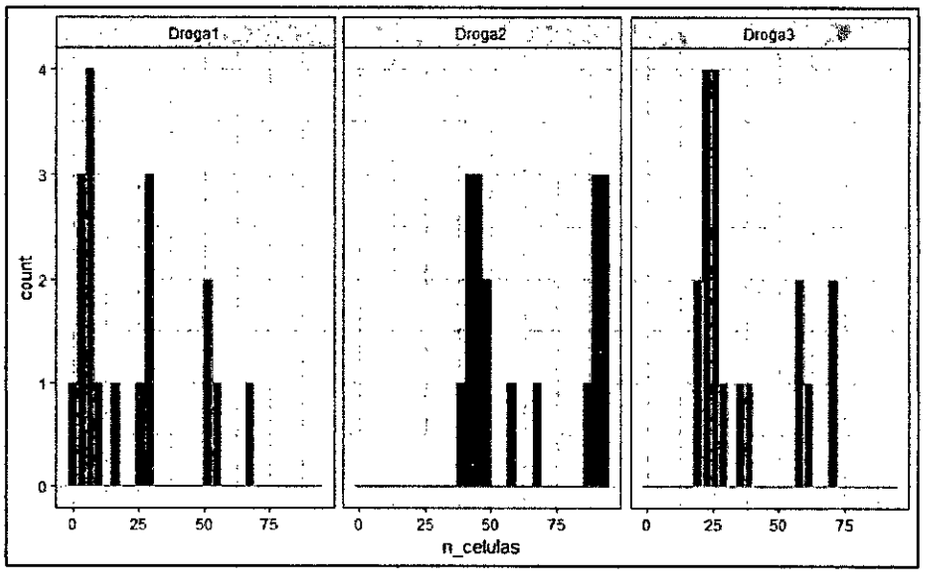
```
> require(ggplot2)
```

```
> ggplot(data = datos, mapping = aes(x = condicion, y = n_celulas
, colour = condicion)) + geom_boxplot() + theme_bw() + theme(lege
nd.position = "none")
```



Evaluamos la normalidad de los datos

```
> ggplot(data = datos, mapping = aes(x = n_celulas, colour = cond
icion)) + geom_histogram() +theme_bw() + facet_grid(. ~ condicion
) + theme(legend.position = "none")
```



La representación gráfica de los datos muestra que las muestras no se distribuyen de forma normal, lo que supone una limitación para emplear

un test ANOVA. Las tres muestras presentan el mismo tipo de distribución, asimetría hacia la derecha, a falta de comprobar la homogeneidad de varianza el test de Kruskal-Wallis es la opción más adecuada.

Condiciones

Homocedasticidad: la varianza debe de ser constante entre todos los grupos.

```
> library(car)
> leveneTest(n_celulas ~ condicion, data = datos, center = "media
n")
Levene's Test for Homogeneity of Variance (center = "median")
      Df F value Pr(>F)
group  2  0.9307 0.4009
      51
```

No hay evidencias en contra de la homogeneidad de varianzas.

Test de Kruskal-Wallis

```
> kruskal.test(n_celulas ~ condicion, data = datos)

        Kruskal-Wallis rank sum test

data:  n_celulas by condicion
Kruskal-Wallis chi-squared = 21.147, df = 2, p-value = 2.559e-05
```

El test encuentra significancia en la diferencia de al menos dos grupos.

Comparaciones post-hoc para saber que dos grupos difieren

Existen diferentes métodos de corrección del nivel de significancia, entre ellos destacan el de *Bonferroni* que es muy estricto y el de *holm*, esteúltimo parece ser más recomendado.

```
> pairwise.wilcox.test(x = datos$n_celulas, g = datos$condicion,
p.adjust.method = "holm")

        Pairwise comparisons using Wilcoxon rank sum test

data:  datos$n_celulas and datos$condicion

       Droga1  Droga2
Droga2 7.2e-05 -
Droga3 0.04124 0.00089
```

```
P value adjustment method: holm
Warning messages:
1: In wilcox.test.default(xi, xj, paired = paired, ...) :
  cannot compute exact p-value with ties
2: In wilcox.test.default(xi, xj, paired = paired, ...) :
  cannot compute exact p-value with ties
```

Las comparaciones por pares encuentran diferencias significativas entre todas las condiciones.

Pruebas no paramétricas

Test de Mann-Whitney

El test de *Mann–Whitney–Wilcoxon* contrasta que la probabilidad de que una observación de la población *X* supere a una observación de la población *Y* es igual a la probabilidad de que una observación de la población *Y* supere a una de la población *X*. Es decir, que los valores de una población no tienden a ser mayores que los de otra.

