UNA

AMBIENTALES
UNA

XLStatistics

ANÁLISIS ESTADÍSTICO CON EXCEL

La información independientemente de lo costosa que haya sido crearla, puede ser replicada y compartida a un costo mínimo o nulo. -- Thomas Jefferson

Jorge Fallas

UNA
UNIVERSIDAD NACIONAL
COSTA RICA

2010

Contenido

¿Qué es estadística?

La estadística es el arte-ciencia de tomar decisiones ante situaciones concretas y a la luz de información o datos parciales. En otras palabras, es tomar decisiones bajo condiciones de incertidumbre. La estadística nació en las sociedades antiguas para resolver un problema muy concreto: colectar datos y crear información sobre aspectos tales como producción, población, e impuestos; elementos esenciales para gobernar una nación o un imperio. Este primer aspecto de la estadística todavía persiste y es lo que se conoce como *estadística descriptiva*.

Desde el punto de vista de su aplicación a trabajos científico-matemáticos, la estadística se originó alrededor de 1925 con la publicación de Fisher "Métodos estadísticos para investigadores". Esta publicación marcó el inicio de la *estadística inferencial*. Gracias a esta publicación y a posteriores trabajos de otros científicos la estadística juega un papel esencial en las ciencias naturales y sociales donde las leyes de causa y efecto no pueden deducirse a partir de observaciones individuales.

Formalmente, la estadística puede definirse como una ciencia con un componente teórico y otro aplicado que consiste en crear, desarrollar, y aplicar técnicas o instrumentos que nos permitan evaluar el grado de incertidumbre o error de nuestras generalizaciones. Por ejemplo, si deseamos conocer el diámetro medio a la altura del pecho (dap) de una plantación de roble de 10 hectáreas, ubicada en San José de la Montaña, debemos decidir acerca del tamaño de la muestra y el método de selección; y una vez colectados los datos hay que seleccionar los estadísticos y tablas a utilizar para generalizar los resultados a la plantación (inferencia). Finalmente, debemos indicar el grado de confiabilidad de nuestros resultados o en otras palabras que tan seguros estamos de nuestra estimación.

No existe ninguna medida perfecta y por tanto, todas las mediciones contienen algún grado de error; de donde se desprende que para extraer la información de las mediciones es necesario analizar los errores. Los errores se agrupan en dos grandes categorías: el sesgo o error sistemático que puede modelarse utilizando una ecuación que describe las mediciones, lo que permite eliminar o reducir significativamente su efecto; y el ruido o error aleatorio, el cual no se puede modelar, pero cuyas propiedades estadísticas se pueden utilizar para optimizar los resultados del análisis.

Exactitud y precisión

Como se mencionó previamente, toda medición tiene un error; sin embargo con frecuencia se confunden los términos *error o sesgo* y *precisión* (ver figura 1).

Exactitud: mide el grado de fidelidad o proximidad de la medición con respecto al valor real de la variable. El **error o sesgo** es igual a valor real-valor medido. Para determinar el error en una medición es necesario conocer el *valor real* de la variable medida.

Precisión: La precisión es una medición de la similitud entre mediciones repetidas de una variable. Para variables con una distribución normal, la **varianza** se utiliza para cuantificar la variación del set de datos con respecto a la media.

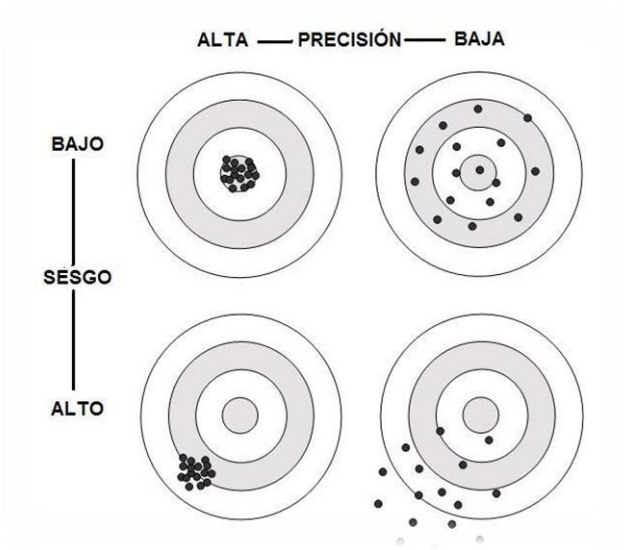


Figura 1: Conceptos de exactitud y precisión. Observe que una medición puede tener una alta precisión y un bajo sesgo o error; sin embargo también puede tener una alta precisión y un alto sesgo o error. Lo deseable es una alta precisión y un error mínimo.

Variables: medición y clasificación

Para utilizar la estadística (descriptiva e inferencial) el investigador(a) debe medir la variable que desea analizar. La *medición* es el procedimiento utilizado para asignar valores a la variable de tal forma que satisfaga las condiciones necesarias para su posterior análisis. La Real Academia Española (http://www.rae.es/rae.html) define el verbo medir como "comparar una cantidad con su respectiva unidad, con el fin de averiguar cuántas veces la segunda está contenida en la primera". La escala de medición es el contexto o marco de referencia bajo el cual se realizan las mediciones; todo medición pertenece a una de las siguientes cuatro escalas: ***nominal***, ***ordinal***, ***intervalo*** y ***razón***. Para decidir cuál prueba estadística puede aplicarse a un set de datos es necesario conocer su escala de medición (**¡Error! No se encuentra el origen de la referencia.**). A su vez, las variables pueden agruparse n cualitativas y cuantitativas.

Variables cualitativas

Estas variables se caracterizan por no expresar una cantidad o magnitud absoluta de lo que se mide y comprenden dos niveles de medición: nominal y ordinal.

Escala nominal

El nivel de medición *nominal* es el más simple; ya que las variables se "miden" utilizando el concepto de igualdad. La especie, el tipo de vegetación y el color de las hojas son ejemplos de mediciones a nivel nominal. Los objetos o eventos se asignan a una u otra clase basados en el concepto de igualdad. Los números o letras asignados a cada categoría son solo códigos y no tienen un orden natural. Por ejemplo, podemos clasificar cinco tipos de uso-cobertura de la tierra de la siguiente manera:

1) bosque seco
2) bosque húmedo
3) mangle
4) pastos y
5) cultivos permanentes

Sin embargo, el valor numérico no indica la precedencia de un tipo de vegetación sobre el siguiente; por ejemplo, el bosque húmedo no es mayor que el bosque seco.

Escala ordinal

En el nivel de medición *ordinal* las variables se miden de acuerdo a su tamaño, valor relativo u orden natural. Esta escala de medición no permite determinar la magnitud de la desigualdad entre categorías contiguas. Por ejemplo, las especies forestales de Costa Rica pueden clasificarse de acuerdo a la densidad de su madera en muy pesadas, pesadas, livianas y muy livianas; sin embargo esta clasificación no indica cuánto más densa es la madera de la primera clase comparada con la segunda o la última. Las variables numéricas o cuantitativas puedes expresarse como variable ordinales utilizando cuantiles, percentiles u otro criterio definido por el usuario(a). Por ejemplo, los datos de la variable "densidad de la madera" pueden dividirse en 5 categorías utilizando quintiles y de esta manera saber cuánto más densa o menos densa es una madera de una categoría con respecto a cualquier otra.

La *escala de actitud de Likert* es un caso especial de una escala de medición **ordinal** que con frecuencia es analizada como una variable cuantitativa. La escala, formada por cinco clases o categorías, fue diseñada con el fin de que las valoraciones sigan una progresión aritmética como se muestra a continuación:

Cuadro 1: Escala de actitud de Likert.

Valores de la escala	Valores de la escala	Valores de la escala
-2 Totalmente en desacuerdo	5 Totalmente en desacuerdo	A Totalmente en desacuerdo
-1 En desacuerdo	4 En desacuerdo	B En desacuerdo
0 Indiferente, indeciso o neutro	3 Indiferente, indeciso o neutro	C Indiferente, indeciso o neutro
1 De acuerdo	2 De acuerdo	D De acuerdo
2 Totalmente de acuerdo	1 Totalmente de acuerdo	E Totalmente de acuerdo

Observe que a diferencia de las variables numéricas o cuantitativas, en la cual los números tienen un orden natural, en la escala de Likert los números o letras asignados a cada categoría son solo códigos y no tienen un orden natural; aunque sí expresan una progresión aritmética; donde se podría considerar la respuesta "Indiferente, indeciso o neutro" como el "cero" de la escala.

Variables cuantitativas

Estas variables se caracterizan por expresar una cantidad o magnitud de lo que se mide y comprenden dos niveles de medición: intervalo y razón.

Escala intervalo y razón

Las escalas de medición de intervalo y razón se diferencian en que la primera no tiene un cero (0) verdadero y la segunda sí. Por ejemplo, variables como temperatura, índices de inteligencia, latitud y fecha se miden a un nivel de intervalo; en tanto que variables como distancia, área y volumen se miden a un nivel de razón. Una temperatura de 0^{o}C no significa la ausencia de temperatura; en tanto que una distancia de 0 mm sí indica la ausencia de distancia. El método más simple para distinguir observaciones entre dichas escalas es aplicar la prueba de razón o proporción a dos valores cualquiera.

El cociente de una razón para observaciones a un nivel de medición de intervalo no tienen sentido o explicación lógica. Por ejemplo, una temperatura de 30°C no es dos veces más caliente que una de 15°C, en tanto que un árbol de 30 metros sí es dos veces más alto que uno de 15 metros. En ambos casos, el cociente es 2 (30/15=2); sin embargo el cero (0) en la escala de grados centígrados es ficticio o sea un punto arbitrario en tanto que en la escala lineal es verdadero. Cualquier operación matemática puede utilizarse e interpretarse en observaciones a un nivel de medición de razón. Para observaciones a un nivel de intervalo sólo tienen sentido la suma, la resta y la multiplicación (cuadro 2).

Cuadro 2: Escalas de medición y operaciones matemáticas que las caracterizan.

Escala de medición	Operaciones matemáticas permitidas
Razón	1. Equivalencia (=) 2. Desigualdad (<, >) 3. Razón de dos intervalos tiene sentido. (a_b/c_d= e) 4. Razón de dos valores tiene sentido (a/b= c)
Intervalo	1. Equivalencia (=) 2. Desigualdad (<,>) 3. Razón de dos intervalos tiene sentido. (a_b/c_d =e)
Ordinal	1. Equivalencia (=) 2. Desigualdad (<,>)
Nominal	1. Equivalencia (=)

Variables circulares

Las variables circulares son un tipo especial de variables cuantitativas que representan ciclos. En estas variables, el valor más grande y el más pequeño se encuentra uno al lado del otro y el punto cero es arbitrario. Algunos ejemplos de variables circulares son: hora del día (0-24), meses del año (enero a diciembre) y la dirección de la brújula (0°-360°). Si se utiliza solo parte del ciclo, una variable circular se convierte en una variable lineal. Por ejemplo, cuando usted utiliza la variable tiempo y la mide como el número días entre dos eventos.

Si su variable es realmente circular (e.g. distancia y dirección de vuelo de las aves), existen pruebas estadísticas diseñadas especialmente para este tipo de variable tales las herramientas de Matlab para estadística circular http://www.kyb.tuebingen.mpg.de/bs/people/berens/circStat.html y el programa comercial Oriana http://www.kovcomp.com/oriana/.

Los métodos estadísticos aplicados a un nivel de medición nominal y ordinal se denominan "**no paramétricos**", en tanto que los aplicados a datos a un nivel de medición de intervalo y razón se denominan "**paramétricos**". Observaciones a un nivel de medición de intervalo y razón pueden transformarse a una escala ordinal o aun nominal. Por ejemplo, si tenemos 10 observaciones de densidad de roble, podemos ordenarlas en forma ascendente, de tal forma que el primer valor es mayor que el segundo, el segundo mayor que el tercero, y así sucesivamente. Luego se asigna un valor de 1 a 10 a cada observación, estos nuevos "valores" se conocen con el nombre de *órdenes*. Esto permite aplicar técnicas no paramétricas a datos medidos originalmente a un nivel apropiado para aplicar técnicas paramétricas.

El proceso de investigación

Existen muchas definiciones del término investigación; sin embargo en el contexto del presente documento la definiremos como el *camino o ruta* que usted sigue para responder a sus preguntas o

someter a prueba sus hipótesis. En el proceso de investigación se pueden reconocer los siguientes elementos (ver figura 2):

A. *Mundo real*: La real que se estudia.
B. *Preguntas*. Lo que deseamos responder mediante el proceso de investigación. Involucra la revisión de estudios previos (revisión del estado del conocimiento). Formulación de preguntas e hipótesis de trabajo.
C. *Poblaciones/muestras/variables*: Transformación del mundo real en elementos estadísticos que puedan ser medidos.
D. *Medición*: Proceso e instrumentos utilizados para recabar datos del objeto de interés.
E. *Análisis de datos*: Métodos y procedimientos utilizados para transformar los datos en información. Someter a prueba hipótesis (diseño experimental/Observacional).
A. *Decisión*: Posición del investigador(ara) frente a los resultados de su estudio y reformulación de hipótesis.
B. *Aplicación*: Acción sobre el mundo real

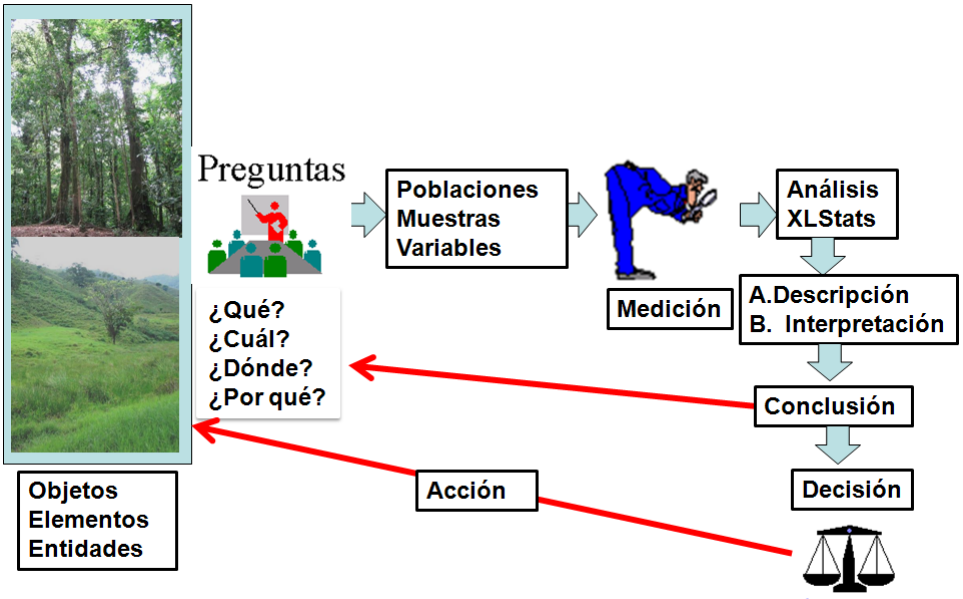


Figura 2: El proceso de investigación.

Terminología

En el análisis de datos estadísticos con frecuencia se utilizan las palabras “variable independiente“ y “variable dependiente”; sin embargo en diferentes disciplinas dichos términos pueden tener diferentes acepciones como se muestra a continuación:

Variable independiente, variable explicativa, variable control, variable manipulada, variable predictiva, regresor, variable de exposición, insumo

Variable dependiente, variable respuesta, variable medida, variable observada, variable explicada, variable resultado, variable experimental, producto.

• Variable "independiente" responde a la pregunta "¿Qué puedo cambiar?"
• Variable "dependiente" responde a la pregunta "¿Qué observo?
• Variable de control responde a la pregunta "¿Qué mantengo constante?"

• Variables externas responden a la pregunta "¿Cuáles variables no consideradas en el análisis pueden mediar en el efecto de la variable “independiente” en la variable dependiente? "

Dado que la mayoría de las variables presentan algún grado de correlación es preferible utilizar los términos *variable respuesta* y *variable explicativa*.

Mediciones repetidas: Las mediciones repetidas se obtienen en un grupo de sujetos o muestras que se miden antes y después de aplicar un tratamiento (e.g. se mide variable respuesta en el sujeto antes de un tratamiento y luego después de aplicar el tratamiento). Por lo tanto, cada sujeto o muestra actúa como su propio control y por esta razón las dos mediciones no son independientes.

Los grupos son establecidos por el emparejamiento de sujetos o muestras utilizando como referencia algún tipo de relación natural entre los sujetos o muestras de cada par. Por ejemplo, las competencias de cada trabajador en el área social podrían compararse con las de su padre/madre. Cada persona podría ser emparejada con su padre/madre en este diseño y debido a esta relación natural, las mediciones no son independientes.

Un diseño emparejado de participantes se emplea cuando los participantes se emparejan basados en resultados similares en una prueba previa. Una persona de cada pareja es asignada al grupo experimental (por ejemplo, el grupo que recibió algún tratamiento) y el otro es asignado al grupo control (que no recibe el tratamiento). El objetivo del diseño es controlar por diferencias en los individuaos. Por ejemplo, podríamos querer controlar la variable inteligencia en la evaluación de los efectos de una droga sobre la memoria. En primer lugar, se podría medir el IQ de los participantes y luego compararlas en base a su índice de inteligencia antes de asignar aleatoriamente a uno a la droga y el otro a la condición sin la droga.

Población: Es el total o universo al cual se desea aplicar la inferencia o conclusión del estudio.

Muestra: Es una parte o porción de la realidad bajo estudio.

Deducción: A partir del todo (población) se deriva una afirmación que aplica a una condición particular (muestra).

Inducción: A partir de una porción de la realidad (muestra) se hace una afirmación sobre el todo (población).

Unidad experimental: Individuo, objeto, grupo o conjunto de sujetos experimentales a los cuales se les aplica un determinado tratamiento. Por ejemplo, la unidad experimental puede ser una parcela en una plantación, un grupo de semillas, un persona a la cual se entrevista, un árbol que se mide, etc. En algunos textos se le denomina a la unidad experimental “caso”.

Observación: Es la medición realizada en una unidad experimental.

Medición: Proceso de asignar un valor numérico ó no numérico a un fenómeno, proceso u objeto.

Tratamientos o variables: Procesos o acciones cuyos efectos serán medidos en el material experimental y posteriormente comparados entre sí para determinar si existen diferencias estadísticamente significativas. Los tratamientos pueden ser cualitativos ó cuantitativos.

Testigo: Tratamiento de referencia utilizado para determinar si los tratamientos tienen un efecto estadísticamente discernible sobre el material experimental.

Variable respuesta: Es aquella propiedad o cualidad de la unidad experimental que se mide. Para mayor detalle sobre el tema ver pág.6.

Repetición: Réplica estadísticamente independiente de un tratamiento. Cuando el tratamiento es aplicado a varias unidades experimentales independientes; cada aplicación brinda una estimación independiente de la respuesta del sujeto experimental al tratamiento. Cuantas más réplicas se tenga mejor será la estimación del error experimental. En la mayoría de los casos se recomienda un mínimo de tres observaciones independientes por tratamiento. La *seudo replicación* es el resultado de muestrear dos o más veces la misma condición (muestras no independientes). Por ejemplo, al evaluar la densidad de peces en dos ríos; uno contaminado y otro no, si se muestrean 5 sitios al azar en cada uno de ellos, dichas muestras no representan réplicas ya que se está muestreando el mismo río. En el sentido estadístico para que se consideren réplicas debería de elegirse al azar dos o más ríos por condición (contaminado-no contaminado) y luego obtener muestras independientes de cada uno de ellos. Esto permitiría estimar la variabilidad natural de cada uno de los sistemas acuáticos en los cuales viven los peces que se muestrean. Aun cuando el análisis de los datos presupone la existencia de réplicas independientes, en la mayoría de los estudios en el área de recursos naturales no es posible cumplir con este supuesto.

Cuasi o seudo experimento: Estudio en el cual se utilizan los principios propuestos por Fisher para el diseño de experimentos; sin embargo, por diversas razones prácticas, no es posible asignar los tratamientos en forma aleatoria. Este tipo de estudios es común en el área de ecología y en general en estudios de tipo observacional.

Significancia estadística: Esta es una regla que permite afirmar que la diferencia observada entre dos o más tratamientos es el resultado del efecto del tratamiento y no del azar. Con frecuencia se declaran como significativas aquellas diferencias que tienen una probabilidad inferior a 0.05 (o sea 5%) de ocurrir en forma aleatoria. En algunos textos de estadística se recomienda utilizar un asterisco (*) para designar las diferencias significativas a un 5% ($P<0.05$), dos asteriscos (**) para designar diferencias significativas al 1% ($P<0.01$) y tres asteriscos (***) para designar diferencias significativas al 0.1% ($P<0.001$). Sin embargo, dado que los paquetes estadísticos le brindan el valor de “p” se recomienda reportar dicho valor y dejar que el lector juzgue por sí mismo(a) la intensidad de la significancia de la prueba.

Consistencia: Un método de análisis estadístico es consistente cuando la significancia de la prueba depende exclusivamente de: 1) la diferencia entre los dos estimadores, 2) el error estándar de las diferencias, 3) el número de grados de libertad del error, y 4) el nivel de significancia al cual se hace la prueba.

Aleatorización: Asignación aleatoria de los tratamientos a los sujetos o unidades experimentales. Esto elimina cualquier sesgo conocido o desconocido en la asignación de los tratamientos.

Error experimental: Variación natural o innata del material experimental no controlado por el investigador(a). Este no es un error adrede o derivado de la aplicación errónea de técnicas de medición sino simplemente un componente propio del material experimental.

Análisis

Un Análisis en sentido amplio es la descomposición de un todo en partes para poder estudiar su estructura y/o sistemas operativos y/o funciones (http://es.wiktionary.org/wiki/).

La acción y el efecto de separar un todo en los elementos que lo componen con el objeto de estudiar su naturaleza, función o significado (http://es.wiktionary.org/wiki/).

La acción y el efecto de identificar, distinguir y clasificar diferentes aspectos integrantes de un campo de estudio, examinando qué relaciones guardan entre ellos y como quedaría modificado el conjunto si se eliminara o se añadiera algún aspecto a los previamente identificados (http://es.wiktionary.org/wiki/).

Documento que revisa, separa o hace un resumen de los elementos o principios de un tema o de una obra (http://es.wiktionary.org/wiki/).

Distinción y separación de las partes de un todo hasta llegar a conocer sus principios o elementos (http://www.rae.es/rae.html).

Ciencia
Conocimiento estructurado y sistemático de las cosas por sus principios y causas; Conjunto de conocimientos que constituyen una rama del saber humano (http://es.wiktionary.org/wiki/).

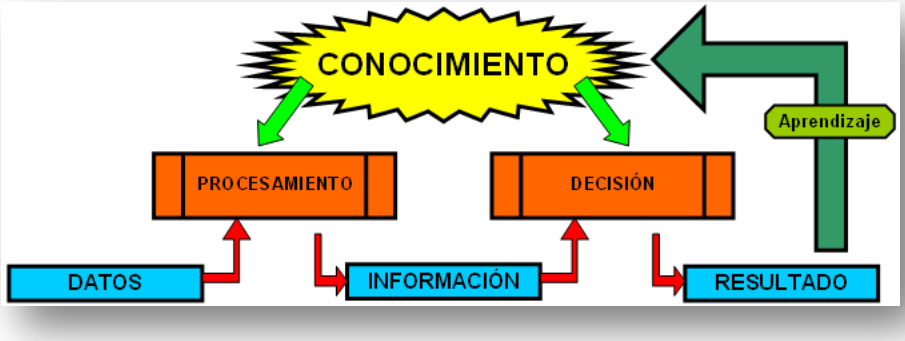
Conjunto de conocimientos obtenidos mediante la observación y el razonamiento, sistemáticamente estructurados y de los que se deducen principios y leyes generales (http://www.rae.es/rae.html).

Conocer
Saber de la existencia de una cosa (http://es.wiktionary.org/wiki/).

Averiguar por el ejercicio de las facultades intelectuales la naturaleza, cualidades y relaciones de las cosas (http://www.rae.es/rae.html).

Conocimiento
Resultado de la acción de conocer. Comprensión, entendimiento, inteligencia, razón (http://es.wiktionary.org/wiki/)

Acción y efecto de conocer; entendimiento, inteligencia, razón natural ello (http://www.rae.es/rae.html).



Esquema sobre el conocimiento desde el punto de vista de las ciencias de la información, como se genera y como se aplica. Fuente: http://es.wikipedia.org/wiki/Saber

Convicción
Acto o efecto de estar seguro sobre algo (http://es.wiktionary.org/wiki/).

Una convicción es una creencia de la que un cierto individuo opina que dispone de suficiente evidencia para considerarla cierta. La diferencia entre una simple creencia y una convicción, es que en el primer caso el individuo puede no tener evidencia suficiente para justificar su veracidad, mientras que en el segundo el individuo si la considera probada, con independencia de que exista evidencia científica o intersubjetiva incontrovertible de que dicha convicción es verdadera (http://es.wikipedia.org/wiki/Saber).

Idea religiosa, ética o política a la que se está fuertemente adherido.
(http://www.rae.es/rae.html).

Creencia
Algo en lo que se cree, confianza en que algo existe o que es cierto.
(http://es.wiktionary.org/wiki/).

Una creencia es una proposición o conjunto de ellas, que un cierto individuo considera ciertas, pero para la que en general no existe evidencia intersubjetiva suficiente para considerarla conocimiento propiamente dicho. Una creencia puede ser acertada o equivocada. Sin embargo, aunque en el uso contidiano al oponer "creencia" y "conocimiento", el primero se usa frecuentemente con el sentido de proposiciones que alguien considera ciertas, pero de la que existe evidencia de estar equivocadas o ser indemostrables(http://es.wikipedia.org/wiki/Saber).

Firme asentimiento y conformidad con algo; completo crédito que se presta a un hecho o noticia como seguros o ciertos; religión, doctrina (http://www.rae.es/rae.html).

Criterio
Norma para conocer la verdad (http://www.rae.es/rae.html).
Juicio o discernimiento (http://www.rae.es/rae.html).

Evaluar
Señalar el valor de algo (http://www.rae.es/rae.html).
Estimar, apreciar, calcular el valor de algo (http://www.rae.es/rae.html).
Estimar los conocimientos, aptitudes y rendimiento de los alumno (http://www.rae.es/rae.html).

Evaluación
Valoración de los conocimientos que se da sobre una persona o situación basándose en una evidencia constatable (http://es.wiktionary.org/wiki/).

Evidencia
Certeza clara y manifiesta de la que no se puede dudar (http://www.rae.es/rae.html).
Prueba determinante en un proceso (http://www.rae.es/rae.html).

Instrumento
Objeto o aparato, normalmente artificial, que se emplea para facilitar o posibilitar un trabajo, ampliando las capacidades naturales del cuerpo humano. Sinónimos: herramienta, utensilio, útil (http://es.wiktionary.org/wiki/).

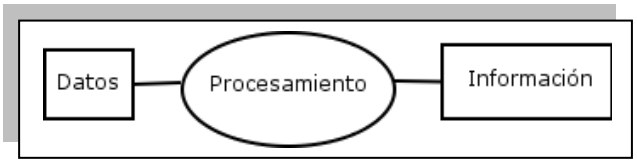
Aquello que sirve de medio para hacer algo o conseguir un fin; Conjunto de diversas piezas combinadas adecuadamente para que sirva con determinado objeto en el ejercicio de las artes y oficios (http://www.rae.es/rae.html).

Información
Comunicación o adquisición de conocimientos que permiten ampliar o precisar los que se poseen sobre una materia determinada (http://www.rae.es/rae.html).

Conocimientos así comunicados o adquiridos (http://www.rae.es/rae.html).

La información es un fenómeno que proporciona significado o sentido a las cosas. En sentido general, la información es un conjunto organizado de datos procesados, que constituyen un mensaje sobre un determinado ente o fenómeno. Los datos se perciben, se integran y generan la información necesaria para producir el conocimiento que es el que finalmente permite tomar decisiones para realizar las acciones cotidianas que aseguran la existencia. La sabiduría consiste en determinar correctamente cuándo, cómo, dónde y con qué objetivo emplear el conocimiento adquirido (http://es.wikipedia.org/wiki/Informaci%C3%B3n).

Principales características de la información
Significado (semántica)
Importancia (relativa al receptor)
Vigencia (en la dimensión espacio-tiempo)
Validez (relativa al emisor)
Valor (activo intangible volátil)
Polimorfismo



Herramienta
Objeto o aparato, normalmente artificial, que se emplea para facilitar o posibilitar un trabajo, ampliando las capacidades naturales del cuerpo humano. Sinónimos: instrumento, utensilio (http://es.wiktionary.org/wiki/).

Instrumento, por lo común de hierro o acero, con que trabajan los artesanos (http://www.rae.es/rae.html).

Método
Procedimiento, técnica o manera de hacer algo, en especial si se hace siguiendo un plan, o de forma sistemática, ordenada y lógica. (http://es.wiktionary.org/wiki/).

Lista ordenada de partes o pasos (avance logrado para la consecución de una tarea.) para lograr un fin (http://es.wiktionary.org/wiki/).

Procedimientos y técnicas característicos de una disciplina o rama del saber (http://es.wiktionary.org/wiki/).

Procedimiento que se sigue en las ciencias para hallar la verdad y enseñarla (http://www.rae.es/rae.html).

Percepción
La percepción es la función psíquica que permite al organismo, a través de los sentidos, recibir, elaborar e interpretar la información proveniente de su entorno (http://es.wikipedia.org/wiki/Percepci%C3%B3n).

Acción y efecto de percibir (http://www.rae.es/rae.html).
Sensación interior que resulta de una impresión material hecha en nuestros sentidos. (http://www.rae.es/rae.html).

Conocimiento, idea (http://www.rae.es/rae.html).

Procedimiento
El o un procedimiento es el modo de ejecutar determinadas acciones que suelen realizarse de la misma forma, con una serie común de pasos claramente definidos, que permiten realizar una ocupación o trabajo correctamente. (http://es.wiktionary.org/wiki/)

Método de ejecutar algunas cosas (http://www.rae.es/rae.html).

Saber
Conjunto de conocimientos, adquiridos mediante el estudio o la experiencia, sobre alguna materia, ciencia o arte. Sinónimo: sabiduría, erudición (http://es.wiktionary.org/wiki/).

Conocer algo, o tener noticia o conocimiento de ello (http://www.rae.es/rae.html)

Sabiduría
Conocimiento de las ciencias y artes (http://es.wiktionary.org/wiki/).

Prudencia en la forma de actuar (http://es.wiktionary.org/wiki/).

Grado más alto del conocimiento; conducta prudente en la vida o en los negocios; conocimiento profundo en ciencias, letras o artes (http://www.rae.es/rae.html).

Técnica
Conjunto de habilidades para aplicar determinados conocimientos (http://es.wiktionary.org/wiki/).

Conjunto de procedimientos y recursos de que se sirve una ciencia o un arte. (http://www.rae.es/rae.html).

Teoría
Conocimiento especulativo considerado con independencia de toda aplicación. (http://www.rae.es/rae.html).

Serie de las leyes que sirven para relacionar determinado orden de fenómenos. (http://www.rae.es/rae.html).

Hipótesis cuyas consecuencias se aplican a toda una ciencia o a parte muy importante de ella (http://www.rae.es/rae.html).

Una teoría es un sistema lógico compuesto de observaciones, axiomas y postulados, así como predicciones y reglas de inferencia que tienen sirven para explicar de manera económica cierto conjunto de datos e incluso hacer predicciones, sobre que hechos serán observables bajo ciertas

condiciones. Las teorías además permiten ser ampliadas a partir de sus propias predicciones, e incluso ser corregidas, mediante ciertas reglas o razonamientos, siendo capaces de explicar otros posibles hechos diferentes de los hechos de partida de la teoría.
(http://es.wikipedia.org/wiki/Teor%C3%ADa)

En ciencia, se llama teoría también a un modelo para el entendimiento de un conjunto de hechos empíricos. En física, el término teoría generalmente significa una infraestructura matemática derivada de un pequeño conjunto de principios básicos capaz de producir predicciones experimentales para una categoría dada de sistemas físicos. Un ejemplo sería la "teoría electromagnética", que es habitualmente tomada como sinónimo del electromagnetismo clásico, cuyos resultados específicos pueden derivarse de las ecuaciones de Maxwell.

Prueba de hipótesis: Error tipo I y II

El nivel de significancia se designa con la letra griega α e indica cuan rara (muy grande o muy pequeña) deber ser la diferencia entre la media maestral y la media poblacional para rechazar la hipótesis nula dado que ésta sea correcta. En trabajos estadísticos es usual rechazar la hipótesis nula si dicha diferencia tiene una probabilidad de ocurrencia por factores aleatorios inferior o igual al 5% ($\alpha < 0.05$). El valor de α es un indicador del riesgo que el investigador(a) está dispuesto a asumir cuando evalúa Ho. Por ejemplo, para un α de 0.05 y aun cuando Ho sea verdadera, se espera que la misma sea rechazada en un 5% de las veces que se ejecute el experimento. O sea, que en 5 de cada 100 experimentos Ho será rechazada aún cuando sea verdadera. A este tipo de error se le denomina *error tipo I*. Por otro lado, el no rechazar Ho cuando en realidad debe rechazarse se denomina *error tipo II*. Es común observar en reportes científicos valores de α entre 0.1 y 0.001; al primero se le denomina diferencia significativa y al último diferencia altamente significativa.

Cuando se rechaza Ho se utiliza el término "significante" o estadísticamente significativo" y el resultado puede interpretarse en el sentido de "haber aprendido algo nuevo sobre la población". Por otro lado, el término "no significativo" expresa el sentir de no haber aportado nuevo conocimiento de la población en estudio. Al realizar una prueba de hipótesis recuerde que muestras muy grandes tienden a rechazar Ho aunque las diferencias entre grupos sean muy pequeñas. De igual manera, pruebas estadísticas con un gran poder tienden a generar mayor número de resultados significativos (pueden detectar diferencias muy pequeñas como estadísticamente significativas).

Cuadro 1: Error tipo I y II.

		Verdad o estado de la realidad	
		Ho (no existe diferencia entre los dos grupos)	H_I (existe diferencia entre los dos grupos)
Decisión	$H_0: \mu_1 = \mu_2$	No se rechaza Ho (Decisión correcta)	No se rechaza Ho (Decisión incorrecta), Error tipo II, β. No se detectan diferencias verdaderas. Este valor indica los falsos negativos o sea la posibilidad no de rechazar Ho cundo en la realidad exista una diferencia)
	$H_1: \mu_1 \neq \mu_2$	Error tipo I, α (este valor indica los falsos positivos; o sea la posibilidad de rechazar Ho sin que en la realidad exista una diferencia)	Se rechaza Ho (Decisión correcta).

Ho (hipótesis nula) H_I (hipótesis alternativa)

Steiger, J.H., & Fouladi, R.T. 1997. Noncentrality interval estimation and the evaluation of statistical models. Pp. 221-257. In Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.) What if there were no significance tests? Mahwah, NJ: Lawrence Erlbaum Associates. Disponible en: http://www.statpower.net/Steiger%20Biblio/Steiger&Fouladi97.PDF

Sugerencias para el análisis de datos

1. Listar variable(s) a analizar y su respectivo nivel de medición (nominal, ordinal, intervalo, razón).
2. ¿Cuál es el historial y contexto de los datos (origen, métodos de colecta, instrumentos utilizados, temporalidad, limitaciones)?
3. ¿Para qué realiza usted el análisis del set de datos? ¿Qué se desea resaltar del set de datos?
4. Describa el producto esperado o solicitado (e.g. descripción del set de datos, prueba de una hipótesis, comparación de datos, ajustar un modelo).
5. Seleccione el software a utilizar (e.g. Instat, XLSTatistics, PASS, otro) y realice un análisis exploratorio de datos.
    a. Análisis gráfico
    b. Elaboración de tablas
    c. Estadísticos descriptivos
6. Busque valores atípicos o extremos que podrían indicar errores en la digitación de los datos y distribuciones asimétricas o inusuales. ¿Concuerda la distribución de los datos con lo que usted esperaba?.
7. Estadística inferencial. Selecciones las pruebas estadísticas a realizar, defina el valor de significancia a utilizar en las pruebas estadísticas.
8. Conclusiones
9. Retroalimentación

Estadística: Software gratuito

Si usted desea explorar otros programas estadísticos gratuitos, le recomiendo visitar los siguientes sitios.

BioEstat. Análisis estadístico para Windows y Mac. Estadística descriptiva e inferencial paramétrica y no paramétrica, análisis de poder. Interfaz en español. Manual en Portugués. http://www.mamiraua.org.br/download/index.php?dirpath=./BioEstat%205%20Portugues&order=0

Instat Análisis estadístico para Windows. Estadística descriptiva e inferencial paramétrica y no paramétrica. Modulo para aplicaciones climáticas.

LazStats Análisis estadístico para Windows. Estadística descriptiva e inferencial paramétrica y no paramétrica. http://statpages.org/miller/LazStats/

MacAnova Análisis estadístico para Macs y Windows. Estadística descriptiva e inferencial paramétrica y no paramétrica, análisis de poder.

Mstat *Windows* *Mac OSX* *Linux* Análisis estadístico para Windows, Mac y Linux. Estadística descriptiva e inferencial paramétrica y no paramétrica.

OpenEpi produce estadísticas para casos y medidas en estudios descriptivos y analíticos, análisis estratificado con límites de confianza exactos, análisis de datos apareados y de personas-tiempo,

tamaño de la muestra y cálculos de potencia, números aleatorios, sensibilidad, especificidad y otras estadísticas de evaluación, tablas FxC, chi-cuadrados para dosis-respuesta, y enlaces a otros sitios de interés. Este software está orientado al análisis de datos epidemiológicos.

OpenSTat Análisis estadístico para Windows. Estadística descriptiva e inferencial paramétrica y no paramétrica. http://statpages.org/miller/openstat/

PAST Análisis estadístico univariado, multivariado, índices de diversidad. Estadística descriptiva e inferencial paramétrica y no paramétrica.

PSPP. Este es un programa para el análisis estadístico, su funcionalidad es similar al programa comercial SPSS http://www.jdmp.org/misc/related-software/

Remuestreo Sofware para análisis estimación y pruebas de hipótesis utilizando remuestreo.

The R Project for Statistical Computing. Gran variedad de análisis, muy poderoso pero requiere de usuarios experimentados. Opera en base a comandos.

WinIDAMS. Este es paquete de software para la validación, tratamiento y análisis estadístico de datos desarrollado por la Secretaría de la UNESCO en cooperación con expertos de varios países. http://portal.unesco.org/ci/en/ev.php-URL_ID=2070&URL_DO=DO_TOPIC&URL_SECTION=201.html

Referencias

Bryan F.J. Manly. Randomization, Bootstrap and Monte Carlo Methods in Biology, Third Edition. Chapman and Hall/CRC. 388p. 2006.

Bryan F.J. Manly. Statistics for Environmental Science and Management, Second Edition.Chapman & Hall/CRC. 292p. 2008.

Cox Nicholas J. Stata Users' Meeting London June 2004. Circular statistics in Stata, revisited. Department of Geography, University of Durham, Durham City, DH1 3LE, UK http://fmwww.bc.edu/repec/usug2004/dir2004_london.pdf

Dinov, Ivo D. (2006). "Statistics Online Computational Resource". Journal of Statistical Software 16 (1): 1–16. http://www.jstatsoft.org/v16/i11.

Dinov, Ivo D.; Sanchez, Juana; Christou, Nicolas (2008). "Pedagogical Utilization and Assessment of the Statistic Online Computational Resource in Introductory Probability and Statistics Courses". Journal of Computers & Education 50 (1 pages=284–300): 284. doi:10.1016/j.compedu.2006.06.003.

Good, Phillip I. Introduction to statistics through resampling methods and Microsoft Office Excel.West Sussex, Reino Unido (INGLATERRA). John Wiley & Sons. 231p. 2005.

McDonald, J.H. 2009. Handbook of Biological Statistics (2nd ed.). Sparky House Publishing, Baltimore, Maryland. Last revised August 18, 2009. http://udel.edu/~mcdonald/statintro.html.

Steel, R.G.D. y Torrie, J.H. Principles and procedures of Statistics: A Biometrical Approach (2 nd Ed.). McGrawHill, New York. 629p. 1980.

Sokal,R. R.; Rohlf, F.J.. Biometry. 3 Ed. New York, WH Freeman. 887p. 1995

Stata Users' Meeting London June 2004. Circular statistics in Stata, revisited. Nicholas J. Cox Department of Geography, University of Durham, Durham City, DH1 3LE, UK. 4p.

Wonnacott, Thomas H. y Ronald J. Wonnacott.. Introductory statistics. Third Edition. New York, John Willey & Sons. 650p. 1977.

¿Qué es XLSTatistics?

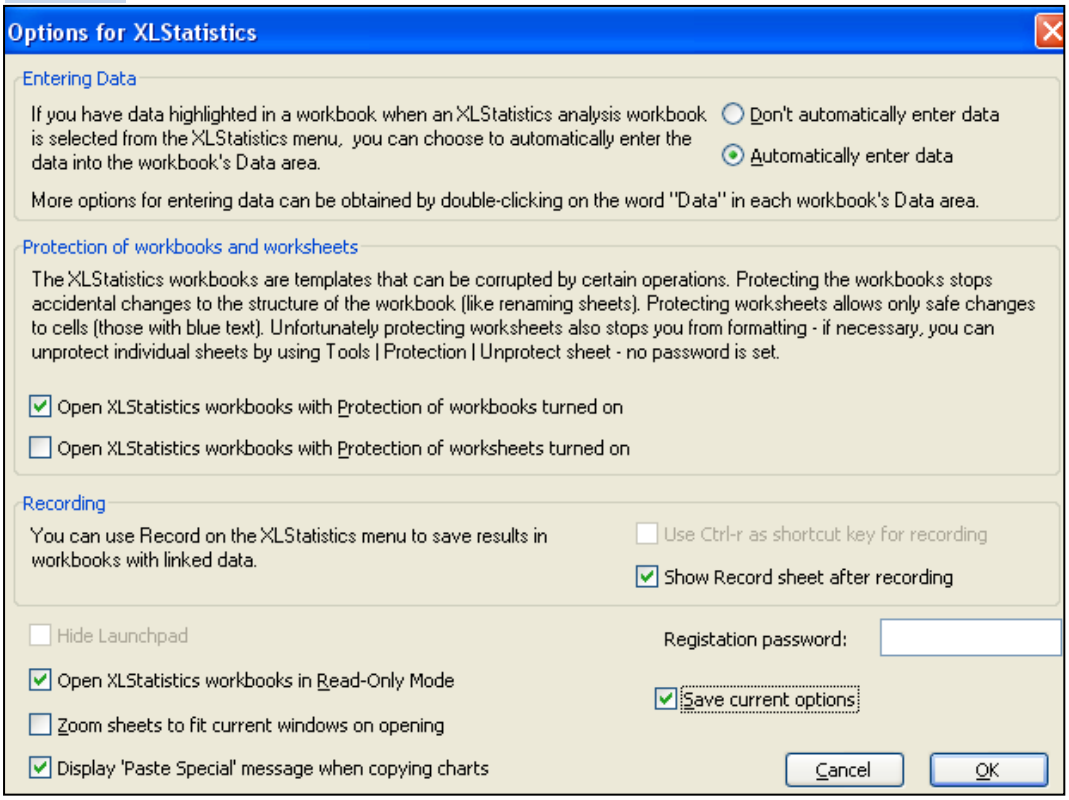
XLSTatistics es un conjunto de 82 libros de Excel creados por el profesor Rodney Carr de la Universidad Deakin, Australia (http://www.deakin.edu.au/~rodneyc/XLStatistics). XLSTatistics opera en Microsoft Excel 97, 2000, 2002, 2003(XP) y 2007.

Citar XLSTatistics de la siguiente manera:

Carr, R., 2009, XLStatistics 09.09.24, XLent Works, Australia. Disponible en: http://www.deakin.edu.au/~rodneyc/XLStatistics. rodneyc@deakin.edu.au

La pestaña de “options” le permite configurar las principales opciones de XLSTatistics.

Options



Automatically enter data: Vincula datos seleccionados en otro archivo de Excel a “datos” del Libro de Análisis de XLStatistics.

“*Open workbooks with protection of workbooks turned on*”: Protección activada.

“*Open XLStatistics workbooks in Read only mode*”: Abrir archivos en modo sólo lectura para evitar que los mismos sean sobrescritos accidentalmente.

“*Display ‘Paste Special’ message when copying charts*” Active esta opción para evitar que los gráficos de Excel sean copiados y pegados en Word como Objetos de Excel.

Instalación

Descargue el archivo **XLS5.zip** (5 Megas). http://www.deakin.edu.au/~rodneyc/XLStatistics/XLS5.zip

Cree un directorio denominado Xlstats y descomprima el contenido del archivo **XLS5.zip.**

Para cargar XLStatistics abra el fichero **XLStatistics.xlam** (Excel 2007) o **XLStats.xls** (Excel 5-2003). Le sugiero crear un acceso directo en su escritorio para dicho archivo.

Guardar sus resultados. ¡No guardar directamente el libro de trabajo de XLStatistics! Si desea copiar los resultados de su análisis a otra hoja de Excel, Word, o Powerpoint simplemente seleccione el área o grafico deseado cópielo y péguelo en el archivo receptor. Si lo desea también puede guardar el libro de trabajo con otro nombre y con la opción de macros habilitada.

Desinstalación

XLStatistics no modifica la configuración del sistema ni la de Excel y por tanto no requiere de desinstalación. Para desactivar XLStatistics haga un clic sobre Quit XLStatistics.

XLStatistics: interfaz grafica y funciones

En Excel 2007 los libros de trabajo se organizan mediante una cinta utilizando como criterio el número y tipo de variable que usted analizará:

1Num: Análisis de una variable numérica (variable cuantitativa discreta ó continua).
1Cat: Análisis de una variable nominal (variable cualitativa).

2Num: Análisis de dos variables numéricas (continuas ó discretas): Correlación y regresión simple.
1Num1Cat: Una variable numérica y otra nominal.
2Cat: Análisis de dos variables nominales (variables cualitativas).

1Num2Cat: Análisis de una variable numérica y dos variables nominales (variables cualitativas).
2Num1Cat: Análisis de dos variables numéricas y una variable nominal (variables cualitativas).

nNum: Análisis de tres o más variables numéricas.
nNum1Cat: Análisis de tres o más variables numéricas y una variable nominal (variables cualitativas).
nCat: Análisis de tres o más variables nominales (variables cualitativas).

A continuación se lista, para cada libro de trabajo, el análisis estadístico que usted puede realizar.

Variable	Análisis estadístico
Una variable numérica (datos de una muestra simple) **1Num** La variable puede ser cuantitativa continuo ó discreta.	Síntesis numérica y gráfica • Estadística descriptiva (tendencia central, variabilidad y forma). • Histograma, polígono de frecuencia simple y acumulada, diagrama de cajas o de Box-Whisker, gráfico de media aritmética y desviación estándar, error estándar o intervalo de confianza. Pruebas estadísticas • Prueba t para μ□ (media) • Intervalo de confianza para μ□ (media) • Prueba de signos e intervalo de confianza para la mediana • Prueba de Chi^2 para la varianza • Gráfico de residuos • Análisis de poder y tamaño de muestra • Prueba de normalidad
Una variable Cualitativa **1Cat**	Síntesis numérica y gráfica • Tabla de frecuencia y proporciones, diagramas de barras, gráfico de pastel o circular., Histograma con barras de error para proporciones. Pruebas estadísticas • Pruebas de proporciones para muestras grandes y pequeñas • Intervalo de confianza para proporciones • Análisis de poder y tamaño de muestra • Prueba de bondad de ajuste/Prueba de corridas

Dos variables numéricas **2Num**	Estadísticos descriptivos por variable, gráfico de dispersión Análisis de correlación y regresión lineal, con o sin término constante • Estimación de parámetros del modelo (a y b) • Intervalos de confianza para “a” y “b”, prueba de hipótesis sobre la significancia de los a los parámetros • Gráfico de dispersión con línea de regresión • Análisis de varianza • Análisis de residuos • Predicción y predicción inversa • Bandas de predicción y de estimación Ajuste de una función definida por el usuario(a) • Funciones linealizables (método mínimos cuadrados) • Ecuaciones polinomiales • Regresión no lineal (método mínimos cuadrados) • Regresión lineal por mínimos cuadrados para detectar un punto de inflexión en la tendencia de la serie Ajuste de una curva de suavizado a los datos • Media móvil (media o mediana) • Medias para datos agrupados con barras / bandas de error • Regresión lineal localmente ponderada - LOWESS Análisis de correlación entre variables ordinales • rho de Spearman • tau-b de Kendall Añadir etiquetas a los puntos en un gráfico de dispersión Análisis de series de tiempo • Media móvil • Ajuste ecuación exponencial

Una variable numérica, una variable cualitativa **1Num1Cat**	Síntesis numérica y gráfica • Estadísticos descriptivos (tendencia central, variabilidad, forma) • Gráfico de de frecuencia (absoluta, relativa, proporciones), histogramas, líneas, barras • Gráfico de de frecuencia por categoría • Gráficos de caja (Box-y-Whisker) • Gráficos de medias con barras de error Pruebas estadísticas • Análisis de la varianza simple o de una vía (efectos aleatorios y fijos) • Prueba de intercepto del modelo • Prueba de Kruskal-Wallis • Análisis de residuos • Prueba de Hartley (comparación de varianzas) • Prueba t para dos muestras e intervalo de confianza para la diferencia entre medias • Prueba de Mann-Whitney • Prueba F de igualdad de varianzas • Análisis de la poder y tamaño de muestra • Prueba de aleatorización para la media de dos grupos
Dos variables cualitativas **2Cat**	Cuadros sinópticos • Frecuencia (absoluta, relativa), proporciones, % por fila o columna, total • Gráficos de proporciones y porciento con barras de error (error estándar, Intervalo de Confianza) Pruebas estadísticas • Prueba de Chi^2 (independencia entre grupos) • Prueba sobre diferencia entre dos proporciones para muestras pequeñas • Prueba sobre diferencia entre dos proporciones para muestras pequeñas y análisis de poder • Análisis de tablas 2x2 (prueba exacta de Fisher)
Una variable numérica, dos variables cualitativas **1Num2Cat**	Cuadros sinópticos • Estadísticos descriptivos (tendencia central, variabilidad, forma), frecuencias (absoluta, relativa) • Gráficos de barras e histogramas • Gráficos de medias con barras o bandas de error Pruebas estadísticas • Análisis del varianza de dos vías (diseño balanceado) para efectos fijos y aleatorios, interacción • Análisis de residuos

Dos variables numéricas, una variable cualitativa **2Num1Cat**	Gráficos de dispersión multiseries con ejes y/o dirección permutable. Gráficos multilínea / multieje Regresión lineal (análisis de covarianza) • Varias opciones de análisis (con o sin término constante) • Diagrama de dispersión con líneas de regresión • Gráficos de pendientes e interceptos con barras de error • Pruebas de hipótesis Ajuste de una función definida por el usuario(a) (grupo por grupo) • Funciones linealizables con regresión lineal por mínimos cuadrados • Ecuaciones polinomiales • Regresión no lineal por mínimos cuadrados • Regresión lineal por mínimos cuadrados para detectar un punto de inflexión en la tendencia de la serie Ajuste de curva de suavizado a los datos (grupo por grupo) • Media móvil (media o mediana) • Medias de grupo de datos con barras / bandas de error • Regresión localmente ponderada – LOWESS

n Variables numéricas (regresión múltiple) **nNum**	Síntesis numérico y gráfico por variable
	• Histogramas y diagramas combinados por frecuencia
	• Gráficos múltiples de cajas (Boxplots)
	• Gráficos de medias
	• Estadística descriptiva
	
	Correlación y regresión simple (2 variables)
	• Diagrama de dispersión y matriz de correlación
	• Regresión lineal de dos variables, estadísticos y pruebas de hipótesis
	• Gráficos multilíneas y multiejes
	
	Regresión múltiple
	• Ecuación de regresión, estadísticos y pruebas de hipótesis
	• Predicción
	• Análisis de residuos
	  • Gráficos de residuos
	  • Prueba de Jarque-Bera para normalidad
	  • Prueba de Ramsey para evaluar omisión de término en el modelo
	  • Prueba de Durbin-Watson para correlación de residuos Correcciones para la correlación serial de orden uno (Cochran-Orcutt)
	  • Prueba de Glejser para heterocedasticidad (varianzas diferentes)
	• Análisis de varianza y comparación de modelos
	
	Análisis de mediciones repetidas
	• Gráfico de caso-por-caso
	• Gráficos de medias
	• Gráficos múltiples (boxplots)
	• Análisis de varianza

n Variables numéricas y 1 variable cualitativas **nNum1Cat**	Estadísticos descriptivos y análisis gráfico categoría-por-categoría Gráficos múltiples de las medias y error
n Variables cualitativas **nCat**	Tablas dinámicas y gráficos por categoría • Frecuencia absoluta • Porcentaje del total, porcentaje por columna ó por fila • Tabla de frecuencia (absoluta, proporciones, porcentajes) para un máximo de ocho variables • Gráfico de media para proporciones, porcentajes con barra de error (desviación estándar, intervalo de confianza)
1 Numérica y n variables cualitativas **1NumnCat**	Tablas dinámicas con síntesis adecuadas a la variable numérica por categoría • Frecuencia absoluta • Mínimo, máximo, media, desviación estándar

La interfaz grafica de cada libro de Excel es prácticamente estándar, lo cual facilita su uso. Al interior de cada libro usted puede utilizar las funciones y herramientas de Excel (e.g. fuentes: tamaño, color, itálico, negrita; colores en gráficos; copiar y pegar tablas y gráficos a otros programas de Office; guardar libros de Excel con macros habilitados.

El (la) usuario(a) solo tiene que digitar, copiar y/o pegar los datos en las columnas de 'Data' de un libro de trabajo y el resto se hace automáticamente: Síntesis numérica y gráfica; análisis estadístico (e.g. pruebas de hipótesis, intervalos de confianza, pruebas de poder, determinación de tamaño de muestra, ajustar modelos de regresión, etc.). Su tarea fundamental es decidir ¿qué prueba debo realizar? o ¿cómo expreso gráficamente los datos?.

Nota: Usted **SOLO** debe modificar el contenido de las celdas con números o textos en color **azul**.

Otros complementos gratuitos para análisis estadístico en Excel

Regresión y diseño de experimentos: otro complemento gratuito para Excel

"Essential Regression and Experimental Design for Chemists and Engineers" es un complemento gratuito para Excel que permite ajustar modelos de regresión simple, múltiple (selección hacia adelante y hacia atrás) y polinomiales, así como diseñar plantillas para experimentos. (http://www.jowerner.homepage.t-online.de/ERPref.html).

Regresión localmente ponderada (LOESS) http://peltiertech.com/WordPress/loess-utility-for-excel/ http://peltiertech.com/WordPress/loess-utility-awesome-update/

Complementos comerciales para Excel

Analizar-se ®

Este es un complemento para Excel que permite realizar análisis exploratorio de datos, calculo de estadísticos descriptivos, ajustar modelos de regresión (simples y múltiples, polinomiales) y realizar pruebas de hipótesis paramétricas y no paramétricas (http://www.analyse-it.com).

XLSTATPro ®

XLSTATPro es un conjunto de libros de trabajo que permiten automatizar una gama muy diversa de análisis y pruebas estadísticas paramétricas y no paramétricas. XLSTAT funciona con diversas versiones de Excel, desde 97 hasta 2007 para Windows y Mac (http://www.xlstat.com).

Remuestreo con Excel (Resampling Stats versión 4)

Remuestreo es un complemento para Excel que implementa los siguientes métodos de remuestreo: bootstrap, permutaciones y los procedimientos de simulación de datos en Excel. La última versión (4.0), ofrece una variedad de opciones para BCA Bootstrap, remuestreo estratificado, iteración de una función personalizada, la capacidad de ejecutar hasta 1.000.000 de iteraciones para cientos de celdas (Excel 2007), muestreo estratificado e histogramas entre otras funciones (http://www.resample.com/content/software/excel/index.shtml). El manual del programa puede descargarse de:

http://www.resample.com/content/software/excel/userguide/RSXLHelp.pdf

XLMiner

XLMiner es un conjunto de herramientas para minería de datos conformado tanto por métodos estadísticos como de aprendizaje automatizado (disciplina científica que diseña y desarrolla algoritmos que le permiten a las computadoras discernir comportamientos basados en datos empíricos). El programa parte de la premisa que un mismo set de datos debe analizarse utilizando diferentes enfoques para luego elegir el modelo que mejor se adapte a los datos. Entre las funciones del programa están: partición del set de datos, diversos métodos de clasificación y predicción, análisis de afinidad, análisis de series de tiempo y exploración y reducción del set de datos. Para mayores detalles sobre la funcionalidad del programa visitar:
http://www.resample.com/xlminer/capabilities.html

Programas gratuitos y en línea para análisis estadístico

- Análisis de regresión simple y múltiple, logística, partición recursiva (clasificación y arboles de decisión)
- Distribuciones estadísticas
- Pruebas de hipótesis (medias, error tipo II y tamaño de muestra)
- Análisis de series de tiempo

STATGRAPHICS Online http://www.statgraphicsonline.com/index.html

Usted puede digitar los datos, leer datos de varios tipos de archivos (incluyendo Excel), copiar y pegar los conjuntos de datos de otras aplicaciones, o analizar los set de datos de StatPoint. El acceso en línea es gratuito para sets de datos de hasta 100 filas (observaciones) por 10 columnas (variables). Para utilizar el programa debe registrarse (es gratuito).

