Resumen de Conceptos de Estadística

1 Estadística

La estadística básicamente estudia cifras y datos y los interpreta. En este sentido, cualquier cifra o dato numérico que refleje cualquier realidad. La idea es obtener un significado correcto para comprender las situaciones del mundo real. Dependiendo del campo de aplicación y del tipo de datos que manejen, existen diversas variantes de estadística. Ejemplos de ello son:

- Estadísticas económicas: producción, precios, parados...
- Estadísticas deportivas: partidos, puntos, goles...
- Estadísticas demográficas: nacimientos, muertes, divorcios...
- Estadísticas meteorológicas: temperatura, precipitación...

1.1 Definición de Estadística

Existen diferentes y diversas definiciones de estadística. Entre las más conocidas podemos destacar las siguientes:

1. La Estadística estudia los métodos científicos para recoger, organizar, resumir y analizar datos (realmente variables), así como para sacar conclusiones válidas y tomar decisiones razonables basadas con tal análisis.

2. Rama de las matemáticas que se ocupa de reunir, organizar y analizar datos numéricos y que ayuda a resolver problemas como el diseño de experimentos y la toma de decisiones.

3. Ciencia que trata de la recopilación, organización presentación, análisis e interpretación de datos numéricos con e fin de realizar una toma de decisión más efectiva.

4. Es la ciencia que se encarga de la recopilación, representación y el uso de datos sobre una o varias características de interés para, a partir de ellos, tomar decisiones o extraer conclusiones generales.

5. Es la ciencia se ocupa en general de fenómenos observables

6. La ciencia se desarrolla observando hechos, formulando leyes que los explican y realizando experimentos para validar o rechazar dichas leyes

7. La estadística se utiliza como tecnología al servicio de las ciencias donde la variabilidad y la incertidumbre forman parte de su naturaleza

La estadística se divide en dos áreas:

- **Estadística descriptiva:** Trata de describir datos aleatorios a partir de pequeñas porciones.

- **Estadística inductiva o inferencial:** Trata de la generalización de los resultados obtenidos en las porciones y de las condiciones bajo las cuales esos resultados son válidos.

Este resumen sólo involucra los conceptos más importantes de la estadística descriptiva, que son los empleados en el curso.

1.2 Para qué sirve la estadística

Principalmente se persiguen 3 aspectos básicos para interpretar una situación del mundo real.

1. Descripción de datos
2. Conocer datos generales a partir de datos específicos
3. Relaciones entre datos

Descripción de datos

- Por ejemplo sexo y edad de 20 personas
- Hay que resumir, de muchas maneras:
  - § Porcentajes
  - § Cantidades medias
  - § Gráficos

Esto se le llama **estadística descriptiva.**

De lo específico a lo general

- ¿Necesitamos tener todos los datos?
- Censo/población frente a muestra
- Con datos sólo de una muestra podemos saber muchas cosas de la población
  - § Descripción
  - § Relaciones

Esto se llama **estadística inferencial.**

Relacionar datos

- ¿Varía una variable en relación con otra?
  Ejemplos
  - § Años de noviazgo y divorcio
  - § Edad al casarse y divorcio

1.3 Pasos de un estudio estadístico

1. Plantear hipótesis sobre una población
   Los fumadores tienen "más bajas" laborales que los no fumadores
   ¿En qué sentido? ¿Mayor número? ¿Tiempo medio?
2. Decidir qué datos recoger (diseño de experimentos)
   Qué individuos pertenecerán al estudio (muestras)?
   Fumadores y no fumadores en edad laboral
   Criterios de exclusión ¿Cómo se eligen? ¿Descartamos los que padecen enfermedades crónicas?
3. Qué datos recoger de los mismos (variables)
   Número de bajas
   Tiempo de duración de cada baja
   ¿Sexo? ¿Sector laboral? ¿Otros factores?
4. Recoger los datos (muestreo)
   ¿Estratificado? ¿Sistemáticamente?
5. Describir (resumir) los datos obtenidos
   tiempo medio de baja en fumadores y no (estadísticos)
   % de bajas por fumadores y sexo (frecuencias), gráficos, ..., etc.
6. Realizar una inferencia sobre la población
   Los fumadores están de baja al menos 10 días/año más de media que los no fumadores.
7. Cuantificar la confianza en la inferencia
   Nivel de confianza del 95%
   Significación del contraste: p=2%

1.4 Precauciones ante la estadística

No todo es tan fácil en la estadística. Hay que tener en cuenta ciertos aspectos que pueden afectar los resultados obtenidos. Entre los más destacados podemos citar a los siguientes:

1. Mitos populares
2. Problemas de calidad de los datos
3. Problema especial en Ciencias Sociales: datos de entrevista

Mitos populares

- Las estadísticas dicen lo que uno quiera que digan
- Las estadísticas pueden manipularse para que produzcan una sensación equivocada, engañosa...

realidad

- Estadística es un "resumen de datos": hay que entender el resumen
- ¿Qué hay detrás de un número?: Cuidado
- Estadística es muy variada: un instrumento para cada situación
- Estadística es necesaria, pero hay que saberla usar bien

Problemas de calidad de los datos

- Entra basura - sale basura
- Datos "de la estantería"
- ¿Son buenos?: Son los mejores
- Ya, ya, pero ¿son buenos?

realidad

- Los datos subjetivos son más vulnerables, p. e. Ciencias Sociales.

Problema especial en Ciencias Sociales: datos de entrevista

Algunos datos son de observación directa otros no.

- Algunos datos son "objetivos" (ej. edad, muertes, etc.)
- Muchos datos son respuestas a preguntas
- La gente olvida, miente, se cansa, se niega a responder, desconfía, oculta....
- Una pregunta no es un termómetro, no es fácil "medir" una respuesta.

2 Introducción a la Estadística Descriptiva

La **estadística descriptiva** analiza series de datos (por ejemplo, edad de una población, altura de los estudiantes de una escuela, temperatura en los meses de verano, etc.) y trata de extraer conclusiones sobre el comportamiento de esos datos.

Primero debemos entender qué y cómo son los datos. Según el tipo de datos podremos realizar ciertos tipos de análisis.

Una **variable** es una característica de cada elemento en un grupo. Se le llama "variable" precisamente porque esa característica "varía" de un elemento a otro. Cada elemento tiene un **valor** para cada variable. Ejemplos de ello son:

- Variable "sexo"; Valores "hombre" y "mujer"
- Variable "edad en su último cumpleaños"; Valores: 0, 1, 2, 3
- Variable "ingresos anuales"; Valores: cualquier número entre 0 y cientos de miles de pesos

Así, el conjunto de valores que puede tomar una variable se llama la **escala** de esa variable

El proceso de definir y medir las variables es crucial. Si hacemos una definición incorrecta o medimos mal, todo lo que venga detrás, toda la estadística que podamos hacer estará mal. Algunas variables no hace falta definirlas ni hay dificultades para medirlas (ejemplo "sexo"). Otras variables aparentemente "obvias" no lo son tanto, ejemplo "estado civil". El definir las variables podría considerarse un "arte" muy complejo; que puede basarse en prueba y error, para definir y medir variables que captan características como "estatus social", "nivel educativo", "ideología política", "religiosidad"...

2.1 Tipos de variables

Según el tipo de valores que toman las variables, distinguimos diferentes tipos de variables. El tipo de variable es importante ya que afecta a lo que podemos hacer con ella, al tipo de análisis que podemos hacer. Los métodos estadísticos que usamos dependen del tipo de variable.

Las **variables** pueden ser de dos tipos:

Variables cualitativas o atributos: no se pueden medir numéricamente.

- Los valores son "categorías"
- Las categorías son valores diferentes por una cualidad, no por una cantidad
- Ningún "valor" se puede decir que sea mayor o menor que otro
- La escala de valores es nominal
- Ejemplos: partido político al que votó; región en que vive; sexo; estado civil; marca de coche que conduce, nacionalidad, color de la piel, sexo...

Variables cuantitativas: tienen valor numérico.

- Los valores de la variable son "números", cada valor posible es menor o mayor que otro valor
- El conjunto de valores forman una escala de intervalo (distancia entre valores).
- Podemos calcular la distancia o intervalo entre cualquier par de valores de la variable
- NOTA: hay "números" que son "etiquetas"; por ejemplo: el código postal; el número de teléfono; el código de una asignatura.
- Ejemplos: edad, ingresos, nota en un examen, número de años de educación, kilómetros de distancia entre trabajo y residencia, precio de un producto, ingresos anuales...

Los métodos para variables cualitativas no se pueden aplicar a variables cuantitativas. Por ejemplo: valor medio de estado civil; o de partido político. Al revés sí. una variable cuantitativa la podemos "transformar" en cualitativa. Por ejemplo edad: niños, jóvenes, adultos, ancianos. Normalmente el método estadístico es el que aprovecha al máximo las características de la variable. Con números podemos calcular valores medios, pero con "categorías" no.

