Apuntes de Probabilidad y Estadística para Ingeniería y Administración

Ignacio Vélez Pareja
Decano
Facultad de Ingeniería Industrial
Politécnico Grancolombiano
Bogotá, Colombia
Octubre, 2002

Conceptos básicos de estadística

Hay tres clases de mentiras: las mentiras, las malditas mentiras y la estadística.
There are three kinds of lies: lies, damned lies, and statistics.
Benjamin Disraeli (1804 - 1881)

Estadístico: Una persona que cree que las cifras no mienten, pero que admite que si se analizan algunas de ellas no tendrían fundamento.
Statistician: A man who believes figures don't lie, but admits that under analysis some of them won't stand up either.
Evan Esar (1899 - 1995), Esar's Comic Dictionary
Estadística: la única ciencia que les permite a expertos diferentes, sacar conclusiones diferentes usando las mismas cifras.
Statistics: The only science that enables different experts using the same figures to draw different conclusions.
Evan Esar (1899 - 1995), Esar's Comic Dictionary
"Hay gente que utiliza la Estadística como un borracho utiliza el poste de la luz; más para apoyarse que para iluminarse".
Andrew Lang (Citado por Thomas y Ronald Wonnacott)

La estadística es un método científico de análisis que se aplica a las ciencias sociales y naturales. Su principal utilización es la inferencia estadística, esto es, que a partir de la información obtenida de una muestra -reducido número de observaciones de un universo- se hacen "inferencias" sobre la población total.

Estadística Descriptiva

Lo primero que se debe hacer con la información obtenida de una muestra, es reducirla a unas cuantas cifras que condensen o concentren la información más importante. Estas cifras se conocen como las estadísticas de la muestra.

Obsérvese la diferencia entre Estadística, área del conocimiento que permite hacer inferencia sobre poblaciones, y la estadística de una muestra, que es una cifra que describe a esa muestra o al universo. También debe distinguirse entre la estadística de una muestra y el parámetro que describe a un universo. Lo que se calcula para una muestra son las estadísticas de la muestra que pueden servir para calcular o hacer una estimación de los parámetros del universo.

Supóngase que se ha obtenido una muestra de la audiencia de una cierta población y se indaga sobre hábitos de lectura. Si la muestra que se obtuvo es de 1.000 personas y 345 de ellas responden que no leen, entonces una forma de describir la muestra es diciendo

que el 34,5% de ella no lee. Esta cifra puede ser utilizada para hacer una inferencia de la población en cuanto a los hábitos de lectura.

Ahora bien, los datos que se obtienen no pueden ser utilizados sin un previo análisis y sin reserva. Por lo general, cuando se toma una muestra se incurre en algún tipo de error estadístico, el cual tiene que ver con el tamaño de la muestra; intuitivamente es obvio que si se tiene un universo muy grande, a mayor información que se obtenga -mayor tamaño de la muestra- más cerca de la realidad van a estar las estadísticas de la muestra, comparadas con las estadísticas del universo. Los técnicos reconocen entonces un margen de error, y se dice que un dato tiene un margen de error. Por ejemplo, los datos de preferencia de votos en una campaña electoral se expresan como que el 65% de la gente votará por el candidato *A*, con un margen de error de ±5%. Esto es, que el verdadero valor se estima que está dentro del intervalo 60%-70% y esta afirmación tiene una determinada probabilidad de que sea cierta; se dice entonces que es un intervalo de *P*% de confianza (por ejemplo, 95% de confianza).

Distribuciones de probabilidad

Un universo tiene unas características estadísticas que, como se dijo arriba, pueden especificarse con unas cuantas cifras. Así mismo, los universos tienen dos características básicas: son discretos cuando los valores que toman las unidades que lo configuran toman valores finitos, por ejemplo, el número de años en que los ingresos son cero, 0, 1, 2, 3, etc. o son continuos lo cual significa que pueden tomar un infinito número de valores, como puede ser el espesor de una lámina de acero o, con mucho rigor, la edad del ser humano.

Histogramas y Tablas

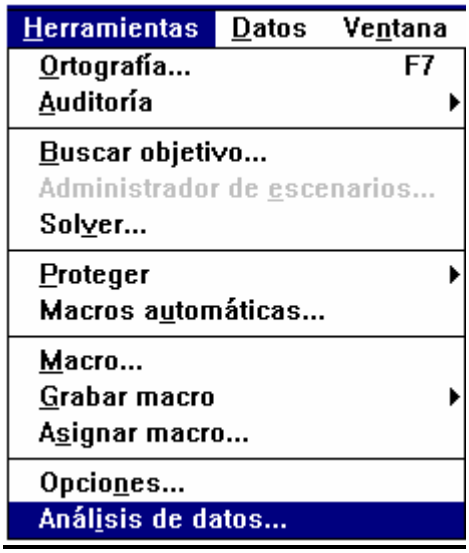
Una manera de visualizar la información de una muestra es tabularla o mostrar la gráfica de los valores obtenidos.

Caso discreto

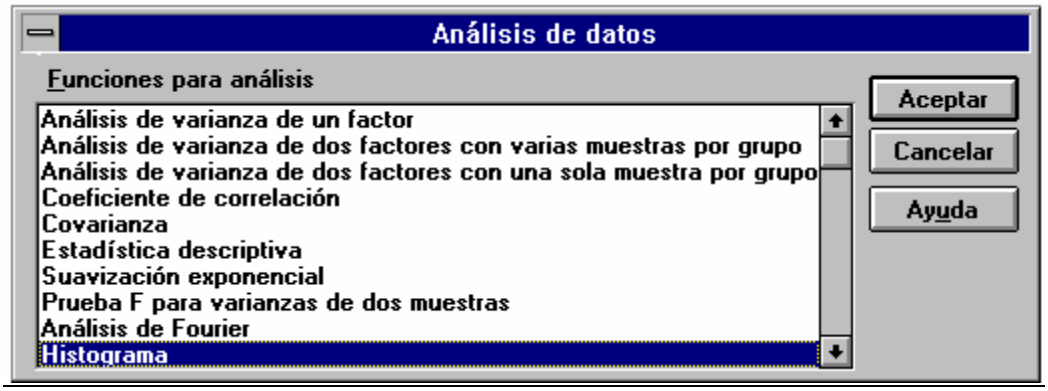
Suponga que se hace una muestra 6.400 viviendas de un país. (Esto puede hacerse con facilidad si se tiene acceso a los formularios de un censo de población o de una manera más compleja, construyendo una muestra aleatoria de las 6.400 viviendas, localizándolas y visitándolas para verificar los datos). La muestra indica que en las viviendas el número de habitaciones es de 1, 2, 3, 4, 5 ó 6.

 En el archivo *Estadística.XLS* en la hoja *DISTRIVI* se encuentran los valores de una muestra de 6.400 viviendas en términos del número de habitaciones de cada vivienda. Estos datos están en el rango *A1:J640* y a ellos se hace referencia en la siguiente tabla. Se debe advertir que según los manuales de *Excel*, esta función no puede manejar más de 6.400 observaciones; sin embargo, el autor ha trabajado con 10.000 y se han obtenido resultados satisfactorios, excepto en la configuración de la fórmula. Para ilustrar el uso de la función se presenta un ejemplo:

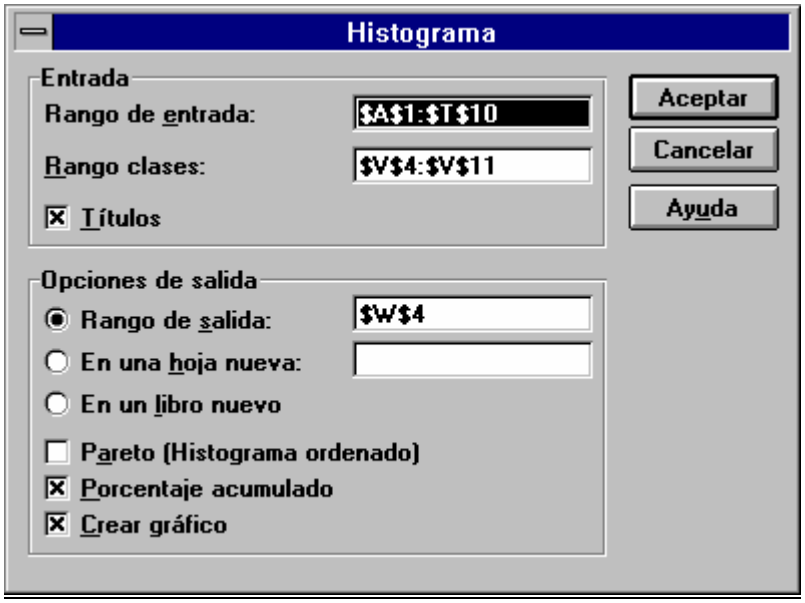
Si se tienen los siguientes datos y se desea calcular cuántas veces aparecen viviendas con 1,2,3,4,5 o 6 habitaciones, se debe usar la función *=FRECUENCIA(Datos;grupos)* o en la Barra de Herramientas de *Excel*, se oprime *Herramientas* y aparece el menú que se muestra a continuación. Allí se escoge *Análisis de Datos*. Esta opción se explicará en detalle**.**



Cuando se hace *click* con el *Mouse* en *Análisis de Datos*, aparece el siguiente cuadro de diálogo:



Allí se selecciona *Histograma* y aparece el cuadro de diálogo siguiente:



Se le debe indicar al programa en qué rango se hallan los datos, dónde insertar los grupos o rango de clases y cuál ha de ser el rango de salida. Además se debe indicar si se desea el porcentaje acumulado, además del histograma de frecuencias absolutas y si se desea hacer una gráfica. Hecho esto, se oprime el botón 

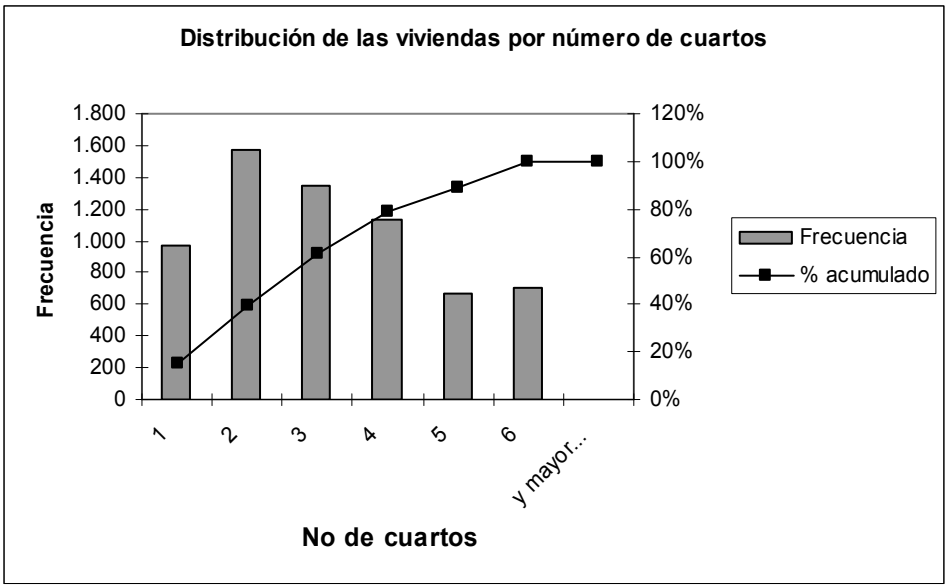
. También se puede pedir que construya una curva de *Pareto*, la cual consiste en ordenar los valores de mayor a menor frecuencia. Todo esto lo hace *Excel* en forma automática.

Con los datos del archivo *Estadística.XLS*, en la hoja *DISTRIVI* se introdujeron los rangos en el cuadro de diálogo anterior.

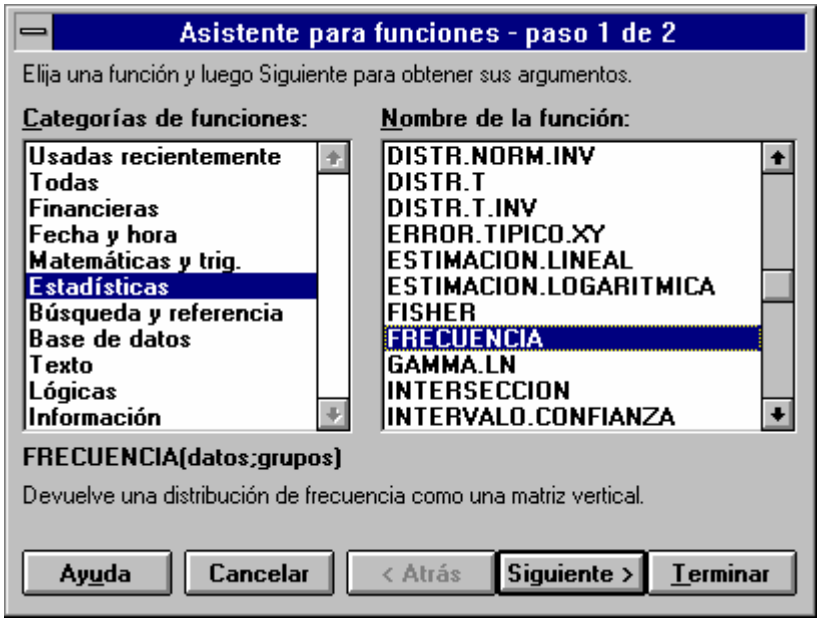
Los resultados con los datos anteriores son los siguientes:
En forma tabular:

No de cuartos	*Frecuencia*	*% acumulado*
1	970	15,16%
2	1.575	39,77%
3	1.349	60,84%
4	1.136	78,59%
5	665	88,98%
6	705	100,00%
y mayor...	0	100,00%

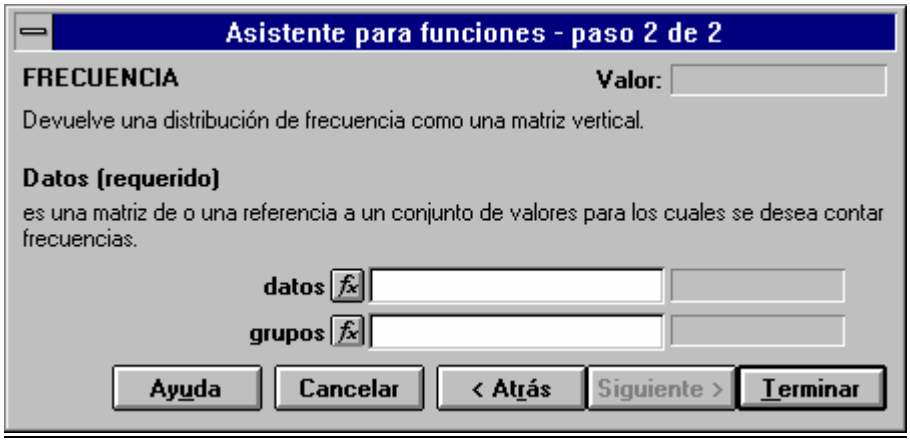
En forma gráfica:



Cuando se usa la función ***frecuencia***, el procedimiento es más complicado y menos espectacular. En la barra de herramientas se encuentra el botón del "*Asistente de Funciones*",  al oprimirlo aparece el siguiente cuadro de diálogo:



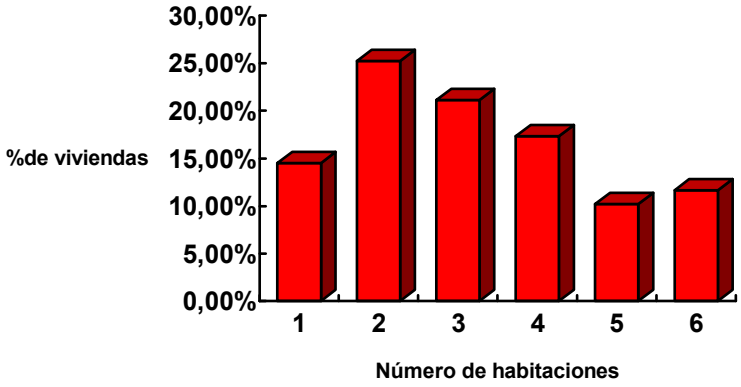
En este cuadro ya se señaló las funciones "*Estadísticas*" y dentro de ellas se escogió "*Frecuencia*". Cuando ya ha sido seleccionada la función que interesa, se oprime el botón  y aparece este cuadro de diálogo:



En la casilla  se introduce el rango donde se encuentran los datos; en el ejemplo, *A1:T10*. En la casilla  se introduce el rango de los grupos en que se desea clasificar los datos; en el ejemplo, *M3:M8*. Al oprimir el botón  aparece el resultado en la celda *O3*; a partir de esa celda se debe señalar el rango *O3:O8*, inmediatamente, estando sobre la celda *O3*, se oprime *F2* y simultáneamente las teclas *CTRL+MAYUSCULAS+ENTRAR (en Windows)* y *COMANDO+ENTRAR (en Macintosh)*. Así se convierte la función (fórmula) *FRECUENCIA*, en una matriz y se obtiene el valor de la frecuencia de ocurrencia para cada grupo.