$$H0:P(X>Y)=P(Y>X)$$

$$H0:P(X>Y)=0.5$$

$$Ha:P(X>Y)\neq P(Y>X)$$

$$Ha:P(X>Y)\neq 0.5$$

condiciones necesarias del test de *Mann–Whitney–Wilcoxon*

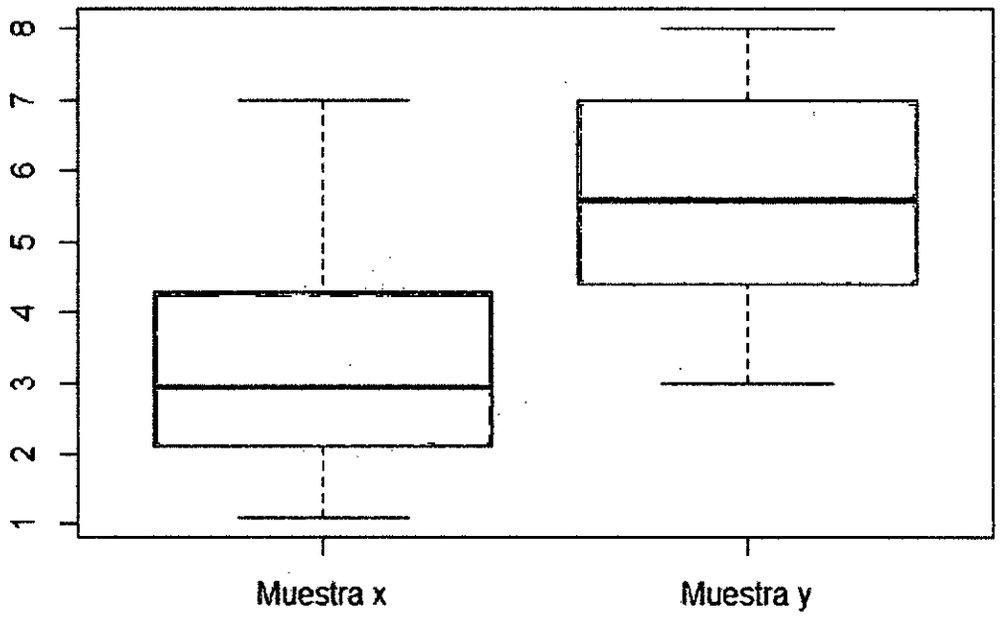
- Los datos tienen que ser independientes.
- Los datos tienen que ser ordinales o bien se tienen que poder ordenar de menor a mayor.
- No es necesario asumir que las muestras se distribuyen de forma normal o que proceden de poblaciones normales. Pero, para que el test compare medianas, ambas han de tener el mismo tipo de distribución (varianza, asimetría...).
- Igualdad de varianza entre grupos (homocedasticidad).

Supóngase que se dispone de dos muestras, de las que no se conoce el tipo de distribución de las poblaciones de origen y cuyo tamaño es demasiado pequeño para determinar si siguen una distribución normal. ¿Existe una diferencia significativa entre poblaciones?

```
> muestraX <- c( 1.1, 3.4, 4.3, 2.1, 7.0 , 2.5 )
> muestraY <- c( 7.0, 8.0, 3.0, 5.0, 6.2 , 4.4 )
```

Generamos un diagrama de caja para observar la distribución de los datos

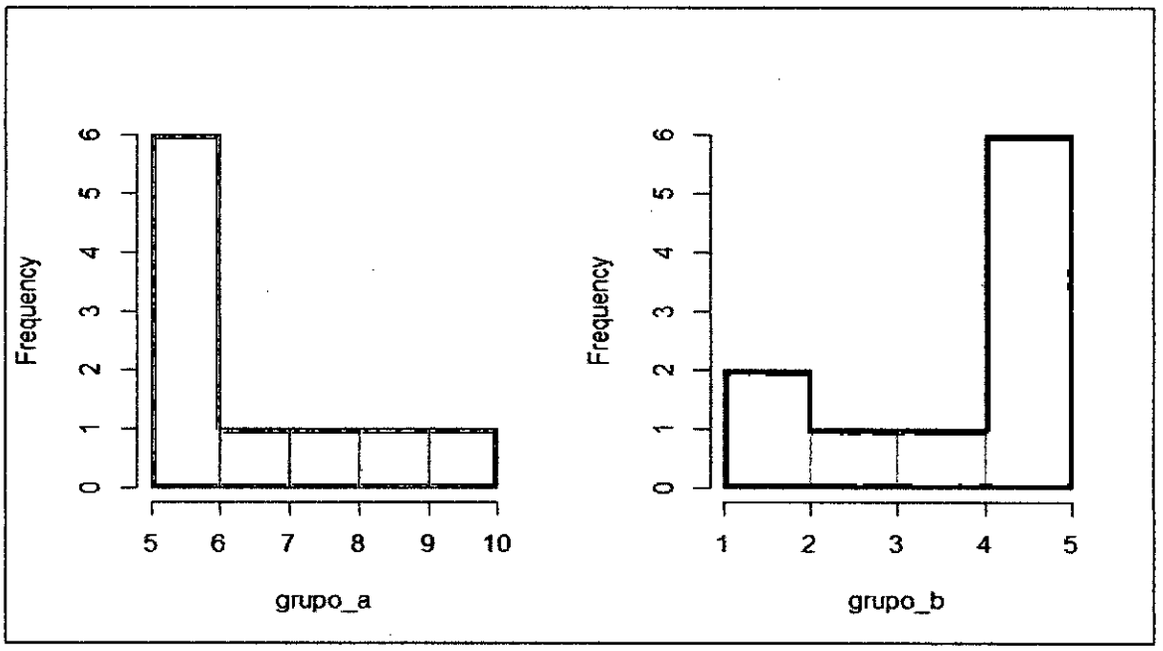
```
> boxplot(muestraX,muestraY,names = c("Muestra x","Muestra y"),co
l=c("blue","green"))
```




```
> median(muestraX)
[1] 2.95
> median(muestraY)
[1] 5.6
```

Pedimos histogramas para observar la normalidad de los datos

```
> par(mfrow = c(1,2))
> hist(grupo_a, col = "cyan4", main = "")
> hist(grupo_b, col = "firebrick", main = "")
```



Para poder aplicar el test de *Mann–Whitney–Wilcoxon* se requiere que la varianza sea igual en los dos grupos. Los test más recomendados para analizar homocedasticidad en estos casos son el test de *Levene* o el test de *Fligner-Killeen*, ambos trabajan con la mediana por lo que son menos sensibles a la falta de normalidad (si se está empleando un test de *Mann–Whitney–Wilcoxon* suele ser porque los datos no se distribuyen de forma normal).

```
> fligner.test(x = list(muestraX,muestraY))

        Fligner-Killeen test of homogeneity of variances

data:  list(muestraX, muestraY)
Fligner-Killeen:med chi-squared = 0.07201, df = 1, p-value = 0.78
84
```

No hay evidencias en contra de la igualdad de varianzas.

