EAN Escuela de **Administración de Negocios**

ANALIZAR

Análisis Exploratorio de Datos

UNIVERSIDAD DE COSTA RICA
LUCEM ASPICIO

Pasos Básicos

1.Explorar y entender la dinámica del proceso.

2.Generar teorías sobre las causas.

3.Verificar y eliminar causas.

Primer Paso – Exploración Inicial

1. Definir el problema
2. Seleccionar las categorías apropiadas al problema
3. Llevar a cabo lluvia de ideas de causas posibles y relacionarlas con las categorias
4. Agrupe las causas para ello utilice un diagrama de causa y efecto
5. Priorize por orden de importancia para ello apóyese en con un diagrama de Pareto
6. Utilice alguna técnica que le ayude a entender el problema, por ejemplo un 5W & 1H
7. Enfocarse en las causas seleccionadas, se espera que sean las de mayor impacto

Herramientas para la Exploración Inicial vistas en DN-0496.

Lluvia de Ideas

Diagrama de Pareto

5W 1H APROACH

06

Why

Doubt everything again. Think over & discuss the above 5 questions by asking more why & then find out the best solution for improvement.

01

WHAT

What should be improved? What is the purpose of improvement? Is there anything else to do?

02

WHERE

Is the workers orientation or method correct when changing the workplace?

03

WHEN

Change the time, perioed or sequence of the operation. Why perform at that time? Will it be better to execute at other time?

04

WHO

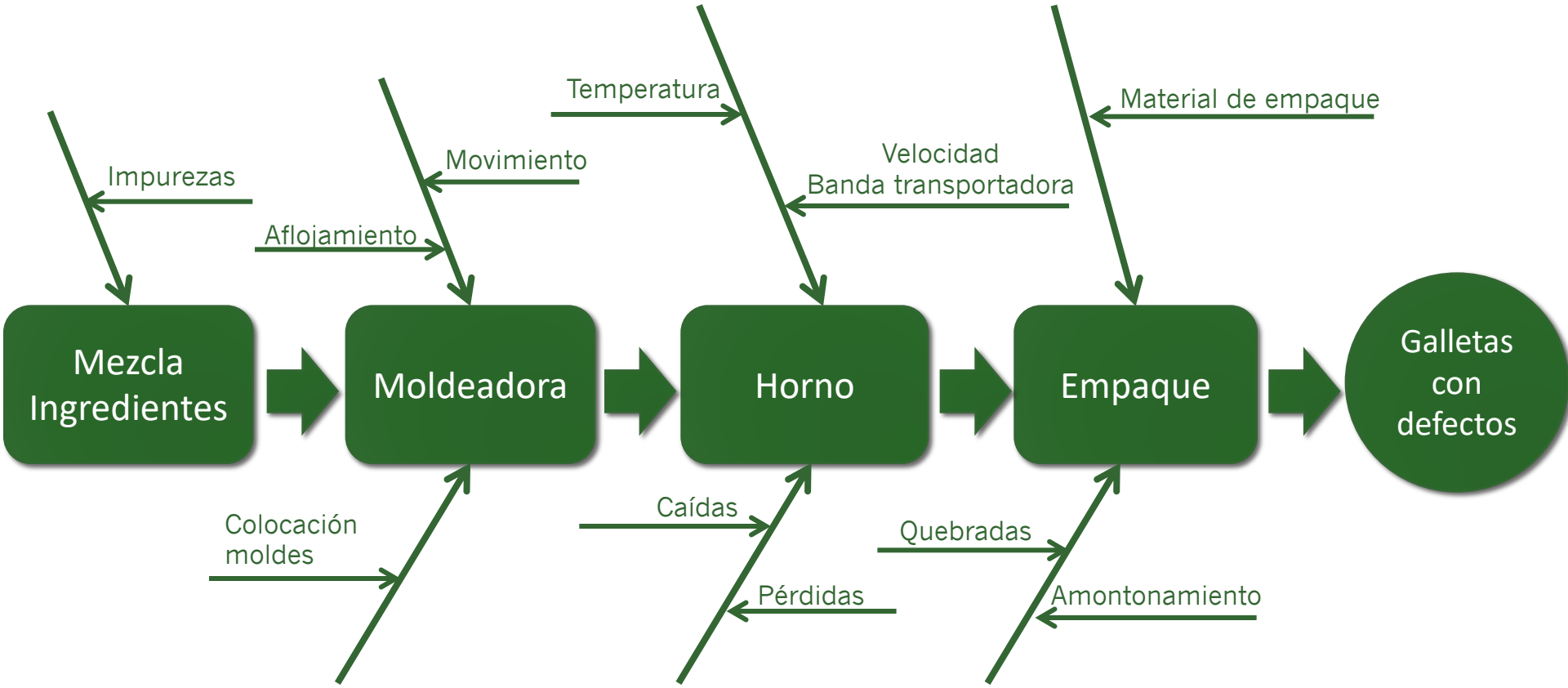
Manpower, collaboration or task assignment. Check & discuss this issue again.

05

How

Change the procedure so that less manpower is needed. Apply the method that requires less degree of proficiency & expense.

Digrama de Ishikawa, método de flujo del proceso



Segundo Paso la Toma de Datos

Tamaño de la muestra

Método de muestreo

Definir el Tamaño de la Muestra

Se debe escoger el nivel de confianza, así como el margen de error máximo tolerable con el que se desea hacer la estimación.

De tal forma que se logre obtener una muestra que sea representativa y permita hacer inferencias correctas sobre la población.

