PRIMER CONGRESO INTERNACIONAL

Y TERCER ENCUENTRO DEPARTAMENTAL

DE MATEMÁTICA EDUCATIVA

ANÁLISIS EXPLORATORIO DE DATOS

JAIRO ALFONSO CLAVIJO M
jclavijo@bunde.tolinet.com.co

IBAGUE, SEPTIEMBRE DE 2002

CONCEPTOS PRELIMINARES

Variable: ente estadístico que representa una magnitud que puede tomar diferentes valores. Edad, peso, estatura, número de hijos, longitud, duración, etc
Variable aleatoria: Aquella variable que toma valores que no pueden ser determinados con anticipación. Con frecuencia se sabe qué valores puede tomar mas no cuáles va a tomar. Ejemplo: la edad en años cumplidos de un estudiante de estadística seleccionado de una lista.

Las variables aleatorias y sus propiedades constituyen el principal objeto de estudio de la estadística. Esta ciencia no se ocupa de las variables determinísticas aunque sí las use en algunas situaciones.

La estadística obtiene información de diferentes maneras. Una de las formas más frecuentes de hacerlo es a través de **experimentos** que pueden ser diseñados (planificados) o no.

Experimento es cualquier acción que produzca un resultado medible. Por ejemplo: Seleccionar un individuo estudiante de un listado y medir su presión sanguínea. El resultado será un valor numérico que representa la presión y que puede ser observado y medido. Un experimento es aleatorio cuando sus resultados no son previsibles con antelación.

Un experimento planeado es aquel en el que, previo a la medición, se han planificado algunos pasos, como por ejemplo, cuántas unidades, considerar, en qué circunstancias, cuáles medir, cómo procesar los datos de las mediciones, etc Una parte de la estadística se ocupa del **diseño de experimentos**. Un experimento no planeado es aquel en el que simplemente se toma la información de una fuente que no obedece a un proceso de planificación previo. Por ejemplo: aplicación de un cuestionario de opinión a algunos miembros de una comunidad.

Todo experimento, planeado o no, involucra una o más variables aleatorias (es decir, es **univariado** o **multivariado**). Variables que son observadas en **varios** individuos los cuales constituyen las observaciones (también llamadas casos o individuos). Las variables son medidas en cada uno de los individuos, dando como resultado una **medición** o **dato.**

Los datos no son necesariamente números. También pueden ser cualidades o categorías de una lista previamente establecida, por ejemplo, sexo (M, F), grado de aceptación

(Mucho, Poco, Nada). Esto produce una primera clasificación de las variables en numéricas y categóricas (o nominales). Estas últimas a su vez, pueden ser nominales puras u ordinales. Sexo es una variable categórica pura mientras que grado de aceptación es variable categórica ordinal.

Las variables numéricas pueden ser continuas o discretas. Son **continuas** cuando sus valores pueden ser cualesquiera dentro de un intervalo. Por ejemplo, la altura en cm de una planta o la duración de una bombilla eléctrica. Variables **discretas** son aquellas que sólo pueden tomar valores de un conjunto finito (enumerable). Por ejemplo, el número de hijos en una familia. El número de estudiantes de un curso.

Los datos que se obtienen al observar y medir una variable, son entonces números (en el caso de variables numéricas) o símbolos que representan una categoría (en el caso de variables categóricas). Así, por ejemplo:

VARIABLE	**DATO O RESULTADO:**
Ingresos	1, 526,315.00 465,000.00 245,315
Número de Sillas en el aula	12 60
Sexo o género	M F Macho Hembra 1 2
Opinión acerca del aborto	De acuerdo En desacuerdo 2 1

Aunque las categorías de una variable categórica (nominal) sean representables mediante símbolos cualesquiera, lo usual es usar **códigos** más prácticos, económicos o eficientes que reemplacen a esos símbolos. Se dice entonces que una variable está *codificada*. Una variable puede ser *recodificada* cuando los códigos son modificados de alguna manera.

Es frecuente el uso de códigos numéricos para la codificación de variables nominales. Por ejemplo: en vez de *Masculino* usar 1, en vez de *Femenino* usar 2. Generalmente se usan los dígitos 1, 2, 3, ..., 9 como códigos y rara vez son necesarios códigos numéricos de dos dígitos.

Cuando se usen variables categóricas ordinales es conveniente y recomendable usar códigos numéricos consecutivos cuya magnitud esté en correspondencia con el orden de la categoría. Por ejemplo: En vez de *Mucho* usar 3, en vez de *Poco* usar 2 y en vez de *Nada* usar 1 (aunque se crea que es más conveniente usar 0 en vez de *Nada*, debemos evitar el uso del código 0 por otras razones, principalmente por el tratamiento de cálculo que tienen algunos paquetes de computador). Resulta evidente que aunque una variable categórica esté codificada numéricamente, con los datos que ella proporcione no es lícito hacer operaciones aritméticas: no tiene sentido, por ejemplo, calcular una media o una varianza.

Como se dijo antes, la información es el resultado que se obtiene en uno o más experimentos y que es expresable en datos, codificados mediante números u otros símbolos adecuados. Para que sea útil y perdurable, esta información debe ser almacenada de una manera organizada y de modo que sea fácilmente accesible. Antiguamente la información se almacenaba en papel donde seguramente estaba bien organizada pero resultaba poco accesible. Con la aparición del computador y los medios magnéticos asociados a él (cintas, discos, etc) es posible almacenar enormes cantidades de información en medios relativamente baratos y de muy alta accesabilidad. Este es el método más usado. Pero ello implica dar una estructura a los archivos de datos de manera que resulten fácilmente accesibles por los paquetes estadísticos.

Muchos paquetes estadísticos (SPSS, Minitab, SAS, SYSTAT, por ejemplo) tienen estructuras que les son propias y generalmente incompatibles entre sí aunque muchos de ellos tienen la posibilidad de transformar la estructura de otro en la suya propia. Es conveniente, sin embargo, utilizar una estructura universal para el almacenamiento de la información, estructura que es compartida por la gran mayoría de paquetes estadísticos y que tiene cierta ventajas adicionales. Es la siguiente:

Un archivo de datos es una gran matriz con la siguiente estructura:

CASOS	VARIABLE 1	VARIABLE 2	...	VARIABLE p
CASO 1	Dato 11	Dato 12	...	Dato 1p
CASO 2	Dato 21	Dato 22	...	Dato 2p
...	...	...	...	...
CASO n	Dato n1	Dato n2	...	Dato np

La zona sombreada es opcional y muchos paquetes estadísticos no la usan.

Generalmente se usan como separadores de los datos los espacios en blanco u otros símbolos como la coma o el slash (/). Se debe buscar que los datos estén alineados por la derecha y que no haya datos faltantes (MD o Missing Data) ya que esto ocasionaría problemas en el momento de procesar información. Existen procedimientos de **Imputación** de datos faltantes, es decir, procedimiento de "llenado de los huecos" cuando hay faltantes.

Aunque los archivos de datos pueden crearse mediante hojas electrónicas como EXCEL, (esto se hace por facilidad y rapidez) es recomendable que su almacenamiento se haga en formato ASCII (American Estándar Code for Interchange of Information) debido a que éste es un código universal (entendible por todos los paquetes) y poco dado a contener

virus informáticos. Un archivo en formato ASCII (pronúnciese "aski") puede ser creado con cualquier editor de texto plano, por ejemplo, EDIT (que viene en todos los computadores compatibles IBM), WordPad de Windows, EDITOR de ESM, El editor de Minitab/DOS, etc. Una manera muy rápida de hacerlo es mediante el uso de una hoja como Excel y su posterior conversión (exportación) a ASCII.

El siguiente ejemplo ilustra la estructura de un archivo de datos. Se trata de un archivo que contiene información (ficticia) sobre 150 fincas ubicadas en diferentes regiones y terrenos de Colombia, dedicadas a diferentes actividades de economía agropecuaria.

EJEMPLO:

El siguiente ejemplo, corresponde a una parte de los datos reales tomados por Leyder Lozano en el Laboratorio de Investigaciones en Parasitología de la Universidad del Tolima, como parte de su trabajo de grado. La información representa medidas morfométricas de tres especies de Rhodnius (Pitos, transmisores del mal de Chagas). En este taller se han considerado 6 de las variables medidas originalmente. Son ellas:

1. **C1** LONGITUD DE CABEZA Y CUELLO (En micras)
2. **C2** ANCHO DEL COLLAR
3. **C3** DISTANCIA ENTRE HUMEROS
4. **C4** LONGITUD DEL TORAX
5. **C5** SEXO. 1 = Hembras 2 = Machos.
6. **C6** PROCEDENCIA. 1 = U de los Andes. 2 = U del Tolima. 3 = Silvestres.

Como puede apreciarse, las cuatro primeras variables son numéricas y las dos últimas son categóricas.

Los análisis que pueden realizarse con ellas depende de su naturaleza: con las variables numéricas podemos realizar análisis que impliquen operaciones aritméticas como cálculo de promedios, varianzas, correlaciones, regresiones, etc. Con las variables categóricas podrían realizarse conteos de frecuencias, tablas de contingencia, pruebas de independencia, análisis de correspondencia, etc.

En este taller, se utilizará con mucha frecuencia el paquete estadístico **NCSS** en una versión de demostración, válida por 30 días y que sólo procesa 100 observaciones como máximo. NCSS es un excelente paquete estadístico de un costo moderado, cuya versión de demostración puede ser usada con el fin de ser evaluada y adquirida por los interesados. Se puede consultar las características del paquete en Internet.

