Facultad
de Ciencias
Económicas

Introducción a la Estadística Judicial: Un enfoque descriptivo

María Gabriela Benedicto

**Facultad
de Ciencias
Económicas**

María Gabriela Benedicto

INTRODUCCION A LA ESTADISTICA JUDICIAL: UN ENFOQUE DESCRIPTIVO

INDICE

PROLOGO

Pocos parecen ser, a priori, los puntos de contacto entre la ciencia estadística y la justicia. Fórmulas, mediciones, gráficos, parecieran estar bastante lejos del proceso judicial que tramita una causa para dar respuesta a un conflicto.

Pero si pensamos a la justicia como un servicio público que debe dar respuestas a la sociedad, y que esas respuestas deben ser rápidas, eficaces y eficientes, es claro que este servicio debe ser monitoreado y evaluado. Desde los años 80, este concepto se ha instalado y es entonces que los sistemas de información y los indicadores de gestión han ido adquiriendo importancia creciente.

Mucho se ha escrito acerca de los indicadores, de su construcción e implementación, pero poco respecto de las técnicas estadísticas que permiten, una vez calculados esos indicadores, resumirlos y graficarlos. Por otro lado, el avance en la implementación de los sistemas de gestión permite contar con información puntual y desagregada que enriquece la descripción de la situación de un organismo judicial en cualquier momento. Esto hace que además se puedan reemplazar antiguos indicadores que se basaban en información agregada, por otros más precisos y actualizables en línea.

El objetivo de este libro es presentar técnicas estadísticas descriptivas básicas, y sus aplicaciones en el ámbito judicial. Está pensado para que el lector pueda hacer una lectura amena de los conceptos y las interpretaciones, sin tener que detenerse en las fórmulas matemáticas que dieron origen a cada uno de los resultados. Estas quedan a disposición de los interesados en el Anexo.

El proceso de construcción de los conceptos se da en forma casi intuitiva y desde la realidad misma del trabajo diario. Esto permite que cualquier operador judicial sin una formación específica en matemática o estadística, pero interesado en el manejo de la información, pueda incursionar en el entorno de la Estadística Judicial.

I - INTRODUCCION

La estadística judicial constituye una herramienta imprescindible para la gestión de los organismos judiciales. El contar con información válida y actualizada, y el procesar adecuadamente dicha información, contribuye a la identificación de fortalezas y debilidades del sistema, a monitorear la gestión, diseñar reformas y adoptar medidas, implementar políticas públicas judiciales, rendir cuentas a la ciudadanía, entre otras cosas.

La información que se genera en los órganos jurisdiccionales tiene ciertas características especiales que obligan a reflexionar sobre las reglas específicas que deben guiar su análisis. En este sentido resulta de vital importancia incorporar los conocimientos y técnicas necesarias para poder construir e interpretar datos judiciales tratados estadísticamente.

El término *"estadística"* es ampliamente escuchado y pronunciado a diario desde diversos sectores y disciplinas. Sin embargo, hay una gran diferencia entre el sentido del término cuando se utiliza en el lenguaje corriente (generalmente al anteceder una cita de carácter numérico) y lo que la estadística significa como ciencia.

Debido a lo extenso y variado del campo cubierto por la estadística, es difícil dar una definición precisa del concepto. Sin embargo, podemos citar tres definiciones que son abarcativas de los distintos aspectos.

El New Collegiate Dictionary de Webster, la define como *"una rama de la matemática que trata de la recopilación, el análisis, la interpretación y la presentación de una cantidad de datos numéricos"*. Para Kendall y *Stuart "es la rama del método científico que trata de los datos reunidos al contar o medir propiedades de alguna población"*. Y Freund sostiene que puede pensarse como *"el conocimiento relacionado con el tomar decisiones en situaciones de incertidumbre".*

Más allá de las definiciones, todos los estadísticos están de acuerdo en clasificar la materia en dos tipos: la Estadística Descriptiva y la Estadística Inferencial, que desempeñan funciones distintas pero complementarias del análisis estadístico.

***Estadística Descriptiva*:** Trata del resumen y descripción de los datos. Este resumen puede ser tabular, gráfico o numérico. El análisis se limita en sí mismo a los datos recolectados y no se hace inferencia alguna o generalización a la totalidad de donde provienen esas observaciones (población) si es que el conjunto de los datos analizados no constituye la totalidad de esa población y se remite sólo a una porción de la misma (muestra).

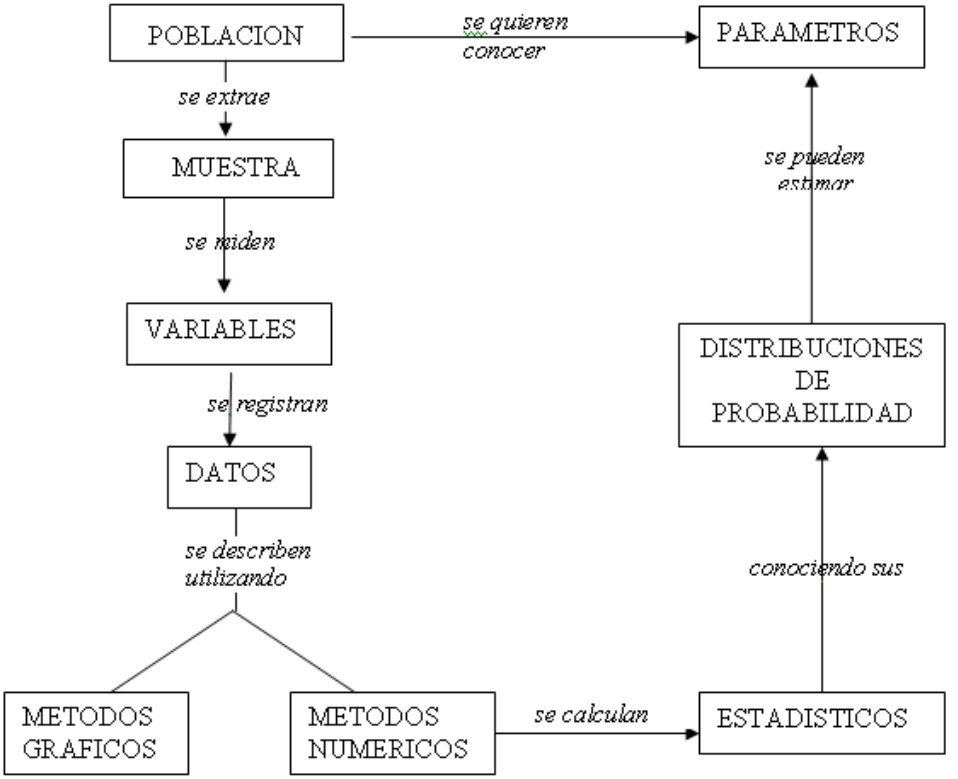
***Estadística Inferencial*:** Si bien la descripción de los datos recolectados es muchas veces en sí misma el fin que se persigue en un estudio, generalmente el objetivo último de la labor estadística es el de extraer conclusiones útiles sobre la totalidad de la población, basándose en las observaciones recolectadas. La inferencia estadística es el proceso de hacer predicciones acerca de un todo o tomar decisiones al basarnos en la información contenida en una muestra.

Se pueden resumir estas dos facetas de la estadística en la siguiente definición:

En la actualidad, la estadística es un campo de la ciencia que comprende técnicas de recolección de datos pertinentes y procedimientos de análisis de los mismos destinados a extraer conclusiones a partir de información incompleta o imperfecta.

Es una disciplina de apariencia abstracta por estar íntimamente relacionada con la matemática, pero utilizada por ciencias de diversas índoles como lo son las biológicas, sociales y económicas, dado que todas ellas precisan recopilar y resumir conjuntos de datos que, una vez analizados adecuadamente, seguramente producirán información útil que mejorará el conocimiento de la realidad bajo estudio.

El siguiente cuadro resume el círculo que vincula a estas dos metodologías de la estadística:



La terminología utilizada amerita dar algunas definiciones de los conceptos que se manejarán en adelante.

Población: Es el conjunto de todos los elementos bajo estudio. Puede ser un conjunto de personas pero también puede tratarse de un conjunto de otras entidades, como por ejemplo, expedientes, juzgados, etc.

Muestra: es un subconjunto representativo de la población. A partir de ella se obtiene la información para describir a la población de la cual fue extraída.

Unidad: cada objeto o persona de la población sobre la cual se hacen las mediciones.

La lista de todas las unidades de la población se denomina **marco muestral**.

Variable: es una característica de interés que se mide en cada unidad de la muestra.

Parámetro: es un número que se calculó usando todas las unidades de la población. En general, es la característica numérica de la población en la cual el investigador está interesado.

Estadístico: es un número calculado usando todas las unidades de la muestra. En base al cálculo de estadísticos adecuados, se realiza la inferencia respecto de los parámetros de interés. Por ejemplo, si se quiere estimar la media de una población, se calculará la media de la muestra. Ese será el puntapié inicial para el proceso de inferencia.

Es importante distinguir entre *población objetivo* y *población muestreada*. La población meta, u objetivo, es aquella sobre la cual se desea hacer inferencia, mientras que la población muestreada es aquella de la que se selecciona realmente la muestra. Estas poblaciones no son necesariamente iguales y hay que tener cuidado a la hora de sacar conclusiones. El asunto clave es si la correspondencia entre la población muestreada y la población objetivo, respecto a los elementos de interés, es suficientemente estrecha como para permitir la ampliación.

Por ejemplo, si se quiere medir el grado de acceso a la justicia en una ciudad determinada, se puede diseñar una encuesta específica. Pero si esta encuesta se distribuye entre algunas personas que acceden a la mesa de entradas de un juzgado, la población que se está muestreando es realmente el conjunto de personas que acceden al juzgado, y no todos los habitantes de la ciudad. Las conclusiones que se extraigan aplicarán sólo a esa población, y no a la que originalmente se intentaba estudiar. (Puede que haya personas que no se acercan porque ni siquiera saben a dónde deben

hacerlo, y eso es parte justamente de lo que se quiere medir: el acceso a la justicia).

En este punto pueden surgir dos preguntas:

- *Cómo se determina la muestra?*
- *Cómo se obtienen los datos?*

Existen distintos métodos de muestreo que responden a la primera cuestión planteada. Dependiendo de las características de la población de interés, del estudio que se quiera realizar y de la accesibilidad que se tenga a la población, convendrá utilizar uno u otro. Lo importante es que, dentro de las posibilidades, el muestreo que se realice no conduzca a sesgos sistemáticos que distorsionen las conclusiones del estudio.

II - TIPOS DE MUESTREO

Los métodos de muestreo que existen pueden dividirse en dos grandes grupos: métodos de muestreo probabilísticos y métodos de muestreo no probabilísticos.

Métodos de muestreo probabilísticos

Los métodos de muestreo probabilísticos son aquellos que se basan en el principio de equiprobabilidad. Es decir, aquellos en los que todos los individuos tienen la misma probabilidad de ser elegidos para formar parte de una muestra y, consecuentemente, todas las posibles muestras de tamaño n tienen la misma probabilidad de ser elegidas. Sólo estos métodos de muestreo probabilísticos aseguran la representatividad de la muestra extraída y son, por lo tanto, los más recomendables. Dentro de los métodos de muestreo probabilísticos se pueden detallar:

Muestreo aleatorio simple: El procedimiento empleado es el siguiente: se asigna un número a cada individuo de la población y luego, a través de algún medio mecánico (bolillas dentro de una bolsa, tablas de números aleatorios, números aleatorios generados con una calculadora o computadora, etc) se eligen al azar tantos sujetos (números) como sea necesario para completar el tamaño de muestra requerido.

Por ejemplo, si se quieren seleccionar 12 expedientes de un total de 30, se ordenan los 30 expedientes en una lista numerada como la siguiente:

Nro de Orden	Expediente
1	3452/06
2	4520/06
3	2354//04
…	
…..	
30	5679/10

y se sortean 12 números del 1 al 30. Estos números indicarán el orden de los expedientes que hay que seleccionar para integrar la muestra.

Este procedimiento, atractivo por su simpleza, tiene poca o nula utilidad práctica cuando la población que se maneja es muy grande y no existe un marco muestral adecuado.

Muestreo aleatorio sistemático: Este procedimiento exige, como el anterior, numerar todos los elementos de la población, pero en lugar de extraer n números aleatorios sólo se extrae uno. Se parte de ese número aleatorio i, que es un número elegido al azar, y los elementos que integran la muestra son los que ocupan los lugares i, i+k, i+2k, i+3k,...,i+(n-1)k. Es decir, se toman los individuos de k en k, siendo k el resultado de dividir el tamaño de la población entre el tamaño de la muestra: k=N/n. El número i que se emplea como punto de partida será un número al azar entre 1 y k.

Este mecanismo es muy útil cuando las unidades están ordenadas en forma cronológica. Si se tienen ficheros con expedientes, ordenados por fecha, un muestreo sistemático no sólo es más simple que un muestreo simple al azar en cuanto a su operatividad, sino que además garantiza recorrer todo el historial del juzgado.

El riesgo de este tipo de muestreo está en los casos en que existen periodicidades en la población ya que al elegir a los miembros de la muestra con una periodicidad constante (k) se puede introducir una homogeneidad que no se da en la población. Si por ejemplo, se quiere seleccionar una muestra sobre listas de 8 individuos en las que los 4 primeros son varones y las otras 4 son mujeres, al emplear un muestreo aleatorio sistemático con k=8 siempre se seleccionarían o sólo hombres o sólo mujeres, es decir que no podría haber una representación de los dos sexos.

