Estadística Inferencial (CE29), ciclo 2014-1

PREGRADO

PROFESOR : **Salomón Acosta**

TÍTULO : **Material de enseñanza**

FECHA : Abril 2014

CURSO : **Estadística Inferencial**

CODIGO : MA148 CE29

ÁREA : Ciencias

CICLO : 2014-1

INDICE

INTRODUCCIÓN

CONCEPTOS PRELIMINARES

Introducción.

La Estadística estudia los métodos científicos para recoger, organizar, resumir y analizar datos, así como para sacar conclusiones válidas y tomar decisiones razonables basadas en el análisis. La Estadística es una ciencia que estudia la recolección, análisis e interpretación de datos, ya sea para ayudar en la toma de decisiones o para explicar condiciones regulares o irregulares de algún fenómeno o estudio aplicado, de ocurrencia en forma aleatoria o condicional.

IMPORTANCIA DE LA ESTADISTICA EN LA ADMINISTRACION

La Estadística es de gran importancia en las diferentes empresas, enfocadas desde cualquier área profesional ya que:

- Ayuda a lograr una adecuada planeación y control apoyados en los estudios de pronósticos, presupuestos etc.
- Incrementan la participación de los diferentes niveles de la organización, cuando existe motivación adecuada.
- Obliga a mantener un archivo de datos históricos controlables.
- Facilita a la administración la utilización óptima de los diferentes insumos.
- Facilita la coparticipación e integración de las diferentes áreas de la compañía.
- Obliga a realizar un auto análisis periódico.
- Facilita el control administrativo.
- Ayuda a lograr una mayor efectividad y eficiencia en las operaciones.
- A través de los pronósticos, se pueden prever las perdidas en los resultados de los estados financieros futuros, y de esta manera se pueden tomar decisiones bien sea la reducción de costos y gastos, planear estrategias que ayuden al mejoramiento de la empresa, y que se cumpla con el objetivo de toda empresa que es obtener utilidades
- Por último nos ayuda a tomar decisiones objetivas como:

  ¿Qué clientes les generan los mayores beneficios?
  ¿Qué zonas o regiones son las que generan mayores ventas?
  ¿Cuál es el nivel de satisfacción de sus clientes?
  ¿Cuál es el nivel de rotación o permanencia de clientes?

Las estadísticas son fundamentales tanto para la administración financiera, como para la administración de operaciones, las ventas, el marketing, las cobranzas, la logística y la gestión de personal entre otras áreas y actividades de toda corporación.

Definiciones

- **Población:** Es el conjunto de todos los elementos que se desean analizar y que presentan una o varias características en común. Dependiendo del número de elementos que lo conforman, una población puede ser finita o infinita.

- **Muestra:** Es un subconjunto representativo de elementos provenientes de una población. La muestra es seleccionada de acuerdo a un plan de muestreo, con el fin de que la muestra represente adecuadamente a la población.

- **Unidad Elemental:** Es cada una de las personas, animales u objetos de las que se requiere información. Estos elementos están afectados por las características que se desea estudiar. Constituye la unidad más pequeña de la población y de las muestra.

- **Variable:** Es todo factor o característica que se desea evaluar de las unidades elementales.
  Las variables pueden ser cualitativas (nominal ó jerárquicas) ó cuantitativas (discreta ó continua).

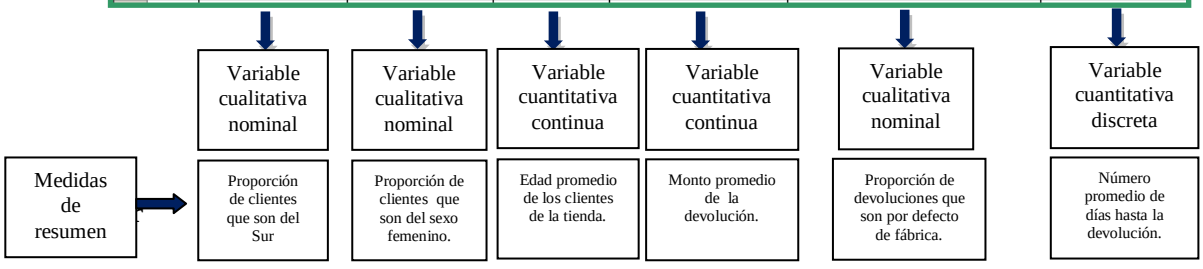
  Por ejemplo: Si nuestra población está conformada por todos los clientes de una gran tienda comercial que realizan cambios ó devoluciones de algún producto, la muestra sería un número determinado de clientes elegidos bajo algún esquema de muestreo. Las variables a estudiar pueden ser las que se muestran parcialmente en la siguiente base de datos:

Microsoft Excel - CASO HIPERMERCADO DATOS

Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ? Adobe PDF

H13

	A	B	C	D	E	F	G
1	No.	UBICACIÓN DE LA TIENDA	GENERO	EDAD	MONTO DE DEVOLUCIÓN O CAMBIO	RAZON DE LA DEVOLUCIÓN O CAMBIO	DIAS HASTA LA DEVOLUCIÓN O CAMBIO
2	1	Zona Norte	MASCULINO	29	69.9	Defecto de Fabrica	1
3	2	Zona Norte	FEMENINO	25	79.9	Cliente equivoco tamño/talla	1
4	3	Zona Norte	MASCULINO	42	9.9	Merc buena, cliente insatisfecho	0
5	4	Zona Sur	FEMENINO	18	23.7	Cliente equivoco tamño/talla	3
6	5	Zona Centro	FEMENINO	25	39.9	Merc buena, cliente insatisfecho	0
7	6	Zona Norte	FEMENINO	45	33.9	Cliente cambio de parecer	2
8	7	Zona Centro	MASCULINO	49	14.9	Cliente equivoco tamño/talla	1
9	8	Zona Centro	FEMENINO	24	49.9	Cliente equivoco tamño/talla	2
10	9	Zona Norte	FEMENINO	25	9.4	Cliente cambio de parecer	0
11	10	Zona Centro	MASCULINO	31	149	Defecto de Fabrica	2
12	11	Zona Centro	MASCULINO	56	29.9	Defecto de Fabrica	1
13	12	Zona Centro	FEMENINO	22	29.9	Cliente equivoco tamño/talla	1
14	13	Zona Sur	MASCULINO	32	29.9	Defecto de Fabrica	2
15	14	Zona Norte	FEMENINO	18	59.9	Cliente cambio de parecer	2
16	15	Zona Sur	FEMENINO	45	29.9	Cliente equivoco tamño/talla	2
17	16	Zona Sur	FEMENINO	32	23.35	Cliente cambio de parecer	1
18	17	Zona Norte	FEMENINO	32	2.9	Otros	2
19	18	Zona Norte	MASCULINO	19	9.4	Cliente cambio de parecer	0
20	19	Zona Sur	MASCULINO	37	39.9	Cliente equivoco tamño/talla	4



- **Pámetro:** Es una medida que resume la información de la(s) característica(s) de interés de la población.

  Características:
  - Es un valor único.
  - Generalmente desconocido.
  - Para hallar su valor se necesita de todos los elementos de la población.

- **Estadígrafo**: Es una medida que resume la información de la(s) característica(s) de interés de la muestra..

  Características:
  - No es un valor único si no variable. Su valor cambia de muestra a muestra.
  - Para hallar su valor se necesita sólo de los elementos de la muestra.
  - También se le conoce como estimador puesto que estima al parámetro poblaiconal.

  Las notaciones utilizadas para un parámetro y su respectivo estimador puntual son las siguientes:

Parámetro	Estimador puntual
μ	$\bar{x}$
σ^2	S^2
p	$\hat{p}$

  Nota: Tanto el parámetro como el estadígrafo son medidas de resúmenes, la diferencia radica que uno usa todos los elementos de la población mientras que el otro los elementos muestrales.

Ramas de Estadística:

Estadística Descriptiva

Es la rama de Estadística que se ocupa de la recolección, clasificación y simplificación de la información. La información recolectada se resume en cuadros (tablas) y gráficos los cuales deben describir en forma apropiada el comportamiento de la información recolectada.

Estadística Inferencial

Es la rama de Estadística que se ocupa de los procesos de estimación (puntual y por intervalos), análisis y pruebas hipótesis. La finalidad de la estadística inferencial es llegar a conclusiones que brinden una adecuada base científica para la toma de decisiones, considerando la información muestral recolectada.

En otras palabras la estadística inferencial se ocupa del análisis, interpretación de los resultados y de las conclusiones a las que se puede llegar a partir de la información obtenida de una muestra con el fin de extender sus resultados a la población bajo estudio. La generalización de las conclusiones obtenidas en una muestra a toda la población esta sujeta a riesgo por cuanto los elementos de la muestra son obtenidos mediante un muestreo probabilístico.

La estadística inferencial provee los procedimientos para efectuar la inferencia inductiva y medir la incertidumbre de las conclusiones que se van a generalizar. Los problemas más importantes en este proceso son:

- **Estimación Puntual:** Es la estimación del valor del parámetro por medio de un único valor obtenido mediante el cálculo o evaluación de un estimador para una muestra específica.

  Por ejemplo: Si se quiere determinar en cuál de las ciudades, Lima o Arequipa, el sueldo semanal promedio de un empleado es mayor

- **Estimación por intervalos:** Es la estimación del valor de un parámetro mediante un conjunto de valores contenidos en un intervalo. Para la obtención de intervalos de confianza se debe considerar el coeficiente de confianza que es la probabilidad de que el intervalo contenga al parámetro poblacional.

- **Prueba de Hipótesis:** Es el procedimiento estadístico de comprobación de una afirmación y se realiza a través de las observaciones de una muestra aleatoria.

El objetivo de la inferencia estadística es hacer inferencias acerca de una población basada en la información contenida en una muestra. Ahora considerando que las poblaciones están caracterizadas por medidas descriptivas numéricas llamadas parámetros., a la inferencia estadística le corresponde hacer inferencias acerca de los parámetros poblacionales.

A continuación muestra la notación de dos parámetros con sus respectivos estimadores puntuales.

Parámetro	Estimador puntual	
$\mu_1 - \mu_2$	$\overline{X}_1 - \overline{X}_2$	$\overline{X}_1 - \overline{X}_2$ estima puntualmente a $\mu_1 - \mu_2$
σ_1^2 / σ_2^2	S_1^2 / S_2^2	S_1^2 / S_2^2 estima puntualmente a σ_1^2 / σ_2^2
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\hat{p}_1 - \hat{p}_2$ estima puntualmente a $p_1 - p_2$

CAPÍTULO I

PRUEBA DE HIPÓTESIS PARA UN PARÁMETRO

- PRUEBA DE HIPÓTESIS PARA UN PROMEDIO
- PRUEBA DE HIPÓTESIS PARA UNA PROPORCIÓN

La planificación de una investigación estadística usualmente tiene por propósito verificar si los supuestos que se tienen sobre la población en estudio se pueden aceptar como válidos o deben ser considerados falsos.

Esta sección tiene como finalidad presentar los conceptos y aplicaciones de las principales pruebas de hipótesis.

1.1 Conceptos generales

Hipótesis estadística: Es cualquier afirmación o conjetura que se hace acerca de la distribución de una o más poblaciones. Por ejemplo: la longitud media de un tipo de objeto es de 20 centímetros, es decir, $\mu = 20$; afirmar que el porcentaje de objetos defectuosos producidos por cierto proceso sea menor al 4%, es decir, $p < 0,04$.

Hipótesis nula (H_o): A partir de la información proporcionada por la muestra se verificará la suposición sobre el parámetro estudiado. La hipótesis que se contrasta se llama hipótesis nula

Hipótesis alterna (H_1): Es la hipótesis que debe ser aceptada si se rechaza la hipótesis nula. Es la conclusión a la que se llegaría si hubiera sufuciente evidencia en la información de la muestra para decidir que es improbable que la hipótesis nula sea verdadera. El hecho de no rechazar la hipótesis nula no implica que ésta sea cierta, significa simplemente que los datos de la muestra son insuficientes para inducir un rechazo de la hipótesis nula.

Tipos de errores: Cuando usamos los datos de una muestra para tomar decisiones acerca de un parámetro existe el riesgo de llegar a una conclusión incorrecta. De hecho se pueden presentar dos tipos diferentes de error cuando se aplica la metodología de la prueba de hipótesis.

		Decisión estadística	
		No rechazar H_0	**Rechazar H_0**
Situación	**H_0 verdadera**	Confianza $(1 - \alpha)$	**Error tipo I** (α)
	H_0 falsa	**Error tipo II** (β)	Potencia $(1 - \beta)$

Error Tipo I: Ocurre cuando se rechaza una hipótesis H_0 que es verdadera.
La probabilidad de cometer un error de tipo I es denotada por α.

Nivel de significancia (α): Es la probabilidad de cometer el error tipo I. El valor α es fijado por la persona que realiza la investigación (por lo general 5%).

Error Tipo II: Ocurre cuando no se rechaza una hipótesis H_0 que es falsa.
La probabilidad de cometer un error de tipo II es denotada por β.

Poder de prueba (1 - β): Es la probabilidad de rechazar una hipótesis nula que es falsa.

Ejemplo:
Un investigador cree haber descubierto una vacuna contra el SIDA. Para verificar su hallazgo hará una investigación de laboratorio. De acuerdo con el resultado, se decidirá lanzar o no la vacuna al mercado. La hipótesis nula que propone es: "La vacuna no es efectiva"

a) Según el enunciado propuesto, redacte en qué consiste el error de tipo I y tipo II.

b) ¿Cuál sería el error más grave de cometer? Sustente su respuesta.

Pasos a seguir en una Prueba de Hipótesis:

- ✓ Establecer las hipótesis nula y alterna.
- ✓ Determinar el nivel de significación.
- ✓ Elegir el estadístico apropiado de prueba a utilizar.
- ✓ Especificar los supuestos necesarios para la validez de la prueba.
- ✓ Establecer los valores críticos que separan la región de rechazo y no rechazo.
- ✓ Recolectar los datos y calcular el valor del estadístico de prueba apropiado.
- ✓ Tomar la decisión estadística y exprezar la conclusión en términos del problema.

Supuestos para las pruebas de hipótesis:

Para las diferentes pruebas de hipótesis se deben cumplir los siguientes supuestos:

- Para pruebas de hipótesis para una media poblacional (μ)
  - La muestra es aleatoria.
  - La muestra proviene de una distribución normal o el tamaño de muestra es grande.

- Prueba de hipótesis para una proporción (π).
  - La muestra es aleatoria.
  - El tamaño de muestra es grande.

- Para pruebas de hipótesis para la diferencia de medias poblacionales $(\mu_1 - \mu_2)$ y razón de variancias poblacionales $\left(\frac{\sigma_2^2}{\sigma_1^2}\right)$
  - Las muestras son aleatorias.
  - Las muestras provienen de distribuciones normales.
  - Las poblaciones son independientes

- Prueba de hipótesis para la diferencia de proporciones $(\pi_1 - \pi_2)$.
  - Las muestras son aleatorias.
  - Los tamaños de muestras son grandes.
  - Las poblaciones son independientes

- Prueba de hipótesis para datos pareados ó muestras relacionadas
  - La muestra es aleatoria.
  - La diferencia de las primeras observaciones con respecto a las segundas observaciones (o viceversa) provienen de una distribución normal.

1.2 Prueba de hipótesis para una media poblacional (μ)

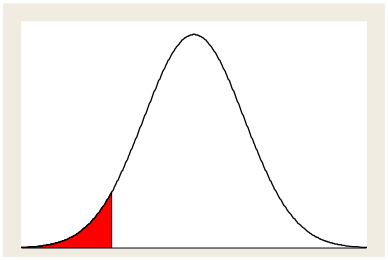
Cuando la muestra proviene de una población Normal y la varianza poblacional (σ^2) es desconocida

Hipótesis:

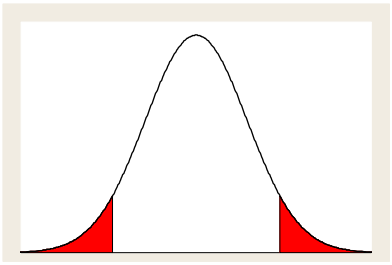
$H_0 : \mu \geq \mu_0$	$H_0 : \mu = \mu_0$	$H_0 : \mu \leq \mu_0$
$H_1 : \mu < \mu_0$	$H_1 : \mu \neq \mu_0$	$H_1 : \mu > \mu_0$

Estadístico de prueba: $$T_c = \frac{\overline{X} - \mu_0}{S / \sqrt{n}}$$

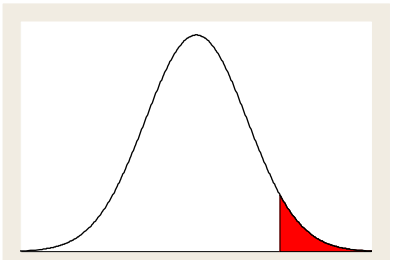
Región de Rechazo: Está representada por la zona sombreada



Prueba unilateral de extremo izquierdo



Prueba bilateral



Prueba unilateral de extremo derecho

Se rechaza Ho si el valor calculado del estadístico de prueba cae en la zona de rechazo

Sobre el estadístico de prueba:

$\overline{X}$: Es la media muestral.

μ_0 : Es el valor supuesto de la media poblacional en la hipótesis nula.

S : Es la desviación estándar de la muestra.

n: Es el tamaño de la muestra.

$t_{(n-1)}$: Denota la distribución t de Student con $n - 1$ **grados de libertad**.

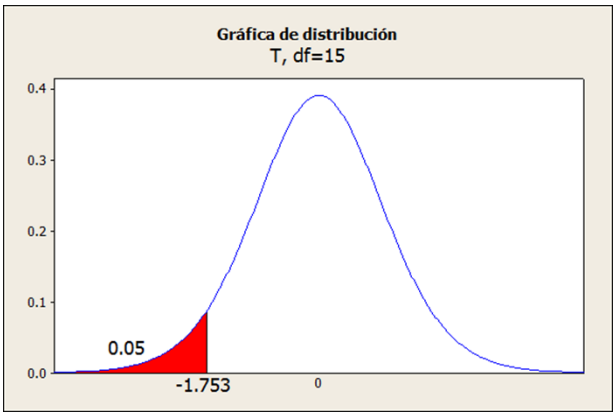
α es el nivel de significación de la prueba

Si la **población** es **finita de tamaño *N***, entonces se debe agregar el factor de corrección para población finita, con lo cual se obtiene:

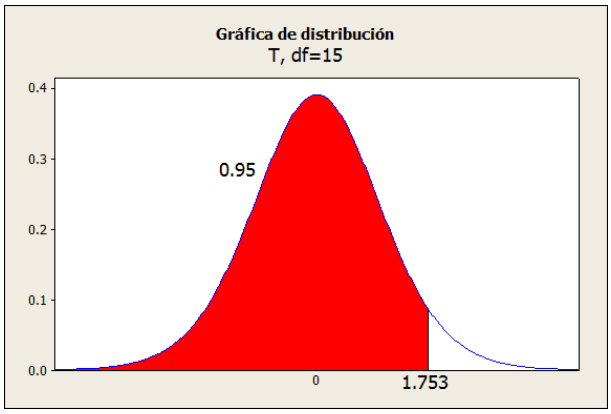
$$T_c = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$$

El **VALOR CRÍTICO** divide la gráfica en zona de rechazo y no rechazo. Para hallar su valor en EXCEL, usaremos la siguiente función:

INV.T(área a la izquierda, grados de libertad)



INV.T(0.05,15)



INV.T(0.95,15)

Ejemplo:

Star América es una línea aérea de capital compartido (peruano-americano) que tiene más de 10 años laborando en el Perú. El gerente de marketing de aerolíneas Star América desea realizar un estudio considerando como segmentos de interés a los pasajeros nacionales y extranjeros. Para realizar dicho estudio se seleccionan al azar muestras aleatorias e independientes de los registros de pasajeros peruanos y extranjeros. Algunas de las características que desea analizar el gerente son las que se muestran en la siguiente tabla:

- ✓ Origen del pasajero: peruano o extranjero.
- ✓ Género: masculino o femenino.
- ✓ Opinión sobre el servicio de la aerolínea en el último viaje: Pésima, Mala, Regular, Buena o Muy Buena.
- ✓ Edad del pasajero (en años)
- ✓ Peso del equipaje en el último viaje (en kg).

Origen	Género	Opinión	Edad	Peso
Extranjero	Mujer	Regular	17	18.1
Extranjero	Hombre	Regular	62	17.9
Extranjero	Hombre	Regular	50	21.2
Extranjero	Mujer	Regular	48	19.1
Extranjero	Mujer	Regular	39	19.7
Extranjero	Hombre	Mala	44	21.3
Extranjero	Mujer	Regular	40	19.3
Extranjero	Mujer	Mala	37	18.8
Extranjero	Mujer	Muy buena	25	17.8
Extranjero	Hombre	Muy buena	7	16.3
Extranjero	Hombre	Regular	7	22.5

Peruano	Mujer	Mala	29	24
Peruano	Hombre	Buena	56	16.2
Peruano	Hombre	Muy buena	44	19.4
Peruano	Hombre	Buena	7	20.6
Peruano	Hombre	Regular	51	22.2
Peruano	Hombre	Mala	41	18
Peruano	Hombre	Regular	46	20.6
Peruano	Hombre	Buena	41	19
Peruano	Mujer	Regular	30	18
Peruano	Hombre	Buena	45	23.5
Peruano	Mujer	Regular	46	21.7
Peruano	Hombre	Regular	22	17.2
Peruano	Mujer	Muy buena	8	20.7
Peruano	Hombre	Regular	64	19.4
Peruano	Mujer	Mala	16	17.9
Peruano	Hombre	Muy buena	41	16.4
Peruano	Mujer	Buena	43	21.3
Peruano	Hombre	Buena	12	22.5

Con la información presentada y usando un nivel de significación del 7% responda lo siguiente:

¿El peso promedio de los equipajes es menor de 21 Kgr?

PASO 1: HIPÓTESIS

PASO 2: NIVEL DE SIGNIFICANCIA

PASO 3: ESTADÍSTICO DE PRUEBA

PASO 4: REGIONES CRÍTICAS Y CRITERIO DE DECISIÓN

PASO 5: VALOR CALCULADO DEL ESTADÍSTICO DE PRUEBA

PASO 6: DECISIÓN
PASO 7: CONCLUSIÓN

Ejemplo:
Una empresa que embotella yogurt cuenta con una máquina programada para llenar botellas de 1180 ml. Sin embargo, debido a variación natural y desgaste, el volumen medio por botella puede cambiar en cualquier momento, razón por la cual se implementa el siguiente sistema de control: Seleccionar una muestra de 20 botellas, obtener de dicha información el volumen medio y la desviación estándar, luego, parar la producción y revisar la máquina si se encuentra evidencia en la muestra de que el volumen medio de llenado es inferior a 998 ml. Con los datos que se muestran a continuación, y con un nivel de significación de 2%, ¿cuál será su decisión? Asuma que el contenido de las botellas se distribuye normalmente.

