UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES
"ACATLÁN"

Análisis comparativo de la exactitud del software estadístico
SAS v9.0, SPSS v15.0 y las herramientas estadísticas de EXCEL v2003

TESINA

QUE PARA OBTENER EL TÍTULO DE

LICENCIADO EN
MATEMÁTICAS APLICADAS Y COMPUTACIÓN

PRESENTA

CARLOS JAVIER RODRÍGUEZ ABAD

Asesor: Fis.Mat. Jorge Luis Suárez Madariaga

MAYO 2008

Universidad Nacional Autónoma de México
UNAM
Biblioteca Central
Dirección General de Bibliotecas de la UNAM

UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Agradecimientos

Si bien para mí la titulación de la licenciatura representa alcanzar el primer gran objetivo de mi formación profesional, también significa volver realidad un sueño familiar del cual soy el principal protagonista. Aquí, haciendo un ejercicio de retrospectiva desde la antesala de este culminante acontecimiento, puedo decir que me siento satisfecho y orgulloso del camino andado, por ello ofrezco mis agradecimientos.

Agradezco a mi familia. Son parte de mi propio ser y siempre los llevo conmigo, los amo: Abuelita Elena, abuelo Laco, abuelita Chabela, tío Neto, mamá Antonieta, tío José, María, tío Toño, Asael, Gely, Erick, Evelyn, Edgar, Luz Elena, Marina y Asaelito.

Agradezco a mis amigas y amigos. Es un privilegio conocer personas como ustedes y compartir emociones, sentimientos y momentos tan especiales lado a lado, es un placer. Mi cariño y aprecio.

Agradezco a las personas que han cruzado por mi camino y han dejado huella por brindarme una enseñanza, un apoyo o simplemente darme la oportunidad de darles algo de mí. A mis vecinos, profesores, compañeros, conocidos, etc. Mi respeto y admiración.

Un agradecimiento especial para la Maestra en Estadística María del Socorro Romero Hernández por el invaluable apoyo en la realización de este trabajo.

No puedo dejar de mencionar que dejé pasar aproximadamente siete años antes de decidirme a realizar el trabajo de titulación, lo cual reconozco me resulta embarazoso, sin embargo he tenido la satisfacción de enfrentar y vencer retos de índole laboral.

Índice

Capítulo IV

Introducción

Frecuentemente durante el curso de la licenciatura cuando se comienza a utilizar el software estadístico e incluso después en el desempeño de la práctica profesional cuando el uso del software estadístico es cotidiano, surgen interrogantes como ¿cuál es el mejor?, ¿cuál es el más potente?, ¿cuál es el más sencillo de manejar?, ¿cuál es el más versátil?, etc., sin embargo la interrogante ¿cuál es el más exacto?, no surge como una pregunta realmente inquisidora porque se da por sentada la exactitud del software estadístico comercial.

Pese a la trascendencia de la exactitud son muy pocos los usuarios de software estadístico que tienen la inquietud y una referencia concreta de como evaluarla. Existen artículos al respecto que se han publicado en revistas especializadas de estadística en Estados Unidos, pero son poco conocidos. Con el propósito de realizar un análisis comparativo entre la exactitud en las estimaciones de SAS, SPSS y EXCEL, en este trabajo se expone de manera breve, sencilla y práctica dos metodologías complementarias para evaluar la exactitud de las estimaciones de software estadístico, así como la forma de presentar e interpretar los resultados obtenidos.

Las metodologías incluyen:

1) Por una parte la evaluación de la media, la desviación estándar, el análisis de varianza de un factor (ANOVA) y la regresión lineal simple utilizando los Conjuntos de Datos de Referencia Estadística (StRD) del Instituto Nacional de Estándares y Tecnología (NIST).

2) Por otra parte la evaluación de las distribuciones estadísticas: normal, ji cuadrada, t de Student y F empleando los resultados del software Distribuciones Elementales (ELV) de la Universidad de Munich.

Como antesala a la exposición de las metodologías se repasa el marco teórico estadístico, la relación que existe entre la estadística y la informática, así como los productos comerciales más recientes y los conceptos involucrados como son la exactitud, la precisión, el error de redondeo, el error de truncamiento, el error numérico y como medirlo, entre otros.

Capítulo I

1 MARCO TEÓRICO ESTADÍSTICO

1.1 Media muestral

La media es el promedio aritmético de una muestra de datos, representa el valor central de esta muestra, y es el estimador de la media poblacional (μ).

Se define como: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Donde:

n = número de observaciones de la muestra

x_i = valor de cada observación de la muestra

Las propiedades de la media muestral son:

1. Todo conjunto de datos de nivel de intervalo y de nivel de razón tiene un valor medio.

2. Al evaluar la media se incluyen todos los valores.

3. Un conjunto de valores sólo tiene una media.

4. La media es la única medida de ubicación donde la suma de las desviaciones de cada valor con respecto a la media, siempre es cero:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

5. Cuando el número de observaciones de la muestra es pequeño, la media es muy sensible a valores extremos.

6. La media no es necesariamente igual a un valor de la muestra.

1.2 Desviación estándar muestral

Para medir la variabilidad o dispersión de los datos, se toma un valor de los datos como punto de referencia, respecto del cual se cuantifica la variabilidad. Este punto de referencia es la media aritmética.

Sea $X = (x_1, x_{2,}, \ldots, x_n)$ los valores de una variable que pertenecen a una muestra, se define la varianza muestral, como el promedio de los desvíos respecto a la media elevados al cuadrado.

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

La varianza cuantifica la dispersión de los datos en la unidad de medida de la variable pero al cuadrado, provocando que sea bastante difícil el poder imaginar tal variabilidad, aunque no por eso deja de ser útil, ya que ésta tiene un sinnúmero de aplicaciones.

Se define la desviación estándar como la raíz cuadrada de la varianza. Se pueden tener desvíos positivos, si el dato es mayor que la media aritmética; o desvíos negativos, si el dato es menor que la media aritmética.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

La desviación estándar juega un rol trascendental en el análisis de los datos, ya que, así como para medir longitudes se tiene el metro como unidad de medida, para cuantificar pesos se tiene el kilogramo como unidad de medida, en el campo de la estadística, interesa cuantificar la dispersión de los datos, teniendo en este caso como unidad de medida a la desviación estándar.

1.3 Distribuciones estadísticas

1.3.1 Distribución normal

La distribución normal es la más utilizada en la práctica y una de las distribuciones teóricas mejor estudiadas, también llamada distribución Gaussiana. Su importancia se debe fundamentalmente a la frecuencia con la que distintas variables asociadas a fenómenos naturales y cotidianos siguen, de manera aproximada, esta distribución. Además de que se puede demostrar que bajo ciertas condiciones, la distribución normal puede ser usada para aproximar una gran variedad de distribuciones en muestras grandes, a través del Teorema del Límite Central.

La distribución de una variable aleatoria normal está completamente determinada por dos parámetros, media y desviación estándar, denotadas generalmente por μ y σ. Con esta notación, la función densidad de la normal viene dada por la ecuación:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\};-\infty < x < \infty$$

Esta función determina la curva en forma de campana. Así, se dice que una variable aleatoria X sigue una distribución normal de media (μ) y varianza (σ^2) y se denota como:

$$X \approx N(\mu, \sigma)$$

La distribución normal posee ciertas propiedades importantes que conviene destacar:

1. Tiene una única moda, que coincide con su media y su mediana.

2. La curva normal es asintótica al eje de abscisas. Por ello, cualquier valor entre $-\infty$ y $+\infty$ es teóricamente posible. El área total bajo la curva es, por tanto, igual a 1.

3. Es simétrica con respecto a su media (μ).

4. La distancia entre la línea trazada en la media y el punto de inflexión de la curva es igual a una desviación estándar (σ). Cuanto mayor sea σ, más aplanada será la curva de la densidad.

5. La forma de la campana de Gauss depende de los parámetros μ y σ. La media indica la posición de la campana, de modo que para diferentes valores de μ, la gráfica es desplazada a lo largo del eje horizontal. Por otra parte, la desviación estándar determina el grado de apuntamiento de la curva. Cuanto mayor sea el valor de σ, más se dispersarán los datos en torno a la media y la curva será más plana. Un valor pequeño de este parámetro indica, por tanto, una gran probabilidad de obtener datos cercanos al valor medio de la distribución.

Como se deduce, no existe una única distribución normal, sino una familia de distribuciones con una forma común, diferenciadas por los valores de su media y su varianza. De entre todas ellas, la más utilizada es la distribución normal estándar, que corresponde a una distribución de media = 0 y varianza = 1. Así, la expresión que define su función de densidad es:

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right); -\infty < z < \infty$$

Es importante saber que, a partir de cualquier variable X que siga una distribución $N(\mu,\sigma)$, se puede obtener otra variable Z con una distribución normal estándar, con la transformación:

$$Z = \frac{x-\mu}{\sigma}$$

De aquí se tiene que:

$$\mathrm{P}(Z \le z) = \mathrm{P}\left(\frac{x-\mu}{\sigma} \le z\right) = \mathrm{P}(X \le z\sigma + \mu) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z\sigma+\mu} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt$$

con $t = \frac{x-\mu}{\sigma}$

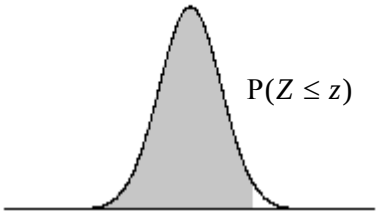


Figura 1.1

Esta propiedad es importante en la práctica, ya que la función $e^{-\frac{t^2}{2}}$ no tiene una antiderivada que pueda escribirse en términos de la función y no se puede integrar directamente, para lograrlo se cambian los términos a coordenadas polares. Sin embargo, existen tablas publicadas a partir de las que se puede obtener de modo sencillo la probabilidad de observar un dato menor o igual a un cierto valor z, y que permitirán resolver preguntas de probabilidad acerca del comportamiento de variables de las que se sabe o se asume que siguen una distribución aproximadamente normal.

Las probabilidades más usadas son aquellas que indican que un dato esta contenido a 1, 2 ó 3 desviaciones estándar de la media y son:

$$\mathrm{P}(|X-\mu|\leq\sigma)=\mathrm{P}(|Z|\leq 1)=0.6826$$

$$\mathrm{P}(|X-\mu|\leq 2\sigma)=\mathrm{P}(|Z|\leq 2)=0.9544$$

$$\mathrm{P}(|X-\mu|\leq 3\sigma)=\mathrm{P}(|Z|\leq 3)=0.9974$$

1.3.2 Distribución ji cuadrada

La distribución ji cuadrada, sólo tiene un parámetro *k* que representa los grados de libertad de la variable aleatoria:

La función de densidad de la ji cuadrada χ_k^2 está dada por:

$$f(x)=\frac{1}{\Gamma\left(\frac{k}{2}\right)2^{\frac{k}{2}}}x^{\left(\frac{k}{2}-1\right)\left(e^{\frac{-x}{2}}\right)}$$

Donde:

$x\geq 0 \;\; y \;\; f(x)=0 \;\; para \;\; x\leq 0$

Entonces $\mathrm{P}(X\leq x)=\int_0^x \frac{1}{\Gamma\left(\frac{k}{2}\right)2^{\frac{k}{2}}}x^{\left(\frac{k}{2}-1\right)\left(e^{\frac{-x}{2}}\right)}dx$

Sus principales aplicaciones se derivan en el análisis de varianzas y pruebas de bondad de ajuste.

Las propiedades de la distribución ji cuadrada son:

1. Si $Z \sim N(0,1)$ entonces $Z^2 \sim \chi_1^2$ esto es, el cuadrado de una variable aleatoria normal estándar es una variable aleatoria con distribución ji cuadrada y un grado de libertad. De forma general se tiene que: $X = Z_1^2 + \ldots Z_K^2 \sim \chi_k^2$

2. Si $X_1, \ldots, X_n$ son independientes y $X_i \sim \chi_{pi}^2$, entonces: $X_1 + \ldots + X_2 \sim \chi_{p1+p2+\ldots+pn}^2$

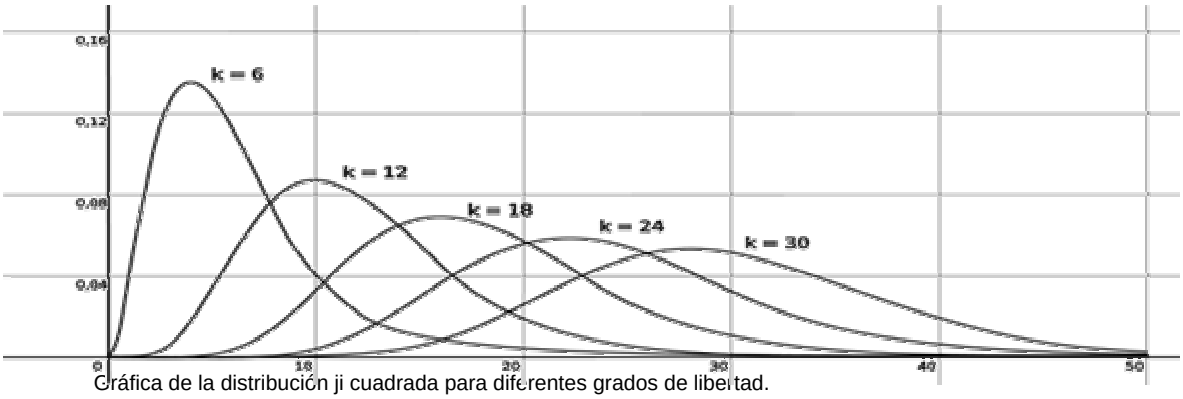
3. Sean X_1 y X_2 dos variables aleatorias. Si X_1 tiene una distribución ji cuadrada con *k* grados de libertad, y $X_1 + X_2$ tiene otra distribución ji cuadrada con *j* > *k* grados de libertad, entonces X_2 tiene una distribución ji cuadrada con *j* - *k* grados de libertad.

4. La media y varianza son K y 2k respectivamente.

5. Si $\bar{x}$ y S^2 son la media y la varianza de una muestra aleatoria tomada de una población normal con media (μ) y varianza (σ^2), entonces $\bar{x}$ y S^2 son independientes. Y la variable aleatoria: $\frac{(n-1)S^2}{\sigma^2}$ tiene una distribución ji cuadrada con $n-1$ grados de libertad.

6. Cuando el tamaño de la muestra aumenta, se aproxima a la distribución normal.

La función de distribución no puede calcularse en forma analítica; sin embargo, ha sido tabulada para diferentes valores de la probabilidad acumulada, y para varios grados de libertad, véase la figura 1.2.



Gráfica de la distribución ji cuadrada para diferentes grados de libertad.

Figura 1.2

1.3.3 Distribución t de Student

La distribución t de Student se construye como un cociente entre una variable normal estándar y la raíz de una variable con distribución ji cuadrada siendo ambas independientes.