	M	N	O
1	Número de habitaciones.	Frecuencia absoluta acumulada.	Frecuencia relativa.
2			
3	1	{=FRECUENCIA(A1:J640;M3:M8)} [937]	=N3/$N$9 [14,64%]
4	2	{=FRECUENCIA(A1:J640;M3:M8)} [1,603]	=N4/$N$9 [25,05%]
5	3	{=FRECUENCIA(A1:J640;M3:M8)} [1,363]	=N5/$N$9 [21,30%]
6	4	{=FRECUENCIA(A1:J640;M3:M8)} [1,109]	=N6/$N$9 [17,33%]
7	5	{=FRECUENCIA(A1:J640;M3:M8)} [650]	=N7/$N$9 [10,16%]
8	6	{=FRECUENCIA(A1:J640;M3:M8)} [738]	=N8/$N$9 [11,53%]
9	TOTAL	=SUMA(N3:N8) [6,400]	=SUMA(O3:O8) [100,00%]

Los valores obtenidos se pueden mostrar en una gráfica que se llama histograma de frecuencia.



En el archivo *Estadística.XLS* en la hoja *DISTRIVI* construir la tabla y gráfica anteriores.

Caso continuo

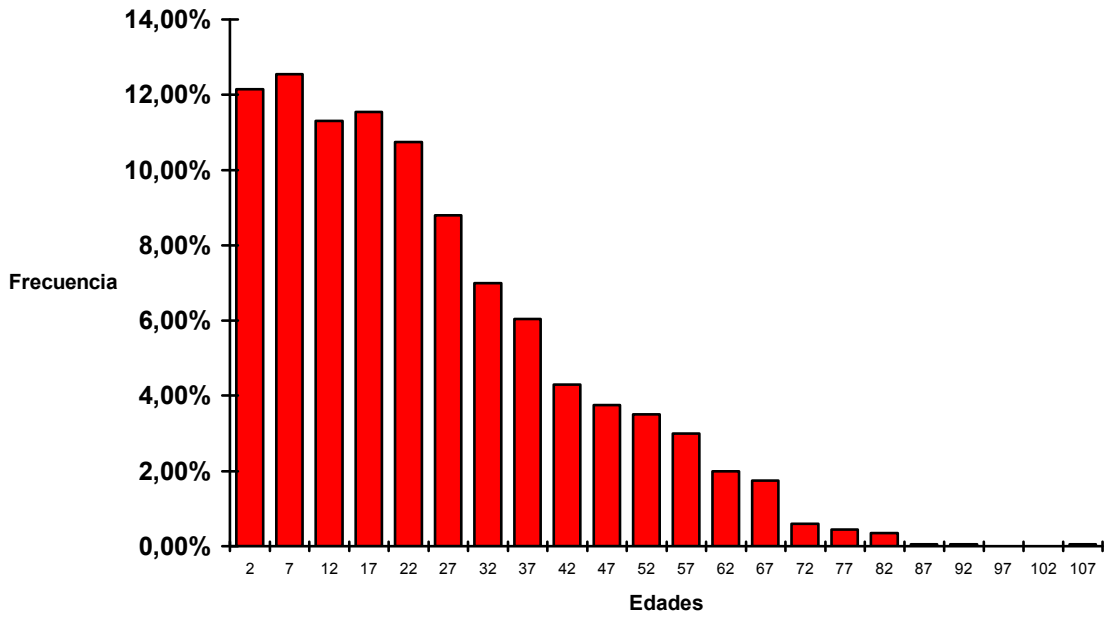
Nuevamente, si se toman los datos de un censo de población y se obtiene una muestra de 2.000 personas, las edades se clasifican en intervalos y no en valores puntuales. Estrictamente, la edad de una persona se comporta como una variable continua, a pesar de que en la práctica la gente "redondea" su edad en números enteros y casi nunca, o nunca, se dice que alguien tiene 22 años, 3 meses, 27 días, 4 horas, etc. También en la práctica nadie tiene la misma y exacta edad de otra persona. Por consiguiente carece de sentido hablar de valores concretos, antes bien, se habla de rangos de edad. Más aun, en el caso de las edades, se acostumbra a definir los rangos con extremos enteros, por ejemplo, se habla del grupo de edad de 0-4 años o de 5-9 años. La muestra indica que las personas tienen edades entre 0 y 100 años (en la práctica, aunque se puede exceder esa cifra). La muestra se puede clasificar de acuerdo con los grupos de edad quinquenales (cinco años) y su tabulación se puede presentar así:

Grupos de Edad (años)	Frecuencia absoluta	Frecuencia relativa
0-4	243	12,15%
5-9	251	24,70%
10-14	226	36,00%
15-19	231	47,55%
20-24	215	58,30%
25-29	176	67,10%
30-34	140	74,10%
35-39	121	80,15%
40-44	86	84,45%
45-49	75	88,20%
50-54	70	91,70%
55-59	60	94,70%
60-64	40	96,70%
65-69	35	98,45%
70-74	12	99,05%
75-79	9	99,50%
80-84	7	99,85%
85-89	1	99,90%
90-94	1	99,95%
95-99	0	99,95%
100-104	0	99,95%
105-109	1	100,00%
Total	0	100,00%

Siempre surge la pregunta de: ¿cuántos intervalos se deben construir? La respuesta es que esto depende de los datos que se deseen analizar y no deben ser, ni muchos, ni pocos. Se puede considerar que entre 5 y 15 intervalos sería razonable. En cuanto al punto medio de cada intervalo, es preferible considerar un número entero.

Si se analizan las edades de la muestra y de la población de manera estricta, se tendría un patrón de muchos valores continuos con una concentración en los primeros 40 años, que se iría reduciendo a medida que se aumenta la edad.

El histograma de frecuencias de las edades sería así, considerando el valor central de cada intervalo como el valor del mismo:



En el archivo Estadística.XLS y en la hoja DISTRIED construir la tabla y gráfica anteriores. Mostrar también la gráfica de la frecuencia relativa.

Estadísticas de una distribución

Arriba se mencionó que la distribución de un universo se podía representar por las estadísticas de la muestra o del universo. Las estadísticas más comunes son aquellas que muestran la tendencia central o valor alrededor del cual se agrupan los elementos del universo y el grado de dispersión. Estas dos ideas se ilustrarán con el caso discreto de las habitaciones de las viviendas de la muestra seleccionada.

Tendencia central de la distribución

Con esta estadística se trata de examinar hacia qué valor se concentran los valores de la distribución. Las estadísticas más conocidas que miden la tendencia central son: La *moda*, la *mediana* y la *media* o *valor esperado.*

La moda

La moda se define como el valor más frecuente. Esto es, aquel valor que tiene mayor frecuencia. Debido a que los datos pueden agruparse de manera arbitraria —en el caso de la distribución continua— la moda no es la mejor medida de tendencia central. También puede suceder que haya dos "modas" iguales, en ese caso se dice que la distribución es bimodal y se presenta una ambigüedad. La forma más fácil de determinar la moda es utilizando el histograma de frecuencias.

Por medio del histograma de frecuencias de los ejemplos anteriores, identificar la moda. *Excel* tiene la fórmula para ello, *=MODA(Datos)* pero está restringida a un número reducido de observaciones; para 400 observaciones calcula el valor, para 10.000 arroja error.

La mediana

La mediana es aquel valor que divide la distribución en partes iguales, o sea que el número de observaciones por encima de la mediana es igual al número de observaciones por debajo de ella. Se conoce también como el valor medio o percentile 50.

 Con los datos de los ejemplos anteriores debe usted identificar la mediana. Excel tiene la fórmula para ello, *=MEDIANA(Datos)*.

La media o valor esperado

El valor esperado o media indica la tendencia central de los datos. Esto significa que es el valor alrededor del cual tienden a agruparse los datos de una distribución. En el caso de una variable aleatoria discreta, se calcula multiplicando cada valor posible por su probabilidad y sumando sus resultados. En el caso de una variable aleatoria continua, se debe recurrir al concepto de integral que se estudia en el cálculo integral. Generalmente se expresa por medio de la letra griega μ (parámetro) para el universo y por la notación $E(\ )$ o $\overline{X}$ (estadística) para una muestra.

 Con los datos de los ejemplos anteriores calcular la media. En Excel la fórmula es *=PROMEDIO(Datos)*.

Medidas de la dispersión de la distribución

Las estadísticas que describen a una muestra o universo muestran qué tan dispersas están las observaciones o los elementos del universo. Las más comunes son la varianza, la desviación estándar (es la raíz cuadrada de la varianza) y el rango. Intuitivamente se puede pensar en medir las diferencias entre cada observación y el valor central, por ejemplo, el valor esperado o media. Eso va a producir valores negativos y positivos y al sumarse entre sí deben cancelarse y producir el valor cero. Se puede obviar este inconveniente trabajando con el valor absoluto, en *Excel*, *=DESVPROM(Datos)* o con los cuadrados de las diferencias en *Excel*, *=DESVIA2(Datos)*. Cuando se desea medir las *variaciones entre dos o más variables*, entre sí, entonces se habla de la *covarianza*.

Varianza

Una medida de la dispersión de unos datos es la varianza. Esta se calcula así:

$$\sigma^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n} \qquad (12)$$

O sea que es el promedio del cuadrado de las diferencias de cada dato con el promedio. Esta expresión se aplica para la distribución y la muestra; cuando se refiere a la población o universo, se utiliza la letra griega sigma σ^2 (parámetro) y s^2 (estadística), cuando se trata de la muestra. Sin embargo, cuando se trata de estimar la varianza de un universo o distribución a partir de la varianza de una muestra de tamaño n, entonces la fórmula debe modificarse así:

$$\sigma^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1} \qquad (13)$$

Con los datos de los ejemplos anteriores calcular la varianza de los datos obtenidos y de la distribución o universo de donde procedieron esos datos. En *Excel* la fórmula es *=VAR(Datos)* cuando se trata de medir la varianza de la muestra, y *=VARP(Datos),* cuando se trata, a partir de la muestra, calcular la varianza del universo.

Desviación estándar

La desviación estándar (σ) es la raíz cuadrada de la varianza. Se puede demostrar que si $X_1, X_2, X_3....X_n$ son variables aleatorias independientes con media μ_i y desviación estándar σ_i, entonces la suma de esas variables tendrán una distribución normal con media μ_i y desviación estándar $nx\sigma_i^2$.

Con los datos de los ejemplos anteriores, calcular la desviación estándar de los datos obtenidos y de la distribución o universo de donde procedieron esos datos. En *Excel* la fórmula es *=DESVEST(Datos),* cuando se trata de medir la desviación estándar de la muestra (con esta función se hace una estimación del parámetrto del universo), y *=DESVESTP(Datos),* cuando se trata de calcular la desviación estándar del universo, a partir de la totalidad de los datos de ese universo. (DESVEST parte de la hipótesis de que los argumentos representan la muestra de una población. Si sus datos representan la población total, utilice DESVESTP para calcular la desviación estándar).

Rango

Otra manera de estimar la dispersión de unos datos es medir su rango. Esta es la diferencia entre el valor máximo y el valor mínimo.