Tamaño de la muestra para estimar la media poblacional

La primera se aplica para poblaciones infinitas y la segunda para poblaciones finitas y conocidas.

$$n = \frac{(Z_{\alpha/2})^2 \sigma^2}{E^2}$$

$$n = \frac{Z_{\alpha/2}^2 N \sigma^2}{\sigma^2 Z_{\alpha/2}^2 + (N-1)E^2}$$

En donde:
$Z_{\alpha/2}$ = Nivel de confianza
E= error máximo tolerable
N= tamaño de la población

Tamaño de la muestra para estimar la media poblacional

Cuando σ se desconoce, se puede elegir mediante uno de los siguientes procedimientos:

1- Usar la desviación muestral de una muestra anterior de las mismas unidades

2- Llevar a cabo un estudio piloto para seleccionar una muestra preliminar de unidades.

3- Usar el juicio o un “estimado mejor” del valor de σ

4- Si no se aplica alguna de las alternativas anteriores, divida entre cuatro la diferencia entre el máximo y el mínimo de los datos.

Ejercicio

Un ingeniero de calidad analiza la resistencia a la compresión del concreto. Se desea estimar esa resistencia con un error menor que 15 psi para un nivel de confianza del 99% . ¿Qué tamaño de muestra debe emplearse para este fin si se sabe que la σ^2 de la resistencia es de 1000 psi?

Tamaño de la muestra para estimar la proporción poblacional

La primera se aplica para poblaciones infinitas y la segunda para poblaciones finitas y conocidas.

$$n = \frac{(Z_{\alpha/2})^2 pq}{E^2}$$

$$n = \frac{Z^2_{\alpha/2} Npq}{pqZ^2_{\alpha/2} + (N-1)E^2}$$

En donde:
$Z_{\alpha/2}$ = Nivel de confianza
E= error máximo tolerable
N= tamaño de la población
p=proporción de elementos que poseen la característica de interés

Tamaño de la muestra para estimar la proporción poblacional

Cuando p se desconoce, se puede elegir mediante uno de los siguientes procedimientos:

1- Usar la proporción muestral de una muestra anterior de las mismas unidades

2- Llevar a cabo un estudio piloto para seleccionar una muestra preliminar de unidades.

3- Usar el juicio o un “estimado mejor” del valor de p

4- Si no se aplica alguna de las alternativas anteriores, usar p=0,5

Ejercicio

En los lotes fabricados en la última semana se obtuvo un 5% de producto defectuoso. ¿De que tamaño debe ser la muestra si se desea estimar la proporción defectuosa poblacional con un margen de error de 0.02 con un nivel de confianza del 99%?

MÉTODOS DE MUESTREO

Muestreo Aleatorio Simple

Escoge al azar los miembros del universo hasta completar el tamaño muestral previsto

En teoría se enumeran previamente todos los elementos y de acuerdo con una tabla de números aleatorios se van escogiendo

El procedimiento puede darse con o sin reemplazos y esta condición afectará posteriormente el análisis

Muestreo Aleatorio Sistemático

En el universo (N) se elige el primer elemento al azar

Luego los demás se escogen cada cierto intervalo (k), hasta completar el tamaño muestral (n).

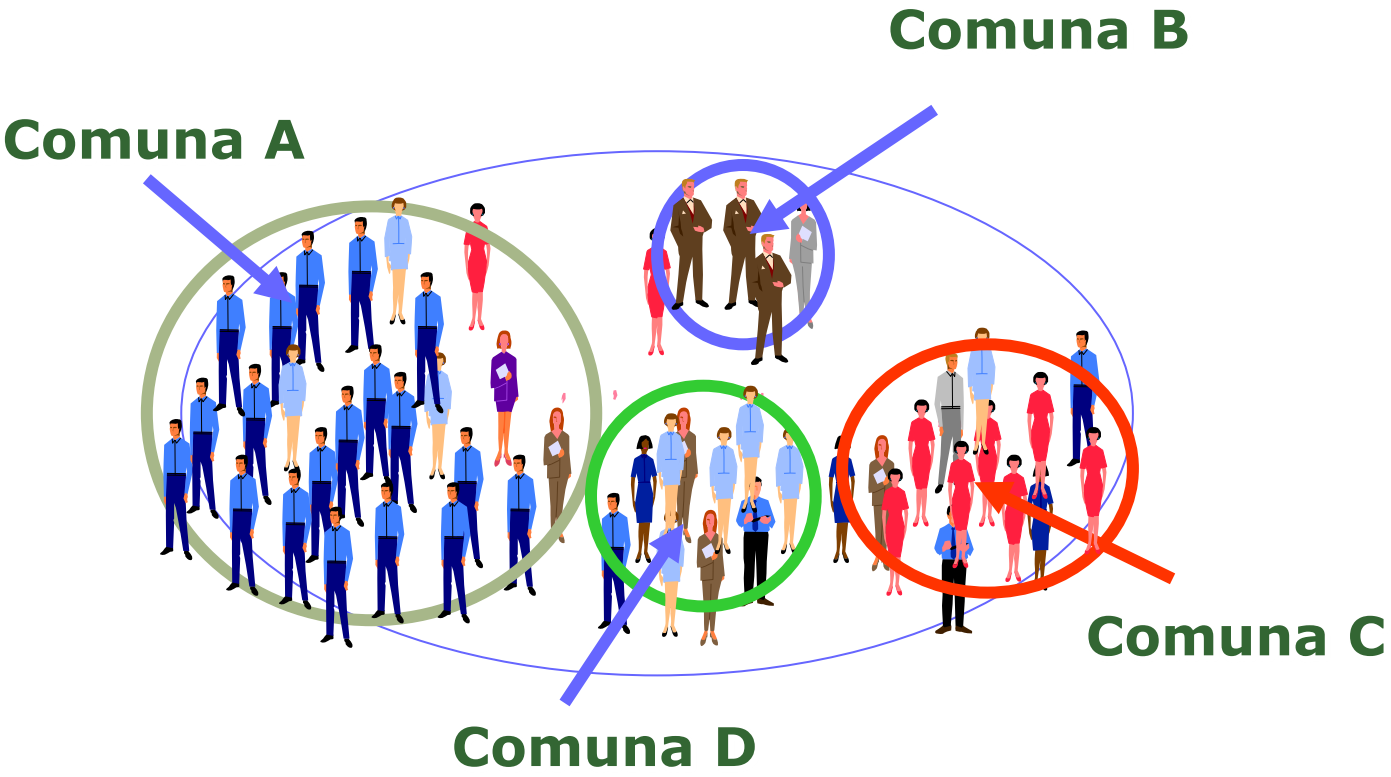
El tamaño del intervalo (k) se calcula así: $k = N/n$