Sea $Z \sim N(0,1)$ y $Y \sim \chi_n^2$, entonces la distribución t de Student con *n* grados de libertad t_n está dada por:

$$T = \frac{Z}{\sqrt{\frac{1}{n}Y}} \sim t_n$$

La función de densidad de t_n es:

$$f_T(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}; -\infty < x < \infty$$

Las propiedades de la distribución t de Student son parecidas a la $N(0,1)$:

1. Si $X_0, X_1, \ldots, X_n$ son variables aleatorias idénticamente distribuidas, entonces:

$$\frac{X_0}{\sqrt{\frac{1}{n}(X_1^2 + X_2^2 + \ldots + X_n^2)}} \sim t_n$$

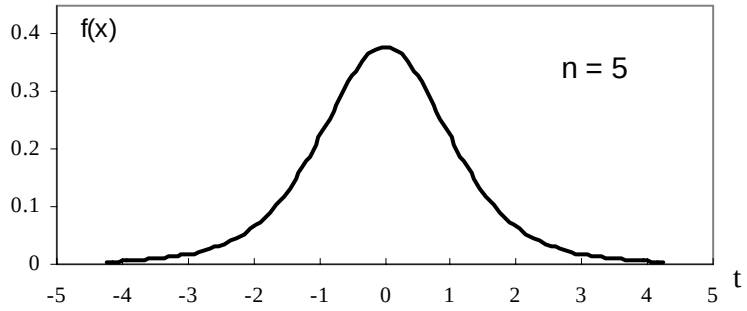
2. Es de media cero y simétrica con respecto a la misma.

3. Es algo más dispersa que la normal, pero la varianza decrece hasta 1 cuando el número de grados de libertad aumenta.

4. Para un número alto de grados de libertad se puede aproximar la distribución t de Student por la normal. Esto es: $t_n \to N(0,1), \ \textit{cuando} \ n \to \infty$

5. Para calcular las probabilidades asociadas a la distribución se tiene que:

$$P(T \le t) = F_T(t) = \int_{-\infty}^{t} f_T(x)dx = \int_{-\infty}^{t} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} dx$$



Gráfica de la distribución t de Student.

Figura 1.3

1.3.4 Distribución F

Surge del cociente de dos variables ji cuadrada independientes, cada una dividida por sus respectivos grados de libertad.

Sea $Z \sim \chi^2_m$ y $Y \sim \chi^2_n$, entonces la distribución F con *m* y *n* grados de libertad respectivamente $F_{m,n}$ está dada por:

$$X = \frac{Z/m}{Y/n} \sim F_{m,n}$$

La función de densidad de esta variable aleatoria X está dada por la siguiente expresión:

$$f(x) = \frac{\Gamma\left[\frac{n+m}{2}\right]}{\Gamma\left[\frac{n}{2}\right]\Gamma\left[\frac{m}{2}\right]}\left[\frac{m}{n}\right]^{\frac{m}{2}} . x^{\frac{n}{2}-1}\left[1+\frac{mx}{n}\right]^{-\frac{n+m}{2}}, \quad x \geq 0$$

Esta distribución tiene dos parámetros: *m* y *n* que representan los *grados de libertad*, y estos números coinciden con los grados de libertad de las variables ji cuadrada del numerador y denominador, respectivamente.

Las propiedades de la distribución F son:

1. La variable F toma valores de 0 a $+\infty$, por ser cociente de dos variables que asumen valores positivos.

2. Si $Z_1, Z_2, \ldots, Z_m$ son variables aleatorias independientes idénticamente distribuidas $N(0,1)$ y $Y_1, Y_2, \ldots, Y_n$ también son variables aleatorias independientes idénticamente distribuidas $N(0,1)$, entonces:

$$\frac{\frac{1}{m}(Z_1^2 + Z_2^2 + Z_3^2 + \ldots + Z_m^2)}{\frac{1}{n}(Y_1^2 + Y_2^2 + Y_3^2 + \ldots + Y_n^2)} \sim F_{m,n}$$

3. Si $X \sim F_{m,n}$, entonces su media y variancia son:

$$E(X) = \frac{n}{n-2}, n > 2 \qquad V(X) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}, n > 4$$

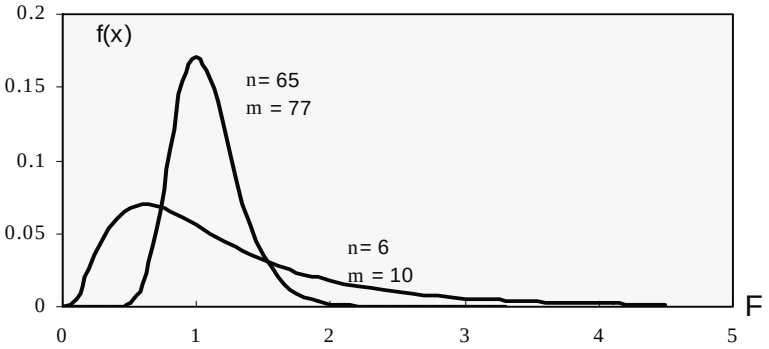
4. De la segunda propiedad resulta que: si $Z_1, Z_2, \ldots, Z_m$ son variables aleatorias independientes idénticamente distribuidas $N(\mu_z, \sigma_z)$ y $Y_1, Y_2, \ldots, Y_n$ también son variables aleatorias independientes idénticamente distribuidas $N(\mu_y, \sigma_y)$, entonces con Z_i y Y_j independientes $\forall$ i,j entonces:

$$\frac{\frac{1}{m}\sum_{i=1}^{m}\left[\frac{Z_i - \mu_z}{\sigma_z}\right]^2}{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i - \mu_y}{\sigma_y}\right]^2} \sim F_{m,n} \qquad \textit{ya que} \quad \sum_{i=1}^{m}\left[\frac{Z_i - \mu_z}{\sigma_z}\right]^2 \sim \chi_m^2$$

$$\textit{y} \quad \sum_{i=1}^{n}\left[\frac{Y_i - \mu_y}{\sigma_y}\right]^2 \sim \chi_n^2$$

5. La distribución F de Snedecor es positivamente asimétrica, pero su asimetría se reduce a medida que aumentan los grados de libertad m y n.

6. Si $X \sim F_{m,n}$, *entonces* $1/X \sim F_{n,m}$



Gráfica de la distribución F, para diferentes grados de libertad.

Figura 1.4

1.4 Análisis de varianza de un factor (ANOVA)

El análisis de varianza de un factor o en un sentido es un método estadístico para comparar más de dos medias poblacionales. El término de *un factor* o *sentido* indica que existe una característica o tratamiento que distingue las diferentes poblaciones entre sí.

El procedimiento para comparar estos valores está basado en la varianza global observada en las poblaciones a comparar.

La hipótesis nula a probar, con un nivel de significancia de α es:

Ho: $\mu_i = \mu_2 = \ldots = \mu_a$

Ha: al menos una media poblacional es diferente.

Para obtener resultados verosímiles, las poblaciones deben cumplir los siguientes supuestos estadísticos:

1. Las a poblaciones son independientes entre sí, están normalmente distribuidas y tienen varianzas iguales.
2. La variable dependiente debe medirse al menos a nivel de intervalo.
3. Se tiene una muestra aleatoria de tamaño n_1 de la población normal N_1 con una media μ_1 y desviación estándar σ_1, de igual forma una muestra de tamaño n_2 de la población normal N_2 con una media μ_2 y desviación estándar σ_2, y así sucesivamente para las poblaciones restantes, no es necesario que el tamaño de la muestra sea el mismo para cada población, como se muestra en la tabla 1.1.

Tratamiento	Observaciones	Suma por tratamiento	Media por tratamiento
1	$y_{11}\ \ y_{12}\ \ \cdots\ \ y_{1n_1}$	$y_{1\bullet} = \sum_{j=1}^{n_1} y_{1j}$	$\bar{y}_{1\bullet} = \frac{y_{1\bullet}}{n_1}$
2	$y_{21}\ \ y_{22}\ \ \cdots\ \ y_{2n_2}$	$y_{2\bullet}$	$\bar{y}_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	$y_{a1}\ \ y_{a2}\ \ \cdots\ \ y_{an_a}$	$y_{a\bullet}$	$\bar{y}_{a\bullet}$

Tabla 1.1

La suma total de todos los tratamientos está dada por:

$$y_{\bullet\bullet} = \sum_{i=1}^{a} y_{i\bullet} = \sum_{i=1}^{a} \sum_{j=i}^{n_j} y_{ij}$$

Y la media general está dada por:

$$\bar{y}_{\bullet\bullet} = \frac{y_{\bullet\bullet}}{\sum_{i=1}^{a} n_i}$$

Bajo las condiciones anteriores y considerando que la capacidad promedio de respuesta de todas las unidades experimentales o de medición antes de aplicar los tratamientos es la misma, es decir μ y que si todas se observaran en condiciones lo más similares posible las respuestas serían las variables aleatorias con el modelo lineal siguiente:

$$y_{ij} = \mu_i + e_{ij}; \quad j = 1,\ldots,n_i; \quad i = 1..,a$$

Donde los errores de medición e_{ij} son independientes, todos con distribución normal, esto es:

$$e_{ij} \sim N(0,\sigma^2) \ \ para \ \ j = 1,\ldots,n_i; \ \ i = 1..,a$$

El modelo lineal anterior se puede escribir de otra manera atendiendo a la forma en que se realiza un diseño experimental. Se sigue que al aplicar el tratamiento *i-ésimo* a un grupo de tamaño $n_{i,}$ se introduce un efecto τ_i de ese tratamiento en las variables por observar. El modelo puede entonces reescribirse como:

$$y_{ij} = \mu_i + \tau_i + e_{ij}; \quad j = 1,\ldots,n_i; \ \ i = 1..,a; \ \ e_{ij} \sim N(0,\sigma^2)$$

Para probar la hipótesis nula, y por el supuesto de que las *a* poblaciones o tratamientos tienen varianza común esto es $\sigma_1^2 = \sigma_2^2 = ... = \sigma_a^2 = \sigma^2$, el procedimiento descansa en una partición de la variabilidad de las observaciones. Para obtener la partición se debe considerar la siguiente igualdad:

$$Y_{ij} - \bar{y}_{\bullet\bullet} = (Y_{ij} - \bar{y}_{i\bullet}) + (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})$$

Donde la desviación total de una observación respecto de la media general del experimento ha sido descompuesta en dos fuentes de variación:

$(Y_{ij} - \bar{y}_{i\bullet})$ Esta diferencia se debe exclusivamente a la variabilidad experimental entre diferentes repeticiones del mismo tratamiento y por lo tanto es una desviación debido al error experimental.

$(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})$ Es una desviación entre la media del i-ésimo tratamiento y la media general del experimento, es una desviación que refleja las diferencias entre los tratamientos, ya que si todos los tratamientos tuvieran el mismo efecto, se debería esperar que las medias $\bar{y}_{i\bullet}$ fuesen iguales a la media general, por esta razón a esta diferencia se le llama desviación debida a los tratamientos.

$$Y_{ij} - \bar{y}_{\bullet\bullet} = (Y_{ij} - \bar{y}_{i\bullet}) + (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})$$

Desviación total | Desviación debida al error | Desviación debida al tratamiento

Como la igualdad anterior se cumple para todas las observaciones del experimento se puede escribir lo siguiente:

$$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij} - \bar{y}_{\bullet\bullet})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij} - \bar{y}_{i\bullet})^2 + \sum_{i=1}^{a} n_i(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2$$

Suma de cuadrados totales (SCT) | Suma de cuadrados del error (SCE) | Suma de cuadrados de los tratamientos (SCTr)

Cuando se divide SCE entre los correspondientes grados de libertad se obtiene el cuadrado medio del error (CME):

$$CME = \frac{SCE}{n-a}$$

Y al dividirse la SCTr entre los correspondientes grados de libertad se obtiene el cuadrado medio de los tratamientos (CMTr):

$$CMTr = \frac{SCTr}{a-1}$$

SCT representa la variación total de los datos y esta dada por: $SCT = SCE + SCTr$.

Y sus grados de libertad correspondientes son: $n-1 = (n-a) + (a-1)$

La partición de la variabilidad permite justificar una tabla de análisis de varianza, utilizando una prueba F ya que $\frac{SCE}{\sigma^2}$ y $\frac{SCTr}{\sigma^2}$ siguen una distribución ji cuadrada, véase tabla 1.2.

Tabla de ANOVA

Fuente de variación	g.l.	Suma de cuadrados	Media de cuadrados	F
Tratamientos	$a-1$	SCTr	$CMTr = \frac{SCTr}{a-1}$	$\frac{CMTr}{CME}$
Error	$N-k$	SCE	$CME = \frac{SCE}{n-a}$	
Total	$N-1$	SCT		

Tabla 1.2

CME y CMTr proporcionan una estimación de la varianza común, y si ambas varianzas son iguales entonces F se aproxima a 1. Si existe una diferencia entre poblaciones, por ejemplo si la varianza dentro de los grupos CMTr es mayor a la varianza entre los grupos CME entonces F será mayor a 1. El estadístico F sigue una distribución F con $a-1$ y $n-a$ grados de libertad en el numerador y denominador respectivamente, de manera más general $F_{a-1,n-a}$.

1.5 Regresión lineal simple

La regresión lineal se refiere al análisis que se hace de la relación entre dos variables, que puede ser una dependencia funcional de una con la otra.

Se intenta investigar la naturaleza de la relación y construir modelos que la describan, con el propósito de predecir el comportamiento de una de ellas a partir de los valores de la otra.

Mediante el modelo de regresión lineal simple se pretende describir la relación entre dos variables, una llamada independiente o predictora y otra llamada dependiente, además de realizar inferencias sobre el comportamiento de la variable dependiente.

La magnitud de una de ellas (la variable dependiente) se supone que queda determinada por la magnitud de la segunda (la independiente), mientras que en el sentido inverso no se cumple. En lo sucesivo se denotará por *X* a la variable independiente y por *Y* a la dependiente.

Hablar de dependencia no implica una relación de causa–efecto. Las relaciones de causa–efecto entre la variable independiente y la variable dependiente no pueden ser establecidas en el análisis de regresión.

Si se supone que la relación entre X y Y puede "modelarse" con una línea recta, se tiene la ecuación:

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

Donde: β_0 : *ordenada al origen*

β_1 : *pendiente de la recta*

En una ecuación como ésta, que "modela" la relación entre dos variables, todas las observaciones (x_i, y_i) de las variables deben localizarse sobre la gráfica de la función, de tal forma que para un valor dado de *X* el valor de *Y* está perfectamente determinado. Si esto ocurre se dice que la relación entre las dos variables es de tipo determinístico.

Los modelos determinísticos son de utilidad limitada cuando se trata de representar las relaciones entre dos variables y al menos una de ellas se mide por medio de un experimento aleatorio. En estos casos, deben utilizarse modelos estadísticos, por ejemplo:

$$Y = \beta_0 + \beta_1 X + \varepsilon_i;\ \ i = 1,2,\ldots,n$$

Este modelo expresa simbólicamente una relación lineal imperfecta entre las variables X y Y. Dado que la relación no es determinística, se intenta encontrar la recta que mejor represente la tendencia exhibida por los datos en cada caso, es decir estimadores de β_0 y β_1 que permitan predecir con cierta confiabilidad el comportamiento de Y a partir de valores de X.

Los supuestos del modelo son:

1. Y es una variable aleatoria cuya distribución probabilística depende de X, es decir, para cada valor de X, Y es una variable aleatoria cuya media depende del valor de X.
2. Las varianzas de las distribuciones de Y son idénticas para todos los valores de X.
3. Los valores de Y deben ser estadísticamente independientes.
4. La distribución de Y para cualquier valor de X es normal.