Una vez comprobadas las condiciones necesarias para que el test de *Mann–Whitney–Wilcoxon* sea válido se procede a realizar el test de Mann Whitney. R contiene una función llamada wilcox.test() que realiza un test de *Mann–Whitney–Wilcoxon* entre dos muestras cuando se indica que

paired = False y además genera el intervalo de confianza para la diferencia de localización.

```
> wilcox.test(x = muestraX, y = muestraY, alternative = "two.side
d", mu = 0,paired = FALSE, conf.int = 0.95)

        Wilcoxon rank sum test with continuity correction

data:  muestraX and muestraY
W = 6.5, p-value = 0.07765
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -4.9000274  0.4000139
sample estimates:
difference in location
             -2.390035

Warning messages:
1: In wilcox.test.default(x = muestraX, y = muestraY, alternative
= "two.sided",  :
  cannot compute exact p-value with ties
2: In wilcox.test.default(x = muestraX, y = muestraY, alternative
= "two.sided",  :
  cannot compute exact confidence intervals with ties
```

En la salida de vuelta por la función, el valor *W* equivale a *U*.

Cuando hay ligaduras o *ties*, la función wilcox.test() no es capaz de calcular el *p-value* exacto por lo que devuelve una aproximación asumiendo que *U* se distribuye de forma ~ normal. En estos casos, o cuando los tamaños muestrales son mayores de 20 y se quiere la aproximación por la normal, es más recomendable emplear la función wilcox_test() del paquete **coin**.

```
> install.packages("coin")
> library(coin)
> wilcox_test(valores ~ grupo, data = datos, distribution = "exac
t", conf.int=0.95)
        Exact Wilcoxon-Mann-Whitney Test
data:  valores by grupo (A, B)
Z = -1.8447, p-value = 0.06926
alternative hypothesis: true mu is not equal to 0
95 percent confidence interval:
 -4.9  0.4
sample estimates:
difference in location
                 -2.4
```

Resultado

La diferencia entre las probabilidades de que observaciones de una población superen a las de la otra no difiere de forma significativa (*p-value = 0.06926*). El tamaño del efecto observado es grande (0.53) pero no significativo.

Prueba de Friedman con STATA:

La prueba de Friedman es el equivalente no paramétrico del ANOVA de un factor para muestras relacionada, cuando la VD esta medida al menos a nivel ordinal.

Ejemplo Friedman con STATA

La siguiente data corresponde a tres medidas hechas a o individuos, correspondientes a su nivel de ansiedad (medida ordinal) frente a tres circunstancias laborales distintas, a saber, A = amenaza de suspensión, B = amenaza de multa y C = amenaza de despido.

id	A	B	C
1	1	2	3
2	2	3	4
3	1	3	3
4	2	3	4
5	3	3	4
6	1	2	4
7	2	2	3
8	2	2	4

Para correr la prueba Friedman es STATA es necesario descargarla (findit Friedman). Luego se usan los siguientes comandos:

xpose, clear

friedman v1 –v8

Friedman = 0.5000

Kendall = 0.0625

P-value 0 0.4795

El primer comando invierte la base de datos a la forma necesaria para que el comando "Friedman" pueda operar. El resultado p=0.4795 indica que no hay diferencia significativa de los niveles de ansiedad entre grupos según amenaza laboral.

Prueba Chi – Cuadrado

Procedimiento:

(1) Planteamiento de hipótesis:

Hp: El comportamiento aleatorio de la variable en estudio se ajusta a la distribución "A".

Ha: El comportamiento aleatorio de la variable en estudio no se ajusta a la distribución "A".

(2) Elección de α

(3) Cálculo de las frecuencias esperadas. Para obtener las frecuencias esperadas se procede de la siguiente manera:

$e_i = n\,\pi_i$, donde: $n = \sum_{i=1}^{k} e_i = \sum_{i=1}^{k} o$

Donde: π_i = Probabilidad de ocurrencia del intervalo o categoría i

n = Tamaño de muestra

k = Número de intervalos o categorías.

De manera similar a la prueba anterior, si se tienen categorías con frecuencias esperadas menores de 5, se deben agrupar categorías contiguas hasta lograr que las frecuencias sean mayores o iguales a 5.

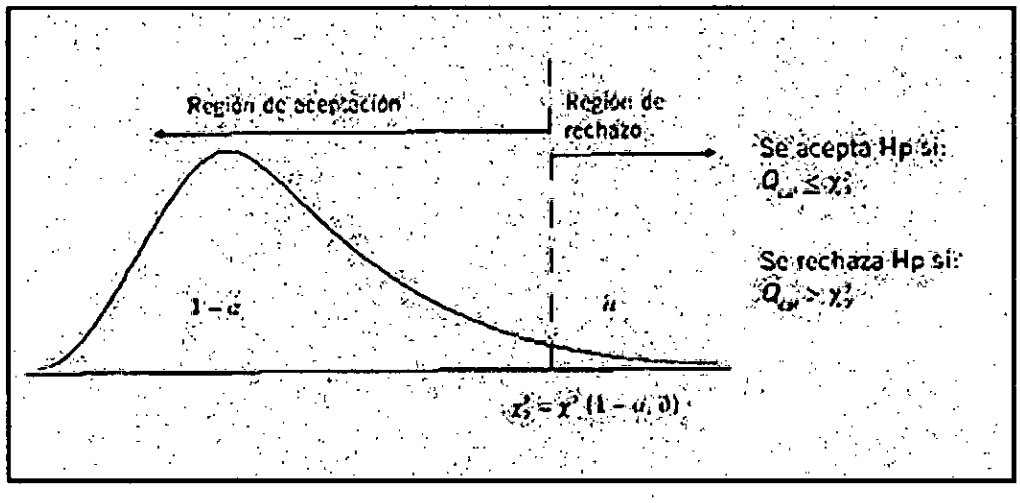
(4) Estadístico de prueba. Considerando que δ=k-m-1 grados de libertad, y siendo m=número de parámetros estimados, se tiene:

$$Q = \sum_{i=1}^{k} \frac{(|o_i - e_i| - 0.5)^2}{e_i} \underset{aprox}{\approx} X^2_{(\delta)}, \; si\; \delta = 1 \;\; \text{y/o} \;\; n<50$$

$$Q = \sum_{i=1}^{k} \frac{(|o_i - e_i| - 0.5)^2}{e_i} \underset{aprox}{\approx} X^2_{(\delta)}, \; si\; \delta = 1 \;\; \text{y/o} \;\; n\geq 50$$

Variable que tiene una distribución aproximadamente chi-cuadrado con δ grados de libertad.