Muestreo aleatorio estratificado: Trata de superar las dificultades que presentan los anteriores ya que simplifican los procesos y suelen reducir el error muestral para un tamaño dado de la muestra. Consiste en considerar categorías típicas diferentes entre sí (estratos) que poseen gran homogeneidad respecto a alguna característica (se puede estratificar, por ejemplo, según el tipo de proceso, la circunscripción, el sexo, reincidencia, etc). Lo que se pretende con este tipo de muestreo es asegurarse de que todos los estratos de interés estén representados adecuadamente en la muestra. Cada estrato funciona independientemente, pudiendo aplicarse dentro de ellos el muestreo aleatorio simple o el sistemático para elegir los elementos concretos que formarán parte de la muestra. En ocasiones las dificultades que plantean son demasiado grandes, pues exige un conocimiento detallado de la población.

Muestreo aleatorio por conglomerados: Los métodos presentados hasta ahora están pensados para seleccionar directamente los elementos de la población, es decir, las unidades muestrales son elementos de la población. En el muestreo por conglomerados la unidad muestral es un grupo de elementos de la población que forman una unidad, a la que se denomina conglomerado. Los juzgados, las localidades, las provincias, etc, son conglomerados naturales. Cuando los conglomerados son áreas geográficas suele hablarse de "muestreo por áreas".

El muestreo por conglomerados consiste en seleccionar aleatoriamente un cierto número de conglomerados (el necesario para

alcanzar el tamaño muestral establecido) y en investigar después todos los elementos pertenecientes a los conglomerados elegidos.

Si por ejemplo, se quiere estudiar el nivel de satisfacción de los empleados de un Poder Judicial, se puede diseñar un estudio y confeccionar encuestas para que sean respondidas por empleados de todas las categorías. Aplicando un muestreo por conglomerados, se pueden seleccionar aleatoriamente algunos juzgados y encuestar a todos los empleados del organismo. Cada organismo repite la estructura general de las categorías del poder judicial, y de esta manera estarían todas representadas, optimizando al mismo tiempo el estudio ya que se ahorran tiempo y esfuerzos: si se hiciera un muestreo aleatorio estratificado, habría que seleccionar algunos empleados de cada categoría y recorrer luego todos los juzgados.

Para finalizar con esta exposición de los métodos de muestreo probabilísticos es necesario comentar que ante lo compleja que puede llegar a ser la situación real de muestreo con la que uno se enfrente es muy común emplear lo que se denomina *muestreo polietápico*. Este tipo de muestreo se caracteriza por operar en sucesivas etapas, empleando en cada una de ellas el método de muestreo probabilístico más adecuado. Por ejemplo, en el muestreo por conglomerados, se puede hacer un muestreo en dos etapas: una vez seleccionados los conglomerados, tomar una muestra aleatoria de las unidades dentro de cada uno.

Métodos de muestreo no probabilísticos

A veces, para estudios exploratorios, el muestreo probabilístico resulta excesivamente costoso y se acude a métodos no probabilísticos, aún siendo conscientes de que no sirven para realizar generalizaciones, pues no se tiene certeza de que la muestra extraída sea representativa. Esto es porque no todos los sujetos de la población tienen la misma probabilidad de ser elegidos. En general se seleccionan a los sujetos siguiendo determinados criterios, de manera que la muestra obtenida represente bastante fehacientemente a la población.

Muestreo por cuotas: También denominado en ocasiones "accidental". Se asienta generalmente sobre la base de un buen conocimiento de los estratos de la población y/o de los individuos más "representativos" o "adecuados" para los fines de la investigación. Mantiene, por tanto, semejanzas con el muestreo aleatorio estratificado, pero no tiene el carácter de aleatoriedad de aquél.

En este tipo de muestreo se fijan "cuotas" que consisten en un grupo de individuos que reúnen determinadas condiciones, por ejemplo: 20 individuos de 25 a 40 años, de sexo femenino y residentes en una determinada localidad. Estas cuotas se establecen respetando la distribución de la variable en la población original (en este caso, el porcentaje de mujeres en ese rango de edad). Una vez determinada la cuota se eligen los primeros individuos que se encuentren que cumplan esas características, siguiendo algún criterio preestablecido. Este método se utiliza mucho en las encuestas de opinión.

Muestreo opinático o intencional: Este tipo de muestreo se caracteriza por un esfuerzo deliberado de obtener muestras "representativas" mediante la inclusión en la muestra de grupos supuestamente típicos. Por ejemplo, se puede utilizar para recabar opinión respecto del sistema carcelario, buscando intencionalmente a ex –convictos para conformar la muestra.

Muestreo casual o incidental: Se trata de un proceso en el que el investigador selecciona directa e intencionadamente los individuos de la población. El caso más frecuente de este procedimiento es utilizar como muestra a los individuos a los que se tiene fácil acceso (los médicos emplean con mucha frecuencia a sus propios pacientes). Un caso particular es el de los voluntarios.

Bola de nieve: Se localiza a algunos individuos, los cuales conducen a otros, y estos a otros, y así hasta conseguir una muestra suficiente. Este tipo de muestreo se emplea muy frecuentemente cuando se hacen estudios con poblaciones "marginales", delincuentes, sectas, determinados tipos de enfermos, etc.

III - FUENTES DE DATOS Y EL USO DE LA INFORMACION

Fuentes Existentes

La mayoría de los datos estadísticos utilizados por los Poderes Judiciales son ***datos internos***. Estos son recopilados en los diferentes registros básicos generando una diversidad de bases de datos acerca de las causas tramitadas, empleados, inventario, etc.

Para posicionarse en relación a otros Poderes Judiciales, o para relativizar algunos datos respecto de datos provinciales (por ejemplo, asignación presupuestaria) se necesitan datos que, para el propio organismo son ***datos externos***. Las Direcciones de Estadística provinciales son, por ejemplo, fuentes de datos externos en relación a valores económicos, demográficos, etc.

Una publicación de datos originales se denomina ***fuente primaria***, cuando ha debido recolectar el dato. Cuando la publicación contiene y analiza datos inicialmente recopilados y publicados por otra entidad, se llama ***fuente secundaria***.

Experimentos y Estudios

A veces los datos necesarios para un estudio determinado no están disponibles y por lo tanto es necesario realizar un relevamiento de los mismos mediante un estudio estadístico. Estos estudios pueden ser *experimentos* o *estudios observacionales*.

En un experimento se identifica la variable de interés, y se manipulan variables adicionales aplicando activamente un tratamiento a las unidades para observar la respuesta.

En un estudio observacional no se trata de controlar las variables de interés ni de influir sobre ellas, simplemente se observa a las unidades y se

registran los valores de las variables. El tipo más común de estudio observacional es la encuesta.

Aún cuando las fuentes de información sean confiables, y los estudios hayan estado bien diseñados, las conclusiones que se obtengan pueden no ser válidas. Hay muchos factores que influyen en la calidad de los datos y que por lo tanto conducen a conclusiones erróneas. La mala interpretación de la información también es causal de malas decisiones e implementación de políticas incorrectas.

Cuando uno tiene en sus manos los resultados de un estudio, es muy importante conocer cómo fue realizado, antes de aceptarlo y aplicar los resultados ciegamente.

Si los datos fueron recolectados con un cuestionario mal estructurado, difícilmente provean la información buscada. Los respondentes pueden no interpretar las preguntas o falsear las respuestas, dependiendo de la forma en que estén formuladas; incluso un mal ordenamiento de las preguntas puede condicionar las respuestas.

Una encuesta telefónica o por correspondencia (e-mail masivos enviados a una base de datos) generalmentemente tendrá un alto porcentaje de no respuesta, introduciendo un sesgo en las estimaciones: las personas que responden seguramente tienen características diferentes de las que no lo hacen.

El lenguaje es, como en todas las disciplinas, la cuestión básica en la que todos (investigadores, planificadores, tomadores de decisiones) tienen que estar de acuerdo. Manejar un lenguaje común y definir claramente los conceptos (variables, indicadores) es la cuestión primaria de cualquier estudio. No se puede avanzar sin haber “puesto sobre la mesa” las definiciones de cada concepto que va a ser referenciado. De no ser así, los datos van a resultar incomparables: si en una circunscripción disminuye la cantidad de jueces respecto del año anterior, esta disminución no es tal si la circunscripción estaba definida hasta ese momento de manera diferente (puede haberse desdoblado en dos circunscripciones debido, por ejemplo, al incremento poblacional).

Variables y Datos

Los ***datos*** son los hechos y los números que se reúnen, analizan y resumen para su presentación e interpretación. Se obtienen como resultados de observaciones realizadas. Pueden provenir de recuentos (como la cantidad de causas ingresadas en un determinado año) o de mediciones (como la duración del proceso desde la denuncia del hecho hasta la elevación a juicio).

La siguiente tabla muestra un conjunto de datos.

Nro Expediente	Año	Circunscripción	Objeto del Proceso	Sentencia Definitiva	Duración (días)
302	2008	A	Ejecutivo	0	-
93	2008	A	Ejecutivo	1	222
120	2008	A	Ejecutivo	1	221
282	1999	D	Sumario	1	3409
747	2006	B	Cobro de pesos	1	769
298	2008	C	Desalojo	0	-
175	2006	D	Ejecución Fiscal	1	687
456	2006	C	Cobro de pesos	1	636
76	2006	C	Cobro de alquiler	1	651
135	2004	A	Ejecutivo	1	599

Las ***unidades*** son las entidades acerca de las cuales se reúnen los datos. Para el conjunto de datos de la tabla anterior, cada expediente individual es una unidad. Hay en total 10 unidades en este conjunto de datos.

Una ***variable*** es una característica de interés de los elementos. Toma diferentes valores en cada unidad. En la tabla anterior se muestran 5 variables:

- Año: año de inicio del expediente;
- Circunscripción: circunscripción donde se inició el expediente
- Objeto del Proceso
- Sentencia Definitiva: vale 1 si tiene sentencia definitiva dictada, 0 si no

- Duración: mide la duración, en días, desde el inicio del expediente hasta el dictado de sentencia definitiva.

Según sea su naturaleza, las variables, pueden ser clasificadas en ***cuantitativas*** y ***cualitativas***.

Las variables cualitativas son las que identifican un atributo o nombre de la unidad observada. Se refieren a una clasificación, como el objeto del proceso, la circunscripción, el estado de sentencia, el sexo de una persona, etc.

Las variables cuantitativas devuelven valores numéricos, acompañados por una cantidad de medida. Por ejemplo, la duración de un proceso, el monto de una ejecución, la edad de un imputado, etc.

Para medir variables necesitamos una **"escala de medición"**, o simplemente una **"escala"**. En términos técnicos, una escala es *"un conjunto de reglas para cuantificar o asignar calificaciones numéricas a una variable determinada"* (Tuckman, 1972: 142). En términos más simples, las escalas son los instrumentos que usamos para ordenar elementos con el objeto de medirlos y compararlos. En realidad no todas las escalas son numéricas. Diferentes escalas miden distintos tipos de variables. El nivel de sofisticación de un análisis estadístico depende en buena parte de la naturaleza de las variables que se desea analizar. Algunas variables no permiten el uso de escalas de medición muy complejas. El tipo de escala que se use determinará también el nivel de complejidad y sofisticación que podrá alcanzar el análisis estadístico.

Existen cuatro tipos de escalas: Nominal, Ordinal, de Intervalo, y de Razón.

- **Escala Nominal o clasificatoria:** Los números de la escala son meras etiquetas usadas para identificar a las unidades como pertenecientes a una determinada categoría. No hay relación ninguna entre las posibles respuestas a obtener.

Tampoco es imprescindible el empleo de números. Basta con que cada clase se identifique con una etiqueta, como por ejemplo CON

SENTENCIA y SIN SENTENCIA. En particular, esta es una variable dicotómica ya que arroja únicamente dos posibles resultados. Por simplicidad para el análisis se utilizan números en lugar de rótulos. En nuestro conjunto de datos se asignó el valor 1 a la categoría CON SENTENCIA y 0 a SIN SENTENCIA. Sin embargo, esto no la convierte en variable cuantitativa.

- **Escala Ordinal:** Las posibles respuestas son categorías ordenadas de acuerdo a algún criterio. No importa los números utilizados para identificar cada categoría de elementos iguales, basta que preservemos el orden entre ellos. Por ejemplo, al calificar la atención en una Mesa de Entradas, un encuestado podría seleccionar entre las opciones Excelente, Muy Bueno, Bueno, Regular, Malo. Usando esta clasificación es evidente la relación de orden entre las diversas categorías posibles.

- **Escala lineal o de intervalos:** Cuando una escala de medida posee las características de una escala ordinal y además las distancias entre dos intervalos sucesivos es de una determinada medida (constante), se dice que es una escala de intervalos.

Es de un nivel superior de refinamiento. En este caso las diferencias entre valores consecutivos de la escala son siempre iguales. La escala preserva no sólo el orden sino también la distancia entre los elementos con distinta medición. El cero en la escala es por convención. El ejemplo más gráfico de esta escala es el de la temperatura. Que en un lugar haya 0 grados, no significa que no haya temperatura; tampoco tienen sentido los cocientes: si en una ciudad el termómetro marca 22º y en otra 11º, no quiere decir que en la primera ciudad haga 'el doble de calor' que en la segunda. El año de radicación de un expediente también es una variable medida en escala de intervalo.

- **Escala de Razón:** *S*e distingue de la de intervalo porque posee un cero que no es arbitrario*,* es decir*,* que representa la ausencia absoluta de la cualidad que se está midiendo. Ello permite comparar **'razones'** (de ahí el nombre de la escala) o relaciones numéricas del tipo, por ejemplo, "La circunscripción A tiene el triple de superficie que la

circunscripción B", sobre la simple base de que en la escala el valor que corresponde a A es tres veces mayor que el de B.