Muestreo Estratificado

Considera que al interior del universo existen estratos (subgrupos internamente homogéneos pero cualitativa y cuantitativamente diferentes entre sí), y que no se cumple la condición de selección aleatoria pues los miembros del grupo mayoritario tienen una mayor probabilidad de ser seleccionados en la muestra.

ESTRATOS

Homogéneos en su interior; diferentes entre sí en propiedades y tamaño



Los estratos más grandes
Tienen mayor probabilidad de ser representados

¿Cómo grantizar la aleatoriedad en universos estratificados... ?

Puede usarse alguna de las siguientes técnicas:

1. **Muestreo Estratificado Proporcional**
2. **Muestreo Estratificado No Proporcional**
3. **Alocación óptima de los estratos.**

Muestreo Aleatorio por Conglomerados

Los miembros del grupo mayoritario tienen una mayor probabilidad de ser seleccionados en la muestra

No se cumple la aleatoridad

En Las Unidades de observación se eligen aleatoriamente al interior de los conglomerados

El error de la medición (error muestral) no se de al interior del conglomerado sino entre los conglomerados

Antes de hacer inferencias, el analista deberá examinar la variabilidad interna de cada conglomerado y la variabilidad entre ellos, pues es posible que algunos de los conglomerados no sean representativos del universo.

Tercer Paso Explorando los Datos

Análisis de Normalidad
Pruebas de hipótesis

Conceptos Generales

I. Hipótesis: Enunciado que se quiere demostrar.

II. Prueba de Hipótesis: Procedimiento para determinar si se debe rechazar o no una afirmación acerca del valor de un parámetro de la población.

III. Hipótesis nula: Ho, hipótesis que propone un valor tentativo acerca de una parámetro poblacional.

Conceptos Generales

IV. Hipótesis alternativa: Ha, hipótesis opuesta a la hipótesis nula.

V. La hipótesis que se investiga por lo común se expresa como la alternativa.

VI. Se puede llegar a la conclusión de que la hipótesis que se investiga es verdadera si se rechaza la hipótesis nula.

Caso 1: Una investigación

- Se desea probar si un nuevo sistema de inyección mejora el rendimiento de un modelo de vehículo cuyo rendimiento promedio es de 24 millas/galón.
    - Ho: $\mu \leq 24$
      Ha: $\mu > 24$

La Ha se formula de tal modo que el rechazo de Ho respalde la conclusión que se propone.

Caso 2: Una afirmación

- Un fabricante de bebidas afirma que el contenido de sus botellas no es menor a 2 litros.
    - Ho: $\mu \geq 2$
      Ha: $\mu < 2$

La Ha se formula de tal modo que el rechazo de Ho proporcione evidencia estadística de que la afirmación es incorrecta.

Caso 3: Una decisión

- Se prueban unas piezas de un embarque para saber si están o no dentro de especificaciones, en donde un valor menor o mayor a 2 causa problemas de ensamblaje.
    - Ho: $\mu = 2$
      Ha: $\mu \neq 2$

En estos casos, se toma tanto una decisión si la hipótesis nula se rechaza como si no se pueda rechazar.

Tipos de Error

Hipótesis nula: H_0: $\mu_1 = \mu_2$ (las dos medias son iguales)
Hipótesis alternativa: H_1: $\mu_1 \neq \mu_2$ (las dos medias no son iguales)

		Atendiendo a los resultados del estadístico se decide:	
		No rechazar	Rechazar
H_0	Verdadera	Decisión correcta	Error tipo I (α)
	Falsa	Error tipo II (β)	Decisión correcta

Nivel de Significancia
Error tipo I: α

En la práctica, la persona que efectúa la prueba de hipótesis especifica la máxima probabilidad permisible, llamada nivel de significancia para la prueba, de cometer un error tipo I.

Se acostumbran los valores de 0,05 y 0,01 para α.

Un pequeño valor de α significa un alto grado de confianza en que sea correcta la conclusión de rechazar Ho y de que Ha es verdadera.

Error Tipo II: β

Aunque en la mayoría de las aplicaciones de pruebas de hipótesis se controla la probabildad de cometer un error tipo 1, no siempre se controla la probabilidad de incurrir en un error tipo II.

Debido a la incertidumbre de cometer un error tipo II, en estadística se recomienda usar la redacción “no rechazar Ho” en lugar de “aceptar Ho” porque si se acepta Ho se corre el riesgo de cometer un error tipo II.

Prueba de Normalidad

La inferencia estadística acerca de una media o proporción poblacional a través de intervalos de confianza requiere que los datos se comporten normalmente.

Las pruebas de normalidad se conocen como pruebas de bondad de ajuste a partir de una hipótesis formulada así:

Ho: f(x) = distribución normal

Ha: f(x) ≠ distribución normal

¿Cómo reconocer la normalidad de los datos?