1074.27	938.74	979.68	938.74	986.9	966.59	1010.9	934.64	1096.88	1160.43
953.17	1040.01	940.42	931.83	998.72	981.65	1038.48	1109.49	897.59	1009.8

Solución:

Siendo X: Volumen de llenado
Procesando la información con Excel:

Media	999.4465
Error típico	15.44193147
Mediana	984.275
Moda	938.74
Desviación estándar	69.05841694
Varianza de la muestra	4769.06495
Cuenta	20

Hipótesis:

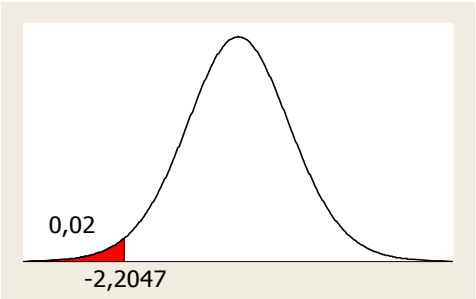
H_0: $\mu \geq 998$
H_1: $\mu < 998$

Nivel de significación:

$\alpha = 0.02$

Estadístico de prueba: $T_c = \dfrac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$

Regiones de aceptación y rechazo para Ho:



Se rechazará Ho si: $T_c < -2.2047$

Reemplazando datos, obtenemos el valor calculado del estadístico de prueba:

$$T_C = \frac{999.4465 - 998}{69.0584 / \sqrt{20}} = 0.0937$$ (este valor se ubica en la zona de no rechazo)

Decisión: No se rechaza Ho

Conclusión: No existe suficiente evidencia estadística, con un nivel de significación del 2%, para afirmar que el volumen medio de llenado es inferior a 998 ml. Con este resultado, no se procederá a parar la máquina para revisión.

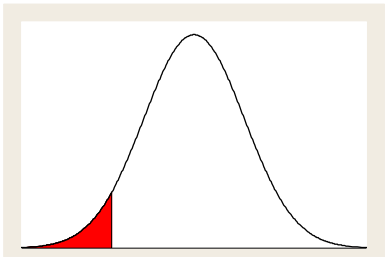
1.3 Pruebas de hipótesis para una proporción poblacional (p)

Hipótesis:

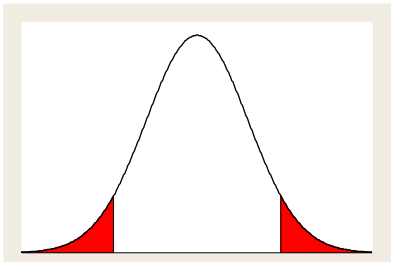
$H_0 : p \geq p_0$	$H_0 : p = p_0$	$H_0 : p \leq p_0$
$H_1 : p < p_0$	$H_1 : p \neq p_0$	$H_1 : p > p_0$

Estadístico de prueba: $$Z_c = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

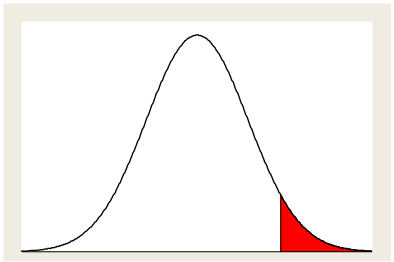
Región de Rechazo: Está representada por la zona sombreada



Prueba Unilateral de extremo izquierdo



Prueba Bilateral



Prueba Unilateral de extremo derecho

Se rechaza Ho si el valor calculado del estadístico de prueba cae en la zona de rechazo

Sobre el estadístico de prueba:

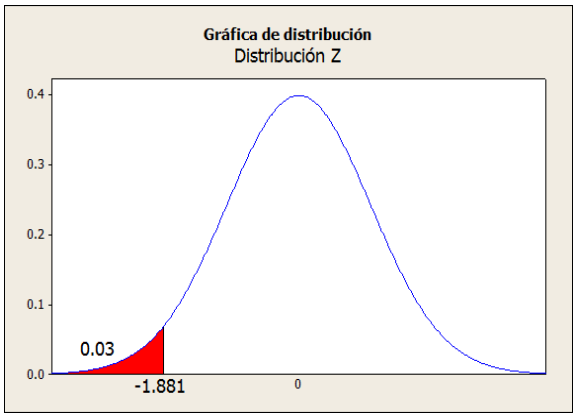
$\hat{P}$: Es la proporción muestral.
p_0 : Es el valor supuesto de la proporción poblacional en la hipótesis nula.
n: Es el tamaño de la muestra.
Z denota la distribución normal estándar.
α es el nivel de significación de la prueba.

Si la **población** es **finita de tamaño *N*** se debe agregar el factor de corrección para poblaciones finitas, con lo cual la fórmula se transforma en:

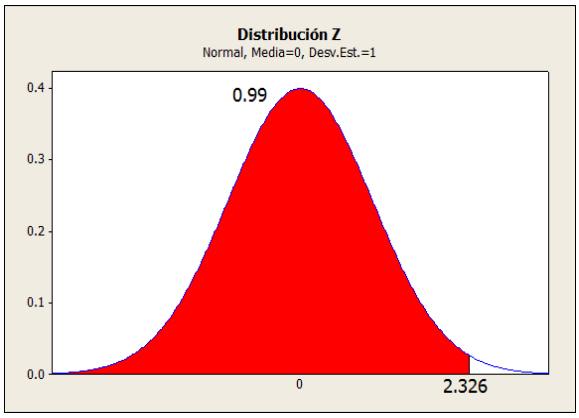
$$Z_c = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}\frac{N-n}{N-1}}}$$

El **VALOR CRÍTICO** divide la gráfica en zona de rechazo y no rechazo. Para hallar su valor en EXCEL, usaremos la siguiente función:

INV.NORM.ESTAND(área a la izquierda)



INV.NORM.ESTAND(0.03)



INV.NORM.ESTAND(0.99)

Ejemplo:

Star América es una línea aérea de capital compartido (peruano-americano) que tiene más de 10 años laborando en el Perú. El gerente de marketing de aerolíneas Star América desea realizar un estudio considerando como segmentos de interés a los pasajeros nacionales y extranjeros. Para realizar dicho estudio se seleccionan al azar muestras aleatorias e independientes de los registros de pasajeros peruanos y extranjeros. Algunas de las características que desea analizar el gerente son las que se muestran en la siguiente tabla:

- ✓ Origen del pasajero: peruano o extranjero.
- ✓ Género: masculino o femenino.
- ✓ Opinión sobre el servicio de la aerolínea en el último viaje: Pésima, Mala, Regular, Buena o Muy Buena.
- ✓ Edad del pasajero (en años)
- ✓ Peso del equipaje en el último viaje (en kg).

Origen	Género	Opinión	Edad	Peso
Extranjero	Mujer	Regular	17	18.1
Extranjero	Hombre	Regular	62	17.9
Extranjero	Hombre	Regular	50	21.2
Extranjero	Mujer	Regular	48	19.1
Extranjero	Mujer	Regular	39	19.7
Extranjero	Hombre	Mala	44	21.3
Extranjero	Mujer	Regular	40	19.3
Extranjero	Mujer	Mala	37	18.8
Extranjero	Mujer	Muy buena	25	17.8
Extranjero	Hombre	Muy buena	7	16.3
Extranjero	Hombre	Regular	7	22.5
Peruano	Mujer	Mala	29	24
Peruano	Hombre	Buena	56	16.2

Peruano	Hombre	Muy buena	44	19.4
Peruano	Hombre	Buena	7	20.6
Peruano	Hombre	Regular	51	22.2
Peruano	Hombre	Mala	41	18
Peruano	Hombre	Regular	46	20.6
Peruano	Hombre	Buena	41	19
Peruano	Mujer	Regular	30	18
Peruano	Hombre	Buena	45	23.5
Peruano	Mujer	Regular	46	21.7
Peruano	Hombre	Regular	22	17.2
Peruano	Mujer	Muy buena	8	20.7
Peruano	Hombre	Regular	64	19.4
Peruano	Mujer	Mala	16	17.9
Peruano	Hombre	Muy buena	41	16.4
Peruano	Mujer	Buena	43	21.3
Peruano	Hombre	Buena	12	22.5

Con la información presentada y usando un nivel de significación del 2% responda lo siguiente:
¿La proporción de pasajeros que consideran el servicio muy bueno, es inferior al 27%?

PASO 1: HIPÓTESIS

PASO 2: NIVEL DE SIGNIFICANCIA
PASO 3: ESTADÍSTICO DE PRUEBA

PASO 4: REGIONES CRÍTICAS Y CRITERIO DE DECISIÓN

PASO 5: VALOR CALCULADO DEL ESTADÍSTICO DE PRUEBA

PASO 6: DECISIÓN
PASO 7: CONCLUSIÓN

Ejercicios Propuestos.

TEMA: Prueba de Hipótesis para un parámetro.

1. Star América es una línea aérea de capital compartido (peruano-americano) que tiene más de 10 años laborando en el Perú.
   El gerente de marketing de aerolíneas Star América desea realizar un estudio considerando como segmentos de interés a los pasajeros nacionales y extranjeros.
   Para realizar dicho estudio se seleccionan al azar muestras aleatorias e independientes de los registros de pasajeros peruanos y extranjeros. Algunas de las características que desea analizar el gerente son las que se muestran en la siguiente tabla:

   - Origen del pasajero: peruano o extranjero.
   - Género: masculino o femenino.
   - Opinión sobre el servicio de la aerolínea en el último viaje: Pésima, Mala, Regular, Buena o Muy Buena.
   - Edad del pasajero (en años)
   - Peso del equipaje en el último viaje (en kg).

Origen	Genero	Opinion	Edad	Peso
extranjero	mujer	regular	17	18.1
extranjero	hombre	regular	62	17.9
extranjero	hombre	regular	50	21.2
extranjero	mujer	regular	48	19.1
extranjero	mujer	regular	39	19.7
extranjero	hombre	mala	44	21.3
extranjero	mujer	regular	40	19.3
extranjero	mujer	mala	37	18.8
extranjero	mujer	muy buena	25	17.8
extranjero	hombre	muy buena	7	16.3
extranjero	hombre	regular	7	22.5
peruano	mujer	mala	29	24.0
peruano	hombre	buena	56	16.2
peruano	hombre	muy buena	44	19.4
peruano	hombre	buena	7	20.6
peruano	hombre	regular	51	22.2
peruano	hombre	mala	41	18.0
peruano	hombre	regular	46	20.6
peruano	hombre	buena	41	19.0
peruano	mujer	regular	30	18.0
peruano	hombre	buena	45	23.5
peruano	mujer	regular	46	21.7
peruano	hombre	regular	22	17.2
peruano	mujer	muy buena	8	20.7
peruano	hombre	regular	64	19.4
peruano	mujer	mala	16	17.9
peruano	hombre	muy buena	41	16.4
peruano	mujer	buena	43	21.3
peruano	hombre	buena	12	22.5

Con la información presentada y usando un nivel de significación del 5% responda lo siguiente:

a) ¿La edad promedio del pasajero extranjero es superior a 32 años?
b) ¿La proporción de equipajes que pesan menos de 17 kg , excede al 12%?

2. La directora de mercadotecnia de A&B Cola está preocupada porque el producto no atrae a suficientes consumidores jóvenes. Para probar su hipótesis, encuesta aleatoriamente a 100 consumidores de A&B Cola. Se obtuvo como resultado una media de 35 años con una desviación estándar de 10 años.

   Al nivel de significación del 5%, ¿estos hechos son suficientes para concluir que los consumidores de A&B Cola posen una edad promedio mayor a 32 años?

   Respt: Prueba unilateral derecha, Tcal = 3.00, Tcrit = 1.6604, Decisión: RHo

3. El fabricante de la motocicleta Ososki anuncia en una propaganda de televisión que su vehículo posee un rendimiento promedio de 87 millas por galón en viajes largos. Los millajes (recorrido en millas) observados en ocho viajes prolongados fueron: 88, 82, 81, 87, 80, 78, 79, y 89. Al nivel de significación del 5% ¿el millaje medio es menor que el anunciado?

   Respt: Prueba unilateral izquierda, Tcal = -2.61, Tcrit = -1.895, Decisión: RHo

4. En una encuesta aleatoria de 1000 hogares realizada en Lima, se encontró que 90 de los hogares tenían al menos un miembro de familia con educación superior. ¿Este resultado refuta la aseveración de que en los hogares de Lima esta proporción es al menos 12%? Use un nivel de significación del 5%.

   Respt: Prueba unilateral izquierda, Zcal = -2.919, Zcrit = -1.6449, Decisión: RHo

5. Se realizó una investigación de mercadotecnia para estimar la proporción de amas de casa que pueden reconocer la marca de un producto de limpieza con base a la forma y color del recipiente. De las 1400 amas de casa, 420 fueron capaces de identificar la marca del producto. ¿Se puede afirmar, a un nivel de significación del 5%, que la proporción de amas de casa que reconocen la marca del producto, es superior al 25%?

   Respt: Prueba unilateral derecha, Zcal = 4.32, Zcrit = 1.64485, Decisión: RHo

CAPÍTULO II

PRUEBA DE HIPÓTESIS DE DOS PARÁMETROS

- PRUEBA DE HIPÓTESIS PARA DOS VARIANZAS
- PRUEBA DE HIPÓTESIS PARA DOS PROMEDIOS
- PRUEBA DE HIPÓTESIS PARA DOS PROPORCIONES

2.1 Pruebas de hipótesis para dos varianzas poblacionales

Para esta prueba de hipótesis solo desarrollaremos el caso bilateral debido a que esta prueba indicará si dos muestras independientes provienen de poblaciones con varianzas homogéneas o heterogéneas lo que será necesario saber al realizar prueba de hipótesis para dos promedios.

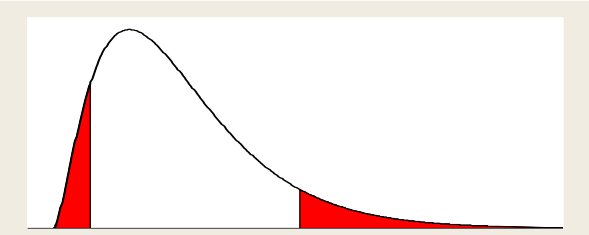
Hipótesis:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Prueba bilateral

Estadístico de prueba: $F_{\text{calculado}} = \dfrac{S_1^2}{S_2^2}$

Región de Rechazo: Está representada por la zona sombreada



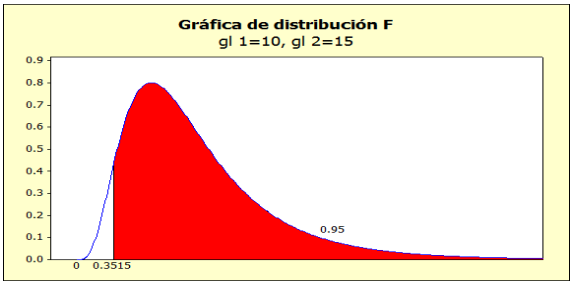
Se rechaza H_0 si el valor calculado del estadístico de prueba cae en la zona de rechazo

Sobre el estadístico de prueba:

n_1: Es el tamaño de la muestra proveniente de la población 1.
n_2: Es el tamaño de la muestra proveniente de la población 2.
S_1^2 : Es la varianza de la muestra de la población 1.
S_2^2 : Es la varianza de la muestra de la población 2.
$F_{(n_1-1,\, n_2-1)}$: Es la distribución F con n_1-1 y n_2-1 grados de libertad.
α es el nivel de significación de la prueba.

El **VALOR CRÍTICO** divide la gráfica en zona de rechazo y no rechazo. Para hallar su valor en EXCEL, usaremos la siguiente función:

INV.F.CD(área a la derecha, grados de libertad 1, grados de libertad 2)



Ejemplo:

Se está realizando un estudio comparativo sobre tiempo de atención en dos restaurantes. Se han registrado los tiempos que demora en ser atendidos algunos pedidos, los cuales se muestran:

A (1)	6,15	5,63	5,58	6,91	4,63	5,53	5,05	5,45	5,03	6,09	
B (2)	4,96	5,04	4,75	4,61	4,47	5,02	5,35	3,6	5,26	5,41	5,42

¿Se puede afirmar que los tiempos de atención en ambos restaurants no tienen la misma variabilidad? Use un nivel de significación del 6%.

PASO 1: HIPÓTESIS

PASO 2: NIVEL DE SIGNIFICANCIA
PASO 3: ESTADÍSTICO DE PRUEBA

PASO 4: REGIONES CRÍTICAS Y CRITERIO DE DECISIÓN

PASO 5: VALOR CALCULADO DEL ESTADÍSTICO DE PRUEBA

PASO 6: DECISIÓN

PASO 7: CONCLUSIÓN

Ejemplo:

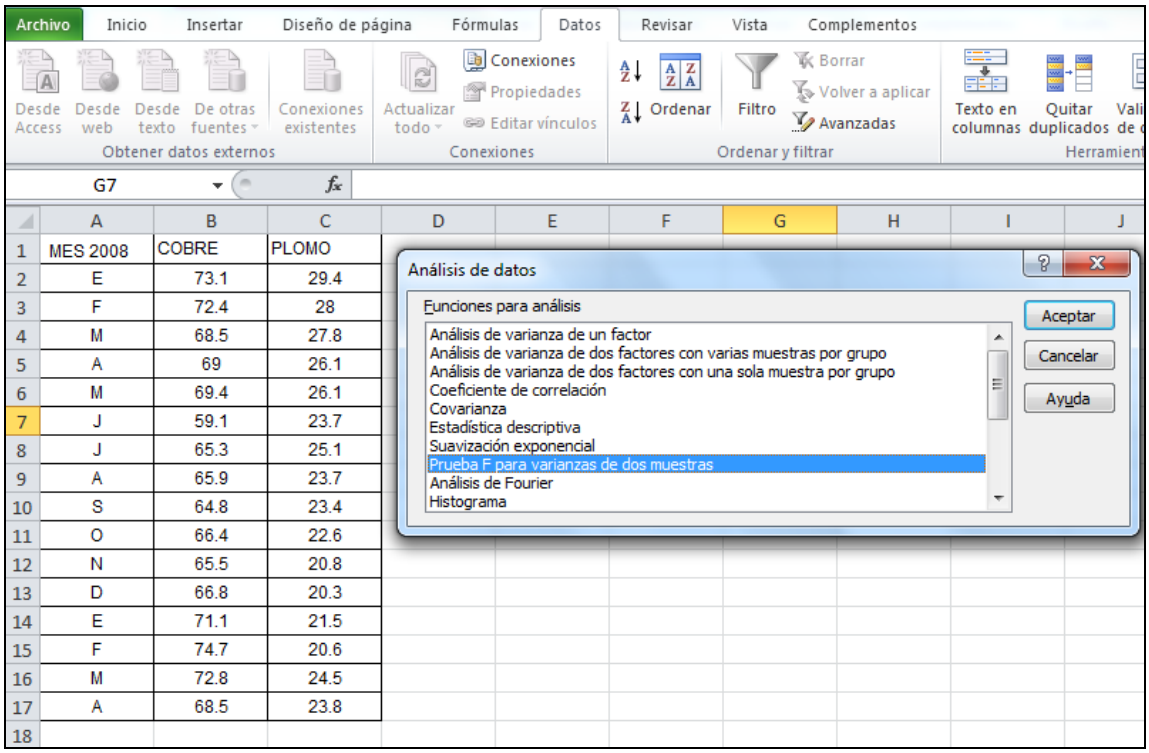
Un empresario minero desea saber si existen diferencias respecto a las variaciones de las cotizaciones observadas de plomo y cobre para los años 2010 y 2012. Use un nivel de significación del 8%.
A continuación se presenta la tabla de cotizaciones de los años indicados:

MES 2010	E	F	M	A	M	J	J	A	S	O	N	D
COBRE	73.1	72.4	68.5	69	69.4	59.1	65.3	65.9	64.8	66.4	65.5	66.8
PLOMO	29.4	28	27.8	26.1	26.1	23.7	25.1	23.7	23.4	22.6	20.8	20.3

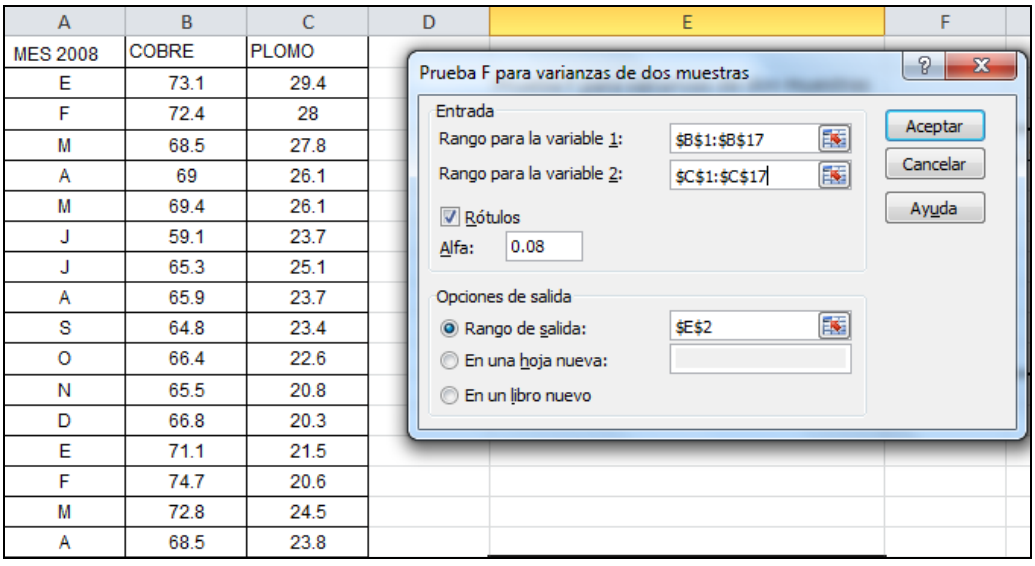
MES 2012	E	F	M	A
COBRE	71.1	74.7	72.8	68.5
PLOMO	21.5	20.6	24.5	23.8

Solución:

Procesando la información con Excel:



	A	B	C
1	MES 2008	COBRE	PLOMO
2	E	73.1	29.4
3	F	72.4	28
4	M	68.5	27.8
5	A	69	26.1
6	M	69.4	26.1
7	J	59.1	23.7
8	J	65.3	25.1
9	A	65.9	23.7
10	S	64.8	23.4
11	O	66.4	22.6
12	N	65.5	20.8
13	D	66.8	20.3
14	E	71.1	21.5
15	F	74.7	20.6
16	M	72.8	24.5
17	A	68.5	23.8
18			



Resultados:

Prueba F para varianzas de dos muestras

	COBRE	PLOMO
Media	68.33125	24.2125
Varianza	15.7342917	7.5625
Observaciones	16	16
Grados de libertad	15	15
F	2.08056749	
P(F<=f) una cola	0.08372709	
Valor crítico para F (una cola)	2.10856159	

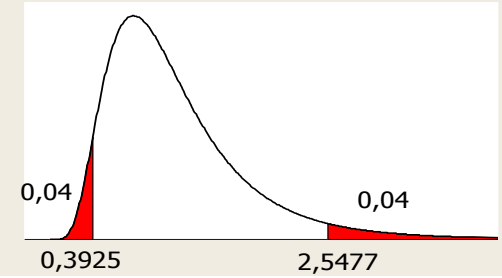
Hipótesis:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Nivel de significación: $\alpha = 0.08$

Estadístico de prueba: $F_{\text{calculado}} = \dfrac{S_1^2}{S_2^2}$

Valor calculado del estadístico de prueba: $F_{\text{calculado}} = \dfrac{15.7343}{7.5625} = 2.0806$



Valor crítico inferior:

INV.F.CD(0.96; 15, 15) = 0.3925

Valor crítico superior:

INV.F.CD(0.04; 15, 15) = 2.5477

El valor calculado cae en la zona de no rechazo

Decisión: No se Rechaza Ho

Conclusión: No existe suficiente evidencia estadística, con un nivel de significación del 5%, para concluir que las varianzas son diferentes. Por lo tanto, se puede afirmar que las varianzas son homogéneas

Ejemplo:

Una empresa de bebidas energizantes posee dos tipos de bebidas en el mercado: Energy Aid y Energy Pro. El ingeniero de control de calidad desea evaluar el contenido de refresco en los dos tipos de energizantes, para el análisis se seleccionó 17 latas de refresco Energy Aid que posee una media de 17.2 onzas, con una desviación estándar de 3.2 onzas, y trece refrescos

Energy Pro de donde se obtuvo una media de 18.1 onzas y una desviación estándar de 2.7 onzas. Asumiendo que el contenido de refrescos se distribuye normalmente, ¿se puede afirmar con 6% de significación que las varianzas de los contenidos son iguales?

