Escuela Superior de Física y Matemáticas

ESTADÍSTICA

Autor:
Roberto Carlos Santos Alonzo

Febrero 2021

Contenido

Capítulo 1

ESTADÍSTICA DESCRIPTIVA

1.1. INTRODUCCIÓN.

La importancia que estudiemos estadística, es porque es una ciencia que en tiempos reciente ha sido de gran apoyo a otros campos del conocimiento y de gran apoyo, justamente para los avances tecnológicos científicos y humanos que hemos tenido. Un problema estadístico se involucra con el estudio de alguna característica asociada a un grupo de objetos. Al grupo de objetos en cuestión, suele conocerse como, **unidad experimental**

Ejemplo 1.1.1

En la actualidad, el feminicidio ha crecido en forma descontrolada. Si se desea saber la edad promedio de las mujeres que han fallecido por este problema, mediante un estudio estadístico, entonces la ***unidad experimental*** *involucrada en el estudio sería* ***las mujeres que han fallecido por este problema***

Con frecuencia los periódicos contienen reportes de**estudios estadísticos**

Ejemplo 1.1.2

El lunes 2 de mayo de 2005 la mayoría de los diarios del país publicaron la siguiente noticia:
En la Ciudad de México, hay el mayor número de personas infectadas con VIH-SIDA
Esta información fue proporcionada por un especialista, quien completó la información diciendo que cerca de 8 millones de habitantes de la Ciudad de México registran más de la cuarta parte de casos con esa enfermedad
Un dato preocupante en este estudio indica que alrededor del $10\,\%$ *de la gente infectada por el virus de inmunodeficiencia humana en México no sabe de su condición, por lo que no se puede controlar la epidemia*

En la actualidad, existen muchos organismos, sean nacionales o internacionales, que están realizando estudios estadísticos con fines comparativos, así como para evaluar el avance y desarrollo de un país

Ejemplo 1.1.3

En las publicaciones de los estudios que lleva a cabo la Organización para la Cooperación y el Desarrollo Económico (OCDE), se reportó que México ocupa los siguientes lugares

- *La última posición de 21 países miembros de la OCDE en una tabla que describe el gasto en educación por estudiantes;*
- *La última posición de los 27 países en una tabla que describe el número de investigadores por cada mil empleados,y;*

- *La segunda posición entre 28 países de una tabla que describe el porcentaje de la población que padece obesidad:*

Definición 1.1.1 (ESTADÍSTICA)

La estadística es una rama de las matemáticas que trata del análisis e interpretación de un conjunto de datos.

Definición 1.1.2 (FENÓMENO ALEATORIO)

Es un fenómeno que tiene más de un posible resultado, que no se puede predecir y que depende del azar

Definición 1.1.3 (FENÓMENO DETERMINISTA)

Es un fenómeno que tiene más de un posible resultado, que no se puede predecir y que depende del azar

Algunas vivencias que nos son familiares pueden ayudarnos a comprender los conceptos de **población** y **muestra**.

Ejemplo 1.1.4

- *Cuando vamos a un laboratoiro para que nos practiquen un análisis de sangre a fin de evaluar nuestro estado de salud, sólo toman una muestra de sangre y no toda la sangre, es decir,**la población**.Por supuesto, se sabe lo que ocurriía si se sacara toda la sangre, por lo que con una muestra se tendrá una buena aproximación al diagnóstico del estado de salud.*

- *Cuando queremos comprar nueces en un tianguis, le pedimos al vendedor que nos permita tomar una **muestra**. No dejamos que él nos dé la muestra, y sacamos del costal un par de nueces*

- *Así también, cuando pedimos una probadita de algún producto, como un helado, el vendedor nos dará un poco de helado con una cucharita para que decidamos si lo compramos o no.Lo que nos dan es una muestra, pues el vendedor no nos dará todo el bote de helado, que es la **población***

PROBLEMA 1.1.0.1 (MEJORAR UNA MEDICINA)

Elaborar medicinas más eficientes para curar una gripe es una tarea que involucra a muchas personas, tales como a los bioquímicos que la formula, a los dueños de laboratorios que la producen, a los médicos que las prescriben y a los enfermos que las toman

- *La idea es considerar una **población** de personas mayores de 16 años y enfermas de gripe.*

- *Se observará el **tiempo** en días que tarda una persona en restablecerse después de haber iniciado un tratamiento*

Preguntas sobre la naturaleza del problema

- *Por nuestra experiencia todos sabemos que una gripe dura varios días.*

- *En muchos casos, las personas acuden a medicinas que se venden sin receta, y por uso y costumbre se evalúa cuál es más efectiva.*

- *Pero otras siguen algún tratamiento médico.*

¿cómo valoramos la eficiencia de un medicamento?

- *Una respuesta sencilla a esta pregunta es saber en cuántos días se alivian las personas de los síntomas*
- *También se puede averiguar si con el tratamiento seguido disminuyeron sus molestias*

Obtener información

¿ Cómo conseguir esta información?

- *Nuestra meta es producir información sobre el efecto de un tratamiento para curar la gripe. En principio, se sabe que el mercado día a día incorpora nuevos medicamentos para disminuir los síntomas y malestares de esa enfermedad.*
- *Un procedimiento natural para generar información en este caso, es aplicar un tratamiento usando la medicina X a un grupo de personas con gripe. Con el fin de ampliar nuestro conocimientos sobre el tema, se anotará si disminuyeron las molestias, si el medicamento produjo sueño o no. Sin embargo, el reto está aún presente y se plantean las siguientes preguntas:*

  1. ***¿ Cómo se planea realizar el estudio***
  2. ***¿ Qué personas recibirán seguimiento?***
  3. ***¿ Cómo medir el efecto de una medicina?***
  4. *Si sólo a unas cuantas personas se les da la medicina y se observan los resultados,* ***¿ éstos serán igualmente válidos para más personas?***

Procedimiento para recabar información

1. *Identificar a la* ***población*** *que será objeto de estudio*
   La población está constituida, en sentido amplio, por individuos u objetos. A éstos se les ***observa*** *algún atributo, el cual es la finalidad de cualquier estudio. En resumen, estas observaciones generan un conjunto de* ***datos*** *y se convierten en a información que se requiere para conocer sobre un tema específico*

En resumen, estas observaciones generan un conjunto de **datos** y se convierten en al información que se requiere para conocer sobre un tema específico. En ocasiones, los datos son producto de resultados **experimentales**. En muchas áreas de aplicación, las **poblaciones** son fáciles de identificar, en otras, no es tan sencillo.**La población debe estar bien definida antes de iniciar un estudio**

En el problma 1.1, la población de búsqueda fueron las personas enfermas de gripe. Pero, **¿ Cómo la definimos?**
Se considera a las personas con gripe y la población puede limitarse a una escuela, una colonia, una delegación o un hospital.
Una vez definida la **población**, el siguiente paso es obtener una **muestra** de la población.

Definición 1.1.4 (Muestra)

- *Una muestra consiste en seleccionar una parte de la población para realizar el estudio*
- *La muestra es una parte (subconjunto) de la población que se estudiará para conocer las características de la población*

A continuación se realiza un análisis de la información generada a partir de la muestra.

¿Por qué tomar una muestra?

Se ha dicho que la población representa el todo y la muestra es sólo una parte de la población. Un pregunta que surge sobre ellos es **¿ Por qué procurar examinar una muestra cuando lo que realmente se desea es estudiar la población.?** La mayoría de las veces no se puede estudiar a la población, por lo que podemos usar a muestra como una guía.

- Reducir el tiempo de estudio. Podría tomar mucho tiempo estudiar una población.
- Reducir los costos de estudio. Resulta muy costoso, en cuanto a esfuerzo, estudiar una población.
- En ocasiones, es difícil identificar a todos los miembros de una población
- Si estudiamos a toda la población no debemos dejar a alguien fuera del estudio, por lo que en resumen diremos que:

Definición 1.1.5 (POBLACIÓN)

Una población consiste en una colección de individups u objetos de interés a los que se les observa una característica particular que será objeto de estudio.

En investigación científica se le define como la totalidad de elementos sobre los cuales recae la investigación. A cada elemento se le llama ***unidad estadística****, ésta se le observa o se le somete a una experimentación, estas unidades son medidas pertinentemente. Es el conjunto total de nuestro estudio (el universo total que se va analizar).*
Si representamos mediante X, una variable aleatoria bajo investigación, al estudiar esta variable en la población, como resultado tendemos valores:

$$\{X_1, X_2, \ldots, X_N\}$$

Donde ***N*** *es el total de elementos de la población El propósito de un estudio estadístico es extraer conclusiones acerca de la naturaleza de la población, pero resulta que las poblaciones son grandes o por razones de ética, recursos financieros, metodológicos, u otros no será posible, entonces, se debe trabajar con una* ***muestra*** *extraída de la* ***población*** *bajo estudio.*
Dependiendo del número de sus elementos, ***una población puede ser finita o infinita***

- ***Una población es finita*** *si esta consta de una cantidad finita de elementos. es decir, donde la contar sus elementos siempre garantizamos llegar hasta el último de estos.*
- ***Una población infinita*** *es aquella en la que al contar sus elementos es imposible llegar hasta el último de ellos*

Definición 1.1.6 (PARÁMETRO) *En estadística se refiere a los valores o medidas que caracteriza una* ***población****, como por ejemplo la* ***media*** *y la* ***desviación típica*** *de una* ***población****. Son cantidades indeterminadas constantes o fijas respecto a una condición o situación que caracteriza a un fenómeno en un momento dado que ocurre una población.*
Se suele representar a un ***parámetro*** *mediante letras griegas, por ejemplo, la* ***media poblacional*** *se representa mediante μ_x y se lee comomedia poblacional de la variable aleatoria X, la* ***Varianza Poblacional*** *se representa mediante σ_x^2 y se lee como varianza poblacional de la variable X En términos prácticos un* ***parámetro*** ES UN VALOR QUE RESULTA AL EMPLEAR LOS VALORES QUE SE OBTIENE DE UNA POBLACIÓN *Son todos aquellos pasos que nos permiten saber a partir de datos, algunas características que puedan tener una* ***población*** *en particular. El problema que tenemos es que, las* ***poblaciones*** *pueden no ser sujeto de estudio exhaustivo por varias razones:*

1. *Una población puede ser muy grande como para que podamos atacarla.*
2. *Las poblaciones están localizadas en sitios que son sumamente inaccesibles.*
3. *LA población esta compuesta de elementos que físicamente no estén en disposición.*

Dicho lo anterior, por eso es importante que se trabaje con **MUESTRAS**

1.2. MUESTRAS

Es toda parte representativa de una **población** cuyas características debe de reproducir en pequeño, lo más exactamente posible. Los número que obtenemos cuando se realiza una muestra se llaman **estadísticos o índices**.
En estadística se trabaja con muestras para analizar poblaciones, porque la mayoría de las veces no es posible trabajar con todos los elementos de la población
Las **muestras** son la clave para los conocimientos estadísticos precisos. Tienen dos principales características: **aleatoriedad** y **representativa**

- **Aleatoriedad:** Cuando cada miembro se elige al azar.
- **Representativa** Refleja fielmente a toda la población

Para seleccionar una **MUESTRA**, tenemos que cuidar bastantes elementos:

1. Debemos tener muy bien definida nuestra población y tener muy bien definida la variable de estudios que nos interesa analizar. De esta manera, podremos tener más confianza en que los resultados que nos arroje la **muestra**, efectivamente, puedan representar a la **población**
2. Un segundo elemento para considerar es el **tamaño de la muestra**

.

- Si tenemos **muestras muy pequeñas**, estas pueden no ser representativas del comportamiento de la variable dentro de la **población**
- Si tenemos **muestras muy grandes** pueden llevar a costos muy elevados que hagan el estudio estadístico algo inviable.

1.2.1. MUESTRA REPRESENTATIVA

Es aquella que tiene características muy similares a la población.
La construcción de muestras adecuadas, representativas, es uno de los aspectos más delicados de la Estadística.
La estadística trabaja con muestras, y estas deben ser lo más parecido a una muestra representativa.
un censo es el estudio sobre toda la población

1.2.2. MUESTRAS ALEATORIAS

Una muestra aleatoria de tamaño **n**, de la **función de distribución de la variable aleatoria X**, es ua colección de **n** variables aleatorias independientes.

$$\{X_1, X_2, \ldots, X_n\}$$

Cada una con la misma **función de distribución de la variable aleatoria**

1.2.3. MUESTREO ALEATORIO SIMPLE

En este tipo de muestreo lo que estamos tratando de garantizar es que, de principio, cada elemento de la **población** , pueda ser seleccionado para la**muestra**. Una forma de poder llevar acabo un esquema de **muestreo aleatorio simple**, es tener una hoja de cálculo en la que estén en la lista todos los elementos de la **población** y pedir una selección aleatoria, para los cuales, existen funciones dentro de las hojas de cálculo. Esto equivale a que si pusiéramos el nombre de todos y cada una de las personas o de los elementos que componen a la **población** en un pedazo de papel , los colocáramos en una urna y empezáramos a extraer uno a uno de manera completamente aleatoria.

1.2.4. MUESTRA ALEATORIA APLICADA

Una muestra aleatoria de tamaño **n** es un conjunto de **n** observaciones $\{x_1, x_2, \ldots, x_n\}$ sobre las variables $\{X_1, X_2, \ldots, X_n\}$
Hay tres formas de considerar una muestra aleatoria:

- Como un conjunto de unidades seleccionadas y que son sometidas a estudio
- Como un conjunto de variables aleatorias teóricas asociadas con esas unidades
- Como un conjunto de valores numéricos tomadas por las variables

1.3. ILUSTRACIÓN DE POBLACIÓN Y MUESTRA

1. Elaborar productos más eficientes, en la búsqueda de nuevos remedios para tratar la caspa, la calvicie, la obesidad, etc.
   En el tratamiento de la caspa muchas personas están involucradas -como en el problema 1.1, esto es, los bioquímicos que hacen formulas para diferentes champús u otros productor, los dueños de los laboratorios que las producen y las personas que requieren tratamiento.

   ¿ Quién sería la población y quién sería la muestra?

   Aquí la **población** quedará integrada por las personas que tiene caspa. Si bien, en este caso puede resultar complicado identificar a las personas,. será necesario limitar el estudio a un universo más especifico.
   En consecuencia, la **muestra** será una parte de personas con caspa

2. El área de control de calidad busca mejorar los productos. El control de calidad es una actividad importante en la mayoría de las empresas que manufacturan diferentes productos. El seguimiento de estos productos está a cargo de administradores, ingenieros industriales, químicos, mecánicos, electrónicos, por mencionar algunos.
   Un fabricante de " chips"para computadoras desea monitorear la calidad del producto. Dado que se produce una gran cantidad de chips cada día, sólo se tomará una muestra de éstos. Por lo cual, la **muestra** será sólo una parte de los " chips"de ese lote de producción. Existen mecanismos estadísticos para seleccionar los " chips "de la muestra.Tales mecanismos se indicarán más adelante.

3. Encuestas para las elecciones
   En la actualidad, diferentes empresas encuestadoras realizan sondeos para conocer la preferencia de los votantes por algún candidato. Ellos seleccionan una muestra del directorio telefónico para conocer la opinión de la gente

¿ Cual es la población?

La población, en este caso, es el número de personas con teléfono y que aparezcan en ese directorio .

¿ Cuál es la muestra?

Una muestra será escoger unos cuantos números que hay en ese directorio. Más adelante se indicará cómo seleccionar los individuos de la muestra

RAMAS DE LA ESTADÍSTICA

La estadística se divide en dos ramas, que son:

1. **Estadística descriptiva:** Aquella que se encarga de organiL zar, describir e interpretar un conjunto de datos
2. **Estadística inferencial:** Es aquella que se encarga de analizar un conjutno de datos de una muestra para después poder hacer inferencias sobre la población

Definición 1.3.1

Muestreo: *En al referencia estadística se conoce como muestreo a la técnica para la selección o recolección de una muestra a partir de una población estadística.*
El muestreo puede ser de ***tipo aleatorio*** *o de* ***tipo no aleatorio***

Unidad experimental: *Es la entidad específica que es de interés de un estudio estadístico*

Una ***Variable:*** *Es una característica que se mide a cada unidad experimental en un estudio estadístico.*

Un ***Dato u observación:*** *es el valor que toma la variable para una unidad experimental*

La colección de observaciones que toma una o más variables representa el ***conjunto de datos***

1.4. OBTENCIÓN DE INFORMACIÓN ESTADÍSTICAS

Si se va a preguntar a las personas tiene que ser:

- Que las preguntas no se presten a interpretaciones.
- La pregunta tiene que ser clara
- Tener en cuenta el tiempo que llevará levantar la información.
- A qué hora se hará
- En que espacio geográfico

Si lo que vamos a hacer es tomar medidas, tenemos que tener consideración:

- El instrumento con el que vamos a hacer la medición y las maneras que estás serán levantadas.

La **toma de medidas** es otra forma de recoger información estadísticas.
Una manera adicional, es la de recurrir a **fuentes históricas**, a datos previos, a **muestras** que se hayan levantado. Actualmente el **INTERNET** es una forma de poderlas entender.
Dos conceptos que tenemos que tener claro en estadística son **censo** y **encuesta**

- **Censo:** Aplicación del instrumento estadístico para obtener información al total de la población.

- **Encuesta:** Aplicación del instrumento estadístico para obtener información a una **muestra**, es decir, solamente aplicamos el instrumento a una parte de la **población** y no al total de la misma

1.5. CLASIFICACIÓN DE DATOS

La **estadística** es básicamente una ciencia que implica la **recopilación de datos**, la **interpretación de datos**, y finalmente, **la validación de datos**. El análisis de datos estadístico es un procedimiento para realizar diversas operaciones estadísticas.

1.5.1. TIPOS DE DATOS EN ANÁLISIS ESTADÍSTICOS

1. **Categóricos:** Los datos categóricos representan grupos o categorías.

2. **Numéricos:** Los datos numéricos representan números.

    - Discretos: Los datos discretos generalmente se pueden escribir en una materia finita.
    - Continuos: Los datos continuos es infinito e imposible de contar

1.5.2. POR SU NIVEL DE MEDICIÓN

1. **Cualitativos**

    - Nominal: Representa categorías que no se pueden poner en ningún orden.
    - Ordinal: Representa categorías que se pueden ordenar.

2. **Cuantitativos**
   Ambos representan "números"

    - Intervalo: Proporciona información sobre el orden y también poseen intervalos iguales. Construir bajo estos niveles de medición más profunda de principios matemáticos y estadísticos. Sin embargo, es importante comprender los diferentes niveles de medición al utilizar e interpretar escalas.
    - Razón: Además de poseer las cualidades de las escalas nominal, ordinal y de intervalo, una escala de razón tiene un **cero absoluto** (un punto donde no existe ninguna de las cualidades que se están midiendo). Utilizar una **escala de razón** permite hacer comparaciones como sr el doble de alto, o la mitad de alto de una persona.

1.6. FRECUENCIA ESTADÍSTICA

1.6.1. FRECUENCIA ABSOLUTA

En **estadística**, la **frecuencia absoluta** de un evento es el número de veces en que dicho evento se repite durante un experimento o **muestra estadística**

1.6.2. DISTRIBUCIÓN DE FRECUENCIAS

A la forma en que las observaciones se acomodan con respecto a los valores que toman, les vamos a llamar **distribución.** En **estadística**, se denomina **distribución de frecuencias** a la agrupación de datos, generalmente representada en una tabla, en categóricas excluyentes que concentran el número de veces que tales datos se repiten, es decir, su **frecuencia** de aparición en cierto conjunto. La intención es, observar de manera más sencilla el número de datos existentes en cada categoría de la **distribución**

1.6.3. TABLAS DE DISTRIBUCIÓN DE FRECUENCIAS

Es una manera más ordenada de representar la información y a partir de la cual podemos hacer lecturas un poco más amplias.

- También podemos ver la **tendencia** (hacia donde van los valores) y la **variabilidad** que pueden tener nuestras observaciones dentro de las muestras.

La **tabla de distribución de frecuencias** nos permite darnos cuenta de la **tendencia** en los valores de la variable, de su **distribución** y de su**variabilidad**.

- Cuando tenemos una tabla valor por valor, decimos que tenemos una **presentación de datos no agrupados**.

- También es posible tener una **presentación de datos por intervalos.**. Cuando trabajamos con un conjunto grande de números, generalmente no podemos sacar conclusiones, pues solemos concentrarnos en los valores mayores o en los más frecuentes, sin afirmar algo categóricamente. En este caso, decimos que tenemos **datos**, pero no **información**

1.6.4. AGRUPACIÓN DE DATOS EN UNA DISTRIBUCIÓN DE FRECUENCIAS

¿ Cómo agrupar correctamente un conjunto de datos?

Cuando la cantidad de posible respuesta es grande, es necesario trabajar la información de una manera simplificada: Para ellos, lo que se acostumbra es dividir el espectro de valores posibles en intervalos, llamados **clases** que luego se registran en una tabla de frecuencias conocida como **Distribución de Frecuencias Agrupadas**

1. Ordenar los **n** datos de menor a mayor.

2. **Rango de valores:** Ordena los datos e identifica el dato mayor y el dato menor, con ello vamos a determinar el rango.

$$rango = Dayo\,mayor - Dato\,menor$$

3. **Número de clases:** Cada uno de los intervalos que utilizaremos en la **distribución de frecuencias agrupadas** se llaman **clases**
   Una manera de determinar el **número de clases** es usando la siguiente formula:

$$k = \sqrt{n}$$

   - Si el no. de clases da en forma decimal se redondea el resultado al entero más próximo

4. **Amplitud:** Todos los intervalos o clases deben tener el mismo ancho, de manera que cada dato caiga dentro de solamente una clase.

$$amplitud = \frac{rango}{k} = \frac{Dato\ mayor - Dato\ Menor}{\sqrt{n}}$$

Este resultado, redondearlo al siguiente entero, decimal, centésima, milésima, etc. Dependiendo si el conjunto de datos consta de puros enteros, puros decimales con una cifra,decimales con dos cifras, decimales con tres cifras significativas,etc,respectivamente

5. **Hacer el pre-cálculo**

$$precalculo = Dato\ menor + k \cdot (amplitud)$$

Este resultado se compara con el dato mayor. El pre-cálculo debe superar al dato mayor en una cantidad par o impar.

Ejemplo

a) Si el valor del pre-cálculo supera el dato mayor en 4 unidades, **cantidad par**, entonces dividimos este exceso entre dos

$$\frac{4}{2} = 2$$

De esta forma, el valor del límite inferior de la primer clase de la distribución de frecuencia será:

$$lim\ inf = Dato\ menor - 2$$

b) Si el valor del pre-cálculo supera el dato mayor en 3 unidades, **cantidad impar**, entonces encontramos dos enteros, lo más cercanos posible, de tal forma que sea igual a 3
En este caso los números enteros serían 1 y 2.
De esta forma, el valor del límite inferior de la primera clase de la distribución de frecuencia será:

$$lim\ inf = Dato\ menor - 1$$

NOTA: Estos límites inferiores ilustrados son casos particulares, **NO** generales.

6. **Construir las clases**

a) **Clases No reales**
Para construir las clases.(en esta clase hacen saltos de una unidad, o sea NO hay continuidad)
Supongamos que nuestros datos son puros números enteros, en donde;

$$Dato\ menor = 6\ ,\ Dato\ menor = 26\ ,\ amplitud = 5$$

pre-cálculo$= 6 + 5 \cdot (5) = 31$, el cual supera al dato mayor en 5 unidades: 5=2+3
Entonces el límite inferior de mi primer clase será:

$$lim\ inf = 6 - 2 = 4$$

Para determinar el límite superior de cada clase. al límite inferior se le suma el valor $Amplitud - 1 = 4$

Clases no. reales	f_i	f_{ai}	$f_r\,\%$	$f_{rai}\,\%$
4 a 8				
9 a 13				
14 a 18				
19 a 23				
24 a 28		n		
	$\sum f_i = n$		$\sum f_r\,\% = 100\,\%$	

Donde

- f_i Denota a la frecuencia absoluta y se define como la cantidad de datos que cae en cada clase
- f_{ai} Denota a la frecuencia acumulada de las frecuencias absolutas
- $f_r\,\%$ Denota a la frecuencia relativa porcentual y se define como

$$f_r\,\% = \frac{f_i}{n} \cdot 100\,\%$$

- $f_{rai}\,\%$ Denota a la frecuencia relativa porcentual acumulada

b) **Clases reales**
Para construir las clases reales(En esta clase SI hay continuidad en la construcción) **Supongamos** que nuestros datos son inicialmente número enteros, en donde:

$$Dato\ menor = 6\ ,\ Dato\ mayor = 26\ ,\ k = 5\ ,\ amplitud = 5$$

pre-cálculo$= 6 + (5) \cdot (5) = 31$, el cual supera el dato mayor en 5 unidades $5 = 2 + 3$, entonces utilizamos el 2 para construir nuestra primer clase. El límite inferior de mi primer clase será

$$lim\ inf = 6 - 2 = 4$$

Para determinar el limite superior de cada clase, el limite inferior se le suma el valor de la $amplitud = 5$

Clases reales	f_i	f_{ai}	$f_r\,\%$	$f_{rai}\,\%$
$4 < 9$				
$9 < 14$				
$14 < 19$				
$19 < 24$				
$24 < 29$		n		100
	$\sum f_i = n$		$\sum f_r\,\% = 100\,\%$	

¿ Cuando se debe hacer una distribución de frecuencias con clases reales y cuando con clases no reales?

R: Cuando el conjunto de datos consta de decimales se utiliza una distribución de frecuencias con clases reales y cuando trabajamos con datos discretos (no. enteros) lo aconsejable es hacer una distribución de frecuencias con clases No reales.
En otras palabras, si los datos pertenecen a una **variable aleatoria continua** entonces en estos casos se debe hacer una **distribución de frecuencia con clases reales** de lo contrario **distribución de frecuencias con clases no reales**

Ejercicio 1.6.1 *Agrupar los siguientes conjuntos de datos en una distribución de frecuencias. Adaptando la que sea más adecuada.*
El siguiente conjunto de datos corresponde a la medida del diámetro en un engrane de pulgadas

$$\begin{pmatrix} 6.00 & 6.00 & 6.25 & 6.25 & 6.25 \\ 6.25 & 6.25 & 6.50 & 6.50 & 6.50 \\ 6.50 & 6.50 & 6.50 & 6.50 & 6.65 \\ 6.65 & 6.70 & 6.70 & 6.75 & 6.75 \\ 6.75 & 6.75 & 6.75 & 6.75 & 6.75 \\ 6.75 & 6.75 & 7.00 & 7.00 & 7.00 \\ 7.00 & 7.00 & 7.00 & 7.00 & 7.10 \\ 7.10 & 7.15 & 7.15 & 7.25 & 7.25 \end{pmatrix}$$

Solución:
Son ***Datos continuos*** *con dos cifras.*

a) *Los datos ya están ordenados*

b) ***Rango***

$$Rango = 7.25 - 6 = 1.25$$

c) ***Número de clases:***

$$k = \sqrt{40} = 6.32 \approx 6$$

d) ***Amplitud de clase:***

$$\frac{Rango}{k} = \frac{1.25}{6} = 0.2083 \approx 0.21$$

(Se aproxima a la próxima centésima, esto es ya que los datos tienen 2 decimales)

e) ***Pre-cálculo:***

$$precalculo = 6 + (6) \cdot (0.21) = 7.26$$

Comparando con el dato mayor, se pasa por

$$0.01 \longrightarrow 0.01 = 0.00 + 0,01$$

Construimos nuestras clases, son clases reales porque tienen decimales.

Clases reales	f_i	f_{ai}	$f_r\,\%$	$f_{rai}\,\%$
6 a < 6.21	*2*	*2*	*5*	*5*
6.21 a < 6.42	*5*	*7*	*12.5 %*	*17.5 %*
6.42 a < 6.63	*7*	*14*	*17.5 %*	*35 %*
6.63 a < 6.84	*13*	*27*	*32.5 %*	*67.5 %*
6.84 a < 7.05	*7*	*34*	*17.5 %*	*85 %*
7.05 a < 7.26	*6*	*n=40*	*15 %*	*100 %*
	$\sum f_i = n$		$\sum f_r\,\% = 100\,\%$	

El limite inferior de la primer clase dista a 0 unidades del dato menor y el limite superior dista a .1 centésima del dato mayor

Recuerda que la distribución de frecuencias nos sirve para interpretar el comportamiento de nuestro conjunto de datos

Ejercicio 1.6.2 *Se toma el pulso (latidos por minuto)a 40 mujeres en un centro de salud y se obtuvieron los siguientes resultados: Agrupar los datos en una distribución de frecuencias*

$$\begin{pmatrix} 72 & 64 & 80 & 65 & 76 & 60 & 64 & 88 \\ 80 & 65 & 76 & 72 & 88 & 72 & 65 & 88 \\ 60 & 65 & 76 & 80 & 96 & 72 & 64 & 80 \\ 72 & 65 & 88 & 124 & 72 & 76 & 80 & 60 \\ 88 & 64 & 80 & 76 & 72 & 104 & 76 & 72 \end{pmatrix}$$

Solución
Para este caso utilizaremos la distribución de frecuencia para clases no reales

a) *Ordenar los* ***n*** *datos de menor a mayor:*

$$\begin{pmatrix} 60 & 60 & 60 & 64 & 64 & 64 & 64 & 65 \\ 65 & 65 & 65 & 65 & 72 & 72 & 72 & 72 \\ 72 & 72 & 72 & 72 & 76 & 76 & 76 & 76 \\ 76 & 76 & 80 & 80 & 80 & 80 & 80 & 80 \\ 80 & 88 & 88 & 88 & 88 & 96 & 104 & 124 \end{pmatrix}$$

b) ***Rango de valores:***

$$rango = dato\ mayor - dato\ menor = 124 - 60 = 64$$

c) ***número de clase:***

$$k = \sqrt{n} = \sqrt{40} = 6.32 \approx 6$$

d) ***amplitud:***

$$amplitud = \frac{rango}{k} = \frac{64}{6} = 10.66 \approx 11$$

Recordar que esto se hace respecto a los datos (como todos los datos son números enteros, esto se aproxima al siguiente número entero)

e) ***pre-cálculo:***

$$precalculo = dato\ menor + k(amplitud) = 60 + 6(11) = 126$$

$$exceso = precalculo - dato\ mayor = 126 - 124 = 2$$

*el exceso se pasa por 2 unidades (****cantidad par****) entonces el*

$$lim_{inf} = dato\ menor - \frac{exceso}{2} = 60 - 1 = 59$$

Construimos nuestras clases, son no reales por los datos dados

Clases reales	f_i	f_{ai}	$f_r\,\%$	$f_{rai}\,\%$
59 a 69	*12*	*12*	*30 %*	*30 %*
70 a 80	*20*	*32*	*50 %*	*80 %*
81 a 91	*5*	*37*	*12.5 %*	*92.5 %*
92 a 102	*1*	*38*	*2.5 %*	*95 %*
103 a 113	*1*	*39*	*2.5 %*	*97.5 %*
114 a 124	*1*	*40*	*2.5 %*	*100 %*
	$\sum f_i = n$		$\sum f_r\,\% = 100\,\%$	

Ejercicio 1.6.3

Los siguientes datos muestran las edades de personas que tenían a la hora de engañar a sus parejas. Construir una tabla de frecuencia para estos datos y analicen sus hallazgos. ***¿ Entre que edad hay más personas infieles?***

$$\begin{pmatrix} 32 & 39 & 40 & 47 & 40 \\ 55 & 48 & 31 & 51 & 43 \\ 48 & 46 & 32 & 42 & 33 \\ 39 & 36 & 37 & 62 & 32 \\ 37 & 76 & 43 & 53 & 42 \end{pmatrix}$$

Solución: *Para este caso utilizaremos la distribución de frecuencia para clases no reales*

a) *Ordenar los* $\boldsymbol{n}m$ *datos de menor a mayor*

$$\begin{pmatrix} 31 & 32 & 32 & 32 & 33 \\ 36 & 37 & 37 & 39 & 39 \\ 40 & 40 & 42 & 42 & 43 \\ 43 & 46 & 47 & 48 & 48 \\ 51 & 53 & 55 & 62 & 76 \end{pmatrix}$$

b) ***rango de valores:***

$$rango = dato\ mayor - dato\ menor = 76 - 31 = 45$$

c) ***número de clases:***

$$k = \sqrt{n} = \sqrt{5} \approx 5$$

d) ***amplitud:***

$$amplitud = \frac{rango}{k} = \frac{45}{5} = 9 \approx 10$$

Recuerda que esto se hace respecto a los datos(como todos los datos son números enteros, entonces se aproxima al siguiente número entero

e) ***pre-cálculo:***

$$precalculo = dato\ menor + k(amplitud) = 31 + 5(10) = 81$$

$$exceso = precalculo - dato\ mayor = 81 - 76 = 5$$

el exceso se pasa por 5 unidades ***cantidad impar****, entonces se buscan dos números enteros tal que la suma de 5*

$$\longrightarrow 5 = 2 + 3$$

Se agarra el primer número, entonces con esto tenemos el limite inferior

$$lim_{inf} = dato\ menor - 2 = 31 - 2 = 29$$

Construimos nuestras clases, son clases no reales por los datos dados

Clases reales	f_i	f_{ai}	$f_r\,\%$	$f_{rai}\,\%$
29 a 38	*8*	*8*	*32 %*	*32 %*
39 a 48	*12*	*20*	*48 %*	*80 %*
49 a 58	*3*	*23*	*12 %*	*92 %*
59 a 68	*1*	*24*	*4 %*	*96 %*
69 a 78	*1*	*25*	*4 %*	*100 %*
	$\sum f_i = n$		$\sum f_r\,\% = 100\,\%$	

A partir de la tabla, podemos observar que hay más infieles de los 39 a los 48 años

Ejercicio 1.6.4 *Como control de la ética publicitaria se requiere que el rendimiento, en millas por galón de gasolina, que los fabricantes de automóviles usan con fines publicitarios este basado en un buen número de pruebas efectuadas en diversas condiciones. Al tomar una muestra de 50 automóviles se registraron las siguientes observaciones en millas por galón.*

$$\begin{pmatrix} 27.9 & 34.2 & 35.6 & 28.5 & 30 & 31.2 & 30.5 & 28.7 & 33.4 & 30.1 \\ 29.3 & 32.7 & 31 & 27.5 & 28.7 & 29.5 & 31.3 & 30.4 & 30.5 & 30.3 \\ 31.8 & 26.5 & 28 & 29.8 & 32.2 & 28.7 & 24.9 & 31.3 & 30.6 & 29.6 \\ 22.5 & 26.4 & 33.7 & 31.2 & 30.5 & 23 & 26.8 & 32.7 & 35.1 & 31.4 \\ 34.2 & 32.6 & 32 & 28.7 & 27.9 & 30.1 & 29.9 & 30.3 & 28.6 & 32.4 \end{pmatrix}$$

Solución:

a) *Ordenar los* $\mathbf{n}$ *datos de menor a mayor*

$$\begin{pmatrix} 22.5 & 23 & 24.9 & 26.5 & 26.5 & 26.8 \\ 27.5 & 27.9 & 27.9 & 28 & 28.5 & 28.6 \\ 28.7 & 28.7 & 28.7 & 28.7 & 29.3 & 28.5 \\ 29.6 & 29.8 & 29.9 & 30 & 30.1 & 30.1 \\ 30.3 & 30.3 & 30.4 & 30.5 & 30.5 & 30.5 \\ 30.6 & 31 & 31.2 & 31.2 & 31.3 & 31.3 \\ 31.4 & 31.6 & 31.8 & 32 & 32.2 & 32.4 \\ 32.7 & 32.7 & 33.5 & 33.7 & 34.2 & 34.2 \\ 35.1 & 35.6 & & & & \end{pmatrix}$$

b) ***Rango de valores:***

$$rango = dato\ mayor - dato\ menor = 35.6 - 22.5 = 13.1$$

c) ***No. de clases:***

$$k = \sqrt{50} = 7.07 \approx 7$$

d) ***Amplitud de clase:***

$$amplitud = \frac{rango}{no.\ de\ clases} = \frac{13.1}{7} = 1.87 \approx 1.9$$

e) ***Pre-cálculo:***

$$precalculo = dato\ menor + k(amplitud) = 22.5 + 7(1.9) = 3.8$$

$$exceso = precalculo - dato\ mayor = 35.8 - 35.6 = .2$$

El exceso se pasa por .2 unidades ***cantidad par****, entonces el límite inferior nos queda de la sig. manera*

$$lim_{inf} = dato\ menor - \frac{exceso}{2} = 22.5 - \frac{.2}{2} = 22.5 - .1 = 22.4$$

Construimos nuestras clases, son clases reales por los datos

Clases reales	f_i	f_{ai}	$f_r\,\%$	$f_{rai}\,\%$
22.4<24.3	*2*	*2*	*4 %*	*4 %*
24.3<26.2	*1*	*3*	*2 %*	*6 %*
26.2<28.1	*7*	*10*	*14 %*	*20 %*
28.1<30	*11*	*21*	*22 %*	*42 %*
30<31.9	*18*	*39*	*36 %*	*78 %*
31.9<33.8	*7*	*46*	*14 %*	*92 %*
33.8<35.7	*4*	*50*	*8 %*	*100 %*
	$\sum f_i = n$		$\sum f_r\,\% = 100\,\%$	

Ejercicio 1.6.5

$$\begin{pmatrix} 2 & 3 & 4 & 5 & 6 & 6 & 7 & 7 & 8 & 9 \\ 2 & 3 & 4 & 5 & 6 & 6 & 7 & 7 & 8 & 9 \\ 2 & 4 & 5 & 5 & 6 & 6 & 7 & 7 & 8 & 9 \\ 2 & 4 & 5 & 5 & 6 & 6 & 7 & 8 & 8 & 9 \\ 3 & 4 & 5 & 6 & 6 & 6 & 7 & 8 & 9 & 9 \end{pmatrix}$$

Solución:

a) *Ordenar los* ***n*** *datos de menor a mayor*

$$\begin{pmatrix} 2 & 2 & 2 & 2 & 3 & 3 & 3 & 4 & 4 & 4 \\ 4 & 4 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 6 \\ 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 \\ 7 & 7 & 7 & 7 & 7 & 7 & 7 & 7 & 8 & 8 \\ 8 & 8 & 8 & 8 & 9 & 9 & 9 & 9 & 9 & 9 \end{pmatrix}$$

b) ***Rango de valores:***

$$rango = dato\ mayor - dato\ menor = 9 - 2 = 7$$

En este caso como el rango ≤ 10 *y los datos dados son las clases dadas son no reales, entonces se recomiendo hacer una* ***DISTRIBUCIÓN DE CLASES NO REALES SIMPLE*** *con esto hacemos nuestra tabla.*

Clases reales	f_i	f_{ai}	$f_r\,\%$	$f_{rai}\,\%$
2	*4*	*4*	*8 %*	*8 %*
3	*3*	*7*	*6 %*	*14 %*
4	*5*	*12*	*10 %*	*24 %*
5	*7*	*19*	*14 %*	*38 %*
6	*11*	*30*	*22 %*	*60 %*
7	*8*	*38*	*16 %*	*76 %*
8	*6*	*44*	*12 %*	*88 %*
9	*6*	*50*	*12 %*	*100 %*
	$\sum f_i = n$		$\sum f_r\,\% = 100\,\%$	

1.7. GRÁFICAS ESTADÍSTICAS

Los distintos tipos de gráficas que veremos son:

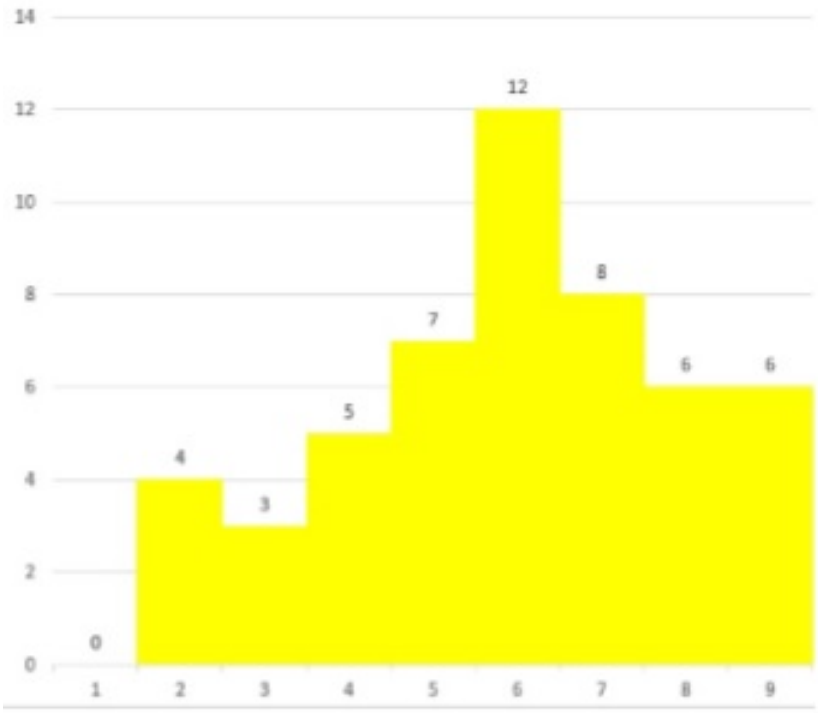
- Diagrama de pastel

- Diagrama de lineas
- Diagrama de barras
- Histograma
- Polígono de Frecuencias
- Ojiva