Así mismo las **variables** también se pueden clasificar en:

Variables unidimensionales: sólo recogen información sobre una característica (por ejemplo: edad de los alumnos de una clase).

Variables bidimensionales: recogen información sobre dos características de la población (por ejemplo: edad y altura de los alumnos de una clase).

Variables pluridimensionales: recogen información sobre tres o más características (por ejemplo: edad, altura y peso de los alumnos de una clase).

Por su parte, las **variables cuantitativas** se pueden clasificar en discretas y continuas según el número de valores que tengan en la escala:

Discretas: sólo pueden tomar valores enteros (1, 2, 8, -4, etc.).

- El número de valores es finito (números enteros), tienen un inicio y un fin.
- Las que son el resultado de contar, valores son números enteros
- Ejemplo: personas en el hogar, número de hermanos (puede ser 1, 2, 3...., etc., pero nunca podrá ser 3,45).

Continuas: pueden tomar cualquier valor real dentro de un intervalo.

- El número de valores en la escala es infinito (números con decimales), no hay inicio ni final.
- Son resultado de medir.
- Ejemplos: altura, peso, tamaño del piso, edad, velocidad...etc.

Lo anterior es la definición "**teórica**". En la **práctica**, la diferencia está difuminada y la realidad es:

- Las variables son resultado de "medir" pero redondeamos y convertimos en número finito de valores enteros (ej. edad)
- Las variables teóricamente "continuas" las convertimos en discretas (ejemplo: escala ideológica de izquierda a derecha, valores 1 a 7)
- Puede que las variables "discretas" tengan muchísimos valores diferentes: ingresos, población de un municipio, etc.
- Puede que las variables "discretas" tengan pocos valores distintos (ej. escala ideológica)
- Puede que las variables "continuas" tengan muchos valores distintos (ej. ingresos)

3 Conceptos básicos

Cuando se estudia el comportamiento de una variable hay que distinguir los siguientes conceptos:

Individuo: cualquier elemento que porte información sobre el fenómeno que se estudia. Así, si estudiamos la altura de los niños de una clase, cada alumno es un individuo; si estudiamos el precio de la vivienda, cada vivienda es un individuo.

Población: conjunto de todos los individuos (personas, objetos, animales, etc.) que porten información sobre el fenómeno que se estudia. Por ejemplo, si estudiamos el precio de la vivienda en una ciudad, la población será el total de las viviendas de dicha ciudad.

Muestra: subconjunto que seleccionamos de la población. Así, si se estudia el precio de la vivienda de una ciudad, lo normal será no recoger información sobre todas las viviendas de la ciudad (sería una labor muy compleja), sino que se suele seleccionar un subgrupo (muestra) que se entienda que es suficientemente representativo.

Tamaño muestral: Es le número de elementos u observaciones considerados. Se denota por n ó N.

Dato: Cada uno de los individuos, cosas, entes abstractos que integran una población o universo determinado. Dicho de otra forma, cada valor observado de la variable.

4 Representación de datos

4.1 Distribución de frecuencias

La **distribución de frecuencia** es la representación estructurada, en forma de tabla, de toda la información que se ha recogido sobre la variable que se estudia. La tabla resume los datos de una variable. Es una manera "sencilla" de agrupar mucha información en unos pocos datos comprensibles. Su contenido básico son los valores que toma la variable, y qué proporción de los sujetos tiene cada valor. Puede ser tanto para muestras como para poblaciones. Lo hacemos de manera diferente para variables discretas y continuas.

En general, se compone de la siguiente manera:

Variable	Frecuencias absolutas		Frecuencias relativas	
(Valor)	Simple	Acumulada	Simple	Acumulada
X1	n1	n1	f1 = n1 / n	f1
X2	n2	n1 + n2	f2 = n2 / n	f1 + f2
...	...	...	...	...
Xn-1	nn-1	n1 + n2 +..+ nn-1	fn-1 = nn-1 / n	f1 + f2 +..+fn-1
Xn	nn	Σn	fn = nn / n	Σf
Siendo **X** los distintos valores que puede tomar la variable.				
Siendo **n** el número de veces que se repite cada valor.				
Siendo **f** el porcentaje que la repetición de cada valor supone sobre el total				

Para ello definimos los conceptos de frecuencia:

Frecuencia absoluta: Es el número de repeticiones que presenta una observación. Se representa por $\mathbf{n_i}$.

Las frecuencias absolutas suman el total de elementos.

