CAPITULO 1
CONCEPTOS BASICOS DE ESTADISTICA

1.1. ESTADÍSTICA DESCRIPTIVA

1.1.1. Tablas y Gráficas.

VARIABLES. La estadística trabaja con *datos* de característica variabilidad conocidos por ello como *variables*. Tradicionalmente, las variables se han clasificados en variables cuantitativas y variables cualitativas. Las variables cuantitativas también se conocen como variables propiamente dichas, mientras que las cualitativas se conocen como atributos o lo que los ingleses llaman "categorías" y SAS llama "clases". Una posterior división de las variables cuantitativas es en continuas y discontinuas o discretas.

Variables Cualitativas o Atributos o Variables Categóricas

Discretas

Variables Cuantitativas

Continuas

El sexo de un animal es un atributo, mientras que la producción de leche de una vaca es una variable cuantitativa. Las variables (cuantitativas) se miden, los atributos se cuentan. Por ejemplo, diremos que un tambo tiene 119 vacas y 2 toros, pero que una vaca produce 8 lts. de leche por día (variable continua) y que tuvo 4 terneros en su vida (variable discreta). Por esta razón el análisis de atributos a veces se llama análisis de conteos.

Lotus distingue entre rótulos y valores. SAS distingue entre variables alfanuméricas y numéricas. Las variables alfanuméricas son automáticamente cualitativas, las variables numéricas son por omisión consideradas cuantitativas pero pueden mediante una declaración ser definidas como variables de clasificación.

ESCALAS. Una escala de medición que ha tenido cierta aceptación en los últimos tiempos es:

i. Valores nominales. La escala más rudimentaria es la nominal, donde los objetos se distinguen en base a un nombre, muchas veces dado por un número. Por ejemplo en el sexo de animales, se puede acordar un número para simbolizar a cada sexo, pero ese número es arbitrario y un investigador puede definir macho como 0 y hembra como 1, mientras que otro puede utilizar exactamente lo opuesto. Las escalas nominales se usan en atributos.

ii. Valores ordinales. Las mediciones en una escala ordinal solo indican orden ("ranking"). Los objetos en una escala ordinal se distinguen, pues, en base a la cantidad relativa de una característica que poseen. Ejemplos de esto son los grados usados en la medición de estado corporal y la conformación en vacas (pobre, regular, buena, excelente). Una escala es: 0, 1, 2, 3, 4, y 5, pero puede haber otras diferentes que distingan igualmente el orden de preferencia de los animales.

iii. Valores por intervalos. Cuando las diferencias entre objetos tiene sentido, es decir que la unidad de medida es fija. Generalmente tienen un cero, aunque éste es arbitrario, como en el caso de la temperatura medida en grados centígrados, donde el cero no indica ausencia de temperatura. No tiene sentido acá decir que una temperatura de 60 grados es doble que una de 30.

iv. Valores racionales. Cuando los cocientes (razones) de valores tienen sentido la escala es racional. Un ejemplo es el peso, donde un objeto que pese 60 kg. pesa el doble de uno que pesa 30 kg.

TABLAS. Muchas veces, al comienzo de un trabajo de análisis de datos se cuenta con un gran volumen de información en bruto. Una de las primeras tareas es organizar esa información y tabularla. El propósito de la tabulación es resumir la información hasta llegar, a veces, a un par de valores (la media y la varianza por ejemplo) que encierran toda la utilidad de la información. Conviene destacar la posible diferencia que puede existir entre un experimento diseñado, donde se supone que el investigador sabe a donde quiere llegar y ya tiene decidido todo el análisis. Si el conjunto de datos surge de una realidad sin predeterminada organización ("observational studies") puede que no haya una idea previa y el curso de acción a seguir dependerá del caso particular. En todos los casos, cierto tipo de gráfica o figura ayudará a la interpretación de los datos.

Ejemplo 1.1. Consideremos los siguientes datos de peso de 60 animales:

234 225 234 225 234 204 225 231 245 202 213 222 231 245 193 202 213 222 229 243
254 193 202 213 220 229 243 254 193 200 211 218 227 243 254 265 184 191 197 211
216 227 240 250 263 274 145 177 188 197 209 216 227 236 247 256 272 288 304 210

El autor ordenó los datos del siguiente modo:

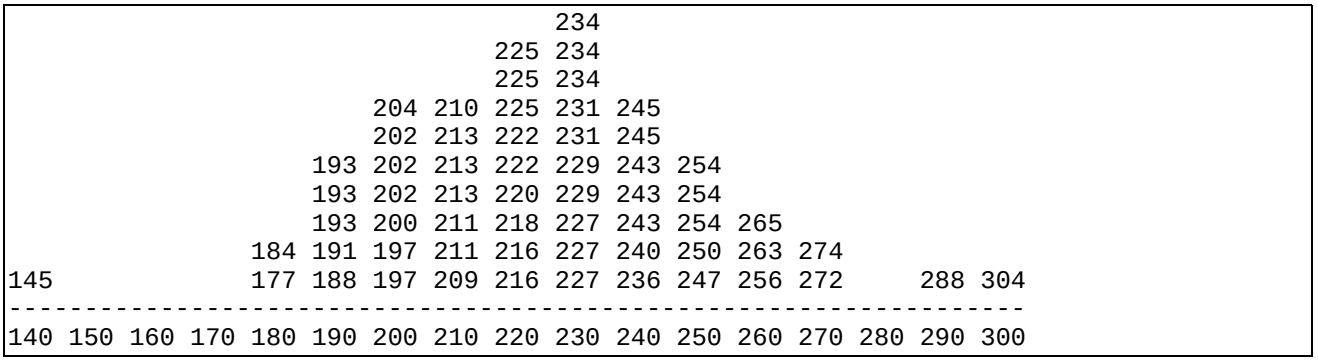


Figura 1.1. Representación de los datos.

En la figura se ve con naturalidad la idea de clases y de gráfico de barras. Los valores entre 176 y 185 se consideran una clase, los entre 186 y 195 otra y así sucesivamente.

Tabla 1.1. Tabulación de los datos del ejemplo 1.1.

Límites .....de la clase	Marca de la clase	n	Observaciones comprendidas en la clase	Media de la clase	"Steam and leaf"
136 145	140	1	145	145	14 5
146 155	150	0			15
156 165	160	0			16
166 175	170	0			17 7
176 185	180	2	177 184	180	18 4 8
186 195	190	5	188 191 193 193 193	191	19 1 3 3 3 7 7
196 205	200	7	197 197 200 202 202 202 204	200	20 0 2 2 2 4 9
206 215	210	7	209 210 211 211 213 213 213	211	21 0 1 1 3 3 3 6 6 8
216 225	220	9	216 216 218 220 222 222 225 225 225	221	22 0 2 2 5 5 5 7 7 7 9 9
226 235	230	10	227 227 227 229 229 231 231 234 234 234	230	23 1 1 4 4 4 6
236 245	240	7	236 240 243 243 243 245 245	242	24 0 3 3 3 5 5 7
246 255	250	5	247 250 254 254 254	251	25 0 4 4 4 6
256 265	260	3	256 263 265	261	26 3 5
266 275	270	2	272 274	273	27 2 4
276 285	280	0			28 8
286 295	290	1	288	288	29
296 305	300	1	304	304	30 4

En la tabla 1.1 se presenta una forma habitual de tabular datos como esos. Una columna con los límites de cada clase, una con la marca de la clase (es decir el valor que representa la clase, generalmente el punto medio), y la frecuencia absoluta n_i. Esta última es el número de observaciones que caen en cada clase. Un concepto relacionado es el de frecuencias relativas, que es el número de observaciones de cada clase dividido por el total de observaciones, simbolizado por f_i.

GRAFICAS. HISTOGRAMAS. En la figura 1.2 se representa la frecuencia (absoluta o relativa) de cada clase con la altura de la barra. Estas gráficas se llaman gráficos de barras.

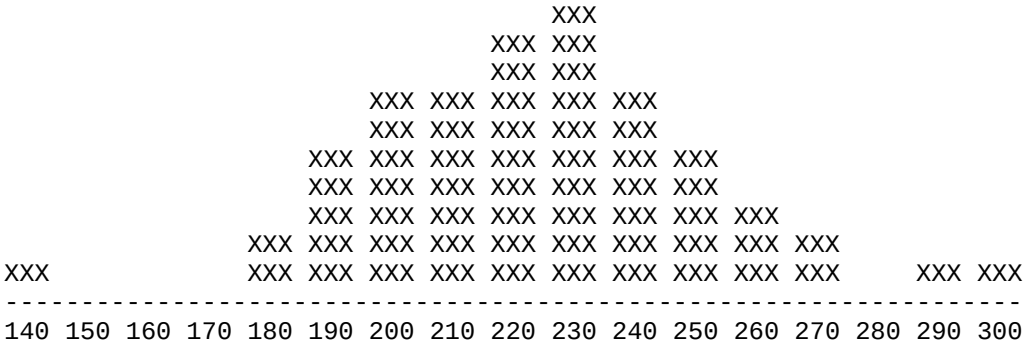


Figura 1.2. Gráfico de barras con los datos del ejemplo 1.1.

El histograma es una representación gráfica en la que la frecuencia (absoluta o relativa) de la clase está representada por el área de la barra. Si todas las clases tienen igual amplitud, la frecuencia de la clase está representada por la altura de la barra y el gráfico se conoce como gráfico de barras. Mucha gente no reconoce las diferencias y llama histograma a los gráficos de barras.

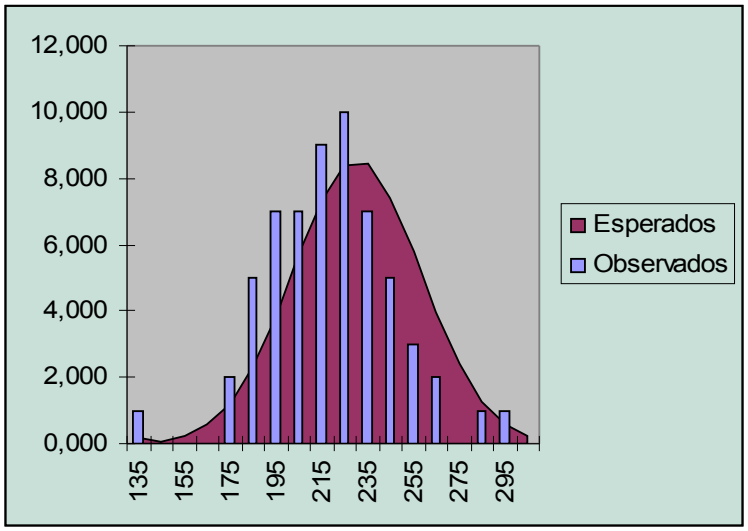


Figura 1.3. Grafico de barras con distribución normal superpuesta de distribución de los pesos de toritos.

STEAM-AND-LEAF. Si en lugar de representar cada valor por una marca cualquiera lo representamos por el dígito que lo identifica, no perdemos ese dato lo que puede ser de utilidad para el cálculo de ciertas cantidades como la media, tal cual se verá más adelante. Este es un caso de herramientas llamadas semi-gráficas.

1.1.2 Medidas de posición.

MEDIA ARITMETICA. La media es la suma de los valores dividido el número de valores. Si la media pertenece a una población se representa con la letra griega μ, si pertenece a una muestra con el símbolo de la variable con una barra encima ([1]):

Muestral	Poblacional
$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$	$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$

Ejemplo 1.1 (Cont.) La media de peso de los 60 toritos del ejemplo 1.1 es 225,2666 como puede comprobarse sumando los 60 valores y dividiendo por 60.

Ejemplo 1.2. Consideremos los siguientes datos de fertilizaciones usadas en un experimento: 80, 80, 240, 240, 160, 0, 320, 160, 160, 0, 0, 320, 320. La media de estos valores es: (80, 80, 240, 240, 160, 0, 320, 160, 160, 0, 0, 320, 320)/13=160

MEDIA DE DATOS AGRUPADOS. Nótese en el ejemplo 1.2 que algunos valores se repiten. En estos casos de datos repetidos no los sumamos sino que los multiplicamos: (0*3+80*2+160*3+240*2+320*3)/10=2080/13=160 como antes. La única particularidad del cálculo de la media si los datos están agrupados es, pues, que los valores deben multiplicarse por la frecuencia en que cada dato ocurre:

$\overline{X} = \frac{\sum_{i=1}^{m} X_i n_i}{n}$. Para calcular la media a partir de las frecuencias relativas se usa la fórmula :

$\overline{X} = \sum_{i=1}^{m} X_i f_i$

MEDIA PONDERADA. La media de datos agrupados se considera una media ponderada por la frecuencia de las observaciones, pero no es el único caso de media ponderada.