Los datos mencionados son los siguientes:

```
4131.0 1526.8 4365.9 4941.0  1 1      4442.8 1680.7 4544.1 5305.5  2 2
4090.5 1494.4 4025.7 4600.8  1 1      4147.2 1563.3 4139.1 4884.3  2 2
4001.4 1478.2 3936.6 4503.6  1 1      4511.7 1733.4 4673.7 5483.7  2 2
4325.4 1514.7 4228.2 4908.6  1 1      4382.1 1595.7 4317.3 5062.5  2 2
4390.2 1603.8 4519.8 5143.5  1 1      4333.5 1749.6 4446.9 5394.6  2 2
4098.6 1547.1 4195.8 4811.4  1 1      3900.1 1506.6 3859.6 4455.0  2 2
3960.9 1543.0 4179.6 4730.4  1 1      4264.6 1636.2 4203.9 5001.7  2 2
4066.2 1466.1 4013.5 4536.0  1 1      4544.1 1753.6 4544.1 5265.0  2 2
4195.8 1640.2 4313.2 5175.9  1 1      4284.9 1609.6 4422.6 5260.9  2 2
4179.6 1539.0 4301.1 4949.1  1 1      4110.7 1636.2 4297.0 4981.5  2 2
4139.1 1563.3 4284.9 4839.7  1 1      4021.6 1531.0 4082.4 4746.6  2 2
3969.0 1591.6 4179.6 4783.0  1 1      4321.3 1717.2 4511.7 5321.7  2 2
4357.8 1660.5 4600.8 5159.7  1 1      4159.3 1628.1 4220.1 4876.2  2 2
3928.5 1498.5 4199.8 4892.4  1 1      3977.1 1571.4 4094.5 4730.4  2 2
3952.8 1506.6 4203.9 4779.0  1 1      4033.8 1611.9 4187.7 4908.6  2 2
4195.8 1410.3 4212.0 4657.5  2 1      4094.7 1459.8 4065.5 3108.7  1 3
4098.6 1474.2 3863.7 4426.6  2 1      4007.1 1459.8 4029.0 3077.6  1 3
4066.2 1482.3 4102.6 4872.1  2 1      4262.6 1583.8 4379.4 3241.1  1 3
4074.3 1474.2 3815.1 4560.8  2 1      4218.8 1459.8 4160.4 3178.8  1 3
4212.0 1389.1 4090.5 4803.3  2 1      3904.9 1459.8 4123.9 3311.2  1 3
3912.3 1486.3 3859.6 4450.9  2 1      3948.7 1459.8 4182.3 3217.7  1 3
3981.1 1489.5 3993.3 4564.3  2 1      4116.6 1547.3 4109.3 2890.6  1 3
3928.5 1482.3 3879.0 4633.2  2 1      4116.6 1459.8 4233.4 2960.7  1 3
4005.4 1481.2 4070.2 4665.6  2 1      3943.1 1401.4 3919.5 2875.0  1 3
4009.5 1492.4 3847.5 4422.6  2 1      4167.7 1547.3 4299.1 2773.7  1 3
3717.9 1482.3 3798.9 4390.2  2 1      4080.1 1459.8 4043.6 3015.2  1 3
3827.2 1421.5 3794.8 4394.2  2 1      4043.6 1437.9 3963.3 2945.1  1 3
3985.2 1474.2 4098.6 4746.6  2 1      4087.4 1474.3 3846.5 2890.6  1 3
3977.1 1530.9 4171.5 4779.0  2 1      4189.6 1532.7 4087.4 3030.9  1 3
4058.1 1510.6 4114.8 4860.0  2 1      4233.4 1547.3 4262.6 3139.8  1 3
4698.0 1822.5 4965.3 5532.3  1 2      4029.0 1459.8 3904.9 2750.3  2 3
4576.5 1741.5 4722.3 5475.6  1 2      3985.2 1474.3 3839.2 2516.5  2 3
4337.5 1644.3 4179.6 5013.9  1 2      3948.7 1459.8 4014.4 2991.9  2 3
4236.3 1636.2 4746.6 5499.9  1 2      4050.9 1401.4 3934.1 2477.6  2 3
4511.7 1709.1 4633.2 5508.0  1 2      4102.0 1518.1 3999.8 2695.7  2 3
4665.6 1696.9 4981.5 5815.8  1 2      4182.3 1547.3 4021.7 2758.1  2 3
4839.7 1717.2 4754.7 5629.5  1 2      4145.8 1488.9 3992.5 2306.2  2 3
4114.8 1660.5 4430.7 5354.1  1 2      3904.9 1408.7 3664.0 2890.6  2 3
4357.8 1717.2 4600.8 5386.5  1 2      4080.1 1518.1 3977.9 2882.8  2 3
4762.6 1826.5 5062.5 5888.7  1 2      4080.1 1481.6 3999.8 2882.8  2 3
4459.0 1684.8 4625.1 5410.8  1 2      4109.3 1503.5 4014.4 2781.5  2 3
4552.2 1749.6 4779.0 5572.8  1 2      4036.3 1540.0 3992.5 2968.5  2 3
5005.8 1741.5 4536.0 5443.2  1 2      4116.6 1510.8 3985.2 2859.4  2 3
4706.1 1773.9 4754.7 5544.4  1 2      4138.5 1525.4 3956.0 2804.8  2 3
4507.6 1798.2 4973.4 5726.7  1 2      4116.6 1547.3 3970.6 2758.1  2 3
```

El archivo de datos puede ser creado, como ya se dijo, con un editor de texto, con Excel o simplemente digitando la información dentro de la hoja que para tal fin presenta el programa NCSS.

Aunque aquí se presentó en dos grandes columnas de 45 observaciones cada una, el archivo debe ser escrito como una matriz de 90 filas (casos o individuos) por 6 columnas (variables).

ANALISIS DE DATOS

Con las variables categóricas, a nivel elemental, es poco lo que puede hacerse: se puede contar cuántas ocurrencias de cada modalidad se presentan, qué porcentaje representa cada modalidad y se pueden ilustrar estos resultados con algunos gráficos que ayudan a globalizar la información, como se ve enseguida. Un análisis un poco más profundo de este tipo de variables pretende medir el grado de dependencia de dos variables categóricas y la asociación que existe entre sus categorías o modalidades. Esto será tema de estudio más adelante.

Con las variables de tipo numérico es posible hacer más análisis a nivel elemental. Aparte de los conteos usuales de casos, uno de los análisis iniciales en cualquier estudio estadístico tiene por fin indagar sobre el comportamiento de los datos. Se quiere saber de una manera global si los datos representan una población simétrica, qué tan fuerte es el grado de dispersión, cómo es la forma de su distribución, Cuánto valen aproximadamente los estadísticos descriptivos más importantes (media y varianza), si existen o no valores extremos, etc. El conocimiento de esta información permite entrar en etapas más avanzadas del análisis con una "actitud" ante los datos.

El conjunto de técnicas que estudia los datos desde el punto de vista anterior es conocido como **análisis exploratorio de datos** (Exploratory Data Analysis o **EDA**). En estas notas veremos algunas de las técnicas más usuales para explorar datos numéricos. Pero antes tendremos que precisar algunos términos.

En primer lugar qué se entiende por **población**. Hemos dicho que una variable aleatoria numérica toma valores numéricos, que pueden ser continuos o discretos. Por ejemplo, puede ser el peso de una persona o puede ser el número de hermanos que ella tenga. En el caso del peso podríamos decir que la variable *puede* tomar valores entre 0 y 120 Kg. En el caso de los hermanos *puede* tomar valores enteros 0, 1, 2, ..., 20, por ejemplo. Nótese que la variable ***puede*** tomar esos valores. Esto no significa que los tome. Y los valores que asume la variable no son igualmente probables. Por ejemplo, quizás sea más probable que la variable NUMERO DE HERMANOS tome el valor 2 que el valor 10 y éste con más probabilidad que el valor 20. De igual manera, la variable PESO toma

valores entre 5 y 80 kg con más probabilidad que entre 90 y 120 Kg, al menos en un grupo de personas "normales".

Dada una variable aleatoria ***X***, se define la *población asociada* a ella como el conjunto de valores numéricos que ***X*** puede tomar. Nótese que, según esta definición, una población esta asociada a una variable. No es algo independiente de ella. Por otra parte, una población será discreta o continua, según como sea la variable asociada a ella. No se debe cometer el error (demasiado frecuente, por desgracia) de creer que la población está formada por un conjunto de personas, animales o cosas. Estos individuos son objetos de medición o de observación y, como tal, son portadores de los valores de una población pero no son la población estadística misma. Por esto, es que en el mismo grupo de individuos puede haber poblaciones diferentes. Por ejemplo, sus pesos y número de hermanos son dos poblaciones muy distintas observadas sobre los mismos individuos. En razón a la definición que se acaba de dar, con mucha frecuencia, consideramos población y variable como una dupla indisoluble y al hablar de cualquiera de ellas se estará hablando de la otra.

Se observa cómo, al decir que ***X*** toma valores en una población, algunos valores son tomados con más probabilidad que otros en la mayoría de los casos. El lenguaje común con frecuencia lo expresa así. Se dice, por ejemplo, que es más probable que un paisa tenga un número alto de hermanos que una persona de otro lugar de Colombia. O, por ejemplo, que es más probable que un norteamericano sea más alto que un colombiano. Excepciones hay, pero la regla general es válida en la mayoría de los casos.

Lo que se acaba de decir en el párrafo anterior da una idea de lo que se quiere mencionar cuando hablamos de la **distribución de probabilidad** de una variable aleatoria, o simplemente, para abreviar la expresión, cuando se habla de la **distribución de una población,** o de **la distribución de *X*.** Se entiende entonces como distribución de una variable aleatoria, *X*, la probabilidad de que *X* tome cada valor dentro de la población. La probabilidad es una medida de la aparición de uno o más números como valores de *X*. Comúnmente, la probabilidad se mide como una frecuencia y, como tal, puede medirse por medio de un porcentaje. Sin embargo en la práctica se mide como una fracción entre 0 y 1, correspondiente a un porcentaje. Por ejemplo, 0.2315 corresponde a 23.15%. De esta manera una probabilidad de cero (correspondiente a una frecuencia de 0%) expresa que un evento no sucederá con toda seguridad. Una probabilidad de 1 (correspondiente a una frecuencia de 100%) indica que un evento sucede con plena seguridad. Entre estos dos extremos se encuentran las demás medidas de probabilidad y, por ejemplo, una probabilidad de 0.3416 (correspondiente a una frecuencia de 34.16%) indica que el evento ocurre el 34.16% de las veces.

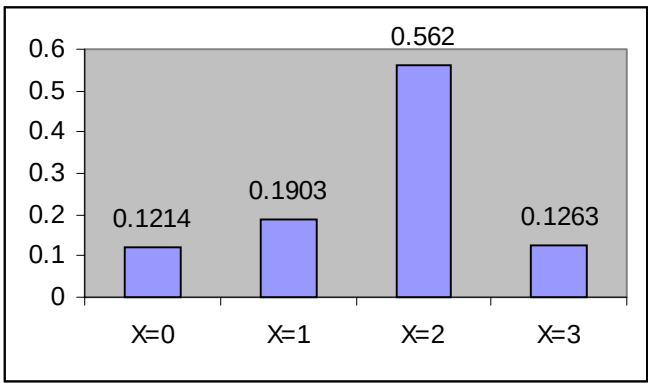
La probabilidad, aplicada a los valores que toma una variable aleatoria, mide la frecuencia con que un valor es asumido por una variable. Por ejemplo, si una variable aleatoria discreta, *X*, sólo puede tomar los valores 0, 1, 2, 3, al decir que P(*X*=2) = 0.5620,

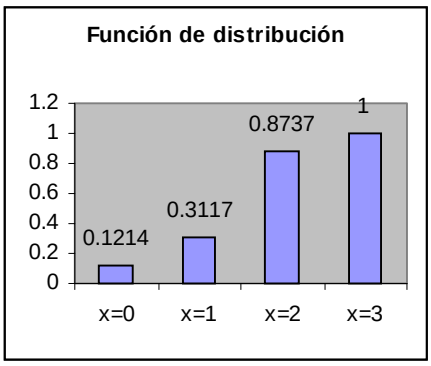
estamos afirmando que el valor 2 es asumido con una frecuencia de 56.2%. Es decir que si observamos dicha variable 500 veces, por ejemplo, ***se espera*** que 281 veces tome el valor 2, (ya que 281 = 0.562x500). Puesto que es seguro que la variable aleatoria toma uno de los cuatro valores 0, 1, 2, 3, la suma de las respectivas probabilidades debe ser 1. Esto es, $P(X=0) + P(X=1) + P(X=2) + P(X=3) = 1$. Se podría tener, a manera de ejemplo, lo siguiente:

$$P(X=0) = 0.1214$$
$$P(X=1) = 0.1903$$
$$P(X=2) = 0.5620$$
$$P(X=3) = 0.1263$$

De esta manera hemos descrito la **distribución (de probabilidad**) de la variable *X*.

Con frecuencia se representan los valores anteriores mediante barras cuyas alturas sean iguales a los valores de probabilidad y se tiene entonces una gráfica de la distribución de la variable discreta *X*. Igualmente es usual acumular los valores anteriores con lo cual se tiene la función de distribución de la variable. Algo como esto:





Para el caso de las variables aleatorias continuas, la situación es ligeramente más complicada, debido a que la variable puede tomar infinitos valores dentro de un intervalo. Por esta razón la probabilidad de que tome exactamente un valor es infinitamente pequeña y se toma igual a cero, sin que esto quiera decir que sea seguro que no tome ese valor. De ser así, la variable no tomaría ningún valor dentro del intervalo lo que, evidentemente es contradictorio. Esta es una de las tanta paradojas causadas por el infinito. Para evitar el problema, la probabilidad se mide por subintervalos. Es decir, hablamos por ejemplo, de la probabilidad de que la variable aleatoria *X* tome valores entre 50 y 80 (pensemos en la variable PESO), de que tome valores por debajo de 30, de que sea mayor a 100, etc. Lo que escribimos como: $P(50 \leq X \leq 80)$, $P(X<30)$, $P(X>100)$.