Solución:

Sean X_1: Contenido de una lata de refresco Energy Aid (onzas), $X_1 \sim N(\mu_1, \sigma_1^2)$

X_2: Contenido de una lata de refresco Energy Pro (onzas), $X_2 \sim N(\mu_2, \sigma_2^2)$

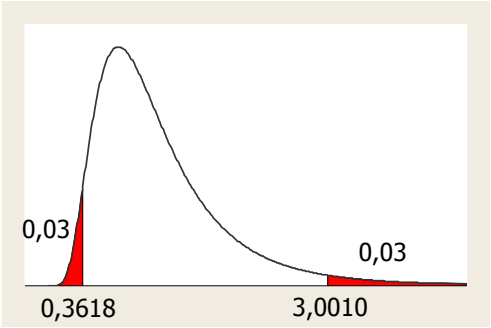
Hipótesis:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Nivel de significación: $\alpha = 0.06$

Estadístico de prueba: $F_{\text{calculado}} = \dfrac{S_1^2}{S_2^2}$

Valor calculado del estadístico de prueba: $F_{\text{calculado}} = \dfrac{3.2^2}{2.7^2} = 1.4066$



Valor crítico inferior:

INV.F.CD(0.97; 16, 12) = 0.3925

Valor crítico superior:

INV.F.CD(0.04; 16, 12) = 2.5477

Decisión: No se Rechaza Ho

Conclusión: No existe suficiente evidencia estadística, con un nivel de significación del 5%, para afirmar que las varianzas son diferentes. Es decir, existe homogeneidad de varianzas.

2.2 Pruebas de hipótesis para la diferencia de dos medias poblacionales (μ_1-μ_2): muestras independientes

Cuando las muestras provienen de poblaciones Normales, las varianzas poblacionales σ_1^2 , σ_2^2 son desconocidas y además:

Caso 1: Varianzas Iguales ($\sigma_1^2 = \sigma_2^2$)

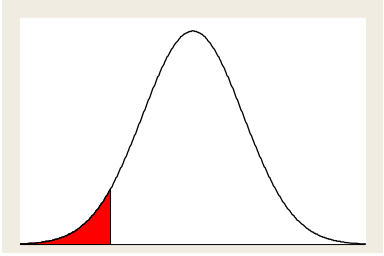
Hipótesis:

$H_0 : \mu_1 \geq \mu_2$ $\quad$ $H_0 : \mu_1 = \mu_2$ $\quad$ $H_0 : \mu_1 \leq \mu_2$

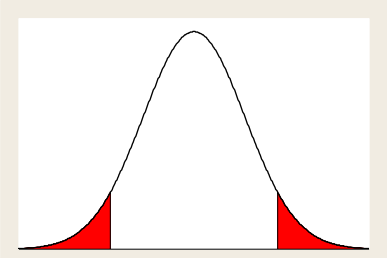
$H_1 : \mu_1 < \mu_2$ $\quad$ $H_1 : \mu_1 \neq \mu_2$ $\quad$ $H_1 : \mu_1 > \mu_2$

Estadístico de prueba: $$T_c = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

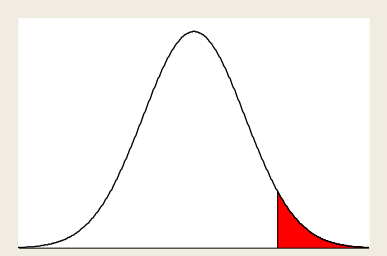
Región de Rechazo: está representada por la zona sombreada



Prueba Unilateral de extremo izquierdo



Prueba Bilateral



Prueba Unilateral de extremo derecho

Se Rechaza Ho si el valor calculado del estadístico de prueba cae en la zona de rechazo

Sobre el estadístico de prueba: $$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}$$

$\overline{X}_1$, $\overline{X}_2$: media de la muestra 1 y 2 respectivamente

S_1^2 , S_2^2 : varianza de la muestra 1 y 2 respectivamente

S_p^2 : Es la varianza muestral ponderada.

n_1: Es el tamaño de la muestra 1.

n_2: Es el tamaño de la muestra 2.

$t_{(n_1+n_2-2)}$: Es el valor de la distribución *t* de Student con $n_1 + n_2 - 2$ g de l.

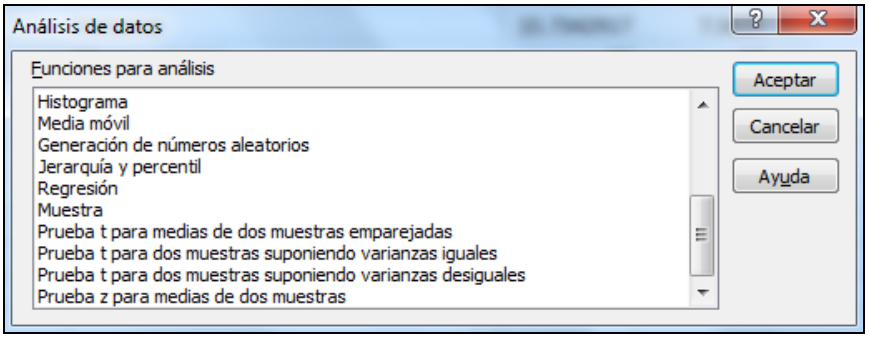
α es el nivel de significación de la prueba.

- **NOTA**: Si la hipótesis nula propone alguna diferencia específica entre los promedios poblacionales sometidos a prueba, y denotamos esta diferencia por k, entonces el estadístico de prueba será:

$$T_c = \frac{\overline{X}_1 - \overline{X}_2 - k}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

El **VALOR CRÍTICO** y el **VALOR CALCULADO** del estadístico de prueba los hallaremos usando EXCEL con la siguiente función:

DATOS, ANÁLSIS DE DATOS; Prueba t para dos muestras suponiendo varianzas iguales



Ejemplo:

Se está realizando un estudio comparativo sobre tiempo de atención en dos restaurantes. Se han registrado los tiempos que demora en ser atendidos algunos pedidos, los cuales se muestran:

A (1)	6,15	5,63	5,58	6,91	4,63	5,53	5,05	5,45	5,03	6,09	
B (2)	4,96	5,04	4,75	4,61	4,47	5,02	5,35	3,6	5,26	5,41	5,42

¿Se puede afirmar que el restaurante A se demora, en promedio, más en atender que el restaurante B? Use un nivel de significación del 6%.

PASO 1: HIPÓTESIS

PASO 2: NIVEL DE SIGNIFICANCIA

PASO 3: ESTADÍSTICO DE PRUEBA

PASO 4: REGIONES CRÍTICAS Y CRITERIO DE DECISIÓN

PASO 5: VALOR CALCULADO DEL ESTADÍSTICO DE PRUEBA

PASO 6: DECISIÓN

PASO 7: CONCLUSIÓN

Ejemplo:

Un grupo de empresarios inauguró el año pasado dos restaurantes en las zonas más representativas de Lima. Después de un año de actividades deciden medir y comparar, el nivel de ingresos de ambos locales para lo cual eligen muestras aleatorias de los ingresos mensuales. La información se presenta en la siguiente tabla:

LOCAL 1	315	263	258	391	163	253	205	245	203	309	
LOCAL 2	196	204	175	161	147	202	235	60	226	241	242

Se puede afirmar que el local 1 tiene ingresos promedio mayores que los del local 2. Asuma que el consumo mensual tiene distribución normal. Use un nivel de significación del 6%.

Solución:

Sean X_1: Ingreso mensual del local 1
X_2: Ingreso mensual del local 2

Dado que las varianzas poblacionales son desconocidas, el primer paso consiste en realizar una prueba de hipótesis para determinar si las varianzas son homogéneas o no.
En Excel: Datos, Análisis de datos, Prueba F para varianzas de dos muestras
Resultados:

Prueba F para varianzas de dos muestras

	LOCAL 1	LOCAL 2
Media	260.5	189.909091
Varianza	4283.83333	2881.69091
Observaciones	10	11
Grados de libertad	9	10
F	1.48656933	
P(F<=f) una cola	0.27224846	
Valor crítico para F (una cola)	2.83576412	

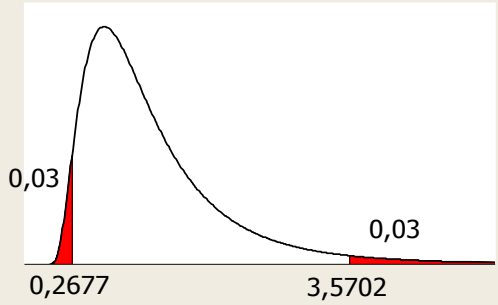
Hipótesis:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Nivel de significación: $\alpha = 0.06$

Estadístico de prueba: $F_{\text{calculado}} = \dfrac{S_1^2}{S_2^2}$

Reemplazando datos: $F_{\text{calculado}} = \dfrac{4283.8333}{2881.6909} = 1.4866$



Valor crítico inferior:
INV.F.CD(0.97; 9, 10) = 0.2677

Valor crítico superior:
INV.F.CD(0.03; 9, 10) = 3.5702

Decisión: No se Rechaza Ho

Conclusión: No existe suficiente evidencia estadística, con un nivel de significación del 6%, para afirmar que las varianzas son diferentes. Con este resultado afirmamos que existe homogeneidad de varianzas.

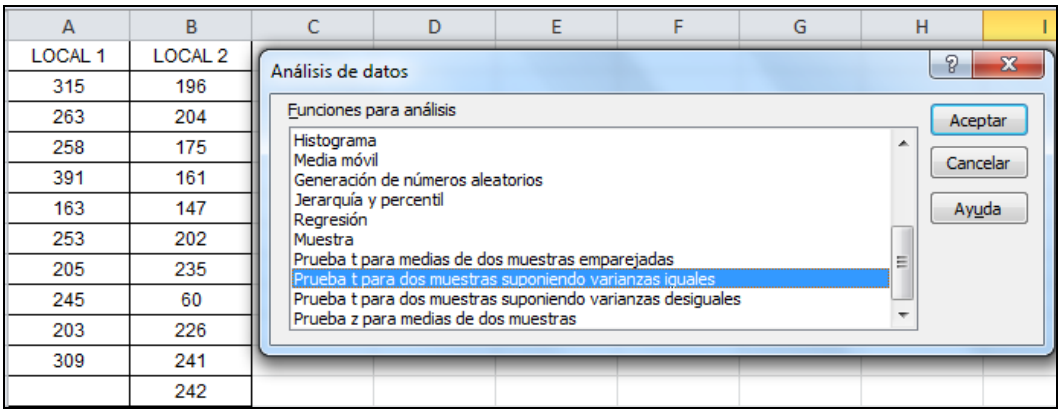
Habiendo probado que las varianzas son homogéneas, ahora pasamos a probar si el local 1 tiene ingresos promedio mayores que los del local 2

Hipótesis:

$$H_0 : \mu_1 \leq \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

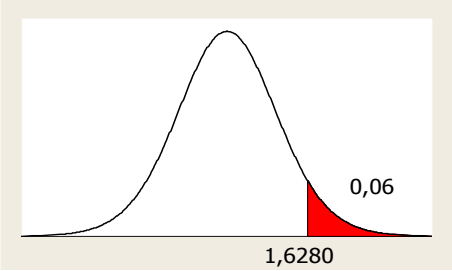
Nivel de significación: $\alpha = 0.06$

Para hallar el valor calculado del estadístico de prueba y el punto crítico usaremos las funciones del Excel:
Herramientas, Análisis de datos, Prueba t para dos muestras suponiendo varianzas iguales:



Resultados:

Prueba t para dos muestras suponiendo varianzas iguales		
	LOCAL 1	LOCAL 2
Media	260.5	189.909091
Varianza	4283.83333	2881.69091
Observaciones	10	11
Varianza agrupada	3545.86364	
Diferencia hipotética de las medias	0	
Grados de libertad	19	
Estadístico t	2.71315406	
P(T<=t) una cola	0.00689602	
Valor crítico de t (una cola)	1.62797232	
P(T<=t) dos colas	0.01379204	
Valor crítico de t (dos colas)	2.00001747	



Se rechazará Ho si: $T_{calculado} > 1.6280$

El valor calculado del estadístico de prueba, $T_{calculado}$ = 2.7132, cae en la región de rechazo.

Decisión: Se Rechaza Ho

Conclusión: Existe suficiente evidencia estadística, con un nivel de significación del 6%, para afirmar que el local 1 tiene en promedio, mayores ingresos mensuales que el local 2.

NOTAS:

Caso 2: Varianzas Diferentes ($\sigma_1^2 \neq \sigma_2^2$)

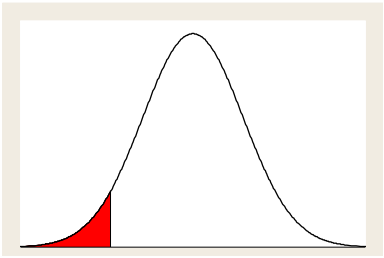
Hipótesis:

$H_0 : \mu_1 \geq \mu_2$ $\quad$ $H_0 : \mu_1 = \mu_2$ $\quad$ $H_0 : \mu_1 \leq \mu_2$

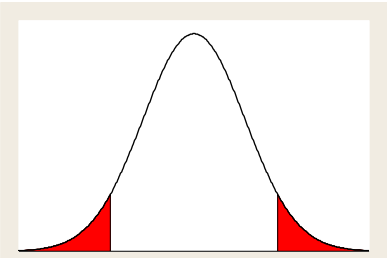
$H_1 : \mu_1 < \mu_2$ $\quad$ $H_1 : \mu_1 \neq \mu_2$ $\quad$ $H_1 : \mu_1 > \mu_2$

Estadístico de prueba: $T_c = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$

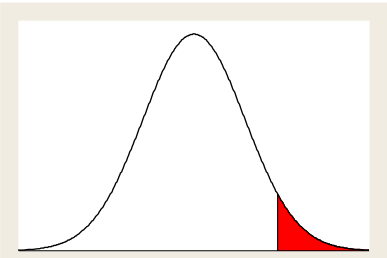
Regió de Rechazo: Representada por la zona sombreada



Prueba Unilateral de extremo izquierdo



Prueba Bilateral



Prueba Unilateral de extremo derecho

Se rechaza Ho si el valor calculado del estadístico de prueba cae en la zona de rechazo

Sobre el estadístico de prueba, $v = \dfrac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{S_1^2}{n_1}\right)^2}{(n_1 - 1)} + \dfrac{\left(\dfrac{S_2^2}{n_2}\right)^2}{(n_2 - 1)}}$

$\overline{X}_1$: Es la media de la muestra 1, $\overline{X}_2$: Es la media de la muestra 2.

S_1^2: Es la varianza de la muestra 1, S_2^2: Es la varianza de la muestra 2.

n_1: Es el tamaño de la muestra 1, n_2: Es el tamaño de la muestra 2.

$t_{(v)}$: Es la distribución *t* de Student con $\boldsymbol{v}$ grados de libertad.

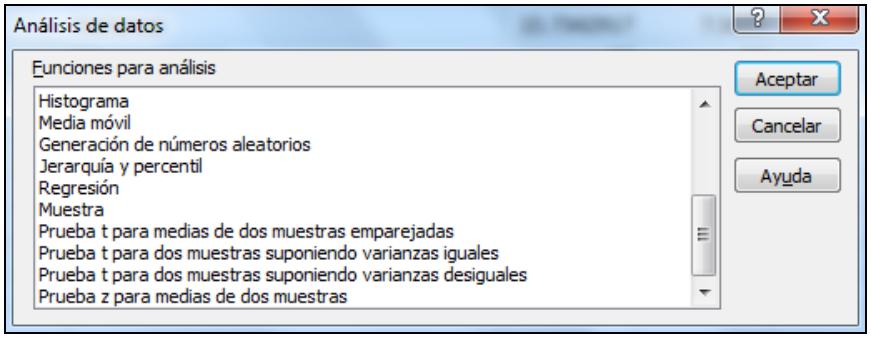
α es el nivel de significación de la prueba.

- **NOTA**: Si la hipótesis nula propone alguna diferencia específica entre los promedios poblacionales sometidos a prueba, y denotamos esta diferencia por k, entonces el estadístico de prueba será:

$$T_c = \frac{\overline{X}_1 - \overline{X}_2 \; - \; k}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$$

El **VALOR CRÍTICO** y el **VALOR CALCULADO** del estadístico de prueba los hallaremos usando EXCEL con la siguiente función:

DATOS, ANÁLSIS DE DATOS; Prueba t para dos muestras suponiendo varianzas desiguales



Ejemplo:

Una empresa fabrica polos deportivos y compra los hilos de dos proveedores (Proveedor 1 y 2). Para verificar la conveniencia de comprar a uno de ellos, compara la resistencia promedio de los hilos adquiridos de estos proveedores. Se toma muestras de piezas de cada clase de hilo y se registra la resistencia en condiciones similares. Los datos en kilogramos, se muestran en la siguiente tabla.

Usando un nivel de significación del 4% y asumiendo heterogeneidad en las varianzas, ¿se puede decidir por el proveedor 2?

Proveedor 1	Proveedor 2
59	84
75	83
82	86
74	79
64	83
58	87
69	86
70	85

PASO 1: HIPÓTESIS

PASO 2: NIVEL DE SIGNIFICANCIA

PASO 3: ESTADÍSTICO DE PRUEBA

PASO 4: REGIONES CRÍTICAS Y CRITERIO DE DECISIÓN

PASO 5: VALOR CALCULADO DEL ESTADÍSTICO DE PRUEBA

PASO 6: DECISIÓN

PASO 7: CONCLUSIÓN

Ejemplo:

Una empresa fabrica, en sus dos plantas situadas en Atlanta y Dallas, impresoras y faxes. Con el fin de medir los conocimientos que tienen los empleados de estas plantas acerca de la calidad de los productos producidos, se toma una muestra aleatoria de empleados de cada fábrica y se les aplica una evaluación de calidad. Los resultados se muestran en el siguiente cuadro. ¿Se puede afirmar que la puntuación promedio obtenida en el examen de calidad no es la misma para las dos fábricas? Use α=0.05

Atlanta	78,0	75,0	80,0	76,0	74,0	82,0	80,0	76,0	74,0	
Dallas	91,0	95,0	73,0	74,0	73,0	82,0	73,0	74,0	73,0	76,0

Solución:

Sean X_1: puntaje obtenido por los trabajadores en la primera planta.

X_2: puntaje obtenido por los trabajadores en la segunda planta.

Dado que las varianzas poblacionales son desconocidas, el primer paso consiste en realizar una prueba de hipótesis para determinar si las varianzas son homogéneas o no: Resultados hallados con Excel:

Atlanta		**Dallas**	
Media	77,2222222	Media	78,4
Desviación estándar	2,90593263	Desviación estándar	8,22192192
Varianza de la muestra	8,44444444	Varianza de la muestra	67,6
Curtosis	-1,24720518	Curtosis	0,69896971
Cuenta	9	Cuenta	10

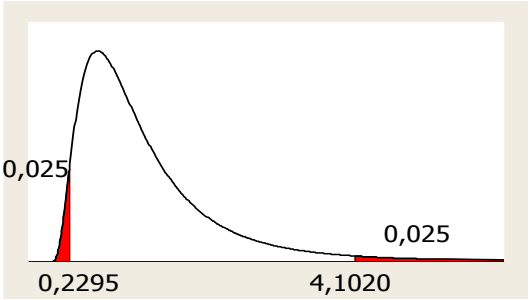
Hipótesis:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Nivel de significación: $\alpha = 0.05$

Estadístico de prueba: $F_{\text{calculado}} = \dfrac{S_1^2}{S_2^2}$

Reemplazando datos: $F_{\text{calculado}} = \dfrac{8.4444}{67.6} = 0.1249$



Valor crítico inferior:
INV.F.CD(0.975; 8, 9) = 0.2295

Valor crítico superior:
INV.F.CD(0.025; 8, 9) = 4.1020

Decisión: Se Rechaza Ho

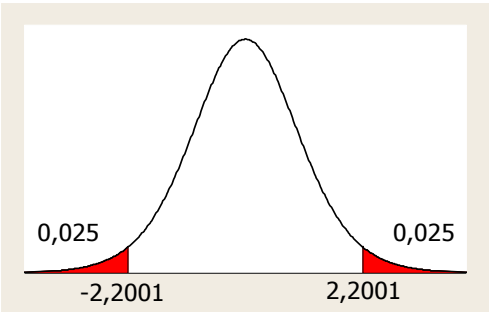
Conclusión: Existe suficiente evidencia estadística, con un nivel de significación del 5%, para afirmar que las varianzas son heterogéneas.