Ejemplo 1.3. Las notas de un curso son el resultado de ponderar el promedio de los exámenes parciales por 0,4 y la nota del examen final por 0,6. Un estudiante que tuvo las siguientes notas:

Primer parcial	50
Segundo parcial	60
Tercer parcial	100
Examen final	80

tiene el siguiente promedio de parciales (media simple): (50 + 60 + 100)/3=70, y su nota final del curso (media ponderada) es: 0,4*70 + 0,6*80 = 76

[1] Los valores poblacionales se llaman parámetros mientras que los muestrales se llaman estadísticos o estadígrafos

PROPIEDADES DE LA MEDIA. Tomando la convención $x=X-\overline{X}$, llamada variable centrada, podemos resumir algunas propiedades de la media así:

i. La suma de los desvíos respecto de la media es cero: $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} (X_i - \overline{X}) = 0$

Ejemplo 1.2 (Cont.) Los desvíos acá son:[(-52)(2) + (-32)(1) +(-12)(1) + (8)(1) + (28)(5)]/10=0

ii. La suma de los cuadrados de los desvíos es menor con respecto a la media que con respecto a cualquier otro valor: $\sum_{i=1}^{n} x_i^2 \leq \sum_{i=1}^{n} (X_i - a)^2$ para cualquier a

iii. La media de una variable más una constante es igual a la media de la variable más la constante: $\overline{X} = a + \overline{d}$

iv. La media del producto de una constante por una variable es igual a la media de la variable por la constante: $b\,X = b\,X$

Las propiedades iii y iv se pueden resumir así: $(a+bX) = a + b\,X$

v. La media de la suma de variables es igual a la suma de las medias: $(X+Y) = X + Y$

MEDIANA. La mediana es el valor de la variable que divide la distribución de tal modo que la mitad de los valores son iguales o menores que ella y la otra mitad son iguales o mayores. Si los datos no se repiten y no están agrupados para calcular la mediana basta con ordenarlos y contarlos: el que ocupe el lugar del medio es la mediana. Si hay un número par, muchos definen la mediana como el promedio de los dos valores intermedios. Si los datos están agrupados aunque sea fácil identificar a la clase que contiene a la mediana, el valor no está unívocamente definido y se puede interpolar (ver Spiegel, 1969).

Comparación de la media y la mediana.

MODA. La moda o modo es el valor más frecuente de la variable. Si los valores no se repiten no hay una moda única. Una distribución puede tener más de una moda, si tiene una sola es unimodal, de lo contrario bimodal etc. La moda se usa generalmente en atributos. En el ejemplo 1.2 la moda es el valor 80 que es el más frecuente. En el ejemplo 1.1 la clase modal es la que va de 226 a 235, de modo que para muchos efectos se considera que la moda es 230. Opcionalmente, se puede interpolar como hace Spiegel (1969).

OTRAS MEDIDAS DE POSICION. ***Media Geométrica.*** La media geométrica es la raíz de orden n del producto de los n valores. Eso equivale al antilogaritmo del promedio de los logaritmos.

Media Armónica. La media armónica es la inversa del promedio de las inversas. Es decir que se toman las inversas de las observaciones, se las promedia y se invierte el valor obtenido.

Media Cuadrática. La media cuadrática es la raíz cuadrada de la media de los cuadrados. Es decir que los valores se elevan al cuadrado, se promedian los cuadrados y luego se toma la raíz cuadrada.

Aunque estas últimas parecen artificiales complicaciones tienen aplicación en determinadas circunstancias aunque no con mucha frecuencia.

Ejemplo 1.4. Los siguientes valores muestran una característica de los dos padres y del promedio de los hijos. Los autores observan que la media geométrica se acerca más al valor de la descendencia que la media aritmética.

Padre 1	Padre 2	Descendencia	Media Geométrica	Media Aritmética
54.1	1.1	7.4	7.7	27.6
57.0	1.1	7.1	7.9	29.1
173.6	1.1	8.3	13.8	87.4
53.0	5.1	23.0	16.4	29.1
150.0	12.4	47.5	43.1	81.2

Ejemplo 1.5. Un ejemplo de media armónica está dado por el siguiente problema: Un coche recorre los 500 km. (aproximadamente) entre Salto y Montevideo en 8 horas al ir y en 6 horas al volver. ¿Cuál fue la velocidad media en el viaje de ida? ¿Cuál fue la velocidad media en el viaje de vuelta? ¿Cuál fue el promedio de la velocidad? La velocidad en el viaje de ida es 500 km/8 h = 62,50 km/h. La velocidad en el viaje de vuelta fue de 500 km/6 h = 83,33 km/h. La velocidad promedio fue de 1000 km/14 h = 71,43 km/h. Esta velocidad no es la media aritmética de 62,50 y 83,33 (que es 72,92) sino la media armónica entre ellas: [1/2][(1/62,50)+(1/83,33)] = [1/2][1/(0,016 + 0,012)] = 71,43. Aunque la diferencia no es muy grande se puede apreciar que son dos cosas diferentes. Otro ejemplo de media armónica se presenta en la salida de SAS con datos desbalanceados de la sección 3.4.3.

Ejemplo 1.6. Si se analizan los desvíos con respecto a la media (en el ejemplo 1.2 resulta fácil) se puede concluir que su media cuadrática tiene un significado muy especial, como veremos a continuación se le conoce como desviación estándar.

Transformación de variables. Como vimos la media geométrica corresponde con una transformación logarítmica y la media armónica corresponde con una transformación inversa. Hay otras transformaciones que son frecuentes en el análisis estadístico.

1.1.3.Medidas de dispersión.

VARIANZA. La varianza de una muestra (^{2}) se define como: $S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}$, mientras que la varianza de una población finita, de N elementos, se define como: $\sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{N}$. Cuando se refiere a la varianza sin especificar si es muestral o poblacional se utiliza la expresión Var o V.

El numerador de la varianza se le conoce como suma de cuadrados y se calcula generalmente como: $\sum (X - \overline{X})^2 = \sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}$. Al denominador (n-1) se le conoce como grados de libertad. Se usa el valor (n-1) porque si se usa n se subestima la varianza poblacional (ver sección 1.2.3.1). La varianza como medida de dispersión nos sirve para determinar si desviaciones observadas son usuales o notorias.

Ejemplo 1.1 (Cont.) La varianza de los datos del ejemplo 1.1 es 778,1955.

VARIANZA PARA DATOS AGRUPADOS. La varianza para datos agrupados se define como: $S^2 = \frac{\sum_{i=1}^{m} (X_i - \overline{X})^2 n_i}{\sum_{i=1}^{m} n_i - 1}$

PROPIEDADES DE LA VARIANZA. Algunas propiedades de la varianza son:

1. La varianza es invariante respecto a un cambio de origen (sumarle una cantidad igual a todos los valores), pero no es invariante respecto a un cambio de escala (multiplicar por una constante los valores): $V[a+bX]^2 = b^2 V[X]$

2. La varianza de una suma (o de una diferencia) de variables es la suma de las varianzas, más (o menos) el doble de la covarianza: $V[X \pm Y] = V[X] + V[Y] \pm 2\, Cov[X,Y]$. Al principio asombra de que la diferencia de variables sume las varianzas, pero es claro que restar variables no puede disminuir la variabilidad sino al contrario. Restar variables suma las varianzas.

[2] Algunos definen como varianza a la suma de cuadrados dividida por n y le dan el nombre de 'cuasi-varianza' cuando se divide por n-1.

DESVIACIÓN ESTANDAR. La desviación estándar o desviación típica es la raíz cuadrada (positiva) de la varianza: $S=+\sqrt{S^2}$. La desviación estándar tiene la ventaja de que se expresa en las mismas unidades que la variable en estudio, pero no tiene las propiedades matemáticas de la varianza, por lo que la consideramos un subproducto de la varianza.

Ejemplo 1.1 (Cont.): La desviación estándar de los datos del ejemplo 1.1 es: $S = \sqrt{}\ 778.196 = 27.897$.

COEFICIENTE DE VARIACIÓN. El coeficiente de variación es el cociente entre la desviación estándar y la media: $CV = Sx\ *100/\overline{X}$. Muchas veces el coeficiente de variación se expresa en porcentaje: $CV = Sx\ *100/\overline{X}$. Ejemplo 1.1 (Cont.). El coeficiente de variación del ejemplo 1.1 es: 27,89*100 /225=12,39%

El coeficiente de variación se utiliza para comparar la variabilidad de características que tienen diferentes unidades de medidas. Supongamos que a un investigador le interesa saber si dos variedades de un cultivo varían más en rendimiento o en número de plantas. Resulta difícil comparar kg. contra plantas, por lo que acude al coeficiente de variación. En general en la investigación en Uruguay el coeficiente de variación ha recibido una atención desproporcionada a su importancia.

CUANTILES. Los cuantiles, de los cuales los más usados son los percentiles, son valores de la variable que dividen la distribución en determinadas partes, por ejemplo los percentiles en 100. Constituyen una extensión del concepto de la mediana, que divide la distribución en dos por lo que es el percentil 50. Por supuesto que también se puede decir que la mediana es un caso particular de percentil. Por la forma que definen la distribución constituyen medidas de dispersión al mismo tiempo que de posición.

MOMENTOS. Otras medidas incluyen los coeficientes de asimetría y de curtosis, que son de utilidad especialmente para comprobar la normalidad de variables y que no se discutirán en este trabajo (sección 1.3.4 pg. 29). En general, se habla de momentos, que son los promedios (mas adelante diremos los valores esperados) de potencias de las variables. El momento ordinario de tercer orden ($\frac{\sum X^3}{n}$) mide la asimetría y el de cuarto orden ($\frac{\sum X^4}{n}$) la curtosis (es decir el achatamiento) de la distribución.

1.1.4.Medidas de Covariación y Otras.

COVARIANZA. Una parte importante de describir un conjunto de datos es proporcionar la relación que existe entre dos o más variables cuantitativas. Este tema será discutido con más detalle en el próximo capítulo pero acá presentamos a la covarianza.

Figura 1.5. Cambio de coordenadas y coeficiente de correlación.

Si se tiene un conjunto de valores como las que se grafican en la figura 1.5, puede verse mediante el cambio de variables: $x= X - \overline{X}$ e $y = Y - \overline{Y}$, que se logra un cambio en los ejes coordenados, porque el nuevo sistema (x,y) tiene su origen en el punto $(\overline{X}, \overline{Y})$ del anterior. Los valores de X mayores que la media tendrán x mayor a cero, estando ubicados a la derecha de la gráfica; en tanto, los valores de Y mayores que su media tendrán valores de y positivos, estando por encima del eje $\overline{Y}$. Tomando los productos de ambas variables reducidas x*y, observamos que tienen signo positivo en el primer y tercer cuadrantes, mientras que tienen signo negativo en el segundo y cuarto; tomando la sumatoria de esos productos para cada par de valores X,Y se puede visualizar su alineación. La sumatoria será positiva en caso de alineación del primer al tercer cuadrante, será negativa en caso de alineación del segundo al cuarto, y nula si la distribución es uniforme. La estadística así obtenida presenta dos inconvenientes. El primero es que depende del tamaño de la muestra, lo que se soluciona tomando el cociente entre la sumatoria de productos y el tamaño de la muestra, con lo que se obtiene la covarianza muestral.

Figura 1.6. Valores del coeficiente de correlación.

COEFICIENTE DE CORRELACIÓN. El segundo inconveniente que mencionábamos es la dependencia de las unidades de medida, lo que se soluciona dividiendo por las desviaciones estándar de ambas variables. El coeficiente así obtenido se conoce como coeficiente de correlación o de Pearson, debido a Karl Pearson que lo propuso. El coeficiente de correlación es pues la covarianza dividida por el producto de las desviaciones estándares. El coeficiente de correlación vara entre -1 y +1. Cuando los puntos se alinean perfectamente con pendiente negativa vale -1, cuando la alineación es perfecta con pendiente positiva es +1 y los casos intermedios corresponden a diagramas de dispersión elípticos.

Ejemplo 1.7. Los datos que se muestran en la tabla corresponden a dos variables, llamémosle X y Y. X es la nota que obtuvo un estudiante en un examen de ingreso a la universidad, Y es el promedio de notas de su primer año en la universidad.