Sin embargo debe quedar claro que, por ejemplo, P(X=50)=0, P(X=100)=0, a pesar de que X puede tomar los valores 50 o 100

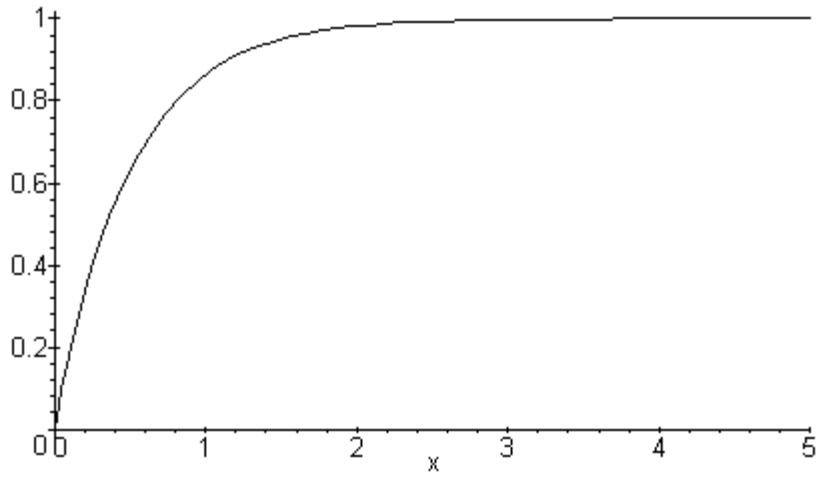
En el caso de una variable continua la probabilidad ya no puede darse mediante una tabla sino que se hace mediante una función F que proporcione la probabilidad de que X tome valores menores o iguales que un número arbitrario. Es decir para cada número real x se define $P(X \leq x) = F(x)$. Una tal función, se llama **función de distribución de X** o función acumulativa de probabilidad para X.

Por ejemplo, una variable aleatoria X, podría tener como función de distribución, la siguiente:

$$F(x) = \begin{cases} 1 - e^{-2x} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}$$

en cuyo caso, por ejemplo, $P(X < 2) = P(X \leq 2) = F(2) = 1 - e^{-4} = 0.9817$

La gráfica de esta función es la siguiente:



En esta gráfica se aprecia que cuando más grandes sean los valores de x mayor es la probabilidad de que X tome valores inferiores a x, sin que tal probabilidad exceda a 1.

Un estudio más completo y detallado de las funciones de distribución para variables aleatorias, será tema de un curso superior. Por ahora es suficiente conocer el concepto que

hemos esbozado. Pero se debe advertir que la estadística inferencial hace mucho uso de las distribuciones de probabilidad. En este curso volveremos sobre el tema más adelante.

Otro concepto importante es el de **muestra**. Una muestra es simplemente un subconjunto de la población. Es por tanto, un conjunto de valores, generalmente finito, que son extraídos de la población. Todo subconjunto es una muestra pero no todo subconjunto es una "buena" muestra. Una característica importante de una buena muestra es que sea aleatoria. Quiere esto decir que se debe haber extraído mediante un proceso que garantice que procede de *toda* la población y no sólo de ciertos sectores de ella. En un curso de muestreo se define con más precisión el concepto de ***muestra aleatoria***. El proceso de seleccionar una muestra se conoce con el nombre de *muestreo*.

Una pregunta natural es: ¿Por qué se hace muestreo en estadística? Hay varias respuestas a esta pregunta. Una de ellas es: por economía. Muchas veces resulta demasiado costoso o aún imposible examinar todos los valores de una población para obtener alguna información sobre ella. Piénsese, por ejemplo, cómo sería de costoso y aún imposible entrevistar a todos los habitantes de Ibagué. Otra respuesta es: Porque no hay otra solución. Esto sucede, por ejemplo, cuando el proceso de observación es destructivo: para medir la duración de un tipo de llantas no es posible acabar con toda la producción. Aún hay otra respuesta: por precisión. Muchas veces una población es tan grande que prácticamente ningún computador corriente podría albergar toda la información correspondiente a ella o procesarla sin producir errores de redondeo.

El estadístico **espera** que una buena muestra refleje las propiedades de la población de donde fue extraída. De esta manera espera poder inferir cómo es la población, examinando solamente la muestra. Una buena muestra debe dar información aproximada acerca de la forma de la distribución de probabilidad de la población, debe reflejar propiedades y características de dicha población, por ejemplo, simetría, valores más probables, valores atípicos, tendencias, etc y finalmente, una buena muestra debe producir valores aproximados de los correspondientes parámetros poblacionales (que son costantes desconocidas). En otras palabras: si nos fuera dado conocer el valor de la media poblacional, dicho valor debería ser muy próximo al que se obtiene promediando los valores de la muestra. Por esta razón se dice que la media muestral es una ***estimación*** de la media poblacional. Igualmente debe suceder con cualquier otro valor que se calcule usando los datos de la muestra. Cuando todo esto se puede garantizar en una muestra, se dice que dicha muestra es ***representativa*** de la población. Mejor dicho: una muestra representativa de una población es como un retrato de dicha población. Es como tener la población en miniatura para poderla examinar. Dice Sharon Lohr en su libro de muestreo (1999) que una buena muestra es como el pueblo de Grandview en la película *Magic Town*, el cual tenía exactamente las mismas características que todo Estados Unidos: exactamente la misma proporción de personas que votaban por los republicanos, la misma proporción de personas en la pobreza, la misma proporción de mecánicos de autos, etc.

Así pues bastaba -en la película- entrevistar a las personas de Grandview para saber cuál era la opinión de toda la Unión Americana. Una muestra representativa es entonces una versión a menor escala de la población.

En la práctica las muestras perfectas no existen. Sin embargo cuando una muestra es seleccionada atendiendo a las normas dictadas por un correcto muestreo, se obtienen muestra buenas, tanto más buenas cuanto más regular sea la población y más riguroso el método de muestreo. La selección de una buena muestra es, por tanto, un paso importantísimo antes de cualquier análisis estadístico. La selección de una muestra representativa de una población se hace atendiendo simultáneamente varias preguntas:
¿Cuántos elementos seleccionar? - (Tamaño de la muestra)
¿Cuáles elementos seleccionar? - (Principio de aleatoriedad)
¿Dónde (en qué parte de la población) seleccionar?
¿Cómo (con qué método) seleccionar?
¿Qué tanto error estamos dispuestos a admitir en las estimaciones?
¿Con qué grado de confiabilidad queremos hacer estimaciones?
¿Qué tan costoso resulta seleccionar la muestra y cuánto dinero poseemos para ello?

Como se ve, no es fácil obtener muestras representativas de una población (aunque en muchos trabajos se diga que se ha usado una de tales muestras). Aprender a hacerlo es lo que se estudia en los cursos de muestreo.

En este taller supondremos siempre que estamos en presencia de una muestra representativa de una población y no nos preocuparemos por saber cómo fue seleccionada. El objetivo que perseguimos es el de explorar los datos de la muestra para adquirir algún conocimiento acerca de la población. En esto el EDA es una de las herramientas más preciadas.

Cómo se dijo antes las muestras son siempre finitas. Así pues, podemos suponer que una muestra está conformada por n valores (números) que podemos enumerar así:

$$\text{Muestra} = \{y_1, y_2, \cdots, y_n\}$$

Dentro de estos valores podría haber algunos repetidos, razón por la cual a menudo se dan los diferentes valores que conforman la muestra y se dice cuál es su **frecuencia absoluta** u observada, esto es, cuántas veces aparece cada uno de ellos. De igual manera, se define la **frecuencia relativa** para cada observación como la frecuencia absoluta dividida entre *n*. Tanto la una como la otra se pueden ir acumulando frente a cada observación, obteniéndose las frecuencias absoluta acumulada y relativa acumulada. Se acostumbra presentar estos cuatro conceptos en una única tabla llamada **TABLA DE FRECUENCIAS**,

cuyo uso es importante y básico aunque a veces desconocido. Tales tablas están conformadas así:

Observación	**Frec. Abs.**	**Frec. Relat.**	**Frec.Abs.Acu**	**Frec.Rel.Ac**
x_i	f_i	$h_i = \frac{f_i}{n}$	$F_i = \sum_{k=1}^{i} f_k$	$H_i = \sum_{k=1}^{i} h_k$
x_1	f_1	h_1	F_1	H_1
x_2	f_2	h_2	F_2	H_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_r	f_r	h_r	F_r	H_r

En estas tablas son particularmente importantes la primera, segunda y última columnas como veremos en seguida.

A partir de lo anterior se definen los siguientes estadígrafos (expresiones calculadas con los valores muestrales):

a. La **media muestral**. Definida por cualquiera de las expresiones siguientes:

$$\bar{x} = \sum_{i=1}^{r} h_i x_i = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{\sum_{i=1}^{r} f_i x_i}{\sum_{i=1}^{r} f_i}$$

b. La **varianza muestral**. Definida por:

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

Esta expresión es una medida de la dispersión de los datos. Su raiz cuadrada se conoce con el nombre de **desviación estándar** o desviación típica.

c. El **p-ésimo momento central**. Definido para p=1, 2, 3, ... así:

$$m_p = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{x})^p = \sum_{i=1}^{r} h_i (x_i - \bar{x})^p$$

como puede observarse, la varianza es el segundo momento central.

Se debe hacer una observación al respecto de los momentos centrales. Con frecuencia estos valores se definen también tomando $n-1$ en cambio de n en el denominador de la primera sumatoria. Esto tiene su razón de ser cuando se está haciendo teoría de la estimación, porque, como puede probarse, tales expresiones constituyen estimaciones no sesgadas de los respectivos parámetros.

d. **La mediana**. Definida como aquel valor tal que el 50% de las observaciones son menores o iguales que él y el otro 50% son mayores o iguales que él. Naturalmente, si n es impar la mediana es la observación central. Si n es par la mediana es el promedio de las dos observaciones centrales.

e. **Los cuartiles**. Definidos como aquellos datos que dividen las observaciones en cuatro grupos tales que cada uno de ellos tiene el 25% de las observaciones. El primer cuartil, Q_1 separa el primer grupo del segundo. Por debajo de él hay un 25% de las observaciones. El segundo cuartil Q_2 es la mediana y el tercer cuartil Q_3 es tal que por debajo de él está el 75% de las observaciones.

f. **Los deciles**. Son valores tales que dividen la muestra en 10 partes iguales, siguiendo el mismo esquema de los cuartiles.

g. **Los percentiles**. Son valores que dividen la muestra en 100 partes iguales. La idea es similar a la de los cuartiles y los deciles. Los percentiles se denotan $P_1, P_2, \cdots, P_{100}$. Es claro que $Q_1 = P_{25}$ y $Q_3 = P_{75}$

h. **La(s) moda(s).** Se define una moda como una observación de máxima frecuencia absoluta. Puede haber una, más de una o ninguna. Esto último sucede cuando todas las observaciones tienen la misma frecuencia absoluta. Lo deseable es que una población sea *unimodal* (una sola moda) y esto debe reflejarse en la muestra.

i. **Coeficiente de Asimetría**. Definido como $a_3 = \frac{m_3}{s^3}$.