Frecuencia relativa: Es la frecuencia absoluta dividida por el número total de datos, se suele expresar en tanto por uno, siendo su valor **-iésimo**

$$f_i = \frac{n_i}{n}$$

La suma de todas las frecuencias relativas, siempre debe ser igual a la unidad.

Frecuencia absoluta acumulada: es la suma de los distintos valores de la frecuencia absoluta tomando como referencia un individuo dado. La última frecuencia absoluta acumulada es igual al n∫ de casos:

$N_1 = n_1$
$N_2 = n_1 + n_2$
$N_n = n_1 + n_2 + \ldots\ldots + n_{n-1} + n_n = n$

Frecuencia relativa acumulada, es el resultado de dividir cada frecuencia absoluta acumulada por el número total de datos, se la suele representar con la notación: F_i

De igual forma, también se puede definir a partir de la frecuencia relativa, como suma de los distintos valores de la frecuencia relativa, tomando como referencia un individuo dado. La última frecuencia relativa acumulada es igual a la unidad.
Así tenemos en la tabla tenemos reflejada la siguiente información:

- Tabla con valores de la variable, frecuencias absolutas y frecuencias relativas
- Una variable X con un número N de observaciones (sujetos).
- La variable tiene una serie de valores diferentes
- La frecuencia absoluta de cada valor.
- La frecuencia relativa de cada valor.
- Las frecuencias absolutas suman el total de elementos.
- Las frecuencias relativas suman 1 (1 = 100%)
- Frecuencias relativas acumuladas permiten calcular:
    - Frecuencia relativa valores menores que x
    - Frecuencia relativa valores mayores que x ($1-F_i$)
    - Frecuencia relativa entre dos valores (F_i-F_j)

Veamos un ejemplo:

Medimos la altura de los niños de una clase y obtenemos los siguientes resultados (cm):

Alumno	Estatura	Alumno	Estatura	Alumno	Estatura
Alumno 1	1.25	Alumno 11	1.23	Alumno 21	1.21
Alumno 2	1.28	Alumno 12	1.26	Alumno 22	1.29
Alumno 3	1.27	Alumno 13	1.30	Alumno 23	1.26
Alumno 4	1.21	Alumno 14	1.21	Alumno 24	1.22
Alumno 5	1.22	Alumno 15	1.28	Alumno 25	1.28
Alumno 6	1.29	Alumno 16	1.30	Alumno 26	1.27
Alumno 7	1.30	Alumno 17	1.22	Alumno 27	1.26
Alumno 8	1.24	Alumno 18	1.25	Alumno 28	1.23
Alumno 9	1.27	Alumno 19	1.20	Alumno 29	1.22
Alumno 10	1.29	Alumno 20	1.28	Alumno 30	1.21

Si presentamos esta información estructurada obtendríamos la siguiente tabla de frecuencia:

Variable	Frecuencias absolutas		Frecuencias relativas	
(Valor)	Simple	Acumulada	Simple	Acumulada
1.20	1	1	3,3%	3,3%
1.21	4	5	13,3%	16,6%
1.22	4	9	13,3%	30,0%
1.23	2	11	6,6%	36,6%
1.24	1	12	3,3%	40,0%
1.25	2	14	6,6%	46,6%

1.26	3	17	10,0%	56,6%
1.27	3	20	10,0%	66,6%
1.28	4	24	13,3%	80,0%
1.29	3	27	10,0%	90,0%
1.30	3	30	10,0%	100,0%

Si los valores que toma la variable son muy diversos y cada uno de ellos se repite muy pocas veces, entonces conviene agruparlos por intervalos, ya que de otra manera obtendríamos una tabla de frecuencia muy extensa que aportaría muy poco valor a efectos de síntesis. (tal como se verá en la siguiente lección).

4.1.1 Distribuciones de frecuencia agrupada

Supongamos que medimos la estatura de los habitantes de una vivienda y obtenemos los siguientes resultados (cm):

Habitante	Estatura	Habitante	Estatura	Habitante	Estatura
Habitante 1	1.15	Habitante 11	1.53	Habitante 21	1.21
Habitante 2	1.48	Habitante 12	1.16	Habitante 22	1.59
Habitante 3	1.57	Habitante 13	1.60	Habitante 23	1.86
Habitante 4	1.71	Habitante 14	1.81	Habitante 24	1.52
Habitante 5	1.92	Habitante 15	1.98	Habitante 25	1.48
Habitante 6	1.39	Habitante 16	1.20	Habitante 26	1.37
Habitante 7	1.40	Habitante 17	1.42	Habitante 27	1.16
Habitante 8	1.64	Habitante 18	1.45	Habitante 28	1.73
Habitante 9	1.77	Habitante 19	1.20	Habitante 29	1.62
Habitante 10	1.49	Habitante 20	1.98	Habitante 30	1.01