VassarStats: Website for Statistical Computation http://faculty.vassar.edu/lowry/VassarStats.html

• **Clinical Research Calculators** (Calculators 1-3. For prevalence, sensitivity, specificity, predictive values, likelihood ratios, etc.,Kaplan-Meier Survival Probability Estimates,Kappa as a Measure of Concordance in Categorical Sorting,Chi-Square, Cramer's V, and Lambda for a Rows by Columns Contingency Table,McNemar's Test for Correlated Proportions in the Marginals of a 2x2 Contingency Table, Simple Logistic Regression [the plain-vanilla version],Simple ROC Curve Analysis)

•**Probabilities** (Randomness and the Appearence of Pattern [Demo], For Sequential Sampling: Pascal (Negative Binomial) Probabilities,Backward Probability Template, Bayes' Theorem: Conditional Probabilities,Bayes' Theorem: Adjustment of Subjective Confidence,[See also: Clinical Research Calculators.)

•**Distributions** (Sampling Distribution Generators - Binomial Distributions, Poisson Distributions, Chi-Square Distributions, t-Distributions, Distributions of the Pearson Product-Moment Correlation Coefficient, Normal Distributions, Distributions of Sample Means, Distributions of Sample Mean Differences-, Central Limit Theorem [Text & Demo],z to P Calculator, Standard Error of Sample Means, Standard Error of the Difference Between the Means of Two Samples.

•**Frequency Data** (Exact Binomial Probability Calculator, Binomial z-Ratio Calculator, Poisson Approximation of Binomial Probabilities, Fitting an Observed Frequency Distribution to the Closest Poisson Distribution, For Sequential Sampling: Pascal (Negative Binomial) Probabilities, Chi-Square "Goodness of Fit" Test, Kolmogorov-Smirnov One-Sample Test, Fisher Exact Probability Test, Phi Coefficient of Association, Rates, Risk Ratio, Odds, Odds Ratio, Log Odds, 2x2Chi-Square,

McNemar's Test for Correlated Proportions in the Marginals of a 2x2 Contingency Table, Fisher Exact Probability Test for Tables Larger than 2x2, Chi-Square, Cramer's V, and Lambda for a Rows by Columns Contingency Table, Log-Linear Analysis for a 3-Way Contingency Table, Kappa as a Measure of Concordance in Categorical Sorting

•**Proportions** (The Confidence Interval of a Proportion, The Confidence Interval for the Difference Between Two Independent Proportions, Significance of the Difference Between Two Independent Proportions, McNemar's Test for Correlated Proportions in the Marginals of a 2x2 Contingency Table)

•**Ordinal Data** (Rank-Order Correlation, Mann-Whitney Test,Wilcoxon Signed-Ranks Test, Kruskal-Wallis Test, Friedman Test)

•**Correlation & Regression** (Basic Linear Correlation & Regression, Matrix of Intercorrelations, Multiple Regression,0.95 and 0.99 Confidence Intervals for r, Estimating the Population Value of rho on the Basis of Several Observed Sample Values of r, Test for the Heterogeneity of Several Values of r,The Significance of an Observed Value of r, Significance of the Difference Between Two Independent Values of r, Significance of the Difference Between an Observed Value of r and a Hypothetical Value of rho, First- and Second-Order Partial Correlations, Phi Coefficient of Association, Point Biserial Coefficient, Correlation for Unordered Pairs: Eta2, Intraclass Correlation, & Resampling of r, Simple Logistic Regression)

•**t-Tests & Procedures** (t-Tests for the Significance of the Difference Between the Means of Two Samples (independent or correlated), Single Sample t-Test,0.95 Confidence Interval for the Estimated Mean of a Population)

•**ANOVA** (One-Way ANOVA for Independent or Correlated Samples, Two-Way Factorial ANOVA for Independent Samples, Two-Factor ANOVA with Repeated Measures on One Factor, Two-Factor ANOVA with Repeated Measures on Both Factors,2x2x2 ANOVA for Independent Samples, Orthogonal Latin Square Designs for $n=j^2$)

•**ANCOVA** (One-Way ANCOVA for Independent Samples, Two-Way Factorial ANCOVA for Independent Samples)

•**Miscellanea** (Basic Sample Stats, Resampling Probability Estimates for the Difference Between the Means of Two Independent Samples,The Power of the Chi-Square "Goodness of Fit" Test [Text & Demo]

Free Statistics and Forecasting Software http://www.wessa.net/
Descriptive Statistics Software, Regression Software, Statistical Hypothesis Testing Software, Time Series Analysis (R modules), Multiple Regression Software (C module), Scientific Forecasting Software (C modules), Sample Size Software, Factor Analysis, Clustering Software.

Sitios de interés

http://www.statsoft.com/textbook/nonparametric-statistics/

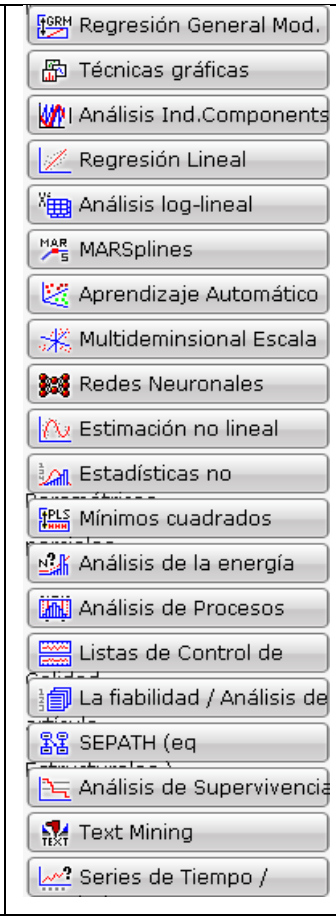
Según la compañía StatSoft este es el único recurso de Internet sobre Estadística recomendado por la Enciclopedia Británica. El libro trata, de manera concisa, la mayoría de las técnicas de análisis estadístico, incluye un glosario de términos estadísticos así como una lista de referencias.

Citar como:
(Versión electrónica): StatSoft, Inc. (2010). Libro de texto electrónico de Estadística. Tulsa, OK: StatSoft. WEB: http://www.statsoft.com/textbook/.

(Versión impresa): Hill, T. & Lewicki, P. (2007). ESTADÍSTICAS Métodos y Aplicaciones. StatSoft, Tulsa, OK

En este sitio usted puede encontrar otros programas en línea para el análisis de datos
http://www.psychnet-uk.com/experimental_design/online_calculators.htm

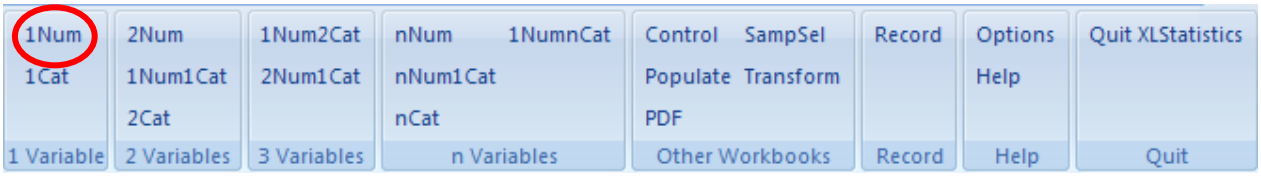
Descripción y análisis grafico de datos



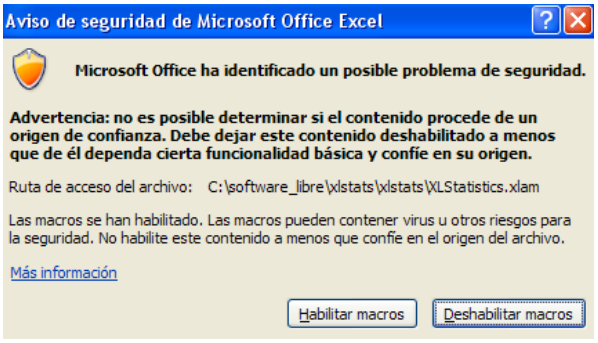
Temas tratados en el libro digital.
http://www.statsoft.com/textbook/

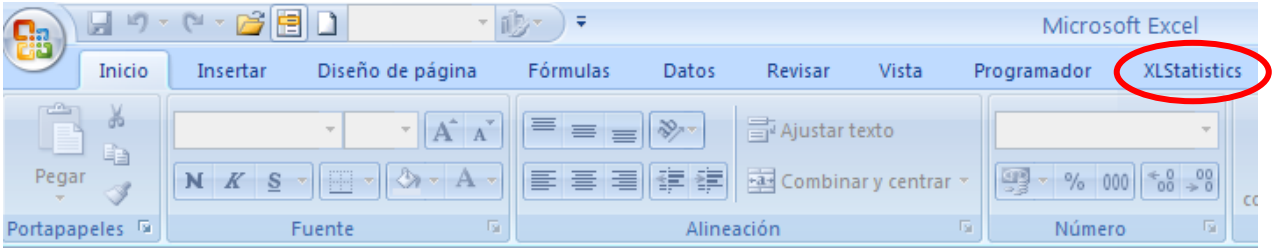
Análisis de una variable numérica (variable cuantitativa discreta ó continua)

En la presente sección se ilustra el uso de XLSTatistics para el análisis de una variable numérica. Algunos ejemplos de variables cuantitativas son: peso, volumen, precipitación, temperatura, biomasa, conteos (e.g. numero de arboles en una parcela, numero de animales por aparto). En el menú grafico de XLStatisitics corresponde a **1Num**.



Para iniciar la sesión haga un doble clic sobre el archivo XLStatistics.xlam si usted utiliza Excel 2007 o sobre XLStats.xls si usted utiliza Excel 2000 y habilite los MACROS.





Una vez activados los macros, observará una nueva pestaña “XLStatitics”

1. Abra el archivo “xlstats_tutorial.xlsx” y seleccione de la hoja de cálculo **1Num** la variable precipitación anual de la estación Moravia de Chirripó (1200 msnm).
2. Haga un clic sobre XLStatistics y luego otro clic sobre 

. Observe que el programa abre el libro de trabajo **1Num.xls** y copia los datos seleccionados a dicho libro.

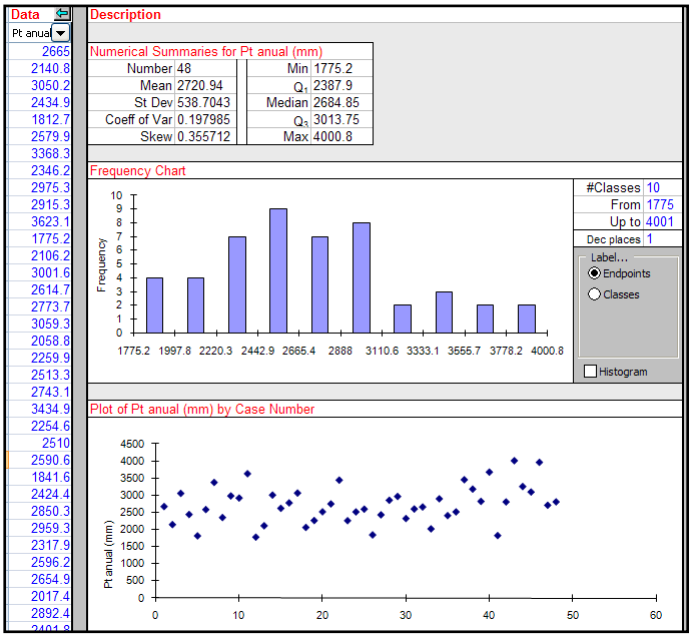
En la sección inferior del libro de de trabajo **1Num.xls** usted observará cinco hojas de cálculo:

1) *Information*: Información. Esta hoja contiene información general sobre la organización de este libro de trabajo.
2) *Data and Description*: Datos y estadísticos descriptivos
3) *Summaries*: Síntesis de datos (Tabla de frecuencia y gráficos)

4) *Tests*: Pruebas estadísticas
5) *Extra Tools*: Herramientas adicionales

Nota: Usted **SOLO** debe modificar el contenido de las celdas con números o textos en color azul.

Data and Description: Datos y su descripción

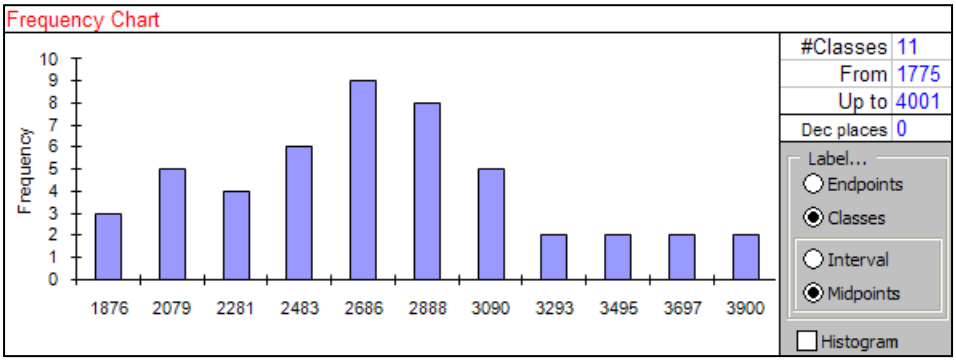


Data (Datos): Variable que se analiza, en este caso precipitación anual (mm)

Description (Describir):
Estadísticos descriptivos (número de observaciones, media, desviación estándar, Coef. Variación, asimetría, mínimo, primer cuartil, mediana, tercer cuartil, máximo).

Grafico de frecuencia (barras ó histograma). Usted puede elegir número de clases, límite inferior y superior de la serie estadística.

Grafico de datos individuales.



¿Cuál es la precipitación media de la serie?_____

¿Cuál es la precipitación mediana de la serie?_____

¿Cuál es el coeficiente de variación de la serie?_____

¿Cuántas clases debe seleccionar para un intervalo de clase de 200 mm?_____

¿Observa alguna tendencia en la gráfica de datos individuales?

Nota: si desea observar los valores graficados por el programa, seleccione con el puntero del ratón la gráfica y observe a la derecha de la hoja de cálculo las celdas seleccionadas. Ahora seleccione dichas celdas y asígneles un color al texto (color de la fuente)  .

***Summaries*: Síntesis de datos (Tabla de frecuencia y gráficos)**

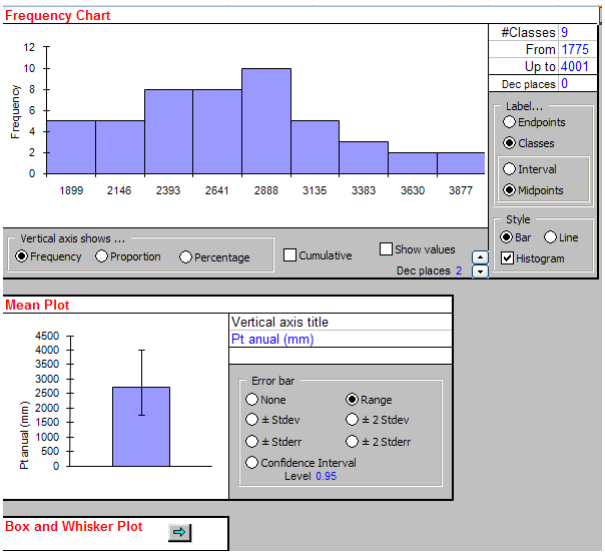


Gráfico de frecuencia (barras, histograma ó línea). Usted puede elegir límite inferior y superior de la serie estadística, decimales, número de clases, título del eje X (punto final de clase, punto medio de clase, intervalo). También puede crear una OJIVA (Gráfico de frecuencia acumulada).

Gráfico de media aritmética: Si lo desea puede modificar el título del eje vertical. Puede adicionar al gráfico de barra el rango, desviación estándar, error estándar e intervalo de confianza. Recuerde que usted puede elegir el nivel de confianza.

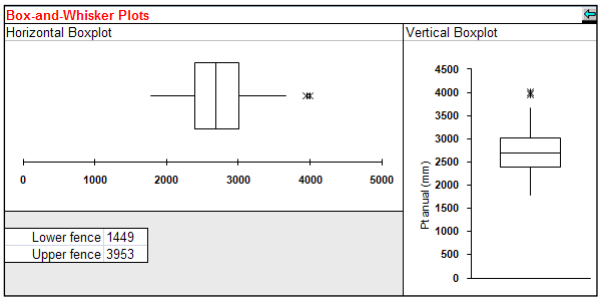
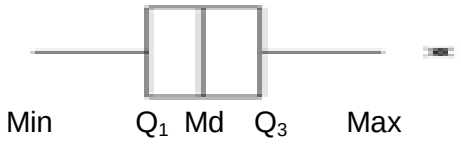


Gráfico de cajas (Box-Whisker). Usted puede elegir el gráfico vertical o horizontal. Si lo desea puede adicionar un título al eje X de la gráfica.

Haga un clic sobre la flecha azul para regresar a la hoja de cálculo anterior.

Observe que el gráfico muestra un valor atípico o extremo para la serie, el cual corresponde al año 2002.

El gráfico de Box-Whisker muestra los siguientes valores de la serie estadística:



El "*" indica un valor extremo, atípico o no esperado para la serie estadística y corresponden a datos menores que Q_1-1.5*(Q_3 - Q_1) o mayores que Q_3+1.5*(Q_3 - Q_1). Estos valores deben analizarse cuidadosamente ya que pueden representar errores de transcripción o condiciones particulares de dicho dato (e.g. formar parte de otra población estadística). Por ejemplo, si usted analiza datos de

lluvia y en un año determinado se presenta un huracán, la lluvia de dicho año debe considerarse como un evento de otra población estadística.

Tests: Pruebas estadísticas

Prueba "t" para una muestra (requisitos muestra independiente y normalidad)

Tests on the Mean (μ) (t-tests)

Sample Data	
Sample Size	48
Mean	2727.379
Standard Deviation	531.4104
SE Mean	76.70248

Hypothesis Tests	
H_0: μ =	3000
Alternative	≠ > <
H_1: μ ≠	3000
T	-3.5543
DF	47
p-value =	0.00088

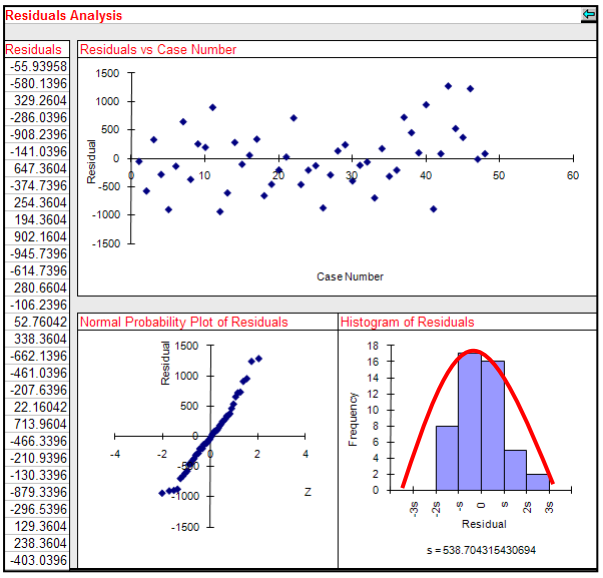
Confidence Intervals for μ		
Type (2,U,L)	2	
Confidence Level	0.95	
ME	Lower	Upper
154.3055	2573.074	2881.685

Pruebas estadísticas sobre la media de una población (Prueba t de Estudiante). Puede realizar pruebas de una y dos colas (superior e inferior). Recuerde que usted debe fijar el nivel de alfa antes de realizar la prueba.

El ejemplo ilustra la siguiente prueba de hipótesis:
Ho: μ = 3000m
Ha: μ ≠ 3000m
Para un alfa de 0.05, Ho debe rechazarse y por tanto la media es diferente de 3000 mm.

Calculo de Intervalo de Confianza para la media para un nivel de confianza de 95%. Observe que el valor 3000 mm se encuentra fuera del IC.

Análisis de Residuos



Análisis de residuos
El análisis de residuos permite probar por el supuesto de normalidad del set de datos.

Para datos normales, los residuos deben ajustarse a una recta y el histograma debe mostrar una forma de campana.

El término "error" o residuo se refiere a la diferencia entre cada valor y la media del grupo. Los resultados de una prueba "t" son validos cuando la dispersión de los residuos es al azar o sea el factor causante de que un valor sea muy grande o muy pequeño afecta sólo a dicho valor.

Observe que a partir de la observación 23 (1995) los residuos tienden a ser positivos.

Ejercicio: Utilizando las viables Pt y Año, realice un análisis de tendencia en tiempo (Pt Vs año) como el que se muestra en "Cambio de pendiente (tendencia) en el set de datos" para determinar si efectivamente existe un cambio en la pendiente para dicho año.

Nota: si desea observar los valores graficados por el programa, seleccione con el puntero del ratón la gráfica y observe a la derecha de la hoja de cálculo las celdas seleccionadas. Ahora seleccione dichas celdas y asígneles un color al texto (color de la fuente)  . A continuación se muestran los datos para la gráfica de probabilidad normal para los residuos y para el histograma de residuos.

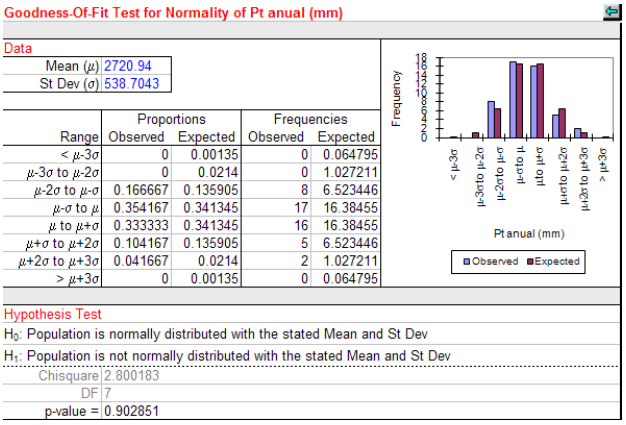
ResNA	ResSort	Z	ZNA			
-62.3792	-952.179	-2.04539	-2.04539	-1594.23	-3s	0
-586.579	-914.679	-1.74129	-1.74129	-1062.82	-2s	0
322.821	-885.779	-1.54458	-1.54458	-531.41	-s	8
-292.479	-709.979	-1.39417	-1.39417	0	0	17
-914.679	-697.379	-1.27001	-1.27001	531.41	s	16
-147.479	-668.579	-1.16283	-1.16283	1062.82	2s	5
640.921	-621.179	-1.06757	-1.06757	1594.23	3s	2
-381.179	-586.579	-0.98113	-0.98113			0
247.921	-472.779	-0.90145	-0.90145	2727.38		48
187.921	-467.479	-0.82713	-0.82713	48		
895.721	-409.479	-0.75712	-0.75712	s = 531.410363914637		
-952.179	-381.179	-0.69063	-0.69063	1NRP.xls		
-621.179	-325.579	-0.62707	-0.62707	FALSO		
274.221	-302.979	-0.56595	-0.56595	Batan.xlsx		
-112.679	-292.479	-0.50687	-0.50687			
46.3208	-216.679	-0.44951	-0.44951			
331.921	-214.079	-0.3936	-0.3936			
-668.579	-147.479	-0.33889	-0.33889			
-467.479	-136.779	-0.28517	-0.28517			
-214.079	-131.179	-0.23227	-0.23227			
15.7208	-117.379	-0.18001	-0.18001			
707.521	-112.679	-0.12824	-0.12824			
-472.779	-72.4792	-0.07681	-0.07681			
-697.379	-62.3792	-0.02558	-0.02558			
-136.779	-22.6792	0.02558	0.02558			
-885.779	15.7208	0.07681	0.07681			
-302.979	46.3208	0.12824	0.12824			
122.921	73.4208	0.18001	0.18001			
231.921	74.9208	0.23227	0.23227			
-409.479	89.8208	0.28517	0.28517			
-131.179	122.921	0.33889	0.33889			
-72.4792	165.021	0.3936	0.3936			
-709.979	187.921	0.44951	0.44951			
165.021	231.921	0.50687	0.50687			

ResNA: Residuos no ordenadas (Pt mm - Media de Pt mm)
ResSort Residuos ordenadas (Pt mm - Media de Pt mm) de menor a mayor
Z: Residuos normalizado (valor de Z para residuo ajustado a una distribución normal)

Limite superior	Desv. Estándar	Frecuencia
-1594.23	-3s	0
-1062.82	-2s	0
-531.41	-s	8
0	0	17
531.4104	s	16
1062.821	2s	5
1594.231	3s	2

Datos utilizados en el histograma de residuos.

Prueba de Normalidad



Goodness-Of-Fit Test for Normality of Pt anual (mm)

Data

Mean (μ)	2720.94
St Dev (σ)	538.7043

Range	Proportions Observed	Proportions Expected	Frequencies Observed	Frequencies Expected
< μ-3σ	0	0.00135	0	0.064795
μ-3σ to μ-2σ	0	0.0214	0	1.027211
μ-2σ to μ-σ	0.166667	0.135905	8	6.523446
μ-σ to μ	0.354167	0.341345	17	16.38455
μ to μ+σ	0.333333	0.341345	16	16.38455
μ+σ to μ+2σ	0.104167	0.135905	5	6.523446
μ+2σ to μ+3σ	0.041667	0.0214	2	1.027211
> μ+3σ	0	0.00135	0	0.064795

Hypothesis Test

H_0: Population is normally distributed with the stated Mean and St Dev

H_1: Population is not normally distributed with the stated Mean and St Dev

Chisquare 2.800183
DF 7
p-value = 0.902851

Esta prueba se utiliza para determinar si los datos de la muestra provienen de una distribución normal.

La hipótesis nula es la que la muestra proviene de una población normal, la hipótesis alternativa es que la muestra proviene de una población no normal.

El valor de p (nivel de significancia o alfa) permite evaluar Ho. En este caso y para un alfa de 0.05, no se rechaza Ho porque P= 0.903.

Recuerde que usted debe fijar el nivel de alfa antes de realizar cualquier prueba estadística.

Análisis de poder y determinación de tamaño de muestra

Al realizar cualquier prueba de hipótesis, su decisión está sujeta a dos errores estadísticos: el error tipo I y el error tipo II. A continuación se describe cada uno de ellos.

Tabla de decisión estadística para una prueba de hipótesis sobre dos grupos.

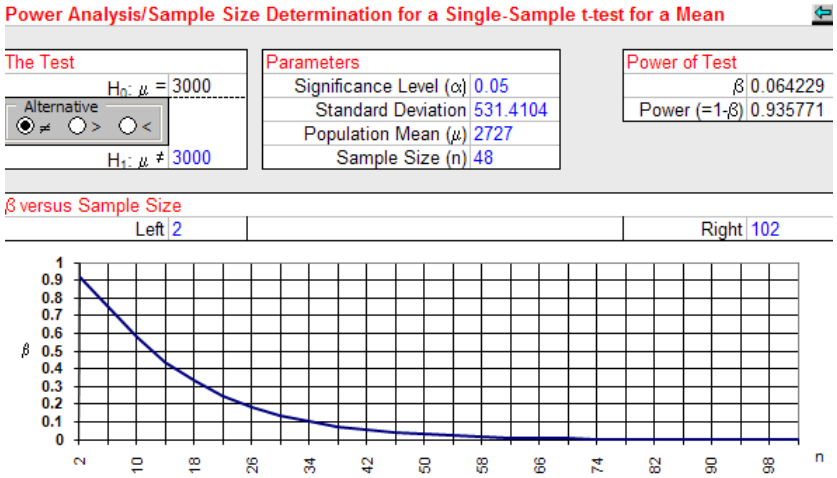
	Hipótesis	Verdad o estado de la realidad	
		Ho (no existe diferencia entre los dos grupos)	H_I (existe diferencia entre los dos grupos)
Decisión	H_0: $\mu_1 = \mu_2$	No se rechaza Ho (Decisión correcta)	No se rechaza Ho (Decisión incorrecta) y por tanto no se detectan diferencias verdaderas. A este error se le denomina “error tipo II” y se representa con la letra β. Este valor indica los falsos negativos o sea la posibilidad no de rechazar Ho cundo en la realidad exista una diferencia.
	H_1: $\mu_1 \neq \mu_2$	Bajo esta condición el error tipo I es igual a α (este valor indica los falsos positivos; o sea la posibilidad de rechazar Ho sin que en la realidad exista una diferencia).	Se rechaza Ho (Decisión correcta).
El **poder** de una prueba estadística es igual a **1- β**.			

Ho (hipótesis nula)
H_I (hipótesis alternativa)

Nota: Muestras muy grandes tienden a rechazar Ho aunque las diferencias entre los grupos sean muy pequeñas. Pruebas estadísticas con un gran poder tienden a generar mayor número de resultados

significativos (pueden detectar diferencias muy pequeñas como estadísticamente significativas). Recuerde que una diferencia estadísticamente significativa no necesariamente es una diferencia importante para la toma de decisiones.

Steiger, J.H., & Fouladi, R.T. 1997. Noncentrality interval estimation and the evaluation of statistical models. Pp. 221-257. In Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). What if there were no significance tests? Mahwah, NJ: Lawrence Erlbaum Associates. Disponible en: http://www.statpower.net/Steiger%20Biblio/Steiger&Fouladi97.PDF



Esta hoja de trabajo le permite realizar un análisis de poder de la prueba" t" para una media obtenida de una muestra independiente.

Recuerde que usted puede modificar los números en color azul.

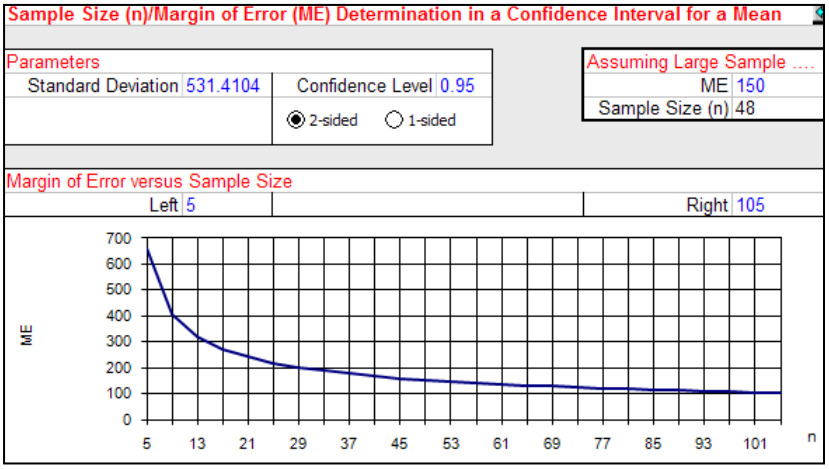
¿Cuál es el error tipo II de la prueba?

¿Es el poder de la prueba alto o bajo?

¿Cuál es la implicación práctica del análisis de poder?

¿Cuál es el tamaño de muestra requerido para un poder de 0.010?

Análisis de error marginal para IC de la media



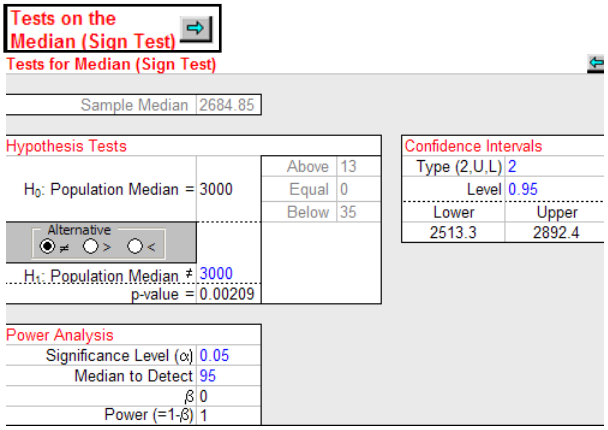
Esta hoja de trabajo le permite realizar un análisis de error marginal para el intervalo de confianza para media obtenida de una muestra independiente.

Recuerde que usted puede modificar los números en color azul.
¿Cuál sería el tamaño de muestra requerido para un error marginal de 150 mm?

Pruebas no paramétricas

Mediana (Prueba de signos, prueba del signo)
Prueba de Wilcoxon para muestras pareadas
Prueba de Chi-2 para la varianza

Prueba de signos (prueba de mediana)



Prueba sobre la Mediana (prueba de signos). Esta prueba no paramétrica utiliza la distribución binomial para determinar si el número de datos sobre y bajo la mediana es el mismo.

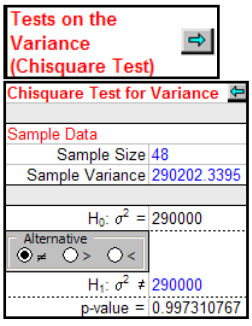
Describa la prueba de hipótesis.

¿Cuál es el poder de la prueba?

Wilcoxon Signed Rank Test

La prueba de Wilcoxon de los rangos con signo permite comparar nuestros datos con una mediana teórica (por ejemplo un valor publicado en un artículo). También se utiliza en muestras pareadas ó para mediciones repetidas en el mismo sujeto o muestra. Esta prueba es el análogo no paramétrico de la "t" de Estudiante y no requiere que la diferencia entre los pares de observaciones sigan una distribución normal.

Prueba de Chi-2 para la varianza



Prueba de varianza (Ji-cuadrado). Usted puede comparar la varianza observada con un valor predeterminado (e.g. Usted puede comparar el valor de una muestra con un valor de la población).

¿Dado el valor de "p" y para un alfa de 0.05 se rechaza Ho?_____

Recuerde que usted puede modificar los números en color azul.

Intervalos de tolerancia y de predicción

Tolerance Interval
Prediction Interval

Tolerance and Prediction Intervals	
Sample Data	
Sample Size	48
Mean	2720.939583
Standard Deviation	538.7043154
Tolerance Interval	
Type (2,U,L)	2
Confidence Level	0.95
Coverage	0.99
Lower	Upper
1324.172279	4117.706888
Prediction Interval	
Type (2,U,L)	2
Confidence Level	0.95
Lower	Upper
1625.975609	3815.903558

Datos

Intervalo de tolerancia para una confianza de 95% y una cobertura de 99%.

Intervalo de predicción para una confianza de 95%.

Recuerde que usted puede modificar los números en color azul.

Extra Tools **Herramientas adicionales**

Grouped Data **Datos agrupados 1NumGD.xls**

Data & Description | Summaries | Tests

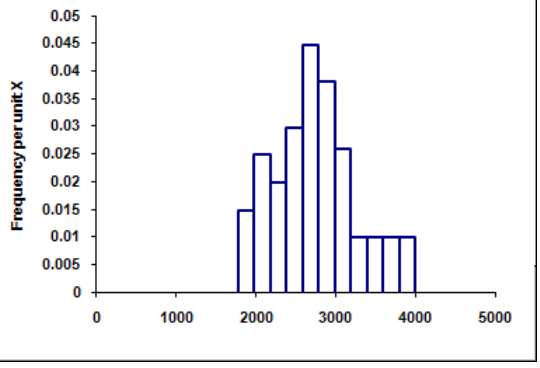
Este libro de cálculo (1NumGD.xls) le permite analizar datos agrupados (calcular estadísticos descriptivos, graficar datos y realizar una prueba t de Estudiante sobre la media).

Data	Variable X	
Lower	Upper	Frequency
1775	1977	3
1978	2179	5
2180	2381	4
2382	2584	6
2585	2786	9
2787	2997	8
2998	3190	5
3191	3393	2
3394	3595	2
3596	3797	2
3798	4001	2

Description

Numerical Summaries			
Number	48	Min	1775
Mean	2741.271	Q_1	2382
St Dev	521.3104	Median	2719
Coeff of Var	0.190171	Q_3	3036.4
		Max	4001

Frequency Chart

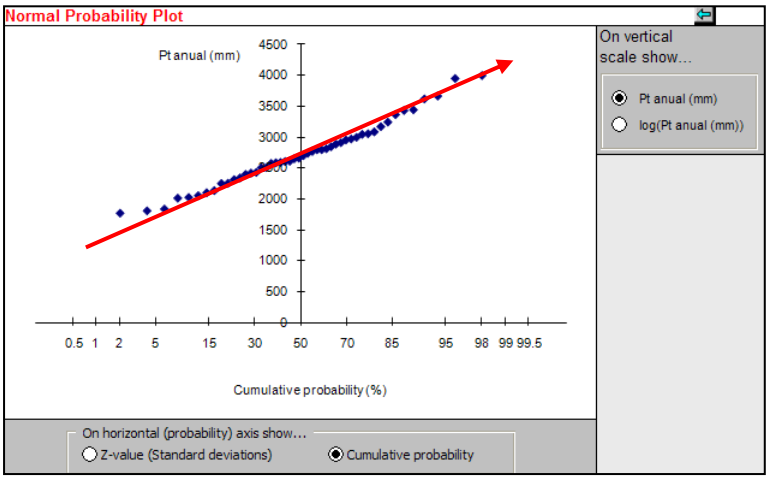


Si desea utilizar este libro puede digitar los siguientes datos en su libro de Excel.

1775	1977	3
1978	2179	5
2180	2381	4
2382	2584	6
2585	2786	9
2787	2997	8
2998	3190	5
3191	3393	2
3394	3595	2
3596	3797	2
3798	4001	2

Normal Probability Plots

Gráfico de probabilidad normal



Bajo el supuesto de distribución normal la tendencia de los datos debe ajustarse a una recta.

Si los datos provienen de una distribución lognormal la tendencia de los logaritmos de los datos (log Pt anual) debe ajustarse a una recta.

Usted puede elegir entre graficar desviaciones estandarizadas (valores de Z) ó una frecuencia acumulada.

Análisis de una variable cualitativa (nominal-ordinal)

En la presente sección se ilustra el uso de XLSTatistics para el análisis de una variable cualitativa (nominal ó ordinal). Algunos ejemplos de variables cualitativas son: uso-cobertura de la tierra (e.g. bosque, pasto, urbano), sexo (masculino, femenino), color (e.g. azul, rojo, blanco) y cultivos (eg. maíz, arroz, frijol, yuca), escala de Likert (. En el menú grafico de XLStatisitics corresponde a **1Cat**.

> **Nota**: Esta sección del tutorial asume que usted ya activó el complemento de XLStatisticis (XLStatistics.xlam ó XLStats.xls); de no ser así, ver página 23.

1. Abra el archivo "xlstats_tutorial.xlsx" y seleccione de la hoja de cálculo **1Cat** la variable ENOS.
2. Haga un clic sobre XLStatistics y luego otro clic sobre 

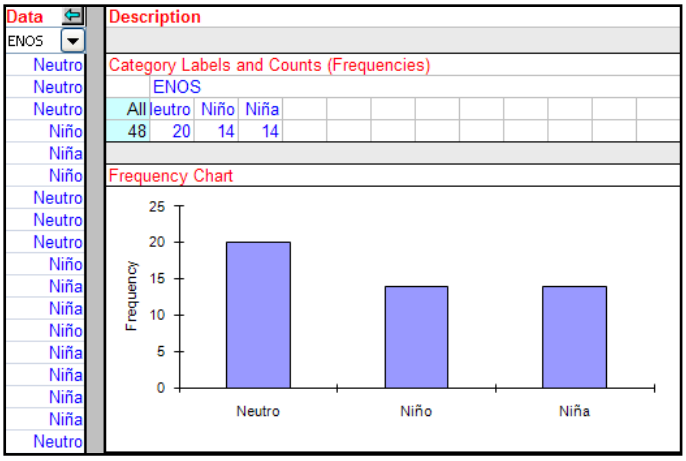
. Observe que el programa abre el libro de trabajo **1Cat.xls** y copia los datos seleccionados a dicho libro.

En la sección inferior de la hoja de cálculo **1Cat.xls** usted observará cuatro hojas de cálculo:

1) *Information*: Información. Esta hoja contiene información general sobre la organización de este libro de trabajo.
2) *Data and Description*: Datos y estadísticos descriptivos
3) *Summaries*: Síntesis de datos (Tabla de frecuencia y gráficos)
4) *Tests*: Pruebas estadísticas

Nota: Recuerde que **SOLO** debe modificar el contenido de las celdas con números o textos en color azul.

Data and Description: Datos y su descripción



Data (datos): Variable que se analiza, en este caso ENOS (Niña, Niño, Neutro).

Description (describir):
Frecuencia absoluta para cada valor de la variable ENOS (Niña, Niño, Neutro)

Grafica de frecuencia absoluta (barras).

Summaries: Tabla de frecuencia y gráficos

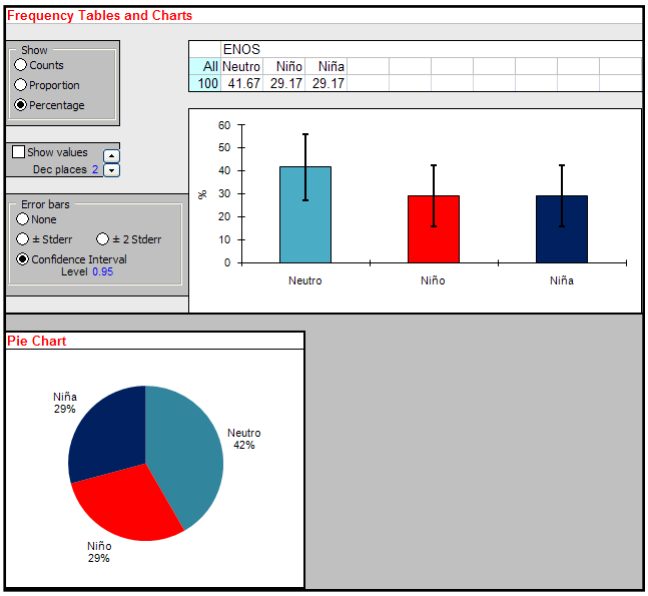
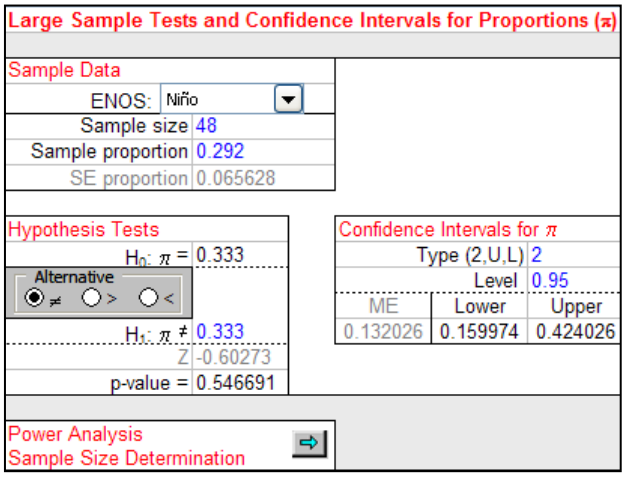


Gráfico de frecuencia (barras). Usted puede elegir número de decimales, tipo de frecuencia (absoluta, porciento, proporción). Puede adicionar al gráfico de barra ± 1 Error estándar, ± 2 Errores estándares, o un intervalo de confianza. Recuerde que usted puede elegir el nivel de confianza.

Gráfico de pastel: Si lo desea puede modificar el título del eje vertical.
Para series estadísticas con tres o más categorías ordenar las frecuencias de mayor a menor.

Tests: Intervalo de confianza y prueba de hipótesis para proporciones (N grande)



Inferencia sobre proporciones (muestras grandes-al menos 5 observaciones por celda), valor de "p" basado en aproximación a la distribución normal).

1. Prueba de hipótesis
2. Intervalo de confianza
3. Análisis de poder y determinación de tamaño de muestra

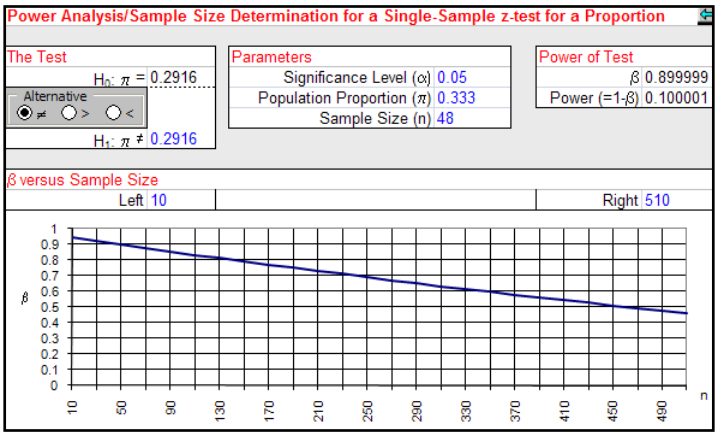
En este caso se somete a prueba la hipótesis de que la frecuencia de años Niño es igual a 0.333 (o sea no existe predominio de años Niños sobre Niñas o Neutros).

Observe que el intervalo de confianza contiene el valor 0.333. ¿Qué le indica esto?

¿Cuál es el resultado de la prueba de hipótesis?

Realice una prueba de hipótesis para los años Niña.

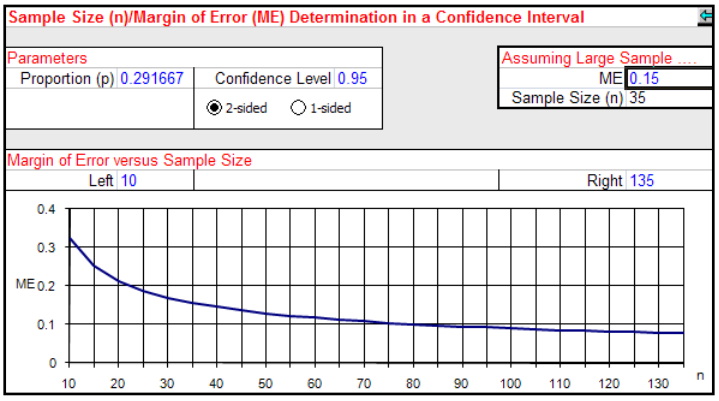
Análisis de poder y determinación de tamaño de muestra



Esta hoja de trabajo le permite realizar un análisis de poder de la prueba Z para proporciones.

Recuerde que usted puede modificar los números en color azul.

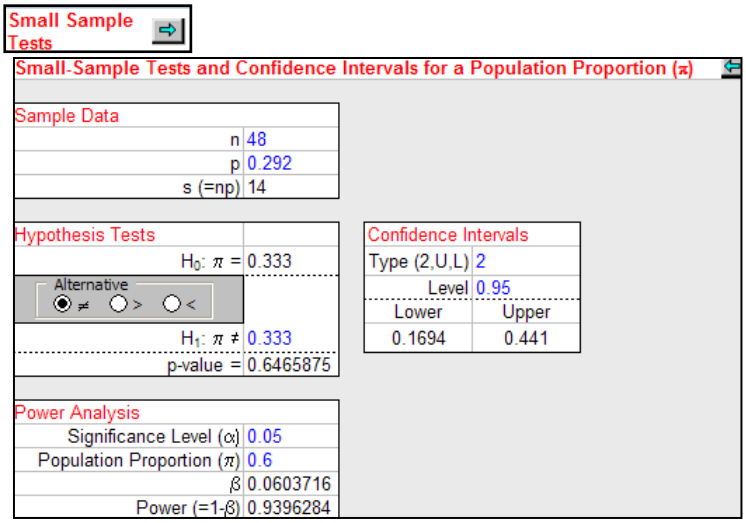
Análisis de error marginal para IC de proporciones



Esta hoja de trabajo le permite realizar un análisis de error marginal para el intervalo de confianza para una proporción.

Recuerde que usted puede modificar los números en color azul.

Intervalo de confianza y prueba de hipótesis para proporciones (muestras pequeñas)



1. Prueba de hipótesis
2. Intervalo de confianza
3. Análisis de poder

Recuerde que usted puede modificar los números en color azul.

En este caso se somete a prueba la hipótesis que para la serie la proporción esperada de años Niños, Niñas y Neutros es de 0.333.

Observe que el IC incluye el valor 0.333.

¿Cuál es el resultado de la prueba de hipótesis? ¿Cuál es el poder de la prueba?

Goodness Of Fit Test **Prueba de Bondad de ajuste (ji-cuadrado)**

Goodness-Of-Fit Test

Data

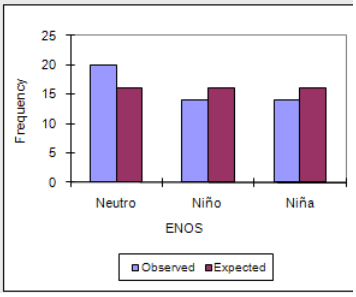
Total Count	48		
Proportions		Counts	
Observed	Expected	Observed	Expected
0.41666667	0.333	20	15.984
0.29166667	0.333	14	15.984
0.29166667	0.333	14	15.984

Hypothesis Test

H_0: Population is as expected

H_1: Population is not as expected

Chi-square	1.50154955
DF	2
p-value =	0.472000725



El programa estima la proporción observada para las variables y usted debe digitar la proporción esperada o teórica. En este caso se desea probar que la proporción entre años Niño (0.333), Niña (0.333) y Neutro (0.333) es la misma.

Para un alfa de 0.05, ¿cuál es el resultado de la prueba de hipótesis?

Runs Test **Prueba de corridas o de Wald–Wolfowitz**

(http://www.quantitativeskills.com/sisa/statistics/ordhlp.htm)

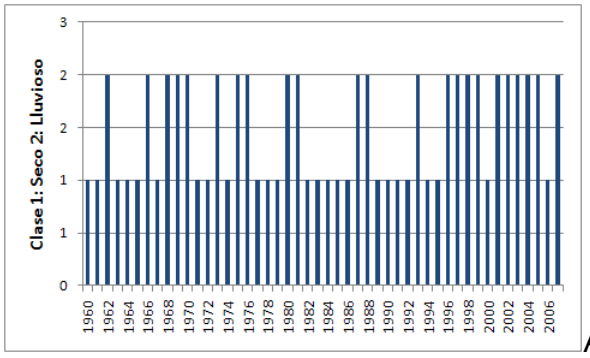
Esta prueba es aplicada a un set de datos conformado por una secuencia de dos valores (e.g. "++++−−−+++−−++++++++−−−−") y se utiliza para probar por la aleatoriedad en la secuencia del set de datos (la hipótesis nula es que los elementos de la secuencia son independientes entre sí). A continuación se ilustra el resultado para el análisis de la frecuencia de años lluviosos (mayor que la media) Vs años secos (menores que la media) para la precipitación anual de la estación Moravia de Chirripó.

Runs Test

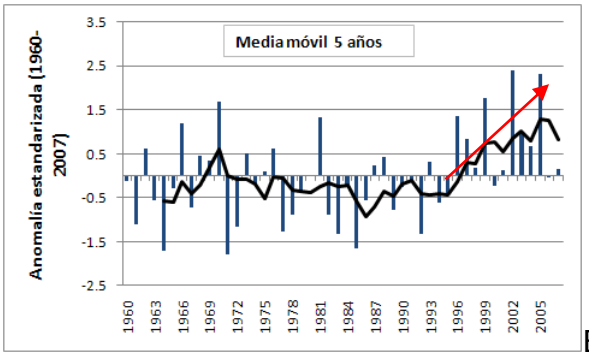
Observed number of runs (R)	22
Expected number of runs if H_0 is true (ρ)	24.9583
H_0: Data occurs in a random order, ρ =	24.9583
Alternative: ≠ > <	
H_1: Data occurs in a non-random order, $\rho \neq$	24.9583
p-value =	0.38717

Los resultados indican que existen 22 secuencias de seco-lluvioso (o lluvioso-seco) y que para un set de datos aleatorio dado el tamaño de muestra (41 años) se esperarían 24.95 secuencias.

El resultado de la prueba indica que efectivamente los datos ocurren en un orden aleatorio; o sea no existe un patrón de años más lluviosos o más secos en el set de datos.



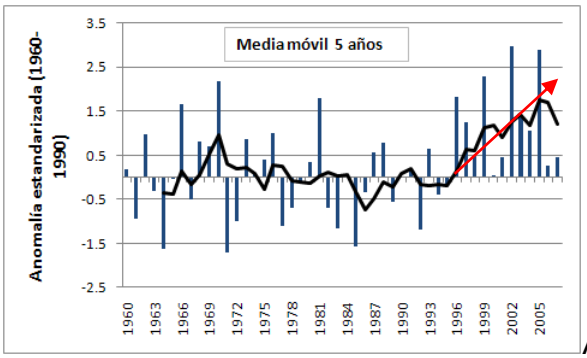
A



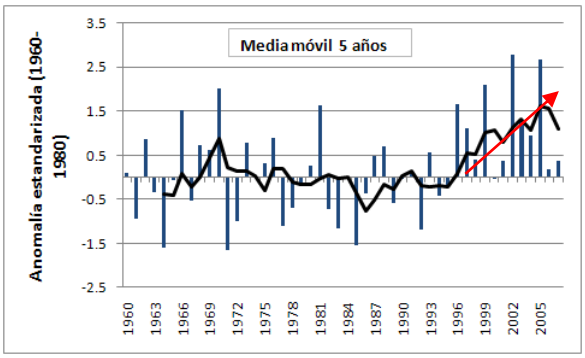
B

A. Secuencia de años clasificados como secos (1) y lluviosos (2)
B. Pt anual expresado como un valor estandarizado (anomalía estandarizada basado en el periodo 1960-2007).

Observe que a partir de 1996 existe una mayor frecuencia de años “lluviosos”; sin embargo a partir de 2006 parece que se inicia otro ciclo de menor precipitación.



A.



B.

A. Anomalías estandarizadas basado en el periodo 1960-1990.

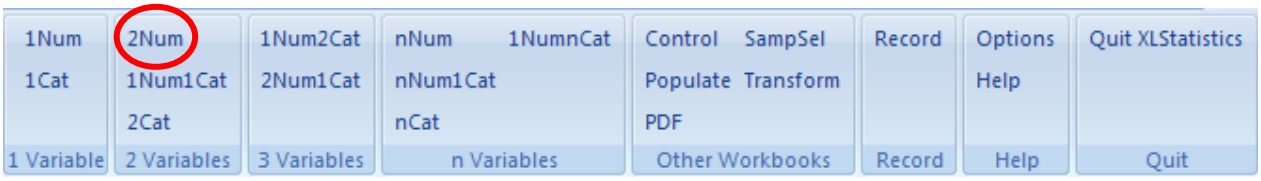
B. Anomalías estandarizadas basado en el periodo 1960-1980.

La anomalía estandarizada es: (Pt año-Pt media)/desviación estándar

Observe que a partir de 1996 existe una mayor frecuencia de años “lluviosos”; sin embargo a partir de 2006 parece que se inicia otro ciclo de menor precipitación.

Análisis de dos variables numéricas (continuas ó discretas): Correlación y regresión simple

En la presente sección se ilustra el uso de XLSTatistics para el análisis de dos variables cuantitativas. En el menú grafico de XLStatisitics corresponde a **2Num.**



> **Nota**: Esta sección del tutorial asume que usted ya activó el complemento de XLStatisticis (XLStatistics.xlam ó XLStats.xls); de no ser así, ver página 23.

A continuación analizaremos los datos de las variables año y precipitación anual. A esta serie se le conoce como una serie temporal porque involucra a la variable tiempo (años).

> **Nota**: Dado que las dos variables son cuantitativas, XLStatistics ofrece herramientas para realizar un análisis de correlación y regresión simple. En Excel la primera columna corresponde al eje "Y" y la segunda al Eje "X".
>
> En este caso deseamos analizar el comportamiento de la lluvia anual en el tiempo y por tanto la variable Y es PT anual y la variable X el tiempo.

1. Abra el archivo "xlstats_tutorial.xlsx" y seleccione de la hoja de cálculo **2Num** las variables Año (variable predictora) y Moravia Pt. anual (mm) (variable dependiente).
2. Haga un clic sobre XLStatistics y luego otro clic sobre 

. Observe que el programa abre el libro de trabajo **2Num.xls** y copia los datos seleccionados a dicho libro.

En la sección inferior de la hoja de cálculo **2Num.xls** usted observará cuatro hojas de cálculo:

5) *Information*: Información. Esta hoja contiene información general sobre la organización de este libro de trabajo.
6) *Data and Description*: Datos y estadísticos descriptivos
7) *Corr &Linear Regress*: Correlación y regresión simple
8) *Extra Tools*: Herramientas adicionales

Nota: Recuerde que **SOLO** debe modificar el contenido de las celdas con números o textos en color azul.

Conceptos básicos sobre regresión lineal simple

La ecuación de regresión de “Y” en “X” muestra el valor medio esperado de “Y” dado un valor de “X”. La variable “Y” se denomina *dependiente* y la variable “X” *predictora*. Una distribución condicional es la distribución de una variable (Y) dado un valor particular de otra variable (X). Por ejemplo, dado un bosque, la distribución condicional de la variable altura (h) dada un valor particular de la variable diámetro (d) mostraría todos los posibles valores de altura para cada uno de los valores de diámetro. Dicha relación se simboliza como h|d. Dado que no es posible conocer todas las posibles distribuciones condiciones de “h” dada cada uno de los posibles valores de “d”; se asume que el muestreo independiente realizado de las variables “h” y “d” representa dicha relación. Bajo esta suposición y asumiendo que la distribución de “h” es normal para cada valor de “d”, la mejor estimación de “h” dado “d” sería la media de la distribución como se muestra en la figura 1.