Con los datos de los ejemplos anteriores calcular el rango de los datos obtenidos. En *Excel* la fórmula para el valor máximo es *=MAX(Datos)* y para el valor mínimo es *=MIN(Datos).*

Covarianza

La covarianza indica en qué medida dos variables se mueven al unísono. Si se observa el comportamiento de la rentabilidad de las acciones en la Bolsa, se encontrará que algunas de ellas aumentan al mismo tiempo y otras disminuyen mientras las otras aumentan. El cálculo de la covarianza relaciona las diferencias entre las variables y sus medias, unas con otras, así:

$$\sigma_{ij}^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(X_i - \overline{X_i}\right)\left(X_j - \overline{X_j}\right)}{n} \qquad (14)$$

o

$$\sigma_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{n} P_{ij}\left(X_i - E(X_i)\right)\left(X_j - E(X_j)\right) \qquad (15)$$

También se puede expresar como:

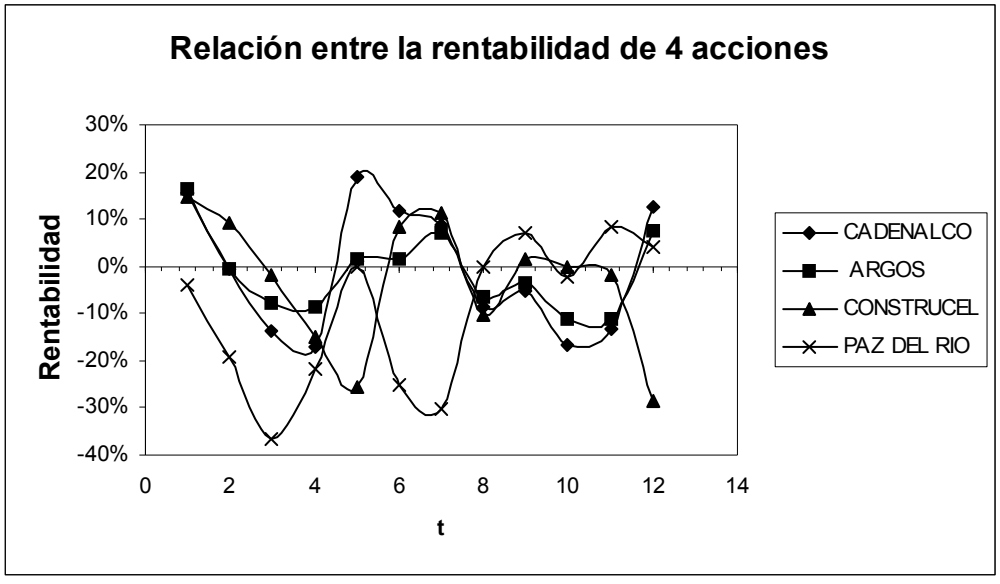
$$\sigma_{ij} = r_{ij}\sigma_i\sigma_j \tag{16}$$

El resultado del cálculo de la covarianza es una tabla como la siguiente:

$$\begin{matrix} \sigma_{11} & \sigma_{21} & \dots & \dots & \sigma_{n1} \\ \sigma_{12} & \sigma_{22} & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \sigma_{nn-1} \\ \sigma_{1n} & \dots & \dots & \sigma_{n-1n} & \sigma_{nn} \end{matrix} \tag{17}$$

Hay que observar que los datos de la diagonal son las varianzas de cada variable. Las demás son las covarianzas y son simétricas. Por ejemplo, si se desea saber cómo varían cuatro acciones de la Bolsa de Bogotá, se tiene:

Mes	CADENALCO	ARGOS	CONSTRUCEL	PAZ DEL RIO
1	15,75%	16,63%	14,87%	-3,91%
2	-0,47%	-0,36%	9,42%	-19,33%
3	-13,65%	-7,77%	-1,68%	-36,61%
4	-17,00%	-8,77%	-14,81%	-21,58%
5	18,87%	1,51%	-25,58%	0,04%
6	11,78%	1,50%	8,57%	-25,22%
7	9,00%	6,90%	11,42%	-30,23%
8	-8,61%	-6,54%	-10,32%	0,00%
9	-5,31%	-3,57%	1,71%	7,22%
10	-16,73%	-11,06%	0,00%	-2,25%
11	-13,21%	-11,33%	-1,88%	8,52%
12	12,65%	7,72%	-28,44%	4,22%



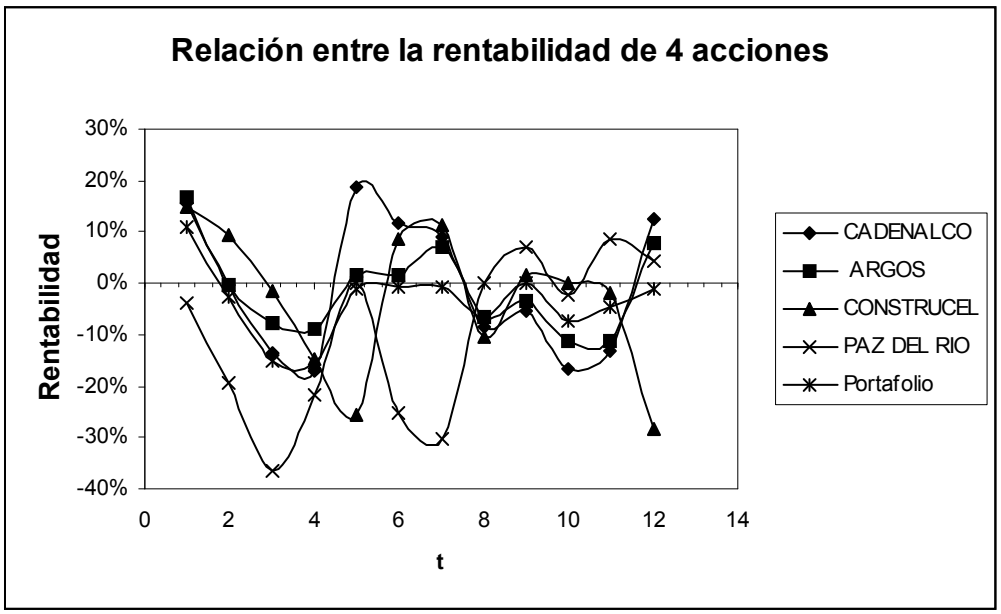
La covarianza se puede calcular en *Excel* con la opción de menú *Herramientas* y allí se selecciona *Análisis de datos.* En el cuadro de diálogo que aparece, se escoge *Covarianza* y se indica el rango donde están los datos para los cuales se desea calcular la covarianza. La matriz de covarianza de las rentabilidades de las cuatro acciones es:

	CADENALCO	*ARGOS*	*CONSTRUCEL*	*PAZ DEL RIO*
CADENALCO	0,0182927	0,01022173	-0,00046223	0,00106384
ARGOS	0,01022173	0,00738926	0,00233332	-0,00041683
CONSTRUCEL	-0,00046223	0,00233332	0,01999673	-0,00041683
PAZ DEL RIO	0,00106384	-0,00041683	-0,00827034	0,02463414

La covarianza de *Argos* y *Argos* es su varianza; en forma similar, la covarianza entre Cadenalco y Cadenalco es su varianza. La covarianza de Construcel y Construcel es su varianza; en forma similar, la covarianza entre Paz del Río y Paz del Río es su varianza.

En estas gráficas se observa que mientras dos de las acciones tienden a seguir el mismo comportamiento (suben o bajan ambas) las otras dos se comportan de manera contraria (mientras una sube, la otra baja). La medida del grado de coincidencia en el comportamiento está dada por la covarianza entre ellas.

Si se mezclan las cuatro acciones en partes iguales (si se construye un portafolio de las cuatro acciones por partes iguales), gráficamente se tiene lo siguiente:



Y los resultados de la media y la desviación estándar son los siguientes:

	CADENALCO	ARGOS	CONSTRUCEL	PAZ DEL RIO	PORTAFOLIO
PROMEDIO	-0,58%	-1,26%	-3,06%	-9,93%	-3,71%
DESVEST	12,95%	8,23%	13,54%	15,03%	6,74%

Obsérvese cómo se redujo la desviación estándar (la variabilidad del portafolio), debido a la combinación de cuatro variables (rentabilidad) y a la inclusión de acciones con covarianzas negativas.

Correlación

El dato que proporciona la covarianza no es fácil de interpretar cuando se expresa en las unidades de las variables analizadas; se necesita un indicador que sea independiente de las unidades de las variables analizadas; para evitar este problema, se puede encontrar la correlación o índice de correlación entre dos variables, escalándolo o "normalizándolo" con las desviaciones estándar, así:

$$r_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \tag{18}$$

Esta expresión se conoce como coeficiente de y está entre -1 y 1. Sirve para medir el grado de asociación entre dos variables u observaciones. En el ejemplo de las acciones se tiene, para Construcel y Paz del Río, desviaciones estándar:

	CADENALCO	*ARGOS*	*CONSTRUCEL*	*PAZ DEL RIO*
CADENALCO	0,0182927	0,01022173	-0,00046223	0,00106384
ARGOS	0,01022173	0,00738926	0,00233332	-0,00041683
CONSTRUCEL	-0,00046223	0,00233332	0,01999673	-0,00827034
PAZ DEL RIO	0,00106384	-0,00041683	-0,00827034	0,02463414

	CONSTRUCEL	PAZ DEL RIO
Varianza	0,01999673	0,02463414
Desviación estándar	0,14140981	0,15695266

$$\rho(c, p) = \frac{-0{,}00827034}{0{,}14140981 \times 0.15695266} = -0{,}37262751$$

Este valor indica que están correlacionadas negativamente; o sea, cuando la rentabilidad de una acción aumenta, la rentabilidad de la otra tiende a bajar.

Variable aleatoria

Una variable aleatoria es el valor que se le asigna a un determinado evento. Por ejemplo, en el ejemplo de la inversión a tres años, se le puede asignar un valor monetario a cada evento y un *Valor Presente Neto* (VPN) a la inversión; el VPN es una variable aleatoria.

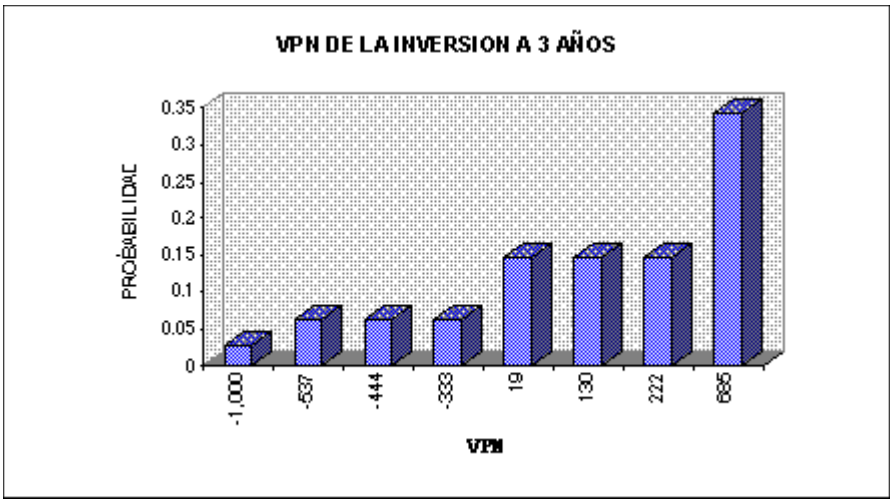
Si se retoma el ejemplo de la inversión a tres años, se habían previsto ciertos resultados y si se supone que cada año sin ingreso se le asocia el valor cero y cada año con ingreso se le asocia el valor $600, al año sin ingreso se le asocia una probabilidad de 30% y se tiene una tasa de descuento de 20% anual, entonces se tiene lo siguiente:

	A	B	C
1	$P(N)_1$=	30%	Inversión
2	$P(N)_2$=	30%	$1,000
3	$P(N)_3$=	30%	Tasa de descuento anual.
4	Valor de un año con ingreso.	$600	20%
5	Valor de un año sin ingreso.	$0	
6	Evento combinado.	Probabilidad total.	Variable aleatoria (VPN).
7	(*NNN*) = m_1	=B1*B2*B3 [0,027]	-1,000
8	(*NNS*) = m_2	=B1*B2*(1-B3) [0,063]	=B4/(1+C4) ^3-C2 [-537,04]
9	(*NSN*) = m_3	=B1*(1-B2)*B3 [0,063]	=B4/(1+C4) ^2-C2 [-444,44]
10	(*NSS*) = m_4	=B1*(1-B2)*(1-B3) [0,147]	=B4/(1+C4) ^2+B4/(1+C4) ^3-C2 [18,52]
11	(*SNN*) = m_5	=(1-B1)*B2*B3 [0,063]	-333,33
12	(*SSN*) = m_6	=(1-B1)*(1-B2)*B3 [0,147]	222,22
13	(*SNS*) = m_7	=(1-B1)*B2*(1-B3) [0,147]	129,63
14	(*SSS*) = m_8	=(1-B1)*(1-B2)*(1-B3) [0,343]	685,19

La distribución de la variable aleatoria será:

	A	B	C
13	Variable aleatoria (VPN)	Probabilidad	
14	-1.000,00	2,7%	
15	-537,04	6,3%	
16	-444,44	6,3%	
17	-333,33	6,3%	
18	18,52	14,7%	
19	129,63	14,7%	
20	222,22	14,7%	
21	685,19	34,3%	
22	Total	100,0%	

Con los datos de la tabla anterior construir el histograma que aparece a continuación.



Las variables aleatorias se representarán por X y los valores específicos por x. En el ejemplo X es el *Valor Presente Neto* (VPN) de la inversión a tres años. Se dice entonces que esta variable aleatoria X toma los siguientes valores:

X_1=	-1.000,00
X_2=	-537,04
X_3=	-444,44
X_4=	-333,,33
X_5=	18,,52
X_6=	129,63
X_7=	222,22
X_8=	685,19

En general, se escribiría $X=x$. La probabilidad se escribiría *P(X=222,22), P(X=685,19),* etc.

Asociada a toda variable aleatoria existe una función de distribución acumulada. Si se define el evento: variable aleatoria X menor o igual que b, como $E = (X< b)$, entonces $P(E)$ es la probabilidad de este evento y se denomina probabilidad acumulada de b, $F(b)$. La función acumulada de probabilidad es una función numérica definida para todos los valores posibles de b y tiene las siguientes propiedades:

- $F(b)$ es una función no decreciente de b. Esto es, que a medida que b aumenta, $F(b)$ aumenta o permanece igual y nunca disminuye.
- La variable aleatoria puede tomar valores entre menos infinito (-∞) y más infinito (+∞).

Por lo tanto:

- $F(-\infty) = 0$ y $F(\infty) = 1$ (19)

De acuerdo con estas definiciones y propiedades, se puede establecer el valor de la probabilidad que una variable aleatoria se encuentre entre dos valores dados como:

$P(a < X < b) = F(b) - F(a)$ (20)

Esto significa que cuando se tiene una variable continua, la probabilidad de un valor preciso es cero.

La distribución de probabilidad discreta

Una distribución de probabilidad es discreta cuando para cada valor existe una probabilidad de ocurrencia. Ejemplos de fenómenos que se clasifican como discretos, son el valor del lanzamiento de un dado, de una moneda, el número de hijos de una pareja, etc.

En términos generales, la distribución acumulada de esta clase de leyes probabilísticas está dada por:

$$F(b) = P(x \le b) = \sum_{i=-\infty}^{b} P(x = x_i) \quad para\ todas\ las\ x_i \le b \tag{21}$$

Esto significa que la probabilidad de que ocurra un valor menor o igual a *b*, es igual a la suma de todas las probabilidades de los valores de X menores que *b*. Por ejemplo, la probabilidad de que el resultado del lanzamiento de un dado de seis caras sea menor o igual que 3 es igual a la suma de las probabilidades de que el valor del lanzamiento sea 1, 2 ó 3.

Por otro lado, la media y la varianza se calculan así:

Media de la población: $\mu = \sum_{x} xp(x)$ (22)

Varianza de la población: $\sigma^2 = \sum_{x} (x - \mu)^2 p(x)$ (23)

La Distribución Binomial

Existen muchas leyes de probabilidad discretas; la más común es la binomial. Esta distribución se utiliza en situaciones con un número fijo de pruebas o ensayos, cuando los resultados de un ensayo son sólo éxito o fracaso, cuando los ensayos son independientes y cuando la probabilidad de éxito es constante durante todo el experimento. Por ejemplo, se puede calcular la probabilidad que dos de los próximos tres bebés que nazcan de un pareja sean hombres. A continuación se muestran algunos fenómenos regidos por la distribución binomial:

Fenómeno	Éxito	Fracaso	*p* (probabilidad de éxito)	*n* (casos totales)	*r* (éxitos)
Lanzar una moneda.	Cara.	Sello.	0,50	n lanzamientos.	Número de caras.
Nacimientos en una familia.	Niño.	Niña.	0,50	Tamaño de la familia.	Número de niños en la familia.
Lanzamiento de 3 dados.	8 puntos.	Cualquier otro resultado.	21/216	*n* lanzamientos.	Número de 8's.
Resultado de inversiones.	Ingreso.	No ingreso.	Asignada según el caso.	Períodos futuros en que pueden ocurrir los ingresos.	Períodos con ingresos.
Nacimientos de terneros.	Ternera.	Ternero.	0,50	Número de partos.	Número de terneras.
Escogencia de un votante en una encuesta de opinión política.	Liberal.	Otros partidos.	Proporción de Liberales en la población.	Tamaño de la muestra.	Número de liberales en la muestra.
Lanzamiento de un dado.	1 punto.	Cualquier otro valor.	1/36	n lanzamientos.	Número de 1's.
Aprobación de leyes en el Congreso.	Aprobada.	Rechazada.	Proporción de congresistas a favor de la ley.	Número de congresistas.	Aprobación de la ley.

