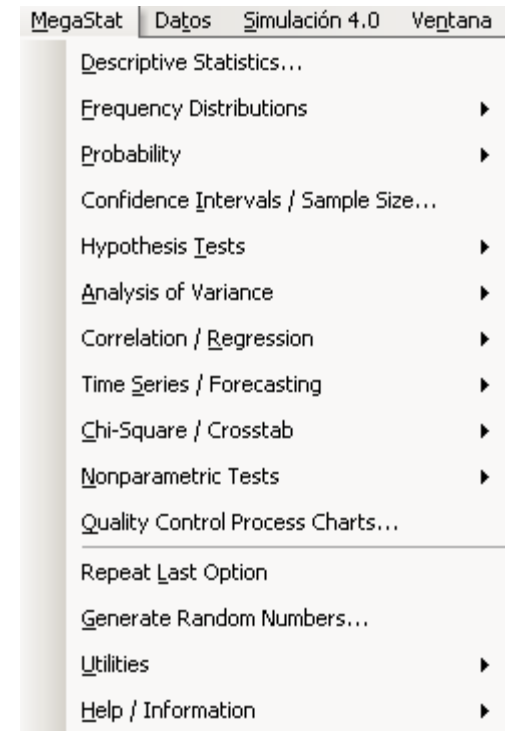


Manual de bolsillo del MegaStat *



* MegaStat es un complemento estadístico para el Excel elaborado por el profesor J. B. Orris de Butler University.

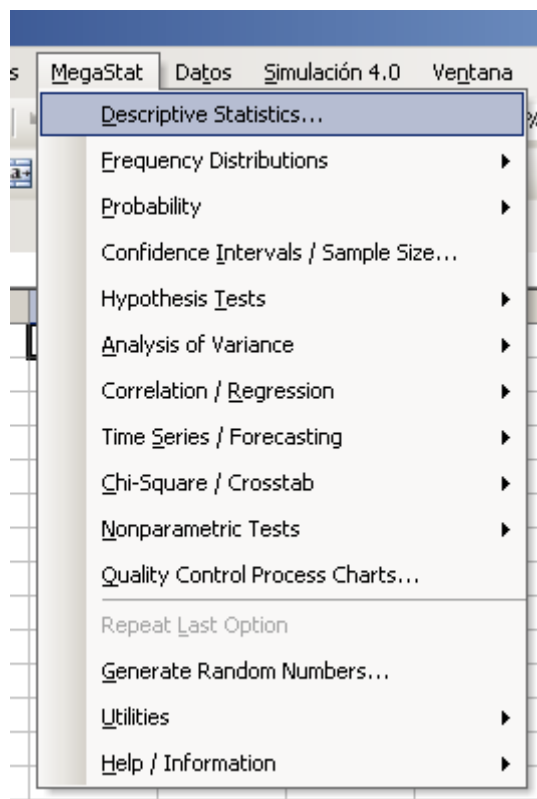
Estadísticas con MegaStat

AgeCat	Gender	Seconds
1	2	50.1
1	1	53.0
2	2	43.2
1	2	34.9
3	1	37.5

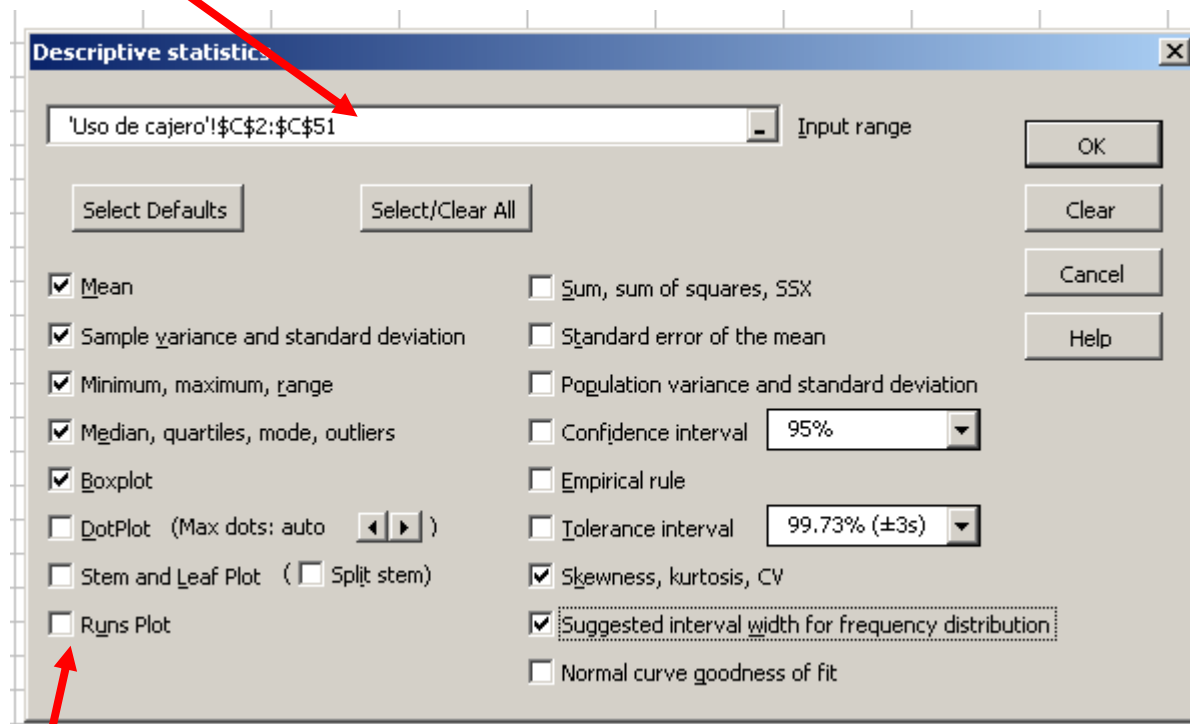
Para ver la utilidad del MegaStat, primero presentaremos el ejemplo con el cual vamos a trabajar.

Se trata del uso de un cajero automático de un banco cualquiera, la variable **AgeCat** es la clasificación por edad del usuario de este cajero, 1 si es menor de 30 años, 2 si tiene entre 30 y 50, y 3 si tiene mas de 50. La variable **Gender** es el genero (sexo) del usuario de este cajero, 1 si es hombre y 2 si es mujer. Y **Seconds** es el tiempo en segundo del uso de este cajero.

Estadísticas descriptivas



Ingresar en rango de datos que están en el Excel.

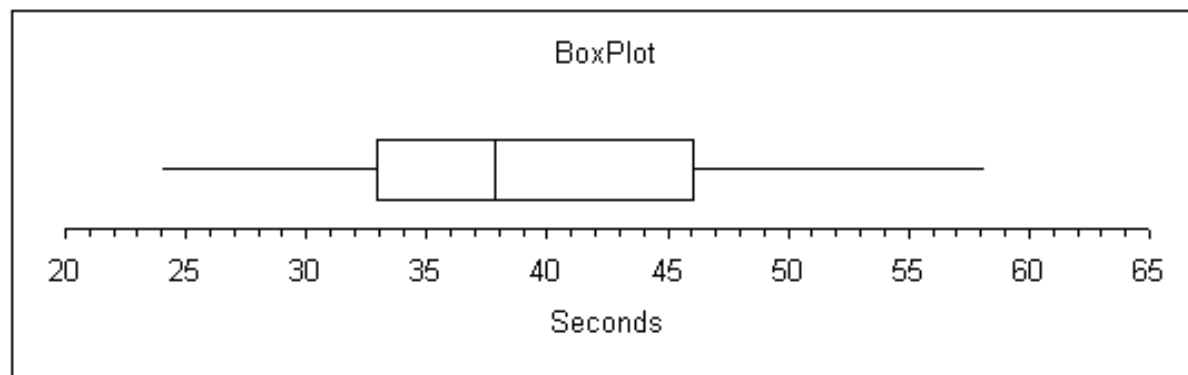


Seleccionar las estadísticas que se desean.

Estadísticas descriptivas

Descriptive statistics

	<i>Seconds</i>
count	50
mean	39.756
sample variance	79.488
sample standard deviation	8.916
minimum	24.1
maximum	58.1
range	34
skewness	0.235
kurtosis	-0.868
coefficient of variation (CV)	22.43%
1st quartile	32.975
median	37.800
3rd quartile	46.050
interquartile range	13.075
mode	37.800
low extremes	0
low outliers	0
high outliers	0
high extremes	0
suggested interval width	5



Estadísticas descriptivas

The image shows a 'Descriptive statistics' dialog box with various options checked. Red arrows point from Spanish labels to these options:

- Media** points to the Mean option.
- Varianza y desviación estándar de la muestra** points to the Sample variance and standard deviation option.
- Máximo, mínimo, rango** points to the Minimum, maximum, range option.
- Mediana, cuartiles, moda y datos fuera de lugar.** points to the Median, quartiles, mode, outliers option.
- Grafico de caja** points to the Boxplot option.
- Grafico tallos y hojas** points to the Stem and Leaf Plot option.
- Error estándar de la media** points to the Standard error of the mean option.
- Varianza y desviación estándar de la población** points to the Population variance and standard deviation option.
- Asimetría, curtosis y coeficiente de variación.** points to the Skewness, kurtosis, CV option.
- Ancho de intervalo sugerido para la distribución de frecuencias.** points to the Suggested interval width for frequency distribution option.

The dialog box also includes buttons for 'OK', 'Clear', 'Cancel', and 'Help', and a 'Confidence interval' dropdown set to '95%'.

Distribución de frecuencias cuantitativas

Ingresar en rango de datos que están en el Excel.

Frequency Distributions - Quantitative

InputRange: 'Uso de cajero!':\$C\$2:\$C\$51

Equal width intervals | Custom intervals | Options

(Leave these fields blank for auto estimation.)

5 interval width

24 lower boundary of first interval

Histogram
 Polygon
 Ogive

OK
Clear
Cancel
Help

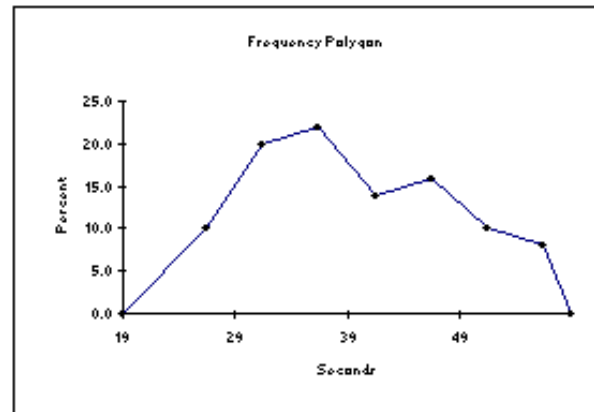
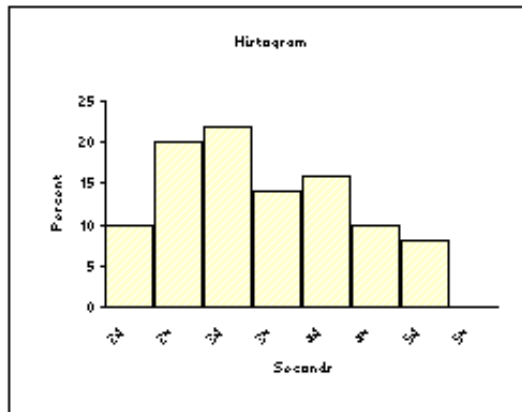
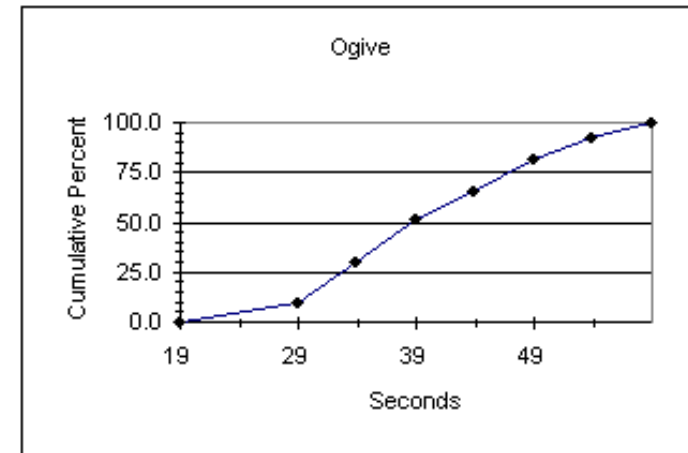
El ancho del intervalo se obtiene de las estadísticas descriptivas.