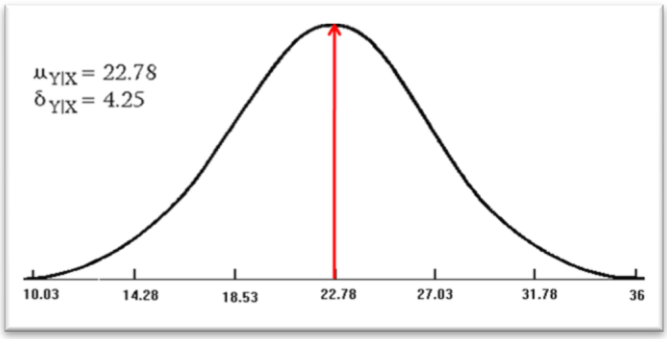


Figura 1: Distribución hipotética de la variable altura (h) en metros para un valor de diámetro de 40 cm. La distribución de valores de altura es normal con una media de 22.78 m y un error de estimación es 4.25 m.

Ahora supongamos que existen otros valores de altura y diámetro como se muestra en la figura 2. La ecuación de regresión ajusta una recta a través de las cimas de cada una de las distribucioes de h|d. Sin embargo, observe que el ajuste no es perfecto, o sea existe un error de estimación; dicho error se utiliza para calcular los intervalos de confianza y realizar las pruebas de hipótesis sobre la validez estadística del modelo. Cuanto más pequeño sea dicho error mayor capacidad predictiva y descriptiva tendrá el modelo.

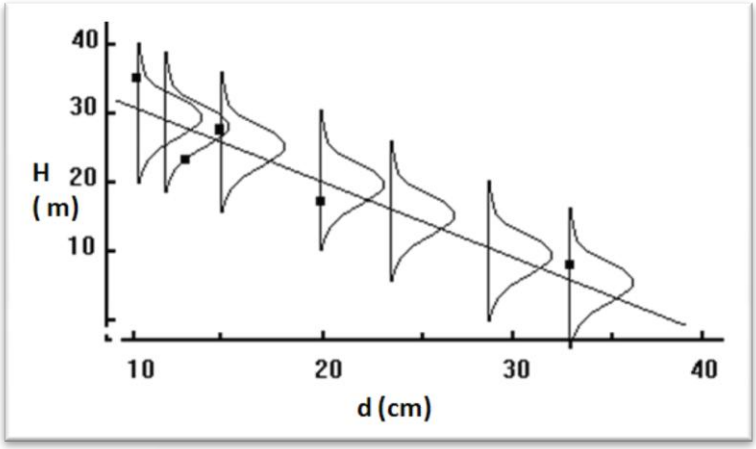
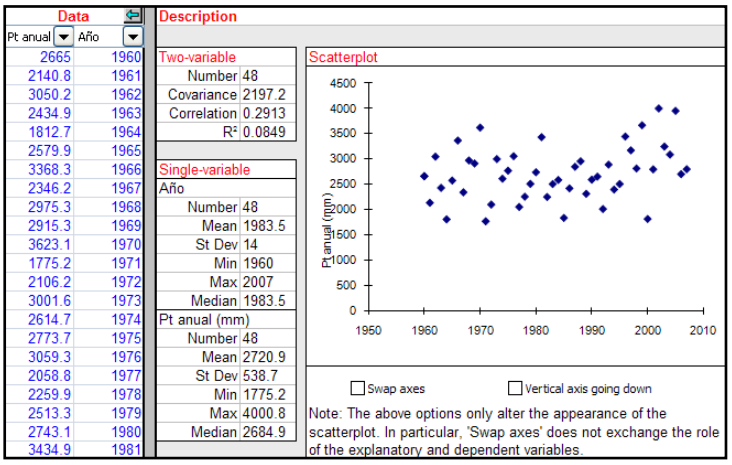


Figura 2: Distribución hipotética de valores de altura (h, m) y diámetro (cm). El error de estimación es una función de la amplitud de cada una de las distribuciones de “h” dado un valor de “d”.

Data and Description: Datos y su descripción



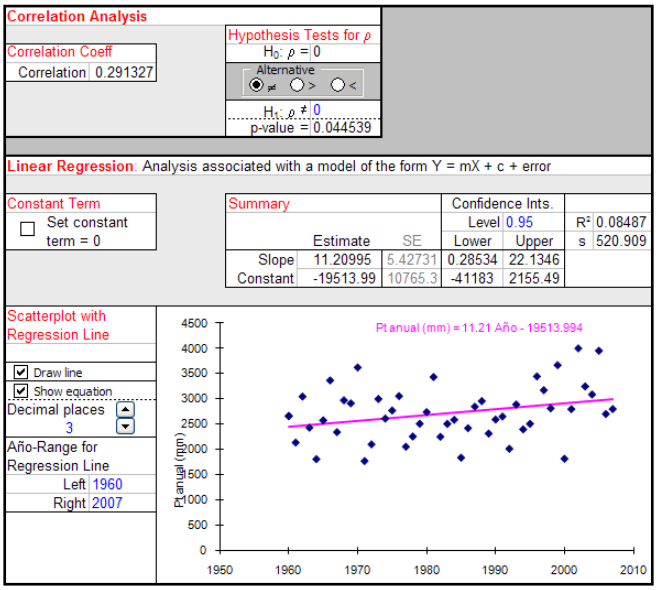
Data: Listado de datos

Two-variable: Análisis de correlación (r) y coeficiente de determinación (R^2).

Single variable: Estadísticos descriptivos para las variables Y y X.

Scatterplot: Gráfico de dispersión

Análisis de correlación y regresión lineal



Análisis de correlación

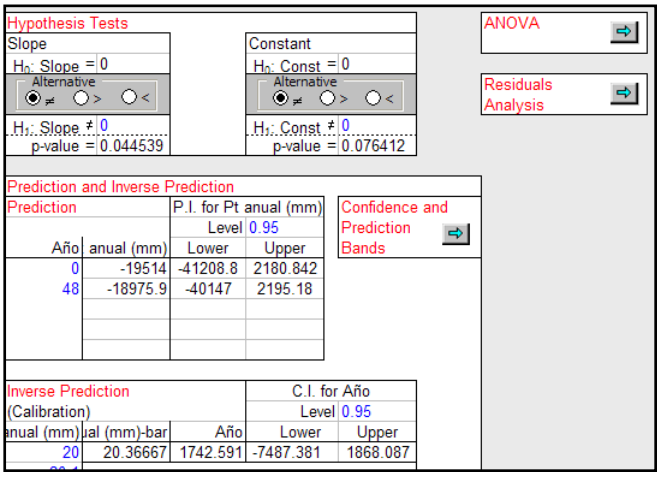
1. Coeficiente de correlación lineal
2. Prueba de hipótesis sobre rho.

Para un alfa 0.05 ¿se rechaza Ho?____

Regresión lineal simple

1. Estimadores (intercepto y pendiente), error estándar e intervalos de confianza.
2. Coef. determinación (R^2).
3. Error estándar de estimación (Syx) = √CME.
4. Diagrama de dispersión.

Prueba de hipótesis (intercepto, pendiente)



Prueba de hipótesis

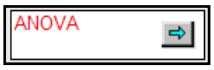
A. Pendiente. Para un alfa 0.05 ¿se rechaza Ho?____

B. Intercepto. Para un alfa 0.05 ¿se rechaza Ho?____

Predicción: estimar valores de Y a partir de valores de X e intervalo de predicción.

Predicción inversa: estimar valores de X a partir de valores de Y.

Análisis de varianza: significancia del modelo



La hipótesis nula es que no existe regresión de Y en X; o sea no existe covariación entre Y y X.

Analysis of Variance

ANOVA Table

Source	DF	SS	MS	F	p-value
Regression	1	1157607.2	1157607.2	4.2661708	0.0445391
Residual	46	12481903	271345.71		
Total (corrected)	47	13639510			
Mean	1	355368586			
Total (uncorrected)	48	369008096			

Tabla de análisis de varianza

DF: Grados de libertad
SS: Suma de cuadrados
MS: Cuadrado medio
F: valor de F
p-value: valor de "p" corresponde al valor de F en la distribución F.

A un nivel de significancia de 0.05, la regresión de precipitación en tiempo es estadísticamente significativa (p=0.044).

Análisis de residuos: Evaluación de los supuestos del modelo

Los supuestos del análisis de regresión son las siguientes:

1- Modelo
   a. *El modelo es completo*. La ecuación posee todos las variables requeridas para describir la relación entre Y y X. Ejemplo: *Y=a+bX+e.* No existe ninguna prueba estadística que le indique directamente si el modelo posee todas las variables requeridas para describir la relación entre Y y X. Le sugiero utilizar sus conocimientos teóricos (e.g. publicaciones previas, consulta con expertos, colegas, etc.) para evaluar este criterio.
   b. *Lineal*. La relación entre las variables X y Y es lineal. Principio de linealidad de la regresión.
   c. *Aditivo*. La relación entre las variables es aditiva. Este supuesto es especialmente crítico en regresiones múltiples (dos o más variables predictoras); ya que el efecto de una variable puede estar afectado por otra variable. Por ejemplo, el efecto de la fertilización en el crecimiento de un bosque segundario puede depender del uso previo del sitio, de la profundidad del suelo ó del pH. No existe ninguna prueba estadística que le indique directamente si el modelo requiere de términos no aditivos. Si usted sospecha que la relación entre las variables es no aditiva le sugiero los siguiente:
      d. Utilizar sus conocimientos teóricos (e.g. publicaciones previas, consulta con expertos, colegas, etc.) para evaluar este criterio.
      e. Ajustar otras funciones y comparar los valores de R^2 ó del cuadrado medio del error.
      f. Ajustar una regresión separada para los diferentes grupos utilizando variables de clasificación.
      g. Incluir un término multiplicativo en el modelo (e.g. X_1*X_2)

2- Variables
   a. *Nivel de medición de intervalo o razón*.
   b. *Variables se miden sin error*. Se supone que las variables solo son afectadas por el error aleatorio. Si los datos incluyen un error sistemático, el error de estimación del modelo incrementará y el valor de la pendiente será diferente.

3- Error del modelo

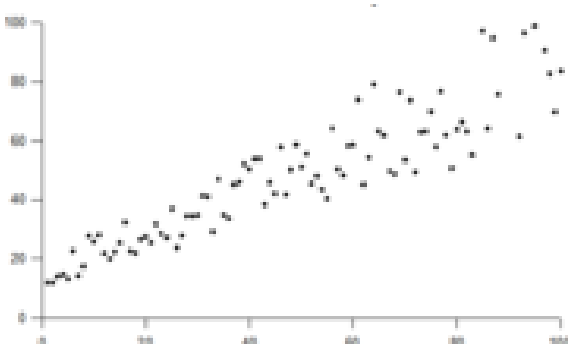
a. *Distribución normal.* Las pruebas estadísticas dependen de este supuesto; sin embargo la inferencia estadística puede se robusta aun cuando se viole este supuesto. Graficar residuos y evaluar su distribución, realizar prueba de hipótesis para normalidad.

b. Su valor esperado es 0. Graficar residuos y evaluar su distribución alrededor de cero.

c. Los errores son independientes entre sí. La inferencia estadística asume que cada caso u observación es independiente de la otra. Por ejemplo, la medición de lluvia diaria en el mes de octubre no es independiente, ya que es probable que los días lluviosos no ocurran al azar.

> La no independencia entre las observaciones sesga la estimación del error estándar, ya que la fórmula asume que las observaciones son independiente ("n muestras independientes"). Si las observaciones están correlacionadas, el valor de "n" utilizado es más grande que el real y por tanto el error estándar estimado será más pequeño y las pruebas estadísticas tendrán una mejor significancia (se tenderá a rechazar Ho). Esto también puede sesgar las estimaciones de la intercepción y la pendiente del modelo. La no linealidad es un caso especial de los errores correlacionados.

d. Homocedasticidad de los errores (varianzas son constantes) con respecto a la variable predictora. La no homogeneidad de la varianza sesga la estimación del error estándar y por ende la significancia estadística. Si sus datos no cumplen con este supuesto puede utilizar una matriz ponderada. (e.g. 1/X). La ecuación original $Y=a+bX+e$ se convierte en $Y/X=a/X+b+e/X$. En este nuevo modelo el error para los valores grandes de X será más pequeño (e/X). Para ajusta esta nueva regresión, usted debe crear la variable Y/X (su nueva variable dependiente) y 1/X (su nueva variable predictora). Para este nuevo modelo, el valor del intercepto (a) será ahora la pendiente y el valor de la pendiente (b) será la intersección de la nueva ecuación.



> Observe que la dispersión de los puntos es mayor conforme aumenta el valor de X. Este es un claro ejemplo de no homogeneidad de varianzas.

e. La variable predictora es independiente de los errores. Los errores representan la variabilidad de "Y" no explicada por la variable predictora "X" y por tanto representan todos aquellos factores que influyen en "Y" y que no se incluyeron en la ecuación de regresión. Si alguna variable omitida está correlacionada con X, este supuesto es violado. El error será **SIEMPRE** *independiente de "X"* y por lo tanto no hay manera de establecer el verdadero error.

f. En un *sistema de ecuaciones interrelacionadas los errores son independientes*.

Si las condiciones anteriores se cumplen y los datos provienen de un muestreo probabilístico, entonces el ajuste por mínimos cuadrados provee estimadores insesgados, eficientes y consistentes.

A. Insesgados: en promedio, el estimador es igual a la media de la población. Ejemplo: $E(b)=\beta$

B. Eficientes: Estima el parámetro de interés con el menor error posible. Ejemplo: el error estándar de estimación será un mínimo.

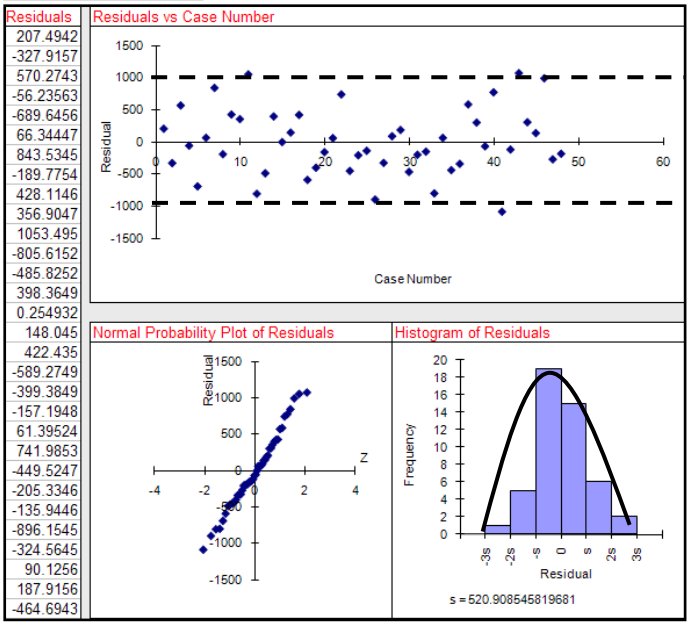
C. Consistentes: Conforme el tamaño de muestra aumenta el error del estimador tiende a cero.

Si alguno de estos supuestos es violado, entonces las estimaciones de los parámetros del modelo, las predicciones, las pruebas de hipótesis, los intervalos de confianza y la relación entre las variables indicada por el modelo de regresión puede ser, en el mejor de los casos, ineficiente o peor aun, estar gravemente sesgadas o ser engañosas. Normalmente los residuos se utilizan para probar por los siguientes supuestos del modelo:

Supuesto	Prueba
La relación entre las variables X y Y es lineal	Diagrama de dispersión, gráfico de valores observados Vs valores estimados, gráfico de residuos Vs valores estimados ((.$\hat{y}_i$ En caso de violar este supuesto debe utilizar un modelo linealizable ó no lineal, transformar las variables ó utilizar variables de clasificación (dividir del set de datos en segmentos lineales).
Los errores son independientes	Para series temporales graficar errores vs tiempo. Para otros datos utilizar el gráfico de autocorrelación de residuos de resago 1. Es deseable que la mayoría de las autocorrelaciones residuales estén dentro de las bandas de confianza del 95% en torno a cero, las cuales se encuentran aproximadamente a ± 2/(n)$^{0.5}$, donde "n" es el tamaño de la muestra. Así, si el tamaño de la muestra es de 48, las autocorrelaciones debe estar entre +/- 0,29. Ponga especial atención a las autocorrelaciones significativas de resago uno y dos. El estadístico Durbin-Watson (d) es utilizado para probar por autocorrelación de resago-1 entre residuos: "d" es aproximadamente igual a 2*(1-r) donde "r" es la autocorrelación residual: un valor de 2,0 indica ausencia de autocorrelación. Como regla general, cualquier valor de "d" inferior a 1 indica autocorrelación grave entre los residuos. El valor de "d" siempre se encuentra entre 0 y 4. Si el estadístico de Durbin-Watson es sustancialmente menor que 2, existe evidencia de correlación serial positiva. Un estadístico robusto y más poderoso que DW es la prueba LM de autocorrelación de residuos de Breusch-Godfrey. $d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$. Relación entre autocorrelación "r" y 2*(1-r) r — 2*(1-r) 0.05 — 1.9 0.1 — 1.8 0.2 — 1.6 0.3 — 1.4 0.4 — 1.2 **0.5 — 1** 0.6 — 0.8 0.7 — 0.6 0.8 — 0.4 0.9 — 0.2 1 — 0

Homocedasticidad de los errores (varianzas son constantes) con respecto a la variable predictora.	Gráfico de residuos versus variable predictora (en este caso “año”).
Los errores tienen una distribución normal.	Gráfico de probabilidad normal e histograma de residuos. Prueba de normalidad.
El valor esperado de los residuos es cero	Graficar residuos y evaluar su distribución alrededor de cero.

El libro 2NRP.xls de XlStatistics le ofrece las siguientes pruebas para evaluar los supuestos del modelo.



Residuos versus observación. Los valores son gráficos según su orden (1 a n). Esta gráfica permite determinar si existe algún patrón en el orden en que se midieron los datos (e.g. error sistemático). Observe que los residuos se encuentran dentro de una banda de ± 1000 mm.

Grafico de probabilidad normal. Para cumplir con el supuesto de normalidad, los residuos deben ajustarse a una recta y el histograma debe mostrar una forma de campana.

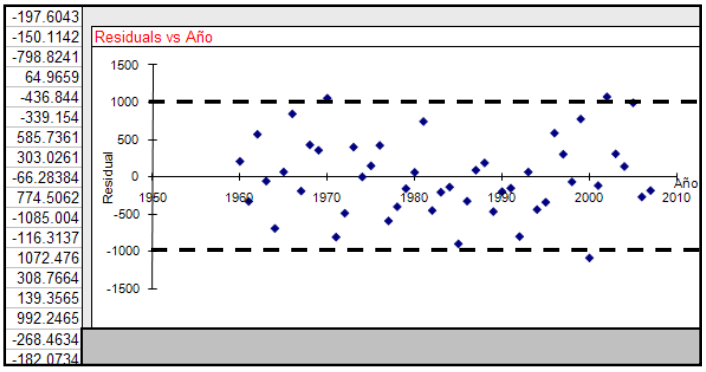
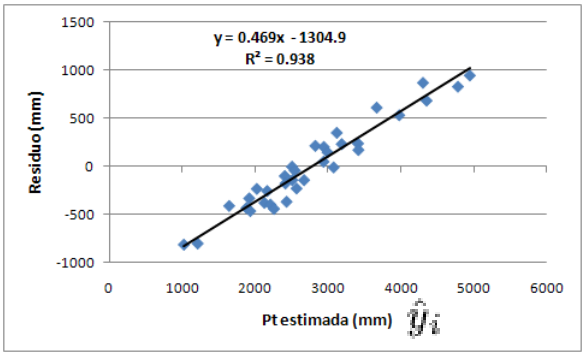


Gráfico de errores versus variable predictora (X). Para cumplir con el supuesto de homocedasticidad la varianza de los residuos debe ser constante con respecto a la variable predictora. Observe que los residuos se encuentran dentro de una banda de ± 1000 mm. El error será **SIEMPRE** *independiente de "X"* y por lo tanto no hay manera de establecer el verdadero error.

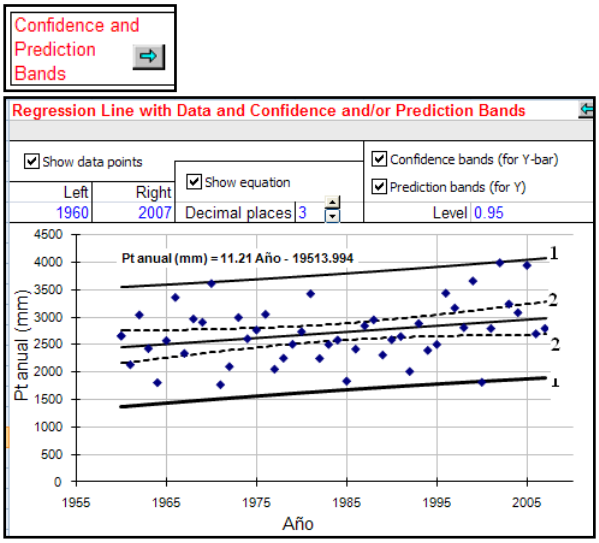
Para probar por el supuesto de linealidad usted puede graficar los residuos versus los valores estimados, para esto debe crear dos nuevas variables como se muestra a continuación.

Pt Estimado	Residuos
2832	217.3
3127	353.0
3675	615.2
1650	-408.6
2029	-230.8
2513	-0.8
2949	205.7
4309	874.2
1925	-329.4
2413	-97.4
2550	-40.1
1029	-812.4
.	.



Observe que los datos se ajustan a una recta, por cuanto puede asumirse que cumplen con el supuesto de linealidad.

Intervalo de predicción y banda de predicción



Línea de regresión y bandas de predicción para una confianza de 95%.
1: Banda de predicción de Yi.
2. Banda de predicción de media de Y (Yμ).

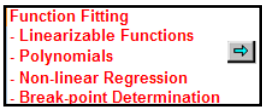
Observe que la banda de predicción de la media de Y es siempre más angosta que para un valor particular de Y.

Herramientas adicionales

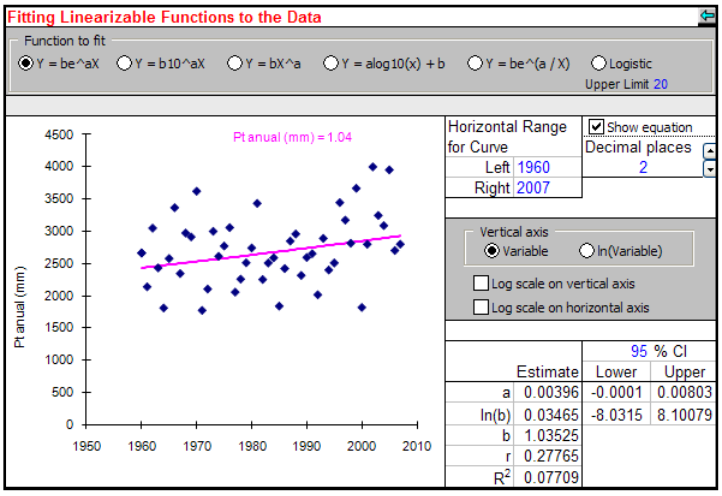
Esta hoja de trabajo le brinda acceso a las siguientes funciones:

1. Ajuste de funciones: ecuaciones linealizables, regresión polinomial, regresión no lineal y determinación de puntos de inflección en la tendencia del set de datos.
2. Suavizado de tendencia en el set de datos. Media móvil, media con barras de error y regresión localmente ponderada.
3. Síntesis numérica y gráfica para cada una de las variables en el set de datos.
4. Análisis de correlación no paramétrico: Coeficientes de correlación de Spearman y de Kendall.
5. Análisis de series de tiempo. Ajuste de función exponencial y suavizado de datos.
6. Adicionar etiquetas a datos

Ajuste de funciones

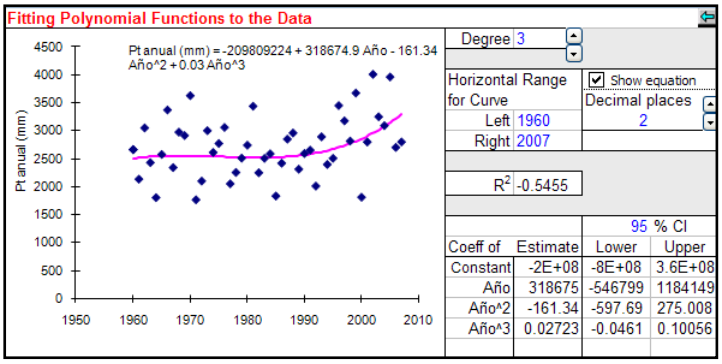


Transformaciones lineales



Esta hoja de trabajo permite ajustar ecuaciones no lineales en su forma original pero que pueden linealizarse mediante el uso de logaritmos (naturales y de base 10).
El programa ofrece estimaciones de "a" e IC, "b" e IC, r y R^2; así como la ecuación de regresión.
También puede ajustar una regresión logística.

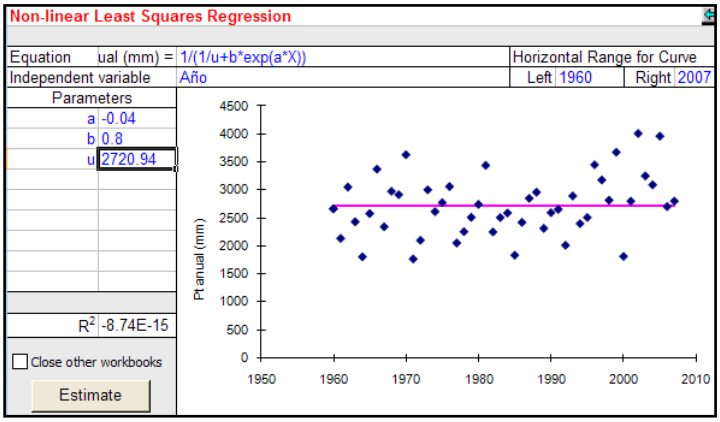
Ecuación polinomial



Esta hoja de trabajo le permite ajustar un polinomio de grado "n" al set de datos.

El programa ofrece estimaciones de "a" e IC, "b" e IC y R^2; así como la ecuación de regresión.

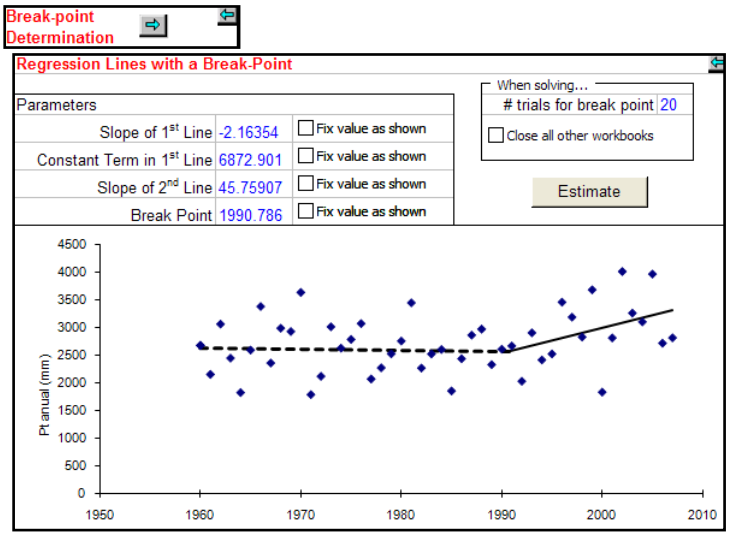
Regresión no lineal



El texto azul corresponde a los parámetros que usted desea utilizar para ajustar la ecuación que usted seleccionó. El único estadístico que ofrece el programa es el R^2.

Nota: Esta función utiliza el complemento "Solver" de Excel.

Cambio de pendiente (tendencia) en el set de datos

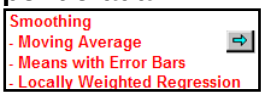


El texto azul corresponde a los parámetros que el programa utiliza para determinar si existe un punto donde cambia la pendiente de la recta ajustada al set de datos.

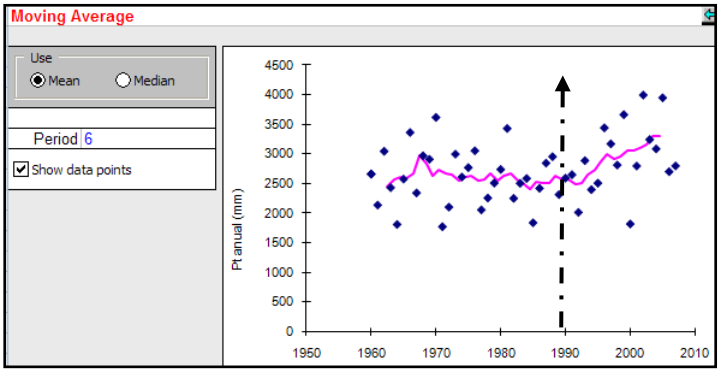
En este caso, el programa encontró un cambio de tendencia a partir de 1990. La pendiente de la primer recta es negativa (-.2.16 mm por año) en tanto que para la segunda es positiva (45.8mm por año).

Notas:
1. Esta función utiliza el complemento “Solver” de Excel.
2. El punto de cambio en la pendiente del set de datos puede variar con el número de eventos utilizados (*# trials for breaking point*).

Suavizado de tendencia: media móvil, media con barras de error, regresión localmente ponderada



Media móvil



Media móvil
La media móvil es utilizada con series de tiempo para determinar si existe algún patrón en el set de datos. Usted puede elegir entre la media ó la mediana; así como el número de años a utilizar en el cálculo.

Observe que en este caso aparentemente existe una tendencia a mayor precipitación a partir de 1990.

Media con barras ó bandas de error

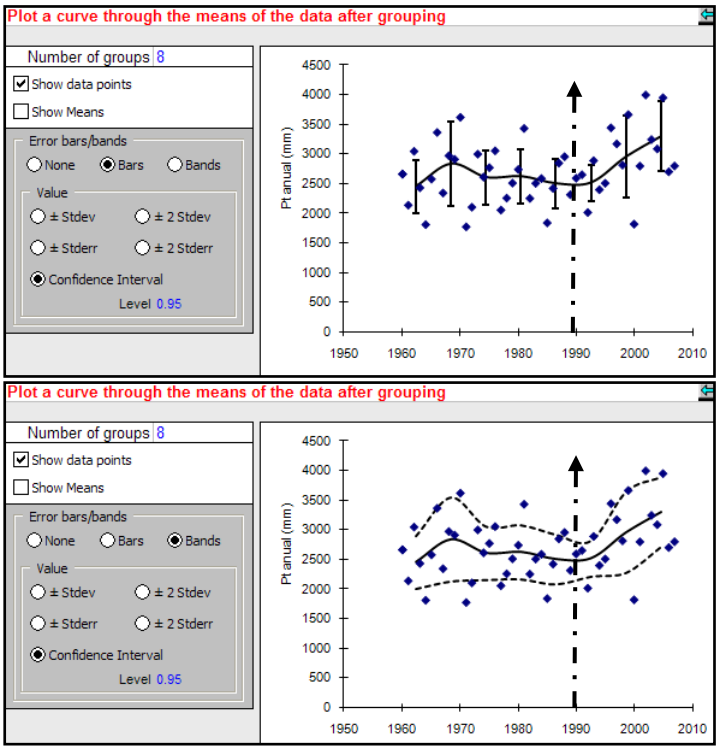
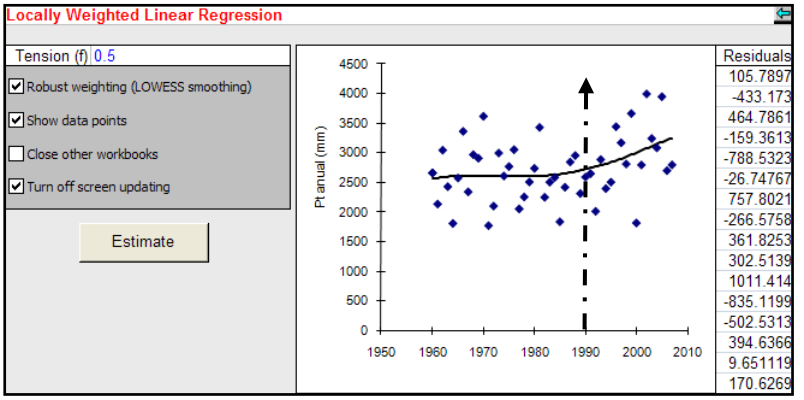


Gráfico de puntos medios para “n” subgrupos de datos. En este caso el set de datos de dividió en 8 grupos (aproximadamente 5 observaciones por grupo). Las líneas verticales corresponden a un IC de 95%. Usted puede elegir además entre ± 1S, ± 2S, ± 1 Error estándar y ± 2 Errores estándares.

El grafico inferior muestra bandas de confianza al 95%.

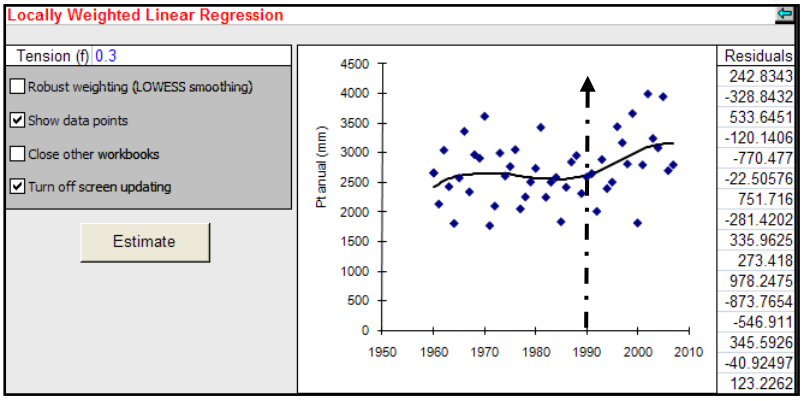
Observe que en este caso aparentemente existe una tendencia a mayor precipitación a partir de 1990 así como un pico de precipitación en 1970; el cual fue muy lluvioso en la costa Caribe.

Regresión lineal localmente ponderada



Regresión robusta localmente ponderada(LOWESS)

El factor de tensión (f) define la forma de la curva ajustada al set de datos. Valores cercanos a cero tienden a generar líneas más irregulares. La columna de la izquierda muestra los residuos.



Regresión localmente ponderada(LOWESS)

El factor de tensión (f) define la forma de la curva ajustada al set de datos. Valores cercanos a cero tienden a generar líneas más irregulares. La columna de la izquierda muestra los residuos.

LOWESS y LOESS

El término "LOWESS" proviene del inglés "**Lo**cally **We**ighted **S**catter plot **S**moothing"; sin embargo algunos autores también utilizan la palabra "LOESS" como sinónimo. Ambos métodos utilizan regresiones lineales ponderadas localmente con el objetivo de suavizar la tendencia local en el set de datos. Por defecto las funciones **lowess** y **loess** realizan un ajuste localmente lineal y localmente cuadrático, respectivamente.

LOESS, es un método de regresión polinomial localmente ponderada propuesto originalmente por Cleveland (1979) y desarrollado posteriormente por Cleveland y Devlin (1988). Dicho método de regresión se caracteriza por ser una técnica más descriptiva que predictiva. Para cada punto del conjunto de datos se ajusta un polinomio de bajo grado, utilizando los valores de la variable explicativa más cercanos de dicho punto (entorno de X). El polinomio utiliza mínimos cuadrados ponderados, dando más peso o importancia a los puntos más cercanos al punto cuya respuesta se está estimando y menos peso a los puntos más alejados.

La principal ventaja de LOESS es que el usuario(a) no debe especificar ningún modelo que deba ajustarse a todos los datos de la muestra. En su lugar, el o la analista solo tienen que proveer un valor de parámetro de suavizado y el grado del polinomio local. LOESS es una técnica de análisis muy flexible y por tanto es ideal para analizar procesos para los cuales no existen modelos teóricos. La técnica LOESS permite calcular la incertidumbre asociada al modelo de predicción y de calibración así como aplicar la mayoría de las pruebas y procedimientos utilizados para validar los modelos de regresión basados en mínimos cuadrados.

Entre las desventajas del método tenemos que hace un uso menos eficiente de los datos que otros métodos de mínimos cuadrados. Sin embargo, dados los resultados que el método proporciona, sin duda podría ser más eficiente en general, que otros métodos de estimación como el de mínimos cuadrados no lineales. Otra desventaja de LOESS es que no produce una función de regresión y por tanto no puede transferirse a otros usuarios(as).

LOESS, al igual que cualquier otro método de mínimos cuadrados, es afectado por valores atípicos o extremos en el set de datos. La versión iterativa y robusta de LOESS (Cleveland (1979) reduce dicha sensibilidad; sin embargo, valores muy atípicos o extremos pueden superar incluso el método de análisis más robusto.

Si usted lo desea puede descargar de http://peltiertech.com/WordPress/loess-utility-awesome-update/ un complemento gratuito para Excel que ajusta una función LOESS a sus datos.

http://peltiertech.com/WordPress/loess-utility-what-the-buttons-mean/ Descripción de cada uno de los botones en el complemento

Referencias

Cleveland, W.S. (1979). "Robust Locally Weighted Regression and Smoothing Scatterplots". *Journal of the American Statistical Association* **74** (368): 829–836. MR0556476. JSTOR 2286407.

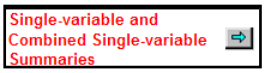
Cleveland, W.S.; Devlin, S.J. (1988). "Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting". *Journal of the American Statistical Association* **83** (403): 596–610. JSTOR 2289282.

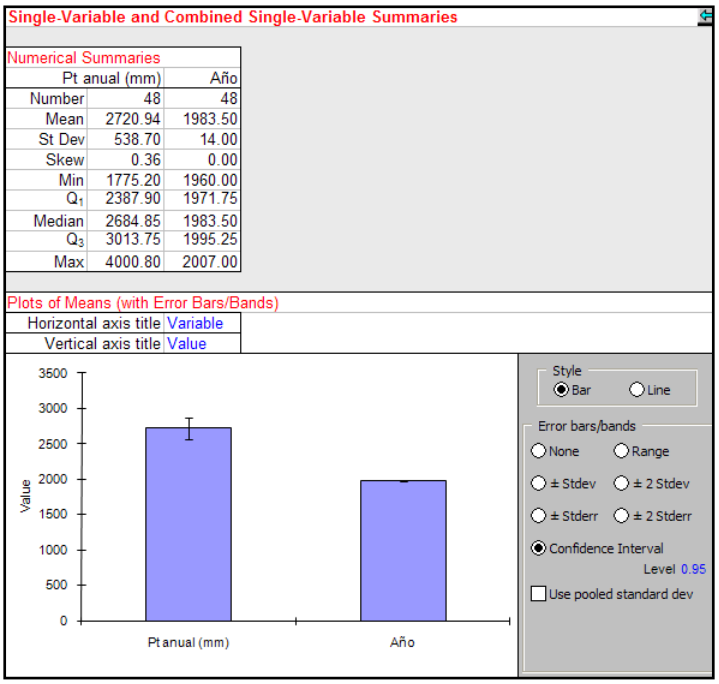
http://www.stat.purdue.edu/~wsc/papers/localregression.principles.ps

NIST Engineering Statistics Handbook Sección sobre LOESS

Local Polynomial Regression Fitting utilizando R.
http://stat.ethz.ch/R-manual/R-patched/library/stats/html/loess.html

Estadísticos descriptivos y gráficos para la variable predictora (X) y la dependiente (Y)





Estadísticos descriptivos y gráficos (medias con líneas/bandas de error, frecuencia absoluta y relativa, diagrama de cajas) para las variables precipitación anual y tiempo en años.

Observe que usted puede adicionar al diagrama de barras una línea vertical indicando:
- Rango de los datos
- Desviación estándar
- Error estándar
- Intervalo de confianza

Nota: Para este set de datos solo los estadísticos de Pt anual son de interés.

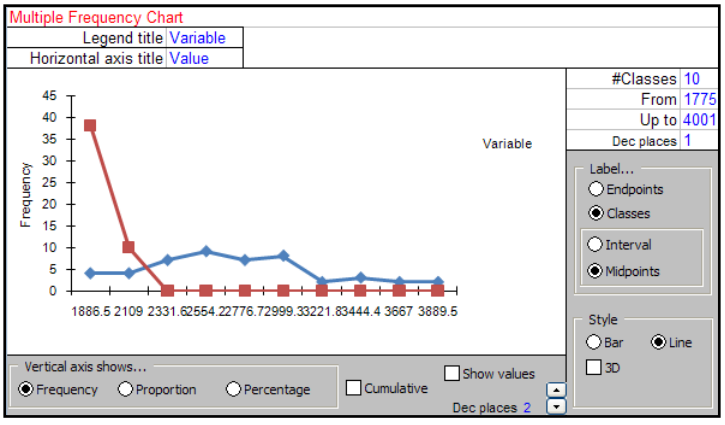


Grafico de frecuencia
Etiquetas eje Y: frecuencia absoluta, proporciones, porcentaje, acumulada.

Etiquetas eje X:
Límite superior de clase
Clases: punto medio, intervalo de clase

Estilo: Barras, 3D y líneas.

Recuerde que usted puede modificar el texto y los números de color azul.

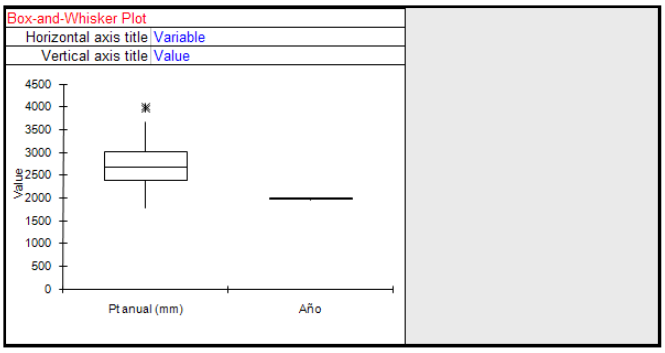
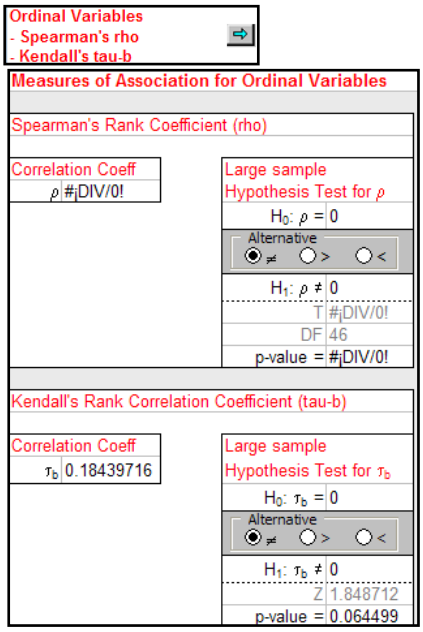


Diagrama de cajas o de Box-Whisker.

Recuerde que usted puede modificar el texto y los números de color azul.

El * indica un valor atípico o extremo.

Análisis de correlación no paramétrico

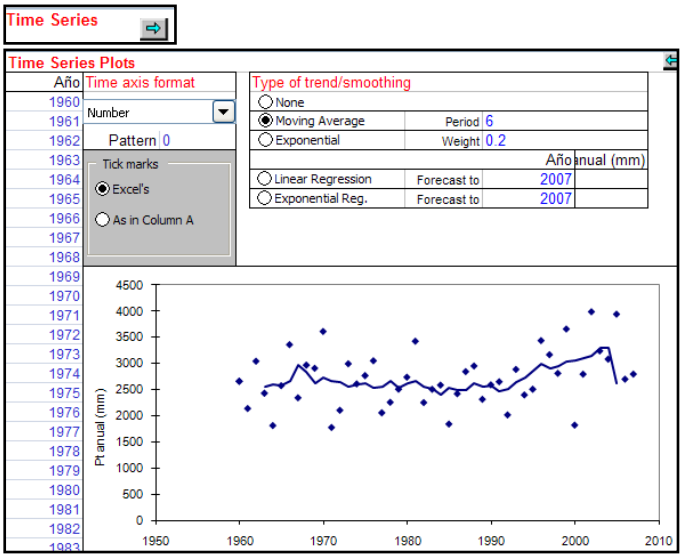


Análisis de correlación no paramétrico. Correlación entre órdenes de los datos.

Coeficiente de correlación de Spearman (rho) y prueba de hipótesis sobre significancia de la correlación.

Coeficiente de correlación de Kendall (tau-b) y prueba de hipótesis sobre significancia de la correlación.

Series de tiempo

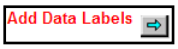


Esta hoja de cálculo asume que usted está analizando una serie de tiempo y por esta razón solo observa la columna de tiempo.

Funciones de tendencia o suavizado

A. Media móvil, usted debe elegir el periodo o número de años.
B. Exponencial: usted debe elegir el peso.

Predicciones utilizando una regresión lineal. Digite el año.
Predicciones utilizando una regresión exponencial. Digite el año.



Esta hoja de trabajo adiciona etiquetas a los datos.

Ejercicio

Si usted desea practicar lo expuesto en esta sección puede utilizar los datos de lluvia anual de las estaciones Batán (15 msnm) y Moravia de Chirripó (1200 msnm) (hoja 2Num del archivo xlstats_tutorial.xlsx).

Modelos matemáticos para datos cuantitativos bivariados

	Modelo	Trans. X	Trans. Y
Simple	$Y = \alpha_0 + \alpha_1 X$	$t(x) = x$	$t(y) = y$
Expon.	$Y = \exp(\alpha_0 + \alpha_1 X)$	$t(x) = x$	$t(y) = \ln(y)$
Recípr. Y	$Y = \frac{1}{\alpha_0 + \alpha_1 X}$	$t(x) = x$	$t(y) = \frac{1}{y}$
Recípr. X	$Y = \alpha_0 + \alpha_1 \frac{1}{X}$	$t(x) = \frac{1}{x}$	$t(y) = y$
Rec Doble	$Y = \frac{1}{\alpha_0 + \alpha_1 / X}$	$t(x) = \frac{1}{x}$	$t(y) = \frac{1}{y}$
Logar. X	$Y = \alpha_0 + \alpha_1 \ln(X)$	$t(x) = \ln(x)$	$t(y) = y$
Multipl	$Y = \alpha_0 X^{\alpha_1}$	$t(x) = \ln(x)$	$t(y) = \ln(y)$
Raíz C. X	$Y = \alpha_0 + \alpha_1 \sqrt{X}$	$t(x) = \sqrt{x}$	$t(y) = y$
Raíz C. Y	$\sqrt{Y} = \alpha_0 + \alpha_1 X$	$t(x) = x$	$t(y) = \sqrt{y}$
Curva S	$Y = \exp\left(\alpha_0 + \frac{\alpha_1}{X}\right)$	$t(x) = \frac{1}{x}$	$t(y) = \ln(y)$

Precauciones al realizar un análisis de correlación-regresión

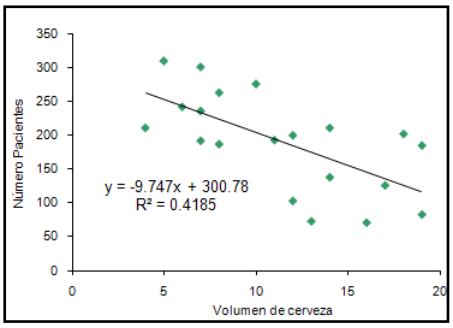
Existen múltiples razones por las cuales su análisis de correlación-regresión puede indicar que no existe relación entre las variables analizadas.

- Relación entre variables es no lineal
- El tamaño de muestra es muy pequeño
- Se omite una variable predictora

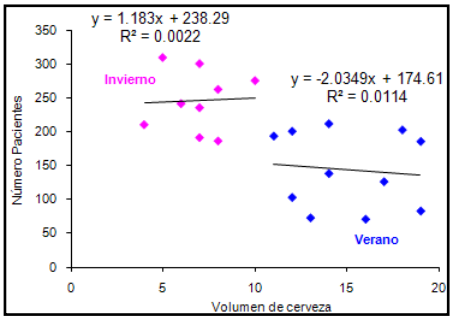
A continuación se ilustran tres ejemplos sobre errores en el análisis de los datos.

1. Relación aparente pero no real entre variables

En este caso si se analizan los datos sin considerar el efecto de estacionalidad, la ecuación de regresión indica que existe una relación entre el número de pacientes y el consumo de cerveza; sin embargo si el análisis se realiza por estación (variable predictora), la ecuación de regresión no es significativa.



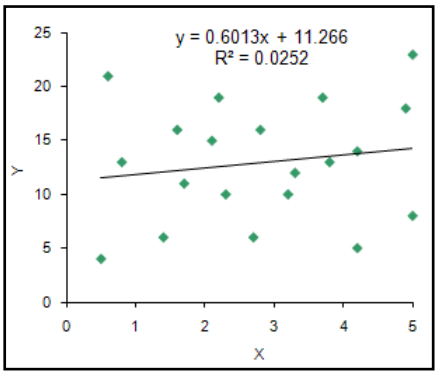
A



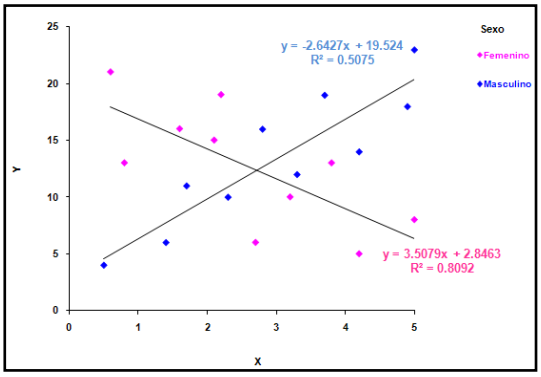
B

2. Relación no aparente pero real entre variablesl

En este caso, si se omite la variable de clasificación “sexo”, la ecuación de regresión es no significativa; sin embargo al utilizar dicha variable, se observa que la respuesta de la variable “Y” se explica por la covariación de la variable “sexo”.



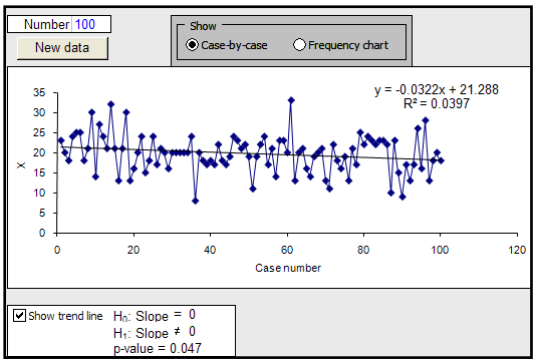
A



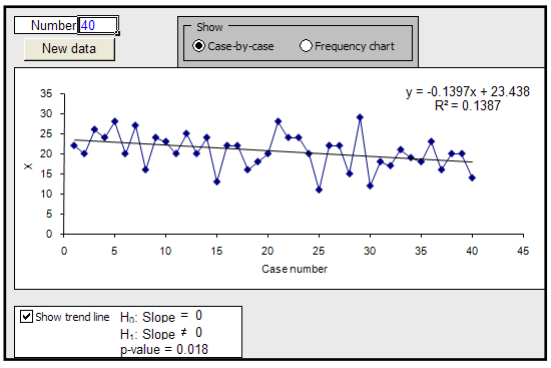
B

3. Patrones en datos al azar

Es posible obtener ecuaciones de regresión significativas aún para series estadísticas creadas al azar. Esto se ilustra en las siguientes gráficas para una muestra de 100 observaciones (A) y otra de 40 observaciones. Observe que en ambos casos la tendencia es negativa y significativa para una alfa de 5%. Esta es una de las principales dificultades para detectar tendencias cuando se analizan series de tiempo.



A



B

Ejemplo de análisis de regresión utilizando software en línea

A continuación se presentan los resultados del análisis de correlación y regresión para los mismos datos analizados con Excel pero utilizando software en línea disponible en http://www.wessa.net/.

De la página principal de dicho sitio, seleccione la opción *Descriptive Statistics Software.*

Descriptive Statistics Software: Central Tendency, Average, Mean, Median, Variability, Interquartile Range, Concentration, Lorenz Curve, Gini Coefficient, Skewness, Kurtosis, Quartiles, Percentiles, Notched Boxplot, Histogram, Correlation, Partial Correlation, Rank Correlation (Spearman and Kendall), **Simple Regression**, Kernel Density Estimation, Harrell-Davis Quantiles, Bivariate KDE, Correlation Matrix, Stem-and-leaf plot, Explorative Data Analysis

De las opciones de menú seleccione **Simple regresision** http://www.wessa.net/slr.wasp.

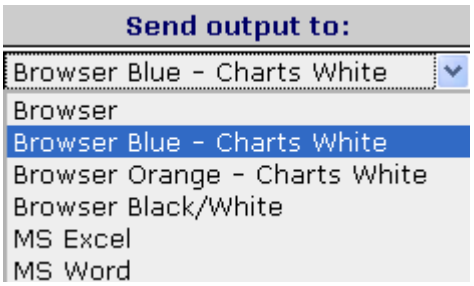
Citar como: Wessa, P. (2010), Free Statistics Software, Office for Research Development and Education, version 1.1.23-r6, URL http://www.wessa.net/

Paso 1: Copiar de la hoja de Excel los datos de la variable “Y” (Pt anual de Moravia) y pegarla en la columna “Data Y”. Etiquete el eje Y (Label y-axis) como Moravia. Asegúrese que no quede ningún espacio vacío al final de la columna de datos.

Paso 2: Copiar de la hoja de Excel los datos de la variable “X” (año) y pegarla en la columna “Data X”. Etiquete el eje Y (Label x-axis) como Año. Asegúrese que no quede ningún espacio vacío al final de la columna de datos.

Paso 3: Title: Digite Regresión Lineal Simple.

Paso 4: Enviar resultado a:



Usted puede elegir entre:

Browser: Explorador, gráficos color azul
Browser Blue-Charts White: Página del explorador, gráficos de color blanco.
Browser Orange-Charts White: Página del explorador, gráficos de color azul.
Browser Black/White: Página del explorador, gráficos de color blanco. (***Recomendada***)
MS Excel, MS Word: Hoja de Excel ó archivo de Word.

Paso 4: Haga un clic sobre *compute.*

Resultados

Enter (or paste) two different data series delimited by hard returns.

X = exogenous variable / Y = endogenous variable

Send output to:	
Browser Black/White	
Data Y:	**Data X:**
2615 2774 3059 2059 2260 2513	1974 1975 1976 1977 1978 1979
Chart options	
Width:	600
Height:	400
Title:	Regresión Lineal Simple
Label y-axis:	Moravia
Label x-axis:	Año
Compute	

Configuración de la pantalla de resultados y entrada de datos.
"Width" y "Height": Usted puede define el ancho y alto de las graficas.