El coeficiente de asimetría mide la simetría general de la distribución

Este coeficiente es independiente de las unidades de medida y vale 0 en distribuciones simétricas. El signo de este coeficiente está en correspondencia con el sesgo de las distribuciones: es positivo en distribuciones cuya cola es cargada a la derecha (como los salarios, "muchos con poco y pocos con mucho") y es negativo en distribuciones cuya cola se carga hacia la izquierda.

j. **Coeficiente de curtosis**. Definido como $a_4 = \frac{m_4}{s^4}$.

Este coeficiente mide el apuntamiento o curtosis de una distribución. Se toma como patrón la distribución normal estándar (se verá formalmente después) en la que este coeficiente vale 3. Cuando una distribución tiene curtosis inferior a 3 se dice que es plana o platicúrtica. Cuando tiene curtosis superior a 3, se dice que es leptocúrtica o puntiaguda. Algunos paquetes como SAS, miden un coeficiente modificado: $a_4 - 3$, llamado ***exceso de curtosis***. Este último puede ser negativo.

k. **Rango y Rango intercuartílico**. Son medidas de dispersión definidas respectivamente como $R = y_{max} - y_{min}$ y $Q = Q_3 - Q_1$

Todas las cantidades definidas anteriormente, al ser calculadas en la muestra, constituyen estimaciones de los respectivos valores (parámetros) en la población.

A manera de ejemplo, examinaremos el comportamiento de la primera variable del archivos de datos dado anteriormente. Esta variable representa la longitud de la cabeza junto con el cuello de una muestra de 90 individuos. Los estadísticos descriptivos, calculados con ESM, pueden apreciarse en las páginas siguientes.

VALORES DE ESTADISTICOS PARA LA VARIABLE: CABEZA

1.	Número de observaciones:	N =	90
2.	Suma de observaciones:	äx =	376043.1999999999
3.	Suma de Cuadrados:	äxý =	1576298380
4.	Observación Máxima:	MAX =	5005.8
5.	Observación Mínima:	MIN =	3717.9
6.	MEDIA muestral:	m =	4178.257777777777
7.	Error estándar de la media:	Em =	25.21556209426596
8.	VARIANZA MV Muestral (GL=n): ...	Soý =	56588.38688395064
9.	VARIANZA INSESG.Pobl (GL=n-1):..	Sý =	57224.21145568042
10.	Desviación estándar muestral: ..	So =	237.8831370315068
11.	Desviación Estándar Poblacional:	S =	239.2158260978576
12.	Tercer Momento Central:	M3 =	15648786.56472517
13.	Cuarto Momento Central:	M4 =	13586344945.15262
14.	Coeficiente Asimetría:	A3 =	1.162491259252346
15.	Coeficiente Curtosis:	K =	4.242755534304991
16.	Coeficiente Variación:	CV =	5.725252936047556 %
17.	Coef. Aprox Normal (25G2):	C =	32.67095974492804
18.	Rango Muestral:	R =	1287.9
19.	Mediana Muestral:	Q2 =	4115.700000000001
20.	Primer cuartil (aprox)	Q1 =	4008.3
21.	Tercer cuartil (aprox)	Q3 =	4274.750000000001
22.	Rango Intercuart¡lico	Q3-Q1 =	266.4500000000007
23.	Moda(s):	4116.6	

TABLA DE FRECUENCIAS PARA LA VARIABLE: LCABEZA
(DATOS NO AGRUPADOS)

OBSERVACION	FREC.ABS	FREC.REL	FREC.ACUM	F.REL.ACM
3717.90000	1	0.011111	1	0.011111
3827.20000	1	0.011111	2	0.022222
3900.10000	1	0.011111	3	0.033333
3904.90000	2	0.022222	5	0.055556
3912.30000	1	0.011111	6	0.066667
3928.50000	2	0.022222	8	0.088889
3943.10000	1	0.011111	9	0.100000
3948.70000	2	0.022222	11	0.122222
3952.80000	1	0.011111	12	0.133333
3960.90000	1	0.011111	13	0.144444
3969.00000	1	0.011111	14	0.155556
3977.10000	2	0.022222	16	0.177778
3981.10000	1	0.011111	17	0.188889
3985.20000	2	0.022222	19	0.211111
4001.40000	1	0.011111	20	0.222222
4005.40000	1	0.011111	21	0.233333
4007.10000	1	0.011111	22	0.244444
4009.50000	1	0.011111	23	0.255556
4021.60000	1	0.011111	24	0.266667
4029.00000	1	0.011111	25	0.277778

4033.80000	1	0.011111	26	0.288889
4036.30000	1	0.011111	27	0.300000
4043.60000	1	0.011111	28	0.311111
4050.90000	1	0.011111	29	0.322222
4058.10000	1	0.011111	30	0.333333
4066.20000	2	0.022222	32	0.355556
4074.30000	1	0.011111	33	0.366667
4080.10000	3	0.033333	36	0.400000
4087.40000	1	0.011111	37	0.411111
4090.50000	1	0.011111	38	0.422222
4094.70000	1	0.011111	39	0.433333
4098.60000	2	0.022222	41	0.455556
4102.00000	1	0.011111	42	0.466667
4109.30000	1	0.011111	43	0.477778
4110.70000	1	0.011111	44	0.488889
4114.80000	1	0.011111	45	0.500000
4116.60000	4	0.044444	49	0.544444
4131.00000	1	0.011111	50	0.555556
4138.50000	1	0.011111	51	0.566667
4139.10000	1	0.011111	52	0.577778
4145.80000	1	0.011111	53	0.588889
4147.20000	1	0.011111	54	0.600000
4159.30000	1	0.011111	55	0.611111
4167.70000	1	0.011111	56	0.622222
4179.60000	1	0.011111	57	0.633333
4182.30000	1	0.011111	58	0.644444
4189.60000	1	0.011111	59	0.655556
4195.80000	2	0.022222	61	0.677778
4212.00000	1	0.011111	62	0.688889
4218.80000	1	0.011111	63	0.700000
4233.40000	1	0.011111	64	0.711111
4236.30000	1	0.011111	65	0.722222
4262.60000	1	0.011111	66	0.733333
4264.60000	1	0.011111	67	0.744444
4284.90000	1	0.011111	68	0.755556
4321.30000	1	0.011111	69	0.766667
4325.40000	1	0.011111	70	0.777778
4333.50000	1	0.011111	71	0.788889
4337.50000	1	0.011111	72	0.800000
4357.80000	2	0.022222	74	0.822222
4382.10000	1	0.011111	75	0.833333
4390.20000	1	0.011111	76	0.844444
4442.80000	1	0.011111	77	0.855556
4459.00000	1	0.011111	78	0.866667
4507.60000	1	0.011111	79	0.877778
4511.70000	2	0.022222	81	0.900000
4544.10000	1	0.011111	82	0.911111
4552.20000	1	0.011111	83	0.922222
4576.50000	1	0.011111	84	0.933333
4665.60000	1	0.011111	85	0.944444
4698.00000	1	0.011111	86	0.955556
4706.10000	1	0.011111	87	0.966667

4762.60000	1	0.011111	88	0.977778
4839.70000	1	0.011111	89	0.988889
5005.80000	1	0.011111	90	1.000000

TABLA DE FRECUENCIAS PARA DATOS AGRUPADOS EN 13 CLASES

INFER (-----] SUPER		MARCAS	F.ABS	F.RELAT	F.ACUM	FREL.ACU
3717.90	3816.97	3767.43	1	0.01110	1	0.01110
3816.97	3916.04	3866.50	5	0.05560	6	0.06670
3916.04	4015.11	3965.57	17	0.18890	23	0.25560
4015.11	4114.18	4064.64	21	0.23330	44	0.48890
4114.18	4213.25	4163.71	18	0.20000	62	0.68890
4213.25	4312.32	4262.78	6	0.06670	68	0.75560
4312.32	4411.38	4361.85	8	0.08890	76	0.84440
4411.38	4510.45	4460.92	3	0.03330	79	0.87780
4510.45	4609.52	4559.99	5	0.05560	84	0.93330
4609.52	4708.59	4659.06	3	0.03330	87	0.96670
4708.59	4807.66	4758.13	1	0.01110	88	0.97780
4807.66	4906.73	4857.20	1	0.01110	89	0.98890
4906.73	5005.80	4956.27	1	0.01110	90	1.00000

ESTADISTICOS APROXIMADOS SEGUN AGRUPAMIENTO ANTERIOR:

NUMERO DE OBSERVACIONES 90
MEDIA ESTIMADA: 4178.0215
VARIANZA MUESTRAL 56377.03652782389
VARIANZA POBL. ESTIM..... 57010.48637645112
DESVIACION TIPICA MUESTRAL ... 237.4384899880891
DESVIACION TIPICA POBLAC. 238.7686880150978

CLASE(S) MODAL(ES):
4 ... (4015.1 , 4114.17]

HISTOGRAMA DE DISTRIBUCION DE VARIABLE LCABEZA
(Agrupamiento en 13 clases)

MARCAS:
FREC.REL(%)

```
      3767.4 ³|                                                    1.11
      3866.5 ³|||||||||                                           5.56
      3965.6 ³|||||||||||||||||||||||||||||                     18.89
      4064.6 ³|||||||||||||||||||||||||||||||||||||||           23.33
      4163.7 ³|||||||||||||||||||||||||||||                     20.00
      4262.8 ³||||||||||||                                       6.67
      4361.9 ³|||||||||||||||                                    8.89
      4460.9 ³|||||                                              3.33
      4560.0 ³|||||||||                                          5.56
      4659.1 ³|||||                                              3.33
      4758.1 ³|                                                  1.11
      4857.2 ³|                                                  1.11
      4956.3 ³|                                                  1.11
```

ESM no calcula algunos estadísticos que dependen de percentiles pero es un buen ejercicio hacer los cálculos a mano.

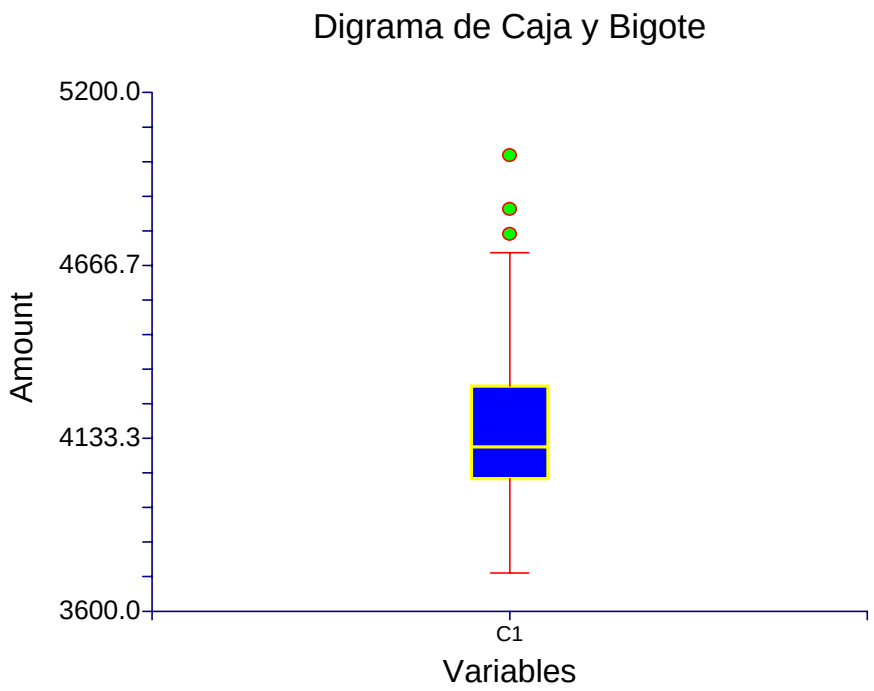
Existen también algunas técnicas exploratorias de carácter gráfico que son excelentes auxiliares para averiguar el comportamiento y características de las variables numéricas. Aquí presentaremos las siguientes (aunque existen muchos más con diversos propósitos):