(5) Criterios de decisión



(6) Cálculo del valor Q_{cal}

(7) Decisión y conclusiones

Pruebas de Bondad de Ajuste

En muchas aplicaciones estadísticas es asumir que la variable en análisis tiene un comportamiento probabilístico específico, y en estos casos resulta de interés conocer si es posible considerar que dicho supuesto es o no estadísticamente aceptable.

En general, las pruebas de bondad de ajuste se aplican cuando se desea verificar si una variable en estudio tiene un comportamiento probabilístico que se ajuste o se aproxima a cierta distribución teórica en particular; es decir, sobre la base de cierta información muestral se desea evaluar si es posible concluir que variable en análisis tiene una distribución de

probabilidades que se asemeja a la distribución teórica especificada en la hipótesis que se desea probar.

Diferencias entre las pruebas de Independencia y de Homogeneidad

1. Las pruebas de independencia se usan para probar si dos características son independientes entre sí. Las pruebas de homogeneidad se usan para probar si una característica tiene un comportamiento homogéneo en dos o más localidades.
2. Las pruebas de independencia suponen una sola muestra. Las pruebas de homogeneidad suponen dos o más muestras independientes.
3. En las pruebas de independencia las frecuencias marginales de filas y columnas son valores aleatorios. En las pruebas de homogeneidad, una de las frecuencias marginales (filas o columnas) son valores fijados pues corresponden a los tamaños de muestra de cada localidad.

Regla de Cochran

La prueba Q de Cochran es una técnica estadística utilizada en los diseños de investigación con tres o más muestras dependientes o relacionadas entre so[es decir, la población que se utiliza en la investigación interviene como su propio control. Así, existe un periodo de medición inicial y otros ulteriores, además de que el tipo de escala es nominal.

El valor que se calcula en la prueba Q de Cochran se distribuye igual que la ji cuadrada, por lo que se simboliza X_Q^2.

La ecuación de esta prueba estadística es:

$$X_Q^2 = \frac{(K-1)[\sum Gn^2 - (\sum Gn^2)]}{K\sum L_C - \sum {L_C}^2}$$

Simbología

X_Q^2 Valor estadístico de ji cuadrado en la prueba Q de Cochran.

K Número de tratamientos.

Gn Número total de respuestas de cambio de cada tratamiento columna

Lc Número total de respuestas de cambio por individuos de la muestra o fila.

$\sum$ Sumatoria

Algoritmo

1. Arreglar la muestra individualmente con sus respuestas de cambio.
2. Efectuar las sumatorias de cambios por cada tratamiento o columna (Gn y $\sum$Gn)
3. Efectuar la sumatoria de cambio por cada fila y la sumatoria de los cuadrados del cambio por cada fila ($\sum L_C - \sum {L_C}^2$)
4. Aplicar la ecuación de la prueba Q de Cochran, de modo que se obtenga el valor de X_Q^2
5. Calcular los grados de libertad con K tratamientos -1
6. Comparar el valor estadístico X_Q^2 obtenido, considerando los grados de libertad.
7. Tomar las decisiones de aceptar y rechazar hipótesis.

Prueba Exacta de Fisher

Esta prueba estadística denominada probabilidad exacta de Fischer y Yates se utiliza frecuentemente como alternativa, cuando no se puede aplicar la prueba ji cuadrado de Pearson. Es un procedimiento más potente y eficiente en la escala nominal con dos muestras independientes, ya que se calcula directamente la probabilidad de una serie de arreglos de frecuencias observadas en una tabla de contingencia de 2 x 2, dada una

distribución hipergeométrica. Para calcular es necesario seguir los pasos del algoritmo.

La ecuación para calcular la probabilidad exacta de Fisher y Yates es:

$$p = \frac{(A+B)!\,(C+D)!\,(A+C)!\,(B+D)!}{GT!\,(A!\,B!\,C!\,D!)}$$

Algoritmo

1. Arreglar las frecuencias observadas en una tabla de contingencia de 2 x 2, con el formato que se muestra a continuación:

Grupos	*Variables*		
	X	*Y*	*Total*
Grupo 1	*A*	*B*	*(A + B)*
Grupo 2	*C*	*D*	*(C + D)*
Total	*(A + C)*	*(B + D)*	Gran total: *(GT = A+B+C+D)*

2. Obtener los totales de las filas, de las columnas y el gran total.
3. Obtener los valores factoriales de los totales de las filas y columnas y después multiplicarlos.
4. Calcular el factorial del gran total y multiplicarlo por todos los factoriales de las casillas de la tabla de contingencia.
5. Dividir el primer valor del producto de factoriales entre el segundo. Este resultado es la probabilidad exacta de Fischer y Yates.
6. Tomar las decisiones de aceptar y rechazar hipótesis con base en la probabilidad.

Muestreo Aleatorio Simple (M.A.S.)

Es un método de selección fundamentado en la extracción aleatoria de "n" unidades de una población con "N" unidades de muestreo, de modo tal que cada una de las muestras posibles tiene la misma probabilidad de ser elegida. La selección de los elementos de la muestra se puede realizar mediante los siguientes criterios:

a.1) Selección con reemplazo y probabilidades iguales

Consiste en obtener una muestra mediante "n" extracciones aleatorias sucesivas, restituyendo a la población cada unidad elegida. En este caso cada unidad elegida no puede ser seleccionada posteriormente.

Considerando que en este caso el número de muestras posibles de tamaño "n" que se puede obtener "N^n", entonces la probabilidad de seleccionar una muestra cualquiera es $1/N^n$.

a.2) Selección sin reemplazo y probabilidades iguales

Consiste en obtener una muestra mediante "n" extracciones aleatorias sucesivas, sin restituir a la población cada unidad elegida. En este caso cada unidad elegida no puede ser seleccionada posteriormente.