El limite inferior del primer intervalos de clase, se obtiene del menor dato (obtenido con las estadísticas descriptivas – 24.1) y se le resta algo.

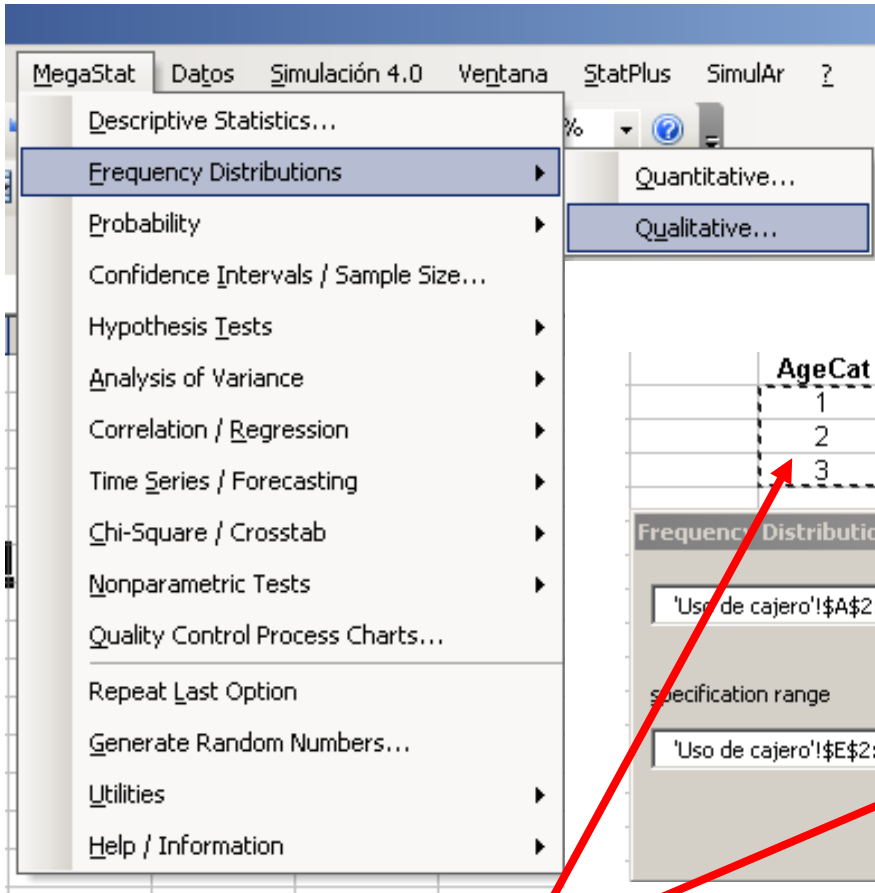
Distribución de frecuencias cuantitativas

Frequency Distribution - Quantitative

Seconds					<i>cumulative</i>			
<i>lower</i>	<i>upper</i>	<i>midpoint</i>	<i>width</i>	<i>frequency</i>	<i>percent</i>	<i>frequency</i>	<i>percent</i>	
24	< 29	27	5	5	10.0	5	10.0	
29	< 34	32	5	10	20.0	15	30.0	
34	< 39	37	5	11	22.0	26	52.0	
39	< 44	42	5	7	14.0	33	66.0	
44	< 49	47	5	8	16.0	41	82.0	
49	< 54	52	5	5	10.0	46	92.0	
54	< 59	56	5	4	8.0	50	100.0	
				50	100.0			

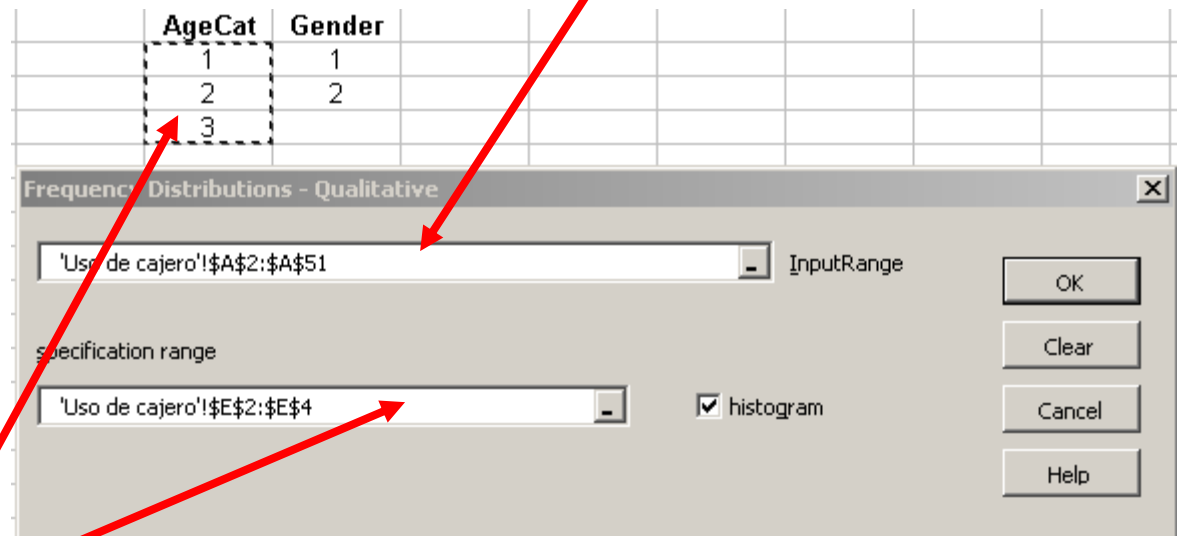


Distribución de frecuencias cualitativas



Trabajaremos con la variable **Agecat**

Ingresamos el rango de datos de la variable **Agecat**.

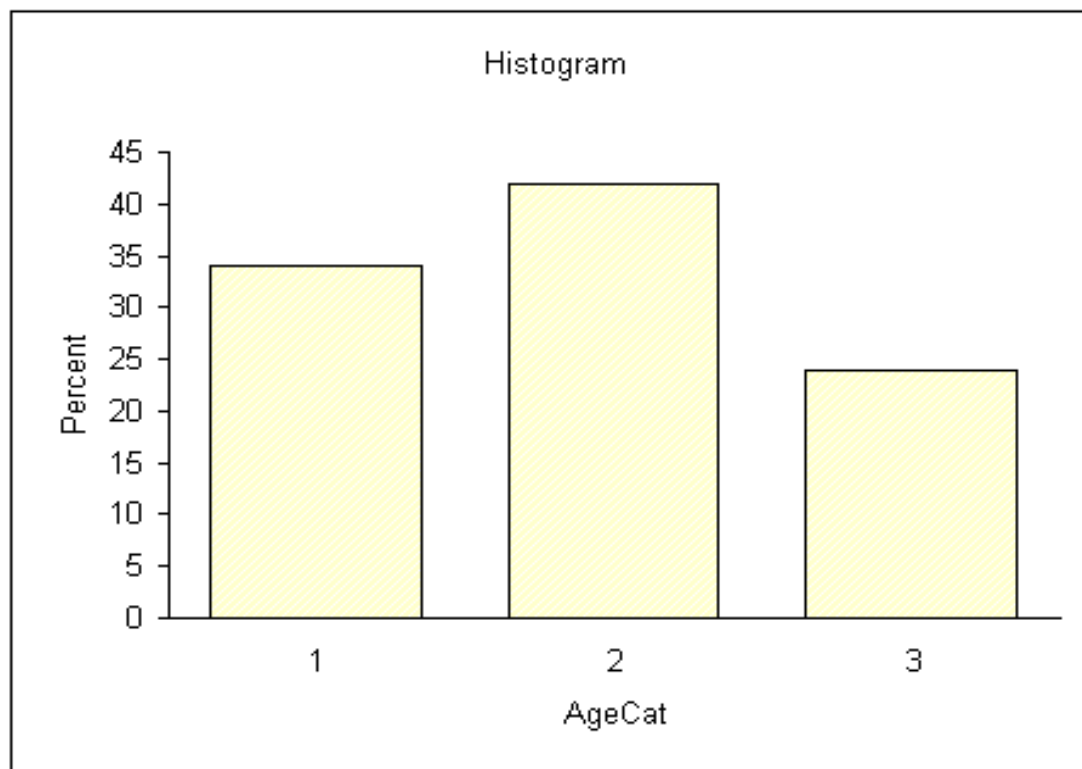


Ingresamos como se clasifica la variable **Agecat**, en 1, 2 y 3 (se acuerdan en menores de 30, entre 30 a 50 y mayores de 50)

Distribución de frecuencias cualitativas

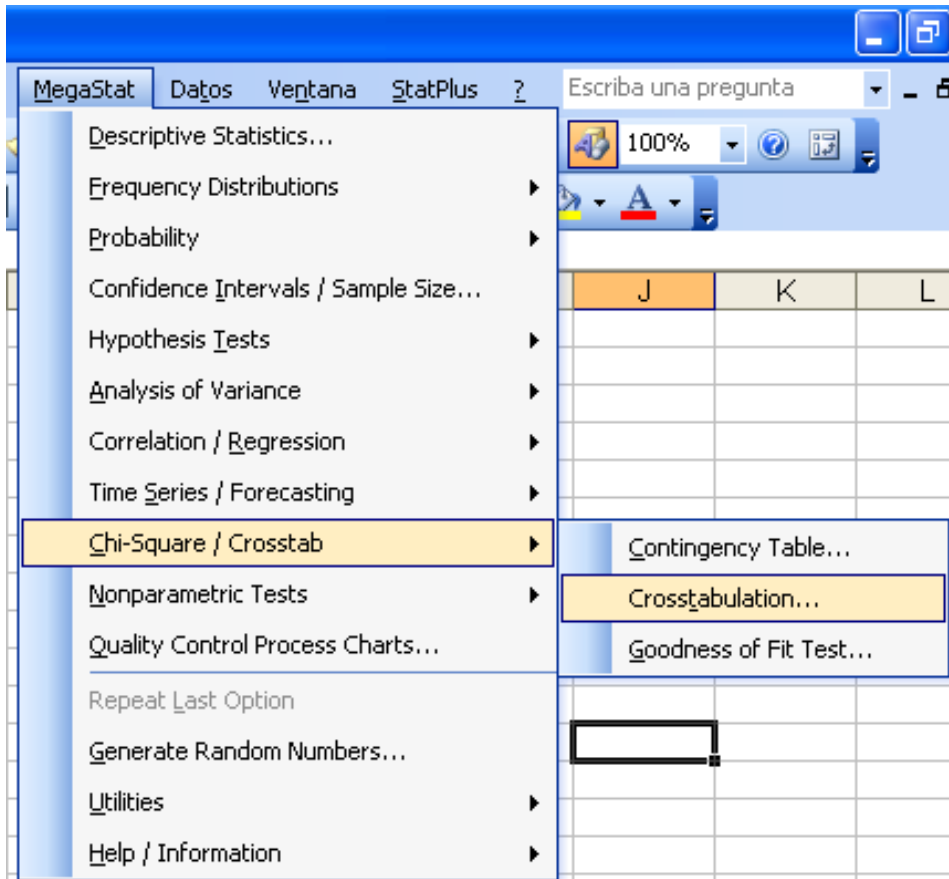
Frequency Distribution - Qualitative

<i>AgeCat</i>	<i>frequency</i>	<i>percent</i>
1	17	34.0
2	21	42.0
3	12	24.0
	50	100.0



Tablas de contingencias o tablas cruzadas

En muchas ocasiones, cuando analizamos variables cualitativas, necesitamos cruzar la información de estas variables. A esto lo llamamos tablas cruzadas.