Habiendo probado que las varianzas no son iguales, ahora pasamos a probar si la puntuación promedio es la misma:

Hipótesis:

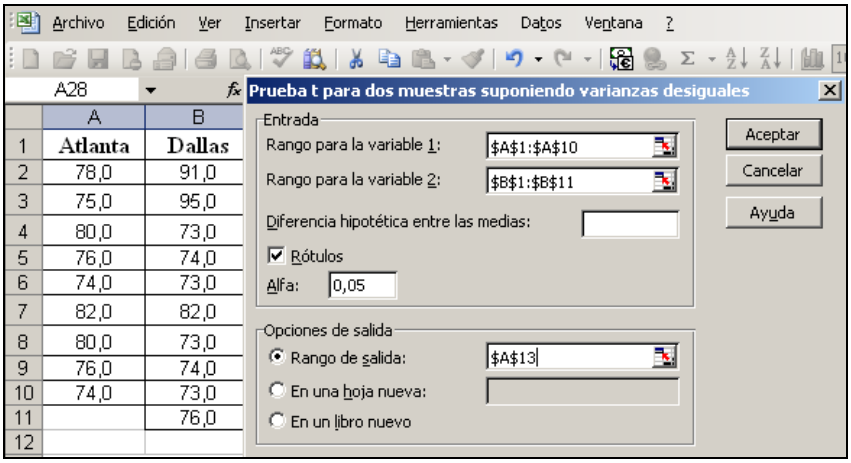
$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Nivel de significación: $\alpha = 0.05$



Se rechazará Ho si: $T_{\text{calculado}} > -2.2001$ *ó* $T_{\text{calculado}} < 2.2001$

Para hallar el "Estadístico de prueba" usaremos las funciones del Excel: Herramientas, Análisis de datos, Prueba t para dos muestras suponiendo varianzas desiguales:



Prueba t para dos muestras suponiendo varianzas desiguales

	Atlanta	*Dallas*
Media	77,22222222	78,4
Varianza	8,444444444	67,6
Observaciones	9	10
Diferencia hipotética de las medias	0	
Grados de libertad	11	
Estadístico t	-0,424489294	
P(T<=t) una cola	0,339696247	
Valor crítico de t (una cola)	1,795884814	
P(T<=t) dos colas	0,679392494	
Valor crítico de t (dos colas)	2,200985159	

El valor calculado del estadístico de prueba es $T_{\text{calculado}}$ = -0.4245 cae en la región de no rechazo.

Decisión: No se Rechaza Ho

Conclusión: No existe suficiente evidencia estadística, con un nivel de significación del 5%, para afirmar que los promedios son diferentes. Es decir, el puntaje promedio es el mismo.

NOTAS:

NOTAS:

Ejercicio

Una empresa grande de corretaje de acciones desea determinar qué tanto éxito han tenido sus nuevos ejecutivos de cuenta en la consecución de clientes. Después de haber terminado su entrenamiento, los nuevos ejecutivos pasan varias semanas haciendo llamadas a posibles clientes, tratando de conseguir prospectos para abrir cuentas con las empresas. Los datos siguientes dan el número de cuentas nuevas que fueron abiertas durante las primeras dos semanas por diez ejecutivas y ocho ejecutivos de cuenta escogidos aleatoriamente.

Ejecutivas	12	11	14	13	13	14	13	12	14	12
Ejecutivos	13	10	11	12	13	12	10	12		

A un nivel del 5%, ¿Parece que las mujeres son más efectivas que los hombres para conseguir nuevas cuentas?

PRUEBA DE HIPÓTESIS PARA DETERMINAR SI SON VARIANZAS HOMOGENEAS O HETEROGÉNEAS

PASO 1: HIPÓTESIS

PASO 2: NIVEL DE SIGNIFICANCIA

PASO 3: ESTADÍSTICO DE PRUEBA

PASO 4: REGIONES CRÍTICAS Y CRITERIO DE DECISIÓN

PASO 5: VALOR CALCULADO DEL ESTADÍSTICO DE PRUEBA

PASO 6: DECISIÓN

PASO 7: CONCLUSIÓN

PRUEBA DE HIPÓTESIS DE LA DIFERENCIA DE MEDIAS

PASO 1: HIPÓTESIS

PASO 2: NIVEL DE SIGNIFICANCIA

PASO 3: ESTADÍSTICO DE PRUEBA

PASO 4: REGIONES CRÍTICAS Y CRITERIO DE DECISIÓN

PASO 5: VALOR CALCULADO DEL ESTADÍSTICO DE PRUEBA

PASO 6: DESICIÓN

PASO 7: CONCLUSIÓN

2.3 Pruebas de hipótesis para la diferencia de dos medias poblacionales (D): muestras relacionadas

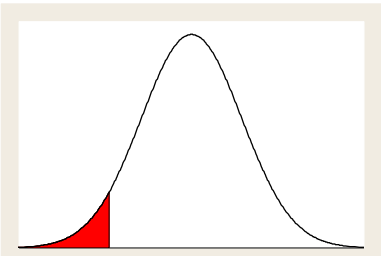
Considere dos poblaciones relacionadas con medias y variancias desconocidas desde las cuales se extrae una muestra aleatoria bivariada de tamaño n (X_1, Y_1), (X_2, Y_2), …, (X_n, Y_n). Defina la variable $D_i = X_i - Y_i$. Entonces esta prueba se reduce a la prueba para una media considerando a la variable D.

Hipótesis:

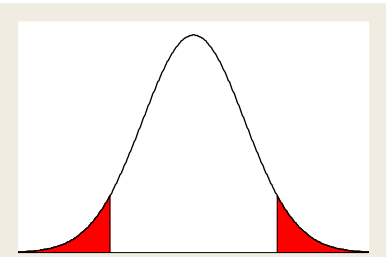
$H_0: \mu_d \geq 0$	$H_0: \mu_d = 0$	$H_0: \mu_d \leq 0$
$H_1: \mu_d < 0$	$H_1: \mu_d \neq 0$	$H_1: \mu_d > 0$

Estadístico de prueba: $T_c = \dfrac{\bar{d}}{S_d / \sqrt{n}}$

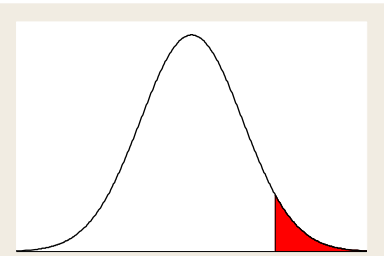
Región de Rechazo: Representada por la zona sombreada



Prueba Unilateral de extremo izquierdo



Prueba Bilateral



Prueba Unilateral de extremo derecho

Se Rechaza Ho si el valor calculado del estadístico de prueba cae en la zona de echazo

Sobre es estadístico de prueba:

$\bar{D}$: Es la media muestral de las diferencias.
S_D : Es la desviación estándar muestral de las diferencias.
n: Es el tamaño de la muestra.
$t_{(n-1)}$: Es la distribución t de Student con $n - 1$ grados de libertad.
α es el nivel de significación de la prueba.

- **NOTA**: Si la hipótesis nula propone alguna diferencia específica entre las proporciones poblacionales sometidas a prueba, y denotamos esta diferencia por k, entonces el estadístico de prueba será:

$$T_c = \frac{\bar{d} - k}{S_d / \sqrt{n}}$$

Ejemplo:

El gerente de un gimnasio afirma que un nuevo programa de ejercicio reducirá la medida de la cintura de una persona en un período de cinco días. Las medidas de cinturas de seis hombres que participaron en este programa de ejercicios se registraron antes y después del período de cinco días en la siguiente tabla:

	Hombres					
	1	**2**	**3**	**4**	**5**	**6**
Medida de cintura antes	90,4	95,5	98,7	115,9	104,0	85,6
Medida de cintura después	91,7	93,9	97,4	112,8	101,3	84,0

¿La afirmación del gimnasio es válida al nivel de significación de 5%? Suponga que la distribución de las diferencias de medidas de cintura antes y después del programa es aproximadamente normal.

Solución:

Sea X_1: Medida de cintura antes (cm.), X_2: Medida de cintura después (cm.)

d = antes – después, Procesando la información con Excel:

Prueba t para medias de dos muestras emparejadas

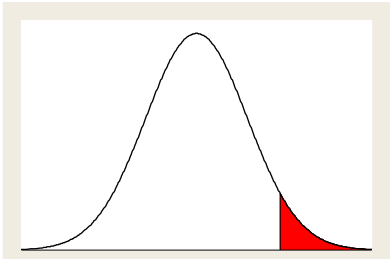
	Medida antes	*Medida después*
Media	98.35	96.85
Varianza	114.787	94.971
Observaciones	6	6
Coeficiente de correlación de Pearson	0.993095074	
Diferencia hipotética de las medias	0	
Grados de libertad	5	
Estadístico t	2.381652558	
P(T<=t) una cola	0.031517895	
Valor crítico de t (una cola)	2.015048373	
P(T<=t) dos colas	0.063035791	
Valor crítico de t (dos colas)	2.570581836	

Hipótesis:

H_0: $\mu_d \leq 0$

H_1: $\mu_d > 0$

Nivel de significación: 0.05



2.0150

Valor calculado del estadístico de prueba: $T_c = \dfrac{\bar{d}}{S_d / \sqrt{n}} = 2.3817$ (cae en la zona de rechazo)

Decisión: Se Rechaza Ho

Conclusión: Existe suficiente evidencia estadística, con un nivel de significación del 5%, para afirmar es cierto lo que afirma el gerente del gimnasio.

2.4 Prueba de hipótesis para la diferencia de dos proporciones poblacionales (p_1-p_2).

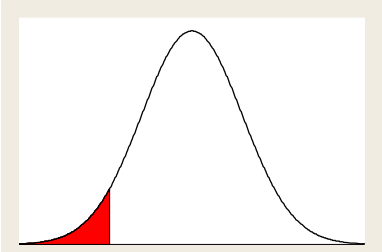
Hipótesis:

$H_0 : p_1 \geq p_2$ $\quad$ $H_0 : p_1 = p_2$ $\quad$ $H_0 : p_1 \leq p_2$

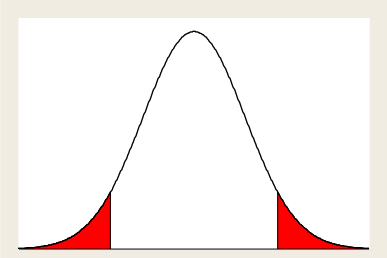
$H_1 : p_1 < p_2$ $\quad$ $H_1 : p_1 \neq p_2$ $\quad$ $H_1 : p_1 > p_2$

Estadístico de prueba: $$Z_C = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\overline{P}(1-\overline{P})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

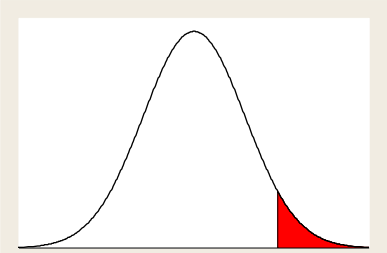
Zona de rechazo: Representada por la zona sombreada



Prueba Unilateral de extremo izquierdo



Prueba Bilateral



Prueba Unilateral de extremo derecho

Se rechaza Ho si el valor calculado del estadístico de prueba cae en la zona de rechazo.

Sobre es estadístico de prueba, $\overline{P} = \frac{n_1\hat{P}_1 + n_2\hat{P}_2}{n_1 + n_2}$, además:

$\hat{P}_1$: Es la proporción de la muestra 1.

$\hat{P}_2$: Es la proporción de la muestra 2.

n_1: Es el tamaño de la muestra 1.

n_2: Es el tamaño de la muestra 2.

- **NOTA**: Si la hipótesis nula propone alguna diferencia específica entre las proporciones poblacionales sometidas a prueba, y denotamos esta diferencia por k, entonces el estadístico de prueba será:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - K}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}}$$

Ejemplo:

Un patrocinador de un programa especial de televisión afirma que el programa representa un atractivo mayor para los televidentes hombres que para las mujeres. Si una muestra aleatoria de 300 hombres y otra de 400 mujeres reveló que 120 hombres y 120 mujeres estaban viendo el programa especial de televisión. Al nivel de significación del 5%, ¿se podría decir que el patrocinador tiene la razón?

PASO 1: HIPÓTESIS

PASO 2: NIVEL DE SIGNIFICANCIA

PASO 3: ESTADÍSTICO DE PRUEBA

PASO 4: REGIONES CRÍTICAS Y CRITERIO DE DECISIÓN

PASO 5: VALOR CALCULADO DEL ESTADÍSTICO DE PRUEBA

PASO 6: DESICIÓN

PASO 7: CONCLUSIÓN

Ejemplo:

En una prueba de preferencia de dos comerciales de televisión se pasó cada uno en un área de prueba seis veces, durante un período de una semana. La semana siguiente se llevó a cabo una encuesta telefónica para identificar a quiénes habían visto esos comerciales. A las personas que los vieron se les pidió definieran el principal mensaje en ellos. Se obtuvieron los siguientes resultados:

Comercial	Personas que lo vieron	Personas que recordaron el mensaje principal
A	150	63
B	200	60

Use $\alpha = 0.06$ para probar la hipótesis de que no hay diferencia en las proporciones que recuerdan los dos comerciales.

Solución:

Sean

p_1: Proporción de personas que recordaron el mensaje principal del comercial A.

p_2: Proporción de personas que recordaron el mensaje principal del comercial B.

Hipótesis:

H_0: $P_1 = P_2$

H_1: $P_1 \neq P_2$

Nivel de significación: $\alpha = 0.06$

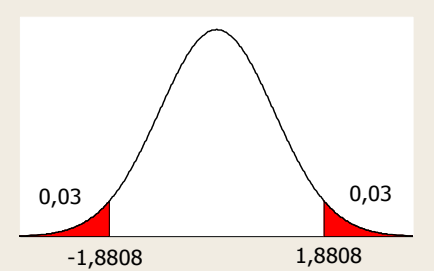
Estadístico de prueba:

$$Z_C = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\overline{P}(1-\overline{P})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Reemplazando datos:

$$\hat{p}_1 = \frac{63}{150} = 0.42, \quad \hat{p}_2 = \frac{60}{200} = 0.30, \quad \overline{P} = 0.3514$$

$$Z_C = \frac{0.42 - 0.30}{\sqrt{0.3514 * (1-0.3514) * \left(\frac{1}{150}+\frac{1}{200}\right)}} = 2.3271$$



El valor calculado del estadístico de prueba cae en la zona de rechazo

Decisión: Se Rechaza Ho

Conclusión: Existe suficiente evidencia estadística, con un nivel de significación del 5%, para afirmar que las proporciones de recordación son diferentes.

Ejercicio

Una empresa realiza un estudio para determinar si el ausentismo de los trabajadores en el turno de día es diferente al de los trabajadores en el turno nocturno. Se realiza una comparación de 100 trabajadores de cada turno. Los resultados muestran que 27 trabajadores diurnos han faltado por lo menos cinco veces durante el año anterior, mientras que 49 trabajadores nocturnos han faltado por lo menos cinco veces.

Con un nivel de significación del 2%, ¿existen diferencias significativas entre las proporciones de trabajadores de los turnos que faltaron cinco veces o más al año?

Respt: Prueba bilateral, Zcal = - 3.2051, Zcrit = ± 2.3263, Decisión: Rho

PASO 1: HIPÓTESIS

PASO 2: NIVEL DE SIGNIFICANCIA

PASO 3: PRUEBA ESTADÍSTICA

PASO 4: REGIONES CRÍTICAS Y CRITERIO DE DECISIÓN

PASO 5: VALOR CALCULADO DEL ESTADÍSTICO DE PRUEBA

PASO 6: DESICIÓN

PASO 7: CONCLUSIÓN

Ejercicios Propuestos.

1. Star América es una línea aérea de capital compartido (peruano-americano) que tiene más de 10 años laborando en el Perú.
   El gerente de marketing de aerolíneas Star América desea realizar un estudio considerando como segmentos de interés a los pasajeros nacionales y extranjeros.
   Para realizar dicho estudio se seleccionan al azar muestras aleatorias e independientes de los registros de pasajeros peruanos y extranjeros. Algunas de las características que desea analizar el gerente son las siguientes:

   - Origen del pasajero: peruano o extranjero.
   - Género: masculino o femenino.
   - Opinión sobre el servicio de la aerolínea en el último viaje: Pésima, Mala, Regular, Buena o Muy Buena.
   - Edad del pasajero (en años)
   - Peso del equipaje en el último viaje (en kg).

Origen	Genero	Opinion	Edad	Peso
extranjero	mujer	regular	17	18.1
extranjero	hombre	regular	62	17.9
extranjero	hombre	regular	50	21.2
extranjero	mujer	regular	48	19.1
extranjero	mujer	regular	39	19.7
extranjero	hombre	mala	44	21.3
extranjero	mujer	regular	40	19.3
extranjero	mujer	mala	37	18.8
extranjero	mujer	muy buena	25	17.8
extranjero	hombre	muy buena	7	16.3
extranjero	hombre	regular	7	22.5
peruano	mujer	mala	29	24.0
peruano	hombre	buena	56	16.2
peruano	hombre	muy buena	44	19.4
peruano	hombre	buena	7	20.6
peruano	hombre	regular	51	22.2
peruano	hombre	mala	41	18.0
peruano	hombre	regular	46	20.6
peruano	hombre	buena	41	19.0
peruano	mujer	regular	30	18.0
peruano	hombre	buena	45	23.5
peruano	mujer	regular	46	21.7
peruano	hombre	regular	22	17.2
peruano	mujer	muy buena	8	20.7
peruano	hombre	regular	64	19.4
peruano	mujer	mala	16	17.9
peruano	hombre	muy buena	41	16.4
peruano	mujer	buena	43	21.3
peruano	hombre	buena	12	22.5

Con la información que se muestra y usando un nivel de significación del 6% responda lo siguiente:

a. Verifique el supuesto de homogeneidad de varianzas en la edad para las personas de género femenino y masculino.
Respt: Prueba bilateral, Fcal = 0.4494, Fcrit = 0.3185 y 2.8052, Decisión: No RHo

b. ¿Existen diferencias significativas entre los pesos promedio de los equipajes de ambas géneros?
Respt: Prueba bilateral, Fcal = 0.4494, Fcrit = 0.3185 y 2.8052, Decisión: No RHo

c. ¿Se puede afirmar, que el porcentaje de viajeros de género femenino que opinan que el servicio es malo es diferente al porcentaje de viajeros de género masculino con tal opinión?

2. Se llevó a cabo una encuesta entre los miembros del Club del libro del mes, para determinar si pasan más tiempo viendo televisión que leyendo. Suponga que en una muestra de 12 encuestados se obtuvieron las horas semanales que se dedican a ver televisión y las que se dedican a la lectura. Con un nivel de significación del 5%, ¿se puede llegar a la conclusión de que los miembros del Club del libro del mes pasan más tiempo, en promedio, viendo televisión que leyendo? Asuma Normalidad de las variables en estudio.

Encuestado	1	2	3	4	5	6	7	8	9	10	11	12
Televisión	11	19	8	5	16	8	4	12	10	14	15	18
Leyendo	6	10	3	10	5	8	7	14	14	8	10	10

Respt: Prueba unilateral derecha, Tcal = 1.847, Tcrit = 1.79588, Decisión: RHo

3. Se realiza un estudio en la North Central University para medir el efecto del cambio ambiental en estudiantes extranjeros. Uno de los aspectos del estudio es una comparación del peso de los alumnos al ingresar a esa universidad, un año después se midió el peso de los estudiantes. Se sospecha que los alimentos estadounidenses más nutritivos provocan aumento de peso. Los datos para una muestra de estudiantes se dan a continuación.

Nombre	**Peso al inicio**	**Peso un año después**
Nassar	124	142
O'Toole	157	157
Oble	98	96
Silverman	190	212
Kim	103	116
Gross	135	134

Con 1% de nivel de significación, ¿los alimentos estadounidenses más nutritivos provocan aumento de peso?

CAPÍTULO III

PRUEBAS NO PARAMÉTRICAS: PRUEBAS JI-CUADRADO

- ✓ PRUEBA DE INDEPENDENCIA DE VARIABLES
- ✓ PRUEBA DE HOMOGENEIDAD DE PROPORCIONES
- ✓ PRUEBA DE BONDAD DE AJUSTE

Introducción

Como se ha visto en la sección anterior uno de los supuestos en el que se basa muchas de las pruebas estadísticas (conocidas como pruebas paramétricas) es el supuesto de normalidad.
Una parte de esta sección contempla el desarrollo de una prueba para verificar la normalidad de un conjunto de datos que se encuentra agrupado en una tabla de frecuencia.

Las pruebas a desarrollar son conocidas como pruebas no paramétricas. Están desarrolladas sobre la base de un estadígrafo que no hace referencia a ningún parámetro poblacional.
Este tipo de técnicas no utiliza directamente la información muestral recogida sobre la variable objeto de estudio, si no más bien la frecuencia con que aparecen dichos valores en la muestra.

Las pruebas a estudiar en esta sección son:

- Prueba de Independencia
- Prueba de Homogeneidad de proporciones.
- Prueba de Bondad de Ajuste.

Tabla de Contingencia

Es una tabla de frecuencia simple de dos vías (bidireccional). Sus r filas y columnas se usan para resumir y anotar los resultados de datos recolectados y jerarquizados de dos variables.

		Variable 2			
		Columna 1	**Columna 2**	**. . .**	**Columna c**
Variable 1	**Fila 1**	f_{11}	f_{12}		f_{1c}
	Fila 2	f_{11}	f_{11}		f_{11}
	. . .				
	Fila r	f_{r1}	f_{r1}	…	f_{rc}

3.1 Prueba de independencia

Una de las pruebas donde se utiliza la distribución Ji Cuadrada es cuando se desea probar que dos variables categóricas son independientes entre sí. Estas variables categóricas reciben el nombre de factores. El factor 1 o factor fila tiene "r" categorías y el factor 2 o factor que se muestra en la columna tiene "c" categorías.
En la prueba de independencia se prueba la hipótesis nula de que la variable fila y la variable de columna de una tabla de contingencia no están relacionadas. (La hipótesis nula es la proposición de que las variables de filas y de columna son independientes)

Ejemplo:

Para determinar si existe una relación entre el aprovechamiento de un empleado en el programa de capacitación y su rendimiento real en el trabajo, se tomó una muestra de 400 registros y se obtuvo las frecuencias observadas que se presentan en la siguiente tabla de contingencia:

Rendimiento (calificación del empleador)	**Aprovechamiento en el programa de capacitación**			
	Debajo del promedio	**Promedio**	**Sobre el promedio**	**Total**
Deficiente	23	60	29	112
Promedio	28	79	60	167
Muy bueno	9	49	63	121
Total:	60	188	152	400

Con el nivel de significación 0,01, ¿La calificación del rendimiento del trabajador está asociada con la calificación en aprovechamiento del programa de capacitación?

Solución:

Los factores que se muestran en la tabla son:

Variable 1: Calificación del rendimiento real en el trabajo, con 3 categorías: Deficiente, promedio y muy bueno.

Variable 2: Calificación en el programa de entrenamiento, con 3 categorías: Debajo del promedio, promedio o sobre el promedio.

La prueba de Independencia compara las frecuencias observadas, frente a otras llamadas frecuencias esperadas.