Considerando que en este caso pueden ser extraídas sin reemplazo C_n^N muestras de tamaño "n", de una población de N elementos, entonces la probabilidad de seleccionar una muestra cualquiera es $1/C_n^N$.

En la práctica, el procedimiento usado con mayor frecuencia para obtener una muestra simple aleatoria es el de una selección sin reemplazo y con probabilidades iguales. Este procedimiento equivale a obtener "n" números aleatorios diferentes entre 1 y "N", con el objetivo de identificar las unidades de muestreo que se deben elegir del marco de muestreo.

CAPÍTULO 6

ESTIMACIÓN DEL TAMAÑO DE MUESTRA

Se suele trabajar con una muestra, no con toda la población. Como no suele ser factible, por motivos prácticos, determinar o medir la característica en todas las personas de la población, se usará solo un subgrupo, que ser denomina muestra para, a partir de ella, describir la población.

Además, esto no supone perder mucha información.

Las muestras representativas escasean. En la vida real ninguna muestra es verdadera y estrictamente representativa de una población. ¿Qué problemas provoca esto? Las consecuencias pueden ser graves cuando el objetivo del estudio es responder a preguntas descriptivas (¿cuál es el colesterol medio en la población?, ¿qué porcentaje de mujeres usan el método sintotérmico?, etc. El objetivo de las investigaciones descriptivas no es realizar comparaciones, sino calcular medias o proporciones. Exigen representatividad.

Capacidades adquiridas:

- ✓ Utilizar las diferentes herramientas estadísticas para calcular tamaño de muestra
- ✓ Recolectar los datos de acuerdo a una muestra representativa
- ✓ Determinar la técnica estadística para contrastar la hipótesis nula.

6.1 Conceptos básicos y generalidades del muestreo

Población diana. Conjunto de elementos o individuos al que hace referencia la pregunta principal u objetivo del estudio. Es la población a la que se desearía generalizar los resultados. Se define principalmente por sus características clínicas y demográficas generales.

6.2 Población objetivo, tipos de muestreo, marco muestral.

Subconjunto de la población diana al que se tiene la intención de estudiar. Se define por los criterios de selección establecidos en el protocolo y presenta determinadas características geográficas y temporales que la hacen accesible a los investigadores.

Muestra. Conjunto de individuos realmente estudiados. En la mayoría de las ocasiones, el número de sujetos necesarios para la realización del estudio es
mucho menor que el de candidatos que forman la población de estudio, por lo que, por razones de eficiencia y disponibilidad de recursos (viabilidad)

Unidad de análisis. Es el objeto sobre el cual se realiza una medición. Esta es la unidad básica de observación, a veces llamado elemento. En los estudios de poblaciones humanas, con frecuencia ocurre que las unidades de observación son los individuos.

Marco muestral. Es la lista de las unidades de muestreo. Por ejemplo, si queremos hacer una encuesta telefónica a los médicos residentes, el marco muestral podría ser una lista de todos los números telefónicos de los médicos residentes, para las entrevistas personales, una lista del nombre y el hospital donde laboran los médicos; para encuesta sobre la situación estructural de los hospitales, una lista de todos los hospitales o un mapa donde se señalan los hospitales.

Cálculo del tamaño de la muestra

Para calcular el tamaño de la muestra debe conocerse:

• **La *variabilidad* del parámetro que se desea estimar**. Si no se conoce, puede obtenerse una aproximación a partir de datos propios o de otras

investigaciones, o un estudio piloto. En el caso de las variables cuantitativas se mide por la varianza, y en el de las cualitativas, por el producto P (1 – P).

• **La *precisión***. con que se desea obtener la estimación, es decir, la amplitud del IC. Cuanto más precisa se desee, más estrecho deberá ser este intervalo, y más sujetos deberán ser estudiados. La precisión debe fijarse previamente, en función de la finalidad de la estimación. En algunos casos puede requerirse una gran precisión, mientras que en otros, si sólo se necesita conocer aproximadamente entre qué valores se encuentra el parámetro, se requerirá una menor precisión y,
consecuentemente, menos sujetos.

• **El *nivel de confianza***. Que habitualmente se fija en el 95%, correspondiente a un valor a de 0,05. Indica el grado de confianza que se tiene de que el verdadero valor del parámetro en la población se sitúa en el intervalo obtenido. Cuanta más confianza se desee, menor será el valor de a, y más elevado el número de sujetos necesario.

Cálculo de la muestra

Para estimar la proporción poblacional (Variable Cualitativa)

a. Cuando no se conoce N

$$n_o = \frac{Z^2(p)(1-p)}{E^2}$$

b. Cuando la población se conoce N

$$n_o = \frac{N\,Z^2 p(1-p)}{(N-1)E^2 + Z^2 p(1-p)}$$

Para estimar promedios poblacionales (Variable Cuantitativa)

a. Cuando no se conoce N

$$n_o = \frac{Z^2 S^2}{E^2}$$

$$n_o = \frac{N\, Z^2 S^2}{(N-1)E^2 + Z^2 S^2}$$

b. Cuando la población se conoce N

Tipos de muestreo

1. Muestreo probabilístico Los métodos de muestreo probabilísticos son aquellos que se basan en el principio de equiprobabilidad. Es decir, aquellos en los que todos los individuos tienen la misma probabilidad de ser elegidos para formar parte de una muestra y, consiguientemente, todas las posibles muestras de tamaño n tienen la misma probabilidad de ser seleccionadas. La elección se hace al azar. Sólo estos métodos de muestreo probabilísticos nos aseguran la representatividad de la muestra extraída y son, por tanto, los más recomendables. Los tipos de muestreo probabilístico más utilizados son:

TIPO	DESCRIPCION
Simple	✓ Útil en poblaciones pequeñas. Se requiere marco muestral. ✓ Puede usarse Tabla de números aleatorios o por sorteo.
Sistemático	✓ Útil en poblaciones grandes. No requiere marco muestral ✓ Se usa un intervalo = Tamaño Población / Tamaño Muestra (y nos puede resultar cada 5, cada 10, cada 20, etc)
Estratificado	✓ Divide a la población en estratos (sub grupos) y de cada uno de ellos elige una muestra aleatoria simple. ✓ Puede estratificarse por edad, sexo, gravedad, por servicios hospitalarios, por grupos profesionales, etc.
Conglomerados	✓ Frecuente en investigaciones comunitarias donde un área geográfica debe separarse en áreas o conglomerados. ✓ Puede conglomerarse por Región, provincia, distrito, comunidades, manzanas, familias, etc.

aleatorio simple, aleatorio sistemático, aleatorio estratificado y aleatorio por conglomerados.