En nuestro conjunto de datos:

VARIABLE	TIPO	ESCALA DE MEDICION
Año	Cuantitativa	De Intervalo
Circunscripción	Cualitativa	Nominal
Objeto del Proceso	Cualitativa	Nominal
Sentencia Definitiva	Cualitativa	Nominal
Duración	Cuantitativa	De Razón

Dependiendo con qué tipos de datos se cuente, y cuál sea el objeto de interés, se puede hacer un estudio ***transversal***, o de ***sección cruzada*** (los datos están registrados en un mismo instante de tiempo, y corresponden a diferentes unidades muestrales), ***longitudinal*** o ***de serie de tiempo,*** o ***cronológica*** (cuando se hacen registros de una misma unidad muestral a lo largo del tiempo) o de ***panel de datos*** (el conjunto de información contiene tanto características de datos de sección cruzada como de serie temporal).

Si por ejemplo, para un año determinado se registra en un juzgado la cantidad de causas ingresadas, la cantidad de sentencias dictadas y la proporción de causas resueltas por modos anormales de terminación, se está realizando un estudio de tipo transversal para ese juzgado en ese año determinado.

Si una de esas variables es medida a través de los años, por ejemplo la cantidad de causas ingresadas, se puede estudiar su evolución (y eventualmente realizar pronósticos) utilizando un análisis de serie de tiempo sobre esa variable.

Y si además de la evolución de los ingresos se registra también el comportamiento cronológico de las sentencias dictadas y de la proporción de causas finalizadas por modos anormales, la evolución de la unidad (juzgado) puede ser estudiada mediante un análisis de panel.

IV - REPRESENTACIÓN TABULAR Y GRAFICA DE LA INFORMACION

RESUMEN DE DATOS CUALITATIVOS

Una vez obtenidos los datos, es necesario poder representarlos globalmente, sin detenerse en cada dato individual, de manera de poder describir el comportamiento general o analizar diferentes cuestiones como por ejemplo la existencia de subgrupos, datos aislados, o por el contrario, si las observaciones son homogéneas.

Distribución de Frecuencias

La primera forma de sintetizar esa información es a través de la *distribución de frecuencias*, una tabla que ordena los valores de la variable y muestra cuántas veces se repite cada uno de estos valores.

Formalmente se puede definir a la Distribución de Frecuencias como un *resumen tabular de un conjunto de datos que muestra la frecuencia (o cantidad) de unidades en cada una de varias clases que no se superponen.*

Objeto del Proceso

	Frecuencia	Frecuenia Relativa	Porcentaje
Daños y perjuicios	8	,029	2,9
Div. vincular por present. conjunta	2	,007	,7
Ejecución fiscal	31	,114	11,4
Ejecutivo	34	,125	12,5
Prepara vía ejecutiva	5	,018	1,8
Quiebra	1	,004	,4
Sucesión ab-intestato	191	,700	70,0
Tercería de dominio	1	,004	,4
Total	273	1,0	100,0

La **frecuencia relativa** de una clase es la *proporción de la cantidad total de datos que pertenecen a esa clase*. Se calcula dividiendo la frecuencia de la clase entre el total de elementos del conjunto de datos. El porcentaje se calcula multiplicando la frecuencia relativa por 100 y en este caso el todo representa el 100% de las observaciones.

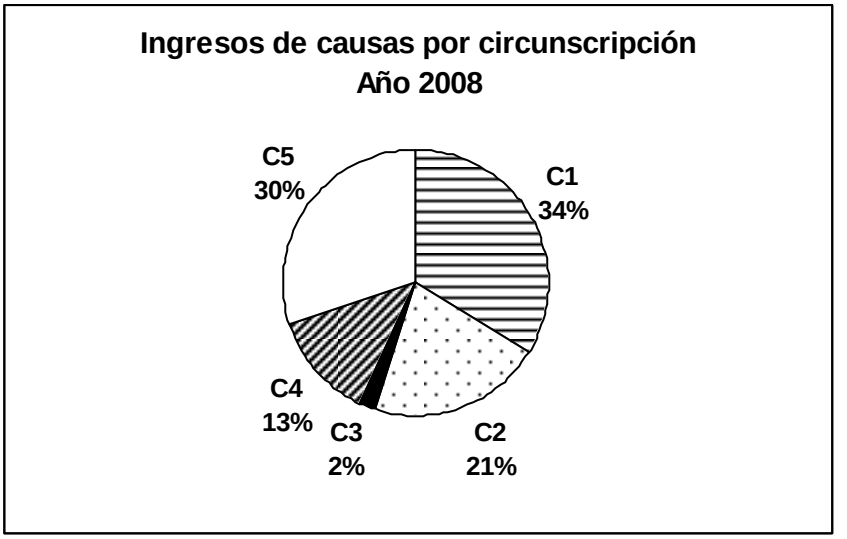
En general, la cantidad de clases en una distribución de frecuencias coincide con la cantidad de categorías en los datos. Sin embargo, cuando aparecen muchas categorías con pocos datos, se recomienda agruparlas en una nueva clase denominada “Otros”. Esto no es muchas veces correcto si en la clase “otros” queda oculta una categoría relevante. Si lo que se quiere mostrar son los tipos de delitos que se comenten en una ciudad, probablemente los homicidios sean poco frecuentes, pero es importante que queden identificados en la tabulación.

Gráficos

Aunque una tabla encierra toda la información disponible, es muchas veces conveniente traducirla a un resumen visual que lo da un buen gráfico. Según el tipo de variable o característica estudiada, se utilizan varios tipos de representación gráfica. A continuación se presentan algunos de los gráficos más utilizados.

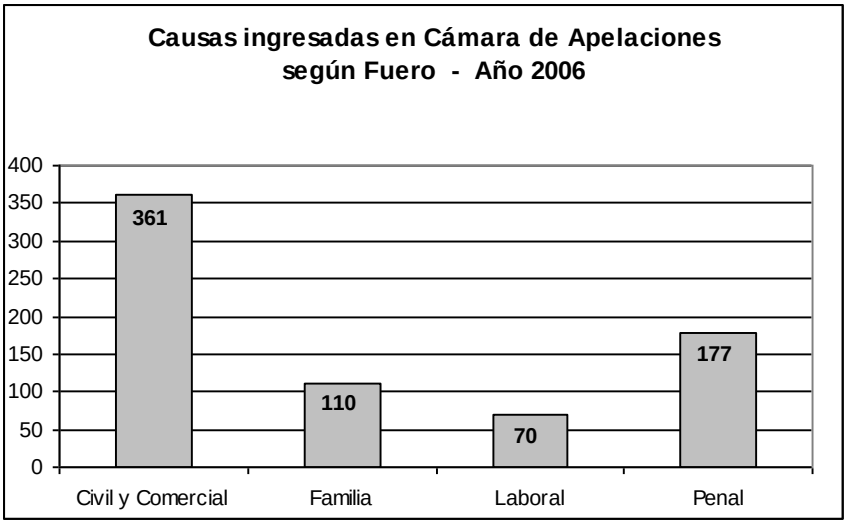
Para mostrar los resultados de variables cualitativas, medidas en escala nominal u ordinal, sirven tanto los diagramas circulares (o de sectores) como los de barras.

Cuando un todo es dividido en sus partes componentes es conveniente un diagrama circular. Como ejemplo:

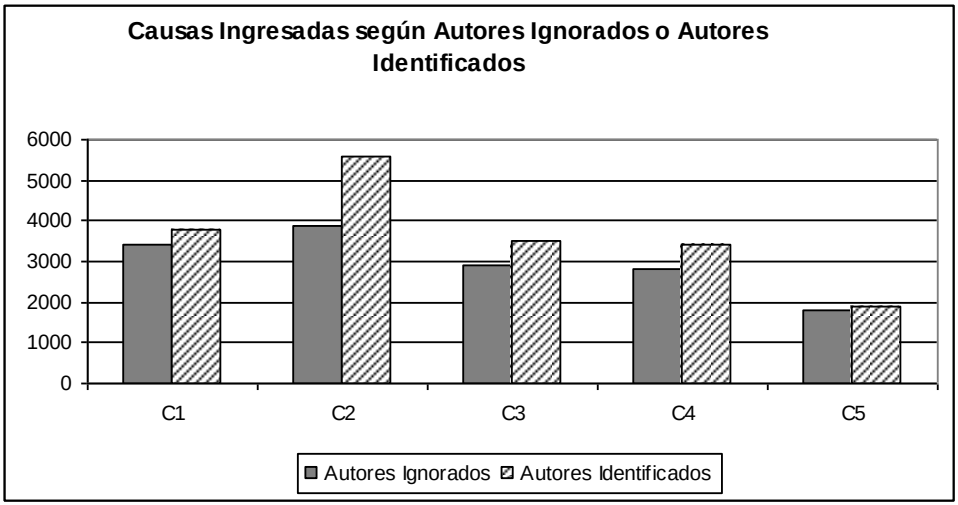


En un diagrama circular, cada uno de los sectores representa una categoría de la variable resumida, y su área es proporcional a la frecuencia de dicha categoría.

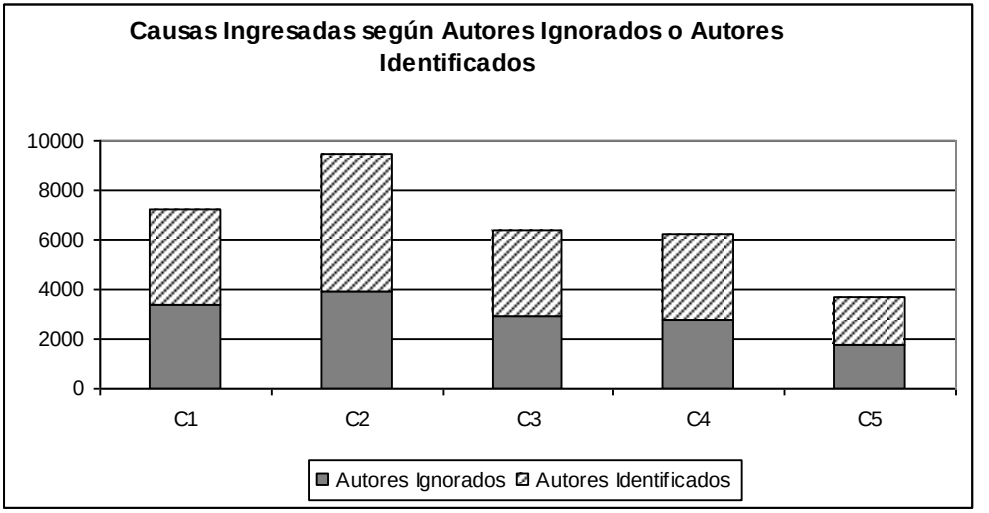
Cuando se quieren representar frecuencias absolutas se recomienda utilizar un gráfico de barras. En este caso se utiliza un sistema de ejes cartesianos; en el eje de abscisas se colocan los valores de la variable (categorías) y en el de ordenadas las frecuencias de los diferentes valores.



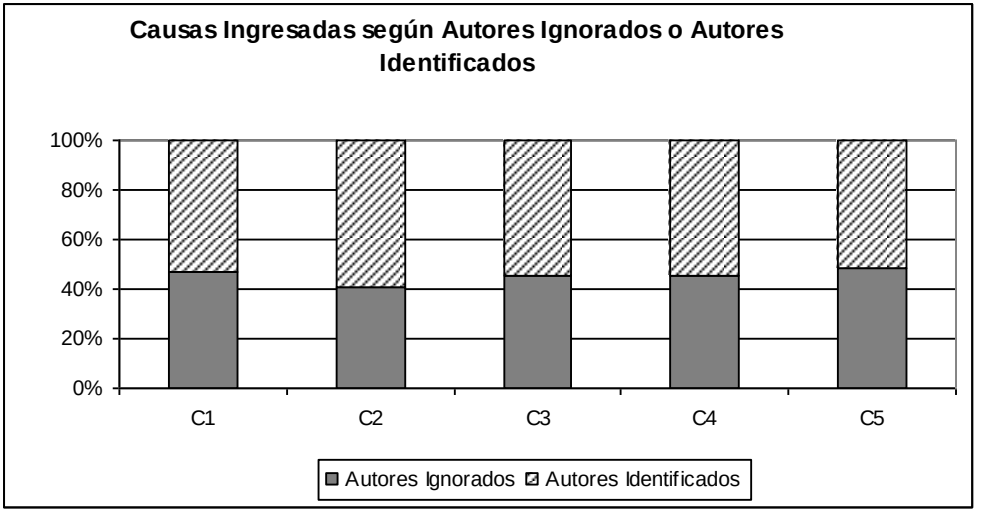
Para comparar un conjunto de datos en 2 o más aspectos se utilizan los diagramas de barras agrupadas



o barras apiladas



Esta relación se puede ver también en un gráfico donde las barras representan el 100% de cada categoría. Esto permite comparar más fehacientemente la distribución de la segunda variable involucrada.



Si bien los dos gráficos muestran la misma información, el segundo resulta más claro a la hora de interpretarla: C5 tiene una mayor incidencia de casos con autores ignorados que C2 o C4. Esto no saltaba a simple vista en el primer gráfico.

El diagrama de barras apiladas también permite mostrar un todo dividido en sus partes, análogo a lo que muestra el diagrama de sectores.

RESUMEN DE DATOS CUANTITATIVOS

Distribución de Frecuencias

La definición de una tabla de distribución de frecuencias, es válida tanto para datos cualitativos como para datos cuantitativos. Sin embargo, la naturaleza misma de los datos hace que el tratamiento sea algo diferente.