X	Y	x	y	x^2	y^2	xy
37	97	-13	18,86	169	355,59	-245,14
36	30	-14	-48,14	196	2317,73	674,00
97	97	47	18,86	2209	355,59	886,29
27	77	-23	-1,14	529	1,31	26,29
55	63	5	-15,14	25	229,31	-75,71
84	87	34	8,86	1156	78,45	301,14
14	96	-36	17,86	1296	318,88	-642,86
350	547	0	0	5580	3656,86	924,00
50	78,1	0	0	797,14	522,41	132,00

Las respectivas medias son: $\overline{X} = 50$ y $\overline{Y} = 78,1$. En la tercer y cuarta columnas se presentan los desvíos con respecto a las medias de los valores de X y de Y, se puede verificar que suman cero. Finalmente, en la quinta columna se presentan los productos. La covarianza es el promedio de esos productos de desvíos con respecto a la media:

$$S_{XY} = \frac{\sum xy}{n} = \frac{924}{7} = 132$$

El coeficiente de correlación es la covarianza dividida por el producto de las desviaciones estándares:

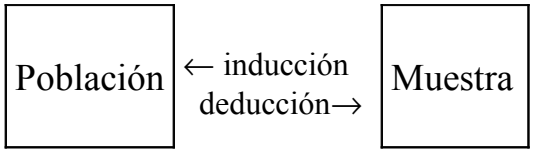
$$r = \frac{\text{Co var}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{132,0}{\sqrt{(797,14)(522,41)}} = 0,20$$

La correlación y regresión se estudiarán con más detalle en el capítulo 2. Al coeficiente de correlación de Pearson se le llama también “coeficiente de producto momento”.

1.2.INFERENCIA ESTADÍSTICA

1.2.1.Conceptos Generales de Inferencia

POBLACIÓN Y MUESTRA. Conviene distinguir entre población y muestra El conjunto de todos los datos en consideración se denomina *poblacional* o *universo* y un subconjunto es una *muestra.*



Sacar conclusiones de la población para la muestra (de lo general a lo particular) es hacer *deducción*, mientras que sacar conclusiones de la muestra para la población (de lo particular a lo general) es inducción o *inferencia*. La parte de la estadística que describe como hacer inferencia es la *inferencia estadística*. Nosotros trabajaremos con *muestras aleatorias* (obtenidas al azar) como representativas de la población. Si la muestra no es aleatoria los resultados de la teoría estadística no son válidos. Considerando los ejemplos 1.1 y 1.2, vemos que en el primer caso se puede intentar sacar inferencias sobre todos los animales de esa raza y condición (la población), mientras que en el segundo caso no tiene sentido decir que esos valores son una muestra representativa de alguna población, ya que si el investigador hubiese querido podría haber usado otras fertilizaciones a voluntad.

Ejemplo 1.11. En la siguiente tabla se presentan los resultados de una famosa experiencia en genética. Son dos poblaciones parentales (P1 y P2) que se cruzaron y se obtienen dos generaciones derivadas (F1 y F2). La F2 proviene de autofecundar la F1. Los datos representan la frecuencia con que se presentan diferentes alturas de las cuatro poblaciones de plantas en centímetros.

Cm	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	n	Media	Varianza
P1	4	21	24	8														57	6,63	0.65
P2									3	11	12	15	26	15	10	7	2	101	16,80	3.53
F1				1	12	12	14	17	9	4								69	12,12	2.28
F2		1	10	19	26	47	73	68	68	39	25	15	9	1				401	12,89	5.06

Los autores dicen que las poblaciones F1 y F2 no difieren en media pero sin en varianza. Mientras que las poblaciones P1 y P2 difieren en media pero no en varianza.

MODELOS. Como las poblaciones son muchas veces conceptuales (no reales) o infinitas se las define en un modelo. En el ejemplo anterior los posibles rendimientos que pueden proporcionar parcelas del cultivo constituyen una población infinita o imaginaria. Un conjunto de supuestos con una estructura de predicción constituye un modelo. Al decir que una raza de animales tiene una distribución normal con media 600 y varianza 3.600, estamos adoptando un modelo.

Modelo lineal aditivo. Muchas veces se postula que cada observación es la suma de una media mas un error aleatorio: $Y_i = \mu + \varepsilon_i = \bar{Y} + e_i$. Este tipo de modelo se conoce como aditivo porque la variable Y se explica por la suma de μ y ε. Se le llama lineal debido a que ninguno de los parámetros está sometidos a multiplicaciones con otros parámetros. Cuando usemos modelos más complejos la característica de lineal (en oposición a cuadrático, exponencial, etc.) aparecerá más clara.

ESTIMACION DE PARÁMETROS. Los modelos incluyen parámetros, como la media, la varianza y la proporción, que resultan desconocidos por lo que se intenta estimarlos a través de estadísticos, llamados estimadores por tal razón.

Las estimaciones pueden ser: *puntuales* o *por intervalos*. En las estimaciones puntuales se toma un valor para el parámetro, por ejemplo el valor mas probable. En las estimaciones por intervalos se toma un intervalo en el que se estima que el parámetro estará comprendido. Las primeras tienen mayor facilidad de uso en ciertas ocasiones, las segundas tienen una probabilidad conocida de ser correctas.

En el ejemplo tenemos que la media de la variedad P1 puede ser estimada puntualmente por la media obtenida 6,63 o en intervalo diciendo que está entre 5,82 y 7,44 (como se calcula más adelante).

PROPIEDADES DE LOS ESTIMADORES. Existen una serie de propiedades deseables en un estimador:
1 - ***Insesgamiento***. Asegura que los investigadores que usan este método no se equivocan en promedio.
2 - ***Eficiencia***. Dice que si se usa un método A que, por ejemplo, tiene 110% la eficiencia de B, entonces B necesita 110% el número de observaciones que necesita A para tener igual precisión.
3 - ***Consistencia***. Indica que la estimación mejora con el aumento del tamaño de muestra.
4 - ***Distribución conocida***. Posibilita construir estimadores por intervalos.

Precisión y exactitud de una estimación. La exactitud de una estimación se refiere a la cercana entre la cantidad que se desea estimar, por ejemplo μ, y su estimador, en este caso $\overline{X}$. La precisión de una estimación, $\overline{X}$, en este caso, se mide por el error estándar del estimador. Generalmente se escribe (ver Mood y Graybill [1976]):

$$E[\hat{\theta} - \theta]^2 = E[\hat{\theta} - E(\theta)]^2 + [E(\hat{\theta}) - \theta]^2$$

Exactitud = Precisión + Sesgo

METODOS DE ESTIMACIÓN. Existen diferentes métodos de estimación, es decir métodos de encontrar estimadores, de los que mencionaremos:

1 - Método de los momentos. Consiste en igualar los momentos de la población con los momentos muestrales. Por ejemplo, se puede decidir estimar la media de la población por la media de la muestra, la varianza poblacional por la varianza muestral o la correlación poblacional por la correlación muestral.
2- Método de la ***máxima verosimilitud***. Propone considerar como estimaciones los valores más probables del parámetro.
3 - Método de los ***mínimos cuadrados***, uno de los más importantes a nuestros efectos. acá. Propone como estimador el valor que haga mínima la suma de cuadrados de los desvíos.
4- Mínimo χ^2. Para algunas situaciones en que se usa χ^2, se puede proponer como estimador el valor que minimice ese χ^2.

DISTRIBUCIONES EN EL MUESTREO. Supongamos que de la población extraemos sucesivamente m muestras todas de tamaño n, como se observa en la figura 1.5.

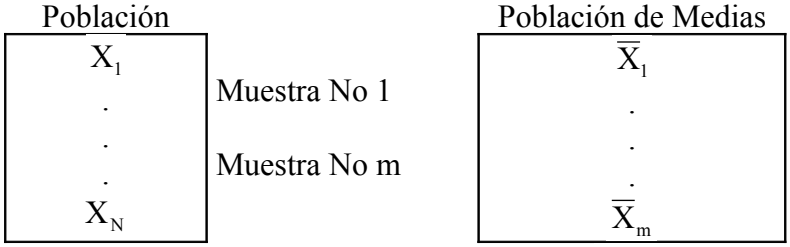
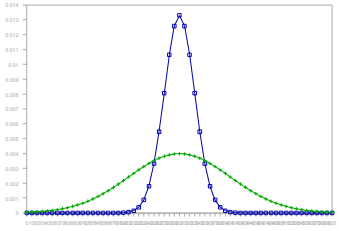


Figura 1.5. Población y población de medias muestrales.

$\mu_{\overline{X}} = \mu_X$ y $\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$. A la desviación estándar de las estadísticas se les llama error estándar. Así, por ejemplo, el error estándar de $\overline{X}$ es $\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$

DISTRIBUCIÓN DE LAS MEDIAS MUESTRALES. TEOREMA DEL LIMITE CENTRAL. Las medias de muestras aleatorias tienen distribución normal si provienen de poblaciones normales o tienden a distribuirse normalmente al aumentar el tamaño de las muestras si la distribución no es normal. La media de la población de medias muestrales es la media de la población, y la varianza es una enésima parte de la varianza poblacional.

Estandarización de la distribución de medias. Como toda distribución normal, la de las medias se puede estandarizar: $\frac{\overline{X} - \mu}{\sigma_{\overline{X}}} = z \sim N(0,1)$ donde $\sigma_{\overline{X}}$ es la desviación estándar de la variable medias muestrales llamada también el error estándar de la media.



Distribución t de Student. Si no conocemos σ no podemos utilizar la distribución normal pero W. S. Gosset ("Student") construyó tablas con la distribución que tiene el cociente $(\overline{X} - \mu)\sqrt{n} / s$ denominado por ello con el seudónimo que él utilizó "t de Student". La distribución de Student tiene un nuevo parámetro, los grados de libertad, y al aumentar éstos tiende a la distribución normal. Por lo tanto se puede considerar a la normal una t con infinitos grados de libertad (ver figura 1.8 en página 24).

OTRAS DISTRIBUCIONES EN EL MUESTREO. ***Distribución X^2***.

$$\chi^2_{(n)} = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma_X} = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{\sigma_x} + \frac{(\overline{x} - \mu)^2}{\sigma_{\overline{x}}} = \chi^2_{(1)} + \chi^2_{(n-1)}$$

La distribución χ^2 tiene una propiedad reproductiva, de que una variable con distribución χ^2 y n grados de libertad, más otra con la misma distribución y m grados de libertad tiene distribución χ^2 con m+n grados de libertad.

1.2.2. Inferencia sobre la media.

1.2.2.1. INTERVALOS DE CONFIANZA PARA LA MEDIA POBLACIONAL.

Observando: $P[-z_\alpha < t < z_\alpha] = 1-\alpha$, podemos reescribir:

$$P[-z < \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} < z] = P[\overline{X} - z\sigma_{\overline{X}} < \mu < \overline{X} + z\sigma_{\overline{X}}] = 1 - \alpha$$

Esto se puede interpretar diciendo que esos límites aleatorios encierran la media poblacional un $(1-\alpha)$% de las veces, de modo que un par dado la encierran con una confianza del $1-\alpha$. A ese par de valores determinado se le conoce como los límites de confianza para la media.

Ejemplo 1.8 (Cont.) Un intervalo de confianza para la media de P1 en el ejemplo 1.8 será: $6{,}63 \pm 1{,}96\ (0{,}81/\sqrt{57})$ es decir $P[6{,}42 < \mu < 6{,}84] = 0{,}95$ o sea que la media poblacional estará entre 6,42 y 6,84 con un 95% de probabilidad.

Tamaño de la muestra. Tomando la expresión anterior, y llamando d a la diferencia entre la media muestral y la media poblacional $d = |\overline{X} - \mu|$, podemos escribir: $z = \frac{d}{s/\sqrt{n}}$, de donde despejamos $n \geq \frac{z^2\sigma^2}{d^2}$. De modo que para obtener una precisión (es decir una diferencia máxima entre la media y su estimación) d, la muestra tiene que tener un tamaño mínimo dado por la expresión anterior. Nótese que si el cálculo proporciona un valor fraccionario (caso frecuente en la práctica) se tiene que utilizar el número entero inmediato mayor para asegurar la precisión deseada.

Intervalo de confianza para μ en caso de σ^2 desconocida. Cuando la varianza es desconocida la fórmula anterior no se puede utilizar pero una expresión adecuada es:

$$P[-t < \frac{\overline{X} - \mu}{s_{\overline{X}}} < t] = P[\overline{X} - ts_{\overline{X}} < \mu < \overline{X} + ts_{\overline{X}}] = 1 - \alpha$$

de modo que la única diferencia está en que en este caso se debe utilizar la variable t en lugar de la z.

Tamaño de muestra en caso de varianza desconocida. Del mismo modo la expresión para el cálculo del tamaño de muestra mínimo se debe ajustar al uso de la variable t. Pero como en este caso necesitamos saber el tamaño de la muestra para definir los grados de libertad de la t a utilizar, debe recurrirse a un proceso iterativo.

1.2.2.2.PRUEBA DE HIPOTESIS SOBRE LA MEDIA. Si se plantea el supuesto (=hipótesis) de que la media de la población tiene un valor dado y si este cae en el intervalo de confianza para la media poblacional, se puede aceptar la hipótesis. Si el valor hipotético de la media poblacional no está entre los valores posibles (intervalo de confianza de la media poblacional) se rechaza la hipótesis.