Estimación de parámetros del modelo de regresión (intercepto y pendiente)

Simple Linear Regression - Ungrouped Data				
Parameter	Value	S.E.	T-STAT	Notes
Constant	-43642.097326			
Beta	23.322995	8.016451	2.909392	H0: beta = 0
Elasticity	16.685486	5.735043	2.735025	H0: elast. = 1

Estimador de parámetro β_0 (intercepto, a)

$$b_0 = \bar{y} - b_1\bar{x}$$

Estimador de parámetro β_1 (pendiente, b)

$$b_1 = \frac{C(yx)}{V(x)} = \frac{S_{yx}}{S_{xx}}$$

$$b_1 = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

Elasticidad con respecto a la media

$$\varepsilon\left(b_1\right) = b_1 \times \left(\frac{\bar{x}}{\bar{y}}\right)$$

$$b_1 = \frac{C(yx)}{V(x)}$$

$$C(yx) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right)$$

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

Tabla de análisis de varianza (ANOVA)

Simple Linear Regression - Analysis of Variance			
ANOVA	DF	Sum of Squares	Mean Square
Regression	1.000000	1.78012e+6	1.78012e+6
Residual	32.000000	6.72967e+6	210302.235495
Total	33.000000	8.50979e+6	257872.346702
F-TEST	8.464560		

SST : Sum of Squares - Total

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

$$SST = \underbrace{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}_{SST} = \underbrace{\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2}_{SSM} + \underbrace{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}_{SSE}$$

SSM : Sum of Squares - Model

$$SSM = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2$$

SSE : Sum of Squares - Error

$$SSE = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

Suma de cuadrados totales (*Total*), del modelo (*Regression*) y del error (*Residual*)

DFT : Degrees of Freedom - Total

$$DFT = n - 1$$

$$DFT = \underbrace{[n-1]}_{DFT} = \underbrace{[1]}_{DFM} + \underbrace{[n-2]}_{DFE}$$

DFM : Degrees of Freedom - Model

$$DFM = 1$$

DFE : Degrees of Freedom - Error

$$DFE = n - 1$$

Grados de libertad (DF)

MS : Mean Square

$$MS = \frac{SS}{DF} = \frac{\text{Sum of Squares}}{\text{Degrees of Freedom}}$$

Cuadrado medio

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

$$F = \frac{MSM}{MSE} = \frac{SSM/DFM}{SSE/DFE}$$

$$F \sim F_{1,n-2}$$

Prueba F (significancia del modelo)

Fuente: http://www.xycoon.com/anova.htm

Análisis de autocorrelación

Simple Linear Regression - Autocorrelation	
Statistic	Value
Durbin-Watson	2.011281
Von Neumann Ratio	2.072229
rho - Least Squares	-0.019447
rho - Maximum Likelihood	-0.019197
rho - Serial Correlation	-0.019640
rho - Goldberger	-0.019322

Estadística Descriptiva - Regresión lineal simple – Autocorrelación

$DW = \frac{\sum_{i=2}^{n} \left(e_i - e_{i-1}\right)^2}{\sum_{i=1}^{n} e_i^2} = \frac{\sum_{i=2}^{n} \left(e_i - e_{i-1}\right)^2}{SSE}$ --- $e_i = y_i - \hat{y}_i$ $e_i = y_i - b_0 - b_1 x_i$ Estadístico de Durbin-Watson	$VNR = \frac{\sum_{i=2}^{n} \left(e_i - e_{i-1}\right)^2}{\sum_{i=1}^{n} e_i^2} \times \frac{n}{n-1}$ $VNR = \frac{\sum_{i=2}^{n} \left(e_i - e_{i-1}\right)^2}{SSE} \times \frac{n}{n-1}$ --- $e_i = y_i - \hat{y}_i$ $e_i = y_i - b_0 - b_1 x_i$ Razón de Von Neumann
$\rho = \frac{\sum_{i=2}^{n} e_i \times e_{i-1}}{\sum_{i=2}^{n} e_{i-1}^2}$ --- $e_i = y_i - \hat{y}_i$ $e_i = y_i - b_0 - b_1 x_i$ Coeficiente de autocorrelación (Rho). Estimación por mínimos cuadrados (LS-estimte).	$\rho = \frac{\sum_{i=2}^{n} e_i \times e_{i-1}}{\sum_{i=1}^{n} e_i^2} \times \frac{n}{n-1}$ --- $e_i = y_i - \hat{y}_i$ $e_i = y_i - b_0 - b_1 x_i$ Autocorrelación serial de resago 1 (Rho).
$\rho = \frac{\sum_{i=2}^{n} e_i \times e_{i-1}}{\sum_{i=2}^{n} e_i^2}$ --- $e_i = y_i - \hat{y}_i$ $e_i = y_i - b_0 - b_1 x_i$ Coeficiente de autocorrelación (Rho). Estimación máximo verosimilitud.	$\rho = \frac{\sum_{i=2}^{n} e_i \times e_{i-1}}{\sqrt{\sum_{i=2}^{n} e_i^2} \sqrt{\sum_{i=2}^{n} e_{i-1}^2}}$ --- $e_i = y_i - \hat{y}_i$ $e_i = y_i - b_0 - b_1 x_i$ Coeficiente de autocorrelación (Rho) de Goldberger

Fuente: http://www.xycoon.com/lsautocorrelation.htm

Borghers,E; Wessa, P. Statistics - Econometrics - Forecasting (Descriptive Statistics - Simple Linear Regression - Autocorrelation), Office for Research Development and Education (Publisher), http://www.xycoon.com/ (URL), (visitado 11 mayo 2010). Facilities, development, and design: Office for Research, Development, and Education

Estadísticos descriptivos

Simple Linear Regression - Descriptive Statistics	
Statistic	Value
Mean X	1990.500000
Biased Variance X	96.250000
Biased S.E. X	9.810708
Mean Y	2782.323529
Biased Variance Y	250287.865917
Biased S.E. Y	500.287783
Mean F	2782.323529
Biased Variance F	52356.350157
Biased S.E. F	228.815100
Mean e	0.000000
Biased Variance e	197931.515760
Biased S.E. e	78.647059

$V(x) = s_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)$ Varianza sesgada	$V(x) = s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)$ Varianza insesgada
$HM = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$ $HM^{-1} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}$ Media armónica	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ Media aritmética
$GM = \sqrt[n]{x_1 x_2 \ldots x_n} \qquad \forall_i : x_i > 0$ $GM = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{\prod_{i=1}^{n} x_i}$ --- $\ln GM = \frac{1}{n}\left(\ln x_1 + \ln x_2 + \ldots + \ln x_n\right) = \frac{1}{n}\sum_{i=1}^{n}\ln x_i$ $GM = \exp\left[\ln GM\right] = e^{\ln GM}$ ln stands for $\log_e$. Media geométrica	$WM = \sum_{i=1}^{n}\frac{w_i}{W}x$ --- $W = \sum_{i=1}^{n} w_i$ Media ponderada

Fuente: http://www.xycoon.com/

Student Distribution Probability (mathematical equation plotter)	
T-Test	2.90939
D.F.	32
Tails (1 or 2)	2
Student Probability	

Estadístico “t”

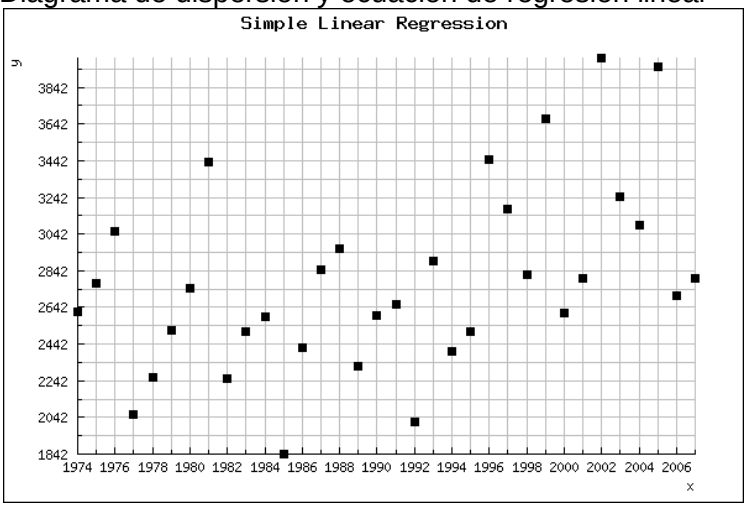
Student Distribution Probability (mathematical equation plotter)	
T-Test	2.735025
D.F.	32
Tails (1 or 2)	2
Student Probability	

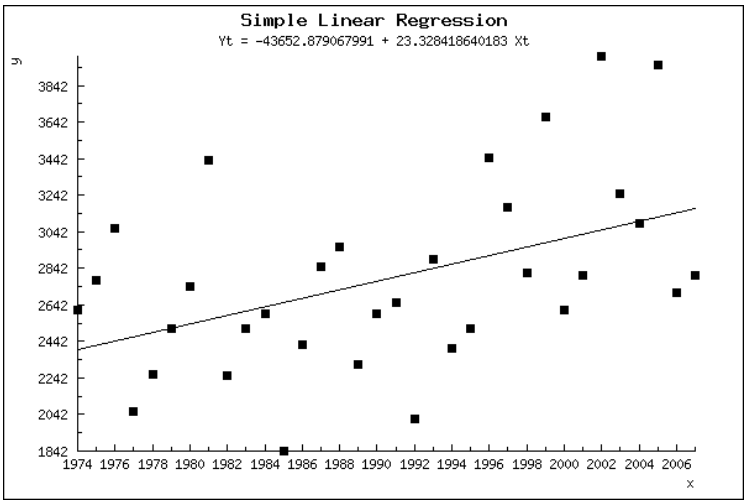
Estadístico “t”

Fisher Distribution Probability (mathematical equation plotter)	
F-Test	8.464
D.F. numerator	1
D.F. denominator	32
Fisher Probability	

Estadístico “F”

Diagrama de dispersión y ecuación de regresión lineal





Análisis de una variable numérica y otra nominal (1Num1Cat)

En la presente sección se ilustra el uso de XLSTatistics para el análisis de dos variables: una cuantitativa y otra nominal. En el menú grafico de XLStatisitics corresponde a **1Num1Cat.**

Nota: Esta sección del tutorial asume que usted ya activó el complemento de XLStatisticis (XLStatistics.xlam ó XLStats.xls); de no ser así, ver página 23.

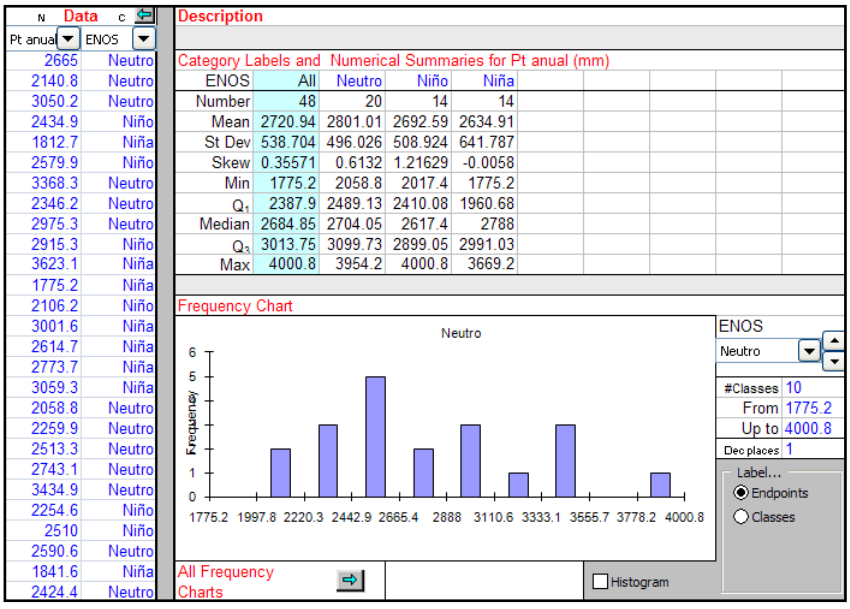
1. Abra el archivo "xlstats_tutorial.xlsx" y seleccione de la hoja de cálculo **1Num1Cat** las columnas Pt anual (mm) y ENOS, haga un clic sobre XLSTatistics y luego otro clic sobre 1Num1Cat. Observe que el programa abre el libro **1Num1Cat** y copia los datos seleccionados a dicho libro.

En la sección inferior del libro **1Num1Cat** usted observará cuatro hojas de cálculo:

1) *Information*: Información. Esta hoja contiene información general sobre la organización de este libro de trabajo.
2) *Data and Description*: Datos y estadísticos descriptivos
3) *Summaries*: Síntesis gráfica de datos
4) *Tests-n categories*: Pruebas estadísticas para n categorías
5) *Tests-2 categories*: Pruebas estadísticas para 2 categorías

Nota: Recuerde que **SOLO** debe modificar el contenido de las celdas con números o textos en color azul.

Datos y estadísticos descriptivos



Estadísticos descriptivos por episodio de ENOS (Niña, Niño, Neutro).

En la gráfica de frecuencia, usted puede elegir entre barras e histograma; así como el episodio de ENOS a graficar: Niño, Niña o Neutro.
Usted puede elegir:
de clases, límite superior e inferior de los datos, límite superior de la clase, clases (intervalo, punto medio).

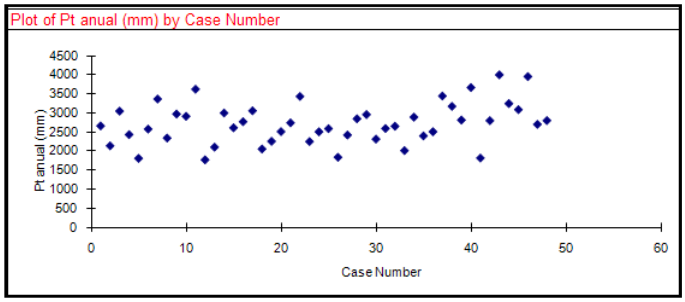
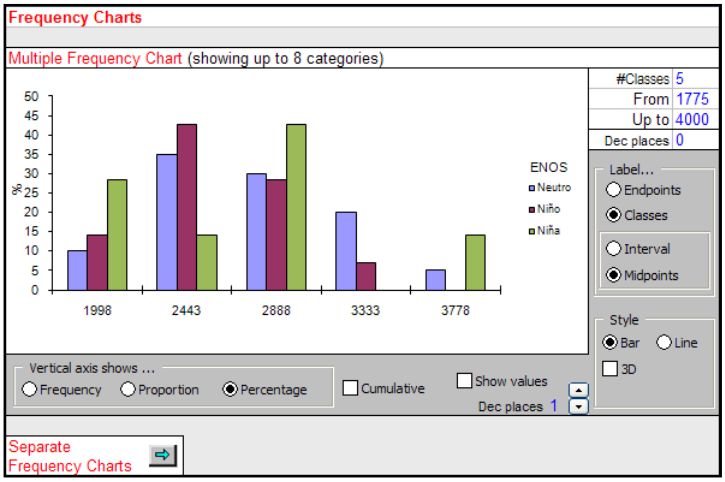


Gráfico de valores de lluvia ordenados según año.

Grafico de frecuencia comparativo por categoría



Esta hoja de cálculo muestra en el mismo grafico la distribución de frecuencia (absoluta y relativa) por valor de la variable ENOS (Niño, Niña, Neutro).

En la gráfica de frecuencia, usted puede elegir entre barras, histograma y líneas; así como la frecuencia acumulada.

Eje X
de clases, límite superior e inferior de los datos, límite superior de la clase, clases (intervalo, punto medio).

Eje Y
Frecuencia absoluta y relativa, proporciones

Tipo de gráfico: Barras, líneas, 3D

Grafico de frecuencia por categorías

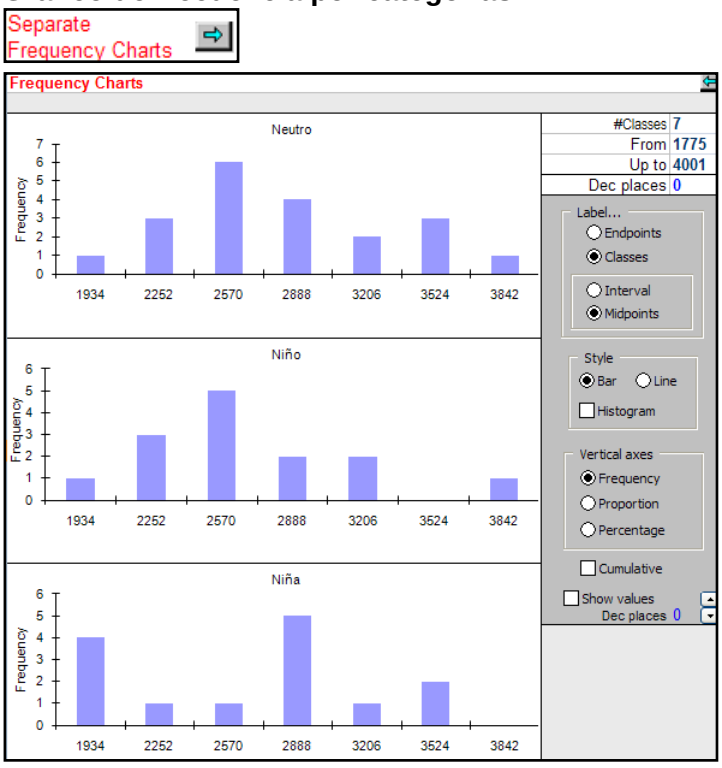


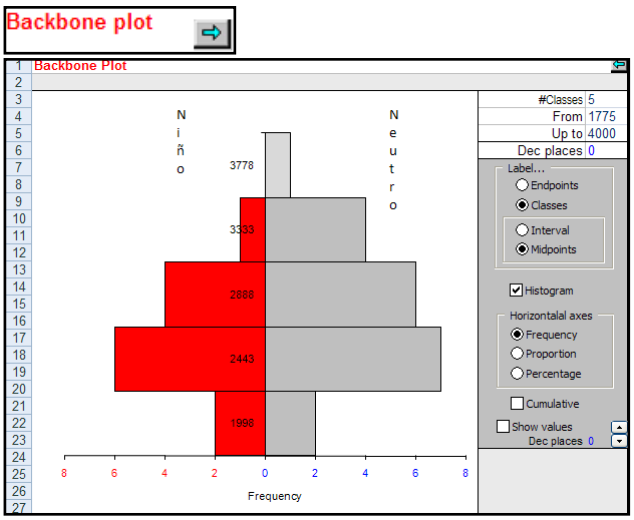
Grafico de frecuencia (absoluta, relativa, proporciones) por episodio de ENOS. En la gráfica de frecuencia, usted puede elegir entre barras, histograma y líneas; así como la frecuencia acumulada.

Eje X
de clases, límite superior e inferior de los datos, límite superior de la clase, clases (intervalo, punto medio).

Eje Y
Frecuencia absoluta y relativa, proporciones, acumulada.

Tipo de gráfico: Barras, líneas, histograma

Grafico de Columna



Nota: El grafico solo mostrará las dos primeras categorías de la variable ENOS; en este caso Niño y neutro (ver datos).

Grafico de cajas por categoría

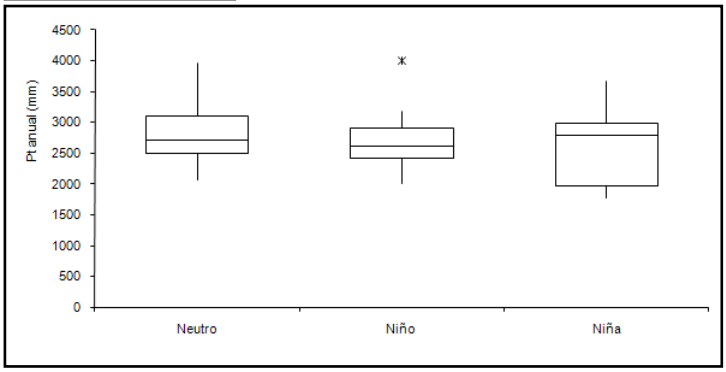
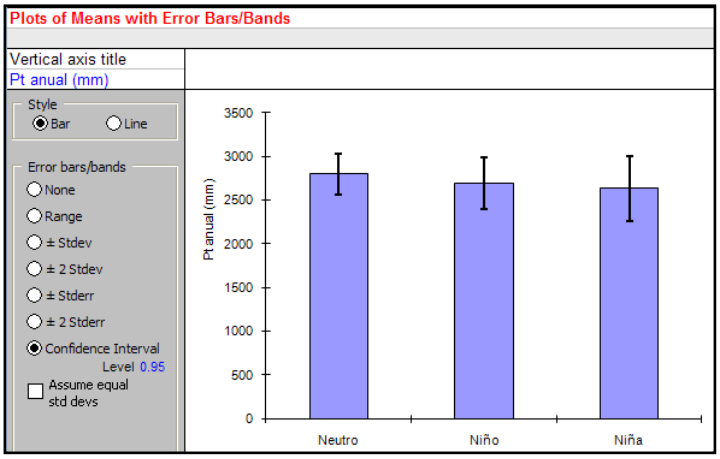


Diagrama de caja.

El asterisco (*) indica que la lluvia del año 2000 es atípico (valor muy grande para la serie estadística.

Nota: todo valor atípico debe investigarse ya que puede indicar un error en los datos ó condiciones particulares del dato.

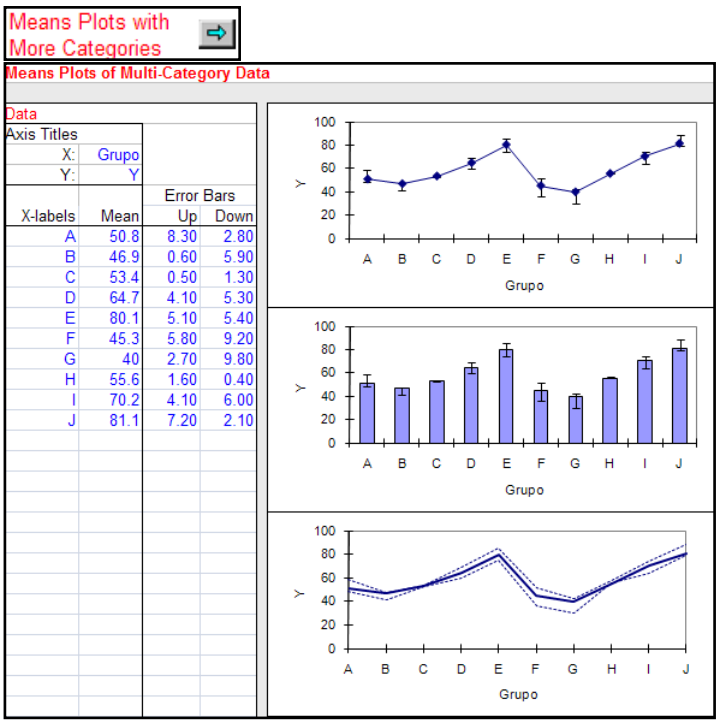
Grafico de medias con barra de error



Esta hoja de cálculo muestra un grafico de la media por valor de la variable ENOS (Niño, Niña, Neutro). Usted puede elegir entre barras ó líneas y adicionar a cada categoría una medida de variación/error:

- Rango
- Desviación estándar
- Error estándar
- Intervalo de confianza (usted define el nivel de confianza)

Grafico de medias con mas categorías



Usted debe digitar los valores correspondientes a cada clase así como sus respectivos valores de error (e.g. desviación estándar, error estándar).

Análisis de varianza paramétrico de 1 vía

El análisis de varianza-ANDEVA (ANOVA en inglés) permite comparar las medias de tres o más grupos o tratamientos. En este caso se comparan las estimaciones de la varianza entre tres o más muestras (tratamientos) y al interior de las muestras (tratamientos) para determinar mediante una prueba F sí se trata de una misma población (Ho: las muestras provienen de la misma población) o sí por el contrario provienen de poblaciones diferentes. Una vez probado que al menos un par de medias son diferentes se procede a realizar una prueba de comparaciones múltiples a los promedios para determinar cuáles de ellos son estadísticamente diferentes. Un método de análisis alternativo es realizar comparaciones planeadas; para las cuales no es necesario realizar primero una prueba F.

El análisis de varianza permite analizar el efecto de una o más variables ó categorías en un grupo experimental. Por ejemplo, podemos investigar el efecto de diferentes métodos de plantación (bolsa, raíz desnuda, plantón) en la sobrevivencia de los árboles; en este caso se trabaja con una variable predictora *método de plantación* y estamos interesados es determinar sí al menos uno de los métodos utilizados provee resultados estadísticamente significativos ó sí por el contrario el método de plantación (tratamiento) no tiene efecto en la sobrevivencia y por lo tanto podemos asumir que todos los métodos de plantación son iguales. Ahora, supongamos que utilizamos plántulas procedente de 3 viveros y que además utilizamos 2 métodos de plantación (bolsa y raíz desnuda). En este caso tenemos un experimento factorial de 2^3, formado por 6 tratamientos: 2 métodos de plantación y 3 procedencias de los árboles. En un análisis de varianza factorial cada tratamiento puede tener varias observaciones, o sea varias muestras son obtenidas bajo las mismas condiciones (e.g. 20 plántulas por tratamiento); ó por el contrario tener una única observación por tratamiento (e.g. No. de semillas que germinan por cada 100 semillas).

La hipótesis nula a probar es la siguiente: $Ho: \mu_1 = \mu_2 = \mu_3 = \mu_4 = ...$ y la hipótesis alternativa es que al menos dos de las medias sean diferentes. A diferencia de la prueba “t”, en este caso no aplica el concepto de direccionalidad en la prueba de hipótesis; ya que Ho puede ser falsa por diversas razones. Por ejemplo, dados tres tratamientos μ_1 puede ser diferente de μ_2 pero a su vez igual a μ_3.

Bajo los supuestos del diseño irrestricto al azar o completamente al azar el valor esperado de cualquier observación X_{ij} puede estimarse mediante el modelo:

$X_{ij} = \mu + \beta_i + \varepsilon_{ij} \quad \text{con } i = 1,\ldots., a; j = 1, \ldots.n; \varepsilon_{ij} = (0,\sigma^2)$, en donde:

μ : Media de la población
β_i : Efecto del tratamiento i
ε_{ij}: Residuo o error con una distribución normal $(0,\sigma^2)$

$\hat{\mu} = \bar{\bar{X}}..$ Estimación de la media poblacional.

$\hat{\beta}_i = (\bar{\bar{X}}_i - \bar{\bar{X}}..)$ Estimación de β_i, variación explicada por el tratamiento i.

$\hat{\varepsilon}_{ij} = X_{ij} - \hat{\beta}_i - \hat{\mu}$ Estimación del error o residuo, variabilidad no explicada por los tratamientos.

La tabla de análisis de varianza de una vía contiene los siguientes elementos para el modelo de efectos fijos (modelo tipo I):

Fuente de variación	Gl	SC	CM	F	CM esperado
Total	N-1	SC_T	----------	----------	---------
Entre (K)	k-1	SC_k	$SC_k / k-1$	CM_k / CM_E	$\sigma^2 + n\sigma^2_k$
Dentro	N-k	SC_w	$SC_w / N-k$		σ^2

SC_T: suma de cuadrados total
SC_k: suma de cuadrados de tratamientosl
SC_k: suma de cuadrados al interior de los tratamientos
CM: Cuadrado medio
k: número de tratamientos
N: Total de observaciones (tamaño de muestra)
F: Prueba F

En un modelo de efectos fijos o modelo “I” se eligen todos los posibles niveles de interés de la variable nominal y asume que los datos provienen de poblaciones normales, las cuales podrían diferir únicamente en sus medias. En el modelo “II” o de efectos aleatorios solo se eligen algunos de los posibles niveles de interés de la variable nominal y por esta razón la partición de varianza es de mayor interés que determinar qué grupos son diferentes.

Supuestos del ANOVA

Independencia: Errores independientes al interior de los tratamientos y entre tratamientos. La correlación positiva incrementa el valor del error estándar y por tanto, la estimación de las medias para los tratamientos es más exacta que lo indicado por los errores estándares.

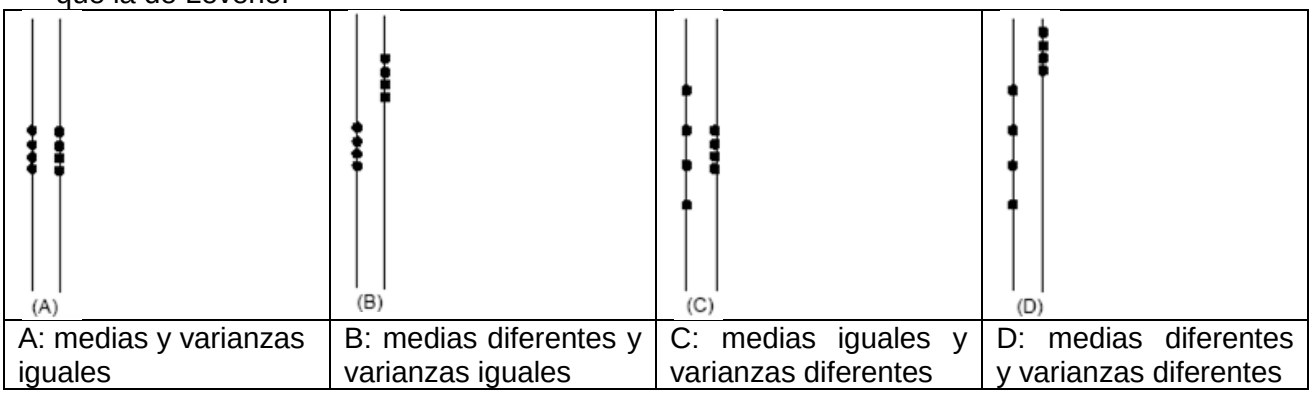
Pruebas: autocorrelación entre errores, Prueba de Durbin-Watson (autocorrelación de resago 1).

Normalidad. La dependiente o respuesta debe ser normal. Este es un requisito para las pruebas estadísticas. Sin embargo, la prueba F (razón de dos varianzas) es muy robusta a datos no normales, especialmente en modelos de *efectos fijos* o cuando el *tamaño de muestra es superior a 50*. Cuando se viola seriamente este supuesto los resultados de las pruebas estadísticas son muy laxas (mayor tendencia a rechazar Ho) y decrece el poder y la eficiencia de la prueba. Si sus datos son no normales puede transformarlos, utilizar un equivalente no paramétrico (Kruskal-Wallis ó Friedman) ó utilizar una prueba F basada en aleatorización.

Pruebas: Shapiro-Wilk's W (compara el cociente de la desviación estándar con la varianza multiplicada por una constante con el valro uno), pruebas de bondad de ajuste: Kolmogorov-Smirnov D, Cramer-von Mises W^2 y Anderson-Darling A^2, gráficos de distribución normal, histograma, gráfico de cajas.

Homogeneidad de varianzas. El análisis de varianza se basa en la partición de la variabilidad total asociada al set de datos (SCtotal) en dos *estimaciones independientes* de varianza: la varianza dentro de los grupos experimentales y la varianza entre los grupos experimentales. Dado que Ho sea verdadera, ambas estimaciones de varianza son una estimación de la varianza poblacional σ^2 y por lo tanto, la razón de varianzas debería ser cercana a uno. La prueba F es muy robusta a la violación de este supuesto, especialmente para el modelo de *efectos fijos* y para diseños balanceados (igual tamaño de muestra por tratamiento); sin embargo los contrastes entre diferencias para tratamientos sí son afectados (resultados no son muy confiables).

Pruebas: Para probar por la igualdad de varianzas pude utilizar la gráfica de cajas para cada una de los tratamientos ó la gráfica de desviación estándar por tratamiento (la varianza de cada media debe ser equivalente). También puede utilizar la prueba de *Levene*, la cual calcula una ANOVA de 1 vía utilizando el valor absoluto (o algunas veces al cuadrado) de los residuos, |yij - Yi| con t-1, N - t grados de libertad); esta es una prueba robusta cuando se utiliza con datos no normales pero es demasiado conservadora. Si sus datos son normales o casi normales, se recomienda utilizar la prueba de *Bartlett*, ya que, bajo estas condiciones, es más poderosa (capacidad para detectar varianzas desiguales cuando las mismas son realmente diferentes) que la de Levene.



A: medias y varianzas iguales	B: medias diferentes y varianzas iguales	C: medias iguales y varianzas diferentes	D: medias diferentes y varianzas diferentes

Figura 1: Ejemplo de muestras con varianzas iguales y diferentes.

Los supuestos de normalidad e igualdad de varianzas son críticos para realizar las pruebas de hipótesis; sin embargo el análisis de varianza es robusto a dichas violaciones; especialmente conforme aumenta el tamaño de la muestra. En la práctica se pueden aceptar violaciones a dichos supuestos sin invalidad los resultados. También es posible transformar los datos (e.g. raíz cuadrada, logaritmos, recíproca, angular) para hacerlos compatibles con los supuestos del análisis de varianza; sin embargo dichas transformaciones afectarán la magnitud de las interacciones. Otra alternativa es utilizar pruebas no paramétricas (Kruskal-Wallis ó Friedman) ó utilizar una prueba F basada en aleatorización como la que realiza el programa "resampling" http://www.uvm.edu/~dhowell/StatPages/Resampling/Resampling.html).

Análisis de varianza paramétrico de 1 vía con XLSTatistics

Analysis of Variance (Comparison of Means of Pt anual

Independent variable (ENOS) is a
(●) Fixed effect ( ) Random effect

H_0: All population means (of Pt anual (mm)) are equal
H_1: Not all population means (of Pt anual (mm)) are equal
p-value = 0.66721

ANOVA Table

Source	DF	SS	MS	F
ENOS	2	243103	121551	0.4083
Error	45	1.3E+07	297698	
Total	47	1.4E+07		

Test for Intercept Term ⇨ | Residuals Analysis ⇨

Análisis de varianza paramétrico
Usted puede elegir entre efectos fijos y aleatorios.

ENOS tiene solo tres clases y por tanto se eligió “efectos fijos”.

La prueba de hipótesis plantea que la precipitación media es independiente de la fase de ENOS, o sea que las tres medias son iguales. Para un nivel de significancia de 0.05, Ho no se rechaza dado el valor de es p=0.408.

Prueba sobre intercepto en modelo

Test for the Intercept Term in the Model

H_0: Intercept Term in model = 0
H_1: Intercept Term in model not = 0
p-value = 4.85E-34

ANOVA Table

Source	DF	SS	MS	F	p-value
Intercept	1	3.55E+08	3.55E+08	1193.722	4.85E-34
ENOS	2	243102.8	121551.4	0.408304	0.667214
Error	45	13396407	297697.9		
Total	48	3.69E+08			

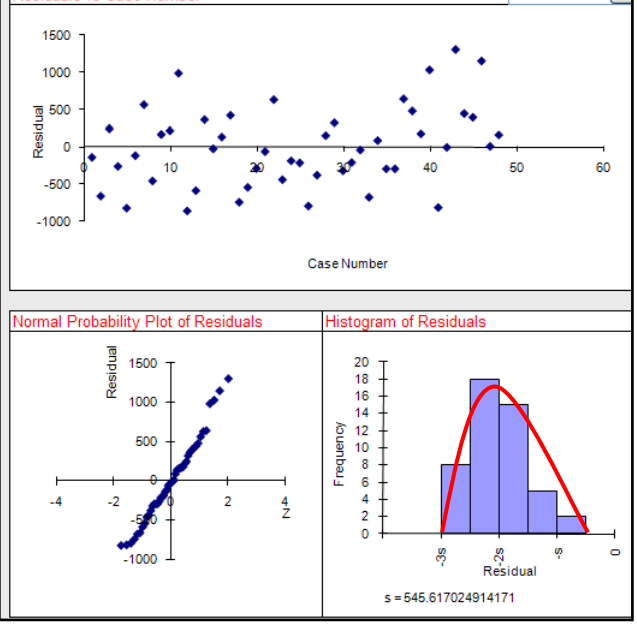
La tabla de análisis de varianza indica que el intercepto es diferente de cero (p= 4.85E-34) en tanto que el factor ENOS no es significativo (p= 0.667).

Supuestos: distribución normal y varianzas homogéneas.

Análisis de residuos

Residuals Analysis

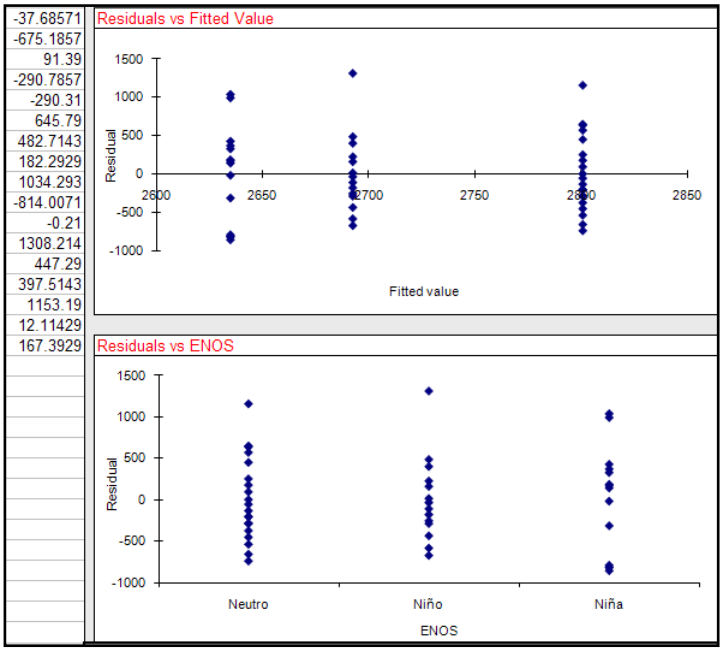
Residuals
-136.01
-660.21
249.19
-257.6857
-822.2071
-112.6857
567.29
-454.81
174.29
222.7143
988.1929
-859.7071
-586.3857
366.6929
-20.20714
138.7929
424.3929
-742.21
-541.11
-287.71
-57.91
633.89
-437.9857
-182.5857
-210.41
-793.3071
-376.61
157.7143
324.3929
-317.0071



Los residuos se utilizan para evaluar los siguientes supuestos del modelo de regresión utilizado en el análisis de varianza.

Distribución de variable respuesta es normal (en este caso la precipitación anual).

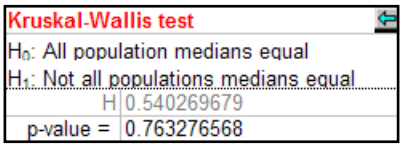
El histograma y el gráfico de probabilidad normal permiten concluir que los datos cumplen con el supuesto de normalidad.



Residuos Vs valores ajustados. Homocedasticidad. La varianza de la variable respuesta debe ser la misma para los diferentes grupos (fases de ENOS). La gráfica de residuos vs datos estimados ($\hat{y}_i$) permite evaluar este supuesto.

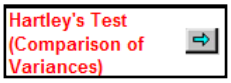
Residuos Vs variable predictora (X, ENOS). El gráfico indica que la dispersión (varianza) para cada una de las fases de ENOS entre las permite deducir si la heterocedasticidad o la ausencia de linealidad en el modelo son debidas a la variable explicativa o predictora (Neutro, Niño, Niña).

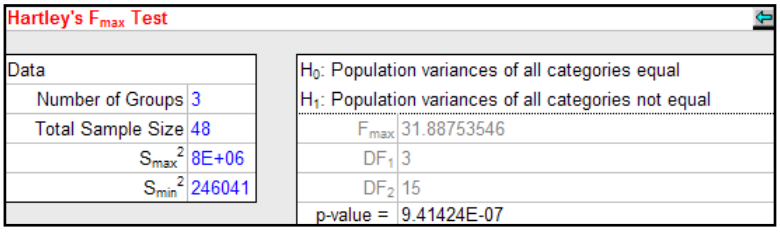
Análisis de varianza no paramétrico de 1 vía (Kruskal-Wallis)



La prueba de hipótesis plantea que la precipitación mediana es independiente de la fase de ENOS, o sea que las tres medianas son iguales. Ho no se rechaza dado el valor de p=0.763.

Prueba de Hartley (comparación de varianzas)





La prueba indica que no todas las varianzas son iguales (p=9.414E-07); o sea los datos no cumplen con el requisito de homogeneidad de varianzas. Sin embargo las pruebas de Bartlett y Levene indican que las varianzas son iguales (ver siguiente sección).

Hipótesis nula: todas las σ^2_j son iguales

Hipótesis alternativa: no todas las σ^2_j son iguales

La prueba de Hartley, conocida también como prueba F_{max} o prueba F_{max} de Hartley, calcula la proporción de la varianza más grande ($\max(s_j^2)$ con respecto a la varianza más pequeña ($\min(s_j^2)$ y compara el resultado con los valores de una tabla de valores críticos tabla de F_{max}. Esta prueba requiere que los datos sean normales y que el número de observaciones por nivel de tratamiento sea el mismo.

El estadístico de Bartlett también es utilizado para probar por la homogeneidad de varianzas entre grupos y puede utilizarse tanto con grupos con igual número de observaciones como con grupos con diferente número de observaciones. Otra prueba alternativa es el estadístico de Levene (ver siguiente sección).

Análisis de varianza de una vía en línea con StatGraphics

http://www.statgraphicsonline.com/SGOnline.aspx;
http://www.statgraphicsonline.com/OnewayANOVA.aspx)

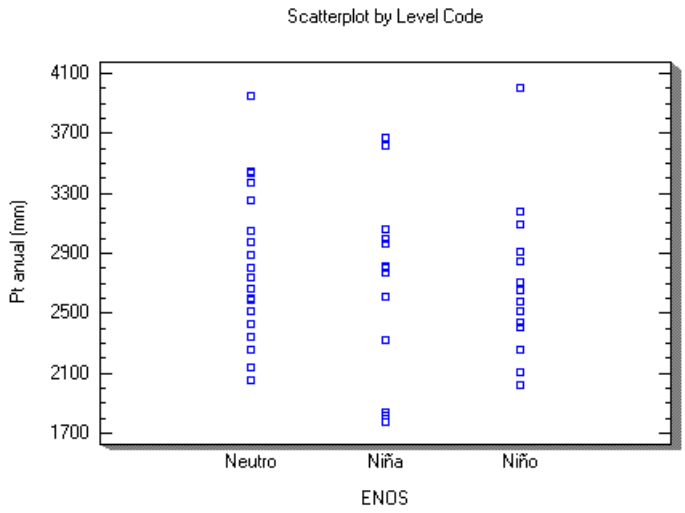
Para utilizar esta versión en línea de StatGraphics usted debe registrarse (es gratuito). Los datos utilizados se encuentran en el archivo anova_statgraphics.xls y corresponden a los analizados con XLSTatistics.

Esta barra de menú le permite configurar las opciones que ofrece el programa.

- ✓ *Data Input*: Insumo de datos (incluye archivos de Excel)
- ✓ *Tables and Graphs*: Tablas y gráficos que desea como parte de la ANOVA
- ✓ *Results to Save*: Resultados que desea guardar
- ✓ *Preferences*: Preferencias de fuentes, color de gráficos (fondo y primer plano), tipo de puntos y líneas.

Una vez leídos los datos y configurado las opciones de análisis que desea haga un clic sobre *Calculate*.

Gráfico de dispersión



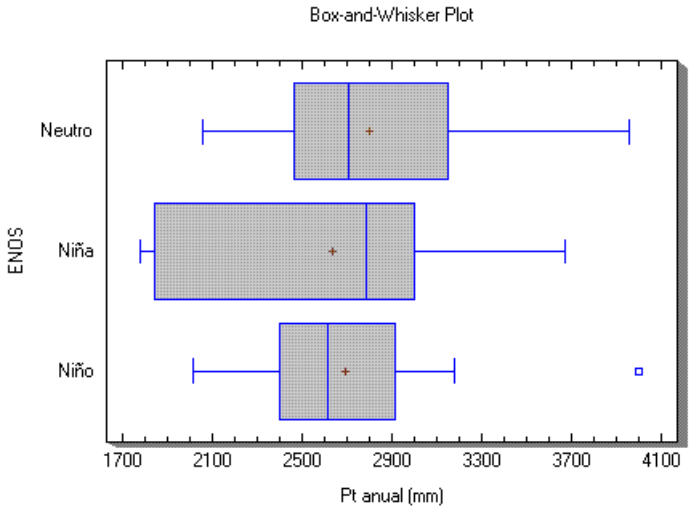
Este gráfico muestra la precipitación (mm) por episodio de ENOS. Observe que la dispersión entre grupos es similar; lo cual es un indicador de la igualdad de varianzas entre grupos; esto se confirma al observar el coeficiente de variación (entre 17 y 24%).

Summary Statistics for Pt anual (mm) (Estadísticos descripivos)

ENOS	Count	Average	Standard deviation	Coeff. of variation	Minimum	Maximum	Range
Neutro	20	2801.01	496.026	17.71%	2058.8	3954.2	1895.4
Niña	14	2634.91	641.787	24.36%	1775.2	3669.2	1894
Niño	14	2692.59	508.924	18.90%	2017.4	4000.8	1983.4
Total	48	2720.94	538.704	19.80%	1775.2	4000.8	2225.6

ENOS	Stnd. skewness	Stnd. kurtosis
Neutro	1.11955	-0.0581878
Niña	-0.0088079	-0.724125
Niño	1.85791	1.78562
Total	1.00611	-0.0788969

Gráfico de Box-Whisker



El signo "+" indica la ubicación de la media para cada grupo. Los cuadrados pequeños indican valores atípicos que se encuentran a más de 1.5 veces el rango intercuartil por encima o por debajo del primer (Q1) ó del tercer cuartil (Q3), respectivamente.

ANOVA Table (Tabla de ANOVA)

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Between groups	243103	2	121551	0.41	0.6672
Within groups	1.33964E+07	45	297698		
Total (Corr.)	1.36395E+07	47			

La tabla de ANOVA descompone la varianza total de Pt annual (mm) en dos fuentes de variación: variación entre grupos y variación dentro de grupos. El valor del estadístico F (0.408304), es el cociente entre la variación entre grupos y la variación dentro de grupos. Para un nivel de significancia de 5% no existe una diferencia significativa entre las medias de precipitación por episodios de ENOS ya que el valor de "p" es 0.667.

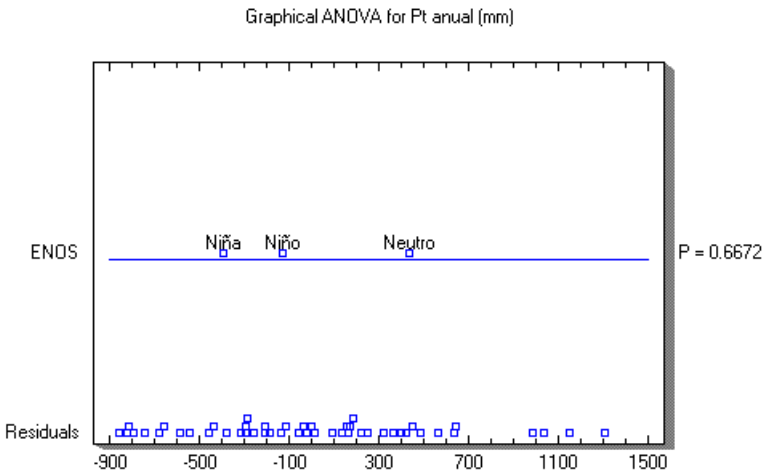
ANOVA gráfica

La representación gráfica de ANOVA se utiliza para mostrar gráficamente la importancia de las diferencias entre los niveles del factor experimental (en este caso ENOS). Se trata de una grafica de los efectos del factor de escala, donde el "efecto" es igual a la diferencia entre la media para un nivel de dicho factor y la media global. Cada uno de los efectos es multiplicado por un factor de escala dado por:

$$\sqrt{\frac{\nu_R n_i}{\nu_T \overline{n}}}$$

donde v_R son los grados de libertad del residuo, v_T son los grados de libertad para el factor, n_i es igual al número de observaciones en el nivel i-ésimo del factor, y $\overline{n}$ es el número medio de observaciones para todos los niveles del factor. Estos valores escalan los efectos de manera que

la variación natural de los puntos en el diagrama es comparable a la de los residuos, que se muestran en la parte inferior del gráfico.

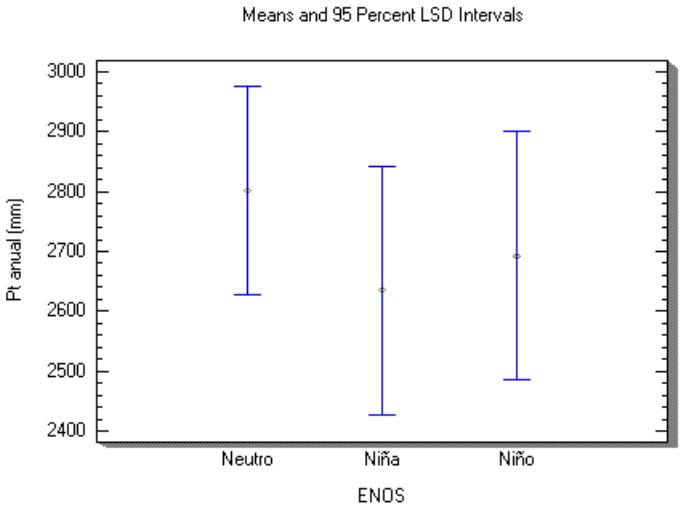


ANOVA Gráfica: El valor de "p" es obtenido de la tabla de análisis de varianza.

Para un nivel de significancia de 5% no existe una diferencia significativa entre las medias de precipitación por episodios de ENOS ya que el valor de "p" es 0.667.

La representación gráfica de ANOVA muestra los efectos de los grupos reescalados para que puedan compararse con la variabilidad de los residuales. La gráfica muestra las desviaciones de la media de cada grupo con respecto a la media global. Si las diferencias entre grupos son grandes en relación con la variabilidad de los residuos, entonces es posible que las mismas sean importantes (estadísticamente significativas).

Gráfico de media anual de Pt (mm) por episodio de ENOS e intervalo LSD



Esta gráfica muestra la media anual de Pt (mm) por episodio de ENOS. Los intervalos para la media que se muestran se basan en la diferencia mínima significativa de Fisher (LSD) y están construidos de tal manera que si dos medias son iguales, sus intervalos se solaparán 95% del tiempo. Los pares de intervalos que no se superponen verticalmente corresponden a medias estadísticamente diferentes.

$$\bar{Y}_j \pm \frac{\sqrt{2}M}{2}\sqrt{\frac{MS_{within}}{n_j}}$$

Fórmula utilizada para calcular los intervalos. Donde M corresponde a la pruebas de rango múltiple seleccionada (en este caso LSD).

Multiple Range Tests for Pt anual (mm) by ENOS (Comparaciones múlitiples)

Esta tabla muestra el resultado de aplicar el método LSD (*Diferencia mínima significativa de Fisher*) para determinar si alguno de los pares de comparaciones es significativo. Normalmente esta prueba se aplica cuando el ANOVA indica que existen diferencias significativas entre las medias de los grupos.

Method: 95 percent LSD (*Diferencia mínima significativa de Fisher)*

ENOS	Count	Mean	Homogeneous Groups
Niña	14	2634.91	X
Niño	14	2692.59	X
Neutro	20	2801.01	X

La X indica que no existe una diferencia significativa entre ninguna de las medias comparadas como se aprecia en la siguiente tabla:

Contrast	Sig.	Difference	+/- Limits
Neutro-Niña		166.103	382.94
Neutro-Niño		108.424	382.94
Niña-Niño		-57.6786	415.36

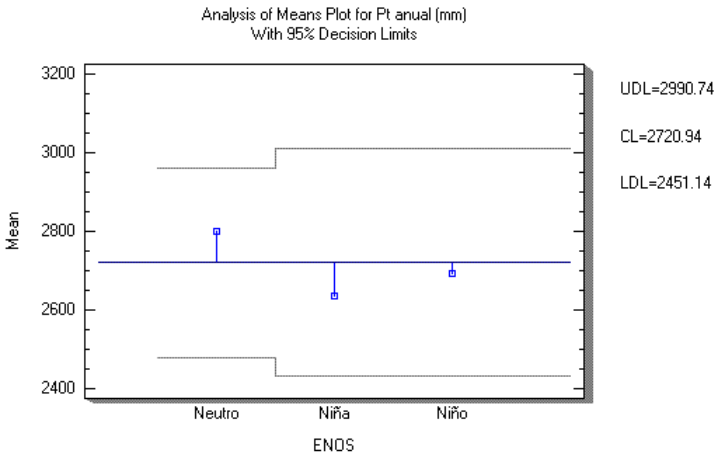
* *Sig*: indica diferencia estadísticamente significativa
Difference: diferencia entre medias
+/- Limits: límites para el contraste dado por:

$$\hat{\Delta}_{j_1 j_2} \pm M\sqrt{MS_{within}\left(\frac{1}{n_{j_1}}+\frac{1}{n_{j_2}}\right)}$$

donde: M es una constante que depende del procedimiento seleccionado (En el presente ejemplo LSD de Fisher).

Según la diferencia mínima significativa de Fisher y para un nivel de significancia de 5%, no se encontró ninguna diferencia significativa entre pares de medias; o sea no existe diferencia estadística entre la precipitación media por episodio de ENOS. Para esta prueba el error tipo I es de 5%.

Analysis of Means (ANOM) Plot (Gráfico de medias)



Este gráfico muestra la media de cada una de las tres muestras (episodios de ENOS). También se muestra la gran media y los límites de decisión al 95%. Dado que ninguna de las medias queda fuera de los límites de decisión, no hay diferencia significativa entre ellas para un alfa de 0.05.

$$\bar{Y} \pm h_{n-q,1-\alpha}\sqrt{\frac{MS_{within}}{n_j}\left(\frac{q-1}{q}\right)}$$ Fórmula utilizada para calcular los límites de decisión. Donde “h” es un valor crítico obtenido de una tabla de la distribución t multivariada.

Variance Check (Prueba de homogeneidad de varianzas de Levene)

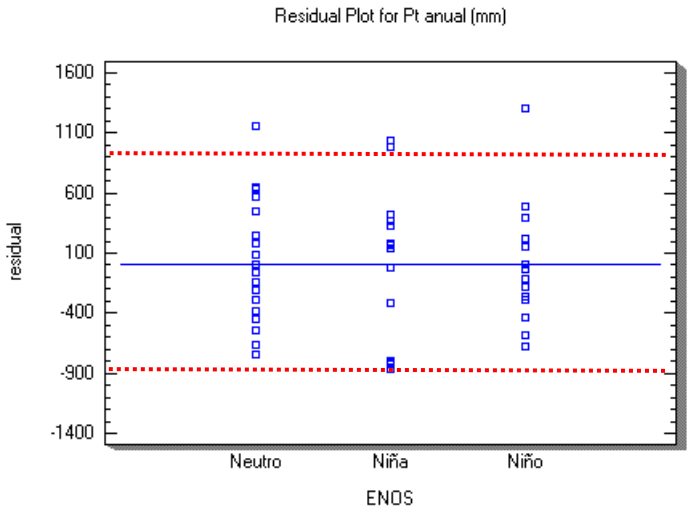
	Test	P-Value
Levene's	0.586176	0.560643

El estadístico de Levene somete a prueba la siguiente hipótesis nula: la desviación estándar de la variable Pt anual (mm) es la misma para los 3 episodios de ENOS. Para un alfa de 0.05, no existe una diferencia estadísticamente significativa entre las desviaciones estándares entre los episodios de ENOS (p=0.56) y por tanto se concluye que existe homogeneidad de varianzas.

Comparison	Sigma1	Sigma2	F-Ratio	P-Value
Neutro / Niña	496.026	641.787	0.597347	0.2988
Neutro / Niño	496.026	508.924	0.949954	0.8954
Niña / Niño	641.787	508.924	1.59029	0.4140

Esta tabla muestra la comparación de las desviaciones estándares para cada par de muestras. Para un alfa de 0.05, no existe ninguna diferencia estadísticamente significativa entre las desviaciones estándares comparadas (p>=0.05).

Residual Plots (Gráfico de residuos)



Este gráfico muestra los residuos para cada uno de los episodios de ENOS. Los residuos son iguales a los valores observados de Pt anual (mm) menos la media anual de Pt (mm) del respectivo grupo. Observe que la dispersión entre grupos es similar; lo cual es un indicador de la igualdad de varianzas entre grupos.

Kruskal-Wallis Test for Pt anual (mm) by ENOS (Prueba no paramétrica)

ENOS	Sample Size	Average Rank
Neutro	20	21.85
Niña	14	20.6429
Niño	14	32.1429

Test statistic = 5.95164 P-Value = 0.0510056.

La prueba de Kruskal-Wallis evalúa la hipótesis nula de que las medianas de Pt anual (mm) para cada uno de los tres episodios de ENOS es la misma. Para un nivel de significancia de 5%, dado que el valor “p” es mayor o igual a 0,05, no existe una diferencia estadísticamente significativa entre las medianas de los episodios de ENOS. Sin embargo, observe que el valor de “p” es muy cercano al “p” crítico y posiblemente con un tamaño de muestra mayor el resultado sería significativo al menos para la diferencia Niño-Niña.

Mood's Median Test for Pt anual (mm) by ENOS (Prueba no paramétrica)
Total n = 48
Grand median = 2684.85

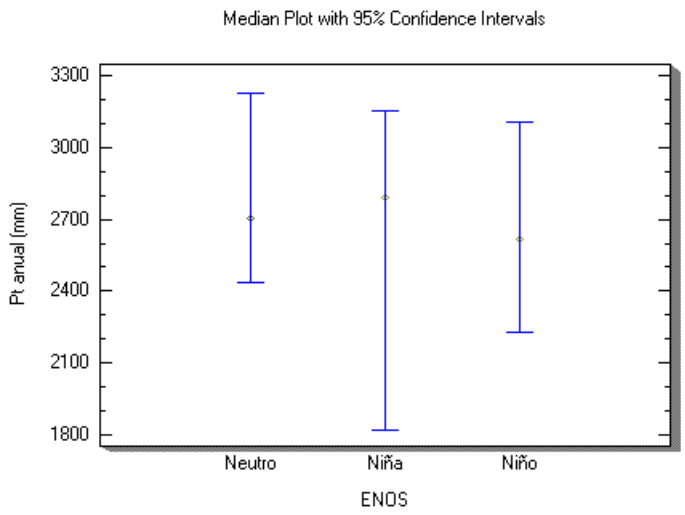
ENOS	Sample Size	n<=	n>	Median	95% lower CL	95% upper CL
Neutro	20	10	10	2704.05	2434.45	3225.23
Niña	14	6	8	2788	1819.54	3152.85
Niño	14	8	6	2617.4	2229.98	3104.24

Test statistic = 0.571429 P-Value = 0.751477

El estadístico de Mood somete a prueba la siguiente hipótesis: las medianas de las tres muestras son iguales. Dado que el valor “p” para la prueba de chi-cuadrado es mayor o igual a 0,05, las medianas de las muestras no son significativamente diferentes al nivel de confianza del 95%. También se incluyen los intervalos de confianza al 95% para cada mediana. Esta prueba es menos sensible a los valores atípicos que la prueba de Kruskal-Wallace, pero también es menos potente cuando los datos proceden de una distribución normal.

Nota: Observe que la conclusión (no diferencia entre Pt media o mediana por episodio de ENOS) es consistente en todas las pruebas realizadas.