En el modelo de línea recta los supuestos son:

$$Y = \beta_0 + \beta_1 X + \varepsilon_i;\ \ i = 1,2,\ldots,n$$

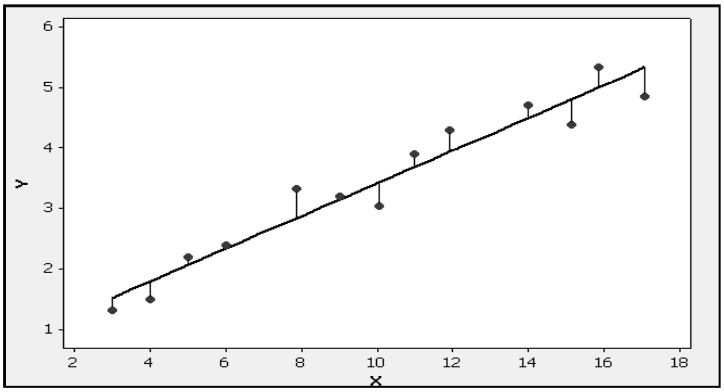
1. $x_1, x_{2,}, \ldots, x_n$ son observaciones de la variable independiente controladas por el experimentador.
2. β_0 y β_1 son parámetros desconocidos que determinan la recta de regresión.
3. ε_i son variables aleatorias no observables, independientes, distribuidas normalmente, con media cero y varianza σ^2.

$$\varepsilon_i = NID(0,\sigma^2);\ \ i = 1,2,\ldots,n$$

Los parámetros β_0 y β_1 son desconocidos y deben ser estimados usando datos de la muestra, para n pares de datos $(x_1, y_1),(x_2, y_2),\ldots,(x_n, y_n)$.

El principio de mínimos cuadrados es estimar β_0 y β_1 de tal manera que la suma de los cuadrados de la diferencia entre la observación y la línea recta sea el mínimo, como se muestra en la figura 1.5.

Al minimizar $S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$ se obtienen las ecuaciones normales de mínimos cuadrados:



Gráfica de regresión lineal simple.

Figura 1.5

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \qquad \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

Con solución:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \left(\frac{1}{n}\right)\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{\left(\sum_{i=1}^{n} x_i^2\right) - \left(\frac{1}{n}\right)\left(\sum_{i=1}^{n} x_i\right)^2}$$

Donde: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Por simplicidad se define:

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

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

De manera que: $\hat{\beta}_1 = \frac{SC_{xy}}{SC_{xx}}$

El modelo ajustado de regresión simple es:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

La diferencia entre el valor y_i observado y el correspondiente valor ajustado $\hat{y}_i$ es un residual. Matemáticamente el *i*-ésimo residual es:

$$e_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x\right); \quad i = 1,2,\ldots,n$$

Las propiedades del modelo ajustado por mínimos cuadrados son:

1. La suma de los residuales en cualquier modelo de regresión que contiene la ordenada al origen β_0, siempre es cero:

$$\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right) = \sum_{i=1}^{n} e_i = 0$$

2. La suma de los valores observados y_i es igual a la suma de los valores ajustados $\hat{y}_i$.

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$$

3. La línea de regresión de los mínimos cuadrados tiene el punto $\left(\bar{x}, \bar{y}\right)$ de centroide.
4. La suma de los residuales por el valor correspondiente de la variable independiente es igual a cero.

$$\sum_{i=1}^{n} x_i e_i = 0$$

5. La suma de los residuales por el valor ajustado correspondiente es igual a cero.

$$\sum_{i=1}^{n} \hat{y}_i e_i = 0$$

Además de β_0 y β_1 es necesario tener un estimador de σ^2 para realizar pruebas de hipótesis y construir intervalos de confianza para el modelo.

Si no se tiene información acerca de σ^2, se obtiene de los residuales o de la suma del cuadrado del error:

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

Un estimador insesgado de σ^2 es:

$$\hat{\sigma}^2 = \frac{SC_E}{(n-2)gl} = MC_E$$

Cualquier violación a los supuestos sobre el error y en las especificaciones del modelo puede afectar seriamente la utilidad de $\hat{\sigma}^2$ como estimador de σ^2.

Prueba de hipótesis para la pendiente β_1.

En este procedimiento se requiere asumir que los errores de modelo son normales e independientemente distribuidos con media 0 y varianza σ^2.

$$\varepsilon_i \approx NID\left(0, \sigma^2\right)$$

Se desea probar la siguiente hipótesis:

Ho: $\beta_1 = \beta_{10}$
Ha: $\beta_1 \neq \beta_{10}$

De manera que el estadístico se distribuye $N(0,1)$ si la hipótesis nula es cierta.

$$\hat{\beta}_1 \approx N\left(\beta_1, \frac{\sigma^2}{SC_{xx}}\right)$$

Si σ^2 es conocida se puede utilizar Z_0 para realizar la prueba de hipótesis:

$$Z_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{\sigma^2}{SC_{xx}}}}$$

Si σ^2 es desconocida, la media cuadrada de los residuales es un estimador insesgado de σ^2, entonces podemos utilizar el estadístico t_0 el cual tiene una distribución t de Student con $n-2$ grados de libertad si la hipótesis nula es cierta.

El estadístico t_0 es usado para probar Ho, comparando el valor observado con el valor teórico y rechazando la hipótesis nula si:

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MC_E}{SC_{xx}}}}$$

Con un nivel de confianza α.

$$|t_0| > t_{\alpha/2, n-2}$$

Para probar:

Ho: $\beta_0 = \beta_{00}$
Ha: $\beta_0 \neq \beta_{00}$

Se tiene que:

$$\hat{\beta}_0 \approx N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SC_{xx}}\right)\right)$$

Y el estadístico usado es:

$$t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MC_E\left(\frac{1}{n} + \frac{\bar{x}^2}{SC_{xx}}\right)}}$$

Se rechaza Ho si:

$$|t_0| > t_{\alpha/2, n-2}$$

Un caso importante de pruebas de hipótesis es:

Ho: $\beta_1 = 0$
Ha: $\beta_1 \neq 0$

Esta hipótesis se relaciona con la *significancia* de la regresión. Cuando no se rechaza implica que no existe relación lineal entre *X* y *Y*, como se ve en la figura 1.6.

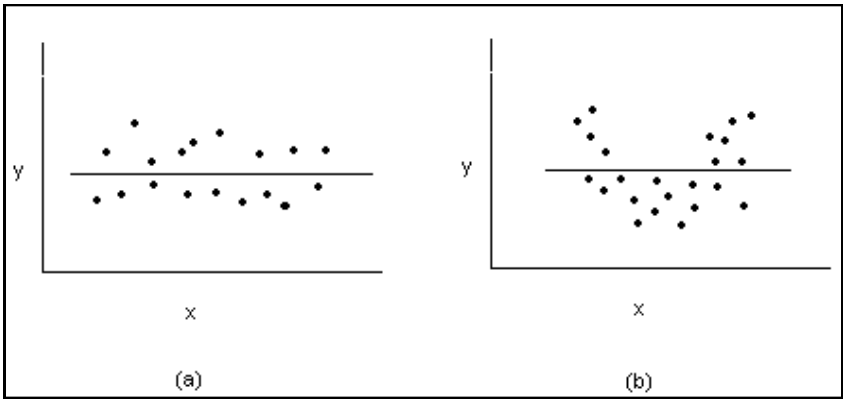


Figura 1.6

Alternativamente, si Ho se rechaza, esto implica que *X* sirve para explicar la variabilidad en *Y*. Esto se observa en la figura 1.7.

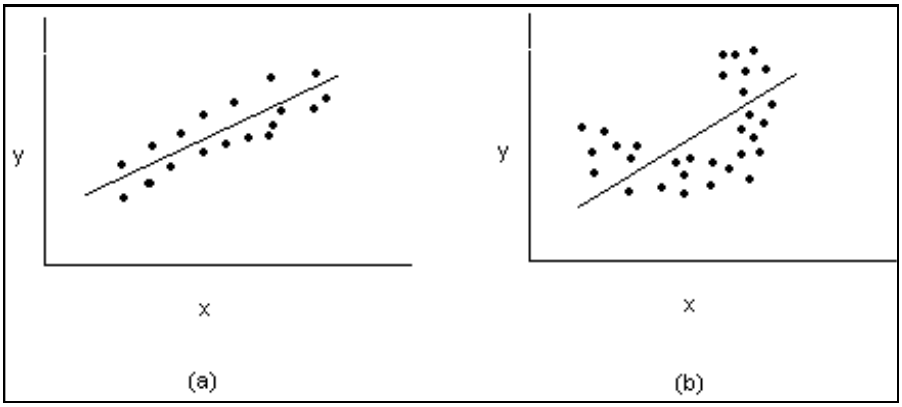


Figura 1.7

Rechazar Ho puede significar que:

a) Un modelo lineal es adecuado.

b) Se pueden obtener mejores resultados si se trabaja con polinomios de mayor orden en los términos de *X*.

Capítulo II

2 SOFTWARE ESTADÍSTICO

2.1 La estadística y la informática

La estadística es la ciencia que trata de la recopilación, organización, presentación, análisis e interpretación de datos numéricos con el fin de realizar una toma de decisión más efectiva.

El advenimiento de la computadora y el nacimiento de la informática abrieron nuevas fronteras para toda la ciencia, la estadística no fue la excepción. La necesidad de enfoques sistemáticos para el desarrollo y mantenimiento de productos de software se patentó en la década de 1960, pero fue hasta el año 1968 que se convocó una reunión en Garmisch, Alemania Oriental estimulándose el interés hacia los aspectos técnicos y administrativos utilizados en el desarrollo y mantenimiento del software, y fue entonces donde se utilizó el término “Ingeniería del Software”. Entonces surgió como una necesidad el hecho de que los productos de software deben ser desarrollados con base en la implementación de estándares mundiales, modelos, sistemas métricos, capacitación del recurso humano y otros principios y técnicas de la ingeniería de software que garanticen la producción de software de calidad y competitividad a nivel local e internacional.

El acelerado avance tecnológico de la información incrementa considerablemente la cantidad y la complejidad de los productos de software, así como también la exigencia en la funcionalidad y confiabilidad; es por esto que la calidad y la productividad se están constituyendo en las grandes preocupaciones tanto de gestores como de desarrolladores de software.

La estadística dentro de este contexto ha seguido un proceso largo de desarrollo y evolución, en nuestros días la relación entre la estadística y la informática proporciona el llamado software estadístico, que es un conjunto de programas y subprogramas integrados en una aplicación computacional con el objetivo de aplicar un método estadístico a un conjunto de datos y obtener un resultado. Aunque a nivel comercial existe software estadístico que permiten no sólo aplicar un método, sino metodologías y procedimientos estadísticos complejos, además de permitir el manejo de grandes volúmenes de información por medio de bases de datos y “data warehouse”; y por si fuera poco, brindan herramientas para manipulación, presentación y almacenamiento de los resultados.

El software estadístico se convierte en una herramienta analítica muy poderosa ya que además de recabar, resumir, reportar y almacenar datos permite la obtención de conclusiones para la toma de decisiones. Desde la década de los ochentas que la industria fijó su atención en el mejoramiento de la calidad por medio del uso de métodos estadísticos para crear una atmósfera que permitiera la manufactura de productos de alta calidad hasta la actualidad, además de la industria, las organizaciones, las instituciones, las universidades, etc; se han percatado de la necesidad de evaluar aspectos implicados en los diferentes procesos que les competen, asimismo se han percatado de la utilidad de las herramientas estadísticas para acceder a un mejor conocimiento de la información contenida en los datos

a través de metodologías y procesos de recogida, análisis e interpretación. Entre los beneficios otorgados por el software estadístico se cuentan un importante ahorro en tiempo, mayor precisión y una mejor interpretación de resultados con un manejo fácil y versátil.

La selección del software estadístico va de acuerdo con las necesidades e intereses que se tengan, algunas recomendaciones para elegir un software estadístico serían:

a) El costo de la licencia.
b) El nivel de sofisticación del usuario: idealmente tener conocimientos de estadística y de programación.
c) Tamaño del conjunto de datos: las computadoras modernas permiten manipular eficientemente conjuntos de datos cada vez más extensos.
d) Técnicas estadísticas incluidas en el software: por ejemplo, técnicas de muestreo, de análisis multivariado, pruebas no paramétricas multivariadas, variedad de gráficas etc.
e) Ergonomía: es el aspecto que tal vez encarezca determinado software estadístico, incluye contar con ayudas interactivas o tutoriales, representar gráficamente datos seleccionando y haciendo "clic", y exportar automáticamente los resultados a varios formatos, por ejemplo: html, pdf, etc.

Gran parte del software carece de un verdadero lenguaje de programación que resulte eficiente, es por eso que Chambers(2000) propone una lista de cinco condiciones básicas que debería cumplir toda herramienta informática empleada en estadística:

1. Especificación fácil de tareas sencillas
2. Capacidad de refinamiento gradual de las tareas
3. Posibilidades ilimitadas de extensión mediante programación
4. Desarrollo de programas de alta calidad
5. Posibilidad de integrar los resultados de los puntos 2 a 4 como nuevas herramientas informáticas.

2.2 Software estadístico disponible

En la actualidad la oferta de software estadístico dentro del mercado es bastante amplia y variada. Podemos encontrar software de propósito general como hojas de cálculo con herramientas de análisis estadístico sencillas, software con herramientas para técnicas muy específicas y limitadas, software con herramientas muy versátiles, software que puede añadir funcionalidad por medio de programación y software que ofrece todo un conjunto de productos adicionales a las herramientas estadísticas. A continuación se muestran algunas descripciones del software estadístico más usado, reconocido e importante:

EXCEL.

Es una herramienta eficaz usada para crear y aplicar formato a hojas de cálculo, y para analizar y compartir información para tomar decisiones mejor fundadas. La interfaz de usuario Fluent que muestra inmediatamente las herramientas más importantes, la visualización de datos enriquecida y las vistas de tabla dinámica permiten crear, de un

modo más sencillo, gráficos de aspecto profesional y de fácil uso. Ofrece mejoras significativas para compartir datos con más seguridad.

Principales características:

- Presenta la interfaz Fluent, un dispositivo que presenta los comandos organizados en un conjunto de fichas para ayudarle a encontrar herramientas muy eficaces cuando las necesite. Los menús y las barras de herramientas tradicionales se han sustituido por la cinta de opciones.
- Importar, organizar y explorar datos masivos con hojas de cálculo ampliadas significativamente.
- Usar el motor de gráficos rediseñado para comunicar los análisis en gráficos de aspecto profesional.
- Disfrutar de mayor y mejor compatibilidad para trabajar con tablas.
- Crear y trabajar interactivamente con vistas de tablas dinámicas fácilmente.
- "Ver" tendencias importantes y buscar excepciones en los datos.
- Con Excel Services permite compartir hojas de cálculo con mayor seguridad.
- La Ayuda garantiza al usuario y a la organización trabajar con la información empresarial más actual.
- Reducir el tamaño de las hojas de cálculo y mejorar la recuperación de archivos dañados a la vez.
- Ampliar las inversiones en inteligencia empresarial ya que es totalmente compatible con Microsoft SQL Server 2005 Analysis Services.

Aunque no es un software estadístico propiamente, forma parte del conjunto de herramientas básicas de la mayoría de las computadoras personales. Ofrece un conjunto de herramientas para el análisis de los datos denominado "Herramientas para Análisis", que es de gran importancia ante la falta de un software estadístico. Proporciona los estadísticos básicos para un análisis sencillo como lo son: estadísticas descriptivas, ANOVA, covarianza, regresión, prueba para dos medias, para dos varianzas y diversas distribuciones.

MINITAB.

Ofrece herramientas precisas y fáciles de usar para aplicaciones estadísticas generales y muy especialmente para control de calidad.