Prueba de hipótesis simple contra una alternativa simple. La mecánica usual de una prueba de hipótesis $\mu = \mu_0$ versus $\mu = \mu_1$ se muestra como sigue. Supongamos que tenemos que decidir entre:

H_0: $\mu = 60$

H_1: $\mu = 80$

Figura 1.11.

sabiendo que ambas poblaciones tienen varianza igual a 334, con muestras de tamaño 4.

Notemos que para valores menores a 70 es mas probable que la muestra provenga de la población A mientras que para valores mayores a 70 es mas probable que venga de la población B. Por lo tanto podemos tomar como criterio de decisión: si la $\overline{X}$ es menor a 75 decidimos que la población es A y si $\overline{X}$ es mayor a 70 decidimos que la población es B. Se pueden cometer dos errores:

		Realidad	
		Ho es verdadera	Ho es falsa
Decisión	Rechazo Ho	Error I	Decisión correcta
	Acepto Ho	Decisión correcta	Error II

A la probabilidad de cometer error de tipo I se le llama *nivel de significación* y se simboliza con α. A la probabilidad de cometer error de tipo II se la simboliza como ß, y a 1 - ß se le conoce como la *potencia* de la prueba de hipótesis. Notemos que en todo este proceso se tomó como base de trabajo la hipótesis de que $\mu=60$. A la hipótesis de trabajo se la denomina *hipótesis nula* o de nulidad, mientras que a la opcional se la denomina *hipótesis alternativa*. **Prueba de hipótesis simple contra una alternativa compuesta unilateral.** Es más común que la hipótesis alternativa no sea simple sino compuesta[3], por ejemplo decidir entre $\mu=60$ y $\mu>60$. En este tipo de hipótesis no se puede conocer ß; la prueba tiene una potencia que dependerá del verdadero valor de μ, el cual es desconocido. Por lo tanto se hace lo siguiente: Si $\overline{X}$ es mayor que 75,03 decidimos que la población es B y si es menor decidimos que la población es A. Si μ es 60 la probabilidad de que una media muestral, proveniente de una muestra aleatoria, sea mayor de 75,03 es menor de 0,05. En base a esto podemos proponer como criterio de decisión: si la muestra tiene media mayor a 75,03 rechazaremos que provenga de la población con media $\mu=60$, con una probabilidad de error del 0,05. Como ya dijimos, el error de tipo II tiene una probabilidad de ocurrencia desconocida. **Prueba de hipótesis simple contra una alternativa compuesta bilateral.** La situación más general es: H_0: $\mu = 60$ vs H_1: $\mu \neq 60$, entonces la región crítica tendrá dos partes.

[3] A las hipótesis de alternativa simple se les llama problemas de clasificación.

PROCEDIMIENTO USUAL PARA UNA PRUEBA DE HIPÓTESIS.

Paso 1. Plantear las hipótesis.

H_0: $\mu = 10$

H_A: $\mu > 10$

El problema de la alternativa unilateral se puede plantear en los siguientes términos: i) algunos dicen que la alternativa depende de lo que el investigador quiere probar, se pone en la alternativa lo que se quiere probar; ii) otros dicen que la prueba de hipótesis es una cosa o la otra, por tanto, solo podemos hacerla unilateral cuando sabemos que no puede ser bilateral. En la práctica afecta la región crítica.

Paso 2. Elegir el nivel de significación. Los niveles de significación más usuales son α = 0,05 y 0,01, algunas veces se usan 0,10 o 0,001

Paso 3. Elegir la variable pívot. Un panorama es:

	Una	Dos	Tres o mas
Medias	z o t[1]	z o t[1]	F anova
Varianzas	χ^2	F	Test de Bartlett
Proporciones	z	z o χ^2	χ^2

[1]Que la variable pívot sea z o t depende de si conocemos la varianza o no. Algunos autores dicen que si la muestra es grande es como conocer la varianza, pero el problema se transforma en determinar cuando se considera que una muestra es grande. Entonces sigue la discusión con los que consideran muestra grande mayor de 30 observaciones, pero el valor de t para 30 grados de libertad no es igual a z.

Paso 4. Determinar la región crítica.

La región crítica va a depender de: i) la hipótesis alternativa, ii) el nivel de significación y iii) la variable pívot.

Paso 5. Hacer los cálculos.

Paso 6. Tomar la decisión. Se rechaza o no la hipótesis nula. Rechazar la hipótesis nula en el corto plazo es aceptar la alternativa, pero también puede ser falta de potencia de la prueba. Si rechazamos la H_0 debemos decir "con probabilidad de error tipo I de 0,005", pero si no la rechazamos podemos no conocer la probabilidad de error (que sería b).

Comentarios sobre este esquema. Hay gente que critica este esquema diciendo que esquematiza al alumno, no caigan en eso.

p-value. El software moderno no da directamente el valor p para un valor calculado. Si p es menor que 0,05 entonces es significativo a ese nivel. Algunos autores dicen proceder de esta manera afecta el nivel de significación.

Intervalos de confianza y prueba de hipótesis. Muchas veces la mecánica de ambos procedimientos es muy semejante. La principal diferencia es que en la prueba de hipótesis se toma una decisión, ahí es donde aparecen los errores I y II. El que no decide nada no se equivoca. Tiene que existir una concordancia entre la prueba de hipótesis y el intervalo de confianza, pero no siempre es total. Harville y otros argumentan que el intervalo de confianza es más general.

1.2.3.Inferencia sobre varianzas

UNA VARIANZA. Al igual que con respecto a la media se pueden realizar inferencias sobre la varianza de una población. En esto tiene importancia fundamental la propiedad: $\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2=\sum_{i=1}^{n}\left(\frac{X-\overline{X}}{\sigma}\right)^2+n\left(\frac{\overline{X}-\mu}{\sigma}\right)^2$

Estimadores de la varianza. Como $E(S^2)=[(n-1)/n]\sigma^2$ la varianza muestral es un estimador sesgado de la poblacional por lo que se propone el nuevo estimador (*) que es insesgado:

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

Intervalos de Confianza para σ^2 . En la siguiente reformulación de la variable χ^2 se puede construir intervalos de confianza para σ^2 (Recordemos que $\Sigma x^2=\sum(X-\overline{X})^2$).

$$P\left[\chi^2_I<\chi^2<\chi^2_S\right]=P\left[\chi^2_{\alpha/2}<\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\sigma^2}<\chi^2_{1-\alpha/2}\right]=P\left[\frac{\sum(X-\overline{X})^2}{\chi^2_{(1-\alpha/2)}}<\sigma^2<\frac{\sum(X-\overline{X})^2}{\chi^2_{\alpha/2}}\right]=1-\alpha.$$

Ejemplo 1.9. Supongamos que estamos estudiando los tratamientos cuyos datos se proporcionan abajo.

Tratamiento	1	2
	62	163
	86	208
	117	154
	125	154
	132	183
		212
		169
Totales	522	1243
Medias	104.4	177.57

Las sumas de cuadrados son: SC = SC sin corregir - C. Para el tratamiento 1 tenemos $57978 - 522^2/5 = 3481{,}20$ por lo tanto la estimación de la varianza es: 3481,20/4 = 870,30. Un intervalo de confianza para la varianza del tratamiento 1 es:

$P[\,3481{,}20/11{,}14 < \sigma^2 < 3481{,}20/0{,}484] = P[312{,}50 < \sigma^2 < 7192{,}56] = 0{,}95$

Similarmente para el tratamiento 2 tenemos: $SC = 224259 - 1243^2/7 = 3537{,}71$; con lo que la estimación de la varianza es: 3537,71/7 = 589,62: $P[251{,}44 < \sigma^2 < 2093{,}32] = 0{,}95$

Pruebas de Hipótesis Ejemplo 1.15 (Cont.) Supongamos que queremos probar la hipótesis de que el tratamiento 1 proviene de una población con varianza 500. Tenemos H_0: $\sigma^2 = 500$, contra H_1: $\sigma^2 \neq 500$. Entonces tomando como variable pívot

$$P[\chi^2_{\alpha/2} < \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{\sigma^2} < \chi^2_{1-\alpha/2}]$$

PRUEBA DE HOMOGENEIDAD DE VARIANZAS. El cociente de dos estimaciones de la varianza de una población tiene una distribución F: $F = \frac{\hat{\sigma}^2_1}{\hat{\sigma}^2_2}$. La distribución F tiene dos parámetros que son los grados de libertad del numerador y los del denominador.

Ejemplo 1.9 (Cont.) Si queremos comparar las varianzas de los dos tratamientos, tenemos: F = 870,30 / 589,62 = 1,48; como el valor de tablas para la F con 4 grados de libertad en el numerador y 6 en el denominador es: F(4,6) = 4,53 se concluye que el cociente de varianzas no es significativo, es decir que a los efectos prácticos tomamos las varianzas como iguales.

Intervalos de confianza para un cociente de varianzas. En genética interesa el cociente de varianzas por medio de un intervalo de confianza. La situación es del siguiente tipo: $P\{0,2 < F < 2,1\} =$

$$P\left\{0,2 < \frac{\hat{\sigma}^2_1 / \sigma^2_1}{\hat{\sigma}^2_2 / \sigma^2_2} < 2,1\right\} = P\left\{0,2 < \frac{\hat{\sigma}^2_1}{\hat{\sigma}^2_2}\frac{\sigma^2_2}{\sigma^2_1} < 2,1\right\} = P\left\{0,2\frac{\hat{\sigma}^2_1}{\hat{\sigma}^2_2} < \frac{\sigma^2_2}{\sigma^2_1} < 2,1\frac{\hat{\sigma}^2_1}{\hat{\sigma}^2_2}\right\}$$

1.2.4.Contraste de Medias

MUESTRAS NO INDEPENDIENTES: OBSERVACIONES APAREADAS. Si deseamos comparar las medias de observaciones apareadas el problema se transforma en uno de una sola muestra usando la variable diferencia: $d_i = X_i - Y_i$. La media de las diferencias es la diferencia de las medias: $\bar{d} = \bar{X} - \bar{Y}$ y la varianza de las diferencias es:

$s_L^2 = s_{\bar{X}}^2 + s_{\bar{Y}}^2 - 2Cov[\bar{X}, \bar{Y}] = \dfrac{\sum_{i=1}^{m} \left(d_i - \bar{d}\right)^2}{m-1}$. El problema se redujo a una sola muestra de diferencias y se analiza como tal.

Ejemplo 1.10. Analizaremos la diferencia entre las medias de dos muestras dadas abajo.

Muestra 1	Muestra 2	d	d^2
161,30	149,64	11,66	135,96
148,26	163,30	-15,04	226,20
142,99	152,68	- 9,69	93,90
184,47	161,64	22,83	521,21
146,69	157,69	-11,00	121,00
164,11	146,08	18,03	325,08
162,31	170,04	- 7,73	59,75
171,22	173,27	- 2,05	4,20
170,08	146,70	23,38	546,62
161,27	157,89		

$t_{(7)} = \dfrac{\bar{d} - \mu_d}{\sigma_{\bar{d}}} = = 0.65197$ (ns) ya que:

$\hat{\sigma}_{\bar{d}} = \sqrt{\dfrac{\sum_{i=1}^{m} \left(d_i - \bar{d}\right)^2}{m(m-1)}} = \sqrt{\dfrac{\sum d_i^2 - \left(\sum d_i\right)^2}{m(m-1)}} = [2033.922-(30,39)^2/9]/9x8 = \sqrt{}\ 26.82369$

CONTRASTE DE 2 MEDIAS INDEPENDIENTES. Generalmente en estadística no interesa tanto el efecto de un tratamiento como la comparación de efectos de dos o más tratamientos. En esta sección estudiaremos las comparaciones o contrastes entre medias. Para realizar este tipo de comparación nos valdremos de la propiedad reproductiva de la distribución normal que nos dice que toda función lineal de variables normales es normal. Se aplica como vemos:

$Y_1 \sim N(\mu_1 ; \sigma_1^2) \rightarrow Y \sim N(\mu_1 ; \sigma_{Y1}) |$

$> Y_1 - Y_2 \sim N(\mu_1 - \mu_2 ; \sigma_{Y1-Y2})$

$Y_2 \sim N(\mu_2 ; \sigma_2^2) \quad Y_2 \sim N(\mu_2 ; \sigma_{Y2}) |$

En el caso que las muestras sean independientes (es decir que la covarianza es cero) y si suponemos que tienen igual varianza poblacional:

$$s_L^2 = s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \qquad \text{donde:} \quad s_p^2 = \frac{\sum_{i=1}^{n1} x^2 + \sum_{j=1}^{n2} y^2}{n_1 + n_2 - 2} = \frac{s_1^2 + s_2^2}{2}$$

la última igualdad vale si las muestras tienen igual tamaño. Si $n_1 = n_2$, entonces $\sigma_{\hat{L}}^2 = \frac{2\sigma^2}{n}$. Si, por el contrario, $n_1 \neq n_2$, entonces $\frac{1}{n} = \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$ la media armónica representa bien al promedio de los tamaños muestrales.

