ESTADÍSTICA DESCRIPTIVA

Capítulo 1. INTRODUCCIÓN

1.1 ¿Qué es la estadística?

ESTADÍSTICA es el arte de realizar inferencias y sacar conclusiones a partir de datos imperfectos.

Los datos son generalmente imperfectos en el sentido que aún cuando posean información útil no nos cuentan la historia completa. Es necesario contar con métodos que nos permitan extraer información a partir de los datos observados para comprender mejor las situaciones que los mismos representan.

Algunas técnicas de análisis de datos son sorprendentemente simples de aprender y usar más allá del hecho que la teoría matemática que las sustentan puede ser muy compleja.

Todos, aún los estadísticos, tenemos problemas al enfrentarnos con listados de datos. Existen muchos métodos estadísticos cuyo propósito es ayudarnos a poner de manifiesto las características sobresalientes e interesantes de nuestros datos que pueden ser usados en casi todas las áreas del conocimiento.

Los métodos estadísticos pueden y deberían ser usados en todas las etapas de una investigación, desde el comienzo hasta el final. Existe el convencimiento de que la estadística trata con el ANÁLISIS DE DATOS (quizás porque esta es la contribución más visible de la estadística), pero este punto de vista excluye aspectos vitales relacionados con el DISEÑO DE LAS INVESTIGACIONES. Es importante tomar conciencia que la elección del método de análisis para un problema, se basa tanto en el tipo de datos diponibles como en la forma en que fueron recolectados.

1.2 ¿Por qué estudiar estadística?

Porque los datos estadísticos y las conclusiones obtenidas aplicando metodología estadística ejercen una profunda influencia en casi todos los campos de la actividad humana. En particular, la estadística invade cada vez más cualquier investigación relativa a salud pública. Este crecimiento, probablemente relacionado con el interés por aumentar la credibilidad y confiabilidad de las investigaciones, no garantiza que en todos los casos la metodología estadística haya sido correctamente utilizada, o peor aún, que sea válida.

¿Por qué debe preocuparnos la aplicación incorrecta de métodos estadísticos en un trabajo científico o en un informe técnico?

- Porque las conclusiones pueden ser incorrectas.
- Porque no todos los lectores están en condiciones de detectar el error, y esto genera un importante "ruido" en la bibliografía científica (Aunque este argumento tiende a sobredimensionar la importancia de un paper, existe considerable evidencia que los lectores sin formación metodológica tienden a aceptar como válidas las conclusiones

de los trabajos publicados, en especial si se encuentran publicados en revistas prestigiosas).

El estudio de la Estadística y el modo de pensamiento que se genera a partir del mismo, capacita a la persona para evaluar objetiva y efectivamente si la información que recibe (vía tablas, gráficos, porcentajes, tasas, etc.) es relevante y adecuada. Por supuesto, la interpretación de cualquier problema requiere, no sólo de conocimientos metodológicos sino también, de un profundo conocimiento del tema.

Aún cuando una persona no esté interesada en especializarse en estadística, un entrenamiento básico en el tema permite una mejor comprensión de la información cuantitativa.

1.3 Áreas de la estadística

Describiremos brevemente cada una de las áreas en que puede dividirse la estadística:

I. **Diseño**: Planeamiento y desarrollo de investigaciones.

II. **Descripción**: Resumen y exploración de datos.

III. **Inferencia**: Hacer predicciones o generalizaciones acerca de características de una población en base a la información de una muestra de la población.

I. Diseño

Es una actividad crucial. Consiste en definir como se desarrollará la investigación para dar respuesta a las preguntas que motivaron la misma. La recolección de los datos requiere en general de un gran esfuerzo, por lo que, dedicar especial cuidado a la etapa de planificación de la investigación ahorra trabajo en las siguientes etapas. Un estudio bien diseñado resulta simple de analizar y las conclusiones suelen ser obvias. Un experimento pobremente diseñado o con datos inapropiadamente recolectados o registrados puede ser incapaz de dar respuesta a las preguntas que motivaron la investigación, más allá de lo sofisticado que sea el análisis estadístico.

Aún en los casos en que se estudian datos ya registrados, en que estamos restringidos a la información existente, los principios del buen diseño de experimentos, pueden ser útiles para ayudar a seleccionar un conjunto razonable de datos que esté relacionado con el problema de interés.

II. Descripción

Los métodos de la *Estadística Descriptiva o Análisis Exploratorio de Datos* ayudan a presentar los datos de modo tal que sobresalga su estructura. Hay varias formas simples e interesantes de organizar los datos en gráficos que permiten detectar tanto las características sobresalientes como las características inesperadas. El otro modo de describir los datos es resumirlos en uno o dos números que pretenden caracterizar el conjunto con la menor distorsión o perdida de información posible.

Explorar los datos, debe ser la primera etapa de todo análisis de datos. ¿Por qué no analizarlos directamente? En primer lugar porque las computadoras no son demasiado hábiles (sólo son rápidas), hacen aquello para lo que están programadas y actúan sobre los datos que les ofrecemos. Datos erróneos o inesperados serán procesados de modo inapropiado y ni usted, ni la computadora se darán cuenta a menos que realice previamente un análisis exploratorio de los datos.

III. Inferencia

Inferencia Estadística hace referencia a un conjunto de métodos que permiten hacer predicciones acerca de características de un fenómeno sobre la base de información parcial acerca del mismo.

Los métodos de la inferencia nos permiten proponer el valor de una cantidad desconocida (**estimación**) o decidir entre dos teorías contrapuestas cuál de ellas explica mejor los datos observados (**test de hipótesis**).

El fin último de cualquier estudio es aprender sobre las poblaciones. Pero es usualmente necesario, y más práctico, estudiar solo una muestra de cada una de las poblaciones.

Definimos:

POBLACIÓN $\Rightarrow$ total de sujetos o unidades de análisis de interés en el estudio

MUESTRA $\Rightarrow$ cualquier subconjunto de los sujetos o unidades de análisis de la población, en el cual se recolectarán los datos

Usamos una muestra para conocer o estimar características de la población, denominamos:

PARÁMETRO $\Rightarrow$ una medida resumen calculada sobre la población

ESTADÍSTICO $\Rightarrow$ una medida resumen calculada sobre la muestra

La calidad de la estimación puede ser muy variada, y generalmente las estimaciones estadísticas son erróneas, en el sentido que no son perfectamente exactas. La ventaja de los métodos estadísticos es que aplicados sobre datos obtenidos a partir de muestras aleatorias permiten cuantificar el error que podemos cometer en nuestra estimación o calcular la probabilidad de cometer un error al tomar una decisión en un test de hipótesis.

Finalmente, cuando existen datos para toda la población (CENSO) no hay necesidad de usar métodos de estadística inferencial, ya que es posible calcular exactamente los parámetros de interés. En el censo poblacional, por ejemplo, se registra el sexo de todas las personas censadas, que son prácticamente toda la población, así que es posible conocer exactamente la proporción de habitantes de los dos sexos.

Capítulo 2. TIPOS DE DATOS

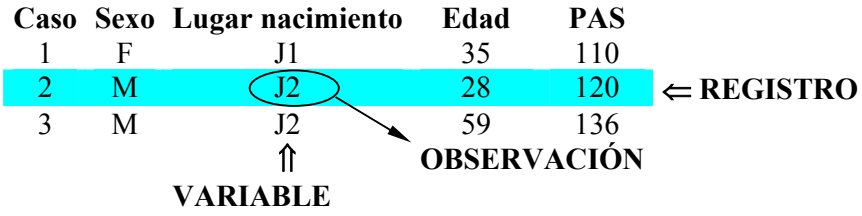
En este capítulo presentaremos los distintos tipos de datos o variables que podemos encontrar en una investigación e comentaremos algunas estrategias para el manejo de datos con una computadora.

2.1 CARACTERÍSTICAS DE LOS CONJUNTOS DE DATOS.

En lo que sigue denominaremos

- UNIDAD DE ANÁLISIS O DE OBSERVACIÓN al objeto bajo estudio. El mismo puede ser una persona, una familia, un país, una región, una institución o en general, cualquier objeto.
- VARIABLE a cualquier característica de la unidad de observación que interese registrar, la que en el momento de ser registrada puede ser transformada en un número.
- VALOR de una variable, OBSERVACIÓN o MEDICIÓN, al número que describe a la característica de interés en una unidad de observación particular.
- CASO o REGISTRO al conjunto de mediciones realizadas sobre una unidad de observación.

Consideremos el siguiente ejemplo:



Sexo, lugar nacimiento, edad, presión arterial sistólica son variables que describen a una persona, *su* sexo, *su* lugar de nacimiento, *su* edad, etc. son los valores que estas variables toman para esta persona.

Cuando se diseña una investigación, se intenta estudiar de qué modo una o más variables (*variables independientes*) afectan a una o más variables de interés (*variables dependientes*). Por ejemplo en un experimento, el investigador impone a los sujetos condiciones (variable independiente) y estudia el efecto de la misma sobre una característica del sujeto (aparición de una cierta característica, modificación de una condición, etc.).

Un paso importante al comenzar a manejar un conjunto de datos es identificar *cuántas variables* se han registrado y *cómo* fueron registradas esas variables, lo que permitirá definir la estrategia de análisis. En el ejemplo anterior algunas de las variables son números y otras son letras que indican categorías. A continuación se presenta una clasificación de los distintos tipos de datos que podemos encontrar. Debe notarse que distintos autores usan distintos criterios para clasificar datos por lo que presentaremos aquí un criterio que resulta útil desde el punto de vista de seleccionar el método de análisis estadístico más apropiado para los mismos.

2.2 TIPOS DE DATOS

2.2.1 DATOS CATEGÓRICOS O CUALITATIVOS

Las *variables categóricas* resultan de registrar la presencia de un atributo.

Las categorías de una variable cualitativa deben ser definidas claramente durante la etapa de diseño de la investigación y deben ser mutuamente excluyentes y exhaustivas. Esto significa que cada unidad de observación debe ser clasificada sin ambigüedad en una y solo una de las categorías posibles y que existe una categoría para clasificar a todo individuo.

En este sentido, es importante contemplar todas las posibilidades cuando se construyen variables categóricas, incluyendo una categoría tal como No sabe / No contesta, o No registrado u Otras, que asegura que todos los individuos observados serán clasificados con el criterio que define la variable.

Los datos categóricos se clasifican en *dicotómicos, nominales* y *ordinales*.

a) Dos categorías (DICOTÓMICOS)

El individuo o la unidad de observación puede ser asignada a solo una de dos categorías. En general, se trata de *presencia - ausencia* del atributo y es ventajoso asignar código 0 a la ausencia y 1 a la presencia.

Ejemplos:

1) varón – mujer
2) embarazada - no embarazada
3) fumador - no fumador
4) hipertenso – normotenso

Debe notarse que los ejemplos 1) y 2) definitivamente cubren todas las categorías, mientras que 3) y 4) son simplificaciones de categorías más complejas. En 3) no está claro donde se asignan los ex-fumadores, en tanto que en 4) fue necesario establecer un criterio de corte para armar una variable categórica a partir de una variable numérica.

b) Más de dos categorías

CATEGORÍAS NOMINALES $\Rightarrow$ No existe orden obvio entre las categorías.

Ejemplos: país de origen, estado civil, diagnóstico.

CATEGORÍAS ORDINALES $\Rightarrow$ Existe un orden natural entre las categorías.

Ejemplos:

1) Tabaquismo: No fuma / ex-fumador / fuma $\leq$ 10 cigarrillos diarios / fuma > 10 cigarrillos diarios
2) Severidad de la patología: Ausente / leve / moderado / severo.

Aún cuando los datos ordinales puedan ser codificados como números como en el caso de estadios de cáncer de mama de I a IV, no podemos decir que una paciente en el estadio IV

tiene un pronóstico dos veces más grave que una paciente en estadio II, ni que la diferencia entre estadio I y II es la misma que entre estadio III y IV. En cambio, cuando se considera la edad de una persona, 40 años es el doble de 20 y una diferencia de 1 año es la misma a través de todo el rango de valores.

Por esta razón, debemos ser cuidadosos al tratar variables cualitativas, especialmente cuando se han codificado numéricamente, ya que no pueden ser analizadas como números sino que deben ser analizados como categorías. Es incorrecto presentar, por ejemplo, el estadio promedio de cáncer en un grupo de pacientes.

En la práctica clínica se usan escalas para definir grados de un síntoma o de una enfermedad, tales como 0, +, ++, +++. Es importante definir operativamente este tipo de variables y estudiar su confiabilidad de modo de asegurar que dos observadores puestos frente al mismo paciente, lo clasificarán en la misma categoría.

2.2.2 DATOS NUMÉRICOS

Una variable es numérica cuando el resultado de la observación o medición es un número. Se clasifican en:

a) Discretos. La variable sólo puede tomar un cierto conjunto de valores posibles. En general, aparecen por conteo.

Ejemplo: número de miembros del hogar, número de intervenciones quirúrgicas, número de casos notificados de una cierta patología.

b) Continuos. Generalmente son el resultado de una medición que se expresa en unidades. Las mediciones pueden tomar teóricamente un conjunto infinito de valores posibles dentro de un rango. En la práctica los valores posibles de la variable están limitados por la precisión del método de medición o por el modo de registro.

Ejemplos: altura, peso, pH, nivel de colesterol en sangre.

La distinción entre datos discretos y continuos es importante para decidir qué método de análisis estadístico utilizar, ya que hay métodos que suponen que los datos son continuos.

Consideremos por ejemplo, la variable edad. Edad es continua, pero si se la registra en años resulta ser discreta. En estudios con adultos, en que la edad va de 20 a 70 años, por ejemplo, no hay problemas en tratarla como continua, ya que el número de valores posibles es muy grande. Pero en el caso de niños en edad preescolar, si la edad se registra en años debe tratarse como discreta, en tanto que si se la registra en meses puede tratarse como continua.

Del mismo modo, la variable número de pulsaciones/min. es una variable discreta, pero se la trata como continua debido al gran número de valores posibles.

Los datos numéricos (discretos o continuos) pueden ser transformados en categóricos y ser tratados como tales. Aunque esto es correcto no necesariamente es eficiente y *siempre* es preferible registrar el valor numérico de la medición, ya que esto permite:

- Analizar la variable como numérica $\Rightarrow$ Análisis estadístico más simple y más potente.
- Armar nuevas categorías usando criterios diferentes.

Sólo en casos especiales es preferible registrar datos numéricos como categóricos, por ejemplo, cuando se sabe que la medición es poco precisa (número de cigarrillos diarios, número de tazas de café en una semana).

2.2.3 OTRO TIPO DE DATOS

a) Porcentajes

Los porcentajes son el resultado de tomar el cociente entre dos cantidades. Ejemplos: reducción porcentual de la presión arterial luego de la aplicación de una droga, o peso corporal relativo (peso observado/peso deseable). En el primer caso las cantidades que forman el cociente se miden simultaneamente, en tanto que en el segundo caso el denominador es un valor estándar preexistente.

Aunque los porcentajes pueden pensarse como variables continuas pueden causar problemas en el análisis, especialmente cuando pueden tomar valores mayores y menores que 100% (ejemplo: de peso corporal relativo) o cuando pueden dar valores negativos (ejemplo: reducción porcentual de la PA. En este último caso, un paciente con PAS en 150 mm Hg con un 20% de aumento en la PAS llegará a 180 mmHg, pero una posterior disminución del 20% lo llevará a 144 mm Hg). Se debe tener cuidado al analizar estos datos.

b) Escalas analógicas visuales

Cuando se necesita que una persona indique el grado de alguna característica no medible, tal como satisfacción, dolor, bienestar, agrado, acuerdo, etc. una técnica que permite obtener *categorías ordinales* es la escala analógica visual. Se presenta al encuestado una línea recta (generalmente de 10 cm.) cuyos extremos indican estados extremos y se les pide que marquen una posición en la recta que represente la percepción de su estado.

Ejemplo. Interesa estimar grado de satisfacción con un tratamiento, se puede usar la siguiente escala.

Estas escalas son muy útiles para valorar cambios en el mismo individuo. Aún cuando un puntaje de 3.7 no dice nada en si mismo, una reducción de 2 puntos en un paciente si nos da información. Debe tenerse cuidado al tratar este tipo de datos ya que, a diferencia de los datos numéricos, aún cuando se registren como números la escala subyacente no necesariamente es la misma para dos sujetos distintos.

c) Scores

Los scores son indicadores de la condición de un individuo basados en la observación de varias variables, generalmente categóricas. En clínica los scores se construyen en base a síntomas y signos, asignándole a cada uno de ellos un puntaje y calculando un puntaje total o score, que es un indicador de la condición del paciente.

Un ejemplo es el score Apgar usado como indicador de pronóstico en el recién nacido.

	Puntaje		
Signo	**0**	**1**	**2**
Latidos	Ausente	< 100	≥ 100
Respiración	Ausente	Llanto débil, hiperventilación	Llanto fuerte
Tono muscular	Fláccido	Leve	Buena flexión
Reflejos	Ausente	Leve	Llanto
Color	Azul, pálido	Cuerpo rosado, extremidades azules	Totalmente rosa

El recién nacido es evaluado en los minutos 0 y 5 de vida. Cada signo recibe un puntaje de 0 a 2, los cuales se suman y el score resultante es un número entre 0 a 10. Se considera que un score ≥ 7 es de buen pronóstico, y que un Apgar ≤ 3 es de muy mal pronóstico.

No es de interés aquí discutir la validez de este particular score, pero remarcaremos tres características que son comunes a este tipo de scores:

- en la evaluación de cada signo está presente cierto nivel de subjetividad,
- al transformar las categorías en números, estamos valorando las diferencias entre 0 y 1 y entre 1 y 2 como equivalentes,
- los cinco signos son igualmente importantes en la construcción del score.

Los scores deberían tratarse en el análisis tal como se los trata en la práctica, como criterios para definir categorías ordinales y no como variables numéricas.

d) Datos censurados

Una observación censurada es aquella que no pudo ser medirla exactamente, pero que se sabe que está más allá de un cierto límite, es decir, conocemos una cota inferior o superior para el dato.

Ejemplos.

- Cuando se miden elementos traza, el nivel del elemento en la muestra puede ser menor que el límite de detección de la técnica. Este es un dato con censura izquierda ya que no se conoce el verdadero valor, pero si se conoce una cota superior.
- Estudios de seguimiento en los que interesa el tiempo de supervivencia. En los pacientes que se mantienen vivos finalizar el estudio, se desconoce el tiempo real de supervivencia, pero se sabe que éste es mayor que el tiempo de permanencia en el estudio. El tiempo de supervivencia está censurado a derecha, sólo conocemos una cota inferior para el mismo.
- Un estudio de seguimiento en que interesa estudiar el tiempo transcurrido hasta la recidiva de una patología. En aquellos sujetos que se pierden del estudio (por abandono, por muerte por otras causas o por cualquier otra razón) pero que sabemos que estuvieron libres de la patología mientras permanecieron en el estudio (hasta el último control), el dato de tiempo transcurrido hasta la recidiva está censurado a derecha.

¿Por qué es importante identificar el tipo de datos?

Porque el tipo de datos DETERMINA el método de análisis apropiado y válido y cada método de análisis estadístico es específico para un cierto tipo de datos. La distinción más importante es entre datos numéricos y categóricos.

2.3 USANDO UNA COMPUTADORA PARA PROCESAR DATOS

Las computadoras nos ahorran los aspectos tediosos del análisis estadístico y en principio producen cálculos correctos, pero *no garantizan que obtendremos resultados válidos y correctos.* Consideraremos brevemente las ventajas y desventajas de usar una computadora para procesar datos y consideraremos algunas formas de armar archivos de datos.