Es una herramienta informática compacta, versátil y de fácil manejo enfocada al análisis de datos complejos y a la identificación y resolución de problemas relativos a procesos, por ello se ha convertido en un instrumento fundamental para todas aquellas compañías con procesos productivos que requieren de un software de análisis para poder controlar fácilmente esos procesos o mejorar el rendimiento de sus cadenas de producción

Usado en instituciones universitarias, mencionado en publicaciones de estadística, es la herramienta predilecta en las industrias. La confiabilidad de sus algoritmos estadísticos y la sólida base de la combinación de potencia y simplicidad de manejo le han hecho merecer la confianza de los usuarios.

Principales herramientas estadísticas:

- Estadística básica y avanzada.
- Regresión y ANOVA.
- SPC.
- DOE - Diseño de experimentos.
- Gage R&R.
- MINITAB Análisis de fiabilidad.
- Tamaño de muestra y capacidad.
- Series de tiempo y predicción.
- Potente importación, exportación y manipulación de datos.
- Lenguaje de macros.

R.

Es un lenguaje y ambiente para cómputo estadístico y gráfico. Provee una amplia variedad de estadísticos (modelado lineal y no lineal, pruebas estadísticas clásicas, análisis de series de tiempo, clasificación, conglomerado…) y técnicas gráficas. Es altamente escalable, además provee una ruta de fuente libre para participar en investigación en metodología estadística.

Está disponible como software de libre acceso bajo los términos del Free Software Foundation's GNU General Public License en forma de código fuente. Se puede compilar y ejecutar en un extensa variedad de plataformas de UNIX y sistemas similares (incluyendo BSD Libre y LINUX), Windows y MacOS.

Integrado con un conjunto de facilidades de software para manejo de datos, cálculos y presentaciones gráficas que incluyen:

- Facilidad para el manejo y el almacenamiento de datos en forma efectiva.
- Un conjunto de operadores para cálculos con arreglos, en particular matrices.
- Una colección integrada, coherente y extensa de herramientas intermedias para análisis de datos.
- Facilidades gráficas para análisis de datos y presentación en pantalla o en copia digital.
- Un lenguaje de programación bien desarrollado, simple y efectivo que incluye condicionales, ciclos, funciones recursivas definidas por el usuario y facilidades para importación y exportación de datos.

El código desarrollado en C, C++ y Fortran se puede ligar y llamar en tiempo de ejecución para realizar tareas computacionales demandantes. Los usuarios avanzados pueden desarrollar código en C para manejar directamente las librerías de objetos incluidas.

SAS (Statistical Analysis Software).

Es un sistema de entrega de información que permite tomar la ventaja del acceso a cualquier fuente de datos, incluyendo archivos planos, archivos jerárquicos y los más importantes manejadores de bases de datos relacionales. También cuenta con un “data warehouse” propio para el manejo y almacenamiento de la información utilizada que soporta un amplio rango de aplicaciones. Los resultados pueden guardarse hasta como archivos planos y usarse para alimentar las aplicaciones de otros procesos o como fuente de

entrada para futuras ejecuciones. El lenguaje de programación añade mayor flexibilidad para personalizar el manejo, el análisis y el acceso a la base de datos.

SAS/STAT provee herramientas para necesidades analíticas especializadas e integrales, desde el clásico análisis de varianza y modelado predictivo hasta métodos exactos y técnicas de visualización estadística. Cada vez más organizaciones están recurriendo al software de análisis estadístico para dirigir los procesos de toma de decisiones. Por medio de la utilización de las técnicas estadísticas adecuadas se obtiene información innovadora que mejora los procesos; genera nuevas oportunidades de negocio y valor agregado; y ayuda a conservar el valor del producto y la satisfacción del cliente. Impulsa el proceso de avance científico aplicando las más recientes técnicas estadísticas y brindando soporte técnico a nivel doctorado Diseñado para analistas de negocios, estadísticos, investigadores e ingenieros. Cuenta con la aprobación corporativa y gubernamental

Principales herramientas estadísticas:

- Análisis de varianza.
- Regresión paramétrica y no paramétrica.
- Análisis de datos categóricos.
- Análisis multivariado.
- Análisis de sobrevivencia.
- Análisis psicométrico.
- Análisis de conglomerados.
- Análisis no paramétrico.
- Análisis para valores “missing”.

S-PLUS.

Herramienta de análisis estadístico gráfico que incorpora un potente lenguaje de programación orientado a objetos.

Es la solución para análisis exploratorio de datos y modelación estadística más avanzada; con más de 4,200 funciones de análisis de datos, incluido el mejor conjunto de métodos modernos y robustos; de manera sencilla se pueden importar datos, seleccionar las funciones estadísticas y mostrar los resultados; cuando el análisis requiere nuevos métodos se pueden modificar modelos existentes o desarrollar nuevos a través del lenguaje de programación S; provee de gran facilidad para examinar, visualizar y ejecutar funciones.

Algunas de las áreas de aplicación son la industria farmacéutica y biotecnológica, estadísticos medio ambientales y espaciales, negocios y finanzas, análisis financiero y econometría, ciencias matemáticas estadísticas e investigación operativa.

Principales características:

- Facilidad de manejo.
- Fácil acceso de datos.
- Librería de estadísticos.
- Integración con Microsoft Excel.

- Programación con el Lenguaje S.
- Interfaces para otros lenguajes.

SPSS (Statistical Product and Service Solutions).

Es una línea de productos modular, altamente integrada y con todas las funcionalidades necesarias para llevar a cabo cada paso del proceso analítico: planificación, recogida de datos, acceso y manejo de los datos, análisis, creación de informes y distribución de los mismos. En combinación con sus módulos adicionales así como otros módulos independientes que lo complementan proporcionan las más potentes herramientas para seguir el proceso analítico. La interfaz gráfica de usuario lo hace sencillo de utilizar dado que le proporciona toda la gestión de los datos, los estadísticos y los métodos de creación de informes que necesita para realizar todo tipo de análisis.

Puede generar información para la toma de decisiones de forma rápida utilizando potentes procedimientos estadísticos, comprender y representar de forma efectiva sus resultados en tablas y gráficos de alta calidad y compartir sus resultados con otros, utilizando una gran variedad de métodos de generación de informes, incluyendo una publicación en la web de forma segura. Permite tomar mejores decisiones más rápidamente, descubriendo factores clave, patrones y tendencias.

Utilizado para el análisis de bases de datos y minería de datos, investigación de mercados e investigaciones de todo tipo para resolver problemas reales de empresas e investigadores utilizando métodos estadísticos.

STATA.

Es un paquete estadístico completo, integrado, que ofrece todo lo que se requiera para el análisis, la administración de datos y su graficación. Integrado para Windows, Macintosh y Unix, diseñado para investigadores profesionales.

Es utilizado por investigadores médicos, bioestadísticos, epidemiólogos, economistas, sociólogos, científicos, políticos, especialistas en geografía, sicólogos, científicos sociales y otros investigadores profesionales que necesitan analizar datos o información estadística.

Lo mejor del mercado para análisis econométrico y epidemiológico, un paquete estadístico que ofrece un gran número de análisis estadísticos para investigadores en muchas disciplinas. Es especialmente práctico para profesionales que trabajan en investigación médica y económica, es tan programable que los desarrolladores y usuarios añaden cada día nuevas características, de ese modo responden a la creciente demanda de los investigadores. Además está al día en la web, se pueden compartir conjuntos de datos, programas, actualizaciones, ficheros de ayuda, etc. Podrá actualizarse siempre que esté conectado a Internet y cada vez que haya algo nuevo.

STAT/TRANSFER es la forma más fácil de mover datos entre hojas de cálculo, bases de datos y paquetes estadísticos

Principales herramientas estadísticas:

- Modelos Multinivel Mixtos
- Estadísticas Exactas
- Variables Endógenas
- Métodos Multivariados
- Estudios de panel
- Encuestas y datos recolectados
- Análisis con variables de Tiempos / Fechas
- Análisis de sobrevivencia
- Series de tiempo
- Métodos longitudinales

Dentro de las características del software estadístico, quizá la más importante sea la exactitud, sin embargo los fabricantes y vendedores rara vez proporcionan información que sirva como una referencia válida para evaluar este aspecto, mucho menos proporcionan detalles con respecto a la implementación de los algoritmos. Si acaso demuestran el funcionamiento con valores dentro de los límites de confiabilidad del producto o muestran cualquier otra evidencia de su exactitud.

Considerando el caso en el que un investigador soluciona un mismo problema usando dos aplicaciones de software estadístico diferentes. Por supuesto que el investigador no tiene manera de saber cual resultado es correcto o si alguno de los dos resultados es correcto. Otro caso posible es que el investigador únicamente utilice una sola aplicación de software estadístico y dé por sentada la exactitud al no tener evidencia para cuestionársela.

Así como no se habla de la exactitud tampoco se habla del concepto de error, que resulta inseparable al cómputo numérico y que es necesario conocer para comprender mejor el concepto de exactitud.

2.3 Errores en la computación numérica

Generalmente cuando las computadoras se utilizan para el cómputo numérico incurren en dos tipos de errores: por un lado tenemos el error de redondeo y por el otro el error de truncamiento. Buena parte del error inherente al cómputo numérico es atribuible a la representación aproximada de los números mismos, en tanto que el resto recae en la naturaleza aproximada de las propias soluciones.

2.3.1 Conceptos previos

Cifras significativas.

El concepto de cifras significativas o dígitos significativos se ha concebido para designar de manera formal la confiabilidad de un valor numérico. Determinar las cifras significativas de un número es un procedimiento sencillo, aunque en algunos casos genera cierta confusión. A veces los ceros no siempre son cifras significativas, ya que pueden usarse sólo para ubicar el punto decimal. Las cifras significativas de un número son la

primera no nula y todas las siguientes leídas de izquierda a derecha, considerando el punto decimal. Por ejemplo:

El número 0.20022008 tiene nueve cifras significativas.
Mientras que el número 0.00020022008 también tiene nueve cifras significativas.

Exactitud y precisión.

Los errores en cálculos se pueden caracterizar con respecto de su exactitud y su precisión. La exactitud se refiere a que tan cercano está el valor calculado del valor verdadero. La precisión se refiere a que tan cercanos se encuentran diversos valores calculados, unos de otros.

En la figura 2.1 se ilustran gráficamente los conceptos en analogía con una diana de práctica de tiro. Los agujeros en cada blanco representan los resultados de un cálculo; mientras que el centro del blanco representa el valor verdadero. a) Inexacto e impreciso; b) exacto e impreciso; c) Inexacto y preciso; d) exacto y preciso.

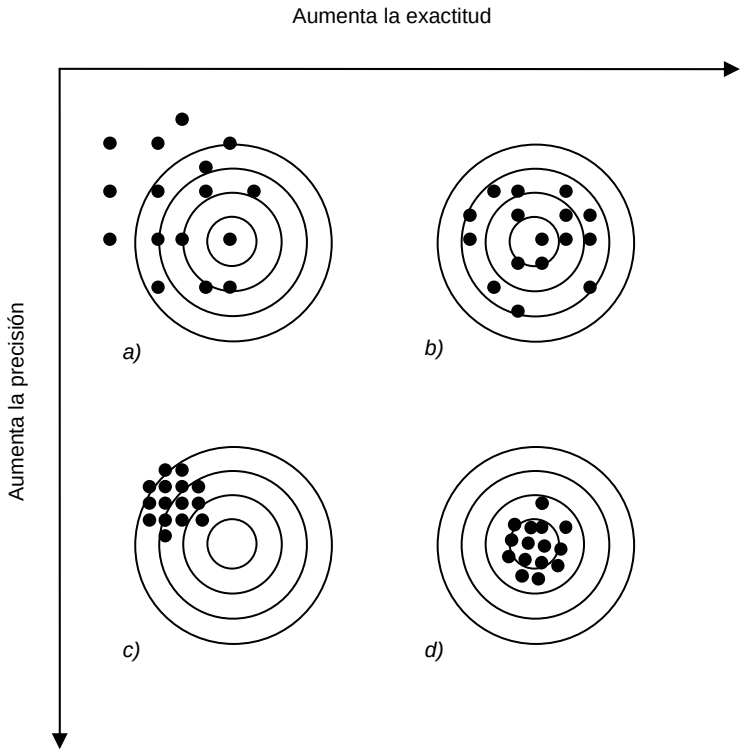


Figura 2.1

Sistemas numéricos.

Un sistema numérico es simplemente una convención para representar cantidades. Debido a que se tienen diez dedos en las manos y diez dedos en los pies, el sistema de numeración que nos es más familiar es el decimal o de base 10. Una base es el número que se usa como referencia para construir un sistema. El sistema de base 10 utiliza diez dígitos

(0, 1, 2, 3, 4, 5, 6, 7, 8, 9) para representar números, tales dígitos son satisfactorios por sí mismos para contar de 0 a 9.

Para grandes cantidades se usa la combinación de estos dígitos básicos; con la posición o valor de posición se especifica su magnitud. El dígito en el extremo derecho de un número entero representa un número del 0 al 9. El segundo dígito a partir de la derecha representa un múltiplo de 10. El tercer dígito a partir de la derecha representa un múltiplo de 100 y así sucesivamente. Por ejemplo si se tiene el número 86,409 se tienen ocho grupos de 10,000; seis de 1,000; cuatro de 100; cero de 10; y nueve unidades, o bien:

$$(8 \times 10^4) + (6 \times 10^3) + (4 \times 10^2) + (0 \times 10^1) + (9 \times 10^0) = 86{,}409$$

A este tipo de representación se llama notación posicional.

Debido a que el sistema decimal resulta ser tan familiar, no es común darse cuenta de que existen otras alternativas. Por ejemplo, si el ser humano tuviera ocho dedos en las manos y ocho en los pies, se tendría, sin duda, una representación en un sistema octal o de base 8. En tal sentido la computadora es como un animal que tiene dos dedos, limitado a dos estados: 0 ó 1. Esto se relaciona con el hecho de que las unidades lógicas fundamentales de las computadoras digitales sean componentes electrónicos de apagado / encendido. Por lo tanto, los números en la computadora se representan con un sistema binario o de base 2. Del mismo modo que con el sistema decimal, las cantidades pueden representarse usando la notación posicional. Por ejemplo, el número binario 11 en el sistema decimal es equivalente a:

$$(1 \times 2^1) + (1 \times 2^0) = 2 + 1 = 3$$

Representación binaria y precisión finita.

La representación que la computadora hace de los números está relaciona de manera directa con la forma en que los almacena en la memoria. La unidad fundamental mediante la cual se representa la información se llama palabra. Ésta es una entidad que consiste en una cadena de dígitos binarios o bits. La representación binaria del número decimal 0.1 es 0.000110011<u>0011</u>, donde los dígitos subrayados indican una secuencia periódica infinita. Las computadoras manejan una precisión finita y no pueden representar un número infinito de dígitos, de hecho, a la mayoría de los equipos personales, estaciones de trabajo y muchos servidores se les implementa el estándar IEEE-754[1] para la aritmética computacional, según el cual:

[1] IEEE corresponde a las siglas de The Institute of Electrical and Electronics Engineers, el Instituto de Ingenieros Eléctricos y Electrónicos, una asociación técnico-profesional mundial dedicada a la estandarización. IEEE-754 el título completo es: IEEE Standard for Binary Floating-Point Arithmetic (ANSI/IEEE Std 754-1985), el estándar de la IEEE para aritmética binaria de punto flotante define formatos para la representación de números en punto flotante y valores desnormalizados, así como valores especiales como infinito y NaN, con un conjunto de operaciones en punto flotante que trabaja sobre estos valores. También especifica cuatro modos de redondeo y cinco excepciones.