1. **Diagrama de Box-Whiskers**. Utilizado para averiguar la simetría de una variable y la presencia de datos atípicos.
2. **Histogramas de barras** para datos agrupados. Cuyo uso principal es detectar la "forma" de la distribución
3. **Diagrama P-P** (papel probabilistico o probabilistic plot). Utilizado para detectar el ajuste de las observaciones a una distribución teórica. Casi siempre a una distribución normal
4. **Diagramas de dispersión de dos variables.** Utilizado con el fin de detectar dependencias funcionales entre dos variables numéricas

1. Un diagrama Box-Whiskers (Caja y bigotes), también llamado Box-Plot, consta de una caja rectangular cuyo largo es proporcional al rango intercuartílico, cuyo bigote

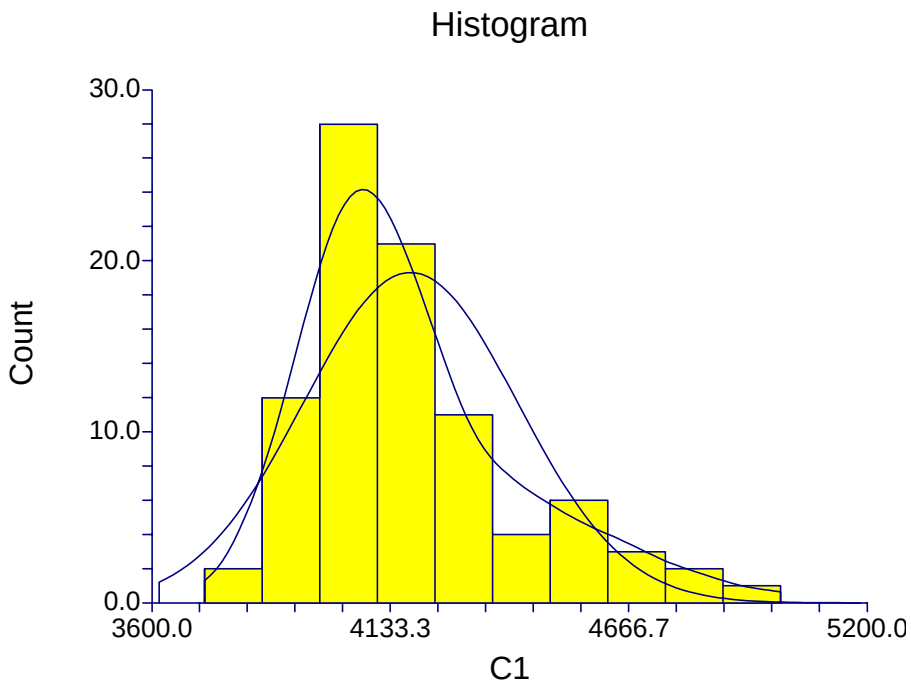
izquierdo o inferior es proporcional a la diferencia entre el primer cuartil y el valor mínimo, cuyo bigote derecho o superior es proporcional a la diferencia entre el máximo y el tercer cuartil y en la cual se han señalado la media con una x y la mediana con una línea paralela a la base y que atraviesa la caja a lo ancho. Cuando la variable examinada es simétrica la caja también lo es y entonces media y mediana coinciden. Cualquier asimetría se refleja en una asimetría más o menos pronunciada en la caja.

Para el caso de la variable analizada en el ejemplo anterior la gráfica Box-Whiskers es la siguiente:



Como se ve, hay una ligera asimetría positiva, reflejada en el hecho de que bigote superior es más largo que el inferior. Los puntos verdes indican la presencia de algunos valores extremos cuyo valor se considera un poco fuera de lo usual.

En segundo lugar, trataremos de averiguar la forma de la distribución. Para ello agrupamos los datos en 10 clases y dibujaremos un histograma. En este caso se ha dibujado una curva normal sobre el histograma para determinar el grado de ajuste de los datos a dicha distribución.



En este caso la tabla de frecuencias se construye de una manera similar a la ya vista para datos no agrupados, pero las frecuencias corresponden a la cantidad de observaciones que se encuentren dentro de cada clase.

La tabla de datos agrupados se construye dividiendo el rango de variación de los datos (valor máximo menos valor mínimo) en 10 partes iguales y construyendo luego intervalos de clase cuyos extremos son puntos situados de tal manera que cubran un intervalo de longitud igual al valor obtenido anteriormente. Los puntos medios de cada intervalo se llaman marcas de clase y son, por así decirlo, los representantes de cada clase. La frecuencia de clase está constituída por el número de observaciones que se encuentran en cada clase.

Con las frecuencias anteriores pueden calcularse frecuencias relativas y acumuladas de una manera similar a como se hizo con los datos anteriormente.

TABLA DE FRECUENCIAS PARA DATOS AGRUPADOS EN 10 CLASES

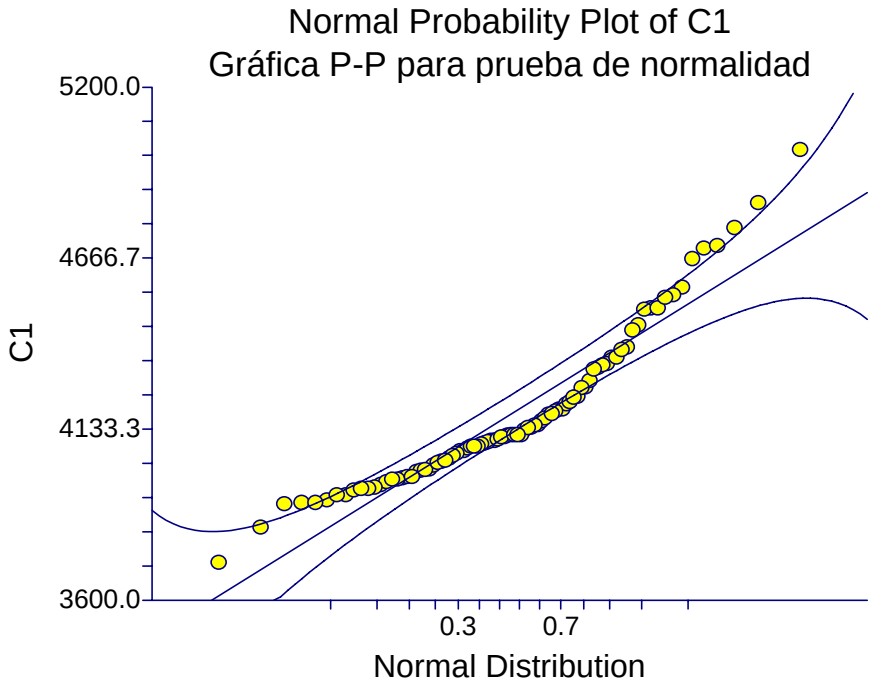
INFER (-----] SUPER		MARCAS	F.ABS	F.RELAT	F.ACUM	FREL.ACU
3717.90	3846.69	3782.30	2	0.02220	2	0.02220
3846.69	3975.48	3911.09	12	0.13330	14	0.15560
3975.48	4104.27	4039.88	28	0.31110	42	0.46670
4104.27	4233.06	4168.67	21	0.23330	63	0.70000
4233.06	4361.85	4297.46	11	0.12220	74	0.82220
4361.85	4490.64	4426.25	4	0.04440	78	0.86670
4490.64	4619.43	4555.04	6	0.06670	84	0.93330
4619.43	4748.22	4683.83	3	0.03330	87	0.96670
4748.22	4877.01	4812.62	2	0.02220	89	0.98890
4877.01	5005.80	4941.41	1	0.01110	90	1.00000

Todavía se lee en algunos textos que el agrupamiento de datos se hace con el fin de reducir cálculos cuando el número de observaciones es grande. Esto era válido cuando no existían los computadores. Hoy en día se usa con otros fines. Uno de ellos es la construcción de histogramas que ayuden a identificar la distribución a la que los datos se ajustan. El agrupamiento debe hacerse de una manera sencilla y natural sin los misteriosos y complicados procedimientos que a veces se encuentran en algunos textos anacrónicos, procedimientos que nada aportan al resultado final. Casi todos los paquetes producen histogramas de barras como herramienta gráfica exploratoria. ESM-PLUS también lo hace después de haber agrupado los datos en un cierto número de clases que el usuario haya decidido (entre 3 y 15, dependiendo de la cantidad de datos). La conocida fórmula de Sturges para definir el número de clases en que agrupan los datos está dada por $k = 1 + 3.322\,\mathrm{Log}(n)$. Puede usarse, si se quiere, pero su uso no agrega valor científico a un análisis. Lo mismo sucede con ciertas tablas que pretenden definir el número de clases en función del tamaño de muestra (Kelley, Walker y Lev). El sentido común es, en estos casos, el mejor consejero.

2. Diagrama P-P para verificar el ajuste de los datos a una distribución normal. Debido a que la mayor parte de los resultados estadísticos concernientes a la estimación de parámetros descansan en el supuesto de normalidad de las variables involucradas, resulta importante saber si una variable es normal o no, ya que de este hecho va a

depender la validez de las inferencias. Una fuerte violación del supuesto de normalidad debe producir dudas sobre la validez de los resultados que dependan de este supuesto.

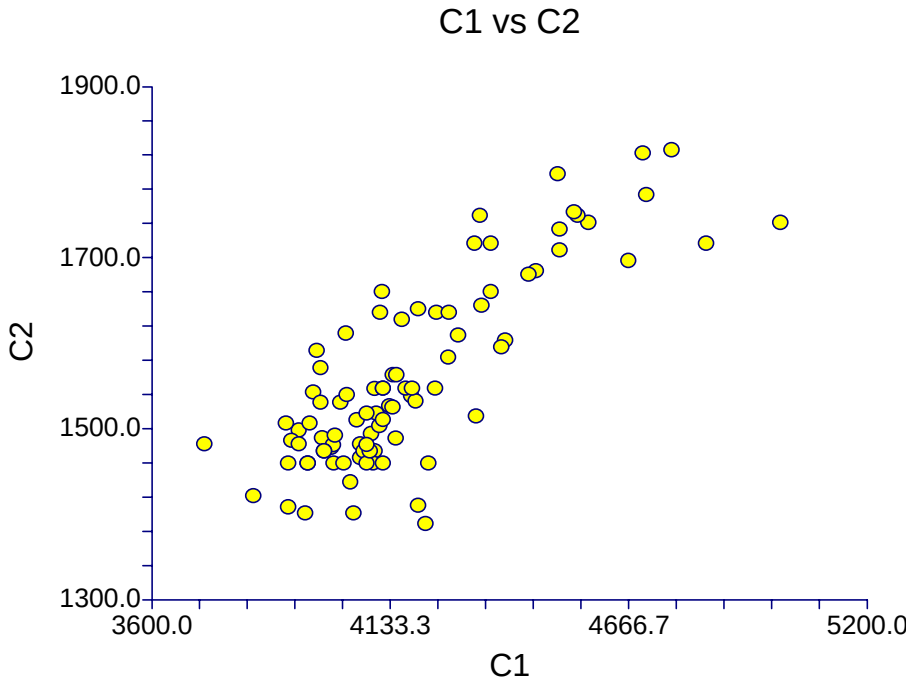
La gráfica obedece al principio general de dibujar parejas de puntos de la forma $(y,F(x))$ donde y representa los valores de probabilidad observados en la tabla de frecuencias y $F(x)$ representa los valores de probabilidad calculados por el modelo teórico al que supuestamente se ajustan los datos. Cuando éstos se ajustan al modelo teórico se tendrá $y = F(x)$ y, por tanto, la gráfica será una línea recta. En nuestro ejemplo se ha obtenido la primera gráfica de la página siguiente. En este caso se han construido límites de significancia del 5%



Como puede apreciarse, la variable es aproximadamente normal. La falta de normalidad no es acentuada, lo que se deduce del hecho de que los puntos que la representan no abandonan de una manera severa la región de significancia del 5%.