Para determinar si un grupo de datos corresponde a una distribución normal o no:

1. Histograma: permite observar si los datos se agrupan alrededor de un valor central.
2. Curtosis: indica qué tan aguda (>0) o achatada (<0) es la curva
3. Coeficiente de Asimetría, indica qué tan a la izquierda de la media están los datos (<0) o qué tan a la derecha (>0)
4. Si los dos indicadores están entre -1 y 1, los datos se aproximan bastante a una distribución normal.
5. Prueba de Bondad de Ajuste, empleando el estadístico Chi-cuadrado o la prueba de Kolmogorov-Smirnof (técnicas no paramétricas)

La Prueba de Gráfico de Cuantiles Teóricos

Consiste en comparar los cuantiles de la distribución observada con los cuantiles teóricos de una distribución normal con la misma media y desviación estándar que los datos. Cuanto más se aproximen los datos a una normal, más alineados están los puntos entorno a la recta.

Baje el archivo de excel Sesión 08 Analizar

Cuarta pregunta (20 pts.) Los contenidos de lactosa, en miligramos, de una barra de chocolate de cierta marca se registraron de la siguiente manera:

1.09	1.74	1.58	2.11	1.64	1.79	1.37	1.75
1.92	1.47	2.03	1.86	0.72	2.46	1.93	1.63
2.31	1.97	1.70	1.90	1.69	1.88	1.40	2.37
1.79	0.85	2.17	1.68	1.85	2.08	1.64	1.75
2.28	1.24	2.55	1.51	1.82	1.67	2.09	1.69

a. ¿Se podría afirmar con un nivel de significancia de 0.01 que la lactosa incluida en las barras de chocolate siguen una distribución normal con $\mu = 1.8$ y $\sigma = 0.4$?
b. Si las tolerancias de la barra de chocolate son 1.5 ± 0.5 construya una gráfica de capacidad de este proceso y dé una primera opinión sobre la capacidad.

Pruebas de Bondad de Ajuste

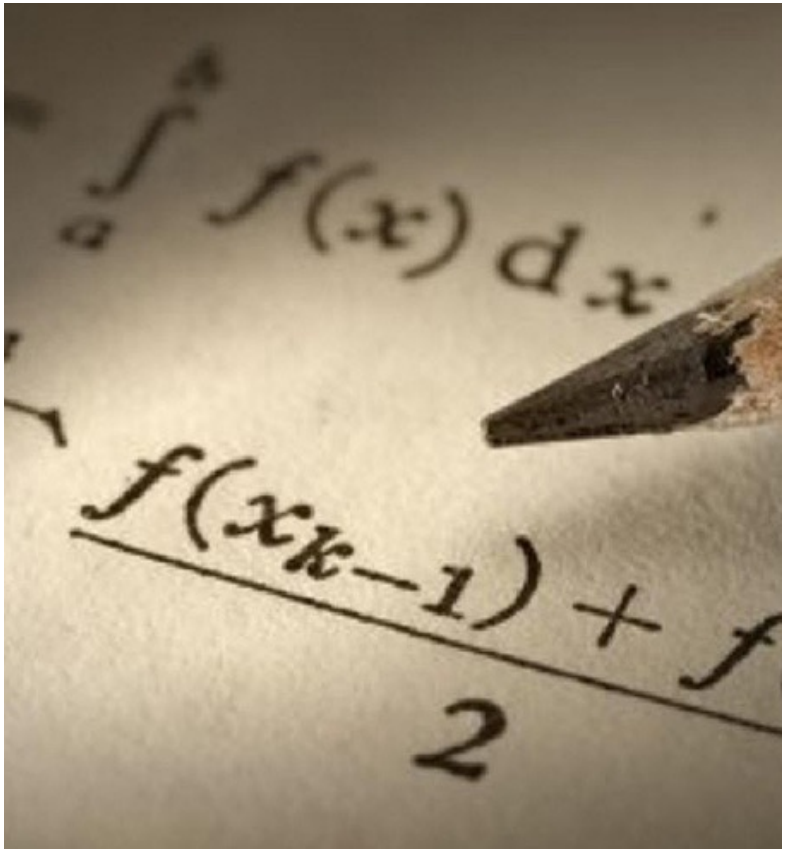
Una de las bases fundamentales del control estadístico de la calidad es la inferencia estadística. Por ello, la determinación del tipo de distribución correspondiente a un conjunto de datos provenientes del estudio es absolutamente necesario. La prueba de bondad de ajuste permite probar el ajuste de los resultados de un experimento a una distribución de probabilidad teórica sujeto a un error o nivel de confianza.

Pruebas de Bondad de Ajuste

El método en cuestión se basa en la comparación de las frecuencias absolutas observadas y las frecuencias absolutas esperadas, calculadas a partir de una distribución teórica en análisis.

Tipos de Pruebas de Bondad de Ajuste

1. Chi Cuadrado
2. Kolmogorov Smirnof
3. Anderson Darling
4. Shapiro – Wilk
5. Prueba D'Agostino para el Sesgo y para la Curtosis
6. Prueba Omnibus D'Agostino



La Prueba de Bondad de Ajuste de Chi Cuadrado

El test de bondad de ajuste chi cuadrado puede ser utilizado para trabajar tanto con distribuciones discretas como, por ejemplo, la Distribución de Poisson o la Distribución Binomial como así también con distribuciones continuas (por ejemplo, Distribución Normal, Distribución Exponencial, etc). Esto a diferencia de las pruebas de bondad de ajuste Kolmogorov Smirnov y Anderson Darling que sólo pueden ser utilizados para trabajar con distribuciones continuas.

La Prueba de Bondad de Ajuste de Chi Cuadrado

Una desventaja potencial del test de chi cuadrado es que requiere una muestra suficientemente grande de modo que la aproximación de chi cuadrado sea válida.