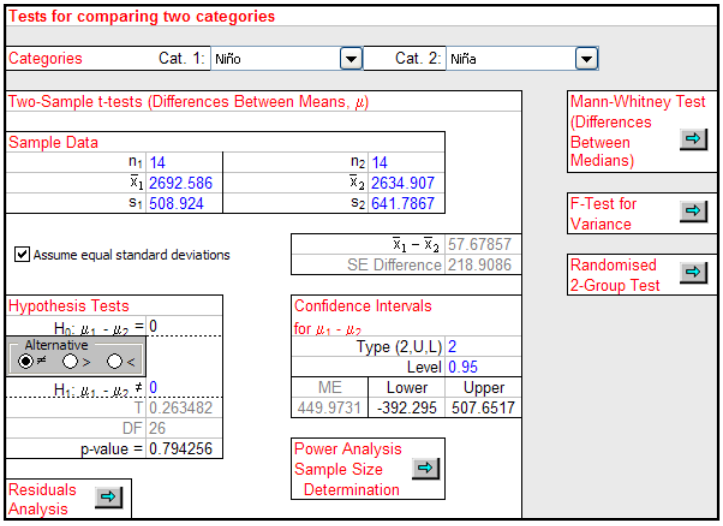
Grafico de medianas



Este gráfico muestra la mediana de cada grupo con su respectivo intervalo de confianza al 95%. Los pares de intervalos que no se superponen verticalmente corresponden a medianas estadísticamente diferentes. Observe que este caso todos los intervalos se traslapan y por tanto todas las medianas son iguales.

Comparación de dos grupos (2 muestras independientes)

El análisis de varianza se utiliza cuando se desea saber si la media de tres o más grupos es diferente; sin embargo dado que usted rechace Ho solo sabe que al menos un par de medias es diferente. No es recomendable llevar a cabo todas las posibles pruebas una vez realizado el ANOVA ya que eso aumenta su error tipo I. Usted debe planear las pruebas de hipótesis que desea realizar en su fase de diseño de la investigación y en todo caso antes de realizar el ANOVA. Sin embargo, si por alguna razón imprevista usted debe realizar pruebas no planeadas, una forma de controlar el *error tipo I* (declarar falsas significancias) es utilizar la corrección de Bonferroni (http://mathworld.wolfram.com/BonferroniCorrection.html), la cual consiste en dividir para cada prueba el nivel de significancia original entre el número de pruebas que está realizando. Por ejemplo, si realiza 5 pruebas y desea utilizar un nivel de significancia de 5%, podría realizar cada prueba con un nivel de significancia de 0.05/5 ≈ 0,01.



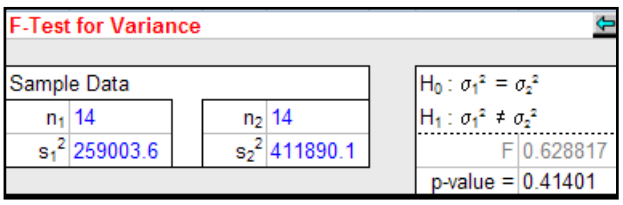
Esta libro de cálculo le permite realizar una prueba "t" de Estudiante para dos categorías; en este caso años Niño Vs año Niña.

Se eligió varianzas iguale porque la prueba F así lo indica (ver siguiente sección). Para un nivel de significancia de 5%, el valor de "p" (0.794) indica que no existe una diferencia significativa entre la precipitación anual de los años Niña y Niño. El intervalo de confianza al 95% para la diferencia entre medias contiene el valor cero y por tanto la misma no es diferente de cero.

Nota: Estos sitios le permiten realizar un análisis de poder y determinar el tamaño de muestra para la prueba "t" (independiente y dependiente o pareada).
http://www.quantitativeskills.com/sisa/calculations/power.htm
http://www.quantitativeskills.com/sisa/calculations/samsize.htm

Weisstein, Eric W. "Bonferroni Correction." From MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/BonferroniCorrection.html
http://www.itl.nist.gov/div898/handbook/prc/section4/prc473.htm

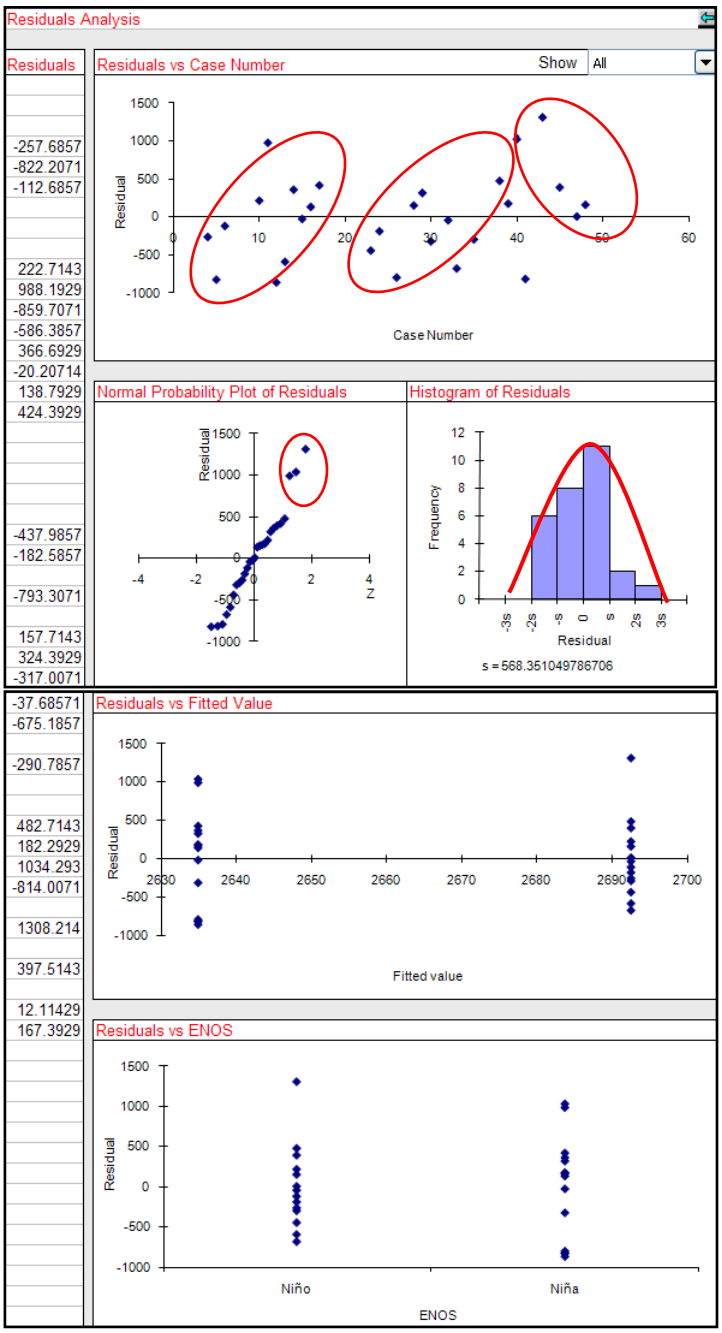
Prueba F de igualdad de varianzas (Niño Vs Niña)



Para un nivel de significancia de 5%, la prueba F indica que las varianzas son iguales (p=0.414).

Nota: ¿Son las varianzas iguales entre Niño-Neutro y Niña-Neutro?

Análisis de residuos



El análisis de residuos tiene como objetivo verificar los supuestos de la prueba de hipótesis:
Normalidad e Igualdad de varianzas.

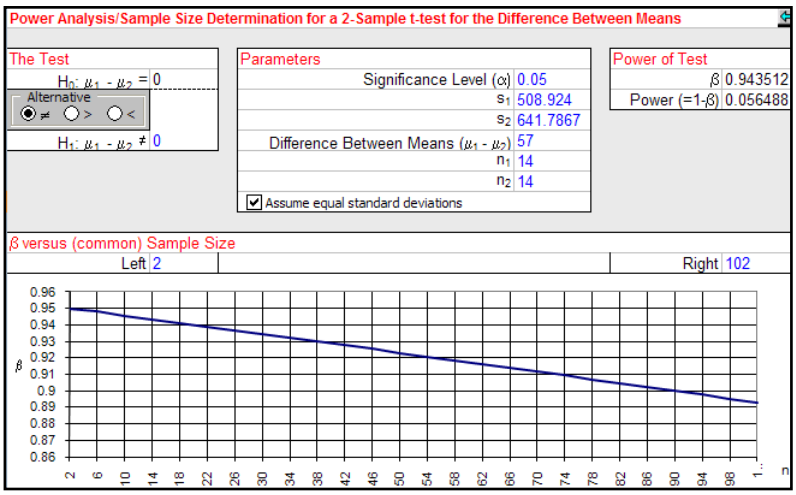
Cuando los datos son graficados en orden cronológico (como en este caso), cualquier patrón en los mismos podría indicar una influencia externa.

Observe los residuos en el círculo, parece que se apartan del patrón lineal de los otros residuos. En un caso de análisis real de datos, dichos valores deberían investigarse.

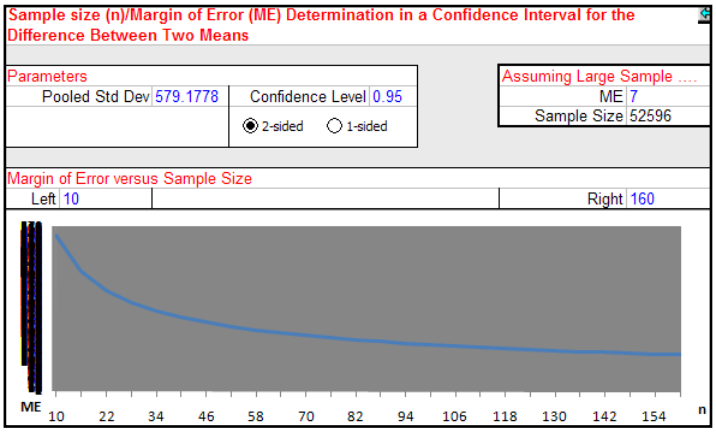
Residuos Vs valores ajustados
Este gráfico es útil para determinar si las varianzas son diferentes entre los grupos. Cuando esto ocurre se viola el supuesto de homogeneidad de varianzas. Si sus datos no cumplen con este supuesto debe transformarlos o utilizar una prueba no paramétrica.

Residuos Vs variable predictora
Este gráfico es útil para evaluar la variabilidad entre grupos. Cuando esto ocurre se viola el supuesto de homogeneidad de varianzas.

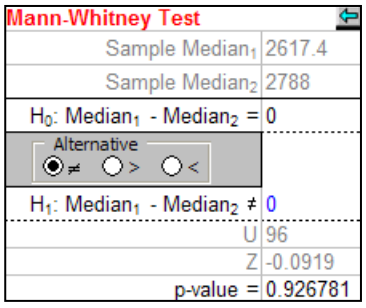
Análisis de poder de la prueba t



Análisis de poder marginal de la prueba t



Prueba de Mann-Witney (diferencia entre medianas de 2 muestras independientes)

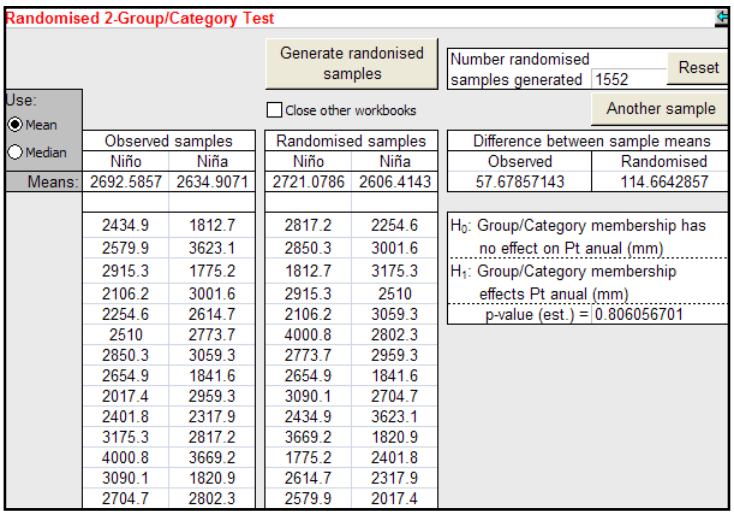


Prueba no paramétrica utilizada para evaluar diferencia entre dos medianas obtenidas de muestras independientes.

Para un nivel de significancia de 5%, el valor de "p" (0.92) indica que no existe una diferencia significativa entre la precipitación mediana anual de los años Niña y Niño.

Nota: Usted puede someter a prueba los otros contrastes: Niña-Neutro; Niña-Niño. Recuerde el efecto de las comparaciones múltiples sobre el valor de alfa (ver pág. 78).

Prueba de hipótesis entre dos medias utilizando el método de aleatorización



Randomised 2-Group/Category Test

Generate randonised samples | Number randomised samples generated 1552 | Reset | Another sample | Close other workbooks

Use: (•) Mean ( ) Median

	Observed samples		Randomised samples		Difference between sample means	
	Niño	Niña	Niño	Niña	Observed	Randomised
Means:	2692.5857	2634.9071	2721.0786	2606.4143	57.67857143	114.6642857
	2434.9	1812.7	2817.2	2254.6		
	2579.9	3623.1	2850.3	3001.6		
	2915.3	1775.2	1812.7	3175.3		
	2106.2	3001.6	2915.3	2510		
	2254.6	2614.7	2106.2	3059.3		
	2510	2773.7	4000.8	2802.3		
	2850.3	3059.3	2773.7	2959.3		
	2654.9	1841.6	2654.9	1841.6		
	2017.4	2959.3	3090.1	2704.7		
	2401.8	2317.9	2434.9	3623.1		
	3175.3	2817.2	3669.2	1820.9		
	4000.8	3669.2	1775.2	2401.8		
	3090.1	1820.9	2614.7	2317.9		
	2704.7	2802.3	2579.9	2017.4		

H_0: Group/Category membership has no effect on Pt anual (mm)

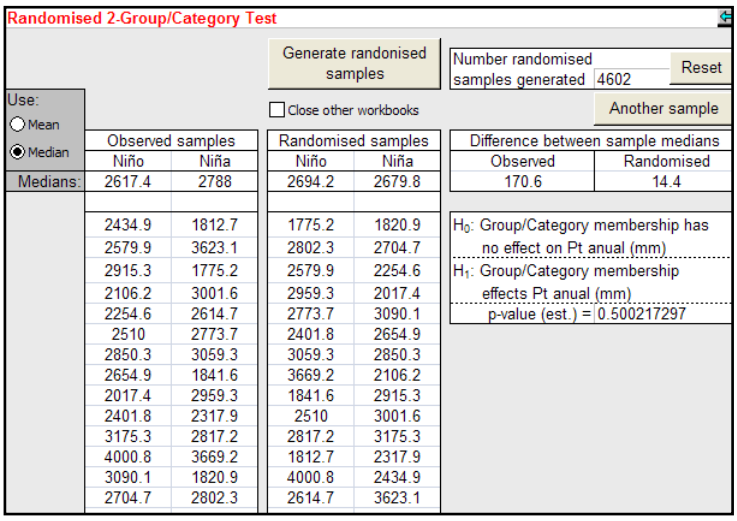
H_1: Group/Category membership effects Pt anual (mm)

p-value (est.) = 0.806056701

Este libro de trabajo realiza una prueba de hipótesis para la media utilizando la técnica de aleatorización. Usted puede elegir el número de muestras aleatorias que desee; sin embargo se sugieren al menos 1000.

Para un nivel de significancia de 5%, el resultado de la prueba de hipótesis sobre la media de la precipitación anual entre años Niña y Niño indica que no son estadísticamente diferentes (p=0.806). El valor de “p” para la prueba “t” fue de 0.794; el cual es muy similar al obtenido por el método de aleatorización.

Prueba de hipótesis entre dos medianas utilizando el método de aleatorización



Randomised 2-Group/Category Test

Generate randonised samples | Number randomised samples generated 4602 | Reset | Another sample | Close other workbooks

Use: ( ) Mean (•) Median

	Observed samples		Randomised samples		Difference between sample medians	
	Niño	Niña	Niño	Niña	Observed	Randomised
Medians:	2617.4	2788	2694.2	2679.8	170.6	14.4
	2434.9	1812.7	1775.2	1820.9		
	2579.9	3623.1	2802.3	2704.7		
	2915.3	1775.2	2579.9	2254.6		
	2106.2	3001.6	2959.3	2017.4		
	2254.6	2614.7	2773.7	3090.1		
	2510	2773.7	2401.8	2654.9		
	2850.3	3059.3	3059.3	2850.3		
	2654.9	1841.6	3669.2	2106.2		
	2017.4	2959.3	1841.6	2915.3		
	2401.8	2317.9	2510	3001.6		
	3175.3	2817.2	2817.2	3175.3		
	4000.8	3669.2	1812.7	2317.9		
	3090.1	1820.9	4000.8	2434.9		
	2704.7	2802.3	2614.7	3623.1		

H_0: Group/Category membership has no effect on Pt anual (mm)

H_1: Group/Category membership effects Pt anual (mm)

p-value (est.) = 0.500217297

Este libro de trabajo realiza una prueba de hipótesis para la mediana utilizando la técnica de aleatorización. Usted puede elegir el número de muestras aleatorias que desee; sin embargo se sugieren al menos 1000.

Para un nivel de significancia de 5%, el resultado de la prueba de hipótesis sobre la mediana de la precipitación anual entre años Niña y Niño indica que no son estadísticamente diferentes (p=0.500).

Remuestreo

La palabra remuestreo (resampling) describe aquellas técnicas de simulación empleadas para estimar parámetros y realizar pruebas de hipótesis, a partir de los datos observados y de la generación de muestras simuladas de igual tamaño que la muestra original (denominadas remuestras). Los métodos de aleatorización y las estimaciones "bootstrap" son considerados como casos particulares de la metodología de simulación conocida como Monte Carlo (por su relación con los juegos de azar de Monte Carlo, Mónaco).

El procedimiento de remuestreo puede implementarse mediante diferentes técnicas analíticas como Bootstrap (muestreo con reemplazo, literalmente “cordones de las botas”, Jackknife (traducido como “la cuchilla de mil usos”), aleatorización y permutaciones (pruebas exactas basadas en todas las posibles muestras). Las técnicas de remuestreo responden a la siguiente pregunta: "¿Qué tan probable es que si la hipótesis nula es cierta, se observaría un valor tan extremo como el medido por casualidad?"

Dado dos muestras aleatorias, el **método de "bootstrap"** parte de la premisa que los datos de ambas muestras provienen de una misma población y que por tanto es posible crear una seudopoblación conformada por las observaciones de ambas muestras. A partir de este nuevo set de datos se obtienen 1000 o más *muestras con reemplazo* y se calcula la media para cada muestrea. Finalmente se reordenan los resultados y se obtienen los percentiles 2.5 y 97.5; dichos percentiles corresponden a un intervalo de confianza para la media de 95%. Este método se utiliza para:

- Valorar el sesgo y el error estándar de un estadístico calculado a partir de una muestra.
- Establecer un intervalo de confianza para un parámetro estimado.
- Realizar pruebas de hipótesis respecto a uno o más parámetros poblacionales.

El procedimiento de **aleatorización o asignación al azar**, también utiliza el set de datos compuesto (seudopoblación), pero en lugar de extraer muestras con reemplazo obtiene *muestras sin reemplazo*, los datos se reordenan sistemática o aleatoriamente al menos unas 1000 veces y en cado reordenamiento se asignan n1 observaciones al primer grupo y n2 observaciones al segundo grupo y luego se calcula el estadístico de prueba deseado para cada reordenamiento (e.g. media, mediana, desviación estándar).

El *método de Jackknife* (traducido como "la cuchilla de mil usos") consiste en crear muestras eliminando una observación "X_i" cada vez que se obtiene una nueva muestra e igual que en los casos anteriores para cada muestra se calcula el valor del parámetro de interés (e.g. media), el proceso se repite para *i=1* hasta *N*. De esta manera se obtienen los denominados pseudovalores, los cuales se utilizan para calcular la media, la cual se denomina estimador jackknife. A partir de la distribución obtenida se calcula su varianza y el intervalo de confianza para el mismo. "Jackknifing" es similar a "bootstrapping" y se utiliza en la estadística inferencial para estimar:

- Valorar el sesgo y el error estándar de un estadístico calculado a partir de una muestra.
- Establecer un intervalo de confianza para un parámetro estimado.

El método de **permutaciones** (aunque realmente se trabaja con las combinaciones) se utiliza para someter a prueba la hipótesis nula asumiendo que se obtienen todas las muestras posibles correspondientes a las permutaciones en el set datos (W), el cual está dado por:

$$W = \frac{N!}{n_1!\, n_2!}$$ donde N: es el número total de observaciones, n_1 y n_2 el número de observaciones en la muestra 1 y 2, respectivamente y el símbolo ! indica el operador factorial. Por ejemplo, el número de formas en que 45 datos pueden dividirse en tres grupos de 15 observaciones cada uno es 5.34 $*10^{19}$. Obviamente este número de muestras no es posible de calcular ni necesaria y por tanto se opta por utilizar un método de remuestreo como aleatorización ó las estimaciones "bootstrap".

$$W = \frac{45!}{15!\ 15!\ 15!} = 5.34 \times 10^{19}$$

Las pruebas basadas en permutaciones se pueden utilizar con cualquier estadístico, independientemente de si su distribución es conocida o no. Esto permite elegir el estadístico que mejor discrimina entre la hipótesis nula y la alternativa y que minimiza las pérdidas. Las pruebas basadas en permutaciones pueden utilizarse para el análisis de diseños balanceados y no balanceados. Las permutaciones someten a prueba una hipótesis sobre la distribución de los datos; en tanto que "bootstraps" realiza una prueba de hipótesis sobre los parámetros de la distribución. Por esta razón, sus supuestos son menos exigentes, pero a cambio, las pruebas de hipótesis no son exactas.

Procedimiento para realizar una prueba de hipótesis mediante remuestreo utilizando el procedimiento de aleatorización con 20 observaciones procedentes de dos muestras de 10 observaciones cada una (http://davidmlane.com/hyperstat/B163479.html**)**.

1) Calcule la media o mediana para cada grupo y la diferencia "D1" entre las medias o medianas de los grupos. Esta es la **diferencia observada** en las muestras.

2) Mezcla las dos muestras y asigne al azar 10 datos a la muestra 1 y 10 datos a la muestra 2 y calcule la diferencia entre las medias o medianas de dichos grupos (D2). El numero de formas (muestras) en que las 20 observaciones pueden dividirse en dos grupos esta dado por:

$W = \frac{N!}{n_1!\, n_2!}$ donde N: es el número total de observaciones (20), n_1 y n_2 el número de observaciones en la muestra 1 y 2, respectivamente y el símbolo ! indica el operador factorial.

3) Repita el paso (2) un gran número de veces (e.g. 1000) para crear un conjunto de muestras de las diferencias entre las medias o medianas (D) y de esta manera crear una distribución empírica de "D" dado que su asignación a los dos grupos fuese aleatoria (hipótesis nula: bajo la hipótesis nula el grupo/categoría no ejerce ninguna influencia en el valor de X). A estas estimaciones se les conoce como remuestreo mediante aleatorización.

4) Determine cuántas veces las diferencias entre las medias o medianas obtenidas por remuestreo son iguales o superiores a la diferencia obtenida entre los datos reales. Esta proporción representa el valor de "p" estimado a partir del remuestreo y equivale al valor de "p" obtenido en las pruebas tradicionales de hipótesis. Si el valor de "p" es pequeño (e.g. menor que 0.05) se puede concluir que la media o mediana de las dos muestras es diferente a dicho nivel de significancia; para una prueba de 1 cola el valor de "p" debe ser 0.025. En otras palabras, la conclusión es que existe una diferencia entre la media ó mediana de los grupos (prueba de dos colas).

En resumen, una prueba de remuestreo se base en los datos obtenidos por el investigador(a) y no en una distribución teórica como la "t" o la normal. La prueba compara un estadístico calculado (la diferencia entre medias o medianas en este ejemplo) con el valor de ese estadístico obtenido de otros cientos o miles de arreglos del set de datos. El valor de probabilidad (p) es la proporción de veces que el estadístico estimado a partir de diferentes muestras aleatorias obtenidas del set de datos es mayor o igual al valor obtenido de los datos reales. Las pruebas de permutaciones sólo tienen un inconveniente: son imprácticas para muestras de moderadas a grandes. Por ejemplo, el número de formas en que 45 datos pueden dividirse en tres grupos de 15 observaciones cada uno es 5.34 *1019.

$$W = \frac{45!}{15!\ 15!\ 15!} = 5.34 \times 10^{19}$$

La alternativa a esta limitación es utilizar las técnicas de "bootstrap" ó aleatorización, las cuales calcula una proporción de todas las posibles muestras y a partir de dichos datos estima el valor de "p". Para que el resultado sea confiable se recomienda realizar entre 10000 y 5000 remuestreos.

Good, Phillip I. Introduction to statistics through resampling methods and Microsoft Office Excel.West Sussex, Reino Unido (INGLATERRA). John Wiley & Sons. 231p. 2005.

Ejercicio

El archivo efecto_borde.xls o efecto_borde.xlsx puede utilizarse para evaluar el efecto de muestras de diferentes tamaños en las pruebas de hipótesis de dos grupos o categorías.

Análisis de dos variables nominales (variables cualitativas)

En la presente sección se ilustra el uso de XLSTatistics para el análisis de dos variables nominales. Algunos ejemplos de variables nominales o cualitativas son: uso-cobertura de la tierra (e.g. bosque, pasto, urbano), sexo (masculino, femenino), color (e.g. azul, rojo, blanco) y cultivos (eg. maíz, arroz, frijol, yuca). En el menú grafico de XLStatisitics corresponde a **2Cat**.

Nota: Esta sección del tutorial asume que usted ya activó el complemento de XLStatisticis (XLStatistics.xlam ó XLStats.xls); de no ser así, ver página 23.

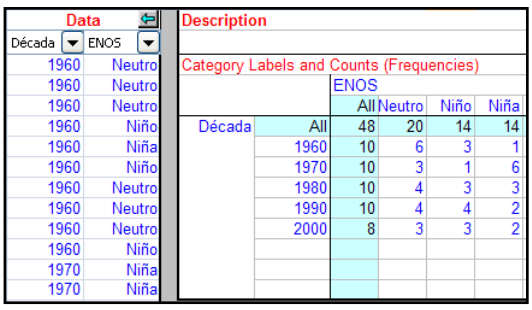
1. Abra el archivo "xlstats_tutorial.xlsx" y seleccione las columnas Década y ENOS, haga un clic sobre XLStatistics y luego otro clic sobre 2Cat. Observe que el programa abre la el libro de cálculo **2cat.xls** y copia los datos seleccionados a dicho libro.

En la sección inferior del libro de cálculo **2Cat.xls** usted observará seis hojas de cálculo:

1) *Information*: Información. Esta hoja contiene información general sobre la organización de este libro de trabajo.
2) *Data and Description*: Datos y estadísticos descriptivos
3) *Summaries*: Síntesis de datos (Tabla de frecuencia y gráficos)
4) *Tests rxc*: Prueba de Ji-cuadrado (de Pearson). Prueba de independencia.
5) *Tests rx2:* Prueba de diferencia entre dos proporciones
6) *Extra tolos*. Otras herramientas

Nota: Recuerde que **SOLO** debe modificar el contenido de las celdas con números o textos en color azul.

Datos y estadísticos descriptivos



Data

Década	ENOS
1960	Neutro
1960	Neutro
1960	Neutro
1960	Niño
1960	Niña
1960	Niño
1960	Neutro
1960	Neutro
1960	Neutro
1960	Niño
1970	Niña
1970	Niña

Description

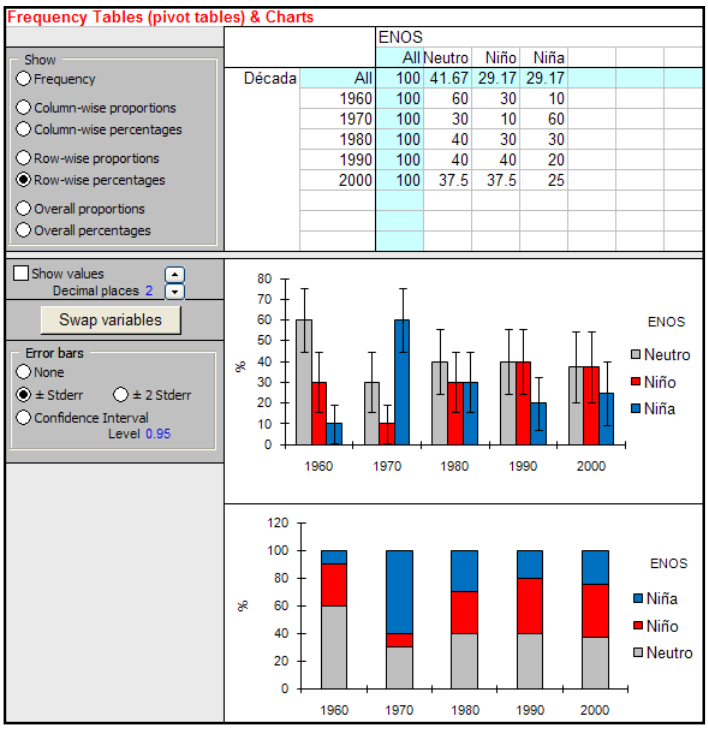
Category Labels and Counts (Frequencies)

		ENOS			
		All	Neutro	Niño	Niña
Década	All	48	20	14	14
	1960	10	6	3	1
	1970	10	3	1	6
	1980	10	4	3	3
	1990	10	4	4	2
	2000	8	3	3	2

Esta hoja de caculo le muestra los datos y un cuadro de frecuencia para las variables década y ENOS.

Nota: observe que las décadas están declaradas como texto y no como números.

Tabla y grafico de frecuencia



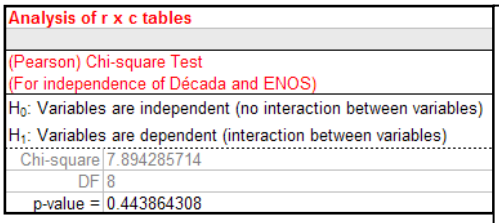
Esta hoja de caculo le muestra un cuadro de frecuencia para las variables década y ENOS así como gráficos de barras comparativos y compuestos.

Los gráficos muestran el tipo de frecuencia que usted desea ilustrar en su trabajo.

- Frecuencia absoluta
- Proporciones/porcentajes por columna.
- Proporciones/porcentajes por fila.
- Proporciones/porcentajes para todo el set de datos.

Nota: Seleccionar cuidadosamente la proporción que usted desea mostrar.

Prueba de Ji-cuadrado (de Pearson): Prueba de independencia



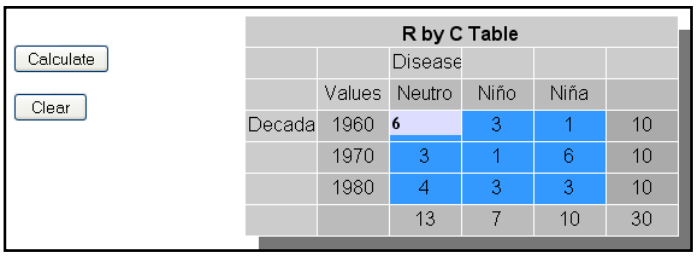
Esta prueba responde a la siguiente pregunta: ¿Existe relación entre la frecuencia de episodios de ENOS y la década? La prueba indica que las variables son independientes ya que el valor de "p" es 0.4.

Nota: Cochran recomienda aceptar el valor de ji-cuadrado si:

- ✓ Un máximo de 20% de las celdas tienen una frecuencia esperada <5.
- ✓ Ninguna celda tiene un valor esperado <1.

Observe que en este caso el set de datos no cumple con este criterio.
Fuente: http://www.openepi.com/RbyC/RbyC.htm.

Nota: "Tables" es un programa gratuito que permite realizar pruebas estadísticas para tablas de 2*2, 3*3 y 2*7. http://www.quantitativeskills.com/downloads/winprog/tabsetup.exe



Interfaz grafica de Open Epi para prueba de fila por columna. http://www.openepi.com/RbyC/RbyC.htm.

Chi Square for R by C Table

Chi Square=	6.02
Degrees of Freedom=	4
p-value=	0.1977

Cochran recomienda aceptar el valor de ji-cuadrado si:

- ✓ Un máximo de 20% de las celdas tienen una frecuencia esperada <5.
- ✓ Ninguna celda tiene un valor esperado <1.

Para esta tabla:

Las 9 celdas tienen un valor esperado inferior a < 5.
Ninguna celda tiene un valor esperado inferior < 1.

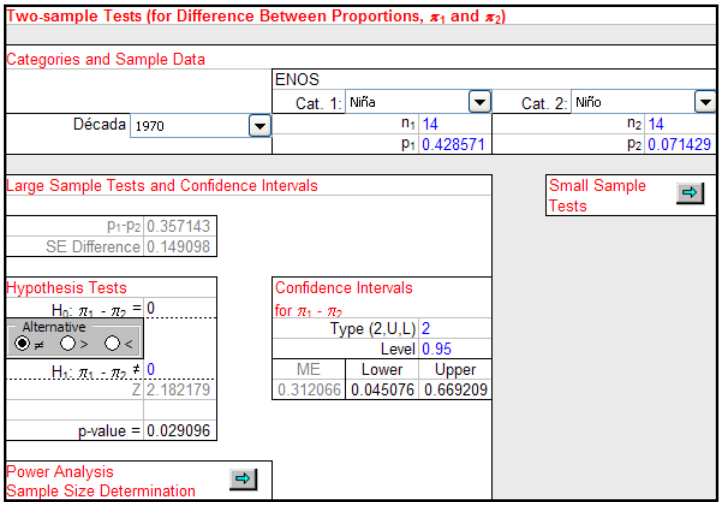
Conclusión: La tabla no cumple con el criterio de Cochran.

Valor esperado = toral fila row l*total columna /gran total

Rosner, B. Fundamentals of Biostatistics. 5th ed. Duxbury Thompson Learning. 2000; p. 395

Resultado de OpenEpi, Version 2, open source calculator--RbyC ttp://www.openepi.com/RbyC/RbyC.htm Source file last modified on 05/16/2009 11:08:58

Prueba de diferencia entre dos proporciones (muestras grandes)



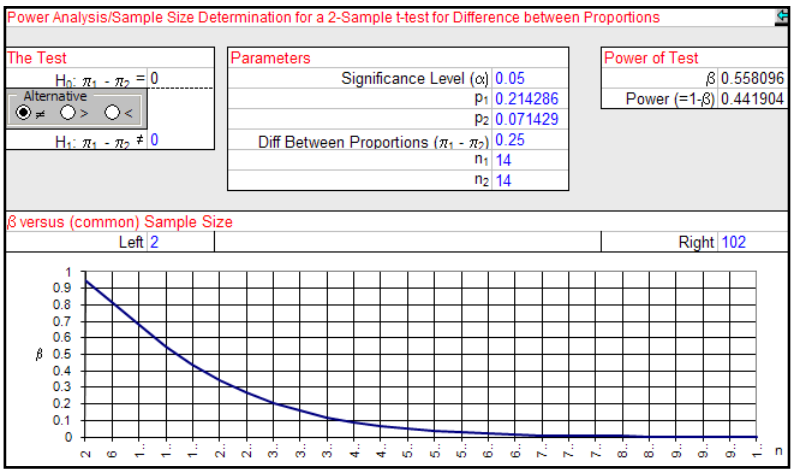
En esta hoja de cálculo usted prueba someter a prueba la siguiente hipótesis:

¿Es la frecuencia de dos categorías (eg. Niño Vs Niña por década) diferente?

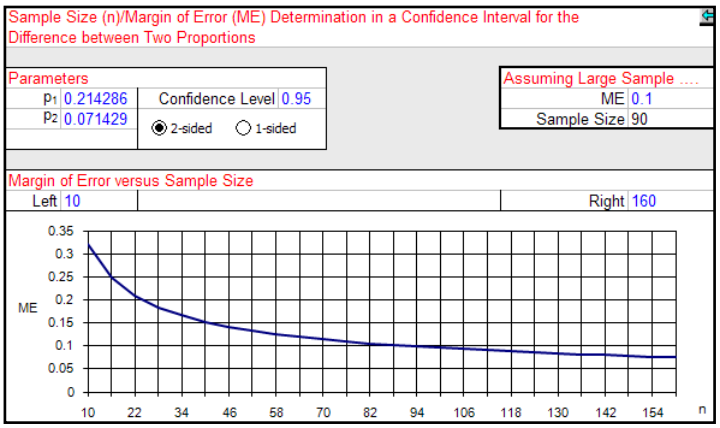
El ejemplo muestra la prueba de hipótesis para la década del 70. El resultado es que la frecuencia de años Niño y Niña es estadísticamente diferente (p=0.029).

Nota: El valor de "p" es calculado utilizando una aproximación a la distribución normal (muestra grande). Observe que el programa también ofrece la alternativa de muestras pequeñas (*small sample test*)

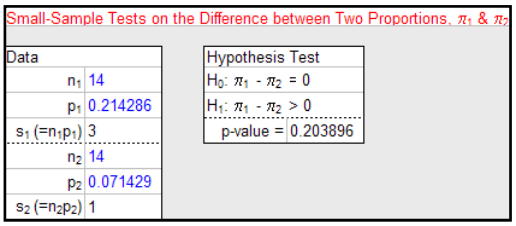
Análisis de poder de la prueba (diferencia entre dos proporciones)



Análisis de poder marginal de la prueba (diferencia entre dos proporciones)



Diferencia entre dos proporciones (muestras pequeñas)



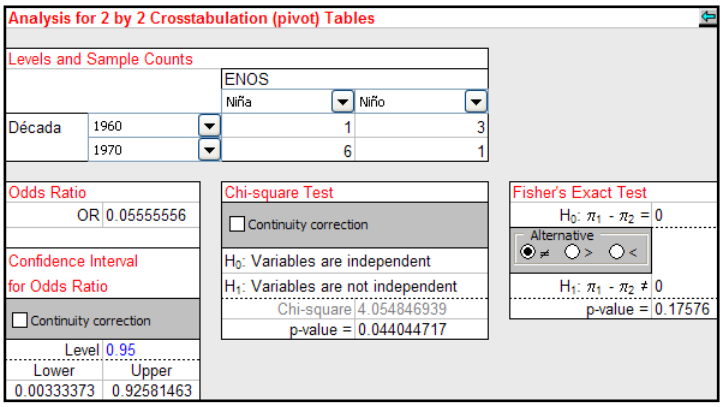
Prueba de hipótesis para dos proporciones:
¿Es la frecuencia de dos categorías (e.g. Años Niño Vs años Niña) diferente?

El resultado es que la frecuencia de años Niño y Niña no es estadísticamente diferente (p=0.20).

Nota: El valor de “p” es calculado utilizando una aproximación a la distribución normal de muestras pequeñas (*small sample test*)

Otras Herramientas

Análisis de tabla de 2x2



La prueba χ^2 de Pearson con *corrección por continuidad.*

La prueba Ji-cuadrado de Yate es equivalente al Ji-cuadrado de Pearson con Corrección por Continuidad. Se cree que la prueba de Ji-cuadrado de Mantel-Haenszel es la más cercana al "verdadero" valor de Ji-cuadrado para muestras pequeñas.

Ver: http://www.quantitativeskills.com/sisa/statistics/twoby2.htm
Prueba de Yate's= 1.856 (p= 0.1731)
Prueba de Mantel Haenszel= 3.686 (p= 0.0548)
Prueba exacta de Fisher p= 0.175757.

Nota: En ninguno de las pruebas la diferencia es significativa (p>0.05).

Tabla y grafico de frecuencia por variables

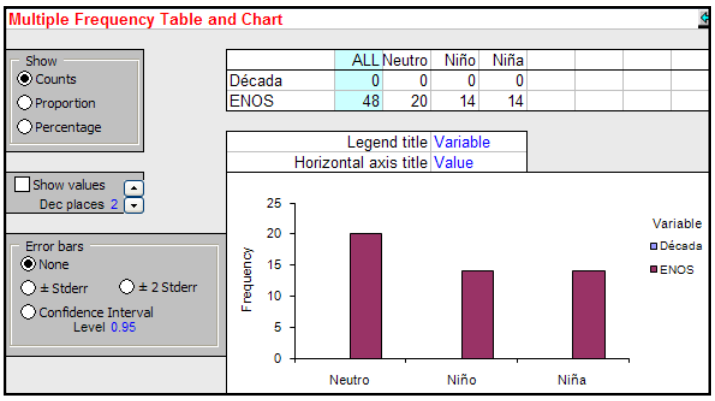
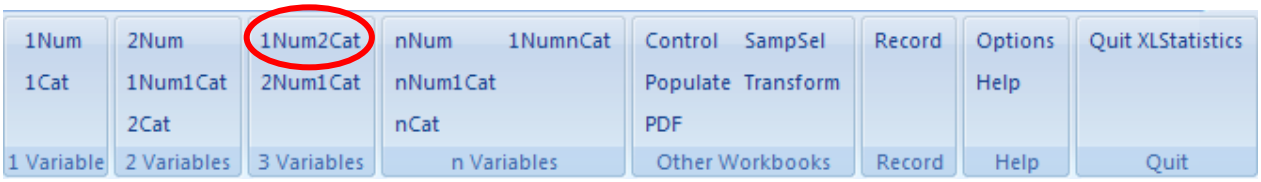


Tabla:
- ✓ Frecuencia absoluta, porcentaje, proporción

Grafico de barras de error
Ninguna
- ✓ ± 1 Error estándar
- ✓ ± 2 Errores estándares
- ✓ Intervalo de confianza

Análisis de una variable numérica y dos nominales

En la presente sección se ilustra el uso de XLSTatistics para el análisis de tres variables: una numérica y dos nominales. En el menú grafico de XLStatisitics corresponde a **1Num 2Cat**.



> **Nota**: Esta sección del tutorial asume que usted ya activó el complemento de XLStatisticis (XLStatistics.xlam ó XLStats.xls); de no ser así, ver página 23.

1. Abra el archivo “xlstats_tutorial.xlsx” y seleccione de la hoja **1Num2Cat** las columnas Pt anual (mm), ENOS y Década.

2. Haga un clic sobre XLStatistics y luego otro clic sobre 

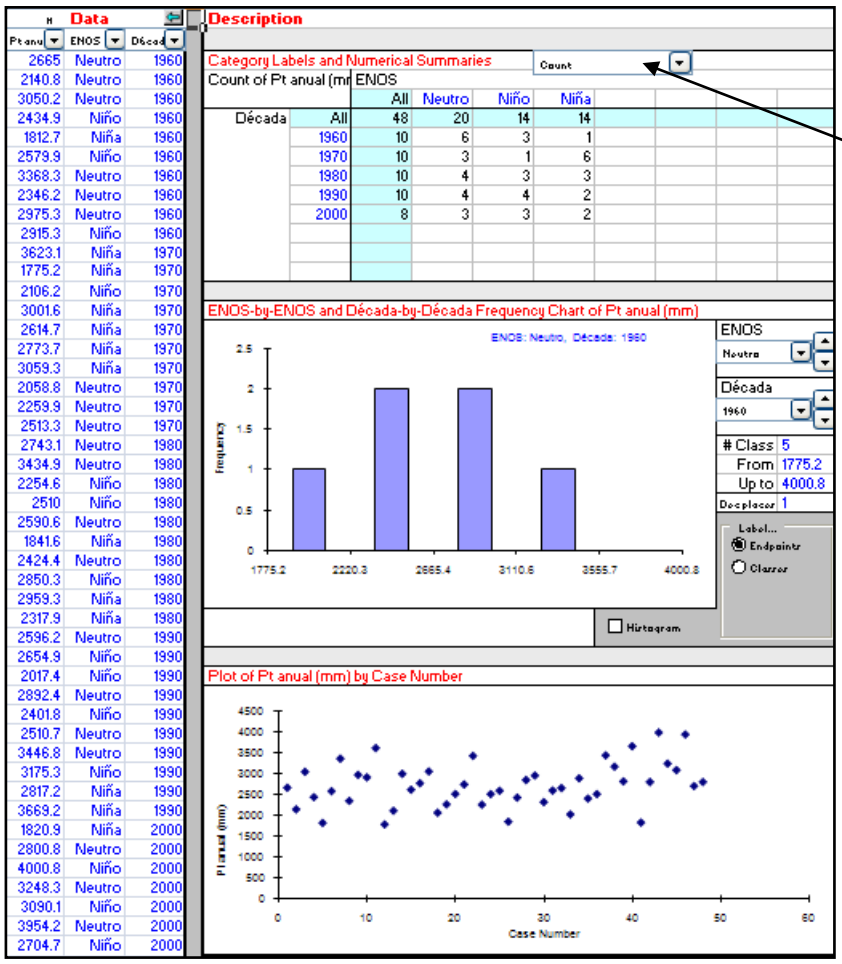
. Observe que el programa abre el libro de cálculo **1Num2cat.xls** y copia los datos seleccionados a dicho libro.

En la sección inferior de la hoja de cálculo **1Num2cat.xls** usted observará cuatro hojas de cálculo:

1) *Information*: Información. Esta hoja contiene información general sobre la organización de este libro de trabajo.
2) *Data and Description*: Datos y estadísticos descriptivos
3) *Summaries*: Síntesis de datos (Tabla de frecuencia y gráficos)
4) *Tests:* Análisis de varianza para muestras balanceadas

Nota: Recuerde que **SOLO** debe modificar el contenido de las celdas con números o textos en color azul.

Datos y estadísticos descriptivos



Cuadro de frecuencia de ENOS Vs Década. Usted debe elegir el estadístico que sea incluir en el cuadro.

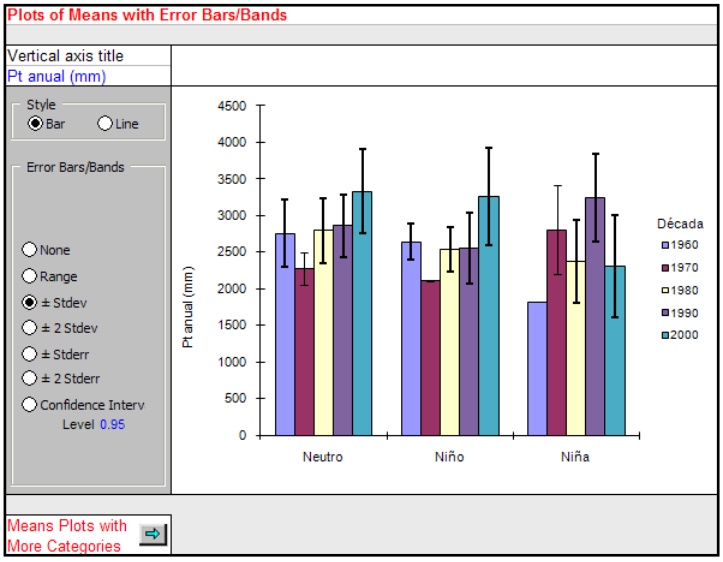
- ✓ Frecuencia
- ✓ Media
- ✓ Des. Estándar
- ✓ Asimetría
- ✓ Mínimo
- ✓ Primer cuartil (Q1)
- ✓ Mediana (Segundo cuartil Q2)
- ✓ Tercer cuartil (Q3)
- ✓ Máximo

Grafico de barras para década Vs episodio de ENOS.
Eje X:

- ✓ Límite superior de clase
- ✓ Clase (intervalo, punto medio)

Histograma

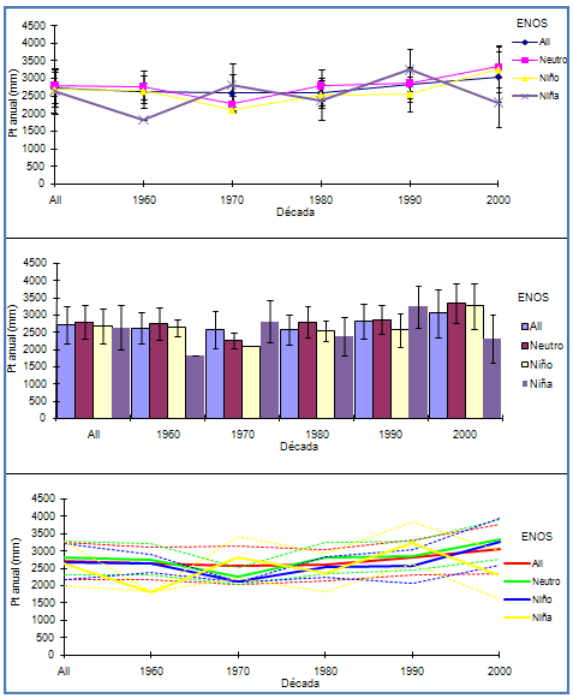
Gráfico de medias y bandas/barra de error



Precipitación anual media (mm) por década y fase de ENOS.
Usted puede adicionar al grafico:

- ✓ Rango de datos
- ✓ Desviación estándar
- ✓ Error estándar
- ✓ Intervalo de confianza

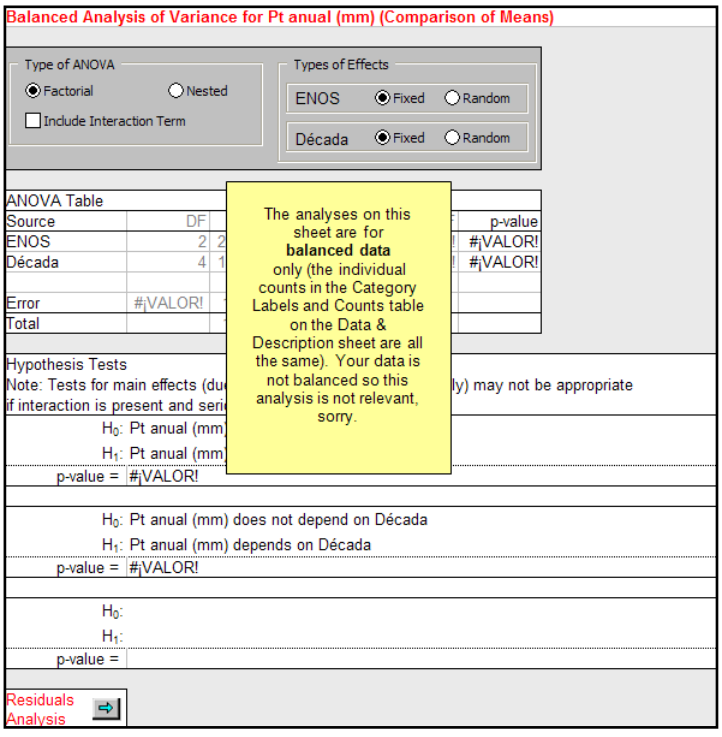
Gráfico de medias por categoría



Usted debe digitar o copiar y pegar en cada una de las hojas de cálculo del libro 1N2CMP.xls los datos a graficar:

- ✓ Media para cada grupo.
- ✓ Valor que desea adicionar a la media (e.g. desviación estándar).
- ✓ Valor que desea restar a la media. (e.g. desviación estándar).
- ✓ Valor superior: Excel calcula este valor.
- ✓ Valor inferior: Excel calcula este valor.

Análisis de varianza para muestras balanceadas



Esta hoja de cálculo realiza un análisis de varianza factorial ó anidado para muestras balanceadas (igual número de observaciones por tratamiento).

El set de datos seleccionado no es balanceado y por tanto el programa le advierte que no puede utilizar esta prueba estadística.

Nota: A continuación utilizamos las variables Pt anual (mm), Estación y Década de la hoja de cálculo **1Num2Cat** del archivo "xlstats_tutorial.xlsx". Los efectos que se desean analizar son: lluvia por década y por estación; en otras palabras ¿es la cantidad de lluvia medida durante la década del 70 y 80 independiente de la estación?

Supuestos de ANOVA una vía y factorial y de la prueba t para muestras independientes

1. **La distribución de la variable dependiente debe ser normal**. Los datos en cada celda deben ser aproximadamente normales. Compruebe la normalidad de los datos utilizando histogramas, coeficientes de asimetría y curtosis. Realice prueba de hipótesis de normalidad. Analice los residuos para verificar los supuestos del modelo.

2. **Homogeneidad de varianzas**. La varianza en cada celda debe ser similar. Compruebe a través de prueba de Levene de homogeneidad o de otros análisis de varianza que generalmente se produce como parte de la producción de promedio. Analice los residuos para verificar los supuestos del modelo.

3. **Tamaño de la muestra.** Es preferible que el tamaño de muestra sea superior a 20 observaciones por celda, esto fortalece la robustez de la prueba ante la violación de los supuestos de normalidad e igualdad de varianzas. Al aumentar el tamaño de la muestra aumenta el poder de la prueba.

4. **Las observaciones deben ser independientes**. Dos observaciones son independientes cuando el resultado de una variable o de un grupo no depende de otra variable o grupo (por lo general se garantizan en el diseño del estudio).

Resultado del análisis de varianza (diseño balanceado)

El diseño es balanceado cuando el número de observaciones por combinación de tratamientos es igual. En este caso los tratamientos o factores son: estación y década.

Balanced Analysis of Variance for Pt anual (mm) (Comparison of Means)

Type of ANOVA: (•) Factorial ( ) Nested; [ ] Include Interaction Term

Types of Effects: Estación (•) Fixed ( ) Random; Decada (•) Fixed ( ) Random

ANOVA Table

Source	DF	SS	MS	F	p-value
Estación	3	12289277	4096425.7	15.528224	5.972E-08
Decada	1	2433891.6	2433891.6	9.2260955	0.0032814
Error	75	19785387	263805.16		
Total	79	34508555			

Hypothesis Tests

Note: Tests for main effects (due to Estación and Decada separately) may not be appropriate if interaction is present and serious

H_0: Pt anual (mm) does not depend on Estación

H_1: Pt anual (mm) depends on Estación

p-value = 5.972E-08

H_0: Pt anual (mm) does not depend on Decada

H_1: Pt anual (mm) depends on Decada

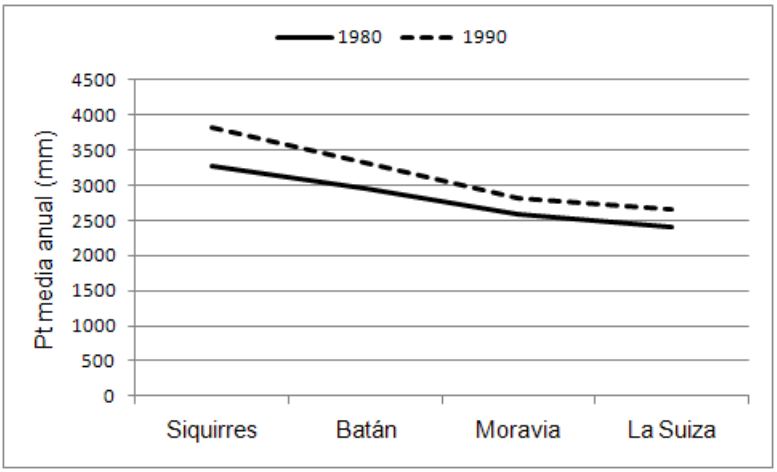
p-value = 0.0032814

Para un alfa de 0.05, existe una diferencia significativa tanto entre las décadas como entre los episodios de ENOS.

Ver gráfico de interacción entre factores.

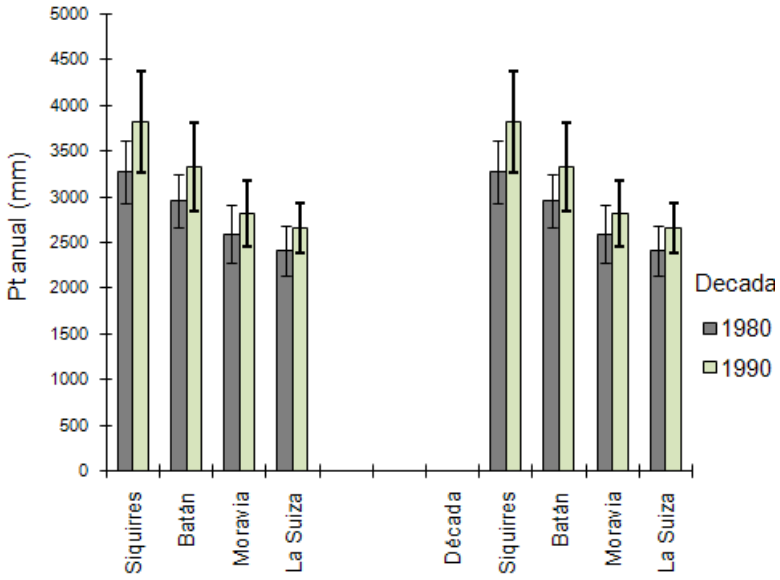
Se eligió efectos fijos porque se asume que no existen otras estaciones ni tampoco otras décadas para el análisis; en otras palabras se utilizaron todas las opciones disponibles para las variables estaciones y décadas. Sin embargo, esta decisión podría cuestionarse pues se podría asumir que los sitios donde se mide la lluvia son solo algunos de todos los posibles sitios donde podría ubicarse una estación meteorológica. Bajo este argumento el efecto estación debería ser aleatorio.

Gráfico de interacción entre factores (estación-década)



El gráfico muestra que no existe interacción entre década y estaciones. Para que exista interacción entre los factores las líneas deben entrecruzarse.