- La precisión sencilla tiene 6 ó 7 dígitos de precisión.
- La precisión doble tiene 15 ó 16 dígitos de precisión.

Por defecto se asume una precisión sencilla que es más fácil de usar que la precisión doble, es decir que la computadora representa números en forma binaria con 24 dígitos binarios excluyendo los primeros ceros y considerando el punto decimal. Por ejemplo:

a) El número 0.00011001100110011001100110 es la representación binaria a precisión sencilla del número decimal 0.1. Al convertir este número binario de vuelta a notación decimal se obtiene 0.099999964, lo cual es preciso a siete posiciones decimales. Esto es lo que la computadora interpreta cuando se captura el número decimal 0.1.

b) El número 11000011010100000.000000 es la representación binaria a precisión sencilla del número decimal 100,000, que al convertirlo de vuelta se obtiene 100,000 exactamente.

c) El número 11000011010100000.000110 es la representación binaria a precisión sencilla de la suma de los dos números binarios anteriores, es claro que únicamente se dispone de siete posiciones a la derecha del punto decimal para representar el número decimal 0.1. Reconvertido a decimal, este número no es 100,000.1 sino 100,000.09375, el cual es preciso únicamente a dos posiciones decimales. Así en este caso, sumando un número preciso a siete decimales más un número exacto resulta una suma menos precisa que cualquiera de sus sumandos.

d) Por otra parte, restar el número mayor no mejora los resultados, para una precisión sencilla la computadora interpreta (100,000.1 – 100,000.0) como 0.09375. Esto comprueba que para números muy grandes que difieren únicamente por decimales, la computadora no es capaz de representar precisamente las diferencias entre los números.

Algoritmos iterativos y procesos finitos.

Existen algoritmos que por naturaleza solamente producen un resultado exacto para un número infinito de iteraciones. Si consideramos el cálculo del seno de x:

$$seno(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \ldots$$

La solución consiste de una suma infinita de términos. Las computadoras manipulan procesos finitos y no pueden ejecutar algoritmos para un número infinito de iteraciones, por tanto suponen un número predeterminado de iteraciones. La computadora ejecuta algoritmos para un número de iteraciones k, éste número es determinado por los recursos disponibles, por las necesidades del problema a resolver o por parámetros o criterios preestablecidos. La solución calculada de seno de x es presentada en la tabla 2.1.

Es notorio que la aproximación a la solución mejora al aumentar el número de términos calculados o del número de iteraciones realizadas.

No. de Términos	Seno (π)
K = 3	0.524043913417169
K = 6	-0.000445160238209
K = 9	0.000000022419510
K = 12	-0.000000000000170

Tabla 2.1

2.3.2 Error de redondeo

El error de redondeo es una limitación del hardware derivada del hecho que la computadora tiene exclusivamente una determinada cantidad de bits para representar cualquier número, esto implica que los números son representados de una manera aproximada acotando el número de cifras significativas. A la discrepancia que existe entre el número exacto y la representación aproximada se le llama error de redondeo.

Suponiendo dos números:

- x = 1,000,000
- y = 0.000001

Ambos pueden ser representados en precisión sencilla, mas no así la suma:

- z = x + y

En precisión sencilla el resultado sería:

- z = 1,000,000

Perdiendo las cifras significativas del número menor a causa del error de redondeo. En precisión doble, la suma puede ser exactamente representada:

- z = 1,000,000.000001

El error de cancelación es un tipo de error de redondeo que se presenta al calcular la resta de dos números casi iguales, dando como resultado los dígitos de la extrema derecha, justamente los dígitos más propensos al error de redondeo. Consecuentemente en una serie de cálculos los errores de redondeo lejos de cancelarse, se acumulan y el límite del error total es proporcional al número de cálculos. En ocasiones el error de redondeo acumulado únicamente afecta a los dígitos de la extrema derecha del resultado final, sin embargo otras veces el error total puede ser tan grande como abrumador, arrojando un resultado sin un solo dígito de precisión.

Una consecuencia adicional del error de redondeo es que dos fórmulas algebraicamente equivalentes, talvez no sean numéricamente equivalentes. Considerando las dos fórmulas siguientes:

$$\sum_{n=1}^{10,000} n^{-2} \qquad \sum_{n=1}^{10,000} (10,001 - n)^{-2}$$

En la primera fórmula, la secuencia de números es ascendente e inversamente los sumandos son cada vez más pequeños, dando pie a que por consecuencia del error de redondeo se pierdan al ser sumados. Mientras que en la segunda fórmula, el orden es descendente y los sumandos junto con la suma crecen de manera directa y proporcional. De hecho, el error numérico de la primera fórmula es 650 veces mayor que el error en la segunda.

2.3.3 Error de truncamiento

El error de truncamiento es una limitación del software originada por la condicionante que la computadora cuenta únicamente con un lapso de tiempo finito para ejecutar cualquier algoritmo y obtener un resultado, por lo tanto algunos resultados son calculados en forma aproximada al restringir el algoritmo de cálculo a un número finito de iteraciones. A la discordancia surgida entre el resultado exacto y el cálculo aproximado como consecuencia de truncar el número de iteraciones a calcular se le denomina error de truncamiento.

2.3.4 Medición de la exactitud

Con respecto a los dos tipos de errores mencionados, la relación entre el valor verdadero y el valor aproximado está dada por la expresión:

$$valor \cdot verdadero = valor \cdot aproximado + error$$

En donde podemos observar que el error numérico es la diferencia entre los valores verdaderos y aproximados:

$$error = valor \cdot verdadero - valor \cdot aproximado$$

Que se representa como:

$$E_v = c - q$$

También llamado error absoluto, donde:

$\boldsymbol{E_v}$ se utiliza para denotar el error absoluto verdadero
$\boldsymbol{c}$ representa el valor verdadero
$\boldsymbol{q}$ representa el valor aproximado

Un problema que se tiene al utilizar esta definición es que la magnitud no nos permite cuantificar la importancia del error. Por ejemplo, no es lo mismo cometer un error de un centímetro en la estimación de la longitud de un puente que en la estimación de la longitud de un remache. Para resolver esa posible deficiente interpretación de error se recurre a la normalización del mismo usando el valor verdadero como referente:

$$error \cdot relativo \cdot fraccionario = \frac{error \cdot verdadero}{valor \cdot verdadero}$$

A menudo el signo del error no tiene la relevancia de la magnitud, y para fines de poder comparar los errores de un cálculo contra los de otro, se prefiere utilizar sus correspondientes valores absolutos, teniendo así que:

$$E_r = \frac{|q - c|}{|c|} \qquad c \neq 0$$

$\boldsymbol{E_r}$ se utiliza para denotar el error relativo fraccionario
$\boldsymbol{c}$ representa el valor verdadero
$\boldsymbol{q}$ representa el valor aproximado

El concepto de cifras significativas es en realidad solo una simple regla de representación de precisión relativa. Frecuentemente se usa el logaritmo del error relativo (LRE[2]) como una estimación más exacta del número de cifras significativas de una medida de precisión, aunque en realidad no es una definición de cifras significativas. Por ejemplo, un número con tres cifras significativas como 1.23, representa el intervalo de 1.225 a 1.235, con un error relativo de 0.005 / 1.23 = 0.004. Entonces, un número con cuatro cifras significativas tendría aproximadamente un décimo de ese error relativo. Por lo tanto, se puede definir el número de cifras significativas de manera más precisa como el negativo del logaritmo común de este error relativo, en este caso 2.39:

$$LRE = -\log\left[\frac{|q - c|}{|c|}\right]$$

$$LRE = -\log\left[\frac{0.005}{1.23}\right] = -\log(0.004) = 2.39$$

El LRE es indefinido cuando sucede que $q = c$ exactamente, sin embargo se debe considerar que el LRE debe ser igual al número de cifras significativas en c. En algunas ocasiones, el valor verdadero es cero como en varios casos de pruebas aplicadas para errores estándar de regresión lineal, en consecuencia el LRE es indefinido. En esas instancias se debe utilizar el logaritmo del error absoluto (LAR).

$$LAR = -\log[|q|]$$

[2] LRE corresponde a las siglas de Logarithm of Relative Error.

Por comodidad en la exposición no se hará distinción entre LRE y LAR, refiriéndose a ambos o a cada uno como LRE. El símbolo λ con un subíndice apropiado denotará el LRE de una cantidad computada.

a) El LRE tiene una interpretación específica:

   - LRE = 2.70.

   Significa que q concuerda con los primeros 9 ($\approx 2.7 / \log10\ 2$) bits de c.

b) El LRE es una medida del número correcto de cifras significativas sólo cuando q es "cercano" a c. Para ejemplificar esto, considere:

   - q = 165.89
   - c = 2.7070
   - LRE = 1.78.

   No es poco común que se obtengan valores de q "distantes" de c, en especial cuando se realizan pruebas para problemas no lineales. Por lo tanto, cada valor aproximado debe ser comparado con el valor verdadero para asegurarse que sólo difiera por un factor menor que dos; de otra manera el LRE debe ser igual a cero.

c) El LRE puede exceder el número de cifras significativas que tiene c; por ejemplo:

   - LRE = 11.4.

   Incluso a pesar que c tiene 11 dígitos. Esto es causado en parte, porque una computadora con doble precisión "rellena" los once dígitos con ceros. No resulta práctico tratar de corregir tales inconsistencias, por lo que el LRE debe ser igual al número de cifras significativas en c.

d) Cualquier LRE menor que uno debe ser igual a cero.

Capítulo III

3 METODOLOGÍAS PARA LA EVALUACIÓN DE LA EXACTITUD

3.1 Objetivo

Analizar la exactitud del software estadístico SAS v9.0, SPSS v15.0 y las herramientas estadísticas de EXCEL v2003 a través de la evaluación de las estimaciones obtenidas utilizando dos metodologías:

1) Los Conjuntos de Datos de Referencia Estadística (StRD) del Instituto Nacional de Estándares y Tecnología (NIST).

2) El software de Distribuciones Elementales (ELV) de la Universidad de Munich.

3.2 Software a comparar

Se eligió realizar el análisis comparativo para tres productos. Dos de ellos son SAS y SPSS, los paquetes de software estadístico más reconocidos, líderes en el mercado y más costosos. El restante es EXCEL una hoja de cálculo muy utilizada para propósitos diversos, sin embargo su alta disponibilidad en las computadoras personales la convierten en un producto de fácil acceso para el análisis de datos. Desafortunadamente las versiones más recientes de estos productos no están al alcance de este trabajo.

Las pruebas se llevarán a cabo en una laptop comercial de caracteristicas ordinarias: sistema operativo Microsoft Windows XP SP2, procesador centrino y memoria RAM de 512MB de capacidad.

3.2.1 SAS

The SAS Base Programming System for Windows; Release 9.0 TS Level 01M0. Esta versión de software permite el manejo de cifras con 8 posiciones decimales.

3.2.2 SPSS

SPSS 13.0 for Windows; Release 13.0 (1 Sep 2004). Esta versión de software permite el manejo de cifras con 16 posiciones decimales.

3.2.3 EXCEL

Microsoft Office Excel 2003 (11.8134.8132) SP2; Parte de Microsoft Office Professional Edition 2003. Esta versión de software permite el manejo de cifras con 16 posiciones decimales.

3.3 Conjuntos de datos de referencia estadística del NIST[3]

El Instituto Nacional de Estándares y Tecnología (NIST) es un organismo federal no regulador que forma parte de la Administración de Tecnología del Departamento de Comercio[4] de los Estados Unidos de América. Fue fundado en 1901 con el nombre de Buró Nacional de Estándares (NBS[5]), y para 1988 cambió a NIST.

La misión del NIST consiste en promover la innovación y la competitividad industrial con el propósito de aumentar la estabilidad económica y mejorar la calidad de vida desarrollando soluciones para los problemas críticos en el área de las ciencias de la medición; el desarrollo y uso de estándares y de tecnología.

El proyecto de los conjuntos de datos de referencia estadística surge originalmente como una respuesta a la preocupación industrial sobre la exactitud numérica de los cálculos derivados del software estadístico, es aquí donde el NIST se dio a la tarea de proveer una referencia con conjuntos de datos y resultados computacionales certificados que permiten la evaluación objetiva del software estadístico para una variedad de métodos estadísticos, con el propósito de mejorar la exactitud numérica de los cálculos.

Los conjuntos de datos y los valores certificados son empleados para evaluar la exactitud del software en la estimación de los estadísticos univariados, la regresión lineal, la regresión no lineal y el análisis de varianza. La colección incluye conjuntos de datos generados que están diseñados para cuestiones computacionales específicas y conjuntos de datos de observaciones de fenómenos reales para diferentes niveles de dificultad. Además incluyen los conjuntos de datos tradicionales de Wampler[6] para probar algoritmos de regresión lineal y los conjuntos de datos de Simon y Lesage[7] para probar algoritmos de análisis de varianza. Los datos de observaciones de fenómenos reales contendrían conjuntos de datos tan demandantes como los conjuntos de datos Longley[8] para regresión lineal y más conjuntos de datos iniciales tales como los datos de Daniel y Wood[9] para regresión no lineal. Los valores certificados son las mejores soluciones disponibles y el procedimiento de certificación es descrito en páginas web para cada método estadístico.

Los conjuntos de datos están ordenados por los niveles de dificultad: bajo, medio y alto. De manera rigurosa, el nivel de dificultad de un conjunto de datos depende de la complejidad del algoritmo. Estos niveles son meramente provistos como un consejo preliminar para el usuario, por lo que al obtener resultados satisfactorios en todos los conjuntos de datos de mayor dificultad no implicaría que el software pasara todos los conjuntos de datos de nivel medio o incluso de nivel bajo de dificultad. Similarmente, obtener resultados para todos los conjuntos de datos en esta colección no implica que el

[3] NIST corresponde a las siglas de National Institute of Standards and Technology.
[4] Department of Commerce.
[5] NBS corresponde a las siglas de National Bureau of Standards.
[6] Referencia bibliográfica No. 25.
[7] Referencia bibliográfica No. 26.
[8] Referencia bibliográfica No. 27
[9] Referencia bibliográfica No. 28

software lo haga igual para cualquier otro conjunto de datos particular. No obstante, se tendrá algún grado de certeza, en el sentido que el software obtendrá resultados satisfactorios para conjuntos de datos conocidos.

Además se actualizará la colección con conjuntos de datos para métodos estadísticos adicionales, utilizando para ello la retroalimentación con el usuario final.

El procedimiento de certificación implica que los cálculos se realizaron con una precisión de 500 dígitos utilizando preprocesamiento y el paquete de subrutinas en Fortran de Bailey[10] (1995). En otras palabras, en la lectura, en los cálculos y en los resultados, los datos tuvieron una precisión de 500 dígitos, lo que representa el logro que se alcanzaría si los cálculos se hicieran sin redondeo u otros errores. Los resultados se redondearon a quince cifras significativas. Cualquier algoritmo numérico típico introducirá imprecisiones computacionales y producirá resultados que difieren ligeramente de estos valores certificados.