Presentamos ahora el diagrama de dispersión de la variable C1 (CABEZA) que representa la longitud de la cabeza y la variable C2 (COLLAR) que representa el ancho de collar de cada animal. Esta gráfica no es más que el dibujo de los puntos de la forma (x, y) donde

x es valor de la primera variable y *y* es el correspondiente valor de la segunda variable. La forma de la curva resultante (si es que hay alguna dependencia funcional) indica la posible relación entre las dos variables al expresar a la segunda como función de la primera.



Como se ve, existe una mediana tendencia a una relación de tipo lineal entre las dos variables. Se puede pensar en una recta que atraviese los puntos a lo largo de la nube. Ella sería la **recta de regresión** y es tal que si se miden las distancias de cada punto al correspondiente punto estimado por la recta de regresión y se suman sus cuadrados el resultado es mínimo (por esta razón se dice también que esa recta es la recta de mínimos cuadrados). Un objetivo estadístico es estimar su ecuación

TRANSFORMACIONES DE DATOS

Frecuentemente se hace necesario transformar variables por muy diversas razones: reducir sus rangos de variabilidad, modificar su distribución a fin de ajustarla a otra de mejor comportamiento, etc.

Toda transformación de variables produce una nueva variable con distribución diferente, dependiendo fundamentalmente del tipo de transformación. Algunas de las transformaciones más usuales son las siguientes:

1. **Transformaciones de tipo lineal**: en las cuales la variable X se transforma en una nueva variable $Y = aX + b$ donde a, b son constantes. Un ejemplo, sería tomar la mitad de la edad y sumar 10.

2. **Transformaciones de tipo polinomial**. Constituyen una generalización de la anterior. En ellas $Y = a_0 + a_1 X + \cdots + a_k X^k$. Donde los coeficientes son constantes. Los exponentes podrían en principio ser cualquier real distinto de cero.

3. **Transformación logarítmica**. Como su nombre lo indica, en este caso se ha de tener:

$$Y = \mathrm{Log}(X)$$

4. **Estandarización**. Una de las transformaciones más importantes usadas en estadística es la estandarización la cual consiste en una translación de la población seguida de un cambio de escala. Es usual denotarla mediante la letra Z y se define así:

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

donde $\mu = \mathrm{E}(X)$ y $\sigma^2 = \mathrm{V}(X)$ son respectivamente la media y la varianza de la variable X. En la práctica se estandarizan los datos usando la media y la varianza muestrales. Existe la creencia errónea de que la estandarización normaliza los datos.

Existen otras muchas transformaciones, tantas como fórmulas matemáticas pueda imaginarse, pero no todas son igualmente importantes. Por su sencillez, una de las más importantes es la transformación lineal ya mencionada. Igualmente es importante la transformación logarítmica que ayuda a "normalizar" datos exponenciales o de otras distribuciones asimétricas positivas. Esto es, datos cuyo comportamiento es asimétrico positivo se convierten en nuevos datos cuyo comportamiento es más ajustado a una normal. ESM-PLUS produce varias transformaciones de datos.

La estandarización siempre produce una variable numérica adimensional de media 0 y de varianza 1. A causa de esto muchas personas creen erróneamente que al estandarizar una variable se la está convirtiendo en variable normal y no es así. Una variable no normal al ser estandarizada seguirá siendo no-normal y una variable normal estandarizada sigue

siendo normal pero en otra escala y centrada sobre el eje X. El comando en Minitab para estandarizar datos es CENTER. Su uso debe ser consultado en el manual o en las ayudas del paquete.

LA DISTRIBUCION NORMAL

Las siguientes funciones corresponden a las densidades de variables aleatorias normales, es decir, variables continuas, cuya distribución de probabilidad se ajusta a una normal (distribución de Gauss o de DeMoivre), la primera no estandarizada y la segunda estandarizada:

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma} \text{Exp}\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2 \right] \text{ con } t, \mu, \sigma \in \Re \quad \sigma > 0$$

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \text{Exp}\left(-\frac{1}{2}t^2\right) \quad \text{con } t \in \Re$$

En la práctica sólo se usa la segunda de estas funciones, estandarizando previamente la variable que se esté usando.

La distribución normal tiene una gran importancia tanto teórica como práctica. Es una de las distribuciones de mayor aplicación en estadística. Es importante desde el punto de vista teórico porque gran parte de la teoría estadística ha sido deducida para variables aleatorias continuas normales (poblaciones normales) lo que implica que para otras variables que no sean normales, muchos resultados son apenas aproximados y, tanto más inexactos cuanto más "anormales" sean tales variables. Desde el punto de vista práctico es importante porque el comportamiento de muchas variables de la vida real se ajusta a una distribución normal. Tal es el caso, por ejemplo, de la estatura, el peso, la talla, el coeficiente intelectual.

Las fórmulas anteriores corresponden a las curvas de densidad de distribuciones normales. Ellas dicen cómo es la frecuencia de aparición de los valores que toma una variable normal (forma de la distribución). Sin embargo, como se dijo anteriormente, las medidas de probabilidad para tales valores deben calcularse con las funciones acumulativas. Esto es:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

Se puede demostrar que la estandarización de la variable permite usar sólo la segunda fórmula para el cálculo de las probabilidad. En efecto, se cumple:

$$F(x) = \Phi(z) = \int_{-\infty}^{z} \varphi(t)dt \quad \text{donde} \quad z = \frac{x-\mu}{\sigma}$$

es decir, en la práctica sólo es necesario conocer la distribución normal estándar para calcular probabilidades con cualquier distribución normal.

A manera de ejemplo: supóngase que una variable aleatoria X tiene media 25 y varianza 81 y que se quiere conocer la probabilidad $P(20 < X \leq 36)$. Puede pensarse, por ejemplo, que se trata de una variable que representa la edad de una comunidad humana y se desea saber cuál es la probabilidad de que al seleccionar aleatoriamente un miembro de esa comunidad, su edad esté comprendida entre 20 y 36 años. De acuerdo con lo visto, anteriormente, si *F* es la respectiva función de distribución, tal probabilidad sería igual a $F(36) - F(20)$, lo que implicaría el cálculo de la integral de la primera función mencionada antes, entre los límites 20 y 30, habiendo reemplazado previamente los valores μ, σ por 25 y 9 respectivamente. Esta integral no es inmediata y requiere de la implementación de un método numérico para su cálculo. A cambio de esto, previamente se estandariza X lo que produce $Z = \frac{X-5}{9}$. Los límites de la integral se transforman entonces en $\frac{20-5}{9} = 1.6667$ y $\frac{30-5}{9} = 2.7778$. En consecuencia, bastará calcular la integral de la segunda función -más sencilla- entre estos dos límites. El cálculo de la integral tampoco es simple, pero sus valores se encuentran en tablas ya elaboradas. En ellas se encuentran los valores $\Phi(2.7778)$ y $\Phi(1.6667)$, correspondientes a la función de distribución, cuya diferencia produce como resultado el valor: 0.04504. esta es la probabilidad buscada.

Hoy en día comienzan a caer en desuso las tablas de probabilidades pues muchos programas de computador permiten su cálculo. El programa ESM-PLUS, por ejemplo, presenta el cálculo de probabilidades con 16 funciones de distribución, 4 de ellas discretas

(binomial, binomial negativa, Poisson y geométrica) y 12 continuas (normal en sus dos formas, t, F, Ji cuadrado, exponencial, exponencial negativa, Laplace, Weibull, Gumbel, Cauchy y Gamma). A estas rutinas se tiene acceso siguiendo la secuencia 3 → 9 → 2.

Igualmente el paquete NCSS tiene una calculadora de probabilidades bajo diferentes distribuciones. Estas rutinas pueden ser usadas para calcular probabilidades cuando sea necesario, dejando en la obsolescencia las tablas impresas que sólo presentan la probabilidad para determinados valores.

LA DISTRIBUCION BINOMIAL

Una de las distribuciones discretas más importantes es la distribución binomial, asociada con experimentos que presenten sólo dos resultados. Por ejemplo, presencia o ausencia de enfermedades, apto o no apto para el desempeño de un cargo, estado de un circuito eléctrico: abierto o cerrado, éxito o fracaso en un examen, hembra o macho en el sexo de un animal, etc. Los dos resultados de un experimento de este tipo (llamados experimentos de Bernoulli) se denominan **ÉXITO** y **FRACASO**. Esto es un simple nombre y no importa cuál de ellos sea favorable a nuestros intereses para que sea considerado como éxito. Lo esencial es identificar a cual de ellos le pondremos el mote de ***éxito***. Es costumbre denotar por p la probabilidad de que ocurra un éxito en un experimento de Bernoulli. Por supuesto, la probabilidad de un fracaso será entonces $q = 1 - p$.

Supóngase ahora que se realizan en forma independiente n experimentos de Bernoulli y nos preguntamos por la probabilidad de que en esos n experimentos ocurran exactamente x éxitos (naturalmente x es uno de los valores 0, 1, 2, ..., n). En otras palabras, estamos preguntando cuánto vale $P(X = x)$. En este caso se dice que la variable aleatoria X que **mide el número de éxitos en n experimentos de Bernoulli**, tiene distribución binomial de parámetros p y n.

Se puede probar que

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

En consecuencia, esta expresión define la función de densidad para una distribución binomial de parámetros p y n.

La correspondiente función de distribución está dada entonces por:

$$F(x) = P(X \le x) = \sum_{t=-\infty}^{x} f(t) = \sum_{t=0}^{x} f(t)$$

Los valores de probabilidad para n variando de 1 hasta 80 pueden calcularse con el programa ESM-PLUS siguiendo la secuencia: 3 → 9 → 2

Por ejemplo, usando ESM-PLUS, podemos calcular la probabilidad de que una variable aleatoria binomial con parámetros **p = 0.32** y **n = 20** tome el valor 12 o que tome valores entre 5 y 13. Es decir: $P(X = 12) = f(12)$ y $P(5 < X \leq 13) = F(13) - F(5)$. Se obtiene: $f(12) = 0.066395$ y $F(13) - F(5) = 0.999448 - 0.342615 = 0.656833$.

Nótese que en el caso discreto la inclusión de los límites cambia los resultados. Por ejemplo: con la misma distribución anterior, no es lo mismo $P(5 < X \leq 13)$ que $P(5 < X < 13)$. Esta última equivale a $P(5 < X \leq 12) = F(12) - F(5) = 0.997525 - 0.342615$

Otro ejemplo: Supóngase que en el servicio de urgencias de un hospital se ha establecido que el 12% de los pacientes que llegan a solicitar el servicio los viernes en la noche requiere del uso de unidades de cuidados intensivos por más de un día. El hospital cuenta con tres unidades de cuidados intensivos disponibles para el servicio de urgencias. Un médico se pregunta un viernes en la tarde cuál es la probabilidad de que al llegar 20 pacientes esa noche, el hospital pueda atender cuidados intensivos sin dificultades durante las 12 horas siguientes.