Para poder hacer una de estas gráficas, es necesario contar con una distribución de frecuencias.

a) PARA DISTRIBUCIÓN DE CLASES REALES OCUPAREMOS

1) **HISTOGRAMA:**
La gráfica del histograma es similar a la gráfica de barras.
La diferencia entre estas dos gráficas es que en el diagrama de barras, las barritas están separadas y en el histograma las barras están pegadas.
La gráfica del histograma es adecuada cuando se tiene una **distribución frecuencias con clases reales**. Las frecuencias adecuadas para elaborar el histograma es " f_i o $f_r\,\%$ "

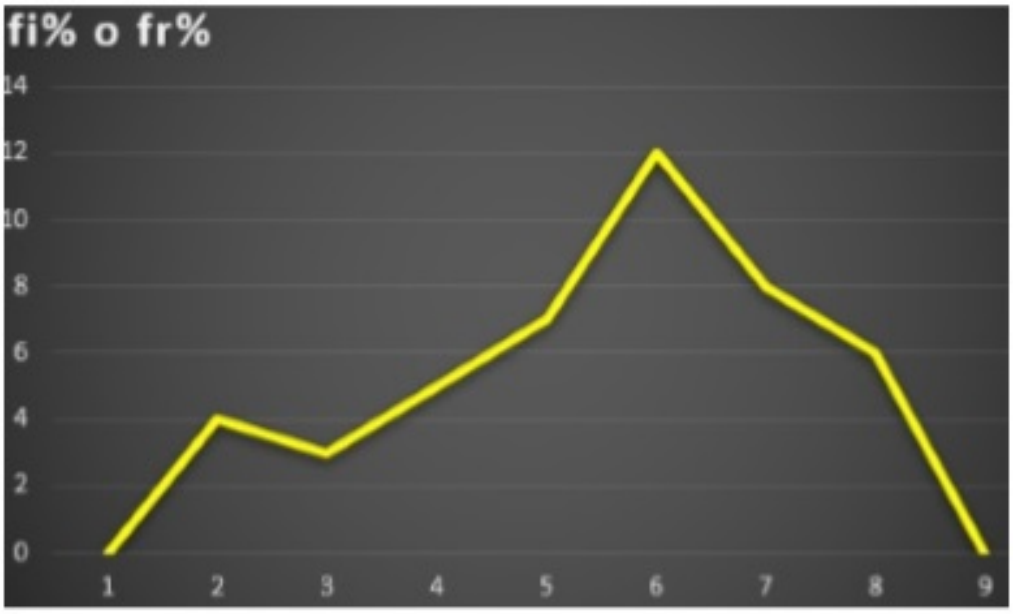


2) **POLÍGONO DE FRECUENCIAS**
Para construir este tipo de gráfica se debe calcular de antemano el valor de la **marca de clase** m_i de cada intervalo de la distribución.
La marca de clase no es más que el punto medio de cada clase:

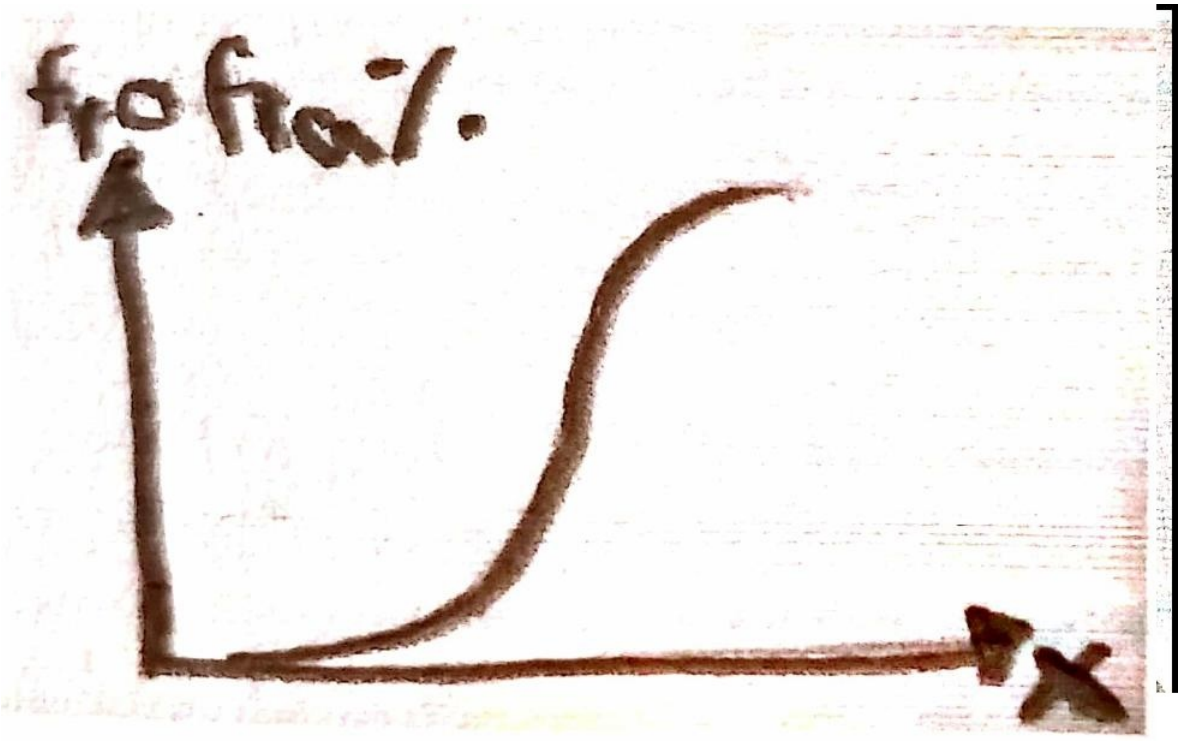
$$m_i = \frac{lim_{inf} + lim_{sup}}{2}$$

de la clase i-ésima.
Para este tipo de gráfica se sugiere utilizar la " f_i o la $f_r\,\%$ " de una distribución

de frecuencias son clases reales.
En cada clase se dibuja un punto en la posición de clase a la altura de la frecuencia correspondiente.
En la posición del eje " x " , se consideran dos clases extras, de la misma amplitud que las demás, una al inicio y otra al final y se dibujan los puntos en su correspondiente..
Finalmente se unen los puntos dibujados previamente, mediante lineas rectas, quedando la gráfica de un polígono cerrado.



3) **OJIVA**
La ojiva es la única gráfica que utiliza *frecuencias acumuladas*, f_{ai} o $f_{rai}\,\%$. Para su construcción se utiliza un sistema de ejes coordenados, en donde en el eje " x " se ubican los limites inferiores y superiores de ls clases, y en el eje " y " se toma la escala de 20 en 20, hasta el 100 %
A la altura del eje " x " en el limite inferior de la primera clase se dibuja el primer punto .
En la posición del limite superior de la primera clase, a la altura de la primera frecuencia acumulada de la distribución, se dibuja un segundo punto.
Para cada limite superior faltante de la distribución se procede de la misma manera.
Finalmente se unen los puntos con una curva suave, quedan siempre una gráfica creciente.
Este tipo de gráfica se recomienda cuando se tiene una **distribución de frecuencias acumuladas con clases reales**



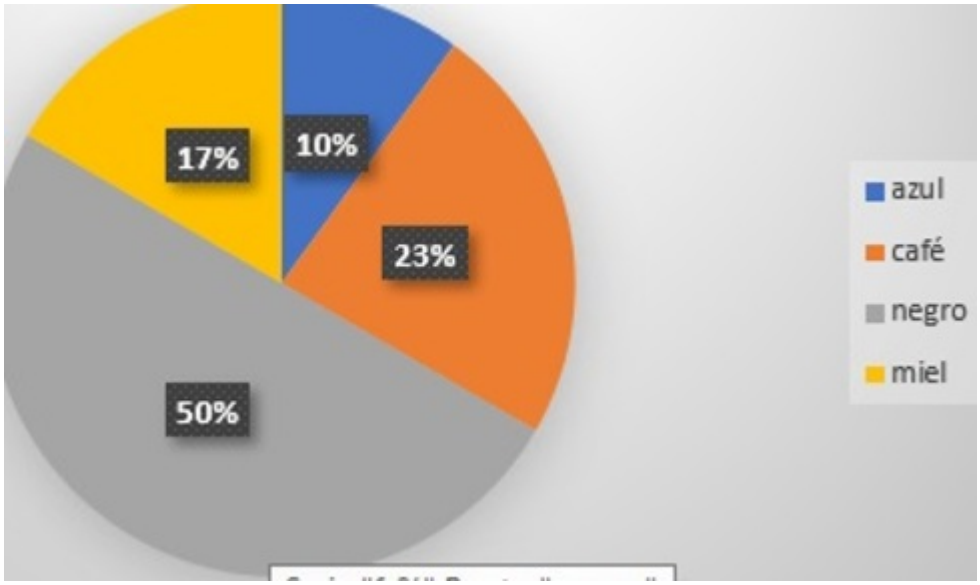
PARA DISTRIBUCIÓN DE CLASES NO REALES OCUPAREMOS

1) **DIAGRAMAS DE PASTEL:**
El tipo adecuado para utilizar un diagrama de pastel, es aquel que se conoce como **distribución de frecuencias con clases categóricas**, como la que se muestra a continuación:

Clases reales	f_i
azul	3
café	7
negro	15
miel	5

donde la variable es **color de ojos de una persona**. También se puede emplear en el caso de tener distribuciones con clases no reales.
La frecuencia más adecuada para elaborar diagrama de pastel es **" fr %¨**. Por lo cual, de acuerdo a la distribución anterior,consideremos la distribución de frecuencias relativas porcentuales.

Clases reales	$fr\,\%$
azul	10
café	23.33
negro	50
miel	16.66



Para realizar el diagrama de pastel, se debe efectuar la siguiente secuencia:

a' Dibujar una circunferencia

b' Dividir a la circunferencia en tanto sectores como clases haya en la distribución, en nuestro caso, esta dividido en 4 sectores.

Cada sector debe corresponder a un área proporcional al porcentaje de la frecuencia . Si la gráfica se hace manualmente,se debe calcular el número de grados correspondientes a cada sector de tal forma que la suma de los grados correspondientes a cada sector sea de 360°.
Por ejemplo, para determinar el número de grados correspondientes al sector de ojos azules, se debe utilizar una regla de tes, como se indica a continuación:

$$360^{\circ} \longrightarrow 100\ \%$$

$$x \longrightarrow 10\ \%$$

$$x = \frac{360^{\circ}(10\ \%)}{100\ \%} = 36^{\circ}$$

Por lo tanto al sector correspondiente al color de ojos azules corresponde a una abertura de 36° . Para los otros 3 sectores se procede de forma similar. Para crear el diagrama de pastel en forma manual, se debe contar con un compás, un transformador, una regla y una caja de colores (iluminar cada sector de distinto color).

2) **DIAGRAMAS DE LINEAS**

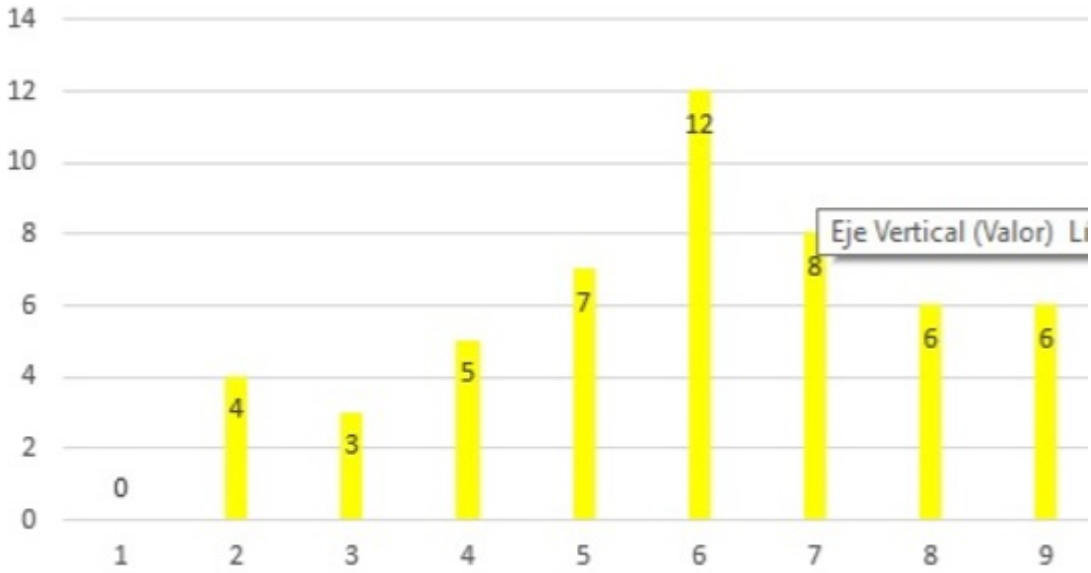
Mediante este gráfico se puede comprobar rápidamente el cambio de tendencia de los datos. El diagrama lineal se suele utilizar con variables cuantitativas, para ver su comportamiento en el transcurso del tiempo. Ejemplo:

Clases reales	$f_i\,\%$
1	0
2	4
3	3
4	5
5	7
6	11
7	8
8	6
9	6

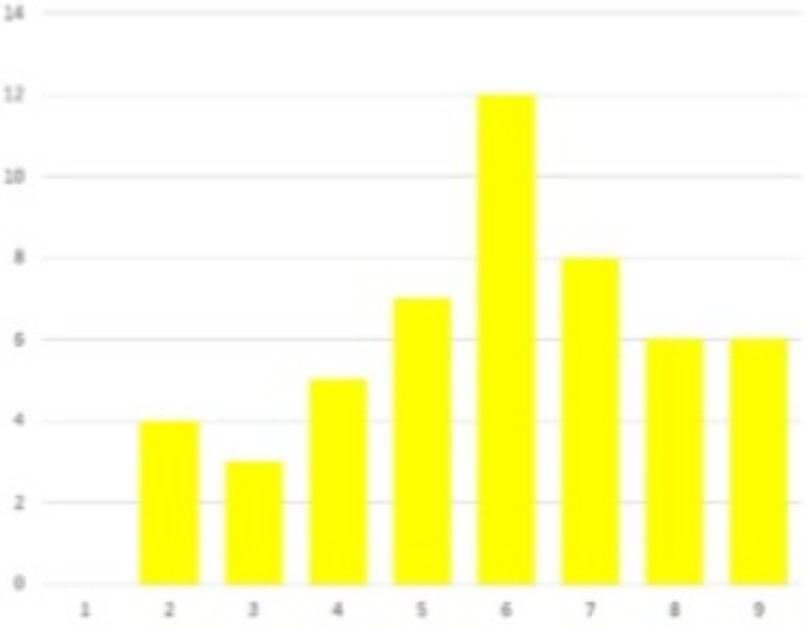
Donde la variable es " el número de días en que una persona se recupera de su gripe " Las frecuencias adecuadas para elaborar el diagrama de líneas es " f_i, $f_r\,\%$,f_{ai} y $f_{rai}\,\%$ " a modo de ejemplo, consideremos la distribución de frecuencias absoluta mostrada anteriormente. Para la elaboración del diagrama de lineas, consideremos la siguiente secuencia:

a' Dibuje un sistema de ejes coordenados, en donde el eje " x " represente el número de días en que una persona se recupera de la gripe (valores de las distintas clases) y el eje " y " representará la frecuencia absoluta " f_i "

b' En cada valor de la variable levantar una línea de altura f_i, según el valor correspondiente de la distribución de frecuencia, como se muestra a continuación



3) **DIAGRAMA DE BARRAS:**
El diagrama de barras es similar, al diagrama de lineas sólo que en lugar de líneas se utilizan barras. se sugiere utilizar el diagrama de barras para el caso de una **distribución de frecuencias con clases no reales**
Por cada clase se utiliza una barra cuya longitud de su base sea igual a la amplitud de la clase y de altura igual a valor de frecuencia correspondiente



Construir una distribución de frecuencias para cada uno de los siguientes problemas.

Ejemplo 1.7.1

Como control de la ética publicitaria se requiere que el rendimiento, en millas por galón de gasolina, que los fabricantes de automóviles usan con fines publicitarios este basado en un buen número de pruebas efectuadas en diversas condiciones. Al tomar una muestra de 50 automóviles se registraron las siguientes observaciones en millas por galón.

$$\begin{pmatrix} 27.9 & 34.2 & 35.6 & 28.5 & 30.0 & 31.2 & 30.5 & 28.7 & 33.4 & 30.1 \\ 29.3 & 32.7 & 31.0 & 27.5 & 28.7 & 29.5 & 31.3 & 30.4 & 30.5 & 30.3 \\ 31.8 & 26.5 & 28.0 & 29.8 & 32.2 & 28.7 & 24.9 & 31.3 & 30.6 & 29.6 \\ 22.5 & 26.4 & 33.7 & 31.2 & 30.5 & 23.0 & 26.8 & 32.7 & 35.1 & 31.4 \\ 34.2 & 32.6 & 32.0 & 28.7 & 27.9 & 30.1 & 29.9 & 30.3 & 28.6 & 32.4 \end{pmatrix}$$

Solución:
Para este caso utilizaremos la distribución de frecuencias para clases reales.

1. *Ordenar los* ***n*** *datos de menor a mayor*

$$\begin{pmatrix} 22.5 & 23.0 & 24.9 & 26.4 & 26.5 & 26.8 & 27.5 & 27.9 & 27.9 & 28.0 \\ 28.5 & 28.6 & 28.7 & 28.7 & 28.7 & 28.7 & 29.3 & 29.5 & 29.6 & 29.8 \\ 29.9 & 30.0 & 30.1 & 30.1 & 30.3 & 30.3 & 30.4 & 30.5 & 30.5 & 30.5 \\ 30.6 & 31.0 & 31.2 & 31.2 & 31.3 & 31.3 & 31.4 & 31.8 & 32.0 & 32.2 \\ 32.4 & 32.6 & 32.7 & 32.7 & 33.4 & 33.7 & 34.2 & 34.2 & 35.1 & 35.6 \end{pmatrix}$$

2. ***rango de valores:***

$$rango = dato\ mayor - dato\ menor = 35.6 - 22.5 = 13.1$$

3. ***número de clases:***

$$k = \sqrt{n} = \sqrt{50} = 7.071067812 \approx 7$$

4. ***amplitud:***

$$amplitud = \frac{rango}{k} = \frac{13.1}{7} = 1.871428571 \approx 1.9$$

Se aproxima al siguiente decimal(esto se basa por los datos)

5. ***pre-cálculo:***

$$pre - calculo = dato\ menor + k \cdot amplitud = 22.5 + (7) \cdot (1.9) = 35.8$$

$$exceso = pre - calculo - dato\ mayor = 35.8 - 35.6 = 0.2$$

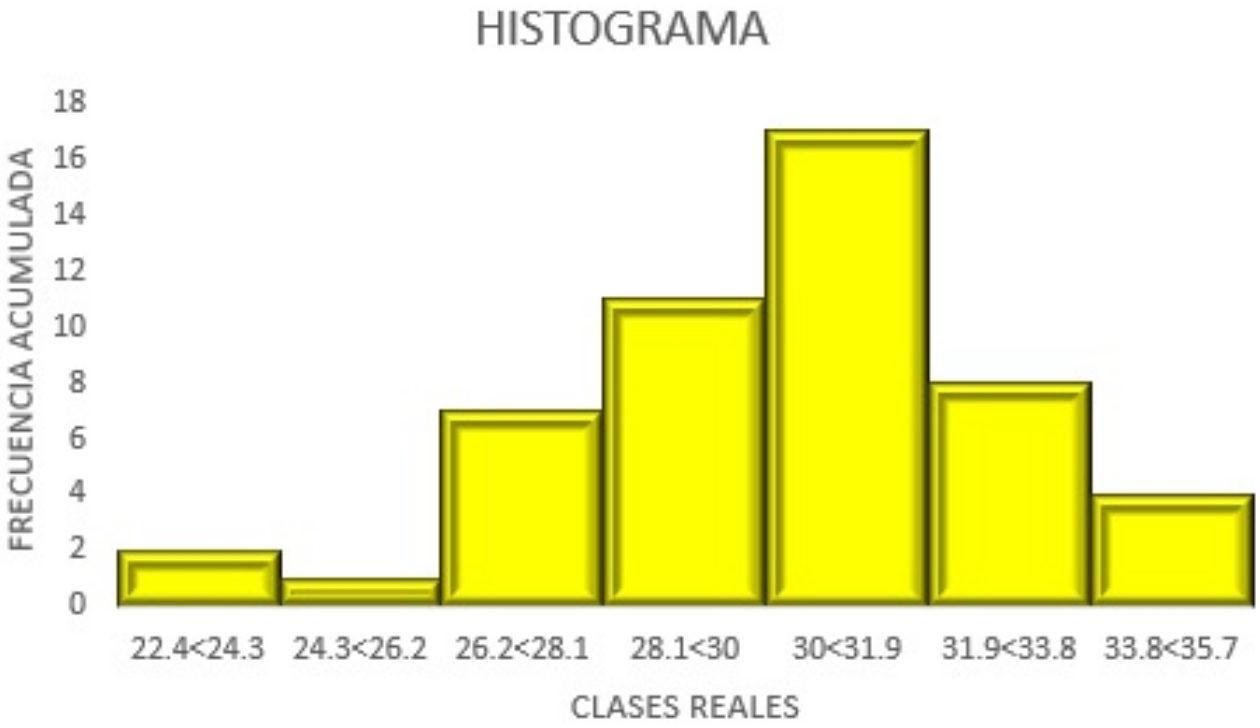
El valor de pre-cálculo supera el dato mayor en 0.2 *décimas (****cantidad par****, entonces dividimos este exceso entre dos.*

$$\frac{exceso}{2} = \frac{0.2}{2} = 0.1$$

$$lim_{\,inf} = dato\ menor - 0.1 = 22.5 - 0.1 = 22.4$$

Para sacar los limites superiores, al limite inferior se le suma la amplitud

Clases reales	f_i	f_{ai}	$f_r\,\%$	$f_{rai}\,\%$
$22.4 < 24.3$	*2*	*2*	$4\,\%$	$4\,\%$
$24.3 < 26.2$	*1*	*3*	$2\,\%$	$6\,\%$
$26.2 < 28.1$	*7*	*10*	$14\,\%$	$20\,\%$
$28.1 < 30.0$	*11*	*21*	$22\,\%$	$42\,\%$
$30.0 < 31.9$	*17*	*38*	$34\,\%$	$76\,\%$
$31.9 < 33.8$	*8*	*46*	$16\,\%$	$92\,\%$
$33.8 < 35.7$	*4*	*50*	$8\,\%$	$100\,\%$
	$\sum f_i = n$		$\sum f_r\,\% = 100\,\%$	



Ejemplo 1.7.2

Un cobrador de una empresa ha registrado el número de días que tarda en cobrar cada una de sus cuentas de crédito. Se ha obtenido los siguientes registros. (Las cuentas con más de seis semanas (46 días) de retraso se consideran incobrables y se envían al departamento legal.

$$\begin{pmatrix} 17 & 21 & 6 & 12 & 45 & 57 & 11 & 20 & 32 & 8 \\ 10 & 7 & 105 & 28 & 19 & 35 & 72 & 20 & 13 & 21 \\ 26 & 5 & 14 & 19 & 38 & 3 & 86 & 42 & 28 & 20 \end{pmatrix}$$

solución: *Para este caso utilizaremos la distribución de frecuencias para* ***clases no reales****.*

1. *Ordenar los* ***n*** *datos de menor a mayor*

$$\begin{pmatrix} 3 & 5 & 6 & 7 & 8 & 10 & 11 & 12 & 13 & 14 \\ 17 & 19 & 19 & 20 & 20 & 20 & 21 & 21 & 26 & 28 \\ 28 & 32 & 35 & 38 & 42 & 45 & 57 & 72 & 86 & 105 \end{pmatrix}$$

2. ***Rango de valores:***

$$rango = dato\ mayor - dato\ menor = 105 - 3 = 102$$

3. ***número de clases:***

$$k = \sqrt{n} = \sqrt{30} = 5.47 \approx 5$$

4. ***amplitud:***

$$amplitud = \frac{rango}{k} = \frac{102}{5} = 20.4 \approx 21$$

En este caso se aproxima al siguiente entero (Ya que los datos dados son enteros)

5. ***pre-cálculo:***

$$pre - calculo = datomenor + k \cdot amplitud = 3 + 5 \cdot 21 = 108$$

$$exceso = precalculo - dato\ mayor = 108 - 105 = 3$$

El valor del pre-cálculo supera el dato mayor en 3 unidades ***cantidad impar****, entonces buscamos dos enteros lo más cercanos posiible, de tal forma que sea igual a 3.*

$$3 = 1 + 2$$

agarramos el primer número y se lo restamos al limite inferior

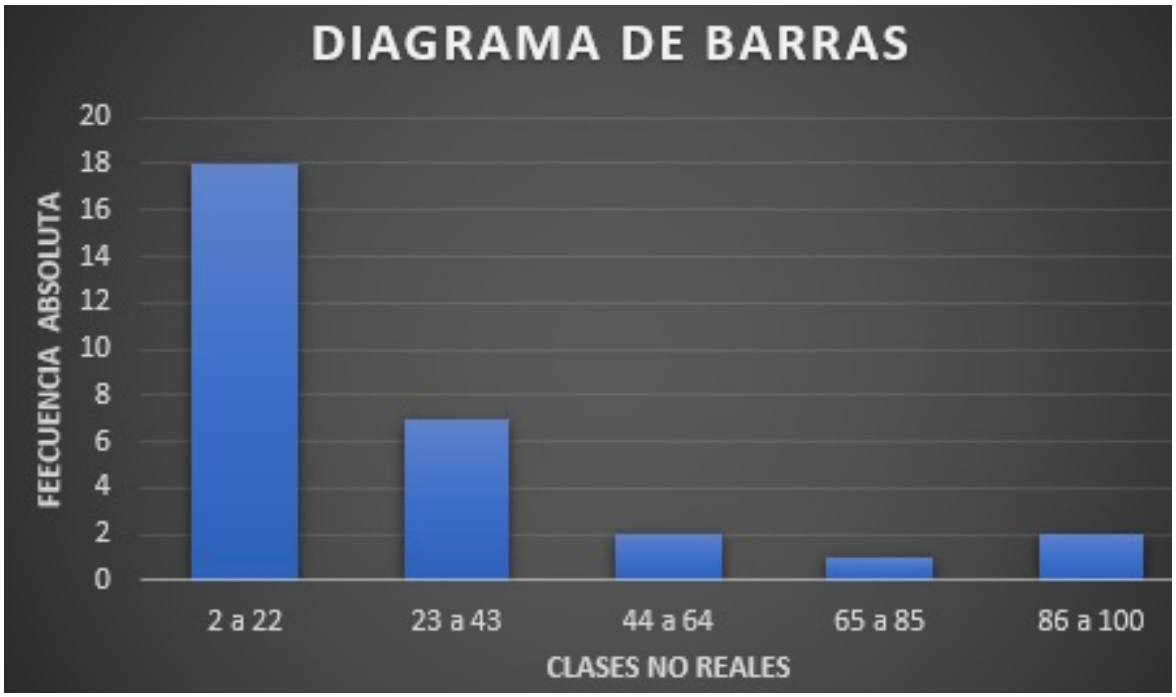
$$lim_{\,inf} = dato\ menor - 1 = 3 - 1 = 2$$

para sacar los limites superiores, al limite inferior se le suma la amplitud-1

Clases reales	f_i	f_{ai}	$f_r\,\%$	$f_{rai}\,\%$
2 *a* 22	*18*	*18*	60 %	60 %
23 *a* 43	*7*	*25*	*23.3 %*	*83.3%*
44 *a* 64	*2*	*27*	*6.6 %*	90 %
65 *a* 85	*1*	*28*	*3.3%*	*93.3%*
86 *a* 106	*2*	*30*	*6.6%*	100 %
	$\sum f_i = n$		$\sum f_r\,\% = 100\,\%$	

¿ Qué proporción de las cuentas tendrán que ser enviadas al departamento legal?: *Hay 4 datos de 30(total de datos), con más de 46 días de retraso, entonces la proporción de las cuentas que tendrán que ser enviadas al departamento legal es:*

$$\frac{4}{30} \cdot 100\,\% = 13.3\,\%$$



Ejemplo 1.7.3

A 81 estudiantes se les aplico la prueba Miller de personalidad, obteniendo las siguientes calificaciones.

$$\begin{pmatrix} 16 & 17 & 17 & 18 & 18 & 18 & 18 & 19 & 19 & 20 \\ 20 & 21 & 21 & 21 & 21 & 21 & 21 & 22 & 22 & 22 \\ 22 & 22 & 22 & 22 & 23 & 23 & 23 & 23 & 23 & 23 \\ 23 & 24 & 24 & 24 & 24 & 24 & 24 & 24 & 24 & 25 \\ 25 & 25 & 25 & 25 & 25 & 25 & 25 & 26 & 26 & 26 \\ 26 & 26 & 26 & 26 & 26 & 26 & 27 & 27 & 27 & 27 \\ 28 & 28 & 28 & 28 & 28 & 28 & 29 & 29 & 29 & 29 \\ 29 & 30 & 30 & 30 & 31 & 31 & 31 & 33 & 33 & 33 \\ 33 & & & & & & & & & \end{pmatrix}$$

Solución:

Para este caso utilizaremos la distribución de frecuencias para ***clases no reales***

1. *Ordena los* ***n*** *datos de menor a mayor: Los datos ya están ordenados*

2. ***rango de valores:***

$$rango = dato\ mayor - dato\ menor = 33 - 16 = 17$$

3. ***número de clases:***

$$k = \sqrt{81} = 9$$

4. ***Amplitud:***

$$amplitud = \frac{rango}{k} = \frac{17}{9} = 1.7 \approx 2$$

Esto se hace ya que se redondea al siguiente entero, decimal, etcétera dependiendo de los datos.

5. ***pre-cálculo:***

$$precalculo = dato\ menor + k \cdot amplitud = 16 + 9 \cdot 2 = 34$$

$$exceso = precalculo - dato\ mayor = 34 - 33 = 1$$

El valor de pre-cálculo supera el dato mayor en 3 unidades ***cantidad impar****, entonces buscamos dos enteros lo más cercanos posible, de tal forma que sea igual a 3.*

$$1 = 0 + 1$$

agarramos el 0

$$lim_{\,inf} = 16 - 0 = 16$$

Clases no reales	f_i	f_{ai}	$f_r\,\%$	$f_{rai}\,\%$
16 *a* 17	*3*	*3*	*3.7037%*	*3.7037%*
18 *a* 19	*6*	*9*	*7.40741%*	*11.1111%*
20 *a* 21	*8*	*17*	*9.87654%*	*20.9877%*
22 *a* 23	*14*	*31*	*17.284%*	*38.2716%*
24 *a* 25	*16*	*47*	*19.7531%*	*58.0247%*
26 *a* 27	*13*	*60*	*16.0494%*	*74.0741%*
28 *a* 29	*11*	*71*	*13.5802%*	*87.6543%*
30 *a* 31	*6*	*77*	*7.40741%*	*95.0617%*
32 *a* 33	*4*	*81*	*4.93827%*	*100%*
	$\sum f_i = n$		$\sum f_r\,\% = 100\,\%$	

Ejemplo 1.7.4

Un grupo de 50 empleados del departamento de contabilidad de una gran compañía recibe un curso intensivo de programación. De los varios ejercicios distribuidos durante el curso, he aquí el número de ejercicios completados satisfactoriamente por los miembros del grupo

$$\begin{pmatrix} 2 & 6 & 9 & 9 & 11 & 12 & 12 & 14 & 15 & 18 \\ 3 & 6 & 9 & 10 & 11 & 12 & 13 & 14 & 16 & 18 \\ 5 & 7 & 9 & 10 & 11 & 12 & 13 & 15 & 16 & 18 \\ 5 & 8 & 9 & 10 & 11 & 12 & 13 & 15 & 16 & 19 \\ 5 & 8 & 9 & 11 & 12 & 12 & 13 & 15 & 17 & 21 \end{pmatrix}$$

solución:*Para este caso utilizaremos la distribución de frecuencias para* ***clases no reales***

1. *Ordenar los* ***n*** *datos de menor a mayor*

$$\begin{pmatrix} 2 & 3 & 5 & 5 & 5 & 6 & 6 & 7 & 8 & 8 \\ 9 & 9 & 9 & 9 & 9 & 9 & 10 & 10 & 10 & 11 \\ 11 & 11 & 11 & 11 & 12 & 12 & 12 & 12 & 12 & 12 \\ 12 & 13 & 13 & 13 & 13 & 14 & 14 & 15 & 15 & 15 \\ 15 & 16 & 16 & 16 & 17 & 18 & 18 & 18 & 19 & 21 \end{pmatrix}$$

2. ***Rango de valores:***

$$rango = 21 - 2 = 19$$

3. ***número de clases:***

$$k = \sqrt{n} = \sqrt{50} = 7.071067812 \approx 7$$

4. ***amplitud:***

$$amplitud = \frac{rango}{k} = \frac{19}{7} = 2.714285714 \approx 3$$

Esto se hace ya que se redondea al siguiente entero,decimal, etcétera dependiendo de los datos

5. ***pre-cálculo:***

$$precalculo = dato\ menor + k \cdot amplitud = 2 + 7 \cdot 3 = 24$$

$$exceso = precalculo - dato\ mayor = 24 - 21 = 3$$

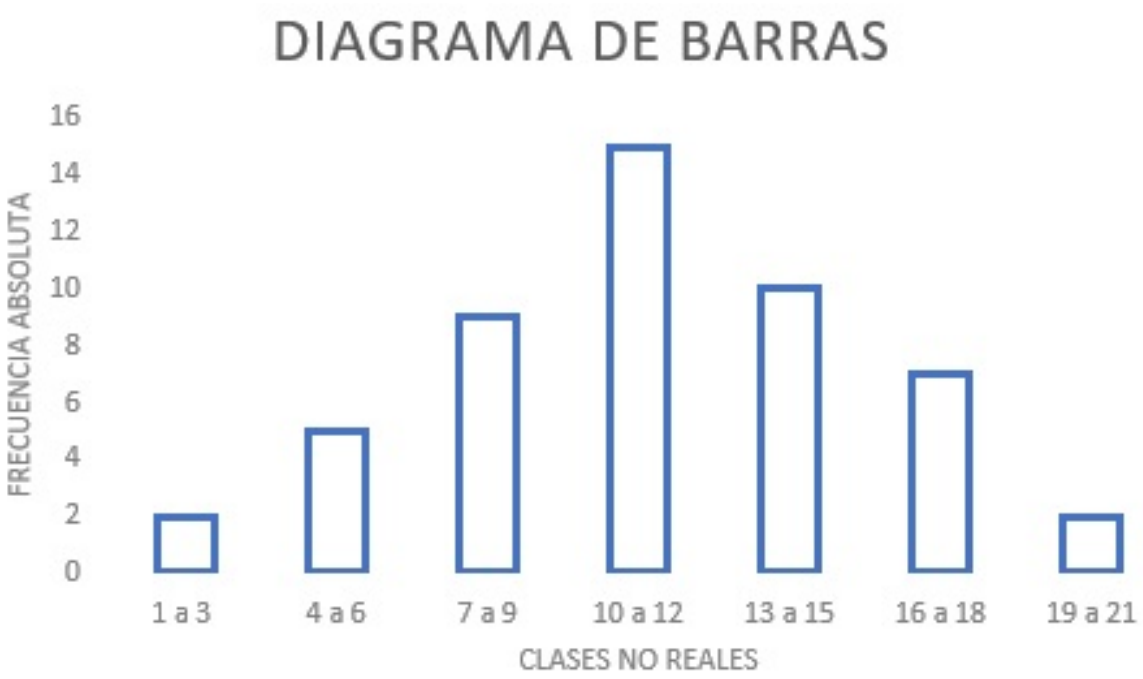
El valor de pre-cálculo supera al dato mayor en 3 unidades ***cantidad impar****, entonces buscamos dos enteros lo más cercanos posible, de tal forma que sea igual a 3.*

$$3 = 1 + 2$$

agarramos el 1

$$lim_{inf} = 2 - 1 = 1$$

Clases no reales	f_i	f_{ai}	$f_r\,\%$	$f_{rai}\,\%$
1 *a* 3	*2*	*2*	*4%*	*4%*
4 *a* 6	*5*	*7*	*10%*	*14%*
7 *a* 9	*9*	*16*	*18%*	*32%*
10 *a* 12	*15*	*31*	*30%*	*62%*
13 *a* 15	*10*	*41*	*20%*	*82%*
16 *a* 18	*7*	*48*	*14%*	*96%*
19 *a* 21	*2*	*50*	*4%*	*100%*
	$\sum f_i = n$		$\sum f_r\,\% = 100\,\%$	



Ejemplo 1.7.5

$$\begin{pmatrix} 6.00 & 6.00 & 6.25 & 6.25 & 6.25 \\ 6.25 & 6.25 & 6.50 & 6.50 & 6.50 \\ 6.50 & 6.50 & 6.50 & 6.50 & 6.65 \\ 6.65 & 6.70 & 6.70 & 6.75 & 6.75 \\ 6.75 & 6.75 & 6.75 & 6.75 & 6.75 \\ 6.75 & 6.75 & 7.00 & 7.00 & 7.00 \\ 7.00 & 7.00 & 7.00 & 7.00 & 7.10 \\ 7.10 & 7.15 & 7.15 & 7.25 & 7.25 \end{pmatrix}$$

Solución:

1. *Los datos ya están agrupados, y son* ***Datos continuos*** *con dos cifras.*

2. ***Rango:***

$$Rango = Dato_{menor} - Dato_{mayor} = 7.25 - 6 = 1.25$$

3. ***No. de clases:***

$$k = \sqrt{n} = \sqrt{40} = 6.3245 \approx 6$$

En el caso de ***no. de clases*** *recordar que se* ***REDONDEA*** *al entero más próximo*

4. ***Amplitud***

$$Amplitud = \frac{Rango}{k} = \frac{1.25}{6} = 0.20833 \approx 0.21$$

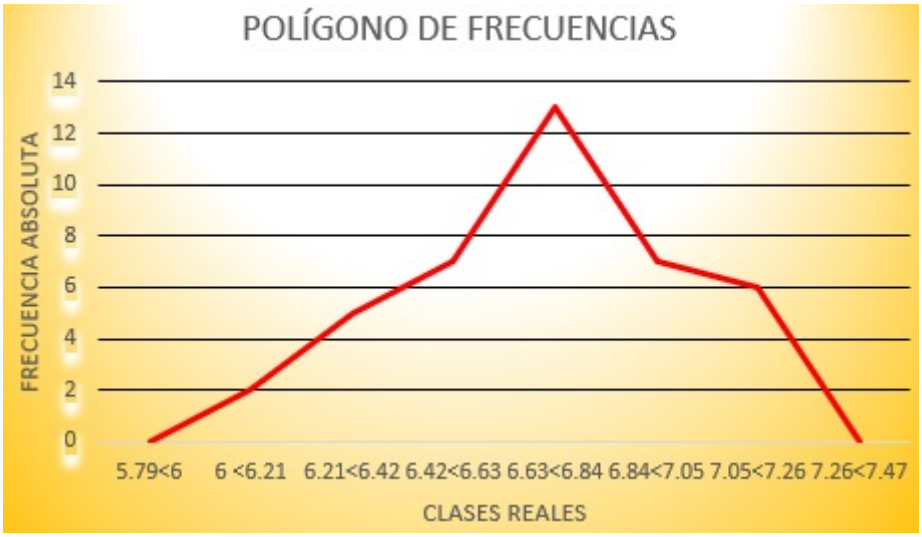
En este caso recordar que se ***APROXIMA*** *al siguiente entero, decimal, etc.* ***dependiendo de los datos***

5. ***Pre-cálculo***

$$\begin{aligned} precalculo &= Dato_{menor} + k \cdot Amplitud = 6 + 6 \cdot 0.21 = 7.26 \\ Exceso &= precalculo - Dato_{mayor} = 7.26 - 7.25 = 0.01 \\ cantidad\ impar\ 0.01 &\longrightarrow 0.01 = 0.00 + 0.01 \\ lin_{inf} &= 6.00 + 0.00 = 6.00 \end{aligned}$$

Como nuestros datos traen decimales, entonces en este caso lo conveniente es construir una ***Distribuciones de frecuencias con clases reales***

Clases reales	f_i	f_{ia}	f_r	f_{ra}	***marca de clase***
$6.00 < 6.21$	2	2	5 %	5 %	6.105
$6.21 < 6.42$	5	7	12.5 %	17.5 %	6.315
$6.42 < 6.63$	7	14	17.5 %	35 %	6.525
$6.63 < 6.84$	13	27	32.5 %	67.5 %	6.735
$6.84 < 7.05$	7	34	17.5 %	85 %	6.945
$7.05 < 7.26$	6	40	15 %	100 %	7.155
	$\sum f_i = n$		$\sum f_r$ %		



1.8. MEDIDAS DESCRIPTIVAS

Una razón importante para calcular estadísticas es describir y resumir un conjunto de datos. Un conjunto de números no es por lo general muy informativo, así que tenemos que encontrar la manera de abstraer la información clave que nos permite presentar los datos en una forma clara y compresible.
Dado un conjunto de datos, deseamos ahora resumirlos median te un número que describa a este. Este número es conocido como una **medida descriptiva**. Las medidas descriptivas se clasifican como:

1. Medidas de tendencia central
2. Medidas de variabilidad
3. Medidas de posición.

1.8.1. MEDIDAS DE TENDENCIA CENTRAL

Una medida de tendencia central es un número obtenido de un conjunto de datos, que tiende a posicionar al centro del conjunto de datos.
Este tipo de medidas son representativas del conjunto de datos.
Las medidas de tendencia central más importantes son:

1. **La media Aritmética o Promedio**

   a) Para el caso de un conjunto de **n** datos pertenecientes a una **muestra**, la media aritmética se denota como $\bar{x}$, y se define de la siguiente manera:

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

   donde x_i son los datos en consideración y **n es el tamaño de la MUESTRA**

   b) Para el caso de un conjunto de **N** datos pertenecientes a una **población** , la media aritmética se denota como μ, y se define de la siguiente manera:

   $$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

   Ventajas e inconvenientes:

   - La media aritmética viene expresada en las mismas unidades que la variable
   - En su cálculo intervienen todos los valores de la distribución.
   - Es el centro de gravedad de toda la distribución, representando a todos los valores observados
   - es única
   - Su principal inconveniente es que se ve afectada por los valores

2. **La moda**
   En este caso sin importar si los datos son de una muestra o de una población, la moda del conjunto de datos se denotará como Mo y se define de la siguiente manera:
   La **moda** es el dato que más se repite en el conjunto de datos en consideración.
   Ventajas e inconvenientes:

- Su cálculo es sencillo.
- Es de fácil interpretación
- Es la única medida de posición central que puede obtenerse en las variables de tipo cualitativo.
- En su determinación no intervienen todos los valores de la distribución.
- Puede ser que no exista, y en caso de que exista puede que no sea única. En caso de existir, puede ser unimodal, bimodal tridomal, etc.