Anteriormente se vio que existen variables cuantitativas discretas y variables cuantitativas continuas. Si la variable es discreta, y la cantidad de valores diferentes que toma es pequeña, cada uno de esos valores puede considerarse como una categoría de la variable y la tabla de frecuencias se construye de manera similar a la correspondiente a datos cualitativos.

Cantidad de Imputados por Causa	Frecuencia	Frecuencia Relativa	Frecuencia Acumulada	Frecuencia Relativa Acumulada	Porcentaje
1	47	0,95	47	0,595	59,5
2	19	0,241	66	0,835	24,1
3	7	0,089	73	0,924	8,9
4	5	0,063	78	0,987	6,3
5	1	0,013	79	1	1,3
Total	79	1			100

En esta tabla se presentan dos nuevas columnas que no aparecían en la tabla de frecuencias para una variable cualitativa: Frecuencia

Acumulada y Frecuencia Relativa Acumulada. Estos valores indican cuántas observaciones (o qué proporción) se acumulan o son menores que un determinado valor. Por ejemplo, a partir de la tabla se puede decir que existen 7 causas con 3 imputados cada una, y que existen 73 causas con 3 imputados o menos. Por diferencia también se puede concluir que el 40.5% de las causas registradas tienen más de un imputado.

Si la variable bajo estudio es continua o discreta con muchos valores, conviene definir Intervalos de Clase y luego contar la frecuencia de cada intervalo.

Un intervalo está definido por un Límite Inferior y un Límite Superior, y la frecuencia de cada intervalo cuenta el número de observaciones que están entre esos dos valores.

Hay que tener cuidado al definir las clases *no solapadas* que se usan en la distribución de frecuencias. Por ejemplo, en la siguiente tabla se muestra la duración de un proceso, en días, desde el ingreso al juzgado hasta el dictado de la sentencia definitiva (en los casos que este dictamen exista).

Duración desde ingreso hasta sentencia (días)

	Frecuencia	Frecuencia relativa	Porcentaje acumulado
Menos de 200	88	,32	,32
200 a 400	62	,23	,55
400 a 600	35	,13	,68
600 a 800	24	,09	,77
800 a 1000	15	,05	,82
1000 a 1200	7	,03	,85
1200 a 1400	10	,04	,88
1400 a 1600	5	,02	,90
1600 a 1800	7	,03	,93
1800 a 2000	4	,01	,94
2000 a 2200	1	,00	,95
2200 a 2400	2	,01	,95
2400 a 2600	4	,01	,97
2600 a 2800	2	,01	,97
2800 a 3000	1	,00	,98
Más de 3000	6	,02	1,00
Total	273	1,00	

Si no se establecen bien los límites de las clases, puede haber interpretaciones confusas. Si una causa tiene una duración de 400 días, ¿en qué intervalo se la contabiliza?, ¿en el segundo o en el tercero? Una alternativa es redefinir las clases de la siguiente manera

Clase	Frecuencia
Menos de 200	91
200 a 399	63
400 a 599	39
600 a 799	26
800 a 999	18

Para construir una tabla de frecuencias para datos cuantitativos es necesario determinar la cantidad de clases, definir el ancho de cada clase y luego especificar los límites.

Como lineamiento general, se sugiere utilizar no menos de 5 y no más de 20 clases, dependiendo de la cantidad de datos que se tengan. El objetivo es usar las suficientes para mostrar la variación de los datos, pero no tantas como para que algunas contengan unos cuantos elementos.

En cuanto al ancho de las clases, se sugiere que todas tengan el mismo ancho. Esto reduce la probabilidad de malinterpretación por parte del usuario. Una vez establecida la cantidad de clases, el ancho queda perfectamente determinado por la siguiente relación:

$$\text{Ancho de clases} = \frac{\text{Valor Máximo de los datos - Valor Mínimo de los datos}}{\text{Cantidad de clases}}$$

El ancho de clases puede ajustarse a un valor conveniente y, en general, conviene redondear al entero siguiente. Por ejemplo, si un ancho de clase resulta en 9,3, se podría ajustar a 10, que es un valor más cómodo para trabajar.

En el conjunto de datos que generó la tabla de frecuencias anterior, el valor mínimo era 1 día y el máximo 7251 días. Había 6 datos superiores a 3000 pero muy diferentes entre sí, más bien muy alejados uno de otro, por lo que cualquier definición de ancho de clase iba a generar muchos intervalos

vacíos. Por eso se decidió tomar una clase extrema que incluyera a todos ellos, “Más de 3000”, y así manejar el rango de 1 a 3000. Al establecer arbitrariamente la utilización de 15 intervalos, la amplitud de cada uno de ellos resultó en

$$\frac{3000-1}{15} = 199{,}93$$

es decir, redondeando, una amplitud de 200 días para cada clase.

Es importante notar que dos personas con diferente criterio, pueden generar distintas tablas de frecuencias, debido a que optan por otro número de intervalos y consecuentemente otra amplitud de clase. Más aún, una misma persona puede construir diferentes tablas y quedarse con aquella que mejor le representa la distribución de los datos en cuestión.

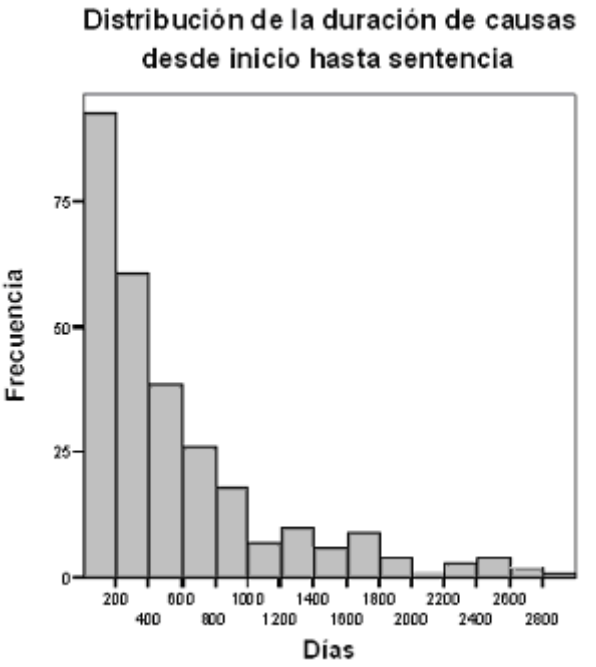
Gráficos

La tabla de frecuencias puede ser representada gráficamente utilizando distintos tipos de gráfico según la información que se quiera mostrar.

Si la variable es cuantitativa discreta con pocos valores, y por lo tanto no están agrupados, se utiliza un diagrama de barras como en el caso de la variable cualitativa.

Cuando se quieren mostrar las frecuencias absolutas o relativas de una variable continua, se puede utilizar un *Histograma* o un *Polígono de Frecuencias*.

Un histograma es a simple vista, similar a un gráfico de barras, pero se diferencia en que en el eje de abscisas se representan los intervalos, uno a continuación del otro, y las barras son rectángulos adyacentes. Tiene en cuenta la continuidad de los datos.

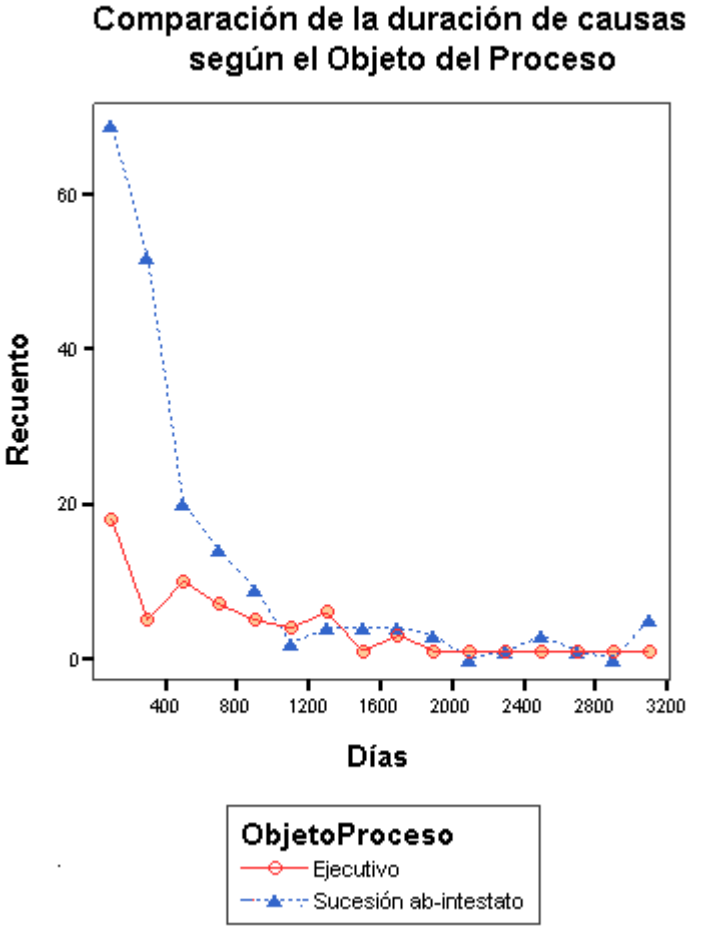


Si uno quisiera sacar conclusiones a partir del histograma podría decir que los datos están concentrados en el extremo izquierdo del gráfico. Esto significa que lo más frecuente es encontrar causas que duren menos de 200 días desde su inicio hasta el dictado de sentencia. Duraciones más largas son cada vez menos frecuentes, y hay causas, las menos, que tardan mucho en ser resueltas de ese modo, digamos, más de 2000 días.

Hay seis valores superiores a 3000 que no se han mostrado en el gráfico ya que al ser muy diferentes entre sí en un rango muy amplio, los intervalos que restan quedan vacíos o con un único valor. No tiene sentido llevar el eje de abscisas hasta el valor del máximo ya que el gráfico no mostraría absolutamente nada.

Si se quiere comparar la distribución de dos conjuntos de datos, conviene utilizar un polígono de frecuencias absolutas. Los puntos indican la

frecuencia de cada intervalo y cada intervalo está representado por el punto medio (que se denomina “marca de clase”). Estos puntos se unen mediante una poligonal, de allí el nombre del gráfico.



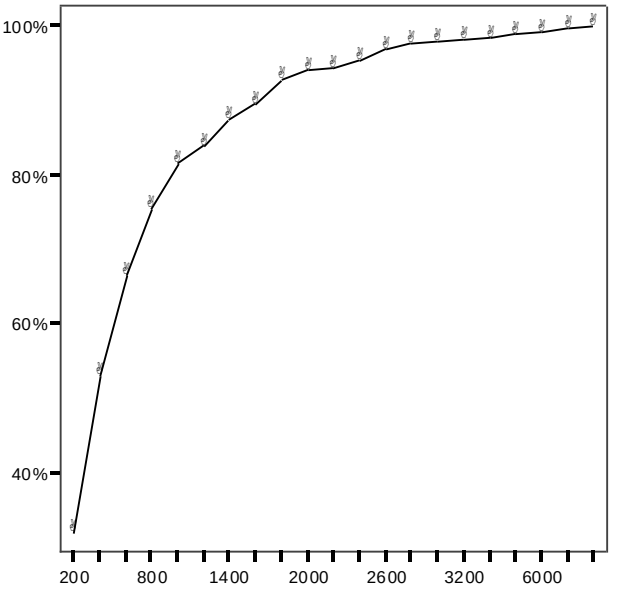
Al observar el gráfico se puede concluir que los juicios ejecutivos son en general más cortos que las sucesiones (notar que a partir de un valor cercano a 2000 la frecuencia es cero), y que dentro de ese rango (hasta

2000 días) la distribución de la duración es similar. Cuando se tienen conjuntos con diferente cantidad de datos es conveniente construir un polígono de frecuencias relativas que simplifica la interpretación.

Otro gráfico muy útil a la hora de describir una variable cuantitativa es el polígono de frecuencias relativas acumuladas, denominado *Ojiva*. También se puede construir en términos de porcentaje.

Ojiva de la variable "Duración"

Porcentaje acumulado



Este gráfico permite ver rápidamente el porcentaje de causas que duran “menos que” una determinada cantidad de días.

En este ejemplo, el 60% de las causas duran menos de 450 días.

ANALISIS DESCRIPTIVO BIVARIADO

Hasta ahora se ha focalizado en los métodos tabulares y gráficos que se utilizan para describir los datos relativos a una única variable, pero muchas veces es interesante comprender la relación entre dos variables ya sean ambas cualitativas, cuantitativas, o una cualitativa y una cuantitativa.

Es decir, las tabulaciones cruzadas y los diagramas de dispersión sirven para resumir los datos en forma tal que ayude a revelar la relación entre dos variables, mostrar su comportamiento simultáneo.

Tablas de Contingencia

Suponga que se quiere mostrar la cantidad de causas ingresadas por fuero (penal o no penal) en una provincia, para un determinado año, y además ver cómo se distribuye esa cantidad de acuerdo a las diferentes circunscripciones.

Se está ante dos variables cualitativas medidas en escala nominal: *Fuero* (con sus categorías Penal y No Penal), y *Circunscripciones*, que en este caso tomará 5 valores: C1, C2, C3, C4 y C5.

En el año 2006 ingresaron en total 43777 causas, distribuidas de la siguiente manera según las variables de interés:

CIRCUNSCRIPCION	FUERO		Total
	Penal	No Penal	
C1	2903	2234	**5137**
C2	6787	9015	**15802**
C3	7268	6759	**14027**
C4	4785	3141	**7926**
C5	551	334	**885**
Total	**22294**	**21483**	**43777**

.