Ejemplo 1.9 (Cont.) Supongamos que deseamos comparar las medias de los dos tratamientos mencionados antes, el tratamiento 1 y el 2. Tenemos: t =(104.4 - 177.5714)/

ANALISIS DE LA VARIANZA. El análisis de varianza es una técnica introducida por Fisher que sirve, entre otras cosas, para probar la hipótesis de que varias medias son iguales. Esto nos proporciona una segunda manera de realizar la prueba de diferencia de medias. Descomponemos la varianza de los datos del siguiente modo:

$$SC = \sum_{i=1}^{n}\sum_{j=1}^{k}\left(Y_{ij} - \overline{Y}_{..}\right)^2 = \sum_{i=1}\sum\left(Y_{ij} - \overline{Y}_{.j}\right)^2 + \sum\sum\left(\overline{Y}_{.j} - \overline{Y}_{..}\right)^2 = SCE + SCT$$

La varianza total de los datos se descompuso en dos partes: una es la varianza dentro de las muestras (al no tener causa aparente de variación solo decimos que es error experimental) y otra atribuible a la diferencia entre las muestras o tratamientos. De modo que si las diferencias entre los tratamientos son solamente debidas al azar las dos variaciones son del mismo orden y su cociente vale mas o menos 1. Se busca en tablas F si las relaciones obtenidas son aceptables como cercanas a 1 o no. Este procedimiento tiene la diferencia sobre el de Student en que es aplicable a mas de dos medias simultáneamente. El resultado se expresa generalmente en un cuadro de análisis de varianza:

Fuente Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	F
Tratamientos		3		
Error o residuo		16		
TOTAL		19		

Ejemplo 1.11. Ejemplo de análisis de varianza

Tratamiento	1	2	3	4
	62	163	60	137
	86	208	62	137
	117	154	72	159
	125	154	75	132
	132	183	52	126
Totales	522	1243	321	949
Medias	104.4	177.57	64.2	135.57

$$SC\,\text{sin corr} = \sum\sum Y_{ij}^2 \quad sSC\,\text{sin corr} = \sum\sum Y_{ij}^2 \quad SCTrt = \frac{\sum T_j^2}{n} - C$$

Con estos comentarios, por supuesto, no agotamos todo lo que hay para decir acerca del análisis de varianza, una de las técnicas mas usadas e importantes en análisis de datos. En secciones posteriores (3.2 por ejemplo) desarrollaremos otras visiones y otras aplicaciones del análisis de varianza.

1.3. ANALISIS DE VARIABLES CATEGORICAS.

1.3.1.Una variable discreta.

VARIABLES CUALITATIVAS. Ya comentamos que las variables que no se miden son cualitativas. Por ejemplo la presencia de una enfermedad o que una vaca tenga o no ternero. Una variable Bernoulli es una variable que tiene dos resultados posibles, generalmente uno se considera éxito y el otro fracaso, o se simbolizan con 1 y 0. Por ejemplo, tirar una moneda es un experimento de Bernoulli. Una variable cualitativa que tenga muchos resultados posibles siempre se puede dicotomizar. Por ejemplo el estado civil puede ser casado, soltero, viudo, divorciado, unión libre. Pero lo podemos categorizar en "casados" o "no". Los valores múltiples darán origen a distribuciones multinomiales y los dicotómicos a la binomial.

BINOMIAL. Si un experimento de Bernoulli se repite n veces y la probabilidad de éxito no cambia, la suma de éxitos tiene una distribución binomial. $P[X = x] = \binom{n}{x} p^x (1-p)^{n-x}$. Por ejemplo, tirar 5 monedas constituye un experimento donde el número de caras sigue una distribución binomial pues en cada moneda la probabilidad es la misma.

Media y varianza de la Binomial. La media y la varianza de una distribución binomial con parámetros n y p, que se simboliza con B(n,p), es: $E(X) = np$ y $V(X) = np(1-p)$. La distribución binomial está asociada a experimentos de muestreo con repetición o muestreo de poblaciones infinitas (casos en que la probabilidad de éxito no cambia).

POISSON. Si la probabilidad de éxito está dada por la siguiente expresión: $P[X = x] = \frac{e^{-\lambda}\lambda^x}{x!}$ la distribución se conoce como de Poisson, por el nombre del autor que la introdujo. El parámetro λ es la media y la varianza simultáneamente de la distribución. La distribución de Poisson se conoce como distribución de los sucesos raros, consideraremos suceso raro a uno que tenga una probabilidad de ocurrencia menor a 0,10 para simplificar.

HIPERGEOMÉTRICA. La distribución hipergeométrica se usa en el muestreo sin reposición de poblaciones finitas. Si llamamos N al tamaño de la población, n al tamaño de la muestra y p=A/N a la probabilidad de que la primera extracción tenga una determinada característica, tenemos que la probabilidad de éxito está dada por la siguiente expresión: $P[X = x] = \frac{C_x^A C_{n-x}^{N-A}}{C_n^N}$ la distribución se conoce como Hipergeométrica. Notemos que la distribución hipergeométrica tiene tres parámetros N, n y A[4]. La media de la distribución es np y la varianza es $np(1-p)\left(\frac{N-n}{N-1}\right)$. Al factor $\left(\frac{N-n}{N-1}\right)$ se le conoce como factor de corrección por población finita. Si la población es infinita vale 1.

RELACIONES ENTRE LAS DISTRIBUCIONES DISCRETAS. Las variables discretas mencionadas se relacionan entre si a través del siguiente cuadro.

Bernoulli	p constante BINOMIAL	p < 0,1 se aproxima por POISSON
		p > 0,1 se aproxima por NORMAL
	p no constante: HIPERGEOMETRICA	

Las aproximaciones se usan para valores grandes de n, digamos mayores a 50.

Distribución	Parámetros	Media	Varianza
Bernoulli	p	p	p(1-p)
Binomial	n,p	np	np(1-p)
Poisson	λ	λ	λ
Hipergeométrica	p	np	np(1-p)(N-n)/(N-1)

[4] Si consideramos que P=A/N se puede decir que el parámetro es P, la proporción de éxito en la primer sacada.

1.3.2. Estudio de proporciones.

Se ha comentado antes que las variables pueden ser cuantitativas o cualitativas también llamadas atributos (o categorías o clases). Decíamos que las variables se miden y los atributos se cuentan. Por esta razón el análisis de atributos muchas veces se denomina "análisis de conteos". En un estudio de atributos (pero no necesariamente ahí) se necesita en realidad estudiar las proporciones (o porcentajes) de observaciones con una determinada característica.

Ejemplo 1.12. Si nos interesa saber que porcentaje de los estudiantes de la Universidad de Uruguay son mujeres disponemos de dos caminos: un estudio exhaustivo (censo) o un muestreo. Para estimar por muestreo el porcentaje de estudiantes que son mujeres, se toma una muestra al azar de estudiantes y se cuenta cuantos son mujeres. Supongamos que se tomaron 11 personas al azar y hay 6 mujeres: el mejor estimador del porcentaje de mujeres es 6/11. De este modo, casi intuitivamente tenemos un estimador puntual de la proporción poblacional. Este estimador se obtuvo por el método de los momentos: el estimador del parámetro es la correspondiente estadística.

DISTRIBUCIÓN APROXIMADA DE PROPORCIONES. Habíamos comentado que un característica deseable en un estimador era conocer su distribución. También habíamos analizado la idea de aproximar la binomial por la normal. La variable "número de casos" tiene una distribución binomial que se puede aproximar por una normal. Es frecuente considerar entonces que las proporciones $\hat{p}$ [5] se distribuyen normalmente con media en la proporción poblacional P y varianza PQ/n, donde Q=1-P, lo que se representa: $\hat{p} \sim N(P; PQ/n)$.

INTERVALOS DE CONFIANZA PARA PROPORCIONES. Para los intervalos de confianza se presentan dos posibilidades: un intervalo aproximado usando p en lugar de P

$$\hat{p} \pm z\sqrt{\frac{PQ}{n}}$$

Supongamos que queremos hacer un intervalo de 95% de confianza para la proporción de mujeres entre los estudiantes de la Universidad de Uruguay con los datos manejados anteriormente:

$p = \hat{P} = 6/11 = 0{,}545$ por lo tanto $\hat{Q} = 1 - 0{,}545 = 0{,}455$. El intervalo de confianza queda: $0{,}545 \pm 1{,}96\sqrt{\frac{(0{,}545)(0{,}455)}{11}}$ o sea: $0{,}545 \pm (1{,}96)(0{,}150)$ resultando (0,251; 0,839). Es decir que la verdadera proporción de mujeres entre los estudiantes de la Universidad de Uruguay está entre 0,251 y 0,839 con un 95% de confianza.

Si observamos que PQ en la fórmula anterior depende de la P desconocida, vemos que el intervalo de confianza es aproximado. Un intervalo exacto sería más complejo. Por ejemplo, Collett (1986) y Mood & Graybill (1986) sugieren usar el siguiente sistema:

$$p_L = \frac{6}{6 + (6)(2{,}69)} \text{ y } p_S = \frac{7}{7 + \frac{5}{2{,}85}} = 0{,}80$$

[5] Usaremos indistintamente $\hat{p}$ "p-gorro" o p minúscula para la proporción muestral.

PRUEBA DE HIPOTESIS SOBRE PROPORCIONES. Supongamos que interesa probar la hipótesis de que la proporción de mujeres en la Universidad es del 50%. Según el procedimiento usual hacemos lo siguiente:

1) Planteo de las hipótesis. H_0: $P = 0,50$

H_A: $P \neq 0,50$

2) Elección del nivel de significación. Elegimos $\alpha = 0,05$ por ejemplo

3) Determinación de la variable pivot. Elegimos la variable: $z = \dfrac{p - P}{\sqrt{\dfrac{PQ}{n}}}$

4) Determinamos la región crítica. Para el α=0,05 es |z|=1,96. O sea que se rechazaran los valores de z que sean mayores que 1,96 o menores que -1,96

5) Cálculo de valores $z = \dfrac{0,545 - 0,500}{\sqrt{\dfrac{(0,500)(0,500)}{11}}} = 0,045/\ 0,151 = 0,298$

6) Toma de decisión. Como 0,298 no pertenece a la región crítica no se rechaza la hipótesis nula. Bien puede ser cierto que la mitad de los estudiantes de la Universidad sean mujeres de acuerdo a la información que nos proporciona la muestra. Notemos que la discrepancia entre los valores observados y los postulados es mínima: es decir que en un total de 11 personas 6 sean mujeres es lo mas cerca de la mitad que se puede pedir. Por lo tanto es fácil ver aún sin consultar a tablas que la hipótesis no se rechazaría.

TAMAÑO DE MUESTRA PARA ESTUDIO DE PROPORCIONES. Una objeción que se puede hacer a la situación anterior es que la muestra puede ser muy chica para detectar una discrepancia con la hipótesis. Otra manera de decirlo es que el intervalo de confianza es muy amplio es decir poco preciso. Entonces se nos puede decir que interesa calcular el tamaño de muestra necesario para estimar la proporción con un margen de error por ejemplo del 0,10: $z\sqrt{\dfrac{PQ}{n}} = 0,10$ por lo tanto $n = \dfrac{z^2 PQ}{d^2}$ lo que es una manera adaptada de la forma dada en la sección 1.2.2. Notemos que necesitamos conocer P para calcular el tamaño de muestra, lo que no tenemos. El razonamiento mas comúnmente seguido es que el máximo de PQ se da cuando P=Q=0,5 y PQ=0,25 por lo que en el peor de los casos: $n = z^2 / 4d^2$. En el presente caso, con γ=0,95 z=1,96 y d=0,10: $n = 1,96^2 / (4*0,10) = 96,04$. Por lo tanto el tamaño mínimo que cumple con ese requisito es 97 observaciones.

***Estudio de Proporciones con Muestras Chicas*.** Se puede ver que el numerador y el denominador no son independientes por lo tanto no tiene sentido plantearse una prueba t para proporciones. Eso, a su vez, implica que no existe un estudio de proporciones para muestras chicas con los procedimientos que estamos viendo. Recordemos que muchos autores consideran muestras chicas a las que son mayores a 30 observaciones, mientras que otros llevan a 100 el límite para considerar grande a la muestra. (Ver sección 1.3.5).

DIFERENCIA DE PROPORCIONES. Si las proporciones se distribuyen normal la diferencia también lo hace, es decir: La diferencia de proporciones se distribuye normal con media en la diferencia de proporciones poblacional y con varianza que es la suma de las varianzas:

$$(\hat{p}_1 - \hat{p}_2) \sim N\left(P_1 - P_2; \sigma_{\hat{p}_1 - \hat{p}_2}\right) \text{ donde } \sigma^2_{\hat{p}_1 - \hat{p}_2} = PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Prueba de Hipótesis sobre Diferencia de Proporciones. Supongamos que tenemos la siguiente situación de prueba de hipótesis.