Esto lo hacemos con el MegaStat.

Seguiremos trabajando con nuestro ejemplo. Supongamos que necesitamos hacer la tabla cruzada entre:

AgeCat y Gender.

Siempre la primera variable corresponde a las filas y la segunda variable corresponde a las columnas.

Se ingresa el rango de datos de la variable que va en la fila, en nuestro ejemplo: Agecat.

Se ingresa el rango de la calificación de la variable Agecat (1, 2 y 3)

Se selecciona lo que nos interesa saber para las estadísticas descriptivas

Crosstabulation

Row variable

'Uso de cajero'!\$A\$2:\$A\$51 Data range

'Uso de cajero'!\$E\$2:\$E\$4 Specification range

Column variable

'Uso de cajero'!\$B\$2:\$B\$51 Data range

'Uso de cajero'!\$F\$2:\$F\$3 Specification range

Output Options

chi-square Expected values Phi coefficient

% of row O - E Coefficient of contingency

% of column (O - E)² / E Cramér's V

% of total % of chi-square Fisher Exact Test

OK

Clear

Cancel

Help

Se ingresa el rango de la calificación de la variable Gender (1 y 2)

Se ingresa el rango de datos de la variable que va en la columna, en nuestro ejemplo: Gender.

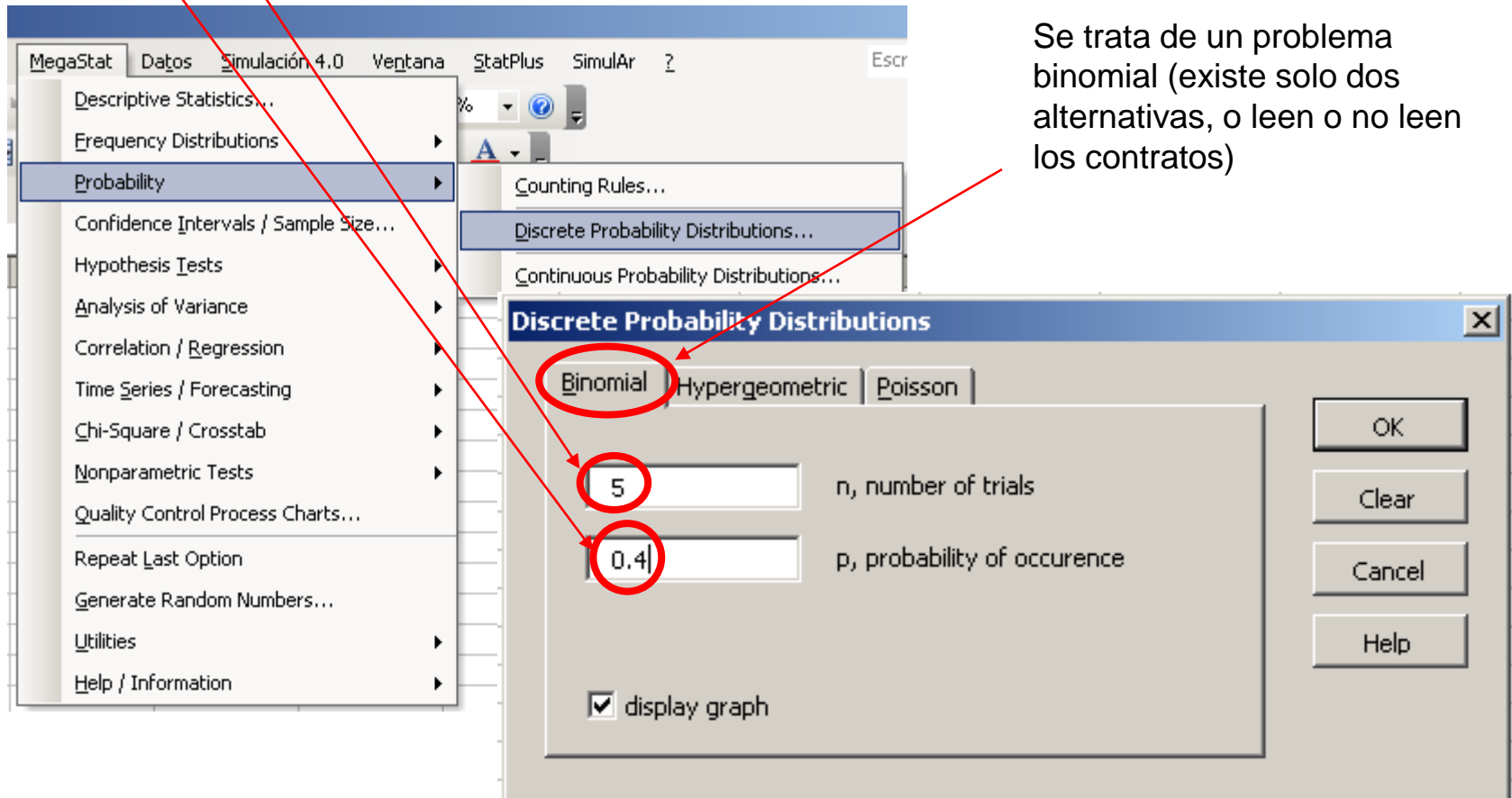
Tablas de contingencias o tablas cruzadas

Crosstabulation

		Gender		Total
		1	2	
AgeCat	1 Observed	7	10	17
	% of row	41.2%	58.8%	100.0%
	% of column	25.0%	45.5%	34.0%
	% of total	14.0%	20.0%	34.0%
	2 Observed	13	8	21
	% of row	61.9%	38.1%	100.0%
	% of column	46.4%	36.4%	42.0%
	% of total	26.0%	16.0%	42.0%
	3 Observed	8	4	12
	% of row	66.7%	33.3%	100.0%
	% of column	28.6%	18.2%	24.0%
	% of total	16.0%	8.0%	24.0%
Total Observed	28	22	50	
% of row	56.0%	44.0%	100.0%	
% of column	100.0%	100.0%	100.0%	
% of total	56.0%	44.0%	100.0%	

Distribución binomial.

El 40% de los peruanos leen su contrato de trabajo, incluyendo las letras pequeñas. Suponga que el número de empleados que leen su contrato se pueden modelar utilizando una distribución binomial. Considerando a un grupo de 5 empleados. ¿Cuál es la probabilidad de que al menos 3 lean su contrato?



The image shows a screenshot of the MegaStat software interface. The 'Probability' menu is open, and the 'Discrete Probability Distributions...' option is selected. The 'Discrete Probability Distributions' dialog box is displayed, with the 'Binomial' tab selected. The number of trials (n) is set to 5, and the probability of occurrence (p) is set to 0.4. The 'display graph' checkbox is checked. The 'OK', 'Clear', 'Cancel', and 'Help' buttons are visible on the right side of the dialog box. Red circles highlight the values 40%, 5, and 0.4 in the original image, with red arrows pointing from the text above to these values.

Se trata de un problema binomial (existe solo dos alternativas, o leen o no leen los contratos)

Distribución binomial.

Binomial distribution

5 n
0.4 p

<i>X</i>	<i>p(X)</i>	<i>cumulative probability</i>
0	0.07776	0.07776
1	0.25920	0.33696
2	0.34560	0.68256
3	0.23040	0.91296
4	0.07680	0.98976
5	0.01024	1.00000

Se quiere saber $P(x \geq 3)$

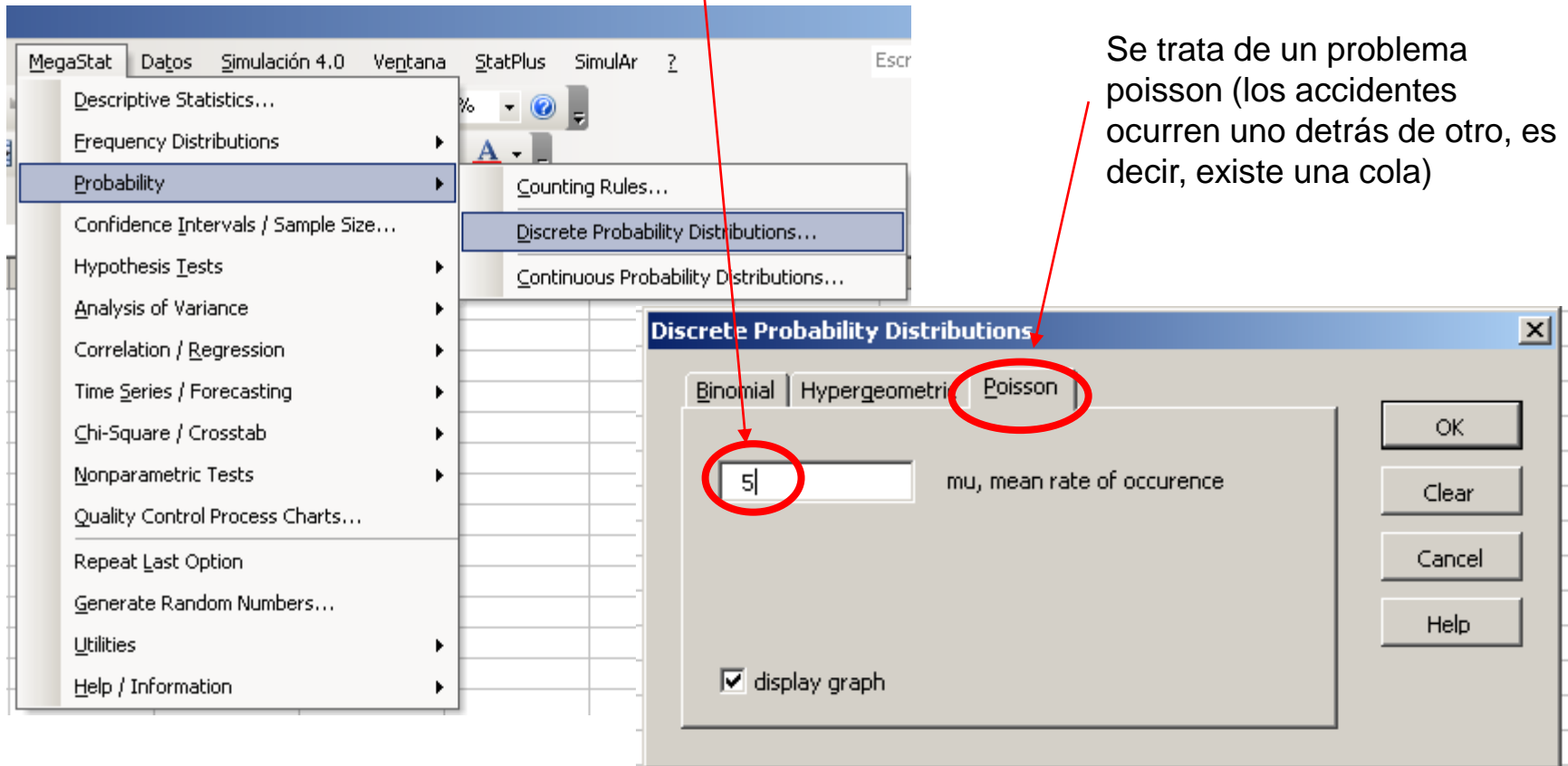
Para esto se debe sumar las probabilidades binomiales de 3, 4 y 5

$$P(x \geq 3) = 0.23040 + 0.07680 + 0.01024$$

$$P(x \geq 3) = 0.31744 = 31.74\%$$

Distribución poisson

Supongamos que estamos investigando la seguridad de una peligrosa intersección de calles, los registros policíacos indican un media de 5 accidentes mensuales en esta intersección. El número de accidentes está distribuido de acuerdo con una distribución de Poisson y el departamento de seguridad vial desea que calculemos la probabilidad de que en cualquier mes ocurra exactamente 3 accidentes.