Un análisis del ejemplo muestra que el servicio se podrá prestar si de los 20 pacientes no hay más de tres que requieran cuidados intensivos. Esto es, si el número de tales pacientes es a lo sumo 3. Un paciente requiere o no el servicio, así que podemos denominar éxito el caso en que se requiera y fracaso el caso en que no. En consecuencia el médico está interesado en el valor de $P(X=0) + P(X = 1) + P(X = 2) + P(X=3) = P(X \leq 3)$, bajo una distribución binomial de parámetros $p = 0.12$ y $n = 20$. Esto es: 0.07756 + 0.21153 + 0.27403 + 0.22421 = 0.78734

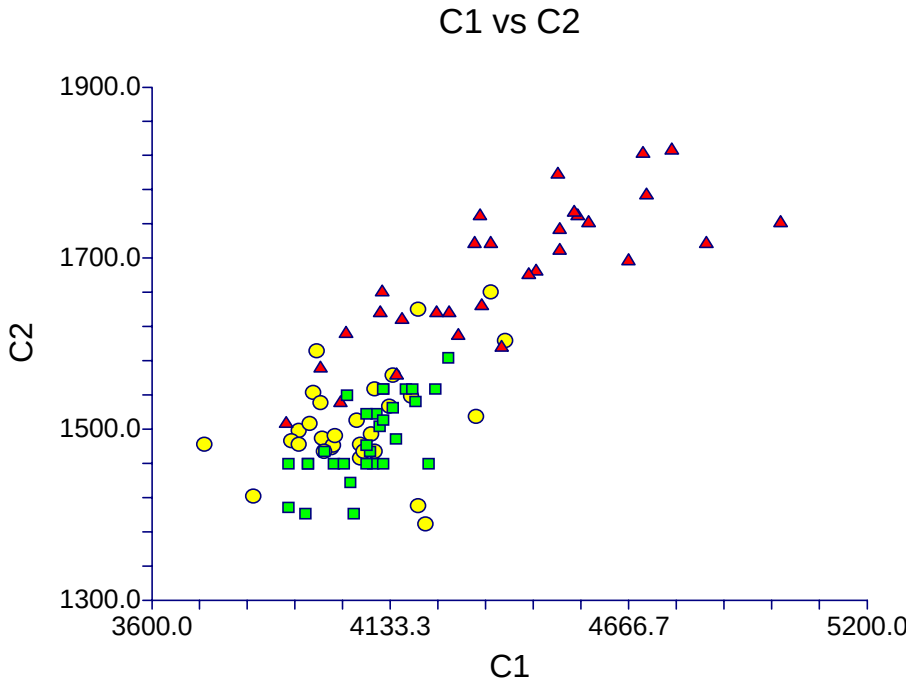
ANALISIS EXPLORATORIOS PARA DOS VARIABLES

Aunque en estadística suele suceder que se presenten simultáneamente muchas variables de análisis (análisis multivariados), a nivel introductorio sólo podemos presentar casos de dos variables., las cuales pueden ser ambas numéricas, ambas categóricas o una numérica y la otra categórica.

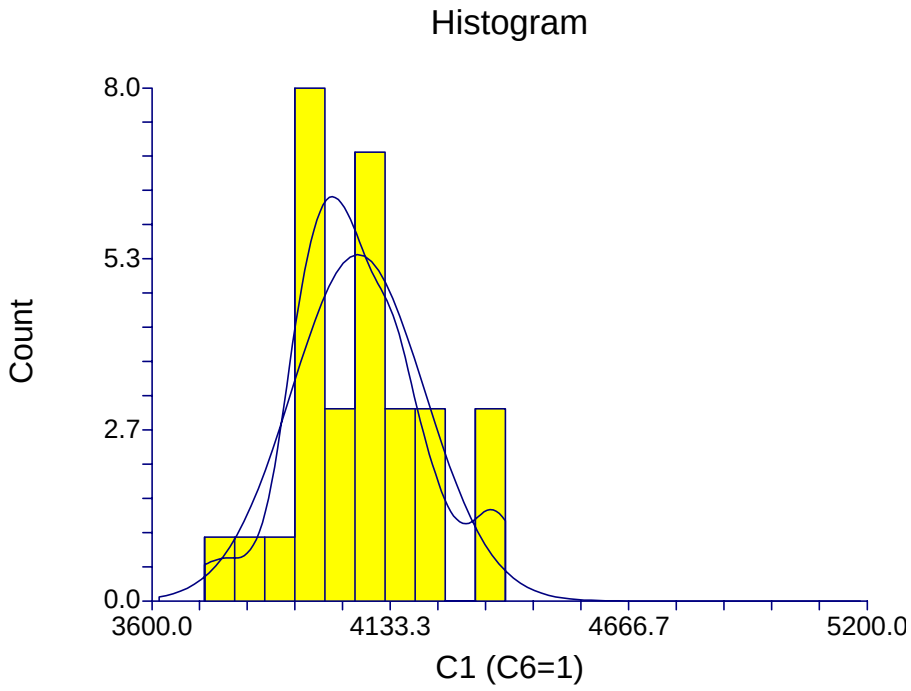
Ya mencionamos el caso en que se tienen dos variables numéricas y existe interés en averiguar si los datos reflejan una dependencia funcional entre ellas. Es el tema conocido

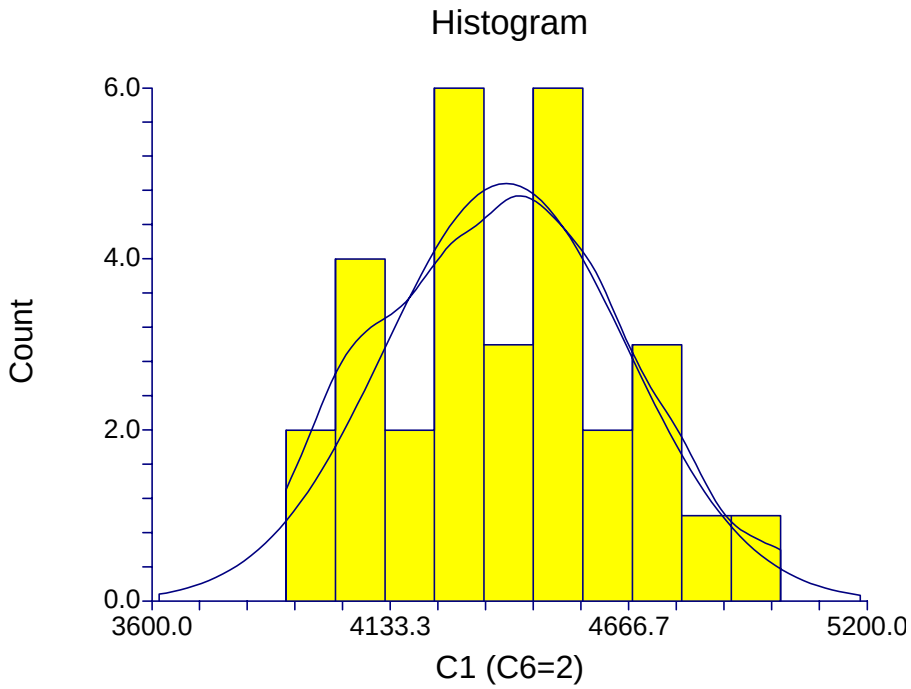
como regresión y comúnmente basta con la construcción de un diagrama de dispersión para poner en evidencia una relación de este tipo. En estos casos la estadística cuenta con herramientas que permiten estimar las funciones que relacionan tales variables. El caso más sencillo y uno de los más interesantes es la regresión lineal en la cual las dos variables aleatorias X y Y están ligadas por una relación del tipo $Y = \beta_0 + \beta_1 X$ Los paquetes estadísticos proporcionan herramientas que permiten estimar los coeficientes y analizar la validez del modelo así obtenido.

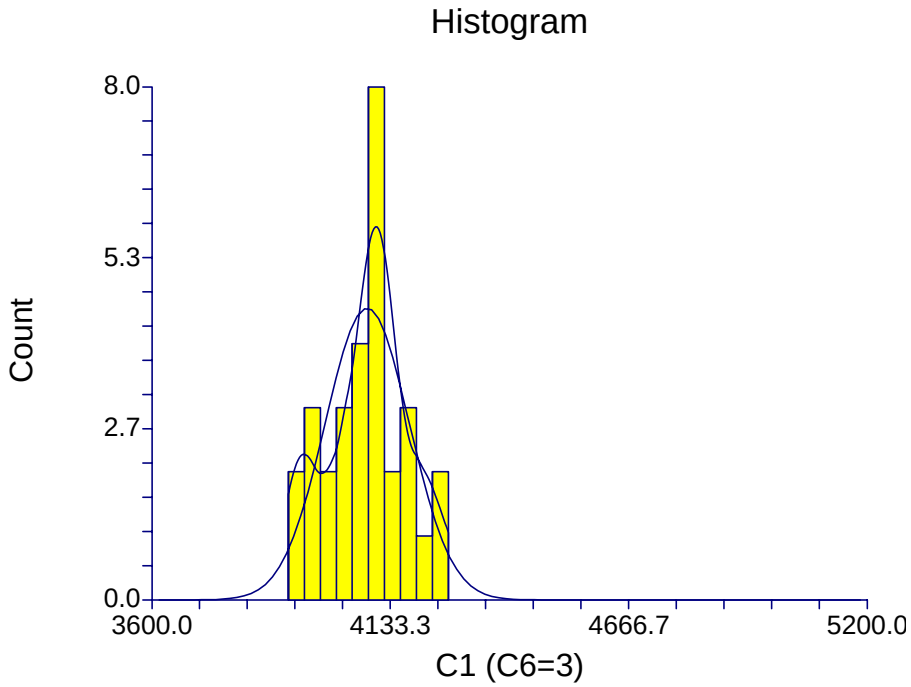
Cuando se tienen dos variables y una de ellas es categórica y la otra numérica se puede decir que los valores de esta última se encuentran clasificados por los valores de aquella. Es posible en estos casos separar los valores de la variable numérica correspondientes a cada valor de la categórica y hacer un análisis separado para cada conjunto de datos. El siguiente gráfico muestra el diagrama de dispersión clasificando la información para cada uno de los tres grupos de procedencia:



De igual manera, es posible hacer histogramas de distribuciones, para cada grupo y obtener algo como lo siguiente:




Histogram
6.0
4.0
2.0
0.0
Count
3600.0
4133.3
4666.7
5200.0
C1 (C6=2)

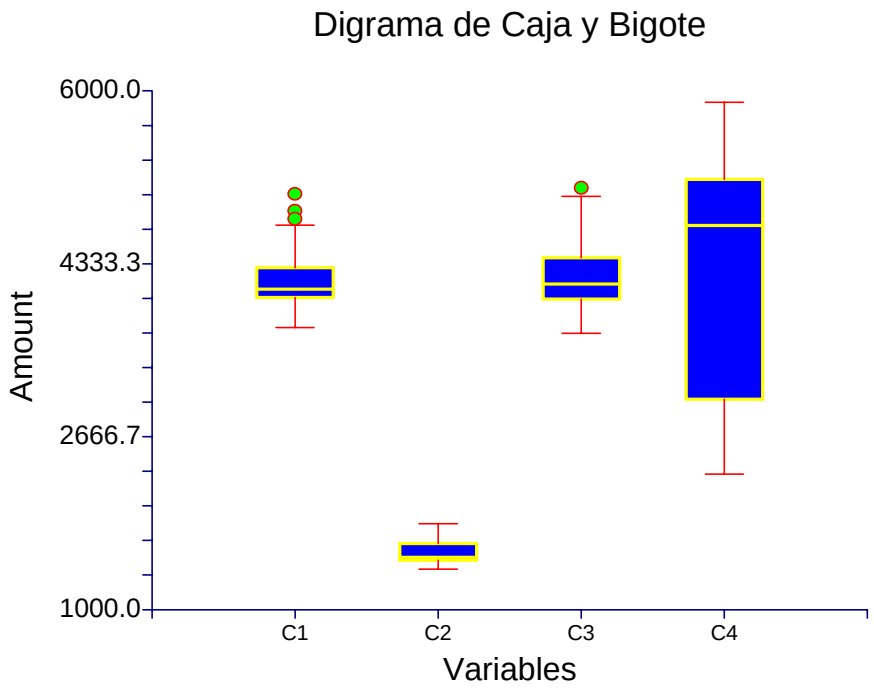

Histogram
8.0
5.3
2.7
0.0
Count
3600.0
4133.3
4666.7
5200.0
C1 (C6=3)

En general, podría decirse que la muestra se ha repartido en tantas submuestras como indiquen las categorías de la variable categórica y que cada submuestra recibe un tratamiento estadístico por separado.