En las columnas figuran las categorías de la variable Fuero y en las filas las categorías de la variable Circunscripción. Cada casilla del cuerpo de la tabla muestra la cantidad de causas que responde simultáneamente a dos características, una de cada variable. Los totales de filas y columnas se denominan *totales marginales.* Así, por ejemplo, en la circunscripción C4 ingresaron 4785 causas penales, en C3 6759 no penales y en total ingresaron 885 causas en la circunscripción C5.

Diagramas de Dispersión

Un diagrama de dispersión es una representación gráfica de la relación entre dos variables cuantitativas. Según el patrón que determine se puede inferir si las dos variables están relacionadas o no, y qué tipo de relación presentan.

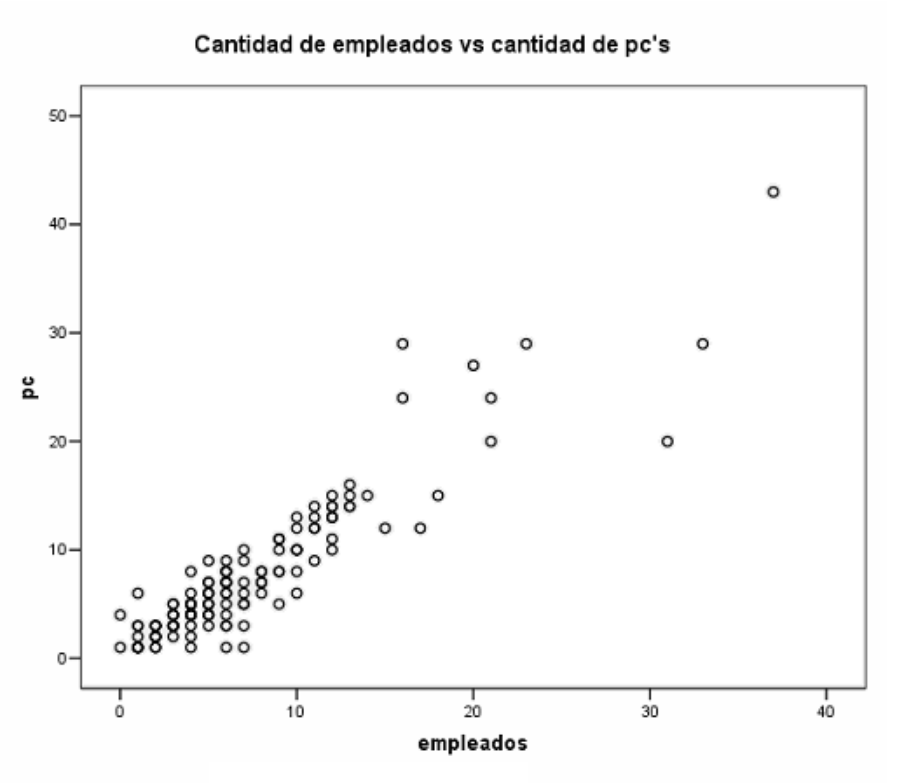
Es un gráfico con dos ejes, uno horizontal y otro vertical. En cada uno de ellos se ubican los valores de cada variable, cada eje representa una de las dos variables cuantitativas involucradas.

Como ejemplo, se toman la cantidad de Pc's que hay en cada organismo y la cantidad de empleados. Sería interesante ver si existe alguna relación entre estas dos variables. El sentido común sugiere que si los

recursos están bien distribuidos, en los organismos con más empleados debería haber más computadoras.

Organismo	Cantidad de pc	Cantidad de empleados
J1	8	6
J2	15	12
J3	5	9
J4	13	12
J5	10	10
J6	8	8
J7	6	7
J8	24	21
J9	16	13
J10	5	7
J11	1	2
J12	6	6
J13	8	9
J14	9	7
J15	7	6
J16	6	5
J17	6	5
J18	5	3
J19	12	11
J20	14	12
J21	15	13
J22	11	9
J23	13	12
J24	24	16
J25	11	12
J26	5	4
J27	43	37
J28	5	7
J29	7	8

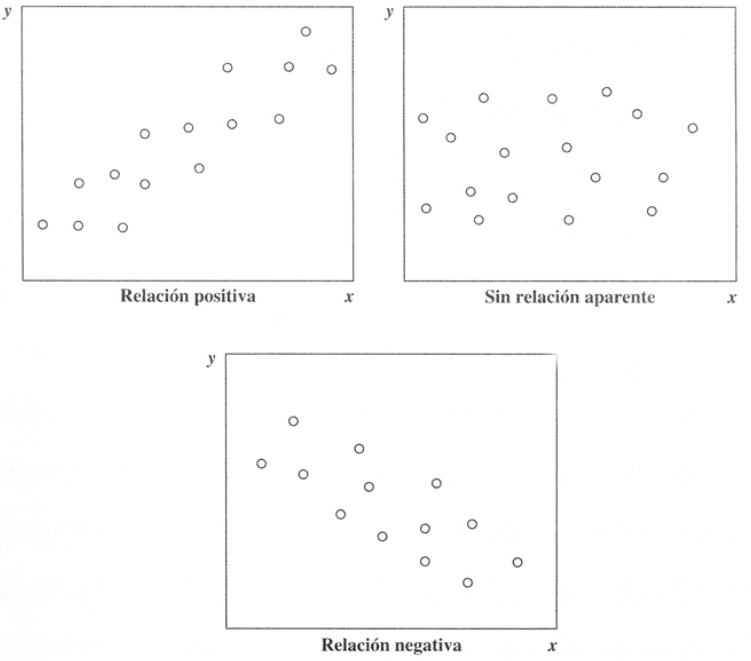
Ubicando la variable “Cantidad de empleados” en el eje horizontal y “Cantidad de pc’s” en el eje vertical, se obtiene el siguiente gráfico. Cada punto representa un organismo que tiene una determinada cantidad de empleados y una cantidad determinada de pc’s.



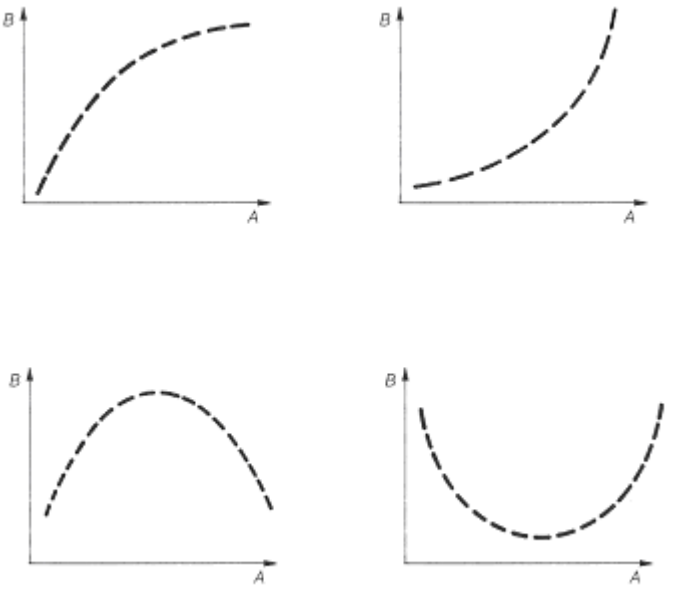
A partir del gráfico se ve que hay una relación directa entre ambas variables. Esto es, a medida que la cantidad de empleados en un organismo aumenta, también aumenta la cantidad de computadoras asignadas.

La relación no es perfecta porque no todos los puntos están sobre una línea recta. Sin embargo, el comportamiento general de los puntos sugiere que la relación general es positiva.

A continuación se muestran algunos patrones que pueden darse a partir de la relación entre dos variables. El primero muestra una relación lineal positiva, semejante a la que se observó en el ejemplo de las pc's por organismo. El segundo no muestra relación aparente entre las variables. El tercero muestra una relación lineal negativa, es decir, donde una variable tiende a disminuir a medida que la otra aumenta.



Las relaciones no necesariamente son de tipo lineal. Algunos patrones no lineales pueden tener las siguientes formas



V - MEDIDAS ESTADISTICAS DESCRIPTIVAS DE LA INFORMACION

Los métodos descriptivos gráficos tienen como función principal hacer que el usuario de un informe aprecie de manera rápida cómo están distribuidas las observaciones. Sin embargo estas técnicas gráficas presentan limitaciones en cuanto a la descripción y análisis de un conjunto de datos. Por otra parte las técnicas gráficas no son apropiadas para hacer inferencias, aunque sí pueden servir como punto de partida.

Se van a estudiar tres tipos de medidas que caracterizan un conjunto de datos: las medidas de *posición*, las medidas de *variabilidad* y las medidas de *asociación*.

Si estas medidas se calculan a partir de los datos de una muestra, se denominan *estadísticos*, y si se calculan a partir de los datos de la población completa se llaman *parámetros*.

Uno de los objetivos finales de la estadística es estimar el valor de un parámetro a partir de un estadístico.

MEDIDAS DE POSICIÓN

Media Aritmética o Promedio

El promedio es una medida conocida por todos, que no implica grandes algoritmos para calcularlo. Simplemente se obtiene sumando todos los valores y dividiendo esta suma por la cantidad de datos.

Si se quiere calcular el promedio de los siguientes valores, que representan las edades de los empleados de un juzgado: 26, 34, 33, 38, 40, 33, 22, 24, se debe hacer:

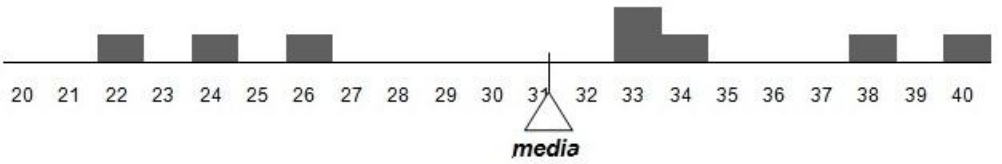
$$\mu = \frac{26+34+33+38+40+33+22+24}{8} = \frac{250}{8} = 31.25$$

La letra griega μ (mu) se utiliza para denotar la media aritmética, esto si se trata de la media de la población completa (parámetro). Si se hace referencia a la media de una muestra (estadístico), se la denota con el símbolo $\bar{x}$.

La edad promedio de los empleados de ese juzgado es de 31.25 años.

Como se ve, no necesariamente la media aritmética debe coincidir con un valor del conjunto de datos, ni ser un valor que tenga sentido estricto respecto de la variable medida. Si en lugar de edades, esos fueran valores correspondientes a cantidades de escritorios en 8 organismos diferentes, nadie podrá decir que hay "un cuarto de escritorio" en alguna oficina. El promedio nos da idea de alrededor de qué valor está centrado el conjunto de datos.

En este caso, el conjunto de datos puede representarse de la siguiente manera sobre un eje numérico:



La media puede pensarse como el punto de equilibrio del conjunto: hay algunos datos menores que ella, otros mayores, pero en el promedio el conjunto "está en equilibrio".

No siempre el promedio es el mejor representante del conjunto de datos. Se puede pensar en dos situaciones diferentes.

Suponga tener que dar un promedio para la duración de los procesos penales, desde el ingreso del caso hasta el archivo. Si se trata con casos de Autores Ignorados, puede ser que en 15 días el caso esté archivado porque no hay fundamentos para formalizar una investigación, pero si un caso se investiga puede pasar muchísimo tiempo hasta que se archive (sin tener en cuenta las diversas formas en que puede llevar adelante el proceso, esto es a los únicos efectos de ejemplificar sencillamente la situación).

Si se toma una muestra de casos, y se registra la duración en días desde el ingreso hasta el archivo, uno se podría encontrar con, por ejemplo, los siguientes datos:

150 5 12 176 200 13 187 220 201 15

Calculando la media de estos 10 datos, se llegaría a la conclusión de que, en promedio, un proceso penal tarda 117.9 días en llegar a archivo. Pero este valor no es para nada representativo de la realidad. Claramente hay en los datos dos conjuntos diferentes: casos con autores ignorados y casos con autores identificados, y como tales, deben ser tratados como conjuntos diferentes a la hora de buscar un valor representativo de la duración del proceso.

Así, se podría concluir que los casos con autores ignorados tardan, en promedio, 11.25 días en ser archivados, mientras que en los casos con autores identificados la demora promedio es de 189 días.

Otra situación en que la media no es un buen representante de un conjunto de datos es en el caso en que existen valores extremos o “outliers”. Estos valores distorsionan el centro de equilibrio y la media se desplaza hacia uno de los extremos.

Suponga que, además de los datos anteriores, hay un proceso que demora 500 días en ser archivado. Ahora el promedio de duración de los casos con autores identificados es de

$$\frac{150+176+200+187+220+201+500}{7} = 233.43$$

La media, que anteriormente valía 189 días, se desplazó a 233.43, debido a un único valor muy diferente a los demás. Es representativo este valor? No; más aún, es superior a todos los valores que estaban en el conjunto original de datos.

La media aritmética, si bien es de muy fácil interpretación y cálculo, es muy sensible a la presencia de datos extremos. Esto implica que hay que ser muy prudente al utilizarla como representante de un conjunto de datos.

Si hay datos anómalos, es aconsejable utilizar otra medida descriptiva de tendencia central, que a continuación vamos a presentar.

Mediana

La mediana es otra medida de posición, que se encuentra precisamente en el centro del conjunto de datos ordenado en forma ascendente. Es el valor del elemento intermedio.

Esto significa que divide al conjunto de datos ordenado, en dos porciones iguales en cuanto a cantidad de datos: al menos el 50% de los valores es inferior o igual a la mediana, y al menos otro 50% es superior o igual a la mediana.