Ejemplo 1.13. En dos facultades se obtuvieron apoyo para una determinada iniciativa en las siguientes proporciones: en la facultad A 14 apoyaron en 25 encuestados, en la facultad B 3 apoyan la iniciativa en 30 personas encuestadas. Queremos saber si la diferencia entre facultades es significativa al 0,05. Los pasos son:

1) Definir las hipótesis. H_0: $P_1 - P_2 = 0$

H_A: $P_1 - P_2 \neq 0$

2) Elegir el nivel de significación, por ejemplo tomamos α=0,05

3) Elegir la variable, $$z = \frac{(\hat{p}_1 - \hat{p}_2) - (P_1 - P_2)}{\sigma_{\hat{p}_1 - \hat{p}_2}}$$

4) Definir el valor crítico y la región crítica. Para el nivel de significación 0,05 el valor es 1,96; por lo tanto la región crítica es el conjunto de valores mayores que 1,96 y menores de -1,96

5) Hacer los cálculos $$z = \frac{\left(\frac{14}{25} - \frac{3}{30}\right) - 0}{\sqrt{\left(\frac{14+3}{25+30}\right)\left(1 - \frac{14+3}{25+30}\right)\left(\frac{1}{25} + \frac{1}{30}\right)}} = \frac{0{,}56 - 0{,}10}{\sqrt{(0{,}309)(0{,}691)(0{,}073)}} = 3{,}68$$

6. Tomar la decisión. Se rechaza la hipótesis nula, las dos proporciones son diferentes.

Intervalos de Confianza para Diferencia de Proporciones. La construcción de un intervalo de confianza para la diferencia de proporciones se hace como siempre, por ejemplo para el 95% de confianza: $(P_1 - P_2) \pm 1{,}96\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$ en el ejemplo:

$\left(\frac{14}{25} - \frac{3}{30}\right) \pm 1{,}96\sqrt{\frac{0{,}56 * 0{,}44}{25} + \frac{0{,}10 * 0{,}90}{30}}$ = 0,46 ± 0,22. O sea que el intervalo de confianza es: 0,24 – 0,68.

Tamaño de Muestra. Para determinar el tamaño de muestra se sigue una metodología igual a la mostrada antes. Suponiendo igual tamaño de muestra n_1=n_2=n, y suponiendo P=Q=0,50 tenemos: $n \geq \frac{2z^2PQ}{d^2} = \frac{2}{d^2}$

Ejemplo 1.14. Un fabricante desea estimar la diferencia en el porcentaje de piezas defectuosas entre dos procesos de producción de fusibles con una precisión de 0,06 y una probabilidad de 0,95. Cuantos fusibles debe elegir de cada proceso?

1.3.3.Cuadros de contingencia.

La distribución chi-cuadrado proporciona otro modo muy conocido de estudiar la diferencia entre proporciones.

Ejemplo 1.15. Un grupo de 300 estudiantes de ambos sexos fueron consultados si preferían matemáticas, ciencias sociales o humanísticas. La tabla siguiente presenta los resultados:

Sexo	Matemáticas	C. Sociales	Humanidades	Total
Mujeres	35	72	71	178
Varones	37	41	44	122
Total	72	113	115	300

La filosofía básica consiste en calcular los valores esperados si los dos criterios fueran independientes. Por tanto, el enfoque consiste en calcular un valor esperado para cada celda del siguiente modo: la probabilidad de que un encuestado sea mujer esta dada por 178/300 la probabilidad de que a una persona tenga preferencia por las matemáticas es 72/300. Estas probabilidades se conocen como marginales, ya que se calculan a partir de los márgenes de la tabla. Por lo tanto la probabilidad de que una persona al azar sea mujer y le guste las matemáticas es:

P[mujer y guste matemáticas] = P[mujer].P[guste matemáticas] = $\left(\frac{178}{300}\right)\left(\frac{72}{300}\right)$ =

Por lo tanto el numero esperado de mujeres es la probabilidad de que sea mujer por el numero de personas encuestadas (el tamaño de la muestra): Número esperado de mujeres que gustan matemáticas (178*72)/300. Similarmente, numero esperado de hombres que gustan matemáticas =122*72/300

Si la diferencia entre lo observado y lo esperado es grande el supuesto de independencia no se cumple por lo que los dos sexos tienen diferente porcentaje de preferencia por las materias.

Luego tenemos diferentes opciones de variable pívot, la mas usada es la siguiente (estadística de Pearson): $\chi^2 = \sum \frac{(o-e)^2}{e}$. Se considera que la distribución de esta estadística es aproximadamente χ^2 en una serie de condiciones, especialmente si todos los esperados son mayores de 5. Los grados de libertad de χ^2 son el número de filas menos 1 por el número de columnas menos 1 simbolizado por (f-1)(c-1).

Otras estadísticas que se pueden utilizar son la de Neyman o el test de razón de verosimilitudes (Freeman, 1987, página 39).

Se puede ver que la metodología presentada se puede aplicar a la prueba de hipótesis de diferencia entre dos o mas de dos proporciones.

Equivalencia entre χ^2 y la prueba z. Ya habíamos dicho que la relación entre ambas era muy estrecha: una normal es una χ^2 con un solo grado de libertad o una χ^2 es una suma de normales elevadas al cuadrado. Por lo tanto no debe sorprender que ambas pruebas sean equivalentes en la presente situación.

Ejemplo 1.16. En la tesis de Joaquín Azanza se estudiaron vacas a la sombra y al sol para medir su efecto sobre la preñez. Se obtuvieron los siguientes datos:

	Preñadas	Vacías	
Sombra	8	6	14
Sol	5	10	15
	13	16	29

Por chi cuadrado, calculamos los esperados así:

	Preñadas	Vacías	Totales
Sombra	6,28	7,72	14
Sol	6,72	8,28	15
Total	13	16	29

		Observado	Esperado	Desvíos	
Sombra	Preñadas	8	6,28	1,72	0,471
	Vacías	6	7,72	-1,72	0,383
Sol	Preñadas	5	6,72	-1,72	0,440
	Vacías	10	8,28	1,72	0,357
		29	29	0	1,652

Por tanto el chi cuadado es: 1,65 lo que es no significativo.

Por z.

$$z = \frac{\frac{8}{14} - \frac{5}{15} - 0}{\sqrt{\left(\frac{13}{29}\right)\left(\frac{16}{29}\right)\left(\frac{1}{14} + \frac{1}{15}\right)}}$$

Una observación adicional es que el método de χ^2 no proporciona lo que se conoce como capacidad de separar medias: es decir que si detectamos diferencias entre materias eso implica "no todas las materias tienen igual grado de preferencia" pero no sabemos si son todas diferente de todas o alguna en particular difiere de las demás.

1.3.4.Ajuste De Modelos

La distribución χ^2 tiene un uso muy difundido en los llamados problemas de bondad de ajuste a modelos. Los modelos mas comunes, son de ajuste a una determinada distribución teórica como binomial, normal o Poisson, pero puede haber otros casos, por ejemplo a una teoría genética. Notemos que la cantidad (o-e) mide la discrepancia entre lo observado y lo postulado por el modelo. Lo menores que sean esas discrepancias lo menor que resulta el χ^2. Se eleva al cuadrado para que los desvíos no se anulen y se divide por lo esperado para relativizar el resultado, es decir que si dos estudios tienen distinto número de observaciones no influya. Por esas razones las pruebas de χ^2 de este tipo son siempre a una sola cola. Estas hipótesis ("un dado es correcto", los "factores son independientes", etc.) no afirman valores acerca de un parámetro.

BINOMIAL. Un ejemplo de cómo hacer un ajuste a una distribución binomial (Ejemplo 1.15) se puede realizar con los siguientes datos:

Número de Hijos	Número de familias	Probabilidad	Número esperado	Desvios^2/ esperado.
0	8	$C_0^5(0,456)^0(0,544)^5$		
1	16	$C_0^5(0,456)^0(0,544)^5$		
2	38	$C_0^5(0,456)^0(0,544)^5$		
3	22	$C_0^5(0,456)^0(0,544)^5$		
4	10	$C_0^5(0,456)^0(0,544)^5$		
5	6			

La tarea consiste en calcular los valores que se esperarían si la distribución fuera binomial exacta y compararlos con los observados. Como la media era 2,28=np y n=5, se dedujo que p=2,28/5= 0,456 y por tanto q=0544. Para calcular los esperados
Los cálculos se resumen en la siguiente tabla:

Xi	o	Prob	e	(o-e)2/e
0	8	0,048	4,764	2,198
1	16	0,200	19,968	0,788
2	38	0,335	33,475	0,612
3	22	0,281	28,060	1,309
4	10	0,118	11,761	0,264
5	6	0,020	1,972	8,231
	100	1,000	100,000	13,401

NORMAL Ejemplo 2. Ajústense los siguientes datos a una distribución normal.

Límites de la clase	Observados	Esperados	$(o-e)^2/e$	Probabilidad
135-145	1			
145-155	0			
155-165	0			
165-175	0			
175-185	2			
185-195	5			
195-205	7			
205-215	7			
215-225	9			
225-235	10			
235-245	7			
245-255	5			
255-265	3			
265-275	2			
275-285	0			
285-295	1			
295-305	1			

Una de las características de este tipo de problemas es la laboriosidad, el trabajo que dan y el tiempo que consumen. La media de los datos es: 109,82 y la desviación estándar es: 22,22025. Con esos valores calculamos la probabilidad de que se encuentren en una determinada clase si la distribución fuera normal y comparamos con los datos observados.

		esp	obs			
0,001	0,003	0,164	1	0,836	0,698	4,253
0,002	0,001	0,084	0	-0,084	0,007	0,084
0,006	0,004	0,233	0	-0,233	0,054	0,233
0,015	0,009	0,569	0	-0,569	0,324	0,569
0,036	0,020	1,224	2	0,776	0,602	0,492
0,074	0,039	2,320	5	2,680	7,184	3,097
0,139	0,065	3,871	7	3,129	9,793	2,530
0,234	0,095	5,687	7	1,313	1,723	0,303
0,356	0,123	7,359	9	1,641	2,692	0,366
0,496	0,140	8,386	10	1,614	2,607	0,311
0,636	0,140	8,414	7	-1,414	2,000	0,238
0,760	0,124	7,435	5	-2,435	5,931	0,798
0,857	0,096	5,786	3	-2,786	7,761	1,341
0,923	0,066	3,965	2	-1,965	3,860	0,974
0,963	0,040	2,392	0	-2,392	5,724	2,392
0,984	0,021	1,271	1	-0,271	0,074	0,058
0,994	0,010	0,595	1	0,405	0,164	0,276
0,998	0,004	0,245	0	-0,245	0,060	0,245
	0,997	60	60			18,56

KOLMOGOROFF Y OTRAS PRUEBAS DE ADHERENCIA. El ajuste de una serie de datos a una distribución normal se hizo por la prueba χ^2, pero no es la única ni la mejor. Otras opciones son la prueba de Kolmogoroff, la prueba de Shapiro Wilk, etc. El SAS usa esta última con la idea que es la mejor.

PRUEBA EXACTA DE FISHER. La χ^2 no es la única prueba que se le puede aplicar a un cuadro de contingencia para docimar el modelo de independencia. Otra opción es la llamada "prueba exacta de Fisher" que consiste en calcular la probabilidad de que valores iguales a los observados o más extremos que esos ocurran por azar. Generalmente se utilizaban cuando las tablas son pequeñas pero con la actual disponibilidad de software se pueden utilizar para tablas grandes.

DISEÑO DE EXPERIMENTOS CON RESPUESTA CUALITATIVA. CASO CONTROL vs COHORTES. El cuadro de contingencia puede aparecer en dos tipos de situaciones: por ejemplo el investigador tiene un grupo de vacas que han quedado preñadas y otro que no quedaron preñadas. Entonces estudia que factores pueden haber incidido en ese resultado, por ejemplo la alimentación. El chi cuadrado significativo no le asegura que la alimentación es la razón por la cual los animales han quedado preñados[6]. Esta situación se llama de caso-control. La otra situación es cuando el investigador organiza dos grupos de animales, con baja y alta alimentación (dos cohortes) y luego del periodo de parición analiza los datos. Si el chi cuadrado le dio significativo el poder como evidencia es muy superior. Por tanto, las cuentas son iguales, pero la lógica experimental es diferente. Este tipo de análisis se ha desarrollado con seres humanos (en el área médica) pero se aplica a la agronomía.

Algunas notas sobre las distribuciones. Si partimos de la binomial, cuando n crece y np permanece constante la distribución se aproxima por Poisson.