Chi-Cuadrado

$$\chi^2 = \sum \frac{(fo - fe)^2}{fe}$$

- En donde:
- fo =Frecuencia observada de datos discretos
- fe =Frecuencia esperada de la distribución teórica
- Los grados de libertad se emplea (k-1) y luego se resta un grado adicional de libertad para cada parámetro de población que tenga que ser estimado de los datos de la muestra

Media con datos Agrupado

$$\overline{X} = A + \frac{(\Sigma\ fo \times d) \times i}{n}$$

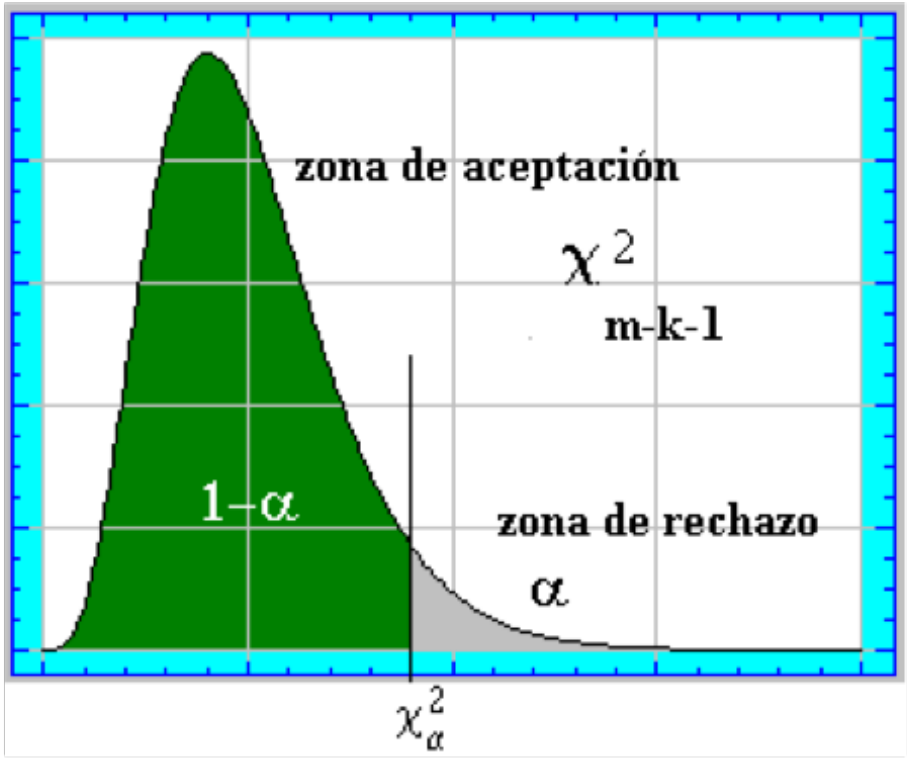
- En donde:
- d= desviación del punto medio con respecto a la posición de la media supuesta, es medida en unidades de intervalo de clase
- i= amplitud o intervalo de clase
- A= punto medio de la clase que contiene la media supuesta (clase de d=0)
- fo = Frecuencias observadas en número de clases
- n = tamaño de la muestra

Desviación estándar con datos Agrupados

$$S = i\sqrt{\frac{\Sigma fo \times d^2}{n} - \frac{(\Sigma fo \times d)^2}{n^2}}$$

Chi-Cuadrado

- Se acepta la hipótesis nula cuando:
- Chi-calculado < Chi-tabular



Pregunta número 1(20 pts.) Proveedores de Equipo para la Industria Petrolera S.A., es un productor de equipo industrial para empresas que se dedican a la extracción de petróleo y venta de gasolina al detallista a nivel mundial. Actualmente acaba de diseñar una nueva bomba expendedora de gasolina la cual ya pasó las primeras pruebas de calidad satisfactoriamente. En estos momento Evelyn, presidenta de la corporación está analizando el cartel de oferta de este producto a Texaco y Shell, importantes clientes de la compañía. Evelyn solo tiene duda en cuanto a un punto específico de la grantía del producto. Ya que los ingenieros de diseño afirman que esta nueva bomba expendedora puede trabajar 50 dias las 24 horas sin parar sin descalibrarse, a Evelyn le gustaría hacer más pruebas de calidad antes de poner esta afirmación por escrito, pero primero debe determinar qué clase de comportamineto estadístico se tiene para decidir que herramienta se va a utilizar. Los siguientes datos representan el periodo de duración, en días, de 60 bombas de combustible que lograron trabajar de manera consecutiva sin descalibrarse.

23	60	48	32	57	74	52	27	82	36
49	77	37	95	41	65	92	59	55	55
52	10	64	55	58	23	80	98	58	67
41	45	50	54	45	72	88	62	43	43
60	58	39	76	84	48	54	90	15	60
34	67	17	60	69	74	63	52	49	61

Se puede afirmar con un nivel de significancia de 0.05 que el proceso de descalibración de las máquinas sigue un comportamiento normal. Utilice el estadístico de Chi cuadrado

La Prueba de Bondad de Ajuste de Chi Cuadrado

Ho: el proceso de descalibración sigue un comportamiento normal
Ha: el proceso de descalibración NO sigue un comportamiento normal

Li	Ls	Fo	Xk	d	Fod	Fod^2	Pi	Fe

Kolmogorov-Smirnov

- Es un simple método no paramétrico para probar si hay una diferencia significativa entre una distribución observada y una distribución teórica de frecuencia.
- La hipótesis nula se rechaza sí:

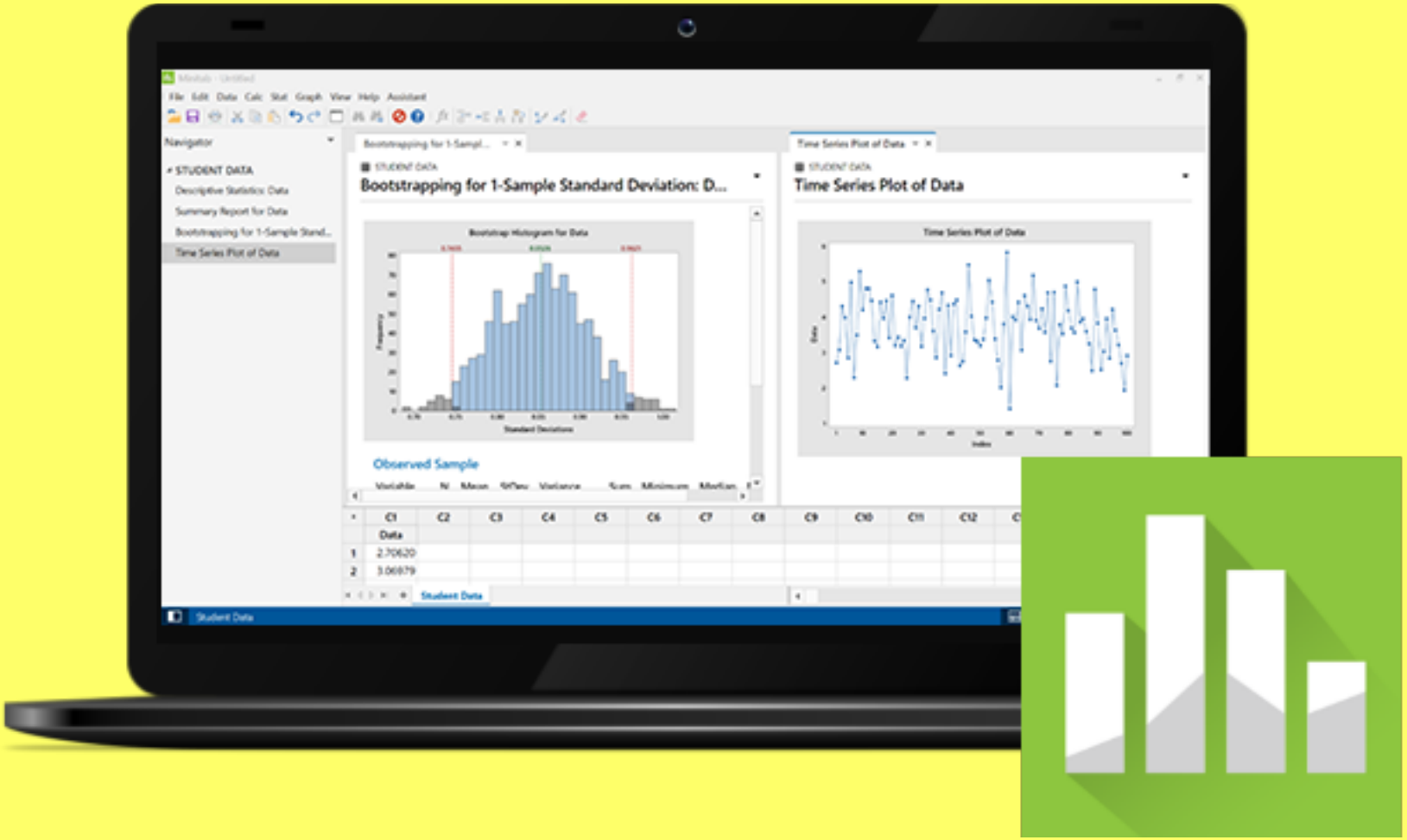
- * $Dn \geq d_t, 1-\alpha$
- En donde Dn es el valor máximo.

Pregunta número 1.(25 pts.) Se supone que una máquina mezcla maní, avellanas, macadamias, y pacanas en la relación 5:2:2:1. Una lata contiene 750 gramos de estos frutos mezclados. La gerencia de calidad desea controlar la calidad del proceso de fabricación de estas latas por medio de gráficas de control. Sin embargo, primero debe determinar si la distribución de frecuencias del peso en gramos de cacahuates puede aproximarse con una μ de 350 gramos y una σ de 70 gramos.

Frecuencias Observadas de los Gramos de Cacahuates Encontrados en la Muestra	
Límites de Clase	**Frecuencia Observada**
145-195	2
195-245	1
245-295	4
295-345	15
345-395	10
395-445	5
445-495	3

En el nivel de significancia de 0.05 pruebe la normalidad de los datos utilizando el estadístico de Komolgorov-Smirnov.

Li	Ls	Fo	Fo acumulada	Fo acumulada relativa	Fe acumulada	D Ifo-feI
145	195	2				
195	245	1				
245	295	4				
295	345	15				
345	395	10				
395	445	5				
445	495	3				



Bondad de Ajuste en Minitab

Pregunta número 1(20 pts.) Proveedores de Equipo para la Industria Petrolera S.A., es un productor de equipo industrial para empresas que se dedican a la extracción de petróleo y venta de gasolina al detallista a nivel mundial. Actualmente acaba de diseñar una nueva bomba expendedora de gasolina la cual ya pasó las primeras pruebas de calidad satisfactoriamente. En estos momento Evelyn, presidenta de la corporación está analizando el cartel de oferta de este producto a Texaco y Shell, importantes clientes de la compañía. Evelyn solo tiene duda en cuanto a un punto específico de la grantía del producto. Ya que los ingenieros de diseño afirman que esta nueva bomba expendedora puede trabajar 50 dias las 24 horas sin parar sin descalibrarse, a Evelyn le gustaría hacer más pruebas de calidad antes de poner esta afirmación por escrito, pero primero debe determinar qué clase de comportamineto estadístico se tiene para decidir que herramienta se va a utilizar. Los siguientes datos representan el periodo de duración, en días, de 60 bombas de combustible que lograron trabajar de manera consecutiva sin descalibrarse.