The image shows a screenshot of the MegaStat software interface. The 'StatPlus' menu is open, and the 'Discrete Probability Distributions...' option is selected. The 'Discrete Probability Distributions' dialog box is displayed, with the 'Poisson' tab selected. The input field for the mean rate of occurrence (mu) contains the value '5'. The 'display graph' checkbox is checked. A red circle highlights the number '5' in the input field, and another red circle highlights the 'Poisson' tab. A red arrow points from the text '5' in the paragraph above to the circled '5' in the dialog box. Another red arrow points from the text 'Se trata de un problema poisson...' to the 'Poisson' tab.

Se trata de un problema poisson (los accidentes ocurren uno detrás de otro, es decir, existe una cola)

Distribución poisson

Poisson distribution

5 mean rate of occurrence

<i>X</i>	<i>p(X)</i>	<i>cumulative probability</i>
0	0.00674	0.00674
1	0.03369	0.04043
2	0.08422	0.12465
3	0.14037	0.26503
4	0.17547	0.44049
5	0.17547	0.61596
6	0.14622	0.76218
7	0.10444	0.86663
8	0.06528	0.93191
9	0.03627	0.96817
10	0.01813	0.98630
11	0.00824	0.99455
12	0.00343	0.99798
13	0.00132	0.99930
14	0.00047	0.99977
15	0.00016	0.99993
16	0.00005	0.99998
17	0.00001	0.99999
18	0.00000	1.00000
19	0.00000	1.00000
20	0.00000	1.00000
21	0.00000	1.00000
22	0.00000	1.00000
	1.00000	

Se quiere saber $P(x = 3)$

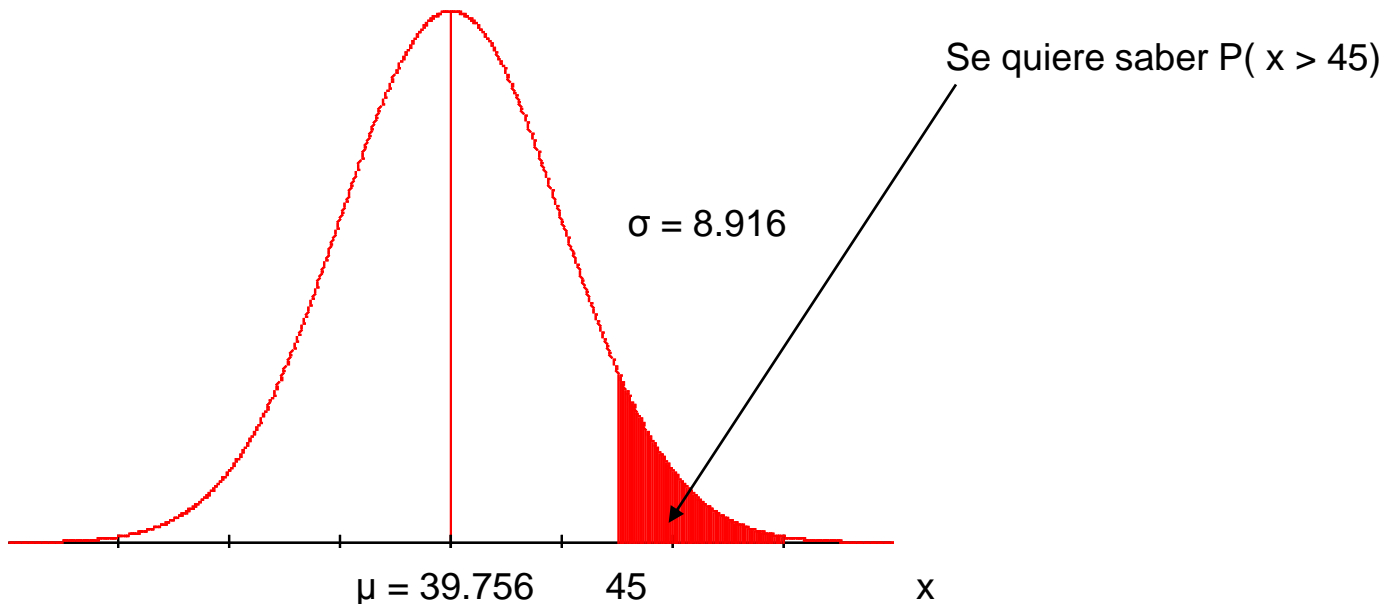
Esto es la probabilidad poisson cuando x es 3

$$P(x = 3) = 0.14037 = 14.04\%$$

Distribución normal

Supongamos que deseamos saber si escogemos a un cliente al azar, ¿Cuál es la probabilidad de que el cliente utilice más de 45 segundos en cajero?

Como el tiempo de uso es una variable cuantitativa continua, corresponde a una distribución normal. Para esto necesitamos conocer la media y la desviación estándar del tiempo del uso del cajero (en las estadísticas descriptivas las calculamos $\mu = 39.756$ seg. y la $\sigma = 8.916$ seg.)



Distribución normal

MegaStat Datos Simulación 4.0 Ventana StatPlus SimulAr ? Escr

Descriptive Statistics...
Frequency Distributions
Probability
Confidence Intervals / Sample Size...
Hypothesis Tests
Analysis of Variance
Correlation / Regression
Time Series / Forecasting
Chi-Square / Crosstab
Nonparametric Tests
Quality Control Process Charts...
Repeat Last Option
Generate Random Numbers...
Utilities
Help / Information

Counting Rules...
Discrete Probability Distributions...
Continuous Probability Distributions...

Continuous Probability Distributions

normal distribution | t distribution | F distribution | chi-square distribution

z 0.59 x 45 calculate probability given x
 calculate x given probability

mean 39.756
standard deviation 8.916

p(lower) .7218 p(upper) .2782 Preview

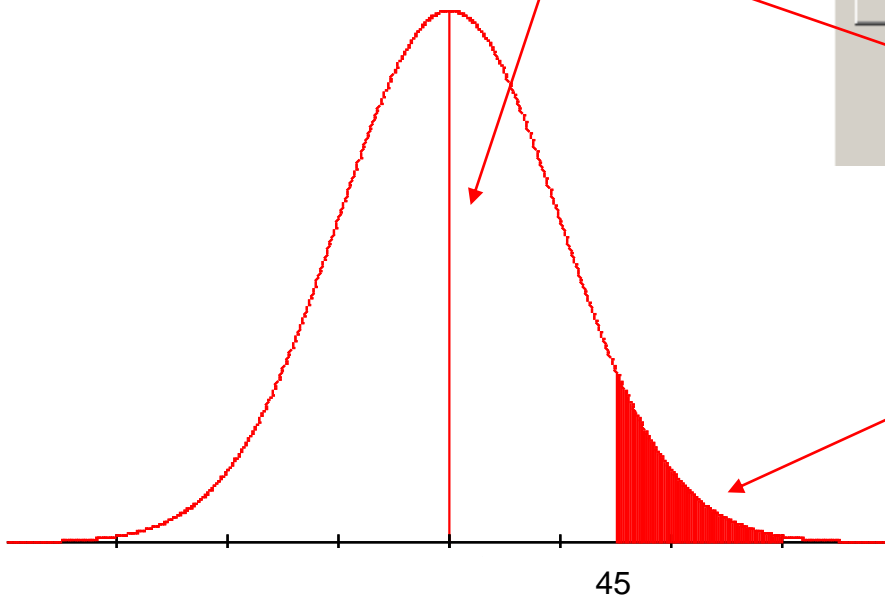
OK
Clear
Cancel
Help

Primero se ingresa el valor de la media, luego el valor de la desviación estándar, luego el valor de x y por último presionamos Preview

Distribución normal

Si observamos el MegaStat me da dos probabilidades: Lower (0.7218) y Upper (0.2782)

La probabilidad **lower** corresponde a la $P(x < 45)$, es decir todo a la izquierda de 45.



Continuous Probability Distributions

normal distribution | t distribution | F distribution | chi-square distribution

z: 0.59 x: 45 calculate probability given x
 calculate x given probability

mean: 39.756
standard deviation: 8.916

p(lower): .7218 p(upper): .2782

Preview

La probabilidad **upper** corresponde a la $P(x > 45)$, es decir todo a la derecha de 45.

Entonces, como nos están pidiendo $P(x > 45)$, esta es 0.2782

$$P(x > 45) = 0.2782 = 27.82\%$$

Intervalos de confianza

Supongamos que deseamos conocer el intervalo de confianza del tiempo de uso del cajero con un 95% de nivel de confianza.

Para esto debo conocer la media, la desviación estándar y el tamaño de la muestra del tiempo de uso del cajero. Esto se obtiene con las estadísticas descriptivas. Media es 39.756, la desviación estándar es 8.916 y el tamaño de la muestra es 50.

The screenshot shows the MegaStat software interface. The 'Confidence Intervals / Sample Size' dialog box is open. The 'Confidence interval - mean' option is selected and circled in red. The 'Mean' field contains 39.756, the 'Std. Dev.' field contains 8.916, and the 'n' field contains 50. The 'Confidence Level' is set to 95%. The 'z' test is selected, also circled in red. The 'lower' bound is 37.285 and the 'upper' bound is 42.227. The 'Preview' button is visible at the bottom left of the dialog box.

Se utiliza prueba z porque el tamaño de la muestra es \geq que 30.

Se puede decir que el promedio de uso de cajero de la población se encuentra entre 37.28 y 42.23 segundos con un 95% de confianza.

Prueba de hipótesis de una muestra

Supongamos que deseamos saber ¿si existe evidencia para aceptar que el tiempo promedio del uso de cajeros es menor a 30 segundos?

Ho	=	≥	≤
Ha	≠	<	>
N° colas	2	1	

Esto es una prueba de hipótesis.

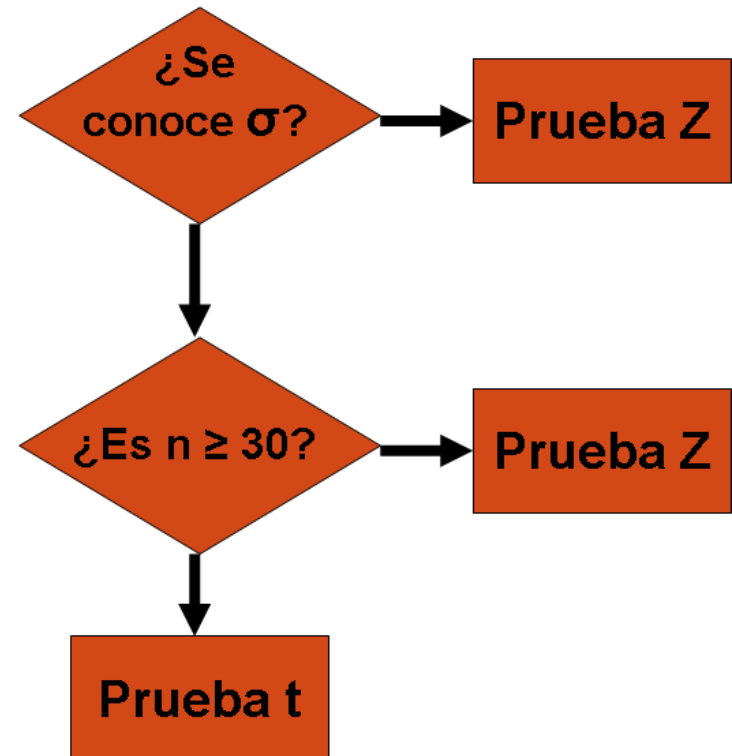
Planteamos primero las hipótesis.