La distribución binomial típica es el lanzamiento de una moneda: *S* = número de caras en *n* lanzamientos.

Se dice que son *n* lanzamientos independientes y para cada lanzamiento hay un éxito (cara) o un fracaso (sello) y las probabilidades son *p* para cara y *(1-p)* para sello.

Si se tiene 1 dado, la distribución binomial que rige este fenómeno se puede deducir examinando el caso mencionado de obtener cierto número de *1's* en *n* lanzamientos. Si se estipula que sean tres *1's* en 6 lanzamientos, los casos posibles en que aparecen tres *1's* son:

AAAFFF	AFAFAF	FAAAFF	FAFFAA
AAFAFF	AFAFFA	FAAFAF	FFAAAF
AAFFAF	AFFAAF	FAAFFA	FFAAFA
AAFFFA	AFFAFA	FAFAAF	FFAFAA
AFAAFF	AFFFAA	FAFAFA	FFFAAA

El resultado acertado (1) se indicará con la letra *A* y no acertado (diferente a 1) se indicará con la letra *F*.

La probabilidad de cualquier evento, por ejemplo *AFFAFA*, es:

px(1-p)x(1-p)xpx(1-p)xp

Esta probabilidad es igual para cualquiera de los eventos señalados, pues sólo cambia el orden. Como estos eventos son excluyentes, entonces la probabilidad de obtener tres *1's* en seis lanzamientos es la suma de cada una de esas probabilidades. O sea:

$20x(1/6)^3(1-1/6)^3 = 20x(1/216)x(125/216) = 2,500/46,656 = 0,05358$

En general para calcular el número de *r* casos exitosos de *n* intentos, en cualquier orden, se tiene:

$$C_r^n = \frac{n!}{r!(n-r)!} \qquad (24)$$

n se define como *n* factorial y es igual a *1x2x3x4...x(n-1)xn* y el caso especial de 0 es igual a 1.

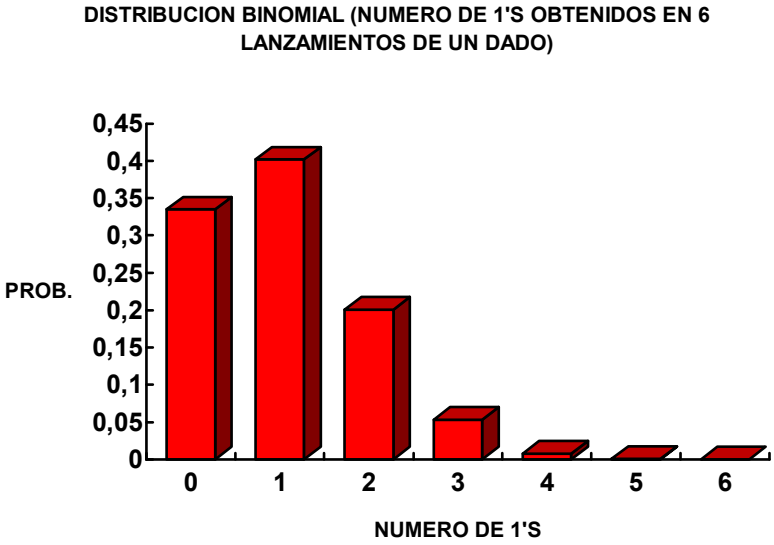
En *Excel =DISTR.BINOM(num_éxitos (r);intentos (n);prob de éxito (p);acumulado).* En el argumento acumulado, "falso"=valor de la densidad de probabilidad "verdadero"= valor acumulado de la probabilidad.

En el ejemplo, obtener tres unos en seis lanzamientos del dado:

	A	B
19	*p*=	=1/6 [0,1667]
20	*n*=	6
21	*r*=	3
22	*P*=	***=DISTR.BINOM(B21;B20;B19 "FALSO")*** [0,05358368]

Si se analizan todos los casos posibles, se tiene:

	A	B	C	D	E
22	n	r	p	ACUMULADO = "VERDADERO"	ACUMULADO = "FALSO"
23	6	6	0.1667	***=DISTR.BINOM(A23;B23;B23; "VERDADERO")*** 1	***=DISTR.BINOM(A23;B23;B23 "FALSO")*** [2,1433E-05]
24	6	5	0.1667	**=DISTR.BINOM(A24;B24;B24; "VERDADERO")** [0,99997857]	**=DISTR.BINOM(A24;B24;B24 "FALSO")** [0,000643]
25	6	4	0.1667	[0,99933556]	0,00803755
26	6	3	0.1667	0,99129801	0,05358368
27	6	2	0.1667	0,93771433	0,20093879
28	6	1	0.1667	0,73677555	0,40187757
29	6	0	0.1667	0,33489798	0,33489798
30				TOTAL	1



Construir en la hoja de cálculo la tabla y la gráfica anteriores.

La distribución binomial tiene los siguientes parámetros:
Media = p
Varianza = pq = p(1-p)

La distribución de probabilidad continua

Cuando la variable que se está analizando puede tomar cualquier valor entre -∞y ∞de una manera "continua", esto es, que se admite cualquier valor, entero o no dentro de esos límites, entonces se dice que es una variable continua. A diferencia de la distribución discreta, donde cada valor tiene asociada una probabilidad, en este caso cada valor tiene asociado un valor que se llama función de densidad de probabilidad. Esta función de densidad de probabilidad no es un histograma de frecuencia, sino de una curva. La probabilidad se le asigna a un rango de valores y se mide en términos de la proporción del área bajo la curva entre esos dos valores y el área total.

La distribución acumulada es en este caso:

$$F(b) = P(x \le b) = \int_{-\infty}^{b} P(x)dx \qquad (25)$$

La media y la varianza serán:

Media de la población, $\mu = \int_{-\infty}^{+\infty} xp(x)dx$ (26)

Varianza de la población, $\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 p(x)dx$ (27)

La Distribución de Probabilidad Normal

También existen muchas leyes de probabilidad continuas; la más conocida y frecuente es la que se conoce como distribución normal o de Gauss. Esta ley de probabilidad rige muchos fenómenos de la naturaleza.

Esta distribución de probabilidad (función de densidad de probabilidad) se expresa así:

$$P(x) = \frac{e^{-\left(\frac{1}{2}\right)\left(\frac{x-\mu}{\sigma}\right)^2}}{\sqrt{2\pi\sigma}} \qquad (28)$$

Donde:

x = Variable aleatoria con distribución normal.

e = Base de los logaritmos naturales, 2.71828183.

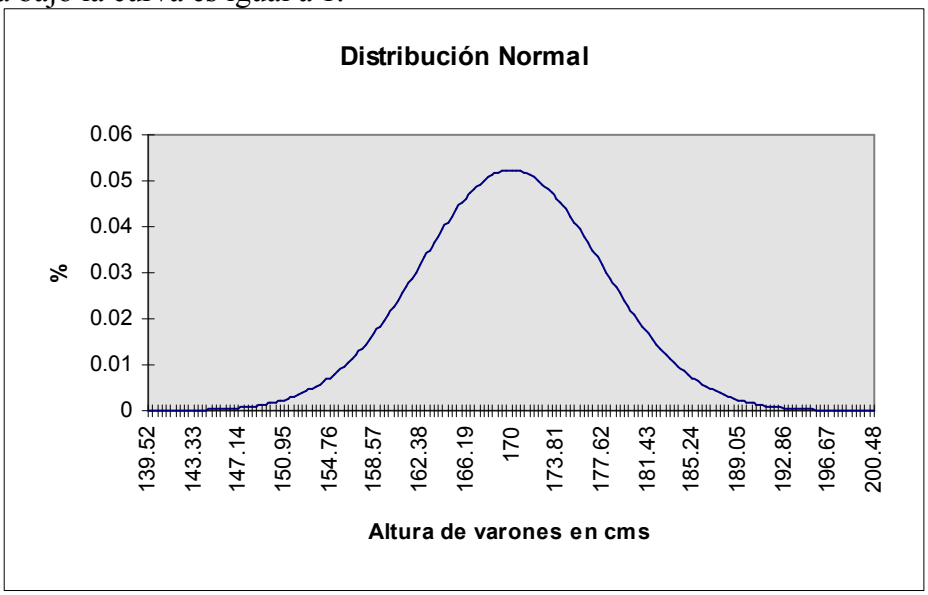
π = Número pi 3.141516....

μ = Media de la distribución.

σ = Desviación estándar de la distribución.

Esta distribución tiene unas características que la hacen muy especial:

- La moda, la media y la mediana son iguales.
- Es simétrica alrededor de la media.
- La curva tiene dos puntos de inflexión en la media ± una desviación estándar.
- Es asintótica en cero alejándose de media.
- El área bajo la curva es igual a 1.



Hay un caso especial que consiste en "estandarizar" la distribución normal; esto consiste en cambiar el origen de la distribución y suponer que la media es cero y que la desviación estándar es 1. Para lograr esto se hace la siguiente transformación:

$$Z = \frac{x - \mu}{\sigma} \qquad (29a)$$

Z es la variable normal estandarizada.

Cuando se trata del promedio de la variable $\bar{x}$ entonces la expresión (29a) se convierte en

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \qquad (29b)$$

En este caso la variable queda definida con un valor esperado de 0 y una desviación estándar de $\frac{\sigma}{\sqrt{n}}$, es decir, que su varianza es $\frac{\sigma}{n}$.

Esto es muy utilizado cuando no se usan hojas de cálculo. Con la normal estandarizada es como se calculan las tablas de la distribución normal acumulada.

Cuando se trabaja con distribuciones continuas se manejan áreas, ya que se supone que cualquier valor puede ocurrir y que entonces los intervalos son infinitesimales. Sin embargo, para hablar de probabilidad se debe referir a un intervalo y, como se estudió en las características de una variable aleatoria, la probabilidad entre -∞ y ∞ es igual a 1. Como se dijo también arriba, cuando se trata de distribuciones continuas, la probabilidad debe calcularse como el área bajo la curva entre dos valores. El área bajo la curva entre -∞ y ∞ es 1. Todo esto significa que la probabilidad de un valor exacto, por ejemplo la probabilidad de que una persona tenga 32 años, 7 meses, 4 días, 3 horas, 3 minutos 22 segundos (inclusive se puede llegar a expresar esto de manera infinetisimal) es cero, puesto que el área entre dos valores iguales (esto es, el mismo valor) es cero.

Muchos fenómenos de la naturaleza pueden ser descritos por la distribución normal o de *Gauss*. Inclusive, algunos fenómenos que no siguen esta ley de probabilidad pueden ser analizados suponiendo que siguen esa distribución, y los resultados, en términos prácticos son bastantes aceptables.

La distribución normal tiene ciertas características que sirven para hacer más fácil su manejo; una de ellas es la simetría —ya mencionada— y que el área debajo de la curva es proporcional, independientemente del fenómeno que se esté analizando y de los valores de sus parámetros (valor esperado o media y desviación estándar) a la probabilidad de los valores que la limitan a cada lado.

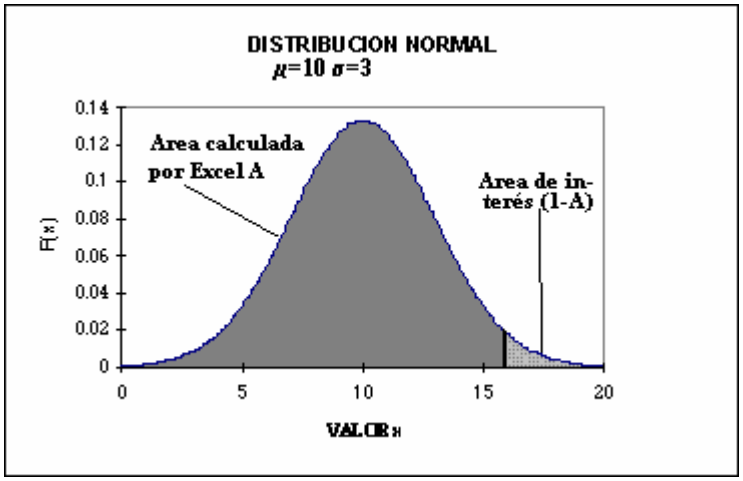
Esto significa que si se tiene información sobre el área bajo la curva normal con parámetros $\mu = 0$ *y* $\sigma = 1$ esta información se puede utilizar para otra distribución normal con parámetros diferentes, si se hacen ciertas transformaciones. Ahora bien, esta transformación era válida cuando era necesario recurrir a tablas por la dificultad del cálculo. Hoy, las hojas electrónicas como *Excel*, por ejemplo, traen funciones que no sólo manejan la distribución estandarizada con parámetros $\mu = 0$ *y* $\sigma = 1$ -la función es *=DISTR.NORM.ESTAND(z)*- sino que tiene funciones que manejan directamente el valor de la probabilidad deseada, incluyendo los parámetros de la distribución que se estudia. Para este caso, también se utiliza el *Asistente de Funciones* y se aplica la función *=DISTR.NORM(Valor que interesa x;media;desv.estándar;acum).* Si se escribe *Acumulado* en *acum*, entonces arroja el valor acumulado entre -∞ y el valor que interesa, *x*; si no se escribe *Acumulado*, arroja el valor de la densidad de probabilidad, o sea el valor de la función $P(x) = \frac{e^{-\left(\frac{1}{2}\right)\left(\frac{x-\mu}{\sigma}\right)^2}}{\sqrt{2\pi\sigma}}$.

Ejemplo

Si se tiene una distribución normal con $\mu = 10$ *y* $\sigma = 3$ y se desea calcular la probabilidad que la variable en cuestión tome un valor mayor que 16.

Ahora bien, como se desea saber la probabilidad de que la variable sea mayor que 16 y lo que arroja la función es la probabilidad de que sea menor que 16, entonces se debe restar esta última de 1.

	A	B
1	Valor x.	16
2	Media.	10
3	Desviación estándar.	3
4	Probabilidad de que *x* sea menor que 16.	*=DISTR.NORM(B1;B2;B3;VERDADERO)* [97,72%]
5	Probabilidad de que *x* sea mayor que 16.	=1-B4 [2,28%]



Si se estandariza, esto es que la media se convierte en 0 y la desviación estándar en 1:

$$z = \frac{16-10}{3} = 2$$

Esto es, se supone que la media se traslada a 0 y se calcula cuántas veces está $\sigma = 3$ en el intervalo entre 10 y 16. Esta transformación equivale a trabajar con una distribución normal con parámetros $\mu = 0$ y $\sigma = 1$, y se debe calcular la probabilidad de que z sea mayor que 2. Si la probabilidad de que sea menor que 2 —o 16 en el problema original— entonces la probabilidad que sea mayor que 2 -o sea, mayor que 16 en la variable original- será 1 – 0,9773 = 0,0227 o sea 2,27%.