Para calcular las frecuencias esperadas se utiliza la siguiente fórmula:

$$\begin{pmatrix}\text{Frecuencia}\\ \text{esperada}\end{pmatrix} = \frac{(\text{Total de la columna}) \times (\text{Total del renglón})}{\text{Gran total}}$$

La siguiente tabla muestra: frecuencias observadas y esperadas (entre paréntesis)

Rendimiento en el trabajo (calificación del empleador)	**Aprovechamiento en el programa de capacitación**			
	Debajo del promedio	Promedio	Sobre el promedio	Total
Deficiente	23 (16,80)	60 (52,64)	29 (42,56)	112
Promedio	28 (25,05)	79 (78,49)	60 (63,46)	167
Muy bueno	9 (18,15)	49 (56,87)	63 (45,98)	121
Total:	**68**	**188**	**152**	**400**

Pasos para realizar la prueba de Independencia de variables
Valores críticos

1. Los valores críticos se encuentran de la tabla de contingencia con los grados de libertad $(r-1)(c-1)$
   Donde *r* es el número de renglones o filas y *c* es el número de columnas de la tabla

2. Las pruebas de hipótesis en tablas de contingencia, solo implican regiones críticas a la derecha

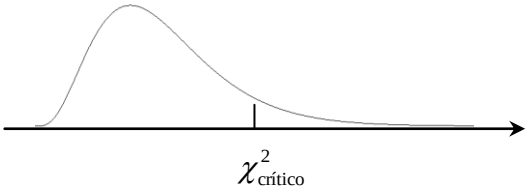
En la realización de la prueba de hipótesis los pasos sugeridos son:

1. Formular las hipótesis
2. Escoger α
3. La estadística de prueba Chi-cuadrado aproximada es:

$$\chi_c^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-e_{ij})^2}{e_{ij}}$$

O_{ij} representa las frecuencias observadas
e_{ij} representa las frecuencias esperadas

4. Establecer las regiones críticas y los criterios de decisión



Si $\chi_c^2 > \chi^2_{crítico}$, se rechaza la H_0.

5. Selección de la muestra y calcular la estadística de prueba
6. Aplicar los criterios de decisión y concluir.

NOTA: *El tamaño de muestra **n** total general debe ser suficientemente grande para asegurar que las frecuencias esperadas e_{ij} sean mayores o iguales a 5. Esto Asegura que la aproximación en la prueba sea buena.*

En el ejemplo:

Formulación de las hipótesis

H_0: La calificación del rendimiento real de un empleado en el trabajo es independiente del aprovechamiento en el programa de capacitación.
H_1: La calificación del rendimiento real de un empleado en el trabajo no es independiente del aprovechamiento en el programa de capacitación.

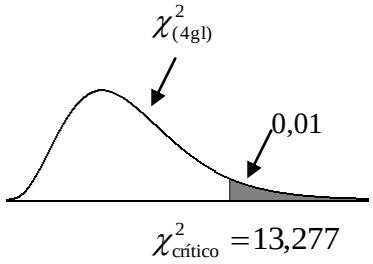
Fijación del nivel de significación: 0,01

Estadístico de prueba

$$\chi_c^2 = \sum_{i=1}^{r}\sum_{j1}^{c}\frac{(O_{ij}-e_{ij})^2}{e_{ij}} \sim \chi_{0,01}^2 \quad \text{con } (3\text{-}1)(3\text{-}1) = 4(\text{gl})$$

Criterios de decisión

Si el valor calculado $\chi^2 > 13{,}277$ se rechaza H_0
Si el valor calculado $\chi^2 \leq 13{,}277$ No se rechaza H_0



Resultado de la prueba:

$$\chi_c^2 = \frac{(23-16{,}80)^2}{16{,}80} + \frac{(28-25{,}05)^2}{25{,}5} + \ldots + \frac{(63-45{,}98)^2}{45{,}98} = 20{,}18$$

Con nivel de significación 0,01 se rechaza la hipótesis nula, por lo tanto la calificación del rendimiento real de un empleado en el trabajo depende (no es independiente) de la calificación en el programa de entrenamiento.

Nota:

- En Excel se puede hacer uso de la función **PRUEBA.CHICUAD** donde se debe ingresar las frecuencias observadas y las frecuencias esperadas. El resultado de la aplicación de esta funcion es el p-valor el cual es comparado directamente con el nivel de significación para dar las conclusiones.

- **(Corrección de Yates)**

  Cuando la muestra es menor de 50, o cuando algunas o todas las frecuencias esperadas son menores que 5, o cuando el grado de libertad es igual a 1, es recomendable aplicar la corrección de Yates; entonces el estadístico de prueba es el siguiente:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(|o_i - e_i| - 0.5\right)^2}{e_i} \sim \chi^2_{[(r-1)(c-1),\alpha]}$$

Ejemplo:

De acuerdo con una encuesta de participación en los deportes de la Asociación Nacional del Deporte de Estados Unidos, publicada en "American Demographics", las actividades deportivas en las que participa la gente está relacionada con el género. La siguiente tabla proporciona los resultados de una encuesta que incluía a 767 personas, clasificados por actividad deportiva (que practican con regular frecuencia) y por sexo. ¿La evidencia que

proporcionan estos datos es suficiente para inferir que el sexo y la actividad deportiva están relacionados? Use α=0,05

	Actividad deportiva			
Sexo	Ciclismo	Aeróbicos	Caminata	Natación
Hombres	85	28	60	179
Mujeres	81	138	106	90

PASO 1: HIPÓTESIS

PASO 2: NIVEL DE SIGNIFICANCIA

PASO 3: PRUEBA ESTADÍSTICA

PASO 4: REGIONES CRÍTICAS Y CRITERIO DE DECISIÓN

PASO 5: VALOR CALCULADO DEL ESTADÍSTICO DE PRUEBA

PASO 6: DECISIÓN

PASO 7: CONCLUSIÓN

Ejemplo

Un estudio de usuarios y no usuarios de cinturón de seguridad produjo los datos de muestra aleatoria que se resumen en la tabla adjunta. Pruebe la aseveración de que la cantidad de cigarrillos fumados es independiente del uso de cinturón de seguridad. Una teoría verosímil es que las personas que fuman más se preocupa menos por su salud y seguridad y, por tanto, tiene una menor inclinación a usar cinturón de seguridad.¿Los datos de muestra apoyan esta teoría?

	Número de cigarrillos fumados al día			
	0	1-14	15-34	35 o más
Usan cinturón de seguridad	175	20	42	6
No usan cinturón de seguridad	149	17	41	9

a) Realice la prueba respectiva, con un nivel de significación del 5%, usando el enfoque clásico
b) Realice la prueba respectiva, con un nivel de significación del 5%, usando el enfoque del valor *p*.

PASO 1: HIPÓTESIS

PASO 2: NIVEL DE SIGNIFICANCIA

PASO 3: PRUEBA ESTADÍSTICA

PASO 4: REGIONES CRÍTICAS Y CRITERIO DE DECISIÓN

PASO 5: VALOR CALCULADO DEL ESTADÍSTICO DE PRUEBA

PASO 6: DECISIÓN

PASO 7: CONCLUSIÓN

3.2 Prueba de homogeneidad de proporciones (o subpoblaciones)

Es una prueba estadística aproximada que se usa para determinar si las frecuencias esperadas en una fila son proporcionales a las frecuencias esperadas de cada uno de las otras filas de la tabla de contingencia o si las frecuencias en una columna son proporcionales a las frecuencias esperadas de las otras columnas de la tabla de contingencia.

Ejemplo:

La enfermería de un colegio llevó a cabo un experimento para determinar el grado de alivio proporcionado por tres remedios para la tos. Cada remedio se suministró a 50 estudiantes y se registraron los siguientes datos:

Efecto	Remedio para la tos		
	NyQuil	Robitussin	Triaminic
Sin alivio	11	13	9
Cierto alivio	32	28	27
Alivio total	7	9	14

Pruebe la hipótesis, con un nivel de significación del 5%, que los tres remedios para la tos son igualmente efectivos.

PASO 1: HIPÓTESIS

PASO 2: NIVEL DE SIGNIFICANCIA

PASO 3: PRUEBA ESTADÍSTICA

PASO 4: REGIONES CRÍTICAS Y CRITERIO DE DECISIÓN

PASO 5: VALOR CALCULADO DEL ESTADÍSTICO DE PRUEBA

PASO 6: DECISIÓN
PASO 7: CONCLUSIÓN

Ejemplo:

Muestras de tres tipos de materiales, sujetos a cambios extremos de temperatura, produjeron los resultados que se muestran en la siguiente tabla:

	Material A	**Material B**	**Material C**	**Total**
Desintegrados	41	27	22	90
Permanecieron intactos	79	53	78	210
Total	120	80	100	300

Use un nivel de significancia de 0.05 para probar si, en las condiciones establecidas, la probabilidad de desintegración es la misma para los tres tipos de materiales.

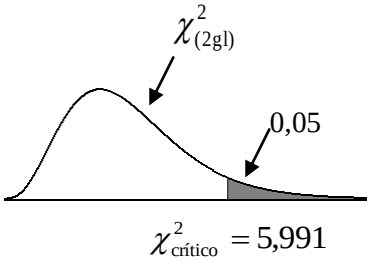
Pasos para realizar la Prueba de homogeneidad de proporciones

Formulación de las hipótesis

H_0: La probabilidad de desintegración es la misma para los tres tipos de materiales.
H_1: La probabilidad de desintegración no es la misma para los tres tipos de materiales.
Fijación del nivel de significación: 0,05

Estadístico de prueba

$$\chi_c^2 = \sum \frac{(O_{ij} - e_{ij})^2}{e_{ij}} \sim \chi_{0,05}^2 \quad \text{con } (2-1)(3-1) = 2(\text{gl})$$



Criterios de decisión.
Si $\chi_c^2 > 5{,}991$ se rechaza H_0
Si $\chi_c^2 > 5{,}991$ no se rechaza H_0

Cálculos previos:

	Material A	Material B	Material C	Total
Desintegrados	41 (36)	27 (24)	22 (30)	90
Permanecieron intactos	79 (84)	53 (56)	78 (70)	210
Total	120	80	100	300

$$\chi_c^2 = \frac{(41-36)^2}{36} + \frac{(79-84)^2}{84} + ... + \frac{(78-70)^2}{70} = 4{,}575$$

Con nivel de significación 0,05 no se rechaza la hipótesis nula.
Por lo tanto no se puede afirmar que la probabilidad de desintegración no es la misma para los tres tipos de materiales.

Usando el enfoque del valor *p*

p = 0,101es mayor que el nivel de significación $\alpha = 0,05$.

No rechazamos la hipótesis nula

Nota:

En Excel existe la función PRUEBA.CHICUAD que permite obtener el p-valor de la prueba estadística. Solo se requiere de la tabla de valores observados y valores esperados.

Ejercicios Propuestos:

1. Se realizó una encuesta para saber si existe una "brecha de género" en la confianza que la gente tiene en la policía. Los resultados de muestra se listan en la tabla adjunta. Use un nivel de significación del 0,05 para probar la afirmación de que sí existe una relación entre el género y la confianza en la policía.

Género	Confianza en la policía		
	Mucha	Regular	Muy poca o ninguna
Hombres	115	56	29
Mujeres	175	94	31

$$\chi_c^2 = 2,195; \quad \text{P} - \text{Valor} = 0,334$$

2. En un estudio de los sistemas lectores de cajas registradoras, se usaron muestras de compras para comparar los precios leídos con los precios anunciados. En la tabla adjunta se resumen los resultados para una muestra de 819 artículos. Cuando las tiendas usan lectores para registrar las compras, ¿las tasas de error son las mismas para los artículos a precio normal y los artículos en oferta?¿Cómo podría cambiar la conducta de los consumidores si creen que ocurre un número desproporcionado de cobros de más con los artículos en oferta? Use un nivel de significación del 6%

	Artículos normales	Artículos en oferta
Cobro de menos	20	7
Cobro de más	15	29
Precio correcto	384	364

$$\chi_c^2 = 10,814; \quad \text{P} - \text{Valor} = 0,004$$

3. Se realiza un estudio para determinar la relación entre el tipo de crimen y siel criminal es un extraño o no. La tabla adjunta lista los resultados de una encuesta practicada a una muestra aleatoria de víctimas de diversos crímenes. Con un nivel de significación de 0,05, pruebe la Hipótesis respectiva.

	Homicidio	Asalto	Agresión
El criminal era un extraño	12	379	727
El criminal era un conocido o pariente	39	106	642

$\chi_c^2 = 119.330; \quad P - Valor = 0{,}0000$

4. Un estudio de accidentes automovilísticos seleccionados al azar y conductores que usan teléfonos celulares proporcionó los datos de muestra adjuntos. Se desea saber si existe alguna relación entre la ocurrencia de accidentes y uso de teléfonos celulares. Con base en estos resultados, realice la prueba correspondiente con un nivel de significación del 5%.

	Tuvo accidente el año pasado	No tuvo accidente el año pasado
Usa teléfono celular	23	282
No usa teléfono celular	46	407

$\chi_c^2 = 1{,}505; \quad P - Valor = 0{,}220$

5. La tabla adjunta lista datos de muestra que el estadístico Karl Pearson usó en 1909. ¿Cree usted que el tipo de delito esté relacionado con el hecho de que el criminal beba o se abstenga? ¿Hay delitos aparentemente asociados al hábito de beber?

	Incendio provocado	Violación	Violencia	Robo	Falsificación	Fraude
Bebedor	50	88	155	379	18	63
Abstemio	43	62	110	300	14	144

$\chi_c^2 = 49{,}731; \quad P - Valor = 0{,}000$

6. Una de las preguntas del estudio de suscriptores de 1996 de Bussiness Week fue: "Durante los últimos 12 meses, en viajes de negocios, ¿qué tipo de boleto de avión compró con más frecuencia?" Las respuestas obtenidas se muestran en la siguiente tabla:

		Tipo de vuelo	
		Nacional	Internacional
Tipo de boleto	Primera clase	29	22
	Clase de negocios o ejecutiva	95	121
	Clase económica	518	135

Usando nivel de significación 0,05, pruebe la independencia del tipo de vuelo y tipo de boleto. $\chi_c^2 = 100{,}434; \quad P - Valor = 0{,}000$

7. En el estudio de un taller, se obtuvo un conjunto de datos para determinar si la proporción de artículos defectuosos producidos por los trabajadores era la misma durante el día, la tarde o la noche. Se encontraron los siguientes resultados:

Condición	TURNO		
	Día	Tarde	Noche
Defectuosos	45	55	70
No defectuosos	905	890	870

Utilice un nivel de significación del 2,5% para determinar si la proporción de artículos defectuosos es la misma para los tres turnos.

$$\chi_c^2 = 6{,}234; \quad \text{P} - \text{Valor} = 0{,}044$$

8. La enfermería de un colegio llevó a cabo un experimento para determinar el grado de alivio proporcionado por tres remedios para la tos. Cada remedio se suministró a 50 estudiantes y se registraron los siguientes datos:

Efecto	Remedio para la tos		
	NyQuil	Robitussin	Triaminic
Sin alivio	11	13	9
Cierto alivio	32	28	27
Alivio total	7	9	14

Pruebe la hipótesis, con un nivel de significación del 5%, que los tres remedios para la tos son igualmente efectivos.

$$\chi_c^2 = 3{,}810; \quad \text{P} - \text{Valor} = 0{,}432$$

3.3 Prueba de bondad de ajuste

Caso 1: Población multinomial

El procedimiento de prueba para los experimentos multinomiales es bastante semejante al descrito para el caso de independencia y homogeneidad, el mayor cambio viene con el planteamiento de la hipótesis nula. Puede ser un planteamiento verbal.

- ✓ El valor critico es determinado por el nivel de significación asignado (α) y el numero de grados de libertad es una unidad menos que el número de clases o categorías (k) en que se dividen los datos. (gl = k –1).

- ✓ Cada frecuencia esperada e_i se determina al multiplicar el número total de ensayos n por la probabilidad correspondiente $e_i = n\,p_i$

Población multinomial: cuando cada elemento de una población se asigna a una y sólo una de varias categorías.
La distribución multinomial de probabilidad se puede concebir como una ampliación de la distribución binomial para el caso de tres o más categorías de resultados.

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \sim \chi^2_{(k-1)}$$

Donde:
o_i: frecuencia observada para la categoría i.
e_i: frecuencia esperada para la categoría i.
k: Número de categorías.
Nota: Las e_i deben ser cinco o más para todas las categorías.

Ejemplo:
A continuación se presentan las preferencias de grupos de consumidores hacia tres aparadores de tienda.

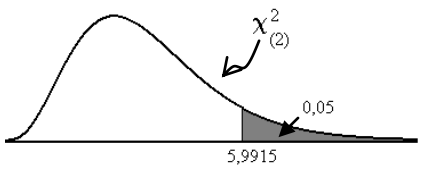
Aparador A	Aparador B	Aparador C
43	53	39

Use nivel de significación 5% para probar si hay alguna diferencia de preferencia hacia los tres aparadores.

Solución:
H_0: La preferencia de consumidores es la misma para cada aparador
H_1: La preferencia de consumidores no es la misma para cada aparador
Nivel de significación de la prueba: 0,05

Estadística de prueba: $\chi_c^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \sim \chi^2_{(2)}$

Regiones críticas y criterios de decisión:



Si χ_c^2, es mayor que 5,9915; se rechaza la H_0

Cálculos:

A continuación se muestra la tabla que contiene las frecuencias observadas, las frecuencias esperadas entre otros valores que se requieren para esta prueba.

Aparador	*o*i	*p*i	*e*i = *np*i	$(o_i-e_i)^2/e_i$
A	43	1/3	45	0,08888889
B	53	1/3	45	1,42222222
C	39	1/3	45	0,8
Total	135	1	135	**2,31111111**

Observe que las probabilidades de preferencia para cada aparador deben ser las mismas, pues debe tenerse igual frecuencia teóricas en el supuesto de que las preferencias son las mismas para cada aparador.

$$\chi_c^2 = \sum_{i=1}^{3} \frac{(o_i - e_i)^2}{e_i} = 2{,}3111$$

Conclusión: Como χ_c^2 es menor que 5,9915; no rechazamos la H0.

Bajo un nivel de significación del 5% no podemos afirmar que las preferencias de los consumidores a los aparadores A, B y C es la misma.

Caso 2: Distribución de Poisson

Hipótesis

H_0: La población tiene distribución de probabilidad de Poisson.

H_1: La población no tiene distribución de probabilidad de Poisson.

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \sim \chi^2_{(k-1-m)}$$

o_i: frecuencia observada para la categoría i.

e_i: frecuencia esperada para la categoría i.

k: número de categorías.

m: número de parámetros a estimar

Nota: Las e_i deben ser cinco o más para todas las categorías.

Ejemplo:

Se cree que el número de accidentes automovilísticos diarios en determinada ciudad tiene una distribución de Poisson. En una muestra de 80 días del año pasado se obtuvieron los datos de la tabla adjunta. ¿Apoyan estos datos la hipótesis de que el número diario de accidentes tiene una distribución de Poisson? Use nivel de significación 0,05.

N° accidentes	Frecuencia observada
0	34
1	25
2	11
3	7
4	3

Solución:

H_0: La población tiene distribución de probabilidad de Poisson

H_1: La población no tiene distribución de probabilidad de Poisson

Nivel de significación de la prueba: 0,05

Calculando la media (un parámetro a estimar)

N° accidentes (*xi*)	Frecuencia observada (O_i)	$O_i x_i$
0	34	0
1	25	25
2	11	22
3	7	21
4	3	12
	80	80

$$\hat{\mu} = \bar{x} = \frac{\sum o_i x_i}{n} = \frac{80}{80} = 1$$

A continuación tenemos otros cálculos que nos permiten realizar la prueba y obtener los grados de libertad de la estadística de prueba.

N° accidentes (X)	Probabilidad de Poisson	$e_i = np_i$
0	0,3679	29,43
1	0,3679	29,43
2	0,1839	14,72
3	0,0613	**4,91**
4	0,0613	**1,52**
	1,0000	80,00

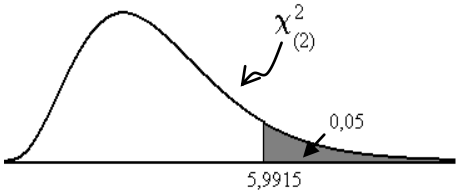
Observe que las tres últimas clases tienen frecuencias menores a cinco

Tenemos la siguiente tabla que resulta de unir las tres últimas clases. Los grados de libertad para la distribución Chi- cuadrado de la prueba son: $k - m$-1 = 4-1-1 = 2 grados de libertad.

Frecuencia observada (oi)	Frecuencias esperadas (ei)	(oi - ei)2/ei
34	29,43	0,7096
25	29,43	0,6668
11	14,72	0,9401
10	6,42	1,9963
TOTAL: 80	TOTAL: 80	TOTAL: 4,3129

Estadística de prueba:

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \sim \chi^2_{(2)}$$



Regiones críticas y criterios de decisión:

Si χ_c^2, es mayor que 5,9915; se rechaza la H_0

$$\chi_c^2 = \sum_{i=1}^{3} \frac{(o_i - e_i)^2}{e_i} = 4{,}3129$$

Conclusión: Como χ_c^2 es menor que 5,9915; no rechazamos la H_0.

Bajo un nivel de significación del 5% no podemos afirmar que la población no tiene distribución de probabilidad de Poisson con una media de un accidente por día.

Ejercicios:

1. En un estudio de Análisis de Mercado realizado por "Compañía de Investigación de Mercado(CIM)", se observa que la participación de tres empresas competidoras era del 30% de la Compañía A, 50% de la Compañía B y 20% de la Compañía C. Si la Compañía C introdujo un nuevo producto de "Calidad Extra Blanca", se producirá una modificación en el comportamiento del mercado. ¿Se modificará la participación de cada una de las empresas competidoras en el mercado?
   Para tratar de responder a esta pregunta, CIM llevó a cabo una encuesta practicada a 200 clientes a fin de averiguar sobre su preferencia de compra en las tres compañías.

   La encuesta arrojó los siguientes resultados:

   48 indicaron que prefieren el producto de la Compañía A;
   98 indicaron que prefieren el producto de la Compañía B y
   54 indicaron que prefieren el producto de la Compañía C.
   Realice la prueba correspondiente con un nivel de significación del 5%.

2. Se distribuyó el número de incidencias resueltas diariamente por un joven administrador durante sus primeros 126 días de práctica, de la siguiente manera:

N° De clientes	0	1	2	3	4	5
N° De días	24	36	28	18	12	8

   Pruebe si el número de incidencias resueltas sigue una distribución Poisson.

Ejercicios Propuestos:

1. Durante las primeras 13 semanas de la temporada de televisión, se registraron las audiencias de sábado por la noche, de 8:00 p.m. a 9:00 pm. Como sigue: ABC 29%, CBS 28%, NBC 25% y otros 18%. Dos semanas después, una muestra de 300 hogares arrojó los siguientes resultados de audiencia: ABC 95 hogares, CBS 70 hogares, NBC 89 hogares y otros 46 hogares. Pruebe, con nivel de significación 0,05, si han cambiado las proporciones de telespectadores.