Ho: $\mu \geq 30$

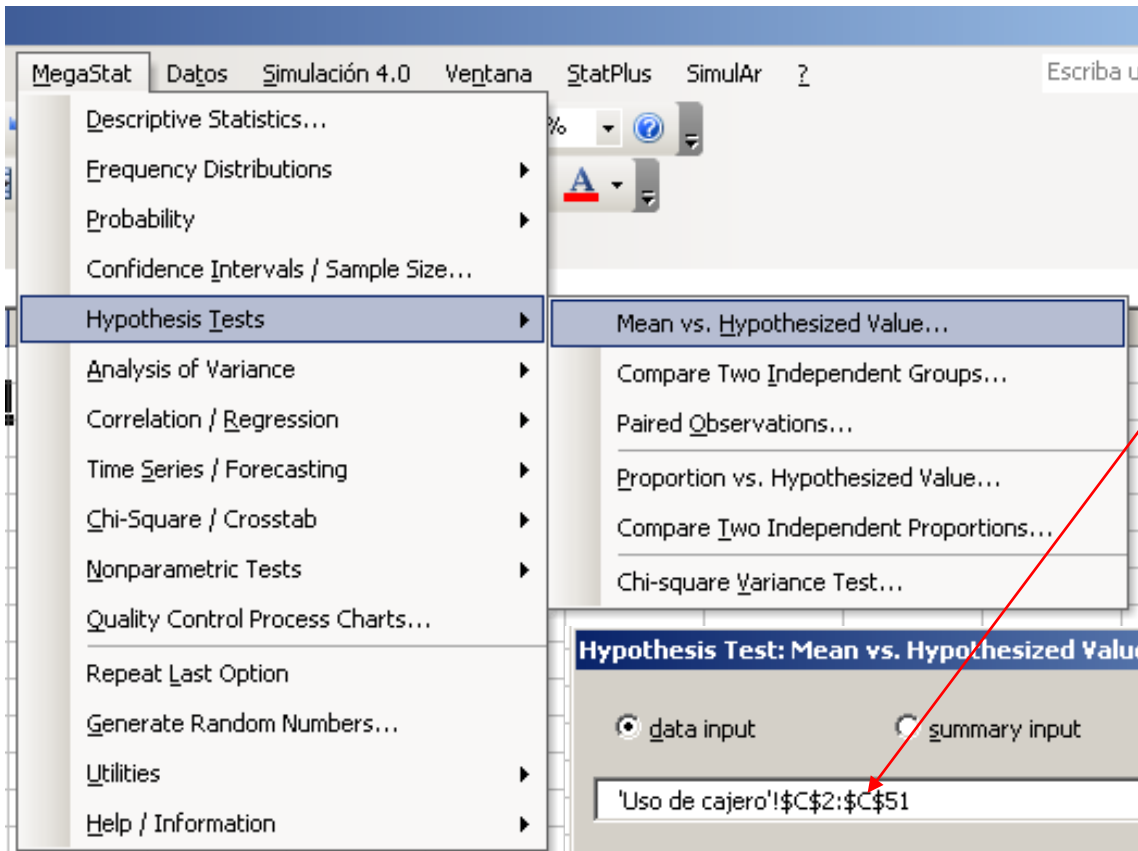
Ha: $\mu < 30$

Ahora tenemos que escoger que prueba utilizamos, para eso tenemos la siguiente regla:

Como no conocemos la σ , pero el tamaño de la muestra es 50 y es mayor que 30, utilizamos prueba z

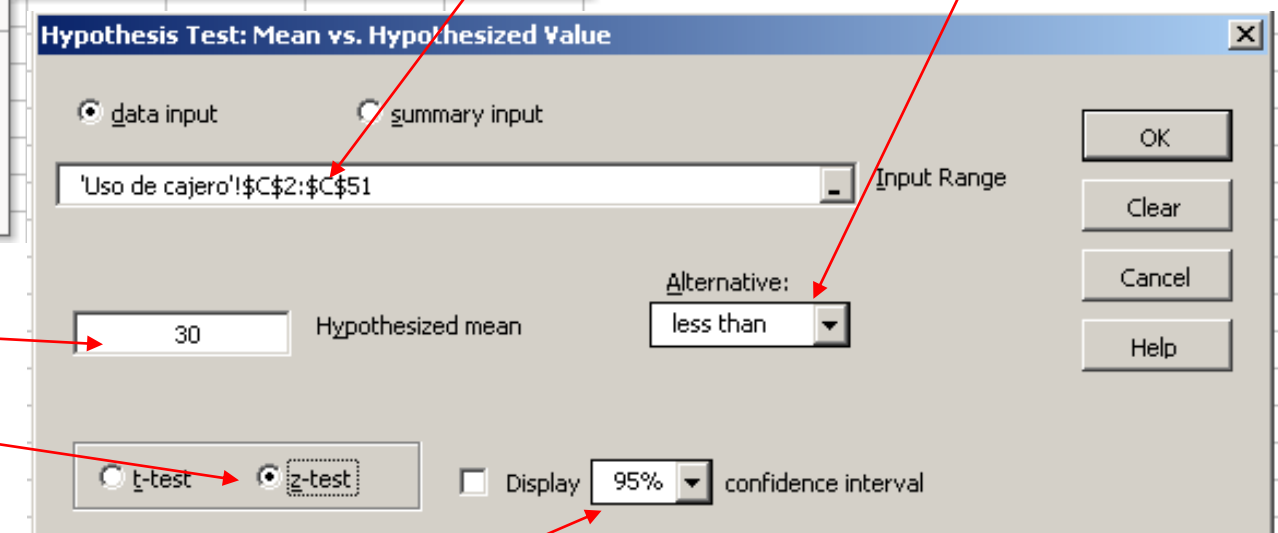


Prueba de hipótesis de una muestra



Ingresamos en rango de datos

La $H_a: \mu < 30$, entonces seleccionamos **less than** (menor que)



Como $H_0: \mu \geq 30$, ingresamos 30

Seleccionamos prueba Z

En este caso $\alpha = 0.05$ (nivel de significancia), le corresponde un nivel de confianza del 95%

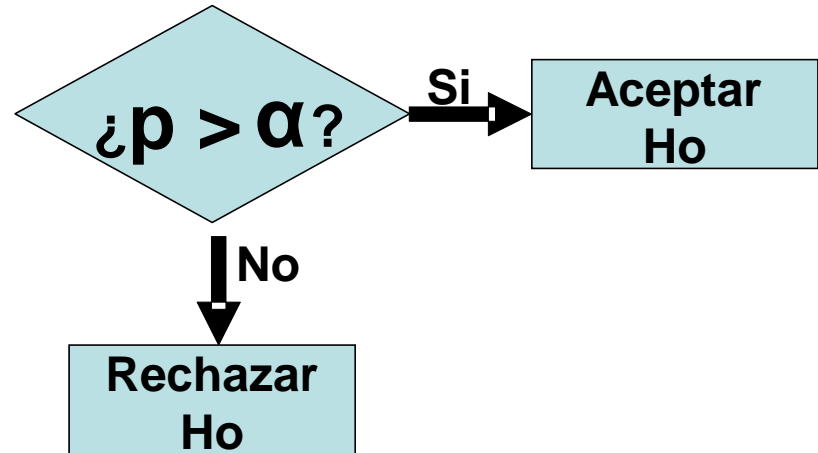
Prueba de hipótesis de una muestra

Hypothesis Test: Mean vs. Hypothesized Value

Usaremos la siguiente regla:

30.0000 hypothesized value
39.7560 mean Seconds
8.9156 std. dev.
1.2609 std. error
50 n

7.74 z
1.0000 p-value (one-tailed, lower)



En este caso p es 1.00 y es mayor que α . Por lo tanto se acepta la H_0 .

$H_0: \mu \geq 30$ Aceptar

$H_a: \mu < 30$ Falso

¿Existe evidencia para aceptar que el tiempo promedio del uso de cajeros es menor a 30 segundos?

Por lo tanto, NO EXISTE evidencia para aceptar que el tiempo promedio del uso del cajero sea menor que 30 segundos.

Prueba de hipótesis de dos muestras.

Supongamos que deseamos saber ¿si existe evidencia para aceptar que **existe** diferencia en el tiempo promedio del uso de cajeros entre hombres y mujeres?

Ho	=	≥	≤
Ha	≠	<	>
N° colas	2	1	

Esto es una prueba de hipótesis de dos muestras independientes.

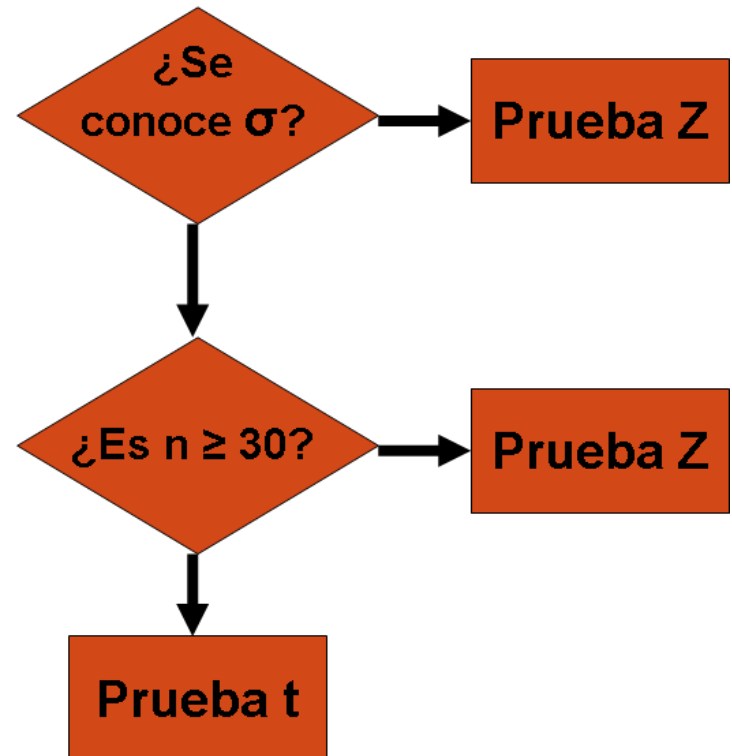
Planteamos primero las hipótesis.

$$H_0: \mu_H = \mu_M$$

$$H_a: \mu_H \neq \mu_M$$

Ahora tenemos que escoger que prueba utilizamos, para eso tenemos la siguiente regla:

Como no conocemos la σ , pero los tamaños de las muestras son menores que 30, utilizamos prueba t