3.3.1 Cálculo del logaritmo del error relativo

Como se comentó en el Capítulo II, frecuentemente se usa el logaritmo del error relativo (LRE[11]) como una estimación más exacta del número de cifras significativas de una medida de exactitud, aunque estrictamente hablando no es una definición de cifras significativas:

$$LRE = -\log\left[\frac{|q-c|}{|c|}\right]$$

c representa el valor certificado por la NIST
q representa el valor estimado por el software estadístico

El LRE es indefinido cuando sucede que $q = c$ exactamente, sin embargo se debe considerar que el LRE debe ser igual al número de cifras significativas en c. En algunas ocasiones, el valor verdadero es cero como en varios casos de pruebas aplicadas para errores estándar de regresión lineal, en consecuencia el LRE es indefinido. En esas instancias se debe utilizar el logaritmo del error absoluto (LAR).

$$LAR = -\log[|q|]$$

Por comodidad en la exposición no se hará distinción entre LRE y LAR, refiriéndose a ambos o a cada uno como LRE. El símbolo λ con un subíndice apropiado denotará el LRE de una cantidad estimada.

Las siguientes son las consideraciones necesarias para la interpretación del LRE:

[10] Disponible en NETLIB: http://www.netlib.org.
[11] LRE corresponde a las siglas de Logarithm of Relative Error.

a) El LRE tiene una interpretación específica, por ejemplo si el $LRE = 2.70$, significa que *q* concuerda con los primeros nueve bits de *c* ($9 \approx 2.7 / \log 10\ 2$), de manera más práctica significa que *q* es exacto a dos cifras significativas con respecto a *c*.

b) El LRE es una estimación adecuada del número de cifras significativas sólo cuando *q* es "cercano" a *c*. Para ejemplificar esto, considere $q = 165.89,\ c = 2.7070,\ LRE = 1.78$. Es claro que el LRE no es una estimación adecuada del número de cifras significativas, sin embargo no es poco común que se obtengan valores de *q* "distantes" de *c*, en especial cuando se realizan pruebas para problemas no lineales. Por lo tanto, cada valor estimado debe ser comparado con el valor certificado para asegurarse que no difiera por un múltiplo mayor o igual que dos; de otra manera el LRE debe ser igual a cero.

c) El LRE puede exceder el número de cifras significativas que tiene *c*, por ejemplo si el $LRE = 11.4$ cuando *c* tiene once dígitos. Esto es causado en parte, porque una computadora con doble precisión "rellena" los once dígitos con ceros. No resulta práctico tratar de corregir tales inconsistencias, por lo que el LRE debe ser igual al número de cifras significativas en *c*.

d) Cualquier LRE menor que uno debe ser igual a cero.

Un ejemplo práctico del cálculo del LRE es el siguiente:

Considere el conjunto de datos Michelso, que consiste de 100 observaciones de un experimento sobre la velocidad de la luz en el aire; con un nivel de dificultad medio. El valor certificado del NIST y la estimación del software estadístico W para la desviación estándar son:

- NIST : 0.0790105478190518
- Software W : 0.078614502891384

$$\lambda_\sigma = -\log\left[\frac{|q-c|}{|c|}\right] = -\log\left[\frac{0.0003\ldots}{0.0790\ldots}\right] = -\log(0.0050\ldots) \approx 2.30$$

Tomando en cuenta que *c* tiene dieciséis cifras significativas y *q* sólo tiene dos, esto sugiere que el software estadístico W utiliza uno de los algoritmos menos eficientes para la estimación de la desviación estándar.

3.3.2 Conjuntos para la media y la desviación estándar

Nueve son los conjuntos de datos que provee el NIST para evaluar la media y la desviación estándar y están agrupados en los conjuntos para estadísticos univariados, cuentan con tres niveles de dificultad y el número de observaciones va desde 3 hasta 5,000, como se detalla en la tabla 3.1

Para cada conjunto de datos se proporcionan los valores certificados a 15 cifras significativas para la media (μ) y la desviación estándar (σ); el LRE para cada uno puede ser presentado como: λ_μ y λ_σ .

Conjunto De Datos	Nivel de Dificultad	Descripción	Fuente
PiDigits	Bajo	Los primeros dígitos de la constante matemática π . Mathematics of Computation. January 1962, page 76. $n = 5{,}000$	Observados
Lottery	Bajo	Números de lotería desde sep-03-'89 a abr-14-'90. Un número diario, con excepción de algunos días. $n = 218$	Observados
Lew	Bajo	Las deformaciones sufridas por vigas de acero y concreto al ser sometidas a una presión periódica. $n = 200$	Observados
Mavro	Bajo	Medidas de transmisión desde un filtro con valor nominal de 2. $n = 50$	Observados
Michelso	Bajo	Medidas de la velocidad de la luz en el aire. Dorsey, Ernest N. (Referencia bibliográfica No. 29). $n = 100$	Observados
NumAcc1	Bajo	Números enteros de ocho dígitos que difieren en la última posición decimal. Simon, Stephen D. and Lesage, James P. (Referencia bibliográfica No. 26). $n = 3$	Generados
NumAcc2	Medio	Números decimales con un dígito entero y un dígito decimal que difieren en la posición decimal. Simon, Stephen D. and Lesage, James P. (Referencia bibliográfica No. 26). $n = 1001$	Generados
NumAcc3	Medio	Números decimales con siete dígitos enteros y un dígito decimal que difieren en la posición decimal. Simon, Stephen D. and Lesage, James P. (Referencia bibliográfica No. 26). $n = 1001$	Generados
NumAcc4	Alto	Números decimales con ocho dígitos enteros y un dígito decimal que difieren en la posición decimal. Simon, Stephen D. and Lesage, James P. (Referencia bibliográfica No. 26). $n = 1001$	Generados

Tabla 3.1

3.3.3 Conjuntos para el análisis de varianza de un factor

Once son los conjuntos de datos que provee el NIST para evaluar el análisis de varianza de un factor (ANOVA) y están agrupados en los conjuntos para análisis de varianza, cuentan con tres niveles de dificultad y el número observaciones va desde 25 hasta 18,009, como se detalla en la tabla 3.2

Para cada conjunto de datos se proporcionan los valores certificados a 15 cifras significativas para la suma de cuadrados, los cuadrados de la media y el estadístico F.

Puesto que la mayoría de los valores cerificados se usan en el cálculo del estadístico F, únicamente se presenta el LRE del estadístico F representado como λ_F .

Conjunto de Datos	Nivel de Dificultad	Descripción	Fuente
SiRstv	Bajo	Ehrstein, James and Croarkin, M. Carroll. Conjunto de datos sin publicar del NIST. $n = 25$	Observados
SmLs01	Bajo	Simon, Stephen D. and Lesage, James P. (Referencia bibliográfica No. 26). $n = 189$	Generados
SmLs02	Bajo	Simon, Stephen D. and Lesage, James P. (Referencia bibliográfica No. 26). $n = 1{,}809$	Generados
SmLs03	Bajo	Simon, Stephen D. and Lesage, James P. (Referencia bibliográfica No. 26). $n = 18{,}009$	Generados
AtmWtAg	Medio	Powell, L.J., Murphy, T.J. and Gramlich, J.W. (Referencia bibliográfica No. 30). $n = 48$	Observados
SmLs04	Medio	Simon, Stephen D. and Lesage, James P. (Referencia bibliográfica No. 26). $n = 189$	Generados
SmLs05	Medio	Simon, Stephen D. and Lesage, James P. (Referencia bibliográfica No. 26). $n = 1{,}809$	Generados
SmLs06	Medio	Simon, Stephen D. and Lesage, James P. (Referencia bibliográfica No. 26). $n = 18{,}009$	Generados
SmLs07	Alto	Simon, Stephen D. and Lesage, James P. (Referencia bibliográfica No. 26). $n = 189$	Generados
SmLs08	Alto	Simon, Stephen D. and Lesage, James P. (Referencia bibliográfica No. 26). $n = 1{,}809$	Generados
SmLs09	Alto	Simon, Stephen D. and Lesage, James P. (Referencia bibliográfica No. 26). $n = 18{,}009$	Generados

Tabla 3.2

3.3.4 Conjuntos para la regresión lineal simple

Tres son los conjuntos de datos que provee el NIST para evaluar la regresión lineal simple y están agrupados en los conjuntos para regresión lineal, cuentan con dos niveles de dificultad; el número de observaciones va desde 3 hasta 36; y los coeficientes beta a estimar pueden ser la pendiente de la recta y la ordenada al origen ó únicamente la pendiente de la recta, como se detalla en la tabla 3.3.

Para cada conjunto de datos se proporcionan los valores certificados a 15 cifras significativas para los coeficientes estimados, los errores estándar de los coeficientes, el residuo de la desviación estándar del modelo, R^2 y el análisis de varianza común para la tabla de regresión lineal, la cual incluye el residuo de la suma de cuadrados. Para evaluar la

regresión lineal simple se presenta el LRE de los coeficientes estimados, de los errores estándar de los coeficientes y del residuo de la desviación estándar del modelo representados como λ_β, λ_σ y λ_r.

Conjunto de Datos	Nivel de Dificultad	Descripción	Fuente
Norris	Bajo	Norris, J., NIST; Calibración de los monitores de Ozono. $n = 36$	Observados
NoInt1	Medio	Eberhardt, K., NIST. $n = 11$	Generados
NoInt2	Medio	Eberhardt, K., NIST. $n = 3$	Generados

Tabla 3.3

La tabla 3.4 para el conjunto de datos de Norris muestra el LRE para los coeficientes y errores estándar producidos por el software estadístico W. Este exceso de información puede ser resumida convenientemente recurriendo al principio "el eslabón más débil de la cadena", usando el mínimo LRE.

Variable	λ_β	λ_σ
β_0	4.87	6.44
β_1	5.14	6.23

Tabla 3.4

Entonces, para software estadístico W y este conjunto de datos $\lambda_\beta = 4.9$ y $\lambda_\sigma = 6.2$.

3.4 Software de distribuciones elementales

La evaluación de las distribuciones estadísticas se realiza a través del programa Elementare Verteilungen (ELV[12]) que fue desarrollado en 1989 por el profesor Leo Knüsel, quién pertenece al Departamento de Estadística de la Universidad de Munich. El software provee el valor crítico dado un percentil de las distribuciones normal estándar, gamma, ji cuadrada, beta, F, t de Student, Poisson, Binomial e hipergeométrica; además de manejar la función inversa que provee el percentil dado un valor crítico de cada una de ellas; e incluye una versión no central que provee el valor crítico dado un percentil para las distribuciones gamma, ji cuadrada, beta, F y t de Student.

Las probabilidades de todas las distribuciones son procesadas con 6 cifras significativas para probabilidades tan pequeñas como 100^{-100}; las probabilidades menores que 100^{-100} se

[12] ELV corresponde a las iniciales en Alemán de ELementare Verteilungen. http://www.stat.uni-muenchen.de/~knuesel.

redondean a cero. Los percentiles mayores y menores se procesan para las nueve distribuciones disponibles para las probabilidades de $100^{-12} \leq p \leq 1/2$. Las distribuciones que requieren el parámetro de grados de libertad admiten valores de.$1 \leq n \leq 2^{27}$.

3.4.1 Cálculo del error relativo

Como se comentó en el Capítulo II, para cuantificar la importancia del error se utiliza la normalización del mismo usando el valor verdadero como referente:

$$error \cdot relativo \cdot fraccionario = \frac{error \cdot verdadero}{valor \cdot verdadero}$$

A menudo el signo del error no tiene la relevancia de la magnitud, y para fines de poder comparar los errores de un cálculo contra los de otro, se prefiere utilizar sus correspondientes valores absolutos, teniendo así que:

$$E_r = \frac{|q - c|}{|c|} \qquad c \neq 0$$

E_r se utiliza para denotar el error relativo fraccionario
c representa el valor certificado por el ELV
q representa el valor estimado por el software estadístico

Un ejemplo práctico del cálculo del E_r es el siguiente:

Considere evaluar la distribución F para la variable aleatoria x = 0.7 con n_1= 201 y n_2= 10,001 grados de libertad. El valor certificado del ELV y la estimación del software estadístico W son:

- ELV : 0.00045779
- Software W : 0.00045785

$$E_r = \frac{|q - c|}{|c|} = \frac{6E - 8}{0.0004...} = 1E - 4$$

Tomando en cuenta que E_r = 1E-4, entonces las primeras tres cifras significativas del percentil estimado concuerda con el valor exacto.

3.4.2 Valores para las distribuciones estadísticas

Los resultados del ELV son considerados como los valores certificados y los resultados del software estadístico a evaluar son considerados como los valores estimados.

Para la distribución normal se prueba con la siguiente secuencia básica de percentiles (BSP[13]): {z: 0.0001, 0.001, 0.01, 0.1, 0.9, 0.99, 0.999, 0.9999} para evaluar las probabilidades estimadas $P\{Z < z\}$.

Para la distribución ji cuadrada se prueban los siguientes valores para la variable aleatoria *x*: {2E4, 2E5, 2E6, 4E6, 6E6, 8E6, 1E7} con {n: 2E4, 2E5, 2E6, 4E6, 6E6, 8E6, 1E7} grados de libertad para evaluar las probabilidades estimadas $P\{X < x\}$.

Para la distribución F se prueban los siguientes valores para la variable aleatoria *x*: {1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.38, 0.37, 0.3} con n_1= 201 y n_2= 10,001 grados de libertad para evaluar las probabilidades estimadas $P\{X < x\}$.

Para la distribución t de Student se prueban los siguientes valores para la variable aleatoria *x*: {-9.7, -9.8, -9.9, -10.0, -10.1, -10.2} con n=100 grados de libertad para evaluar las probabilidades estimadas $P\{X < x\}$.

El error relativo fraccionario E_r es utilizado para medir la exactitud de las probabilidades. Esencialmente, si E_r < 1E-4, entonces las primeras cuatro cifras significativas de la probabilidad estimada concuerda con el valor exacto.

[13] BSP corresponde a las siglas de Basic Sequence of Percentiles.

Capítulo IV

4 RESULTADOS

4.1 Resultados con los conjuntos de datos de referencia estadística

A continuación se presentan los resultados obtenidos a partir de los conjuntos de datos de referencia estadística del NIST para la evaluación de la exactitud de la estimación de la media, la desviación estándar, el análisis de varianza de un factor (ANOVA) y la regresión lineal simple por medio del cálculo del logaritmo del error relativo.

4.1.1 Resultado para la media

Conjunto De Datos	Nivel de Dificultad	Cifras Significativas	SAS λ_μ	SPSS λ_μ	EXCEL λ_μ
PiDigits	Bajo	15	4.2	15	15
Lottery	Bajo	15	2.5	15	15
Lew	Bajo	15	3.0	15	15
Mavro	Bajo	15	6.3	15	15
Michelso	Bajo	15	7.1	15	15
NumAcc1	Bajo	15	7.3	15	15
NumAcc2	Medio	15	15	15	14.0
NumAcc3	Medio	15	15	15	15
NumAcc4	Alto	15	15	15	14.0

Tabla 4.1

La evaluación de la exactitud de la estimación de la media a través de los conjuntos de datos univariados (véase tabla 4.1) arroja que la estimación obtenida por:

SPSS es la más exacta con 15 cifras significativas para todas las pruebas.

EXCEL obtuvo una estimación muy cercana con 15 cifras significativas para siete pruebas y 14 cifras significativas para las dos restantes, una de nivel medio y una de nivel bajo.

SAS mostró sólo para tres pruebas una estimación de 15 cifras significativas y fueron las dos pruebas de nivel medio y la prueba de nivel alto, para las seis pruebas de nivel bajo las cifras significativas de las estimaciones fueron pobres y oscilaron entre 2 y 7.

4.1.2 Resultado para la desviación estándar

La evaluación de la exactitud de la estimación de la desviación estándar a través de los conjuntos de datos univariados (véase tabla 4.2) es la siguiente:

SPSS es la más exacta con 15 cifras significativas para cuatro pruebas, tres de nivel bajo y una de nivel medio, para el resto de las pruebas las cifras significativas fueron buenas y oscilaron entre 9 y 13.