Si hay una cantidad impar de valores, la mediana será el valor central, pero si la cantidad es par, hay que buscar un valor intermedio entre los dos centrales.

Volviendo al ejemplo de la duración de los procesos, se calcula la mediana. Los datos son

150 176 200 187 220 201

que ordenándolos resultan en

150 176 187 200 201 220

Este conjunto tiene 6 datos, por lo que el valor central debería dejar antes y después de él 3 valores.

150 176 187 | 200 201 220

Mediana

Por convención, se promedian los dos valores centrales, y así la Mediana es

$$\frac{187+200}{2}=193.5$$

Este valor no difiere tanto de la media, que era de 189 días.

Sin embargo, considérese ahora el conjunto con el dato agregado de 500 días. La media se había desplazado al valor 233.43, qué ocurre con la mediana?

150 176 187 **200** 201 220 500

Ahora hay 7 valores en el conjunto, por lo tanto el valor central es el que está en la posición 4, esto es, el 200.

La mediana, que originalmente era de 193.5 días, ahora es de 200 días. Sufrió una leve modificación, pero sigue cumpliendo con su papel de "buen representante" del conjunto de datos. Esto no ocurría con el valor de la media.

La mediana es un estadístico "robusto" a la presencia de datos extremos. Esto significa que cuando hay valores extremos, conviene informar a partir de la mediana y no del promedio. O, informar ambos para complementar el análisis.

Moda

La Moda es el valor de los datos que tiene mayor frecuencia, es decir, el que se repite más cantidad de veces.

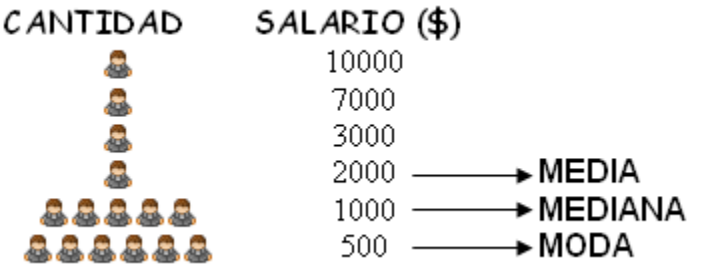
Puede ser que en un conjunto no haya ninguna moda o, por lo contrario, que haya más de una. En el ejemplo de la duración de los procesos no hay ninguna moda, pero en el ejemplo de las edades de los empleados, el valor modal es 33.

Muchas veces se confunde el sentido de la moda y se lo asocia al comportamiento de la mayoría. Pero esto no es así: el hecho de que la moda sea 33 no significa que la mayoría de los empleados tiene 33 años, sino que la edad más común de ese grupo es de 33 años.

Las tres medidas descriptas hasta ahora son todos estadísticos descriptivos y se utilizan cuando se trata de facilitar un “retrato” rápido de un conjunto de datos mediante una indicación aproximada del centro del conjunto. Se diferencian entre sí en que cada uno se obtiene utilizando una definición ligeramente diferente del término “centro” (aunque en el caso de la moda, no necesariamente el valor más típico habrá de ubicarse en el centro). Ya se expusieron las ventajas y desventajas del uso de cada una de ellas, y de aquí en más es cada usuario el que debe decidir cuál es el más adecuado a la hora de informar: “El valor seleccionado, proporciona una descripción adecuada de la situación que presentan los datos?”

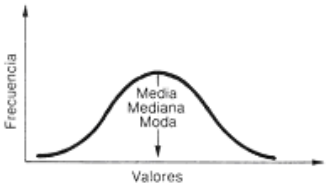
También es necesario estar atento a los informes que han sido publicados por otras personas, ya que el desconocimiento o la falta de honestidad pueden llevar a utilizar una cifra “promedio” que resulte engañosa.

Observando la siguiente figura, esto queda claro. Dependiendo de si el que quiere informar es la prensa o el Poder Judicial, seguramente intentarán usar diferentes estadísticos para dar un valor representativo del sueldo de los integrantes de un organismo.

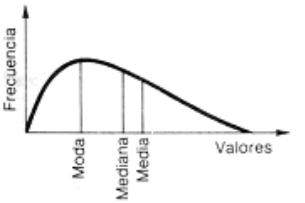


Cuando se estudiaron los gráficos estadísticos, se observó que las distribuciones podían tener infinidad de formas, de acuerdo a la naturaleza de los datos.

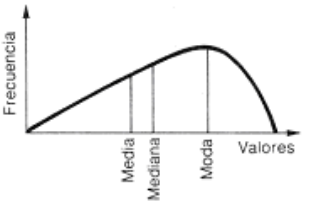
De acuerdo a la forma de la distribución es la relación que existe entre la media, mediana y moda, por lo que el observar la forma de la distribución también puede ayudar a identificar el estadístico adecuado.



simétrica



asimétrica a derecha



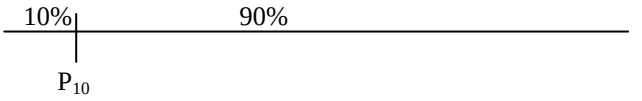
asimétrica a izquierda

Percentiles

Los percentiles son, en cierta forma, una extensión de la mediana. Son aquellos números que dividen al conjunto de datos ordenado, en partes que contienen un determinado porcentaje de las observaciones.

Así, el percentil 10 (P_{10}) deja por debajo de él al menos un 10% de las observaciones, y por encima un 90% o más.

Al conjunto de datos ordenados, lo divide de la siguiente manera:



Si al calcular los percentiles de la variable "salario" resulta que P_{10} =800, esto significa que alrededor del 10% de los empleados ganan $800 o menos, y alrededor del 90% tiene un salario igual o superior a ese valor.

Si se tienen 10 datos ordenados

2 5 8 10 12 12 21 23 30

un solo dato representa el 10%, entonces el percentil 10 es el valor "2". Significa que el 10% de los datos es menor o igual que 2 y el 90% es mayor o igual que 2.

Un conjunto particular de percentiles, son los que dividen al conjunto de datos en "cuartos". Se denominan **cuartiles**. El 1er Cuartil coincide con P_{25} y deja por debajo de él al 25 % de las observaciones. El 2do Cuartil (Q_1) coincide con P50 y la Mediana, y deja por debajo y por encima el 50% de las observaciones. El 3er Cuartil (Q_3) coincide con P_{75} y es superior al 75% de los datos; por encima de él queda un 25%.

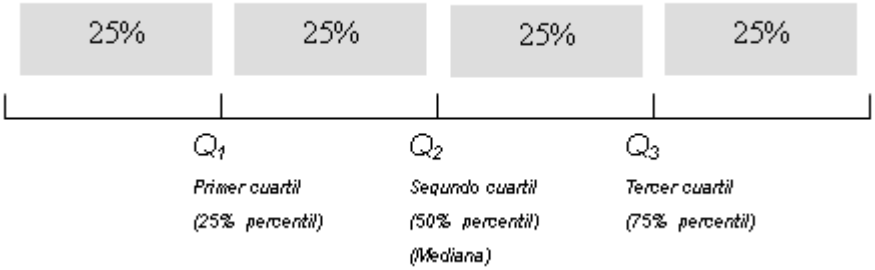
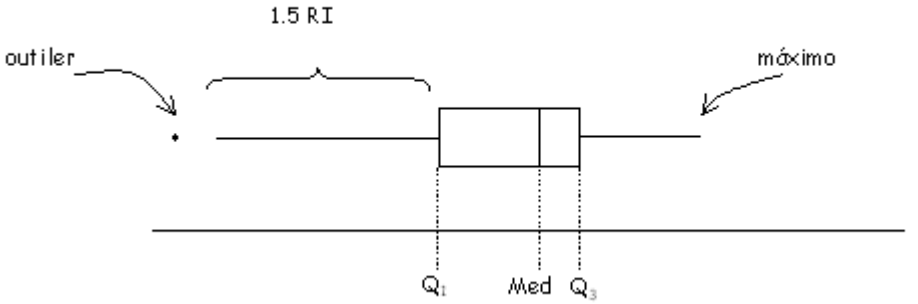


Diagrama de Caja (Box-Plot)

Un *diagrama de caja* es una forma de resumir en una gráfica los datos. La base para construir este diagrama es el cálculo de la Mediana y los cuartiles Q_1 y Q_3. También se usa el rango intercuartil RI = $Q_3 - Q_1$, que permitirá identificar los valores atípicos.

En la caja central del gráfico están contenidas el 50% de las observaciones. El borde inferior coincide con el primer cuartil y el superior con el tercero. La línea que aparece en el interior de la caja representa la Mediana.

Las líneas que salen de la caja se denominan *bigotes* y se extienden hasta 1.5RI (una vez y media el rango intercuartil) hacia cada lado. Si el valor mínimo o máximo se encuentran dentro de esos límites, el bigote llega sólo hasta ese punto.



En este caso se graficó en forma horizontal, pero también puede orientarse en sentido vertical.

Una aplicación de los conceptos desarrollados

El programa estadístico SPSS (como otros tantos software específicos o inclusive, y en algunos casos, una planilla de cálculo como Excel) tiene incorporadas todas estas herramientas, por lo que no debería preocupar el cálculo manual de los estadísticos o la construcción de los gráficos, sino más bien el saber qué tipos de procedimientos solicitar e interpretar correctamente las salidas que el software provea.

Tomemos un ejemplo:

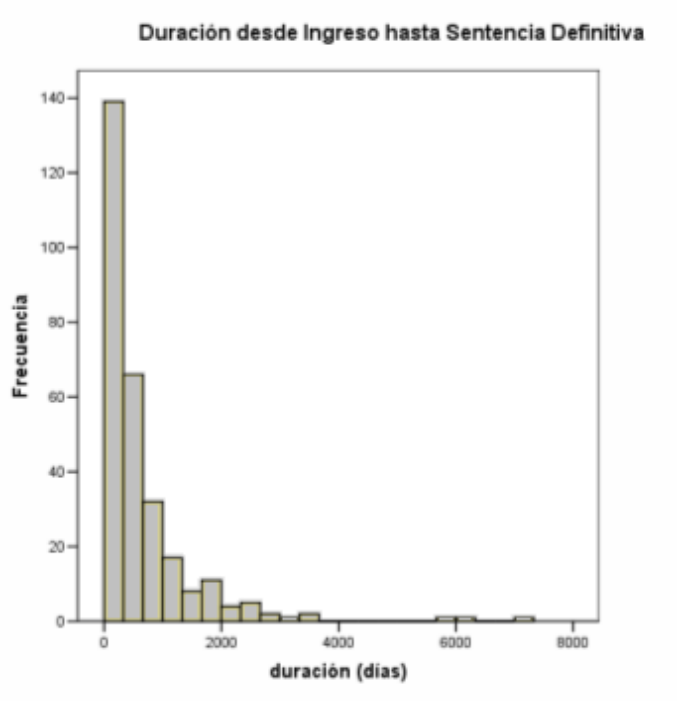
Supongamos que tenemos 290 expedientes provenientes de un juzgado civil, para los cuales se midió la duración desde el ingreso al juzgado hasta el dictado de sentencia definitiva.

Para visualizar rápidamente los datos, se genera una tabla de frecuencias como la que sigue:

Duración desde Ingreso a Sentencia (días)

	Frecuencia	Porcentaje	Porcentaje acumulado
Menos de 200	91	31,4	31,4
200 a 399	63	21,7	53,1
400 a 599	39	13,4	66,6
600 a 799	26	9,0	75,5
800 a 999	18	6,2	81,7
1000 a 1199	7	2,4	84,1
1200 a 1399	10	3,4	87,6
1400 a 1599	6	2,1	89,7
1600 a 1799	9	3,1	92,8
1800 a 1999	4	1,4	94,1
2000 a 2199	1	,3	94,5
2200 a 2399	3	1,0	95,5
2400 a 2599	4	1,4	96,9
2600 a 2799	2	,7	97,6
2800 a 2999	1	,3	97,9
3000 ó más	6	2,1	100,0
Total	290	100,0	

Un histograma para estos datos es



Calculando los estadísticos descriptivos se obtiene:

Estadísticos

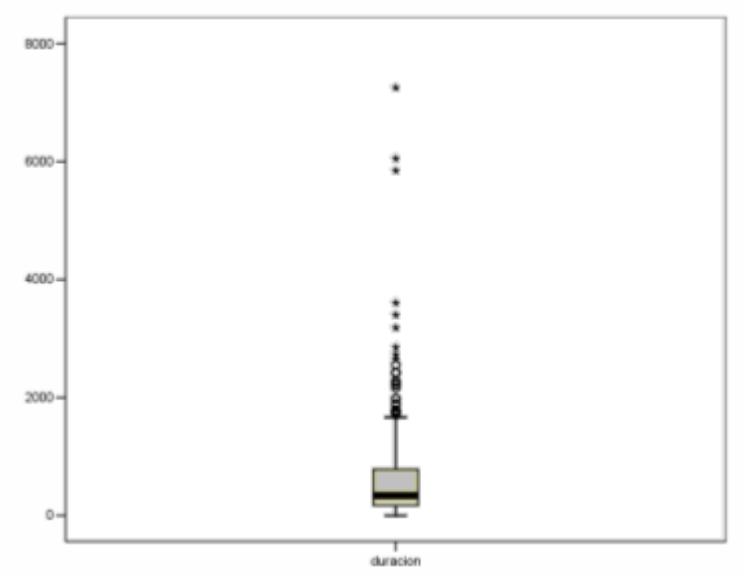
duracion

N	Válidos	290
	Perdidos	0
Media		663,43
Mediana		340,50
Moda		163
Mínimo		1
Máximo		7251
Percentiles	25	169,00
	50	340,50
	75	785,00

Algunas de las conclusiones que podemos extraer de acá son:

- El valor de la Media es mucho mayor al de la Mediana. Esto se debe a la presencia de valores extremos a la derecha (altos).
- Un 25% de las causas tienen una duración menor o igual a 169 días.
- Un 25% tiene una duración mayor o igual a 785 días. Considerando que el máximo es de 7251 días, esto sugiere la presencia de valores muy extremos. Se condice con lo observado en el histograma. También es coherente con la diferencia observada entre la media y la mediana.