2.3.1 VENTAJAS Y DESVENTAJAS DE USAR UNA COMPUTADORA.

a) Ventajas

- *Exactitud y velocidad*. Cuando el software es de calidad se obtienen resultados correctos rápidamente.
- *Versatilidad*. Se tiene acceso a un amplio rango de técnicas estadísticas. Muchas más de las que es posible describir en cualquier curso de estadística.
- *Gráficos.* Se pueden producir representaciones de los datos originales o de los resultados obtenidos que permiten una mejor visualización.
- *Flexibilidad.* Una vez que se ha construido la base de datos, se pueden realizar pequeños cambios y repetir el análisis. Por ejemplo, es posible excluir algunos casos, hacer análisis por subgrupos o estratos, etc.
- *Nuevas variables.* Es simple generar nuevas variables. Ejemplo: diferencia entre mediciones antes y después de un tratamiento, calcular edad como diferencia de fechas, crear variables categóricas a partir de variables numéricas, recategorizar variables cualitativas, realizar transformaciones, etc.
- *Volumen de datos.* Algunos programas pueden procesar un número de registros o de variables ilimitado.

b) Desventajas.

- *Errores en el software.* Muchos paquetes estadísticos de uso corriente presentan errores en algunos procedimientos. Los más seguros son: SAS, S-PLUS, STATA y SPSS. Si no se tiene seguridad acerca de la calidad del software que se está usando debería chequearse comparando los resultados de cada procedimiento con ejemplos de libro o con software de primer nivel.
- *Versatilidad.* Esta ventaja se transforma en desventaja porque al haber tantos métodos estadísticos disponibles es fácil usar uno inapropiado. Es importante que el usuario tenga en claro sus limitaciones en conocimientos estadísticos y use sólo los métodos que comprende. Si el problema parece requerir métodos que no son familiares, es aconsejable consultar a un estadístico.

- *Caja Negra.* Se puede perder el contacto con los datos. Si el análisis se realiza automáticamente, se corre el riesgo de no advertir las características más relevantes de los datos, o de perder la información acerca de individuos con comportamiento atípico.
- *Los resultados dependen de la calidad del archivo de datos.* Si los datos están mal registrados o tienen inconsistencias y el investigador no lo advierte, los resultados serán incorrectos más allá de lo sofisticado y elegante que sea el método de análisis estadístico que se utilice.

2.3.2 ESTRATEGIA PREVIA EL ANÁLISIS DE DATOS

a) Definición y codificación de las variables. Carga de datos.

Es recomendable usar un formato estandarizado para registrar la información. Esto vale tanto para estudios en los que los datos serán obtenidos a partir de registros existentes (por ejemplo historias clínicas) así como para estudios prospectivos.

Algunas variables tienen varias respuestas posibles no mutuamente excluyentes. En este caso es necesario ofrecer la opción si – no para cada posible respuesta. Ejemplo: Durante la última semana consumió: pescado si-no, legumbres si-no, carnes rojas si-no, carnes de ave: si-no, etc.

Las variables numéricas deberían ser registradas con la misma exactitud con que fueron obtenidas, no redondear. No categorizar variables numéricas para registrarlas.

Cuando el mismo sujeto es observado más de una vez, por ejemplo durante el control de embarazo o a lo largo de un ensayo, se obtienen *medidas repetidas sobre el mismo individuo*. No debe considerarse cada visita de un sujeto como un registro independiente. Es incorrecto tratar registros múltiples de un individuo como si fueran registros de distintos individuos. Este tipo de datos requiere de métodos estadísticos específicos que se conocen como *técnicas para medidas repetidas*.

Asignar un nombre de no más de 10 letras a cada variable. El nombre completo de la variable puede asignarse a través de una etiqueta (label). Algunos paquetes aceptan nombres de variables de a lo sumo 8 letras truncando las letras finales. Algunos caracteres no son permitidos en los nombres de variables, por ejemplo el punto. No deben dejarse espacios en blanco en el nombre de las variables.

La carga de datos se hace más simple, rápida y exacta si se *codifican todas las variables categóricas.* Es conveniente usar números para codificar las categorías de las distintas variables categóricas y asignar una etiqueta (label) a cada categoría de modo de identificarlas sin dificultad y de hacer más amigable las salidas de los procedimientos estadísticos.

Cuando se trata de *fechas* es importante definir el formato que se usará para la variable: día/mes/año, mes/día/año, día-mes-año, etc. Algunos paquetes no reconocen cualquier formato para las fechas y en consecuencia tratan a los valores de la variable como caracteres alfanuméricos (texto). Cuando ésto ocurre las fechas no pueden ser utilizadas en operaciones algebraicas ya que no son consideradas números sino caracteres.

b) Chequeo de los datos (Consistencia)

Pueden producirse errores cuando se toman las mediciones, cuando se registran los datos originales (ejemplo en la historia clínica), cuando se transcribe de la fuente original a una planilla, o cuando se tipean los datos para armar la base.

Usualmente no podemos saber si los datos son correctos, pero deberíamos asegurar que son plausibles. Esta etapa corresponde a lograr la CONSISTENCIA del archivo. No esperamos solucionar todos los errores, pero esperamos detectar los errores más groseros.

La consistencia de los datos intenta IDENTIFICAR y de ser posible RECTIFICAR errores en los datos.

El primer paso es chequear si el tipeo ha sido correcto. Cuando el archivo es pequeño se imprime y se controla. Cuando es grande, conviene tipearlo dos veces y comparar ambas versiones (EpiInfo lo hace con el procedimiento VALIDATE y produce un listado de diferencias).

Datos categóricos.

En este caso es simple chequear si todos los valores de la variable son plausibles, ya que hay un conjunto fijo de valores posibles para la variable. Ejemplo: Grupo sanguíneo: 0, A, B, AB. Es suficiente con producir una tabla de frecuencias para cada variable categórica en la que se controla que las categorías coinciden con las categorías definidas. Algunos paquetes diferencias letras mayúsculas de minúsculas, por lo tanto consideran que la categoría “a” de grupo sanguíneo es diferente de la “A”.

Es aconsejable hacer un listado de todas las tablas de frecuencia de las variables categóricas antes de comenzar con el análisis estadístico de los datos.

Datos numéricos.

Para cada variable debería proponerse el rango de valores esperado o posible. Ejemplo: Edad materna al parto: 12 a 50 años, Presión arterial sistólica: 70 a 250 mg de Hg.

Un error frecuente es colocar mal la coma o el punto decimal. Valores fuera del rango esperado no necesariamente son incorrectos. Existen valores que son poco probables y valores que son imposibles, lamentablemente el límite entre ambos es difícil de definir. Valores poco probables pero posibles deberían ser corregidos sólo cuando hay evidencia de error.

Cuando la base ha sido importada desde un programa (software) diferente al que se está usando es impotante controlar que durante la exportación se haya respetado el tipo de variable. En particular, que las variables que originalmente estaban definidas como numéricas, no hayan sido transformadas a texto durante la transformación porque no se reconoce el indicador de símbolo decimal (coma, punto). Cuando la variable es de tipo texto no es posible realizar operaciones albegraicas con ella.

Chequeo lógico.

Hay cierta información que sólo se releva en ciertos casos. Por ejemplo, número de embarazos es relevante si sexo = femenino, pero para sexo = masculino, esta variable debería ser '. ' o "no corresponde".

Los datos deben satisfacer los criterios de inclusión y exclusión del estudio. Ejemplo: Estudio de agentes anti-hipertensivos, los pacientes que entran en el estudio deben tener valores de la presión arterial dentro de un cierto rango al ingreso.

Evaluar la consistencia de los datos es algo más complicado cuando existen valores de algunas variables que dependen de valores de otras variables. Existen combinaciones de valores de ciertas variables que son inaceptables, aún cuando cada una de ellas se encuentre dentro de límites razonables.

El investigador debe proponer chequeos lógicos que permitan detectar aberraciones en los datos. Ejemplos: es poco probable que un sujeto se ubique en el percentil 5 de presión diastólica y en el percentil 95 de presión sistólica, o es poco probable que un niño nacido con 30 semanas de gestación pese 3800 g.

Cuando una variable se mide varias veces en la misma unidad de observación puede graficarse a lo largo del tiempo para ver si el comportamiento es acorde a lo esperado.

Fechas.

Son la base para calcular tiempo transcurrido entre eventos. *Ejemplos*: edad del paciente al momento de la consulta, tiempo de supervivencia, etc.

Un criterio de consistencia es chequear si las fechas caen dentro de intervalos de tiempo razonables. *Ejemplo*s: fechas de evaluación dentro del período de desarrollo de la investigación, fechas de nacimiento consistentes con criterios de inclusión y exclusión para edad, etc.

Finalmente, es importante controlar que las fechas siguen una secuencia correcta para cada sujeto. Ejemplo: nacimiento, internación, muerte.

Datos faltantes

Otro problema es el manejo de los *datos missing* (perdidos o faltantes). Cuando al cargar la información se deja un blanco debe tenerse en cuenta que algunos paquetes estadísticos asignan al blanco un cero. En ocasiones se asigna a los datos perdidos valores imposibles como 99999 o un valor negativo para datos que sólo pueden ser positivos. El problema es que si no se excluyen los registros con estos valores atípicos en el momento del análisis, el resultado será erróneo ya que cualquier programa aceptará el valor 0 o el valor 99999 como verdaderos.

En particular, EpiInfo indica los datos missings con un punto, con lo cual se evita este problema.

EpiInfo provee un procedimiento denominado CHEK que permite hacer consistencia de datos a medida que se cargan los mismos.

c) Análisis exploratorio de los datos

Antes de analizar los datos es importante producir gráficos y tablas, los que permitan detectar rápidamente datos anómalos o comportamientos atípicos. Dedicaremos el siguiente capítulo a tratar este tema.

2.3.3 MALOS USOS O ABUSOS DE LA COMPUTADORA

Hemos descripto algunas desventajas de usar computadoras para manejar nuestros datos, agregamos aquí algunos malos usos y abusos que deberían evitarse.

a) Pescar en los datos

En estudios con objetivos pobremente definidos, en los que se registra información porque "puede ser interesante", suelen realizarse gran número de análisis estadísticos buscando que aparezca alguna diferencia entre grupos o asociaciones entre pares de variables. Debe tenerse en cuenta que en este tipo de análisis existe buena chance de encontrar relaciones significativas sólo debidas al azar, cuando en realidad no existe tal relación en la población.

Los análisis exploratorios son muy útiles para ayudar a *proponer nuevas hipótesis* que deberán ser contrastadas en *otro estudio posterior*. Un mismo estudio no puede ser usado para proponer hipótesis y para verificarlas.

b) Análisis estadísticos complejos

Aunque es tentador, no es una buena práctica someter a los datos a análisis estadísticos complejos sólo porque se encuentren disponibles en el software. El análisis debe ser el mínimo requerido para responder sus preguntas. Una razón importante para hacer análisis simples es que las conclusiones son más fáciles de interpretar y de comunicar.

c) Precisión espuria

Las salidas de los programas estadísticos producen resultados con gran cantidad de cifras decimales. Sin embargo, los resultados deben ser comunicados con adecuada precisión.

Ejemplo: Un porcentaje calculado como (17/45)*100 = 37.778% debería informarse como 38% ya que la ocurrencia de un caso más modifica el porcentaje en más del 2%, (18/45) *100 = 40%.

Capítulo 3. ESTADÍSTICA DESCRIPTIVA. GRÁFICOS.

La *estadística descriptiva* o *análisis exploratorio de datos* ofrece modos de presentar y evaluar las características principales de los datos a través de tablas, gráficos y medidas resúmenes. En este capítulo presentaremos formas simples de resumir y representar gráficamente conjuntos de datos.

El objetivo de construir gráficos es poder apreciar los datos como un todo e identificar sus características sobresalientes. El tipo de gráfico a seleccionar depende del tipo de variable que nos interese representar por esa razón distinguiremos en la presentación gráficos para variables categóricas y para variables numéricas.

3.1 PRESENTACIÓN DE DATOS CATEGÓRICOS

3.1.1 TABLA DE FRECUENCIA

El modo más simple de presentar datos categóricos es por medio de una tabla de frecuencias. Esta tabla indica el número de unidades de análisis que caen en cada una de las clases de la variable cualitativa.

Consideremos los casos de meningitis notificados durante el año 2000 al SI.NA.VE (Argentina) clasificados según tipo de meningitis.

Tabla 1. Notificaciones de meningitis en la Argentina, año 2000. Fuente: SI.NA.V.E.

	Notación	Número de notificaciones (frecuencia)	Frecuencia relativa (%)
Meningitis bacteriana sin aislar	BSA	446	22.85 %
Haemophilus infuenzae	HI	34	1.74 %
Meningitis tuberculosa	MTB	17	0.87 %
Neisseria meningitidis	NM	489	25.05 %
Otros gérmenes	OG	89	4.56 %
Sin especificar	SE	228	11.68 %
Streptococo neumoniae	SN	304	15.57 %
Total viral	TV	345	17.67 %
	Total país	1952	100.00 %

La primer y segunda columna de la Tabla 1 muestran las categorías de la variable (tipo de meningitis y la sigla correspondiente), la tercer columna presenta el número de casos de meningitis de cada tipo notificados, es decir la *frecuencia o frecuencia absoluta*, en tanto que la última columna presenta la *frecuencia relativa o el porcentaje* de casos notificados de cada tipo de meningitis. Por ejemplo, la frecuencia relativa de la categoría BSA se calcula del siguiente modo:

$$fr_{BSA} = \frac{\text{números de casos de BSA}}{\text{número total de casos}} \cdot 100 = \frac{f_{BSA}}{n} \cdot 100 = \frac{446}{1952} \cdot 100 = 22.85\%$$

La representación gráfica de una distribución de frecuencias puede realizarse a través de un gráfico de barras o de un gráfico de tortas. A continuación presentamos ambos métodos.

3.1.2 GRÁFICO DE BARRAS

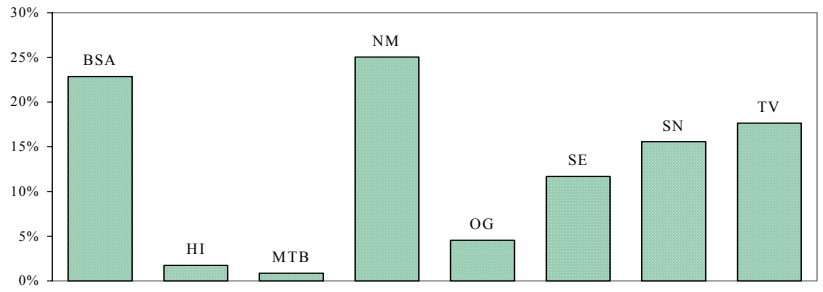
Este gráfico es útil para representar datos categóricos nominales u ordinales. A cada categoría o clase de la variable se le asocia una barra cuya *altura* representa la frecuencia o la frecuencia relativa de esa clase. Las barras difieren sólo en altura, no en ancho.

La escala en el eje horizontal es arbitraria y en general, las barras se dibujan equiespaciadas, por esta razón este tipo de gráfico sólo debe usarse para variables categóricas.

Es importante que el eje vertical comience en cero, de modo que no se exageren diferencias entre clases.

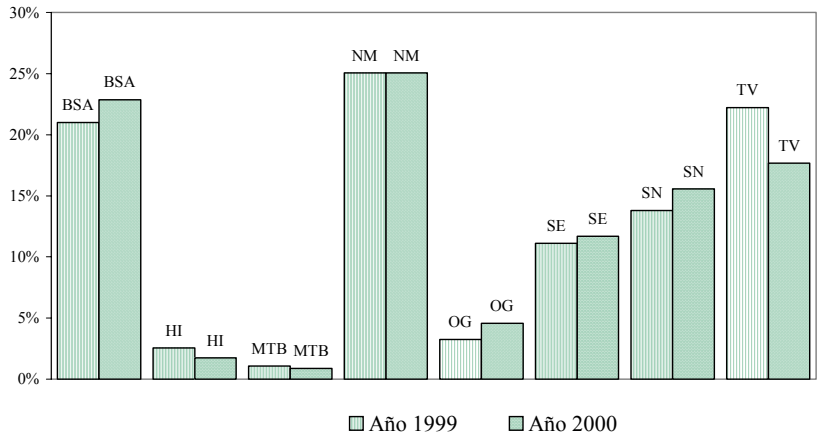
En un gráfico de barras, así como en cualquier tipo de gráfico se debe indicar el número total de datos ya que el gráfico sólo muestra porcentajes o frecuencias relativas y la fuente de la que se obtuvieron los mismos.

Figura 1. Notificaciones de meningitis en la Argentina. Año 2000. Fuente: SINAVE.



Cuando se desea comparar dos o más distribuciones cualitativas, el modo más sencillo de representarción es el *gráfico de barras combinadas*. En la Figura 2 se presentan las distribuciones de casos notificados de meningitis en Argentina para los años 1999 y 2000.

Figura 2. Notificaciones de meningitis en la Argentina. 1999 y 2000. Fuente: SINAVE.



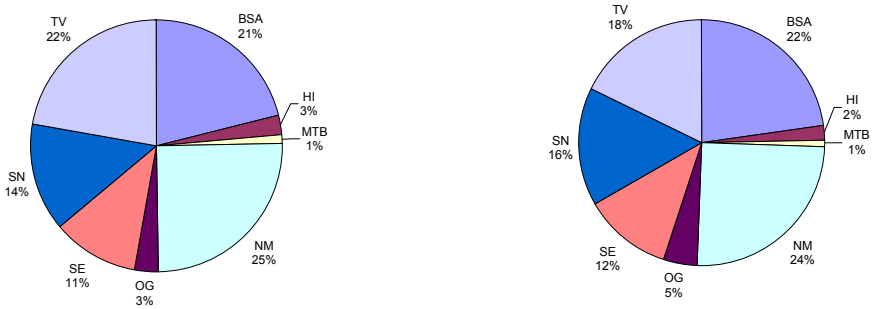
3.1.3 GRÁFICO DE TORTAS

En este gráfico, ampliamente utilizado, se representa la frecuencia relativa de cada categoría como una porción de un círculo, en la que el ángulo se correponde con la frecuencia relativa correspondiente. Como en todo gráfico es importante indicar el número total de sujetos.

Esta representación gráfica es muy simple y permite comparar la distribución de una variable categórica en 2 o más grupos.

Las Figura 3 muestra los datos sobre meningitis presentados en la Figura 2.

Figura 3. Notificaciones de meningitis en la Argentina. 1999 y 2000. Fuente: SINAVE.



¿Cuál preferir: gráfico de barras o de tortas?

La información que brindan los dos tipos de gráficos es equivalente, sin embargo, el gráfico de barras resulta más natural para comparar las distribuciones de dos grupos, debido a que nuestro ojo percibe mejor diferencias en longitudes que en ángulos. Por otra parte, en el gráfico de barras todas las barras comienzan al mismo nivel, lo que facilita la comparación.

3.2 REPRESENTACIÓN GRÁFICA DE UN ÚNICO CONJUNTO DE DATOS NUMÉRICOS

Comenzaremos representando el conjunto de datos más simple posible: un único grupo de números. Trataremos de responder a preguntas tales como:

¿Son los valores medidos casi todos iguales?

¿Son muy diferentes unos de otros?

¿En qué sentido difieren?

¿Cómo podemos describir cualquier patrón o tendencia?

¿Son un único grupo? ¿Hay varios grupos de números?

¿Difieren algunos pocos números notablemente del resto?

Usaremos distintos tipos de gráficos para representar a los datos de modo de hacer visibles sus características más importantes. Mirando un gráfico, es posible ver más allá de los detalles que presenta un listado de números y formarse una impresión de la estructura general.

3.2.1 GRÁFICO DE TALLOS Y HOJAS (STEM AND LEAF)

Esta técnica gráfica desarrollada por Tukey es muy sencilla y permite mostrar la *forma de la distribución de una variable numérica*.

Es apropiada para conjuntos de observaciones no muy extensos, se construye con poco esfuerzo por lo que es muy simple de realizar con lápiz y papel.