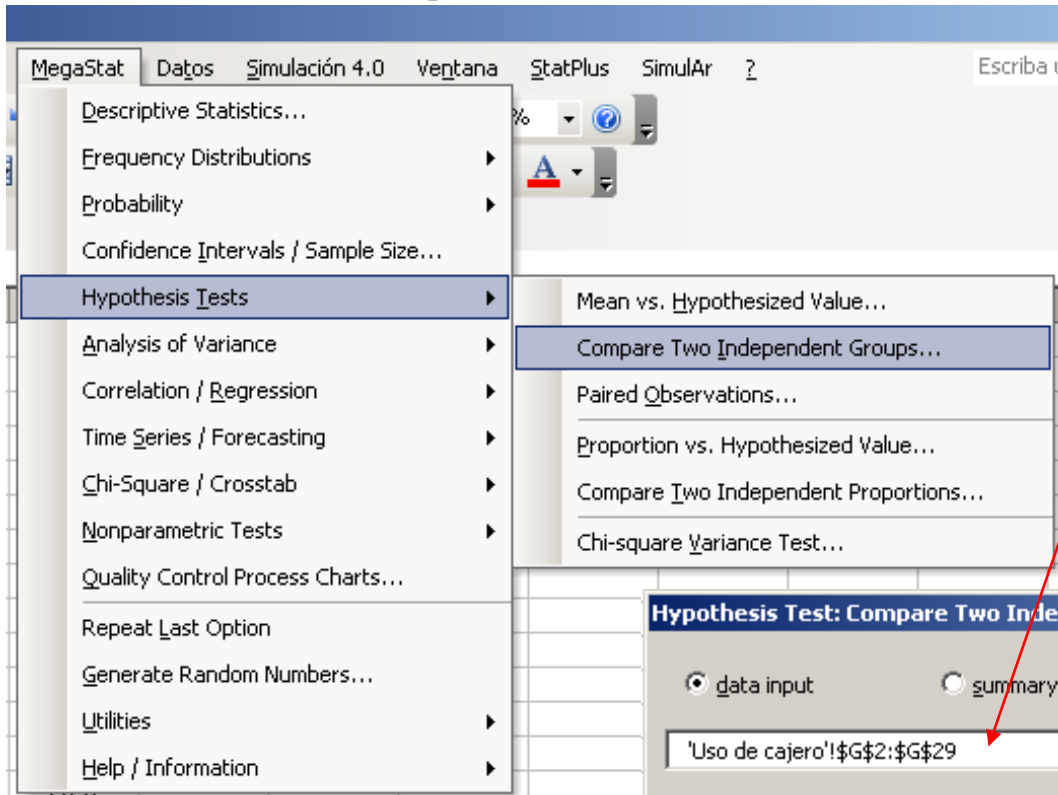


Prueba de hipótesis de dos muestras

	A	B	C	D	G	H
1	AgeCat	Gender	Seconds		Hombres	Mujeres
2	1	1	53.0		53.0	50.1
3	3	1	37.5		37.5	43.2
4	2	1	37.8		37.8	34.9
5	3	1	49.4		49.4	27.6
6	1	1	50.5		50.5	55.6
7	3	1	48.1		48.1	50.8
8	3	1	43.6		43.6	37.7
9	1	1	35.4		35.4	46.4
10	3	1	44.7		44.7	56.3
11	2	1	39.5		39.5	37.8
12	2	1	31.0		31.0	34.4
13	2	1	44.8		44.8	43.6
14	2	1	30.3		30.3	29.9
15	2	1	33.6		33.6	40.1
16	2	1	32.9		32.9	29.1
17	1	1	26.3		26.3	54.9
18	2	1	30.8		30.8	29.4
19	1	1	24.1		24.1	42.9
20	2	1	26.3		26.3	44.3
21	3	1	47.7		47.7	45.0
22	2	1	33.2		33.2	47.6
23	2	1	37.7		37.7	32.0
24	3	1	37.1		37.1	
25	1	1	58.1		58.1	
26	2	1	36.9		36.9	
27	1	1	27.5		27.5	
28	3	1	42.3		42.3	
29	2	1	34.1		34.1	
30	1	2	50.1			
31	2	2	43.2			
32	1	2	34.9			

Para poder utilizar el MegaStat, debemos previamente ordenar los tiempos en función al sexo de los cliente (Gender), esto lo hacemos con el Excel. Y luego copiamos los tiempos en dos columnas, una para los hombres y otra para las mujeres.

Prueba de hipótesis de dos muestras

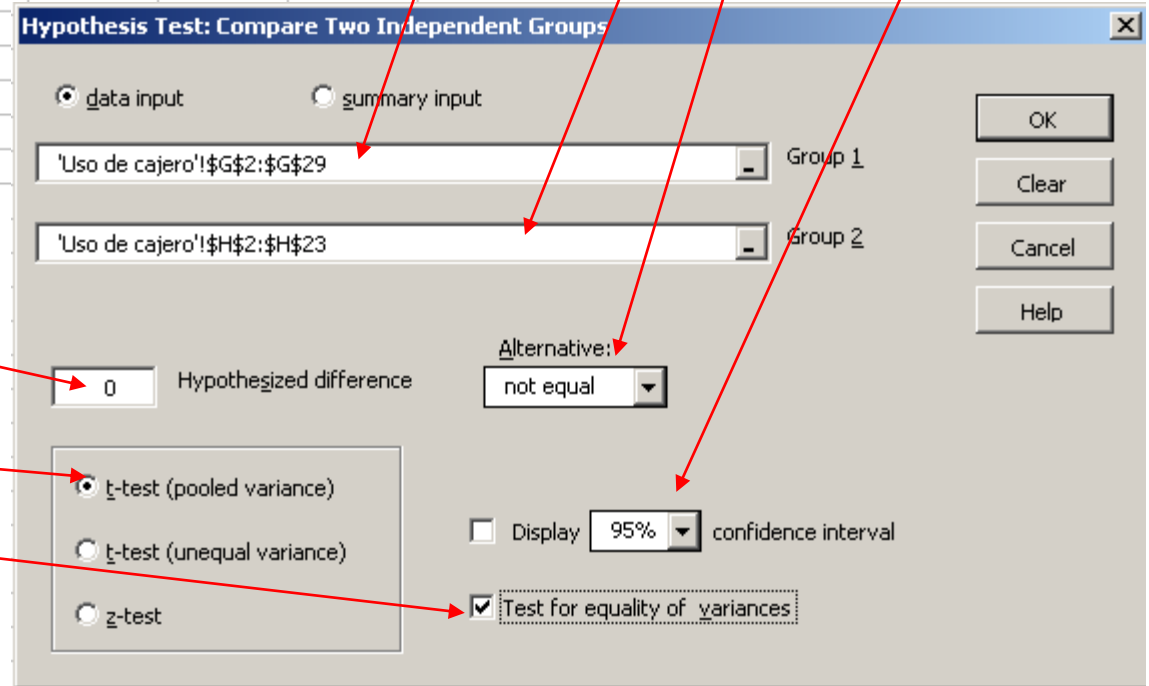


Se ingresa el rango de datos de hombres

Se ingresa el rango de datos de mujeres

La H_a es \neq (no igual)

$\alpha = 0.05$



La diferencia entre los dos grupos es 0 (cero)

Prueba t

Como son muestras independientes, hay que hacer la prueba de igualdad de varianzas para ver si vienen de la misma población.

Prueba de hipótesis de dos muestras

Hypothesis Test: Independent Groups (t-test)

Hombres	Mujeres	
38.364	41.527	mean
8.779	8.973	std. dev.
28	22	n

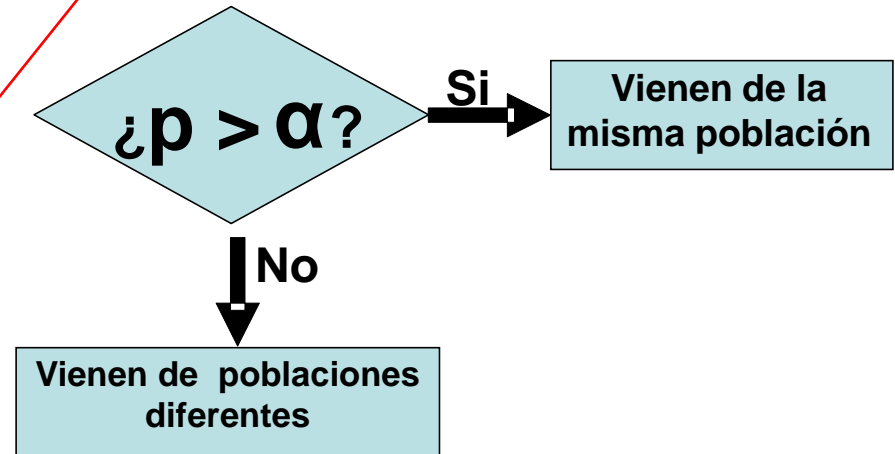
48 df
 -3.1630 difference (Hombres - Mujeres)
 78.5760 pooled variance
 8.8643 pooled std. dev.
 2.5255 standard error of difference
 0 hypothesized difference

-1.25 t
 .2165 p-value (two-tailed)

F-test for equality of variance

80.508 variance: Mujeres
 77.073 variance: Hombres
 1.04 F
 .9028 p-value

Primero evaluamos la igualdad de varianzas para ver si las muestras provienen de la misma población.



Como p es 0.9028 y es mayor que α (0.05), las muestras provienen de la misma población.

Si hubiera sido que viene de diferentes poblaciones, tendríamos que utilizar la prueba t para varianzas diferentes

t-test (pooled variance)

t-test (unequal variance)

z-test

Prueba de hipótesis de dos muestras

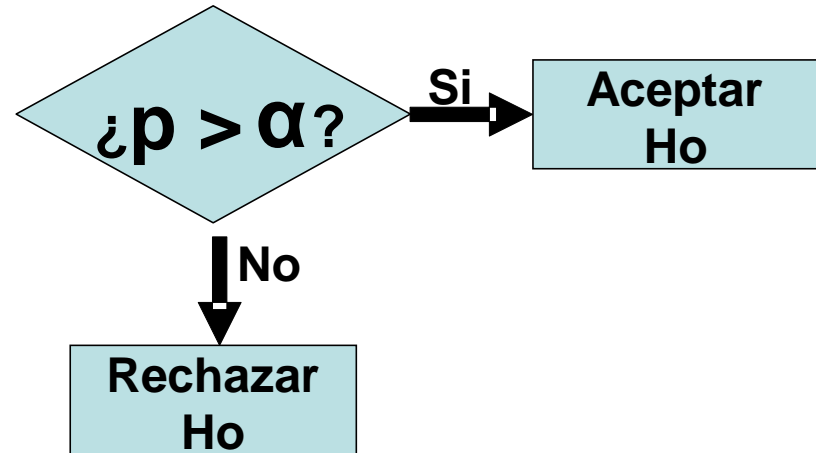
Hypothesis Test: Independent Groups (t-test)

Ahora si evaluamos la prueba de hipótesis.

Hombres	Mujeres	
38.364	41.527	mean
8.779	8.973	std. dev.
28	22	n

48 df
 -3.1630 difference (Hombres - Mujeres)
 78.5760 pooled variance
 8.8643 pooled std. dev.
 2.5255 standard error of difference
 0 hypothesized difference

1.25 t
 .2165 p-value (two-tailed)



Como p es 0.2165 y es mayor que α (0.05), se acepta la H_0 .