EXCEL obtuvo una estimación con altibajos con 15 cifras significativas para cuatro pruebas de nivel bajo, en las dos últimas pruebas, de nivel medio y alto obtuvo malos resultados con 0 y 1 cifras significativas respectivamente, para el resto de las pruebas obtuvo 8. 9 y 11 cifras significativas.

SAS logró resultados muy deficientes que incluyen una prueba de nivel bajo sin cifras significativas, el resto de las pruebas mostró desde 2 hasta 4 cifras significativas.

Conjunto de Datos	Nivel de Dificultad	Cifras Significativas	SAS λ_σ	SPSS λ_σ	EXCEL λ_σ
PiDigits	Bajo	15	4.2	13	15
Lottery	Bajo	15	2.9	15	15
Lew	Bajo	15	2.6	15	15
Mavro	Bajo	15	2.0	12.1	9.4
Michelso	Bajo	15	2.3	12.4	8.3
NumAcc1	Bajo	15	0	15	15
NumAcc2	Medio	15	3.3	15	11.6
NumAcc3	Medio	15	3.3	9.5	1.1
NumAcc4	Alto	15	3.3	9.5	0

Tabla 4.2

4.1.3 Resultado para el análisis de varianza de un factor

Conjunto de Datos	Nivel de Dificultad	Cifras Significativas	SAS λ_F	SPSS λ_F	EXCEL λ_F
SiRstv	Bajo	15	0	12.6	8.5
SmLs01	Bajo	15	2.2	0	14.3
SmLs02	Bajo	15	3.3	1.5	12.5
SmLs03	Bajo	15	4.3	1.5	12.6
AtmWtAg	Medio	15	1.3	8.5	1.8
SmLs04	Medio	15	2.2	0	1.7
SmLs05	Medio	15	3.3	1.5	1.1
SmLs06	Medio	15	4.3	1.5	NA
SmLs07	Alto	15	2.4	0	NA
SmLs08	Alto	15	2.9	1.5	NA
SmLs09	Alto	15	3.0	1.5	NA

Tabla 4.3

La evaluación de la exactitud de la estimación del análisis de varianza de un factor (ANOVA) a través de los conjuntos de datos para análisis de varianza (véase tabla 4.3) se presenta a continuación:

SAS es la más consistente, ya que únicamente para una prueba de nivel bajo no obtuvo cifras significativas, logró alcanzar la mayor exactitud de los tres productos para las pruebas de nivel alto con 2 y 3 cifras significativas y para el resto de las pruebas estuvo entre 1 y 4 cifras significativas.

SPSS mostró resultados pobres en general con tres pruebas sin cifras significativas, una de nivel bajo, una de nivel medio y una de nivel alto, además de seis pruebas de 1 cifras significativas, dos de nivel bajo, dos de nivel medio y dos de nivel alto y en contraste

obtuvo 8 y 12 cifras significativas para una prueba de nivel medio y para una de nivel bajo respectivamente.

EXCEL obtuvo una estimación variante con cuatro pruebas no logró obtener estimación alguna, una de nivel medio y las tres de nivel alto, para las restantes tres de nivel medio mostró 1 cifras significativa, sin embargo para las pruebas de nivel bajo presentó 8, 12 y 14 cifras significativas.

4.1.4 Resultado para la regresión lineal simple

Conjunto de Datos	Nivel de Dificultad	Cifras Significativas	SAS λ_β	SPSS λ_β	EXCEL λ_β
Norris	Bajo	15	1.3	12.3	1.3
NoInt1	Medio	15	2.4	14.7	2.4
NoInt2	Medio	15	2.1	15	2.1
			SAS λ_σ	SPSS λ_σ	EXCEL λ_σ
Norris	Bajo	15	1.3	10.2	1.3
NoInt1	Medio	15	1.2	12.5	1.2
NoInt2	Medio	15	0	14.3	0
			SAS λ_r	SPSS λ_r	EXCEL λ_r
Norris	Bajo	15	1.8	10.2	1.8
NoInt1	Medio	15	1.0	13.1	1.0
NoInt2	Medio	15	0	14.5	0

Tabla 4.4

La evaluación de la exactitud de las estimaciones de la regresión lineal simple a través de los conjuntos de datos para regresión lineal (véase tabla 4.4) se muestra en seguida:

SPSS presenta las estimaciones más exactas con cifras significativas que van de las 10 a las 15 para las tres pruebas en las tres estimaciones β, σ y γ .

SAS y EXCEL muestran muy pobres e idénticas cifras significativas, con dos estimaciones de una prueba de nivel medio sin cifras significativas, cinco estimaciones de dos pruebas de nivel bajo y medio con 1 cifras significativas y dos estimaciones de dos pruebas de nivel medio con 2 cifras significativas.

4.2 Resultados con el software de distribuciones elementales

En seguida se presentan los resultados obtenidos a partir del software de distribuciones elementales (ELV) para la evaluación de la exactitud de la estimación de las probabilidades para las distribuciones estadísticas: normal, ji cuadrada, t de Student y F por medio del cálculo del error relativo.

4.2.1 Resultado para la distribución normal

La evaluación de la exactitud de la estimación de las probabilidades para la distribución normal (véase tabla 4.5) expone que la estimación obtenida por:

SPSS y EXCEL son las más exactas, obteniendo 6 cifras significativas en todas las probabilidades estimadas.

SAS mostró una muy buena estimación, en dos probabilidades obtuvo 6 cifras significativas y en las restantes seis logró 5 cifras significativas.

Z	Cifras Significativas	SAS E_r	SPSS E_r	EXCEL E_r
0.0001	6	0	0	0
0.001	6	2E-6	0	0
0.01	6	2E-6	0	0
0.1	6	4E-6	0	0
0.9	6	0	0	0
0.99	6	4E-6	0	0
0.999	6	4E-6	0	0
0.9999	6	1E-6	0	0

Tabla 4.5

4.2.2 Resultado para la distribución ji cuadrada

X	*n*	Cifras Significativas	SAS E_r	SPSS E_r	EXCEL E_r
2E+4	2E+4	6	0	6E-3	NA
2E+5	2E+5	6	2E-6	0	NA
2E+6	2E+6	6	6E-6	0	NA
4E+6	4E+6	6	8E-6	0	NA
6E+6	6E+6	6	6E-6	0	NA
8E+6	8E+6	6	8E-6	0	NA
1E+7	1E+7	6	2E-6	0	NA

Tabla 4.6

La evaluación de la exactitud de la estimación de las probabilidades para la distribución ji cuadrada (véase tabla 4.6) es la siguiente:

SAS es la más consistente, con una estimación de 6 cifras significativas y con las restantes seis de 5 cifras significativas.

SPSS mostró resultados muy buenos en general con seis estimaciones de 6 cifras significativas, sin embargo en una estimación obtuvo 2 cifras significativas.

EXCEL no produjo resultados en la estimación de probabilidades.

4.2.3 Resultado para la distribución t de Student

La evaluación de la exactitud de la estimación de las probabilidades para la distribución t de Student (véase tabla 4.7) se muestra a continuación:

SAS es la más exacta y consistente, ya que en tres estimaciones obtuvo 5 cifras significativas y en las tres restantes logró 4 cifras significativas.

SPSS dejó mucho que desear, ya que sólo produjo resultados sin cifras significativas.

EXCEL no produjo resultados en la estimación de probabilidades.

x	*n*	Cifras Significativas	SAS E_r	SPSS E_r	EXCEL E_r
-9.7	100	6	2E-5	1	NA
-9.8	100	6	3E-5	1	NA
-9.9	100	6	5E-6	1	NA
-10.0	100	6	8E-6	1	NA
-10.1	100	6	3E-6	1	NA
-10.2	100	6	1E-5	1	NA

Tabla 4.7

4.2.4 Resultado para la distribución F

x	n_1	n_2	Cifras Significativas	SAS E_r	SPSS E_r	EXCEL E_r
1.0	201	10,001	6	2E-6	2E-6	5E-2
0.9	201	10,001	6	3E-5	1E-6	1
0.8	201	10,001	6	6E-6	2E-6	1
0.7	201	10,001	6	0	0	1
0.6	201	10,001	6	5E-5	2E-6	1
0.5	201	10,001	6	3E-6	7E-4	1
0.4	201	10,001	6	7E-6	1	1
0.38	201	10,001	6	8E-6	1	1
0.37	201	10,001	6	2E-5	1	1
0.3	201	10,001	6	1E-6	1	1

Tabla 4.8

La evaluación de la exactitud de la estimación de las probabilidades para la distribución F (véase tabla 4.8) se aborda en seguida:

SAS es la más exacta, con una estimación de 6 cifras significativas, con seis más de 5 cifras significativas y con las tres remanentes de 4 cifras significativas.

SPSS mostró cuatro estimaciones sin cifras significativas, también logró una estimación con 6 cifras significativas, cuatro con 5 cifras significativas y una con 3 cifras significativas.

EXCEL obtuvo resultados muy pobres, únicamente en una estimación obtuvo 1 cifras significativas y para las restantes nueve no logró cifras significativas.

Conclusiones

SPSS mostró mayor exactitud en las pruebas en general, sin embargo hubo excepciones en algunas pruebas que es importante mencionar como: el análisis de varianza de un factor (ANOVA), donde en nueve de las once pruebas tuvo muy malas estimaciones del estadístico F; la distribución t de Student, donde sólo obtuvo estimaciones de probabilidades sin cifras significativas y la distribución F, que en cuatro probabilidades estimadas no obtuvo cifras significativas.

SAS logró mayor consistencia, ya que obtuvo resultados aproximados en todas las pruebas. Si bien en la mitad de los resultados presentó una pobre exactitud, es en parte atribuible a la limitante del número de posiciones decimales que maneja la versión utilizada. Cabe mencionar que en las pruebas de las distribuciones estadísticas obtuvo estimaciones muy buenas en cada una de ellas.

EXCEL presentó varios matices, en las pruebas para la distribución normal y para la media obtuvo excelente y muy buena exactitud respectivamente; en las pruebas para la desviación estándar y para el análisis de varianza de un factor (ANOVA) mostró buena exactitud para las pruebas de nivel bajo y mala exactitud para las pruebas de nivel medio y alto; en las pruebas para la regresión lineal simple y para la distribución F tuvo muy pobre exactitud y finalmente para las pruebas de las distribuciones ji cuadrada y t de Student no obtuvo estimación de probabilidades para ninguna de las dos.

La evaluación de la exactitud es un proceso laborioso que consume tiempo, esfuerzo y recursos, por eso generalmente no se cuestiona la exactitud del software estadístico, ni de las herramientas estadísticas disponibles en el mercado. Se da como un hecho la exactitud de estos productos, más cuando los respaldan marcas conocidas que gozan de prestigio. En nuestro caso se puede ver claramente que, contrario a lo que pudiera pensarse, una versión para estudiantes de SAS, probablemente más económica, sacrifica la eficiencia en los cálculos repercutiendo directamente en la exactitud de las estimaciones.

Por otro lado, con respecto a las metodologías se utilizan dos maneras diferentes para evaluar la exactitud de las estimaciones obtenidas:

1) El logaritmo del error relativo para las pruebas del NIST.
2) El error relativo para las pruebas con el software ELV.

Aunque la evaluación de la exactitud de las estimaciones por medio del error relativo resulta más confusa por varias razones por ejemplo:

- Que para la lectura de las cifras significativas es necesario restar una unidad al exponente de la notación científica o contar los ceros en caso de notación numérica posicional.
- Que cuando el error relativo es cero no significa que se tienen cero cifras significativas, sino que el error relativo es nulo, lo que se interpreta como una estimación exacta.

- Que cuando el error relativo es uno no significa que se tiene una cifra significativa, sino que el valor estimado es nulo.

Por estás razones resulta más práctico interpretar el número de cifras significativas por medio del logaritmo del error relativo, ya que la lectura de las cifras significativas es casi directa. Con fines de comparación se calculó de manera adicional el logaritmo del error relativo para las tablas de resultados de las distribuciones estadísticas y se muestran en el Anexo B.

Finalmente, lo más importante desde la perspectiva del usuario no es saber cuál producto estadístico es el mejor, sino que necesidades se tienen y con que recursos se cuenta. Cada usuario debe conocer la complejidad del análisis de datos que pretende y la exactitud que lo satisface, sólo entonces se debe abocar a la tarea de encontrar el mejor producto para él, aunque no sea ni lo más exacto, ni lo más avanzado, ni lo más novedoso, ni lo más rápido, ni lo más amigable, etc., bastará con que sea lo más adecuado.

Bibliografía

1) McCULLOUGH, B.D. “Statistical Computing Software Reviews; Assessing The Reliability of Statistical Software: Part I”. The American Statistician 52, no 4 (noviembre 1998); 358 – 365.
2) McCULLOUGH, B.D. “Statistical Computing Software Reviews; Assessing the Reliability of Statistical Software: Part II”. The American Statistician 53, no 2 (mayo 1999); 149 – 158.
3) QUINTANA, Pedro; VILLALOBOS, Eloísa; CORNEJO, Ma. Del Carmen. “Métodos Numéricos con aplicaciones en Excel”. Barcelona: Editorial Reverté, S. A. 2005.
4) CHAPRA, Steven C; CANALE, Raymond. “Métodos numéricos para ingenieros, Cuarta Edición”. México, D. F: McGraw-Hill Interamericana Editores, S.A. de C.V. 2003.
5) HUERTA, Antonio; SARRATE, Josep; RODRÍGUEZ, Antonio; FERRAN. “Métodos numéricos, Introducción, aplicaciones y programación”. Barcelona: Editions UPC. 2001.
6) CHAMBERS, J.M. Users, programmers, and statistical software. Journal of Computational and Graphical Statistics 9, no 3 (septiembre 2000); 404-422.
7) WIKIPEDIA. Wikimedia Foundation, Inc. http://es.wikipedia.org/wiki/IEEE (consultado en noviembre 21, 2007).
8) WIKIPEDIA. Wikimedia Foundation, Inc. http://es.wikipedia.org/wiki/IEEE_punto_flotante (consultado en noviembre 21, 2007).
9) ADDLINK. Addlink Software Científico, S.L. http://www.addlink.es/productos.asp?pid=44 (consultado en noviembre 22, 2007).
10) ADDLINK. Addlink Software Científico, S.L. http://www.addlink.es/productos.asp?pid=42 (consultado en noviembre 22, 2007).
11) R-PROJECT. Department of Statistics and Mathematics of the WU Wien. http://www.r-project.org (consultado en noviembre 22, 2007).
12) SOFTWARE shop. Software shop. http://www.software-shop.com/in.php?mod=fabricantes&manID=5 (consultado en noviembre 22, 2007).
13) SPSS. SPSS INC. http://www.spss.com/la/productos/base/base.htm (consultado en noviembre 22, 2007).
14) SAS. SAS Institute Inc. http://www.sas.com/technologies/analytics/statistics/stat/index.html (consultado en noviembre 22, 2007).
15) EXCEL HOMEPAGE. Microsoft Corporation. http://office.microsoft.com/es-es/excel/FX100487623082.aspx (consultado en noviembre 22, 2007).
16) EUMEDNET. Fundación Universitaria Andaluza Inca Garcilaso. http://www.eumed.net/cursecon/libreria/drm/0.htm (consultado en noviembre 22, 2007).
17) “MONOGRAFIAS.COM”. Lucas Morea / Monografias.com S.A. http://www.monografias.com/trabajos10/esta/esta.shtml (consultado en noviembre 22, 2007).
18) GESTIOPOLIS. Carlos López / Webprofit Ltda. http://www.gestiopolis.com/recursos/experto/catsexp/pagans/eco/21/estadistica.htm (consultado en noviembre 22, 2007).
19) THE MATH FORUM. Drexel University. http://mathforum.org/dr.math (consultado en diciembre 03, 2007).
20) StRD. Instituto Nacional de Estándares y Tecnología (NIST). http://www.itl.nist.gov/div898/strd/index.html (consultado en octubre 31, 2007).
21) ELV. Departamento de Estadística de la Universidad de Munich. http://www.stat.uni-muenchen.de/~knuesel (consultado en octubre 31, 2007).
22) WALPOLE, Ronald E; MYERS, Raymond H; MYERS, Sharon L. “Probabilidad y estadística para ingenieros, Sexta Edición”. México, Prentice-Hall Hispamoamericana, S.A. 1999.
23) DEVORE, Jay L. “Probabilidad y estadística para ingeniería y ciencias, Quinta Edición”. Internacional Thomson Editores, S.A. de C.V. 2004.
24) HILNES, William W; MONTGOMERY, Douglas C; GOLDSMAN, David M; BARROS, Connie M. “Probabilidad y estadística para ingeniería, Cuarta Edición en Inglés (Tercera Edición en Español)”. México, D.F. Compañía Editorial Continental, S.A. de C.V. 2005.
25) WAMPLER, R. H. “A Report of the Accuracy of Some Widely-Used Least Squares Computer Programs”. Journal of the American Statistical Association, 65, (1970); 549 – 565.
26) SIMON, Stephen D; LESAGE, James P. “Assessing the Accuracy of ANOVA Caluclations in Statistical Software”. Computational Statistics & Data Analysis 8, (1989); 325 – 332.
27) LONGLEY, J. W. “An Appraisal of Least Squares Programs for the Electronic Computer from the Viewpoint of the User”. Journal of the American Statistical Association 62, (1967); 819 – 841.
28) DANIEL, C; WOOD, F. S. “Fitting Equations to Data, Second Edition”. New York, N. Y. John Wiley and Sons. 1980. 428 – 431.
29) DORSEY, Ernest N. “The Velocity of Light, Part 1”. Transactions of the American Philiosophical Society 34, (1944); 1 – 110, table 22.
30) POWELL, L. J; MURPHY, T. J; GRAMLICH, J. W. “The Absolute Isotopic Abundance & Atomic Weight of a Reference Sample of Silver”. NBS Journal of Research 87, (1982); 9-19.