2. **Muestreo no probabilísticos.** A veces, para estudios exploratorios, el muestreo probabilístico resulta excesivamente costoso y se acude a métodos no probabilísticos, aun siendo conscientes de que no sirven para realizar generalizaciones (estimaciones diferenciales sobre la población), pues no se tiene certeza de que la muestra extraída sea representativa, ya que no todos los sujetos de la población tienen la misma probabilidad de ser elegidos. En general se seleccionan a los sujetos siguiendo determinados criterios procurando, en la medida de lo posible, que la muestra sea representativa. En algunas circunstancias los métodos estadísticos y epidemiológicos permiten resolver los problemas de representatividad aun en situaciones de muestreo no probabilístico, por ejemplo, los estudios de casocontrol, donde los casos no son seleccionados aleatoriamente de la población. Los tipos de muestreo no probabilístico más utilizados son: accidental, de conveniencia, por cuotas y por bola de nieve.

TIPOS	DESCRIPCION
Accidental	• Utilizado en enfermedades raras con incidencia muy baja (Ejm: Enfermedad de Von Reklinhaussen)
Voluntario	• Cuando los sujetos de estudio se someten voluntariamente al trabajo de investigación
Intencional (por conveniencia)	• Cuando el investigador decide a quien o quienes investigar (Ejm: Los 10 primeros, etc)

6.3 Condiciones de una buena muestra

Una de las preguntas que se plantean con mayor frecuencia al iniciar una investigación, y que también es de las más difíciles de contestar, sobre todo por falta de información acerca del problema, es: "¿Cuántas observaciones se deben obtener para que el tamaño de la muestra sea realmente representativo del universo estadístico? En este sentido es necesario considerar que las muestras varían en su composición de una a otra. La magnitud de la variación depende del tamaño de la muestra y de la variabilidad original de la población. Así, el tamaño de la muestra queda determinado por el grado de precisión que se desea obtener y por la variabilidad inicial de la población.

6.4 Muestreo Sistemático

En este tipo de muestreo, las unidades son ordenadas en forma sucesiva, de tal forma que la población se pueda dividir en pequeños intervalos. El punto de partida en el primer intervalo es elegido aleatoriamente, para luego extraer los siguientes elementos de manera ordenada y sistemática.

Ejemplo:

El jefe de control de calidad de una empresa textil que produce pantalones realizó un estudio para evaluar los defectos más frecuentes en el proceso de fabricación de estos. La compañía tiene una línea de producción de 1 000 pantalones diariamente, y en un día determinado se ha fijado el tamaño de la muestra en 90 unidades. Entonces, el intervalo de selección de cada elemento de la muestra será de **11 en 11** (1 000/90=11.1). El primer elemento de la muestra se selecciona aleatoriamente entre los 11 primeros pantalones producidos (por lo obtenido anteriormente). De suponer que el primer elemento es el quinto de la lista, entonces el segundo será el décimo sexto (5+**11**=16); el tercero, el vigésimo séptimo (16+**11**=27); y así sucesivamente hasta completar los 90 pantalones.

6.5 Muestreo Estratificado

Este tipo de muestreo se recomienda cuando la población se divide en grupos, denominados estratos; de tal manera que los estratos sean heterogéneos, y los elementos dentro de los estratos, homogéneos.

Ejemplo:

Se desea conocer el tiempo promedio semanal dedicado al deporte en jóvenes de 15 a 29 años de edad en las áreas rural y urbana de un departamento de la sierra; por lo tanto, para estimar dicho promedio se procederá al muestreo estratificado debido a que, dentro de cada área, los jóvenes presentan características similares y, a la vez, las características de los jóvenes de las áreas rural y urbana son diferentes. Si se estableció el tamaño de muestra en 400 jóvenes, y teniendo en cuenta que en este departamento el 70% de la población joven (entre 15 y 29 años) vive en el área urbana, se considerará una muestra conformada por 280 jóvenes del área urbana y 120 jóvenes del área rural, los mismos que participarán en el estudio.

6.6 Muestreo por Conglomerados

Este tipo de muestreo se recomienda cuando la población se divide en grupos, denominados conglomerados; de tal manera que los conglomerados sean homogéneos, y los elementos dentro de los conglomerados, heterogéneos.

Ejemplo:

En un distrito de Lima se estimará el porcentaje de electores que favorecen al candidato A, minutos después de la hora final del proceso de votación.

La información para estimar el porcentaje será recopilada a partir de una encuesta a boca de urna. Se considerará a cada mesa de votación como un conglomerado, ya que cada mesa posee electores con diversas preferencias y, a su vez estas mesas son homogéneas entre sí.

CAPÍTULO V.