Se podría preguntar muchas cosas. Por ejemplo, cada submuestra representa una población diferente y ¿pueden considerarse iguales las medias de dichas poblaciones? - ¿Se podría afirmar que las tres varianzas son iguales o no?

Preguntas cómo éstas, aunque son legítimas, no se pueden responder en este taller. Pero se debe saber que sí existen métodos de solución.

Cuando existen múltiples variables numéricas como en el ejemplo que venimos manejando, es posible construir gráficos simultáneos para ellas, lo que permite comparar sus comportamientos. Por ejemplo, para las cuatro variables numéricas se tienen los siguientes Box-plots simultáneos:

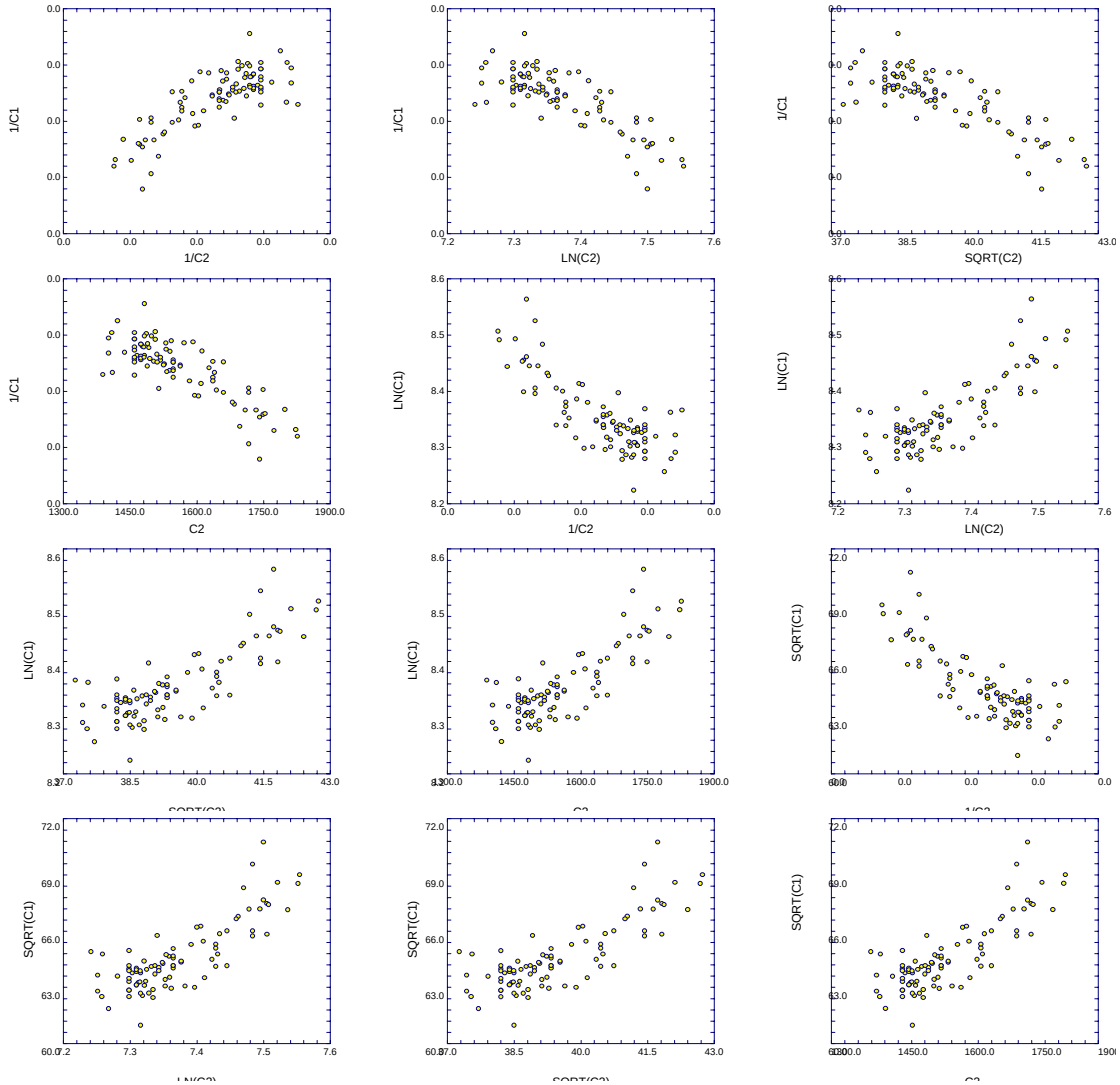


El programa NCSS permite explorar algunas relaciones especiales entre variables como sucede en los gráficos siguientes, donde una relación funcional puede ser puesta en evidencia a través de los gráficos:

Scatter-Plot Matrix Report

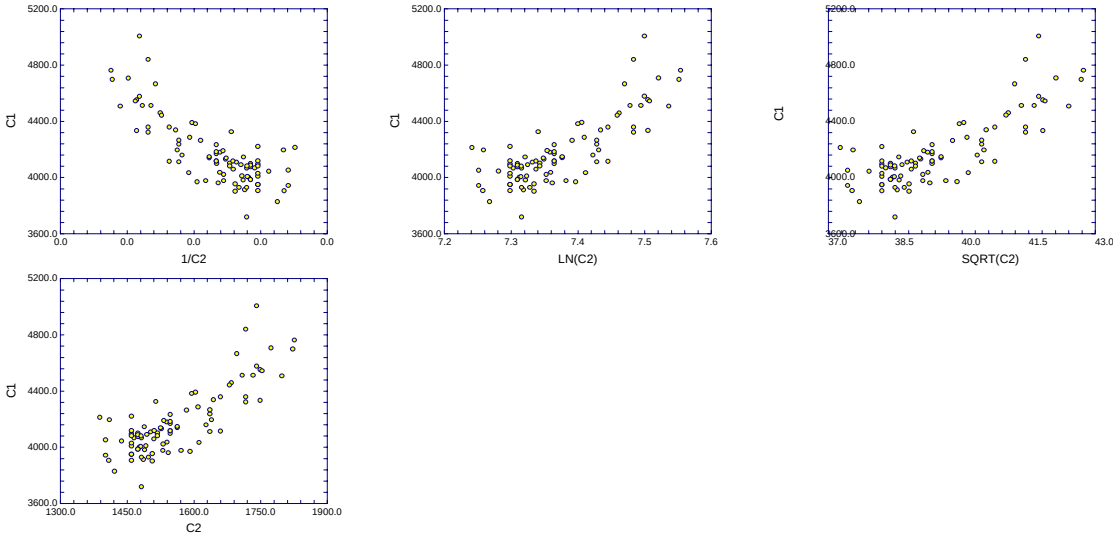
Database C:\datos\TALLER.S0

Plot Section



Scatter-Plot Matrix Report

Database C:\datos\TALLER.S0



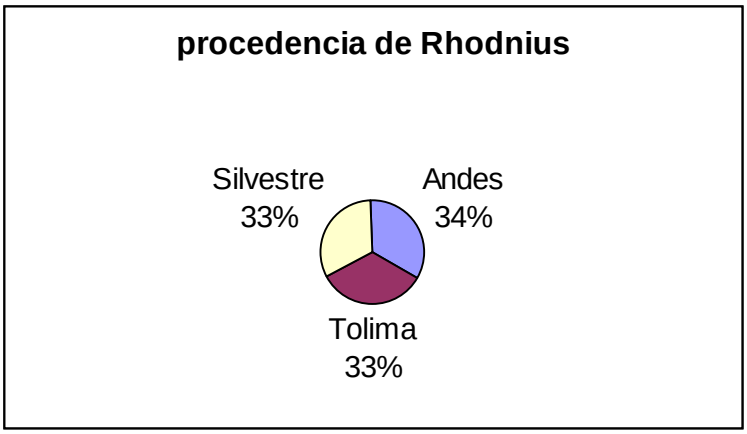
ANÁLISIS EXPLORATORIO DE VARIABLES CATEGÓRICAS

A nivel elemental es poco lo que puede decirse de variables cuyos valores son categorías, tales como las variables C5 y C6 de los datos que hemos estado analizando. Lo más usual son los conteos de frecuencias y la estimación de los porcentajes correspondientes a cada una de las categorías. En estos casos es posible realizar diagramas de tortas o histogramas de barras en los cuales el tamaño (área) de cada uno de los elementos gráficos es proporcional al porcentaje de la correspondiente categoría. Estas gráficas pueden hacerse de una manera sencilla y elegante con Excel.

Por ejemplo, la variable C6 con sus tres categorías produce los siguientes resultados al realizar un conteo de ellas (este cálculo se realizó con ESM ingresando por análisis de encuestas):

Modalidad:	Numero:	Porcentaje:
1	30	33.33
2	30	33.33
3	30	33.33
TOTAL:	90	100.00 %

La gráfica correspondiente, realizada en Excel, es la siguiente:



Finalmente consideremos el caso en que simultáneamente se están examinando dos variables categóricas.

El análisis más frecuente que se hace en estos casos es el conocido **cruce de las variables** en el cual se puede hacer un conteo de cuántos individuos se encuentran simultáneamente en cada una de las categorías de una variable y cada una de las categorías de la otra. Las tablas resultantes reciben el nombre de tablas de contingencia y pueden ser usadas con fines de inferencia para probar la dependencia de dos variables categóricas.

A manera de ejemplo, mostramos la tabla de contingencia resultante al cruzar las variables C5 (SEXO) y C6 (PROCEDENCIA). Dicha tabla se ha realizado con ESM:

```
** CRUCES DE VARIABLES CATEGORICAS - TABLAS DE CONTINGENCIA **
_____________________________________________________________
Cada celda contiene tres valores as¡:
                                        1. Frecuencia observada.
                                        2. Frecuencia esperada.
                                        3. Porcentaje del Total.
```

Continúa...

```
FILAS = SEXO          COLUMNAS = PROCED
_____________________________________________________
Categs: \ PROCED:
SEXO      1         2         3        TOTAL
          _____________________________________________
  1       15        15        15        45
         15.00     15.00     15.00
         16.67     16.67     16.67     50.00

  2       15        15        15        45
         15.00     15.00     15.00
         16.67     16.67     16.67     50.00
_____________________________________________________
TOTAL:    30        30        30        90
  %   :  33.33     33.33     33.33     %100
```

Finalmente diremos que ViSta es uno de los paquetes estadísticos de análisis visual de datos más interesantes que existe. Este programa que es totalmente gratuito y de libre uso en educación puede ser bajado de Internet en la dirección www.visualstats.org .

El archivo de datos TALLERLY.TXT está listo para ser importado dentro de ViSta y poder así examinar los datos.

BIBLIOGRAFIA BASICA:

1. HOGG R. y E. TANIS; *Probability and Statistical Inference*. Maxwell Macmillan. New York. 1989

2. MONSERRAT FREIXA I BLANXART y otros; *Análisis Exploratorio de Datos, Nuevas Técnicas Estadísticas*. PPU. Barcelona. 1992

3. CLAVIJO J. A.; *Curso de Métodos Estadísticos*. Universidad del Tolima. Ibagué. 2001