	A	B
1	Valor *x*.	16
2	Media.	10
3	Desviación estándar.	3
4	Probabilidad de que *x* sea menor que 16.	*=DISTR.NORM(B1;B2;B3;VERDADERO)* [97,72%]
5	Probabilidad de que *x* sea mayor que 16.	=1-B4 [2,28%]
6	z	=(B1-B2)/B3 [2]
7	Probabilidad de que *z* sea menor que 2.	*=DISTR.NORM.ESTAND(V17)* [97,72%]
8	Probabilidad de que *z* sea mayor que 2.	=1-B8 [2,28%]

Prueba de hipótesis z para medias de dos muestras

Es muy frecuente enfrentarse a la discusión de si los promedios o medias de dos muestras son estadísticamente iguales o no. Aunque los resultados de los promedios de las muestras sean numéricamente diferentes puede suceder que esta diferencia no sea significativa desde el punto de vista estadístico. Ese significancia se puede examinar con base en los valores de las medias, de sus varianzas y del tamaño de la muestra. La prueba más sencilla es la de z para medias de dos muestras.

La idea es determinar el valor de z definido como

$$z = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\text{var}_1}{n_1} + \frac{\text{var}_2}{n_2}}} \qquad (30)$$

donde μ_1 y μ_2 son las medias de las muestras, var_1 y var_2 son las varianzas de cada muestra y n_1 y n_2 son los tamaños de cada muestra. Si la hipótesis es que las medias son iguales entonces en una situación ideal el valor de *z* calculado debería ser cero. Obsérvese que el denominador es la desviación estándar de la nueva variable aleatoria definida como la diferencia de las medias.

Podemos entonces examinar la hipótesis de igualdad de las medias y en ese caso estamos haciendo un examen de lo que se conoce como dos colas, es decir, que al considerar la igualdad (o desigualdad) de las medias, esta puede no ocurrir (u ocurrir) porque una es mayor o menor que la otra. También podemos examinar la hipótesis de que una media sea menor o mayor que la otra y en ese caso utilizaremos un examen de una cola.

Supongamos que se tienen los resultados de los ECAES (pruebas o exámenes de calidad de la educación superior en Colombia) para una facultad que tiene sedes en Cartagena y Bogotá. Los datos de las medias, las varianzas y los tamaños de las muestras son los siguientes:

	Bogotá	*Cartagena*
Media	52,80952381	45,77586207
Varianza (conocida)	57,35	48,10
Observaciones	42	58

¿Son los valores 52,81 y 45,78 diferentes desde el punto de vista estadístico como lo indican los números? ¿O simplemente se trata de las desviaciones normales producidas por ser unas muestras de un universo mayor?

Si aplicamos nuestra fórmula (30) tendremos

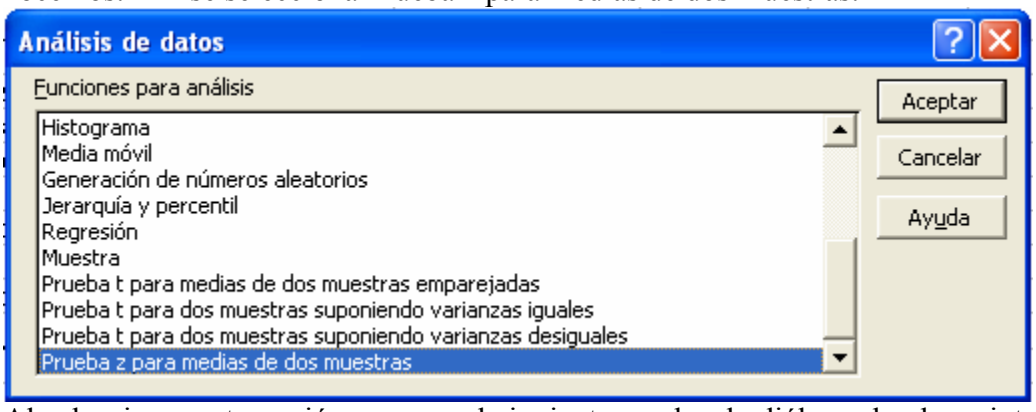
$$z = \frac{52{,}81 - 45{,}78}{\sqrt{\frac{57{,}35}{42} + \frac{48{,}10}{58}}} = 4{,}75$$

En esta prueba suponemos que las variables están distribuidas según la distribución normal.

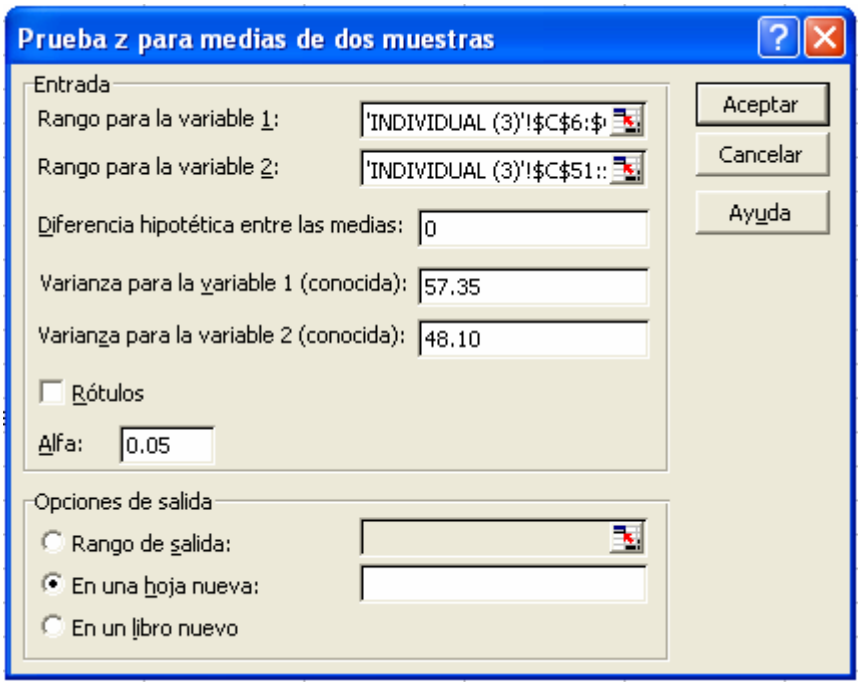
Si intentamos examinar la hipótesis para un nivel de 5% el valor de z, caso de dos colas debería ser menor que 1,96 que es el valor de z para que la suma de las dos colas resulte en 5%. En este caso vemos que el z calculado es mayor que 1,96. Por lo tanto, la evidencia estadística nos indica que podemos rechazar la hipótesis de que las medias son iguales.

Si por el contrario, examinamos la hipótesis de igualdad de las medias, pero tomamos como hipótesis alterna que la media de Cartagena es menor que la de Bogotá, entonces el valor de z para que la cola única tenga un área (probabilidad) de 5% será de 1,64. Esto nos indica que podemos rechazar la hipótesis de igualdad de las medias y no rechazar la hipótesis alterna de que la media de Cartagena es menor que la de Bogotá.

Esta prueba está definida en Excel y se encuentra en Análisis de Datos que ya conocemos. Allí se selecciona Prueba z para medias de dos muestras.



Al seleccionar esta opción aparece el siguiente cuadro de diálogo donde se introducen los rangos donde están los valores, las varianzas, la diferencia hipotética entre las medias (en este caso 0) y se indica si se incluyen rótulos o no, el nivel de significancia y el sitio donde se desea que aparezcan los resultados.



El resultado aparece así

Prueba z para medias de dos muestras		
	Bogotá	*Cartagena*
Media	52,80952381	45,77586207
Varianza (conocida)	57,35	48,10
Observaciones	42	58
Diferencia hipotética de las medias	0	
z	4,747722588	
P(Z<=z) una cola	0,000001030	
Valor crítico de z (una cola)	1,644853476	
P(Z<=z) (dos colas)	0,000002060	
Valor crítico de z (dos colas)	1,959962787	

Observe que los resultados de los cálculos son los mismos.

Ahora consideremos otro ejemplo sencillo.

	Notas del curso A	*Notas del curso B*
Media	2,43	2,39
Varianza (conocida)	0,8464	0,4356
Observaciones	211	33

¿Son los valores 2,43 y 2,39 diferentes desde el punto de vista estadístico como lo indican los números? ¿O simplemente se trata de las desviaciones normales producidas por ser unas muestras de un universo mayor?

Si aplicamos otra vez nuestra fórmula (30) tendremos

$$z = \frac{2{,}43 - 2{,}39}{\sqrt{\frac{0{,}8464}{211} + \frac{0{,}4356}{33}}} = -0{,}3048$$

Los valores críticos para un nivel de 5% son ±1,96 para el caso de dos colas y −1,64 para caso de una cola. En ambos casos el valor calculado de z está dentro de los límites aceptables. En este caso podemos decir que a un nivel de 5% no existe evidencia para rechazar la hipótesis nula que dice que las medias son iguales. Y tampoco existe evidencia estadística para decir que la nota promedio del curso B es menor que la del curso A.

La distribución χ^2 (Chi cuadrado o ji cuadrado)

La distribución χ^2 es muy útil para hacer pruebas de hipótesis, en particular pruebas de bondad de ajuste. Algunas de ellas son las pruebas de normalidad (verificar si una variable tiene distribución normal) y tablas de contingencia que permiten evaluar la independencia de los resultados.

Esta distribución tiene una formulación tan poco amigable como la distribución normal:

$$Y = \frac{e^{-\chi^2/2}\left(\chi^2\right)^{(\nu-2)/2}}{2^{\nu/2}\left(\frac{\nu-2}{2}\right)!} \tag{30}$$

Donde *e* es la base de los logaritmos naturales (2,7148...), χ^2 es la variable *chi* cuadrado y ν son los grados de libertad.

La prueba χ^2 compara las frecuencias obtenidas de la variable con la frecuencia esperada de esa variable. Si las dos frecuencias son "parecidas" no se rechaza la hipótesis de que la frecuencia observada procede de una distribución dada. La forma general para estas pruebas es la siguiente:

$$\sum_{i=1}^{k} \frac{\left(F_i - f_i\right)^2}{f_i} \tag{31}$$

Donde F_i es la frecuencia observada, f_i es la frecuencia teórica o esperada y k es el número de observaciones. En cada caso hay que definir el número de grados de libertad.

La Distribución t de Student

La distribución t de Student fue publicada por William Gosset en 1908. La empresa donde trabajaba le impidió usar su nombre y la presentó bajo el seudónimo de "Student." En la sección anterior dijimos que el estadístico Z tiene una distribución normal. Este estadístico tiene la siguiente formulación

$$Z = \frac{x - \mu}{\sigma} \tag{32a}$$

Z es la variable normal estandarizada.

Dijimos además que cuando se trata del promedio de la variable x entonces la expresión (29a) se convierte en

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \tag{32b}$$

Esta es una situación donde se conocen todos los parámetros de la distribución, es decir, se conoce la media y la varianza o la desviación estándar.

Cuando no se conoce la varianza debemos usar un estimador para la desviación estándar, en este caso **s**. Así, creamos el estadístico

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \tag{33}$$

Donde μ es la media de la población, $\bar{x}$ es la media de la muestra y **s** es el estimador de la desviación estándar de la población, definida como

$$s = \sqrt{\left[\frac{1}{n-1}\sum\left(x_i - \bar{x}\right)^2\right]} \tag{34}$$

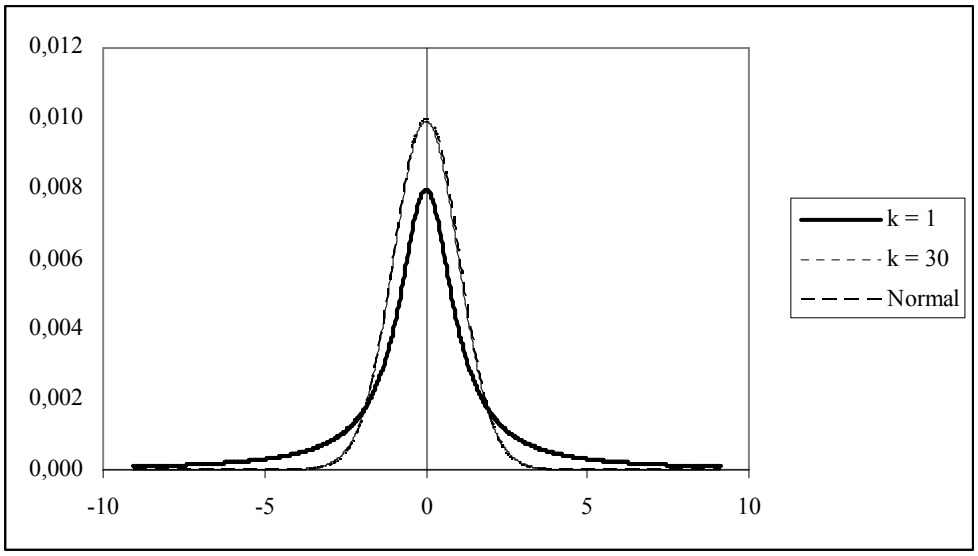
El estadístico **t** tiene una distribución t de Student con n − 1 grados de libertad (k).

En el caso de Z es un estadístico que se comporta según la distribución normal con media cero y varianza 1, mientras que **t** no tiene una distribución normal.

La distribución **t** de Student se define como la distribución de la variable aleatoria **t** la cual es lo "mejor" que podemos obtener cuando no conocemos σ. Cuando se especifica una distribución **t** de Student hay que especificar los grados de libertad. Estos tienen que ver con la desviación estándar **s**, de la muestra.

Cuando t = z la distribución t de Student se convierte en la distribución normal. A medida que n aumenta (los grados de libertad también aumentan) la t de Student tiende a la distribución normal.

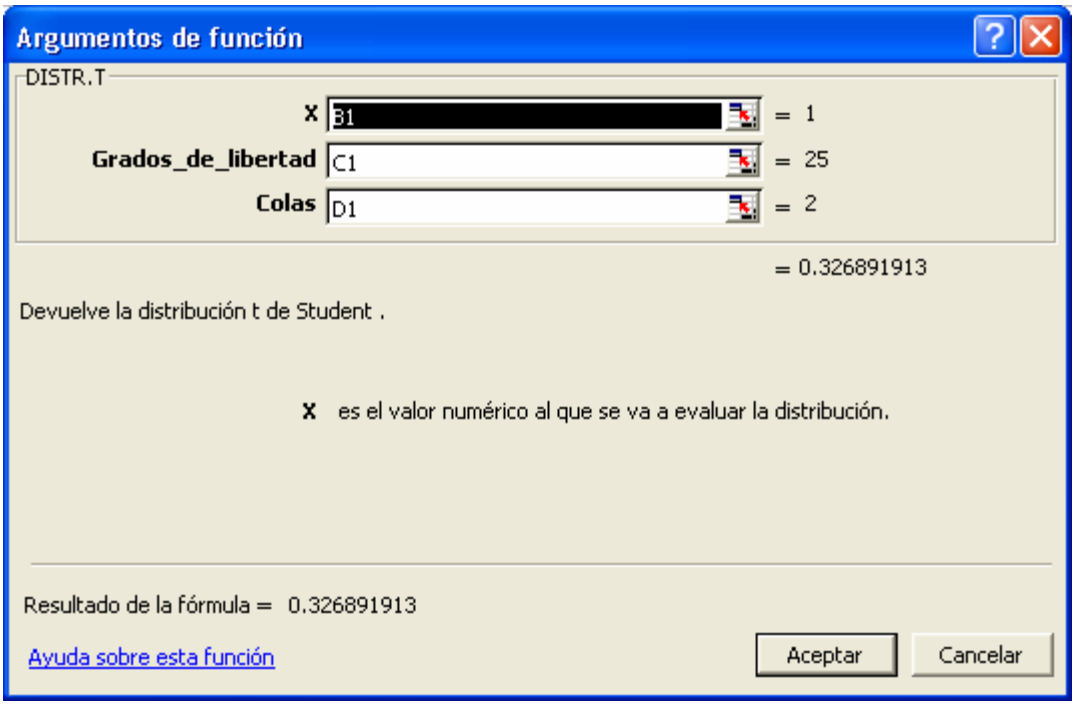
En la siguiente gráfica se puede ver la distribución t para k = 1 y k = 30. Así mismo, se muestra la distribución normal.