Si presentáramos esta información en una tabla de frecuencia obtendríamos una tabla de 30 líneas (una para cada valor), cada uno de ellos con una frecuencia absoluta de 1 y con una frecuencia relativa del 3,3%. Esta tabla nos aportaría escasa información

En lugar de ello, preferimos agrupar los datos por intervalos, con lo que la información queda más resumida (se pierde, por tanto, algo de información), pero es más manejable e informativa:

Estatura	Frecuencias absolutas		Frecuencias relativas	
Cm	Simple	Acumulada	Simple	Acumulada
1.01 - 1.10	1	1	3,3%	3,3%
1.11 - 1.20	3	4	10,0%	13,3%
1.21 - 1.30	3	7	10,0%	23,3%
1.31 - 1.40	2	9	6,6%	30,0%
1.41 - 1.50	6	15	20,0%	50,0%
1.51 - 1.60	4	19	13,3%	63,3%
1.61 - 1.70	3	22	10,0%	73,3%
1.71 - 1.80	3	25	10,0%	83,3%
1.81 - 1.90	2	27	6,6%	90,0%
1.91 - 2,00	3	30	10,0%	100,0%

El número de tramos en los que se agrupa la información es una decisión que debe tomar el analista: la regla es que mientras más tramos se utilicen menos información se pierde, pero puede que menos representativa e informativa sea la tabla.

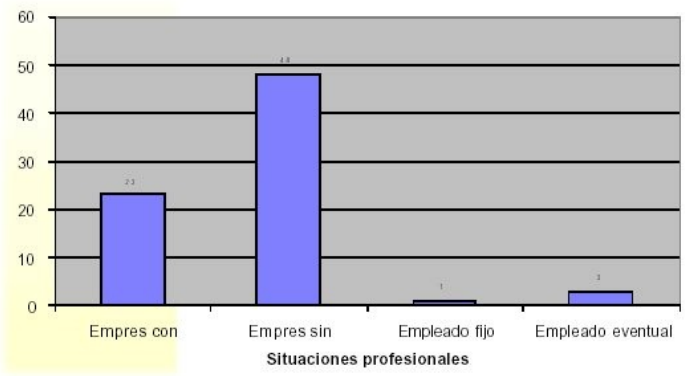
4.2 Representaciones gráficas

Son otra forma de resumir información de una variable mediante imágenes gráficas. Existen diferentes tipos, algunas son dependen del tipo de variable que se intenta reflejar (discretas o continuas). También presentan alguna diferencia entre las variables cualitativas o cuantitativas (orden de los valores). La idea primordial es que forzosamente **deben transmitir información.**

4.2.1 Diagrama de barras

Debe cumplir las siguientes características:

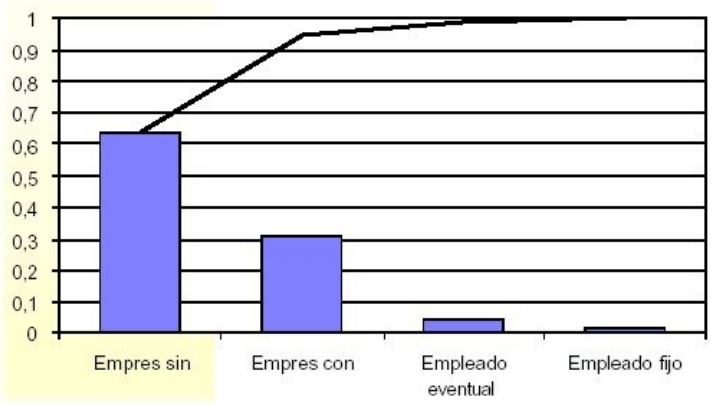
- Sólo variables discretas
- Cada valor de la variable: una barra
- Altura de la barra: frecuencia del valor
- Eje de ordenadas (vertical -Y-): pueden ser frecuencias absolutas o relativas
- Orden de los valores de izquierda a derecha:
  - Variables cualitativas: puede ser cualquiera
  - Variables de escala ordinal, o de intervalo: normalmente hay un "orden natural" que se sigue en el gráfico (como en distribuciones de frecuencias)
- Variables con escala ordinal (las que llevan un orden) o de intervalo (rango) admiten diagrama de barras de frecuencias acumuladas.