Consideremos los datos de la Tabla 2, correspondientes a casos de neumonía notificados (tasa cada 1000 habitantes) por las provincias argentinas durante el año 2000 (Fuente: SI.NA.VE, Argentina). Los datos se presentan ordenados de menor a mayor para simplificar el trabajo.

Tabla 2. Tasas de neumonía cada 1000 habitantes. Año 2000, Argentina. Fuente: SINAVE

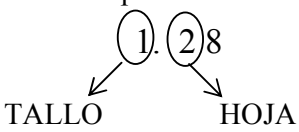
Provincia	Tasa	Provincia	Tasa
Corrientes	0.00	Río Negro	3.86
Córdoba	1.28	La Rioja	3.98
Capital Federal	1.60	Chubut	4.01
Entre Ríos	1.67	Santa Fé	4.22
Tucumán	2.19	Tierra del Fuego	4.38
Catamarca	2.87	Neuquén	4.84
Buenos Aires	3.01	San Juan	4.92
Salta	3.16	Mendoza	5.50
Misiones	3.20	San Luis	7.36
Jujuy	3.21	Formosa	8.07
Santa Cruz	3.33	La Pampa	9.29
Santiago del Estero	3.37	Chaco	10.83

Para construir un gráfico de tallo y hojas procedemos del siguiente modo:

1. Separamos cada observación en dos porciones, TALLO y HOJA. En general, el tallo tendrá tantos dígitos como sea necesario, pero las hojas contendrán un único dígito.

   En nuestro ejemplo podemos elegir el dígito correspondiente a la unidad como tallo y el primer dígito después de la unidad (décima).

   Ejemplo. Consideremos el dato correspondiente a Córdoba:



2. Se listan los tallos verticalmente en orden creciente y se traza una línea vertical a la derecha de los tallos.
3. A continuación de cada tallo se agregan las hojas correspondientes en la misma línea, arreglándolas de menor a mayor.

   Se debe tomar una decisión sobre qué se hará con el dígito posterior a la hoja, si se truncará o se redondeará, poco se pierde truncando y esta última opción hace más simple volver a la lista de datos a partir del gráfico.

   Los tallos que no están acompañados con hojas también se representan, de este modo se respeta la escala de los datos.

Seleccionando como tallo la unidad se obtiene el siguiente gráfico.

0	0
1	2 6 6
2	1 8
3	0 1 2 2 3 3 8 9
4	0 2 3 8 9
5	5
6	
7	3
8	0
9	2
10	8

La altura o extensión de la columna de hojas asociadas a un tallo nos dice con que frecuencia ocurren las observaciones de la magnitud asociada al tallo.

¿Qué información nos brinda este gráfico?

Podemos observar:

- El *rango* de las observaciones y los valores máximos y mínimos.
- La *forma* de la distribución:
  - Si es aproximadamente simétrica o es asimétrica.
  - Cuántos picos o modas tiene la distribución.
- Si existen valores que se aparten notablemente del conjunto, a los que denominaremos datos atípicos o outliers.

¿Cómo elegir el número de tallos?

El número de tallos debe ser tal que permita mostrar una imagen general de la estructura del conjunto de datos. Aunque existen algunos criterios para definir el número de tallos, la decisión depende fundamentalmente del sentido común. Demasiados detalles distraen, demasiado agrupamiento puede distorsionar la imagen del conjunto.

Consideremos el siguiente ejemplo con datos sobre consumo diario per cápita de proteínas en 32 países desarrollados. Los datos se presentan ordenados de menor a mayor por simplicidad.

Tabla 3. Consumo de proteínas per cápita en países desarrollados.

7.83	9.03	10.56
8.06	9.16	10.52
8.45	9.23	10.75
8.49	9.34	10.86
8.53	9.39	10.89
8.60	9.42	11.07
8.64	9.56	11.27
8.70	9.89	11.36
8.75	10.00	11.58
8.92	10.28	11.76
8.93	10.41	

Seleccionando como tallo la unidad obtenemos el gráfico de tallo-hojas de la izquierda de la Figura 4.

Figura 4. Variaciones de los tallos. Datos de consumo de proteínas per cápita.

7	8
8	0 4 4 5 6 6 7 7 9 9
9	0 1 2 3 3 4 5 8
10	0 2 4 5 5 7 8 8
11	0 2 3 5 7

7	8
8	0 4 4
8	5 6 6 7 7 9 9
9	0 1 2 3 3 4
9	5 8
10	0 2 4
10	5 5 7 8 8
11	0 2 3
11	5 7

En este gráfico se acumula un número importante de hojas en cada tallo, por lo que podríamos estar perdiendo información acerca de la estructura de los datos. Dividiremos cada tallo en dos, es decir, representaremos dos veces cada tallo, la primera vez que este aparezca irá acompañado por las hojas 0 a 4 y la segunda vez por las hojas 5 a 9. Obtenemos, entonces, el gráfico de la derecha de la Figura 4.

Como puede observarse, al expandir la escala se observan más detalles y parece haber dos "grupos" de países, uno con mayor consumo per cápita de proteínas y otro con menor consumo, ya que la distribución de la variable tiene dos picos.

El problema de expandir la escala es que comienzan a aparecer detalles superfluos, o simplemente atribuibles al azar.

Gráfico de tallo-hojas espalda con espalda. Comparación de grupos.

Los gráficos de tallo-hojas son útiles para comparar la distribución de una variable en dos condiciones o grupos. El gráfico se denomina tallo-hojas espalda con espalda porque ambos grupos comparten los tallos.

A continuación se muestra un gráfico de la presión arterial sistólica a los 30 minutos de comenzada la anestesia en pacientes sometidos a dos técnicas anestésicas diferentes a las que nos referiremos como T1 y T2.

Figura 5. Comparación de la presión arterial sistólica en pacientes sometidos a dos técnicas anestésicas (30 minutos del inicio de la anestesia).

T1		T2
	5	4 7
	6	2
7 4	7	3 7
9 6 3	8	7 7 8 9 9
6 6 0	9	0 3 5 8
9 6 6 2	10	2 2 2
8 2 1	11	3 7
7 0	12	
2	13	
	14	
4	16	

El gráfico nos muestra las siguientes características de la TAS en los dos grupos de pacientes.

- La distribución de TAS tiene *forma* similar en ambos grupos: Un pico o moda y forma simétrica y aproximadamente acampanada.
- Diferencias en *posición*. Los pacientes del grupo T1 tienen niveles de TAS levemente mayores que los pacientes del grupo T2.
- Similar *dispersión*. Los valores de TAS de los pacientes de ambos grupos se encuentran en rangos aproximadamente iguales, salvo por el valor atípico (*outlier*) que se observa en el grupo T1.

3.2.2 HISTOGRAMA

El histograma es el más conocido de los gráficos para resumir un conjunto de datos numéricos y petende responder a las mismas preguntas que un gráfico de tallo-hojas. Una virtud del gráfico de tallo-hojas es que retiene los valores de las observaciones, sin embargo, esta característica puede ser una desventaja para gran cantidad de datos. Construir manualmente un histograma es más laborioso que construir un gráfico de tallo-hojas, pero la mayoría de los paquetes estadísticos producen histogramas.

Para construir un histograma es necesario previamente construir una *tabla de frecuencias.*

Tabla de frecuencia para datos numéricos.

A partir de una variable numérica es posible construir una *distribución de frecuencias* clasificando los datos en clases o categorías definidas por el investigador.

Las *clases o intervalos de clase* de una tabla de frecuencias deben ser mutuamente excluyentes y exhaustivas, es decir, cada dato debe caer en una y sólo una clase y todos los datos deben tener una clase a la cual pertenecen.

¿Cómo construimos una tabla de frecuencias?

- Se divide el rango total de los datos en clases o intervalos, los que no necesariamente deben tener la misma longitud.
- Se cuenta el número de observaciones que cae en cada clase y se determna la *frecuencia* en cada clase.
- Se calculan las *frecuencias relativas*, *frecuencias acumuladas* y *frecuencias acumuladas relativas* para cada intervalo.

  Notación:

  frecuencia $\Rightarrow$ f_i = número de casos que cae en el intervalo i-ésimo

  frecuencia relativa porcentual $\Rightarrow$ $fr_i = (f_i / n)\cdot 100$ = porcentaje de casos en el intervalo i-ésimo

  frecuencia acumulada $\Rightarrow fa_i = f_1 + f_2 + ... + f_i$ = suma de las frecuencias desde la primer categoría hasta la categoría i-ésima

  frecuencia acumulada relativa porcentual $\Rightarrow far_i = (fa_i / n) \cdot 100$ = suma de las frecuencias relativas desde la primer categoría hasta la categoría i-ésima .

La Tabla 4 muestra la tabla de frecuencias para los datos de tasas de neumonía cada 1000 habitantes presentados en la Tabla 2 (Año 2000, Argentina, Fuente: SINAVE). Se definieron intervalos de longitud igual a 1.

Tabla 4. Distribución de frecuencias. Tasas de notificación de neumonías por provincia, Argentina, 2000.

Intervalo	Frecuencia (f_i)	Frecuencia relativa porcentual (fr_i)	Frecuencia acumulada (fa_i)	Frecuencia relativa acumulada (fra_i)
[0, 1)	1	4.2	1	4.2
[1, 2)	3	12.5	4	16.7
[2, 3)	2	8.3	6	25.0
[3, 4)	8	33.3	14	58.3
[4, 5)	5	20.8	19	79.2
[5, 6)	1	4.2	20	83.3
[6, 7)	0	0.0	20	83.3
[7, 8)	1	4.2	21	87.5
[8, 9)	1	4.2	22	91.7
[9, 10)	1	4.2	23	95.8
[10, 11)	1	4.2	24	100.0

Notación: El intervalo [0, 1) indica el conjunto de números reales entre 0 y 1, inluye el 0 y excluye el 1.

Construcción del histograma

a) Intervalos de clase todos de la misma longitud.

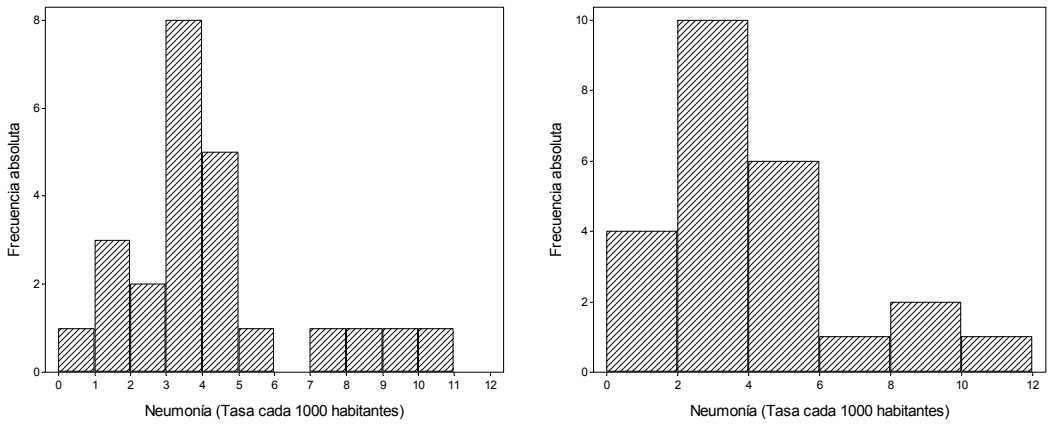
Se trazan dos ejes de coordenadas rectangulares. En el eje horizontal se representan los valores de la variable y en el eje vertical una medida de frecuencia (frecuencia absoluta, frecuencia relativa o frecuencia relativa porcentual.

Indicamos en el eje horizontal los límites de los intervalos de clase. Asociamos a cada clase una columna cuya base cubre el intervalo de clase y cuya altura indica cuantos datos "caen" en une intervalo a través de la frecuencia o la frecuencia relativa de la clase.

El gráfico se construye sin dejar espacio horizontal entre categorías, a menos que una clase esté vacía (es decir tenga altura cero).

La Figura 6 presenta dos histogramas para los datos de tasas de neumonía de la Tabla 2. El primero se construyó con intervalos de longitud unitaria, mientras que el segundo con intervalos de longitud dos.

Figura 6. Histogramas para los datos de tasas de neumonía notificadas por las provincia argentinas , Argentina, año 2000.



¿Qué características observamos en los gráficos anteriores?

- La distribución es asimétrica, con mayor concentración de datos en tasas bajas y algunas provincias con tasas altas.
- Se observan cuatro provincias con tasas de notificación de casos de neumonía más altas que el resto. Ellas son San Luis, Formosa, La Pampa y Chaco. Tal vez podríamos pensar en dos agrupamientos.
- En el histograma de la izquierda observamos un único pico (o moda) pero en el de la derecha aparenta haber dos. Es importante remarcar que características del gráfico que

no se mantienen al modificar levemente la definición de los intervalos de clase pueden ser consideraradas como artificales.

El propósito de un histograma es mostrar la *forma de la distribución* de los datos, por lo que debemos estar atentos a los aspectos visuales de la representación. Como hemos observado en el ejemplo, la forma del histograma depende del número de intervalos de clase que seleccionemos.

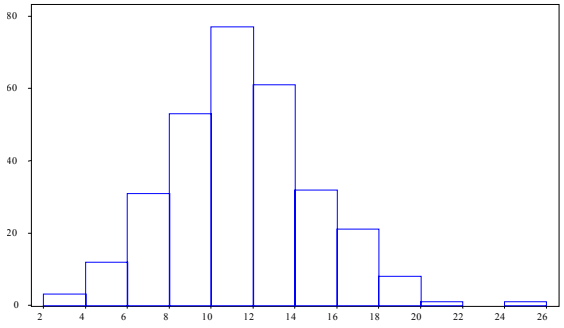
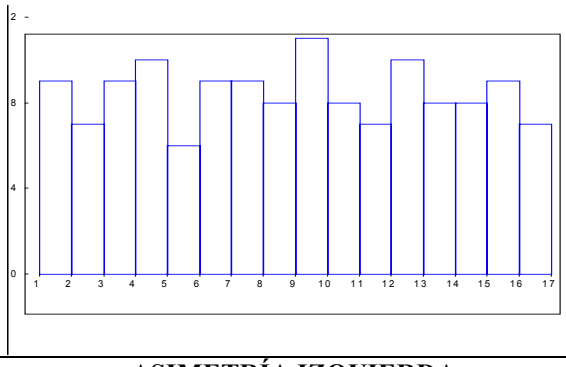
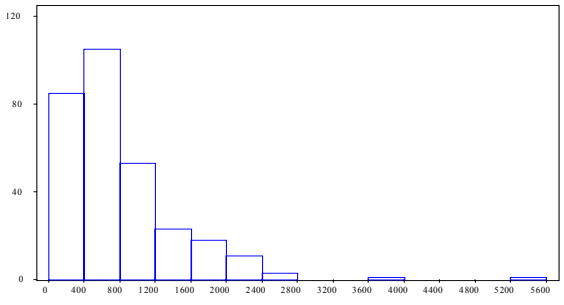
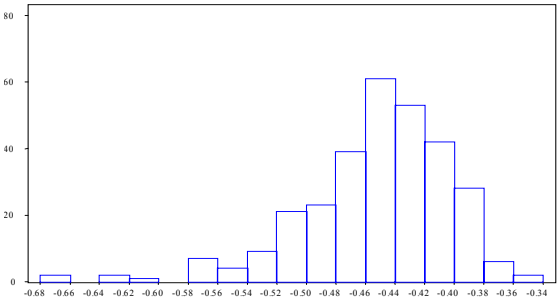
¿Cuántas clases usar?

Existen distintas fórmulas que permiten calcular el número máximo de clases apropiado para un conjunto de datos, en base al rango de datos y al número de datos.

La decisión, tal como ocurre en el gráfico de tallo-hojas, es una solución de compromiso. En general entre 6 y 15 clases resulta ser una buena elección. Muchas intervalos harán que caigan muy pocas observaciones en cada clase, por lo que las alturas de las barras variarán irregularmente. Muy pocas clases producen una gráfica más regular, pero demasiado agrupamiento puede hacer que se pierdan las características prinicipales.

¿Cómo describimos la forma de una distribución?

Los histogramas siguientes representan distintas formas posibles para la distribución de los datos. Los dos primeros muestran distribuciones aproximadamente simétricas, mientras que los dos últimos muestran distribuciones claramente asimétricas.

DISTRIBUCIÓN ACAMPANADA	**DISTRIBUCIÓN UNIFORME**
	
ASIMETRIA DERECHA	**ASIMETRÍA IZQUIERDA**
	

El histograma debería representar la frecuencia asociada a cada clase en el *área de la barra* y no en su altura. Cuando las clases son todas de la misma longitud representar la frecuencia en la altura es equivalente a representarla en el área, ya que en todas las barras el área y la altura son proporcionales.

En ocasiones es necesario construir histogramas con intervalos de clase de distinto tamaño, por ejemplo, cuando se toma información de datos sociales o económicos publicados por el estado. En estos casos, la altura de la barra debe ser tal que el *área* de la barra sea proporcional a la frecuencia. Consideraremos este tipo de histogramas a continuación.

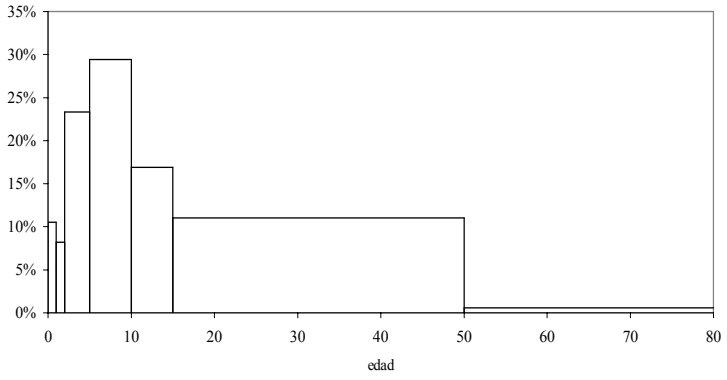
b) Intervalos de clase de diferente longitud.

Los datos de la Tabla 5 presentan los casos de rubéola notificados al SI.NA.VE durante el año 2000 según grupos de edad. Notemos que los intervalos de edad tienen diferente longitud.

Cuando (erróneamente) se construye un histograma considerando como altura de la barra la frecuencia relativa se obtiene la gráfica siguiente. La última categoría de edad se truncó arbitrariamente en 80 años para poder representarla.

Tabla 5. Notificaciones de casos de rubéola. Argentina, año 2000. Fuente: SINAVE

Intervalo (años)	**Frecuencia (f_i)**	**Frecuencia relativa (f_r)**
[0, 1)	497	10.5%
[1, 2)	387	8.2%
[2, 5)	1100	23.3%
[5, 10)	1389	29.4%
[10, 15)	798	16.9%
[15, 50)	521	11.0%
≥ 50	28	0.6%
Total	4720	100.00%



A partir de este gráfico concluiríamos que la proporción de casos es notablemente mayor en los grupo de 2 a 5 años, de 5 a 10 años o de 10 a 15 años que en los grupos de menores de 1 año o de 1 a 2 años. Además, la proporción de casos en el grupo de 15 a 50 años impresiona como notable.

El problema es que en la imagen visual asociamos la frecuencia de casos con el área de la barra, por ello parece haber mas notificaciones de gente de 15 a 50 que de cualquier otro grupo de edad.

¿Cómo construimos el histograma teniendo en cuenta que los intervalos de clase son de distinta longitud?