3. **La mediana:**
La mediana, denotada como Me, es la medida de tendencia central que se posiciona justamente al centro de los datos: Para calcular el valor de la mediana se debe seguir la siguiente secuencia, sin importar si el conjunto de datos es de una muestra o de una población.

   a) Ordenar los datos de menor a mayor

      1) **Si n es impar:** Si el conjunto de datos en consideración es una cantidad impar, entonces la mediana será el valor del dato más central perteneciente al conjunto de datos ordenados, y se encuentra en la posición

$$\frac{(n+1)}{2}$$

      2) **Si n es par:** Si el conjunto de datos en consideración es una cantidad par, entonces la mediana será el valor del promedio de los dos datos más centrales.

   Ventajas o inconvenientes:

   - Es la medida más representativa en el caso de variables que solo admitan la escala ordinal
   - es fácil de calcular.
   - En la mediana sólo influyen los valores centrales y es insensible a los valores extremos.
   - En su determinación no intervienen todos los valores de la variable.

Ejemplo 1.8.1 *Determinar la media, moda y mediana de cada uno de los conjuntos de datos; pertenecientes ya sea una* ***muestra*** *o una* ***población***

1. $7, 11, 6, 9, 7$*: muestra*
Solución

   - ***La media aritmética*** *es:*

$$\bar{x} = \frac{7+11+6+9+7}{5} = \frac{40}{5} = 8$$

   - ***La moda*** *es:*

$$Mo = 7$$

   - ***La mediana****:*
Los datos ordenados son; $6, 7, 7, 9, 11$*. Como representan una cantidad impar(n=5), el dato central se encuentra el la posición 3,que en nuestro ejemplo es* 7*. Así el valor de la mediana es:*

$$Me = 7$$

2. $24, 18, 15, 33, 28, 20$*: población*

- ***La media aritmética*** *es (como en este caso es una población, entonces se representa como* μ *):*
$$\mu = \frac{24 + 18 + 15 + 33 + 28 + 20}{6} = 23$$
- ***La moda*** *no existe(porque recuerda que, la moda es el valor que más se repite):*
- ***La mediana****:*
Los datos ordenados son; $15, 18, 20, 24, 28, 33$*. Como el conjunto representa una cantidad par(n=6), el valor de la* ***mediana*** *será el promedio de los datos más centrales:*
$$Me = \frac{20 + 24}{2} = 22$$

3. $14, 23, 12, 14, 28, 22, 18, 25, 22, 14$*; población:*
Los datos ordenados son: $12, 14, 14, 14, 18, 22, 22, 23, 25, 28$

- ***La media aritmética*** *es:*
$$\mu = \frac{12 + 14 + 14 + 14 + 18 + 22 + 22 + 23 + 25 + 28}{10} = 19.2$$
- ***La moda*** *es:*
$$Mo = 14$$
- ***La mediana****:*
Como el conjunto representan una cantidad par,el valor de la mediana será:
$$Me = \frac{18 + 22}{2} = 20$$

4. 2.3,5.4,7.4,2.3,11.5,2.1,5.8,5.4,7.4*; muestra*
Los datos ordenados son: 2.1,2.3,2.3,5.4,5.4,5.8,7.4,7.4,11.5

- ***La media aritmética*** *es:*
$$\bar{x} = \frac{2.1 + 2.3 + 2.3 + 5.4 + 5.4 + 5.8 + 7.4 + 7.4 + 11.5}{9} = 5.5$$
- ***La moda*** *es trimodal:* 2.3,5.4,7,4
- ***La mediana :***
Como el conjutno de datos representa una cantidad impar, la ***mediana*** *será el valor del dato más central*

1.8.2. MEDIDAS DE VARIABILIDAD

Las medidas de variabilidad nos permite conocer qué tan dispersas se encuentran las observaciones a cada lado del centro en una serie de datos, o bien que tan alejadas se encuentran de la medida de tendencia central.
Las medidas de tendencia central se complementan con un análisis de variabilidad o de dispersión de datos.
Una medida de variabilidad es un número que indica el grado de dispersión (esparcimiento) en un conjunto de datos con respecto a un estadístico de tendencia central (por l general, la media aritmética).

- **Valor pequeño:** Si este valor es pequeño (con respecto de la unidad de medida), entonces hay una gran uniformidad de datos
- **Valor grande:** Si el valor es grande, entonces indica poca uniformidad
- **Valor cero:** Si el valor es cero, indica que todos los datos son iguales.

Las medidas de variabilidad más comunes, son:

- El rango (amplitud o recorrido).
- La desviación absoluta promedio.
- La varianza.
- La desviación estándar.
- La dispersión relativa: el cociente de variación.

1. **Rango (amplitud o recorrido)**:
   Es la medida más elemental de las medidas de variabilidad, y también es fácil de entender y calcular.
   El rango se clasifica como una medida de distancia.
   El recorrido se puede conocer con facilidad a partir de una muestra ordenada de tamaño n, en donde el rango o recorrido es la diferencia que hay entre el valor máximo y el valor mínimo

$$Rango = Dato\ mayor - Dato\ menor$$

   a) Como se analizan únicamente dos datos (el valor más grande y el valor más pequeño), no se obtiene información del comportamiento del resto que forma el conjunto (como se distribuyen los datos)

   b) Como no considera ninguna medida de tendencia central, entonces no informa nada acerca del comportamiento de los datos con respecto al centro

   El rango es un estadístico muy débil y sensible a los valores extremos en una serie de datos, por lo que su utilidad puede ser escasa en muchos caso prácticos e inadecuados si se utiliza el recorrido como medida de dispersión cuando el dato mayor o menor(o ambos) son valores extremos.

2. **Desviación absoluta promedio**
   Como ya se mencionó, la variabilidad de cualquier serie de datos se analiza en términos de la desviación de cada valor observado individual (x) con respecto algún valor central, como μ *o* $\bar{x}$
   Ahora la pregunta es: ¿ Cuánto varía cada dato con respecto a la media?
   La respues ta es: si las desviaciones de todos los datos son pequeñas con respecto a la media, entonces los datos son menos variables o están menos dispersos, a diferencia de cuando las desviaciones son grandes.
   Entonces $(x_i - \bar{x})$ proporciona información del grado de dispersión de una serie de datos. Para calcular la variabilidad es necesario establecer una formula con base en el promedio de las desviaciones de la serie de datos.

   Ejemplo 1.8.2 *Consideremos la serie* $2, 3, 5, 5, 7, 8$*, cuya media aritmética es cinco.*

Núm.	***Dato***	$d_i = (x_i - \bar{x})$
1	*2*	$2 - 5 = -3$
2	*3*	$3 - 5 = -2$
3	*5*	$5 - 5 = 0$
4	*5*	$5 - 5 = 0$
5	*7*	$7 - 5 = 2$
6	*8*	$8 - 5 = 3$
sumatoria	*30*	***cero***

Como se muestra, las desviaciones promedian cero. Esto querría decir que todos los datos sin iguales, lo cual es falso.
La única manera de solucionar este problema es tratar todas las desviación es negativas y positivas como si no tuvieran signo negativo (se tratan igual); esto es, obtener los valores absolutos de cada desviación como se muestra a continuación:

$$Promedio = \frac{(x_i - \bar{x})}{n} = \frac{|3| + |2| + |0| + |0| + |2| + |3|}{6} = \frac{10}{6} = 1.66$$

La desviación absoluta promedio, es una medida de dispersión promedio o desviación promedio, y es igual a la sumatoria de los valores absolutos de las desviaciones entre el total de los datos.
***Las desviaciones se definen** como la diferencie entre cada uno de los datos en el conjunto de estudio y el estadístico de tendencia central usado (media aritmética).*
La desviación absoluta promedio se define matemáticamente como:

$$DAP = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} = \frac{\sum_{i=1}^{n} |d_i|}{n}$$

***Ejemplo 1.8.3** Calcular la desviación absoluta promedio del conjunto de datos:* 3,4,5,6,6,7,8,9,9,10
Solución:

- *Determinar la media aritmética del conjunto de datos =*

$$\bar{x} = 6.7$$

- *Encontrar el valor absoluto de las desviaciones :*

$$|x_i - \bar{x}|$$

$|3-6,7| = 3.7$ $|4-6.7| = 2.7$ $|5-6.7| = 1.7$ $|6-6.7| = 0.7$ $|6-6.7| = 0.7$

$|7-6.7| = 0.3$ $|8-6.7| = 1.3$ $|9-6.7| = 2.3$ $|9-6.7| = 2.3$ $|10-6.7| = 3.3$

Calcular la media de las desviaciones cuyo resultado es la desviación absoluta promedio

$$DAP = \frac{3.7 + 2.7 + 1.7 + 0.7 + 0.7 + 0.3 + 1.3 + 2.3 + 2.3 + 3.3}{10} = \frac{19}{10} = 1.9$$

Si la distribución de estos datos es normal y simétrica, entonces 68 % de las observaciones queda comprendida entre $[\bar{x} - DAP]$ y $[\bar{x} + DAP]$, es decir, entre 4.8 *y* 8.6*. Prácticamente, en este conjunto, 5 de 10 datos (50 %) quedan comprendidos en este intervalo, por lo que **no** hay una dispersión muy grande de los datos alrededor de la media aritmética.*

3. **Varianza**
Esta es un estadístico que se define como el promedio de las desviaciones con respecto a la media, elevadas al cuadrado, y es similar a la desviación absoluta promedio; sin embargo, en este caso se elimina el uso del valor absoluto y se reemplaza por otra alternativa matemática que consiste en elevar al cuadrado todas estas, sean ahora positivas, lo que evita el uso del valor absoluto.
Cuando se utiliza la varianza como medida de variabilidad, resulta que el promedio obtenido de las desviaciones elevadas al cuadrado siempre serán unidades cuadradas.
Así, si el conjunto de datos está medido en kilogramos, la varianza de estos se medirá en kilogramos al cuadrado (kg^2); si es en años, la varianza se medirá en $años^2$

La varianza se simboliza matemáticamente

- Para la **población** con la letra griega

$$\sigma^2$$

Cuando se quiere calcular la varianza considerando todos los datos de una **población**, entonces la relación a utilizar es:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(|x_i - \mu|)^2}{N}) = \frac{\sum_{i=1}^{n}(|d_i|)^2}{N}$$

donde: N=número total de datos de la población.

- Para una **muestra** con

$$s^2$$

Matemáticamente, la varianza se calcula para una muestra mediante la relación siguiente:

$$s^2 = \frac{\sum_{i=1}^{n}(|x_i - \bar{x}|)^2}{n-1}) = \frac{\sum_{i=1}^{n}(|d_i|)^2}{n-1}$$

Donde n=número total de datos de la muestra.
En este caso se utiliza $n-1$ debido a las propiedades de los grados de libertad, es decir, si se conoce x se pierde un grado de libertad, por lo que sólo se necesita conoces $n-$ de los n términos para determinar la observación restante con una resta.

Ejemplo 1.8.4 *Calcular la varianza del conjunto de datos* $3, 4, 5, 6, 6, 7, 8, 9, 9, 10$ *pertenecientes a una muestra.*
Solución:

$$\bar{x} = 6.7$$

Núm.	***Dato***	$d_i = (x_i - \bar{x})$	$d_i^2 = (x_i - \bar{x})^2$
1	*3*	*3-6.7=-3.7*	*13.69*
2	*4*	*4-6.7=-2.7*	*7.29*
3	*5*	*5-6.7=-1.7*	*2.89*
4	*6*	*6-6.7=-0.7*	*0.49*
5	*6*	*6-6.7=-0.7*	*0.49*
6	*7*	*7-6.7=0.3*	*0.09*
7	*8*	*8-6.7=1.3*	*1.69*
8	*9*	*9-6.7=2.3*	*5.29*
9	*9*	*9-6.7=2.3*	*5.29*
10	*10*	*10-6.7=3.3*	*10.89*
Total	*67*	$\sum(x_i - \bar{x}) = 0$	*48.10*

$\therefore$ *el valor de la varianza es*

$$s^2 = \frac{13.69 + 7.29 + 2.89 + 0.49 + 0.49 + 1,69 + 5.29 + 5.29 + 10.89}{10 - 1} = \frac{48.10}{9} = 5.34 unidades^2$$

Esta medida indica que la dispersión de los datos con respecto a la media muestral es 5.34 $unidades^2$
Físicamente, el estadístico no dice mucho acer de la variabilidad de los datos; mientras la media aritmética se mide en unidades (lineales), la varianza se mide en unidades cuadradas.
Al respecto, cabe hacer la siguiente observación:
Si se quiere describir la variablilidad de un solo conjunto de datos, la varianza no es de gran ayuda porque esta no se expresa en la unidades originales, sino en unidades al cuadrado.
Una forma alternativa para calcular la varianza muestral, es:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2$$

4. **Desviación estándar**
Cuando la varianza se usa como una medida descriptiva, se puede considerar como un cálculo intermedio para obtener la desviación estándar (muestral o poblacional).
Este paso intermedio es necesario para regresar a la unidad original, el cual se logra al obtener la raíz cuadrada del valor de la variaza.
La desviación estándar es la medida de dispersión promedio más importante.
Debido a la dificultad de poder medir con la varianza el grado de dispersión en un conjunto de datos, se puede crear este nuevo estadístico, que es igual a la raíz cuadrada positiva de la varianza.
Es decir, en la desviación estándar, las unidades con las que se mide este nuevo estadístico de dispersión serán las mismas que tiene las observaciones y la media aritmética de estas.

- Así, **la desviación estándar para una muestra** es:

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

- **la desviación estándar para una población** es:

$$\sigma = \sqrt{\sigma^2}$$

Ejemplo 1.8.5 *Determinar la desviación estándar de la muestra:* $3, 4, 5, 6, 6, 7, 8, 9, 9, 10$
Solución:
Previamente obtuvimos que

$$s^2 = 5.34 \longrightarrow s = \sqrt{5.34} \approx 2.3108$$

5. **Coeficiente de variación**
La desviación estándar es una medida de variación absoluta que no permite concluir qué tan grande o pequeña es la dispersión de los datos; sin embargo, combinada con la media aritmética da origen a una medida de dispersión relativa llamada **coeficiente de variación** .

El coeficiente de variación (CV) es la medida relativa que permite tener una idea general de la magnitud de la desviación estándar en relación con la magnitud de la media aritmética.
Esta relación expresa la desviación estándar como porcentaje de la media aritmética, y sus unidades se miden en " por ciento "Ėl CV se puede expresar de manera matemática así:

- Para una **población** se calcula:

$$CV = \frac{\sigma}{\mu} x100\ \%$$

- Para una **muestra** se calcula:

$$CV = \frac{S}{\bar{x}} x100\ \%$$

El coeficiente de variación que es, simplemente, el cociente entre la desviación estándar y la media aritmética multiplicado por 100, es una medida relativa de dispersión ya que esa forma de cálculo implica que su valor indica que proporción de la media representa la desviación estándar.

Ejemplo 1.8.6

Determinar el CV de la muestra $3, 4, 5, 6, 6, 7, 8, 9, 10$

previamente se obtuvo que: $\bar{x} = 6.7$, $s \approx 2.3108$

$$\longrightarrow CV = \frac{2.3108}{6.7} x100\ \% = 34.489\ \%$$

$\longrightarrow$ *que la desviación estándar representa el* $34.489\ \%$ *de la media*

El CV también nos puede servir para comparar la variabilidad entre dos grupos de datos que tengan la misma o distinta de medida, (por ejemplo, un conjunto medido en metros y otro en kilogramos).

Ejemplo 1.8.7 *Se consideran los resultados de las calificaciones obtenidas en Estadística de dos grupos distintos, del sexto semestre, en la ESFM*

grupo	***Grupo A***	***Grupo B***
Media(promedio)	*8.2*	*9.5*
Desviación estándar	*0.5*	*0.5*

utilice el CV para determinar cuál fue el grupo con mejor aprovechamiento en estadística.
Solución:
Si solo se analiza la desviación estándar de los dos grupos se debe aceptar que la variabilidad de estos en lo que se refiere a las calificaciones obtenidas es la misma.
Sin embargo, las medias de los dos grupos son diferentes, lo que permite distinguir a cada grupo: es decir, el primer análisis de la desviación estándar debe revisarse.
Una forma de hacerlo es calcular el coeficiente de variación en ambos grupos.

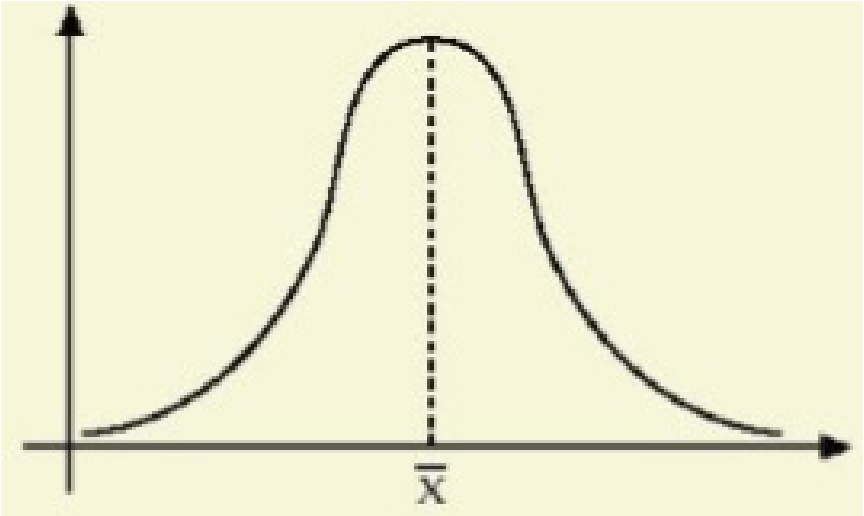
Grupo A	$CV = \frac{0.5}{8.2} \cdot 100\ \%$=*6.1 %*
Grupo B	$CV = \frac{0.5}{9.5} \cdot 100\ \%$=*5.2 %*

Como se sabe, en cualquier medida de variabilidad, a mayor valor, más variabilidad.
En este ejemplo, el grupo A presenta una variabilidad relativa mayor que la del grupo B; es decir, este último fue más homogéneo en lo relativo a rendimiento que el grupo A

1.8.3. MEDIDAS DE FORMA

Las medidas de forma son de gran utilidad para todas las personas que hagan uso de la estadística, ya que se puede describir la forma que toma una distribución de datos.
Las curvas que se utilizan para representar las observaciones de una serie de datos pueden ser simétricas o asimétricas (sesgadas).
Las curvas simétricas son aquellas que trazan una linea vertical desde el punto más alto de ella (la cima) hasta el eje horizontal.
El área de esta curva será divida exactamente en dos partes iguales, siendo la parte derecha espejo de la parte izquierda.
Por ejemplo:

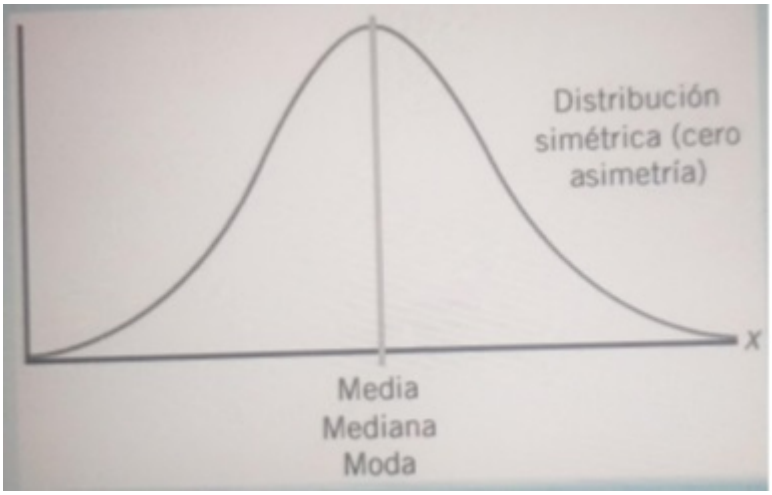
Una distribución simétrica es la curva de distribución norma o de campana



El sesgo permite comprender la relación de la media, la mediana y la moda, en una distribución de una sola cima o moda (unimodal)

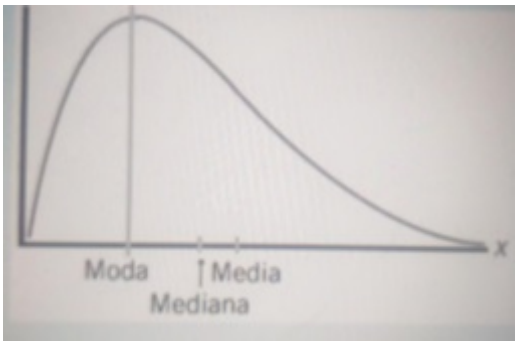
- **CURVA SIMÉTRICA:**

  Cuando la distribución es simétrica (no tiene sesgo), la media, la mediana y la moda se ubicarán en el centro de la distribución. En este caso, estas tiene el mismo valor $\mu = Me = Mo$, como se muestra en la figura



CURVA SESGADA HACIA LA DERECHA :

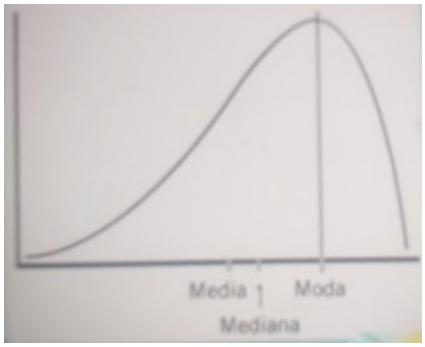
Por otro lado, las curvas de las siguientes figuras son sesgadas porque los valores de sus distribuciones de frecuencias se concentran en el extremo inferior o en el extremo superior de la escala de medición situada sobre el eje delas abscisas, por tanto, los valores no pueden tener una distribución igual(simétrica)

- 

Esta curva esta sesgada hacia la derecha o con simetría positiva, lo que se debe a que disminuye de manera gradual hacia el extremo superior de la escala.
En esta curva, la moda es el punto más alto, la mediana es el punto medio, mientras que la media siempre tiende a ubicarse hacia la cola (derecha) de la distribución.
Esto se debe a que la media siempre se afectará por los valores extremos.
En este caso, la media, la mediana y la moda tienen diferentes valores.

$$(\mu > Me > Mo)$$

- ## CURVA SESGADA HACIA LA IZQUIERDA:



Sucede lo contrario en esta curva. Esta tiene sesgo a la izquierda o asimetría negativa, ya que disminuye de forma gradual el extremo inferior de la escala.
En esta curva, la moda es el punto más alto, la media es el medio, mientras que la media siempre tenderá a ubicarse hacia la cola (izquierda) de la distribución.
En este caso, la media, la media y la moda tienen diferentes valores.

$$(\mu < Me < Mo)$$

COEFICIENTE DE SESGO

La medida estadística que cuantifica el sesgo de un conjunto de datos se llama coeficiente de sesgo (CS)
El CS se define y se denota mediante la relación siguiente

$$CS = \left[\frac{n}{(n-1)(n-2)}\right]\left[\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3\right]$$

Donde
x_i=observación (dato) i
n=número total de datos en la muestra
$\bar{x}$=media aritmética de la muestra.
s= desviación estándar de la muestra
En el cálculo del coeficiente de sesgo, la interpretación debe realizarse de la siguiente manera:

- Si CS=0 entonces los datos (de la curva) se distribuyen de manera simétrica
- Si $CS > 0$, entonces los datos (de la curva) son sesgados a la derecha
- Si CS <0, entonces los datos (de la curva) son sesgados a la izquierda

COEFICIENTE DE SESGO DE FISHER

Observemos que el cálculo del CS es muy tedioso, sobre todo, cuando el conjunto de datos es grande.
Una alternativa para esto, es encontrar el **coeficiente de sesgo de Fisher**, denotado como CAS (coeficiente de asimetría).
Este coeficiente de asimetría se define como:

- Para **muestra**
$$CAS = \frac{3(\bar{x} - Me)}{S}$$
- Para **población**
$$CAS = \frac{3(\mu - Me)}{\sigma}$$
- Si el valor del CAS =0, diremos que es simétrica.
- Si el valor del $CAS < 0$, es sesgada hacia la izquierda (asimétrica)
- Si el valor del $CAS > 0$ es sesgada hacia la derecha (asimétrica)

El valor del CAS cae en el rango $[-3, 3]$
Si el valor de CAS cae en el rango $[-0.1, 0.1]$, diremos que la forma de la distribución es casi simétrica.

Ejemplo 1.8.8

Determinar la forma de la distribución del siguiente conjunto de los datos pertenecientes a una muestra

$$\begin{pmatrix} 6 & 6,5 & 6,65 & 7 & 7,25 \\ 6,1 & 6,5 & 6,7 & 7 & 7,25 \\ 6,25 & 6,5 & 6,7 & 7 & 7,3 \\ 6,5 & 6,5 & 6,75 & 7,1 & 7,5 \\ 6,25 & 6,5 & 6,75 & 7,15 & 8 \\ 6,25 & 6,5 & 7 & 7,15 & 15 \\ 6,5 & 6,65 & 7 & 7,2 & 20 \end{pmatrix}$$

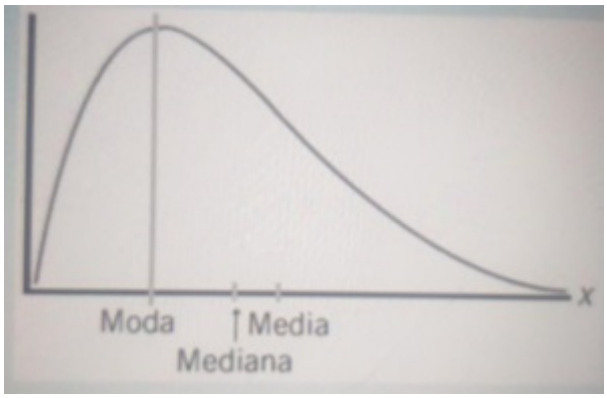
Solución: *En este caso,*

$$\bar{x} = 7,325 \quad Mo = 6,5 \quad Me = 6,75$$

Como

$$\bar{x} > Me > Mo$$

Entonces, la forma de la distribución es Asimétrica, sesgada hacia la derecha.



Determinar el valor del CAS para los datos de la tabla (siendo pertenecientes a una muestra) , e indique la forma de la distribución

$$CAS = \frac{3(\bar{x} - Me)}{S}$$

$$CAS = \frac{3(\bar{x} - Me)}{S} = \frac{3(7,325 - 6,75)}{2,4707} = 0,69818 > 0$$

Ejemplo 1.8.9

En una clase de español con 42 *alumnos, se propuso la aplicación de una prueba de compresión de* 100 *palabras. Las puntuaciones resultantes del examen son las siguientes:*

27	39	41	46	48	48	54	57	57	57	59	59	60	61
61	61	62	63	64	64	64	65	65	66	67	67	67	68
68	68	68	69	71	72	72	75	76	76	78	80	86	94

1. *Elaborar una* ***distribución de frecuencia con clases reales****, que contenga a,* f_i *,* $f_i a$ *,*f_r*,* $f_r a$
   Solución: *Para este caso el problema nos pide que hagamos una* ***distribución de frecuencias para clases reales***

   a) *Ordenar los* ***n*** *datos de menor a mayor.*
   Los datos ya están ordenados

   b) ***Rango de valores:***

   $$Rango = Dato_{mayor} - Dato_{mayor} = 94 - 27 = 67$$

   c) ***No. de clases:***

   $$k = \sqrt{n} = \sqrt{42} = 6.4807 \approx 6$$

   Recordando que el número de clases se ***RENDONDEA al entero más próximo***

d) ***Amplitud***

$$Amplitud = \frac{Rango}{k} = \frac{67}{6} = 11.166 \approx 12$$

Recordando que se ***APROXIMA*** *al siguiente entero, decimal, etc. dependiendo de los datos, y como en este caso los datos son ENTEROS, por eso* 11.166 *se aproxima al siguiente entero, que sería* 12

e) ***Pre-cálculo***

$$precalculo = Dato_{menor} + k(amplitud) = 27 + 6(12) = 99$$
$$Exceso = precalculo - Dato_{mayor} = 99 - 94 = 5$$

Como 5 *es una* ***cantidad impar*** *entonces encontramos dos enteros, lo más cercanos posibles.*

$$5 \longrightarrow 5 = 2 + 3$$

De esta forma, el valor del límite inferior de la primera clase será:

$$lim_{inf} = Dato_{menor} - 2 = 27 - 2 = 25$$

Como es para clases reales, entonces el lim_{sup} *de la primera clase, nos queda*

$$lim_{sup} = lim_{inf} + 12 = 25 + 12 = 37$$

Clases Reales	f_i	f_{ia}	f_r	f_{ra}
$25 < 37$	1	1	$\frac{50}{21}\ \%$	$\frac{50}{21}\ \%$
$37 < 49$	5	6	$\frac{250}{21}\ \%$	$\frac{300}{21}\ \%$
$49 < 61$	7	13	$\frac{50}{3}\ \%$	$\frac{650}{21}\ \%$
$61 < 73$	22	35	$\frac{1100}{21}\ \%$	$\frac{1750}{21}\ \%$
$73 < 85$	5	40	$\frac{250}{21}\ \%$	$\frac{2000}{21}\ \%$
$85 < 97$	2	42	$\frac{100}{21}\ \%$	$100\ \%$
	$\sum f_i = n$		$\sum f_r = 100\ \%$	

2. Para los datos del problema, determinar

a) *La media aritmética del grupo*
En este caso tenemos que el enunciado nos indica de una ***POBLACIÓN****, entonces con esto calculamos la* ***media poblacional****:*

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{27 + 39 + 41 + \cdots + 80 + 86 + 94}{42} = 63.57142857 \approx 63.57$$

b) *La varianza del grupo*
Calcularemos la ***Varianza poblacional***

$$\begin{aligned}\sigma &= \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \\ &= \frac{(27 - 63.57)^2}{42} + \frac{(39 - 63.57)^2}{42} + \cdots + \frac{(86 - 63.57)^2}{42} + \frac{(94 - 63.57)^2}{42} \\ &= 148.5782313 \approx 148.58\end{aligned}$$

3. *Determinar la forma de la distribución, utilizando tres maneras distintas para esto:*
Solución:

a)

$$\mu = 63.57$$

Para calcular la mediana, recordemos que es la medida de tendencia central que se posiciona justamente al centro de los datos. En este caso n es par (42)*, entonces la mediana será el valor promedio de los datos más centrados.*

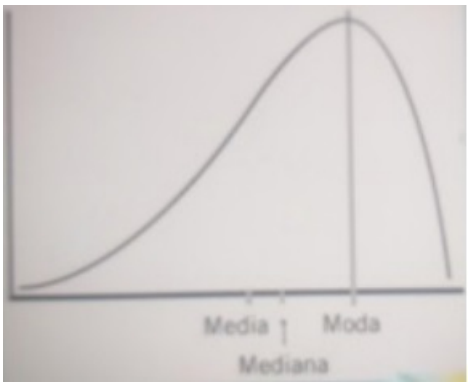
$$Me = \frac{x_{21} + x_{22}}{2} = \frac{64 + 65}{2} = 64.5$$

$$Mo = 68$$

De lo anterior tenemos que:

$$Mo > Me > \mu$$
$$\mu < Me < Mo$$
$$63.57 < 64.5 < 68$$

$\therefore$ ***CURVA SESGADA HACIA LA IZQUIERDA***

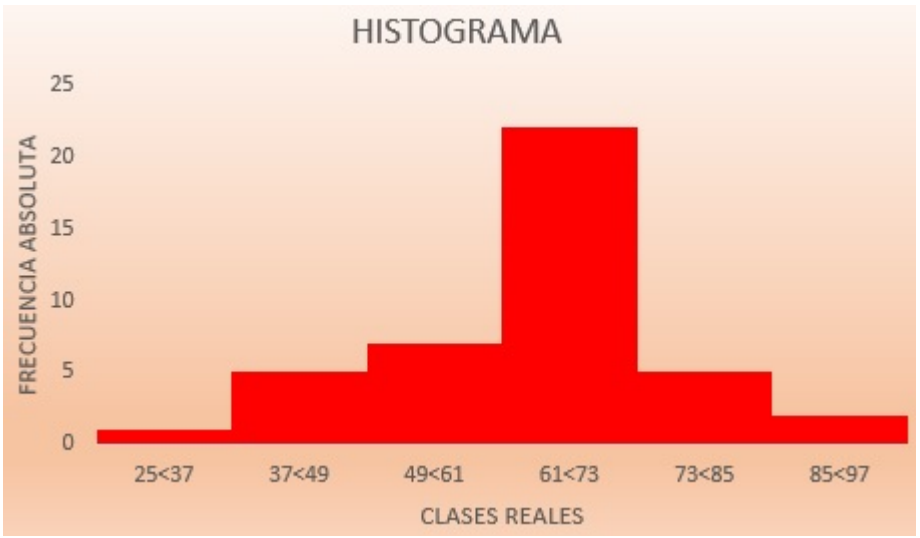


b) ***COEFICIENTE DE SESGO FISHER***
Tenemos que para una ***población***

$$CAS = \frac{3(\mu - Me)}{\sigma} = \frac{3(63.57 - 64.5)}{12.19} = -0.2288 < 0$$

$\therefore$ ***CURVA SESGADA HACIA LA IZQUIERDA***

c) ***HISTOGRAMA***



Por el HISTOGRAMA también nos podemos dar cuenta que es una ***CURVA SESGADA HACIA LA IZQUIERDA***

4. *¿ Qué porcentaje representa la desviación estándar de la media aritmética ?*
   Para este caso utilizaremos el ***COEFICIENTE DE VARIACIÓN*** *, como es una* ***población*** *nuestro caso, entonces este CV se calcula de la siguiente manera:*

$$CV = \frac{\sigma}{\mu} \cdot 100\,\% = \frac{12.19}{63.57} \cdot 100\,\% = 19.17\,\%$$

5. *Demuestre que si* $y_i = kx_i$*, entonces* $\sigma_y^2 = k^2\sigma_x^2, i = 1, 2, \ldots, n$
   Solución:
   Supongamos que $\sigma_y^2 \neq k^2\sigma_x^2$
   Tenemos que:

$$y_1 = kx_1, y_2 = kx_2, y_3 = kx_3, \ldots, y_n = kx_n$$

entonces, por definición de ***media aritmética poblacional***

$$\mu_y = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{n} kx_i}{n} = k\mu_x$$

por definición de ***varianza poblacional***

$$\sigma_y^2 = \frac{\sum_{i=1}^{n}(y_i - \mu_y)^2}{n} = \frac{\sum_{i=1}^{n}(kx_i - k\mu_x)^2}{n} = k^2 \left[\frac{\sum_{i=1}^{n}(x_i - \mu_x)^2}{n} \right]$$

$\therefore$ ***por reducción al absurdo*** *tenemos que* $\sigma_y^2 = k^2\sigma_x^2$

Capítulo 2

INFERENCIA ESTADÍSTICA

2.1. MUESTREO CON O SIN REEMPLAZO DE UNA POBLACIÓN FINITA

El hecho de regresar o no un elemento muestreado a la población antes de extraerotro elemento de esa misma población determina si el muestreo es con o sin remplazo.
El muestreo es con remplazo si en una extracción el elemento extraído sigue participando en otra u otras extracciones posteriores; el muestreo es sin remplazo si el elemento extraído ya no participa en otra u otras extracciones posteriores.
Si la población es muy grande, al extraer varios elementos, uno por uno, el tipo de muestreo es irrelevante.
De hecho, al reemplazar el primer elemento antes de extraer el segundo, las observaciones en la primera y en la segunda extracción serian totalmente independientes.
No obstante, si no se reemplaza el primer elemento, el segundo resultado de la extracción afectará ligeramente la segunda extracción. En poblaciones pequeñas el efecto si es relevante.
El desarrollo matemático es mas sencillo si las observaciones son independientes. A lo largo del muestreo se va a suponer el muestreo aleatorio con reemplazo, el cual también suele llamársele **muestreo aleatorio simple**
Nota:
El número de elementos de la población, llamado tamaño de población, lo representamos con la letra N.
El número de elementos de la muestra, llamado tamaño de muestra, lo representamos con la letra n.
El cálculo del número de muestras posibles de tamaño n, extradías de una población tamaño N, cuando el muestreo es con reemplazo (se permite repetición de elementos y permutaciones de elementos: 1 1, 1 2, 2 1,, etc.) se determina mediante la siguiente expresión:

$$N^n$$

Por otro lado, cuando el muestreo es si remplazo (no se permite repetición de elementos de elementos ni la permutación de elementos: 1 2 = 2 1), el número de muestras posibles que se pueden obtener es;

$$C_{n,N} = \binom{N}{n} = \frac{P_{n,N}}{n!} = \frac{N!}{n!(N-n)!}$$

nota:
No todas las muestras se presentan a generalizaciones válidas acerca de las poblaciones de las cuales provinieron.
De hecho la mayoría de los métodos de inferencia están basados en la suposición de que se

manipulan **muestras aleatorias**
En la práctica, a menudo tratamos con muestras aleatorias de poblaciones finitas, pero lo suficientemente grandes para tratarlas como si fuesen infinitas.
Por consiguiente, la mayor parte de la teoría estadística y la mayoría de los métodos que analizaremos se aplican a muestras tomadas en poblaciones finitas.
nota:
No todas las muestras se presentan a generalizaciones válidas acerca de las poblaciones de las cuales provinieron.
De hech, la mayoría de los métodos de inferencia están basados en la suposición de que se manipulan **muestras aleatorias.**
En la práctica, a menudo tratamos con muestras aleatorias de poblaciones finitas, pero lo suficientemente grandes para tratarlas como si fuesen infinitas.
Por consiguiente, la mayor parte de la teoría estadística y la mayoría de los métodos que analizaremos se aplican a muestras tomadas de poblaciones infinitas.

Definición 2.1.1 (Muestra aleatoria)

Si $x_1, x_2, x_3, \ldots, x_n$ son datos o variables aleatorias independientes o idénticamente distribuidas, decimos que constituyen una ***muestra aleatoria*** *de la población infinita dado por su distribución común.*

nota:

1. A lo largo del tiempo se ha aplicado muestra aleatoria a los valores de las v.a en vez de las v.a mismas.

2. Como los estadísticos son v.a sus valores varían de una muestra a otra y se acostumbra a llamar a sus distribuciones, **distribuciones muéstrales**

En esta sección, se dedicará a las distribuciones muestrales de valores estadísticos que desempeñan papeles importantes en las aplicaciones.