Si tenemos una binomial y n crece la distribución se aproxima por normal, especialmente si p está en la cercanía de ½.

Por esos motivos, las técnicas presentadas en la sección 1.3.2 se basan en la aproximación de la binomial por la normal. Recordemos que insistimos en que no se usa la distribución t para el estudio de proporciones. Cuando usamos la distribución chi cuadrado en la sección 1.3.3 también hacemos uso de la aproximación basada en la normal. Un análisis más complejo estaría basado en la distribución multinomial. Los autores que desarrollan este pensamiento se plantean diferentes posibilidades: i) los totales de columna son fijos (la distribución es Poisson, ii)

GLIM. Por último destaquemos los métodos modernos de modelación de situaciones de variable categórica, conocidos como Modelos Lineales Generalizados.

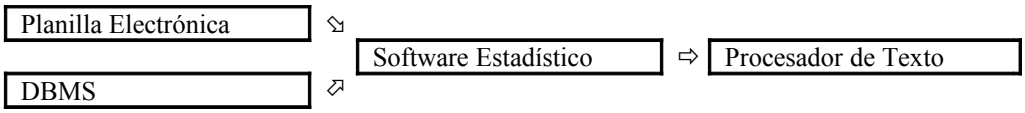
[6] Puede haber otros factores confundidos.

1.4.INTRODUCCION AL SOFTWARE ESTADISTICO

1.4.1. Visión general del software estadístico actual.

ESTADÍSTICA Y PROCESAMIENTO DE DATOS. En el pasado dentro de Estadística se estudiaba todo lo concerniente a análisis de datos. Con la actual disponibilidad de computadores, se presenta una tendencia a unir la problemática del análisis de datos con Computación mas que con Estadística. Por ejemplo el investigador que tiene datos y no sabe que hacer con ellos recibe el consejo "lo primero es entrarlos al computador" ([7]). Veremos en esta sección algunos elementos de procesamiento de datos y de estadística.

SOFTWARE Y ESTADISTICA. El computador tiene dos elementos: el hardware y el software. Actualmente lo que interesa es el software. El procesamiento de los datos se puede diagramar del siguiente modo:



En ese proceso los investigadores deben utilizar los siguientes elementos de software:

i. Planillas electrónicas, como el Lotus 1-2-3 o el Excel.

ii. Data Base Management System (DBMS), por ejemplo dBase III o Access.

iii. Paquetes estadísticos, como SAS, SPSS, Minitab, Genstat o Systat.

iv. Procesador de texto, puede ser el WordStar, WordPerfect o Word for Windows.

Las planillas electrónicas, son excelentes para entrar los datos y hacerles el procesamiento preliminar pre-estadístico, como graficar, ordenar los datos, etc. También son muy eficientes para pasar los datos a otros paquetes. El análisis que nosotros llamamos pre-estadístico, muchas veces es tan o más valioso que el propiamente estadístico y este último resulta para muchos una especie de confirmación de tendencias observadas en los datos. Los DBMS, son especialmente útiles para el almacenamiento y procesamiento primario de grandes volúmenes de datos. Los paquetes estadísticos actuales como el SAS son de gran calidad y divulgación. El procesador de texto es una herramienta de gran utilidad para el investigador, ya que ningún experimento está terminado hasta que se hace un informe sobre él. Una característica de interés es la comunicación entre los diferentes elementos de software que permiten al investigador considerable ahorro de esfuerzo en la elaboración de su informe.

SOFTWARE ESTADISTICO. El software estadístico moderno es excelente. Algunos de los productos que se comercializan a nivel mundial se detallan a continuación.

1 SAS - Statistical Analysis System. Creado en Carolina del Norte, USA. Utilidad general. Uno de los mejores y con gran distribución.

2 SPSS - Statistical Package for the Social Sciences. Creado para uso en ciencias sociales, pero actualmente de utilidad general. Hay versión en español.

3 BMDP - Bio Medical Package. Creado en California, USA. Finalidad de uso biológico y médico.

4 Minitab. Creado en Pennsilvania, USA. Tiene fama de software didáctico y fácil de usar.

5 S-Plus. Creado por ATT, es muy reciente y poderoso, sobre todo en el área de modelos lineales generalizados.

6 SYSTAT - System for Statistics. De origen USA, desarrollado por una empresa privada.

7 GLIM - Generalized Interactive Linear Models. Creado por la Royal Statistical Society de Inglaterra para modelos lineales generalizados.

8 REML - Residual Maximum Likelihood. Programa de uso en agronomía y biología

9 Otros: Harvey, M-STAT (U. de Minnesota) para uso en agronomía, PEST, etc.

---

[7] Tal vez la conversación siga así: "Y luego? No sé, supongo que consultar con alguien que esté en computación".

1.4.2. El sistema SAS.

SAS quiere decir Statistical Analysis System y, como el nombre lo dice, es un sistema para análisis de datos, con un enfoque estadístico. Haremos aquí una breve descripción de algunos aspectos del sistema. Un programa SAS ("SAS job") es un conjunto de sentencias SAS, una sentencia SAS siempre termina en un punto y coma. Todo programa SAS tiene dos partes: el DATA y los PROCs. Eso quiere decir que los datos se preparan con el DATA y se analizan con el correspondiente procedimiento (PROC).

UN PROGRAMA SAS. Como un ejemplo inicial de programa SAS presentamos[8] el análisis de los datos del ejemplo 1.1. El programa es el siguiente:

```
DATA uno1;
INPUT peso @@;
CARDS;
234  225  234  225  234  204  225  231  245  202
213  222  231  245  193  202  213  222  229  243
254  193  202  213  220  229  243  254  193  200
211  218  227  243  254  265  184  191  197  211
216  227  240  250  263  274  145  177  188  197
209  216  227  236  247  256  272  288  304  210
PROC UNIVARIATE;
RUN;
```

Notemos que cada sentencia SAS termina con punto y coma. Identifiquemos tres partes en el programa:

1. Preparación de los datos

DATA uno1; indica que el conjunto de datos recibe el nombre "uno1"

INPUT peso @@; indica que la variable que se le proporciona (en esta caso hay una sola) es peso. El símbolo @@ indica que hay varios datos en cada linea de programa.

CARDS; indica que terminaron las explicaciones y comienzan los datos.

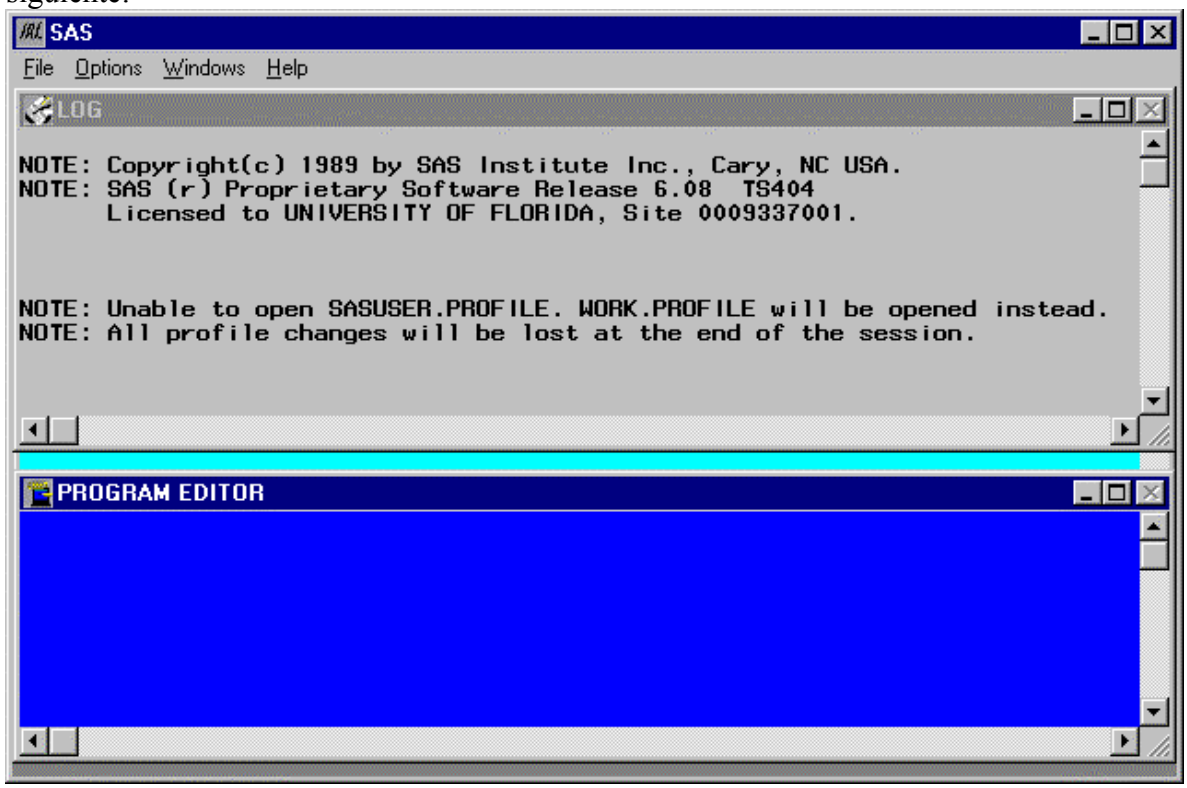
2. Datos. Luego de CARDS vienen los datos. Luego de los datos, a veces se pone un punto y coma aislado indicando que se terminó con la parte de datos.

3. Uso de un procedimiento de SAS. Acá usamos PROC UNIVARIATE; . Los PROC son, como decíamos antes, el centro de la capacidad de SAS. Acá le ordenamos al sistema lo que queremos hacer con los datos. En este caso es el procedimiento de descripción univariada de datos. Como se puede ver en la salida, este procedimiento proporciona extensa información acerca de los datos y se usa especialmente con grandes bancos de datos que no fueron trabajados "a mano" (es decir con el Lotus).

RUN; es la orden de proceder a correr el programa.

[8] Los comandos de SAS (que deben ser escritos textualmente así) los escribimos en mayúscula, los valores opcionales en minúsculas.

EJECUTANDO EL PROGRAMA EN EL SAS. Supongamos que tenemos el programa en un archivo llamado Ejemplo1.SAS. En las versiones modernas de SAS se invoca el sistema clickando sobre el botón correspondiente y el SAS presenta una pantalla como la siguiente:



Entramos en menú de archivos (files) abrimos el archivo con el programa y luego presionamos F8 para ejecutar el programa. La salida estará visible en la ventana output. Para guardar la salida, desde la ventana de output guardamos en forma de archivo. Ese archivo se puede editar, imprimir, etc. Para salir del SAS se presiona en el botón que está en el extremo superior derecho (X) como se hace con todo programa Windows.

La salida es:

EJEMPLO 1 DEL CAPITULO 1 - UNIVARIATE PROCEDURE

Variable=PESO

Moments

N	60	Sum Wgts	60
Mean	225.2667	Sum	13516
Std Dev	28.13157	Variance	791.3853
Skewness	0.166249	Kurtosis	0.811475
USS	3091396	CSS	46691.73
CV	12.48812	Std Mean	3.63177
T:Mean=0	62.02668	Prob>|T|	0.0001
Sgn Rank	915	Prob>|S|	0.0001
Num ^= 0	60		

Quantiles(Def=5) [1.1.4.6]

100% Max	304	99%	304
75% Q3	243	95%	273
50% Med	225	90%	259.5
25% Q1	206.5	10%	193
0% Min	145	5%	186
		1%	145
Range	159		
Q3-Q1	36.5		
Mode	193		

Extremes

Lowest	Obs	Highest	Obs
145(	47)	265(	36)
177(	48)	272(	57)
184(	37)	274(	46)
188(	49)	288(	58)
191(	38)	304(	59)

Algunos elementos que merecen comentarios:

Moments		indica los momentos que el sistema proporciona.
N		el número de observaciones
Sum Wgts		suma de los pesos
Mean	[1.13]	la media
Sum		la suma
Std Dev	[1.1.4]	la desviación estándar
Variance	[1.1.4]	la varianza (el cuadrado del anterior)
Skewness		el coeficiente de asimetría[9],
Kurtosis	[1.1.5.3]	el coeficiente de curtosis o apuntamiento de la distribución.
USS	[1.1.4]	suma de cuadrados sin corregir
CSS	[1.1.4]	suma de cuadrados corregidas
CV	[1.1.4.5]	coeficiente de variación
Std Mean	[pg 38]	error estándar de la media
T:Mean=0	[1.3.2.2]	valor de t para probar que $\mu=0$
Prob>|T|		probabilidad que los resultados se deban al azar si $\mu=0$
Num ^= 0		el número de observaciones diferentes de 0

[9] no decir como algunos que esto es un coeficiente de sesgo

1.4.3. Visión más Detallada de Algunos Procedimientos SAS.

Vamos a analizar el SAS con más detalle. El estudio de SAS se puede seguir en el libro de Cody & Smith (1991) o en el manual de SAS. Básicamente para lidiar con el SAS hay que proporcionarles datos y decirles que hacer con ellos. SAS tiene muchas posibilidades, no obstante lo cual, SAS dice que el 90% de los usuarios usa el 10% de las cosas. De modo que no intentamos ni por cerca agotar las posibilidades. "Esto no es un libro de SAS, sino de Estadística".