La barra debe tener una altura tal que el área (base x altura) sea igual a la frecuencia (o a la frecuencia relativa). Es decir,

$$\text{altura de la barra} = \frac{\text{frecuencia en el intervalo}}{\text{longitud del intervalo}}.$$

De este modo el área de la barra coincide con la frecuencia en el intervalo:

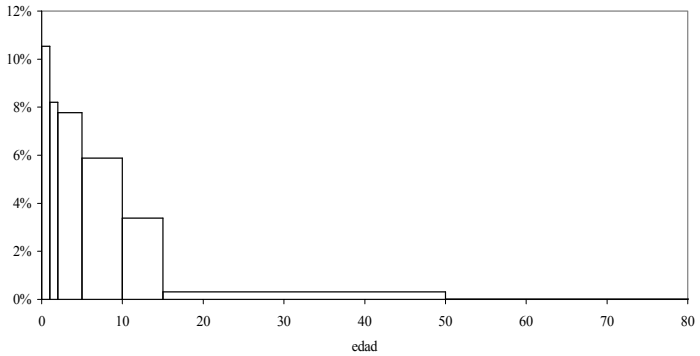
$$\textit{área} = \textit{base} \cdot \textit{altura} = \text{longitud del intervalo}\,\frac{\text{frecuencia en el intervalo}}{\text{longitud del intervalo}} = \textit{frecuencia}$$

La altura de la barra definida de este modo se denomina *escala densidad* porque indica el número de datos por unidad de la variable. La última columna de la Tabla 6 muestra la escala densidad para los datos de la Tabla 5 y la Figura 7 el histograma que se obtiene usando la escala densidad.

Tabla 6. Escala densidad. Notificaciones de casos de rubéola. Argentina, año 2000. Fuente: SINAVE.

Categoría (años)	**Frecuencia (f_i)**	**Frecuencia relativa (f_r)**	**Escala densidad**
[0, 1)	497	10.5%	10.53%
[1, 2)	387	8.2%	8.20%
[2, 5)	1100	23.3%	7.77%
[5, 10)	1389	29.4%	5.89%
[10, 15)	798	16.9%	3.38%
[15, 50)	521	11.0%	0.32%
≥ 50	28	0.6%	0.01%
Total	4720	100.00%	--

Figura 7. Histograma usando escala densidad. Notificaciones de casos de rubéola. Argentina, año 2000. Fuente: SINAVE

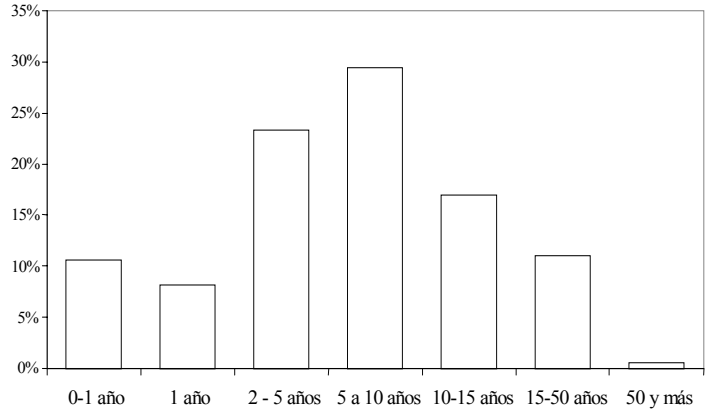


En este gráfico, el porcentaje de casos de rubéola notificados para cada grupo está representado en el *área de la barra*. El histograma muestra que una gran proporción de casos ocurre en menores de 1 año, y que la proporción desciende a medida que aumenta la edad. En este gráfico estamos representando la "densidad de notificaciones" por cada año de edad.

Comentarios

Una práctica común al manejar datos como los del ejemplo es tratar los datos como categóricos y representarlos en un gráfico de barras como el de la Figura 8.

Figura 8. Gráfico de barras. Notificaciones de casos de rubéola. Argentina, año 2000. Fuente: SINAVE



¿En que difieren un gráfico de barras y un histograma?

- El gráfico de barras no tiene en cuenta el hecho de que los intervalos de clase (grupos de edad) tienen distinta longitud.
- El gráfico de barras representa el porcentaje en la altura de la barra. Mientras que en un histograma el porcentaje se representa en el área de la barra.
- En el gráfico de barras, las barras se representan separadas para indicar que no hay continuidad entre las categorías. En un histograma barras adyacentes *deben* estar en contacto indicando que la variable es continua.

¿Cuándo usar cada uno de ellos? ¿Cuál de las dos representaciones es adecuada?

- Depende de lo que se pretenda mostrar con los datos.
- Cuando la variable que define los grupos es categórica corresponde usar un gráfico de barras.
- Cuando la variable que define las categorías es numérica, en general lo que interesa es estudiar la *distribución* de casos en las distintas edades, por lo tanto es preferible el histograma ya que la escala del eje horizontal respeta la escala de la variable de interés.

- En el ejemplo de casos de rubéola, el gráfico de barras da una impresión engañosa de la distribución de casos en las distintas edades.
- Para variables numéricas discretas con pocos valores posibles puede utilizarse un gráfico de barras.

Comentarios.

Una piramide de población es un histograma para la variable edad, con intervalos de edad de 5 años.

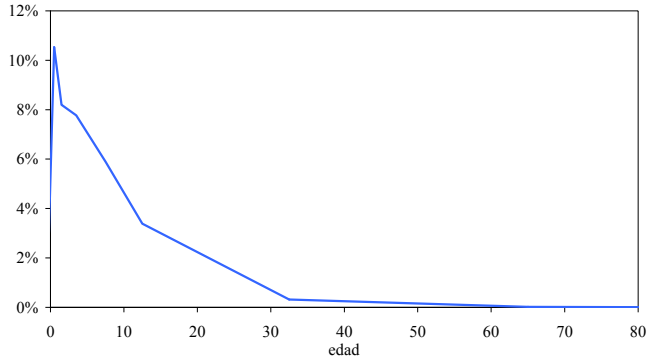
3.2.3 POLÍGONO DE FRECUENCIAS

El polígono de frecuencias es similar al histograma en muchos aspectos, pero pretende dar una imagen aproximada de la "curva" definida por la distribución de la variable.

Para cosntruirlo se usan los mismos ejes que en el histograma. Se indica en la escala horizontal el punto medio de cada intervalo y en la escala vertical la escala densidad para ese intervalo, esto define pares (x, y) en el gráfico que se unen con tramos de líneas rectas. Se marcan además los puntos medios del intervalo que precede al primero y del que sigue al último.

Para los datos de la Tabla 5 (Notificaciones de rubéola) se obtiene el gráfico de la Figura 9.

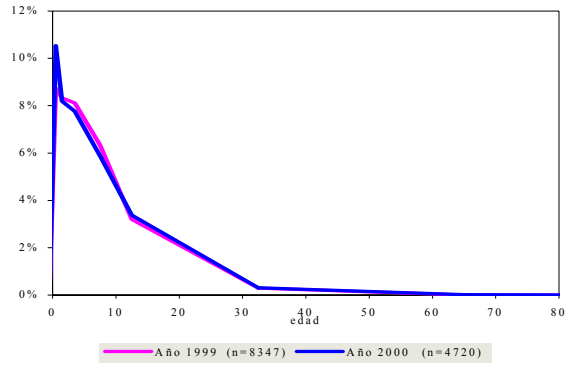
Figura 9. Polígono de frecuencias. Notificaciones de casos de rubéola. Argentina, 2000.



Los dos tipos de gráficos (histograma y polígono) brindan esencialmente la misma información. En ambos gráficos, el área total es 100%.

El polígono de frecuencias es un gráfico útil para comparar dos distribuciones de frecuencias. En la Figura 10 observamos los polígonos de frecuencia de la distribución por edad de los casos de rubéola en el año 1999 y 2000 en Argentina. A pesar de que el número de casos notificados disminuyó casi un 50% en el 2000, la distribución de edad de los casos fue muy similar los dos años.

Figura 10. Casos notificados de rubéola. Argentina, 1999 y 2000. Fuente: SINAVE



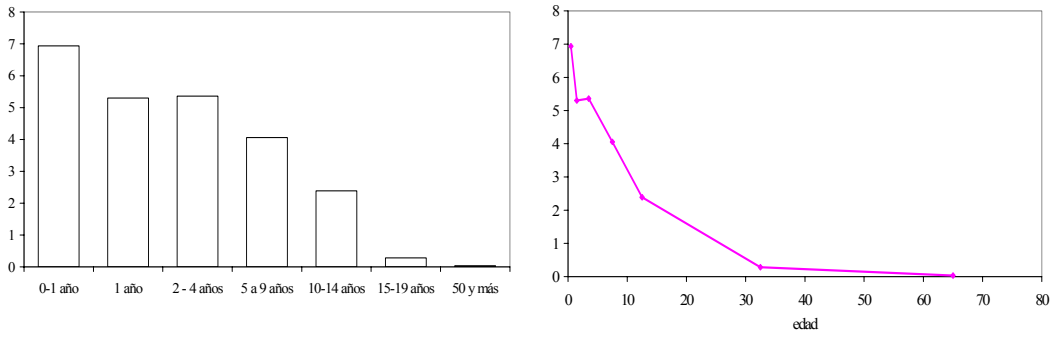
Comentario.

El histograma o el polígono de frecuencias muestran la distribución de edad de los casos de rubéola notificados durante un año, es decir, muestran la proporción del total de los casos que cae en cada categoría de edad. Pero, los distintos grupos de edad tienen distinta composición, por lo tanto, puede ser de interés presentar la *tasa* de casos de rubéola en cada grupos de edad.

Podemos representar las tasas de rubéola cada 1000 habitantes usando:

- un gráfico de barras o
- un gráfico en el que cada tasa se representa como un punto ubicado en el punto medio de la categoría de edad respetando de este modo la "distancia" entre las categorías.

Figura 11. Tasas de rubéola cada 1000 habitantes. Argentina, 2000. Fuente: SINAVE



¿Cuál de las dos representaciones preferir?

Ambas contienen la misma información, sin embargo, la segunda muestra de modo fidedigno la forma en que decae la tasa de notificación de casos de rubéola con la edad, porque preserva la escala de edad en el eje horizontal.

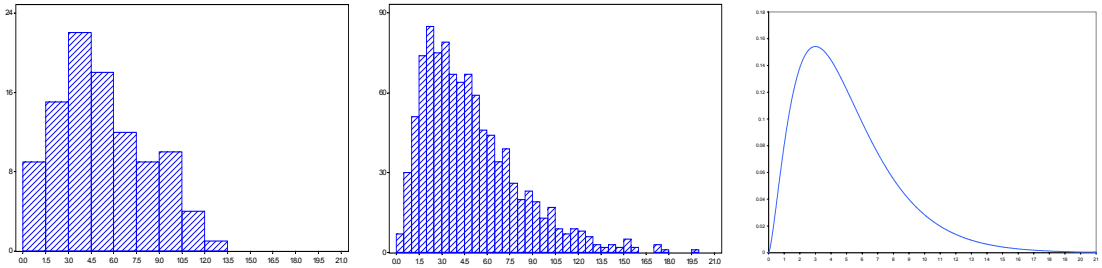
3.2.4 DISTRIBUCIÓN MUESTRAL Y POBLACIONAL.

Las distribuciones de frecuencia y los histogramas de una variable se aplican tanto a datos de una muestra como a datos de toda la población. En el primer caso hablamos de *distribución muestral* y en el segundo caso de *distribución poblacional*. En algún sentido la distribución muestral es una fotografía *borrosa* de la distribución poblacional.

A medida que el tamaño de muestra aumenta la proporción de casos que cae en cada intervalo se parece más y más a la proporción poblacional. La fotografía se torna más y más definida y la distribución muestral luce similar a la distribución poblacional.

Si la población contiene una gran cantidad de unidades de observación y la variable es continua es posible elegir intervalos tan delgados como deseemos para construir el histograma y además hacer crecer el tamaño de muestra indefinidamente. En este caso, la forma del histograma se aproximará a una *curva suave* denomina distribución de la variable en la población.

Figura 12. Histogramas para variables continuas.



Muestra n = 100 Muestra n = 1000 Población

La Figura 12 muestra dos histogramas, el primero basado en una muestra de tamaño 100 y el segundo basado en una muestra de tamaño 1000, y una curva suave que representa la distribución poblacional. Aún cuando la variable sea discreta, una curva suave suele ser una buena aproximación para la distribución poblacional, especialmente cuando el número de valores posibles de la variable es grande.

Comentaremos a continuación y a modo de cierre del tema de estadística descriptiva algunos problemas que aparecen al interpretar gráficos.

3.3 GRÁFICOS ENGAÑOSOS

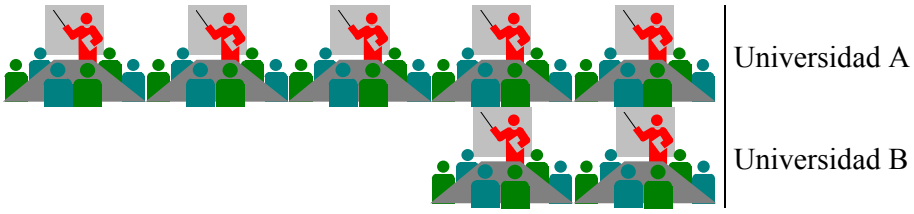
A menudo los gráficos que se presentan son engañosos, es decir, no reflejan adecuadamente los resultados o exageran ciertas características de los datos. Veremos algunas situaciones.

3.3.1 DIBUJOS

En la Figura 13 se representa el número de conferencias organizadas en todos los departamentos de la Universidad A y la Universidad B, en el año 2000. Cada ícono representa 20 conferencias, por lo tanto, el gráfico informa que en la Universidad A se

organizaron aproximadamente 100 conferencias en tanto que en B se organizaron 40. La información que brinda el gráfico es equivalente a la información numérica.

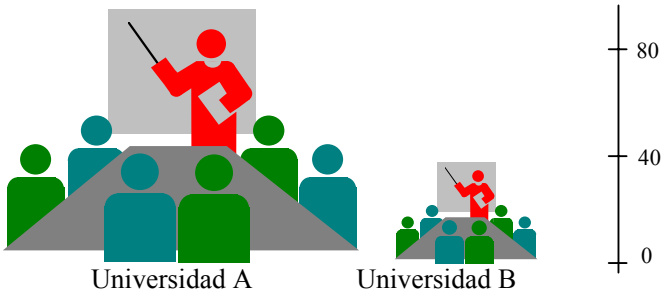
Figura 13. Número de conferencias organizadas por las Universidades A y B en 2000 (*).



(*) *Cada ícono representa 20 conferencias.*

Cuando la representación se realiza utilizando símbolos que cambian de tamaño, la imagen puede resultar engañosa, tal como ocurre al representar los datos anteriores en la la Figura 14. En esta Figura, la altura del ícono indica el número de conferencias. La impresión visual es engañosa porque no está claro cual de las dimensiones de la figura representa la magnitud de la variable. En general, frente a dibujos que no tienen la misma base, tendemos a comparar áreas.

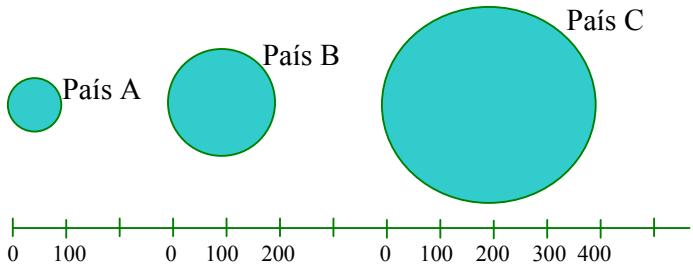
Figura 14. Número de conferencias organizadas por las Universidades A y B en 2000(*).



(*) *El número de conferencias se representa en la altura del ícono.*

La Figura 15 es otro ejemplo de la misma situación. Como las magnitudes se representan en el diámetro, aún cuando el diámetro de B es el doble que el de A, como el área de B es 4 veces la de A, el gráfico produce una impresión engañosa.

Figura 15. Deuda externa de 3 países (en miles de millones de dolares) (a).



(a) *La deuda se representa en el diámetro.*

El punto clave aquí es que aún cuando el gráfico es correcto, sólo será correctamente interpretado por los pocos lectores acostumbrados a leer los detalles de las notas al pie.

Capítulo 4. MEDIDAS RESÚMENES

En el Capítulo anterior presentamos métodos gráficos para datos cualitativos y cuantitativos. En este capítulo introduciremos distintas formas de resumir la distribución muestral o poblacional de una variable NUMÉRICA y finalmente presentaremos un tipo de gráfico que se construye a partir de medidas resúmenes.

Resumir un conjunto de datos es pasar de una visión detallada a una generalización simple e informativa tratando de preservar las características esenciales.

¿Por qué resumir? Para simplificar la comprensión y la comunicación de los datos.

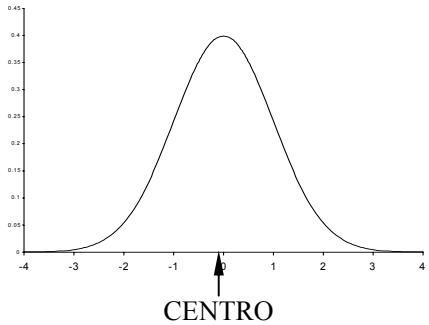
Las medidas resúmenes son útiles para comparar conjuntos de datos cuantitativos y para presentar los resultados de un estudio y se clasifican en dos grupos principales:

Medidas de posición o localización $\Rightarrow$	describen un valor alrededor del cual se encuentran las observaciones
Medidas de dispersión o escala $\Rightarrow$	pretenden expresar cuan variable es un conjunto de datos.

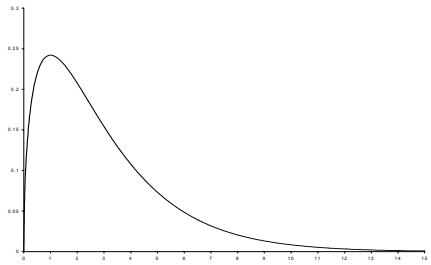
4.1 MEDIDAS DE POSICIÓN O LOCALIZACIÓN

Un modo de resumir un único conjunto de datos numéricos es a través de un número que debería ser *típico* para el grupo. No debería ser ni demasiado grande, ni demasiado pequeño y debería estar tan cerca del "centro" de la distribución como sea posible.

Por lo tanto, una *medida de posición* es un número que pretende indicar dónde se encuentra el *centro* de la distribución de un conjunto de datos. Pero, ¿dónde se encuentra el "centro" de una distribución?



CENTRO



¿CENTRO?

El centro es fácil de identificar si la distribución es simétrica, pero es difícil si la distribución es asimétrica. Por esta razón, no hay una única medida de posición para resumir una distribución. Si la distribución es simétrica diferentes medidas conducirán a similares resultados. Si la distribución es claramente asimétrica diferentes propuestas apuntarán a distintos conceptos de "centro" y por lo tanto los valores serán diferentes.

A los efectos de resumir los datos debemos preguntarnos:

- ¿Qué medida resumen es la más apropiado para la distribución que presentan nuestros datos?

- ¿Qué propuesta permite responder mejor a las preguntas sobre el mundo real que pretendemos responder con estos datos?

4.1.1 EL PROMEDIO O LA MEDIA ARITMÉTICA

Es la medida de posición más frecuentemente usada. Para calcular la *media aritmética* o *promedio* de un conjunto de observaciones se suman todos los valores y se divide por el número total de observaciones.

Definición

Si tenemos una muestra de *n* observaciones y denotadas por X_1, X_2, ..., X_n, definimos la *media muestral* $\overline{X}$ del siguiente modo:

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

El símbolo $\sum_{i=1}^{n} X_i$ indica la suma de todos los valores obesrvados de la variable desde el primero (i = 1) hasta el último (i = n).

Ejemplo.

$X_1 = 10$ $X_2 = 14$ $X_3 = 12$ $X_4 = 11$ $X_5 = 12$ $X_6 = 13$

$$\overline{X} = \frac{X_1 + X_2 + ... + X_6}{n} = \frac{10+14+12+11+12+13}{6} = \frac{72}{6} = 12$$

Media poblacional

Si se dispone de la información de una variable X para las N unidades de análisis de la población, es posible calcular la *media poblacional* a la que denotaremos con la letra griega μ (mu), para distinguirla de la media obtenida en una muestra de *n*

$$\mu = \frac{X_1 + X_2 + ... + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N}.$$

Media de datos agrupados

Supongamos que se dispone de dos conjuntos de datos en los que se conoce la media y el número de datos de cada uno de ellos ($\overline{X}_1, n_1$ y $\overline{X}_2, n_2$). Calculamos la media de los $n_1 + n_2$ datos como el *promedio pesado*

$$\overline{X} = \frac{n_1 \overline{X}_1 + n_2 \overline{X}_2}{n_1 + n_2}$$

Ejemplo. Datos sobre niveles de hierro sérico en niños y niñas con fibrosis cística.