Distribución t de Student y normal

En esta gráfica se puede observar la similitud entre la normal y la distribución t cuando k es grande (k≥30).

La distribución t de Student se puede encontrar en Excel como =DISTR.T(Valor,Grados de libertad,Colas) y al llenar el cuadro de diálogo de la función, aparece así:



La distribución F

Si se tienen dos variables aleatorias independientes con distribución Chi-cuadrado y k_1 y k_2 grados de libertad respectivamente, entonces se puede construir una variable aleatoria de la siguiente manera

$$F = \frac{\chi_1^2 / k_1}{\chi_2^2 / k_2} \tag{35}$$

La variable aleatoria F está definida sólo para valores mayores que cero y su función de densidad de probabilidad es

$$f(F) = \frac{F^{\left(k_1/2 - 1\right)}}{(k_2 + k_1 F)^{(k_1+k_2)/2}} \tag{36}$$

Esta distribución de probabilidad se conoce como la distribución F.

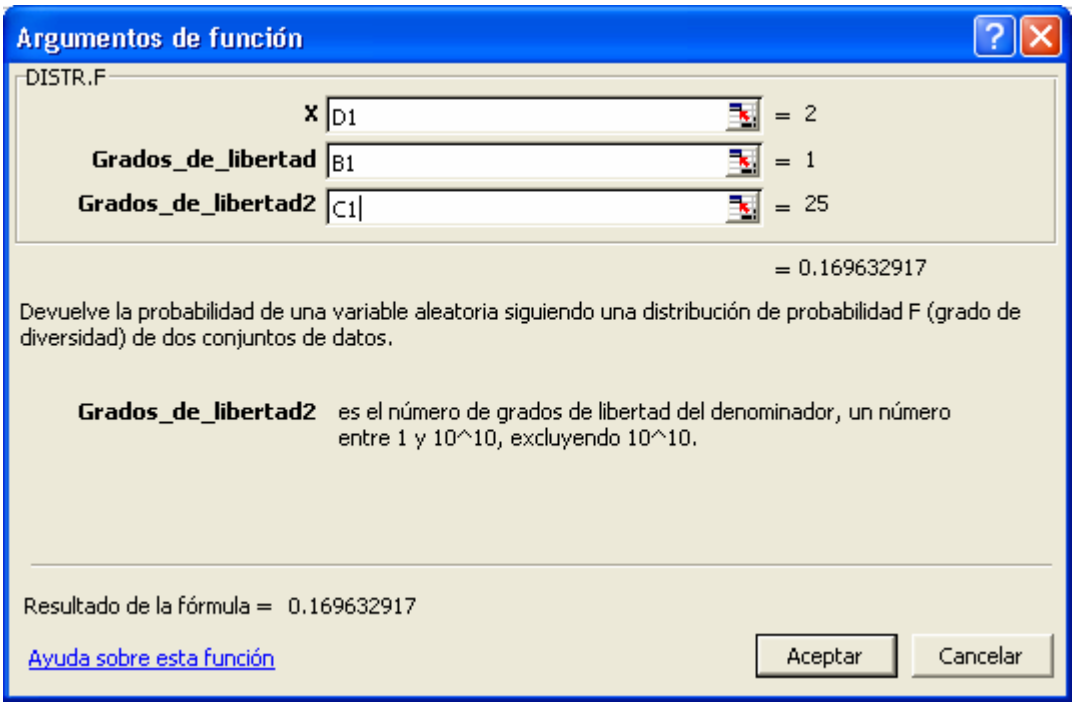
La aplicación más importante de esta distribución es el estudio de variables aleatorias con distribución Chi-cuadrado de tal manera que si χ_1^2 y χ_2^2 son variables aleatorias con distribución Chi-cuadrado con k_1 y k_2 grados de libertad, entonces

$$F = \frac{\chi_1^2 / k_1}{\chi_2^2 / k_2} \tag{37}$$

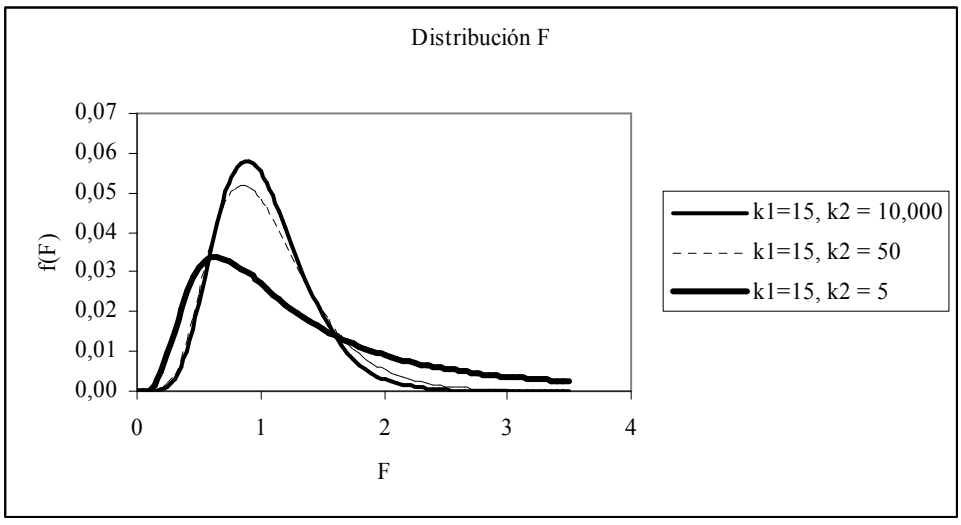
tiene una distribución F con k_1 y k_2 grados de libertad.

La utilidad de esta distribución radica en las decisiones que se pueden tomar respecto de la proporción entre dos varianzas maestrales. Esto se aplcará en la sección de análisis de regresión, más adelante.

La distribución F se puede encontrar en Excel como =DISTR.F(Valor,grados de libertad, grados de libertad 2) y al llenar el cuadro de diálogo de la función, aparece así:



En la siguiente gráfica se puede ver la distribución F para diferentes grados de libertad.



Distribución F

Estadística no paramétrica

Hasta ahora nos hemos ocupado de distribuciones que nos permiten "medir" ciertas estadísticas tales como media, varianza, etc. Existen muchas situaciones en la realidad que requieren más que todo verificar si ciertos resultados son independientes o si siguen

determinada ley de probabilidad (distribución de probabilidad). También es necesario encontrar herramientas que nos permitan evaluar resultados basados en ordinalidad, más que en valores absolutos. De estos temas se ocupa esta sección. La estadística no paramétrica se utiliza cuando no se cumplen ciertas condiciones rigurosas, por ejemplo, en lo que se conoce como estadística paramétrica se puede requerir que las muestras a estudiar sean independientes y que provengan de distribuciones normales con varianzas iguales. Las pruebas de hipótesis no paramétricas son muy populares debido a dos razones: La primera, es que requieren suposiciones menos restrictivas que las paramétricas y con frecuencia los cálculos son cortos y simples. La segunda, es el hecho de que las pruebas no paramétricas son las más adecuadas cuando se trata de analizar información de muestras que sólo pueden ser ordenadas.

Pruebas de normalidad

Muchas pruebas estadísticas están basadas en el supuesto de que la variable que se estudia tenga una distribución normal. Aunque existen varias formas de hacer la prueba de normalidad, sólo se va a presentar la que utiliza la distribución *chi cuadrado*.

En este caso, la frecuencia esperada es la que se obtiene de calcular el valor de la distribución normal entre ciertos valores. Los grados de libertad son *k-3*. Esto se debe a que se tienen que cumplir tres condiciones en el proceso de ajuste:

$$\sum f_i = \sum F_i$$
$$\overline{X'} = \sum \frac{F_i X_i}{n} \qquad (38)$$
$$\sigma'^2 = \sum \frac{F_i x_i^2}{n}$$

Al tener que satisfacer esas tres condiciones se pierden tres grados de libertad por tanto, los grados de libertad para la prueba serán *k-3*.

Ejemplo

Con una distribución normal cuya media sea 10 y la desviación estándar fuera de 3, se han recolectado 250 observaciones de cierta variable y han sido clasificadas en los rangos que indica la siguiente tabla:

Entre	y	Frecuencia absoluta observada	Frecuencia relativa acumulada teórica f_i	Frecuencia relativa teórica	Frecuencia absoluta observada F_i
	-2		0,00%		
-2	1	1	0,13%	0,13%	0,34
1	4	6	2,28%	2,14%	5,35
4	7	33	15,87%	13,59%	33,98
7	10	110	50,00%	34,13%	85,34
10	13	75	84,13%	34,13%	85,34
13	16	15	97,72%	13,59%	33,98
16	19	10	99,87%	2,14%	5,35

Para las pruebas de ajuste *chi cuadrado* se recomienda que por lo menos en cada intervalo o valor, haya por lo menos 5 observaciones. Por tanto, se fusionan los dos primeros intervalos.

Entre	y	Frecuencia absoluta observada	Frecuencia absoluta observada Fi	*Fi--fi*	$\frac{(F_i - f_i)^2}{f_i}$
-2	4	7	5,68751551	1,31248449	0,30287663
4	7	33	33,9762994	-0,975	0,02797906
7	10	110	85,336185	24,663815	7,12832159
10	13	75	85,3361851	-10,3361851	1,25195101
13	16	15	33,9762994	-18,9762994	10,5985627
16	19	10	5,3500237	4,6499763	4,04152968
19	22				
		Prueba chi cuadrado inversa (5%, 5)	11,0704826	Resultado $\sum \frac{(F_i - f_i)^2}{f_i}$	23,3512207

Si los parámetros se estiman a partir de los datos de la muestra, entonces se pierden 3 grados de libertad. Si los parámetros no se estiman, sino que se conocen desde el universo, entonces se pierde un grado de libertad. En este caso los parámetros se conocen, por tanto, el número de grados de libertad será 5 (6-1). La Frecuencia relativa acumulada teórica se calcula con la función de Excel *=DISTR.NORM(valor de X (valor superior del rango); media;desviación estándar;VERDADERO). VERDADERO* indica que se está calculando el valor acumulado entre menos infinito y el valor superior del rango (*X*).

$\frac{(F_i - f_i)^2}{f_i}$

La estadística se debe calcular con base *en las observaciones*, no en la frecuencia relativa. El valor máximo permisible del total es de 11,0704826 y se calcula con la función *=PRUEBA.CHI.INV(probabilidad %; grados de libertad)*. En este ejemplo se utilizó un nivel de 5% y 5 grados de libertad. Como el resultado 23,3512207 es mucho mayor que el permitido, 11,0704826, entonces *se rechaza* la hipótesis de que las observaciones provienen de una distribución normal.

Tablas de contingencia

Las tablas de contingencia muestran asociaciones entre clasificaciones. La forma más simple de tablas de contingencia es la llamada tablas 2x2. Todas las tablas de contingencia se pueden construir con una opción en el menú *Datos de Excel*, bajo *Asistente de tablas dinámicas*.

Tablas 2x2

La siguiente tabla clasifica a una población de adultos entre filiación política y edad, así:

Edad	Partido A	Partido B	Total
Menor de 30 años	77	323	400
Mayor de 30 años	177	223	400
Total	254	546	800

Lo que se pretende con el análisis de las tablas de contingencia 2x2 es verificar si una clasificación es independiente de la otra. ¿Tiene algún efecto la edad de la población en la

escogencia de la filiación política? Esto quiere decir que se trata de una prueba de independencia.

Cuando un elemento se tabula como en la tabla anterior si las frecuencias de cada fila son proporcionales a las de las otras filas o si las frecuencias de cada columna son proporcionales a las de las otras columnas, entonces las dos clasificaciones son independientes una de otra.

Para calcular la frecuencia esperada de cada celda se toma el total de cada columna y se divide en la misma proporción en que están divididos los grandes totales de las filas. Así, la celda *Partido A* – menor de 30 años, tendrá como frecuencia esperada 254x400/800 = 127, y así las demás. Entonces las frecuencias esperadas para cada celda serán:

Edad	Partido A	*Partido B*	*Total*
Menor de 30 años	127	273	400
Mayor de 30 años	127	273	400
Total	254	546	800

El cálculo de $\sum_{i=1}^{k} \frac{(F_i - f_i)}{f_i}$ será:

$(77-127)^2/127 + (177-127)^2/127 \;\; (323-273)^2/273 + (223-273)^2/273 = 57{,}7$

Para determinar los grados de libertad se debe tener en cuenta cuáles son las restricciones que se imponen en la tabla 2x2. Estas son:

$f_{11} + f_{12}$ = Total de la fila 1
$f_{21} + f_{22}$ = Total de la fila 2
$f_{11} + f_{21}$ = Total de la columna 1
$f_{12} + f_{22}$ = Total de la columna 2

Como una de ellas se puede obtener de las otras, entonces los grados de libertad que se pierden son 3 y los grados de libertad para la distribución valen 1 (2-1)(2-1).

El máximo valor permitido con un nivel de 5% es de 3,84, por tanto se rechaza la hipótesis de independencia y se dice que la edad sí es determinante de la filiación política.

Tablas de contingencia rxc

El análisis de las tablas de contingencia se puede generalizar para cualquier número de grupos de clasificación en los dos sentidos. En ese caso se dice que son tablas de contingencia rxc y los grados de libertad serán (r-1)x(c-1). El procedimiento es similar al presentado.

Pruebas de signo y orden

Estas son pruebas típicas no paramétricas. Miran más a las relaciones entre los valores que los valores mismos. Por ejemplo, lo importante no es si los valores de unas muestras son 2 y 5, sino que el valor de la muestra dos es más alto que el de la uno.

Prueba de signos

Cuando se trata de medir la diferencia en la media de dos poblaciones en condiciones paramétricas, se requiere que las dos muestras sean independientes y que provengan de

universos normales con igual varianza. Si alguna de estas dos condiciones no se cumple hay que usar una prueba no paramétrica llamada del signo. Esto es, se comparan los valores y se determina cuál es el mayor y con base en el número de signos positivos (si fuera mayor) o negativos (si fuera menor o viceversa) se analiza la información.