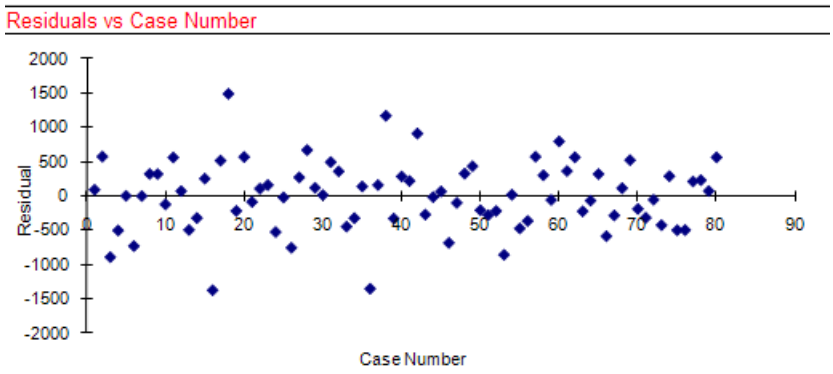
Gráfico de medias e intervalo de confianza al 95%



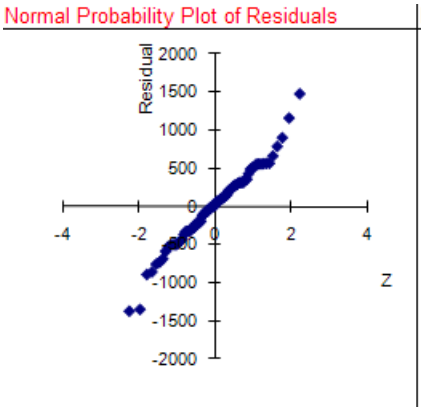
Precipitación media por década y estación. La barra vertical indica un intervalo de confianza al 95%.

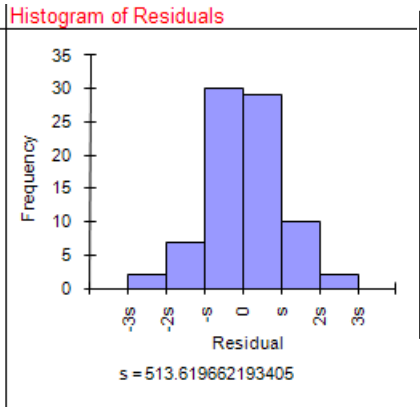
Si dos barras se traslapan entonces las diferencias entre las medias no es significativa a un alfa de 0.05.

Análisis de residuos



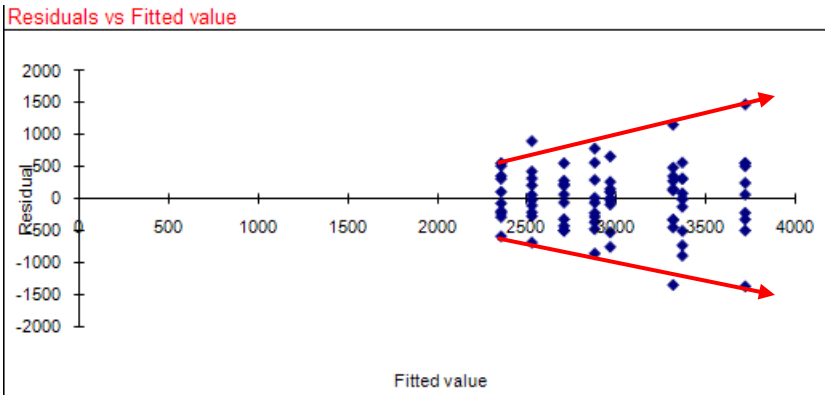
Residuos por observación
Cuando los datos son graficados en orden cronológico (como en este caso), cualquier patrón en los residuos podría indicar el efecto de una influencia externa.
¿Observa usted algún patrón?





Evaluación del supuesto de normalidad de los datos.

Gráfico de probabilidad normal e histograma de residuos. Las gráficas indican que los datos no se apartan de una distribución normal.



Residuos Vs valores estimados. Evaluación del supuesto de igualad de varianzas. Los datos muestras una posible violación de este supuesto pues la variabilidad aumenta al aumentar la precipitación.

Análisis de varianza factorial en línea con StatGraphics

http://www.statgraphicsonline.com/SGOnline.aspx;
http://www.statgraphicsonline.com/MultipleFactorANOVA.aspx

Para utilizar esta versión en línea de StatGraphics usted debe registrarse (es gratuito). Los datos utilizados se encuentran en el archivo anova2vias_balanceado.xls y corresponden a los analizados en la sección anterior (**1Num2Cat)**.

Esta barra de menú le permite configurar las opciones que ofrece el programa.

- ✓ *Data Input*: Insumo de datos (incluye archivos de Excel)
- ✓ *Tables and Graphs*: Tablas y gráficos que desea como parte de la ANOVA
- ✓ *Results to Save*: Resultados que desea guardar
- ✓ *Preferences*: Preferencias de fuentes, color de gráficos (fondo y primer plano), tipo de puntos y líneas.

Una vez leídos los datos y configurado las opciones de análisis que desea haga un clic sobre *Calculate*.

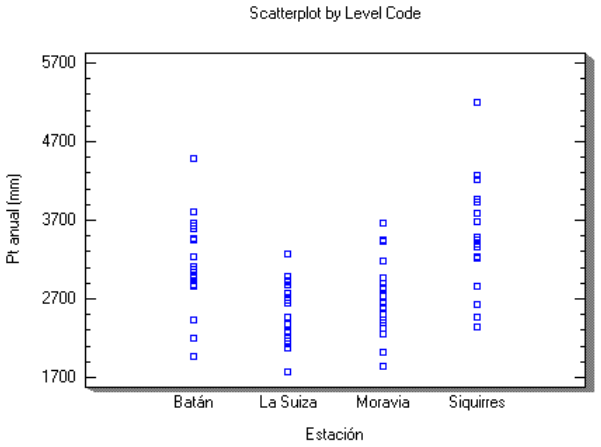


Gráfico de dispersión
Este gráfico muestra la precipitación (mm) para cada una de las estaciones. El gráfico permite evaluar el supuesto de igualad de varianzas. Los datos muestras una posible violación de este supuesto pues la variabilidad aumenta con la cantidad de precipitación.

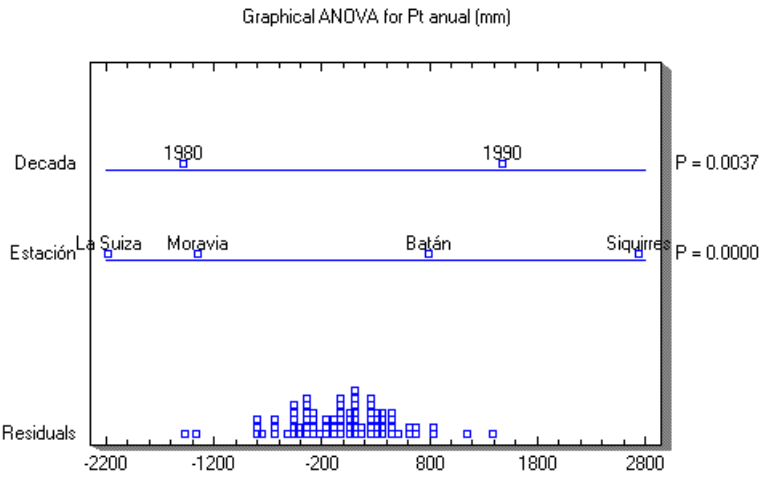
Analysis of Variance for Pt anual (mm) - Type I Sums of Squares (Análisis de varianza Tipo I)

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
MAIN EFFECTS					
A:Estación	1.22893E+07	3	4.09643E+06	15.15	0.0000
B:Decada	2.43389E+06	1	2.43389E+06	9.00	0.0037
INTERACTIONS					
AB	321223	3	107074	0.40	0.7562
RESIDUAL	1.94642E+07	72	270336		
TOTAL (CORRECTED)	3.45086E+07	79			

Los valores de F se calculan utilizando el cuadrado medio del error (270336).

El valor de “p”, inferior a 0.05, indica que el efecto de los factores es estadísticamente significativo en el valor de la precipitación.

ANOVA Gráfica



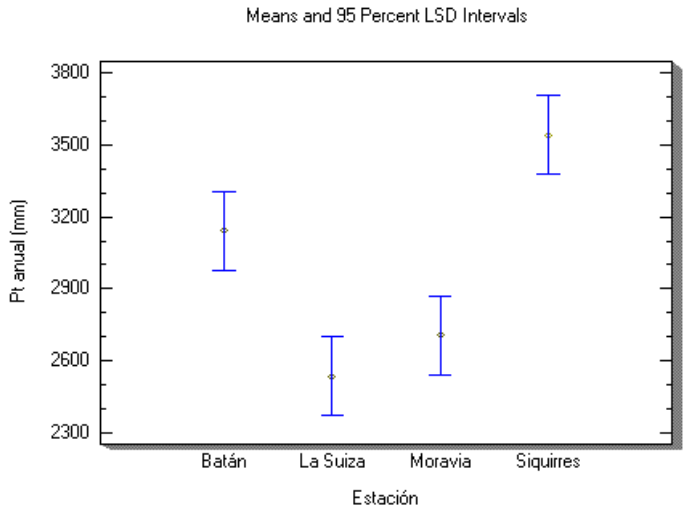
La representación gráfica de ANOVA se utiliza para mostrar gráficamente el efecto de cada factor (en este caso estaciones y década) con respecto a la variabilidad de los residuos. La grafica muestra, para cada factor, las desviaciones medias ajustadas con respecto a la media general del set de datos. Cualquier factor que muestre una variabilidad considerablemente mayor que los residuos es probable que sea un factor importante. Para un alfa de 0.05, los efectos de ENOS y década son significativos ($p<0.05$). Los valores de “p” proceden de la tabla de análisis de varianza.

Table of Least Squares Means for Pt anual (mm) with 95% Confidence Intervals

			Stnd.	Lower	Upper
Level	Count	Mean	Error	Limit	Limit
GRAND MEAN	80	2980.55			
Estación					
Batán	20	3140.69	116.26	2908.93	3372.45
La Suiza	20	2534.8	116.26	2303.04	2766.56
Moravia	20	2705.43	116.26	2473.67	2937.19
Siquirres	20	3541.3	116.26	3309.53	3773.06
Decada					
1980	40	2806.13	82.21	2642.25	2970.01
1990	40	3154.98	82.21	2991.10	3318.86
Estación by Decada					
Batán,1980	10	2956.12	164.42	2628.36	3283.88
Batán,1990	10	3325.26	164.42	2997.5	3653.02
La Suiza,1980	10	2408.2	164.42	2080.44	2735.96
La Suiza,1990	10	2661.4	164.42	2333.64	2989.16
Moravia,1980	10	2592.67	164.42	2264.91	2920.43
Moravia,1990	10	2818.19	164.42	2490.43	3145.95
Siquirres,1980	10	3267.53	164.42	2939.77	3595.29
Siquirres,1990	10	3815.06	164.42	3487.30	4142.82

Esta tabla muestra la precipitación anual media (Pt mm) global y para cada nivel de los factores. También muestra el error estándar de cada media, que es una medida de su variabilidad de muestreo. Las columnas de la derecha muestran los intervalos de confianza al 95% para cada una de las medias.

Means Plot (Gráfico de medias e intervalo LSD al 95%)



Esta gráfica muestra la media anual de Pt (mm) por estación. Los intervalos para la media se basan en la diferencia mínima significativa de Fisher (LSD) y están construidos de tal manera que si dos medias son iguales, sus intervalos se solaparán 95% del tiempo. Los pares de intervalos que no se superponen verticalmente corresponden a medias estadísticamente diferentes.

Multiple Range Tests for Pt anual (mm) by Estación (Pruebas múltiples)

Method: 95 percent LSD (Intervalo confianza LSD de Fisher al 95%)

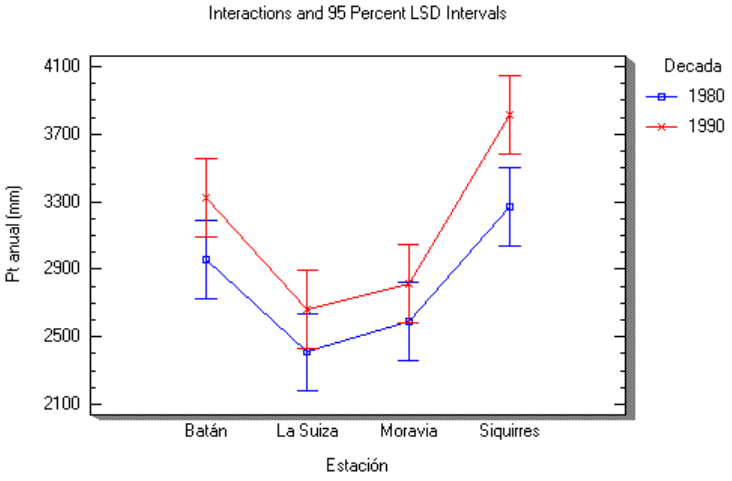
Estación	*Count*	*LS Mean*	*LS Sigma*	*Homogeneous Groups*
La Suiza	20	2534.8	116.262	X..
Moravia	20	2705.43	116.262	X..
Batán	20	3140.69	116.262	.X.
Siquirres	20	3541.3	116.262	..X

Contrast	*Sig.*	*Difference*	*+/- Limits*
Batán - La Suiza	*	605.89	327.764
Batán - Moravia	*	435.26	327.764
Batán - Siquirres	*	-400.605	327.764
La Suiza - Moravia		-170.63	327.764
La Suiza - Siquirres	*	-1006.5	327.764
Moravia - Siquirres	*	-835.865	327.764

* Indica diferencias estadísticamente significativas al 5%

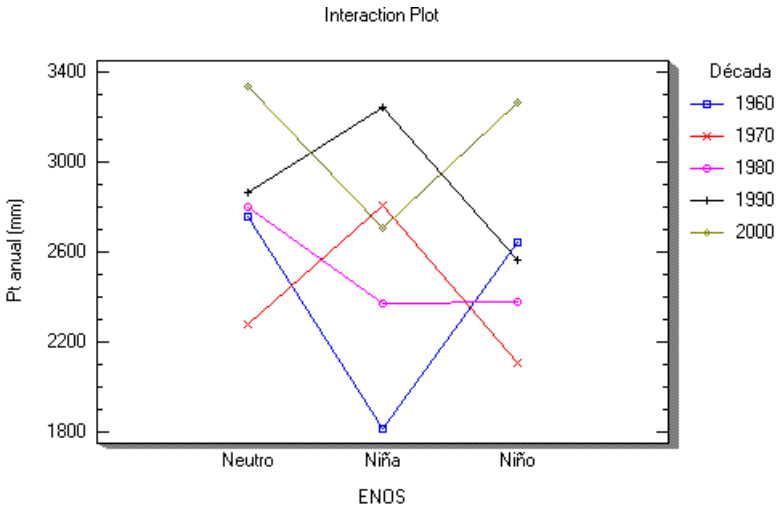
Esta tabla muestra el resultado de aplicar un procedimiento de comparación múltiple para determinar cuáles pares de medias son significativamente diferentes. El asterisco indica que las medias de dichos pares son estadísticamente diferentes a un nivel de confianza del 95%. El método utilizado para discriminar entre medias es la diferencia mínima significativa de Fisher (LSD). Con este método, el error tipo I es de 5% (declarar un contraste como significativo cuando en realidad no lo es).

Gráfico de interacción entre factores (estación-década)



El gráfico muestra que no existe interacción entre las décadas y las estaciones.

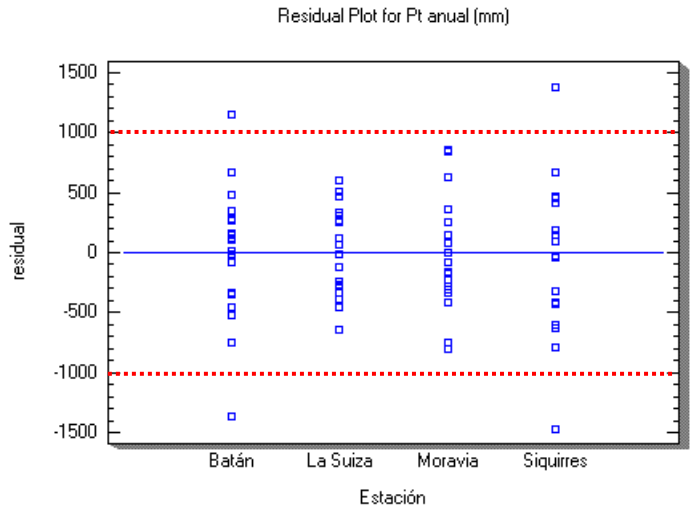
Este gráfico es útil para determinar si existe interacción entre los factores “Estación” y “Década”. Cuando no existe interacción (como en este caso), las líneas son paralelas. El gráfico también incluye los intervalos LSD al 95% para cada media. Si dos intervalos cualesquiera no se traslapan, esto indica que las respectivas medias son estadísticamente diferentes a un alfa 0.05.



Este gráfico ilustra la presencia de interacción entre los factores "ENOS" y "Década" para la Pt anual de la estación Moravia de Chirripó.

Este gráfico es útil para determinar si existe interacción entre los factores "ENOS" y "Década". Las cinco líneas corresponden a los niveles del factor "Década", las cuales conectan las medias estimadas por mínimos cuadrados de los tres niveles de ENOS. Cuando existe interacción (como en este caso), las líneas se traslapan entre sí.

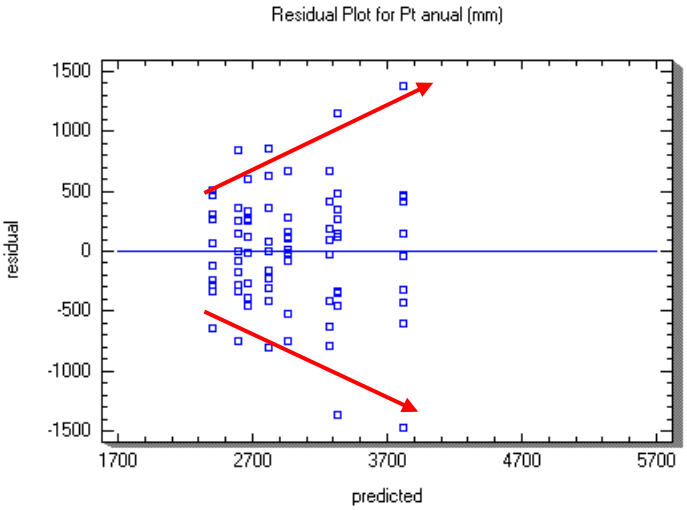
Residuals versus Factor Levels (Análisis de residuos por estación)



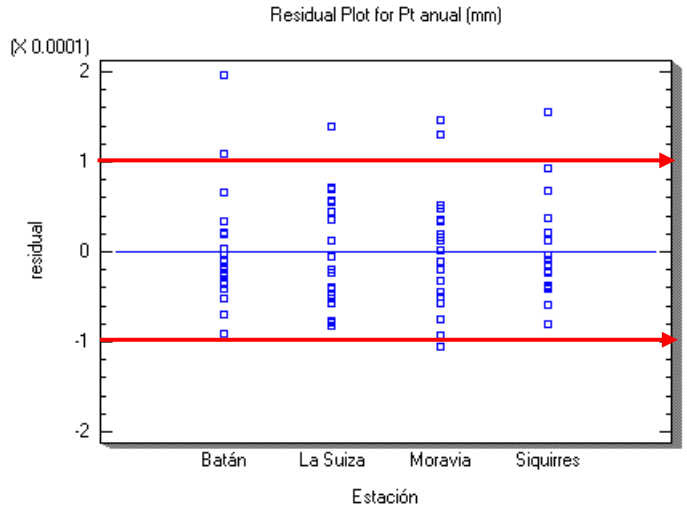
Residuos Vs variable predictora
Este gráfico permite evaluar el supuesto de homogeneidad de varianzas entre estaciones. Cuando las varianzas son heterogéneas se viola dicho supuesto.

Los residuos son iguales a los valores observados de Pt anual (mm) menos la media anual de Pt (mm) del respectivo grupo.

Residuals versus Predicted (Análisis de residuos Vs predicciones)

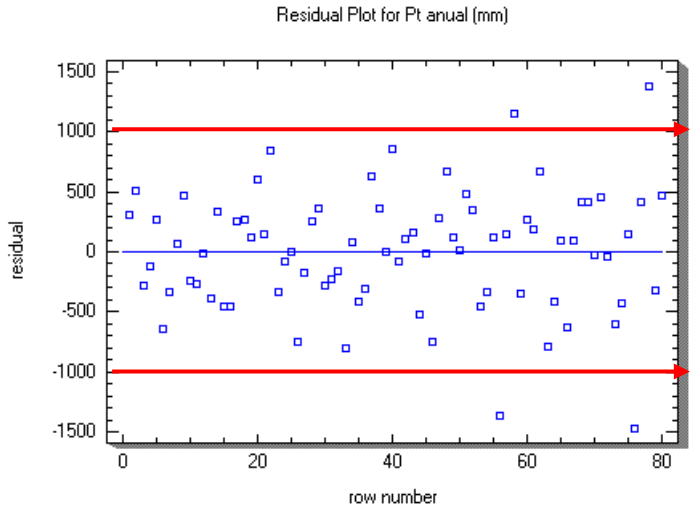


Este gráfico muestra los residuos versus los valores estimados de Pt y se utiliza para evaluar el supuesto de homogeneidad de varianzas de la variable dependiente (Pt). Observe que la dispersión de los datos no es homogénea, pues tiende a incrementar con el valor de Pt. Este tipo de heterocedasticidad pude corregirse utilizando la transformación Log (Pt), raíz cuadrada (Pt) ó 1/Pt.



Transformación 1/Pt. Observe que la dispersión de los datos es homogénea.

Residuals versus Row Number (Análisis de residuos Vs observación)



Este gráfico muestra los residuos versus el número de fila en el archivo. Cualquier patrón que no sea una dispersión aleatoria podría indicar la presencia de correlación serial en los datos.

Análisis de dos variables numéricas y una nominal

En la presente sección se ilustra el uso de XLSTatistics para el análisis de tres variables: una numérica y dos nominales. En el menú grafico de XLStatisitics corresponde a **2Num1Cat**.

Nota: Esta sección del tutorial asume que usted ya activó el complemento de XLStatisticis (XLStatistics.xlam ó XLStats.xls); de no ser así, ver página 23.

1. Abra el archivo "xlstats_tutorial.xlsx" y seleccione de la hoja **2Num1Cat** las columnas Pt anual (mm) , Año y ENOS.

2. Haga un clic sobre XLStatistics y luego otro clic sobre 2Num1Cat. Observe que el programa abre el libro de cálculo **2Num1cat.xls** y copia los datos seleccionados a dicho libro.

En la sección inferior de la hoja de cálculo **2Num1cat.xls** usted observará cuatro hojas de cálculo:

1) *Information*: Información. Esta hoja contiene información general sobre la organización de este libro de trabajo.
2) *Data and Description*: Datos y estadísticos descriptivos
3) *Linear Regression*: Regresión lineal
4) *Extra Tools:* Otras herramientas

Nota: Recuerde que **SOLO** debe modificar el contenido de las celdas con números o textos en color azul.

Datos y estadísticos descriptivos

Data

Pt anu	Año	ENOS
2665	1960	Neutro
2140.8	1961	Neutro
3050.2	1962	Neutro
2434.9	1963	Niño
1812.7	1964	Niña
2579.9	1965	Niño
3368.3	1966	Neutro
2346.2	1967	Neutro
2975.3	1968	Neutro
2915.3	1969	Niño
3623.1	1970	Niña
1775.2	1971	Niña

Description

Category Labels and Numerical Summaries

ENOS	All	Neutro	Niño	Niña
Number	48	20	14	14
Correlation	0.29133	0.38257	0.44257	0.11634
R^2	0.08487	0.14636	0.19586	0.01354

Datos: Lista de variables que se analizarán.

Descripción:
Frecuencia total y por episodio de ENOS.
Coeficientes de correlación para Pt y año por episodio de ENOS
Coeficientes de determinación para Pt y año por episodio de ENOS.

Scatterplot

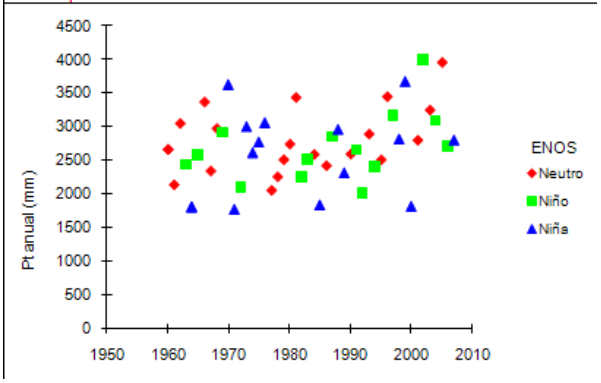
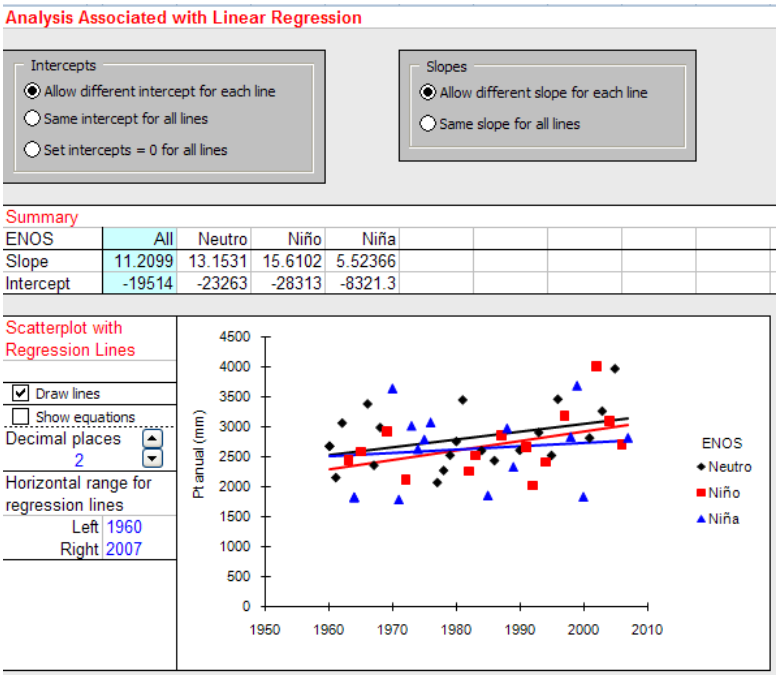


Diagrama de dispersión para Pt Vs año por episodio de ENOS.

¿Observa alguna tendencia en los datos?

Análisis de regresión lineal

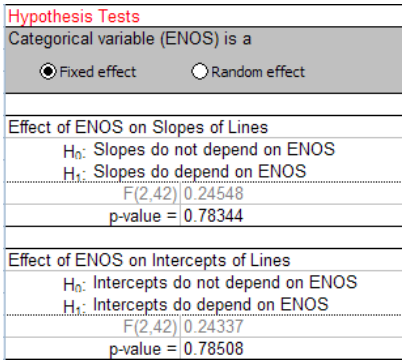


Intercepto:
1. Diferentes para cada regresión
2. Igual para todas las regresiones
3. Igual a cero

Pendiente:
1. Diferentes para cada regresión
2. Igual para todas las regresiones

Resumen:
Pendiente e intercepto para cada modelo de regresión lineal.

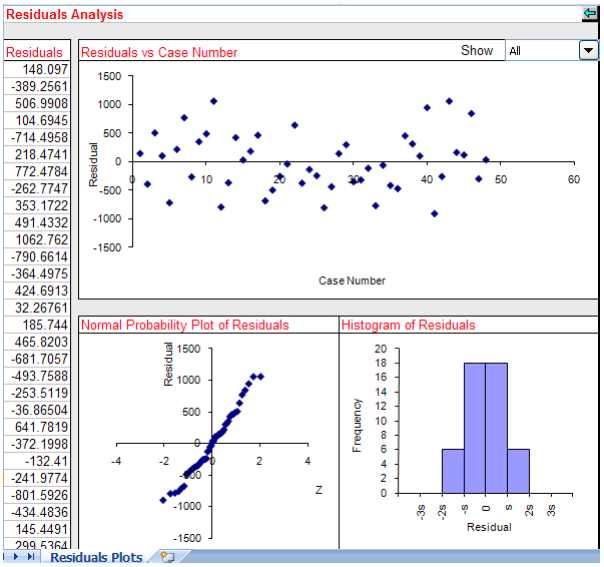
Prueba de hipótesis sobre efecto de ENOS en Pt anual



Se eligió el modelo de efectos fijos pues se evalúan todos los posibles episodios de ENOS.

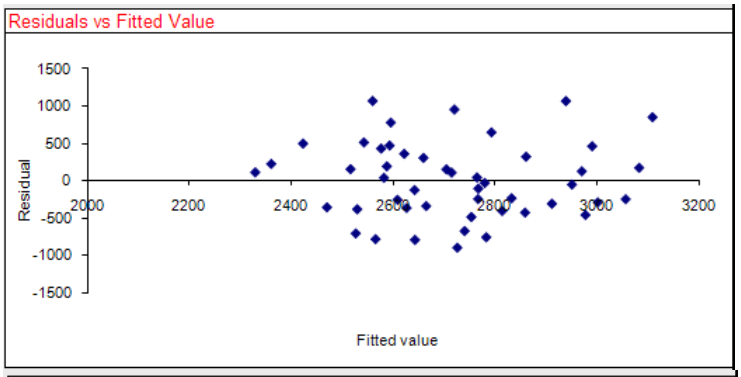
Para un alfa de 0.05 los episodios de ENOS no tienen ningún efecto en la pendiente ni en el intercepto de la precipitación anual.

Análisis de residuos

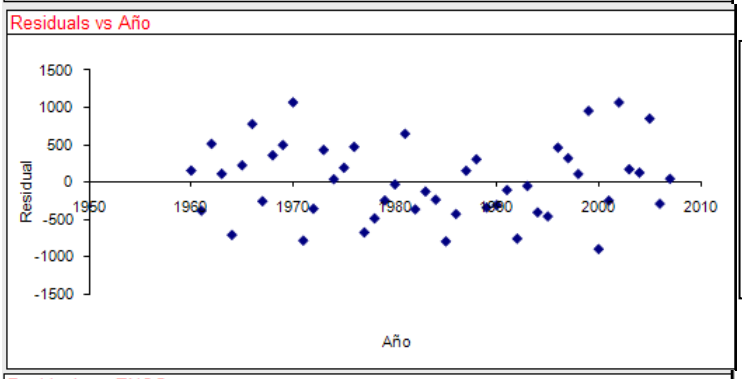


Residuos por observación
Cuando los datos son graficados en orden cronológico (como en este caso), cualquier patrón en los residuos podría indicar el efecto de una influencia externa.

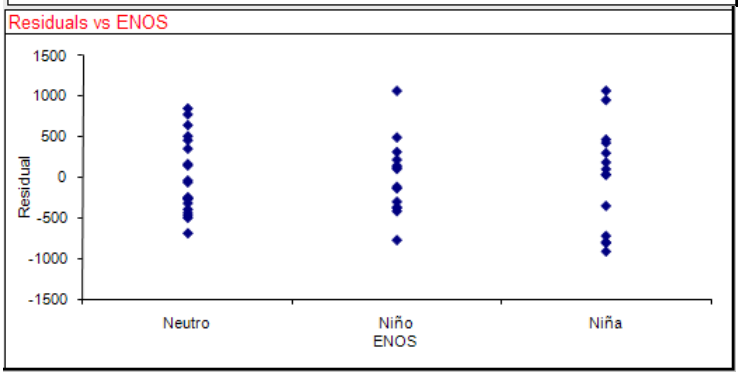
Evaluación del supuesto de normalidad de los datos.
Gráfico de probabilidad normal e histograma de residuos. Las gráficas indican que los datos no se apartan de una distribución normal.



Residuos Vs valores estimados. Evaluación del supuesto de igualad de varianzas. Los datos muestras una posible violación de este supuesto pues la variabilidad aumenta con la cantidad de precipitación.



Residuos Vs variable predictora Este gráfico permite evaluar el supuesto de homogeneidad de varianzas en el tiempo. Para este set de datos, las varianzas son semejantes y por tanto no se viola este supuesto.



Residuos Vs variable de clasificación. Este gráfico permite evaluar el supuesto de homogeneidad de varianzas entre episodios de ENOS. Para este set de datos, las varianzas son semejantes y por tanto no se viola este supuesto.

Estadísticos descriptivos por variable

More Summaries
Multiple Plots of Means

ENOS-by-ENOS Numerical Summaries for Pt anual (mm)

ENOS	ALL	Neutro	Niño	Niña
Number	48	20	14	14
Mean	2720.94	2801.01	2692.586	2634.907
St Dev	538.7043	496.0257	508.924	641.7867
Skew	0.355712	0.613203	1.21629	-0.00577
Min	1775.2	2058.8	2017.4	1775.2
Q1	2387.9	2489.125	2410.075	1960.675
Median	2684.85	2704.05	2617.4	2788
Q3	3013.75	3099.725	2899.05	2991.025
Max	4000.8	3954.2	4000.8	3669.2

ENOS-by-ENOS Numerical Summaries for Año

ENOS	ALL	Neutro	Niño	Niña
Number	48	20	14	14
Mean	1983.5	1981.6	1986.214	1983.5
St Dev	14	14.42731	14.42849	13.51779
Skew	0	0.03001	-0.32276	0.349656
Min	1960	1960	1963	1964
Q1	1971.75	1967.75	1974.5	1973.25
Median	1983.5	1980.5	1989	1980.5
Q3	1995.25	1993.5	1996.25	1995.75
Max	2007	2005	2006	2007

Gráfico de medias

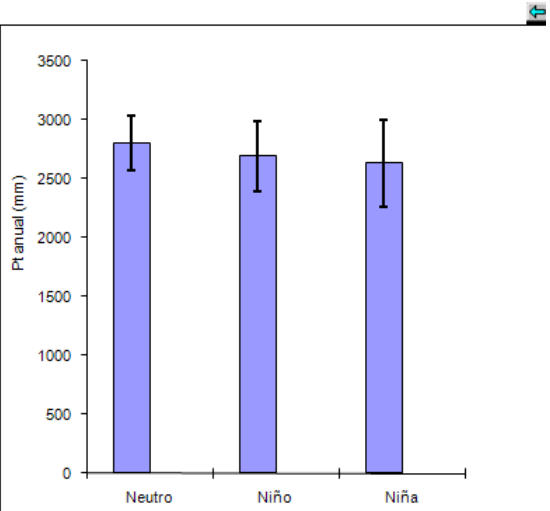
Multiple Plots of Means

Variable	Use axis on		
	Left	Right	None
Pt anual (m	◉	○	○
Año	○	○	◉

Axis Labels
Left: Pt anual (mm)
Right: Año

Style: ◉ Bar ○ Line

Error bars: ○ None ○ Range ○ ± Stdev ○ ± 2 Stdev ○ ± Stderr ○ ± 2 Stderr ◉ Confidence Interval
Level 0.95



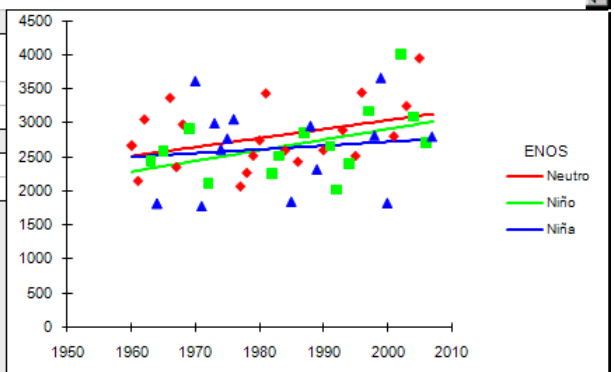
Ajuste de funciones

Function Fitting

Fit Functions

Horizontal Range for curves
Left: 1960
Right: 2007

Independent Variable
Año



ENOS	Equation
Neutro	anual (mm) = 13.15*Año-23263.17
Niño	anual (mm) = 15.61*Año-28312.67
Niña	anual (mm) = 5.52*Año-8321.27

Suavizado

Smoothing
- Moving Average
- Means with Error Bars
- Locally Weighted Regression

Moving Average

Use: ◉ Mean ○ Median

Period: 5

☑ Show data points

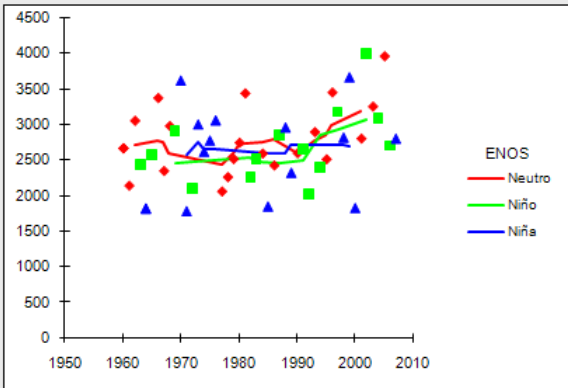
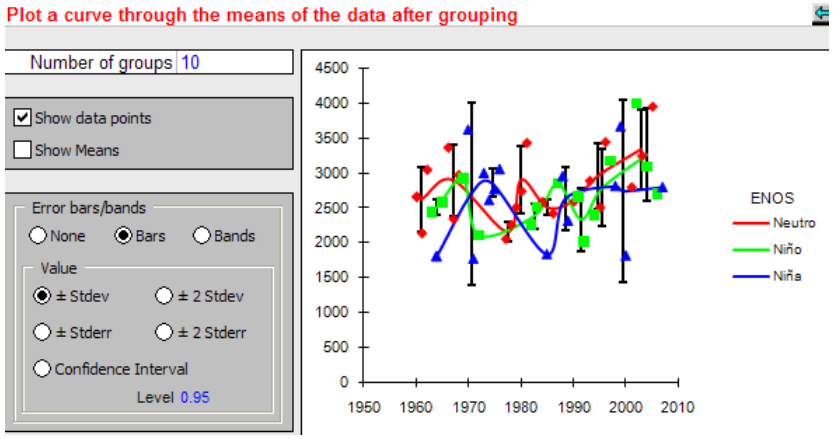
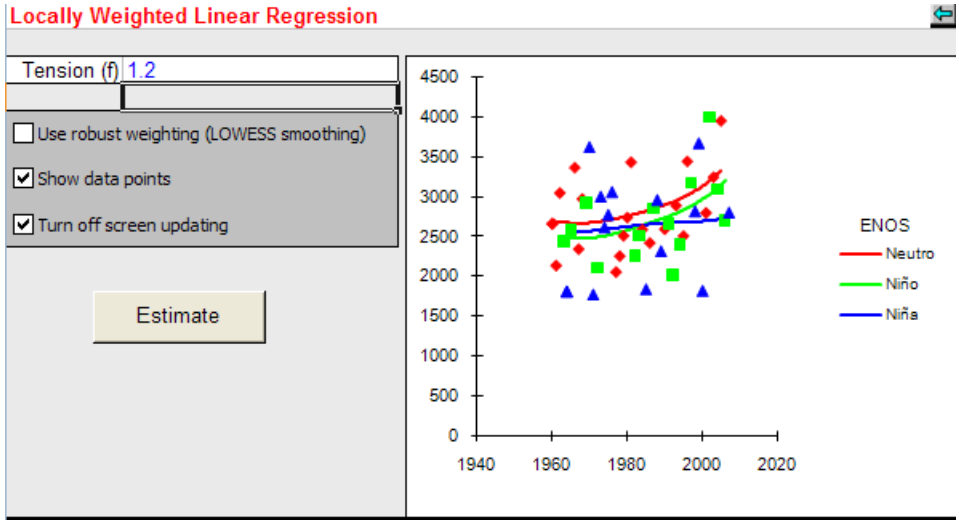


Gráfico de media para grupos



Para mayores detalles ver pág. 52.

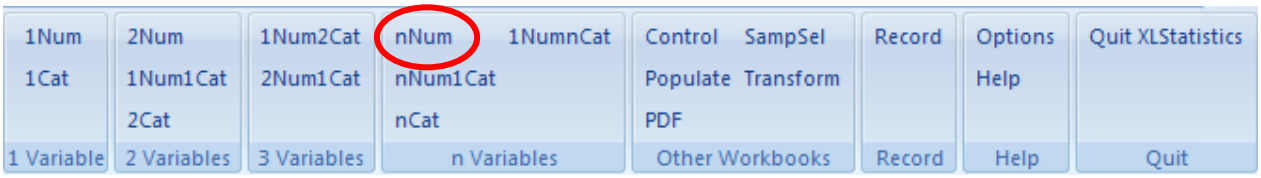
Regresión localmente ponderada



Para mayores detalles ver pág. 52 y 53.

Análisis de tres o más variables numéricas

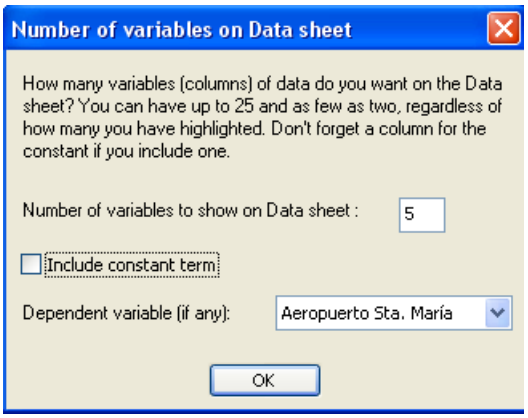
En la presente sección se ilustra el uso de XLSTatistics para el análisis de tres o más variables numéricas. En el menú grafico de XLStatisitics corresponde a **nNum**.



Nota: Esta sección del tutorial asume que usted ya activó el complemento de XLStatisticis (XLStatistics.xlam ó XLStats.xls); de no ser así, ver página 23.

1. Abra el archivo "xlstats_tutorial.xlsx" y seleccione de la hoja **nNum** las columnas Aeropuerto Sta. María, Coronado, C. Quesada, La Marina y Zarcero.

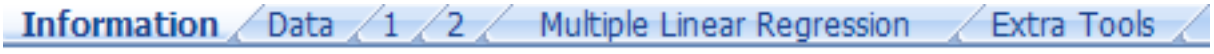
2. Haga un clic sobre XLStatistics y luego otro clic sobre nNum. Observe que el programa abre el libro de cálculo **nNum.xls** y copia los datos seleccionados a dicho libro.



Variable dependiente: Aeropuerto Juan Sta. María

No incluir constante.

En la sección inferior de la hoja de cálculo **nNum.xls** usted observará seis hojas de cálculo:



1) *Information*: Información. Esta hoja contiene información general sobre la organización de este libro de trabajo.
2) *Data*: Datos y estadísticos descriptivos
3) *1:* Estadísticos y grafica por variable
4) *2:* Análisis de correlación y regresión lineal
5) *Multiple Linear regression*: Regresión lineal multiple.
6) *Extra tolos:* Otras herramientas.

Nota: Recuerde que **SOLO** debe modificar el contenido de las celdas con números o textos en color azul.

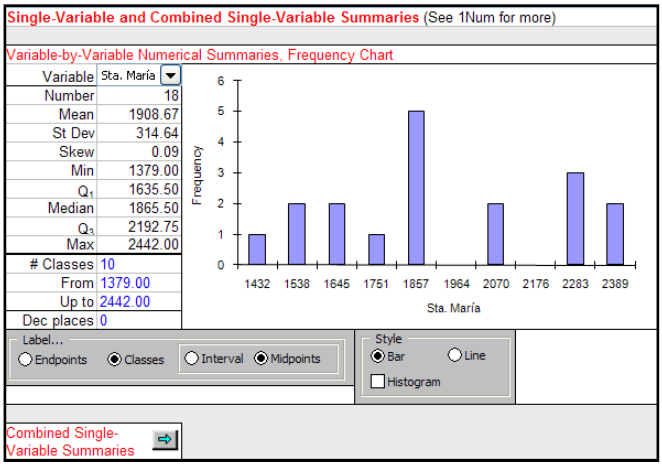
Datos

Lista las variables que usted desea analizar

Data

Aeropue	Coronad	C. Ques	La Marin	Zarcero
2442	3041	5150	4565	2527
1869	2863	5070	3215	2037
1379	1626	4711	2984	2129
1518	1819	2983	2397	1544
1900	2310	4793	3729	2367
1862	1669	3592	2862	2024
2078	2472	4715	4325	1914
1606	2337	4875	3766	1385
1724	2223	3723	3656	1352
1597	2177	4743	4164	1158
1843	2294	6046	6327	1183
1539	2091	3940	4623	1785

1: Estadísticos y grafica por variable



Usted debe elegir la variable que desea visualizar.

Etiqueta eje X
Punto medio de clase
Clases (Intervalo, Punto medio)
Estilo de gráfico
- Barras
- Líneas
- Histograma

Estadísticos descriptivos por variable

Numerical Summaries

	Sta. María	Coronado	Quesada	La Marina	Zarcero
Number	18	18	18	18	18
Mean	1908.667	2329.667	4636.222	4013.333	1847.444
St Dev	314.6442	441.2392	914.0595	915.7002	412.9346
Skew	0.091986	0.26337	0.494067	0.637944	-0.23238
Min	1379	1626	2983	2397	1158
Q_1	1635.5	2112.5	3952.5	3632	1559.75
Median	1865.5	2301	4758	3897.5	1969
Q_3	2192.75	2497.5	4994.25	4523.75	2113.5
Max	2442	3132	6837	6327	2527

Estadísticos descriptivos por variable
¿Cuál es la estación más lluviosa?
¿En cuál estación es más variable la lluvia anual?

Resumen grafico por variable

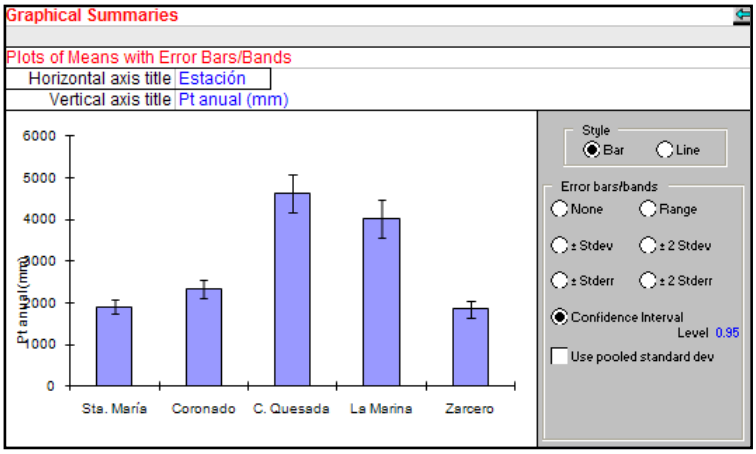


Diagrama de barras para la media.
Estilo del gráfico: Barras, línea, 3D
Barras/bandas de error: ninguna, rango, ± 1S, ± 2S, ± 1 E.E., ± 2E.E., intervalo de confianza (recuerde definir el valor del nivel de confianza)

Utilizar desviación estándar media del set de datos

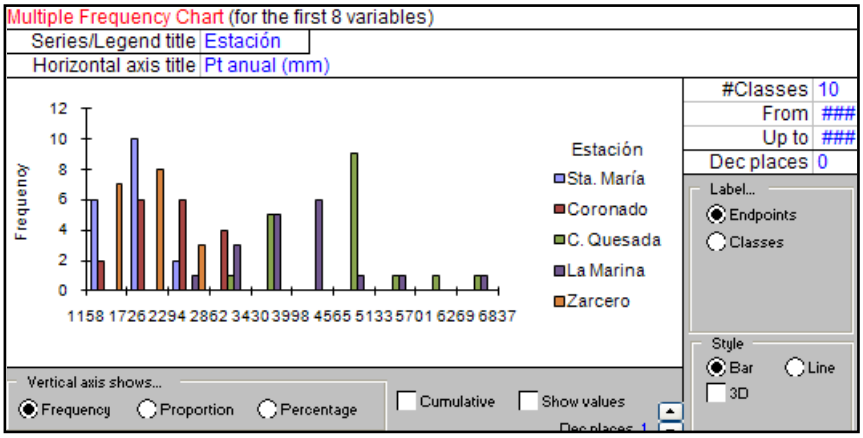


Diagrama de frecuencia multivariable

Etiquetas eje X
Límite superior de clase
Clases: intervalo, punto medio

Etiquetas eje Y
Frecuencia absoluta, porcentaje, proporción, acumulada

2: Análisis de correlación y regresión lineal

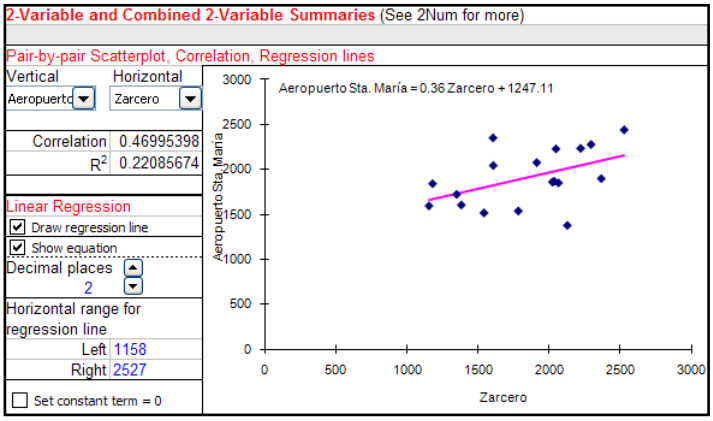
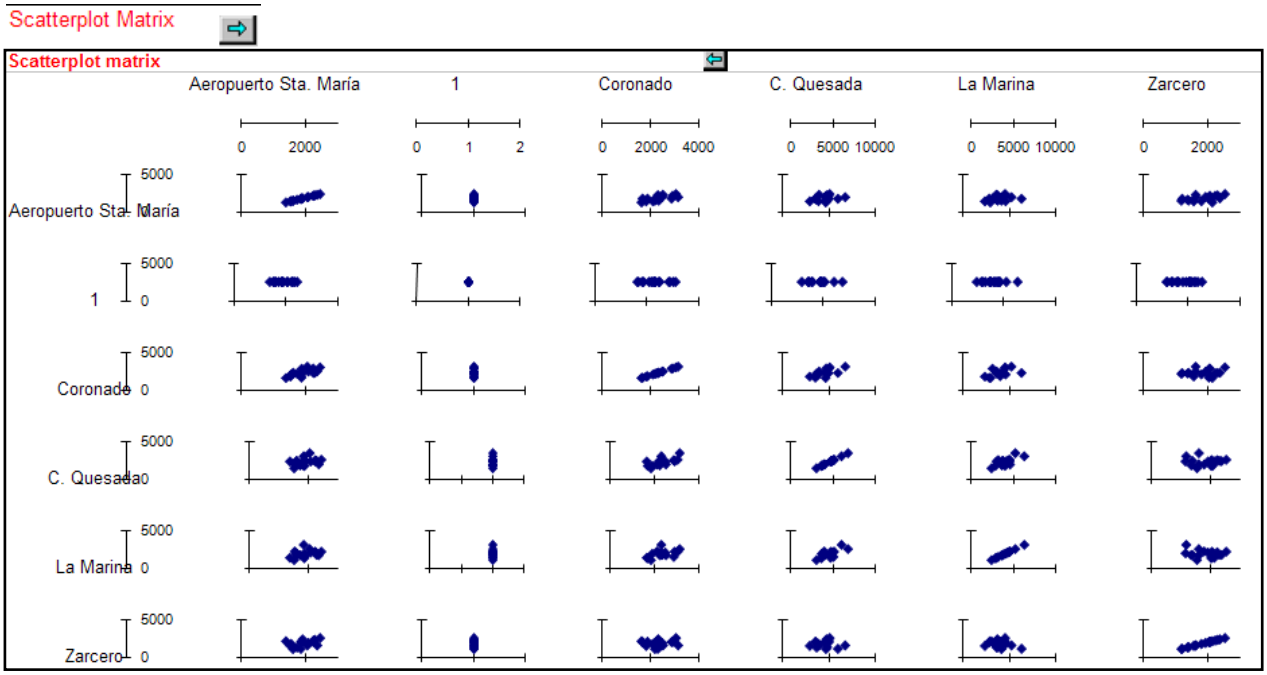


Diagrama de dispersión para dos variables. El programa ajusta una regresión lineal al set de datos y calcula el coeficiente de correlación y de determinación.

Usted debe elegir las variables a graficar y analizar

Matriz de correlación gráfica



Este gráfico permite determinar visualmente el tipo y grado de correlación entre las variables analizadas.

Matriz de correlación lineal (Coef. Pearson)

Correlation Matrix

Correlation matrix

Aeropuerto	Sta. María	Coronado	Quesada	La Marina	Zarcero
Aeropuerto	1	0.643569	0.302826	0.313389	0.469954
Coronado	0.643569	1	0.650061	0.48423	0.120784
C. Quesad	0.302826	0.650061	1	0.699996	-0.01337
La Marina	0.313389	0.48423	0.699996	1	-0.20938
Zarcero	0.469954	0.120784	-0.01337	-0.20938	1

La matriz de correlación lineal permite determinar la intensidad y dirección de la correlación lineal entre las variables analizadas.

¿Cuáles estaciones muestran la mayor correlación?

Diagrama de dispersión multivariable

Multi-Line Plot

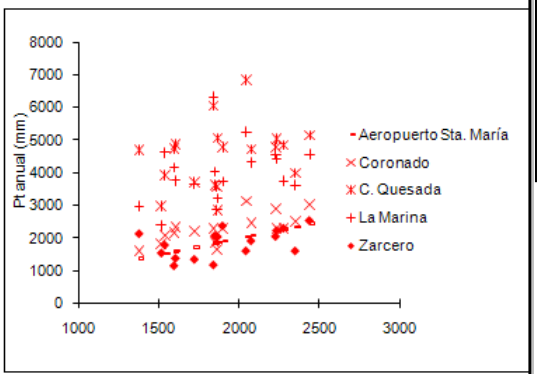
Multi-line Chart with Many Dependent Variables & 1 Independent Variable

Independent variable: Aeropuerto Sta

Left-hand axis label: Pt anual (mm)

Right-hand axis label: Y2

Variable	Use axis on		
	Left	Right	None
Aeropuerto S	◉	○	○
Coronado	◉	○	○
C. Quesada	◉	○	○
La Marina	◉	○	○
Zarcero	◉	○	○



En esta grafica la variable dependiente (eje y) corresponde a la lluvia de la estación Aeropuerto Sta. María y la variable independiente (eje X) a la lluvia en las otras estaciones.

Regresión lineal multiple

Linear Regression Analysis

Regression Equation, etc

Aeropuerto Sta. María = 0.49 Coronado - 0.115 C. Quesada + 0.136 La Marina + 0.407 Zarcero

n	18	s	213.241	d	2.333446
R^2		Adj R^2		Raw R^2	0.990535

Tests, etc, for Individual Coefficients

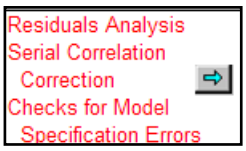
Variable	Coeff. (est.)	Std Err	H_0: Coefficient = 0 / H_1: Coefficient ≠ 0 T	p-value	Confidence Ints. Level 0.95 Lower	Upper
Coronado	0.489927	0.153833	3.18481	**0.006617**	0.159989	0.819865
C. Quesad	-0.11503	0.091058	-1.26326	0.227128	-0.31033	0.08027
La Marina	0.135503	0.078684	1.722119	0.107057	-0.03326	0.304263
Zarcero	0.407149	0.10415	3.909237	**0.001573**	0.183768	0.63053

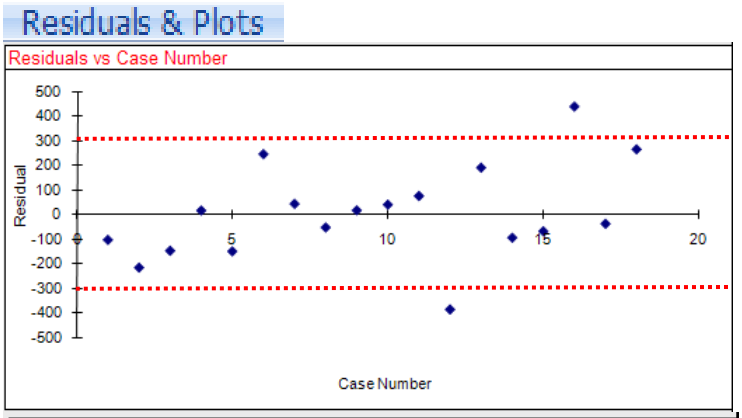
El programa utiliza la primera variable como dependiente y las otras como predictoras.

Para un alfa de 0.05, solo las variables Coronado y Zarcero son estadísticamente significativas. Por tanto, el modelo debería ajustarse nuevamente pero utilizando solo dichas variables.

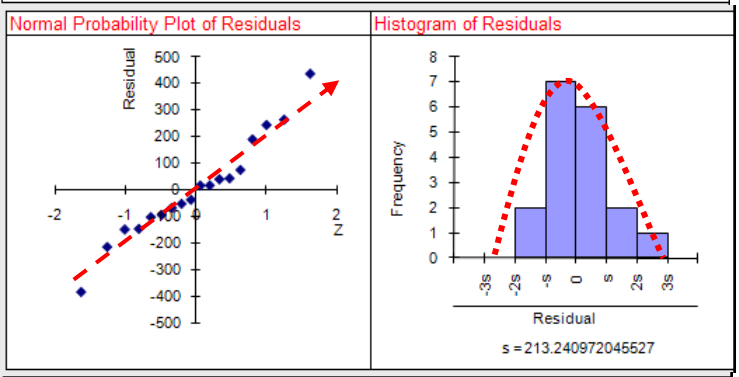
Regresión equation: ecuación de regresión
d: “d” de Durbin-Watson
Raw R^2: coeficiente de determinación (variabilidad explicada por el modelo)
Coeff. (est.): coeficientes de regresión (pendientes)
p-value: valor de “p” para prueba de hipótesis de cada coeficiente.
Confidence Ints: intervalos de confianza al 95% para los coeficientes de regresión.