X = nivel de hierro sérico

	Varones	Mujeres
$\overline{X}$	5.9	6.8
n	13	6

$$\overline{X} = \frac{n_1 \overline{X}_1 + n_2 \overline{X}_2}{n_1 + n_2} = \frac{13 \cdot 5.9 + 6 \cdot 6.8}{13 + 6} = \frac{(76.7 + 40.8)}{19} = \frac{117.5}{19} = 6.18$$

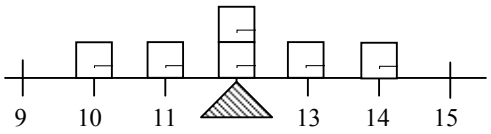
El promedio pesado obtenido aquí es *igual* al que hubiéramos obtenido promediando los datos de los 19 niños.

Características y propiedades de la media.

a) Se usa para datos numéricos.

b) Representa el *centro de gravedad* o el punto de equilibrio de los datos.

Podemos imaginar a los datos como un sistema físico, en el que cada dato tiene una "masa" unitaria y lo ubicamos sobre una barra en la posición correspondiente a su valor. La media representa la posición en que deberíamos ubicar el punto de apoyo para que el sistema esté en equilibrio.



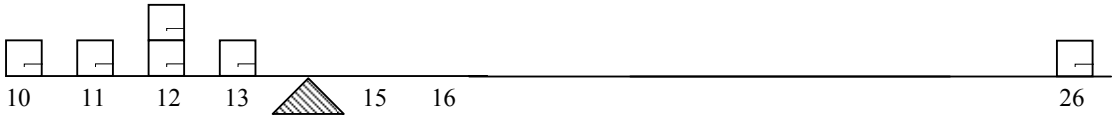
c) La suma de las distancias de los datos a la media es cero. Esta propiedad está relacionada con el hecho que la media es el centro de gravedad de los datos.

En la tabla siguiente comprobamos esta propiedad para los datos del ejemplo anterior.

X_i	$X_i - \overline{X}$
10	-2
14	2
12	0
111	-1
12	0
13	1
Total =	0

d) Es muy sensible a la presencia de datos atípicos (OUTLIERS).

Modificamos 1 dato en el ejemplo anterior $X_2 = 14 \rightarrow X_2 = 26$ y $\overline{X} = 12 \rightarrow \overline{X} = 14$.



Con solo modificar un dato la media se desplazó tanto, que ya no se encuentra entre la mayoría de los datos. Podemos decir que en este caso la media no es una buena medida

de posición de los datos. En consecuencia, la media es una buena medida del centro de la distribución cuando ésta es simétrica.

Aunque la media es una medida simple de tendencia central, otras medidas son más informativas y ocasionalmente más apropiadas.

4.1.2 LA MEDIANA MUESTRAL

La *mediana* es el dato que ocupa la posición central en la muestra ordenada de menor a mayor.

¿Cómo calculamos la mediana de una muestra de n observaciones?

1. Ordenamos los datos de menor a mayor.
2. La mediana es el dato que ocupa la posición $\left(\frac{n+1}{2}\right)$ en la lista ordenada.

   Si el número de datos es *impar*, la mediana $\tilde{X}$ es el dato que ocupa la posición central.

   Si el número de datos es *par*, la mediana $\tilde{X}$ es el promedio de los dos datos centrales.

Ejemplo

- *n impar*

  $X_1 = 10 \qquad X_2 = 14 \qquad X_3 = 12 \qquad X_4 = 18 \qquad X_5 = 11$

  Ordenamos los datos:

  $10 \quad 11 \quad 12 \quad 14 \quad 18$

  La posición de la mediana es $\frac{n+1}{2} = \frac{5+1}{2} = 3$ (tercer dato), es decir $\tilde{X} = 12$.

- *n par*

  $X_1 = 10 \qquad X_2 = 14 \qquad X_3 = 12 \qquad X_4 = 18 \qquad X_5 = 11 \qquad X_6 = 23$

  Ordenamos los datos:

  $10 \quad 11 \quad 12 \quad 14 \quad 18 \quad 23$

  Posición de la mediana $\Rightarrow \frac{6+1}{2} = 3.5$

  Obtenemos la mediana promediando el tercer y cuarto dato: $\tilde{X} = \frac{12+14}{2} = 13$.

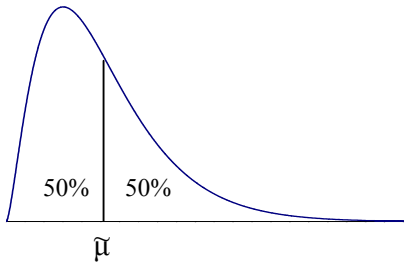
Notar que (n+1)/2 no es la mediana, sino la localización de la mediana en el conjunto ordenado de datos.

Si hay datos repetidos deben ser incluidos en el ordenamiento.

La mediana es muy simple de obtener a partir de un gráfico de tallo-hojas.

Mediana poblacional

La *mediana poblacional* se define de modo equivalente a la mediana muestral y es el valor de la variable por debajo del cual se encuentra a lo sumo el 50% de la población y por encima del cual se encuentra a lo sumo el 50% de la población. La denotamos como $\tilde{\mu}$.



Propiedades de la mediana

a) La mediana puede ser usada no sólo para *datos numéricos* sino además para *datos ordinales*, ya que para calcularla sólo es necesario establecer un orden en los datos.

b) Si la distribución de los datos es aproximadamente simétrica la media y la mediana serán aproximadamente iguales.

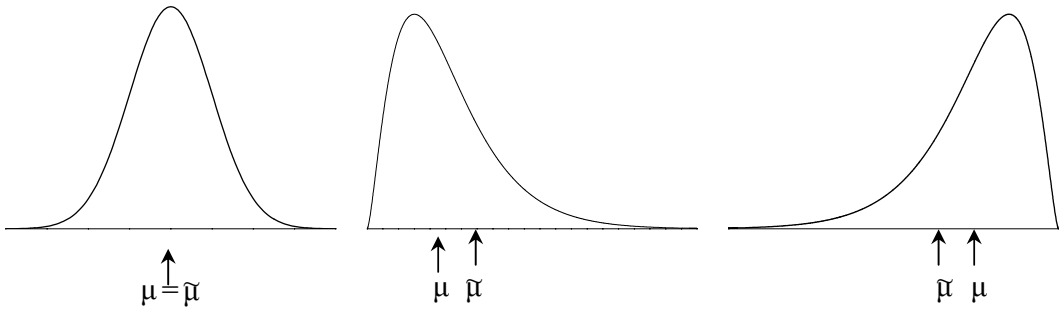
   Si la distribución de los datos es asimétrica, la media y la mediana diferirán según el siguiente patrón:

   Asimetría derecha (cola larga hacia la derecha) $\Rightarrow$ $\overline{X} > \tilde{X}$
   Asimetría izquierda (cola larga hacia la izquierda) $\Rightarrow$ $\overline{X} < \tilde{X}$

   Ejemplos

   1) 12, 13, 14, 15, 16 $\overline{X} = \tilde{X} = 14$
   2) 12, 13, 14, 15, 20 $\overline{X} = 15 > \tilde{X} = 14$
   3) 2, 13, 14, 15, 16 $\overline{X} = 12 < \tilde{X} = 14$

   En la población:



c) La mediana es una medida de posición *robusta*. No se afecta por la presencia de datos outliers, salvo que modifiquemos casi el 50% de los datos menores o mayores de la muestra (la proporción de datos que debemos modificar para modificar la mediana depende del número de datos de la muestra).

Ejemplo

I) 10 11 12 12 13 14 $\overline{X} = 12$ $\tilde{X} = 12$
II) 10 11 12 12 13 26 $\overline{X} = 14$ $\tilde{X} = 12$

d) La mediana es insensible a la distancia de las observaciones al centro, ya que solamente depende del orden de los datos. Esta característica que la hace robusta, es una desventaja de la mediana.

Ejemplo. Todos los conjuntos de datos siguientes tienen mediana 12

I) 10 11 12 13 14
II) 10 11 12 13 100
III) 0 11 12 12 12
IV) 10 11 12 100 100

e) Si hay datos censurados en la muestra no es posible calcular la media, sin embargo, eventualmente puede calcularse la mediana.

Ejemplo

Tiempo de supervivencia (en meses) de pacientes con cierta patología. Los datos que se indican entre paréntesis tienen censura derecha, es decir, se sabe que el paciente sobrevivió ese tiempo, pero no se conoce el tiempo real de supervivencia.

I) 1 5 10 12 18 24 25 28 39 45 (45) 48 50 51 (84) n = 15

Como $n = 15$ la mediana es el octavo dato, por lo tanto $\tilde{X} = 28$. Es posible calcularla aunque haya datos censurados, porque los mismos se encuentran más allá de la posición 8 que define la mediana. Aunque no conocemos exactamente el tiempo que sobrevivió el paciente cuyo dato es (45) sabemos que en esta muestra ese dato ocupará la posición 11 o una superior.

II) 1 5 10 (12) 18 24 25 28 39 45 (45) 48 50 51 (84) n = 15

No es posible calcular la mediana debido al dato indicado como (12). Sabemos que este paciente sobrevivió por lo menos 12 meses, pero desconocemos el verdadero valor, el que puede ocupar cualquier posición entre la quinta y la última.

Comparación de la media y la mediana

	MEDIA	MEDIANA
VENTAJAS	Usa toda la información que proveen los datos. Es de manejo algebraico simple.	Representa el centro de la distribución (en un sentido claramente definido). Robusta a la presencia de outliers. Útil para datos ordinales.
DESVEN-TAJAS	Muy sensible a la presencia de datos outliers.	Usa muy poca información de los datos.

4.1.3 LA MEDIA α-PODADA

La media α-podada es un compromiso entre las dos medidas de posición presentadas. Es una medida más robusta que la media, pero que usa más información que la mediana.

La media α-podada se calcula despreciando $n.\alpha$ datos de cada extremo y promediando las observaciones centrales del conjunto ordenado de datos.

¿Cómo calculamos la media α-podada de una muestra de n observaciones?

1. Ordenamos los datos de menor a mayor.
2. Excluimos los $n.\alpha$ datos más pequeños y los $n.\alpha$ datos más grandes.
3. Calculamos el promedio de los datos restantes y lo denominamos $\overline{X}_\alpha$.

¿Cómo elegimos α?

Depende de cuantos outliers se pretende excluir y de cuán robusta queremos que sea la medida de posición. Cuando seleccionamos $\alpha = 0$ tenemos la media, si elegimos el máximo valor posible para α (lo más cercano posible a 0.5) tenemos la mediana. Cualquier poda intermedia representa un compromiso entre ambas.

Una elección bastante común es $\alpha = 0.10$, que excluye un 20% de los datos.

¿Cuándo usar esta medida?

Cuando se sospecha que hay errores groseros en los datos, pero no tenemos modo de decidir si el dato es erróneo. Esto permite excluir datos aberrantes de un modo menos sesgado, porque estamos excluyendo datos de ambos extremos.

Ejemplo

Calculamos la media 20% podada para los datos siguientes que corresponden a los puntajes asignados a una gimnasta por 5 jueces durante una competencia olímpica.

$X_1 = 85 \qquad X_2 = 98 \qquad X_3 = 99 \qquad X_4 = 95 \qquad X_5 = 98$

1. Ordenamos los datos: 85 95 98 98 99
2. Calculamos el número de datos que podaremos en cada extremo

$$n \cdot \alpha = 5 \cdot 0.20 = 1$$

   Excluimos el primer y el último dato de la muestra ordenada.

3. Promediamos los datos restantes

$$\overline{X}_{0.20} = \frac{95 + 98 + 98}{3} = 97 .$$

Para estos datos el promedio y la mediana resulta ser $\overline{X} = 95$, $\tilde{X} = 98$.

¿Qué ventaja tiene haber usado la media 20% podada? El puntaje final de la gimnasta no se ve afectado por la calificación notablemente baja que le asignara uno de los jueces.

¿Qué hacer cuando el número de datos que debe excluirse no es entero?

Si $n = 37$ y quisiéramos una poda del 10% deberíamos excluir $37 \cdot 0.10 = 3.7$ datos de cada extremo. Las opciones son:

- Seleccionar una poda menor o igual que α. En este caso podamos 3 datos de cada extremo e informamos que se calculó la media 8.1% podada.
- Calculamos la media podando 3 datos y luego la media podando 4 datos de cada extremo y finalmente calculamos un promedio ponderado de estas dos medidas.

¿Cuál de las tres medidas de posición preferir: media, mediana o media α-podada?

Si la distribución de la variable es *simétrica* las tres medidas deberían dar resultados similares. En este caso, es preferible usar la *media* ya que es la que tiene menor error de estimación. Esto es, la distancia entre la media muestral y la verdadera media poblacional en promedio es menor que la distancia entre la mediana o la media α-podada y la media poblacional.

Si la distribución es asimétrica o con outliers generalmente es preferible resumir los datos con la mediana o la media α-podada, ya que la estimación obtenida en una muestra en promedio se encuentra más cercana al correspondiente parámetro (media poblacional y mediana poblacional).

4.1.4 LA MODA

La moda es el dato que ocurre con mayor frecuencia en el conjunto.

Es una medida de poca utilidad salvo para datos categóricos en los que suele interesar identificar la categoría con mayor cantidad de datos. En una muestra de datos numéricos, puede ocurrir que la moda sea un valor que se repite un cierto número de veces, pero que no es típico.

Cuando se considera la distribución poblacional de una variable continua, decimos que esta es UNIMODAL si presenta un pico y BIMODAL si aparecen dos picos claros.

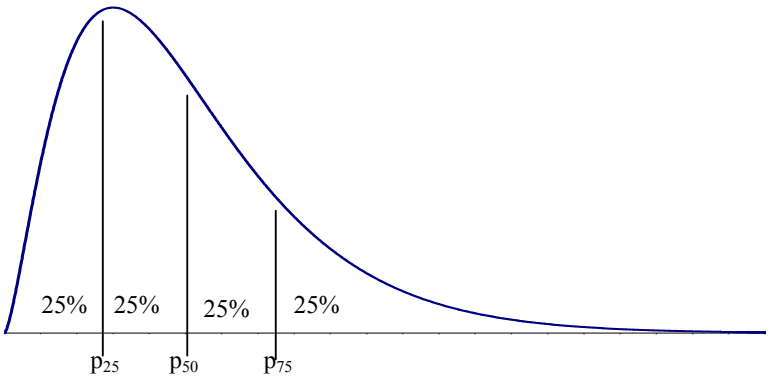
4.1.5 CUARTILES Y OTROS PERCENTILES

Los percentiles son otro modo de resumir una distribución muestral o poblacional.

El *percentil p%* de un conjunto de datos es la observación que deja a lo sumo p% de las observaciones debajo de él y a lo sumo (1 – p)% encima de él.

Como ejemplo, consideremos la distribución de peso de recién nacidos de sexo femenino y 38 semanas de gestación Si se informa que el percentil 10% de esta distribución es 2450 g y el percentil 90% es 3370g, estamos indicando que un 10% de las niñas que nacen en la semana 38 de gestación pesan 2450 g o menos (y en consecuencia, 90% pesan más que 2450 g) y que el 90% de las niñas de esta edad gestacional nacen con peso menor o igual que 3370 g (y sólo el 10% con peso mayor que 3370 g).

La mediana es el percentil 50%. Otros percentiles con nombre propio son el percentil 25% y el percentil 75% que se denominan *cuartil inferior y superior* respectivamente, ya que juntamente con la mediana dividen a la distribución en 4 porciones iguales.



¿Cómo se calculan los cuartiles de una muestra de n observaciones?

1. Ordenar los datos de menor a mayor.
2. El cuartil inferior es el dato que ocupa la posición $(n+1)/4$ en la muestra ordenada.
3. El cuartil superior es el dato que ocupa la posición $3(n+1)/4$ en la muestra ordenada.

   Si la posición resulta ser un número decimal, promediamos los datos que se encuentran a izquierda y derecha de la posición obtenida.

Ejemplo

Consideremos los siguientes datos ordenados (n = 13).

Posición	1	2	3	4	5	6	7	8	9	10	11	12	13
Datos	104	112	134	146	155	168	170	195	246	302	338	412	678

Posición del Cuartil Inferior = (13+1)/4 = 3.5 $\Rightarrow C_I = \frac{134+146}{2} = 140$

Posición de la mediana = (13+1)/2 = 7 $\Rightarrow \tilde{X} = 170$

Posición del Cuartil Superior = 3.(13+1)/4 = 10.5 $\Rightarrow C_S = \frac{302+338}{2} = 320$

Cinco números resúmenes

Un modo de resumir toda la distribución de los datos es informar los siguientes *cinco números resúmenes:*

Mínimo, Cuartil inferior, Mediana, Cuartil superior, Máximo

En nuestro ejemplo:

Mínimo =	104	25%
Cuartil Inferior =	140	25%
Mediana =	170	25%
Cuartil Superior =	320	25%
Máximo =	678	

Comentarios

Los paquetes estadísticos calculan los percentiles usando diferentes métodos, y diferentes criterios para interpolar. El modo de cálculo que presentamos aquí para los cuartiles tiene la ventaja de su simplicidad. Cuando el conjunto de datos es grande los distintos métodos tienden a producir el mismo valor para el percentil, pero para conjuntos pequeños pueden diferir ligeramente.

Los percentiles son modos muy útiles de resumir la distribución de datos censurados. Es posible calcular un percentil siempre que todos los datos tengan el mismo tipo de censura y queden a la derecha (cuando la censura es derecha) o a la izquierda (cuando la censura es izquierda) de la posición que define el percentil.

4.2 MEDIDAS DE DISPERSIÓN O VARIABILIDAD

Las medidas de posición dan una idea de dónde se encuentra el centro de la distribución, pero no nos dicen cuán disperso es el conjunto de datos. Consideremos los siguientes conjuntos de datos:

Muestra A:	55	55	55	55	55	55	55
Muestra B:	47	51	53	55	57	59	63
Muestra C:	39	47	53	55	57	63	71

En todos ellos $\overline{X} = \tilde{X} = 55$, pero las muestras difieren notablemente.

Las *medidas de dispersión o variabilidad* describen cuán cercanos se encuentran los datos entre ellos, o cuán cerca se encuentran de alguna medida de posición. Introduciremos a continuación algunos estadísticos que miden variabilidad del conjunto de datos.

4.2.1 RANGO MUESTRAL

El *rango* de *n* observaciones $X_1, X_2, ..., X_n$ es la diferencia entre la observación más grande y la más pequeña,

$$\text{Rango} = \text{máx}(X_i) - \text{mín}(X_i)$$

Ejemplo

Muestra A:	55	55	55	55	55	55	55	Rango = 55 – 55 = 0
Muestra B:	47	51	53	55	57	59	63	Rango = 63 – 47 = 16

Muestra C: 39 47 53 55 57 63 71 Rango = 71 – 39 = 32

Características y propiedades

- Es muy simple de obtener.
- Es extremadamente sensible a la presencia de datos atípicos. Si hay datos outliers, estos estarán en los extremos, que son los datos que se usan para calcular el rango.
- Ignora la mayoría de los datos.
- En general aumenta cuando aumenta el tamaño de la muestra (las observaciones atípicas tienen más chance de aparecer en una muestra con muchas observaciones).