Anexo A

A. Comandos y funciones utilizadas en SAS, SPSS y EXCEL

La tabla A.1 presenta los comandos utilizados en SAS, SPSS y EXCEL para realizar las estimaciones estadísticas.

SAS	SPSS	EXCEL
	Estadísticos Univariados	
PROC UNIVARIATE data=dataset; var x; **RUN**;	DESCRIPTIVES VARIABLES=x /STATISTICS=MEAN STDDEV .	**PROMEDIO** (**número1**;número2;...) **DESVEST** (**número1**; número2; ...)
	Análisis de varianza	
PROC ANOVA DATA=dataset; CLASS x; MODEL y = x; **RUN**;	ONEWAY x BY y /MISSING ANALYSIS .	Herramientas Análisis de datos **Análisis de varianza de un factor** (Asistente)
	Regresión lineal simple	
PROC REG DATA=dataset; model y= x ; output out=resultado predicted=yhat residual=e stdi=seyhat stdp=semu; **RUN**;	REGRESSION /MISSING LISTWISE /STATISTICS COEFF OUTS R ANOVA /CRITERIA=PIN(.05) POUT(.10) /ORIGIN /DEPENDENT y /METHOD=ENTER x .	Herramientas Análisis de datos **Regresión** (Asistente)
	Distribuciones normal, ji cuadrada, t de Student y F	
PROBNORM(x) PROBchi(x,gl) PROBT(x,gl) PROBF(X,gl_1,gl_2)	CDFNORM(x) CDF.CHISQ(x, gl) CDF.T(x, gl) CDF.F(x, gl_1, gl_2)	**DISTR.NORM.ESTAND** (**z**) **DISTR.CHI** (**x;grados_de_libertad**) **DISTR.T** (**x;grados_de_libertad;colas**) **DISTR.F** (**x;grados_de_libertad;grados_de_libertad2**)

Tabla A.1

Anexo B

B. Resultados obtenidos en las pruebas aplicadas a SAS, SPSS y EXCEL

La tabla B.1 presenta los resultados correspondientes a la media.

Conjunto de Datos	Nivel de Dificultad	NIST	SAS	λ_μ	SPSS	λ_μ	EXCEL	λ_μ
PiDigits	Bajo	4.534800000000000	4.535107	4.2	4.534800000000000	15	4.534800000000000	15
Lottery	Bajo	518.958715596330000	520.603700	2.5	518.958715596330000	15	518.958715596330000	15
Lew	Bajo	-177.435000000000000	-177.256000	3.0	-177.435000000000000	15	-177.435000000000000	15
Mavro	Bajo	2.001856000000000	2.001857	6.3	2.001856000000000	15	2.001856000000000	15
Michelso	Bajo	299.852400000000000	299.852424	7.1	299.852400000000000	15	299.852400000000000	15
NumAcc1	Bajo	10000002.000000000000000	10000002.500000	7.3	10000002.000000000000000	15	10000002.000000000000000	15
NumAcc2	Medio	1.200000000000000	1.200000	15	1.200000000000000	15	1.199999999999990	14.0
NumAcc3	Medio	1000000.200000000000000	1000000.200000	15	1000000.200000000000000	15	1000000.200000000000000	15
NumAcc4	Alto	10000000.200000000000000	10000000.200000	15	10000000.200000000000000	15	10000000.200000100000000	14.0

Tabla B.1

La tabla B.2 presenta los resultados correspondientes a la desviación estándar.

Conjunto de Datos	Nivel de Dificultad	NIST	SAS	λ_σ	SPSS	λ_σ	EXCEL	λ_σ
PiDigits	Bajo	2.867339060288710	2.86754000	4.2	2.867339060289000	13	2.867339060288710	15
Lottery	Bajo	291.699727470969000	291.35895000	2.9	291.699727470969000	15	291.699727470969000	15
Lew	Bajo	277.332168044316000	278.02007000	2.6	277.332168044316000	15	277.332168044316000	15
Mavro	Bajo	0.000429123454003	0.00043350	2.0	0.000429123454003	12.1	0.000429123453846	9.4
Michelso	Bajo	0.079010547819052	0.07941227	2.3	0.079010547819084	12.4	0.079010548233645	8.3
NumAcc1	Bajo	1.000000000000000	0.70710678	0	1.000000000000000	15	1.000000000000000	15
NumAcc2	Medio	0.100000000000000	0.10005004	3.3	0.100000000000000	15	0.100000000000271	11.6
NumAcc3	Medio	0.100000000000000	0.10005004	3.3	0.100000000034647	9.5	0.107238052947636	1.1
NumAcc4	Alto	0.100000000000000	0.10005004	3.3	0.100000000034647	9.5	0.000000000000000	0

Tabla B.2

La tabla B.3 presenta los resultados correspondientes al análisis de varianza de un factor (ANOVA).

Conjunto de Datos	Nivel de Dificultad	NIST	SAS	λ_F	SPSS	λ_F	EXCEL	λ_F
SiRstv	Bajo	1.180462374402550	0.94	0	1.180462374402240	12.6	1.180462378112610	8.5
SmLs01	Bajo	21.000000000000000	20.88	2.2	12.126074498567300	0	21.000000000000100	14.3
SmLs02	Bajo	201.000000000000000	200.89	3.3	195.128963101916000	1.5	201.000000000057000	12.5
SmLs03	Bajo	2001.000000000000000	2000.89	4.3	2071.201884741940000	1.5	2000.999999999490000	12.6
AtmWtAg	Medio	15.946733567793000	15.05	1.3	15.946733613450700	8.5	15.670329670329700	1.8
SmLs04	Medio	21.000000000000000	20.88	2.2	12.126074498567300	0	20.560344827586200	1.7
SmLs05	Medio	201.000000000000000	200.89	3.3	195.128963101916000	1.5	218.076923076923000	1.1
SmLs06	Medio	2001.000000000000000	2000.89	4.3	2071.201884741940000	1.5	NA	NA
SmLs07	Alto	21.000000000000000	20.91	2.4	12.126074498567300	0	NA	NA
SmLs08	Alto	201.000000000000000	200.76	2.9	195.128963101916000	1.5	NA	NA
SmLs09	Alto	2001.000000000000000	1998.90	3.0	2071.201884741940000	1.5	NA	NA

Tabla B.3

Las tablas B.4, B.5 y B.6 presentan los resultados correspondientes a la regresión lineal simple.

Conjunto de Datos	Nivel de Dificultad	*Parámetro*	NIST	SAS	λ_β	SPSS	λ_β	EXCEL	λ_β
Norris	Bajo	B0	-0.262323073774029	-0.27436	1.3	-0.262323073774151	12.3	-0.274362682463845	1.3
		B1	1.002116818020450	1.00213	4.9	1.002116818020450	14.3	1.002134013723230	4.8
NoInt1	Medio	B1	2.074380165289260	2.06665	2.4	2.074380165289260	14.7	2.066651157380480	2.4
NoInt2	Medio	B1	0.727272727272727	0.72131	2.1	0.727272727272727	15.3	0.721311475409836	2.1

Tabla B.4

Conjunto de Datos	Nivel de Dificultad	*Parámetro*	NIST	SAS	λ_σ	SPSS	λ_σ	EXCEL	λ_σ
Norris	Bajo	B0	0.232818234301152	0.24482000	1.3	0.232818234284785	10.2	0.244816777719846	1.3
		B1	0.000429796848200	0.00044563	1.4	0.000429796848170	10.2	0.000445625654689	1.4
NoInt1	Medio	B1	0.016528925619835	0.01559000	1.2	0.016528925619840	12.5	0.015591461566347	1.2
NoInt2	Medio	B1	0.042082731807843	0.06557000	0.3	0.042082731807843	14.3	0.065573770491803	0.3

Tabla B.5

Conjunto de Datos	Nivel de Dificultad	NIST	SAS	λ_r	SPSS	λ_r	EXCEL	λ_r
Norris	Bajo	0.884796396144373	0.89763	1.8	0.884796396083009	10.2	0.897627147419838	1.8
NoInt1	Medio	3.567530340063380	3.23255	1.0	3.567530340063670	13.1	3.232550319362940	1.0
NoInt2	Medio	0.369274472937998	0.51215	0.4	0.369274472937997	14.5	0.512147519731585	0.4

Tabla B.6

La tabla B.7 presenta los resultados correspondientes a la distribución normal.

z	ELV	SAS	*Er*	λ	SPSS	*Er*	λ	EXCEL	*Er*	λ
0.0001	0.500040	0.50004	0	6	0.500039894227973	0	6	0.500039894265737	0	6
0.001	0.500399	0.50040	2E-6	5.7	0.500398942213911	0	6	0.500398944529549	0	6
0.01	0.503989	0.50399	2E-6	5.7	0.503989356314631	0	6	0.503989378888570	0	6
0.1	0.539828	0.53983	4E-6	5.4	0.539827837277029	0	6	0.539827895533667	0	6
0.9	0.815940	0.81594	0	6	0.815939874653240	0	6	0.815939908268087	0	6
0.99	0.838913	0.83891	4E-6	5.4	0.838912940489169	0	6	0.838912938586950	0	6
0.999	0.841103	0.84110	4E-6	5.4	0.841102654358681	0	6	0.841102648921588	0	6
0.9999	0.841321	0.84132	1E-6	5.9	0.841320547786237	0	6	0.841320541997720	0	6

Tabla B.7

La siguiente tabla presenta los resultados correspondientes a la distribución ji cuadrada.

x	*n*	ELV	SAS	*Er*	λ	SPSS	*Er*	λ	EXCEL	*Er*	λ
2E+4	2E+4	0.501330	0.50133	0	6	0.504205244180257	6E-3	2.2	NA	NA	NA
2E+5	2E+5	0.500421	0.50042	2E-6	5.7	0.500420522052898	0	6	NA	NA	NA
2E+6	2E+6	0.500133	0.50013	6E-6	5.2	0.500132980835392	0	6	NA	NA	NA
4E+6	4E+6	0.500094	0.50009	8E-6	5.1	0.500094030441220	0	6	NA	NA	NA
6E+6	6E+6	0.500077	0.50008	6E-6	5.2	0.500076776203059	0	6	NA	NA	NA
8E+6	8E+6	0.500066	0.50007	8E-6	5.1	0.500066487859678	0	6	NA	NA	NA
1E+7	1E+7	0.500059	0.50006	2E-6	5.7	0.500059475253468	0	6	NA	NA	NA

Tabla B.8

La siguiente tabla presenta los resultados correspondientes a la distribución t de Student.

x	***n***	**ELV**	**SAS**	*Er*	λ	**SPSS**	*Er*	λ	**EXCEL**	*Er*	λ
-9.7	100	2.2516E-16	2.2515E-16	2E-5	4.7	2E-16	1	0	NA	NA	NA
-9.8	100	1.3590E-16	1.3590E-16	3E-5	4.5	1E-16	1	0	NA	NA	NA
-9.9	100	8.2023E-17	8.2023E-17	5E-6	5.3	1E-16	1	0	NA	NA	NA
-10.0	100	4.9508E-17	4.9508E-17	8E-6	5.1	0	1	0	NA	NA	NA
-10.1	100	2.9886E-17	2.9886E-17	3E-6	5.5	0	1	0	NA	NA	NA
-10.2	100	1.8043E-17	1.8043E-17	1E-5	5.0	0	1	0	NA	NA	NA

Tabla B.9

La siguiente tabla presenta los resultados correspondientes a la distribución F.

x	***n_1***	***n_2***	**ELV**	**SAS**	*Er*	λ	**SPSS**	*Er*	λ	**EXCEL**	*Er*	λ
1.0	201	10,001	5.1287E-1	5.1287E-1	2E-6	5.7	5.1287E-1	2E-6	5.7	4.8713E-01	5E-2	1.3
0.9	201	10,001	1.5967E-1	1.5967E-1	3E-5	4.6	1.5967E-1	1E-6	5.8	8.4033E-01	1	0
0.8	201	10,001	1.7628E-2	1.7628E-2	6E-6	5.2	1.7628E-2	2E-6	5.7	9.8237E-01	1	0
0.7	201	10,001	4.5779E-4	4.5779E-4	0	6	4.5779E-4	0	6	9.9954E-01	1	0
0.6	201	10,001	1.6421E-6	1.6420E-6	5E-5	4.3	1.6421E-6	2E-6	5.6	1	1	0
0.5	201	10,001	3.7174E-10	3.7174E-10	3E-6	5.6	3.7200E-10	7E-4	3.2	1	1	0
0.4	201	10,001	1.4682E-15	1.4682E-15	7E-6	5.2	0	1	0	1	1	0
0.38	201	10,001	6.2755E-17	6.2754E-17	8E-6	5.1	0	1	0	1	1	0
0.37	201	10,001	1.1716E-17	1.1716E-17	2E-5	4.8	0	1	0	1	1	0
0.3	201	10,001	9.2465E-24	9.2465E-24	1E-6	5.9	0	1	0	1	1	0

Tabla B.10