Ejemplo

Se registra el promedio 25 estudiantes antes y después de tomar un taller de métodos de estudio. Se trata de estudiar si ese taller aumenta o no el promedio.

Efecto de un taller de métodos de estudio sobre el rendimiento

Sujeto	Antes	después	Signo del cambio		
1	3,7	3,8	+	1	1
2	3,7	3,5	-	0	1
3	3,55	3,6	+	1	1
4	3,4	3,35	-	0	1
5	3,7	3,65	-	0	1
6	3,5	3,55	+	1	1
7	3,6	3,65	+	1	1
8	3,7	3,8	+	1	1
9	3,45	3,5	+	1	1
10	3,4	3,5	+	1	1
11	3,55	3,6	+	1	1
12	3,45	3,6	+	1	1
13	3,65	3,45	-	0	1
14	3,35	3,4	+	1	1
15	3,55	3,65	+	1	1
16	3,6	3,65	+	1	1
17	3,6	3,5	-	0	1
18	3,6	3,6	No cambia	0	0
19	3,45	3,6	+	1	1
20	3,65	3,45	-	0	1
21	3,35	3,4	+	1	1
22	3,4	3,5	+	1	1
23	3,45	3,45	No cambia	0	0
24	3,6	3,75	+	1	1
25	3,7	3,65	-	0	1
				16	23

Si el taller no tuviera efecto sobre el promedio se esperaría que la mitad de las veces el promedio aumentara y la otra mitad disminuyera. Las muestras con igual valor de promedio se desechan.

Para una muestra de 23, se espera que la media de cambios de signo sea de:

Media: pxn 11,5.

La desviación estándar sería: $(np(1-p))1/2$ 2,39791576

Si la hipótesis es cierta, entonces el promedio debería ser 11,5. La pregunta ahora es: ¿Cuál es la probabilidad de que el número de signos + sea de 16 o mayor?

Usamos la normal con media 11,5 y desviación estándar 2,39791576; probabilidad de que sea mayor 3,03%.

Esto significa que la probabilidad de que esto ocurra por azar es de 3,03%.

Si aceptamos un error de 5% el valor obtenido está dentro de ese margen de error y no rechazaríamos la hipótesis de que tomar un curso de métodos de estudio influye en el rendimiento.

La binomial se puede aproximar a la normal cuando $n \geq 10$.

Ejemplo

Un panel debe hacer una prueba de sabor de dos productos. Le asigna 1 al sabor que prefiere y 1 al otro.

Jurado	Sabor uno	Sabor dos	Signo del cambio		
A	1	0	+	1	1
B	1	0	+	1	1
C	0	1	-	0	1
D	1	0	+	1	1
E	1	0	+	1	1
F	1	0	+	1	1
G	0	1	-	0	1
H	1	0	+	1	1
I	1	0	+	1	1
J	0	1	-	0	1
K	1	0	+	1	1
L	0	1	-	0	1
				8	12

Si no hubiera preferencia sobre los sabores se esperaría que la mitad de las veces el panel *prefiriera* el uno y la otra mitad el dos. Las muestras con igual valor de calificación se desechan.

Para una muestra de 12 se espera que la media de cambios de signo sea: media $pxn = 6$; la desviación estándar es desviación estándar = $(np(1-p))1/2 = 1{,}73205081$.

Si la hipótesis es cierta, entonces el promedio signos positivos debería ser 6.

La pregunta ahora es: ¿Cuál es la probabilidad de que el número de signos + sea de 8 o mayor?

Usamos la normal con media 6 y desviación estándar 1,7320508.

Probabilidad de que sea mayor: 12,41%.

Esto significa que la probabilidad de que esto ocurra por azar es de 12,41%.

Si aceptamos un error de 5% el valor obtenido no está dentro de ese margen de error y rechazaríamos la hipótesis de que no hay preferencia entre los jurados del panel por ningún sabor.

La binomial se puede aproximar a la normal cuando $n \geq 10$

Prueba U de Mann-Whitney

Cuando las muestras son independientes y tienen varianzas diferentes se puede utilizar esta prueba.

Ejemplo

Se aplica un cierto examen a estudiantes de la jornada diurna y ese mismo examen a estudiantes de la jornada nocturna. Se quiere probar si el rendimiento de los dos grupos es diferente o no.

Nota	Jornada	Nota	Jornada
3,50	Diurno	3,60	Nocturno
3,40	Diurno	3,35	Nocturno
3,65	Diurno	3,70	Nocturno
4,05	Diurno	3,25	Nocturno
3,30	Diurno	3,15	Nocturno
2,80	Diurno	3,85	Nocturno
3,10	Diurno	3,55	Nocturno
3,75	Diurno	3,00	Nocturno
4,15	Diurno	3,80	Nocturno
2,40	Diurno	3,05	Nocturno
		3,20	Nocturno

Esta prueba exige los siguientes pasos:

Paso 1

Asignar un orden a toda la muestra combinada (diurno y nocturno). Así, 1 a la más baja nota, 2 a la siguiente, etc. En el ejemplo, 1 a 2,4, 2 a 2,8, 3 a 3,0, etc.

Nota	Jornada	Orden
2,40	Diurno	1
2,80	Diurno	2
3,00	Nocturno	3
3,05	Nocturno	4
3,10	Diurno	5
3,15	Nocturno	6
3,20	Nocturno	7
3,25	Nocturno	8
3,30	Diurno	9
3,35	Nocturno	10
3,40	Diurno	11
3,50	Diurno	12
3,55	Nocturno	13
3,60	Nocturno	14
3,65	Diurno	15
3,70	Nocturno	16
3,75	Diurno	17
3,80	Nocturno	18
3,85	Nocturno	19
4,05	Diurno	20
4,15	Diurno	21

Paso 2

Se suman todos los rangos de cada grupo:

Nota	Jornada	Orden	Nota	Jornada	Orden
2,40	Diurno	1	3,00	Nocturno	3
2,80	Diurno	2	3,05	Nocturno	4
3,10	Diurno	5	3,15	Nocturno	6
3,30	Diurno	9	3,20	Nocturno	7
3,40	Diurno	11	3,25	Nocturno	8
3,50	Diurno	12	3,35	Nocturno	10
3,65	Diurno	15	3,55	Nocturno	13
3,75	Diurno	17	3,60	Nocturno	14
4,05	Diurno	20	3,70	Nocturno	16
4,15	Diurno	21	3,80	Nocturno	18
			3,85	Nocturno	19
	R1	113		R2	118

$$U = n_1 n_2 + \left(\frac{n_1 (n_1 + 1)}{2} \right) - R_1 \tag{39}$$

o:

$$U = n_1 n_2 + \left(\frac{n_2 (n2 + 1)}{2} \right) - R2 \tag{40}$$

Con la primera fórmula, *n1*= 10, *n2* = 11, *R1* = 113

U= 52

Paso 4

Se determina la media y la desviación estándar de *U*

Media:

$$E(U) = \frac{n_1 n_2}{2} \tag{41}$$

Media: 55.

Desviación estándar:

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \tag{42}$$

Desviación estándar:14,2009389

Si *n1* y *n2* son mayores que 8, entonces se puede aproximar a la normal. La probabilidad de que el valor de *U*≥52, sea producto del azar.

Probabilidad *U*≥ 5258,37%.

Si se acepta un nivel de error de 5%, entonces la prueba indicaría que no hay evidencia de que el rendimiento es el mismo.

Observaciones:

1) Si hay empates, los valores iguales reciben el rango promedio de sus rangos empatados (por ej. Si los valores 5º y 6º están empatados ambos reciben entonces el rango 5,5) y la que sigue recibe el rango siguiente (por ej. 7º)

2) *U* tiene distribución aproximadamente normal sólo cuando *n1* y *n2* son ≥ 8. Si esto no se cumple la normal no sirve.

Prueba H de Kruskal-Wallis

Es una generalización de la *Prueba U de Mann-Whitney*. Se utiliza para examinar la hipótesis nula de que varias muestras independientes pertenecen a poblaciones idénticas. Se asignan los ordenamientos a cada observación teniendo en cuenta todas las muestras. Al menor 1, al siguiente 2, etc. Con esos datos se calcula la estadística *H*:

$$H = \frac{12}{n(n+1)}\left(\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \cdots + \frac{R_k^2}{n_k}\right) - 3(n+1) \qquad (43)$$

Donde *k* es el número de muestras, R_k es la suma de los rangos de la *k*-ésima muestra, y *n* es el número total de observaciones ($n_1+n_2+...+n_k$)

Si se supone que la hipótesis nula es verdadera y que cada muestra consiste de por lo menos 5 observaciones, *H* tiene una distribución que se puede aproximar a la *Chi-cuadrado* con (*k*-1) grados de libertad.

Ejemplo

Se quiere probar una metodología educativa con tres grupos de estudiantes y se mide su mejora porcentual en el promedio acumulado:

Aumento %	Metodología	Aumento %	Metodología	Aumento %	Metodología
21,65%	A	24,79%	B	12,54%	C
18,19%	A	31,39%	B	14,74%	C
16,62%	A	34,22%	B	16,62%	C
27,62%	A	25,42%	B	17,88%	C
23,85%	A	29,19%	B	12,85%	C
20,08%	A	30,76%	B	10,34%	C
		34,85%	B	21,65%	C
				19,14%	C
				9,08%	C

Aumento %	Met.	Orden	Aumento %	Met.	Orden	Aumento %	Met	Orden
16,62%	A	6,5	24,79%	B	15	9,08%	C	
18,19%	A	9	25,42%	B	16	10,34%	C	
20,08%	A	11	29,19%	B	18	12,54%	C	
21,65%	A	12,5	30,76%	B	19	12,85%	C	
23,85%	A	14	31,39%	B	20	14,74%	C	
27,62%	A	17	34,22%	B	21	16,62%	C	
			34,85%	B	22	17,88%	C	
						19,14%	C	
						21,65%	C	1
	R1	70		*R2*	131		*R3*	
n1	6		*n3*	7		*n3*	9	
n=	22							

H = 15,6327875

Valor de *chi cuadrado* a 1% = 9,21035104

Como *H* es mayor entonces se rechaza que las tres metodologías no tienen igual efectividad.

Correlación de orden

El coeficiente de correlación, como ya se sabe, va a permitir medir la asociación entre dos variables, en este caso, de orden de una variable. En este caso se utilizará el coeficiente de correlación de orden o rango de Spearman. Por ejemplo, para medir calificaciones de evaluación de un grupo de personas o productos, a veces puede interesar si los evaluadores han sido coherentes en sus evaluaciones del mismo grupo de sujetos. Por ejemplo, la evaluación que hace un director de departamento comparado con la evaluación del decano.

Ejemplo

Tanto el decano como el director del departamento, han evaluado a los profesores y su resultado ha sido ordenado de acuerdo con las calificaciones de cada uno, así:

	Evaluación de	
Profesor	Decano	Director
A	7	8
B	5	6
C	4	5
D	3	4
E	2	1
F	6	3
G	8	10
H	10	9
I	9	7
J	1	2

El coeficiente r_s de *Spearman* está definido como:

$$r_s = 1 - \frac{6\sum d^2}{n\left(n^2-1\right)} \qquad (44)$$

Donde n es el número de observaciones pareadas y d es la diferencia entre cada par de rangos.

Profesor	Decano	Director	d	d^2
A	7	8	1	1
B	5	6	1	1
C	4	5	1	1
D	3	4	1	1
E	2	1	-1	1
F	6	3	-3	9
G	8	10	2	4
H	10	9	-1	1
I	9	7	-2	4
J	1	2	1	1
10			suma	24
r_s	0,85			

Ejemplo

Se quiere saber si existe alguna relación entre la prueba del ICFES y el rendimiento de los estudiantes.

Estudiante	Promedio ponderado	Puntaje ICFES
1	3,20	253
2	3,96	313
3	3,02	239
4	4,13	327
5	2,69	213
6	3,89	308
7	4,34	344
8	4,66	350
9	4,78	330
10	4,80	321
11	4,05	322
12	3,66	321
13	4,25	318
14	3,51	278
15	3,61	286
16	3,57	261
17	4,40	320
18	3,08	293
19	4,60	322
20	3,15	222
21	4,10	313
22	4,42	338
23	4,75	352
24	4,80	322
25	4,60	339
26	4,16	295

Estudiante	Promedio ponderado	Puntaje ICFES	Orden de promedio	Orden de puntaje	*d*	d^2
24	4,80	322	1	9	8	64
10	4,80	321	1	11,5	10,5	110,25
9	4,78	330	3	6	3	9
23	4,75	352	4	1	-3	9
8	4,66	350	5	2	-3	9
19	4,60	322	6,5	9	2,5	6,25
25	4,60	339	6,5	4	-2,5	6,25
22	4,42	338	8	5	-3	9
17	4,40	320	9	13	4	16
7	4,34	344	10	3	-7	49
13	4,25	318	11	14	3	9
26	4,16	295	12	18	6	36
4	4,13	327	13	7	-6	36
21	4,10	313	14	15,5	1,5	2,25
11	4,05	322	15	9	-6	36
2	3,96	313	16	15,5	-0,5	0,25
6	3,89	308	17	17	0	0
12	3,66	321	18	11,5	-6,5	42,25
15	3,61	286	19	20	1	1
16	3,57	261	20	22	2	4
14	3,51	278	21	21	0	0
1	3,20	253	22	23	1	1
20	3,15	222	23	25	2	4
18	3,08	293	24	19	-5	25
3	3,02	239	25	24	-1	1
5	2,69	213	26	26	0	0

$N = 26$ Suma total = 485,50.

$r_s = 0,83$.

Significancia estadística de r_s

Hipótesis: No existe relación entre los ordenamientos. Si $n \geq 25$, entonces se podría suponer distribución normal. La media de *rs* es 0 y desviación estándar es:

$$\sigma_{r_s} = \frac{1}{\sqrt{n-1}} \qquad (45)$$

= 0,19611614

Probabilidad de que r_s sea mayor en valor absoluto que 0,83 = 2,11379*E*-05.

Esto significa que el coeficiente r_s es significativo al 1%.