En consecuencia, reportar el rango o el máximo y el mínimo de un conjunto de datos, no informa demasiado sobre las características de los datos. A pesar de esto es frecuente encontrar en las publicaciones científicas datos numéricos resumidos a través de una medida de posición acompañada por los valores mínimo y máximo.

4.2.2 DESVIACIÓN ESTÁNDAR Y VARIANZA MUESTRAL

La *desviación estándar* mide cuan lejos se encuentran los datos de la media muestral.

Un modo de medir la variabilidad de los datos de una muestra sería tomar algún valor central, por ejemplo la media, y calcular el promedio de las distancias a ella. Mientras mayor sea este promedio, más dispersión deberían presentar los datos.

Sin embargo, esta idea no resulta útil, ya que las observaciones que se encuentran a la derecha de la media tendrán distancias (o desviaciones) positivas, en tanto que las observaciones menores que la media tendrán distancias negativas y la suma de las distancias a la media será inevitablemente igual a cero.

Un modo de evitar este inconveniente es elevar las distancias al cuadrado y de este modo tener todos sumandos positivos.

Definimos la *varianza* de una muestra de observaciones X_1, X_2, ..., X_n, cuya media es $\overline{X}$, como

$$s^2 = \frac{(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}.$$

La varianza muestral puede pensarse como “promedio” de las distancias a la media al cuadrado.

Sin mebargo, la varianza no tiene las mismas unidades que los datos. Para salvar este inconveniente, definimos la *desviación estándar muestral* como la raíz cuadrada positiva de la varianza

$$s = \sqrt{s^2}\ .$$

Varianza y desviación estándar poblacional

Si se dispone de la información de una variable X para las N unidades de análisis de la población, denotamos con σ^2 y σ (sigma) *la varianza y la desviación estándar de la población* respectivamente y las definimos del siguiente modo:

$$\sigma^2 = \frac{(X_1 - \mu)^2 + \cdots + (X_n - \mu)^2}{N} = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N} \qquad \sigma = \sqrt{\sigma^2}$$

La razón para usar $(n - 1)$ y no n en el denominador de la varianza muestral tiene que ver con el hecho de que el valor de s^2 obtenido en una muestra, se usa para estimar la varianza poblacional σ^2. Definida con $(n - 1)$ en el denominador la varianza muestral posee una propiedad deseable, resulta ser *insesgado*, esto es, en promedio no subestima ni sobrestima el valor de la varianza poblacional.

Ejemplo

Muestra A:	55	55	55	55	55	55	55	$s^2 = 0$	$s_A = 0$
Muestra B:	47	51	53	55	57	59	63	$s^2 = 28$	$s_B = 5.29$
Muestra C:	39	47	53	55	57	63	71	$s^2 = 108$	$s_C = 10.39$

Calculamos la varianza y el desvío estándar para la Muestra B. Se deja como ejercicio verificar que los resultados obtenidos para A y C son correctos.

$$s_B^2 = \frac{(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1} = \frac{(47-55)^2 + (51-55)^2 + \cdots + (63-55)^2}{7-1}$$

$$= \frac{(-8)^2 + (-4)^2 + \ldots + 8^2}{6} = \frac{168}{6} = 28$$

$$s_B = \sqrt{28} = 5.29$$

Comparando las desviaciones estándar de las tres muestras vemos que $s_A < s_B < s_C$. Además observamos que $s_A = 0$, ya que todas las observaciones toman el mismo valor.

Interpretación del valor de la desviación estándar

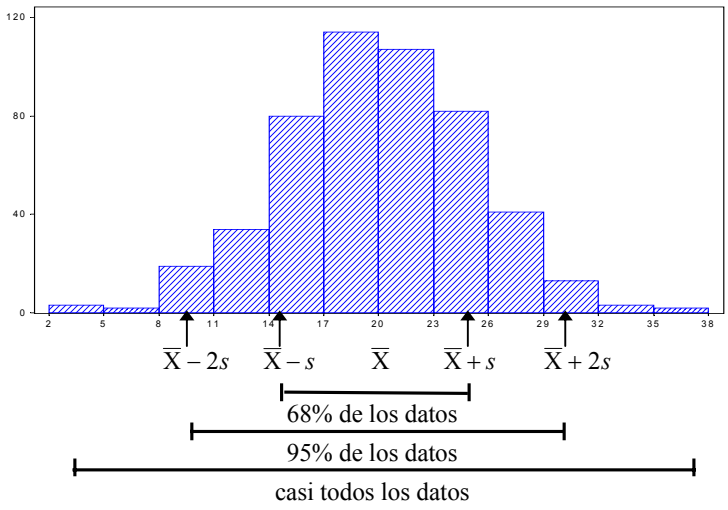
La desviación estándar s es útil para comparar la variabilidad de dos conjuntos de datos en los que la variable a sido medida en las mismas unidades. Si en una muestra $s = 5.4$ y en otra $s = 10.4$ podemos asegurar que los datos de la segunda muestra están más dispersos que los de la primera. Pero ¿cómo interpretamos el valor $s = 5.4$?

La desviación estándar nos da idea de la distancia promedio de los datos a la media (aunque estrictamente hablando no es el promedio). Pero la interpretación de s requiere algún conocimiento de la distribución de los datos.

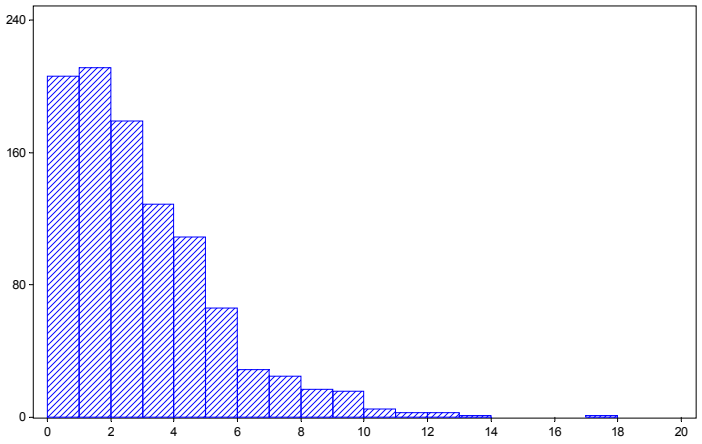
Regla empírica

Si el histograma de los datos es aproximadamente simétrico y acampanado entonces,

- Aproximadamente el 68% de las observaciones caen en el intervalo $\overline{X} - s$ y $\overline{X} + s$.
- Aproximadamente el 95% de las observaciones caen en el intervalo $\overline{X} - 2s$ y $\overline{X} + 2s$.
- Prácticamente todas las observaciones caen en el intervalo $\overline{X} - 3s$ y $\overline{X} + 3s$.



Esta regla es válida para distribuciones no necesariamente acampanadas, pero puede ser errónea cuando se aplica a distribuciones fuertemente asimétricas tales como la que se presenta en el histograma siguiente en el que $\overline{X} = 3$ y s = 2.45. Esta distribución ficticia podría representar la distribución de ingreso mensual (en cientos de pesos) de una muestra de asalariados con cargos no jerárquicos de una provincia Argentina.



¿Es útil nuestra regla empírica para el desvío estándar en datos con esta distribución? En este caso, al restar $2s$ a la media, caemos fuera de la escala de la variable $\overline{X} - 2s = 3 - 2 \cdot 2.45 = -1.9$ y la interpretación que propusimos a través de la regla empírica resulta no ser apropiada.

Cuando la variable sólo puede tomar valores dentro de un cierto rango, tal como ocurre con el ingreso o el tiempo transcurrido hasta un cierto evento que no pueden ser menores que cero, el hecho de obtener valores fuera del rango al aplicar la regla con 1 o 2 desvíos estándar nos indica que la distribución de la variable es fuertemente asimétrica.

Propiedades de la desviación estándar

- s mide la dispersión alrededor de la media, por lo tanto es natural elegir esta medida de dispersión cuando se usa la media como medida de posición.

- $s = 0$ solamente cuando todos los datos son iguales, de otro modo $s > 0$.
- s es una medida de dispersión *muy sensible* a la presencia de datos outliers. De hecho, es más sensible que la media ya que las distancias están elevadas al cuadrado.

Presentaremos a continuación dos medidas de dispersión robustas.

4.2.3 MEDIANA DE LAS DESVIACIONES ABSOLUTAS (MAD)

La MAD (median absolute deviations) es otra medida de dispersión que pretende dar una idea resumen de "distancias a un punto central" tal como ocurre con el desvío estándar. Pero, ¿en qué difiere del desvío estándar?

- Considera la mediana como punto central de la distribución para calcular las desviaciones.
- Toma el valor absoluto de las desviaciones para eliminar el signo (en vez de elevar al cuadrado como hacemos al calcular el desvío estándar).
- Toma la mediana de las distancias (en vez de promediar como hacemos con s).

Definimos la *MAD* de una muestra $X_1, X_2, ..., X_n$ como

$$MAD = mediana\ (|\ X_i - \tilde{X}\ |)$$

¿Cómo calculamos la MAD?

1. Ordenamos los datos de menor a mayor.
2. Calculamos la mediana.
3. Calculamos la distancia de cada dato a la mediana.
4. Despreciamos el signo de las distancias y las ordenamos de menor a mayor.
5. Buscamos la mediana de las distancias sin signo.

Propiedades de la MAD

- Si la distribución es acampanada y simétrica la MAD y el desvío estándar s se relacionan del siguiente modo:

$$s \cong 1.48\ \text{MAD}$$

- La MAD es una medida de dispersión muy robusta a la presencia de datos outliers.

Ejemplo

Consideremos los siguientes datos ordenados (n = 13).

Posición	1	2	3	4	5	6	7	8	9	10	11	12	13
Datos	104	112	134	146	155	168	170	195	246	302	338	412	678

1. Como $n = 13$ la mediana es el dato que ocupa la posición $(13+1)/2 = 7 \Rightarrow \tilde{X} = 170$.

2. Calculamos las diferencias a la mediana

   –66, –58, –36, –24, –15, –2, 0, 25, 76, 132, 168, 242, 508

3. Despreciamos el signo de las distancias y las ordenamos de menor a mayor

   0, 2, 15, 24, 25, 36, 58, 66, 76, 132, 168, 242, 508

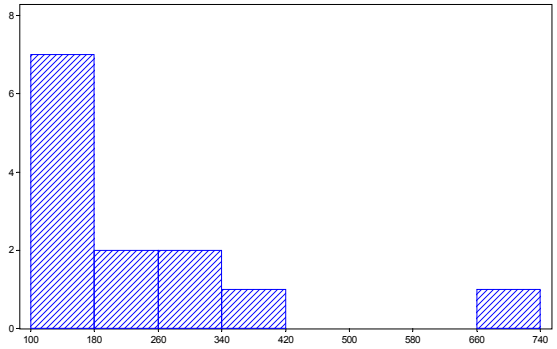
4. Tenemos $n = 13$ diferencias, por lo tanto la mediana es la diferencia que ocupa el séptimo lugar, en consecuencia

$$\text{MAD} = 58$$

Si la distribución fuera simétrica esperaríamos que el desvío estándar fuera

$$s \cong 1.48\ \text{MAD} = 1.48 \cdot 58 = 85.8$$

pero para estos datos $s = 160.48$. Esta gran diferencia nos dice que la distribución es asimétrica. El histograma de estos datos, que se presenta en la figura siguiente confirma este hecho.



4.2.4 DISTANCIA O RANGO INTERCUARTIL

El *rango intercuartil o distancia intercuartil* (D_I) de un conjunto de datos es la distancia entre los dos cuartiles:

$$D_I = C_S - C_I$$

Indica el rango donde se encuentra aproximadamente el 50% "central" de las observaciones.

Propiedades

- Si todos los datos son iguales $D_I = 0$. Pero D_I puedes ser igual a cero aún cuando no todos los datos sean iguales.

  Ejemplo 5 12 12 12 12 12 20 $n = 7$ $C_I = 12$ $C_S = 12$ $D_I = 0$

- Es una medida robusta de dispersión.
- Cuando la distribución es simétrica y acampanada la relación entre la distancia intercuartil y el desvío estándar es la siguiente

$$D_I \cong \frac{4}{3}\, s$$

Para distribuciones muy asimétricas $s > D_I$

Ejemplo

Consideremos nuevamente los datos siguientes.

Posición	1	2	3	4	5	6	7	8	9	10	11	12	13
Datos	104	112	134	146	155	168	170	195	246	302	338	412	678

Posición del Cuartil Inferior = (13+1)/4 = 3.5 $\Rightarrow C_I = \frac{134+146}{2} = 140$

Posición del Cuartil Superior = 3.(13+1)/4 = 10.5 $\Rightarrow C_S = \frac{302+338}{2} = 320$

$$D_I = C_S - C_I = 320 - 140 = 80$$

Concluimos que el 50% central de los datos se encuentra en una distancia de 80 unidades.

Para estos datos s = 160.5. Si la distribución fuera simétrica esperaríamos que $D_I \cong 0.75\, s = 0.75 \cdot 160.5 = 120$. Sin embargo, $D_I = 80$, lo que nos indica que la distribución es asimétrica.

4.3 GRÁFICO DE CAJA (BOX-PLOT)

Concluimos este capítulo presentando un gráfico propuesto por Tukey para presentar datos numéricos, especialmente útil para comparar distribuciones de varios conjuntos de observaciones. Está basado en medidas robustas de posición y dispersión.

¿Cómo se construye un box-plot?

1. Ordenar los datos de menor a mayor
2. Calcular la mediana, el cuartil inferior, el cuartil superior y la distancia intercuartil.
3. Calcular cotas que nos permitirán decidir si un dato es outlier:
   - 2ª cota inferior $= C_I - 3\, D_I$
   - 1ª cota inferior $= C_I - 1.5\, D_I$
   - 1ª cota superior $= C_S + 1.5\, D_I$
   - 2ª cota superior $= C_S + 3\, D_I$

   Cualquier dato que caiga entre la 1ª y 2ª cota inferior o entre la 1ª y 2ª cota superior será declarado *outlier*.

   Cualquier dato que caiga por fuera de la 2ª cota inferior o la 2ª cota superior será declarado *outlier severo.*
4. Dibujar una escala que cubra el rango de variación de los datos y marcar la mediana y los cuartiles. Dibujar una caja que se extienda entre los cuartiles y marcar en ella la posición de la mediana.

5. Partiendo del cuartil inferior trazar una línea (bigote) que llegue hasta el último dato contenido "dentro" de la 1ª cota inferior.

   Partiendo del cuartil superior trazar una línea (bigote) que llegue hasta el último dato contenido "dentro" de la 1ª cota superior.

6. Marcar la posición de los outliers con un símbolo (por ejemplo, *) y de los outliers severos con otro símbolo (por ejemplo, ○).

Ejemplo

Consideremos nuevamente los datos siguientes.

Posición	1	2	3	4	5	6	7	8	9	10	11	12	13
Datos	104	112	134	146	155	168	170	195	246	302	338	412	678

De los ejemplos anteriores sabemos que:

$$C_I = 140 \qquad \tilde{X} = 170 \qquad C_S = 320 \qquad D_I = 320 - 140 = 80$$

Calculamos las cotas:

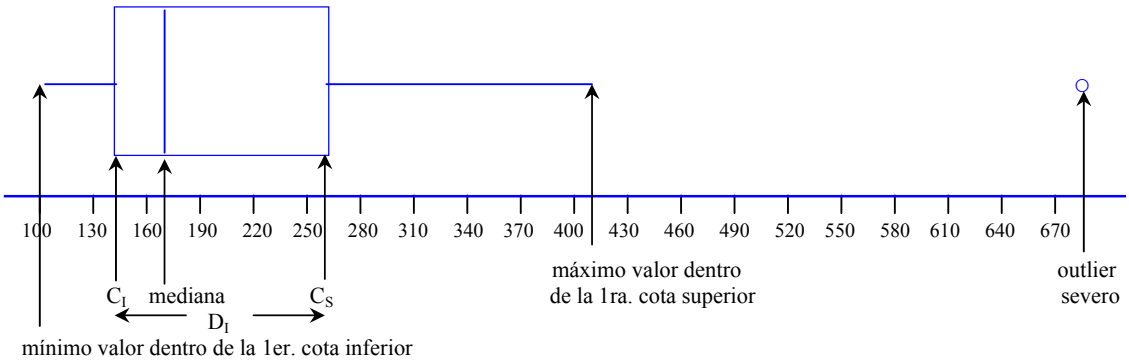
$$2^a \text{ cota inferior} = C_I - 3\,D_I = 140 - 3 \cdot 80 = -100$$

$$1^a \text{ cota inferior} = C_I - 1.5\,D_I = 140 - 1.5 \cdot 80 = 20$$

$$1^a \text{ cota superior} = C_S + 1.5\,D_I = 320 + 1.5 \cdot 80 = 440$$

$$2^a \text{ cota superior} = C_S + 3\,D_I = 320 + 3 \cdot 80 = 580$$

El gráfico de caja resultante se muestra en la figura siguiente.



¿Qué se observa?

- Un dato outlier.
- La distribución de los datos es asimétrica hacia la derecha, la mitad inferior de los datos se distribuye en un rango mucho menor que la mitad superior.

¿Qué características de la distribución de los datos se manifiestan en un box-plot?

- Muestra los cinco números resúmenes
- Muestra una medida de posición robusta ⇒ MEDIANA

- Muestra una medida de dispersión robusta ⇒ DISTANCIA INTERCUARTIL
- Permite estudiar la simetría de la distribución
- Nos da un criterio de detección de datos outliers

Los distintos paquetes estadísticos dibujan box-plots que no siempre se basan en los criterios que hemos detallado aquí, algunos cambian el modo de calcular los cuartiles, otros por ejemplo, ofrecen opciones de indicar la media y no la mediana en la caja.

Estos gráficos son muy útiles para comparar varias distribuciones. La Figura siguiente muestra los datos correspondientes a los resultados de una encuesta que se tomó en cuatro poblaciones diferentes las que se identifican de 1 a 4. La variable que se registró es el grado de satisfacción con el desempeño de los gobernantes en el último año (puntaje de 0 a 100).

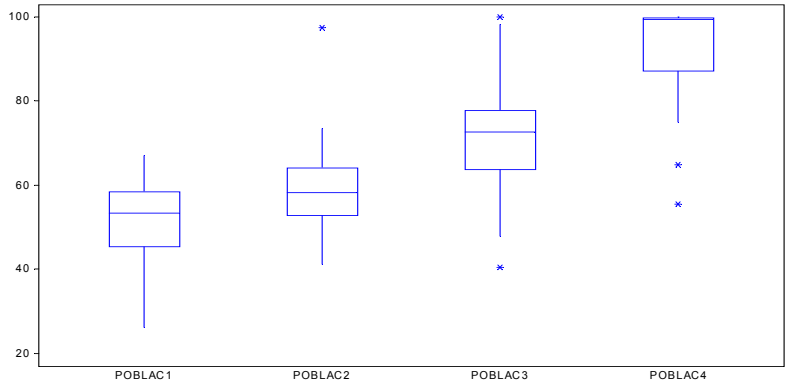
¿Qué concluimos a partir de este gráfico?

La satisfacción de los habitantes de las distintas poblaciones difiere en posición (la mediana cambia notablemente) y en dispersión (la población 3 presenta mayor dispersión que las demás).

Las distribuciónes tienen diferentes formas (Población 4 muy asimétrica, mucha gente está totalmente de acuerdo con el desempeño de sus gobernantes, mientras que en las demás la distribución es simétrica).