Definición 2.1.2 (Distribuciones muestrales de la media)

De acuerdo a lo observado en el tema anterior, el valor de la media muestral $\bar{X}$ varía de una muestra a la otra. Por lo cual $\bar{X}$, además de ser un estimador, ***es una variable aleatoria.*** *De esta forma $\bar{X}$, cuenta con una media, una desviación estándar y una distribución de probabilidad*

Definición 2.1.3 (Distribución muestral de la media)

Una ***Distribución muestral de media*** *es una distribución probabilística que consta de una lista de todas las posibles medidas posibles muestrales extraídas de una población, obtenidas de un muestreo con o sin remplazo y de un tamaño específico, y esta lista acompañada de la probabilidad de ocurrencia asociada con cada media muestral.*
*En este caso, la media de las medias muéstrales (****media esperada****), es ahora la media de todos los posibles valores dfe $\bar{X}$ y se denota por*

$$\mu_{\bar{X}}$$

Está corresponde al parámetro de la distribución de $\bar{X}$
La **desviación estándar de las medias muestrales** $\bar{X}$, se denota por

$$\mu_{\bar{X}}$$

Para ilustrar el concepto de la distribución muestral de medias muestrales, consideremos los siguientes ejemplos

Ejercicio 2.1.1

Consideremos una ***población*** *de solo tres valores*

$$x_1 = 1 \ , \ x_2 = 2 \ , \ x_3 = 3$$

cuya medida y desviación estándar poblacional son

$$\mu = 2 \ \ y \ \ \sigma = 0.8165$$

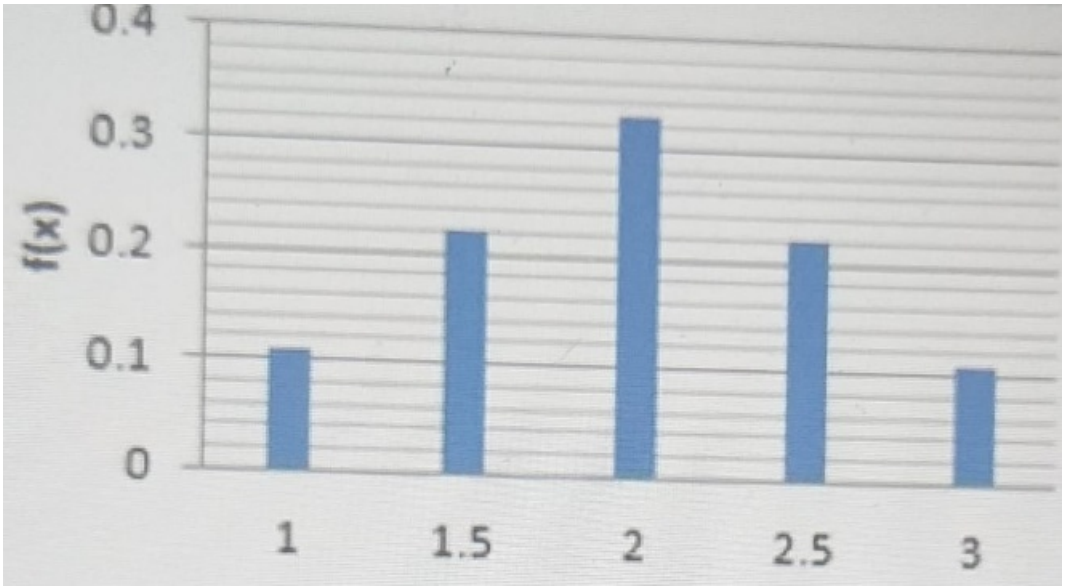
1. Obtener todas las posibles muestras de tamaño 2, con reposición que se puedan extraer a partir de la población dada. Además crear la distribución de las medias muestrales
   Solución:
   El número de todas las muestras posibles de tamaño 2, con reposición, que se pueden extraer de dicha población obviamente son $3^2 = 9$. Estas muestras y su distribución se muestran a continuación.

Muestra	$\bar{X}$
$m_1 = 1,1$	1.0
$m_2 = 1,2$	1.5
$m_3 = 1,3$	2.0
$m_4 = 2,1$	1.5
$m_5 = 2,2$	2.0
$m_6 = 2,3$	2.5
$m_7 = 3,1$	2.0
$m_8 = 3,2$	2.5
$m_9 = 3,3$	2.0
	$\sum 18$

$\bar{X}$	$f(\bar{X})$	$\bar{X}f(\bar{X})$	$\bar{X}^2 f(\bar{X})$
1.0	$\frac{1}{9}$	$1 \cdot \frac{1}{9} = \frac{1}{9}$	$(1)^2 \cdot \frac{1}{9}$
1.5	$\frac{2}{9}$	$\frac{3}{2} \cdot \frac{2}{9} = \frac{1}{3}$	$(\frac{3}{2})^2 \cdot \frac{2}{9} = \frac{1}{2}$
2.0	$\frac{3}{9}$	$2 \cdot \frac{3}{9} = \frac{2}{3}$	$(2)^2 \cdot \frac{3}{9} = \frac{12}{9}$
2.5	$\frac{2}{9}$	$\frac{5}{2} \cdot \frac{2}{9} = \frac{5}{9}$	$(\frac{5}{2})^2 \cdot \frac{2}{9} = \frac{25}{18}$
3.0	$\frac{1}{9}$	$3 \cdot \frac{1}{9} = \frac{1}{3}$	1
	$\sum = 1$	$\mu_{\bar{X}}$	$\frac{13}{3}$

   Observemos que la mayoría de los valores de las medias muestrales individuales difiere del valor de la media poblacional.
   En general se puede afirmar que sin importar de la población que se tenga, las medias muestrales tenderán a estar cerca de media poblacional y muy rara vez tendrán el mismo valor.

2. Hacer la gráfica de la distribución de medias muestrales e identifique lo observado.



Observemos que la gráfica es simétrica.

3. Calcular μ_X y $\sigma_{\bar{X}}$ e indique lo que se observa.

$$\mu_{\bar{X}} = \sum \bar{x} f(\bar{x}) = 2$$

Con lo cual observamos nuevamente que la media de medias muestrales es igual al valor de la media poblacional

$$\mu_{\bar{X}} = \mu$$

Lo cual no ocurre con la desviación estándar de medias muestrales;

$$\sigma_{\bar{X}} = \sigma$$

En este caso guardan la siguiente relación:

$$\sigma_{\bar{X}} \sqrt{n} = \sigma$$

$$\longrightarrow \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

En la expresión anterior se observa que la desviación estándar de medias muestrales disminuye en la medida en que el tamaño de la muestra crece.

Ejemplo 2.1.1 *Resolver el ejercicio 1, para el cso en que la muestra sea sin reposición.*
Solución:
En este caso, el número de muestra sin remplazo, son

$$\binom{3}{2} = 3$$

Todas las posibles muestras de tamaño 2, son:

$$1, 2; 1, 3 \; y \; 2, 3$$

Y la distribución de probabilidad de las medias muestrales es:

muestra	$\bar{X}$	$f(\bar{X})$	$\bar{X}f(\bar{X})$	$\bar{X}^2 f(\bar{X})$
$m_1 = 1, 2$	$\frac{3}{2}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{3}{4}$
$m_2 = 1, 3$	2	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{4}{3}$
$m_3 = 2, 3$	$\frac{5}{2}$	$\frac{1}{3}$	$\frac{5}{6}$	$\frac{25}{12}$
		$\sum 1$	*2*	$\frac{25}{6}$

La gráfica de la distribución presenta iuna distribución uniforme
Por lo cual, esta distribución también es simétrica. Además

$$\mu_{\bar{X}} = 2 \;\; y \;\; \mu_{\bar{X}} = 0.40825$$

Con lo cual observamos nuevamente que la media de medias muestrales es igual al valor de la media poblacional

$$\mu_{\bar{X}} = \mu$$

Lo cual no ocurre con la desviación estándar de medias muestrales

$$\sigma_{\bar{X}} \neq \sigma$$

En este caso guardan la siguiente relación

ERROR ESTÁNDAR DE UNA DISTRIBUCIÓN MUESTRAL

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \cdot \left(\frac{N-n}{N-1} \right)$$

CONCEPTO DE ERROR ESTÁNDAR

Definición 2.1.4

En vez de hablar de la desviación estándar de medias muestrales para describir a la distribución de las medias muestrales, los estadísticos hablan del ***error estándar de la media***

ERROR ESTÁNDAR DE LA MEDIA $\sigma_{\bar{X}}$

El error estándar de la distribución de las medias muestrales mide el grado en que esperamos que las medias de las diferentes muestras varíen por este error accidental en el proceso de muestreo.
Por consiguiente la desviación estándar de la distribución de un estadístico muestral recibe el nombre de ERROR ESTÁNDAR DEL ESTADÍSTICO .
El error estándar índica no solo el tamaño de error accidental que se ha cometido, sino además la exactitud que seguramente alcanzaremos si usamos un estadístico muestral para estimar un parámetro de la población.

Ejemplo 2.1.2

Supongamos que en determinada población se ha seleccionado una muestra con ***reposición*** *de tamaño* $n = 10$ *con* $\sigma_X = 9$. *¿ Cuántas observaciones más necesitamos para reducir el valor de* σ_X, *a*

1. a) 4.5

Solución *Tomando en cuenta que* $\sigma_X = \frac{\sigma}{\sqrt{n}}$*, despejando* σ *obtenemos*

$$\sigma = \sqrt{n}\sigma_X = \sqrt{10}(9) \approx 28.4605$$

Por otro lado, despejando n *de la fórmula de error estándar obtenemos*

$$n = \frac{\sigma^2}{\sigma_X^2}$$

si $\sigma_X = 4.5$

$$n = \frac{(28.46)^2}{(4.5)^2} = 40$$

Por lo tanto, *para reducir de* 9 *a* 4.5*, el tamaño inicial de la muestra se debe* ***aumentar en 30 unidades***

b) 3

Solución:

Si $\sigma_X = 3$

$$n = \frac{(28.66)^2}{3^2} = 89.996$$

Por lo tanto, *para reducir de 9 a 3, el tamaño inicial en la muestra se debe* ***aumentar en 80 unidades***

c) 1

Solución:

Si $\sigma_X = 1$

$$n = \frac{(28.66)^2}{1^2} = 810$$

Por la tanto, *para reducir de 9 1, el tamaño inicial en la muestra se debe* ***aumentar en 800 unidades***

TAMAÑO DE UNA MUESTRA PEQUEÑA

Diremos que una muestra es una muestra pequea si el tamaño de esta es menor que 30

$$n < 30$$

TAMAÑO DE UNA MUESTRA GRANDE

Diremos que una muestra es una muestra grande si el tamaño de esta es mayor o igual que 30

$$n \geq 30$$

DEFINICIONES DE PROBABILIDAD

Definición 2.1.5 *El r-ésimo momento y s-ésimo momento producto con respeto al origen de las v.a X e Y, representado por $\mu'_{r,s}$ es el valor esperado de X^rY^s*

$$\mu'_{r,s} = E\left[X^rY^s\right]$$

Definición 2.1.6 *El r-ésimo y s-ésimo momento producto con respecto a las medias relativas de las v.a X e Y, denotado por $\mu_{r,s}$ es el valor esperado de $(X-\mu_X)^r(Y-\mu_Y)^s$*

$$\mu_{r,s} = E\left[(X-\mu_X)^r(Y-\mu_Y)^s\right]$$

Definición 2.1.7 *μ'_{i1}recibe el nombre de covarianza de X e Y, y se representa por medio de σ_{XY} o $cov(X,Y)$*

Teorema 2.1.1

$$\sigma_{XY} = \mu'_{i1} - \mu_X\mu_Y$$

Teorema 2.1.2 *Si X e Y son v.a independientes, entonces:*

- $E[X,Y] = E[X]E[Y]$
- $Cov(X,Y) = \sigma_{XY} = 0$

Teorema 2.1.3 *Si $X_1, \ldots xX_n$ son v.a y $Y = \sum_{i=1}^{n} a_ix_i$ donde $a_1, \ldots, a_n$ son constantes, entonces:*

- $E(Y) = \sum_{i=1}^{n} a_iE(x_i)$
- $Var(Y) = \sum_{i=1}^{n} a_i^2 \cdot Var(x_i) + 2\sum_i \sum_j a_ia_j \cdot cov(x_i, y_i)$
  donde la doble sumatoria se extiende sobre todos los valores de i y j de 1 a n, para los cuales $i < j$

Colorario 2.1.1 *Si las v.a x_i e y_i son independientes y $Y = \sum_{i=1}^{n} a_ix_i$ entonces:*

$$VAR(Y) = \sum_{i=1}^{n} a_i^2 Var(x_i)$$

Teorema 2.1.4 *Si $X_1, \ldots, X_n$ son v.a independientes y $Y = x_1 + x_2 + \ldots + x_n$ entonces:*

$$M_y(t) = \Pi_{i=1}^{n} M_{xi}(t)$$

donde $M_{xi}(t)$ es el valor de la función generadora de momento de x_i en t.

Teorema 2.1.5 *Si $X_1, \ldots, X_n$ constituyen una muestra aleatoria de una población infinita que tiene media μ y varianza σ^2, entonces:*

$$E[\bar{X}] = \mu_X = \mu$$
$$Var(\bar{X}) = \sigma_X^2 = \frac{\sigma^2}{n}$$

Demostración:
Del teorema, $E(Y) = \sum_{i=1}^{n} a_i E(x_i)$*, como un caso particular, haciendo* $Y = \bar{x}$ *y* $a_1 = \frac{1}{n}$ *(porque* $\bar{x} = \frac{\sum x_i}{n}$*) se obtiene:*

$$\mu_x = E[\bar{X}] = \sum_{i=1}^{n} \frac{1}{n} \cdot \mu = n\left(\frac{1}{n} \cdot \mu\right) = \mu$$

Como $E[x_i] = \mu$*,* $Var(x_i) = \sigma^2$*, entonces por el colorario (1) se tiene que:*

$$Var(\bar{x}) = \sum_{i=1}^{n} \frac{1}{n^2} \cdot \sigma^2 = n\left(\frac{1}{n^2} \cdot \sigma^2\right) = \frac{\sigma^2}{n}$$

Teorema 2.1.6 *Sea* X *una v.a Normal con media* μ *y varianza* σ^2*. Si* $\bar{x}$ *es la media de una muestra aleatoria de tamaño n tomada de esta población. la distribución muestral de* $\bar{x}$ *es una distribución con media* μ *y varianza* $\frac{\sigma^2}{n}$

OBSERVACIÓN

Cuando se tiene un tamaño de muestra grande, sin importar como sea la distribución de la población, la distribución de $\bar{x}$ será aproximadamente normal y para la aplicación de sus cálculos se considera como la **distribución normal**

TEOREMA

Si $x_1, \ldots, x_n$ constituyen una muestra aleatoria de una población infinita que tiene media μ y varianza σ^2, entonces la distribución de la v.a $\bar{x}$ estandarizada

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

Tendrá una distribución norma con $\mu_x = \mu$ y $\sigma_x = \frac{\sigma}{\sqrt{n}}$ cuando $n \longrightarrow \infty$. Más concretamente, diremos que la distribución de $\bar{x}$

1. ## TENDRÁ UNA DISTRIBUCIÓN NORMAL

   sin importar el tamaño de la muestra, siempre y cuando la población original de los datos sea de tipo normal.

2. ## TENDRÁ UNA DISTRIBUCIÓN APROXIMADAMENTE NORMAL

   cuando la población original de datos no tenga una distribución normal, peroo las muestras seleccionadas sean de tamaño mayor o igual a 30.
   Los resultados vistos anteriormente, respecto a una distribución de madias muestrales, se encuentran resumidos en el siguiente teorema:

2.2. TEOREMA DEL LÍMITE CENTRAL

Consideremos una población de datos con media μ y desviación estándar σ. Si está población se extraen todas las diferentes muestras de tamaño **n**, entonces la **distribución muestral de medias:**

a) Tendrá una media esperada $\mu_x = \mu$ sin importar el tipo de muestrep ni el tipo de población que se esté trabajando.

b) Tendrá un error estándar, expresado por medio de la fórmula.

1) Si el muestreo se hace **CON REEMPLAZO**

$$\sigma_X = \frac{\sigma}{\sqrt{n}}$$

2) Si el muestreo es **SIN REEMPLAZO**

$$\sigma_X = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

. En este caso, el factor $\sqrt{\frac{N-n}{N-1}}$ se llama **factor de corrección** para población finita.

c) Tendrá una distribución normal siempre y cuando la población original de datos tenga una distribución norma, por lo contrario, como se menciona antes, tendrá una distribución aproximadamente normal cuando la población original de datos no tenga una distribución normal, pero las muestras seleccionadas sean de tamaño mayor o igual a 30

$$n \geq 30$$

NOTA

1. Cuando en los problemas no sea claro que tipo de formula se debe utilizar para determinar el valor de σ_x se puede utilizar el siguiente criterio.

   a) Si $\frac{n}{N} < 0.05$, entonces se utiliza $\sigma_x = \frac{\sigma}{\sqrt{n}}$

   b) En caso contrario, utilizar $\sigma_x = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$

2. Si no se hace específico que el muestreo es sin reposición, entonces para determinar el valor del error estándar. Utilizar la formula $\sigma_x = \frac{\sigma}{\sqrt{n}}$

2.2.1. APLICACIONES DEL TEOREMA DEL LÍMITE CENTRAL

Ejemplo 2.2.1

Supóngase que de 6 *estudiantes de toda una población, el primer estudiante tiene* 1*$, el segundo tiene* 2*$ y así sucesivamente hasta el sexto estudiantes que tiene* 6*$*

1. *Calcular la media y desviación estándar de los datos.*

$$x_1 = 1 \qquad x_2 = 2 \qquad x_3 = 3 \qquad x_4 = 4 \qquad x_5 = 5 \qquad x_6 = 6$$
$$N = 6$$

$$\mu = \mu_x = \frac{\sum_{i=1}^{N} x_i}{N} = 3.5$$
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} = 1.707825128$$

2. *Si se consideran todas las muestras posibles, con reposición de tamaño 2 extraídas a partir de la población dada. Calcular el valor esperado de* μ_x *y el error estándar.*

$$n = 2$$
$$\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{1.707825128}{\sqrt{2}} = 1.207614729$$

Ejemplo 2.2.2

La media aritmética de las estaturas de $42,000$ *estudiantes de secundaria de una ciudad es de* $1.58m$ *y la desviación estándar es de* $0.08m$*. Si se toman muestras de tamaño* 40 *alumnos, hallar la media esperada de la distribución de medias muestrales y su error estándar.*

$$N = 42,000 \qquad \mu = 1.58m \qquad \sigma = 0.08m$$

por el ***TLC***

$$\mu = \mu_x = 1.58m \qquad \frac{n}{N} = \frac{40}{42,000} = 0.00095238 < 0.05$$
$$\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{0.08}{\sqrt{40}} = 0.01264911064$$

Ejemplo 2.2.3

La distribución de una población de estudiantes que trabajan los fines de semana gana en promedio $1,000$ *al mes, con una desviación estándar de* 100*. Para una muestra de* 75 *estudiantes. ¿ Cuál es la probabilidad de que la media muestral*

1. *¿ Cuál es la probabilidad de que la media muestral esté entre* 1090 *y* 1120 *pesos?*
   Solución:
   La media y la desviación estándar de la población son $\mu = 1,100$ *y* $\sigma = 100$*, respectivamente. El tamaño de la muestra es* $n = 75$ *estudiantes, la cual es suficientemente grande, de modod que el teorema del límite central asegura que la distribución de* $\bar{x}$ *es aproximadamente normal con*

$$\mu_x = 1,100 \qquad \textit{error estándar}\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{75}} = 115$$

Para encontrar la probabilidad pedida, pasamos al proceso de estandarización

$$P(1,090 \leq \bar{x} \leq 1,120)$$
$$P\left(\frac{1,090-1,100}{11.55} \leq \frac{\bar{x}-\mu}{\sigma_x} \leq \frac{1,120-1,100}{11.55}\right)$$
$$= P(-0.866 \leq z \leq 1.732) = 0.958 - 0.1949 = 0.7631$$

2. *¿ Cuál es la probabilidad de que la media muestral sea mayor que* $1,120$ *pesos?*

$$P(\bar{x} \geq 1,120) = P\left(Z > \frac{1,120-1,100}{11.55}\right) = P(Z > 1.732)0.049$$

Ejemplo 2.2.4

Una máquina vendedora de refrescos está programada para que la cantidad de refrescos que se sirva sea una v.a con una media de $200mL$ *y una desviación estándar de* $15mL$*. ¿ Cuál es la probabilidad de que la cantidad de refrescos promedio, servida en una muestra aleatoria de tamaño* 36 *sea cuando menos* $204mL$*.*

Solución:

Por el T.L.C la distribución de $\bar{x}$ *tiene una media* $\sigma_x = \sigma = 200$ *y un error estándar* $\sigma_x = \frac{15}{\sqrt{36}} = 2.5$*. De acuerdo con el teorema de límite central está distribución es aproximadamente normal*

$$P(\bar{x} \leq 204) = P\left(Z \leq \frac{204-200}{2.5} = 1.6\right) = 1 - P(z < 1.6) = 1 - 0.945 = 0.0548$$

2.3. DISTRIBUCIÓN MUESTRAL DE PROPORCIONES

En los estudios estadístico no sólo se trabaja con medias aritméticas, también se puede estar interesado en **conocer la proporción** de algún acontecimiento en la población.

Ejemplo 2.3.1

La proporción de estudiantes de una universidad con ingresos superiores a $1,200$*; la proporción de votantes de un distrito electoral que favorecen a cierto partido político; la proporción de piezas defectuosas de un lote; la porporción de psicóticos en una zona durante una guerra, etc.*

Como **la mayor parte de las investigaciones se hacen a partir de análisis de una parte de la población (MUESTREO)**, necesitamos conocer las relaciones que guardan las proporciones poblacionales y la distribución muestral de proporciones.
Supongamos que hemos sacado una muestra de entidades de una población para averiguar el número de estas que poseen ciertas características de interés.
Supongamos también que se desea determinar por anticipado la probabilidad de que la muestra dé como resultado un número determinado de entidades que posean la característica que se esté estudiando.
Si se satisfacen dichas suposiciones, la **distribución binomial** resulta útil para calcular esas probabilidades.
Generalmente, en la práctica, el investigador hace selección de una muestra de entidades de alguna población finita de grandes muestras (aunque no es raro que la población de muestras sea infinita), con el propósito de hacer inferencia sobre la proporción de entidades de la población que poseen una característica determinada.

2.3.1. PROPORCIÓN MUESTRAL

En tales caso, el valor estadístico que interesa es la proporción muestral dada por:

$$\hat{p} = \frac{\textit{No. de entidades de la muestra que presenta la característica que interesa}}{\textit{No. total de entidades de la muestra}}$$

Ejemplo 2.3.2

Si en una muestra de 500 *electores,* 300 *prefieren al candidato A, la proporción de la muestra que prefieren al candidato A es:*

$$\hat{p} = \frac{300}{500} = 0.60$$

En general

$$\hat{p} = \frac{\textit{No. de entidades de la muestra que presenta la característica que interesa}}{\textit{Tamaño de la muestra}}$$

2.3.2. PROPORCIÓN POBLACIONAL

Similarmente, la proporción poblacional del número de unidades (M) de una característica específica de la población respecto del total de unidades (N), se expresa como:

$$\pi = \frac{M}{N} = \frac{M}{\textit{Tamaño de la población}}$$

En consecuencia, los procedimientos inferenciales dependerán de las distribución muestral de $\hat{p}$.

Definición 2.3.1 (DISTRIBUCIÓN MUESTRAL DE PROPORCIONES) *La definición muestral de proporciones es el* ***conjunto de proporciones de todas las muestrs posibles, del mismo tamaño, que se puede extraer de una determinada población.***

PARA EL CASO DE UNA DISTRIBUCIÓN MUESTRAL DE PROPORCIONES

1. La media de las proporciones de $\hat{p}$es $\mu_{\hat{p}} = E[\hat{p}] = \pi$
2. El error estándar de la distribución de proporciones se determina como:
   a) Si se tiene un población finita de muestras, **MUESTRO SIN REEMPLAZO** donde n = tamaño de la muestra, N = tamaño de la población.
   $$\sigma_{\hat{p}} = \sqrt{\frac{\pi(1-\pi)}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$
   b) Si se tiene una población infinita de muestras, **MUESTREO CON REEMPLAZO**
   $$\sigma_{\hat{p}} = \sqrt{\frac{\pi(1-\pi)}{n}}$$
3. La distribución de proporciones es **SIMÉTRICA**
4. Si $n\pi$ y $(n-\pi)$ son mayores que 5 o si $n \geq 25$ entonces, se afirma que la distribución muestral de $\hat{p}$ es **aproximadamente Normal**

NOTA:

El factor de correción de población finita se puede omitir cuando $\frac{n}{N} \leq 0.05$

Ejemplo 2.3.3

Supóngase que una urna hay 4 esferas blancas (Numeradas del 1 al 4) y seis negras (numeradas del 5 al 10). Si se selecciona una muestra de tamaño 5 y X es la variable aleatoria que representa al número de esferas blancas extraídas, determinar:

1. *La proporción poblacional*
   Solución : *En este caso, la proporción poblacional es:*

$$\pi = \frac{\textit{No. de esferas blancas}}{\textit{Tamaño de la población}} = \frac{4}{10} = 0.4$$

2. *El número de muestras posibles del experimento.*
   Solución:

$$\binom{10}{5} = \frac{10!}{5!(10-5)!} = 252$$

3. *El valor esperado de la proporción.*
   Solución:

Muestra x_i	$\hat{p} = \frac{x_i}{5}$	f_i	$f_r = f(\hat{p})$	$\hat{p}f(\hat{p})$	$(\hat{p}-\pi)^2 f_i$
0 (y 5 N)	0	6	$\frac{6}{252}$	0	0.96
1 (y 4 N)	0.2	60	$\frac{60}{252}$	$\frac{60}{1200}$	2.40
2 (y 3 N)	0.4	120	$\frac{120}{252}$	$\frac{240}{1260}$	0
3 (y 2 N)	0.6	60	$\frac{60}{252}$	$\frac{180}{1260}$	2.40
4 (y 1 N)	0.8	6	$\frac{6}{252}$	$\frac{24}{1260}$	0.96
		$\sum = 252$	$\sum = 1$	$\sum = 0.4$	$\sum$

$$E(\hat{p}) = \mu_{\hat{p}} = \sum \hat{p}f(\hat{p}) = \pi = 0.4$$

4. *La desviación estándar de la proporción*
   Solución:
   Se puede hacer de dos formas

   a*)*

$$\sigma_{\hat{p}} = \sqrt{\frac{\sum(\hat{p}-\pi)^2 f_i}{\textit{No. de muestras}}} = \sqrt{\frac{6.72}{252}} = 0.163299316$$

   b*)*

$$\sigma_{\hat{p}} = \sqrt{\frac{\pi(1-\pi)}{n}} \cdot \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{0.4(0.6)}{5}} \cdot \sqrt{\frac{10-5}{10-1}} = (0.2190910)(0.74535) = 0.163299316$$

Ejemplo 2.3.4

El registro académico de cierta universidad tiene establecido que 40 *% de los estudiantes llevan un CUM de 7 o más. Si se toma una muestra aleatoria simple de tamaño* 100 *del registro, cuál es la probabilidad de que, de esa muestra, la proporción de estudiantes que llevan un CUM de* 7 *o más. Sea.*

$$\pi = 0.4 \qquad n = 100 \textit{muestra grande} \longrightarrow \sigma_{\hat{p}} = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.4(0.6)}{100}}$$

1. 0.35 *o menos*

$$P(\hat{p} \leq 0.35) = P\left(Z = \frac{\hat{p} - \pi}{\sigma_{\hat{p}}} \leq \frac{0.35 - 0.4}{0.049}\right) = P(Z \leq -1.02) = 0.1539$$

2. *Mayor de* 0.32

$$\begin{aligned} P(\hat{p} > 0.32) &= P\left(Z = \frac{\hat{p} - \pi}{\sigma_{\hat{p}}} > \frac{0.32 - 0.4}{0.049}\right) = 1 - P(Z \leq -1.32653061) \\ &= 1 - 0.0516 = 0.9484 \end{aligned}$$

3. 0.47 *o mayor*

$$\begin{aligned} P(\hat{p} > 0.47) &= P\left(Z = \frac{\hat{p} - \pi}{\sigma_{\hat{p}}} \geq \frac{0.47 - 0.4}{0.049}\right) = P(Z \geq 1.428571429 \approx 1.429) \\ &= 1 - 0.9236 = 0.0764 \end{aligned}$$

4. *menor de* 0.52

$$\begin{aligned} P(\hat{p} \leq 0.52) &= P\left(Z = \frac{\hat{p} - \pi}{\sigma_{\hat{p}}} \leq \frac{0.52 - 0.4}{0.049}\right) = P(Z \leq 2.448979592 \approx 2.45) \\ &= 0.9927 \end{aligned}$$

5. *Entre* 0.38 *y* 0.42

$$\begin{aligned} P(0.38 \leq \hat{p} \leq 0.42) &= P\left(\frac{0.38 - 0.4}{0.049} \leq Z = \frac{\hat{p} - \pi}{\sigma_{\hat{p}}} \leq \frac{0.42 - 0.4}{0.049}\right) \\ &= P(-0.4281632653 \approx -0.43 \leq Z \leq 0.4081 \approx 0.41) \\ &= 0.6591 - 0.3336 = 0.3255 \end{aligned}$$

6. *Entre* 0.35 *y* 0.45

$$\begin{aligned} P(0.35 \leq \hat{p} \leq 0.45) &= P\left(\frac{0.35 - 0.4}{0.049} \leq Z = \frac{\hat{p} - \pi}{\sigma_{\hat{p}}} \leq \frac{0.45 - 0.4}{0.049}\right) \\ &= P(-1.0204 \approx -1.02 \leq Z \leq 1.0204 \approx 1.02) \\ &= 0.8461 - 0.1539 = 0.6922 \end{aligned}$$

Ejemplo 2.3.5 *Se sabe que un proceso productivo de cierto tipo de piezas electrónicas produce* 15 *% de piezas defectuosas. ¿ Cuál es la probabilidad de que, al tomar una muestra aleatoria de* 49 *piezas, está contenga* 6.6 *% o más de piezas defectuosas?*
Solución:

$$\pi = 0.15 \qquad n = 49 \textit{ muestra grande} \longrightarrow \sigma_{\hat{p}} = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.15(0.85)}{49}} = 0.051$$

$$\begin{aligned} P(\hat{p} \geq 0.066) &= P\left(Z = \frac{\hat{p} - \pi}{\sigma_{\hat{p}}} \geq \frac{0.066 - 0.15}{0.051}\right) = 1 - P(Z \leq -1.647 \approx -1.65) \\ &= 1 - 0.0495 = 0.9545 \end{aligned}$$

2.4. DISTRIBUCIÓN MUESTRAL DE S^2

La única distribución muestral que ha considerado hasta ahora es la media de la media muestral, pero en muchos problemas prácticos se necesita información acerca de la variabilidad de las medidas.
La variabilidad por lo general, **se mide mediante la varianza** y en algunos casos mediante la **desviación estándar**.
Por lo cual, necesitamos investigar la forma en que se distribuyen todos los valores posibles de S^2 en las muestras de tamaño n.
La distribución muestral de S^2 es particularmente importante en los problemas referidos a la variabilidad en una muestra aleatoria. Recordemos que cuando X esta distribuida Normalmente, entonces $Z = \frac{x-\mu}{\sigma}$ tiene una distribución normal estandarizada. Si tomamos ahora $Z^2 = \frac{(\bar{x}-\mu)^2}{\sigma^2}$, entonces decimos que Z^2 tiene una distribución χ^2 con un grado de libertad.
Por lo cual, si $X_1, X_2, X_3, \ldots, X_n$ es una muestra aleatoria, proveniente de una población de media μ y varianza σ^2, entonces la variable:

$$\begin{aligned}\chi^2_{(n)} &= \frac{(x_1-\mu)^2}{\sigma^2} + \frac{(x_2-\mu)^2}{\sigma^2} + \frac{(x_3-\mu)^2}{\sigma^2} + \cdots + \frac{(x_n-\mu)^2}{\sigma^2} \\ &= \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}\end{aligned}$$

tiene una **DISTRIBUCIÓN JI CUADRADA CON N GRADOS DE LIBERTAD** El resultado anterior, aunque importante, no es apropiado para hacer inferencias, puesto que en su expresión incluye a la media poblacional μ, la cual por lo general también es desconocida.
Puede demostrarse que si sustituimos $\bar{x}$ por μ, entonces se tiene la siguiente igualdad

$$\chi^2_{(n)} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma^2} + \frac{n(\bar{x}-\mu)}{\sigma^2}$$

en donde $\frac{n(\bar{x}-\mu)}{\sigma^2}$ es una variable aleatoria que tiene una distribución ji cuadrada con 1 grado de libertad.
Por lo tanto si despreciamos el último término de la igualdad anterior (por estar en función de μ) entonces se tiene que:

$$\chi^2_{(n-1)} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma^2}$$

pero como

$$S^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma^2}$$

entonces

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = (n-1)S^2$$

$$\therefore \chi^2_{n-1} = \frac{(n-1)S^2}{\sigma^2}$$

la cual tiene una **distribución Ji cuadrado con $n-1$ grados de libertad**
Lo anterior nos está indicando que al sustituir $\bar{x}$ por μ se pierde un grado de libertad.

También es de resaltar que la última expresión no involucra a la media y sólo incluye a la varianza (muestral y poblacional) como parámetro y así mediante el cálculo de S^2 y el conocimiento previo de σ^2 podemos hacer inferencias acerca de la población.

Ejemplo 2.4.1

Supóngase que se ha especificado que los diámetros en pulgadas de ciertas esferas de acero deben tener una varianza $\sigma^2 = 0.001in^2$. ¿ Cuál es la probabilidad de que en una muestra aleatoria de $n = 21$ esferas, su varianza sea mayor o igual que 0.0016?
Solución:

$$v = n - 1 = 21 - 1 = 20 \qquad \sigma^2 = 0.001 \qquad S^2 = 0.0016$$

$$P(S^2 \geq 0.0016) = P\left(\chi^2_{v=20} \geq \frac{20(0.0016)}{0.001}\right) = P(\chi^2_{v=20} \geq 32)$$

Por tabla de la distribución $\chi^2_{v=20}$ obtenemos que:

$$0.025 \leq P(\chi^2_{v=20} \geq 32) \leq 0.05$$

Con lo cual, podemos poner en duda lo que afirma la compañía

Ejemplo 2.4.2

Se sabe que la distribución de espesor de cierto material está Normalmente, distribuida con una desviación estándar de $0.01cm$. Una muestra aleatoria de 25 piezas de este material arroja como resultado una desviación estándar muestral de 0.008. Halle la probabilidad
Solución:

$$v = n - 1 = 24 \qquad S^2 = (0.008)^2 = 0.000064 \qquad \sigma^2 = (0.01)^2 = 0.0001$$

$$P(S \leq 0.008) = P(S^2 \leq 0.000064) = P(\chi^2_{v=24} \leq 15.36) = 1 - P(\chi^2_{v=24} \geq 15.36) = 1 - 0.9 = 0.1$$

2.5. ESTIMACIÓN PUNTUAL

Definición 2.5.1 (PARÁMETRO)

*Un **parámetro** es un valor que describe (parcial o totalmente) a una distribución de probabilidades de una propiedad poblacional de interés.*

Definición 2.5.2 (ESTIMADOR)

*Un **estimador** es una regla, a menudo expresada como una fórmula, que indica cómo calcular el valor de una estimación con base en las mediciones contenidas en una muestra.*

Ejemplo 2.5.1 (MEDIA MUESTRAL)

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

es un posible estimador puntual de la media poblacional μ. Claramente, la expresión para $\bar{X}$ es una regla y una fórmula. Nos indica que sumemos las observaciones muestrales y dividamos entre el tamaño muestral n

Definición 2.5.3 (ESTIMACIÓN PUNTUAL)

*Un **estimador puntual** (o simplemente estimador) es un estadístico muestral que se utiliza con el fin de inferir el valor de un parámetro poblacional desconocido.*

Cuando se realizan inferencias estadísticas, generalmente, se supone que un fenómeno determinado (o la población) **tiene características similares a algún modelo probabilístico**. Por lo tanto, la población tiene una distribución con una *función de densidad o probabilidad* determinada que depende de ciertos parámetros, cuyos valores numéricos necesitamos conocer para asignar probabilidades a los sucesos.

Ejemplo 2.5.2

Si se supone que la población es Normal, entonces, se deberán determinar los valores numéricos de los parámetros μ y σ

Si la población es Poisson, se deberá determinar el valor del parámetro λ

En el proceso de estimación, se realizan dos aproximaciones:

1. Se supone que la población pertenece a cierta familia de distribuciones (Normal, Poisson, etc.)

2. Se estiman los valores numéricos de los parámetros que determinarán el miembro de la familia que caracteriza a la población (Normal estándar ó Normal con $\mu = 5$ y $\sigma = 2$? ? , Poisson con $\lambda = 2$ ó con $\lambda = 5$, etc.)

Definición 2.5.4 (ESTIMACIÓN)

Una estimación es la realización de un estimador en base a datos provenientes de una muestra aleatoria. Es un valor específico observado de un estimador.

Así, con el proceso de estimación puntual se pretende hallar una estimación univaluada del parámetro poblacional de interés, ya que mediante una muestra se estima un único valor del mismo.

Ejemplo 2.5.3

*El estadístico $\bar{X}$ es un "estimador" de la media poblacional, y el valor específico $\bar{x}$ que tome cuando es extraída una muestra aleatoria (es decir, la realización de $\bar{X}$) es la **"estimación puntual"** .*

Ejemplo 2.5.4

*Supongamos que la variable "estatura de los alumnos de la facultad" se distribuye Normalmente. El parámetroμ es desconocido, y su valor podría inferirse a través de un **estimador** como la media muestral $\bar{x}$. Si se extrae una muestra aleatoria de* 30 *alumnos varones*

Varones				
1,68	1,72	1,70	1,71	1,57
1,82	1,86	1,87	1,77	1,65
1,70	1,79	1,72	1,98	1,77
1,93	1,69	1,73	1,80	1,72
1,66	1,73	1,59	1,94	1,74
1,79	1,73	1,65	1,79	1,76

entonces $\bar{x} = 1.75$ *es la* ***estimación puntual*** *de* μ

El proceso de estimación puntual asigna un único valor al parámetro poblacional desconocido. Por lo tanto, *la estimación puntual es correcta o equivocada* y, generalmente, ocurrirá este segundo caso, ya que es muy poco probable que la estimación obtenida a partir de una única muestra coincida con el parámetro poblacional. Por ello, resulta importante asignar alguna medida de riesgo a la estimación, o, mejor aún, indicar un intervalo en el cual puede estar contenido el valor del parámetro.

Suponga que deseamos especificar una estimación puntual para un **parámetro poblacional** al que llamaremos θ. El estimador de θ estará indicado por el símbolo $\hat{\theta}$ .El "sombrero" indica que estamos estimando el parámetro que está inmediatamente bajo él.

Definición 2.5.5 (INSESGADO)

Si $\hat{\theta}$ *es un estimador puntual de un parámetro* θ*, entonces* $\hat{\theta}$ *es un estimador insesgado si* $E(\hat{\theta}) = \theta$*.*

Definición 2.5.6 (SESGADO)

Si $E(\hat{\theta}) \neq \theta$*, se dice que* $\hat{\theta}$ *está sesgado.*

Definición 2.5.7 (SESGO)

El sesgo de un estimador puntual $\hat{\theta}$ *está dado por* $B(\hat{\theta}) = E(\hat{\theta}) - \theta$

Si un estimador es insesgado, entonces su sesgo es nulo. Además, en este caso, el ECM es igual a la varianza. Más que usar el sesgo y la varianza de un estimador puntual para caracterizar su bondad, podríamos emplear $E[(\hat{\theta} - \theta)^2]$, el promedio del cuadrado de la distancia entre el estimador y su parámetro objetivo.
De manera intuitiva, podemos pensar que para que no se produzcan sobre o subestimaciones, la distribución de muestreo del estadístico debería estar centrada en el valor del parámetro. Además, si se cumple lo anterior y el error estándar es pequeño, habrá una mayor probabilidad de que una realización del estadístico esté cerca del valor real del parámetro.

Definición 2.5.8 (ERROR CUADRÁTICO MEDIO)

El error cuadrático medio de un estimador puntual $\hat{\theta}$ es:

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

$$\begin{aligned} MSE(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] \\ &= E[\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2] \\ &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + E[\theta^2] \\ &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \\ &= E[\hat{\theta}^2] + E^2[\hat{\theta}] - E^2[\hat{\theta}] - 2\theta E[\hat{\theta}] + \theta^2 \\ &= E[\hat{\theta}^2] - E^2[\hat{\theta}] + E^2[\hat{\theta}] - 2\theta E[\hat{\theta}] + \theta^2 \\ &= Var(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2 \\ \therefore MSE(\hat{\theta}) &= Var(\hat{\theta}) + [B(\hat{\theta})]^2 \end{aligned}$$

Si un estimador es insesgado

$$\begin{aligned} E[\hat{\theta}] &= \theta \\ E[\hat{\theta}] - \theta &= 0 \end{aligned}$$

entonces su sesgo es nulo $B(\hat{\theta}) = E[\hat{\theta}] - \theta = 0$. Además, en este caso, el $MSE(\hat{\theta})$ es igual a la varianza.