4.2.2 Diagrama de Pareto

Debe cumplir las siguiente características:

- También para variables discretas
- Básicamente es un diagrama de barras pero con algunas peculiaridades
- Orden de los valores: de más frecuente a menos frecuente
- Además de barras contiene una línea que representa las frecuencias acumuladas



4.2.3 Pictograma

También se le conoce como diagrama tarta o diagrama pastel. Debe cumplir las siguientes características.

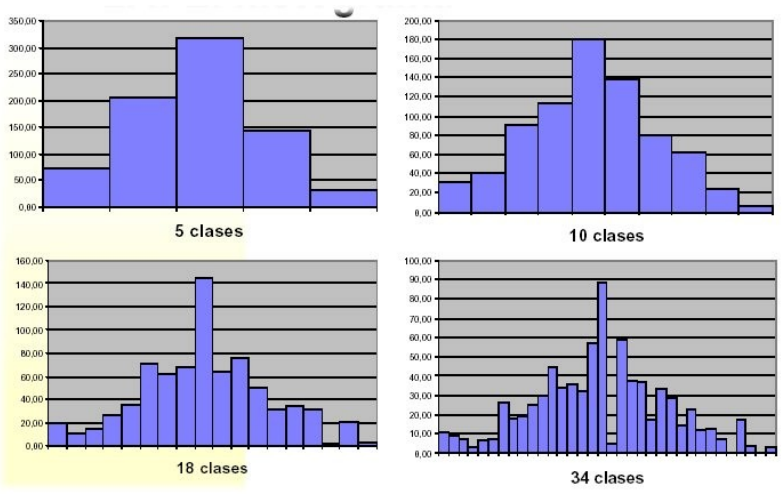
- Se emplea con variables discretas
- Tiene forma circular
- Las secciones del circulo representan los valores de la variable
- Cada sección tiene un color diferente a otras
- Las etiquetas pueden estar dentro o fuera de las secciones

4.2.4 Histograma

Debe cumplir las siguientes características:

- Sólo para variables continuas
- Frecuencias representadas por áreas
- Modelo original:
- Clases de diferente tamaño.
- El eje horizontal contiene rectángulos con base de diferente longitud
- Eje vertical no tiene sentido
- Lo más usual: clases del mismo tamaño --> la altura de los rectángulos proporcional a frecuencia (como en diagrama de barras)
- Similar a diagrama de barras, excepto:
    - Las barras son contiguas
    - Los rótulos no en valores sino en líneas de división
- Como en diagrama de barras, forma no cambia por usar frecuencias absolutas o relativas
- La forma del histograma SÕ cambia según el número de las clases

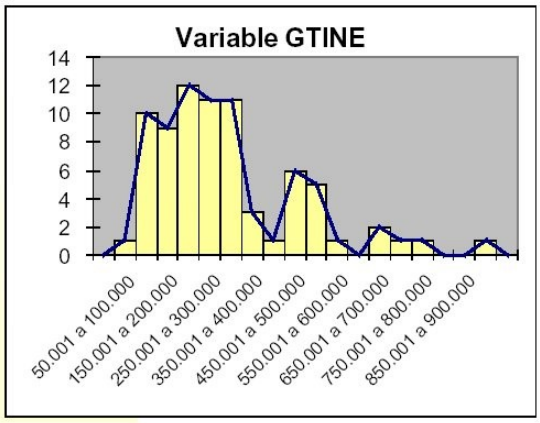
Por tanto a la hora de hacer un histograma es MUY IMPORTANTE la elección de las clases. Una regla básica es la raíz cuadrada de N. Otra regla es empezar con pocas clases y a partir de resultados ir aumentando. Para una población grande el número de clases podrían ser muchas, y muy estrechas. El histograma tiende a formar una curva. Igual que con diagrama de barras se pueden crear datos de frecuencias acumuladas, y usar esos datos para generar un gráfico.