El Box Plot evidencia una distribución asimétrica y la presencia de valores extremos



El comportamiento de la variable “duración” está ligado al objeto de proceso y por eso conviene tal vez hacer un estudio por cada uno de los grupos que quedan determinados según esta segunda variable.

En la siguiente tabla se presenta la cantidad de causas, la media y la mediana de duración, según el objeto de proceso.

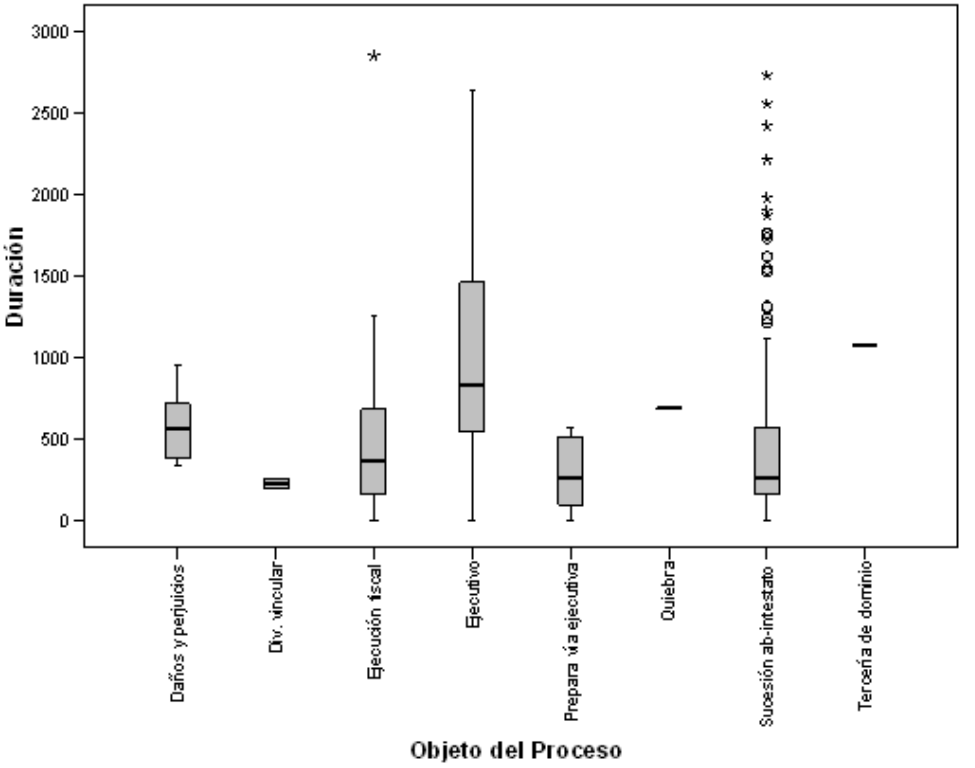
Resúmenes de casos

duracion

ObjetoProceso	N	Media	Mediana
Daños y perjuicios	8	576,63	559,50
Div. vincular por present. conjunta	2	228,00	228,00
Ejecución fiscal	31	528,16	371,00
Ejecutivo	33	1033,06	830,00
Prepara vía ejecutiva	5	291,20	266,00
Quiebra	1	690,00	690,00
Sucesión ab-intestato	186	482,48	260,00
Tercería de dominio	1	1075,00	1075,00
Total	267	556,16	322,00

Hay mucha dispersión entre los valores promedio y también entre las medianas. Esto sugiere que la segmentación de los datos es adecuada. Es decir, que hay que estudiar separadamente las duraciones de los procesos.

El siguiente es un gráfico de cajas comparativo que muestra rápidamente las diferencias entre los grupos. Si bien en este gráfico están todos los objetos representados, hay que tener cuidado con aquellos que presentan un único dato o dos, como es el caso de los Daños y Perjuicios, la Quiebra y la Tercería de Dominio; una única unidad de una población nunca puede representar a la población completa, por lo que no se pueden sacar conclusiones generales respecto de la duración de causas correspondientes a estos objetos del proceso.



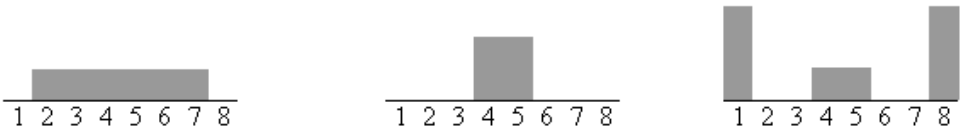
A partir del gráfico se puede observar

- Cuáles son los objetos del proceso más uniformes en cuanto a su duración y cuáles son los menos uniformes: los grupos más dispersos son los que corresponden a los Juicios Ejecutivos y las Sucesiones.
- Los datos anómalos que se observaban a priori, en el conjunto completo, corresponden a las Sucesiones.
- En las Ejecuciones Fiscales hay un dato muy extremo, totalmente diferente del resto del conjunto. Amerita analizar ese valor ya que podría corresponder a un dato mal cargado.

- Los procesos de mayor duración en general son los correspondientes a Juicios Ejecutivos, seguidos por los Daños y Perjuicios. El resto de las cajas se “solapan”, lo que sugiere que la diferencia numérica observada en la media y la mediana puede deberse simplemente al carácter muestral del conjunto de datos y no a una diferencia real atribuida al objeto del proceso.

MEDIDAS DE DISPERSIÓN

En los puntos anteriores se han mostrado algunas maneras de describir conjuntos de datos o distribuciones, dando una ubicación aproximada del “centro” del conjunto. Esta única medición no alcanza para describir acabadamente los datos. Pueden existir conjuntos diferentes con la misma media, por ejemplo, pero cuyas distribuciones tengan forma completamente diferente.



Estos tres conjuntos de datos tienen la misma media (4.5). Sin embargo las distribuciones no son en absoluto parecidas. En un caso (el histograma del centro) las observaciones están concentradas alrededor del promedio, y en los otros dos están más dispersas. Incluso, la dispersión no es la misma en ambos.

Es necesario presentar entonces alguna medida que refleje la ubicación de todos los datos respecto de la media. Esto es, una medida de la dispersión o variabilidad de las observaciones.

En principio, intuitivamente se podría pensar que una variabilidad pequeña es más deseable que una variabilidad grande. Es más fácil describir y sacar conclusiones acerca de conjuntos de datos bastante concentrados, donde las observaciones son parecidas entre si.

Si una variabilidad es cero, significa que todos los datos son iguales.

Así como hay diferentes medidas de posición, existen también diferentes medidas de variabilidad, según el objetivo que se persiga con el estudio.

Amplitud o Rango

Es una medida muy sencilla de variabilidad. El Rango informa acerca de cuántos son, en conjunto, los valores sobre los que se extiende una distribución. Se obtiene simplemente restando el valor máximo menos el mínimo.

Es muy simple de calcular, pero a la vez, el hecho de que sólo tome en cuenta dos valores de todo el conjunto de datos, la hace poco precisa.

En el ejemplo presentado en el punto anterior, el valor mínimo de la duración de los procesos era de 1 día, mientras que el máximo resultó en 7251 días.

Por lo tanto: $Rango = 7251 - 1 = 7250$

Este valor es demasiado grande y no describe adecuadamente la variabilidad de los datos; recuérdese que el 75% de ellos toma valores menores a 785 días.

Desviación Media

La *desviación media* indica cuánto es, en promedio, la distancia entre los valores y la media aritmética.

Suponga que se consideran los números

8 9 10 11 12

La media es 10 y el rango es 4.

El número 8 dista en dos unidades de la media, igual que el número 12. El 9 y el 11 difieren en una unidad. Las distancias o desvíos de cada dato respecto de la media son entonces

2 1 0 1 2

Para dar un único valor que represente a todos los desvíos, se calcula el promedio

$$\frac{2+1+0+1+2}{5}=\frac{6}{5}=1.2$$

Así, 1.2 es, en promedio, la cantidad de unidades que los datos se desvían de la media.

Esta medida, a diferencia del Rango, además de incluir todos los datos, también tiene en cuenta una medida de posición.

Hay que notar, que al calcular las desviaciones no se tuvo en cuenta el signo: como hay algunos valores mayores que la media, y otros menores, las diferencias serán, algunas positivas y otras negativas.

Si se tiene en cuenta el signo de las diferencias, estos errores se compensan y la suma (y por lo tanto el promedio) dan cero. Por eso es necesario tomar el valor absoluto de las diferencias, esto es, todas las diferencias con signo positivo. Además, como lo que interesa es la distancia a la media, el signo no tiene importancia, sólo el valor absoluto del número.

La desviación media no tiene muchas propiedades deseables desde el punto de vista matemático y eso complica su utilización en cuestiones

más avanzadas. Por ello, es universalmente reemplazada por otras medidas que son más “amigables” a la hora de desarrollar y aplicar los algoritmos para el cálculo. Estas medidas son la *Varianza* y el *Desvío Estándar*.

Varianza y Desviación Estándar

Una forma de que los desvíos sean independientes del signo es elevarlos al cuadrado, y esto es lo que se hace al calcular la varianza.

Si se calcula la diferencia entre cada dato y la media (nótese que al definir la desviación media antes se habló de distancia, y no de diferencia, por eso se omitía el signo), obtenemos:

-2 -1 0 1 2

La suma de estos números da cero, por lo tanto los desvíos conjuntamente con su signo, no pueden ser utilizados para calcular una medida de variabilidad (cualquier conjunto de datos tendría variación nula).

Al elevarlos al cuadrado quedan todas medidas positivas

4 1 0 1 4

y el promedio de estos valores, denominado *varianza*, es

$$\sigma^2 = \frac{4+1+0+1+4}{5} = 2$$

Nótese que σ^2 se lee “sigma cuadrado”.

El valor de la varianza es 2, bastante superior al de la desviación media. Esto no parece lógico si se trata del mismo conjunto de datos. Al calcular la varianza, se elevaron todas las diferencias, junto con sus unidades, al cuadrado. Dar conclusiones en “unidades al cuadrado” no tiene sentido. Por otro lado, al tomar el cuadrado, la diferencia real entre el dato y la media se magnifica.

Este hecho conduce a la definición de otra medida de variabilidad que sí puede ser medida en las mismas unidades de los datos, y por lo tanto tiene sentido su interpretación: la *desviación estándar.*

La desviación estándar se calcula simplemente tomando la raíz cuadrada de la varianza.

Es decir, en este caso la desviación estándar es

$$\sigma = \sqrt{2} = 1.4142...$$

Este valor es similar al obtenido cuando se calculó la desviación media.

El proceso de elevar números al cuadrado y extraer luego la raíz cuadrada puede compararse a la operación de observar alguna cosa a través de un microscopio. La lente de aumento hace que el objeto observado parezca mucho mayor; al retirar la lente, se lo interpreta en su perspectiva correcta.

Coeficiente de Variación

El coeficiente de variación es una medida que se emplea fundamentalmente para:

- Comparar la variabilidad entre dos grupos de datos referidos a distintas unidades de medida. Por ejemplo días y años.
- Comparar dos grupos de datos que tienen media muy diferente.

Esta medida relativiza la magnitud de la desviación estándar al valor de la media aritmética.

Se calcula como:

$$CV = \frac{\sigma}{\mu} 100\%$$

Como ejemplo, se tomaron los valores de la duración de procesos en dos juzgados civiles que entienden también en procesos de familia y juicios ejecutivos. En cada uno se midió la duración de los procesos, y se calcularon la media aritmética y el desvío estándar, obteniendose los siguientes valores

	JUZGADO 1	**JUZGADO 2**
Media	250	120
Desvío Estándar	60	60

A simple vista, ambos juzgados presentan el mismo comportamiento en cuanto a la variabilidad de la duración de los procesos. Sin embargo, la media del Juzgado 1 es muy superior a la de Juzgado 2, lo que implica que las variaciones son incomparables, salvo si se las relativiza a dichos valores (60 no representa lo mismo en 250 que en 120).

Los Coeficientes de Variación son:

$$CV_1 = \frac{60}{250}100\% = 24\% \qquad CV_2 = \frac{60}{120}100\% = 50\%$$

Las duraciones de los procesos en el Juzgado 2 resultaron ser más variables que en el Juzgado 1. El Juzgado 1 resultó ser más homogéneo que el Juzgado 2 en cuanto a la duración de los procesos.

Si se amplía el cuadro visto anteriormente donde se comparaba la media de cada tipo de proceso, agregando las medidas de variabilidad presentadas, se obtiene lo siguiente:

Resúmenes de casos

duracion

ObjetoProceso	N	Media	Rango	Desv. Est.	Varianza
Daños y perjuicios	8	576,63	619	214,349	45945,696
Div. vincular por present. conjunta	2	228,00	56	39,598	1568,000
Ejecución fiscal	31	528,16	2848	578,400	334546,3
Ejecutivo	33	1033,06	2641	714,099	509936,9
Prepara vía ejecutiva	5	291,20	573	251,149	63075,700
Quiebra	1	690,00	0	.	.
Sucesión ab-intestato	186	482,48	2730	533,519	284642,2
Tercería de dominio	1	1075,00	0	.	.
Total	267	556,16	2848	578,386	334529,9

- En los casos en que hay un solo valor de la variable no se calcula el desvío estándar ni la varianza (no hay desviaciones de los datos)

- La desviación estándar total es de 578,386 días.