2.5.1. MÉTODOS DE ESTIMACIÓN PUNTUAL

Cuando se extrae una única muestra, y con las observaciones obtenidas con la misma se calcula el valor de un estimador, estamos realizando una **estimación puntual**.
Existen métodos que suelen emplearse para obtener estimadores, es decir, para obtener la función que transforma los datos de una muestra en un número que corresponde a la estimación del parámetro de interés.

Ejemplo 2.5.5

La función para estimar la media poblacional μ es:

$$\frac{X_1 + \ldots + X_n}{n}$$

En lo apartados siguientes mencionaremos algunos métodos que permiten obtener esta fórmula, y, de modo más general, cómo obtener la función g que permite definir al estimador $\hat{\theta} = g(X_1, X_2, \ldots, X_n)$

2.5.2. MÉTODO :MÁXIMA VEROSIMILITUD

Cuando se realizan inferencias estadísticas, generalmente, se supone que un fenómeno determinado (o la población) **tiene características similares a algún modelo probabilístico**. Por lo tanto, la población tiene una distribución con una *función de densidad o probabilidad* determinada que depende de ciertos parámetros, cuyos valores numéricos necesitamos conocer para asignar probabilidades a los sucesos.

Ejemplo 2.5.6

Si se supone que la población es Normal, entonces, se deberán determinar los valores numéricos de los parámetros μ y σ

Si la población es Poisson, se deberá determinar el valor del parámetro λ

Cuando se extrae una muestra de una población que se supone tiene función de densidad $f(x;\theta)$, el interés recae en obtener un valor para el parámetro desconocido θ.
El método de **máxima verosimilitud** consiste en asignar al parámetro aquel valor que **maximiza la probabilidad** de que la muestra a extraer sea la muestra realmente observada.
De modo más general, para cualquier muestra de tamaño n , el método maximizará la probabilidad conjunta de que $X_1 = x_1, X_2 = x_2, \ldots,$ y $X_n = x_n$. Es decir, **maximiza la probabilidad** de que las variables aleatorias que componen la muestra aleatoria tomen los valores que se obtienen en una realización particular de la misma

Definición 2.5.9 (MÁXIMA VEROSIMILITUD)

*Sea $X_1, X_2, \ldots, X_n$ una muestra aletoria de una distribución con una función (densidad) de probabilidad $f(x;\theta)$, y sea la $L(x_1, x_2, \ldots, x_n;\theta)$ la **verosimilitud** de la muestra como función de θ. Si $t = u(x_1, x_2, \ldots, x_n)$ es el valor de θ para el valor de de la **función de verosimilitud es máxima,** entonces $T = u(_1, X_2, \ldots, X_n)$ es el **estimador de máxima verosimilitud de** θ, y t es el **estimador de máxima verosimilitud***

El procedimiento para obtener este tipo de estimadores es (relativamente) directo. Debido a la naturaleza de la función de verosimilitud se escoge, por lo común, maximizar el logaritmo natural de $L(\theta)$. Esto es, en muchas ocasiones es más fácil obtener el estimado MV maximizando $lnL(\theta)$ que $L(\theta)$

Ejemplo 2.5.7 (BINOMIAL)

En un experimento binomial se observan $X = x$ éxitos en n ensayos. Obtener el estimador de máxima verosimilitud del parámetro binomial p
En este caso la función de verosimilitud es idéntica a la probabilidad de que $X = x; deestaforma$

$$L(x;p) = \frac{n!}{(n-x)!x!}p^x(1-p)^{n-x} \qquad 0 \leq q \leq 1$$

Para obtener el valor del estimador de máxima verosimilitud de p, debemos obtener el valor del parámetro p que maximiza la probabilidad (binomial), esto es:

$$\max_p L(x;p) = \max_p \frac{n!}{(n-x)!x!}p^x(1-p)^{n-x}$$

Pero es más fácil obtener el estimado MV maximizando $ln(x;p)$ que $\max_p L(x;p)$ así que lo que nos interesa es:

$$lnL(x;p) = ln\left(\frac{n!}{(n-x)!x!}p^x(1-p)^{n-x}\right)$$

Recordando propiedades de los logaritmos

$$\begin{aligned} log(a \cdot b) &= log(a) + log(b) \\ log(a^b) &= b \cdot log(a) \\ log\left(\frac{a}{b}\right) &= log(a) - log(b) \end{aligned}$$

$$\begin{aligned} lnL(x;p) &= ln\left(\frac{n!}{(n-x)!x!}\right) + ln\left(p^x\right) + ln\left((1-p)^{n-x}\right) \\ &= ln(n!) - ln((n-x)!x!) + x \cdot ln(p) + (n-x) \cdot ln(1-p) \\ &= ln(n!) - ln((n-x)!) - ln(x!) + x \cdot ln(p) + (n-x) \cdot ln(1-p) \end{aligned}$$

Para encontrar el valor de p, para el cual $lnL(x;p)$ tiene un valor máximo, se toma la primera derivada con respecto a p y se iguala a cero:

$$\frac{\partial[lnL(x;p)]}{\partial p} = \frac{x}{p} - \frac{(n-x)}{(1-p)} = 0$$

Resolviendo para p tenemos que:

$$\begin{aligned} \frac{x}{p} - \frac{(n-x)}{(1-p)} &= 0 \\ \frac{x}{p} &= \frac{n-x}{1-p} \\ x - xp &= np - xp \\ \therefore p &= \frac{x}{n} \end{aligned}$$

Después de resolver para p, se obtiene el **estimador** MV de p el cual recibe el nombre de **proporción muestral** $\frac{X}{n}$, y el estimado MV es $\frac{x}{n}$. Para confirmar que este valor maximiza a $lnL(x;p)$, se toma la segunda derivada con respecto a p y se evalúa en $\frac{x}{n}$:

$$\frac{\partial^2[lnL(x;p)]}{\partial p^2} = -\frac{np(1-p) + (x-np)(1-2p)}{[p(1-p)]^2}$$

evaluando en $\frac{x}{n}$ tenemos que:

$$\frac{\partial^2[lnL(x;p)]}{\partial p^2} \Big|_{\frac{x}{n}} = -\frac{x}{\left(\frac{x}{n}\right)^2\left(1-\frac{x}{n}\right)}$$

lo que confirma el resultado, dado que la segunda derivada es negativa.

Ejemplo 2.5.8 (POISSON)

*Sean $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una **distribución de poisson** con una función de densidad de probabilidad*

$$f(x;\lambda) = \exp^{-\lambda}\frac{\lambda^x}{x!}$$

Determinar el estimador de λ

Solución:
En este caso la **función de verosimilitud** es idéntica a la probabilidad de que $X = x$; de la siguiente forma:

$$\begin{aligned} L(x_1, x_2, \ldots, x_n;\lambda) &= \exp^{-\lambda}\frac{\lambda^{x_1}}{x_1!} \cdot \exp^{-\lambda}\frac{\lambda^{x_2}}{x_2!} \cdots \exp^{-\lambda}\frac{\lambda^{x_n}}{x_n!} \\ &= \frac{1}{x_1! \cdot x_2! \cdots x_n!}\exp^{-\lambda n}\lambda^{\sum_{i=1}^{n} x_i} \end{aligned}$$

Para obtener el valor del estimador de máxima verosimilitud de λ, debemos obtener el valor del parámetro λ que maximiza la probabilidad (poisson), esto es:

$$\max_p L(x_1, x_2, \ldots, x_n;\lambda) = \max_p \frac{1}{x_1! \cdot x_2! \cdots x_n!}\exp^{-\lambda n}\lambda^{\sum_{i=1}^{n} x_i}$$

Pero es más fácil obtener el estimado MV maximizando $ln(x;\lambda)$ que $\max_p L(x;\lambda)$ así que lo que nos interesa es:

$$lnL(x_1, x_2, \ldots, x_n; \lambda) = ln\left(\frac{1}{x_1! \cdot x_2! \cdots x_n!} \exp^{-\lambda n} \lambda^{\sum_{i=1}^{n} x_i}\right) = *$$

Recordando propiedades de los logaritmos

$$\begin{aligned} ln(a \cdot b \cdot c) &= ln(a) + ln(b) + ln(c) \\ ln(e^x) &= x \\ ln(a^b) &= b \cdot ln(a) \end{aligned}$$

$$\begin{aligned} * &= ln\left(\frac{1}{x_1! \cdot x_2! \cdots x_n!}\right) + ln\left(\exp^{-\lambda n}\right) + ln\left(\lambda^{\sum_{i=1}^{n} x_i}\right) \\ &= ln\left(\frac{1}{x_1! \cdot x_2! \cdots x_n!}\right) - \lambda n + \sum_{i=1}^{n} x_i \, ln\,(\lambda) \end{aligned}$$

Para encontrar el valor de λ, para el cual $lnL(x;\lambda)$ tiene un valor máximo, se toma la primera derivada con respecto a λ y se iguala a 0:

$$\frac{\partial[lnL(x;\lambda)]}{\partial\lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0$$

Resolviendo para λ tenemos que:

$$\begin{aligned} -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i &= 0 \\ \frac{1}{\lambda}\sum_{i=1}^{n} x_i &= n \\ \therefore \lambda &= \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \end{aligned}$$

Después de resolver para λ, se obtiene que **el estimado MV es :** $\sum_{i=1}^{n} x_i$

Ejemplo 2.5.9 (NORMAL)

*Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una **distribución normal** con una función de densidad de probabilidad.*

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Determinar los estimadores de μ y σ^2

Solución:
Se procede de la misma forma que en el caso de un sólo parámetro. Dado que la función de verosimilitud depende tanto de μ como de σ^2, los estimadores **MV** de μ y σ^2 son los valores para los cuales la función de verosimilitud tiene un valor máximo, esto es:

$$\max_{\mu,\sigma^2} L(x_1, x_2, \ldots, x_n; \mu, \sigma^2) = \max_{\mu,\sigma^2} \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \cdots \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$$

Pero s más fácil obtener el estimador MV maximizando $lnL(x_{1,2}, \ldots, x_n; \mu, \sigma^2)$ que $\text{máx}_{\mu,\sigma^2} L(x_1, x_2, \ldots, x_n; \mu, \sigma^2)$ Con esto tenemos que:

$$\begin{aligned} L(x_1, x_2, \ldots, x_n; \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \cdots \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{(x_n-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \cdots \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{(x_n-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt[2]{(2\pi\sigma^2)^n}} \exp^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} = * \end{aligned}$$

Recordando algunas propiedades:

$$\begin{aligned} (a)^{\frac{1}{n}} \cdot (b)^{\frac{1}{n}} &= (a \cdot b)^{\frac{1}{n}} \\ \sqrt[c]{a^b} &= a^{\frac{b}{c}} \\ \frac{1}{a^b} &= a^{-b} \\ ln(a \cdot b \cdot c) &= ln(a) + ln(b) + ln(c) \\ ln(a^b) &= b \cdot ln(a) \end{aligned}$$

$$\begin{aligned} * &= \frac{1}{\sqrt[2]{(2\pi)^n}} \cdot \frac{1}{\sqrt[2]{(\sigma^2)^n}} \cdot \exp^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}} \cdot \frac{1}{(\sigma^2)^{\frac{n}{2}}} \cdot \exp^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \\ &= (2\pi)^{-\frac{n}{2}} \cdot (\sigma^2)^{-\frac{n}{2}} \cdot \exp^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}} \cdot \frac{1}{(\sigma^2)^{\frac{n}{2}}} \cdot \exp^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \\ ln(L(x_1, x_2, \ldots, x_n; \mu, \sigma^2)) &= ln\left((2\pi)^{-\frac{n}{2}} \cdot (\sigma^2)^{-\frac{n}{2}} \cdot \exp^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}\right) \\ &= ln\left((2\pi)^{-\frac{n}{2}}\right) + \ln\left((\sigma^2)^{-\frac{n}{2}}\right) + ln\left[\exp^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}\right] \\ &= -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \end{aligned}$$

Obteniendo las primeras derivadas parciales con respecto a μ y con respecto a σ^2, se obtiene lo siguiente:

$$\begin{aligned} \frac{\partial[lnL(x_1, x_2, \ldots, x_n; \mu, \sigma^2)]}{\partial\mu} &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0 \\ \frac{\partial[lnL(x_1, x_2, \ldots, x_n; \mu, \sigma^2)]}{\partial\sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \end{aligned}$$

Resolviendo para μ tenemos que:

$$\begin{aligned}
-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu) &= 0 \\
-\sum_{i=1}^{n}(x_i-\mu) &= 0 \\
-\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\mu &= 0 \\
n\mu &= \sum_{i=1}^{n}x_i \\
\mu &= \frac{1}{n}\sum_{i=1}^{n}x_i \\
\therefore \hat{\mu} &= \bar{x}
\end{aligned}$$

Resolviendo para σ^2 tenemos que:

$$\begin{aligned}
-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 &= 0 \\
-n\sigma^2+\sum_{i=1}^{n}(x_i-\mu)^2 &= 0 \\
n\sigma^2 &= \sum_{i=1}^{n}(x_i-\mu)^2 \\
\sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 \\
\therefore \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2
\end{aligned}$$

Ejemplo 2.5.10

Suponga que en una población se desea estudiar cuántos apoyan a cierto candidato político. Para ello, se extrae una muestra de 20 *habitantes, de los cuales* 8 *están a favor del candidato. ¿ Cuál es el valor del estimador de máxima verosimilitud de p?*

Para calcularlo, debemos obtener el valor del parámetro p que maximiza la probabilidad (binomial) de que en una muestra de tamaño 20, se obtengan 8 éxitos. Es decir:

$$\max_{p} \; l(x;p) = \max_{p}\binom{20}{8}p^8(1-p)^{20-8}$$

Para facilitar el procedimiento, se toman logaritmos

$$\begin{aligned}
L(x;p) &= ln[l(x;p)] \\
&= ln\left[\binom{20}{8}p^8(1-p)^{12}\right] = *
\end{aligned}$$

Recordando propiedades de los logaritmos

$$\begin{aligned}
log(a\cdot b) &= log(a)+log(b) \\
log(a^b) &= b\cdot log(a)
\end{aligned}$$

$$
\begin{aligned}
* &= ln\binom{20}{8} + ln(p^8) + ln[(1-p)^{12}] \\
&= ln\binom{20}{8} + 8 \cdot ln(p) + 12 \cdot ln(1-p)
\end{aligned}
$$

Derivando e igualando a cero, tenemos que:

$$\frac{\partial L(x;p)}{\partial p} = \frac{8}{p} - \frac{12}{1-p} = 0$$

Despejando p tenemos que:

$$
\begin{aligned}
\frac{8}{p} &= \frac{12}{1-p} \\
8 - 8p &= 12p \\
p &= \frac{8}{20}
\end{aligned}
$$

El ejemplo anterior puede generalizarse fácilmente: si tomamos una muestra de tamaño n y obtenemos x éxitos, el estimador de máxima verosimilitud de p es $\frac{x}{n}$. Es decir, la proporción muestral.

En la siguiente tabla se ilustran las principales poblaciones utilizadas en la práctica y los estimadores que se obtienen con el método de Máxima Verosimilitud:

Población	Parámetro	Estimador MV
Normal	μ	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
	σ^2	$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$
Binomial		$\bar{p} = \frac{X}{n}$
Poisson	λ	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

2.5.3. MÉTODO: MOMENTOS

El método **iguala algunos momentos muestrales, con los momentos teóricos expresados en términos de los parámetros de una distribución determinada.** Recordando que los parámetros son,en general, funciones de los momentos teóricos.

Definición 2.5.10 (MOMENTOS)

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución con función (densidad) de probabilidad $f(x;\theta)$. El r-ésimo momento alreddor del cero se define como:

$$M_r{}' = \frac{1}{n}\sum_{i=1}^{n} X_i^r$$

El método de los momentos proporciona una alternativa razonable cuando no se pueden determinar los estimadores de máxima verosimilitud.

Ejemplo 2.5.11 (GAMMA)

Si la variable aleatoria X tiene una distribución gamma, entonces:

$$\mu = \alpha\theta \longrightarrow \alpha = \frac{\mu}{\theta}$$
$$\mu_2^{'} = \alpha(\alpha+1)\theta^2$$

De lo anterior tenemos que:

$$\begin{aligned}\mu_2^{'} &= \alpha(\alpha\theta^2+\theta^2)\\ &= \alpha^2\theta^2+\alpha\theta^2\\ &= \alpha^2\theta^2+\alpha\theta\theta\\ \mu_2^{'} &= \mu^2+\mu\theta\\ \longrightarrow \theta &= \frac{\mu_2^{'}-\mu^2}{\mu}\end{aligned}$$

y

$$\longrightarrow \alpha = \frac{\mu^2}{\mu_2^{'}-\mu^2}$$

De esta forma, los dos parámetros de la distribución gamma son funciones de los dos primeros momentos alrededor del cero.
En esencia, el método se implementa igualando tantos momentos muestrales con los correspondientes momentos teóricos tantas veces como sea necesario para determinar un estimador de momentos para un parámetro desconocido. En el ejemplo anterior ***los estimadores de momentos de los parámetros gamma*** *α y θ son:*

$$\hat{\alpha} = \frac{\bar{X}}{M_2^{'}-\bar{X}^2}$$
$$\hat{\theta} = \frac{M_2^{'}-\bar{X}^2}{\bar{X}}$$

2.6. ESTIMACIÓN POR INTERVALO

2.6.1. INTERVALOS DE CONFIANZA PARA LA MEDIA EN MUESTRAS GRANDES

CONCEPTOS PRELIMINARES

Supongamos que deseamos conocer cuánto gana en el mercado laboral un profesionista recién egresado.
Para esto se toma una muestra de tamaño $n = 30$ y a cada uno de ellos se les pregunta su salario.
Bajo nuestras suposiciones, consideremos que con los datos obtenidos se obtiene una media $\bar{x} = 11,200$ de salario. La media muestral $\bar{x}$ es una **estimación puntual** confiable de μ, pero probablemente no corresponde exactamete con la media μ.
En lugar de considerar esta idea de estimación puntual, se puede especificar con una alta probabilidad, digamos de 0.90 o 0.95 que cierto rango cubre o contiene a la verdadera media μ

Ejemplo 2.6.1 *A partir de los datos de la muestra, se puede decir que el intervalo de* $11,100$ *a* $11,300$ *cubre a la media μ con una probabilidad de* 0.95
El intervalo $[11,100, 11,300]$ *es un ejemplo de* ***intervalo de confianza***

INTERVALO DE CONFIANZA

Un intervalo de confianza es un rango (intervalo) de valores que probablemente contiene al valor poblacional que se esta estimando. Los componente de un intervalo de confianza son:

1. Dos límites conocidos como **límites de confiabilidad**. Un límite inferior (LI) y un límite superior (LS). En el ejemplo anterior, $LI = 11,100$ y $LS = 11,300$
2. Un valor de probabilidad, el cual se conoce como **nivel de confianza** y se denota por $1-\alpha$. Para nuestro ejemplo, $1-\alpha = 0.95$ y $\alpha = 0.05$. Expresado en términos de porcentaje se dice que hay un intervalo de 95 % de confianza

LÍMITE DE CONFIABILIDAD

Los límites de confiabilidad son los valores que establecen las fronteras o extremos del intervalo de confianza.
en general, un **intervalo de confianza** para la media poblacional μ presenta la forma:

$$P(LI \leq \mu \leq LS) = 1 - \alpha$$

Intervalos de confianza para la media μ cuando la población es normalmente distribuida y la desviación estándar σ es conocida.
Nuevamente, para nuestro ejemplo, por tablas de la distribución normal estandarizada (Z), para un valor central de área igual a 0.95, le corresponden valores de:

$$LI = -1.96 \qquad\qquad LS = 1.96$$

Además como

$$1 - \alpha = 0.95$$

entonces

$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

donde Z es el valor estandarizado de $\bar{x}$.
Ahora para hallar el intervalo de confianza para μ, necesitamos determinar los valores de los límites inferior LI y superior LS. Como sabemos, la $\bar{x}$ calculada de la muestra no será exactamente igual a la media poblacional μ. Por lo tanto, lo primero que debemos hacer es establecer el tamaño de un cierto imagen de error (E), conocido como error muestral, así:

$$\mu = \bar{x} \pm E$$

De esta manera, los límites inferior y superior del intervalo de confianza son :

$$LI = \bar{x} - E \qquad\qquad LS = \bar{x} + E$$

Como $P(-1.96 \leq Z \leq 1.96) = 0.95$ y $\bar{x}$ tiene una distribución normal, cuyo valor estandarizado Z es:

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

$$P(-1.96 \leq \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96) = 0.95$$

Despejando $\bar{x} - \mu$ de esta fómula, obtenemos

$$P\left(-1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{x} - \mu \leq 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$P\left(\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Con esto, concluimos qu el intervalo de confianza para μ con un nivel de confianza del 95 % es:

$$E = 1.96 \cdot \frac{\sigma}{\sqrt[2]{n}}$$

En genral, para construir un **estimador** de intervalo de un parámetro desconocido θ debemos encontrar dos estadísticos L y U tales que:

$$P(L \leq \theta \leq U) = 1 - \alpha$$

El intervalo resultante

$$L \leq \theta \leq U$$

Se llama intervalo de confianza del $(1 - \alpha)$ por ciento para el parámetro desconocido θ.
La interpretación del intervalo de confianza es que sin se seleccionan muchas muestras aleatorias y se calcula un intervalo de confianza del $(1 - \alpha)$ por ciento para θ en cada muestra, entoncees $(1 - \alpha)$ por ciento de estos intervalos contendrán el valor verdadero de θ

Ejemplo 2.6.2

Si el intervalo fuese del 95 *% de confianza, a la larga sólo el* 5 *% de los intervalos no contendrían el parámetro estimado* θ.
Para determinar los límites de un intervalo de confianza para estimar la media poblacional μ, *se debe considerar lo siguiente:*

1. **SI SE CONOCE EL VALOR DE σ**

   Si $\bar{x}$ es el valor de la media de una muestra aleatoria de tamaño n tomada de una población Normal, el intervalo de confianza $(1 - \alpha)$ para estimar la media poblacional μ cuando se conoce la desviación estándar de la población σ es:

   $$\left[\bar{x} - Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

2. **SI SE CONOCE EL VALOR DE σ y $n \geq 30$,EN SU LUGAR SE PUEDE UTILIZAR EL VALOR DE S. QUEDANDO EN ESTE CASO EL INTERVALO DE CONFIANZA**

   $$\left[\bar{x} - Z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}, \bar{x} + Z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}\right]$$

   Ejemplo 2.6.3

   Consideramos el problema correspondiente a la distancia recorrida por los estudiantes (solo ida, desde su casa a la universidad), en donde su media es desconocida y $\sigma = 6$. *Supóngase que se selecciona una muestra aleatoria de* 100 *valores de distancias registradas y que* $\bar{x} = 10.22$ *millas. Determinar el intervalo de confianza para la estimación de la distancia media que recorren los estudiantes de toda la universidad, para un nivel de significancia del:*

a) 95 %

$$n = 100 muestra\ grande \qquad \sigma = 6 \qquad \bar{x} = 10.22$$

Encontraremos primero el valor crítico $Z_{\frac{\alpha}{2}}$. *Como* $1 - \alpha = 0.95 \longrightarrow \alpha = 0.05 \longrightarrow \frac{\alpha}{2} = 0.025$
Por lo tanto el valor crítico es $Z_{\frac{\alpha}{2}=0.025} = 1.96$ *y el intervalo de confianza es:*

$$\left[\bar{x} - Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

$$\left[10.22 - 1.96\left(\frac{6}{\sqrt{100}}\right), 10.22 + 1.96\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

$$[9.044, 11.396]$$

$$o$$

$$9.044 \leq \mu \leq 11.396$$

b) 90 %
Encontraremos primero el valor crítico $Z_{\frac{\alpha}{2}}$. *Como* $1 - \alpha = 0.90 \longrightarrow \alpha = 0.10 \longrightarrow \frac{\alpha}{2} = 0.05$
Por lo tanto el valor crítico es $Z_{\frac{\alpha}{2}=0.05} = 1.65$ *y el intervalo de confianza es:*

$$\left[\bar{x} - Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

$$\left[10.22 - 1.65\left(\frac{6}{\sqrt{100}}\right), 10.22 + 1.65\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

$$[9.23, 11.21]$$

$$o$$

$$9.23 \leq \mu \leq 11.21$$

c) 88 %
Encontraremos primero el valor crítico $Z_{\frac{\alpha}{2}}$. *Como* $1 - \alpha = 0.88 \longrightarrow \alpha = 0.12 \longrightarrow \frac{\alpha}{2} = 0.06$
Por lo tanto el valor crítico es $Z_{\frac{\alpha}{2}=0.06} = 1.56$ *y el intervalo de confianza es:*

$$\left[\bar{x} - Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

$$\left[10.22 - 1.56\left(\frac{6}{\sqrt{100}}\right), 10.22 + 1.56\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

$$[9.284, 11.156]$$

$$o$$

$$9.284 \leq \mu \leq 11.156$$

Ejemplo 2.6.4

Determine el intervalo de confianza para la media poblacional μ de los salarios de la policía, en donde:

$$1-\alpha = 92\,\% \qquad n = 64 \qquad \bar{x} = 23,228 \qquad S = 8,779$$

Solución:
Encontraremos primero el valor crítico $Z_{\frac{\alpha}{2}}$. Como $1-\alpha = 0.92 \longrightarrow \alpha = 0.08$
$\longrightarrow \frac{\alpha}{2} = 0.04$
Por lo tanto el valor crítico es $Z_{\frac{\alpha}{2}=0.04} = 1.75$ *y el intervalo de confianza es:*

$$\left[\bar{x} - Z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}, \bar{x} + Z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}\right]$$
$$\left[23,228 - 1.75\left(\frac{8,779}{\sqrt{64}}\right), 23,228 + 1.75\left(\frac{8,779}{\sqrt{64}}\right)\right]$$
$$[21,307.59375, 25,148.40625]$$

o

$$21,307.59375 \leq \mu \leq 25,148.40625$$

2.6.2. EL TAMAÑO DE LA MUESTRA NECESARIO PARA ESTIMAR LA MEDIA POBLACIONAL μ

Como sabemos el error máximo de estimación $E = Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$. Si de esta formula despejamos el valor de n, obtenemos el tamaño de muestra mínimo para estimar el valor verdadero de μ. Si $\bar{x}$ se utiliza para estimar el valor de μ, entonces puede tenerse una confianza de $1-\alpha$, en forma particular, de que el error $|\bar{x} - \mu|$ no será mayor que una cantidad específica E cuando el **tamaño de la muestra** sea:

$$n = \left[\frac{Z_{\frac{\alpha}{2}} \cdot \sigma}{E}\right]^2$$

Ejemplo 2.6.5

Determinar el tamaño de la muestra que será necesario para estimar el peso medio de las cajas de un cargamento de importación. Si se requiere que la estimación sea precisa dentro de ***una libra*** *(alrededor de la media muestral), con una confianza de* 95 *%. Supongase que* $\sigma = 3$
Solución:
Como el intervalo de confianza es $1-\alpha = 0.95 \longrightarrow Z_{\frac{\alpha}{2}} = 1.96$. *El error máximo es* $E = 1$

$$n = \left[\frac{Z_{\frac{\alpha}{2}} \cdot \sigma}{E}\right]^2$$
$$n = \left[\frac{1.96(3)}{1}\right]^2 = 34.5744 \approx 35$$

En este caso dejaríamos $n = 35$*, ya que es mejor que se pase el tamaño de la muestra a que nos falte.*

Ejemplo 2.6.6 *Suponga que un gerente de mercadotecnia desea estimar la media poblacional del uso anual de consumibles de las computadoras que utilizan en la empresa, dentro de* $\pm 95,000$ *del valor verdadero. Con base en un estudio realizado el año anterior, se piensa que la desviación estándar puede estimarse como* $150,000$*. ¿ Qué tamaño de muestra será requerido para lograr*

esto, si se desea una confianza del 96 % *?*
Solución

$$1-\alpha = 0.96 \qquad \alpha = 0.08 \qquad \frac{\alpha}{2} = 0.04$$
$$Z_{\frac{\alpha}{2}=0.04} = 2.05$$

$$n = \left[\frac{Z_{\frac{\alpha}{2}} \cdot \sigma}{E}\right]^2$$
$$n = \left[\frac{2.05(150,000)}{95,000}\right]^2 = 10.47 \approx 11$$

2.6.3. DISTRIBUCIÓN MUESTRAL DE $\bar{x}$ CUANDO LA POBLACIÓN ES NORMAL. EL TAMAÑO DE LA MUESTRA ES MENOR QUE 30 ($n \leq 30$) Y σ ES DESCONOCIDA

DISTRIBUCIÓN t STUDENT

Anteriormente se ha visto que tiene en la resolución de problemas el uso de la estandarización siguiente:

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

Generalmente, el objetivo que se persigue el recurrir a este tipo de estandarización es determinar la probabilidad de algún valor específico de $\bar{x}$ suponiendo que la media poblacional es μ, para luego utilizar esta probabilidad en la toma de decisiones.
Cuandp se conoce el valor de μ, el tamaño de la muestra es **menor que 30** $n \leq 30$ y **se desconoce el valor de** σ, es evidente que no se puede utilizar la estandarización Z al depender esta de σ. de aquí la importancia de las distribuciones que no dependen de σ como es el caso de la **t de student**. En el curso de probabilidades se demuestra que la variable

$$\frac{\bar{x}-\mu}{\frac{S}{\sqrt{n-1}}}$$

en estas condiciones, es una **t-student** con $n-1$ grados de libertad, en donde S es la desviación estándar muestral.
Esta propiedad nos permite, fijado el nivel de confianza $1-\alpha$, obtener el valor $t_{\frac{\alpha}{2}}$ tal que:

$$P\left(-t_{\frac{\alpha}{2}} \leq \frac{\bar{x}-\mu}{\frac{S}{\sqrt{n-1}}} \leq t_{\frac{\alpha}{2}}\right) = 1-\alpha$$
$$P\left(\bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n-1}} \leq \mu \leq \bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n-1}}\right) = 1-\alpha$$

$\therefore$ el intervalo pedido es:

$$\left[\bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n-1}}, \bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n-1}}\right]$$

Ejemplo 2.6.7

El número de reclamos recibidos por una empresa que opera un conjunto de salas de cine se registra cada semana. Los datos que se recolectaron durante 10 *semanas son:*

28	35	32	19	48	29	30	43	36	21

Obtener un intervalo de confianza del 95 % *de confianza para la media poblacional del número de reclamos*

Solución: *De los datos tenemos que*

$$S = 8.9498693334 \approx 8.95 \qquad t_{\frac{\alpha}{2},v=n-1} = t_{\frac{0.05}{2},v=9} = 2.26 \qquad n = 10 \qquad \bar{x} = 32.1$$

$\therefore$ *el intervalo pedido es:*

$$\left[\bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n-1}}, \bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n-1}}\right]$$
$$\left[32.1 - 2.26\left(\frac{8.95}{\sqrt{10-1}}\right), 32.1 + 2.26\left(\frac{8.95}{\sqrt{10-1}}\right)\right]$$
$$[25.35766667, 38.84233333]$$

Ejemplo 2.6.8

Se toma una muestra aleatoria de tamaño 20 *con los pesos de bebés nacidos en un hospital durante* 1982*, encontrándose una media de* 6.87*ib, y una desviación estándar de* 1.76*ib. Estimar con una confianza de* 95 % *el peso medio de todos los bebés nacidos en ese hospital de* 1982

Solución:

$$n = 20 \text{ muestra pequeña} \qquad v = n - 1 = 19 \qquad \bar{x} = 6.87 \qquad t_{\frac{\alpha}{2}=0.025,v=19} = 2.09$$
$$S = 1.76$$

$\therefore$ *El intervalo de confianza de* 95 % *es:*

$$\left[\bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n-1}}, \bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n-1}}\right]$$
$$\left[6.87 - 2.09\left(\frac{1.76}{\sqrt{20-1}}\right), 6.87 + 2.09\left(\frac{1.76}{\sqrt{20-1}}\right)\right]$$
$$[6.026117165, 7.713882835]$$

Ejemplo 2.6.9

Un fabricante de pinturas desea determinar el tiempo de secado en promedio de una nueva pintura para interiores. Si en 12 áreas de prueba de igual tamaño el obtuvo un tiempo de secado medio de 63.3 *minutos y una desviación estándar de* 8.4 *minutos, construya un intervalo de confianza del* 95 % *para la media verdadera* μ

Solución:

$$n = 12 \qquad v = n - 1 = 11 \qquad t_{v,\frac{\alpha}{2}} = t_{v=11,\frac{0.05}{2}=0025}$$
$$S = 8.4 \qquad \bar{x} = 63.3$$

$\therefore$ *El intervalo de confianza de* 95 % *es:*

$$\left[\bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n-1}}, \bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n-1}}\right]$$
$$\left[63.3 - 2.20\left(\frac{8.4}{\sqrt{12-1}}\right), 63.3 + 2.20\left(\frac{8.4}{\sqrt{12-1}}\right)\right]$$
$$[57.72807035, 68.87192965]$$

2.6.4. ESTIMAR EL VALOR DE LA VARIANZA POBLACIONAL σ^2(MUESTRA PEQUEÑA)

Dada una muestra aleatoria de tamaño n tomada de una población Normal, se extrae una muestra de tamaño n sobe la que se calcula la **varianza muestral** S^2. El **estadístico de prueba**

$$\frac{(n-1)S^2}{\sigma^2}$$

Sigue una distribución χ^2 con $n-1$ grados de libertad expresándose de la siguiente manera:

$$P\left[\chi^2_{\left(v=n-1,1-\frac{\alpha}{2}\right)} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\left(v=n-1,\frac{\alpha}{2}\right)}\right] = 1-\alpha$$

$$P\left[\frac{(n-1)S^2}{\chi^2_{\left(\frac{\alpha}{2},v=n-1\right)}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{\left(1-\frac{\alpha}{2},v=n-1\right)}}\right] = 1-\alpha$$

$\therefore$ El intervalo de confianza de nivel $(1-\alpha)$ para la varianza de una **distribución normal** con varianza muestral S^2 es:

$$\left[\frac{(n-1)S^2}{\chi^2_{\left(\frac{\alpha}{2},v=n-1\right)}}, \frac{(n-1)S^2}{\chi^2_{\left(1-\frac{\alpha}{2},v=n-1\right)}}\right]$$

Ejemplo 2.6.10

Un fabricante de baterías para automóviles asegura que las baterías que produce duran en promedio 2 *años. con una desviación estándar de* 0.5 *años. Si* 5 *de estas baterías tiene duración* $1.5, 2.5, 2.9, 3.2, 4.0$ *años. Determine un intervalo de confianza del* 95 *% para* σ^2 *indique si es válida la afirmación del fabricante.*

Solución:

$$n = 5 \text{ muestra pequeña} \qquad S^2 = 0.847 \qquad v = n-1 = 4$$

$$1-\alpha = 0.95 \qquad \frac{\alpha}{2} = 0.025$$

$$\chi^2_{\left(\frac{\alpha}{2}=0.025,v=5-1=4\right)} = 11.1 \qquad \chi^2_{\left(1-\frac{\alpha}{2}=0.95,v=5-1=4\right)} = 0.711$$

$$\left[\frac{(n-1)S^2}{\chi^2_{\left(\frac{\alpha}{2},v=n-1\right)}}, \frac{(n-1)S^2}{\chi^2_{\left(1-\frac{\alpha}{2},v=n-1\right)}}\right]$$

$$\left[\frac{(5-1)(0.847)}{11.1}, \frac{(5-1)(0.847)}{0.711}\right]$$

Como el valor de $\sigma^2 \notin [0.305225225, 4.76511955]$*, entonces lo afirmado por el fabricante no está garantizado por los datos muestrales.*

Ejemplo 2.6.11

En 16 *recorridos de prueba, el consumo de gasolina da un motor experimental se distribuye de forma normal con una desviación estándar de* 2.2 *galones. Construye un intervalo de confianza*

del 99 % para σ^2
Solución:

$$n = 16 muestra\ pequeña \qquad S = 2.2 \qquad 1-\alpha = 0.99$$
$$\alpha = 0.1 \qquad \frac{\alpha}{2} = 0.005 \qquad v = n-1 = 16-1 = 15$$

$$\chi^2_{\left(\frac{\alpha}{2}=0.005, v=15\right)} = 32.8$$
$$\chi^2_{\left(1-\frac{\alpha}{2}=0.995, v=15\right)} = 4.6$$

*$\therefore$ El intervalo de confianza del 99 % para la varianza de una **distribución normal** con varianza muestral S^2 es:*

$$\left[\frac{(n-1)S^2}{\chi^2_{\left(\frac{\alpha}{2}, v=n-1\right)}}, \frac{(n-1)S^2}{\chi^2_{\left(1-\frac{\alpha}{2}, v=n-1\right)}}\right]$$
$$\left[\frac{(16-1)(2.2)^2}{32.8}, \frac{(16-1)(2.2)^2}{4.6}\right]$$
$$[2.213414634, 15.7826087]$$
$$\therefore P(2.213414634 \leq \sigma^2 \leq 15.7826087) = 0.99$$

2.6.5. INTERVALO DE CONFIANZA PARA UNA PROPORCIÓN π

Anteriormente hemos inducido la variable aleatoria $\hat{p} = \frac{k}{n}$ como un estimador de π, el parámetro poblacional en una distribución binomial.
El **Estadístico** $\hat{p}$ puede usarse para determinar intervalos de confianza para las proporciones poblacionales en las aplicaciones donde se usa la distribución Binomial, tales como la proporción de personas que fuman cigarros en una población dada, la de electores que están a favor de cierto candidato, o la de elementos defectuosos en determinado proceso de producción.
Como $E[\hat{p}] = \mu_{\hat{p}} = \pi$ y $\sigma_{\hat{p}} = \sqrt{V(\hat{p})} = \sqrt{\frac{\pi(1-\pi)}{n}}$ vemos que ambos están en función de π. Si deseamos estimar el parámetro π, entonces debemos utilizar el siguiente **estadístico.**

$$S_p = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Puesto que la distribución muestral de proporciones tiene un comportamiento Normal para el caso en que $n\pi$ y $n(1-\pi)$ **sean mayores que** 5, o que $n\pi(1-\pi) > 3$, o que $n \geq 25$, entonces para estos casos podemos considerar el valor estandarizado de $\hat{p}$ de la siguiente manera:

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

con lo cual

$$P\left[-Z_{\frac{\alpha}{2}} \leq Z \leq Z_{\frac{\alpha}{2}}\right] = 1-\alpha$$
$$P\left[-Z_{\frac{\alpha}{2}} \leq \frac{\hat{p}-\pi}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq Z_{\frac{\alpha}{2}}\right] = 1-\alpha$$
$$P\left[\hat{p} - Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq \pi \leq \hat{p} + Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] = 1-\alpha$$

$\therefore$ el intervalo para la estimación de π con una confiabilidad de $1-\alpha$ es:

$$I.C = \left[\hat{p} - Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Ejemplo 2.6.12

Hay interés en saber la cantidad de estudiantes de la universidad que trabajan. Se toma una muestra de 900 *estudiantes. A cada uno se le pregunta si trabaja. Los resultados fueron los siguientes:*

Si trabajan: 35 % *No trabajan:* 65 %

Encontrar un intervalo de confianza de 95 % *para la proporción* π *de estudiantes que si trabajan*

Solución:

$$\hat{p} = \frac{35}{100} = 0.35 \qquad n = 900 \qquad Z_{\frac{\alpha}{2}} = 1.96$$

$$S_p = \sqrt{\frac{(0.35)(0.65)}{900}} = 0.01589898669$$

$$\begin{aligned} I.C &= \left[\hat{p} - Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \\ &= [0.35 - (1.96) \cdot (0.01589898669), 0.35 + (1.96) \cdot (0.01589898669)] \\ &[0.3188379861, 0.3811620139] \end{aligned}$$

DETERMINACIÓN DEL TAMAÑO DE LA MUESTRA (n) PARA LA ESTIMACIÓN DE π

El tamaño muestral necesario para la obtención de un intervalo de confianza adecuado para π puede determinarse si se específica de antemano el **error máximo (E)** admisible entre $\hat{p}$ y π, o sea $E = [\hat{p} - \pi]$. En este caso dicho tamaño muestral se obtiene de la siguiente manera:

$$n = 0.25 \cdot \left[\frac{Z_{\frac{\alpha}{2}}}{E}\right]^2$$

Ejemplo 2.6.13

Se hace un estudio para determinar la proporción de votamos de una comunidad cuantificable que favorecen la construcción de una planta generadora de energía nuclear.

1. *Si se tiene que sólo* 140 *de* 400 *votantes seleccionados al azar favorecen el proyecto, obtengan el intervalo de confianza del* 95 % *de la proporción de todos los votantes de esta comunidad que se expresan a favor del proyecto.*

$$n = 400 \qquad \hat{p} = \frac{140}{400} = 0.35 \quad Z_{\frac{\alpha}{2}} = 1.96$$

$$E = Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96 \cdot \sqrt{\frac{(0.35)(0.65)}{400}} = 0.04674302087$$

$\therefore$ *El intervalo de confianza de la proporción poblacional* π *es:*

$$\hat{p} \pm E = 0.35 \pm 0.04674302087 = [0.3032569791, 0.3967430209]$$

2. *Si se desea tomar una muestra de los* 400 *votantes de la población para estimar al parámetro* π *con un error de* ± 10 *por cierto y con un nivel de confianza del* 95 % *. ¿ Qué tamaño de muestra se debe tomar?*
Solución:

$$n = 0.25\left[\frac{Z_{\frac{\alpha}{2}}}{E}\right]^2 = 0.25 \cdot \left[\frac{1.96}{0.1}\right]^2 = 96.04 \approx 97$$

Se aproxima al siguiente entero, ya que mientras más grande el tamaño de la muestra es mejor.