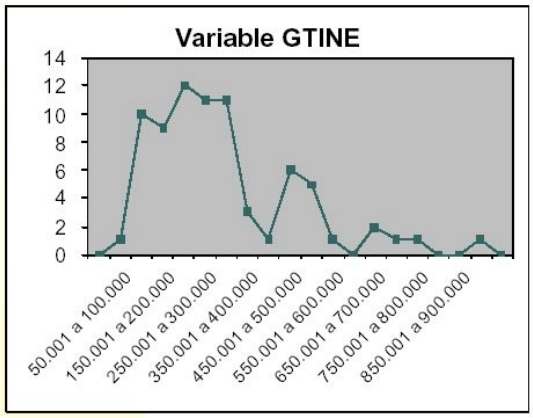


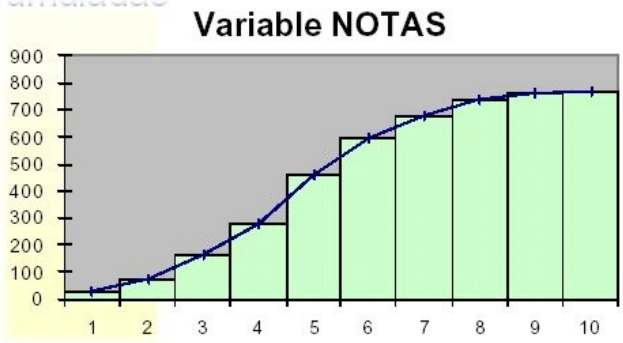
4.2.5 Polígono de frecuencias

Debe cumplir las siguientes características:

- Sólo para variables continuas
- Básicamente equivalente a histograma
- Líneas que unen puntos medios de bases superiores de rectángulos
- También puede hacerse con frecuencias acumuladas







4.2.6 Interpretación de gráficos

- Histogramas y diagramas de barras son los más usados
- Nos dicen mucho sobre la distribución de la variable
- Datos dispersos en muchos valores, o concentrados en pocos valores
- Cuál es el valor más frecuente
- Hay o no valores muy alejados del valor más frecuente
- Distribución es más o menos "simétrica": igual número de casos con valores mayores y menores que el valor más frecuente...

5 Descripciones numéricas

Hasta ahora sólo se ha visto la descripción "visual" con tablas y gráficos. Otra forma muy eficaz de resumir información es emplear valores numéricos sobre:

- la ubicación o centro de los datos (medidas de tendencia o posición) o
- la concentración de los datos en torno al centro (medidas de dispersión).
- Otros rasgos de la distribución

Sólo es posible para variables cuantitativas o para variables de escala ordinal cuando son tratadas como cuantitativas (cuidado, los números son arbitrarios)

5.1 Medidas de tendencia

Las medidas de tendencia nos facilitan información sobre la serie de datos que estamos analizando. Estas medidas permiten conocer diversas características de esta serie de datos. Las **medidas de tendencia** son de dos tipos:

a) Medidas de tendencia central: informan sobre los valores medios de la serie de datos.

b) Medidas de tendencia no centrales: informan de como se distribuye el resto de los valores de la serie.

5.1.1 Medidas de tendencia central

Las principales medidas de tendencia central son las siguientes:

1.- Media: es el valor medio ponderado de la serie de datos. Describe el aspecto más elemental de un conjunto de datos: su "centro". Es la media o promedio de un conjunto de valores. Por ejemplo: la calificación media en un examen, ingreso medio por familia, número de hijos medio por pareja, etc.

Tenemos una variable, que llamamos **X**. Llamamos "**n**" al número de elementos u observaciones de la variable. Los valores que toman las observaciones los llamamos:

$$x_1,\ x_2,\ \ldots\ldots,\ x_{n-1},\ x_n$$

Se pueden calcular diversos tipos de media, siendo las más utilizadas:

a) Media aritmética: se calcula multiplicando cada valor por el número de veces que se repite. La suma de todos estos productos se divide por el total de datos de la muestra:

$$Xm = \frac{(X1 * n1) + (X2 * n2) + (X3 * n3) + ...+ (Xn-1 * nn-1) + (Xn * nn)}{---\\ n}$$

b) Media geométrica: se eleva cada valor al número de veces que se ha repetido. Se multiplican todo estos resultados y al producto final se le calcula la raíz "n" (siendo "n" el total de datos de la muestra).

$$X = (X_1^{n1} * X_2^{n2} * X_3^{n3} * \ldots * X_n^{nn})^{(1/n)}$$

Según el tipo de datos que se analice será más apropiado utilizar la media aritmética o la media geométrica.

La media geométrica se suele utilizar en series de datos como tipos de interés anuales, inflación, etc., donde el valor de cada año tiene un efecto multiplicativo sobre el de los años anteriores. En todo caso, la media aritmética es la medida de posición central más utilizada.

Lo más positivo de la media es que en su cálculo se utilizan todos los valores de la serie, por lo que no se pierde ninguna información. Sin embargo, presenta el problema de que su valor (tanto en el caso de la media aritmética como geométrica) se puede ver muy influenciado por valores extremos, que se aparten en exceso del resto de la serie. Estos valores anómalos (outliers) podrían condicionar en gran medida el valor de la media, perdiendo ésta representatividad.