Análisis de residuos



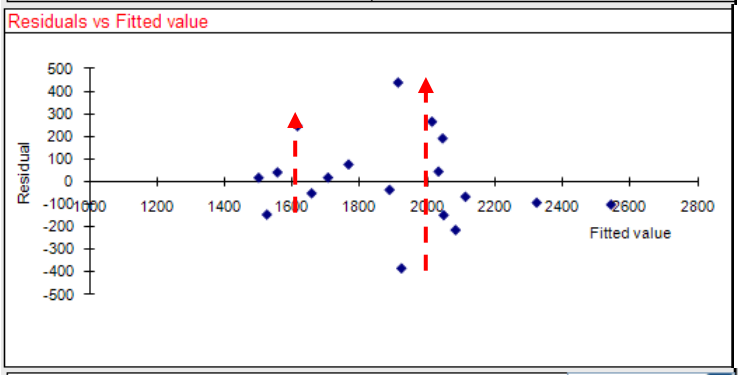


Residuos por observación
Cuando los datos son graficados en orden cronológico (como en este caso), cualquier patrón en los residuos podría indicar el efecto de una influencia externa.

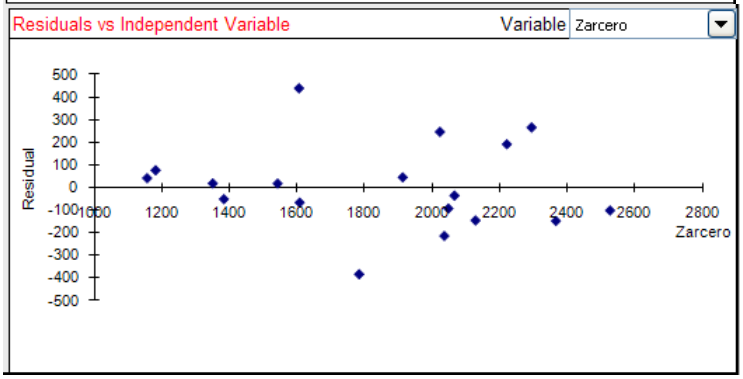


Evaluación del supuesto de normalidad de los datos.

Gráfico de probabilidad normal e histograma de residuos. Las gráficas indican que los datos no se apartan de una distribución normal.



Residuos Vs valores estimados. Evaluación del supuesto de igualad de varianzas. Los datos muestran una posible violación de este supuesto pues la variabilidad es mayor para lluvia anual de aproximadamente 2000 mm.



Residuos Vs variable predictora. Usted debe seleccionar la variable que desea graficar.

Usted debe elegir los residuos de la variable predictora que desea analizar.

Pruebas sobre residuos

Tests & Analysis

Tests and Analyses with Residuals

Jarque-Bera Test for Normality
H_0: Residuals are normally distributed
H_1: Residuals are not normally distributed

JB	0.546
p-value =	0.761

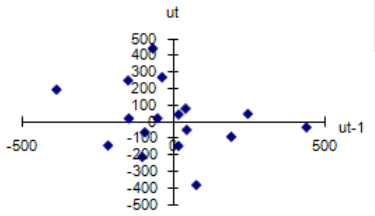
Ramsey RESET Test for Model Specification Error

Serial Correlation of Residuals

Durbin-Watson d Statistic for Serial Correlation
H_0: Residuals are not serially correlated
H_1: Residuals are serially correlated

d	2.3334

Correction for Serial Correlation

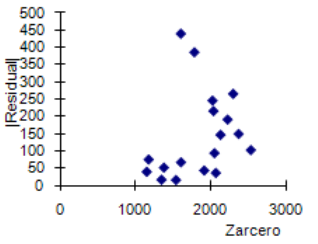


Glejser Test for Heteroscedasticity

Fit: |Residual| = constant + Zarcero 0.5

H_0: Variance of residuals doesn't depend on Zarcero
H_1: Variance of residuals depends on Zarcero

p-value =	0.262



Evaluación del supuesto de normalidad de los datos.

Para un nivel de significancia de 0.05, la prueba de Jarque-Bera indica que los datos son normales.

Evaluación del supuesto de independencia de residuos.

Evaluación del supuesto de igualad de varianzas.
Usted debe realizar la prueba de hipótesis para cada estación.

Corrección por correlación serial (resago 1)

Correction for Serial Correlation

Correction for First Order Serial Correlation

Regression Equation, etc

Aeropuerto Sta. María* = 0.543 Coronado* - 0.173 C. Quesada* + 0.167 La Marina* + 0.421 Zarcero*

Decimal places 3

n	17	s	209.9295	d	1.917192
R^2	0.66552	Adj R^2	0.588332	Raw R^2	0.994383

Tests, etc, for Individual Coefficients

Variable	Coeff. (est.)	Std Err	H_0: Coefficient = 0 / H_1: Coefficient ≠ 0 T	p-value	Confidence Ints. Level 0.95 Lower	Upper
Coronado*	0.54281	0.155811	3.483769	0.003651	0.208628	0.876991
C. Quesac	-0.17331	0.096788	-1.79061	0.094996	-0.3809	0.03428
La Marina	0.167131	0.067797	2.465165	0.027237	0.021721	0.312542
Zarcero*	0.420989	0.089077	4.726147	0.000325	0.229939	0.612039

Coefficient of Autocorrelation: -0.29234

Cochran-Orcutt Iterative Procedure

Another iteration | Reset

Iteration 153

Corrección por autocorrelación de resago 1.

Prueba sobre especificación del error del modelo (Ramsey, 1969).

Ramsey RESET Test for Model Specification Error

Ramsey's RESET test for model specification error

Terms to add	User defined	Fitted values
☑ $(\text{Fitted values})^2$	49.79103892	2479.147556
☐ $(\text{Fitted values})^3$	45.57340053	2076.934836
☐ ln(Fitted values)	39.35838779	1549.08269
☐ $e^{\text{Fitted values}}$	39.71516051	1577.293974
☐ 1/(Fitted values)	45.03572264	2028.216314
☐ User defined	40.67983272	1654.84879
	45.00134539	2025.121087
	41.1425455	1692.70905
	41.90251133	1755.820456
	40.02122183	1601.698197
	41.9433084	1759.24112
	43.95499556	1932.041635
	44.93883875	2019.499228
	47.88886911	2293.343785
	45.55385006	2075.153255
	44.0491573	1940.328259
	43.57262735	1898.573854
	44.68720079	1996.945914

Hypothesis Test

H_0: Added terms are not significant

H_1: Added terms are significant

$F(1, 12)$ 0.3843486

p-value = 0.5468774

Prueba sobre especificación del error del modelo de regresión de Ramsey (Ramsey, 1969)

Ramsey, J.B. 1969. Tests for Specification Errors in Classical Linear Least Squares Regression Analysis. J. Roy. Statist. Soc. B., 31(2), 350–371.

Somete a prueba varias combinaciones no lineales de los valores estimados para determinar si ayudan a explicar la variable endógena. El supuesto de la prueba es que, si las combinaciones no lineales de las variables explicativas tienen algún grado de poder para explicar la variable endógena, entonces el modelo no está correctamente especificado.

Análisis de varianza (significancia de modelo de regresión)

ANOVA Comparison of Models

Analysis of Variance

H_0: All coefficients in model are zero

H_1: Not all coefficients in model are zero

p-value = 5.4E-14

ANOVA Table

Source	DF	SS	MS	F	p-value
Regression	4	66620564	16655141	366.27477	5.4E-14
Residual	14	636603.97	45471.712		
Total (uncorrected)	18	67257168			

Para un nivel de significancia de 0.05, el análisis de varianza indica que el modelo es significativo; o sea existe una relación entre la variable dependiente y las variables predictoras.

Sin embargo la prueba no permite afirmar que todos los coeficientes de regresión sean diferentes de cero.

Comparación de modelos

Comparison of Models

Variables in Models	
Full	Drop
Coronado	☑
C. Quesad	☐
La Marina	☐
Zarcero	☐

Hypothesis Tests

Regression
H_0: Coefficients in reduced model are all zero
H_1: Coefficients in reduced model are not all zero
p-value = 2.235E-14

Dropped Terms
H_0: Aeropuerto Sta. María doesn't depend on dropped terms
H_1: Aeropuerto Sta. María depends on dropped terms
p-value = 0.0066175

ANOVA Table

Source	DF	SS	MS	F	p
Reduced	3	66159344	22053115	484.98536	2.235E-14
Dropped	1	461220.1	461220.1	10.143012	0.0066175
Residual	14	636603.97	45471.712		
Total	18	67257168			

Esta hoja de trabajo le permite probar por la significancia de cada uno de los coeficientes de regresión; así como por combinaciones de variables predictoras (e.g. Coronado y C. Quesada, La Marina y Zarcero, Coronado y Zarcero, etc.)

Comparison of Models

Variables in Models	
Full	Drop
Coronado	☐
C. Quesad	☑
La Marina	☐
Zarcero	☐

Hypothesis Tests

Regression
H_0: Coefficients in reduced model are all zero
H_1: Coefficients in reduced model are not all zero
p-value = 2.146E-14

Dropped Terms
H_0: Aeropuerto Sta. María doesn't depend on dropped terms
H_1: Aeropuerto Sta. María depends on dropped terms
p-value = 0.2271282

ANOVA Table

Source	DF	SS	MS	F	p
Reduced	3	66547999	22182666	487.83442	2.146E-14
Dropped	1	72565.187	72565.187	1.5958314	0.2271282
Residual	14	636603.97	45471.712		
Total	18	67257168			

Predicción o estimación

Para un alfa de 0.05, solo las variables Coronado y Zarcero son estadísticamente significativas y por tanto, en un caso real, el modelo debería ajustarse nuevamente pero utilizando solo dichas variables y luego utilizarlo para realizar predicciones.

Prediction

Prediction

Prediction point	
Variable	Value
Coronado	3041
C. Quesada	5150
La Marina	4565
Zarcero	2527

		P.I. for Aeropuerto Sta. María	
		Level	0.95
	Value	Lower	Upper
Aeropuerto Sta.	2544.900376	2047.913813	3041.886939

		C.I. for Mean Aeropuerto Sta. María	
		Level	0.95
	Value	Lower	Upper
Mean Aeropuert	2544.900376	2350.424748	2739.376004

Digite los valores de "X" para estimar el valor de "Y".

Otras herramientas

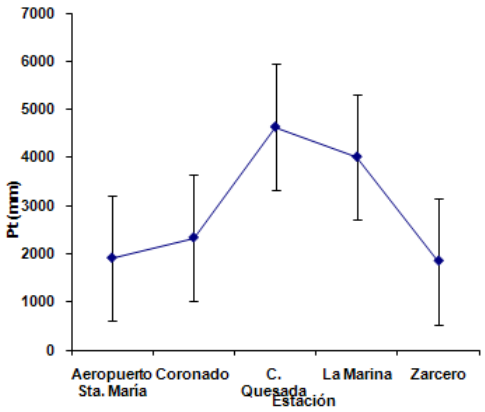
Mediciones repetidas

Repeated Measures

Plots for Repeated Measures

Show
One case
Many cases
Means
Error bars/bands
None
Range
± Stdev
± 2 Stdev
± Stderr
± 2 Stderr
Confidence Interval
Use pooled standard dev
Level 0.95
Show
Bars
Bands
Box-and-whisker

Dependent variable name: Pt (mm)
Within-Subjects Factor name: Estación
Subject Factor name: Año



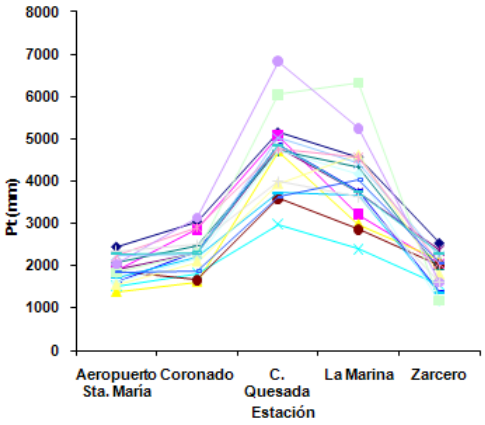
Mediciones repetidas. En cada estación se mide la lluvia durante 18 años.

Intervalo de confianza 95%.

Plots for Repeated Measures

Show
One case
Many cases First 1 Last 18
Means
Box-and-whisker

Dependent variable name: Pt (mm)
Within-Subjects Factor name: Estación
Subject Factor name: Año



Análisis de varianza

Analysis of Variance for Repeated Measures

Dependent variable name: Pt (mm)
Within-Subjects Factor name: Estación
Subject Factor name: Año

ANOVA Table

Source	DF	SS	MS	F	p-value
Estación	4	119858354	29964589	100.98878	7.435E-28
Año	17	16173313	951371.34	3.206379	
Error	68	20176421	296712.07		
Total	89	156208088			

Hypothesis Test

H_0: Pt (mm) does not depend on Estación
H_1: Pt (mm) depends on Estación

p-value = 7.435E-28

Regresión no lineal

Non-linear Regression

Non-linear Least Squares Regression

Equation María = a*(X1-b)*(X2-c)

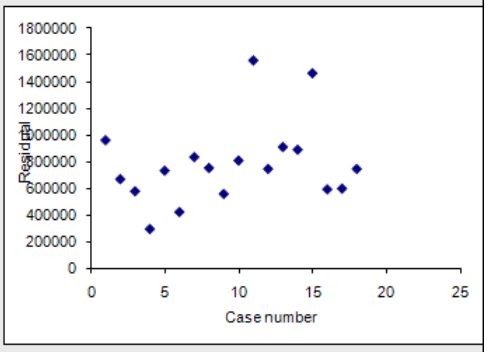
Independent variables
Coronado
C. Quesada
La Marina
Zarcero

Parameters	
a	0.04
b	-50
c	-70

Estimate

Close other workbooks

R^2 -2E+05



Note: Excel's Solver Add-in is used to determine the parameters for the specified equation when Estimate is clicked. Solver must therefore be installed and must appear on the Tools menu. It may also be useful to check that Solver is actually working. Do this by selecting Solver from the Tools menu - you should see Solver's dialog window appear. Then click 'Close'.

Análisis de tres o más variables numéricas y una nominal

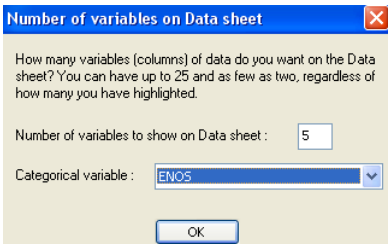
En la presente sección se ilustra el uso de XLSTatistics para el análisis de tres o más variables numéricas y 1 nominal. En el menú grafico de XLStatisitics corresponde a **nNum1Cat**.

Nota: Esta sección del tutorial asume que usted ya activó el complemento de XLStatisticis (XLStatistics.xlam ó XLStats.xls); de no ser así, ver página 23.

1. Abra el archivo "xlstats_tutorial.xlsx" y seleccione las columnas Batán_(Pt mm), Siquirres_Pt (mm), Suiza_Pt (mm), Moravia_Pt (mm) y ENOS.

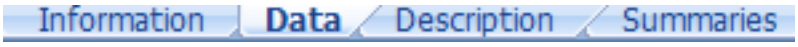
2. Haga un clic sobre XLStatistics y luego otro clic sobre 

. Observe que el programa abre el libro de cálculo **nNum1Cat.xls** y copia los datos seleccionados a dicho libro.



Seleccione ENOS como la variable nominal o de clasificación.

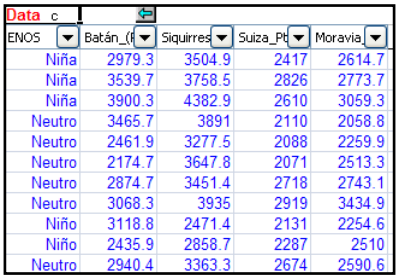
En la sección inferior del libro de cálculo **nNum1Cat.xls** usted observará cuatro hojas de cálculo:



1) *Information*: Información. Esta hoja contiene información general sobre la organización de este libro de trabajo.
2) *Data*: Datos.
3) *Description:* Estadísticos descriptivos y gráfico de barras-histograma.
4) *Resumen:* Gráfico de barras comparativo

Nota: Recuerde que **SOLO** debe modificar el contenido de las celdas con números o textos en color azul.

Datos



Data c

ENOS	Batán_(P	Siquirres	Suiza_Pt	Moravia
Niña	2979.3	3504.9	2417	2614.7
Niña	3539.7	3758.5	2826	2773.7
Niña	3900.3	4382.9	2610	3059.3
Neutro	3465.7	3891	2110	2058.8
Neutro	2461.9	3277.5	2088	2259.9
Neutro	2174.7	3647.8	2071	2513.3
Neutro	2874.7	3451.4	2718	2743.1
Neutro	3068.3	3935	2919	3434.9
Niño	3118.8	2471.4	2131	2254.6
Niño	2435.9	2858.7	2287	2510
Neutro	2940.4	3363.3	2674	2590.6

Lista de los datos

Estadísticos decriptivos y grafico

Description

Category Labels

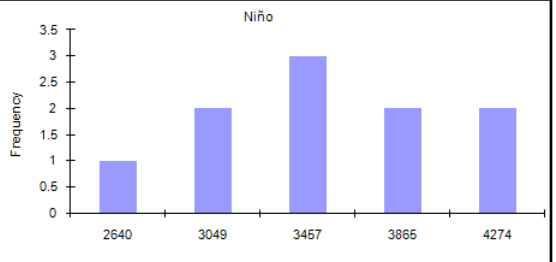
Niña | Neutro | Niño

Variable-by-Variable Numerical Summaries, Frequency Chart

Category: Niño

Variable: Batán_(Pt r

Number	10
Mean	3553.75
St Dev	654.043991
Skew	-0.1009326
Min	2435.9
Q_1	3191.25
Median	3538.6
Q_3	3951.25
Max	4477.9



Classes 5
From 2435.90
Up to 4477.90
Dec places 0

Label... Endpoints (•) Classes ( ) Interval (•) Midpoints

Style (•) Bar ( ) Line [ ] Histogram

Grafico y estadísticos descriptivos por estación y fase de ENOS.

Eje X
Límite superior
Clases
Intervalo de clase
Punto medio

Estilo:
Barra, línea, histograma

Numerical Summaries

All

	Batán_(Pt mm)	res_Pt (mm)	iiza_Pt (mm)	avia_Pt (mm)
Number	34	34	34	34
Mean	3219.82353	3790.87059	2586.02941	2759.12941
St Dev	617.629436	699.02587	413.084369	533.377926
Skew	-0.1635015	0.11994382	0.20543099	0.51809992
Min	1959.5	2337.2	1769	1820.9
Q_1	2949.05	3369.975	2218.75	2445.8
Median	3260.75	3769.8	2657.5	2723.9
Q_3	3618.975	4259.3	2865	3034.3
Max	4477.9	5435.9	3486	4000.8

Niña

	Batán_(Pt mm)	res_Pt (mm)	iiza_Pt (mm)	avia_Pt (mm)
Number	9	9	9	9
Mean	3171.75556	3672.15556	2597.11111	2652.64444
St Dev	488.577571	544.959765	436.159502	591.355443
Skew	-0.5623399	-0.5615544	-0.6074244	0.00493135
Min	2206.7	2632.5	1769	1820.9
Q_1	2979.3	3492.7	2417	2317.9
Median	3080	3680.9	2666	2773.7
Q_3	3539.7	4080.4	2826	2959.3
Max	3900.3	4382.9	3267	3669.2

Neutro

	Batán_(Pt mm)	res_Pt (mm)	iiza_Pt (mm)	avia_Pt (mm)
Number	15	15	15	15
Mean	3026.04667	3789.38667	2567.53333	2818.44667
St Dev	605.312788	631.215298	424.869369	506.719906
Skew	-0.5375826	-0.4615516	0.28784807	0.80615262
Min	1959.5	2337.2	2071	2058.8
Q_1	2668.3	3376.65	2158.5	2512
Median	3068.3	3874.4	2674	2743.1
Q_3	3471.35	4247.9	2917	3070.35
Max	3803.7	4816.2	3393	3954.2

Niño

	Batán_(Pt mm)	res_Pt (mm)	iiza_Pt (mm)	avia_Pt (mm)
Number	10	10	10	10
Mean	3553.75	3899.94	2603.8	2765.99
St Dev	654.043991	938.396094	418.011908	561.804166
Skew	-0.1009326	0.25652267	0.95324453	1.07001559
Min	2435.9	2471.4	2131	2017.4
Q_1	3191.25	3332.175	2281	2428.85
Median	3538.6	3871.9	2559	2679.8
Q_3	3951.25	4273.525	2811	3030.15
Max	4477.9	5435.9	3486	4000.8

Estadísticos descriptivos para todos los datos y por fase de ENOS.

Grafico de barras comparativas

Plots of Means with Error Bars/Bands

Horizontal axis title: Estación

Vertical axis title: Pt anual (mm)

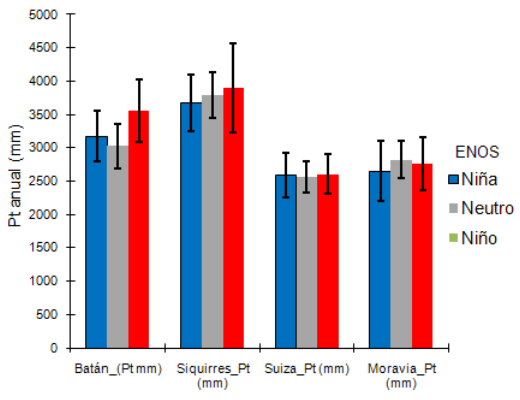
Style (•) Bar ( ) Line

Error bars/bands

Show (•) Bars ( ) Bands

( ) None
( ) Range
( ) ± Stdev
( ) ± 2 Stdev
( ) ± Stderr
( ) ± 2 Stderr
(•) Confidence Interv
Level 0.95

[ ] Assume equal std devs



Este gráfico muestra la precipitación media anual para cada estación por fase de ENOS. Las líneas verticales indican un intervalo de confianza al 95%.

Usted puede elegir un grafico de barras ó de líneas.

Barras de error:

- Rango
- Desv. Estándar
- Error estándar
- Intervalo de confianza

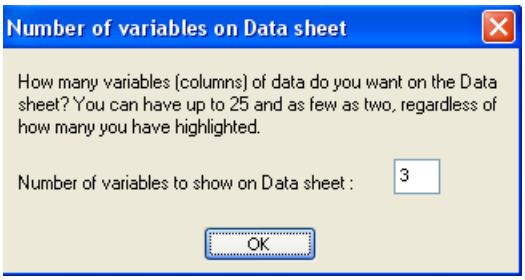
Análisis de tres o más variables nominales

En la presente sección se ilustra el uso de XLSTatistics para el análisis de tres o más variables nominales o cualitativas. En el menú grafico de XLStatisitics corresponde a **nCat**. El programa permite analizar un máximo de 8 variables.

Nota: Esta sección del tutorial asume que usted ya activó el complemento de XLStatisticis (XLStatistics.xlam ó XLStats.xls); de no ser así, ver página 23.

1. Abra el archivo "xlstats_tutorial.xlsx", vaya a la pestaña "ncat" y seleccione las columnas ENOS, Década, Clase.

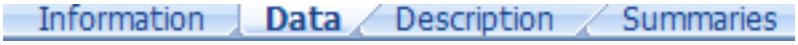
2. Haga un clic sobre XLStatistics y luego otro clic sobre nCat. Observe que el programa abre el libro de cálculo **nCat.xls** y copia los datos seleccionados a dicho libro.



El programa puede analizar un máximo de 25 columnas (variables).

OK, para continuar

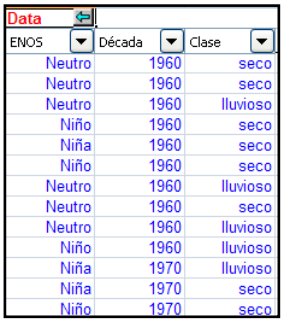
En la sección inferior de la hoja de cálculo **nCat.xls** usted observará cuatro hojas de cálculo:



1) *Information*: Información. Esta hoja contiene información general sobre la organización de este libro de trabajo.
2) *Data*: Datos
3) *Description:* Tabla dinámica
4) *Resumen:* Grafico de barras comparativo

Nota: Recuerde que **SOLO** debe modificar el contenido de las celdas con números o textos en color azul.

Datos


Data

ENOS	Década	Clase
Neutro	1960	seco
Neutro	1960	seco
Neutro	1960	lluvioso
Niño	1960	seco
Niña	1960	seco
Niño	1960	seco
Neutro	1960	lluvioso
Neutro	1960	seco
Neutro	1960	lluvioso
Niño	1960	lluvioso
Niña	1970	lluvioso
Niña	1970	seco
Niño	1970	seco

Lista de los datos

Tabla dinámica

Pivot Tables

Show: (•) Frequency ( ) % of total ( ) Column % ( ) Row %

Decimal places 2

Clase	(Todas)

Frequency	Década					
ENOS	1960	1970	1980	1990	2000	Total general
Neutro	6	3	4	4	3	20
Niña	1	6	3	2	2	14
Niño	3	1	3	4	3	14
Total general	10	10	10	10	8	48

La tabla muestra la frecuencia por episodio de ENOS y Década para la totalidad de las observaciones.

Pivot Tables

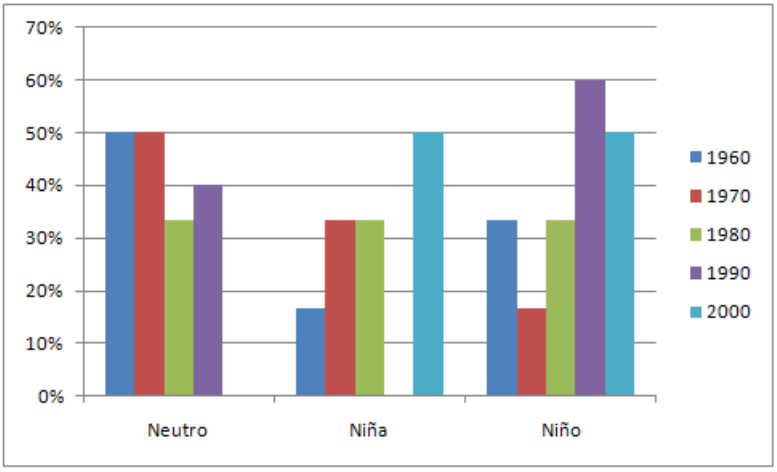
Show: ( ) Frequency ( ) % of total (•) Column % ( ) Row %

Decimal places 0

Clase	lluvioso

Column %	Década					
ENOS	1960	1970	1980	1990	2000	Total general
Neutro	75%	0%	50%	40%	50%	43%
Niña	0%	100%	25%	40%	17%	35%
Niño	25%	0%	25%	20%	33%	22%
Total general	100%	100%	100%	100%	100%	100%

La tabla muestra la frecuencia en porciento por episodio de ENOS y Década para los años clasificados como lluviosos.



Si lo desea puede expresar los resultados en forma gráfica.

Tablas y gráficos multiples

Multiple Frequency Chart and Table (showing up to 8 variables)

Show: ( ) Counts ( ) Proportions (•) Percentages

	ALL	Neutro	Niña	Niño
ENOS	100	41.67	29.17	29.17
Década	#####	#####	#####	#####
Clase	#####	#####	#####	#####

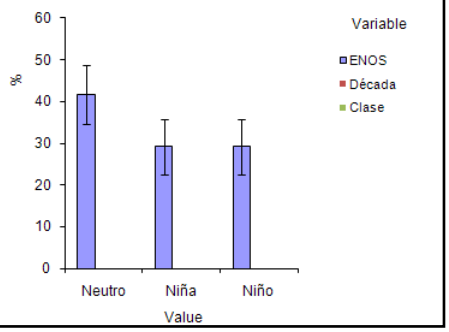
Show values Dec places 2

Legend title: Variable

Horizontal axis title: Value

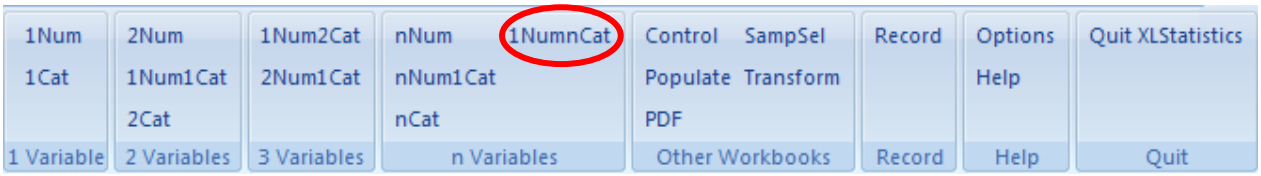
Error bars: ( ) None (•) ± Stderr ( ) ± 2 Stderr ( ) Confidence Interval Level 0.95

Style: (•) Bar ( ) Line



Análisis de una variable numérica y tres o más nominales

En la presente sección se ilustra el uso de XLSTatistics para el análisis de tres o más variables nominales o cualitativas. En el menú grafico de XLStatisitics corresponde a **1NumnCat**. El programa permite analizar un máximo de 8 variables.



Nota: Esta sección del tutorial asume que usted ya activó el complemento de XLStatisticis (XLStatistics.xlam ó XLStats.xls); de no ser así, ver página 23.

1. Abra el archivo “xlstats_tutorial.xlsx”, vaya a la pestaña “1NumnCat” y seleccione las columnas Pt anual (mm), ENOS, Década y Clase.

2. Haga un clic sobre XLStatistics y luego otro clic sobre 

. Observe que el programa abre el libro de cálculo **1NumnCat.xls** y copia los datos seleccionados a dicho libro. El programa puede analizar un máximo de 25 variables.

En la sección inferior del libro de cálculo **1NumnCat.xls** usted observará tres hojas de cálculo:

1) *Information*: Información. Esta hoja contiene información general sobre la organización de este libro de trabajo.
2) *Data*: Datos.
3) *Pivot Tables:* Tablas dinámicas.

Nota: Recuerde que **SOLO** debe modificar el contenido de las celdas con números o textos en color azul.

Datos

Data

Pt anual (r	ENOS	Década	Clase
2665	Neutro	1960	seco
2140.8	Neutro	1960	seco
3050.2	Neutro	1960	lluvioso
2434.9	Niño	1960	seco
1812.7	Niña	1960	seco
2579.9	Niño	1960	seco
3368.3	Neutro	1960	lluvioso
2346.2	Neutro	1960	seco
2975.3	Neutro	1960	lluvioso
2915.3	Niño	1960	lluvioso
3623.1	Niña	1970	lluvioso
1775.2	Niña	1970	seco
2106.2	Niño	1970	seco
3001.6	Niña	1970	lluvioso
2614.7	Niña	1970	seco
2773.7	Niña	1970	lluvioso
3059.3	Niña	1970	lluvioso
2058.8	Neutro	1970	seco

Lista de los datos.

Tablas dinámicas

Pivot Tables

Show
- Frequency
- Minimum
- Maximum
- Mean
- St Dev

Decimal places 0

Clase	(Todas)

Means	Década					
ENOS	1960	1970	1980	1990	2000	Total general
Neutro	2758	2277	2798	2862	3334	2801
Niña	1813	2808	2373	3243	2312	2635
Niño	2643	2106	2538	2562	3265	2693
Total general	2629	2579	2593	2818	3053	2721

Usted puede elegir entre frecuencia, mínimo, máximo, media y desviación estándar por clase y década.

Si lo desea puede utilizar los datos para realizar gráficos.

Pivot Tables

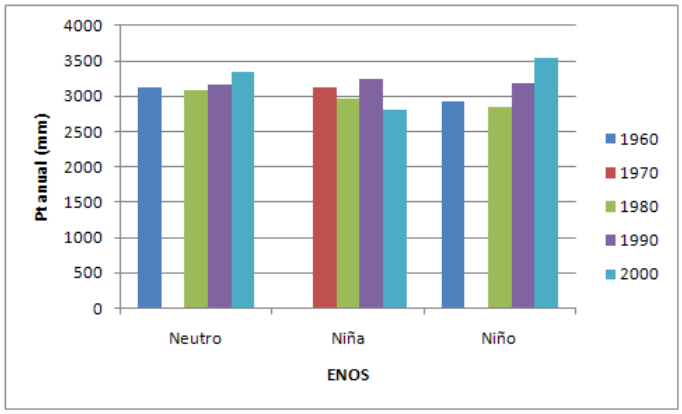
Show
- Frequency
- Minimum
- Maximum
- Mean
- St Dev

Decimal places 0

Clase	lluvioso

Means	Década					
ENOS	1960	1970	1980	1990	2000	Total general
Neutro	3131		3089	3170	3334	3191
Niña		3114	2959	3243	2802	3088
Niño	2915		2850	3175	3545	3206
Total general	3077	3114	2997	3200	3316	3159

La tabla muestra la precipitación media anual por episodio de ENOS y Década para los años clasificados como lluviosos.



Si lo desea puede expresar los resultados en forma gráfica.

Pivot Tables

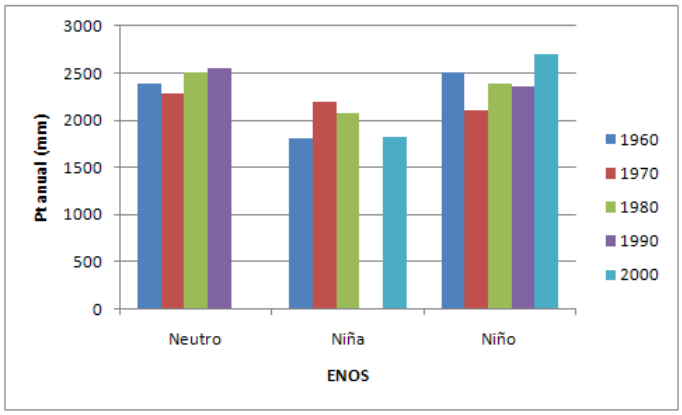
Show
- Frequency
- Minimum
- Maximum
- Mean
- St Dev

Decimal places 0

Clase	seco

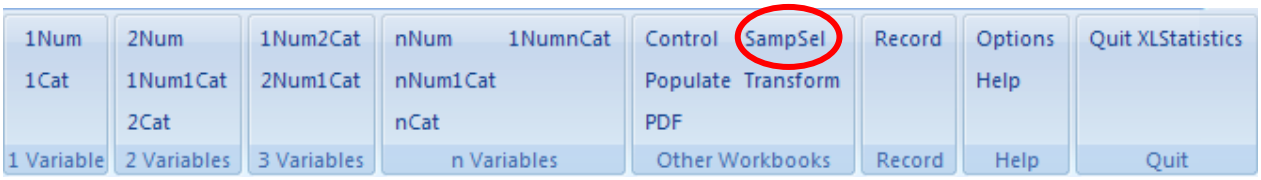
Means	Década					
ENOS	1960	1970	1980	1990	2000	Total general
Neutro	2384	2277	2508	2553		2411
Niña	1813	2195	2080		1821	2031
Niño	2507	2106	2382	2358	2705	2407
Total general	2330	2221	2323	2436	2263	2318

La tabla muestra la precipitación media anual por episodio de ENOS y Década para los años clasificados como secos.



Selección de muestras aleatorias

En la presente sección se ilustra el uso de XLSTatistics en la selección de muestras aleatorias a partir de un set de datos existente. Esta técnica se utiliza con frecuencia para balancear muestras previo a su análisis. En el menú grafico de XLStatisitics corresponde a **SampSel**.



Nota: Esta sección del tutorial asume que usted ya activó el complemento de XLStatisticis (XLStatistics.xlam ó XLStats.xls); de no ser así, ver página 23.

1. Abra el archivo "efecto_borde.xlsx". Este archivo contiene las siguientes variables.
   A. d (cm) interior *Variable cuantitativa.*
   B. d (cm) borde *Variable cuantitativa.*
   C. muestra no balanceada *Variables cuantitativa y cualitativa*
   D. Muestras aleatorias con reemplazo *Variable cuantitativa.*
   E. muestra balanceada *Variables cuantitativa y cualitativa*

2. Haga un clic sobre XLStatistics y luego otro clic sobre SampSel.

2. Seleccione y copie la columna d (cm), luego pegue los datos en la columna borde. Observe que el programa abre el libro de cálculo **SampSel.xls** y copia los datos seleccionados a dicho libro.

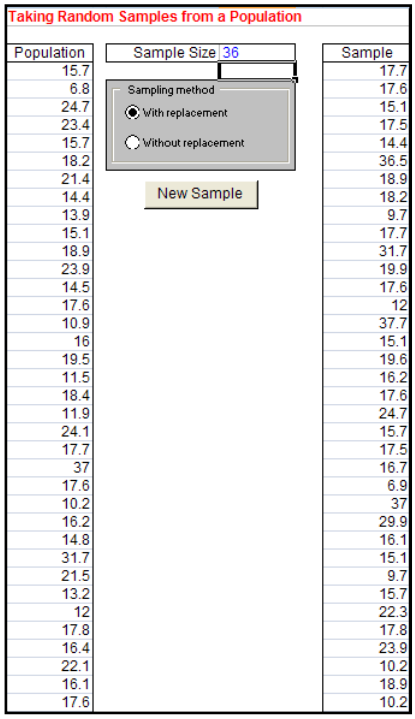
En la sección inferior de la hoja de cálculo **SampSel.xls** usted observará cinco hojas de cálculo:

1) *Information*: Información. Esta hoja contiene información general sobre la organización de este libro de trabajo.
2) *Make Data – 1Var*: Crear set de datos para 1 variable.
3) *Make Data – 2Var*: Crear set de datos para 2 variables.
4) *Random Numbers*: Crea números aleatorios para las siguientes distribuciones: Uniforme, binomial, Poisson, exponencial, lognormal, gamma, beta.
5) *Sample Selection*: Seleccionar una muestra aleatoria

Nota: Recuerde que **SOLO** debe modificar el contenido de las celdas con números o textos en color azul.

A continuación utilizaremos la hoja "*Sample Selection"* para seleccionar una muestra aleatoria con reemplazo de 36 observaciones de las 66 observaciones de arboles de borde.

Usted puede elegir entre una muestra aleatoria con y sin reemplazo; así como el tamaño de la muestra. En este caso hemos elegido 36 observaciones porque existen 36 árboles de borde en la parcela y el objetivo del muestreo es preparar los datos para realizar una prueba de hipótesis sobre el efecto de borde en la parcela.



Usted puede copiar los datos de la columna “simple” a otra hoja de Excel para su posterior análisis.

Presionando “New Sample” obtiene otra muestra aleatoria con reemplazo de 36 observaciones.

Las “muestras aleatorias con reemplazo” del archivo efecto_borde.xls se obtuvieron con esta herramienta de XLStatistics.

Anexo 1: Análisis exploratorio de datos: Estadísticos descriptivos y análisis gráfico

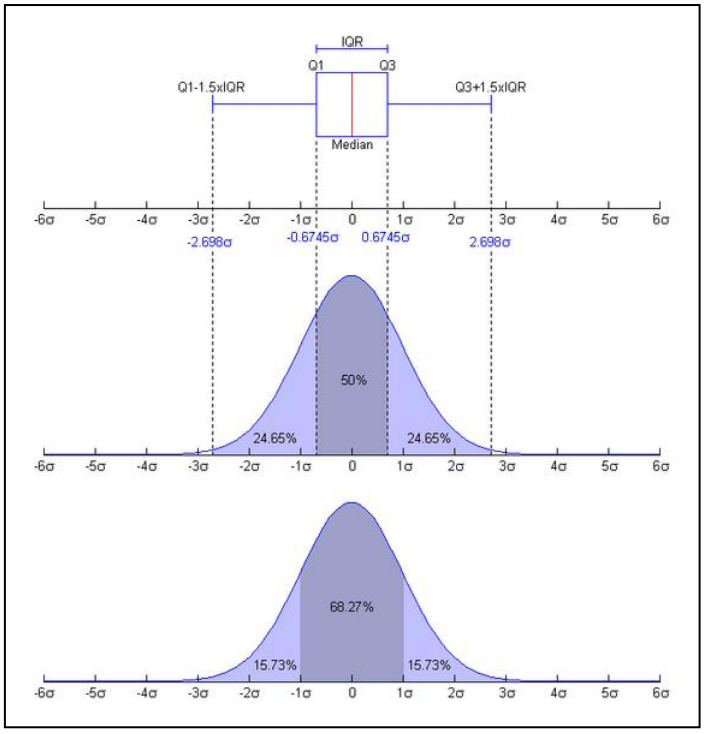
Esta tabla está diseñada para ayudarle a decidir cuál estadístico descriptivo debe utilizar con su set de datos. Para utilizarla, usted debe saber el nivel de medición de sus datos.

Estadístico/gráfico	Variable nominal	Variable cuantitativa	Variable ordinal	Propósito	Notas	Ejemplo
media aritmética Histograma, barra de errores, IC.	-	1	-	Descripción de la tendencia central de los datos. Variable normal.	Asume que datos provienen de distribución normal.	Apropiada para describir la altura total en una plantación forestal o cualquier otra variable con una distribución normal.
Media geométrica		1		Descripción de la tendencia central de los datos. Estimador insesgado.	La media geométrica sólo se aplica a números positivos. Se utiliza a menudo para un conjunto de números cuyos valores serán multiplicados entre sí o son de naturaleza exponencial, tales como datos de crecimiento poblacional o las tasas de interés.	Supongamos que la tasa de crecimiento de un árbol es 10% en el primer año, 8% en el segundo año y 5% en el tercer año. ¿Cuál es la tase de crecimiento media del árbol? La respuesta es la media geométrica de dichos valores.
Media armónica		1		Descripción de la tendencia central de los datos.	Apropiado para calcular la media de tasas (e.g. velocidad) y proporciones (parte entre el todo).	La tasa de cambio puede especificarse por unidad de tiempo, por unidad de longitud, de masa u otra cantidad. El tipo más común de tasa es "por unidad de tiempo", como la velocidad y el flujo. Algunos ejemplos de tasas cuyo denominador no es el tiempo incluyen los tipos de cambio, las tasas de alfabetización y el flujo eléctrico. Por ejemplo, en hidrología, la media armónica se utiliza para estimar valores medios de la conductividad hidráulica.
mediana Gráfico de cajas	-	1	-	Descripción de la tendencia central de los datos. Variable no normal.	Debe utilizarse cuando los datos son no normales (muy asimétricos).	Apropiada para describir la altura total en un bosque natural.
Moda, Gráfico de pastel, barras	1		1	Descripción de la tendencia central de los datos.	También puede utilizarse con datos cuantitativos discretos	Especie más frecuente en una parcela de 1000 m^2.

					(e.g. conteos) y con datos sin decimales.	
rango Gráfico de cajas	-	1	-	Descripción de la dispersión de los datos.	Se utiliza con mayor frecuencia en informes	Apropiada para cualquier variable cuantitativa.
Rango intercuartil Gráfico de cajas		1		Descripción de la dispersión de los datos.	Debe utilizarse cuando los datos son no normales (muy asimétricos). Se utiliza conjuntamente con la mediana.	Se calcula como Q3-Q1. Es un estadístico robusto comparado con el rango.
varianza	-	1	-	Descripción de la dispersión de los datos. Asume datos normales. Estimador insesgado.	Constituye la base de muchas pruebas estadísticas; sus unidades son al cuadrado, por lo que no es muy comprensible.	Apropiada para cualquier variable cuantitativa cuya distribución sea normal.
desviación estándar Barra de errores	-	1	-	Descripción de la dispersión de los datos. Asume datos normales. Estimador sesgado.	Sus unidades son las mismas que los datos originales, es más comprensible que la varianza.	Apropiada para cualquier variable cuantitativa.
Coeficiente de variación		1		Descripción de la dispersión relativa de los datos. Asume datos normales.	El coeficiente de variación (CV) se expresa con frecuencia en porciento.	Apropiada para cualquier variable cuantitativa cuya distribución sea normal.
error estándar de la media Barra de errores	-	1	-	Descripción de la exactitud de una estimación de la media aritmética. Asume datos normales.	Se utiliza para estimar el intervalo de confianza y en prueba de hipótesis.	Apropiada para cualquier variable cuantitativa cuya distribución sea normal.

Referencias

Choosing a statistical test http://udel.edu/~mcdonald/statbigchart.html
Intuitive Biostatistics: Choosing a statistical test. http://www.graphpad.com/www/Book/Choose.htm. Capítulo 37del libro “Intuitive Biostatistics” (ISBN 0-19-508607-4) de Harvey Motulsky. Copyright © 1995 by Oxford University Press Inc.



Relación entre medidas de tendencia central y de dispersión. Diagrama de caja (con un rango intercuartil) y una función de densidad de probabilidad (pdf) de una población normal N (0,1 σ2). Fuente: http://en.wikipedia.org/wiki/File:Boxplot_vs_PDF.png

Análisis exploratorio de datos

El análisis exploratorio de datos incluye el cálculo de estadísticos descriptivos, la tabulación de datos y la creación de gráficos. El objetivo de esta primera fase de análisis es familiarizarse con los datos.

Estadísticos para datos cuantitativos (intervalo y razón)

- **Número de valores utilizados**: número de valores efectivamente utilizados en los cálculos, es decir, los valores no faltantes o perdidos.
- **Número de valores ignorados**: número de valores no utilizados en los cálculos.
- **Mínimo**: valor mínimo de la serie estadística.
- **Máximo**: valor máximo de la serie estadística.
- **Primer cuartil (Q1)**: valor por debajo del cual se encuentran el 25% de los datos.
- **Mediana (Q2, Md)**: valor por debajo del cual se encuentran el 50% de los datos.
- **Tercer cuartil (Q3)**: valor por debajo del cual se encuentran el 75% de los datos.
- **Rango**: diferencia entre el máximo y el mínimo.
- **Total**: suma de todos los valores de la serie estadística (puede ser ponderada)
- **Media aritmética simple**:suma de los valores divididos entre el número de observaciones.
- **Media geométrica**: La media geométrica sólo se aplica a números positivos. Se utiliza con números cuyos valores serán multiplicados entre sí o son de naturaleza exponencial, tales como datos de crecimiento poblacional o las tasas de interés. Esta media es poco sensible a valores altos.
- **Media armónica**: Apropiado para calcular la media de tasas (e.g. velocidad) y proporciones (parte entre el todo). Esta media es poco sensible a unos pocos valores que son mucho más altos que los otros, pero es sensible a valores mucho más pequeños.
- **Curtosis (Pearson)**: coeficiente que representa el pico aplanado o forma de una distribución en comparación con una distribución de Normal o de Gauss.
- **Asimetría (Pearson)**: coeficiente que representa el grado de asimetría de una distribución respecto a su media.

- **Curtosis**: coeficiente de curtosis.
- **Asimetría**: coeficiente de asimetría.
- **CV (desviación estándar / media)**: El coeficiente de variación mide la dispersión relativa de los datos. Este coeficiente es apropiado para comparar la dispersión de variables con unidades de medición diferentes o con medias muy diferentes.
- **Varianza estimada**: Estimación de la varianza para una población a partir de los datos de una muestra. Estimador insesgado: para datos no ponderados, el denominador es n-1, donde n el tamaño de la muestra.
- **Desviación estándar estimada**: Raíz cuadrada de la varianza estimada.
- **Desviación media absoluta**: Medida de dispersión calculada como la media de las desviaciones absolutas de la media.
- **Desviación estándar de la media**: Raíz cuadrada la varianza estimada entre el número de observaciones $(S^2/n)^{0.5}$.

Gráficos para datos cuantitativos

- Histogramas simples, compuestos, comparativos (2d y 3D)
- Diagramas de cajas
- Diagrama de líneas (2d y 3D)
- Diagramas de dispersión univariante
- Diagramas de dispersión bivariado
- Gráficos de Q-Q: Este gráfico permite comparar la función de distribución acumulada (cdf) de la muestra (eje de abscisas) con la función de distribución acumulada de la distribución normal con la misma desviación media y estándar. Si los datos provienen de una distribución normal se graficaran como un recta. Si los datos no se ajustan a una recta entonces la distribución es no normal.
- Gráficos de percentiles
- Diagramas de tallo y hoja
- Grafica de distribución normal

Estadísticas ara los datos cualitativos (nominales y ordinales)

Estos estadísticos se calculan a partir de una tabla simple o doble e involucra el conteo frecuencias.

Estadísticos para la totalidad de las variables

- **Número de categorías o clases**: número de categorías por variable.
- **Moda**: Categoría que presenta la mayor frecuencia, o que tiene el mayor peso (si los datos son ponderados).
- **Frecuencia modal**: Frecuencia de la clase modal.
- **Moda en %**: Frecuencia en porciento de la clase modal.
- **Frecuencia relativa de la moda**: Frecuencia relativa de la clase modal, puede expresarse en % o como una proporción.

Estadísticos por variable

- **Frecuencia absoluta**: Frecuencia absoluta de cada categoría.
- **Frecuencia relativa**: Frecuencia relativa (% o fracción) de cada categoría.
- **Peso o ponderación**: Peso o ponderación de cada categoría.

Gráficos variables cualitativos (nominales, ordinales)

- Barras
- Gráfico de pasteo o circular

Anexo 2: Guía para el análisis de datos

Estadística descriptiva	Variables continuas	Ubicación o tendencia central	Media (Aritmética, Geométrica, Harmónica), Mediana, Moda.
		Dispersión o escala	Rango, Desviación estándar, Coeficiente de variación, Percentiles, desviación absoluta de la media.
		Forma de la distribución	Desviación absoluta de la media, Varianza, Semivarianza, Asimetría, Curtosis, Momentos y L-Momentos
	Variables nominales o categorías	Frecuencia Tabla de contingencia	
Gráficos estadísticos	Barras, Histogramas, líneas, bivariable, cajas, control, correlación, Q-Q, tallo-hoja, radar		
Inferencia estadística y prueba de hipótesis	Inferencia	Intervalo de confianza, Intervalo (inferencia frecuentista) inferencia creible (Inferencia Bayesiana), Nivel de significancia, Meta-análisis	
	Diseño experimental	Experimentos controlados, Experimentos naturales, estudio observacional, replicación, bloques, sensibilidad y especificidad, diseño óptimo.	
	Tamaño de muestra	Análisis de poder, efecto de tamaño de muestra, error estándar	
	Estimación general	Estimador Bayesiano, Máximo verosimilitud, Método de momentos, Estimación de densidad	
	Pruebas particulares	Prueba Z (normal), t de estudiante, Prueba F, Ji-cuadrado, Ji-cuadrado de Pearson, Wald, U de Mann–Whitney, · Shapiro–Wilk, Prueba de ordenes de signos de · Wilcoxon	
Muestreo	Simple al azar, estratificado, jerárquico, conglomerados.		
Correlación y regresión	Correlación	Coef. correlación lineal de Pearson, Coef. de Correlación para ordenes (rho de Spearman y tau de Kendall), correlación parcial, factor o variable de confusión.	
	Regresión lineal	Regresión lineal simple, cuadrados mínimos ordinarios, modelos lineales generales, análisis de varianza, análisis de covarianza	
	Regresión no estándar	Regresión no lineal, no paramétrica, semiparamétrica, robusta.	
	Regresión errores no normales	Modelos lineales generalizados, Binomial, Poisson, Logística	
Análisis de sobrevivencia	Función de sobrevivencia, Kaplan–Meier, Prueba·Logrank, Proporción de fracasos, ·Modelo de proporción de riesgos		
Estadística mulivariante	Regresión multivariante, Compoentes principales, Análisis factorial, Análisis de conglomerados		
Análisis de series de tiempo	Descomposición, estimación de tendencia, Box–Jenkins, Modelos ARMA, Estimación de densidad espectral.		
Estadística social	Censos, Psicometría, muestreo simple y estratificado, sondeos de opinión, cuestionarios, estadísticas oficiales.		

Anexo 3: Prueba de hipótesis: una muestra, dos muestras, tres o más muestras

Análisis univariante			
Número y tipo de variables predictoras	**Tipo de resultado**	**Hipótesis nula (H_0)**	**Prueba(s)**
Ninguna variable predictora	normal	$H_0: \mu = \mu_0$	Prueba “t” para una muestra.
	no-normal	$H_0: \xi = \xi_0$	Prueba del signo; prueba de signos para una muestra (one-sample median/ sign test).
	Categorías (dos clases)	$H_0: \Pi = \Pi_0$	Prueba Z para proporciones para una muestra (dado que *pn*>10 y *qn*>10) de lo contrario utilizar prueba binomial.
		$H_0: \Pi_1 = \Pi_{10}$, $\Pi_2 = \Pi_{20}$, otras	Prueba de ji-cuadrado (Chi-square goodness-of-fit test, chi-squared test; s-test)
Ninguna variable predictora (i.e., se desea conocer si existe asociación o independencia entre dos o más variables).	normal	$H_0: \rho = 0$	Coeficiente de correlación lineal de Pearson
			Análisis factorial
	Una o más no normales	$H_0: \rho_s = 0$	Coeficiente de correlación de órdenes de Spearman
	categorías	$H_0: \phi = 0$	Coeficiente de contingencia
		$H_0: O.R. = 1$	Riesgo relativo; cociente de disparidad (odds ratio).
Una variable predictora cualitativa (2 categorías)	normal	$H_0: \mu_1 = \mu_2$	Prueba “t” para dos muestras independientes. (dado que grupos sean independientes ó que sujetos hayan sido asignadas en forma aleatoria a los grupos).
			Prueba “t” para dos muestras pareadas. (dado que muestras sean pareadas ó mediciones repetidas en los mismos sujetos).
	no-normal	$H_0: \xi_1 = \xi_2$	Prueba de suma de los rangos de Wilcoxon (también se le denomina como prueba de Wilcoxon; prueba U; prueba de Mann-Whitney, prueba de Wilcoxon-Mann-Whitney (Wilcoxon rank sum test ; Wilcoxon's test ; Wilcoxon-Mann-Whitney test ; Mann-Whitney test ; U-test). (dado que grupos sean independientes ó que sujetos hayan sido asignadas en forma aleatoria a los grupos).
			Prueba de Wilcoxon para muestras apareadas (Wilcoxon signed ranks test).

			(dado que muestras sean pareadas ó mediciones repetidas en los mismos sujetos).
	categorías	H_0: $\Pi_{11} = \Pi_{12}$ $\Pi_{21} = \Pi_{22}$ etc.	Prueba de Chi-square test (dado que grupos sean independientes ó que sujetos hayan sido asignadas en forma aleatoria a los grupos).
			Prueba de chi-cuadrado de McNemar (dado que muestras sean pareadas ó mediciones repetidas en los mismos sujetos).
Una variable predictora cualitativa (3 o más categorías)	normal	H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$	ANOVA de una vía (dado que grupos sean independientes ó que sujetos hayan sido asignadas en forma aleatoria a los grupos).
			ANOVA para mediciones repetidas (dado que muestras sean pareadas ó mediciones repetidas en los mismos sujetos),
	no-normal	H_0: $\xi_1 = \xi_2 = \ldots = \xi_k$	Prueba de Kruskal-Wallis (dado que muestras sean independientes ó que hayan sido asignadas en forma aleatoria a los grupos).
	categorías	H_0: $\Pi_{11} = \Pi_{12} = \Pi_{13} = \Pi_{14}$ etc $\Pi_{21} = \Pi_{22} = \Pi_{23} = \Pi_{24}$ etc etc.	Prueba de Chi-cuadrado (dado que muestras sean independientes ó que hayan sido asignadas en forma aleatoria a los grupos).
Una variable predictora continua	normal	Ho: pendiente=0	Regresión lineal simple
	no-normal		Transformar datos
	categorías	H_0: O.R. = 1	Regresión logística

Análisis multivariante: Un resultado y dos o más variables predictoras			
Número y tipo de variables predictoras	**Tipo de resultado**	**Hipótesis Nula (H_0)**	**Prueba(s)**
2 o más variables predictoras cualitativas (2 o más categorías por variable)	Normal	Ho : no existen efectos principales	ANOVA factorial (dado que grupos sean independientes ó que sujetos hayan sido asignadas en forma aleatoria a los grupos).
		Ho : no existen interacción	
	No normal		Transformar datos
	Categorías (cualitativo)	Ho : independencia	Análisis log-lineal
2 o más variables predictoras continuas	Normal	Ho : coeficientes = 0	Regresión lineal múltiple
	No normal		Transformar datos
	Categorías (cualitativo)	H_0: O.R. = 1	Regresión logística
Mezcla de variables predictoras cualitativas y cuantitativas	Normal	Ho : no existen efectos	Análisis de covarianza (ANCOVA)
		Ho : no existen efectos principales o interacciones	Modelo lineal general
	No normal		Transformar datos
	Categorías (cualitativo)	H_0: O.R. = 1	Regresión logística
Dos o más resultados, uno o más variables predictoras			
Tipo de variables predictoras	**Tipo de resultado**	**Prueba**	
Cualitativa	Normal	Análisis de varianza multivariante (MANOVA)	
Continua		Regresión múltiple multivariante	
Mixto cualitativa y cuantitativa		Modelo lineal general multivariante	

Fuente: http://bama.ua.edu/~jleeper/627/revprev.html

Transformaciones comunes utilizadas para normalizar datos

Tipo de dato y distribución	Transformación para normalizar datos
Conteos (C proviene de una distribución de Poisson)	Raíz cuadrada de C
Proporciones (P proviene de una distribución binomial)	Arcoseno de raíz cuadrada de P
Dato proviene de una distribución lognormal	Log de la medición
Tiempo o duración (D)	1/D

Muestras independientes

Datos de muestras independientes							
		Variable producto					
		Nominal	Categorías (>2 Categorías)	Ordinal	Cuantitativa Discreta	Cuantitativa No Normal	Cuantitativa Normal
Insumo	Nominal	χ^2 o Prueba de Fisher (tablas de contingencia y muestras pequeñas)	χ^2	χ^2 tendencia o Mann-Whitney	Mann-Whitney	Mann-Whitney; test de log-rango (prueba de Mantel-Haenszel) (a)	Prueba “t” de Estudiante
	Categorías (>2 categorías)	χ^2	χ^2	Kruskal-Wallis (b)	Kruskal-Wallis (b)	Kruskal-Wallis (b)	Análisis de varianza de una vía (c)
	Ordinal (categorías ordenadas)	χ^2 tendencia o Mann-Whitney	(e)	Orden de Spearman	Orden de Spearman	Orden de Spearman	Coef. Correlación de órdenes de Spearman o regresión lineal (d)
	Cuantitativa Discreta	Regresión logística	(e)	(e)	Orden de Spearman	Orden de Spearman	Coef. Correlación de órdenes de Spearman ó regresión lineal (d)
	Cuantitativa No Normal	Regresión logística	(e)	(e)	(e)	Graficar datos, Coef. Correlación de rangos de Spearman o Coef. Correlación lineal de Pearson.	Graficar datos, Coef. Correlación de Pearson, Coef. Correlación de órdenes de Spearman y regresión lineal

	Cuantitativa Normal	Regresión logística	(e)	(e)	(e)	Regresión lineal (d)	Coef. Correlación lineal de Pearson y regresión lineal

(a) Si los datos son acotados (censored).