2.6.6. INTERVALO DE CONFIANZA PARA ESTIMAR UNA DIFERENCIA DE MEDIAS $\mu_1 - \mu_2$ CON VARIANZAS σ_1^2 y σ_2^2 CONOCIDAS

Si $\bar{x_1}$ y $\bar{x_2}$ son las medias de muestras aleatorias independientes de tamaño $n_!$ y n_2 tomadas de **poblaciones Normles** que tiene medias μ_1 y μ_2 y varianzas σ_1^2 y σ_2^2, entonces $x_1 \bar{-} x_2$ es una variable aleatoria que tiene una distribución Normal con media y varianza:

$$\mu_{\bar{x_1}-\bar{x_2}} = \mu_1 - \mu_2 \qquad \sigma^2_{\bar{x_1}-\bar{x_2}} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$Z = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

tiene una distribución Normal estandarizada. Sustituyendo esta expresión de Z en:

$$P\left[-Z_{\frac{\alpha}{2}} \leq Z \leq Z_{\frac{\alpha}{2}}\right] = 1 - \alpha$$

y despejando $\mu_1 - \mu_2$, obtenemos:

$$P\left[(\bar{x_1} - \bar{x_2}) - Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x_1} - \bar{x_2}) + Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right] = 1 - \alpha$$

Con lo cual, el intervalo de confianza de $\mu_1 - \mu_2$ es:

$$\left[(\bar{x_1} - \bar{x_2}) - Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{x_1} - \bar{x_2}) + Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

Ejemplo 2.6.14

Construya un intervalo de confianza del 94 % *de la diferencia real entre las duraciones en promedio de dos tipos de focos eléctricos, dado que una muestra tomada al azar de* 40 *focos de un tipo duró en promedio* 418 *horas de uso continuo y* 50 *focos de otras clases duraron en promedio* 402 *horas. Las desviaciones estándar de las poblaciones, según se sabe son* $\sigma_1 = 26$ *y* $\sigma_2 = 22$

$$1 - \alpha = 0.94 \longrightarrow \alpha = 0.06 \longrightarrow \frac{\alpha}{2} = 0.03$$

$$Z_{\frac{\alpha}{2}=0.03} = 1.88$$

$$(\bar{x_1} - \bar{x_2}) \pm Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$(418 - 402) \pm 1.88 \left[\sqrt{\frac{(26)^2}{40} + \frac{(22)^2}{50}}\right]$$

$\therefore$ *el intervalo de confianza del* 94 % *para la diferencia de medias es:*

$$[6.3, 25.7]$$

El hecho de que ambos limites de confianza sean positivops sugiere que el primer tipo de foco es superior al segundo tipo

$$6.3 \leq \mu_1 - \mu_2 \leq 25.7$$

$\mu_1 \geq \mu_2 + 6.3$ *La duración de vida del foco 1 es mayor que la duración de vida del foco 2*

Ejemplo 2.6.15

Supóngase que se tiene interés en comprar el éxito académico de alumnos universitarios que pertenecen a una asociación estudiantil con el de estudiantes que no pertenecen a este tipo de organización . Claramente estas dos poblaciones estpan separas, y se debe tomar de ellas muestras independientes De cada población se toman muestras de tamaño 40. *Las medias obtenidas de dichas muestras son* 2.03 *para los alumnos que pertenecen a la asociación y* 2.21 *para los que no pertenecen a la asociación. Si la desviación estándar de ambas poblaciones es* $\sigma = 0.6$. *Determine la estimación con un intervalo de confianza del* 95 % *para la diferencia entre las dos medias independientes.*

Solución :

$$\frac{\alpha}{2} = 0.025$$

$$\begin{aligned}(\bar{x_1} - \bar{x_2}) \pm Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} &= (2.03 - 2.21) \pm (1.96) \cdot \sqrt{\frac{(0.6)^2}{40} + \frac{(0.6)^2}{40}} \\ &= -0.18 \pm 0.2629 \\ &= [-0.4429, 0.0829]\end{aligned}$$

Es decir, con un 95 % *de confianza se estima que la diferencia entre las medias poblacionales está entre* -0.4429 *y* 0.0829

Observación:

Como los extremos del intervalo son de signo contrario, entonces cabe la posibilidad de que $\mu_1 = \mu_2$. *Además se puede establecer las opciones de que* $\mu_1 > \mu_2$ *o* $\mu_1 < \mu_2$

2.6.7. INTERVALO DE CONFIANZA PARA $(\mu_1 - \mu_2)$, SUPONIENDO QUE $\sigma_1 = \sigma_2$ ES DESCONOCIDA.

Si $\bar{x_1}$ y $\bar{x_2}$ son las medias de muestras aleatorias independientes de tamaño n_1 y n_2, tomadas de poblaciones Normales con varianzas desconocidas pero iguales, $\sigma_1^2 = \sigma_2^2$, el intervalo de confianza del $(1 - \alpha)100\,\%$ para $\mu_1 - \mu_2$ está dado por:

$$(\bar{x_1} - \bar{x_2}) - t_{n_1+n_2-2,\frac{\alpha}{2}} \cdot S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, (\bar{x_1} - \bar{x_2}) + t_{n_1+n_2-2,\frac{\alpha}{2}} \cdot S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

donde

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Ejemplo 2.6.16

Se ha realizado un estudio para comparar el contenido de nicotina de dos marcas de cigarrillos. Diez cigarrillos de la marca A tuvieron un contenido de nicotina en promedio de 3.1 *miligramos con una desviación estándar de* 0.5 *miligramos, mientras que* 8 *cigarrillos de marca B tuvieron un contenido de nicotina en promedio de* 2.7 *miligramos con una desviación estándar de* 0.7 *miligramos. Suponiendo que los dos conjuntos de datos son muestras tomadas al azar de poblaciones con distribución Normal y con varianzas iguales, construya un intervalo de confianza del* 95 *% de la diferencia real en el contenido promedio de nicotina de las do marcas de cigarrillos*

MARCA A	**MARCA B**
$n_1 = 10$	$n_2 = 8$
$\bar{x_1} = 3.1$	$\bar{x_2} = 2.7$
$S_1 = 0.5$	$S_2 = 0.7$

$$v = n_1 + n_2 - 2 = 16$$
$$t_{v=16,\frac{\alpha}{2}=0.025} = 2.12$$
$$S_p = \sqrt{\frac{9(0.25) + 7(0.49)}{16}} = 0.596$$

Con base en esto:

$$(3.1 - 2.7) \pm 2.12(0.596)\sqrt{\frac{1}{10} + \frac{1}{8}}$$
$$I.C = [0.8286, 3.1716]$$

De lo cual se puede concluir que:
En realidad, los hombres tardan más que las mujeres en realizar esta actividad.

2.6.8. INTERVALO DE CONFIANZA PARA DIFERENCIAS ENTRE MEDIAS $\mu_1 - \mu_2$ MUESTRAS DEPENDIENTES (DATOS PAREADOS)

Vamos a suponer agora que se tienen dos poblaciones normales $N(\mu_1, \sigma_1^2)$ y $N(\mu_2, \sigma_2^2)$ y de las que se extraen dos muestras que no son independientes.
Nos vamos a restringir al caso en el cual los tamaños n de ambas muestras son iguales entre sí .
Típicamente consideraremos la situación en la cual las muestras no se extraen de forma independiente de cada población, sino que cada muestra consiste en la medida de una característica en los mismos elementos de una población.
Por ejemplo, supongamos que sobre los elementos de una muestra se mide cierta variable, después e aplica un determinado tratamiento a la muestra y, sobre los mismos elementos, se vuelve a medir la misma variable. (ej. temperatura antes y después de aplicar un tratamiento)
Supongamos que tenemos una muestra aleatoria de n pares de observaciones enlazadas procedentes de distribuciones normales de medias μ_x y μ_y
Es decir, sena $x_1, x_2, \ldots, x_n$ los valores de las observaciones de la población que tiene media μ_x e $y_1, y_2, \ldots, y_n$ los valores de las observaciones de la población que tiene media μ_y.
El objetivo en este caso es calcular un intervalo de confianza para la diferencia de medias $\mu_1 - \mu_2$, en dichas muestras.
Para ello se consideran las diferencia $d_i = x_i - y_i$, entre los valores de las variables en cada uno de los elementos de la muestra. Sean $\bar{d}$ y S_D la media y la desviación estándar muéstrales

observadas de las diferencias $d_i = x_i - y_i$.
Para plantear el problema se asume que estas diferencias son los valores de una nueva variable aleatoria D.
Si la muestra es suficientemente grande (en la práctica $n \geq 30$) puede considerarse que dicha variable se distribuye normalmente con media $\mu_D = \mu_1 - \mu_2$ y varianza σ_D^2.
Las estimaciones puntuales de estos parámetros serán respectivamente $\bar{D}$ y S_D^2, que tomarán, para una muestra en particular, los valores concretos:

$$\bar{d} = \frac{\sum d_i}{n} = \frac{\sum (x_i - y_i)}{n} \qquad\qquad S_D^2 = \frac{\sum (d_i - \bar{d})^2}{n-1}$$

El problema se reduce entonces a calcular un intervalo de confianza para la media μ_D de una distribución normal.

MUESTRA GRANDE $n \geq 30$

Por analogía con el intervalo de confianza para una media y aproximando la varianza σ_D^2 por S_D^2 por se la muestra grande, puede escribirse entonces:

$$P\left(\bar{d} - Z_{\frac{\alpha}{2}} \cdot \frac{S_D}{\sqrt{n}} \leq \mu_1 - \mu_2 \leq \bar{d} + Z_{\frac{\alpha}{2}} \cdot \frac{S_D}{\sqrt{n}}\right) = 1 - \alpha$$

donde se ha igualado μ_D a $\mu_1 - \mu_2$. Por lo tanto, el intervalo de confianza de nivel $(1-\alpha)$ para la diferencia de media de observaciones pareadas con $n \geq 30$ puede expresarse como:

$$I.C = \left[\bar{d} \pm Z_{\frac{\alpha}{2}} \cdot \left(\frac{S_D}{\sqrt{n}}\right)\right]$$

MUESTRA PEQUEÑA $n < 30$

En el caso de que la muestra fuera pequeña (n< 30), habría que substituir la distribución normal por una distribución t, siendo el intervalo de confianza

$$I.C = \left[\bar{d} \pm t_{\frac{\alpha}{2}, v=n-1} \cdot \left(\frac{S_D}{\sqrt{n}}\right)\right]$$

Ejemplo 2.6.17

Se aplica un proceso para aumentar el rendimiento en un 10 *fábricas muy diferentes (no dejar tomarse el bocadillo a media mañana). Los rendimientos (en ciertas unidades, como toneladas/día) antes y después son:*

ANTES	13	22	4	10	63	18	34	6	19	43	X_1
Después	15	22	2	15	65	17	30	12	20	42	X_2

Determinar el intervalo de confianza del 95 *% para el aumento del rendimiento.*
Solución:
Definamos las diferencias como $D_i = X_{despues} - X_{antes}$. *Con esto, obtenemos:*

$$D_i = 2, 0, -2, 5, 2, -1, -4, 6, 1, -1$$

$$\hat{d} = \frac{\sum D_i}{n} = \frac{8}{10} = 0.8$$

$$S_D = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}} = 3.08$$

Como tenemos muestra pequeña, $n = 10$, usamos

$$t_{v=9,\frac{\alpha}{2}=0.025} = 2.262$$

El intervalo pedido es:

$$\begin{aligned} I.C &= \left[\bar{d} \pm t_{v,\frac{\alpha}{2}} \cdot \frac{S_D}{\sqrt{n}}\right] = \left[0.8 \pm 2.262\left(\frac{3.08}{\sqrt{10}}\right)\right] \\ &= [0.8 \pm 2.2] = [-1.4, 3] \end{aligned}$$

Ejemplo 2.6.18

Se realiza un estudio médico para comparar las diferencias de eficacia de dos medicamentos para reducir el nivel de colesterol. El grupo de investigación utiliza un enfoque de datos pareados para controlar la variación de la reducción que podría deberse a factores distintos del medicamento. Los miembros de cada par tienen las mismas características de edad, pesos, estilo de vida y otros factores pertinentes. Se administra el medicamento X a una persona seleccionada aleatoriamente en cada par y el medicamento Y a la otra persona par. Tras un determinado período de tiempo, se mide de nuevo el nivel de colesterol de cada persona. Supongamos que se selecciona de las grandes poblaciones de participantes una muestra aleatoria de ocho pares de pacientes que tienen problemas conocidos de colesterol.
La tabla siguiente muestra el número de puntos en que se ha reducido el nivel de colesterol de cada persona, así como las diferencias $d_i = x_i - y_i$, correspondientes a cada par. Estime con un nivel de confianza del 99 % la diferencia media de eficacia entre los dos medicamentos, X e Y, para reducir el colesterol.

Par	***Medicamento X***	***Medicamento Y***	***Diferencia*** $d_i = x_i - y_i$
1	29	26	3
2	32	27	5
3	31	28	3
4	32	27	5
5	32	30	2
6	29	26	3
7	31	33	-2
8	30	36	-6

$\bar{d} = 1.625$ $S_D = 3.777$ $t_{v=7,\frac{\alpha}{2}=0.005} = 3.499$

$\therefore$ *El intervalo de confianza al 99 % tenemos que:*

$$\begin{aligned} 1.625 - 3.499\left(\frac{3.777}{\sqrt{8}}\right) \leq \mu_x - \mu_y \leq 1.625 + 3.499\left(\frac{3.777}{\sqrt{8}}\right) \\ -3.0474 \leq \mu_x - \mu_y \leq 6.2974 \end{aligned}$$

Como el intervalo de confianza contiene el valor de cero, podemos concluir que el medicamento X es igual de eficaz que el medicamento Y.
Así también, $\mu_x - \mu_y$ podría ser positivo, lo que sugeriría que el medicamento X es más eficaz que el medicamento Y.
Finalmente, la diferencia podría ser negativa, lo que sugeriría que el medicamento Y es más eficaz que el medicamento X.

2.6.9. INTERVALO DE CONFIANZA PARA UNA DIFERENCIA DE PROPORCIONES $\pi_1 - \pi_2$ (MUESTRA GRANDE)

Frecuentemente surgen problemas donde es deseable calcular la diferencia entre los parámetros Binomiales π_1 y π_2 sobre la base de muestras aleatorias independientes tomadas de dos poblaciones Binomiales.
Si $\hat{p_1} = \frac{x_1}{n_1}$ y $\hat{p_2} = \frac{x_2}{n_2}$ donde x_1, x_2 y n_1, n_2 son respectivamente el número de éxitos y el tamaño muestral de las proporciones muestrales correspondientes a dos poblaciones Binomiales, la distribución muestral de $\hat{p_1} - \hat{p_2}$ puede utilizarse para estimar a la diferencia de parámetros $\pi_1 - \pi_2$.
Para valores grandes de n_1 y n_2, las distribuciones de x_1 y x_2 pueden aproximarse nuevamente distribuciones Normales, por lo tanto, se deduce que $\hat{p_1} - \hat{p_2}$ tiene aproximadamente una distribución Normal con media y varianza:

$$\mu_{\hat{p_1}-\hat{p_2}} = \pi_1 - \pi_2 \qquad S^2_{\hat{p_1}-\hat{p_2}} = \frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}$$

Además, el valor estandarizado de $\hat{p_1} - \hat{p_2}$ es:

$$Z = \frac{(\hat{p_1} - \hat{p_2}) - (\pi_1 - \pi_2)}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}}$$

Al sustituir este valor de Z en

$$P\left[-Z_{\frac{\alpha}{2}} \leq Z \leq Z_{\frac{\alpha}{2}}\right] = 1 - \alpha$$

se obtiene el intervalo requerido:

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

Ejemplo 2.6.19

Si se tiene que 132 *de* 200 *votantes del distrito A favorecen a un candidato dado para la elección del senado y* 90 *de* 150 *votantes del distrito B se expresan a favor de este mismo candidato, obtenga un intervalo de confianza* 99 *% para* $\pi_1 - \pi_2$*, la diferencia entre las proporciones reales de votantes de los dos distritos favorables al candidato.*
Solución:

$$\hat{p_1} = \frac{132}{200} = 0.66 \qquad \hat{p_2} = \frac{90}{150} = 0.6$$
$$1 - \alpha = 0.99 \qquad \frac{\alpha}{2} = 0.005$$
$$Z_{\frac{\alpha}{2}=0.005} = 2.575 \qquad \hat{p_1} - \hat{p_2} = 0.06$$
$$n_1 = 200 \qquad n_2 = 150$$

$$\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}} = \sqrt{\frac{0.66(0.34)}{200} + \frac{0.6(0.4)}{150}} = 0.05217$$
$$0.06 \pm 2.575(0.05217)$$
$$I.C = [-0.07434, 0.19434]$$

2.6.10. INTERVALO DE CONFIANZA PARA COCIENTE DE VARIANZAS

DISTRIBUCIÓN F DE FISHER

La distribución de la variable aleatoria F es asimétrica sesgada hacia la derecha, y se define como

$$F = \frac{\frac{W}{v_1}}{\frac{Y}{v_2}} = \frac{W \cdot v_2}{Y \cdot v_1}$$

Donde W e Y, son variables aleatorias con distribución χ^2 con grados de libertad correspondiente v_1 & v_2, W e Y son v.a independientes.
Al comparar dos poblaciones es natural que se contrasten sus varianzas o desviaciones estándar. Supóngase que se tienen dos poblaciones normales con varianzas σ_1^2 y σ_2^2, respectivamente. Cuando se seleccionan dos muestras independientes de tamaños n_1 y n_2 de las poblaciones normales 1 y 2, respectivamente, con varianzas muéstrales S_1^2 y S_2^2, entonces el cociente.

$$F = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \frac{S_1^2 \cdot \sigma_2^2}{S_2 \cdot \sigma_1^2}$$

tiene una distribución F con $v_1 = n_1 - 1$ grados de libertad en el numerador y $v_2 = n_2 - 1$ grados de libertad en el denominador. Esto es debido a que $\frac{v_1 S_1^2}{\sigma_1^2}$ está distribuida como χ^2_{v1} y $\frac{v_2 S_2^2}{\sigma_2^2}$ como χ^2_{v2} La forma simbólica para representar a la distribución F es $F_{(v_1, v_2, \alpha)}$

PROPIEDADES DE LA DISTRIBUCIÓN F

1. F toma puros valores no negativos.

2. F es asimétrica, sesgada hacia la derecha.

3. $P\left[F > f(v_1, v_2, \alpha)\right] = \alpha$

4. $f(v_1, v_2, \alpha) = \frac{1}{f(v_2, v_1, 1-\alpha)}$

Ejemplo 2.6.20

Determinar el valor del punto crítico $F(5, 10, 0.05)$
Solución :
Por tablas de F, $F(5, 10, 0.05) = 3.33$.
Es decir, $P(F > 3.33)0.05$

Ejemplo 2.6.21

Determinar el valor del punto crítico $F(5, 10, 0.95)$
Solución:
Observamos que en la tabla de F *no hay valor para* $\alpha = 0.95$
Aplicando la propiedad 4

$$F(5, 10, 0.95) = \frac{1}{f(10, 5, 0.05)} = \frac{1}{4.74} = 0.21097$$

INTERVALO DE CONFIANZA PARA COCIENTE DE VARIANZAS

Si S_1^2 y S_2^2 son las varianzas de muestras aleatorias independientes de tamaño n_1y n_2 tomadas de poblaciones normales, entonces:

$$F = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

Es una v.a que tiene distribución F con $v_1 = n_1$ y $v_2 = n_2 - 1$ grados de libertad.
Al sustituir esta expresión de F en

$$P[F_{v_1,v_2,1-\frac{\alpha}{2}} \leq F \leq F_{v_1,v_2,\frac{\alpha}{2}}] = 1-\alpha$$
$$P\left[F_{v_1,v_2,1-\frac{\alpha}{2}} \leq \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} \leq F_{v_1,v_2,\frac{\alpha}{2}}\right] = 1-\alpha$$
$$P\left[\frac{S_1^2}{S_2^2}\cdot\frac{1}{F_{v_1,v_2,\frac{\alpha}{2}}} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_1^2}{S_2^2}\cdot F_{v_2,v_1,\frac{\alpha}{2}}\right] = 1-\alpha$$

Es decir, si S_1^2 y S_2^2 son los valores de las varianzas de muestras aleatorias independientes de tamaño n_1 y n_2 tomadas de poblaciones normales, un intervalo de confianza del $(1-\alpha)100\,\%$ para $\frac{\sigma_1^2}{\sigma_2^2}$ es:

$$\left[\frac{S_1^2}{S_2^2}\cdot\frac{1}{F_{v_1,v_2,\frac{\alpha}{2}}}, \frac{S_1^2}{S_2^2}\cdot F_{v_2,v_1,\frac{\alpha}{2}}\right]$$

Ejemplo 2.6.22 *Se ha realizado un estudio para comparar el contenido de nicotina de dos marcas de cigarros.* 10 *cigarros de la marca A, tuvieron una desviación estándar de* $0.5mg$*, mientras que* 8 *cigarrillos de la marca B, tuvieron una desviación estándar de* $0.7mg$*. Si los dos conjuntos de datos se tomaron con distribuciones normales, construir un intervalo de confianza del* 98 *% para* $\frac{\sigma_1^2}{\sigma_2^2}$*. ¿ Se puede concluir que* $\sigma_1^2 \neq \sigma_2^2$*?*

Solución:

MARCA A	***MARCA B***
$n_1 = 10$	$n_2 = 8$
$S_1 = 0.5$	$S_2 = 0.7$
$v_1 = 9$	$v_2 = 7$

$$1-\alpha = 0.98$$
$$\frac{\alpha}{2} = 0.01$$
$$F_{\left(v_1=9,v_2=7,\frac{\alpha}{2}=0.01\right)} = 6.72$$
$$F_{\left(v_2=7,v_1=9,\frac{\alpha}{2}=0.01\right)} = 5.61$$

$$\left[\frac{S_1^2}{S_2^2}\cdot\frac{1}{F_{v_1,v_2,\frac{\alpha}{2}}}, \frac{S_1^2}{S_2^2}\cdot F_{v_2,v_1,\frac{\alpha}{2}}\right]$$
$$\left[\frac{(0.5)^2}{(0.7)^2}\cdot\left(\frac{1}{6.72}\right), \frac{(0.5)^2}{(0.7)^2}\cdot(5.61)\right]$$
$$= [0.07592, 2.8622]$$

Con esto podemos concluir que puede ser que $\sigma_1^2 = \sigma_2^2$ *ya que en nuestro intervalo de confianza pasa por el 1*

Capítulo 3

PRUEBA DE HIPÓTESIS

3.1. PRUEBA DE HIPÓTESIS

En las secciones anteriores describimos los procedimientos para hacer estimaciones puntuales y estimaciones por intervalos de los parámetros poblacionales.
Vimos que una de las ventajas de la estimación por intervalo es que puede expresarse la incertidumbre de la estimación respecto al verdadero valor del parámetro poblacional.
Otra ventaja es que los intervalos de confianza sirven para verificar la validez de cualquier suposición hecha a cerca del valor de un parámetro poblacional.
Tal valor supuesto viene a ser lo que se llama **hipótesis estadística.**

La determinación de la validez de una suposición de esta naturaleza se llama **prueba de hipótesis.**
El propósito principal de la prueba de hipótesis es hacer posible una elección adecuada entre dos hipótesis que se refieren al valor de un parámetro poblacional , es decir,**debemos decidir si un parámetro es o no igual a un valor prescrito.**
El propósito de las pruebas de hipótesis incluye muchos de los cálculos o procedimientos vistos anteriormente:

- Cálculo de estadísticos muestrales.
- Variables aleatorias.
- Distribuciones de probabilidad e inferencia estadística

Ejemplo 3.1.1

Supóngase que una compañía produce una cera líquida de limpieza para cocinas que vende en latas que tienen la leyenda " peso neto $300\ gr$ *". Se sabe por larga experiencia que la variabilidad del proceso es estable, de tal forma que* $\sigma = 5gr$.
Una máquina llena de latas y la compañía hace todo esfuero posible por controlar el peso medio neto en el estándar de $300gr$.
Sin embargo, ocurren pequeños errores en los cálculos de la máquina y las partes se desgastan, con lo cual el contenido medio varía algunas veces más y algunas veces menos.
Desviaciones mínimas (del orden de 1 gr o menos) del estándar no son de consecuencia, pero las desviaciones cada vez mayores en uno u otro sentido son de mayor interés.
Al intentar hacer que el contenido medio de las latas sea $300gr.$, *la compañía ha ideado el siguiente procedimiento de inspección.*
Cada hora, durante corridas de producción, la compañía toma una muestra al azar de 25 latas del lote de producción de 1 hora, calcula el peso promedio de la muestra "x" y decide en base a este valor si el proceso está o no "bajo control" el contenido medio es $300gr.$, *como se supone*

que debe ser) o bien "fuera de control" (el contenido medio no es de $300gr.$ *).*
Para tomar esta decisión, la compañía debe aplicar algún criterio inequívoco, o regla que seguir. Según esto, la compañía ha específicado el siguiente criterio:

- *Considérese que el proceso está fuera de control si* $\bar{x} < 297gr.$ *o* $\bar{x} > 303gr.$
- *Considérese que está bajo control si* $297 < \bar{x} < 303$

En el lenguaje de la estadística, la compañía desea demostrar, la relación con cada lote examinado, la hipótesis de que el peso neto (promedio) de la cera es $300gr.$ contra la alternativa de que el contenido medio no es de $300gr.$ y realizará esta prueba sobre la base del criterio siguiente:
Desprecie la " hipótesis (y acepte la alternativa) si $\bar{x} < 297gr.$ o bien $\bar{x} > 303gr$;, en caso contrario, acepte la hipótesis.
Podemos llamar a la hipótesis de que el proceso está en control hipótesis **H** y escribirla como

$$H : \mu = 300gr$$

Y la alternativa de que el proceso está fuera de control como Alternativa **A** y escribirla como

$$H_A : \mu \neq 300gr$$

Es evidente que la hipótesis es verdadera o falsa y siempre que se pruebe el criterio nos llevará a su aceptación o a su rechazo. Desafortunadamente, pese a ello en alguna ocasión cualquiera la compañía puede errar en una u otra de las dos posibles formas(pero no en ambas): primero, la compañía puede decidir que el proceso está fuera de control cuando, de hecho, está bajo control.
Esto sucedería si el peso medio de la lata) es en realidad $300gr.$ pero la media de la muestra es menor que $297gr.(\bar{x} < 297gr)$ o mayor que $303gr;(\bar{x} > 303gr)$; la consecuencia de este error es que en un proceso que opera en el nivel deseado es interrumpido mientras un ingeniero busca un problema inexistente; segundo, la compañía puede decidir que el proceso está bajo control, cuando, de hecho, está fuera de control.
Esto sucedería si el pedo medio de la lata (del lote) no es de 300 gr (es, por ejemplo, sólo 290gr) pero la media de la muestra $\bar{x}$ está entre 297 y 303gr; la consecuencia de este error es que a un proceso que no este operando en el nivel deseado se le permitirá proseguir su operación.
La situación que se describe en el ejemplo es común en la verificación de una hipótesis estadística y se puede resumir en la tabla siguiente:

	Aceptar II	**Rechazar II**
H es verdadera	Decisión correcta	Error de tipo I
H es falsa	Error de tipo II	Decisión correcta

Si la hipótesis es verdadera, la decisión de aceptarla es correcta; por el contrario, si la hipótesis es falsa, la decisión de rechazarla es la correcta.
Por otra parte, si la hipótesis es verdadera y se rechaza, se ha cometido un error; el error que se comete cuando se rechaza una hipótesis verdadera se denomina **error de tipo 1** y la probabilidad de cometerlo se designa por medio de la letra griega

$$\alpha$$

A la inversa, si la hipótesis es falsa y se acepta, también se ha cometido un error; el error que se comete cuando se acepta una hipótesis falsa recibe el nombre de **error de tipo II** y la probabilidad de cometerlo se designa por medio de la letra griega

$$\beta$$

Ejercicio 3.1.1

Con base al ejemplo del contenido de cera, determinar:

1. *El valor de la probabilidad α (error tipo 1)*

2. *El "valor de la probabilidad " β (error tipo II), suponiendo que el contenido medio real es de sólo $296gr$. (con lo cual la hipótesis es falsa y el proceso está fuera de control)*

3. *La probabilidad de aceptar la hipótesis de que $\mu = 300gr$ cuando en realidad μ vale $294, 295, 296, \cdots, 305, 306$*

Solución: *En este caso*

$$\mu = 300 \qquad \sigma = 5gr \qquad n = 25$$

a) α =P(de rechazar la hipótesis si ésta es verdadera)=P($\bar{x} < 297$ o $\bar{x} > 303gr$)
Como la distribución muestral es de tipo Normal. entonces:

$$P(\bar{x} < 297) = P\left(z < \frac{297 - 300}{\frac{5}{\sqrt{25}}}\right) = 0.00134999$$

$$P(\bar{x} > 303) = P\left(z > \frac{303 - 300}{\frac{5}{\sqrt{25}}}\right) = P(z > 3) = 0.0013599$$

$\therefore$*la probabilidad de cometer un error de tipo 1 es:*

$$\alpha = 0.0013499 + 0.0013499 = 0.0026998$$

Es decir, la proabilidad de rechazar equivocadamente la hipótesis de que el proceso está bajo control, y que la compañía busque una falla que no existe, es de 0.0026998.

b)

β = *P(de aceptar la hipótesis falsa, de que el proceso está bajo control)* = $P(297 \leq \bar{x} \leq 303)$
cuando $\mu = 296$

$$= P(297 \leq \bar{x} \leq 303) = \left(\frac{297 - 296}{\frac{5}{\sqrt{25}}} \leq z \leq \frac{303 - 296}{\frac{5}{\sqrt{25}}}\right) = P(1 \leq z \leq 7) = F(7) - F(1) \approx 1 - 0.8413447$$

$=\therefore$ *la probabilidad de cometer un error de tipo II, es:*

$$\beta = 0.1586553$$

c) En el inciso (b), al obtener la probabilidad de aceptar una hipótesis falsa, supusimos que la media del proceso había cambiado de 300 a 296 gr.
Sin embargo, en este problema existe una cantidad infinita de otras alternativas (valores posibles diferente de 300 gr. del peso medio real) y para cada una existe una probabilidad β de que la compañía acepte la hipótesis cuando ésta sea falsa.
A continuación, se da una tabla de valores para algunos caso de las alternativas mencionadas, en donde la segunda columna muestra las probabilidades de aceptar la hipótesis de que $\mu = 300gr$ cuando μ vale en realidad $296, 295, \cdots, 305, 306$:

Valor de μ	β	Probabilidad de aceptar H, cuando μ = 300
294	0.0013	0.0013
295	0.0228	0.0228
296	0.1587	0.1587
297	0.5000	0.5000
298	0.8413	0.8413
299	0.9772	0.9772
300	-----	-----
301	0.9772	0.9772
302	0.8413	0.8413
303	0.5000	0.5000
304	0.1587	0.1587
305	0.0228	0.0228
306	0.0013	0.0013

Si la Hipótesis se formula de tal manera que esta pueda tomar un sólo valor de un parámetro, como por ejemplo $H : \mu = c$, entonces dicha hipótesis se llama **hipótesis simple.**.
Por el contrario, si la hipótesis se fomrula de tal manera que esta pueda tomar más de un posible valor, como por ejemplo

$$H : \mu \neq c$$
$$H : \mu > c$$
$$H : \mu < c$$

Entonces dicha hipótesis se llama **hipótesis compuesta**.
Si una hipótesis es simple, entonces es más fácil comprobar su validez o falsedad comparada con una hipótesis compuesta.
Para poder calcular la probabilidad de cometer un erro de tipo I, se acostumbra a formular las hipótesis que se van a probar como hipótesis simples y en muchos casos esto requiere que se suponga lo contrario de lo que se espera probar.

Ejemplo 3.1.2

1. *Si deseamos evaluar si un nuevo acero con soporte de cobre tiene mayor resistencia que el acero ordinario, formulemos la hipótesis de que las dos resistencias son las mismas.*

2. *Si deseamos demostrar que un método de instrucciones de programación de computadoras es más efectivo que otro, planteamos que los dos métodos son igualmente efectivos.*

3. *Si deseamos demostrar que la proporción de ventas calculada incorrectamente en un departamentos de una tienda es mayor que la de otro departamento, formulamos la hipótesis de que las dos proporciones son idénticas.*

Puesto que en los ejemplos anteriores supusimos que no hay diferencia de resistencia, que no hay diferencia en la efectividad de los dos métodos de instrucción de programación de computadores y que no hay diferencia entre las dos proporciones, a hipótesis como éstas se les conoce como **Hipótesis nulas** y se les representa como H_0**.**
Aunque una probabilidad positiva β de aceptar una hipótesis falsa existe para todos los valores de μ alternativos al valor de prueba, algunas veces podemos evitar por completo cometer un error de tipo II.
Para ilustrar cómo se hace esto, consideremos el siguiente:

Ejemplo 3.1.3

Supóngase que una compañía sabe por experiencia que el número promedio de errores de mecanografía cometidos en ejemplares de la forma A por mecanógrafas que elaboran estas formas es 2.3 *por día con una desviación estándar de* 0.75. *Si se desea confirmar la sospecha de que una mecanógrafa específica comete más errores en promedio que las otras, la compañía decide tomar una muestra de 9 formas escritas por esta mecanógrafa.*

a) Formule una hipótesis nula del problema
b) Formule un criterio de prueba de hipótesis nula, de tal forma que no haya posibilidad de cometer un error de tipo II
Solución:
a) $H_0 : \mu = 2.3$
Es decir, no hay diferencia entre el rendimiento de la mecanógrafa y el de las otras.
b) Rechace la hipótesis nula $\mu = 2.3$ *(y acepte la alternativa* $\mu > 2.3$ *) si la mecanógrafa promedia* 2.3 *o más errores por forma; en caso contrario, resérvese el juicio, pendiente para un mayor estudio.*
Con este criterio no hay posibilidad de cometer un error de tipo II, puesto que cuando se prueba la hipótesis ésta se puede rechazar directamente con lo cual se supone que la mecanógrafa tiene menor habilidad que el promedio); pero en caso contrario, no se aceptará en realidad.

Si la diferencia entre lo que esperamos según la hipótesis y lo que observamos mediante el estudio de una muestra, es demasiado grande para atribuirse razonablemente a la casualidad, entonces decimos que el resultado no es estadísticamente significativo.
Después aceptamos la hipótesis nula o nos reservamos al juicio, según si se requiere una decisión definitiva en una forma u otra.
Una hipótesis estadística es una información relativa a un parámetro de la población sujeta a una verificación.
También puede considerarse como la afirmación acerca de una característica de una población sobre la cual hay inseguridad de una población sobre la cual hay inseguridad en el momento de formularla y que, a la vez es expresada de tal manera que puede ser rechazada.
Una prueba de hipótesis es un procedimiento basado en evidencia de la muestra y la teoría de la probabilidad para determinar si la hipótesis es una afirmación razonable.
Existen dos tipos de hipótesis:

1. **Hipótesis nula(simple)**

2. **Hipótesis alternativa(compuesta)**

3.1.1. HIPÓTESIS NULA

La hipótesis nula es un enunciado relativo al valor de un parámetro poblacional y formulado con el fin de probar evidencia numérica, y se le denota como H_0.
La hipótesis nula **siempre** deberá incluir en su afirmación, respecto al valor de un parámetro, el signo de igualdad.

Ejemplo 3.1.4

$$H_0 : \mu = 10$$

$$H_0 : \mu \geq 10$$

$$H_0 : \mu \leq 10$$

3.1.2. HIPÓTESIS ALTERNATIVA

La hipótesis alternativa se denotará como H_1 o H_a. La hipótesis nula y la alternativa establecen lo contrario, una respecto de la otra, quedando de la siguiente manera:

$$SiH_0 : \mu = 10, entonces\ H_a : \mu \neq 10$$

$$SiH_0 : \mu \geq 10, entonces\ H_a :< \mu 10$$

$$SiH_0 : \mu \leq 10, entonces\ H_a : \mu > 10$$

La hipótesis alternativa es una afirmación, que se acepta si los datos de la muestra ofrecen suficiente evidencia para rechazar la hipótesis nula. La forma de establecer la hipótesis alternativa nos va indicar si la prueba es de uno o dos extremos *Es importante porque lo ocuparemos si rechazamos o aceptamos la hipótesis*

Si la hipótesis alternativa se formula con el signo $\neq$ diremos que la prueba es de dos extremos.
Si se formula con un signo de desigualdad estricto, sea mayor o menor, diremos que se trata de una prueba de extremo.
Diremos que es de extremo izquierdo si el signo es $<$
Diremos que es de extremo derecho si el signo es $>$

Ejemplo 3.1.5

1. *Un artículo indico que el tiempo de uso medio de los aviones comerciales es de 15 años. Para llevar a cabo una prueba estadística con esta afirmación, determinar si la hipótesis nula y la hipótesis alternativa*

   Solución:
   La hipótesis nula representa el estado actual, por lo que se escribe:

   $$H_0 : \mu = 5$$

   La hipótesis alternativa se refiere al hecho de que la afirmación es falsa, por lo que se representa lo contrario a la hipótesis nula :

   $$H_a : \mu \neq 5$$

   Por lo que se dice que la prueba es de dos extremos.

2. *Un fabricante de zapatos está considerando la compra de una nueva máquina automática para estampara calcomanías. Si μ_1 es el número de calcomanías que estampa correctamente la maquina anterior por hora y μ_2 es el promedio correspondiente de la nueva máquina, el fabricante desea demostrar la hipótesis nula $\mu_1 = \mu_2$ contra una alternativa adecuada. Cuál debe ser hipotésis alternativa, si el fabricante no desea compra la nueva máquina a menos que sea definitivamente superior a la anterior.*
   Solución:

   $$H_a : \mu_2 > \mu_1\ o\ H_a : \mu_2 - \mu_1 > 0$$

   prueba de un extremo, extremo derecho.

3. *A un ecologista le agradaría demostrar que su ciudad tiene un problema de contaminación ambiental. Específicamente desea demostrar que el nivel medio de monóxido de carbono en el aire de la ciudad es mayor a 4.9 partes por millón. Establecer la hipótesis nula y alternativa.*

Solución:

$$H_0 : \mu = 4.9(\leq)$$

$$H_a : \mu > 4.9$$

Prueba de un extremo, extremo derecho.

4. *Para tratar de promover la ciudad, la cámara de comercio probablemente esta más deseosa de concluir que el nivel medio de monóxido de carbono en dicha ciudad es menor que 4,9 partes por millón. Establecer las hipótesis nula y alternativa para este enfoque.*

Solución:

$$H_0 : \mu \geq 4.9$$

$$H_a : \mu < 4.9$$

Prueba de un extremo, extremo izquierdo

5. *Suponga que el departamento de empaques de cierta compañía se preocupa porque algunas cajas de cereal exceden considerablemente el peso especifico. El cereal se empaca en cajas de 453 gramos. Establezca las hipótesis nula y alternativa para este caso.*

Solución:

$$H_0 : \mu = 453(\leq)$$
$$H_a : \mu > 453$$

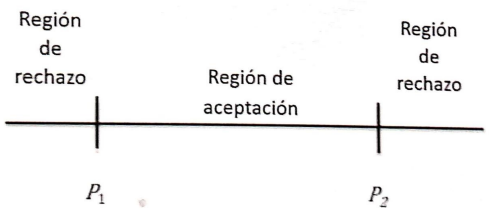
Prueba de un extremo, extremo derecho.
Nota: *Sin importar el planteamiento del problema la hipótesis nula siempre incluirá el signo de igual*

3.1.3. NIVEL DE SIGNIFICANCIA O NIVEL DE RIESGO

El nivel de significancia es la probabilidad (o riesgo) de rechazar la hipótesis nula cuando es verdadera. Este nivel de significancia se denota con α.
Un mismo nivel de significancia no es aplicable a todas las pruebas. Se puede utilizar un nivel entre 0 y 1.
Se puede utilizar un nivel de 0.05 para proyectos de investigación relacionados con consumidores; se utiliza un nivel de 0.01 para control de calidad; y un nivel de 0.10 para encuestas políticas.