Tipos de variables. [sección 1.1]. Los tipos de variables considerados por SAS son numéricas y alfanuméricas (también pueden ser llamados textos o strings). Las variables numéricas clases o continuas.
Una vez que el SAS tiene los datos hay que decirle que hacer con ellos, los procedimientos disponibles en SAS se abrevian PROCS.
Algunos procedimientos comunes se muestran abajo.

Descriptivas	Regresión	Análisis de Varianza	Atributos
MEANS	REG	TTEST	FREQ
SUMMARY	NLIN	ANOVA	CATMOD
UNIVARIATE	GLM	GLM	PROBIT
CORR	CALIS	NESTED	CORRESP
FREQ	LIFEREG	VARCOMP	GENMOD
	LOGISTIC	NPAR1WAY	LOGISTIC
Informes	PROBIT	PLAN	PRINQUAL
PRINT		GENMOD	
CHART	RSREG	LATTICE	
PLOT	RSQUARE	MIXED	
	STEPWISE		
	ORTHOREG		
	TRANSREG		

El PROC PRINT sirve para imprimir los datos en diferentes formas.
Los gráficos se hacen con PROC CHART y PROC PLOT. La descripción de datos con TABULATE, SUMMARY y UNIVARIATE, el único que calcula cuantiles. La descripción por promedios con MEANS y por correlaciones con CORR. Como dijimos en la sección 1.1.2.1, SAS divide a las variables en cualitativas y cuantitativas. Para el estudio de las primeras se dispone de PROC FREQ, CATMOD y PROBIT.
Para comparación de medias se dispone de TTEST, ANOVA, GLM y NESTED. Para métodos no paramétricos se usa el PROC NPAR1WAY. PLAN es útil en el diseño de experimentos (ver capitulo 3). Los contrastes se estudian por medio de una opción dentro del PROC GLM.
La estadística inferencial descripta en la sección 1.3, como intervalos de confianza y prueba de hipótesis para μ se hacen con el PROC TTEST.
La regresión, que veremos en el capitulo 2, se estudia con varios procedimientos: PROC REG, RSREG, RSQUARE, STEPWISE, NLIN y GLM.
Como analizar cada tipo de variable? En parte depende de la clasificación que se usemos. Freeman (19xx) distingue entre observaciones, variables y valores. Nosotros en la sección

3.1 distinguimos entre características y variables. En el manual de SAS se presenta la siguiente tabla que puede se de ayuda en el proceso de decidir que análisis hacerle a cada variable.

Nivel de Medición	Estadísticas Descriptivas	Tablas de Frecuencias	Gráficos de Barras	Análisis Exploratorio
Nominal		X	X	
Ordinal	X	X	X	
Intervalos	X	X	X	X
Racional	X	X	X	X

La X indica que el método es apropiado para ese nivel de medición. Se nota que la tabla de frecuencias y los gráficos de barra son aconsejados para todos los tipos de escala, las estadísticas descriptivas se adaptan a escalas de rango ordinal o superior. El análisis exploratorio de datos ("EDA") se adapta a escalas de intervalos o racional.
El ejemplo de parcelas apareadas se puede procesar así:

```
DATA uno16;
   INPUT  n p i ii total d;
   TITLE 'Ejemplo 1.16';
   CARDS;
    0      0  161.30  149.64  310.94   11.66
   60      0  148.26  163.30  311.56  -15.04
  120      0  142.99  152.68  295.67   -9.69
    0    -80  184.47  161.64  346.11   22.83
   60    -80  146.69  157.69  304.38  -11.00
  120    -80  164.11  146.08  310.19   18.03
    0   -160  162.31  170.04  332.35   -7.73
   60   -160  171.22  173.27  344.49   -2.05
  120   -160  170.08  146.70  316.78   23.38
PROC MEANS MEAN STDERR T PRT;     VAR d;   RUN;
```

La salida es:

Ejemplo 1.16 Analysis Variable : D

```
N Obs     Mean          Std Error         T         Prob>|T|
-----------------------------------------------------------------------
   9   319.1633333     5.9202684    53.9102812     0.0001
-----------------------------------------------------------------------
```

EL PROC ANOVA DE SAS. El programa básico para hacer análisis de varianza en SAS es el PROC ANOVA. La ventana de ayuda se muestra abajo.

DESCRIPCION DEL PROC ANOVA

PROC ANOVA lleva a cabo análisis de varianza para datos balanceados de una amplia variedad de diseños experimentales.

```
PROC ANOVA DATA= OUTSTAT=;
    CLASS variables;      Debe preceder la senencial MODEL
    MODEL dependiente=efectos /NOUNI INTERCEPT;
   MEANS efectos / BON DUNCAN GABRIEL REGWF REGWQ SCHEFFE
       SIDAK SMM|GT2 SNK T|LSD TUKEY ALPHA=p WALLER
     KRATIO= LINES CLM CLDIFF E=efectos;
     ABSORB variables;
     FREQ   variable;
     TEST   H=efectos  E=efectos;
     MANOVA H=efectos  E= efectos  M=ecuaciones  MNAMES=
          PREFIX= / PRINTH  PRINTE ORTH CANONICAL SUMMARY;
    REPEATED nombre-factores  niveles (nombre-niveles)
           CONTRAST (N)| POLINOMIAL|HELMERNT|MEAN
           (N)|PROFILE / NOM NOU PRINTM PRINTH
            PRINTRV PRINTE SUMMARY CANONICAL;
BY variables;
```

Por detalles, refierase a la
GUIA SAS/STAT PARA COMPUTADORES PERSONALES

Un programa para analisis de varianza es el siguiente:

```
DATA uno11;
INPUT raza peso @@;
CARDS;
1     62   1     86   1   117
1   125    1   132    4   136
2   163    2   208    2   154
2   154    2   183    2   212
3    60    3    62    3    72
3    75    3    52    2   169
4   137    4   137    4   159
4   132    4   126    4   122
PROC ANOVA;
  CLASS raza;
  MODEL peso=raza;
  MEANS raza/LSD LINES TUKEY;
  RUN;
```

con la salida SAS:

```
                    Analysis of Variance Procedure
                        Class Level Information
                       Class    Levels    Values
                       RAZA          4    1 2 3 4
                  Number of observations in data set = 24
SAS Analysis of Variance Procedure
Dependent Variable: PESO
                                  Sum of          Mean
Source               DF        Squares        Square   F Value     Pr > F
Model                 3     40682.5298    13560.8433     33.04     0.0001
Error                20      8209.4286      410.4714
Corrected Total      23     48891.9583

               R-Square           C.V.      Root MSE            PESO Mean
               0.832090       16.02116       20.2601            126.45833

Source               DF       Anova SS   Mean Square   F Value     Pr > F
RAZA                  3     40682.5298    13560.8433     33.04     0.0001
SAS Analysis of Variance Procedure - T tests (LSD) for variable: PESO
NOTE: This test controls the type I comparisonwise error rate not the
      experimentwise error rate.
Alpha= 0.05  df= 20  MSE= 410.4714
Critical Value of T= 2.09       Least Significant Difference= 24.746
WARNING: Cell sizes are not equal. Harmonic Mean of cell sizes= 5.833333
Means with the same letter are not significantly different.

T Grouping              Mean      N  RAZA
                        A            177.57      7  2
                        B            135.57      7  4
                        C            104.40      5  1
                        D             64.20      5  3
SAS Analysis of Variance Procedure
Tukey's Studentized Range (HSD) Test for variable: PESO
NOTE: This test controls the type I experimentwise error rate, but
      generally has a higher type II error rate than REGWQ.
Alpha= 0.05  df= 20  MSE= 410.4714
Critical Value of Studentized Range=3.958
Minimum Significant Difference= 33.204
WARNING: Cell sizes are not equal. Harmonic Mean of cell sizes= 5.833333
Means with the same letter are not significantly different.

            Tukey Grouping      Mean       N  RAZA
                  A            177.57      7  2
                  B            135.57      7  4
                  B            104.40      5  1
                  C             64.20      5  3
```

ANÁLISIS DE DATOS CATEGÓRICOS EN SAS.

La gran herramienta para el análisis de datos categóricos con el SAS es el Proc Freq.
Para las tablas de una sola vía los comandos son muy simples: PROC FREQ; TABLES a; RUN; produce una tabla de frecuencia dando los valores de A y la frecuencia de cada valor.

Tablas de dos criterios de clasificación o doble vía, también llamadas tablas de doble entrada, o de un modo general, cuadros de contingencia. Para obtener la tabla de dos variables se dan los nombres de las variables separados por un asterisco. Los valores de la primer variable van a las filas y los de la segunda a las columnas. Por ejemplo, PROC FREQ; TABLES a*b; RUN;

```
data uno15; do i=1 to 2; do j=1 to 3; input n @@; output; end; end;
datalines;
35      72      71
37      41      44
proc freq; tables i*j/chisq; weight n; run;
```

```
                    The FREQ Procedure
                     Table of i by j
          Frequency,
          Percent  ,
          Row Pct  ,
          Col Pct  ,      1,      2,      3,  Total
          ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
               1 ,     35 ,     72 ,     71 ,    178
                 ,  11.67 ,  24.00 ,  23.67 ,  59.33
                 ,  19.66 ,  40.45 ,  39.89 ,
                 ,  48.61 ,  63.72 ,  61.74 ,
          ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
               2 ,     37 ,     41 ,     44 ,    122
                 ,  12.33 ,  13.67 ,  14.67 ,  40.67
                 ,  30.33 ,  33.61 ,  36.07 ,
                 ,  51.39 ,  36.28 ,  38.26 ,
          ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
          Total        72      113      115      300
                    24.00    37.67    38.33   100.00

               Statistics for Table of i by j
        Statistic                     DF       Value      Prob
        ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
        Chi-Square                     2      4.6063    0.0999
        Likelihood Ratio Chi-Square    2      4.5538    0.1026
        Mantel-Haenszel Chi-Square     1      2.5119    0.1130
        Phi Coefficient                       0.1239
        Contingency Coefficient               0.1230
        Cramer's V                            0.1239
                    Sample Size = 300
```

Uno de los aspectos mas polémicos es como medir la asociación entre variables categóricas. La correlación es un concepto que fue desarrollado para variables cuantitativas, y los investigadores desean procedimientos similares para variables cualitativas. Se han propuesto diferentes medidas para esa asociación. Algunas de las comunes son: coeficiente phi, coeficiente de contingencia, indice de Kramer, indice gamma, t de Kendall, t de Stuart, indice lambda, etc. Índice relativo de riesgo, riesgo relativo ajustado, etc. El lector puede encontrar información sobre estos temas en: Manual de SAS y los libros de Freeman (1987) y Fienberg (1977)

1.4.3. Diferentes Formas de Entrada de Datos.

Concepto de base de datos. Una de esas maneras bastante usada es la mostrada en la parte izquierda de la siguiente tabla:

Datos como se presentan originalmente

TRATA	BLOQUE I	BLOQUE II	TOTAL
0-0	161.30	149.64	310.94
60-0	148.26	163.30	311.56
120-0	142.99	152.68	295.67
0-80	184.47	161.64	346.11
60-80	146.69	157.69	304.38
120-80	164.11	146.08	310.19
0-160	162.31	170.04	332.35
60-160	171.22	173.27	344.49
120-160	170.08	146.70	316.78

Base de datos como SAS.

TRATA	B	REND
0-0	1	161.3
60-0	1	148.2
120-0	1	142.9
0-80	1	184.4
60-80	1	146.6
120-80	1	164.1
0-160	1	162.3
60-160	1	171.2
120-160	1	170.0
0-0	2	149.6
60-0	2	163.3
120-0	2	152.6
0-80	2	161.6
60-80	2	157.6
120-80	2	146.0
0-160	2	170.0
60-160	2	173.2
120-160	2	146.7

Los datos están de esa forma porque es más fácil calcular la media del bloque en una planilla electrónica. Pero para el SAS necesitan estar en un formato diferente, acorde al concepto de base de datos, es decir con las variables en una columna y con los casos en una fila. Normalmente la planilla electrónica es el mejor sistema para obtener el arreglo deseado de los datos. Cuando de se termine de arreglarlos los datos quedarían en un formato como el que se muestra en la parte derecha de la tabla (tabla 1.8), ya prontos para ser utilizados por el programa estadístico.