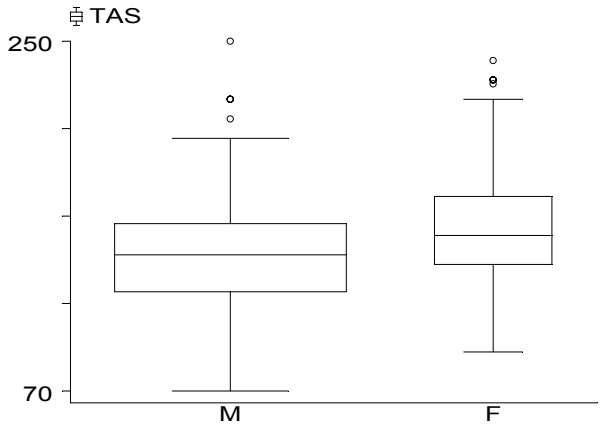
Podemos observar además que el cuartil inferior (percentil 25) del puntaje en la Población 3 es aproximadamente 63 y coincide con el cuartil superior de la Población 2, es decir, en la población 3 el 75% de los encuestados asignaron puntajes de 63 o más, en tanto que en la Población 2 sólo el 25% asignaron puntajes de 63 o más.

Del mismo modo, podemos observar que los encuestados de la Población 4 tienen un grado de satisfacción más alto que practicamente todos los encuestados en las demás poblaciones.



Box-plot con ancho de la caja proporcional al número de observaciones.

Una variación útil de este gráfico consiste en representar las cajas con ancho proporcional al número de observaciones. Como ejemplo se presenta la distribución de presión arterial sistólica en personas adultas de ambos sexos, que concurreron espontáneamente a medir su presión. Se dispone de 934 mediciones en varones y 477 mediciones en mujeres.


TAS
250
70
M
F

Cap 5. RELACIONES ENTRE VARIABLES NUMÉRICAS

Tal como ocurre en el caso univariado, el análisis de datos bivariados (dos variables medidas o registradas en el mismo individuo) comienza con el estudio del patrón o la estructura subyacente en los datos.

Generalmente cuando se estudia la relación entre dos variables, registradas sobre el mismo individuo, una de ellas se considera *variable de respuesta* (efecto o resultado) y la otra se considera la *variable independiente o explicatoria* (potencial factor que afecta la variable respuesta). En este enfoque el objetivo es analizar si *existe relación entre ambas*, y de ser posible, estudiar la naturaleza y la fuerza de la relación que las liga.

Denotando por X la variable independiente y por Y la variable dependiente, un esquema simplificado de las situaciones que podemos encontrar se propone en la tabla siguiente.

	Independiente (X)	**Dependiente (Y)**	**Ejemplo**
A)	Categórica	Categórica	X = hábito de fumar (no / <10 cig / ≥ 10) Y = cáncer de pulmón (si / no)
B)	Categórica	Numérica	X = hábito de fumar Y = nivel de colesterol sérico
C)	Numérica	Categórica	X = nivel de colesterol sérico Y = infarto de miocardio (si / no)
D)	Numérica	Numérica	X = nivel de colesterol Y = presión arterial

En cada una de estas situaciones el enfoque analítico y el modo de resumen y presentación habitual de los datos cambia. Brevemente, el modo de resumir los datos en cada situación se presenta a continuación.

A) Tablas de doble entrada y medidas de asociación (riesgo relativo, odds ratio, etc.).

B) Medidas resúmenes de nivel de colesterol para cada grupo definido por hábito de fumar o box-plots para cada grupo.

C) Un posible modo de resumir es categorizar la variable numérica y presentar la proporción de casos positivos (infarto de miocardio) en los distintos grupos definidos por nivel de colesterol.

D) Gráficos de dispersión y medidas de correlación.

En cualquier caso interesa estudiar si existe asociación entre las dos variables, pero el modo de medir asociación o efecto difiere.

En este capítulo consideraremos únicamente el problema de *representar gráficamente dos variables numéricas* y el modo de *resumir la fuerza de la asociación entre dos variables numéricas*. Finalmente consideraremos el caso en que la variable independiente es el tiempo, que merece un tratamiento especial y se conoce como análisis de seriees de tiempo.

5.1 GRÁFICO DE DISPERSIÓN (SCATTER PLOT)

Es un gráfico muy simple y útil para estudiar relaciones entre dos variables cuantitativas. Se dibuja un sistema de coordenadas cartesianas en el que se representan los valores que toman las dos variables para cada sujeto o unidad de análisis. Se acostumbra asignar la variable independiente al eje horizontal (comúnmente denominado eje X) y la variable dependiente al eje vertical (eje Y).

La nube resultante de puntos permite evaluar si existe relación entre las dos variables y la naturaleza de tal relación. Si es lineal, curvilínea, exponencial, logarítmica, cíclica, creciente, decreciente, etc. o si no hay relación aparente entre las variables.

Para interpretar un gráfico de dispersión debe mirarse el patrón general que siguen los puntos. Este patrón debería revelar la *dirección, forma y fuerza de la relación* entre las dos variables.

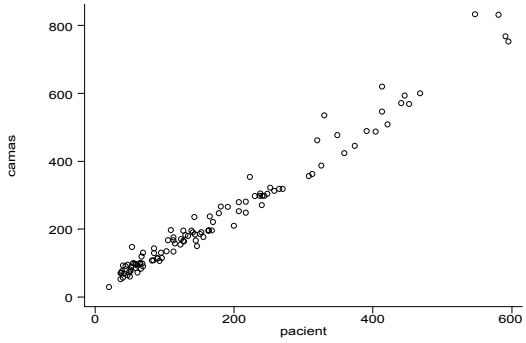
Consideraremos algunos ejemplos.

Los gráficos de la Figura 1 corresponden a datos de una muestra aleatoria de 56 hospitales participantes en el proyecto SENIC (Study on the Efficacy of Nosocomial Infection Control). El objetivo fundamental del Proyecto era determinar si los programas de vigilancia y control de infecciones habían reducido la tasa de infección hospitalaria en los Estados Unidos.

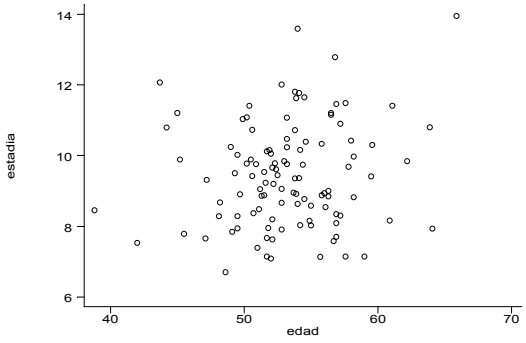
En a) hemos representado el número promedio de camas en el hospital durante el período de estudio y el número promedio de pacientes hospitalizados por día durante el período de estudio. El gráfico b) muestra la relación entre duración promedio de la estadía de todos los pacientes en el hospital (en días) y edad promedio de todos los pacientes del hospital (en años).

Figura 1. Gráficos de dispersión.

a) Número de pacientes versus número de camas en hospitales.



b) Tiempo de internación versus edad del paciente.



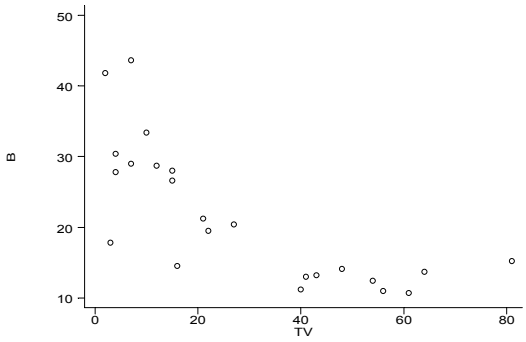
¿Qué nos dicen los gráficos de la Figura 1 acerca de la relación entre las variables?

Figura 1 a) Número de camas y número de pacientes están *fuertemente* relacionados. Cuando una variable aumenta la otra también aumenta, es decir, entre ambas variables existe una *asociación positiva*. Además podemos proponer que la relación entre ambas variables es *lineal* ya que una línea recta aproximaría bastante bien la *tendencia general* de la nube de puntos.

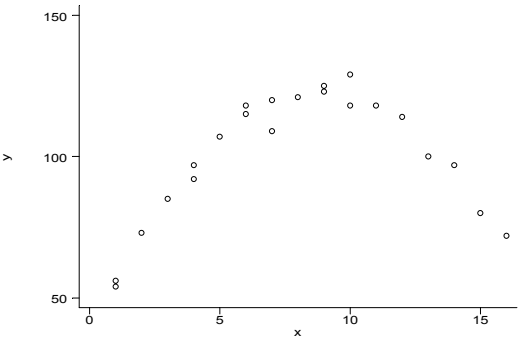
Figura 1 b) No parece haber relación entre el tiempo de internación y la edad del paciente. Si nos ubicamos en alguna edad particular, digamos 50 años, podemos encontrar pacientes cuya internación tuvo una duración de cualquier magnitud. La nube de puntos no presenta una tendencia particular.

Figura 2. Gráficos de dispersión

a) Tasa de natalidad versus número de aparatos de TV



b) Datos ficticios.



En la Figura 2 a) hemos representado la tasa de nacimiento cruda (número de nacimientos cada 1000 habitantes) y el número de televisores cada 100 habitantes para 26 naciones (desarrolladas y en vías de desarrollo). Fuente: Statistical Abstract of the United States, 1995 and Human Development Report, 1995, Oxford University Press.

En la Figura 2 b) se muestran datos ficticios de dos variables X e Y.

¿Qué nos dicen los gráficos de la Figura 2 acerca de la relación entre las variables?

Figura 2 a). La tasa de natalidad está *inversamente* relacionada con el número de televisores cada 100 habitantes. Cuando el número de televisores aumenta, la tasa de natalidad disminuye. Además, el decrecimiento *no es lineal* (una línea recta no es un buen modelo para el tipo de relación que se observa entre las dos variables). Cuando el número de televisores es bajo (cercano a cero), un aumento de 20 televisores por cada 100 habitantes produce una importante disminución de la tasa de natalidad, mientras que si el número de televisores es alto (más de 40), un aumento de la misma magnitud en el número de televisores produce una disminución despreciable en la tasa de natalidad. La relación entre las dos variables podría describirse como exponencial negativa.

Figura 2 b). X e Y están fuertemente relacionadas, podemos proponer que la relación entre ambas es *curvilínea.* No podemos hablar de dirección de la relación ya que es en parte creciente y en parte decreciente.

Al estudiar la relación entre dos variables CUANTITATIVAS. En general interesa:

- ✓ Investigar *si existe asociación* entre las dos variables.
- ✓ Cuantificar la *fuerza de la asociación*, a través de una medida de asociación denominada *coeficiente de correlación*.

- ✓ Estudiar la *forma de la relación* y en lo posible proponer un *modelo matemático* para la relación.
- ✓ *Predecir* una variable a partir de la otra usando el modelo propuesto (REGRESIÓN)

Un MODELO MATEMÁTICO es una función matemática que propone la forma de relación entre la variable dependiente (Y) y la o las variables independientes.

La función más simple para la relación entre dos variables es la FUNCIÓN LINEAL

$$Y = a + b \cdot X$$

Un MODELO DETERMINÍSTICO supone que bajo condiciones ideales, el comportamiento de la variable dependiente puede ser totalmente descripto por una función matemática de las variables independientes (o por un conjunto de ecuaciones que relacionen las variables). Es decir, en condiciones ideales el modelo permite predecir SIN ERROR el valor de la variable dependiente.

- ✓ Ejemplo: Ley de la Gravedad.

Podemos predecir exactamente la posición, en cada instante de tiempo, de un objeto que cae libremente en el vacío.

Un MODELO ESTADÍSTICO permite incorporar un *componente aleatorio* en la relación. Debido a este componente aleatorio, las predicciones obtenidas a través de modelos estadísticos tendrán asociado un *error de predicción*.

- ✓ Ejemplo: Relación de la altura con la edad en niños.

Niños de la misma edad seguramente no tendrán la misma altura. Sin embargo, a través de un modelo estadístico es posible concluir que la altura aumenta con la edad. Es más, podríamos predecir la altura de un niño de cierta edad y asociarle un error de predicción que tiene en cuenta los *errores de medición* y la *variabilidad entre individuos*.

En problemas biológicos, trabajando en "condiciones ideales" es posible evitar los errores de medición, pero no la variabilidad individual, por eso es indispensable incluir el componente aleatorio en los modelos estadísticos.

5.2 COEFICIENTE DE CORRELACIÓN

El *grado de asociación* entre dos variables numéricas puede ser resumido en un estadístico denominado COEFICIENTE DE CORRELACIÓN.

Presentaremos en primer lugar el coeficiente de correlación de Pearson, que mide el grado de asociación lineal entre dos variables y posteriormente un estadístico basado en rangos que estima la correlación sin hacer supuestos sobre el tipo de relación entre las variables.

5.2.1 COEFICIENTE DE CORRELACIÓN DE PEARSON

Supongamos que tenemos dos variables (X, Y) registradas en cada una de los *n* sujetos de una muestra. Sean (X_i, Y_i) las observaciones realizadas para cada variable en el sujeto i-ésimo. Definimos la *covarianza muestral* entre X e Y como:

$$\text{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

donde $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ e $\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$.

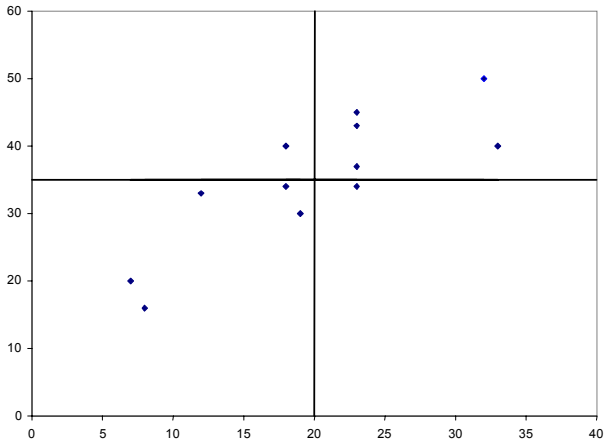
La covarianza es el "promedio" de los productos de las desviaciones de las variables respecto de las correspondientes medias.

¿Cómo se interpreta la covarianza?

En la Figura 3 se representa una nube de puntos correspondiente a los distintos pares (X, Y) observados en una muestra. Trazamos rectas paralelas a los ejes de coordenadas que pasan por $\overline{X}$ e $\overline{Y}$ y dividimos el plano en cuatro cuadrantes I, II, III y IV.

Figura 3

IV **I**



III **II**

Consideremos ul punto en el Cuadrante I.: la diferencia $(X - \overline{X}) > 0$ y la diferencia $(Y - \overline{Y}) > 0$ y lo mismo ocurre con el signo de las diferencias para cualquier punto ubicado en este cuadrante. Por lo tanto, el producto $(X - \overline{X})(Y - \overline{Y}) > 0$. Usando el mismo razonamiento para puntos ubicados en los demás cuadrantes obtenemos la siguiente tabla.

Cuadrante	$(X - \overline{X})$	$(Y - \overline{Y})$	$(X - \overline{X})(Y - \overline{Y})$
I	+	+	+
II	+	–	–
III	–	–	+
IV	–	+	–

Por lo tanto,

- Si la mayoría de los puntos se encuentran en los cuadrantes I y III la covarianza se construirá básicamente con sumandos positivos y por lo tanto será positiva. Este es el caso de los datos de la Figura 3 en la que la Cov(X, Y) = 738.
- Si la mayoría de los puntos se encuentran en los cuadrantes II y IV la mayoría de los sumandos serán negativos y la covarianza será negativa (Figura 4 a, Cov = -1098).
- Si los puntos se encuentran homogéneamente distribuidos por los cuatro cuadrantes, la covarianza será cercana a cero (Figura 4 b, Cov = -15).

Figura 4 a

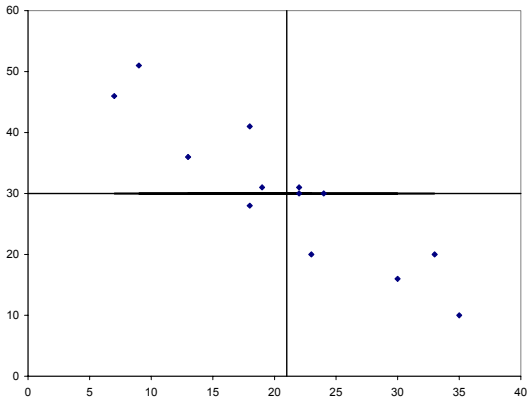
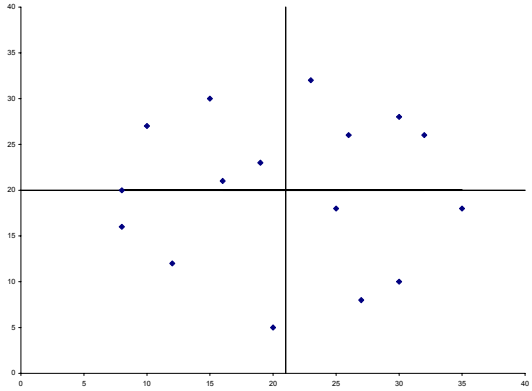


Figura 4 b



Podemos interpretar el signo de la covarianza, pero no su magnitud ya que ésta depende de las unidades de los datos. Si en el Gráfico 4 a, cambiamos las unidades de la variable X dividiendo cada valor por 1000 (si X representara peso, sería equivalente a transformar el peso de gramos a Kg) la covarianza pasa de – 1098 a – 1.098. Por lo tanto, es importante estandarizar la covarianza de modo que no dependa de las unidades de las variables.

Definición

Sean (X_i, Y_i) las observaciones realizadas en cada uno de los *n* sujetos de una muestra de tamaño *n*. Definimos el *coeficiente de correlación muestral de Pearson* entre X e Y como:

$$r = Corr(X,Y) = \frac{\text{cov}(X,Y)}{s_X \; s_Y} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{(n-1) \; s_X \; s_Y}$$

donde s_x y s_y son los desvíos estandares muestrales de las variables X e Y respectivamente.

Ejemplo

	X	Y	$(X - \overline{X})$	$(Y - \overline{Y})$	$(X - \overline{X})(Y - \overline{Y})$
	3	10	-3.86	3.14	-12.12
	6	7	-0.86	0.14	-0.12
	5	9	-1.86	2.14	-3.98
	8	6	1.14	-0.86	-0.98
	9	8	2.14	1.14	2.45
	10	7	3.14	0.14	0.45
	7	8	0.14	1.14	0.16
Media	6.86	7.86		Suma =	-14.14
DS	2.41	1.35			

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)\ s_X\ s_Y} = \frac{-14.14}{(7-1)\ 2.41\ 1.35} = -0.73$$

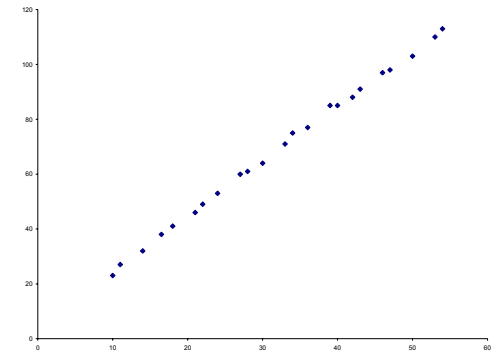
Propiedades del coeficiente de correlación de Pearson

- *r* toma valores entre –1 y 1 ($-1 \le r \le 1$),
- *r* mide la *fuerza* de la asociación LINEAL entre X e Y,
- $r = 0$ implica que no hay relación lineal entre las variables,
- $r = +1$ implica que todos los puntos caen sobre una recta de pendiente positiva (asociación positiva),
- $r = -1$ implica que todos los puntos caen sobre una recta de pendiente negativa (asociación negativa),
- mientras mayor el valor absoluto de *r* mayor la fuerza de la asociación,
- el valor de *r* no depende de las unidades de medición,
- el coeficiente de correlación trata a X e Y simétricamente, no identifica cual es la variable dependiente y cual la independiente.

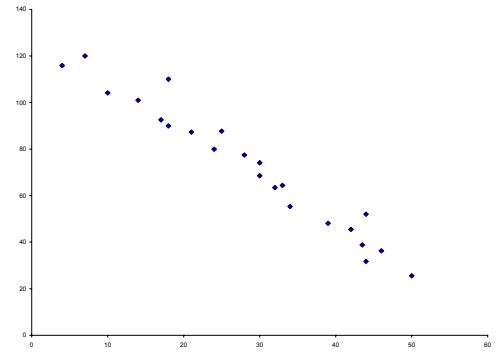
¿Qué mide exactamente el coeficiente de correlación de Pearson?