23	60	48	32	57	74	52	27	82	36
49	77	37	95	41	65	92	59	55	55
52	10	64	55	58	23	80	98	58	67
41	45	50	54	45	72	88	62	43	43
60	58	39	76	84	48	54	90	15	60
34	67	17	60	69	74	63	52	49	61

Se puede afirmar con un nivel de significancia de 0.05 que el proceso de descalibración de las máquinas sigue un comportamiento normal. Utilice el estadístico de Chi cuadrado

Pruebas de Hipótesis

La comparación entre grupos se puede realizar a nivel gráfico, empleando diagramas Box Whisker, Barras Verticales Comparativas, o Líneas de Tendencia comparativas.

También se puede emplear estadísticos descriptivos tales como la **media o mediana** (en caso de que los datos no sean normales), desviaciones estándar, rango, etc.

El empleo de pruebas de hipótesis permite verificar estadísticamente estas diferencias.

Media y Mediana

MEDIA

Empleada en variables numéricas únicamente

Preferible en distribuciones normales

Toma en cuenta todos los valores de los datos

Utilizada en estadística paramétrica

Empleada en inferencia estadística

Gráfico de Barras de Error (compara medias)

MEDIANA

Empleada en variables numéricas y ordinales

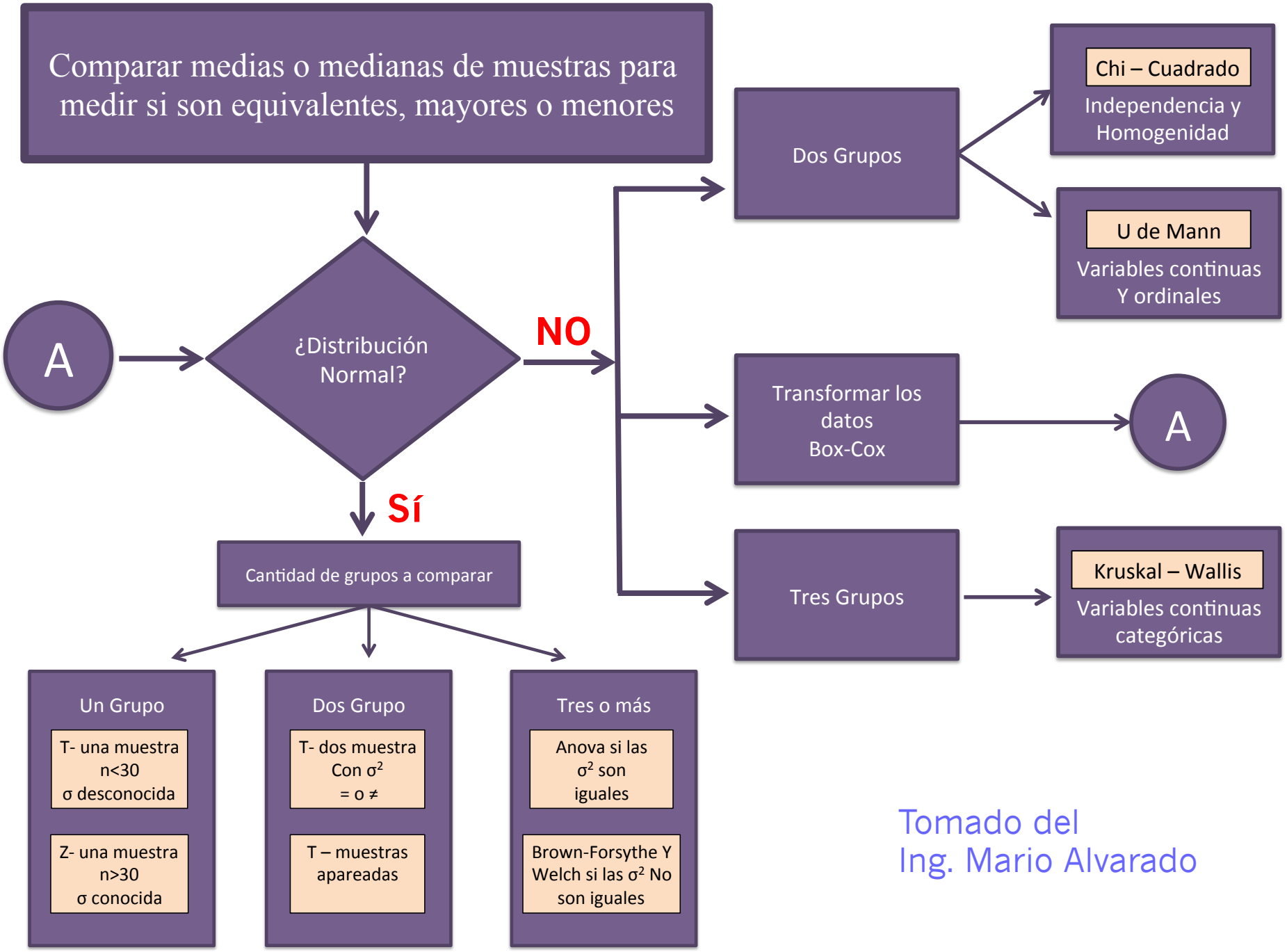
Preferible en distribuciones sesgadas

Elimina valores extremos

Utilizada en estadística no paramétrica

Difícil de emplear en inferencia estadística

Gráfico Boxplot (compara medianas)



Tomado del
Ing. Mario Alvarado

Microsoft®
Excel

Pruebas de Hipótesis en Excel

Tipos de Pruebas en Excel

Excel nos ofrece varios comandos para poder resolver hipótesis para comparar dos medias, tomando el valor de p y comparándolo con α.

Si el valor de p es menor que α o que $\alpha/2$, se concluye que existen diferencias significativas entre los grupos que se comparan.