$H_0: \mu_H = \mu_M$ Aceptar

$H_a: \mu_H \neq \mu_M$ Falso

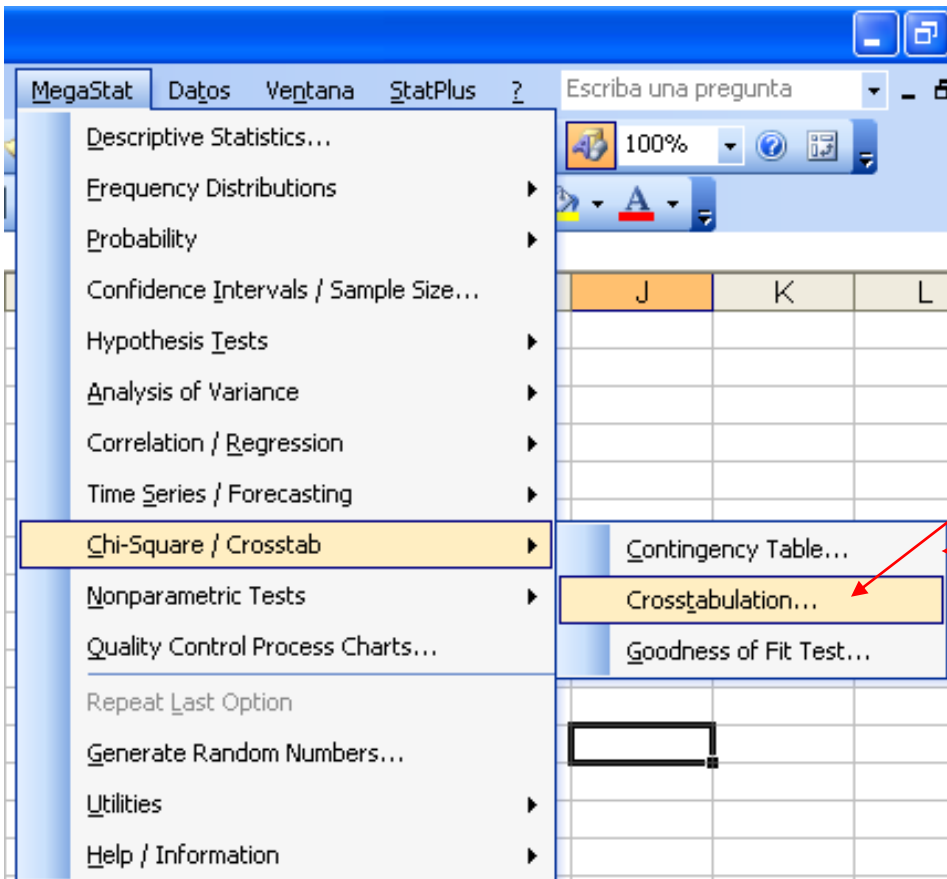
¿si existe evidencia para aceptar que existe diferencia en el tiempo promedio del uso de cajeros entre hombres y mujeres?

Por lo tanto NO EXISTE evidencia para aceptar que hay diferencia entre el tiempo de uso del cajero entre hombre y mujeres. ¡

Prueba de Chi cuadrado de independencia

Supongamos que deseamos saber si existe relación entre las variables **Agecat** (categorías por edad) y **Gender** (sexo), en nuestro ejemplo.

Para poder hacer una prueba de Chi cuadrado, se requiere que las dos variables sean cualitativas (nominal ó ordinal). En nuestro ejemplo, tanto las variables Agecat y Gender son cualitativas.



Como no existe una tabla de contingencias (o tabla cruzada), tenemos que construir la respectiva tabla. Para eso, utilizaremos Crosstabulation.

Si hubiéramos tenido una tabla de contingencia, utilizaremos Contingency Tabla

Se ingresa el rango de datos de la variable que va en la fila, en nuestro ejemplo: Gender.

Se ingresa el rango de la calificación de la variable Gender (1 y 2)

Se selecciona la prueba de Chi -cuadrado

Se ingresa el rango de la calificación de la variable Agecat (1, 2 y 3)

Se ingresa el rango de la calificación de la variable Agecat (1, 2 y 3)

Se ingresa el rango de datos de la variable que va en la columnas, en nuestro ejemplo: Agecat.

Se selecciona la prueba de Chi -cuadrado

Como la variable Agecat es ordinal, se escoge el Coeficiente de contingencia para ver la fuerza de la relación (si las variables son nominales se utiliza Coeficiente Phi)

Prueba de Chi cuadrado de independencia

Crosstabulation

		AgeCat			Total
		1	2	3	
Gender	1	7	13	8	28
	2	10	8	4	22
Total		17	21	12	50

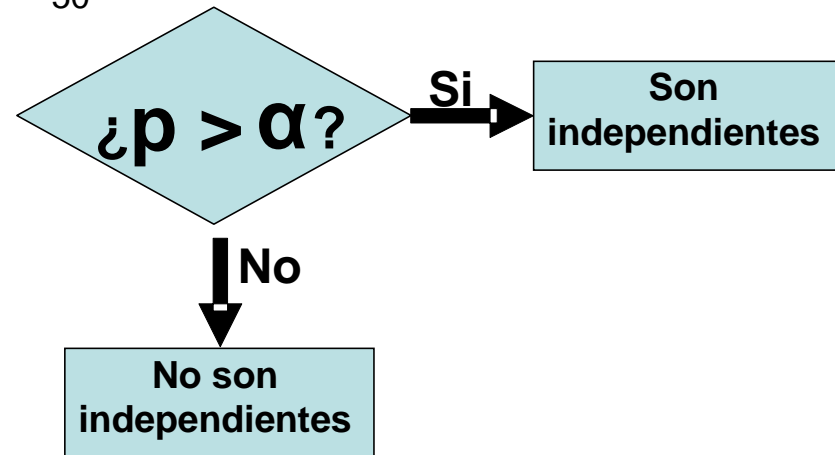
2.37 chi-square

2 df

.3062 p-value

.213 Coefficient of Contingency

Utilizaremos la siguiente regla:



Como p es 0.213 y es mayor que α (0.05), las variables Agecat y Gender son independientes.

Análisis de varianza de un factor

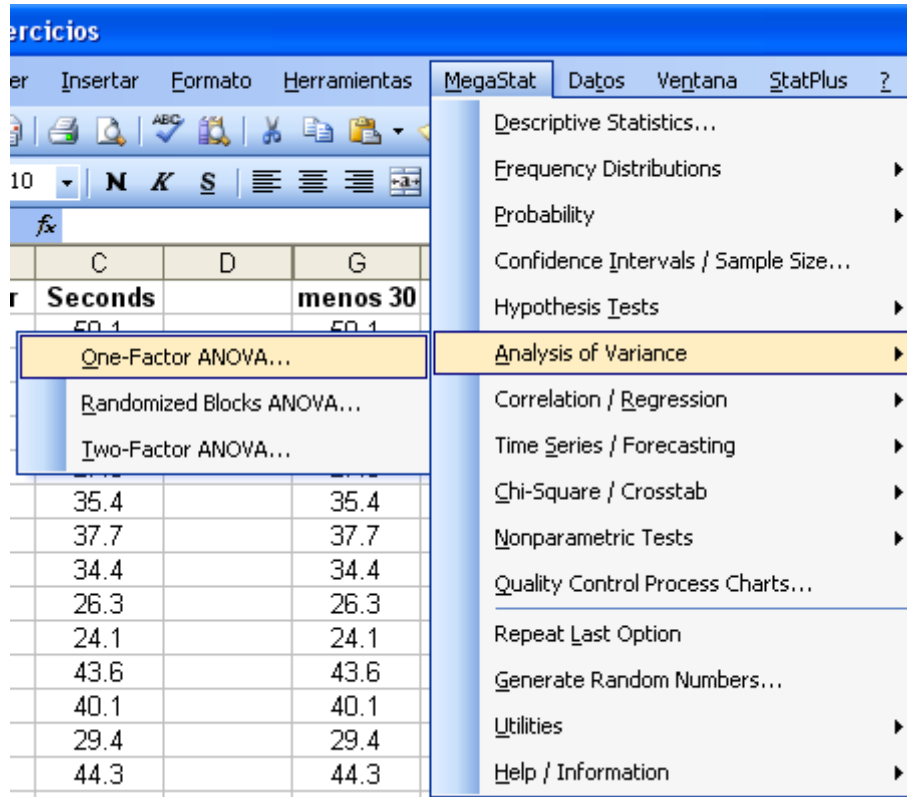
Supongamos que deseamos saber si existe diferencia en el tiempo del uso del cajero de acuerdo a la categoría de edad.

Tenemos una variable cuantitativa (Tiempo) y tres grupos (Egecat), por lo tanto tenemos que utilizar el ANOVA, como solo se evalúa el tiempo, entonces es de un factor.

	A	B	C	D	G	H	I
1	AgeCat	Gender	Seconds		menos 30	30 a 50	mas 50
2	1	2	50.1		50.1	43.2	37.5
3	1	1	53.0		53.0	37.8	49.4
4	1	2	34.9		34.9	50.8	48.1
5	1	1	50.5		50.5	39.5	55.6
6	1	2	27.6		27.6	46.4	43.6
7	1	1	35.4		35.4	31.0	44.7
8	1	2	37.7		37.7	44.8	56.3
9	1	2	34.4		34.4	30.3	47.7
10	1	1	26.3		26.3	33.6	37.1
11	1	1	24.1		24.1	37.8	42.9
12	1	2	43.6		43.6	32.9	45.0
13	1	2	40.1		40.1	30.8	42.3
14	1	2	29.4		29.4	29.9	
15	1	2	44.3		44.3	26.3	
16	1	1	58.1		58.1	33.2	
17	1	2	32.0		32.0	29.1	
18	1	1	27.5		27.5	37.7	
19	2	2	43.2			54.9	
20	2	1	37.8			47.6	
21	2	2	50.8			36.9	

Para poder utilizar el MegaStat, debemos previamente ordenar los tiempos en función a la variable Agecat, esto lo hacemos con el Excel. Y luego copiamos los tiempos en tres columnas, una para los menores de 30 (1), otra para los que están entre 30 y 50 (2) y la ultima para los que tiene mas de 50 (3)

Análisis de varianza de un factor

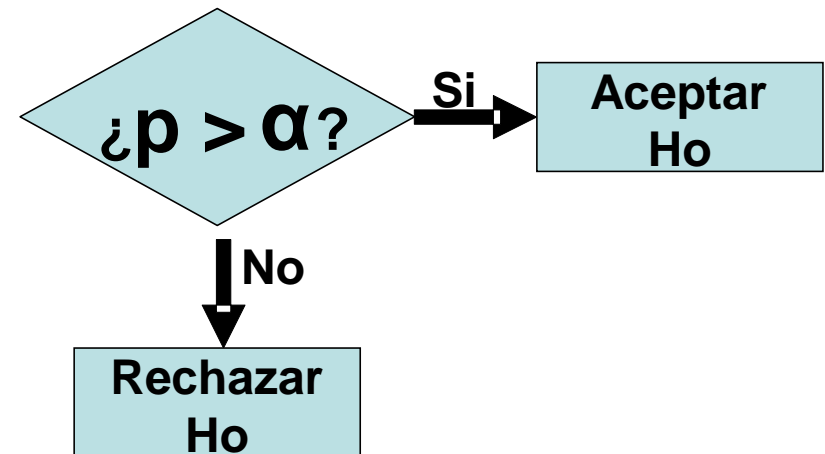


Las hipótesis de una ANOVA, son:

H_0 : Los promedios son iguales

H_a : Al menos una es diferente

Y se utiliza la siguiente regla de decisión:



Análisis de varianza de un factor

Gender	Seconds	menos 30	30 a 50	mas 50
2	50.1	50.1	43.2	37.5
1	53.0	53.0	37.8	49.4

2	27.5	27.5	28.7
2	43.2		54.9
1	37.8		47.6

Analysis of Variance: One-Factor ANOVA

'Uso de cajero!\$G\$2:\$I\$22' Input range

Post-Hoc Analysis

When p < .05 Never Always

Plot Data

OK
Clear
Cancel
Help

Se ingresa el rango de datos que incluye a las tres columnas

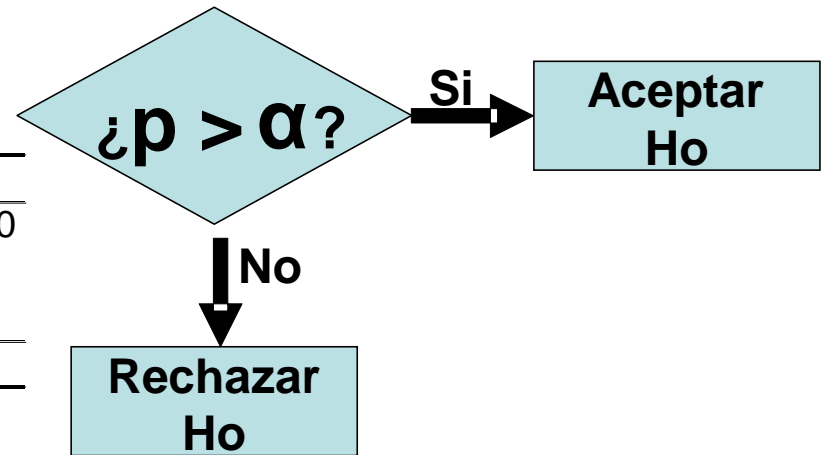
Análisis de varianza de un factor

One factor ANOVA

<i>Mean</i>	<i>n</i>	<i>Std. Dev</i>	
38.18	17	10.291	menos 30
37.55	21	7.779	30 a 50
45.85	12	6.031	mas 50
39.76	50	8.916	Total

ANOVA table

<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p-value</i>
Treatment	590.030	2	295.0151	4.20	.0211
Error	3,304.873	47	70.3164		
Total	3,894.903	49			



Como p vale 0.0211 y es menor que α (0.05), se rechaza la Ho.