Muestreo aleatorio

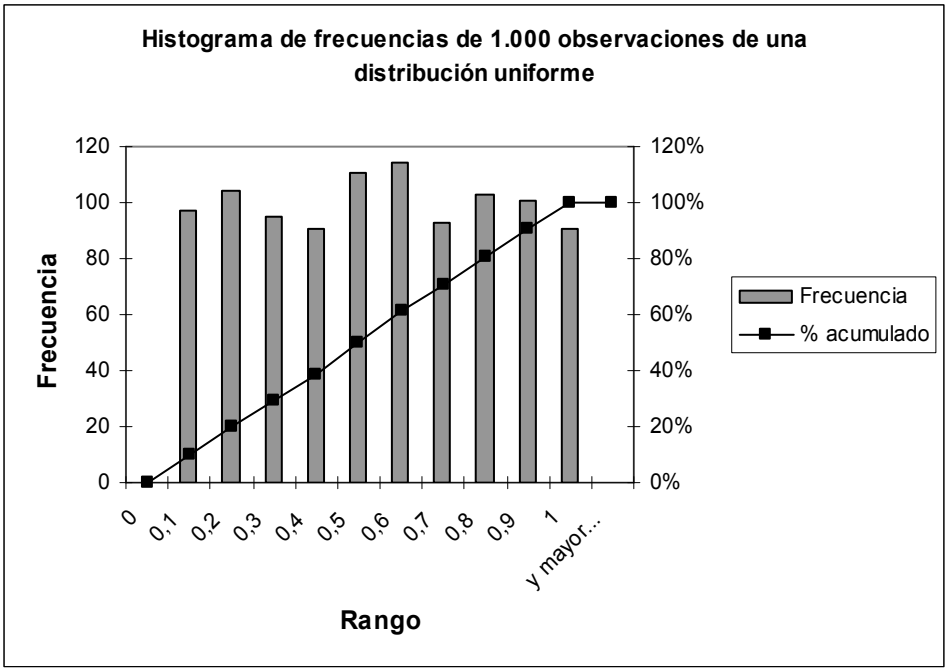
En el archivo *Estadística.XLS* hay una hoja que se llama *aleatorio*. Allí aparecen 1.000 números aleatorios, entre 0 y 1. Un número aleatorio tiene como característica que tiene igual probabilidad de "salir" que cualquier otro. Un ejemplo simple es una bolsa con 10 bolas numeradas del 0 al 9. Esos son 10 dígitos que si sacamos al azar cualquier bola, cualquiera de los números tiene igual probabilidad de salir. En este caso, la probabilidad es de 1/10. Si se reemplaza la bola y siempre sacamos bolas de la bolsa con las 10

numeradas como se indicó, estamos ante un mecanismo de generación, números aleatorios. Esta distribución se conoce como distribución uniforme.

Cuando se hace el histograma de frecuencias de los 1.000 se encuentra lo siguiente:

Rangos	*Frecuencia absoluta*	*% acumulado*
0	0	,00%
0,1	97	9,70%
0,2	104	20,10%
0,3	95	29,60%
0,4	91	38,70%
0,5	111	49,80%
0,6	114	61,20%
0,7	93	70,50%
0,8	103	80,80%
0,9	101	90,90%
1	91	100,00%
y mayor...	0	100,00%

Y su gráfica es:

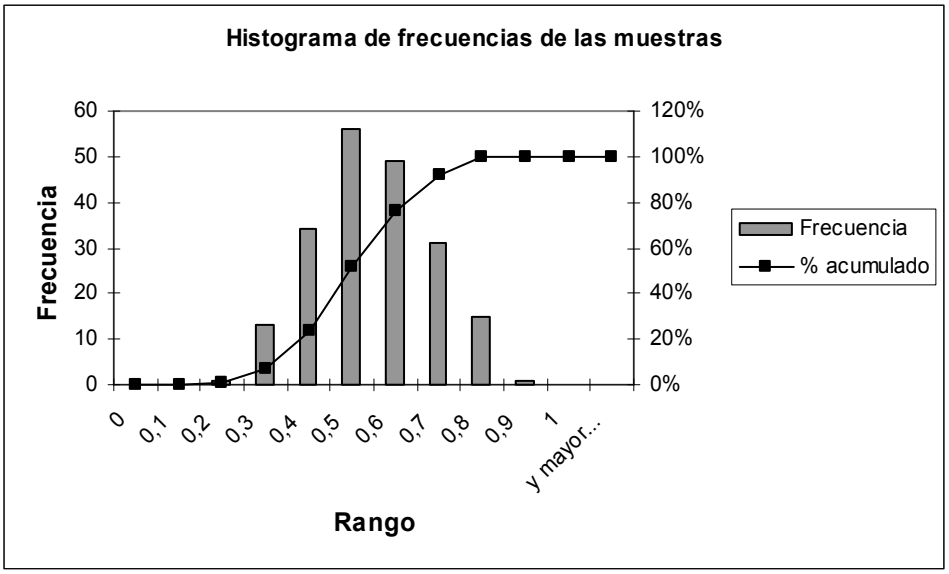


En esta gráfica se observa que la frecuencia es cercana a 100, por tratarse de una distribución uniforme. Es fácil intuir que la media de esta distribución es 0,5.

Si se dibuja el histograma de frecuencias de los promedios de las muestras de tamaño 5, se encuentra lo siguiente:

Rangos	*Frecuencia absoluta*	*% acumulado*
0	0	,00%
0,1	0	,00%
0,2	1	,50%
0,3	13	7,00%
0,4	34	24,00%
0,5	56	52,00%
0,6	49	76,50%
0,7	31	92,00%
0,8	15	99,50%
0,9	1	100,00%
1	0	100,00%
y mayor...	0	100,00%

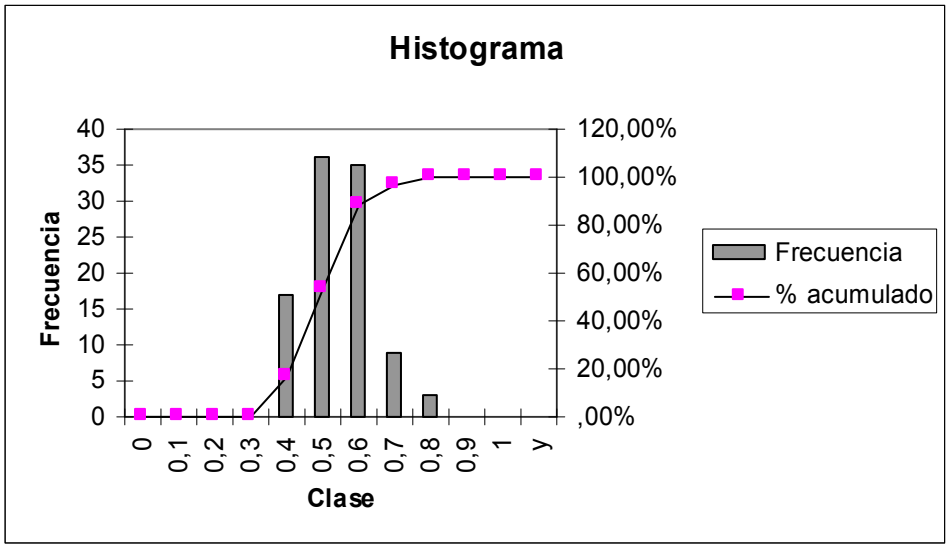
Y en la gráfica:



Si la muestra es de tamaño 10, los resultados son los siguientes:

Clase	Frecuencia	% acumulado
0	0	,00%
0,1	0	,00%
0,2	0	,00%
0,3	0	,00%
0,4	17	17,00%
0,5	36	53,00%
0,6	35	88,00%
0,7	9	97,00%
0,8	3	100,00%
0,9	0	100,00%
1	0	100,00%
y mayor...	0	100,00%

En gráfica:



Observe que ahora el histograma es más estrecho. ¿Qué ha sucedido? Pues que al tomar muestras de una distribución uniforme, y al calcular, además, el promedio de cada muestra, éste tiende a comportarse como una distribución normal.

Esta es una propiedad muy importante que se debe tener en cuenta. Se puede observar además que el rango de variación es mucho menor para los promedios de las muestras que para los datos originales. Esto se puede apreciar en la siguiente tabla:

	Muestras		Valores originales	
Rangos	**Frecuencia absoluta**	*% acumulado*	*Frecuencia absoluta*	*% acumulado*
0	0	,00%	0	,00%
0,1	0	,00%	97	9,70%
0,2	1	,50%	104	20,10%
0,3	13	7,00%	95	29,60%
0,4	34	24,00%	91	38,70%
0,5	56	52,00%	111	49,80%
0,6	49	76,50%	114	61,20%
0,7	31	92,00%	93	70,50%
0,8	15	99,50%	103	80,80%
0,9	1	100,00%	101	90,90%
1	0	100,00%	91	100,00%
y mayor...	0	100,00%	0	100,00%

Observado esto, se puede plantear una ley estadística que dice que la distribución de probabilidad del promedio de las muestras tomadas de una distribución cualquiera, tiende a ser normal con media igual a la media de la distribución original, y la desviación estándar igual a la desviación estándar de la distribución original dividida por la raíz cuadrada del tamaño de la muestra.

Con los datos disponibles se tiene:

Desviación estándar total de observaciones: 0,28472219.

Desviación estándar de la muestra: 0,13323466.

Promedio del total de observaciones: 0,4977773.

Promedio de la muestra: 0,4977773.

Los valores teóricos de la media y de la desviación estándar de la distribución uniforme son:

$$Media\ \mu = \frac{Valor\ máximo + valor\ mínimo}{2} = \frac{1-0}{2} = 0{,}5$$

$$Desviación\ estándar\ \sigma = \frac{Valor\ máximo - valor\ mínimo}{\sqrt{12}} = \frac{1}{\sqrt{12}} = 0{,}28867513$$

Los valores no coinciden exactamente porque los de la tabla fueron calculados de una muestra de 1.000 observaciones y no es el universo total.

$$Desviación\ estándar\ del\ promedio\ de\ la\ muestra\ \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0{,}28867513}{\sqrt{5}} = 0{,}12909944$$

Cuando se toma una muestra, se espera entonces que los resultados obtenidos por la muestra fluctúen alrededor de su media y que cerca del 100% de los valores se encuentren entre $\overline{X} \pm 3\sigma_{\bar{x}}$. Esta variación hacia arriba o hacia abajo de la media es el error permitido, y éste puede ser definido a voluntad. De manera que se puede construir una tabla que muestre el porcentaje de las muestras que caen entre dos valores:

Entre	% observaciones
$\overline{X} \pm \sigma_{\overline{X}}$	68,27
$\overline{X} \pm 2\sigma_{\overline{X}}$	95,45
$\overline{X} \pm 3\sigma_{\overline{X}}$	99,73

Esto indica la confiabilidad del muestreo. De modo que es posible definir la confiabilidad del muestreo (95%, 90%, etc.), el error (*e*) que se acepta y, conociendo (o calculando aproximadamente) la desviación estándar de la población, se puede definir el tamaño de la muestra.

Si se denomina el número de desviaciones estándar por arriba y por debajo de la media como *Z*, entonces se puede expresar el error como:

$$error\ e = Z\sigma_{\overline{X}} = Z\frac{\sigma}{\sqrt{n}}$$
$$n = \frac{Z^2\sigma^2}{e^2} \quad (46)$$

Este valor de *n* es válido cuando se tiene una población infinita, de modo que no hay agotamiento del universo. Cuando se trata de una población finita, se incurre en agotamiento del universo y la probabilidad de la primera muestra es diferente de todas las subsiguientes.

En el caso de un universo finito de tamaño *N*, es necesario hacer un ajuste, así:

$$Tamaño\ de\ la\ muestra\ n_m = \frac{n}{\left(1+\frac{n}{N}\right)} \quad (47)$$

Hay un caso de especial interés y es la distribución binomial que rige experimentos tales como encuestas de opinión o de mercado, donde el resultado a medir es un porcentaje. En ese caso, se sabe ya que los parámetros de la distribución son *p* y *pq*. Entonces el cálculo del tamaño de la muestra será:

$$error\ e = Z\sigma_{\overline{p}} = Z\frac{\sqrt{p(1-p)}}{\sqrt{n}}$$
$$n = \frac{Z^2 p(1-p)}{e^2} \quad (48)$$

En este caso el error *e* es un porcentaje en relación con el porcentaje que se desea medir. Como no siempre se puede calcular el valor de *p*, se adopta una posición conservadora para *p*. Con $p = 0{,}5$ se garantiza la máxima varianza, por tanto, el cálculo de n resulta el mayor posible, dados unos niveles de error y de confiabilidad.

Por ejemplo, si se desea hacer una encuesta sobre un universo finito de 400 elementos, se desea una confiabilidad de 95% y se acepta un error en el cálculo de la respuesta a una determinada pregunta de ±2%. Entonces esto debe interpretarse así: el valor de *Z* debe ser tal que el área bajo la curva normal sea 95%, pues cobija valores por encima y por debajo de la media. Esto quiere decir que se debe encontrar un *Z* tal que el área desde menos infinito y *Z* sea de 97,5%. Esto se encuentra con la función de *Excel*

=DISTR.NORM.ESTAND.INV(0,975). Y arroja como resultado 1,96. ¡De aquí sale el famoso 1,96!.

Por otro lado, decir que se acepta un error de 2% significa que si la respuesta a la pregunta indica 10%, el verdadero valor está entre 8% y 12%, con una probabilidad de 95%. Con estos datos, el tamaño de la muestra, si el universo fuera infinito, sería de 2.401. Para un universo finito de 400 elementos, el tamaño sería de 343. En la siguiente tabla se puede observar que a medida que el decisor está dispuesto a tolerar un mayor error, entonces la muestra se reduce, tanto para el universo infinito como para el finito (400 elementos en este ejemplo):

Error	Muestra (n) universo infinito	Muestra (n_m) universo finito N=400
1%	9.603,61861	384,005788
2%	*2.400,90465*	*342,875599*
3%	1.067,06873	290,938989
4%	600,226163	240,036178
5%	384,144744	195,956039

Un análisis similar puede hacerse fijando el error y variando la confiabilidad. Se deja esto como ejercicio al lector. ¿Qué se espera del tamaño de la muestra si se desea más confiabilidad? ¿menos confiabilidad?

En últimas todo termina siendo un problema de costos: Cuánto se está dispuesto a pagar (por hacer una encuesta con mayor o menor cobertura) por reducir el error o aumentar la confiabilidad.

Referencias

Drake, Alwin W. *Fundamentals of Applied Probability Theory*, McGraw-Hill Book Co., 1967.

Makridakis, S., S.C. Wheelwright, *Forecasting. Methods and Applications*, John Wiley, 1978. Existe tercera edición, 1998).

Wonnacott, Thomas H., Ronald J. Wonnacot, *Introductory Statistics for Business and Economics*, 2ª ed., John Wiley, 1977.

Zuwaylif, Fadil H., *Estadística general aplicada*, Addison-Wesley Iberoamericana, 1987.

Conceptos básicos de estadística 2
Estadística Descriptiva 2
Distribuciones de probabilidad 3
Histogramas y Tablas 3
Caso discreto 3
Caso continuo 8
Estadísticas de una distribución 10
Tendencia central de la distribución 10
La moda 10
La mediana 11
La media o valor esperado 11
Medidas de la dispersión de la distribución 11
Varianza 11

Desviación estándar 12
Rango 12
Covarianza 12
Correlación 15
Variable aleatoria 15
La distribución de probabilidad discreta 18
La Distribución Binomial 18
La distribución de probabilidad continua 21
La Distribución de Probabilidad Normal 22
Prueba de hipótesis z para medias de dos muestras 25
La distribución χ^2 (Chi cuadrado o ji cuadrado) 28
La Distribución t de Student 28
La distribución F 30
Estadística no paramétrica 31
Pruebas de normalidad 32
Tablas de contingencia 33
Tablas 2x2 33
Tablas de contingencia rxc 34
Pruebas de signo y orden 34
Prueba de signos 34
Prueba U de Mann-Whitney 36
Prueba H de Kruskal-Wallis 39
Correlación de orden 40
Significancia estadística de r_s 42
Muestreo aleatorio 42
Referencias 48