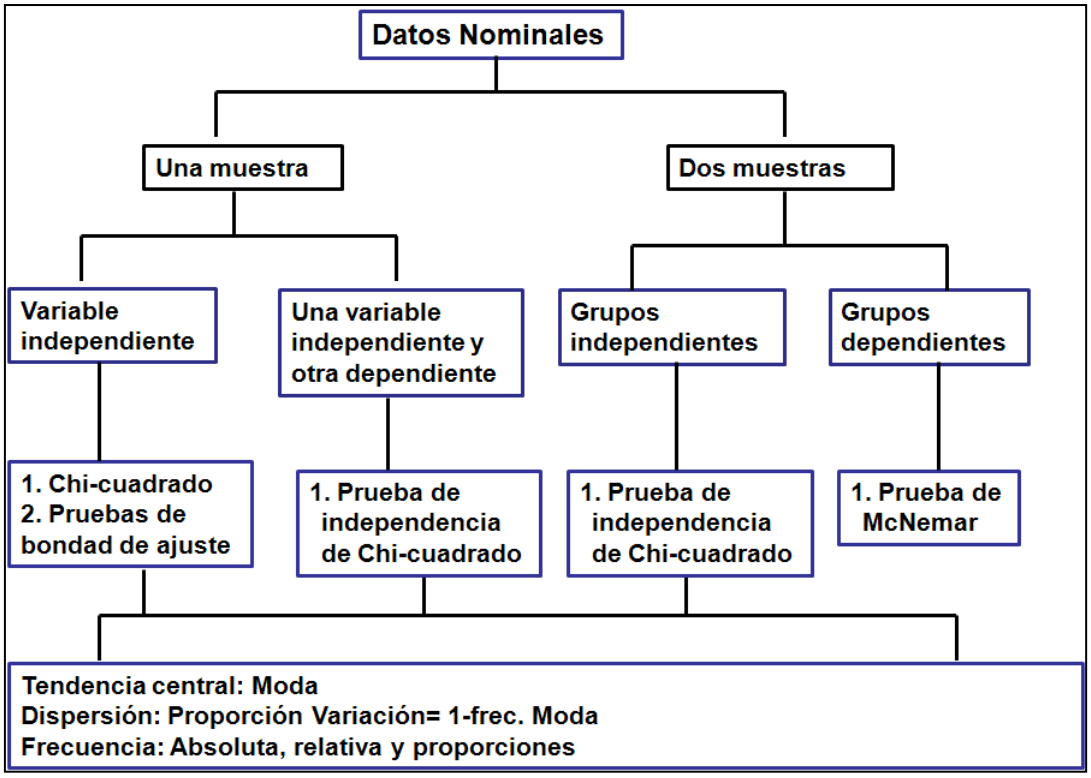
(b) La prueba de Kruskal-Wallis es una generalización de la prueba de U de Mann-Whitney y se utiliza para comparar variables ordinales o nominales

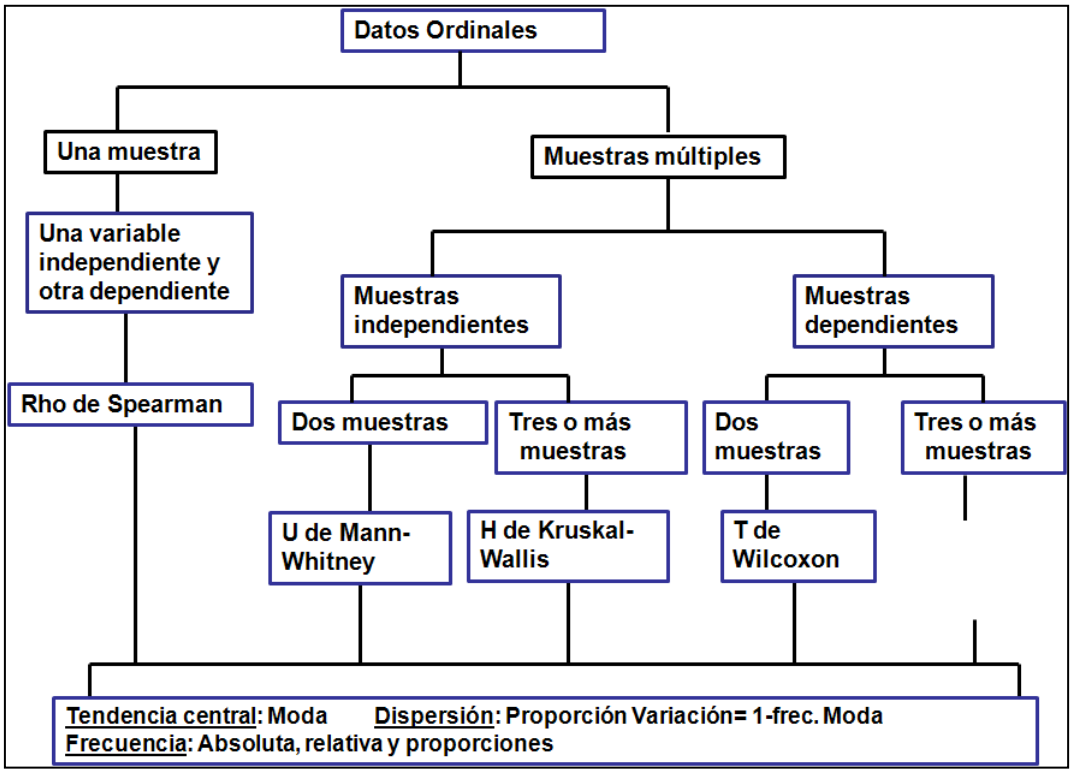
(c) El análisis de varianza de una vía se utiliza para comparar variables con una distribución normal para tres o más grupos. El equivalente no paramétrico es la prueba de Kruskal-Wallis.

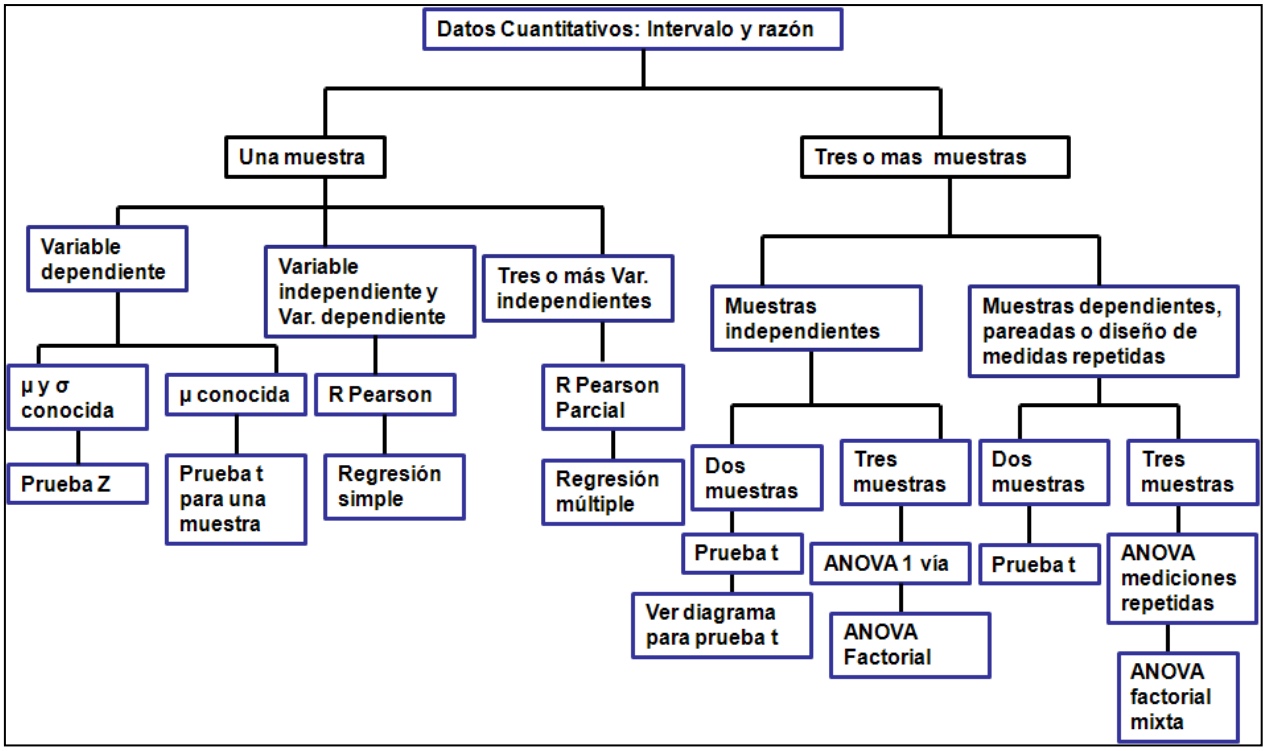
(d) Si la variable de resultado es la variable dependiente, entonces siempre los residuos son plausiblemente normales y por tanto la distribución de la variable independiente no es importante.

(e) Hay una serie de técnicas más avanzadas, tales como la regresión de Poisson, para hacer frente a estas situaciones. Sin embargo, requieren ciertas suposiciones por lo que a menudo es más fácil ya sea expresar la variable resultado como dicotómica o tratarla como una variable continua.

Fuente: http://www.bmj.com/collections/statsbk/13.dtl


Datos Nominales
Una muestra
Dos muestras
Variable independiente
Una variable independiente y otra dependiente
Grupos independientes
Grupos dependientes
1. Chi-cuadrado
2. Pruebas de bondad de ajuste
1. Prueba de independencia de Chi-cuadrado
1. Prueba de independencia de Chi-cuadrado
1. Prueba de McNemar
Tendencia central: Moda
Dispersión: Proporción Variación= 1-frec. Moda
Frecuencia: Absoluta, relativa y proporciones

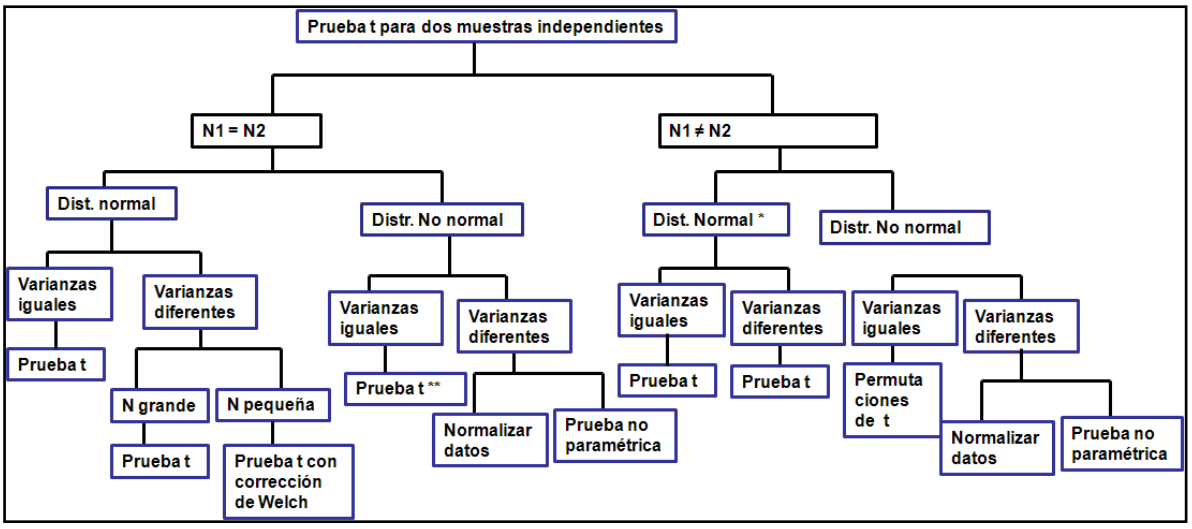

Datos Ordinales
Una muestra
Muestras múltiples
Una variable independiente y otra dependiente
Muestras independientes
Muestras dependientes
Rho de Spearman
Dos muestras
Tres o más muestras
Dos muestras
Tres o más muestras
U de Mann-Whitney
H de Kruskal-Wallis
T de Wilcoxon
Tendencia central: Moda
Dispersión: Proporción Variación= 1-frec. Moda
Frecuencia: Absoluta, relativa y proporciones

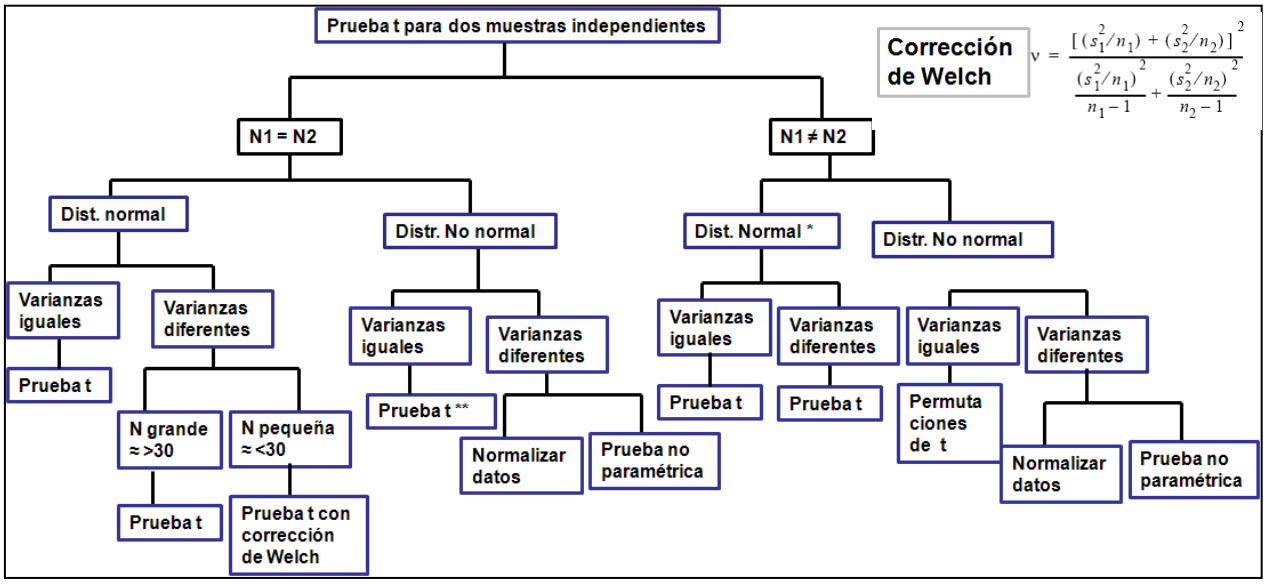


ANOVA 1 vía: Evalúa niveles múltiples de 1 variable o un factor.

ANOVA de 2 vías o Factorial: Evalúa simultáneamente el efecto de dos o más variables independientes en una variable dependiente. Permite evaluar la interacción entre variables.

ANOVA factorial mixta: El modelo mixto combina muestras de un factor independiente y muestras de grupos correlacionados o no independientes.





Sawilowsky, Shlomo S. 2002. Fermat, Schubert, Einstein, and Behrens–Fisher: The Probable Difference Between Two Means When σ1 ≠ σ2. Journal of Modern Applied Statistical Methods, 1(2).

***:** Poder de las pruebas es bajo cuando N1 y N2 son muy diferentes. Evitar utilizar corrección de Welch cuando N1 y N2 son muy diferentes, ya que la prueba tiene poco poder.

****:** La prueba "t" basado en permutaciones es preferible, ya que corrige error tipo I y tiene mas poder.

Pruebas no paramétricas: Wilcoxon-Mann-Witney, Preuba de medianas, Kolmogorov-Smirnov para 2 muestras, otras apropiadas

Kingman A, Zion G. 1994. Some power considerations when deciding to use transformations. Stat Med. 15; 13(5-7):769-83.

Neuhäuser M. 2005. One-sided nonparametric tests for ordinal data. Percept Mot Skills.101(2):510-4.

Sawilowsky, Shlomo S. 2002. Fermat, Schubert, Einstein, and Behrens–Fisher: The Probable Difference Between Two Means When σ1 ≠ σ2 Journal of Modern Applied Statistical Methods, 1(2).

Sawilowsky, S. S., & Hillman, S. B. 1992. Power of the independent samples t test under a prevalent psychometric measure distribution. *Journal of Consulting and Clinical Psychology*, *60*, 240-243.

Correa Juan Carlos, Iral René y Rojas Lucinia. 2006. Estudio de potencia de pruebas de homogeneidad de varianza. Revista Colombiana de Estadística. Volumen 29 No 1. pp. 57 a 76.

Anexo 4: Elección de una prueba estadística

Esta tabla está diseñada para ayudarle a decidir cuál prueba estadística puede utilizar para analizar sus datos. Para utilizarla, usted debe saber el nivel de medición de sus datos.

Prueba	Variables nominales	Variables continuas	Variables ordinales	Propósito	Notas	Ejemplo
prueba exacta de bondad de ajuste	1	-	-	Comparar el grado de ajuste entre frecuencias observadas y esperadas.	Utilizar con muestras pequeñas.	Contar el número de machos y hembras en una muestra pequeña y compararla con la proporción esperada de 1:1.
G-test de bondad de ajuste	1	-	-	Comparar el grado de ajuste entre frecuencias observadas y esperadas.	Utilizar con muestras grandes.	Contar el número de años Niño, Ñina y Neutros y y compararla con la proporción esperada de1/3:1/3:1/3.
prueba de Chi-cuadrado para la bondad de ajuste	1	-	-	Comparar el grado de ajuste entre frecuencias observadas y esperadas.	Utilizar con muestras grandes.	Contar el número de años lluviosos y secos y compararla con la proporción esperada de1:1.
Aleatorización prueba de bondad de ajuste	1	-	-	Comparar el grado de ajuste entre frecuencias observadas y esperadas.	Utilizar con muestras pequeñas y un gran número de categorías.	Contar el número de años Niño, Ñina y Neutros y y compararla con la proporción esperada de1/3:1/3:1/3.
G-prueba de independencia	2 +	-	-	Probar la hipótesis de que las proporciones son las mismas en los diferentes grupos.	Utilizar con muestras grandes.	Contar el número de árboles muertos vs el número de árboles vivos en tres tratamientos de plantación y probar que las proporciones son las mismas.
prueba de Chi-cuadrado de independencia	2 +	-	-	Probar la hipótesis de que las proporciones son las mismas en los diferentes grupos.	Utilizar con muestras grandes.	Contar el número de árboles muertos vs el número de árboles vivos en tres tratamientos de plantación y probar que las proporciones son las mismas.
prueba exacta de Fisher	2	-	-	Probar la hipótesis de que las proporciones son las mismas en los diferentes grupos.	Se utiliza para muestras pequeñas.	Contar el número de árboles muertos vs el número de árboles vivos en dos tratamientos de plantación y probar que las proporciones son las mismas.
Aleatorización prueba de independencia	2	-	-	Probar la hipótesis de que las proporciones son las mismas en los diferentes grupos.	utiliza para muestras pequeñas y un gran número de categorías	Contar el número de árboles por hectárea en tres estados sucesionales en dos categorías de pendiente y probar que las proporciones son las mismas.

Prueba de Mantel-Haenzel	3	-	-	Probar la hipótesis de que las proporciones son las mismas en parejas repetidas de dos grupos.	-	Contar el número de árboles vivos y muertos por hectárea para dos especies (A y B) en varias fincas y probar que las proporciones son las mismas ó que la diferencia es consistente en una dirección (e.g. la sobrevivencia siempre es mayor en especie A que en B).
intervalo de confianza	-	1	-	Describir exactitud de la estimación de la media aritmética.	Se puede utilizar para realizar pruebas de hipótesis.	Apropiada para cualquier variable cuantitativa.
ANOVA de un factor o una vía, modelo I	1	1	-	Probar la hipótesis que los valores medios de la variable cuantitativa son iguales entre los diferentes grupos. El investigador(a) selecciona, basado en su opinión, juicio o prejuicio, K valores de interés de la variable independiente (VI).	El modelo de efectos fijos o modelo “I” asume que los datos provienen de poblaciones normales las cuales podrían diferir únicamente en sus medias.	Comparar la media del contenido de metales pesados en los peces de los ríos Reventazón, Pacuare, Grande de Térraba y Grande de Tárcoles para ver si hay diferencias en el nivel de contaminación.
ANOVA de un factor o una vía, modelo II	1	1	-	Estimar la proporción de la varianza en la variable continua explicada por la variable nominal. El investigador(a) selecciona, utilizando un procedimiento al azar, K de los posibles valores de la variable independiente (VI).	En el Modelo “II” o de efectos aleatorios solo se eligen algunos de los posibles niveles de interés de la variable nominal y por esta razón la partición de varianza es más interesante que determinar qué grupos son diferentes.	Comparar la media del contenido de metales pesados en 5 familias de peces criados bajo las mismas condiciones para determinar si existe una variación heredada con respecto a la fijación de metales pesados.
Método secuencial de Dunn-Sidák	1	1	-	Prueba a posteriori planeada no ortogonal o no independiente entre grupos. Se realiza una vez que el ANOVA	Es deseable realizar comparaciones planeadas ortogonales. Decidir antes de realizar el	Comparar, por ejemplo, la media del contenido de metales pesados en los peces de los ríos Reventazón+Pacuare Vs Grande de Térraba + Grande de Tárcoles.

				encontró diferencias significativas entre las medias. En primer lugar, los valores de P de las diferentes comparaciones se ordenan de menor a mayor. Si hay k comparaciones, el valor más pequeño de P de ser inferior a $1-(1-alfa)^{1/k}$ para ser significativo al nivel de alfa. Por ejemplo, si hay cuatro comparaciones, el valor más pequeño de P debe ser inferior a $1-(1-0.05)^{1/4}$ = 0,0127 para ser significativo a un nivel de 0,05. Si no es significativo, se detiene el análisis. Si el valor menor de P es significativo, el siguiente valor más pequeño debe ser inferior a $1-(1-alfa)^{1/(k-1)}$, que en este caso sería 0,0170. Si este valor es significativo, el siguiente valor de P debe ser inferior a $1-(1-alfa)^{1/(k-2)}$, y así sucesivamente hasta que uno de los valores de P sea no significativo.	análisis las comparaciones de interés para el estudio.	
Método de Bonferroni	1	1		Este método tiene menos poder que el método secuencial de Dunn-Sidák. La diferencia entre el valor del alfa ajustado de Bonferroni y Dunn Sidák es muy pequeña, por lo que su impacto en	El método de Bonferroni utiliza el valor alfa/k como el nivel de alfa ajustado, mientras que el método de Dunn-Sidák utiliza $1-(1-alfa)^{1/k}$. El método de Bonferroni	

				la decisión final no es significativo. Por ejemplo, el valor de alfa ajustado de Bonferroni para cuatro comparaciones es 0,0125, mientras que para Dunn-Sidák es 0,0127.	no es secuencial y por tanto el mismo valor de alfa se aplica a todas las comparaciones.	
Comparación de intervalos de Gabriel Ver libro de Excel anova 1 via_Tukey_Kramer.xlsx	1	1	-	Prueba a posteriori no planeada, una vez que el ANOVA es significativo se realizan comparaciones entre todos los pares de grupos.	Es deseable realizar comparaciones planeadas ortogonales. Decidir antes de realizar el análisis las comparaciones de interés para el estudio.	Comparar, por ejemplo, la media del contenido de metales pesados en los peces de los ríos Reventazón Vs Pacuare, Grande de Térraba Vs Grande de Tárcoles, etc.
Tukey-Kramer método Ver libro de Excel anova 1 via_Tukey_Kramer.xlsx	1	1	-	Prueba a posteriori no planeada, una vez que el ANOVA es significativo se realizan comparaciones entre todos los pares de grupos.	Es deseable realizar comparaciones planeadas ortogonales. Decidir antes de realizar el análisis las comparaciones de interés para el estudio.	Comparar, por ejemplo, la media del contenido de metales pesados en los peces de los ríos Reventazón Vs Pacuare, Grande de Térraba Vs Grande de Tárcoles, etc.
Prueba de Bartlett	1	1	-	Prueba por la igualdad de varianzas entre los diferentes grupos.	Esta prueba se utiliza para evaluar el supuesto de la ANOVA de homogeneidad de varianzas entre grupos.	-
análisis de varianza anidado	2 +	1	-	Prueba la hipótesis que los valores medios de la variable continua son los mismos en diferentes grupos, cuando cada grupo se divide en subgrupos.	Subgrupos debe ser aleatorio (modelo II).	Comparar la media del contenido de metales pesados en los peces de los ríos Reventazón, Pacuare, Grande de Térraba y Grande de Tárcoles; se obtienen varios peces en cada río y se realizan varias mediciones de en cada pez.
ANOVA de dos vías o dos sentidos	2	1	-	Probar la hipótesis de que diferentes grupos, clasificados de dos	Es una prueba más eficiente que la ANOVA de una vía.	Comparar el crecimiento de árboles en parcelas de diferente pendiente (A, B, C) y grado de fertilidad (A, B; C). En

				maneras, poseen la misma media.		este caso los factores son pendiente y fertilidad.
Prueba t pareada	2	1	-	Probar la hipótesis de que las medias de la variable continua son las mismas en datos apareados.	-	Comparar el contenido de nitrógeno en una parcela antes y después de plantar leguminosas.
Regresión lineal	-	2	-	Determinar si la variación en una variable predictora causa una variación en una variable dependiente; estimar el valor de una variable no medida a partir de una variable medida.	-	Medir el diámetro y la altura de árboles y determine si la variación en diámetro explica la variación en altura. Estimar la altura de árboles a partir de la medición de diámetros.
Correlación	-	2	-	Determinar si existe covariación o correlación entre dos variables.	-	Determinar si existe una correlación entre el diámetro y el volumen de un árbol.
Regresión múltiple	-	3 +	-	Ajustar una ecuación que relaciona varias variables predictoras (X) a una variable dependiente (Y).	-	Medir diámetro, altura, volumen de un árbol y analizar cómo se relacionan con su biomasa.
Regresión polinómica	-	2	-	Probar la hipótesis de que una ecuación con términos X^2, X^3, etc. se ajusta mejor a la variable Y que una regresión lineal.	-	Medir el diámetro y el volumen de un árbol y ajustar dos modelos: uno lineal y otro polinómico. Comparar el error de predicción de los modelos.
Análisis de covarianza (ANCOVA)	1	2	-	Probar la hipótesis de que los diferentes grupos tienen la misma línea de regresión.	El primer paso es comprobar la homogeneidad de pendientes (b_1, b_2, b_n) y si no son significativamente diferentes, probar por la igualdad de los interceptos (a_1, a_2, a_n).	Mida la precipitación por año en cuatro estaciones y determine si existe una diferencia por estación en pendiente e intercepto.
Prueba de signo	2	-	1	Prueba por la aleatoriedad en la dirección de la diferencia en datos apareados.	Esta prueba es utilizada a menudo como una alternativa no paramétrica de la	Comparar el contenido de nitrógeno en una parcela antes y después de plantar leguminosas, sólo registre si el nivel es mayor o menor después del

					prueba t pareada de Estudiante.	tratamiento.
Kruskal-Wallis	1	-	1	Probar la hipótesis de que los órdenes son los mismos en todos los grupos.	Esta prueba es utilizada a menudo como una alternativa no-paramétrica a la ANOVA de 1 vía.	Diez plantas de rosas rojas de cinco variedades son juzgadas por su "belleza" utilizando una escala ordinal (valores de 1 a 5), luego los ordenes se comparan entre las variedades.
Coeficiente de correlación de órdenes de Spearman	-	-	2	Determinar si existe correlación entre los órdenes de dos variables.	Esta prueba es utilizada como una alternativa no paramétrica al coef. correlación de lineal de Pearson.	Diez plantas de rosas rojas de cinco variedades son juzgadas por su "belleza" y tamaño utilizando una escala ordinal (valores de 1 a 5), luego los ordenes se comparan para determinar si las rosas más bellas son también las más grandes.

Referencias

Choosing a statistical test http://udel.edu/~mcdonald/statbigchart.html

Intuitive Biostatistics: Choosing a statistical test. http://www.graphpad.com/www/Book/Choose.htm. Capítulo 37 del libro “Intuitive Biostatistics” (ISBN 0-19-508607-4) de Harvey Motulsky. Copyright © 1995 by Oxford University Press Inc.

Statistics at Square One. Study design and choosing a statistical test. http://www.bmj.com/statsbk/13.dtl

Anexo 5: Comparación de paquetes estadísticos

Fuente: http://en.wikipedia.org/wiki/Comparison_of_statistical_packages

Product	Developer	Latest version	Cost (USD)	Open source	Software license	Interface
AcaStat	Acastat		$29	No	Proprietary	GUI
ADaMSoft	Marco Scarno	January 24, 2008	Free	Sí	GNU GPL	CLI/GUI
Analyse-it	Analyse-it		$185–495	No	Proprietary	GUI
ASReml	VSN International	October 2009	>$150	No	Proprietary	CLI
Auguri	Advanced Analytics Group	June 30, 2008	$54 – $799	No	Proprietary	GUI
Autobox	Autobox	March 31, 2010 - Version 6.0	$800 - $34,000	No	Proprietary	GUI
BioStat	AnalystSoft	January 8, 2007	$100[1][2]	No	Proprietary	GUI
BMDP	Statistical Solutions		$1095	No	Proprietary	
BrightStat	Daniel Stricker	February 22, 2008	Free	No	Proprietary	GUI
Dataplot	Alan Heckert	March 2005	Free	Sí	Public domain	CLI/GUI
EasyReg	Herman J. Bierens		Free[3] / $300	No	Proprietary	GUI
Epi Info	Centers for Disease Control and Prevention	August 18, 2008	Free	No	Proprietary	CLI/GUI
EViews	Quantitative Micro Software	April 2010	student: $40 / acad: $425 / comm: $1075	No	Proprietary	CLI/GUI
GAUSS	Aptech systems	May 2008		No	Proprietary	CLI/GUI
GenStat	VSN International	July 2008	>$190	No	Proprietary	CLI/GUI
Golden Helix	Golden Helix	March 2008		No	Proprietary	CLI/GUI
GraphPad Prism	GraphPad Software, Inc.	Feb. 2009	$595	No	Proprietary	GUI
gretl	The gretl Team	January 23, 2009	Free	Sí	GNU GPL	CLI/GUI
JMP	SAS Institute	November 1, 2005	$1190	No	Proprietary	GUI/CLI
jHepWork	SAS Institute	November 1, 2005	-	No	GNU GPL	GUI/CLI
MacAnova	Gary W. Oehlert and Christopher Bingham	August 29, 2005	Free	Sí	GNU GPL	GUI
Maple	Maplesoft	April, 2009	$1895 (commercial), $99 (student)	No	Proprietary	CLI/GUI
Mathematica	Wolfram Research	7.0.1, March 2009	$2,495 (Professional), $1095 (Education), $140 (Student), $69.95 (Student annual license) [4] $295 (Personal)[5]	No	Proprietary	CLI/GUI

MATLAB	The MathWorks	2009b, September 4, 2009	Depends on many things.	No	Proprietary	CLI/GUI
MedCalc	Frank Schoonjans	August 3, 2009	$395	No	Proprietary	GUI
modelQED	marketingqed	June 15, 2007		No	Proprietary	GUI
Minitab	Minitab Inc.	January 10, 2007	$1195[2]	No	Proprietary	CLI/GUI
MRDCL	MRDC	January 04, 2008	$4000[2]	No	Proprietary	CLI/GUI
NCSS	NCSS Statistical Software		>$399	No	Proprietary	GUI
NMath Stats	CenterSpace Software	November 2009	($1295)[2]	No	Proprietary	CLI
NumXL	Spider Financial	October 2009	Lite version (Free), Professional edition ($300) [2]	No	Proprietary	GUI
OpenEpi	A. Dean, K. Sullivan, M. Soe	May 20 2009	Free	Sí	GNU GPL	GUI
Origin	OriginLab		$699	No	Proprietary	GUI
Ox programming language	OxMetrics, J.A. Doornik	August 2009	Free for Academic use	No	Proprietary	CLI
OxMetrics	OxMetrics, J.A. Doornik	August 2009	$1805-...	No	Proprietary	CLI/GUI
Partek	Partek	June 2007		No	Proprietary	GUI
Primer	Primer-E	February 2007	$500-$1000	No	Proprietary	GUI
PSPP	pspp	October 11, 2009	Free	Sí	GNU GPL	CLI/GUI
R	R Foundation	April, 2010	Free	Sí	GNU GPL	CLI/GUI[6]
R Commander[7]	John Fox	August 1, 2006	Free	Sí	GNU GPL	CLI/GUI
RATS	Estima	October 1, 2007	$500	No	Proprietary	CLI/GUI
RKWard[7]	RKWard Community	February 15, 2007	Free	Sí	GNU GPL	GUI
SalStat	Alan James Salmoni	February 2007	Free	Sí	GNU GPL	GUI
SAGE	>100 developers worldwide	4.3, December 2009	Free	Sí	GNU GPL	CLI & GUI
SAS	SAS Institute	March 2008	Commercial: ~$6000 per seat (PC version) / ~$28K per processor (Windows server) first-year fees for BASE, STAT, GRAPH, and ACCESS modules. Modules are licensed individually. Subsequent year fees are roughly half. [2]	No	Proprietary	CLI/GUI
SHAZAM	Elastix Ltd	July 2007	Pro $490 / Std. $390 / Site Lic: / Std. $1200 / Pro $1600	No	Proprietary	CLI/GUI
SOCR	UCLA	October 28, 2008	Free	Sí	none stated	GUI
SOFA Statistics	Grant Paton-Simpson	April 2010	Free	Sí	AGPL	GUI

SPlus	Insightful Inc.	2005	$2399/year	No	Proprietary	CLI
SPSS	SPSS Inc.	2007	$1599[2]	No	Proprietary	CLI/GUI
Stata	StataCorp	July 2009	academic starting at $595[2] / industry starting at $1,245	No	Proprietary	CLI/GUI
Statgraphics	StatPoint	October, 2009	$695 - $1495	No	Proprietary	GUI
STATISTICA	StatSoft		>$695	No	Proprietary	GUI
Statistix	Statistix		$695	No	Proprietary	GUI
StatIt	StatIt		>$395	No	Proprietary	GUI
STATPerl	STATPerl		Free	Sí	GNU GPL	GUI
StatPlus	AnalystSoft	January 7, 2007	$150[1][2]	No	Proprietary	GUI
StatsDirect	StatsDirect		$179	No	Proprietary	GUI
SYSTAT	Systat Software Inc.	February 21, 2007	$1299	No	Proprietary	CLI/GUI
Total Access Statistics	FMS Inc.	May 2008	$599	No	Proprietary	GUI, Access
UNISTAT	Unistat Ltd	March 15, 2005	$895, $495, $300	No	Proprietary	GUI, Excel
The Unscrambler	CAMO Software	September 2008	FREE-to-TRY	No	Proprietary	GUI
VisualStat	VisualStat Computing		$195	No	Proprietary	GUI
Winpepi	J. H. Abramson	June 2008	Free	No	Proprietary	GUI
WinSPC	DataNet Quality Systems	February 2008	$1600	No	Proprietary	CLI/GUI
XLStat	Addinsoft Inc.	February 2009	$395[2]	No	Proprietary	Excel
XploRe	MD*Tech	2006		No	Proprietary	GUI
Product	**Developer**	**Latest version**	**Cost (USD)**	**Open source**	**Software license**	**Interface**

Sistemas operativos

Product	**Windows**	**Mac OS**	**Linux**	**BSD**	**Unix**
AcaStat	Sí	No	No	No	No
ADaMSoft	Sí	Sí	Sí	Sí	Sí
Analyse-it	Sí	No	No	No	No
Auguri	Sí	No	No	No	No
Autobox	Sí	No	Sí	No	Sí
BioStat	Sí	No	No	No	No
BMDP	Sí				
BrightStat	Sí	Sí	Sí	No	Sí
Dataplot	Sí	Sí	Sí	Sí	Sí
EasyReg	Sí	No	No	No	No
Epi Info	Sí	No	No	No	No

EViews	Sí	Terminated (1.1)	No	No	No
GAUSS	Sí	Sí	Sí	No	Sí
Golden Helix	Sí	Sí	Sí	No	Sí
GraphPad Prism	Sí	Sí	No	No	No
gretl	Sí	Sí	Sí	No	No
JMP	Sí	Sí	Sí	No	Sí
JHepWork	Sí	Sí	Sí	Sí	Sí
MacAnova	Sí	Sí	Sí	Sí	Sí
Maple	Sí	Sí	Sí	**?**	Sí
Mathematica	Sí	Sí	Sí	No	Sí
MedCalc	Sí	No	No	No	No
Minitab	Sí	Terminated	No	No	No
modelQED	Sí	No	No	No	No
NCSS	Sí	No	No	No	No
NMath Stats	Sí	No	No	No	No
NumXL	Sí	No	No	No	No
OpenEpi	Sí	Sí	Sí	Sí	Sí
Origin	Sí	No	No	No	No
Partek	Sí	No	Sí	No	No
Primer	Sí	No	No	No	No
PSPP	Sí	Sí	Sí	Sí	Sí
R Commander	Sí[8]	Sí[8]	Sí[8]	Sí[8]	Sí
R	Sí	Sí	Sí	Sí	Sí
RATS	Sí	Sí	Sí	No	Sí
RKWard	No	No	Sí	No	Sí
Sage	No	Sí	Sí	No	Sí
SalStat	Sí	Sí	Sí	Sí	Sí
SAS	Sí	Terminated	Sí	No	Sí
SHAZAM	Sí	No	No	No	No
SOCR	Sí	Sí	Sí	Sí	Sí
SOFA Statistics	Sí	Sí	Sí	Sí	Sí
SPlus	Sí	No	Sí	No	Sí
SPSS	Sí	Sí	Sí	No	No
Stata	Sí	Sí	Sí	No	Sí
Statgraphics	Sí	No	No	No	No
STATISTICA	Sí	No	No	No	No
Statistix	Sí	No	No	No	No
StatIt	Sí	No	No	No	No
STATPerl	Sí	No	No	No	No

StatPlus	Sí	Sí	No	No	No
StatsDirect	Sí	No	No	No	No
SYSTAT	Sí	No	No	No	No
Total Access Statistics	Sí	No	No	No	No
UNISTAT	Sí	No	No	No	No
The Unscrambler	Sí	No	No	No	No
VisualStat	Sí	No	No	No	No
Winpepi	Sí	No	No	No	No
WinSPC	Sí	No	No	No	No
XLStat	Sí	Sí	No	No	No
XploRe	Sí	No	Sí	No	Sí

ANOVA

Product	One-Way	Two-Way	MANOVA	GLM	Post-hoc Tests	Latin Squares Analysis
AcaStat	No	No	No	No	No	No
ADaMSoft	No	No	No	No	No	No
Analyse-it	Sí	Sí	No	No	Sí	No
Auguri	Sí	Sí	No	No	No	No
Autobox	No	No	No	No	No	No
BioStat	Sí	Sí	Sí	Sí	Sí	No
BMDP	Sí	Sí	Sí	Sí	Sí	
BrightStat	Sí	Sí	No	Sí	Sí	No
EasyReg	No	No	No	No	No	No
Epi Info	Sí	Sí	No	No	No	No
EViews	Sí					
GAUSS		No	No	No	No	No
GenStat	Sí	Sí	Sí	Sí	Sí	Sí
Golden Helix	Sí	Sí				
GraphPad Prism	Sí	Sí	No	No	Sí	No
gretl						
Mathematica	Sí	Sí	Sí	Sí	Sí	No
MedCalc	Sí	Sí	No	Sí	Sí	No
Minitab	Sí	Sí	Sí	Sí	Sí	Sí
NCSS	Sí	Sí	Sí	Sí	Sí	Sí
NMath Stats	Sí	Sí	No	No	No	No
OpenEpi	Sí	No	No	No	No	No
Origin	Sí	Sí	No	No	No	No
Partek	Sí	Sí	Sí	Sí	Sí	Sí

PSPP	Sí	Sí	Sí	Sí	Sí	Sí
R	Sí	Sí	Sí	Sí	Sí	
R Commander	Sí	Sí		Sí		
Sage	Sí	Sí	Sí	Sí	Sí	
SAS	Sí	Sí	Sí	Sí	Sí	
SHAZAM	Sí	Sí	No	Sí	Sí	No
SOCR	Sí	Sí	No	No	Sí	Sí
SOFA Statistics	Sí	No	No	No	No	No
Stata	Sí	Sí	Sí	Sí		Sí
Statgraphics	Sí	Sí	Sí	Sí	Sí	Sí
STATISTICA	Sí	Sí	Sí	Sí	Sí	Sí
StatIt	Sí	Sí	Sí	Sí	Sí	No
StatPlus	Sí	Sí	Sí	Sí	Sí	Sí
SPlus	Sí	Sí	Sí	Sí	Sí	Sí
SPSS	Sí	Sí	Sí	Sí	Sí	Sí
StatsDirect	Sí	Sí	No	No	Sí	Sí
Statistix	Sí	Sí	Sí	Sí	Sí	Sí
SYSTAT	Sí	Sí	Sí	Sí	Sí	Sí
Total Access Statistics	Sí	Sí	No	No	No	No
UNISTAT	Sí	Sí	No	Sí	Sí	Sí
The Unscrambler	Sí	No	No	No	No	No
VisualStat	Sí	Sí	Sí	No	Sí	No
Winpepi	No	No	No	No	No	No
WinSPC	Sí	Sí	No	No	No	No
XLStat	Sí	Sí	Sí	Sí	Sí	No

Correlación y regresión

Product	OLS	WLS	2SLS	NLLS	Logistic	GLM	LAD	Stepwise	Quantile regression	Probit	Poisson	MLR
AcaStat												
ADaMSoft	Sí	No	No	Sí	Sí	No	No	No				
Analyse-it	Sí											
Auguri	Sí	Sí	No	Sí								
Autobox	Sí							Sí				Sí
BioStat												
BMDP	Sí				Sí			Sí				
BrightStat	Sí				Sí			Sí				
EasyReg	Sí		Sí						Sí			
Epi Info	Sí	No	No	No	Sí	No	No	No				

EViews	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	
GAUSS							No		No			
Golden Helix					Sí			Sí				
GraphPad Prism	Sí	Sí	No	Sí	No	No	No	No	No	No	No	
gretl	Sí	Sí	Sí	Sí	Sí	No	Sí	Sí	Sí	Sí	Sí	
Mathematica	Sí	Sí		Sí	Sí[9]	Sí[10]			Sí	Sí[11]	Sí	
MedCalc	Sí	No	No	Sí	Sí	No	No	Sí				
Minitab	Sí	Sí	No	No	Sí	No	No	Sí	No			
NCSS	Sí				Sí			Sí				
NMath Stats	Sí	Sí		Sí								Sí
Origin												
Partek	Sí	No	No	No	No	Sí	No	Sí				
PSPP												
R	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí
R Commander[7]	Sí					Sí			No			
RATS	Sí	Sí	Sí	Sí	Sí	No	Sí	Sí	Sí			
Sagata	Sí	Sí	No	No	No	No	Sí	Sí				
Sage	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	
SAS	Sí	Sí		Sí	Sí	Sí		Sí	Sí	Sí	Sí	
SHAZAM	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí
SOCR	Sí	No	No	No	Sí	No	No	No				
SPlus	Sí			Sí	Sí	Sí		Sí	No			
SPSS	Sí	Sí	Sí	Sí	Sí	Sí	No	Sí	No			
Stata	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí
Statgraphics	Sí	Sí	No	Sí	Sí	Sí	No	Sí				
STATISTICA	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	No			
StatIt												
StatPlus	Sí	No	Sí	Sí	Sí	Sí	No	Sí				
StatsDirect	Sí	Sí			Sí			Sí				
Statistix												
SYSTAT	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí	No			
Total Access Statistics	Sí	Sí										
UNISTAT	Sí	Sí	Sí	Sí	Sí	Sí	No	Sí		Sí	Sí	
The Unscrambler												
VisualStat												
Winpepi	Sí				Sí							
XLStat	Sí	Sí		Sí	Sí	Sí				Sí		

Análisis de series de tiempo

Product	ARIMA	GARCH	Unit root test	Cointegration test	VAR	Multivariate GARCH
AcaStat						
Analyse-it						
Auguri					Sí	
Autobox	Sí					
BioStat						
BMDP	Sí					
EasyReg	Sí	Sí	Sí	Sí		
EViews	Sí	Sí	Sí	Sí	Sí	Sí
GAUSS						
GraphPad Prism	No	No	No	No	No	
gretl	Sí	Sí	Sí	Sí	Sí	
Mathematica	Sí	Sí		Sí		
MedCalc	No	No	No	No	No	
Minitab	Sí	No	No	No	No	
NCSS	Sí					
NumXL	Sí	Sí				
NMath Stats						
Origin						
PSPP						
R	Sí	Sí	Sí	Sí	Sí	
R Commander[7]						
RATS	Sí	Sí	Sí	Sí	Sí	Sí
Sage	Sí	Sí	Sí	Sí	Sí	Sí
SAS						
SHAZAM	Sí	Sí	Sí	Sí	Sí	No
SOCR	No	No	No	No	No	
Stata	Sí	Sí	Sí	Sí	Sí	Sí
Statgraphics	Sí	No	No	No	No	
STATISTICA	Sí	No	No	No	No	
StatIt						
StatPlus	Sí	No	No	No	No	
SPlus		Sí			Sí	
SPSS	Sí					
StatsDirect						
Statistix						
SYSTAT	Sí					

Total Access Statistics						
UNISTAT	Sí	No	No	No	No	
VisualStat						
Winpepi						
XLStat						

Gráficos

Chart	**Bar chart**	**Box plot**	**Correlogram**	**Histogram**	**Line chart**	**Scatterplot**
AcaStat	No	No	Sí	Sí	Sí	Sí
ADaMSoft						
Analyse-it						
Auguri	Sí	Sí	Sí	Sí	Sí	Sí
Autobox			Sí		Sí	Sí
BioStat	Sí	Sí	Sí	Sí	Sí	Sí
BMDP				Sí		Sí
BrightStat	Sí	Sí	No	Sí	Sí	Sí
EasyReg						
Epi Info	Sí	No	No	Sí	Sí	Sí
EViews	Sí	Sí	Sí	Sí	Sí	Sí
GAUSS						
Golden Helix						
GraphPad Prism	Sí	Sí	Sí	Sí	Sí	Sí
gretl	Sí	Sí	Sí	Sí	Sí	Sí
Mathematica	Sí[12]	Sí[13]		Sí[14]	Sí[15]	Sí[16][17]
MedCalc	Sí	Sí		Sí	Sí	Sí
Minitab	Sí	Sí	Sí	Sí	Sí	Sí
NCSS	Sí	Sí	Sí	Sí	Sí	Sí
NMath Stats						
Origin	Sí	Sí		Sí	Sí	Sí
Partek						
PSPP						
R	Sí	Sí	Sí	Sí	Sí	Sí
R Commander						
Sage	Sí	Sí	Sí	Sí	Sí	Sí
SAS	Sí	Sí	Sí	Sí	Sí	Sí
SHAZAM	Sí	Sí	Sí	Sí	Sí	Sí
SOCR	Sí	Sí	Sí	Sí	Sí	Sí
Stata	Sí	Sí	Sí	Sí	Sí	Sí
Statgraphics						

STATISTICA	Sí	Sí	Sí	Sí	Sí	Sí
StatIt						
StatPlus	Sí	Sí	Sí	Sí	Sí	Sí
SPlus						
SPSS	Sí	Sí	Sí	Sí	Sí	Sí
StatsDirect						
Statistix						
SYSTAT						
Total Access Statistics						
UNISTAT	Sí	Sí	Sí	Sí	Sí	Sí
The Unscrambler	Sí			Sí	Sí	Sí
VisualStat						
Winpepi						
XLStat	Sí	Sí	Sí	Sí	Sí	Sí

Otros análisis

Producto	**s/w type** [18]	**Descriptive Statistics**		**Nonparametric Statistics**		**Quality Control**	**Survival Analysis**			**Data Processing**	
		base stat. [19]	normality tests [20]	CTA [21]	nonparametric comp., ANOVA			cluster analysis	discriminant analysis	BDP [22]	Ext. [23]
AcaStat	S	+	-	-	-	-	-	-	-	-	-
Analyse-it	X	+	+	+	+	-	-	-	-	+	+
Auguri	S	+	+	-	-	-	-	-	-	+	+
BioStat	S	+	+	+	+	-	+	-	-	+	+
BMDP		+	+	+	+		+	+	+		
BrightStat		+	+	+	+	-	+	-	-	+	+
EasyReg	S	+	-	-	-	-	+?	-	-	+	+
Epi Info	S	+	-	+	+	-	+	-	-	+	+
Gauss	St	+	+	-	-	-	-	-	-	+	+
Golden Helix	S	+	+	-	+	+	-	+	-	-	-
GraphPad Prism	S	+	+	-	+	-	+	-	-	-	-
Mathematica	S	+	+	+	+	-	-	+	-	+	+
MedCalc	S	+	+	+	+	+	+	-	-	+	+

Minitab	S	+	+	+	+	+	+	+	+	+	+
NCSS	S	+	+	+	+	+	+	+	+	+	+
NMath Stats	S	+	+	-	-	-	-	+	-	-	-
OpenEpi	S	+	-	+	-	-	-	-	-	-	-
Origin	S	+	+	-	-	+	+-	-	-	+	+
Partek	S	+	+	+	+	+	+	+	+	+	+
PSPP	S	+	+								
RATS	S	+	+	-	-	-	-	-	-	+	+
SAS	S	+	+	+	+	+	+	+	+	+	+
SHAZAM	S	+	+	-	-	-	-	-	-	+	+
SOCR	S	+	+	+	+	-	+	+	-	+	+
SOFA Statistics	S		-	+	+	-	-	-	-	-	-
Stata	S	+	+	+	+	+	+	+	+	+	+
Statgraphics	S	+	+	+	+	+	+	+	+	+	+
STATISTICA	S	+	+	+	+	+	+	+	+	+	+
StatIt	S	+	+	+	+	+	+	+	+	N/A	N/A
StatPlus	S	+	+	+	+	+	+	-	-	+	+
SPlus	St	+	+	+	+	+	-	+	+	+	+
SPSS	S	+	+	+	+	+	+	+	+	+	+
StatsDirect	S	+	+	+	+	+	+	-	-	N/A	N/A
Statistix	2008	+	+	+	+	+	+	-	-	+	+
SYSTAT	S	+	+	+	+	+	+	+	+	+	+
Total Access Statistics	A	+	+	+	+	-	-	-	-	+	+
UNISTAT	S	+	+	+	+	+	+	+	+	+	+
The Unscrambler	S	+	+								
VisualStat	S	+	+	-	+	-	-	-	-	+	-
Winpepi	S	+	+	+	+	-	+	+	-	-	-
XLStat	X	+	+	+	+	-	+	+	+	N/A	+

Notas

1. ^ [a](#) [b](#) [c](#) Promo price. Check for availability. Regular prices are higher by up to 50%.

2. ^ [a] [b] [c] [d] [e] [f] [g] [h] [i] [j] [k] Academic discounts/bundles available.
3. ^ For noncommercial use
4. ^ "Wolfram Worldwide Web Store". http://store.wolfram.com. Retrieved 2008-11-20.
5. ^ Mathematica Home Edition Released Macworld, Feb 2009
6. ^ R is mainly a command line program but the Windows distribution comes with a GUI component called RGui.
7. ^ [a] [b] [c] [d] GUI interface for the R programming language
8. ^ [a] [b] [c] [d] Using R as platform
9. ^ LogitModelFit Mathematica documentation
10. ^ GeneralizedLinearModelFit Mathematica documentation
11. ^ ProbitModelFit Mathematica documentation
12. ^ BarChart Mathematica documentation
13. ^ BoxWhiskerPlot Mathematica documentation
14. ^ Histogram Mathematica documentation
15. ^ ListLinePlot Mathematica documentation
16. ^ ListPlot Mathematica documentation
17. ^ ListPointPlot3D Mathematica documentation
18. ^ [a] [b] S = Standalone executive; St = Standalone executive, primitive textual (DOS or terminal) interface; A = Access Add-in; X = Excel Plug-In
19. ^ Base Statistics (such as t-test, f-test, etc.)
20. ^ Normality Tests, data exploring
21. ^ Contingency Tables Analysis
22. ^ Base Data Processing, f.ex. sorting
23. ^ Extended (data sampling, transformation)
24. ^ Base Statistics (such as t-test, f-test, etc.)
25. ^ Normality Tests, data exploring
26. ^ Contingency Tables Analysis
27. ^ Base Data Processing, f.ex. sorting
28. ^ Extended (data sampling, transformation)

Anexo 6: Software gratuito

	DESC	FREQ	PROB	ANOVA1	ANOVA+	EXPER	SLR	MLR	LOG	LOGIT	PROBIT	GLM	ANCOVA	NONPAR	LOGLIN	TIME	SURV	PCA	FACT	CCA>	CA	DISCR	CLUST
ADE 4	•			•			•	•										•		•	•	•	•
DATAPLOT	•	•	•	•	•	•	•	•						•		•		•		•		•	
EASYREG	•	•	•				•	•		•	•	•		•		•	•						
GRETL	•	•	•				•	•	•	•	•	•		•		•		•					
INSTAT+	•	•	•	•	•		•	•						•	•	•	•						
MACANOVA	•	•	•	•	•		•	•	•	•		•	•			•		•	•				•
MATRIXER	•	•	•				•	•		•	•			•		•							
MICROSIRIS	•	•	•	•	•		•	•		•					•		•	•	•				•
OPENSTAT	•	•	•	•	•		•	•	•			•	•	•	•			•	•	•		•	•
R	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
TANAGRA	•			•			•	•	•					•				•	•		•	•	•
VISTA	•	•		•	•		•	•							•			•			•		•
WINIDAMS	•	•		•			•	•								•		•	•			•	•

DESC: Descriptive Statistics
FREQ: Distribution Frequencies
PROB: Probability Distributions
ANOVA1: One Way Analysis of Variance
ANOVA+: Two or Three Way Analysis of Variance
EXPER: Experimental Design
SLR: Simple Linear Regression
MLR: Multiple Linear Regression
LOG: Logistic Regression
LOGIT: Logit Model
PROBIT: Probit Model

GLM: Generalized Linear Models
ANCOVA: Analysis of Covariance
NONPAR: Non Parametric Test
LOGLIN: Log-Linear Analysis
TIME: Time Series
SURV: Survival Analysis
PCA: Principal Component Analysis
FACT: Factorial Analysis
CCA: Canonical Correlation Analysis
CA: Correspondence Analysis
DISCR: Discriminant Analysis
CLUST: Cluster Analysis

Referencias

"A Short Preview of Free Statistical Software Packages for Teaching Statistics to Industrial Technology Majors" Ms. Xiaoping Zhu and Dr. Ognjen Kuljaca. Journal of Industrial Technology, (Volume 21-2, April 2005).

http://gsociology.icaap.org/methods/soft.html

http://en.citizendium.org/wiki/Free_statistical_software

http://statpages.org/javasta2.html

Free Statistical Software Comparison http://en.freestatistics.info/comp.php