Ho: Los promedios son iguales Rechaza

Ha: Al menos una es diferente **Verdadero**

¿Existe diferencia en el tiempo del uso del cajero de acuerdo a la categoría de edad?.

Por lo tanto, SI EXISTE evidencia de los tiempo de uso de los cajeros de acuerdo a la categoría de edad, son diferentes.

Análisis de regresión lineal simple

Copy SA, empresa que tiene una gran fuerza de ventas en todo EEUU y Canadá, desea determinar si existe una relación entre el número de llamadas telefónicas de ventas hechas en un mes, y la cantidad de copadoras vendidas durante ese lapso. El gerente selecciona al azar una muestra de 10 representantes, y determina el número de tales llamadas que hizo cada uno en el mes anterior y la cantidad de productos vendidos.

Nº de llamadas	Nº copadoras vendidas
20	30
40	60
20	40
30	60
10	30
10	40
20	40
20	50
20	30
30	70

Deseamos saber, si existe relación entre el N° de llamadas y las copadoras vendidas (ambas variables son cuantitativas). Y si existe relación, como poder pronosticar mis ventas a partir del numero de llamadas.

Esto lo puedo contestar con el análisis de correlación y regresión.

La variable que deseo pronosticar, es la variable dependiente **Y**. En nuestro ejemplo es en N° de copadoras vendidas.

La variable que es mi información , es la variable independiente **X**. En nuestro ejemplo es en N° de llamadas. Como es una sola variable independiente, se utiliza una **regresión lineal simple**.

Análisis de regresión lineal simple

The screenshot shows the MegaStat software interface. The 'MegaStat' menu is open, and the 'Correlation / Regression' option is selected. A sub-menu is displayed, showing 'Regression Analysis...' as the chosen option. In the background, a spreadsheet contains data for two variables: 'Nº copadoras' (Y) and 'Nº llamada' (X).

C	D
Y	X
Nº copadoras	Nº llamada
30	20
60	40
40	20
30	30
30	10
40	10
40	20
50	20
30	20
70	30

Se ingresa el rango de datos de la variable independiente X, el N° llamadas

Regression Analysis

Input ranges:

Hoja3!\$D\$3:\$D\$12 Independent variable(s)

Hoja3!\$C\$3:\$C\$12 Dependent variable

No predictions

predictor values

Options

95% Confidence Level

Variance Inflation Factors

Standardized Coefficients (betas)

Test Intercept Force Zero Intercept

All Possible Regressions

Stepwise Selection 1 best model of each size

Residuals:

Output Residuals

Diagnostics and Influential Residuals

Durbin-Watson

Plot Residuals by Observation

Plot Residuals by Predicted Y and X

Normal Probability Plot of Residuals

	C	D
1	Y	X
2	Nº copadoras	Nº llamadas
3	30	20
4	60	40
5	40	20
6	30	30
7	30	10
8	40	10
9	40	20
10	50	20
11	30	20
12	70	30

Se ingresa el rango de datos de la variable dependiente Y, el N° copadoras vendidas.

Análisis de regresión lineal simple

Regression Analysis

r^2 0.576 n 10
 r 0.759 k 1
Std. Error 9.901 Dep. Var. **Nº copadoras**

ANOVA table

<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p-value</i>
Regression	1,065.7895	1	1,065.7895	10.87	.0109
Residual	784.2105	8	98.0263		
Total	1,850.0000	9			

Regression output

<i>variables</i>	<i>coefficients</i>	<i>std. error</i>	<i>t (df=8)</i>	<i>p-value</i>	<i>confidence interval</i>	
					<i>95% lower</i>	<i>95% upper</i>
Intercept	18.9474	8.4988	2.229	.0563	-0.6509	38.5457
Nº llamadas	1.1842	0.3591	3.297	.0109	0.3560	2.0124

Ahora interpretaremos los resultados.

Análisis de regresión lineal simple

r^2 0.576

r 0.759

r (coeficiente de correlación), es 0.759, lo que me indica una correlación regular entre las variables.

r^2 (coeficiente de determinación), me explica el porcentaje (57.6%) de la variable dependiente (Nº de copadoras vendidas), es explicada por la variable independiente (el Nº de llamadas)

ANOVA table

<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p-value</i>
Regression	1,065.7895	1	1,065.7895	10.87	.0109
Residual	784.2105	8	98.0263		
Total	1,850.0000	9			

La prueba de ANOVA, me sirve para ver si la correlación es real o ficticia. Si la prueba p es menor que α (0.05), la correlación es real, caso contrario es ficticia.

En nuestro caso, p es 0.0109 y es menor que α (0.05), por lo tanto la correlación es real.

Análisis de regresión lineal simple

Se le llama el análisis de regresión lineal simple, porque es la función de una recta del tipo: $Y = a + bX$

Regression output					<i>confidence interval</i>	
<i>variables</i>	<i>coefficients</i>	<i>std. error</i>	<i>t (df=8)</i>	<i>p-value</i>	<i>95% lower</i>	<i>95% upper</i>
Intercept	18.9474	8.4988	2.229	.0563	-0.6509	38.5457
Nº llamadas	1.1842	0.3591	3.297	.0109	0.3560	2.0124

Nº de copadoras = **a** + **b** Nº de llamadas

La función de regresión es:

Nº de copadoras = 18.9474 + 1.1842 Nº de llamadas

Análisis de regresión lineal múltiple

r^2 (coeficiente de determinación), me explica el porcentaje (32.6%) de la variable dependiente (Nº de copadoras vendidas), es explicada por la variable independiente (el Nº de llamadas).

La pregunta es: ¿Ud. estaría conforme con este resultado?

Este valor de r^2 me indica que falta una o mas variables independientes para poder pronosticar el Nº de copadoras vendidas con mayor precisión. Supongamos que agregamos la variable Publicidad.

Nº Copadoras vendidas	Nº llamadas	Publicidad
30	20	25
60	40	50
40	20	35
60	30	50
30	10	40
40	10	50
40	20	50
50	20	60
30	20	40
70	30	80

Como existen dos variables independientes: Nº llamadas (X1) y Publicidad (X2), se utiliza una **regresión lineal múltiple**.

$$Y = a + b X1 + c X2$$

Se ingresa el rango de datos de las variables independientes X1 y X2, el N° llamadas y publicidad

Regression Analysis

Input ranges:

X, Independent variable(s): Hoja3!\$D\$3:\$E\$12

Y, Dependent variable: Hoja3!\$C\$3:\$C\$12

No predictions

predictor values

Options:

- 95% Confidence Level
- Variance Inflation Factors
- Standardized Coefficients (betas)
- Test Intercept
- Force Zero Intercept
- All Possible Regressions
- Stepwise Selection (1 best model of each size)

Residuals:

- Output Residuals
- Diagnostics and Influential Residuals
- Durbin-Watson
- Plot Residuals by Observation
- Plot Residuals by Predicted Y and
- Normal Probability Plot of Residuals

	C	D	E
1	Y	X1	X2
2	Nº copadoras	Nº llamadas	Publicidad
3	30	20	25
4	60	40	50
5	40	20	35
6	30	30	50
7	30	10	40
8	40	10	50
9	40	20	50
10	50	20	60
11	30	20	40
12	70	30	80

Se ingresa el rango de datos de la variable dependiente Y, el N° copadoras vendidas

Análisis de regresión lineal múltiple

Regression Analysis

R ²	0.902	n	10
Adjusted R ²	0.874	k	2
R	0.950	Dep. Var.	Nº copadoras
Std. Error	5.085		

ANOVA table

Source	SS	df	MS	F	p-value
Regression	1,668.9655	2	834.4828	32.27	.0003
Residual	181.0345	7	25.8621		
Total	1,850.0000	9			

Regression output

variables	coefficients	std. error	t (df=7)	p-value	confidence interval	
					95% lower	95% upper
Intercept	-1.7241	6.1137	-0.282	.7861	-16.1808	12.7326
Nº llamadas	0.8448	0.1974	4.280	.0037	0.3780	1.3116
Publicidad	0.5862	0.1214	4.829	.0019	0.2992	0.8732

Ahora interpretaremos los resultados.

Análisis de regresión lineal múltiple

R^2 0.902
Adjusted R^2 0.874
 R 0.950

r (coeficiente de correlación), es 0.950, mejoro tremendamente (antes era 0.7590) lo que me indica una correlación muy buena entre las variables.

r^2 (coeficiente de determinación), es ahora 0.902, que el porcentaje (90.2%) de la variable dependiente (Nº de copadoras vendidas), es explicada por las variables independientes (el Nº de llamadas y la publicidad)

ANOVA table

<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p-value</i>
Regression	1,668.9655	2	834.4828	32.27	.0003
Residual	181.0345	7	25.8621		
Total	1,850.0000	9			

En nuestro caso, p es 0.0003 y es menor que α (0.05), por lo tanto la correlación es real.

Análisis de regresión lineal múltiple

Regression output					<i>confidence interval</i>	
<i>variables</i>	<i>coefficients</i>	<i>std. error</i>	<i>t (df=7)</i>	<i>p-value</i>	<i>95% lower</i>	<i>95% upper</i>
Intercept	-1.7241	6.1137	-0.282	.7861	-16.1808	12.7326
Nº llamadas	0.8448	0.1974	4.280	.0037	0.3780	1.3116
Publicidad	0.5862	0.1214	4.829	.0019	0.2992	0.8732

Si el valor de p de la variable independiente es menor que α (0.05), entonces el aporte de la variable es significativo. En nuestro caso, ambos p son menores que α , por lo tanto el aporte de las variables es significativo.

Regression output					<i>confidence interval</i>	
<i>variables</i>	<i>coefficients</i>	<i>std. error</i>	<i>t (df=7)</i>	<i>p-value</i>	<i>95% lower</i>	<i>95% upper</i>
Intercept	-1.7241	6.1137	-0.282	.7861	-16.1808	12.7326
Nº llamadas	0.8448	0.1974	4.280	.0037	0.3780	1.3116
Publicidad	0.5862	0.1214	4.829	.0019	0.2992	0.8732

$$Y = a + b X_1 + c X_2$$

$$N^{\circ} \text{ copadoras} = -1.7241 + 0.8448 N^{\circ} \text{ llamadas} + 0.5862 \text{ Publicidad}$$