Tipos de Pruebas en Excel

Características de la prueba de medias	Prueba que se emplea en Excell
Se conocen σ_1^2 y σ_2^2	Prueba Z para medias de dos muestras
Comparar antes y después del tratamiento	Prueba t para medias de dos muestras emparejadas
Comparar las medias de dos poblaciones	Prueba t para dos muestras suponiendo varianzas iguales
	Prueba t para dos muestras suponiendo varianzas desiguales
Se supone que $\sigma_1^2 = \sigma_2^2 = \ldots \sigma_n^2$	Análisis de varianza

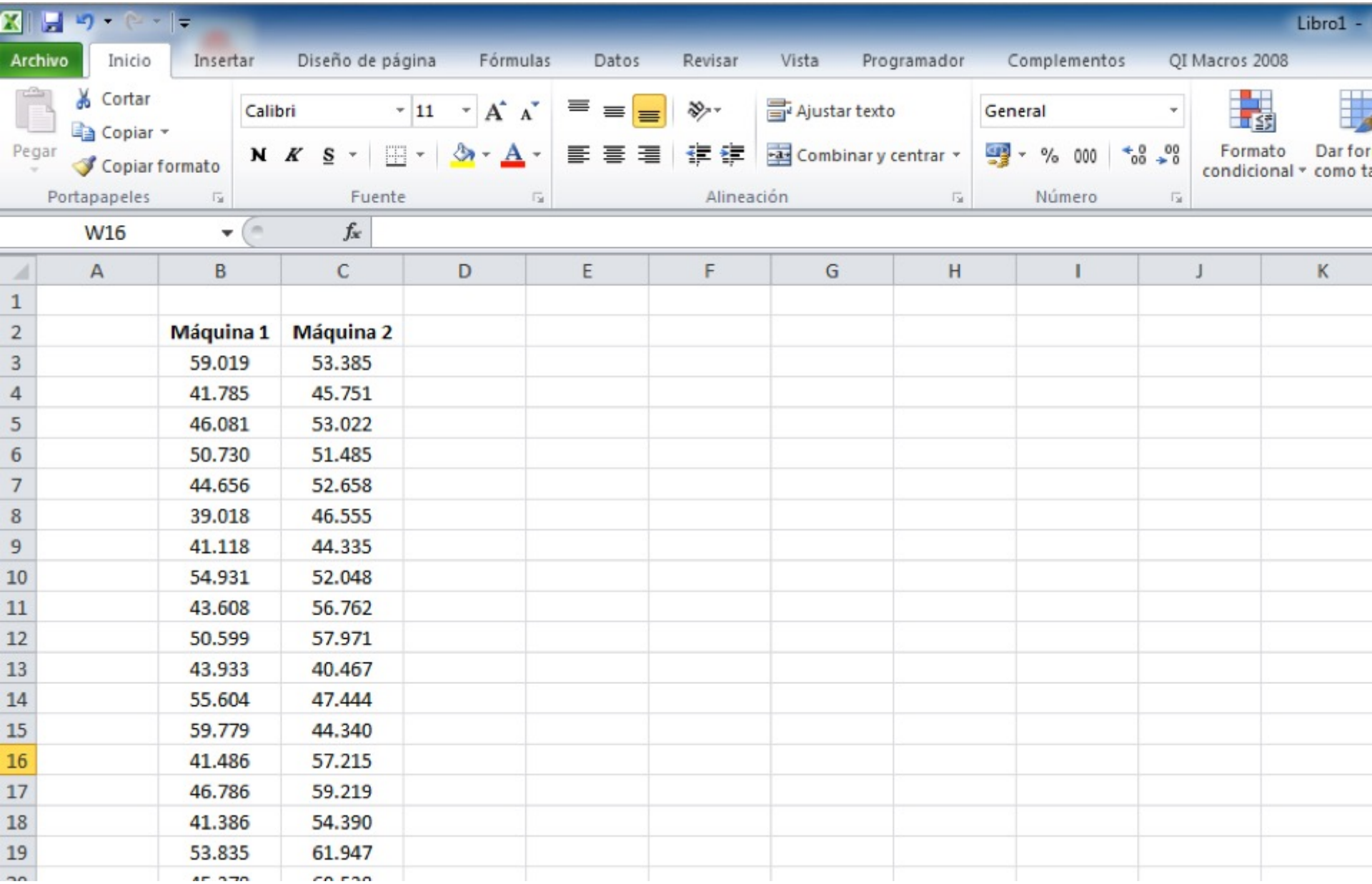
Prueba Z para Dos Medias

Para realizar esta prueba se requiere cumplir una de las dos siguientes condiciones:

1. Que las varianzas poblacionales sean conocidas.
2. Que las varianzas poblacionales sean desconocidas pero el tamaño de la muestra sea mayor a 30 para estimar las varianzas a partir de la muestra.

Ejercicio Los Líquidos S.A.

Ahora suponga que el producto es fabricado por dos máquinas y se desea saber si entre ellas hay una diferencia en el desempeño, para ello se toman dos muestras de 40 unidades cada una y se realiza una prueba de hipótesis de diferencia de medias utilizando la prueba Z con varianzas poblacionales desconocidas.


Libro1 -
Archivo
Inicio
Insertar
Diseño de página
Fórmulas
Datos
Revisar
Vista
Programador
Complementos
QI Macros 2008
Cortar
Copiar
Copiar formato
Pegar
Portapapeles
Calibri
11
Fuente
Ajustar texto
Combinar y centrar
Alineación
General
Número
Formato condicional
W16
A B C D E F G H I J K
1
2 Máquina 1 Máquina 2
3 59.019 53.385
4 41.785 45.751
5 46.081 53.022
6 50.730 51.485
7 44.656 52.658
8 39.018 46.555
9 41.118 44.335
10 54.931 52.048
11 43.608 56.762
12 50.599 57.971
13 43.933 40.467
14 55.604 47.444
15 59.779 44.340
16 41.486 57.215
17 46.786 59.219
18 41.386 54.390
19 53.835 61.947