Cuán cercanos se encuentran los puntos alrededor de una LÍNEA RECTA que indique la tendencia general.

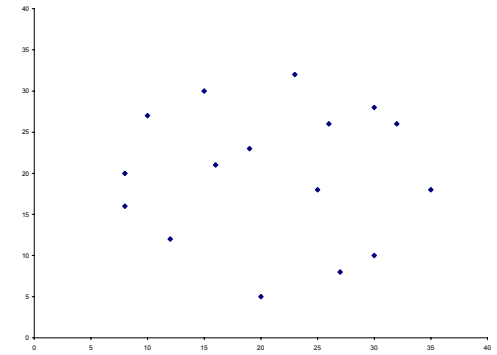
Figura 5. Ejemplos de conjuntos de datos con diferente grado de correlación.



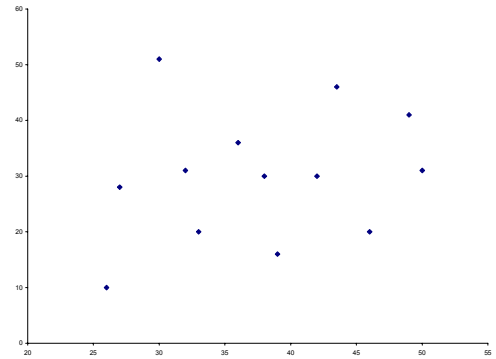
r = 0.9993



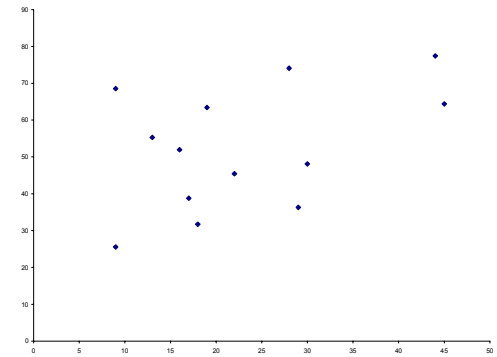
r = -0.9775



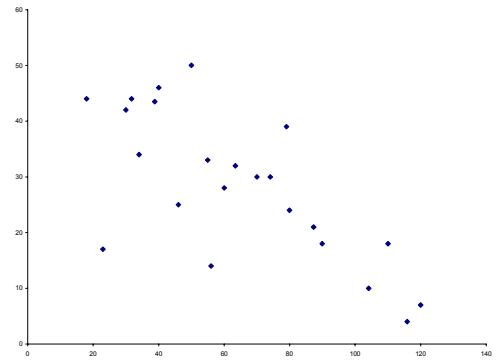
r = 0.0139



r = 0.418



r = 0.2241



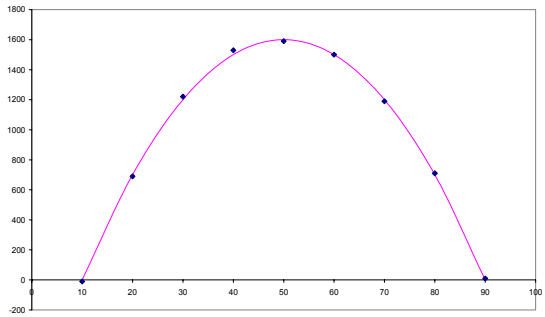
r = - 0.718

¿Qué ocurre si la relación entre las variables no es lineal?

Si el supuesto de linealidad no se cumple el valor del coeficiente de correlación puede ser engañoso. Conisderemos el gráfico de la Figura 6, en que la relación entre las dos

variables es en forma de U. En este caso el coeficiente *r* es cercano a cero, es decir, a partir de *r* concluiríamos que las variables NO están asociadas. Sin embargo, las variables están fuertemente asociadas ya que los valores de Y sigue una relación casi determinística con el valor de X, el problema es que esta relación no es lineal.

Figura 6



¿Cómo afectan los datos outliers al coeficiente de correlación?

En principio el coeficiente de correlación de Pearson puede ser calculado para cualquier conjunto de datos en el que los pares ordenados sigan una relación aproximadamente lineal. Sin embargo, una observación outlier respecto de la relación, que no se encuentra en la tendencia general de los datos, puede influir notablemente en la magnitud del coeficiente. Una observación se denomina *influyente* cuando produce un cambio importante en el coeficiente de regresión lineal o en la recta que se propondría como modelo para la relación entre las variables.

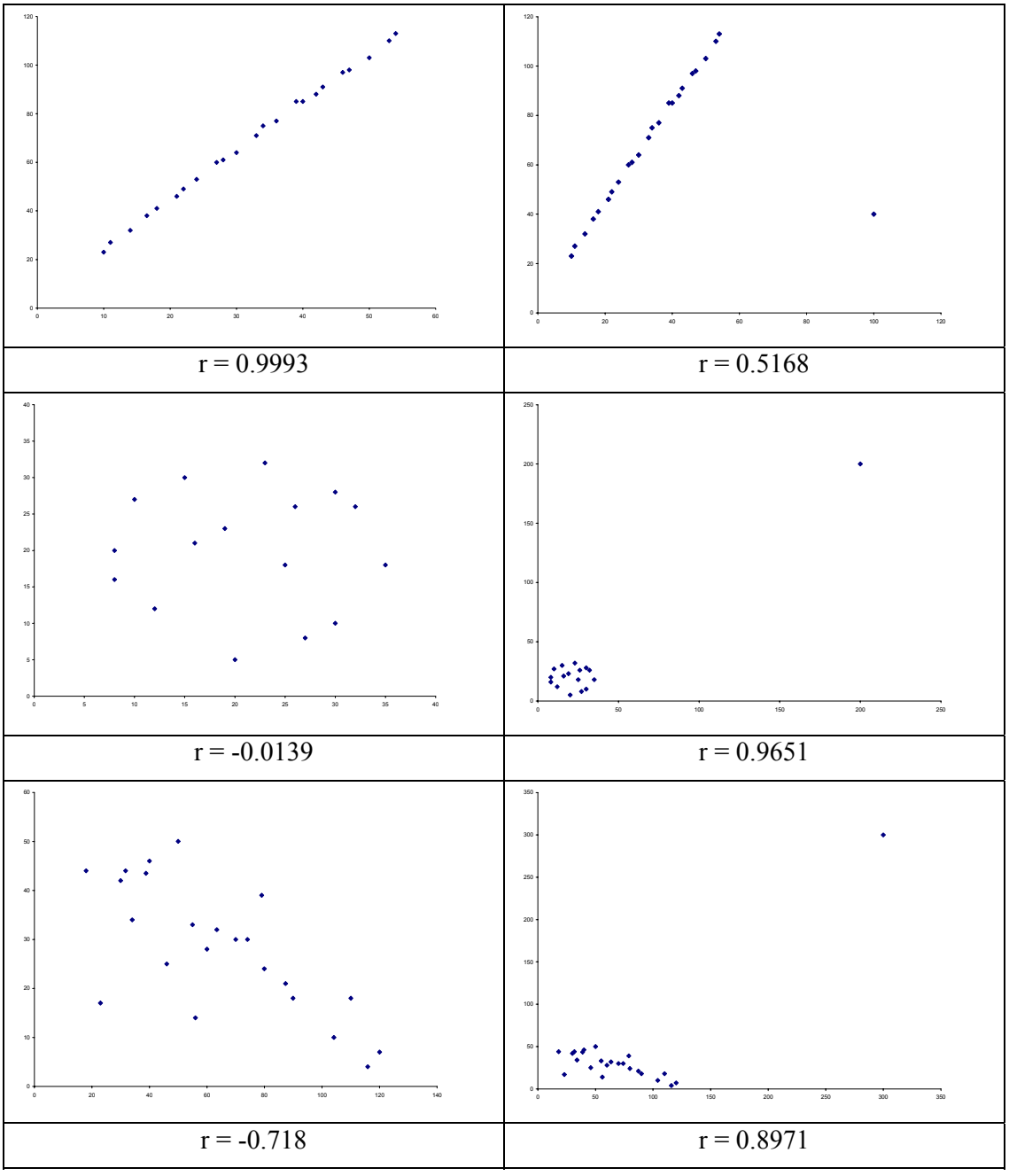
En la Figura 7 se presentan tres de los gráficos de la Figura 5 a los que se les agregó solamente 1 punto, que en cada caso logra modificar notablemente el valor del coeficiente de correlación de Pearson.

En conclusión:

- El coeficiente de correlación de Pearson es una medida *muy sensible* a la presencia de datos influyentes.
- El coeficiente de Pearson cuantifica la fuerza de la relación LINEAL entre las dos variables. Antes de calcularlo es necesario hacer un gráfico para decidir si la relación entre las variables es aproximadamente lineal y si no hay puntos influyentes. En general, el coeficiente de correlación de Pearson es una buena medida resúmen del grado de asociación entre dos variables numéricas cuando el gráfico muestra una nube de puntos elíptica.

Por último, diremos que mostrar que dos variables están asociadas, no implica que exista relación de causalidad entre ellas.

Figura 7. Efectos de datos influyentes sobre el coeficiente de correlación de Pearson.



Resumiendo, una medida de correlación entre dos variables X e Y debería satisfacer los siguientes requerimientos:

- Tomar valores entre –1 y 1.
- Si los valores más grandes de X tienden a aparecer con los valores más grandes de Y y los menores de X con los menores de Y, entonces la medida de correlación debería ser positiva y cercana a 1 cuando la tendencia sea muy fuerte. Decimos entonces que X e Y tienen *correlación positiva*.

- Si los mayores valores de X tienden a aparecer junto con los menores valores de Y y vice cersa, entonces la medida de correlación debería ser negativa, con –1 indicando que la tendencia es fuerte. Decimos entonces que X e Y están *negativamente correlacionadas*.
- Si los valores de X aparecen aleatoriamente apareados con los de Y, la medida de correlación debería ser próxima a cero. Decimos entonces, que X e Y no están correlacionados.

Existen otras medidas para resumir correlación que satisfacen los requerimientos anteriores pero que son robustas a la presencia de datos influyentes. Presentamos a continuación una propuesta alternativa para medir correlación que se construye ordenando los datos.

5.2.2 COEFICIENTE DE CORRELACIÓN DE SPEARMAN

Disponemos de *n* pares de observaciones (X_1, Y_1), ..., (X_n, Y_n). Las variables pueden ser numéricas o categóricas ordinales.

¿Cómo se calcula el coeficiente de Spearman?

1. Se ordenan los valores de cada variable por separado y se reemplaza cada observación por la posición (*rango*) que ésta ocupa en la muestra ordenada.
2. Se calcula el coeficiente de Pearson usando como datos los rangos.

Características

- Como el coeficiente de correlación de Spearman varía entre –1 y 1.
- Mide la fuerza de la correlación entre las dos variables. Valores positivos indican que la relación entre X e Y es creciente. Valores negativos indican que la relación es decreciente. Valores cercanos a cero indican que la relación no es creciente ni decreciente.
- No hace supuestos sobre la forma de la relación entre las dos variables.

Ejemplo

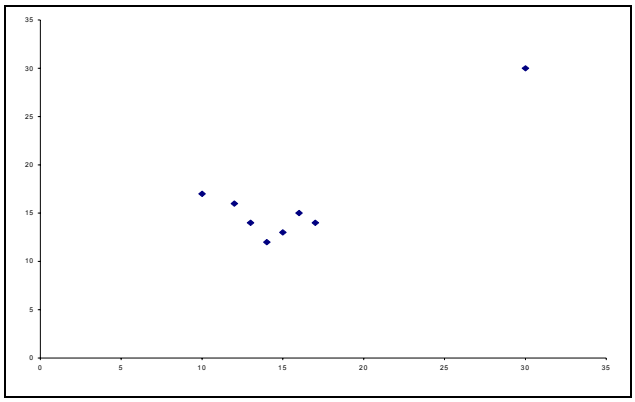
Para los datos de la tabla siguiente calculamos el coeficiente de correlación de Spearman:

$$r_S = \frac{(1-4.5)(7-4.5)+(3-4.5)(3.5-4.5)+\ldots+(8-4.5)(8-4.5)}{2.45\cdot 2.45\cdot(8-1)} = 0.0000$$

Para estos datos el coeficiente de Pearson es $r = 0.8355$. ¿Por qué tanta diferencia entre ambos? La Figura 8 muestra que la diferencia se debe a la presencia de un punto fuertemente influyente.

	X	Y	Rango (X)	Rango(Y)
	10	17	1	7
	13	14	3	3.5
	12	16	2	6
	15	13	5	2
	16	15	6	5
	17	14	7	3.5
	14	12	4	1
	30	30	8	8
Media	15.88	16.38	4.5	4.5
DS	6.13	5.73	2.45	2.43

Figura 8. Efecto de un dato influyente sobre el coeficiente de correlación de Pearson.



¿Cuándo usar coeficiente de Spearman (u otro basado en rangos)?

- Cuando las variables tienen una relación creciente o decreciente pero no necesariamente lineal.
- Cuando hay datos influyentes.
- Cuando la forma de la nube de puntos no es elipsoidal.

5.3 GRÁFICOS ENGAÑOSOS

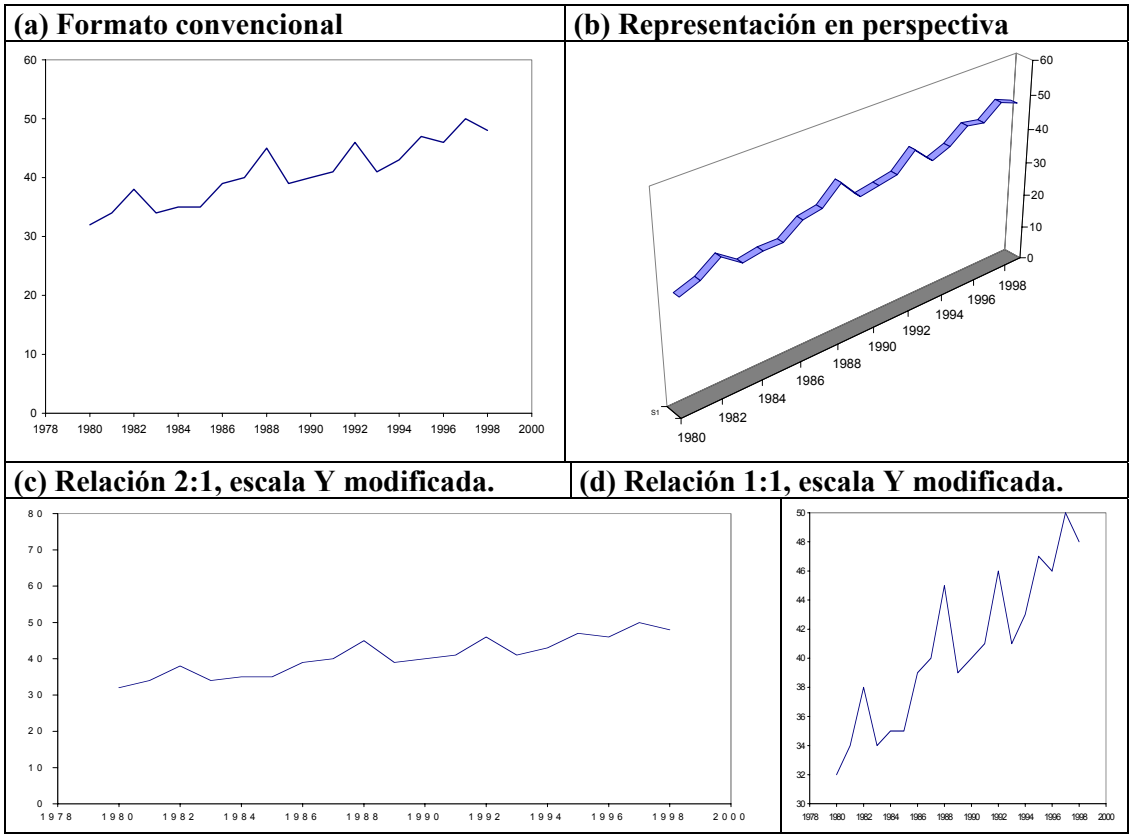
Cuando se trata de gráficos de dispersión o de series, la imagen visual puede modificarse notablemente usando uno o más de los siguientes recursos:

- cambiando la escala de uno o ambos ejes,
- eliminando el cero de la escala vertical en la representación,
- cambiando la relación de longitud entre ambos ejes.

Los gráficos XY por convención se representan respetando una relación 4:3 entre el eje horizontal y el vertical, practicamente todos los paquetes que construyen gráficos respetan esta convención. La Figura 14 muestra cuatro representaciones diferentes de los mismos

datos de una serie anual donde se pretende mostrar como estos cambios pueden afectar la interpretación de la imagen.

Figura 14. Distintos formatos para la misma serie de tiempo



La Figura 14 (a) muestra el gráfico obtenido respetando la relación 4:3 y usando la escala del eje vertical que comienza en cero. Se observa una tendencia moderadamente creciente y fluctuaciones moderadas.

En la Figura 14 (b) se realizó una "bonita" representación en perspectiva, respetando las escalas que se usaron en (a). Este gráfico puede producir una sensación de tendencia más marcada que el gráfico anterior o una impresión de que no hay tendencia, dependiendo del observador.

En (c) modificamos la relación horizontal:vertical, de 4:3 a 2:1, y aumentamos la escala del eje Y. Resultado: la tendencia y las fluctuaciones parecen poco importantes.

Finalmente en el gráfico (d) cambiamos la relación horizontal:vertical a 1:1 y modificamos la escala vertical logrando de este modo magnificar notablemente la tendencia y la importancia de las fluctuaciones.

Todos los gráficos de la Figura 14 son correctos en el sentido que se construyeron usando la misma información (no hemos falseado o modificado los datos para construirlos). Sin embargo, algunos de ellos producen impresiones engañosas amplificando o disimulando diferencias que existen.

INDICE

Capítulo 4. MEDIDAS RESÚMENES

4.1 MEDIDAS DE POSICIÓN O LOCALIZACIÓN

4.1.1 EL PROMEDIO O LA MEDIA ARITMÉTICA

Características y propiedades de la media.

4.1.2 LA MEDIANA MUESTRAL

Mediana poblacional

Propiedades de la mediana

Comparación de la media y la mediana

4.1.3 LA MEDIA α-PODADA

¿Cuál de las tres medidas de posición preferir: media, mediana o media α-podada?

4.1.4 LA MODA

4.1.5 CUARTILES Y OTROS PERCENTILES

4.2 MEDIDAS DE DISPERSIÓN O VARIABILIDAD

4.2.1 RANGO MUESTRAL

4.2.2 DESVIACIÓN ESTÁNDAR Y VARIANZA MUESTRAL

Interpretación del valor de la desviación estándar

Regla empírica

Propiedades de la desviación estándar

4.2.3 MEDIANA DE LAS DESVIACIONES ABSOLUTAS (MAD)

4.2.4 DISTANCIA O RANGO INTERCUARTIL

4.3 GRÁFICO DE CAJA (BOX-PLOT)

Box-plot con ancho de la caja proporcional al número de observaciones.

Cap 5. RELACIONES ENTRE VARIABLES NUMÉRICAS

5.1 GRÁFICO DE DISPERSIÓN (SCATTER PLOT)

5.2 COEFICIENTE DE CORRELACIÓN

5.2.1 COEFICIENTE DE CORRELACIÓN DE PEARSON

Definición

Propiedades del coeficiente de correlación de Pearson

5.2.2 COEFICIENTE DE CORRELACIÓN DE SPEARMAN

¿Cuándo usar coeficiente de Spearman (u otro basado en rangos)?

5.3. GRÁFICOS ENGAÑOSOS