REFERENCIAS BIBLIOGRÁFICAS

1. TIPACTI ALVARADO, César y FLORES RODRÍGUEZ, Néstor. **Metodología de la Investigación en Ciencias Neurológicas. Oficina Ejecutiva de Apoyo a la Investigación y Docencia Especializada.** Perú. Editorial Imprenta Unión. Primera edición. 2012.
2. MENDOZA BELLIDO, Waldo. **Cómo Investigan los Economistas. Guía para elaborar y desarrollar un Proyecto de Investigación.** Perú. Fondo Editorial de la Pontificia Universidad Católica del Perú. 2016.
3. EYSSAUTIER DE LA MORA, Maurice. **Metodología de la Investigación. Desarrollo de la inteligencia.** México. Editorial Edamsa Impresiones, S.A. de C.V. Quinta edición. 2008.
4. DEL CID, Alma; MÉNDEZ, Rosemary y SANDOVAL, Franco. **Investigación. Fundamentos y Metodología**. México. Editorial Pearson Educación. Segunda edición. 2011.
5. ARGIMON, JM. JIMÉNEZ VILLA, J. **Métodos de Investigación Clínica y Epidemiología.** Barcelona, España. Editorial Elsevier. Cuarta Edición. 2013
6. HERNÁNDEZ, R. FERNÁNDEZ, C., Baptista, P. **Metodología de la Investigación.** Editorial McGraw Hill. Sexta Edición. 2014
7. POLIT DF, HUNGLE BP. **Investigación Científica en Ciencias de la Salud: Principios y Métodos.** México. Editorial McGraw Hill. Sexta Edición. 2000.
8. PINEDA EB, DE ALVARADO E.L. **Metodología de la Investigación.** Organización Panamericana de la Salud. 2008
9. TAMAYO y TAMAYO M. **El Proceso de la Investigación Científica.** Editorial Limusa. Cuarta Edición. 2001

10. MARTÍNEZ GONZÁLEZ, Miguel Ángel, SÁNCHEZ VILLEGAS, Almudena, TOLEDO ATUCHA, Estefanía A., FAULIN FAJARDO, Javier. **Bioestadística Amigable.** Barcelona, España. Editorial Elseiver. Tercera Edición. 2014
11. MILLONES, Rosa, BARRENO, Emma, VÁSQUEZ, Féliz, CASTILLO, Carlos. **Estadística Descriptiva y Probabilidades.** Perú. Fondo Editorial. Primera Edición. 2018
12. MONTESINOS RUIZ, Luis, LLANOS MIRANDA, Kelva, CERNA FIGUEROA, Edwin, PAJUELO ROJAS, Silvia, COAQUIRA NINA, Frida. **Estadística Descriptiva e Inferencial.** Perú. Fondo Editorial Universidad San Ignacio de Loyola. Primera Edición. 2017
13. TOMA INAFUKO, Jorge, RUBIO DONET, Jorge Luis. **Estadística Aplicada.** Perú. Universidad del Pacífico. Primera Parte. Segunda Edición. 2017.
14. TOMA INAFUKO, Jorge, RUBIO DONET, Jorge Luis. **Estadística Aplicada.** Perú. Universidad del Pacífico. Segunda Parte. Segunda Edición. 2017.
15. GARCÍA ORÉ, Celestino. **Estadística Descriptiva y Probabilidades.** Perú. Editorial Macro EIRL. Primera Edición. 2017
16. PACHECO CONTRERAS, Johnny. **Guía Práctica Gestión de Datos Gráficos y Tablas Dinámicas con Excel.** Perú. Editorial Macro EIRL. Primera Edición. 2015
17. CASTILLA SERNA, Luis. **Manual práctico de Estadística para las Ciencias de la Salud.** México. Editorial Trillas. Primer Edición. 2011.
18. BERNAL TORRES, César. **Metodología de la Investigación. Administración, economía, humanidades y ciencias sociales.** Colombia. Editorial Pearson. Cuarta Edición. 2016
19. ALTMAN DG, Bland JM. **Stratistics notes: variables and parameters. BMJ** 1999; 318(7199):1667
20. GREENHAGH T: **Stratistic for the non-statistician. I: Different types of data need different statistical tests.** Bmj 1997;315(7104):364-6

21. CANGA N, De Irala J, VARA E. Duaso MJ, FERRER A, MARTÍNEZ-GONZÁLEZ MA. **Intervention study for smoking cessation in diabetic patients: a randomized controlled trial in both clinical and primary care settings. Diabetes Care** 2000;23(10):1455-60
22. CA Paul, AU R, FREDMAN L, MASSARO JM, Seshadri S, Decarli C, et al. **Association of alcohol consumption with brain volume in the Framingham study**. Arch Neurol 2008;65(10):1363-7
23. GREENLAND S. **Analysis of polytomous exposures and outcomes.** En: Rothman KJ, Greenland S.Lash TL, editors. Modern Epidemiology. 3rd ed. Philadephia: Lippincott Williams & Wilkins; 2008. P. 303-4
24. JOLLEY D. **The glitter of the t table. Lancet** 1993; 34(8862):27-9
25. ALTMAN DG, Bland JM. Detecting skewness from summary information. BMJ 1996; 313(7066):1200
26. MARTÍNEZ – GONZÁLES MA, GARCÍA-ARELLANO A, TOLEDO E, SALAS-SALVADÓ J, BUIL-COSIALES P, CORELLA D, et al. A 14-item **Mediterranean diet assessment tool and obesity indexes among high – rish sbjects: the PREDIMED trial. PLoS One** 2012;7(8):e43134
27. ROTHMAN KJ, GREENLAND S, LASH T. **Modern Epidemiology.** 3rd ed. Philadephia: Lippinctt Williams & Wilkins; 2008.
28. GOODMAN SN. Toward **Evidence-Based Medical Statistics.** 1: The P Value Fallacy. Ann Intern Med 1999;130(12):995-1004
29. MA, Hernan, JM. Robins. **Causal Inference. Boca Raton**: CRC Press; 2014. (en prensa).
30. SCHULZ KF, GRIMES DA. **Sample size calculation in randomised trials: mandatory and mystical.** Lance 2005;365:1348-53

VI. APENDICES

Medidas de tendencia central y dispersión variables cuantitativas dependientes de los polifenoles totales (PT) Flavonoides totales (FT) Actividad antioxidante (AA). Catequina (CAT) y Epicatequina (EPICAT)