2. Suponga que los investigadores desean determinar si el patrón de distribución del ingreso familiar en el Perú, ha cambiado significativamente durante los últimos cinco años. Se sabe que hace cinco años la distribución del ingreso familiar para las distintas clases de ingreso era la siguiente:

Clase de Ingreso ($)	**% de todas las familias en la clase**
(1) menos de 3000	9
(2) de 3000 a menos de 5000	11
(3) de 5000 a menos de 7000	12
(4) de 7000 a menos de 10000	22
(5) de 10000 a menos de 15000	27
(6) de 15000 a menos de 25000	15
(7) de 25000 a mas	4
TOTAL	100

Se elige una muestra aleatoria de 1000 familias y se obtiene la siguiente distribución:

Clase de Ingreso ($)	1	2	3	4	5	6	7
Número de familias	70	100	110	200	300	170	50

Con α = 0,05, ¿el patrón actual de distribución del ingreso familiar es significativamente distinto al de hace cinco años?

3. Suponga que el número de llamadas telefónicas que entran al conmutador de una empresa durante intervalos de un minuto tiene una distribución de Poisson. Use nivel de significación 0,10 y los siguientes datos para probar la hipótesis de que las llamadas que entran tiene distribución de Poisson.

N° llamadas que entran c/min., (X)	0	1	2	3	4	5	6
Frecuencia observada, (Oi)	15	31	20	15	13	4	2

CAPÍTULO IV

ANÁLISIS DE VARIANZA

- ✓ **ANÁLISI DE VARIANZA DE UN FACTOR**

Introducción

En la sección 2 se desarrollaron pruebas de hipótesis para comparar dos medias poblacionales (prueba de diferencia de medias) o dos variancias poblacionales (prueba de homogeneidad). Sin embargo en muchas aplicaciones, se desea comparar más de dos poblaciones. En este caso el planteamiento de un Diseño Experimental el cual permite la realización del Análisis de Variancia (ANVA o ANOVA por Analisys of Variance) o la descomposición de la variabilidad total en cada una de sus componentes es de gran utilidad.

Los Diseños Experimentales mediante el ANVA permiten probar si dos o más poblaciones tienen la misma media. Aun cuando el propósito del ANVA es hacer pruebas para hallar las diferencias en las medias poblacionales, implica un examen de las variancias muestrales; de allí el término de análisis de variancia.

4.1 Conceptos Básicos

a) Factor o Variable independiente:
Es una variable externa que afecta los resultados del experimento. El factor en estudio es controlado por el investigador y es de interés estudiarlo. A los distintos valores que puede tomar el factor se le denomina niveles del factor. En un experimento se puede evaluar un solo factor o más de uno.
Ejemplo:

- Factor: Intensidad de publicidad
  Niveles: Bajo, Mediano, Alto

b) Tratamiento:
Corresponde a cada nivel de un factor o también es la combinación de los niveles de varios factores considerados en el experimento.
Ejemplo:

- Factor1: Regiones de venta
  Niveles o Tratamientos: norte, sur, centro, este u oeste
- Factor 2: Categoría de experiencia del vendedor
  Niveles o Tratamientos: Junior, senior

Si se combinan los niveles de ambos factores se pueden obtener los siguientes tratamientos:
Mañana-Junior, Mañana Senior, Tarde-Junior, Tarde Senior, Noche-Junior, Noche Senior,

c) Unidad Experimental:
Es el elemento en el cual se aplica un tratamiento.
Ejemplo:

- Un empleado de una fábrica.

d) Variable respuesta o Variable dependiente:
Es la característica en la cual se evaluaran los efectos de los tratamientos
Ejemplo:

- Puntuaciones obtenidas en una evaluación de capacitación.
- Tiempo (en minutos) de ensamblaje de un producto.

e) Dato u observación:
Es el registro numérico obtenido después de la aplicación del tratamiento a la unidad experimental.

- 15 puntos.
- 18.5 minutos.

f) Diseño Experimental
Es la distribución de los tratamientos (niveles de un factor o combinación de los niveles de varios factores) a las unidades experimentales. Así, también involucra la elección del tamaño muestral y la disposición de las unidades experimentales.
El uso del diseño experimental adecuado permite minimizar el error experimental.

g) Error Experimental
Son las diferencias observadas en los valores de la variable respuesta de cada una de las unidades experimentales por una acción diferente a la de los tratamientos.

h) Análisis de Varianza
El término análisis de varianza describe una técnica mediante la cual se analiza la variación total que existe en una variable respuesta asignando partes de esta variación a componentes representativos (variables independientes y error aleatorio). El objetivo del análisis de varianza consiste en localizar las variables independientes importantes y determinar como afectan la respuesta.

Ejemplo:
El gerente de un establecimiento comercial desea realizar un estudio para comparar el monto de compra (en soles) de sus clientes de acuerdo a la forma de pago (al contado, con tarjeta de crédito ó tarjeta de débito). Durante un día selecciona al azar a 5 clientes de acuerdo a c/u de los tres tipos de forma de pago que admite su establecimiento.

Solución:

- Variable respuesta o Variable Dependiente: Monto de compra (en soles).
- Factor o Variable independiente: Forma de pago.
- Niveles del factor ó tratamientos: al contado, con tarjeta de crédito ó tarjeta de débito
- Unidad experimental: un cliente.

4.2 Diseño de un Factor: Diseño Completamente Aleatorizado (D.C.A.)

Supongamos que el experimentador cuenta con los resultados de *k* muestras aleatorias independientes, cada una de tamaño *n*, provenientes de *k* diferentes poblaciones (esto es, datos relativos a *k* tratamientos, *k* grupos, *k* métodos de producción, etc.) y le interesa probar la hipótesis de que las medias de esas *k* poblaciones son todas iguales.

Muestra	Tratamientos				Total
	Tratam.1	Tratam.2	. . .	Tratam.k	
1	y_{11}	y_{21}	. . .	y_{k1}	$y_{\cdot 1}$
2	y_{12}	y_{22}	. . .	y_{k2}	$y_{\cdot 2}$
3	y_{13}	y_{23}	. . .	y_{k3}	$y_{\cdot 3}$
.	.	.	. . .	.	.
.	.	.	. . .	.	.
.	.	.	. . .	.	.
n_i	y_{1n_1}	y_{1n_2}	. . .	y_{1n_k}	$y_{1n_\bullet}$
Total	$y_{1.}$	$y_{2.}$	. . .	$y_{k.}$	$y_{..}$

Para probar la hipótesis de que las muestras se obtuvieron de *k* poblaciones con medias iguales, haremos varias suposiciones. Con más precisión, supondremos estar trabajando con **poblaciones normales** que tienen **varianzas iguales**.

Modelo Aditivo Lineal: Para un diseño completamente al azar, es el siguiente:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \qquad i = 1, 2, \ldots, k \quad j = 1, 2, \ldots, n_i$$

Y_{ij} : Es el valor observado en el i-ésimo tratamiento y la j-ésima muestra.

μ : Es el efecto de la media general,

τ_i : Es el efecto del i-ésimo tratamiento.

ε_{ij} : Es el efecto del error experimental en el i-ésimo tratamiento y la j-ésima repetición.

Tabla del análisis de varianza (cuadro ANVA):

Fuente de variación	**Grados de libertad**	**Suma de Cuadrados**	**Cuadrado Medio**	**Fc**
Tratamientos	$k - 1$	$SC(Tr) = \sum_{i=1}^{k} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{n_\bullet}$	$CM(Tr) = \frac{SC(Tr)}{k-1}$	$\frac{CM(Tr)}{CME}$
Error	$n. - k$	$SCE = SCT - SC(Tr)$	$CME = \frac{SCE}{n_\bullet - k}$	
Total	$n. - 1$	$SCT = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{..}^2}{n_\bullet}$		

Ejemplo:

Una compañía que fabrica computadoras ha instituido 4 programas diferentes de entrenamiento (Alfa, Beta, Gamma y Sigma) para los empleados que trabajan en operaciones de ensamblado. Veinte trabajadores fueron distribuidos aleatoriamente a los 4 programas para posteriormente evaluar su tiempo de ensamblado (en minutos), obteniéndose los siguientes resultados:

Repetición	Programa			
	Alfa	Beta	Gamma	Sigma
1	59	52	65	64
2	64	58	71	67
3	57	54	63	62
4	62	56	64	64
5	60	58	63	66
Total	302	278	326	323
Promedio	60.4	55.6	65.2	64.6

Pruebe si existen diferencias en los tiempos promedios de los métodos de ensamblado a un nivel de significación de 0.05.

El Modelo Aditivo Lineal es:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \qquad i = 1,2,3,4 \quad j = 1,2,\ldots,5$$

Donde:

Y_{ij} : Es tiempo de ensamblaje obtenido con el i-ésimo método de ensamblaje en la j-ésima repetición.

μ : Es el efecto de la media general del tiempo de ensamblaje

τ_i : Es el efecto del i-ésimo método de ensamblaje.

ε_{ij} : Es el efecto del error experimental en el i-ésimo método de ensamblaje y la j-ésima repetición.

La tabla del análisis de varianza es:

Fuente de variación	Grados de libertad	Suma de Cuadrados	Cuadrado Medio	F_c	F_t
Tratamientos	4 – 1 = 3	296.55	98.85	13.587	3.24
Error	20 - 4= 16	116.4	7.275		
Total	5(4) – 1 = 19	412.95			

Resultados con Excel:

Análisis de varianza de un factor

RESUMEN

Grupos	*Cuenta*	*Suma*	*Promedio*	*Varianza*
Alfa	5	302	60.4	7.3
Beta	5	278	55.6	6.8
Gamma	5	326	65.2	11.2
Sigma	5	323	64.6	3.8

ANÁLISIS DE VARIANZA

FV	*SC*	*GL*	*CM*	*F*	*Probabilidad*	*Fcrit*
Entre grupos	296.55	3	98.85	13.5876289	0.00011516	3.23887152
Dentro de los grupos	116.4	16	7.275			
Total	412.95	19				

H_0: $\mu_1=\mu_2=\mu_3=\mu_4$
H_1: Al menos un μ_i es diferente a los demás i = 1, 2, 3, 4; $\alpha = 0.05$
P-valor=0.00011516 < α , se rechaza H_0

Existe suficiente evidencia estadística a un $\alpha = 0.05$ para rechazar H_0.
Por lo tanto se puede afirmar que si existen diferencias en los tiempos promedios de al menos uno de los métodos de ensamblado.

Hasta el momento se ha encontrado que al menos un tratamiento presenta un efecto diferente al resto, sin embargo no se sabe cual (o cuales) tratamientos son los distintos. Para responder esa interrogante se debe desarrollar las pruebas de comparación.

4.3 Pruebas de comparación

Cuando se produce el rechazo de la hipótesis nula se concluye que las medias de las poblaciones (tratamientos) no son todas iguales. Para determinar entre qué tratamientos existe diferencia de promedios se propone entre otras pruebas, la "Diferencia Mínima de Significación".

Prueba Diferencia Mínimas Significativas (DMS)

Se plantea las siguientes hipótesis:

$$\begin{aligned} H_0 &: \mu_i = \mu_j \\ H_1 &: \mu_i \neq \mu_j \end{aligned} \quad \forall\ i \neq j$$

Se rechaza H_0 si: $$|\bar{y}_i - \bar{y}_j| > t_{(n-k,\ 1-\alpha/2)} \sqrt{CME\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

Ejemplo:

Realice las pruebas de comparación para determinar que método de ensamblaje presentan diferencias significativas. Use un nivel de significación de 0.05

$H_0 : \mu_i = \mu_j$

$H_1 : \mu_i \neq \mu_j$

$\alpha = 0.05$

$$t_{(n-k,\ 1-\alpha/2)} \sqrt{CME\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} = 2.1199\sqrt{7.275\left(\frac{1}{5}+\frac{1}{5}\right)} = 3.616$$

Comparación	Diferencia			Límite	Conclusión	Significación
Alfa-Beta	60.4	55.6	**4.8**	3.616	Se rechaza Ho	*
Alfa-Gamma	60.4	65.2	**4.8**	3.616	Se rechaza Ho	*
Alfa-Sigma	60.4	64.6	**4.2**	3.616	Se rechaza Ho	*
Beta-Gamma	55.6	65.2	**9.6**	3.616	Se rechaza Ho	*
Beta-Sigma	55.6	64.6	**9.0**	3.616	Se rechaza Ho	*
Gamma-Sigma	65.2	64.6	**0.6**	3.616	No se rechaza Ho	n.s.

Resumen

$\bar{y}_2 = 55.6 \qquad \bar{y}_1 = 60.4 \qquad \bar{y}_4 = 64.6 \qquad \bar{y}_3 = 65.2$

Si se desea elegir el método que produce menor tiempo promedio de ensamblaje, este seria el método Beta. Se puede observar que no existen diferencias significativas entre los métodos Gamma y Sigma.

Ejemplo:

Los siguientes datos corresponden a las ventas mensuales (en miles de dólares) para 12 tiendas ubicadas en 4 regiones donde una gran empresa distribuidora realiza sus operaciones.

Región A	Región B	Región C	Región D
0.25	0.18	0.19	0.23
0.33	0.28	0.25	0.30
0.22	0.21	0.27	0.28
0.30	0.23	0.24	0.28
0.27	0.25	0.18	0.24
0.28	0.20	0.26	0.34
0.32	0.27	0.28	0.20
0.24	0.19	0.24	0.18
0.31	0.24	0.25	0.24
0.26	0.22	0.20	0.28
0.20	0.29	0.21	0.22
0.28	0.16	0.19	0.21

Solución:
La tabla del Análisis de Varianza es:

Fuente de variación	Grados de Libertad	Suma de Cuadrados	Cuadrados Medios	F_c	F_t
Región	3	0.0156	0.0052	3.133	2.82
Error	44	0.0728	0.0017		
Total	47	0.0884			

Primero se ordena las cuatro medias en orden creciente:

Región	B	C	D	A
Media	0.227	0.230	0.250	0.272

Prueba Diferencia Mínimas Significativas (DMS)

Región	**Región A**	**Región B**	**Región C**	**Región D**
n_i	12	12	12	12
Media	0.272	0.227	0.230	0.250

$t_{(44,0.975)} =$ 2.015
CME = 0.0017

Comparación		Diferencia de promedios	$t_{(n-k,1-\alpha/2)}\sqrt{CME\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$	
i	j			
A	B	0.045	0.0339	*
	C	0.042	0.0339	*
	D	0.022	0.0339	NS
B	C	-0.003	0.0339	NS
	D	-0.023	0.0339	NS
C	D	-0.020	0.0339	NS

Ejercicios

A continuación se muestra información sobre en nivel de ventas (en miles de dólares) obtenido por las sucursales de una empresa extranjera que opera en nuestro país a nivel nacional. Con el fin de investigar la incidencia del nivel de publicidad sobre las ventas, se asignaron al azar diferente número de tiendas para cada nivel de publicidad:

Nivel de publicidad			
Baja	**Media**	**Media Baja**	**Alta**
65	75	59	94
73	69	78	89
79	83	67	80
81	81	62	88
69	72	83	
	79		

Determine si existen diferencias en el promedio de ventas para los distintos niveles de publicidad. Use α=0.05.

Ejercicios Propuestos

1. El gerente de personal de un banco desea analizar la eficiencia en el tiempo de atención de 4 empleados. Se midió el tiempo que demoran en atender a un cliente cuando cobra un cheque. Los resultados se muestran a continuación (en minutos):

Trabajador			
López	**Valencia**	**Gutierrez**	**Chavez**
3.6	3.7	3.4	3.7
3.6	3.8	3.9	3.9
3.9	4.2	3.8	3.6
3.8	3.9	3.5	3.9
3.8	4.0	3.7	3.6

   ¿Existe evidencia estadística que permita concluir que el efecto de los medicamentos no es el mismo?, Use α=0.05.

2. El director de capacitación de una compañía manufacturera desea comparar tres enfoques de trabajo en equipo. Cada miembro de un grupo de 24 empleados nuevos se asigna al azar a uno de los tres métodos. Terminada la capacitación, se evalúa el tiempo que tardan (en minutos) en ensamblar el producto. Los resultados son:

METODOS		
A	**B**	**C**
8,82	8,21	8,57
9,26	6,65	8,50
8,70	7,44	9,11
8,97	7,95	8,20
8,64	8,20	8,32
8,29	7,75	7,88
9,45	8,84	9,90
9,42	8,40	9,43

   Analice los datos considerando un nivel de significación del 5%. Determine ¿cuál es el método más efectivo?

NOTAS:

CAPÍTULO V

ANÁLISIS DE REGRESIÓN

- ✓ REGRESIÓN LINEAL SIMPLE
- ✓ REGRESIÓN CURVILINEAL
- ✓ REGRESIÓN LINEAL MÚLTIPLE

5.1 Regresión lineal simple

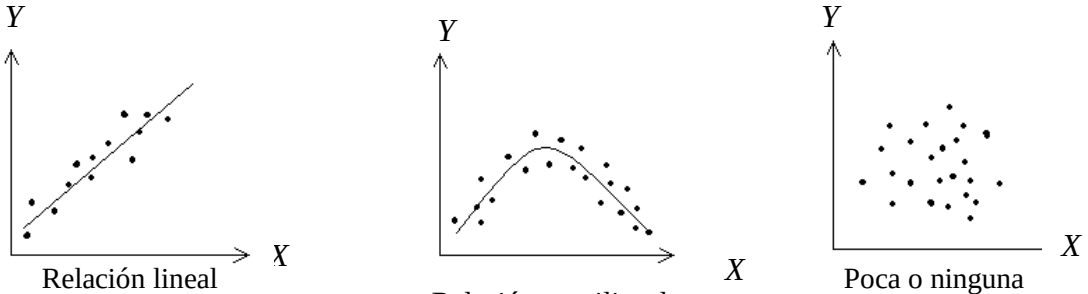
Es el estudio de la relación lineal entre una variable aleatoria Y, llamada ***variable dependiente*** y otra variable *X*, llamada ***variable independiente o explicativa***,

Los datos se pueden representar mediante pares de observaciones o mediciones para cada variable denotadas por (x_i, y_i); $i = 1,2,\ldots,n$ las cuales pueden representarse en un sistema rectangular, lo que genera un ***diagrama de dispersión.***

Según lo que se aprecie en un diagrama de dispersión se puede proponer una relación lineal o no entre las variables.

DIAGRAMA DE DISPERSIÓN

Es una gráfica en la que cada punto representa un par de valores observados (x_i,y_i) de las variables dependientes e independientes. El valor de la variable independiente, *X* se grafica en el eje horizontal, mientras que el valor de la variable dependiente, *Y* en el eje vertical.
El tipo de la relación observada en el diagrama de dispersión puede ser curvilínea (relación no lineal), puede ser lineal o ninguna de las anteriores.



Si el diagrama de dispersión indica una relación de tipo lineal, entonces se estima una línea recta a los datos.

La relación que se proponga entre estas variables no es exacta. Es decir, a un valor dado de *X* no corresponde un valor exacto de *Y*. No es un modelo determinístico.
Para una relación lineal, la ecuación propuesta tiene la forma:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

El término ε es denominado "término aleatorio" o "error".
El modelo propuesto se denomina "Modelo de Regresión Lineal Simple". La linealidad se expresa por que la función es lineal en los parámetros pero no necesariamente es lineal en *X*.

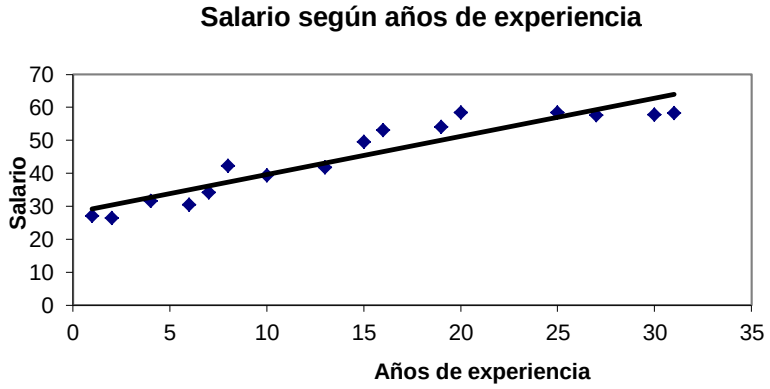
Ejemplo 1

Se llevó a cabo un estudio para determinar la relación entre el número de años de experiencia (X) y el salario mensual, en cientos de soles,(Y) entre los administradores de cierta ciudad. Para ello, se tomó una muestra aleatoria de 17 administradores y se obtuvieron los siguientes datos:

Experiencia	Salario	Experiencia	Salario
13	41,76	20	58,40
16	53,12	1	27,04
30	57,76	4	31,68
2	26,40	10	39,36
8	42,24	27	57,60
6	30,56	25	58,40
31	58,24	7	34,24
19	54,08	15	49,60

Elabore un diagrama de dispersión con la información anterior

Solución:



De la inspección del diagrama de dispersión, se ve que los puntos tienden a agruparse alrededor de una recta, lo que nos hace suponer que la relación entre las dos variables es de carácter lineal.

Por supuesto, esta es una justificación intuitiva para el uso del análisis de regresión lineal, más adelante se continuará con un mayor desarrollo y discusión respecto a la elección del modelo.

Objetivos y supuestos del modelo

Objetivos

El objetivo principal del análisis de regresión es estimar el valor de la ***variable dependiente***, sabiendo que el valor de la ***variable independiente***, es conocido. La variable dependiente se llama también ***variable respuesta*** y la variable independiente también se conoce como ***variable predictora.***

Supuestos

1. Las variables independiente y dependiente se asocian linealmente y la relación funcional entre ellas puede ser expresada mediante el modelo lineal: $y_i = \beta_o + \beta_1 x_i + \varepsilon_i$
2. Los términos aleatorio ε_i son independientes y tienen una distribución normal con media 0 y varianza σ_ε^2
3. Los valores de X son fijados y se miden sin error
4. Para cada valor de *X*, los valores de *Y* tienen una distribución normal con media $\beta_0 + \beta_1 X$ y varianza $\sigma_{y.x}^2$
5. Las distribuciones de *Y* para los diferentes valores de *X*, tienen igual varianza, es decir: $\sigma_{y.x_1}^2 = \sigma_{y.x_2}^2 = \cdots = \sigma_{y.x_n}^2 = \sigma^2$
6. Los valores de *Y*, para cada valor de *X*, son obtenidos mediante una muestra aleatoria.

Estimación de los parámetros del modelo

El modelo propuesto está dado por:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Los parámetros β_0 y β_1 deben ser estimados. Para el proceso de estimación se usa el método de "Mínimos Cuadrados" que propone minimizar la suma de cuadrados del error.

Los parámetros estimados están expresados por:

$$\hat{\beta}_1 = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Si bien es cierto que aquí presentamos las fórmulas para estimar los coeficientes de la línea de regresión, cabe resaltar que Excel tiene una opción en Herramientas / Análisis de Datos / Regresión que facilita estos cálculos.

Coeficiente de correlación

El coeficiente de correlación mide el grado de asociación lineal que existe entre dos variables. El coeficiente de correlación poblacional se denota por ρ y se encuentra dentro del intervalo cerrado de [-1 y 1].