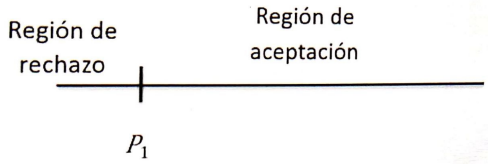
3.1.4. PUNTOS CRÍTICOS EN UNA PRUEBA DE HIPÓTESIS

Los puntos críticos son aquellos que dividen la región de análisis en subregiones de aceptación y rechazo. Los valores de estos puntos se determinan, generalmente, mediante tablas de valores. Al tener una prueba de dos extremos, existen dos puntos críticos P_1 y P_2

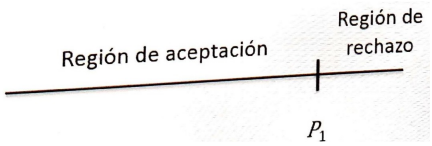


Si se tiene una prueba de un extremo, ya sea izquierdo o derecho, existirá solo un punto crítico P_1.

Extremo izquierdo:



Extremo derecho:



3.1.5. ESTADÍSTICO DE PRUEBA

El estadístico de prueba se define como el valor obtenido a partir de la información de la muestra, y nos sirve para decidir si se acepta o se rechaza la hipótesis nula.
Si el estadístico de prueba se ubica en la escala donde se establecieron los puntos críticos, entonces, la hipótesis nula se aceptará si el estadístico de prueba cae en la región de aceptación, o se rechazará si cae en la región de rechazo

3.1.6. PROCEDIMIENTO PARA PROBAR UNA HIPÓTESIS

Existe un procedimiento de seis paso que sistematiza la prueba de una hipótesis, haciendo que al llegar al paso seis uno este con la posibilidad de rechazara o no la hipótesis.
Este procedimiento se ilustra en el siguiente diagrama:

1. Identificar el tipo de prueba.
2. Establecer H_0 y H_a , e indicar si es prueba de uno o dos extremos.
3. Seleccionar un nivel de significancia
4. Determinar los puntos críticos y establecer las regiones de aceptación y rechazo.
5. Calcular el estadístico de prueba de la muestra.
6. Rechazar o no la H_0 y emitir una conclusión.

3.2. PRUEBA DE HIPÓTESIS PARA UNA MEDIA, MUESTRA GRANDE N $\geq$ 30

La hipótesis se formula como:

$$H_0 : \mu = \mu_0$$

Contra una hipótesis alterna adecuada, según sea el caso.
El estadístico de prueba es:

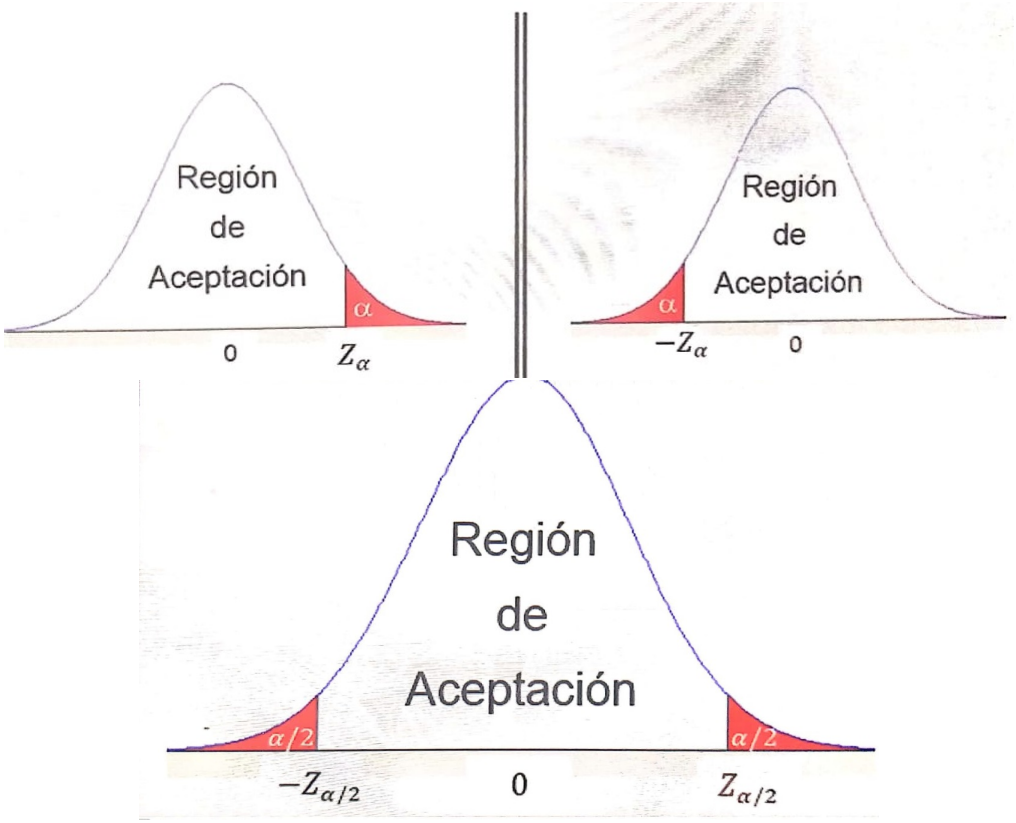
$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Para un nivel de significancia $\alpha = -Z_\alpha$ o $\alpha = Z_\alpha$ si la prueba es de extremo izquierdo o de extremo derecho, respectivamente.
El criterio de prueba es:

Hipótesis alternativa	**Rechazar H_0 si:**	**Aceptar H_0 o reservarse el juicio si:**
$\mu < \mu_0$	$Z < -Z_\alpha$	$Z \geq -Z_\alpha$
$\mu > \mu_0$	$Z > Z_\alpha$	$Z \leq -Z_\alpha$
$\mu \neq \mu_0$	$Z > Z_{\frac{\alpha}{2}}$ ó $Z < -Z_{\frac{\alpha}{2}}$	$-Z_{\frac{\alpha}{2}} \leq Z \leq Z_{\frac{\alpha}{2}}$

3.3. PROCEDIMIENTO PARA PROBAR UNA HIPÓTESIS

3.3.1. GRÁFICAS DE LAS REGIONES DE ACEPTACIÓN Y RECHAZO



Nota: La forma en que se debe concluir cuando se rechaza o no H_0, es la siguiente:

Si la decisión es rechazar H_0, la conclusión debe ser redactada de una manera aproximada como: **" Existe suficiente evidencia para indicar que se acepta H_a " o " el juicio debe reservarse hasta efectuar un mejor análisis del experimento, aumentando la muestra, rediseñando el experimento o teniendo precaución al recabar los datos "**

Si la decisión es aceptar H_0, la conclusión se redactará de la siguiente manera:
"No existe suficiente evidencia para rechazar H_0 "

Ejemplo 3.3.1

Se tomo una muestra de 40 *números de un solo dígito de la tabla de números aleatorios y se obtuvieron los siguientes valores:*

2	8	2	1	5	5	4	0
9	1	0	4	6	1	5	1
1	3	8	0	3	6	8	4
8	6	8	9	5	0	1	4
1	2	1	7	1	7	9	3

Probar si la media real de número aleatorios es de a lo más 4.5*, sabiendo que* $\sigma = 2.87$*. Utilice* $\alpha = 0.1$

Solución: Pasos

1. *Prueba de una media. muestra grande,* $n = 40$
2. *Establecemos las hipótesis*

$$H_0 : \mu = 4.5(\leq)$$
$$H_a : \mu > 4.5 \text{ Prueba de un extremo, extremo derecho.}$$

3. *Tenemos un nivel de significancia a* $\alpha = 0.1$
4. *El punto crítico para el nivel de significancia de* $\alpha = 0.1 \longrightarrow Z_\alpha = Z_{\alpha=0.1} = 1.28$ *imagen*
5. *Estadístico de prueba.*
   A partir de los datos obtenemos: $\bar{x} = 3.975$

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{3.975 - 4.5}{\frac{2.87}{\sqrt{40}}} = -1.1569$$

   El estadístico de prueba cae en la región de aceptación. ***Por lo tanto,*** *se* ***acepta*** H_0
6. ***Conclusión:*** *No existe suficiente evidencia para rechazara la afirmación de que la media real de los números aleatorios es de a lo más* 4.5*. Esto con una confiabilidad del 90 %*

Ejemplo 3.3.2

Un profesor ha registrado las calificaciones de sus estudiantes durante varios semestres y la media de estas es igual a 7.2*. Su grupo actual de* 36 *estudiantes parece tener habilidad superior, por lo que el profesor desea mostrar que, de acuerdo con su media, el grupo actual es mejor que los anteriores. ¿ Constituye el promedio del grupo* $\bar{x} = 7.52$ *suficiente evidencia para respaldar esta afirmación?. Utilizar* $\alpha = 0.05$ *y considerar que* $\sigma = 1.2$

Solución: Pasos

1. Prueba de una media, muestra grande, $n = 36$
2. Tenemos un nivel de significancia a $\alpha = 0.05$
3. Establecemos las hipótesis

$$H_0 : \mu = 7.2(\leq) \textit{El grupo actual no es superior}$$
$$H_a : \mu > 7.2 \textit{El grupo actual es superior}$$

4. El punto crítico es:

$$Z_{(\alpha=0.05)} = 1.64$$

5. El estadístico de prueba es:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{36}}} = \frac{7.52 - 7.2}{\frac{1.2}{\sqrt{36}}} = 1.6$$

Como el estadístico de prueba cae en la región de aceptación.
Decisión:
Aceptar H_0

6. **Conclusión:** El grupo actual no es mejor que los grupos anteriores.
Nota: Como los valores del punto crítico y del estadístico están muy cercanos, la decisión de la prueba pudiese ser de reservarse el juicio.

3.4. PRUEBAS CONCERNIENTES A MEDIAS MUESTRAS PEQUEÑAS N<30

Cuando no conocemos el valor de la desviación estándar de la población y la muestra es pequeña, volveremos a suponer que la población perteneciente al muestreo se aproxima **a una distribución muestral**, y basaremos nuestra decisión con el estadístico:

$$t = \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n}}}$$

Cuya distribución de muestreo es la distribución "t" con $v = n - 1$ grados de libertad.
El criterio de prueba de muestras pequeñas concernientes a medias que se basan en el estadístico "t" para rechazar una hipótesis nula, **es similar al criterio de muestra grande.**

Ejemplo 3.4.1

Supóngase que se desea decidir, con base en una muestra aleatoria de cinco ejemplares, si el contenido de grasa de cierto tipo de helado es menor que el 12 % . Según la muestra, se obtuvo una media de 11.3 % y una desviación estándar de 0.38 % . Realizar una prueba de hipótesis con un nivel de significancia de 0.01
Solución:

1. *Prueba de una media, muestra pequeña.*

2. *Establecemos las hipótesis*

$$H_0 : \mu \geq 12$$
$$H_a : \mu < 12 \text{ prueba de un extremo, extremo izquierdo}$$

3. *Tenemos un nivel de significancia de* 0.01

4. *El punto crítico esta dado por:*

$$t(v = 4, \alpha = 0.01) = -3.75$$

5. *El estadístico de prueba es:*

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{11.3 - 12}{\frac{0.38}{\sqrt{5}}} = -4.12$$

Como el estadístico cae en la región de rechazo.
Decisión:
Rechazar H_0

6. ***Conclusión:*** *La sospecha es cierta. El contenido de grasa del helado es menor que el 12 %. Esto con un nivel de confianza del 99 %*

Ejemplo 3.4.2

Se desea demostrar; en base a una muestra aleatoria de tamaño 6*, si el peso promedio de ciertos muebles es mayor que* $1000Ib$*. Qué podemos concluir a nivel de significancia de* 0.01 *sobre el peso promedio de los muebles, si los valores de la muestra son,* $987, 1146, 995, 1010, 1183, 1075$ *libras.*

Solución: Pasos

1. *Prueba de una media, muestra pequeña,* $n = 6$

2. *Establecemos las hipótesis:*

$$H_0 : \mu \leq 1000 \textit{el peso promedio no es superior}$$
$$H_a : \mu > 1000 \textit{ El peso promedio es superior}$$

3. *Tenemos un nivel de significancia a* $\alpha = 0.01$

4. *El punto crítico esta dado por:*

$$t(v = 5, \alpha = 0.01) = 3.36$$

5. *El estadístico de prueba es:*

$$t = \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n}}} = \frac{1066 - 1000}{\frac{83.17211071}{\sqrt{6}}} = 1.943756046$$

6. *Como* $t < t_\alpha$
   Decisión:*Se acepta* H_0
   Conclusión: *Existe suficiente evidencia para indicar que se acepta* H_a*. Con un nivel de confiabilidad del* 99 %

3.5. PRUEBA DE HIPÓTESIS DE UNA DIFERENCIA DE MEDIAS, MUESTRA GRANDE Y MUESTRAS INDEPENDIENTES.

Cuando se comparan las medias de dos poblaciones, generalemnte se considera la diferencia entre sus medias $\mu_1 - \mu_2$.
Las inferencias que se harán acerca de $\mu_1 - \mu_2$ estarán basadas en las diferencias entre las medias muestrales observadas $\bar{x_1} - \bar{x_2}$.
Tal inferencia pertenece a una distribución muestral, cuyas características se muestran a continuación:

$$H_0; \mu_1 - \mu_2 = D_0$$

Puesto que la distribución muestral de diferencia de medias es aproximadamente normal, se utilizará el estadístico Z en las inferencias.
El estadístico Z se determina de la siguiente manera.

$$z = \frac{(\bar{x_1} - \bar{x_2}) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Ejemplo 3.5.1

Suponga que se tiene interés en comparar el éxito académico de estudiantes universitarios pertenecientes a dos universidades distintas A y B. Los miembros de la universidad A afirman que quienes pertenecen a dicha universidad obtienen logros académicos de un nivel no inferior a los de la universidad B (la medida de éxito académico es el valor de la calificación promedio acumulado). De cada población se toman muestras de tamaño 40. Las medias obtenidas son 2.03 *para la universidad A, y* 2.21 *para la universidad B. Suponga que la desviación estándar poblacional es de* $\sigma = 0.6$ *para ambas poblaciones.*
Realice un contraste para la afirmación de los estudiantes de la universidad A, utilizando un $\alpha = 0.05$.
Solución: Pasos

1. *Prueba de una media diferencia de medias,* $n = 40$
2. *Establecemos las hipótesis:*

   $H_0 : \mu_A \geq \mu_B$ *ó* $H_0 : \mu_A - \mu_B \geq 0$
   $H_0 : \mu_A < \mu_B$ *ó* $H_a : \mu_A - \mu_B < 0$*Se trata de una prueba de un extremo, extremo izquierdo*

3. *El punto crítico es* $Z_\alpha = Z_{\alpha=0.05} = 1.64$
4. *El estadístico de prueba esta dado por:*

$$z = \frac{(\bar{x_1} - \bar{x_2}) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(2.03 - 2.21) - 0}{\sqrt{\frac{(0.06)^2}{40} + \frac{(0.06)^2}{40}}} = -1.34$$

   El estadístico de prueba cae en la región de aceptación.
   Decisión: *Aceptar* H_0
   Conclusión: *La afirmación de que los logros académicos de la universidad A no son inferiores que los de la universidad B es cierta. Esto es con una confiabilidad del 95 %*

Ejemplo 3.5.2

Se tiene interés sobre el contenido de nicotina en los cigarros. Si un experimento de 50 *cigarros de la marca A tuvieron un contenido promedio de nicotina de* $2.61mg$ *con una desviación estándar poblacional de* $0.12mg$*, mientras que para la marca B, para una muestra de* 40 *cigarros, tuvieron un contenido de nicotina promedio de* $2.38mg$ *con una desviación estándar poblacional de* $0.14mg$*. Se desea probar a nivel de confianza del 95 % si la cantidad de nicotina promedio en ambos cigarros es la misma.*
Solución:

1. *Prueba de hipótesis de diferencia de medias para muestras independientes (muestras grandes)*
2. $$H_0 : \mu_1 - \mu_2 = 0$$
   $$H_a : \mu_1 - \mu 2 \neq 0 \text{ prueba de dos extremos}$$
3. *Tiene un nivel de significancia a* $\alpha = 0.025$
4. ***Puntos críticos***

$$z_{\alpha=0.025} = 1.96$$
$$-z_{\alpha=0.025} = -1.96$$

5. *Estadístico de prueba*

$$Z = \frac{(2.61 - 2.38) - 0}{\sqrt{\frac{0.0144}{50} + \frac{0.0196}{40}}} = 8.24$$

6. *El cual cae en la región de rechazo.*
   Decisión: *Se rechaza* H_0
   Conclusión: *Con un nivel de confiabilidad del 95 % podemos afirmar que la cantidad de nicotina promedio en la marca A y en la marca B no son iguales*

3.6. PRUEBA DE HIPÓTESIS DE DIFERENCIA DE MEDIAS Y σ DESCONOCIDA(MUESTRA PEQUEÑA)

Para aplicar este criterio debemos suponer que las dos poblaciones que se muestrean tienen aproximadamente distribuciones Normales con varianzas iguales.
En forma específica, demostraremos la hipótesis nula

$$H_0 : \mu_1 - \mu_2 = D_0$$

contra una alternativa o bilateral apropiada.
En este caso, el estadístco de prueba esta dado por

$$t = \frac{(\bar{x_1} - \bar{x_2}) - D_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Donde:

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

Para los puntos críticos utilizaremos $v = n_1 + n_2 - 2$
El criterio de decisión es similar al de una prueba de μ con muestra pequeña

Ejemplo 3.6.1

Las siguientes son mediciones de la capacidad de reproducción de calor (en millones de calorías por tonelada) de muestras aleatorias de cinco ejemplares de cada una de carbón proveniente de nos minas:

Mina 1	8380	8210	8360	7840	7910
Mina 2	7540	7720	7750	8100	7690

Utilice un nivel de significación de 0.05 *para probar si es importante la diferencia entre las medias de estas dos muestras.*
Solución:

1. *Prueba de hipótesis de diferencia de medias, muestras independientes, muestra pequeña.*

2. *Establecemos las hipótesis:*

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_a : \mu_1 - \mu_2 \neq 0 \text{ prueba de dos extremos}$$

3. *Tenemos un nivel de significancia a* $\alpha = 0.05 \longrightarrow \frac{\alpha}{2} = 0.025, v = n_1 + n_2 - 2 = 8$

4. ***Puntos críticos:***

$$t(8, 0.025) = \pm 2.31$$

5. *Estadístico de prueba:*

$$\bar{x_1} = 8140 \qquad S_1 = 251.8928344$$
$$\bar{x_2} = 7760 \qquad S_2 = 206.5187643$$

$$S_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = 203.33$$
$$t = \frac{(\bar{x_1} - \bar{x_2}) - D_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 2.33$$

6. ***Decisión:***
Rechazar H_0
Conclusión: *Se puede afrimar con un 95 % de confiabilidad que la capacidad promedio de producción en ambas minas es diferente*

3.7. PRUEBAS CONCERNIENTES A DIFERENCIAS ENTRE MEDIAS DE DOS POBLACIONES RELACIONADAS. MUESTRAS DEPENDIENTES(MISMAS MUESTRAS)

En las pruebas anteriores respecto a la diferencia de medias poblacionales. se suposo que estas eran independientes. En este caso se desarrollará un procedimiento para analizar la diferencia entre las medias de dos grupos cuando se obtiene la información muestral de poblaciones que están relacionadas.
Es decir, los resultados del primer grupo no son independientes de los del segundo grupo.
Esta característica de dependencia de los dos grupos se debe a que los artículos o personas están apareados o conjuntados de acuerdo a alguna característica, o por que se obtienen mediciones repetidas del mismo grupo de artículos o personas.
En cualquiera de los dos casos la variable de interés se convierte en la diferencia entre los valores de las observaciones, más que en observaciones mismas.
El primer enfoque al problema de muestras relacionadas incluye el conjuntar artículos o personas de acuerdo a alguna característica de interés.

Ejemplo 3.7.1 ,

Si un gerente quisiera analizar el efecto de diferentes máquinas de llenado sobre la cantidad de cereal derramado y por consiguiente desperdiciado, debe establecer un control para diferenciar los diversos tipos de cereal (que quizá tengan en sí diferentes patrones de derrame).
En esta situación se pueden probar dos cajas de cada tipo de cereal, asignado una de las cajas a la nueva máquina y la otra a la antigua.
El segundo enfoque a los problemas de muestras relacionadas implica tomar mediciones repetidas de los mismos artículos o personas.
De acuerdo a la teoría de que los mismos artículos o personas se comportarán en forma parecida si se tratan en forma similar, el objetivo del análisis es mostrar que cualquier diferencia de dos mediciones de los mismos artículos o personas se debe a diferentes condiciones de tratamiento.

Ejemplo 3.7.2

Suponga que en la investigación de vienes y raíces el estadístico quisiera confirmar el avalúo de las casa para asegurarse de que no hay una diferencia real entre los valuadores al determinar el valor del bien.
Para ello se seleccionará una muestra de doce casas para su evaluación.
Es conveniente hacer que cada casa sea evaluada por las mismas personas en lugar de tomar muestras diferentes de casas.
Este enfoque sirve para reducir la variabilidad en el avalúo y permite centrar la atención en las diferencias entre los dos valuadores.

Independientemente si se utilizan muestras conjuntas (pareadas) o se toman mediciones repetidas, el objetivo es estudiar la diferencia entre dos mediciones al reducir el efecto de la variabilidad debida a los propios artículos o personas.
Para ilustrar el caso en que dos muestras son pareadas(dependientes), consideremos el siguiente ejemplo:

Ejemplo 3.7.3

Se aplicará una prueba para observar si los participantes en un curso de acondicionamiento físico realmente mejoran su nivel de condición. Se anticipa que aproximadamente 500 *personas se inscribirán en tal curso.*

- ***Plan A:*** *Seleccionar* 50 *participantes aleatoriamente de la lista de personas inscritas y aplicarles las pruebas iniciales.*
  Cuando el curso termine, se selecciona otro muestra aleatoria de 50 *y se le aplica las pruebas finales.*

- ***Plan B:*** *Seleccionar* 50 *participantes en forma aleatoria para las pruebas iniciales, y al mismo conjunto de personas de* 50 *personas se le aplican las pruebas finales al terminaer el curso.*

El **plan A** ilustra el muestreo independiente, mientras que el **plan B** utiliza el muestreo dependiente (mismas muestras).
Con el fin de determinar si hay alguna diferencia entre dos grupos relacionados, se tiene que obtener los valores individuales para cada grupo, como se muestra en la siguiente tabla:

Grupos			
Observación	**1**	**2**	**Diferencia**
1	X_{11}	X_{21}	$D_1 = X_{11} - X_{21}$
2	X_{12}	X_{22}	$D_2 = X_{12} - X_{22}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	X_{1n}	X_{2n}	$D_n = X_{1n} - X_{2n}$

De acuerdo al teorema del límite central, la diferencia promedio $\bar{D}$ representa una distribución Normal cuando se conoce la desviación estándar de la diferencia (σ_D) de la población y el tamaño de la muestra es lo suficientemente grande.
Como por lo general sólo se conoce la desviación estándar de la diferencia (S_D) de la muestra, el estadístico de prueba para la diferencia de medias entre dos muestras independientes es:

$$t = \frac{\bar{D} - D_0}{\frac{S_D}{\sqrt{n}}}$$

donde "t" es la distribución "t" de Student con $v = n - 1$, $D_0 = \mu_1 - \mu_2$

$$H_0 : \mu_1 - \mu_2 = D_0$$

Ejemplo 3.7.4

Un nuevo plan para bajar de peso sin ejercicio; para poner a prueba la afirmación de que " perderá peso en dos semanas o le devolvemos su dinero", el profesor de estadística obtuvo los pesos antes y después de 18 personas que usaron este plan. Los datos obtenidos son los siguientes

antes	146	175	150	190	220	157	136	146	128	187	172	138
Después	142	178	147	187	212	160	135	138	132	187	171	135
Diferencia	-4	3	-3	-3	-8	3	-1	-8	4	0	-1	-3

150	124	210	148	141	164
151	126	208	148	138	159
1	2	-2	0	-3	-5

Probar si el plan es efectivo con un un $\alpha = 0.05$ **Solución:**

1. Prueba de diferencia de medias, muestras pareadas.
2. Establecemos las hipótesis

$$\mu_D = \mu_{despues} - \mu_{antes}$$
$$H_0 : \mu_D \geq 0 \ \textit{El programa no es efectivo}$$
$$H_a : H_D < 0 \ \textit{El programa no es efectivo}$$

   Prueba de un extremo, extremo izquierdo

3. Tiene un nivel de significancia a $\alpha = 0.05$
4. Punto crítico

$$t(17, 0.05) = -1.74$$

$$\bar{D} = -1.555 \qquad S_D = 3.484794233$$

5. El estadístico de prueba, nos queda:

$$t = \frac{-1.555 - 0}{\frac{3.484794233}{\sqrt{18}}} = -1.893846$$

   Observación: Como estamos en el caso de muestras pareadas, por lo general la segunda muestra es cuando ya se aplico algún plan para ver el efecto respecto de la primera, entonces la segunda se pone en función de la primera.

6. **Decisión:**
   Rechazar H_0
   Conclusión: Se puede afirmar que la dieta es efectiva. Esto con un nivel de confiabilidad del 95 %

Ejemplo 3.7.1

Se llevó a cabo un estudio para determinar el grado en el cual el alcohol entorpece la habilidad de pensamiento para llevar a cabo determinada tarea. Se seleccionaron aleatoriamente diez personas de distintas características y se les pidió que participaran en el experimento. Después de porporcionarles la información pertinente, cada persona llevó a cabo la tarea sin dad de alcohol en su organismo. Entonces, la tarea volvió a llevarse a cabo, después de que cada persona había consumido una cantidad suficiente de alcohol para tener un contenido en su organimos de 0.1 % *. El tiempo, en minutos, en que cada persona terminó su tarea se muestra a continuación:*

Antes	28	22	55	45	32	35	40	25	37	20
Después	39	45	67	61	46	58	51	34	48	30 ´

Realizar una prueba de hipótesis para ver si se puede concluir que el tiempo promedio "antes" es inferior que el tiempo promedio "después" por más de 10 *minutos. Utilice una confiabilidad del* 95 % .

1. Prueba concerniente a diferencia entre medias de dos poblaciones relacionadas. Muestras dependientes (Mismas muestras).

2. $H_0 : \mu_{antes} \geq \mu_{despues}$
   $H_a : \mu_{antes} < \mu_{despues} \longrightarrow \mu_{antes} - \mu_{despues} < 0$ *prueba de un extremo, extremo izquierdo*

3. Nivel de significancia $\alpha = 0.05$

Antes	28	22	55	45	32	35	40	25	37	20
Después	39	45	67	61	46	58	51	34	48	30´
D	-11	-23	-12	-16	-14	-23	-11	-9	-11	-10

4. Punto crítico

$$t_{v=n-1,\alpha=0.05} = -1.83$$

5. Estadístico de prueba

$$t = \frac{\bar{D} - D_0}{\frac{S_D}{\sqrt{n}}} = \frac{-14}{\frac{5.142416207}{\sqrt{10}}} = -8.61$$

6. Decisión:
   Rechazar H_0.
   Conclusión: Se puede afirmar que el tiempo promedio "antes" es inferior que el tiempo promedio "después". Con un nivel de confiabilidad del 95 %

3.8. PRUEBAS CONCERNIENTES A PROPORCIONES. MUESTRAS GRANDES

En estos casos es aconsejable que el parámetro Binomial **"π"** no sea cercano a los valores extremos cero y uno, $n \cdot \pi \geq 5$ y $n(1 - \pi) \geq 5$, y que el tamaño de la muestra sea grande $(n \geq 30)$.
Puesto que las pruebas concernientes a proporciones suelen basarse en la aproximación de la curva Normal a la distribución Binomial, volveremos a utilizar el estadístico **Z** que nos llevó a obtener el intervalo de confianza de muestra grande de "π", el cual es:

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

La hipótesis nula para esta prueba es:

$$H_0 : \pi = \pi_0$$

Ejemplo 3.8.1

Un ecologista afirma que cuando mucho el 20 *% de todos los automovilistas compran gasolinas de la marca premium. Pruebe esta aseveración en* $\alpha = 0.01$*, si una revisión hecha al azar indica que* 58 *de* 200 *automovilistas compran gasolina de la marca premium.*
Solución:

1. *Prueba de una proporción, muestra grande.*
2. *Establecemos las hipótesis*

$$H_0 : \pi \leq 0.2$$
$$H_a : \pi > 0.2$$

3. *Tiene un nivel de significancia a* $\alpha = 0.01$
4. *Punto crítico*

$$Z_{\alpha=0.01} = 2.33$$

5. *Estadístico de prueba*

$$Z = \frac{\frac{58}{200} - 0.2}{\sqrt{\frac{0.2(0.8)}{200}}} = 3.182$$

Como el estadístico de prueba cae en la región de rechazo:

6. ***Decisión:***
Rechazar H_0
Conclusión: *Podemos afirmar con un 99 % de confiabilidad que más del 20 % de automovilistas compran gasolina premium*

3.9. PRUEBA DE HIPÓTESIS PARA DIFERENCIA DE PROPORCIONES. MUESTRA GRANDE

$$H_0 : \pi_1 - \pi_2 = D_\pi$$

El estadístico de prueba para este caso es:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - D_\pi}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

Ejemplo 3.9.1

De una muestra de 450 votantes hombres, 150 se declararon simpatizantes del candidato A mientras que de 550 mujeres 120 se declararon simpatizantes del candidato A. ¿ Proporcionan estos datos evidencia suficiente como para considerar que las proporciones de los simpatizantes masculinos y femeninos son diferentes? Realizar la prueba para $\alpha = 0.05$*.*
Solución:

1. *Prueba de hipótesis para diferencia de proporciones. Muestra grande*

2. *Establecemos las hipótesis:*

$$H_0 : \pi_1 - \pi_2 = 0$$
$$H_a : \pi_1 - \pi_2 \neq 0$$

Prueba de dos extremos

3. *Tiene un nivel de significancia a* $\alpha = 0.05$

4. *Puntos crítico*

$$Z_{\frac{\alpha}{2}} = 1.96$$
$$Z_{\frac{\alpha}{2}} = -1.96$$

5. *Estadístico de prueba*

$$z = \frac{\left(\frac{105}{450} - \frac{120}{550}\right) - 0}{\sqrt{\frac{\frac{105}{450}\left(1-\frac{105}{450}\right)}{450} + \frac{\frac{120}{550}\left(1-\frac{220}{550}\right)}{550}}} = .5695$$

Como el estadístico de prueba cae en la región de aceptación.

6. ***Decisión:***
Aceptar H_0
Conclusión:
Las proporciones de los simpatizantes masculinos y femeninos son iguales. Esto con una confiabilidad del 95 %

Ejemplo 3.9.2

Los administradores de los hospitales en muchos casos se encargan de obtener y calcular algunas estadísticas que son de suma importancia para los médicos y para los encargados de tomar decisiones en el hospital. En los registros de un hospital se tiene que 52 *hombre en una muestra de* 1000 *hombres y* 23 *mujeres en una muestra de* 1000 *mujeres ingresaron al hospital a causa de alguna enfermedad cardíaca. ¿ Puede o no considerarse que estos datos presentan evidencia suficiente en el sentido de que existe una mayor tasa de afecciones cardíacas en los hombres que ingresaron al hospital?.*
Considerar un $\alpha = 0.05$
Solución:

1. *Prueba de diferencia de proporciones, muestra grande.*
Las proporciones de hombres y mujeres que ingresan al hospital a causa de alguna enfermedad cardíaca las representaremos con π_1 *y* π_2*, respectivamente.*

2. *Establecemos las hipótesis.*

$$H_0 : \pi_1 - \pi_2 \leq 0$$
$$H_a : \pi_1 - \pi_2 > 0$$

3. *Tiene un nivel de significancia a* $\alpha = 0.05$

4. *Puntos críticos*

$$Z_{\alpha=0.05} = 1.64$$

5. *Estadístico de prueba*

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - D_\pi}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} = 3.42$$

Como el estadístico de prueba cae en la región de rechazo.

6. ***Decisión:***
Rechazar H_0
Conclusión: *Los datos proporcionados arrojan evidencia suficiente para considerar que hay una mayor tasa de afección cardíaca en hombres que en mujeres. Esto con una confiabilidad del 95 %*

Ejemplo 3.9.3

A menudo las compañías industriales emplean métodos de "transferencia de riesgos". Utilizan un seguro o las cláusulas de indemnización en los contratos como una técnica de administración de riesgos. En una muestra de 43 *compañías petroleras,* 22 *indicaban que la transferencia de riesgos fue determinante mientras que en una muestra de* 93 *compañías constructoras* 55 *confirmaron lo anterior. Mediante una prueba de hipótesis, ¿ Se puede concluir, con un nivel de confianza del* 95 %*, que la proporción de compañías petroleras que emplean el método de transferencia de riesgos es menor que la proporción de compañías constructoras que lo hacen.*

1. Prueba de hipótesis para diferencia de proporciones. Muestra grande $n \geq 30$

2.

$$H_0 : \pi_{petrolera} \geq \pi_{constructora}$$
$$H_a : \pi_{petrolera} < \pi_{constructora} \longrightarrow \pi_{petrolera} - \pi_{constructora} < 0 \ \textit{Extramo izquierdo}$$

3. Nivel de significancia $\alpha = 0.05$

4. Punto crítico

$$-Z_{\alpha=0.05} = -1.64$$

5. Estadístico de prueba

$$Z = \frac{\hat{p}_{petrolera} - \hat{p}_{constructora}}{\sqrt{\frac{\hat{p}_{petrolera}(1-\hat{p}_{petrolera})}{n_p} + \frac{\hat{p}_{constructora}(1-\hat{p}_{constructora})}{n_{constructora}}}}$$
$$= \frac{\frac{22}{43} - \frac{55}{93}}{\sqrt{\frac{\frac{22}{43}\left(1-\frac{22}{43}\right)}{43} + \frac{\frac{55}{93}\left(1-\frac{55}{93}\right)}{93}}} = -0.8698873722$$

6. **Decisión:**
Aceptar H_0
Conclusión: Los datos proporcionados arrojan evidencia suficiente para considerar que la proporción de compañías petroleras emplean el método de transferencia de riesgos mayor que la proporción de compañías constructoras que lo hacen. Esto con un nivel de confiabilidad del 95 % .

3.10. PRUEBA CONCERNIENTE A UNA VARIANZA. MUESTRA PEQUEÑA

La prueba de hipótesis nula $\sigma = \sigma_0$ se basa en las mismas suposiciones, el mismo estadístico y la misma teoría de muestreo que el intervalo de confianza de muestra pequeña de σ.
Si la muestra aleatoria proviene de una población con distribución aproximadamente Normal, basamos nuestra decisión en el estadístico.

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

con $v = n - 1$ grados de libertad. Donde n es el tamaño de la muestra, S^2 es la varianza de la muestra, y σ_0 es el valor hipotético establecido en la hipótesis nula. El criterio de prueba de muestra en el siguiente cuadro:

Hipótesis Alternativa	**Rechazar la Hipótesis Nula sí:**
$\sigma < \sigma_0$	$\chi^2 < \chi^2(v, 1-\alpha)$
$\sigma > \sigma_0$	$\chi^2 > \chi^2(v, \alpha)$
$\sigma \neq \sigma_0$	$\chi^2 < \chi^2(v, 1-\frac{\alpha}{2})$ *o* $\chi^2 > \chi^2(v, \frac{\alpha}{2})$

Ejemplo 3.10.1

Las especificaciones de soportes producidos en masa de cierto tipo requiere, entre otras cosas, que las desviaciones estándar de sus diámetros exteriores no pasen de $0.005cm$. *Con base en una muestra aleatoria de tamaño* 12 *para la cual* $s = 0.0077cm$, *indicar si las especificaciones de los soportes están dentro de lo requerido,* $\alpha = 0.01$
Solución:

1. *Prueba concerniente a una varianza. Muestra pequeña*

2. *Establecemos las hipótesis:*

$$H_0 : \sigma \leq 0.005$$
$$H_a : \sigma > 0.005$$

Prueba de un extremo, extremo derecho

$$\sigma_0 = 0.005 \qquad \sigma_0^2 = 2.5x10^{-5} \qquad s = 0.0077$$
$$s^2 = 5.929x10^{-5} \qquad n = 12 \qquad v = n - 1 = 11$$

3. *Tiene un nivel de significancia a* $\alpha = 0.01$

4. *Punto crítico*

$$\chi^2(v = 11, \alpha = 0.01) = 24.7$$

5. *Estadístico de prueba*

$$\chi^2 = \frac{(12)(5.929x10^{-5}}{2.5x10^{-5}} = 28.46$$

Ejemplo 3.10.2

La empresa A fábrica una amplia linea de instrumentos de precisión y tiene una buena reputación en el campo por calidad de sus instrumentos. Con el fin de conservar su reputación mantiene un control de calidad en sus productos. No pondrá a la venta una balanza analítica a menos que dicha balanza muestre una variabilidad que esté significativamente debajo de un microgramo, cuando se pesan cantidades de aproximadamente $500gr$*. Una nueva balanza acaba de ser entregada a la división de control de calidad por parte de la línea de producción. Se prueba la nueva balanza pata pesar el mismo peso estándar de* $500gr$ *29 veces distintas. La desviación estándar de la muestra resultó de* 0.73 *microgramos. ¿ Se deberá vender la balanza?. Considere un* $\alpha = 0.01$

Solución:

1. *Prueba de una varianza, muestra pequeña.*

2. $$H_0 : \sigma \geq 1$$
   $H_a : \sigma < 1$ *Prueba de un extremo, extremo izquierdo*

3. *Con un nivel de significancia del* $\alpha = 0.01$

   $\sigma_0 = \sigma_0^2 = 1$ $\quad S = 0.73$ $\quad S^2 = 0.5329$
   $n = 29$ $\quad v = n - 1 = 28$

4. ***Punto crítico***

   $$\chi^2_{(v=28,1-\alpha=0.99)} = 13$$

5. ***Estadístico de prueba***

   $$\chi^2 = \frac{(28)(0.5329)}{1} = 14.92$$

6. *Como el estadístico de prueba cae en la región de aceptación:*
   Decisión:
   Aceptar H_0
   Conclusión:
   No se deberá vender la balanza. Esto con una confiabilidad del 99 %

3.11. PRUEBA CONCERNIENTES A σ y σ^2 CON MUESTRAS GRANDES ($n \geq 30$)

Cuando n es grande podemos basar las pruebas de las hipótesis nula $\sigma = \sigma_0$ en la misma teoría que utilizamos en la construcción de intervalos de confiana de muestra grande. Estos es, utilizamos el estadístico.

$$Z = \frac{S - \sigma_0}{\frac{\sigma_0}{\sqrt{2n}}}$$

cuya distribución muestral es aproximadamente Normal. En este caso, el análisis de la prueba se lleva a cabo en forma similar a como se trató la prueba de una media de muestra grande.

Ejemplo 3.11.1

La variabilidad de las ventas de un almacén en una muestra aleatoria de 50 día la mide la desviación estándar $S = 2250$ *pesos. Utilizar un nivel de significancia de* 0.01 *para rechazar o aceptar la afirmación de que la desviación estándar real es del al menos* 3000 *pesos.*

1. *Prueba concerniente a una varianza (muestra grande).*

2. $$H_0 : \sigma \geq 3000$$
$$H_a : \sigma < 3000 \textit{Extremo izquierdo}$$

3. *Nivel de significancia* $\alpha = 0.01$

4. *Punto crítico* $Z_{\alpha=.01} = -2.33$

5. *Estadístico de prueba*

$$\sigma_0 = 3000 \qquad S = 2250 \qquad n = 50 \qquad \alpha = 0.01$$

$$Z = \frac{2250 - 3000}{\frac{3000}{\sqrt{2(50)}}}$$

6. *Como el estadístico de prueba cae en la región de rechazo:*
   Decisión:
   Rechazar H_0
   Conclusión:
   En realidad, la variabilidad de las ventas es menor a 3000 pesos. Esto con una confiabilidad del 99 %

3.12. PRUEBAS CONCERNIENTES A LA IGUALDAD O COCIENTE DE DOS VARIANZAS (O DESVIACIONES ESTÁNDAR)

En esta ocasión analizaremos una prueba concerniente a la igualdad o cociente de las varianzas de dos poblaciones.
A menudo, esta prueba se utiliza en relación con la prueba de muestra pequeña de la diferencia entre dos medias, que requieren que las varianzaas de las dos poblaciones sean iguales.