Tabla 1. Medidas de resumen de los polifenoles totales, flavonoidestotales, actividad antioxidante, Catequina y Epicatequina según fermentación

Fermentación	Variable	Media	DS	Mediana	Min	Max
3.5 días	PT	4.11	0.87	3.98	2.72	5.60
	FT	3.66	0.80	3.52	2.61	5.24
	AA	0.02	0.00	0.02	0.01	0.02
	CAT	27.9	5.14	26.4	22.7	39.9
	EPICAT	242	56.7	237	158	342
4.5 días	PT	3.40	0.75	3.29	2.35	4.93
	FT	3.52	0.65	3.55	2.35	4.38
	AA	0.02	0.00	0.02	0.01	0.03
	CAT	25.0	4.35	24.7	19.7	31.0
	EPICAT	193.70	38.5	190.84	125.19	247.45
5.5 días	PT	3.12	0.43	3.13	2.39	3.76
	FT	2.69	0.29	2.66	2.10	3.14
	AA	0.03	0.01	0.03	0.02	0.03
	CAT	24.1	4.25	22.8	20.0	32.8
	EPICAT	189	25.5	186	147	235

En los PT: 41.1± 0.87; FT: 340± 0.80; CAT: 27,90± 5.14; EPICAT: 242.18±56.78 en 3.5 días de fermentación presentan los mayores niveles promedios esperados para la variedad proveniente en la zona del Cusco observando que ha mayor tiempo (días) de fermentación de los niveles promedios van disminuyendo.

Para la AA: 0.03±0.01presentan un ligero mayor en el nivel promedio para 5 días de fermentación

Tabla 2. Medidas de resumen de los polifenoles totales, flavonoides totales, actividad antioxidante, Catequina y Epicatequina según secado

Secado	Variable	Media	DS	Mediana	Min	Max
	PT	3.66	0.93	3.47	2.35	5.60
	FT	3.17	0.59	3.10	2.35	4.26
Gradual	AA	0.02	0.00	0.02	0.02	0.03
(0)	CAT	24.56	3.70	23.99	19.75	30.24
	EPICAT	197.81	46.27	188.05	125.19	304.92
	PT	3.42	0.66	3.31	2.39	4.93
	FT	3.40	0.87	3.13	2.10	5.24
Completo	AA	0.02	0.01	0.02	0.01	0.03
(1)	CAT	26.83	5.53	25.74	20.00	39.93
	EPICAT	218.90	48.35	222.31	147.64	342.42

En los PT: 3.66± 0.93 presenta el mayor nivel promedio esperado para el secado gradual, observando una disminución ante el secado de exposición completa al sol.

Para el secado de exposición completa al sol se observa un mayor nivel promedio en FT: 3.40± 0.87; CAT: 26.83±20; EPICAT: 218.90±147.64 comparada con el secado gradual.

Para la Actividad antioxidante AA: 0.02±0. se mantiene constante ante el secado.

Tabla 3. Medidas de resumen de los polifenoles totales, flavonoides totales, actividad antioxidante, Catequina y Epicatequina según proceso

Proceso	Variable	Media	DS	Min	Max
1(A)	PT	3.34	0.52	2.60	4.04
	FT	3.19	0.76	2.35	4.53
	AA	0.02	0.01	0.01	0.03
	CAT	26.73	2.40	23.26	29.98
	EPICAT	238.97	30.83	210.75	304.92
2(B)	PT	3.98	0.95	2.92	5.60
	FT	3.73	0.91	2.75	5.24
	AA	0.02	0.00	0.02	0.03
	CAT	23.15	3.40	20.55	30.24
	EPICAT	202.91	35.07	171.02	253.75
3(C)	PT	3.44	0.91	2.35	5.22
	FT	3.19	0.55	2.61	4.25
	AA	0.02	0.00	0.01	0.03
	CAT	26.06	7.10	19.75	39.93
	EPICAT	211.65	67.68	147.64	342.42
4(D)	PT	3.43	0.70	2.39	4.50
	FT	3.05	0.56	2.10	3.73
	AA	0.02	0.01	0.02	0.03
	CAT	26.83	4.39	20.92	32.88
	EPICAT	179.90	32.22	125.19	226.84

En el proceso 2/(B) de 110°C y 25 min se presenta el valor más alto PT: 3.98± 2.92; FT: 3.73± 2.75; observando que a mayor tiempo de tostado los promedios van disminuyendo.

En el proceso 1/(B) de 110°C y 20 min se presenta el valor más alto de CAT: 26.73± 23.26; EPICAT: 238.97±210.75 observando que a mayor tiempo (25min) de tostado los valores promedios van disminuyendo.

Para la actividad antioxidante permanece casi constante en valore de AA: 0.022±0.

VII. ANEXOS

Cálculo del tamaño de la muestra

Ejemplo N° 1: Variable cualitativa

Se quiere llevar a cabo un estudio sobre el nivel de autoestima que tienen los pacientes con enfermedades de Parkinson que acuden a los consultorios externos del Instituto Nacional de Ciencias Neurológicas (INCN) entre los meses de julio y diciembre del 2011. Según la Oficina de Estadística del INCN se espera que el número de pacientes con diagnóstico de Parkinson que llegarán en ese período de tiempo es de 410 pacientes. Con tal propósito se desea extraer una muestra que tenga una confianza del 95% y un error estimado de 5%; se asume p = 0.5, debido que no se conoce el porcentaje de autoestima en pacientes con diagnóstico de Parkinson.

Solución

Los datos que se tiene son:

N = 410

p = 0.5

q = 0.5

$Z_{a/2}$ = 95% = 1.96

e = 0.05

La fórmula empleada, por ser la variable principal cualitativa, será:

$$n = \frac{Z_{a/2}^2 * p * q * N}{(N-1) * e^2 + Z_{a/2}^2 * p * q}$$

$$n = \frac{1.96^2 * 0.5 * 0.5 * 410}{(410-1) * 0.05^2 + 1.96^2 * 0.5 * 0.5}$$

n = 199