Si ρ esta cerca de cero entonces indicará que no existe relacion lineal significativa entre las variables mientras que cuando se acerca a 1 o a -1 indicará que existe una relacion lineal fuerte, y cuando ρ esta cerca a 1 ó -1 la asociación es perfecta, directa e inversa respectivamente..
El estimador de ρ es "r" y se calcula mediante la siguiente fórmula:

$$r = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left[n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2\right].\left[n\left(\sum_{i=1}^{n} y_i^2\right) - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

Coeficiente de determinación

Es una medida de la bondad del ajuste para una ecuación de regresión. Mide el porcentaje de variación total que es explicada por la ecuación de regresión.
Su rango de valores está entre 0% y 100%.

$$R^2 = \frac{SC\,\mathrm{Re}\,g}{SCTotal} \times 100\%$$

donde: SCReg indica la "Suma de Cuadrados de la Regresión" y SCTotal indica la "Suma de Cuadrados del Total".

Sumas de Cuadrados

$$SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

$$SCR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \hat{\beta}_1^2 \left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \right)$$

$$SCE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \ = SST - SSR$$

Para calcular estos valores usaremos las herramientas estadísticas de Excel.

Prueba de Significación:

Permite verificar si existe relación lineal entre las variables en estudio:

$$H_0 : \beta_1 = 0$$
$$H_0 : \beta_1 \neq 0$$

El estadístico de prueba se obtiene a partir de la construcción de la tabla de ANOVA:

Tabla de ANOVA

Fuente de Variación	Grados de Libertad	Suma de Cuadrados	Cuadrado medio	F_c	F_{tab}
Regresión	1	SSR	CMR	F_{cal}	$F_{(1;\ n-2)}$
Error	n-2	SSE	CME		
Total	n-1	SST			

Se rechazará la hipótesis nula si: $F_c > F_t$
También: se rechaza H_0 si Valor $p < \alpha$

Inferencia sobre la pendiente:

Prueba de hipótesis β_1

$$H_0 : \beta_1 = 0$$
$$H_0 : \beta_1 \neq 0$$

El estadístico de prueba es: $t = \dfrac{\hat{\beta}_1 - \beta_1}{\dfrac{s}{\sqrt{S_{xx}}}} \sim t_{(n-2)}$

El cociente $\dfrac{s}{\sqrt{S_{xx}}}$ es denominado el error tipico de la pendiente y es asi como lo reporta el Excel.

Ejemplo 2

Tome en cuenta el enunciado del ejemplo 1 y determine:

a) La ecuación que nos permita predecir el salario mensual obtenido por un administrador en base a su experiencia. Interprete los coeficientes del modelo
b) El salario de un administrador que tiene 18 años de experiencia
c) El coeficiente de correlación. Interprete este valor
d) El coeficiente de determinación. Interprete este valor
e) La validez del modelo obtenido en la parte a). Use $\alpha = 0,05$

Solución:

Para la solución nos apoyamos en Microsoft Excel
Herramientas > Análisis de datos > Regresión >

Obtenemos el siguiente reporte de Excel:

Resumen	
Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.93988
Coeficiente de determinación R^2	0.88338
R^2 ajustado	0.87505
Error típico	4.32505
Observaciones	16

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	1	1983.8	1983.8	106.1	6E-08
Residuos	14	261.884	18.706		
Total	15	2245.68			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*
Intercepción	28.0619	1.97079	14.239	1E-09	23.84	32.29
Exper.	1.16021	0.11266	10.298	6E-08	0.919	1.402

a) Ecuación: $\hat{y}_i = 28,0619 + 1,1602x_i$

Interpretación de los coeficientes del modelo

$b_0 = \hat{\beta}_0 = 28,0619$: Cuando el administrador no tiene experiencia, el valor de su salario es de aproximadamente 2 806,19 soles mensuales.

$b_1 = \hat{\beta}_1 = 1,1602$: Ante el incremento de un año de experiencia del administrador el salario se ve incrementado en 116,02 soles mensuales aproximadamente.

b) Cuando el administrador tiene 18 años de experiencia su salario aproximado será de $\hat{y}_0 = 28{,}0619 + 1{,}1602(18) = 48{,}9455$ (aproximadamente 4 894,55 soles mensuales aproximadamente)

c) Coeficiente de correlación: 0,93988
Existe una alta correlación de carácter directa entre los años de experiencia del administrador y el salario que percibe

Coeficiente de correlación múltiple	0.93988

d) Coeficiente de correlación: 0,88338
Aproximadamente el 88,34 % de la variación que sufre el salario mensual de un administrador es explicado por sus años de experiencia

Coeficiente de determinación R^2	0.88338

e) Validación del modelo

ANÁLISIS DE VARIANZA					
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	1	1983.8	1983.8	106.1	6E-08
Residuos	14	261.884	18.706		
Total	15	2245.68			

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

El valor de *F* con 1 y 14 grados de libertad es: 4,6 menor que 106,1; por lo tanto se rechaza la hipótesis nula.
Bajo un nivel de significación del 5% podemos afirmar que existe una relación funcional lineal entre los años de experiencia y el salario del administrador.

Observación:
P valor = 0,000006 < 0,05.
Haciendo uso del criterio del de P valor, también llegamos a la misma conclusión que se estableció con el método clásico.

Ejemplo 3

El gerente de operaciones de una empresa aérea desea saber la cantidad de agua (en litros) que deben llevar los aviones en cada uno de sus vuelos. Esto se debe a que si se lleva poca agua los servicios que la requieren podrían no funcionar de manera óptima y si ésta se lleva en exceso implica indirectamente mayor uso de combustible. El gerente cree que una de las variables que puede afectar la cantidad de agua necesaria en los vuelos es el número de pasajeros en el avión. Para despejar su duda registra información de ambas variables. Los resultados se muestran a continuación.

Cantidad de agua	Cantidad de pasajeros
91.7	80
91.8	82
93.2	82
97.7	85
97.8	85
99.2	86
99.9	87
101.5	87
101.7	90
101.8	93
104.8	93
105.2	95
105.6	95
107.0	97
107.7	98
108.5	98

a. Halle la ecuación de regresión lineal. Interprete el coeficiente de regresión del modelo estimado en términos del enunciado del problema.

b. Interpretar el coeficiente de determinación del modelo.

c. Calcular e interpretar el coeficiente de correlación del modelo.

d. Validar el modelo de regresión lineal. Use un nivel de significación del 6%.

e. ¿Se puede afirmar con 8% de nivel de significación que por cada pasajero adicional la cantidad de agua necesaria por vuelo aumenta en más de 0,85 litros?

f. Estimar la cantidad de agua necesaria para un vuelo con 100 pasajeros.

Ejercicios Propuestos:

1. Un economista del Departamento de Recursos Humanos del Estado de Florida está preparando un estudio sobre el comportamiento del consumidor. Él recolectó los datos que aparecen en miles de dólares para determinar si existe una relación entre el ingreso del consumidor y los niveles de consumo. Determine cuál es la variable dependiente.

Consumidor	1	2	3	4	5	6	7	8	9	10	11
Ingreso	24,3	12,5	31,2	28	35,1	10,5	23,2	10	8,5	15,9	14,7
Consumo	16,2	8,5	15	17	24,2	11,2	15	7,1	3,5	11,5	10,7

a. Elabore el diagrama de dispersión para los datos.
b. Calcule e interprete el modelo de regresión. ¿Qué le dice este modelo sobre la relación entre el consumo y el ingreso? ¿Qué proporción de cada dólar adicional que se gana se invierte en consumo?

2. Se reúne datos acerca de la cantidad de familias que ven programas donde se pasan determinado anuncio. Esta observación es de utilidad para los publicistas, por que les dice a cuantos consumidores llegan. Los datos siguientes muestran la cantidad de familias espectadoras, en miles, y la cantidad de veces que salió al aire el anuncio en una semana.

Veces que salió el anuncio	95	46	41	38	29	32	25	21	21	16
Familias espectadoras	758,8	323,0	275,3	241,8	219,9	198,5	193,8	189,7	161,9	160,0

a) Determine la ecuación de regresión para predecir la cantidad de familias espectadoras a partir de la cantidad de veces que apareció el anuncio. Suponga que existe una relación lineal entre la cantidad de familias espectadoras y la cantidad de veces que apareció el anuncio
b) Interprete los coeficientes de regresión de la ecuación de la parte a)
c) Validar el modelo de la parte a). Use un nivel de significación del 5%
d) ¿Se puede afirmar que por cada anuncio adicional que se realice el número de familias espectadoras se incrementa en 6 000? Use un nivel de significación del 5%
e) ¿Qué porcentaje de la variabilidad total de número de familias espectadoras es explicado por el modelo determinado en la parte a)?

3. En un estudio acerca de la talla, medida en centímetros y el peso, medido en kilogramos, de un grupo de 10 personas, se ha recolectado los siguientes valores:

TALLA (cm.)	160	165	168	170	171	175	175	180	180	182
PESO (Kg.)	55	58	58	61	67	62	66	74	79	83

a) Identifique la variable independiente y la variable dependiente
b) Elabore el diagrama de dispersión.
c) Suponga que existe una relación funcional lineal entre la talla y el peso y escriba la ecuación de regresión lineal simple, en correspondencia a la pregunta 1.
d) Interprete los coeficientes de la línea de regresión estimada en la pregunta anterior.
e) ¿Existe relación funcional lineal entre el peso y la talla de las personas? Use un nivel de significación del 5%
f) ¿Se puede afirmar que por cada centímetro adicional en la talla de la persona, los pesos se incrementan en un kilogramo? Use un nivel de significación del 5%
g) Calcule e interprete el coeficiente de determinación
h) Calcule e interprete el coeficiente de correlación
i) ¿En cuanto se estima el peso de una persona, cuando su talla es de 162 centímetros?

4. Los siguientes datos corresponden al cloro residual en una piscina en diversos momentos después de haberse tratado con químicos.

Número de horas	**Cloro residual (partes por millón)**
2	1,8
4	1,5
6	1,4
8	1,1
10	1,1
12	0,9

a) Estime la recta por el método de mínimos cuadrados. Interprete sus coeficientes.
b) Calcule e interprete el coeficiente de determinación.
c) Verifique la existencia de la pendiente del modelo. $\alpha = 0{,}05$.

5. El propietario de una empresa de mudanzas desea desarrollar un modelo de regresión que le permita estimar el tiempo total (horas) empleado para realizar una mudanza en función de la carga transportada (pies cúbicos). Para lograr su propósito recolectó la información que a usted se muestra a continuación.

Y: Horas	24	13.5	26.3	25	9	20	22	11.25	50	12	38.75	40	19.5	18	28
X: Piés cúbicos	545	400	562	540	220	344	569	340	900	285	865	831	344	360	750

a. Determine la ecuación de regresión lineal simple e interprete sus coeficientes.
b. Valide el modelo de regresión usando un nivel de significación del 5%.
c. Interprete el coeficiente de determinación y correlación.
d. ¿Se puede afirmar que por cada pie cúbico adicional en la carga transportada, el tiempo total para realizar la mudanza aumenta en menos de 0.07 horas? Use un nivel de significación del 5%.
e. Estimar el tiempo medio total empleado para realizar una mudanza cuando el transporte tiene una carga de 400 pies cúbicos.

5.2 Regresión curvilineal

Se ha visto que los modelos lineales son útiles en muchas situaciones y aunque la relación entre la variable respuesta y las variables regresoras no sea lineal, en muchos casos, la relación es *"linealizable"* en el sentido de que transformando (tomar logaritmos, calcular la inversa,...) la variable respuesta y/o algunas variables regresoras la relación se vuelve lineal. Sin embargo, existen situaciones en que la relación no es lineal y tampoco es linealizable, por ejemplo, si el modelo de regresión es el siguiente:

$$y_i = \alpha \cdot e^{\sqrt{\beta x_i} + \gamma x_i^2} + \varepsilon_i .$$

En esta sección veremos algunos modelos *linealizables*.

La transformación de datos nos permite *linealizar* la relación entre dos variables, esto se realiza cuando se sospecha (puede ser gráficamente) que no existe dependencia lineal entre las variables en estudio. Las transformaciones que pueden mejorar el ajuste y la capacidad de predicción del modelo son muy numerosas. Aquí se presenta algunas de las trasformaciones.

Forma funcional que relaciona *y* con *x*	**Transformación apropiada**	**Forma de regresión lineal simple**
Exponencial : $y = \beta_0 e^{\beta_1 x}$	$y^* = \ln y$	Regresión de y^* vs x
Potencia: $y = \beta_0 x^{\beta_1}$	$y^* = \ln y;$ $\quad x^* = \ln x$	Regresión de y^* vs x^*
Polinomial: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$		Regresión de y vs x, x^2

Según lo observado en el diagrama de dispersión se usará alguna de estas funciones para luego verificar lo adecuado de la elección.

Pocedimiento para la selección del mejor modelo:

1. Hallar el coeficiente de determinación R^2 de los modelos lineal, cuadrático, exponencial y potencia.
2. Ordenarlos de mayor a menor según su R^2. Esto nos permite priorizar el análisis de los mdelos.
3. Realizar el análisis del modelo que tenga el mayor R^2, verificar si su coeficiente de regresión es significativamente diferente de cero.

4. Si no se demuestra que el coeficiente de regresión modelo que tiene mayor R^2 es significativamente diferente de cero, se debe pasar a evaluar el siguiente modelo con mayor R^2, hasta encontrar un modelo cuyo coeficiente sea significativamente diferente de cero.
   Nota: Solo en el modelo polinomial analizaremos la significancia de β_2

Ejemplo:

Los siguientes datos representan el porcentaje usable de cierto tipo de neumáticos radiales de alto rendimiento después de haber sido empleados el número de millas:

Millas conducidas (en miles) x	Porcentaje usable y
1	98,2
2	91,7
5	81,3
10	64
20	36,4
30	32,6
40	17,1

a. Estime la mejor ecuación para el conjunto de datos.
b. Compruebe la existencia de modelo. Use nivel de significación 0.05.
c. Pronostique con 95% de confianza el porcentaje usable de los neumáticos, luego se recorrer 25000 millas.

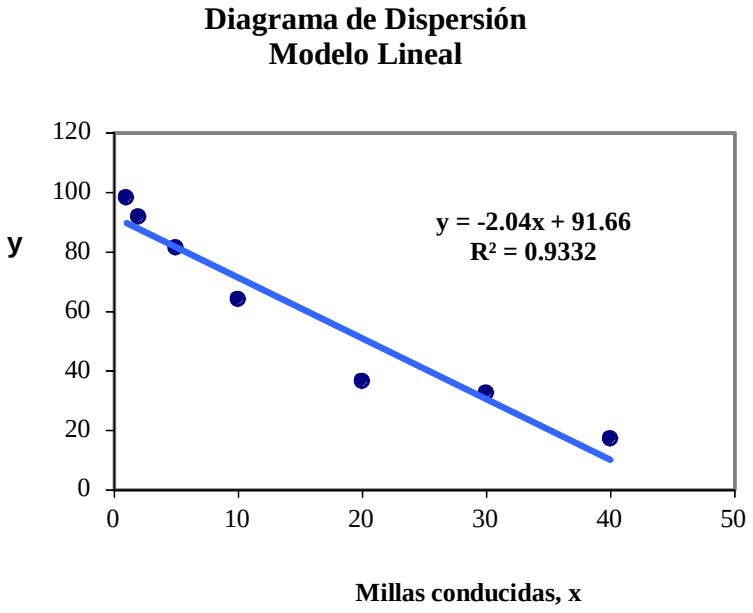
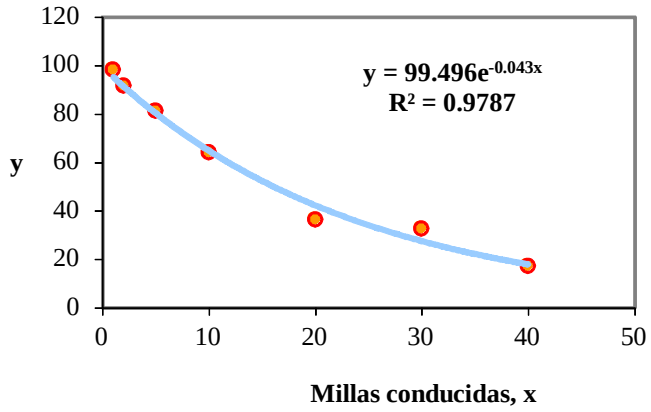


Diagrama de Dispersión
Modelo Exponencial



Resumen del modelo Exponencial	
Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.989301
Coeficiente de determinación R^2	0.9787165
R^2 ajustado	0.9744598
Error típico	0.1041876
Observaciones	7

ANÁLISIS DE VARIANZA

	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	F	Valor crítico de F
Regresión	1	2.4958	2.4958	229.9241	0.0000
Residuos	5	0.0543	0.0109		
Total	6	2.5501			

	Coeficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%
Intercepción	4.6001	0.0587	78.3686	0.0000	4.4492	4.7510
Millas conducidas (miles)	-0.0428	0.0028	-15.1632	0.0000	-0.0500	-0.0355

NOTAS:

Ejemplo:

La tabla que se muestra a continuación registra el número de días que han transcurrido desde que se ha detectado un nuevo virus informático y el número de ordenadores infectados en un país de la comunidad europea.

Número de días	Nro Ordenadores infectados (miles)
1	98.2
2	91.7
4.9	95
4.7	84.6
5	81.3
4	70.4
8	60.5
10	64
20	36.4
30	35
40	26.8

a. Escriba los modelos posibles que permitan estimar el número de ordenadores infectados. Indique además el término que hace posible la detección del primer modelo a ser analizado.

b. Determine la mejor ecuación de regresión para determinar el número de ordenadores infectados a partir del número de días. Use $\alpha = 0{,}01$.

c. Utilice el modelo elegido para estimar el número de ordenadores infectados cuando ha transcurrido 15 días.

NOTAS:

Ejercicio Propuesto:

Una empresa dedicada a la elaboración y distribución de alimentos para perros realizó un estudio para relacionar la utilidad neta por la venta de alimentos para perros con el costo de transporte para un período de 10 años. Los datos obtenidos (en miles de dólares), se presentan a continuación:

Costo de transporte	5	5.6	6.1	6.8	7.4	8.6	9.5	9.8	10.4	11.3
Utilidad Neta	13.2	16.8	20.9	28.8	35.4	42.5	52.6	55.5	66.9	84.1

a. Basados en los datos presentados determine el mejor modelo de regresión. Justifique su respuesta. Considere $\alpha = 0,05$.
b. Estime la utilidad neta si el costo de transporte es de 12 500 dólares.

NOTAS:

5.3 Regresión Lineal múltiple

El análisis de regresión múltiple es el estudio de la forma en que una variable dependiente se relaciona con dos o más variables independientes.
El número de variables independientes se indicará con la letra p.
(El total de parámetros es p+1)

El modelo de regresión múltiple tiene la forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

donde:

y	: variable respuesta que se quiere predecir
$\beta_0, \beta_1, ., \beta_p$	: parámetros del modelo
$x_1, x_2, ..., x_p$	: variables independientes que se miden sin error.
ε	: es un error aleatorio

Ecuación de regresión múltiple estimada

Para estimar la ecuación de regresión anterior, se toma una muestra aleatoria y a partir de ella se estima los parámetros del modelo. Las ecuaciones de los parámetros estimados se obtienen usando el método de Mínimos Cuadrados Ordinarios.
Al igual que para el caso de regresión lineal simple, el método se fundamente en minimizar:

$$\min \sum (y_i - \hat{y}_i)^2$$

donde:

y_i : valor observado de la variable dependiente en la i-ésima observación.

$\hat{y}_i$: valor estimado de la variable dependiente en la i-ésima observación.

Nº	y_i	x_{1i}	x_{2i}	…	x_{pi}
1	y_1	x_{11}	x_{21}	…	x_{p1}
2	y_2	x_{12}	x_{22}	…	x_{p2}
3	y_3	x_{13}	x_{23}	…	x_{p3}
.					
.					
.					
n	y_n	x_{1n}	x_{2n}	…	x_{pn}

El modelo estimado tiene la forma:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p$$

donde:

$\hat{y}$: valor estimado de la variable dependiente

$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \beta_p$: estimaciones muestrales de los parámetros poblacionales

$x_1, x_2, ..., x_p$: son variables predictoras

Supuestos del modelo de regresión

- El error ε es una variable aleatoria cuyo valor medio o esperado es cero.
- La varianza de ε representada por σ^2 es igual para todos los valores de las variables dependientes x1, x2, …, x_p
- Los valores de ε son independientes.
- El error ε es una variable aleatoria con distribución normal.
- Las variables independientes no están correlacionadas (no multicolinealidad)

Coeficiente de regresión

Los valores $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$ se conocen como coeficientes de regresión estimados.

Un coeficiente de regresión estimado específico mide el cambio promedio en la variable dependiente debido a un incremento de una unidad en la variable predictora relevante, manteniendo constantes las otras variables de predicción.

El error estándar de la estimación

El error estándar de la estimación mide la variabilidad, o dispersión, de los valores muestrales y observados alrededor del plano de regresión.

$$s = \sqrt{\frac{SCE}{n-p}} = \sqrt{CME}$$

donde: p es el número de parámetros a estimar.

Coeficiente de determinación múltiple (R^2)

El coeficiente de regresión múltiple mide el porcentaje de la variabilidad de, y, que se puede explicar mediante las variables de predicción.

Un valor de R^2 cercano a uno significa que la ecuación es muy exacta porque explica una gran porción de la variabilidad de y. Se define como:

$$R^2 = \frac{SC\,\mathrm{Re}\,g}{SCTotal} \times 100\%$$

Si embargo si se introducen excesivas variables al modelo el coeficiente de determinación incrementará su valor, por tal razón se suele calcular el coeficiente de determinación ajustado:

$$R^2 ajustado = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

Pruebas de hipótesis

Una vez que se ha recogido una muestra aleatoria se han medido las variables y se ha examinado la matriz de correlación para determinar aquellas combinaciones de variables que son de interés, se analizan los modelos con el mejor potencial. El objetivo es encontrar la mejor ecuación para predecir y después decidir si ésta ecuación satisface las necesidades de exactitud del analista.

Los valores t calculados son de particular importancia en la regresión múltiple porque constituyen la forma principal de detectar multicolinealidad. Si son suficientemente grandes, la correlación entre las dos variables predictoras no es un problema. Si uno o ambos valores t son menores que los valores t de tablas, la multicolinealidad está presente.

Pruebas individuales

Las hipótesis planteada y alternante para las pruebas individuales son:

$H_0 : \beta_i = 0$

$H_1 : \beta_i \neq 0$

Y el estadístico de prueba es:

$$T = \frac{\hat{\beta}_i - \beta_i}{ET\left[\hat{\beta}_i\right]} \sim t_{(n-p-1)}$$

Prueba conjunta

Las hipótesis planteada y alternante para la prueba conjunta son:

$H_o : \beta_1 = \beta_2 = ... = \beta_p = 0$

$H_1 : Al\ menos\ un\ \beta_i\ es\ diferente\ de\ 0$

Y el estadístico de prueba es:

$$F = \frac{CM\ \mathrm{Re}\, g}{CMError} \sim F_{(p,\, n-p-1)}$$

Multicolinealidad

Cuando existe multicolinealidad es difícil distinguir que cantidad del efecto observado se debe a una variable de predicción individual. En otras palabras, si dos variables están altamente correlacionadas, proporcionan casi la misma información en el pronóstico.