- Como las medias son muy diferentes, conviene calcular el coeficiente de variación a fin de determinar cuáles procesos son más homogéneos en su duración.

Objeto del Proceso	Coeficiente de Variación
Daños y perjuicios	37,17%
Div. vincular por present. conjunta	17,37%
Ejecución fiscal	109,51%
Ejecutivo	69,12%
Prepara vía ejecutiva	86,25%
Quiebra	.
Sucesión ab-intestato	110,58%
Tercería de dominio	.
TOTAL	104,00%

Los procesos más variables en cuanto a su duración son las Ejecuciones Fiscales y las Sucesiones, mientras que los más homogéneos son los Daños y Perjuicios (no se tienen en cuenta los divorcios ya que hay sólo dos causas de este tipo). Esto es consecuente con lo que se había observado en el diagrama de cajas.

Medida de Asociación entre dos variables: Coeficiente de Correlación

Hasta ahora se han examinado métodos numéricos que tienen por objetivo resumir los datos de una única variable. Sí se vio, sin embargo, cómo se podía describir gráficamente la relación entre dos variables cuantitativas a partir de un diagrama de dispersión.

Más allá de describirla, muchas veces interesa conocer una medida numérica del grado de asociación lineal, y así poder responder a preguntas como: *Lo que el diagrama de dispersión nos muestra, indica que las variables están fuertemente relacionadas? O se trata de una relación débil? Entre dos o más relaciones descriptas: cuál es la que evidencia mayor dependencia de una variable respecto de la otra?*

Una medida que permite cuantificar esta relación es el *Coeficiente de Correlación Lineal*.

En la construcción de este coeficiente están involucrados los desvíos de cada dato respecto de su propia media.

El coeficiente de correlación puede tomar valores ente –1 y 1.

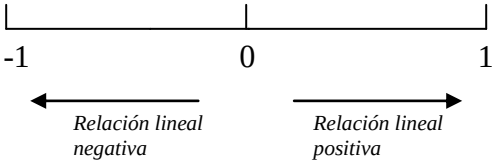
Si vale entre 0 y 1, se trata de una relación lineal positiva: esto es, a medida que una variable aumenta, la otra también.

Si es igual a 1, estamos ante la presencia de una correlación lineal exacta positiva, es decir, todos los puntos están sobre una recta creciente, o con pendiente positiva.

Si, por otro lado, toma valores entre –1 y 0, significa que mientras una variable aumenta, la otra disminuye. Existe una relación negativa.

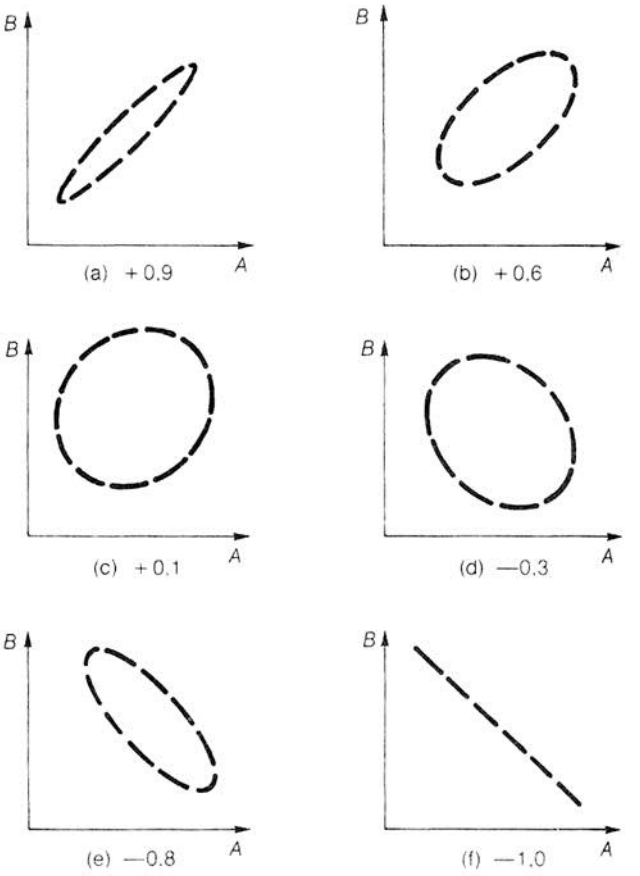
Si vale –1, la relación es exacta: los puntos están ubicados sobre una recta decreciente, o con pendiente negativa.

Cuanto más cercano a los extremos (1 ó –1) esté el coeficiente de correlación, más fuerte es la relación lineal entre las variables. Mientras el coeficiente se acerque más a 0, la relación es más débil.



Notemos que el coeficiente se llama de correlación *lineal*. Esto es porque sólo mide la fuerza de este tipo de relación. Una relación curvilínea no es detectada por este estadístico. Esto significa, que un valor de cero para el coeficiente de correlación no indica que no exista ningún tipo de relación entre las variables, sólo que no existe relación lineal, puede ser curvilínea.

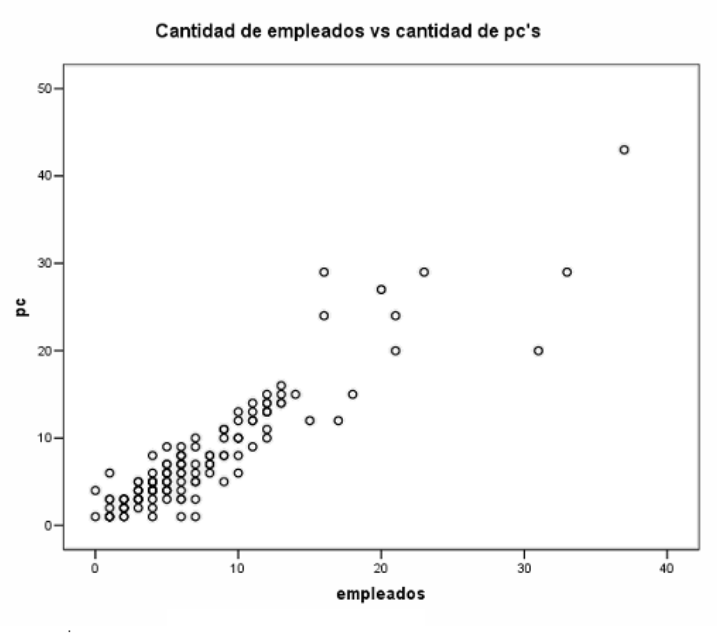
La siguiente figura muestra las nubes de puntos correspondientes a diversos valores de correlación.



No se presenta en este punto la forma en que se calcula el coeficiente de correlación, ya que eso implicaría utilizar notación y conceptos que no han sido desarrollados. Simplemente se utilizará el valor que arroja el software y se lo interpretará.

En el Anexo a este módulo se podrán encontrar las fórmulas matemáticas involucradas en el cálculo de todas las medidas descriptivas introducidas en el presente texto.

Volviendo al ejemplo de la cantidad de pc's por organismo y retomando el diagrama de dispersión que quedaba determinado al graficar ambas variables



Tal como se había analizado en ese momento, se evidencia una relación lineal positiva, pero no se conoce el grado de la misma.

El SPSS arroja para estos datos un valor del coeficiente de correlación de 0.918, lo que indica una alta correlación positiva entre la cantidad de pc's y el número de empleados por organismo. Esto indica que la distribución de pc's por organismo es directamente proporcional a la cantidad de empleados y que se podría ajustar una línea recta, una ecuación matemática, que permita describir exactamente (salvo errores aleatorios) cómo la cantidad de pc's asignadas a un organismo depende de la cantidad de personas que trabajen en él, y así poder pronosticar la cantidad de pc's que serán necesarias en el futuro a partir de la creación de nuevos organismos.

VI- ANEXO FORMULAS ESTADISTICAS

Para formalizar todos los conceptos que se desarrollaron en los apartados anteriores, hay que recordar en principio la diferencia entre parámetro y estadístico.

Un *parámetro* es un valor que se calcula a partir de las unidades de una población. Puede ser un promedio, un porcentaje, un desvío estándar, etc. En general los parámetros se desconocen, porque no se tiene a mano la población completa, ya sea por costos, tiempos o simplemente accesibilidad. Por ello se extrae una muestra y se calculan *estadísticos*, utilizando los datos de la muestra, que permitirán hacer inferencia respecto de los parámetros poblacionales.

Se notará ***N*** a la cantidad de elementos de una población, y ***n*** a la cantidad de elementos de la muestra.

Así, para enumerar a todos los elementos de la población se denota

$$x_1, x_2, x_3, ..., x_N$$

siendo x_1 el primer elemento de la población, x_2 el segundo, y así siguiendo, hasta x_N que se refiere al último elemento de la población.

Para enumerar todos los elemento de una muestra, análogamente se escribe

$$x_1, x_2, x_3, ..., x_n$$

MEDIA ARITMÉTICA

En el caso de la población:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

El símbolo $\sum_{i=1}^{N}$ se lee “sumatoria para *i* desde *1* hasta *N*” e indica la suma de todos los elementos $x_1, x_2, x_3, ..., x_N$ (el subíndice *i* hace referencia a cada *x*, según el valor que tome *i=1, 2,...N*). Es decir

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + ... + x_N}{N}$$

Si se trata de la media de una muestra,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

MEDIANA Y PERCENTILES

Primeramente se calcula la ubicación de estas medidas de posición y luego se identifica el valor.

Para el percentil p , la ubicación en el conjunto de datos ordenado de menor a mayor, es

$\frac{(n+1)}{100}p$ si se trata de una muestra, ó $\frac{(N+1)}{100}p$ en la población.

Así, si se quiere el 1er cuartil, o percentil 25, se debe hacer $\frac{(n+1)}{100}25 = (n+1)0.25$

Si el resultado da un número entero, el percentil buscado es el valor que está exactamente en esa posición. Si por el contrario, resulta un número no entero (con decimales), hay que promediar los dos valores que están en las posiciones inmediatamente anterior e inmediatamente posterior.

Ejemplo: suponga que se quiere identificar la mediana y el 3er cuartil en un conjunto de 13 datos

La posición de la mediana es

(13+1)0.5 = 7

Entonces Med = x_7 , siendo x_7 el valor que está en la ubicación 7, en el conjunto ordenado de menor a mayor.

La posición del 3er cuartil (percentil 75) es

(13+1)0.75 = 10.5

Como no existe un valor que esté en la posición 10.5, se debe promediar los valores ubicados en las posiciones 10 y 11:

$$Q_3 = \frac{x_{10} + x_{11}}{2}$$

DESVIACIÓN MEDIA

$$DM = \frac{\sum_{i=1}^{N} |x_i - \mu|}{N}$$ en el caso de la población

ó

$$dm = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$ si se trata de una muestra

$|x_i - \bar{x}|$ se lee "valor absoluto de los desvíos de cada dato respecto de la media" y significa tomar las diferencias $x_i - \bar{x}$, pero considerando todas con signo positivo. Es decir, lo único que importa es la distancia desde cada dato a la media y no si esa diferencia es positiva o negativa (si el dato es mayor o menor que la media).

Notar que, de no tomar el valor absoluto, $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$.

VARIANZA Y DESVIO ESTANDAR

En la población, la varianza se calcula

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Note que cada uno de los sumandos están elevados a cuadrado; no es lo mismo efectuar la sumatoria completa de los desvíos y luego elevarla al cuadrado (además siempre daría cero!)

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + ...(x_N - \mu)^2}{N}$$

En el caso de querer calcular la varianza en una muestra:

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

El sentido común diría que se debería dividir por *n*, ya que en la varianza poblacional se divide por *N*.

Sin embargo, se puede demostrar (fuera del alcance de este texto) que si se dividiera por *n* a la varianza muestral, se estaría subestimando el verdadero valor de la varianza poblacional. El valor obtenido no sería una buena estimación.

Muchas veces, y a la hora de facilitar los cálculos, se utiliza una "fórmula de trabajo", que está dada por la siguiente expresión:

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

El desvío estándar es la raíz cuadrada de la varianza, por lo tanto, según se trate de una población o de una muestra, es

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \qquad \text{ó} \qquad s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

COEFICIENTE DE CORRELACIÓN

Ahora se está ante la presencia de dos variables, *x* e *y*, medidas para cada unidad *i*. Para cada variable se puede calcular su media: $\bar{x}$ e $\bar{y}$.

El coeficiente de correlación está basado en el producto de los desvíos de los datos respecto de su propia media, y esto se relativiza al producto de los desvíos estándar de cada variable.

$$r = \frac{S_{xy}}{s_x s_y}$$

donde $S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ es la **Covarianza** entre las variables *x* e *y*.

BIBLIOGRAFÍA

- *"Estadística para Administración y Economía"* – Anderson, D., Sweeney, D. Williams- T. Internacional Thomson Editores – 1999

- *"Estadística para las Ciencias Administrativas"* – Chao, L. – Mc Graw Hill – 3rd Ed. 1993

- *"Estadística fácil aplicada a las Ciencias Sociales"* – Clegg, F.- Grupo Editorial Grijalbo – 1984

- *"Statistics"* - Freedman, Pisani, Purves,. Norton & Company. 1980

- *"Estadística Elemental"* - Hoel, CECSA

- *"Estadística básica en Administración"* - Berenson, M., Levine, D - Ed Prentice Hall Internacional. 1983.

- *"Estadística Interactiva"* - Alliaga, M.-