Dadas muestras aleatorias independientes de tamaño n_1 y n_2 de dos poblaciones, por lo general basamos la prueba de la igualdad de las dos varianzas de las poblaciones en las razones $\frac{S_1^2}{S_2^2}$ o bien $\frac{S_2^2}{S_1^2}$, donde S_1^2 y S_2^2 son las varianzas de las muestras.
Si las poblaciones de las cuales provinieron las muestras son aproximadamente Normales, entonces la distribución muestral del cociente de varianzas es la **distribución F**. Recordemos que esta distribución depende de los parámetros $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$, que son los grados de libertad en las estimaciones de las muestras para determinar su varianza respectiva.
Por lo tanto, **el estadístico de prueba concerniente a la igualdad de dos varianzas es:**
La hipótesis nula será siempre

$$H_0 : \sigma_1^2 = \sigma_2^2 \qquad o \qquad H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$

Por otra parte, la hipótesis alternativa podría ser cualquiera de las siguientes, **considerando a la muestra 1 como aquella que tenga la mayor varianza muestral(Re etiquetar en caso necesario):**

1. $H_a : \sigma_1^2 < \sigma_2^2$ Prueba unilateral de extremo izquierdo.
   El estadístico de la prueba es $F = \frac{S_1^2}{S_2^2}$
   Rechazar H_0 si $F < F(v_1, v_2, \alpha)$
2. $H_a : \sigma_1^2 > \sigma_2^2$ Prueba unilateral de extremo derecho.
   El estadístico de la prueba es $F = \frac{S_1^2}{S_2^2}$
   Rechazar H_0 si $F > F(v_1, v_2, \alpha)$
3. $H_a : \sigma_1^2 \neq \sigma_2^2$ Prueba de dos extremos.
   El estadístico de prueba es $F = \frac{S_1^2}{S_2^2}$
   Rechazar H_0 si

$$F \notin \left[\frac{1}{\left(F(v_1, v_2, \frac{\alpha}{2}\right)}, F\left(v_1, v_2, \frac{\alpha}{2}\right) \right]$$

$$\frac{S_1^2}{S_2^2}$$

La varianza muestral mayor deberá ir siempre en el numerador del cociente, e identificará a la población 1.

Ejemplo 3.12.1

La consistencia en el sabor de la cerveza es una cualidad importante para mantener la lealtad de la clientela. La variabilidad en el sabor de una cerveza dada puede verse afectada por la longitud del período de fermentación, variación en los ingredientes y diferencias en el equipo de fermentación. Un fabricante con dos líneas de producción. 1 *y* 2*, ha hecho ligeros cambios a la línea* 2 *buscando reducir la variabilidad, así como el promedio del índice del sabor. Se toman al azar muestras de* $n_1 = 25$ *y* $n_2 = 25$ *vasos de cerveza de cada línea de producción y se determina el índice de sabor con una garganta apropiada, obteniéndose que*

$$\bar{x_1} = 3.0 \qquad S_1^2 = 0.51$$
$$\bar{x_2} = 3.2 \qquad S_2^2 = 1.04$$

Para un nivel de significancia de 0.05*,¿ Presentan estos datos suficiente evidencia para indicar que la variabilidad del proceso es menor para la línea 2?*
Solución:

1. *Prueba concerniente a la diferencia de dos varianzas o desviaciones estándar*

2. *Establecemos las hipótesis:*

$$H_0 : \sigma_1^2 \geq \sigma_2^2$$
$$H_a : \sigma_1^2 < \sigma_2^2 \text{ Prueba de extremo izquierdo}$$

$$m_1(L1) : S_1^2 = 1.04 \qquad n_1 = 25 \qquad v_1 = 24 \qquad \sigma_1^2$$
$$m_2(L2) : S_2^2 = 0.51 \qquad n_1 = 25 \qquad v_1 = 24 \qquad \sigma_2^2$$

3. *Tiene un nivel de significancia a* $\alpha = 0.05$

4. *Punto crítico*

$$F(v_1 = 24, v_2 = 24, \alpha = 0.05) = 1.98$$

5. *Estadístico de prueba*

$$F = \frac{1.04}{0.51} = 2.04$$

6. *Como* $F > F(v_1, v_2, \alpha$
   Decisión:
   Aceptar H_0
   Conclusión:
   La variabilidad de la línea 1 no es mayor que la variabilidad de la línea 2. Esto con una confiabilidad del 95 %

Ejemplo 3.12.2

El riesgo asociado con inversiones alternativas se evalúa generalmente por medio de la varianza de los réditos asociados con cada inversión. Se analizará el riesgo de dos inversiones distintas (AyB). La tasa esperada de redituabilidad para cada inversión es del 17.8 % *, pero con base en los réditos observados en los* 10 *años anteriores para la inversión A y los* 8 *años anteriores de B, las varianzas para los réditos para ambas inversiones son de* 3.21 *y* 7.14 *respectivamente del* 0.1*, ¿ Las varianzas dadas presentan suficiente evidencia para indicar que los riesgos de las inversiones A y B son distintos?*
Solución:

1. *Prueba concerniente a la diferencia de dos varianzas o desviaciones estándar.*

2. *Establecemos las hipótesis:*

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_a : \sigma_1^2 \neq \sigma_2^2 \text{ prueba de dos extremos}$$

$$m_1(lnvB) : S_1^2 = 7.14 \qquad n_1 = 8 \qquad v_1 = 7 \qquad \sigma_1^2$$
$$m_2(lnvA) : S_2^2 = 3.21 \qquad n_1 = 10 \qquad v_1 = 9 \qquad \sigma_2^2$$

3. *Tiene un nivel de significancia a* $\alpha = 0.1$

4. *Puntos crítico*

$$F\left(v_1 = 7, v_2 = 9, \frac{\alpha}{2} = 0.05\right) = 3.29$$
$$\frac{1}{F\left(v_1 = 7, v_2 = 9, \frac{\alpha}{2} = 0.05\right)} = \frac{1}{3.29} = 0.30395$$

5. *Estadístico de prueba:*

$$F = \frac{S_1^2}{S_2^2} = \frac{7.14}{3.21} = 2.22$$

6. *Como el estadístico de prueba cae en la región de aceptación*
   Decisión:
   Aceptar H_0
   Conclusión: *Los riesgos de las inversiones A y B son iguales.*

Ejemplo 3.12.3

El gerente de producción de una compañía afirma que existe igual variabilidad en el rendimiento del segundo turno de trabajo que la variabilidad en el rendimiento del primer turno de trabajo. El jefe de producción cree que la del segundo turno es mayor que la del primero, por lo cual toma una muestra del rendimiento de 21 *obreros del segundo turno cuya varianza es* 4.8 *y toma una muestra del rendimiento de* 16 *obreros del primer turno cuya varianza es* 2.9*. Probar si el jefe de producción tiene la razón o no con un nivel de significancia del 5 %*

1. *Prueba concerniente a la diferencia de dos varianzas o desviaciones estándar.*

2. *Establecemos las hipótesis:*

$$H_0 : \sigma_1^2 \leq \sigma_2^2$$
$$H_a : \sigma_1^2 > \sigma_2^2 \text{ prueba de dos extremos}$$

$$\begin{array}{rllll} m_1(Segundoturno) : S_1^2 = 4.8 & n_1 = 21 & v_1 = 21 - 1 = 20 & \sigma_1^2 \\ m_2(primerturno) : S_2^2 = 2.9 & n_1 = 19 & v_1 = 19 - 1 = 18 & \sigma_2^2 \end{array}$$

3. *Tiene un nivel de significancia a* $\alpha = 0.05$

4. *Punto crítico*

$$F\left(v_1 =, v_2 =, \alpha =\right) = \frac{4.8}{2.9} = 2.195$$

5. *Estadístico de prueba:*

$$F = \frac{S_1^2}{S_2^2} = \frac{4.8}{2.9} = 1.65$$

6. *Como* $F < F(v_1, v_2, \alpha)$
   Decisión:
   Aceptar H_0
   Conclusión: *El gerente no tiene razón. Es decir, la variabilidad en el rendimiento del segundo turno es menor igual a la variabilidad en el rendimiento del primer turno. Esto con una confiabilidad del 95 %*

Ejemplo 3.12.4

3.13. CRITERIO DEL "VALOR P" PARA DECIDIR SI SE ACEPTA O RECHAZA LA HIPÓTESIS NULA

En años recientes por la disponibilidad del software de computadora, con frecuencia se da información relacionada con la seguridad del rechazo a la aceptación.
Es decir, ¿ Cuánta confianza hay en el rechazo de la hipótesis nula?.
Este enfoque indica la probabilidad (en el supuesto de que la hipótesis nula sea verdadera) de obtener un valor estadístico de la prueba por lo menos tan extremo como el valor real obtenido.
Este proceso compara la probabilidad, denominada **valor p**, con el nivel de significancia adecuado.

DEFINICION (VALOR P)

El valor p es la probabilidad de observar el valor muestral tan extremo o más que el valor observado si la hipótesis nula es verdadera.
La determinación del valor p no sólo da como resultado una decisión respecto de H_0, sino que brinda la oportunidad de observar la fuerza de la decisión.
Un valor p muy pequeño como 0.0001, indica que existe poca probabilidad de que H_0 sea verdadera.
Por otra parte, un valor de p de 0.2033 significa que H_0 no se rechaza y que existe poca probabilidad de que sea falsa.

CÁLCULO DEL VALOR P

1. Si se tiene una prueba de dos extremos y T es el valor del estadístico de prueba, entonces el valor p se define como:
$$p = 2p(z \geq T)$$

2. Si se tiene una prueba de extremo izquierdo y $-T$ es el valor del estadístico de prueba, entonces p se define como:
$$p = (z \leq -T)$$

3. Si se tiene una prueba de extremo derecho y T es l alor del estadístico de prueba, entonces p se define como:
$$p = p(z \geq T)$$

CRITERIO PARA ACEPTAR O RECHAZAR H_0

Si $p < \alpha$
Decisión: Rechazar H_0

Si $p > \alpha$
Decisión:No Rechazar H_0

Ejemplo 3.13.1

Determine el valor de p y establezca la decisión de la prueba en cada uno de los siguientes casos:

1. *Se utiliza el nivel de significancia* $\alpha = 0.05$ *para probar la aseveración de que* $\pi > 0.25$*, y los datos muestrales dan por resultado un estadístico de prueba de* $Z = 1.18$
   Solución:

$$H_0 : \pi \leq 0.25$$
$$H_a : \pi > 0.25 \textit{Prueba de extremo derecho}$$

   Por lo cual

$$p = p(z \geq T) = p(z \geq 1.18) = 1 - p(z \leq 1.18)$$
$$= 1 - 0.881 = 0.119$$

   Como

$$p > \alpha = 0.05$$

   Decisión: *No rechazar* H_0

2. *Se utiliza el nivel de significancia* $\alpha = 0.05$ *para probar la aseveración de que* $\pi \neq 0.25$*, y los datos muestrales dan por resultado un estadístico de prueba de* $Z = 2.34$
   Solución:

$$H_0 : \pi = 0.25$$
$$H_a : \pi \neq 0.25 \textit{Prueba de dos extremos}$$

   Por lo cual

$$p = 2p(z \geq 2.34) = 2[1 - p(<\leq 2.34)]$$
$$= 2[1 - 0.9904] = 0.0192$$

   Como

$$p < \alpha = 0.05$$

   Decisión: *Rechazar* H_0 *Utilizar el criterio del valor p para resolver el siguiente problema:*

3. *Un profesor ha registrado las calificaciones de sus estudiantes durante varios semestres y la media de estas es igual a* 7.2*. Su grupo actual de 36 estudiantes parece tener una habilidad superior, por que el profesor desea mostrar que de acuerdo con su media, el grupo actual es mejor que los anteriores.*
   ¿ Constituye el promedio del grupo $\bar{x} = 7.52$ *suficiente evidencia para respaldar esta afirmación?*
   Utilizar $\alpha = 0.05$ *y considerar que* $\sigma = 1.2$

$$H_0 : \mu = 7.2(\leq) \textit{ El grupo actual no es superior.}$$
$$H_a : \mu > 7.2 \textit{ El grupo actual es superior.}$$

   Prueba de un extremo, extremo derecho.
   El estadístico de prueba es:

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{7.52 - 7.2}{\frac{1.2}{\sqrt{36}}} = 1.6$$

$$p = p(z > 1.6) = 1 - p(z \leq 1.6)$$
$$= 1 - 0.9452 = 0.0548$$

   Como $p > \alpha = 0.05$
   Decisión: *No rechazar* $H_0 \therefore$ *aceptamos* H_0
   Conclusión: *El grupo actual no es mejor que los grupos anteriores*

3.14. ANÁLISIS DE VARIANZA ANOVA DE UN FACTOR

En temas anteriores se presentaron métodos para comparar dos medias poblacionales, basadas en muestras aleatorias independientes y en un experimento de diferencias de medias pareadas.
En este nuevo tema se ampliará estos análisis para la comparación de cualquier número de medias poblacionales, usando una técnica llamada análisis de varianza.
ANOVA son las siglas de Analysis of Variance.
Primero nos centraremos en el ANOVA de un factor:

3.14.1. COMPARACIÓN MÚLTIPLE DE MEDIAS

ANOVA de un factor es una técnica estadística que señala si dos varables (una independiente y otra dependiente) están relacionadas en base a si las medias de la variable dependiente son diferentes en las categorías o grupos de la variable independiente.
Es decir, señala si las medias entre dos o más grupos son similares o diferentes. Se le denomina ANOVA de un factor porque a la variable independiente se le conoce como factor.

¿ CUANDO USAR ANOVA DE UN FACTOR ?

Usamos ANOVA de un factor **cuando queremos saber si las medias de una variable son diferentes entre los niveles o grupos de otra variable.**

Ejemplo 3.14.1

Si comparamos el número de hijos entre los grupos o niveles de clase social:

Clase social

Clase baja Clase trabajadora *Clase media-baja* *Clase media-alta* *Clase alta*

Es decir, vamos a comprobar mediante ANOVA si la variable dependiente "número de hijos" está relacionada con la variable independientes " clase social".
' Concretamente, se analizará si la media del número de hijos varía según el nivel de clase social a la que pertenece la persona.
Condición :

- *En ANOVA de un factor solo se relacionan **dos variables**: una variable dependiente (o a explicar) y una variable independiente (que en esta técnica se suele llamar factor).*
- *La **variable dependiente es cuantitativa** (escalar) y la **variable independiente es categórica** (nominal u ordinal).*
- *Se pide que las variables sigan una distribución normal, aunque como siempre esto es difícil de cumplir en investigaciones sociales.*
- *También se pide que las varianzas de cada grupo de la variable dependiente sean similares (fenómeno que se conoce como homocedasticidad). Aunque esto es lo ideal, en realidad esta es difícil de cumplir, pero igualmente se puede aplicar ANOVA.*
- *Los datos se recolectan al azar y son independientes de los demás.*

¿ EN QUÉ SE BASA ANOVA DE UN FACTOR ?

ANOVA de un factor compara las medias de la variable dependiente entre los grupos o categorías de la variable independiente.

Ejemplo 3.14.2

Comparamos las medias de la variable " Número de hijos "según los grupos o categorías de la variable " clase social" .
Si las medias de la variable dependiente son iguales en cada grupo o categoría de la variable independiente, los grupos no difieren en la variable dependiente, y por tanto no hay relación entre las variables.
En cambio, y siguiendo con el ejemplo, si las medias del número de hijos son diferentes entre los niveles de la clase social es que las variables están relacionadas.

¿ QUÉ ESTADÍSTICO SE CALCULAN EN ANOVA ?

Al aplicar ANOVA de un factor se calculan un estadístico o test denominado F y su significación.
El estadístico F o F-test(se llama F en honor al estadístico Ronald Fisher) se obtiene al estimar la variación de las medias entre los grupos de la variable independiente y dividirla por la estimación de la variación de las medias dentro de los grupos. El cálculo del estadístico F es algo complejo de entender, pero lo que hace es dividir la variación entre los grupos por la variación dentro de los grupos.
Si las medias entre los grupos varían mucho y la media dentro de un grupo varía poco, es decir, los grupos son heterogéneos entre ellos y similares internamente el valor de F será más alto, y por tanto, las variables estarán relacionadas.
En conclusión, **cuanto más difieran las medias de la variable dependiente entre los grupos de la variable independiente, más alto será el valor de F.** Si hacemos varios análisis de ANOVA de un factor, aquel con F más alto indicará que hay más diferencias y por tanto una relación más fuerte entre las variable.
Para ilustrar el uso del análisis de varianza, consideremos el siguiente ejemplo:

Ejemplo 3.14.3

Supóngase que se desea verificar si la temperatura es un factor importante que influye en la tasa de producción de una planta industrial.
Para esto se registra el número de unidades producidas en una hora durante horarios periódicos seleccionados de manera aleatoria, cuando el proceso de producción en la fábrica se hallaba en actividad en cada uno de tres niveles de temperatura. Se obtuvieron cuatro registros de producción para dos niveles de temperatura y cinco para el tercer nivel. Los datos se muestran en la siguiente tabla:

Niveles de temperatura.

Muestra obtenida a 68 ***ºF***	***Muestra obtenida a*** 72 ***ºF***	***Muestra obtenida a*** 76 ***ºF***
10	7	3
12	6	3
10	7	5
9	8	4
	7	
$\bar{\chi_1} = 10.25$	$\bar{\chi_2} = 7.0$	$\bar{\chi_3} = 3.75$

n_i	4	5	4
T_i	41	35	15

¿Son suficientes los datos de las tres muestras para indicar una diferencia en las tasas medias de producción entre los tres tipos de temperatura? considere un $\alpha = 0.05$
Solución:
El factor en consideración es "La temperatura". Considerar las siguientes hipótesis

$H_0 : \alpha_{68} = \alpha_{72} = \alpha_{76}$ *(La temperatura no tiene un efecto significativo sobre la tasa de producción)*
H_a : *Al menos una de las medias de producción es distinta a las demás*

El factor a prueba, la temperatura de la planta tiene tres niveles: $68, 72$ ***y*** 76
Para contestar a la pregunta, podríamos comparar las medias por pares:

$$H_0 : \mu_1 = \mu_2 \qquad H_0 : \mu_1 = \mu_3 \qquad H_0 : \mu_2 = \mu_3$$

y repetir la prueba ***t-student*** *vista anteriormente.*
Si detectamos una diferencia entre cualquier pareja de las medias anteriores, entonces concluiríamos que existe una evidencia de que hay por lo menos una diferencia entre las medias y parecería que se hubiera contestado así la pregunta.
El problema con este procedimiento es que hay $\binom{3}{2} = 3$ *pares de valores medios que probar.*
Aunque fueran iguales todas las medias, se tendría una probabilidad α *de rechazar la hipótesis nula de que los valores de un par en particular son iguales*

Al repetir este procedimiento tres veces, la probabilidad de concluir erróneamente que por lo menos un par de valores medios difiera, se incrementa considerablemente (más aún, cuando el numero de muestras es mayor).
Debido a que el riesgo de una decisión equivocada puede ser elevado, buscamos una prueba única de la hipótesis nula de que son iguales las medias de los diferentes grupos o muestras:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_n$$

contra la alternativa de que al menos un par de valores medios es diferente.
Un análisis de varianza para comparar más de dos medias poblacionales formaliza la comparación visual de la variación entre medias con la variación dentro de las muestras.
Esta comparación de dos fuentes de variación se llevará a la prueba F del análisis de varianza.
Las suposiciones que forman la base de los procedimientos de prueba y estimación, para un ANOVA, son similares a las requeridas para la estadística t de Student.
Sin importar el procedimiento de muestreo utilizado para recopilar los datos, se debe suponer que las observaciones pertenecen a una población que se distribuye Normalmente, con una varianza común σ^2

3.14.2. ANÁLISIS DE VARIANZA DE UN SOLO FACTOR PARA MUESTRAS ALEATORIAS INDEPENDIENTES

Supóngase que deseamos comparar k medias poblacionales $\mu_1, \mu_2, \ldots, \mu_k$ basándonos en muestras aleatorias independiente de $n_1, n_2, \ldots, n_k$ respectivamente. Como sabemos, la varianza muestral se define como:

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \tag{3.1}$$

o como

$$S^2 = \frac{\sum x_i^2 - \frac{(\sum \bar{x})^2}{n}}{n-1} \tag{3.2}$$

donde no se requiere el uso de $\bar{x}$.
El procedimiento del ANOVA separará la variación entre todo el conjunto de datos en dos

categorías.
Para lograr esta separación, primero trabaja con el numerador de (1).
El numerador de (1) lo definiremos como suma de cuadrados (SC), es decir:

$$SC = \sum (x_i - \bar{x})^2$$

y al numerador de (2) lo definiremos como suma de cuadrados total, SCT, es decir:

$$SCT = \sum x_1^2 - \frac{(\sum x)^2}{n}$$

El análisis de varianza separa la variación entre el conjunto de datos, SCT, en dos categorías; SCF y çSCE, de tal forma que:

$$SCT = SCF + SCE$$

En este caso, SCF (debida a niveles de factor) se define como:

$$SCF = \sum \left(\frac{T_i^2}{n_i} \right) - \frac{(\sum x_i)^2}{n}$$

Donde T_i representa el total de columna; n_i el número de repeticiones en cada nivel del factor; y n representa el tamaño de muestra total ($n = \sum n_i$).
La SCE se define como:

$$SCE = (\sum x_i^2) - \sum \left(\frac{T_i^2}{n_i} \right)$$

La SCE se debe al error experimental de réplica.
Por otra parte, también se definen los siguientes conceptos:
Los grados de libertad del factor, $v(factor)$, se define como el número de muestras (k) menos uno, es decir;

$$v(factor) = k - 1$$

Los grados de libertad del total, v(Total) o V(SCT), se define como el número total de datos menos uno, es decir:

$$v(SCT) = n - 1$$

Los grados de libertad del error, v(E), son la suma de grados de libertad para todos los niveles puestos a prueba (columnas en la tabla de datos). Cada columna tiene grados de libertad, por lo tanto:

$$v(E) = (n_1 - 1) + (n_2 - 2) + \ldots = (n - k)$$

Con lo que se tiene hasta ahora, para registrar la suma de cuadrados y organizar los datos restantes, se debe utilizar el siguiente formato, comúnmente conocido como tabla de ANOVA:

Fuente de Variación	**SC**	**GI**	**CM**
Factor	SCF	$k-1$	CMF
Error	SCE	$n-k$	CME
Total	SCT	$n-1$	

Con base a los conceptos previos vistos anteriormente, se definen otros nuevos conceptos como los siguientes:
El cuadrado medio del factor, CMF, el cual se denota y define de la siguiente manera:

$$CMF = \frac{SCF}{v(F)}$$

El cuadrado medio del error se denota y define de la siguiente manera:

$$CME = \frac{SCE}{v(E)}$$

Finalmente, el estadístico de prueba se basa en la distribución F definida como: $F = \frac{CMF}{CME}$

3.14.3. CRITERIO DE LA PRUEBA DE F, BASADO EN EL ANOVA DE UN SOLO FACTOR PATA COMPARAR K MEDIAS POBLACIONALES.

1. Hipótesis nula: $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$
2. Hipótesis alternativa: $H_a :$ *Uno o más pares de medias poblacionales difierene*
3. Estadístico de prueba $F = \frac{CMF}{CME}$
4. Punto crítico de comparación $F_\alpha = F(v_1 = k-1, v_2 = n-k, \alpha)$
5. Región de rechazo: Rechazar H_0 si $F > F_\alpha$ donde F_α se encuentra en la cola superior de la distribución $F_\alpha = F(v_1 = k-1, v_2 = n-k, \alpha)$ y satisface la expresión $P(F > F_\alpha) = \alpha$

En dicho criterio se deben considerar las siguientes suposiciones:

1. Se han seleccionado las muestras aleatoria e independientemente de sus poblaciones respectivas.
2. Las poblaciones se distribuyen Normalmente con medias $\mu_1, \mu_2, \mu_3, \ldots, \mu_k$ y varianzas iguales $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2$

Ejemplo 3.14.4

Ojo derecho	***Ojo izquierdo***	***Ambos ojos***
12	10	16
10	17	14
18	16	16
12	13	11
14		20
		21

Solución:

$$H_0 : \mu_D = \mu_I = \mu_A$$
$$H_a : \mu_D \neq \mu_I \neq \mu_A$$

Las repeticiones son los puntos ganados por los tiradores en cada grupo.

$n_1 = 5$ $\quad n_2 = 4$ $\quad n_3 = 6$

$T_1 = 66$ $\quad T_2 = 56$ $\quad T_3 = 98$

$n = \sum n_i = 15$

$k = 3($ *número de medias)*

$$SCF = \sum \frac{T_i^2}{n_i} - \frac{(\sum x_i)^2}{n} = \left(\frac{66^2}{5} + \frac{56^2}{4} + \frac{98^2}{6}\right) - \frac{220^2}{15} = 3255.87 - 3226.67 = 29.2$$

$$SCE = \sum x_i^2 - \sum \frac{T_i^2}{n_i} = 3392 - 3255.87 = 136.13$$

$$CMF = \frac{SCF}{v(F)} = \frac{29.2}{k-1} = \frac{29.2}{2} = 14.6$$

$$CME = \frac{SCE}{v(E)} = \frac{136.13}{n-k} = \frac{136.13}{15.3} = 11.34$$

Punto crítico:
$F_\alpha = F(v_1 = k-1 = 2, v_2 = n-k = 12, \alpha = 0.05) = 3.89$
Estadístico de prueba

$$F = \frac{CMF}{CME} = \frac{14.6}{11.34} = 1.287$$

Como $F < F_\alpha$
Decisión: *Aceptar* H_0
Conclusión: *No hay evidencia suficiente para rechazar la hipótesis de que los tres métodos son igualmente eficaces. Esto con una confiabilidad del 95 %*

Ejemplo 3.14.5

Consideremos 4 compañías A, B, C *y* D*, cuyas acciones cotizan en Bolsa y seleccionamos aleatoriamente las cotizaciones de esas acciones en diferentes instantes del tiempo. Así, para la compañía* A *se observa aleatoriamente la cotización en 5 instantes de tiempo, en la* B *se observa en 4 instantes, en la* C *en 6 y, por último, en la compañía* D *se observa la cotización de las acciones en 5 instantes de tiempo. En la tabla siguiente se muestra la cotización en euros de las diferentes acciones en los instantes de tiempo seleccionado*

Compañías	***Observaciones***
A	670 840 780 610 900
B	600 800 690 650
C	800 810 730 690 750 720
D	970 840 930 790 920

Suponiendo que se verifican las hipótesis de normalidad, aleatoriedad, independencia y homogeneidad de varianzas, contrastar el nivel de significancia del 1 % si la cotización media de las acciones de cada una de las cuatro compañías se pueden considerar iguales.
Solución:

$$H_0 = \mu_A = \mu_B \mu_C = \mu_D$$
$$H_a = \mu_A \neq \mu_B \neq \mu_C \neq \mu_D$$

$n_1 = 5$ $\quad n_2 = 4$ $\quad n_3 = 6$ $\quad n_4 = 5$
$T_1 = 3800$ $\quad T_2 = 2740$ $\quad T_3 = 4500$ $\quad T_4 = 4450$
$n = \sum n_i = 20$
$k = 4$

$$SFC = \sum \frac{T_i^2}{n_i} - \frac{\left(\sum x_i\right)^2}{n} = \left(\frac{3800^2}{5} + \frac{2740^2}{4} + \frac{4500^2}{6} + \frac{4450^2}{5}\right) - \frac{15490^2}{20}$$
$$= 12100400 - 11997005 = 103395$$
$$SCE = \sum x_i^2 - \frac{T_i^2}{n_i} = 12211500 - 12100400 = 111100$$
$$CMF = \frac{SCF}{k-1} = \frac{103395}{3} = 34465$$
$$CME = \frac{SCE}{n-k} = \frac{111100}{16} = 6943.75$$

1. ***Punto crítico:***

$$F_\alpha = F(v_1 = k-1, v_2 = n-k, \alpha) = F(v_1 = 3, v_2 = 16, \alpha = 0.01) = 5.29$$

2. ***Estadístico de prueba:***

$$F = \frac{CMF}{CME} = \frac{34465}{6943.75} = 4.963456346$$

Como $F < F_\alpha$
Decisión: *Aceptar* H_0
Conclusión: *Hay evidencia suficiente para que cada una de las cuatro compañías se pueden considerar iguales, con un nivel de confiabilidad del* 99 %

3.14.4. ANÁLISIS DE VARIANZA DE UN SOLO FACTOR PARA MUESTRAS ALEATORIAS DEPENDIENTES (CORRELACIONADAS)

Un diseño en bloques aleatorizado es una extensión del diseño de diferencias pareadas. Si queremos comparar tres muestras tratadas, debemos hacer la comparación dentro de conjuntos combinados (o bloques), cada uno formado de tres unidades experimentales.
En general, un diseño en bloques aleatorizado, hecho para comparar **k** tratamientos, utilizará **b** bloques de k unidades experimentales combinadas con una y solamente una unidad experimental para cada muestra (tratamiento). Este diseño se muestra en la siguiente tabla:

		Bloques					
		1	2	3	...	b	
Tratamientos	1	x_{11}	x_{12}	x_{13}	...	x_{1b}	T_1
	2	x_{21}	x_{22}	x_{23}	...	x_{2b}	T_2
	3	x_{31}	x_{32}	x_{33}	...	x_{3b}	T_3
	⋮	⋮	⋮	⋮	⋱	⋮	⋮
	k	x_{k1}	x_{k2}	x_{k3}	...	x_{kb}	T_k
		B_1	B_2	B_3	...	B_b	T

Se dice que se trata de un diseño aleatorizado porque se asignan aleatoriamente los tratamientos a las k unidades experimentales dentro de cada bloque.
Si un diseño en bloques aleatorizado involucra la comparación de k tratamientos dentro de cada uno de "b" bloques, el número total de observaciones en el experimento será $n = bk$.
Supóngase que un supervisor de producción quiere comparar el tiempo medio para que un operador de una línea de montaje realice un trabajo al utilizar uno de tres métodos A, B y C.
Cada uno de $b = 5$ operadores tiene que efectuar el trabajo empleando cada uno de los métodos A, B y C.
El objetivo del diseño de bloques es eliminar la variación en los tiempos de montaje debido a las diferencias de un operador a otro, en habilidad manual, motivacional, etc.
Un ANOVA para un diseño de bloques aleatorizado, divide en tres partes la suma total de cuadrados de desviaciones de todos los valores "x" respecto a la media general.
La primera mide la variación entre las medias de los tratamientos, la segunda ide la variación entre las medias de los bloques y la tercera, la variación de las diferencias entre las observaciones de los tratamientos dentro de los bloques (que mide el error experimental). Es decir:

$$SCT = SCF + SCB + SCE \text{ donde } SCB = \text{sumatoria de cuadrados de bloques}$$

Para realizar un ANOVA para un diseño en bloques aleatorizados, utilizaremos la siguiente notación:

$$
\begin{aligned}
k &= \textit{ número de factores(muestras o fuentes de estimación} \\
b &= \textit{ número de bloques (Asignación del método a seguir} \\
n = bk &= \textit{ número total de observaciones en el experimento} \\
T = \sum x &= \textit{ Total de todas las observaciones en el experimento} \\
\bar{x} = \frac{T}{n} &= \textit{ media de todas las observaciones en el experimento.} \\
T_i &= \textit{ Total de todas las observaciones que reciben el tratamiento i, } i = 1, 2, \ldots, k \\
B_j &= \textit{ Total de todas las observaciones en el bloque j, } j = 1, 2, 3, \ldots, b
\end{aligned}
$$

Las fórmulas necesarias para el análisis son las siguientes:

$$
\begin{aligned}
T &= T_1 + T_2 + \cdots + T_k \\
CM &= \frac{T^2}{n} = \textit{ correción para la media} \\
SC(Total) &= \sum x_i^2 - CM \\
SC(Factor) &= \frac{\sum T_i^2}{b} - CM \\
SC(Bloque) &= \frac{\sum B_j^2}{k} - CM \\
SCE &= SCT - SCF - SCB \\
CMF &= \frac{SCF}{k-1}, \textit{ donde } k - 1 = v(CMF) \\
CMB &= \frac{SCB}{b-1}, \textit{ donde } b - 1 = v(CMB) \\
CME &= \frac{SCE}{n-b-k+1}, \textit{ donde } n - b - k + 1 = v(CME)
\end{aligned}
$$

Una tabla ANOVA para el diseño en bloques aleatorizado, con "k" tratamientos y "b" bloques, se debe estructura de la siguiente forma:

Fuente de Variación	**SC**	**gl**	**CM**	**F**
Factor	SCF	$k-1$	CMF	$\frac{CMF}{CME}$
Bloques	SCB	$b-1$	CMB	$\frac{CMB}{CME}$
Error	SCE	$n-b-k+1$	CME	
Total	SCT	$n-1$		

CRITERIO DE PRUEBA PARA UN DISEÑO EN BLOQUES ALEATORIZADO

PARA COMPARAR MEDIAS DE TRATAMIENTOS

1. Hipótesis Nula H_0: las medias de los factores son iguales.
2. Hipótesis Alternativa
   H_a: por lo menos dos de las medias de los factores difieren.
3. Estadística de prueba $F_T = \frac{CMF}{CME}$
4. Punto crítico $F_\alpha = F(v_1 = k-1, v_2 = n-b-l+1, \alpha)$
5. Rechazar H_0 si $F_T > F_\alpha$, donde F_α se localiza en la cola derecha de la distribución F

PARA COMPARAR MEDIAS DE BLOQUES

1. Hipótesis Nula
   H_0 : las medias de los bloques son iguales.

2. Hipótesis Alternativa.
   H_a : por lo menos dos de las medias de los bloques difieren.
   Estadístico de prueba: $F_B = \frac{CMB}{CME}$

3. Punto crítico $F_\alpha = F(v_1 = b-1, v_2 = n-b-k+1, \alpha$

4. Rechazar H_0 si $F_B > F_\alpha$

Ejemplo 3.14.6

Se emplearon 3 ingemieros experimentados para relizar los análisis de los costos, de la estimación y de las cotizaciones para el trabajo en grandes proyectos de construcción. Se pide a cada uno de los estimadores que se analice, estime y proporcione una cotización para cada uno de los 5 proyectos. Con base en esto, se pueden comparar las cotizaciones de los tres estimadores para un mismo proyecto, eliminando con ello la variación en las cotizaciones de proyecto a proyecto. Los datos de las estimaciones de los tres ingenieros se muestran en la siguiente tabla:

Estimador	*I*	*II*	*III*	*IV*	*V*
1	3.52	4.71	3.89	5.21	4.14
2	3.39	4.79	3.83	4.93	3.96
3	3.64	4.92	4.19	5.10	4.20

- *¿ Existe diferencia entre las medis de los tratamientos?*
- *¿ Existe diferencia entre las medias de los bloques ?*

Utilizar un $\alpha = 0.05$ *para esta prueba*

		Bloques					
		1	2	3	...	b	
Tratamientos	1	x_{11}	x_{12}	x_{13}	...	x_{1b}	T_1
	2	x_{21}	x_{22}	x_{23}	...	x_{2b}	T_2
	3	x_{31}	x_{32}	x_{33}	...	x_{3b}	T_3
	⋮	⋮	⋮	⋮	⋱	⋮	⋮
	k	x_{k1}	x_{k2}	x_{k3}	...	x_{kb}	T_k
		B_1	B_2	B_3	...	B_b	T

Solución:

ANOVA de un factor, muestras aleatorias dependientes. Diseño de bloques aleatorizados.

$B_1 = 10.55$ $B_2 = 14.42$ $B_3 = 11.9$ $B_4 = 15.24$ $B_5 = 12.3$

$T_1 = 21.47$ $T_2 = 20.89$ $T_3 = 22.05$ $T = 64.41$

$b = 5$ $k = 3$ $n = b \cdot k5(3) = 15 datos$ $T = 64.41$

$$\sum_{i=1}^{15} x_i^2 = (3.52)^2 + \cdots + (4.2)^2 = 281.6675$$
$$\sum T_i^2 = (21.47)^2 + (20.89)^2 + (22.05)^2 = 1383.5552$$
$$\sum B_j^2 = (10.55)^2 + \cdots + (12.30)^2 = 844.3965$$
$$CM = \frac{(T)^2}{n} = \frac{(64.41)^2}{15} = 276.5765$$
$$SCT = \sum x_i^2 - CM = 281.6675 - 276.5765 = 5.09096$$
$$SCF = \frac{\sum T_i^2}{b} - CM = \frac{1383.5552}{5} - 276.5765 = 0.13456$$
$$SCB = \sum \frac{B_j^2}{k} - CM = \frac{844.3965}{3} - 276.5765 = 4.88896$$
$$SCE = SCT - SCF - SCB = 5.0909 - .13456 - 4.88896 = 0.06744$$
$$CME = \frac{SCE}{g.l(E)} = \frac{SCE}{n-b-k+1} = \frac{0.06744}{8} = 0.00843$$

Punto crítico
$F_\alpha = F(v_1 = k-1, v_2 = n-b-k+1, \alpha) = F(2, 8, 0.05) = 4.46$
Estadístico de prueba

$$F_T = \frac{CMF}{CME} = \frac{0.06728}{0.00843} = 7.981$$

Como $F_T > F_\alpha$
Decisión:
Rechazar H_0
Conclusión:
Existe evidencia suficiente para indicar una diferencia entre por lo menos dos de las medias de los tratamientos.

PARA BLOQUES

$$H_0 : \mu_{B1} = \mu_{B2} = \mu_{B3} = \mu_{B4} = \mu_{B5}$$
$$H_a : \textit{ Al menos una } \mu_B i \textit{ es distinta (i=1,2,3,4,5)}$$
$$CMB = \frac{SCB}{b-1} = \frac{4.88896}{4} = 1.22224$$

Punto crítico

$$F_\alpha = F(v_1 = b-1, v_2 = n-b-k+a, \alpha) = F(4, 8, 0.05) = 3.84$$

Estadístico de prueba

$$F_B = \frac{CMB}{CME} = \frac{1.22224}{0.00843} = 144.99$$

Como $F_B > F_\alpha$
Decisión:
Rechazar H_0
Conclusión :
Existe evidencia suficiente para indicar una diferencia de por lo menos dos de las medias de los bloques.

Ejemplo 3.14.7 *Un experimento de respuesta a estímulo, que comprende tres tratamientos, se presentó en un diseño de bloques aleatorizado usando cuatro personas. La respuesta fue*

el tiempo hasta la reacción, medido en segundos. ¿ Los datos presentan suficiente evidencia para indicar una diferencia en las respuestas medias para estímulos (tratamientos)? ¿ Y en las personas? Use $\alpha = 0.05$

	Persona 1	***Persona 2***	***persona 3***	***Persona 4***
Tratamiento 1	1.7	1.5	0.1	0.6
Tratamiento 2	3.4	2.6	2.3	2.2
Tratamiento 2	2.3	2.1	0.8	1.6

Solución:
ANOVA, de un factor para muestras pareadas por medio de bloques aleatorizados.

- ***PARA TRATAMIENTOS***
  Para comparar medias de tratamientos

$$H_0 : \mu_{T_1} = \mu_{T_2} = \mu_{T3}$$
$$H_a : \text{ Al menos una } \mu_{Ti} \text{ es distinta (i)1,2,3)}$$

$$b = 4 \qquad k = 3 \qquad n = b \cdot k = 4(3) = 12 datos$$
$$T = 21,2$$

	Persona 1	***Persona 2***	***persona 3***	***Persona 4***	***Total***
Tratamiento 1	1.7	1.5	0.1	0.6	$T_1 = 3.9$
Tratamiento 2	3.4	2.6	2.3	2.2	$T_2 = 10.5$
Tratamiento 2	2.3	2.1	0.8	1.6	$T_3 = 6.8$
Total	$B_1 = 7.4$	$B_2 = 6.2$	$B_3 = 3.2$	$B_4 = 4.4$	$T = 21.2$

$$\sum_{i=1}^{12} x_i^2 = (1.7)^2 + \cdots + (1.6)^2 = 48.86$$
$$\sum T_i^2 = (3.9)^2 + (10.5)^2 + (6.8)^2 = 171.72$$
$$\sum B_j^2 = (7.4)^2 + \cdots + (4.4)^2 = 122.79$$
$$CM = \frac{(T)^2}{n} = \frac{(21.2)^2}{12} = 37.4533$$
$$SCT = \sum x_i^2 - CM = 48.86 - 37,4533 = 11.4067$$
$$SCF = \frac{\sum T_i^2}{b} - CM = \frac{171.72}{4} - 37.4533 = 5.4767$$
$$SCB = \frac{\sum B_j^2}{k} - CM = \frac{122.79}{3} - 37.4533 = 3.4767$$
$$SEC = SCT - SCF - SCB = 11.4067 - 5.4767 - 3.4767 = 2.4533$$
$$CMF = \frac{SCF}{k-1} = \frac{5.4767}{2} = 2.73835$$

Punto crítico
$F_\alpha = F(v_1 = k-1, v_2 = n-b-k+1, \alpha) = F(2, 6, 0.05) = 5.14$
Estadístico de prueba

$$F_T = \frac{CMF}{CME} = \frac{2.73835}{0.4088} = 6.6985$$

Como $F_T > F_\alpha$
Decisión:
Rechazar H_0
Conclusión:
existe evidencia suficiente para indicar una diferencia entre por lo menos dos de las medias de los tratamientos.

- ***PARA BLOQUES***

$$H_0 : \mu_{B1} = \mu_{B2} = \mu_{B3} = \mu_{B4}$$
$$H_a : \text{ Al menos un } \mu_{Bi} \text{ es distinta (i=1,2,3,4)}$$
$$CMB = \frac{SCB}{b-1} = \frac{3.4767}{3} = 1.1589$$

Punto crítico:
$F_\alpha = F(v_1 = b-1, v_2 = n-b-k+1, \alpha) = F(3, 6, 0.05) = 4.39$
Estadístico de prueba:
$F_B = \frac{CMB}{CME} = \frac{1.1589}{0.4088} = 2.8348$ *Como* $F_B < F\alpha$
Decisión:
Aceptar H_0
Conclusión: *Existe evidencia suficiente para indicar que las medias de los bloques son iguales*