Cuando dos variables tienen una alta correlación, los coeficientes $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$ estimadores de $\beta_0, \beta_1, \beta_2, ...\beta_k$ no son confiables. La estimación $\hat{\beta}_k$ de β_k puede no ser siquiera cercana al valor del su correspondiente parámetro y en casos extremos puede incluso ser negativo cuando debiera ser positivo.

Regla práctica para seleccionar las variables predictoras en regresión múltiple

- Una variable independiente (X) **debe** tener una correlación fuerte con la variable dependiente (Y).
- Una variable independiente **no debe** tener una correlación demasiado alta con ninguna otra variable independiente. (La correlación entre dos variables predictoras debe estar muy por debajo de la menor de las dos correlaciones entre las variables predictoras y la variable dependiente)
- Cuando se produce la multicolinealidad, si el analista sólo quiere usar el modelo de regresión para hacer pronósticos, la multicolinealidad puede no causar ninguna dificultad seria.

A continuación se presenta una matriz de correlaciones y se desea verificar la multicolinealidad:

	Y	X1	X2	X3
Y	1			
X1	0.96190499	1		
X2	0.95134732	0.96230485	1	
X3	0.95129052	0.94478077	0.98132314	1

$| Corr(X_1, X_2) | = 0.9623$

$| Corr(Y, X_1) | = 0.9619$, $| Corr(Y, X_2) | = 0.9514 \Rightarrow$ Mínima = 0.9514

Existe multicolinealidad entre X1 y X2.

$| Corr(X_1, X_3) | = 0.9448$

$| Corr(Y, X_1) | = 0.9619$, $| Corr(Y, X_3) | = 0.9513 \Rightarrow$ Mínima = 0.9513

No existe multicolinealidad entre X1 y X2.

$| Corr(X_2, X_3) | = 0.9813$

$| Corr(Y, X_2) | = 0.9513$, $| Corr(Y, X_3) | = 0.951291 \Rightarrow$ Mínima = 0.9513

Existe multicolinealidad entre X2 y X3.

Como se detecta multicolinealidad entre variables, se concluye:

X1 y X2 no pueden ingresar juntos al modelo
X2 y X3 no pueden ingresar juntos al modelo

Las consecuencias adversas son:

- Las estimaciones de los coeficientes de regresión fluctúan de manera notoria de una muestra a otra.
- Una variable independiente que tiene una relación positiva con la variable dependiente puede producir un coeficiente de regresión negativo si la correlación con otra variable independiente es alta.
- Con frecuencia se usa la regresión múltiple como una herramienta interpretativa para evaluar la importancia relativa de las distintas variables independientes. Cuando las variables independientes se intercorrelacionan, explican la misma varianza en el pronóstico de la variable dependiente. por esto, es difícil separar la influencia individual de cada variable independiente cuando la multicolinealidad está presente.

Procedimiento para la selección del mejor modelo de regresión

- Analizar la multicolinealidad y descartar aquellos modelos donde existan problemas de multicolinealidad.
- Ordenar los modelos según su R^2 ajustado.
- Evaluar la significancia del (los) coeficiente(s) de regresión.
- Si se demuestra que el(los) coeficientes es (son) significativamente diferentes de cero entonces ese será el mejor modelo, caso contrario se debe pasar a evaluar el siguiente modelo que tenga mayor R2 ajustado hasta encontrar un modelo con coeficientes significativamente diferentes de cero.

Ejemplo:

Una empresa que vende por correo suministros para computadoras personales, software y hardware posee un almacén central para la distribución de los productos ordenados. La administración se encuentra examinando el proceso de distribución desde el almacén y está interesada en estudiar los factores que afectan los costos de distribución del almacén.
Actualmente, un pequeño cargo por manejo se agrega a pedido, independiente de la cantidad por la que se hizo. Se han recolectado datos correspondientes a los 24 meses anteriores y respecto a los costos de distribución del almacén, las ventas y el número de pedidos recibidos.

Costos de distribución (miles de $) (y)
Ventas (miles de $) ($x_1$)
Número de pedidos (x_2)
Tiempo de transporte (x_3)

Los datos del estudio se muestran en la tabla siguiente:

Mes	Ventas	Nº pedidos	Tiempo	Costo
1	386	4015	44	52.95
2	446	3806	63	71.66
3	512	5309	59	85.58
4	401	4262	62	63.69
5	457	4296	55	72.81
6	458	4097	49	68.44
7	350	3213	45	52.46
8	484	4809	52	70.77
9	517	5237	70	82.03
10	503	4732	50	74.39
11	535	4413	55	70.84
12	440	2921	50	54.08
13	372	3977	55	62.98
14	480	4428	58	72.30
15	408	3964	54	58.99
16	491	4582	65	79.38
17	527	5582	64	94.44
18	444	3450	60	59.74
19	515	5079	60	90.5
20	596	5735	69	98.00
21	463	4269	50	69.33
22	389	3708	48	53.71
23	547	5387	66	89.18
24	415	4161	51	62.98

Estime un modelo de regresión para estudiar las ventas semanales.

Matriz de correlaciones

	Ventas	Nº pedidos	Costo	Tiempo
Ventas	1			
Nº pedidos	0,8120	1		
Costo	0,8868	0,9191	1	
Tiempo	0,6512	0,6412	0,7543	1

Resumen	
Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.9620772
Coeficiente de determinación R^2	0.9255925
R^2 ajustado	0.9144314
Error típico	3.9177821
Observaciones	24

ANÁLISIS DE VARIANZA

	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	F	Valor crítico de F
Regresión	3	3818.6833	1272.8944	82.9300	0.0000
Residuos	20	306.9803	15.3490		
Total	23	4125.6637			

	Coeficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%
Intercepción	-24.3719	6.8991	-3.5326	0.0021	-38.7631	-9.9807
Ventas	0.0722	0.0237	3.0525	0.0063	0.0229	0.1215
Nº pedidos	0.0094	0.0020	4.7757	0.0001	0.0053	0.0135
Tiempo	0.3720	0.1498	2.4835	0.0220	0.0595	0.6844

NOTAS:

Ejemplo:

El director de Asuntos Académicos de una universidad, está interesado en determinar la dependencia de las notas del curso de postgrado en función de las notas del examen de Ingreso, el número de clases perdidas en el curso durante el ciclo y las horas de estudio que los estudiantes le dedican al curso durante la semana. Para ello toma una muestra aleatoria de alumnos, la cual se muestra a continuación.

Calificación en curso de postgrado	Calificación en el Examen de Ingreso X1	Clases pérdidas durante el ciclo X2	Horas de estudio a la semana X3
17	13	4	5.5
12	11	6	1.5
12	11	7	2
18	13	2	6
14	11	8	4
16	14	3	5
14	13	3	3.5
18	14	2	5.5
12	11	6	2
18	14	2	6
15	15	1	4.5
13	11	4	2
18	15	2	4.5
18	17	2	5
19	17	5	5.5
15	15	2	3.5
10	13	6	1
19	17	1	5.5
15	15	3	4.5
10	13	6	1

a. Determine si el modelo que relaciona la variable dependiente con todas las independientes propuestas presenta problemas de multicolinealidad? De existir indique entre qué variables se presenta este problema. Sustente su respuesta indicando los valores correspondientes.

b. Considerando la respuesta a la pregunta anterior, escriba los posibles modelos en función de Y con X1, X2, X3.

Modelo	R^2-Ajustado	Prioridad
		1
		2
		3

		4
		5
		6
		7

c. Estime, valide e interprete los coeficientes del mejor modelo que se ajuste a los datos con un nivel de significación del 5%.

d. Pronostique la calificación en el curso de postgrado, para un alumno que obtuvo 15 en el examen de ingreso, tuvo 2 clases perdidas en el curso y le dedica 2 horas a la semana de estudio.

Ejercicios Propuestos:

Para desarrollar las preguntas 1 y 2 considere el siguiente enunciado:

Don Pizzas" es una pizzería de propiedad de Jorge Montoya (JM). En los últimos años el negocio ha ido derivando hacia el Delivery donde obtiene la mayor captación de sus ventas. Las zonas que abarcan para la repartición de los pedidos son A, B y C donde la zona A es la más cerca y C la más alejada. Recientemente, una cadena muy famosa de pizzerías, colocó una pizzería al frente de **Don Pizzas**. Muy pronto, fue evidente que esta nueva pizzería le quitaba clientes a **Don Pizzas**, entonces Jorge comprendió que debía elaborar nuevas estrategias por lo que necesitaba levantar información de las entregas de sus pedidos.

Debido a la gran cantidad de entregas realizadas cada noche, Jorge sabía que no podía vigilar cada una, por lo que decidió tomar una muestra aleatoria de las entregas durante cierto periodo y tomar el mismo las mediciones. El período a considerar es un mes y sólo los viernes, sábados y domingos. Cada día elegía al azar un pedido telefónico, después medía cuidadosamente el tiempo requerido para **preparar** el pedido y el tiempo que éste **esperaba** a que estuviera disponible un repartidor (Manuel, Carlos y Esteban). Luego, Jorge medía con cuidado el tiempo que **demoraban** entre salir de la pizzería y entregar la pizza. Después de regresar, seleccionaba al azar otro pedido y repetía el proceso.

Las variables recolectadas están definidas de la siguiente manera:

Dia: El día de la semana (1 = Viernes, 2 = Sábado, 3 = Domingo)

T_prep: El tiempo requerido (en minutos) para preparar el pedido.

T_esp: El tiempo (en minutos) desde que termina la preparación del pedido hasta que un repartidor está disponible para entregarlo.

T_viaje: El tiempo (en minutos) que tarda el vehículo en llegar al punto de entrega.

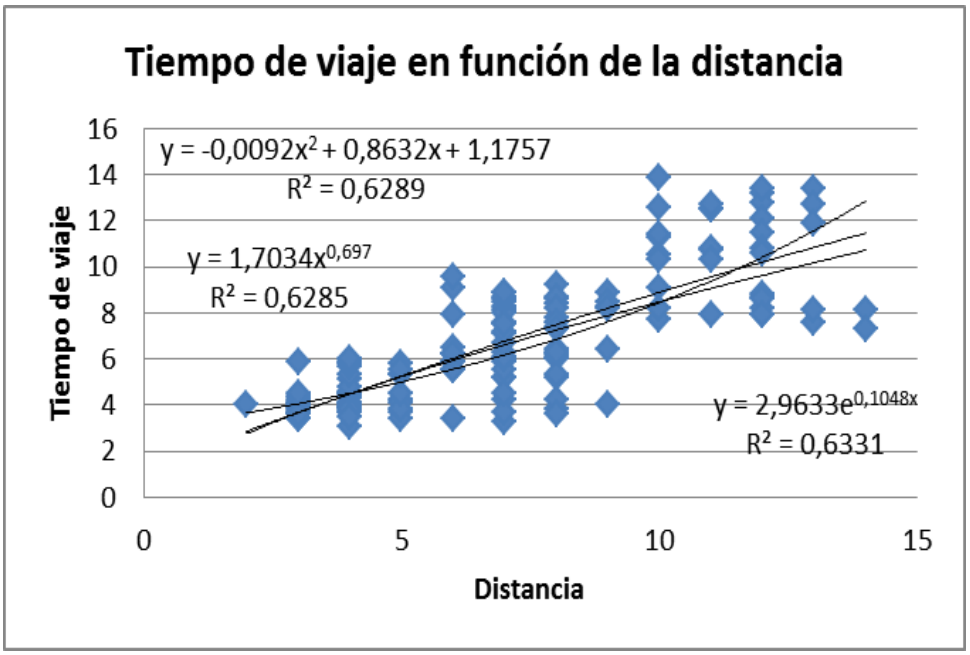
Distancia: La distancia (número de cuadras) de Don Pizzas al punto de entrega.

Zona: La zona donde se debe llevar la pizza (A, B y C)

Repartidor: La persona que realiza la entrega de la pizza al domicilio (Manuel, Carlos y Esteban)

1) Jorge está pensando ofrecer su pizza a la zona D que está en promedio a 15 cuadras de la pizzería. Estima que en promedio el tiempo de preparación de la pizza es de 15 min y el tiempo de espera para que el repartidor esté disponible es 5 min; pero el tiempo de viaje (T_viaje) está en función del número de cuadras y supone que esta relación es lineal.

   Estime el Tiempo de viaje (T_viaje) para entrega del pedido en la zona D, cuando se asume una regresión no lineal. Presente el procedimiento de elección del modelo con su validación correspondiente. Use α=0.05
   Utilice los reportes del Excel que se muestran a continuación.



Exponencial

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*
Intercepción	1.0861	0.0511	21.2451	1.1725E-46
distancia	0.1048	0.0066	15.9218	8.3448E-34

Cuadrático

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*
Intercepción	1.1757	0.8309	1.4150	0.1592
distancia	0.8632	0.2355	3.6650	0.0003
distancia^2	-0.0092	0.0149	-0.6160	0.5388

Potencial

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*
Intercepción	0.5327	0.0851	6.2610	3.979E-09
LNX	0.6970	0.0442	15.7568	2.2003E-33

2) Un especialista en estadística, contratado por JM, que tiene la tarea de estimar el tiempo de la entrega total Y=T_total (preparación+espera+viaje), considera que las variables que influyen para este tiempo son:

X1= Distancia (cuadras)
X2= Calificación del repartidor (1 a 10, donde 10 el mejor calificativo)
X3= Experiencia del repartidor (meses).

A continuación se muestran los reportes obtenidos en Excel. Con toda esta información realice el análisis necesario para *estimar el tiempo total de la entrega* cuando un pedido se

tiene que llevar a 15 cuadras, el calificativo del vendedor es de 8 y su experiencia es de 10 meses.

	Y	*X1*	*X2*	*X3*
Y	1			
X1	0.6108	1		
X2	-0.4209	-0.0742	1	
X3	0.0463	0.0435	0.1007	1

Modelo: YX1X2X3

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.7201
Coeficiente de determinación R^2	0.5185
R^2 ajustado	0.5086
Error típico	2.4937
Observaciones	150

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*
Intercepción	30.0676	1.3623	22.0715	2.3059E-48
X1	0.6763	0.0673	10.0539	2.1347E-18
X2	-1.1266	0.1699	-6.6311	6.041E-10
X3	0.0704	0.0681	1.0331	0.30325885

Modelo: YX1X2

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.7176
Coeficiente de determinación R^2	0.5150
R^2 ajustado	0.5084
Error típico	2.4943
Observaciones	150

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*
Intercepción	30.3988	1.3243	22.9540	1.8487E-50
X1	0.6799	0.0672	10.1180	1.3678E-18
X2	-1.1083	0.1690	-6.5576	8.7063E-10

Modelo: YX1X3

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.6111
Coeficiente de determinación R^2	0.3735
R^2 ajustado	0.3650
Error típico	2.8348
Observaciones	150

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*
Intercepción	22.2610	0.7792	28.5684	7.0651E-62
X1	0.7116	0.0762	9.3341	1.4817E-16
X3	0.0233	0.0770	0.3021	0.76300185

Modelo: YX2X3

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.4303
Coeficiente de determinación R^2	0.1851
R^2 ajustado	0.1740
Error típico	3.2330
Observaciones	150

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*
Intercepción	35.6072	1.6153	22.0444	1.8661E-48
X2	-1.2616	0.2196	-5.7454	5.0978E-08
X3	0.1055	0.0882	1.1966	0.2334021

Modelo: YX1

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.6108
Coeficiente de determinación R^2	0.3731
R^2 ajustado	0.3688
Error típico	2.8261
Observaciones	150

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*
Intercepción	22.4141	0.5901	37.9827	3.3044E-78
X1	0.7126	0.0759	9.3849	1.0443E-16

Modelo: YX2

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.4209
Coeficiente de determinación R^2	0.1772
R^2 ajustado	0.1716
Error típico	3.2377
Observaciones	150

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*
Intercepción	36.1491	1.5527	23.2810	2.4656E-51
X2	-1.2351	0.2188	-5.6454	8.1591E-08

Modelo: YX3

Estadísticas de la regresión	
Coeficiente de correlación múltipl	0.0463
Coeficiente de determinación R^2	0.0021
R^2 ajustado	-0.0046
Error típico	3.5655
Observaciones	150

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*
Intercepción	27.1358	0.7273	37.3078	3.6572E-77
X3	0.0545	0.0968	0.5633	0.5740618

3) A doce piezas de acero reducido en frío con contenidos diferentes de cobre y diferentes temperaturas de recocido se les mide su dureza con los siguientes resultados:
Ajuste una ecuación de la forma $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$, donde x_1 representa el contenido de cobre, x_2 representa la temperatura de recocido e y representa la dureza. Luego, analice el modelo.

Dureza (Rockwell 30-T)	**Contenido de cobre (%)**	**Temperatura del recocido (grados F)**
78.9	.02	1000
65.1	.02	1100
55.2	.02	1200
56.4	.02	1300
80.9	.10	1000
69.7	.10	1100
57.4	.10	1200
55.4	.10	1300
85.3	.18	1000
71.8	.18	1100
60.7	.18	1200
58.9	.18	1300

4) Los datos siguientes presentan el peso, potencia y velocidad al cuarto de milla de doce automóviles deportivos. Suponga que también se conoce el precio de cada vehículo y que el conjunto completo de datos es el siguiente:

Automóvil deportivo	Precio (miles de dólares)	Peso (libras)	Potencia (HP)	Velocidad al cuarto de milla (mi/h)
AUD01	23200	2450	225	91,7
AUD02	24865	2650	305	80,3
AUD03	25035	2577	195	90,7
AUD04	26382	3042	195	89,7
AUD05	40900	2844	189	93,2
AUD06	50144	3246	345	102,1

AUD07	69742	3319	410	116,2
AUD08	93758	3570	305	140,0
AUD09	25035	3520	190	88,0
AUD10	26382	3042	199	91,3
AUD11	40900	2844	189	93,2
AUD12	50200	3500	300	100,2

a. El modelo presenta problemas de multicolinealidad. Sustente.
b. Estime y analice el mejor modelo que ajuste a los datos con un nivel de significación del 5%.
c. Pronostique e interprete con 95% cuál será el precio promedio de un automóvil cuyo peso es 3058 libras, 280 HP y velocidad al cuarto de milla de 115.

5) Use un programa apropiado de computadora para ajustar un plano a los datos siguientes relativos al uso mensual del agua de una planta de producción (miles de galones) a su producción mensual (toneladas), la media de la temperatura ambiental mensual (°F), y el número mensual de días de operación de la planta durante de la planta durante un periodo de 12 meses.

Uso de agua	Producción	Media de la temperatura	Días de operación
2,228	98,5	67,4	19
2,609	108,2	70,3	20
2,678	109,6	82,1	21
2,378	101,0	69,2	21
2,035	83,3	64,5	19
2,124	70,0	63,7	21
2,875	141,1	58,0	22
2,031	84,4	58,1	20
1,984	97,4	36,6	17
3,125	131,8	49,6	23
2,254	82,1	44,3	18
1,645	55,6	44,1	14

a. Estime el uso de agua de la planta durante un mes cuando su producción es 90.0 toneladas, la media de la temperatura es 65°F, y opera por 20 días.
b. El modelo presenta problemas de multicolinealidad. Sustente.
c. Estime y analice el mejor modelo que ajuste a los datos con un nivel de significación del 5%.
d. Estime con 95% de confianza el uso de agua de la planta durante un mes cuando su producción es 91,2 toneladas, la media de la temperatura es 67°F, y opera por 21 días.

6) Suponga que desea desarrollar un modelo para predecir el precio de casas unifamiliares de acuerdo con el área que tiene calefacción, la antigüedad de la casa y el tamaño del lote. Se selecciona una muestra de 15 casas unifamiliares. Se registraron la valuación (en miles de dólares), el área de las casas que tiene calefacción (en miles de pies cuadrados), la antigüedad de las casas (en años) y el tamaño del lote (miles de pies cuadrados) con los siguientes resultados:

Casa	Precio (miles de dólares)	Área con calefacción (miles de pie^2)	Edad (años)	Tamaño del lote (miles de pie^2)
1	70,40	1,60	32,00	2,50
2	79,30	1,39	1,00	1,80
3	75,70	1,45	8,33	1,50
4	79,20	1,50	2,75	2,30
5	74,50	1,54	12,58	1,80
6	75,80	1,55	16,00	2,30
7	78,50	1,59	1,75	1,80
8	76,80	1,59	7,17	1,80
9	77,40	1,71	11,50	2,50
10	85,90	1,76	0,00	1,95
11	84,40	1,85	3,42	3,00
12	83,80	1,89	2,75	2,05
13	86,70	1,90	0,00	2,50
14	79,10	1,93	7,42	2,65
15	85,90	1,93	2,00	3,00

a. Estime el modelo lineal con todas las variables independientes, ¿qué porcentaje de la variabilidad en la valuación de las casas es explicado por el modelo?, ¿este modelo es significativo? Use $\alpha = 0{,}05$.
b. De incluir todas las variables en el modelo para estimar la valuación de la casa, ¿este modelo presentará problemas de multicolinealidad?, ¿qué propone para remediar esto?
c. Estime el mejor modelo para pronosticar la valuación de las casas unifamiliares. Analícelo con 5% de nivel de significación.
d. Pronostique la valuación para una casa que tiene un área con calefacción de 1750 pies cuadrados, 10 años de antigüedad y 2500 pies cuadrados.

7) Un grupo de inversionistas europeos evalúan la posibilidad de invertir en acciones de una empresa cementera de nuestro país para lo cual requiere realizar estimaciones de las ganancias por acción. Las variables de predicción consideradas de mayor importancia por el jefe de inversiones así como los datos recopilados se muestran a continuación.
Las variables involucradas representan:

Y = ganancias por acción,

X1 = ventas (millones $)

X2 = activos (millones $)

X3 = utilidades como porcentaje de inversión (%).

Y	X1	X2	X3
8.72	8510	63688	8.1
3.1	2800	12566	2
3.15	3200	9958	2.85
2.43	2000	8356	1.85
2.01	1820	3124	2.05

4.08	3560	6923	3.55
4.18	4020	4424	3.85
3.86	3950	1116	3.65
14.74	18960	3492	15
6.26	5680	13	6.05
3.42	5360	2782	8
1.91	2300	855	4

a. Analice la existencia de multicolinealidad. Sustente su respuesta.
b. Estime valide e interprete los coeficientes del mejor modelo que permita predecir las ganancias por acción conociendo los valores obtenidos en las ventas, activos y utilidades. Use un nivel de significación del 7%.
c. Use el modelo de regresión anterior para estimar la ganancia por acción obtenidas cuando las ventas son de 3500 millones de dólares, los activos 6000 millones y se tiene un porcentaje de inversión del 4%.