El valor de la media puede verse muy afectado por unas pocas observaciones cuyo valor sea muy diferente de los demás. Por ejemplo 7 sueldos en empresa: 10,200, 10,400, 10,700, 11,200, 11,300, 11,500 y 200.000. Aquí el sueldo medio es 37.900€. Pero un solo valor atípico (fuera de lo normal) "arrastra" la media hacia arriba. Así la media de los seis otros valores es 10.883€. Por lo tanto el valor de la media puede no ser representativo del conjunto de los valores, especialmente en poblaciones o muestras pequeñas, cuando una es muy diferente de las otras

2.- Mediana: es el valor de la serie de datos que se sitúa justamente en el centro de la muestra (un 50% de valores son inferiores y otro 50% son superiores).

No presentan el problema de estar influenciado por los valores extremos, pero en cambio no utiliza en su cálculo toda la información de la serie de datos (no pondera cada valor por el número de veces que se ha repetido).

3.- Moda: es el valor que más se repite en la muestra.

Ejemplo: vamos a utilizar la tabla de distribución de frecuencias con los datos de la estatura de los alumnos que vimos anteriormente.

Variable	Frecuencias absolutas		Frecuencias relativas	
(Valor)	Simple	Acumulada	Simple	Acumulada
1.20	1	1	3,3%	3,3%
1.21	4	5	13,3%	16,6%
1.22	4	9	13,3%	30,0%
1.23	2	11	6,6%	36,6%
1.24	1	12	3,3%	40,0%
1.25	2	14	6,6%	46,6%
1.26	3	17	10,0%	56,6%
1.27	3	20	10,0%	66,6%
1.28	4	24	13,3%	80,0%
1.29	3	27	10,0%	90,0%
1.30	3	30	10,0%	100,0%

Vamos a calcular los valores de las distintas posiciones centrales:

1.- Media aritmética:

$$Xm = \frac{(1.20*1) + (1.21*4) + (1.22 * 4) + (1.23 * 2) + \ldots\ldots\ldots + (1.29 * 3) + (1.30 * 3)}{30}$$

Luego:

$$Xm = 1.253$$

Por lo tanto, la estatura media de este grupo de alumnos es de 1.253 cm.

2.- Media geométrica:

$$X = ((1.20^{ 1}) * (1.21^4) * (1.22^{ 4}) * .....* (1.29^3)* (1.30^3))^{(1/30)}$$

Luego:

$$Xm = 1.253$$

En este ejemplo la media aritmética y la media geométrica coinciden, pero no tiene siempre por qué ser así.

3.- Mediana:

La mediana de esta muestra es 1.26 cm, ya que por debajo está el 50% de los valores y por arriba el otro 50%. Esto se puede ver al analizar la columna de frecuencias relativas acumuladas.

En este ejemplo, como el valor 1.26 se repite en 3 ocasiones, la media se situaría exactamente entre el primer y el segundo valor de este grupo, ya que entre estos dos valores se encuentra la división entre el 50% inferior y el 50% superior.

4.- Moda:

Hay 3 valores que se repiten en 4 ocasiones: el 1.21. el 1.22 y el 1.28, por lo tanto esta seria cuenta con 3 modas.

5.1.2 Medidas de tendencia no central

Las medidas de tendencia no centrales permiten conocer otros puntos característicos de la distribución que no son los valores centrales. Entre otros indicadores, se suelen utilizar una serie de valores que dividen la muestra en tramos iguales:

Cuartiles: son 3 valores que distribuyen la serie de datos, ordenada de forma creciente o decreciente, en cuatro tramos iguales, en los que cada uno de ellos concentra el 25% de los resultados.

Deciles: son 9 valores que distribuyen la serie de datos, ordenada de forma creciente o decreciente, en diez tramos iguales, en los que cada uno de ellos concentra el 10% de los resultados.

Percentiles: son 99 valores que distribuyen la serie de datos, ordenada de forma creciente o decreciente, en cien tramos iguales, en los que cada uno de ellos concentra el 1% de los resultados.

Ejemplo: Vamos a calcular los cuartiles de la serie de datos referidos a la estatura de un grupo de alumnos (lección 2ª). Los deciles y centiles se calculan de igual manera, aunque haría falta distribuciones con mayor número de datos.

Variable	Frecuencias absolutas		Frecuencias relativas	
(Valor)	Simple	Acumulada	Simple	Acumulada
1.20	1	1	3,3%	3,3%
1.21	4	5	13,3%	16,6%
1.22	4	9	13,3%	30,0%
1.23	2	11	6,6%	36,6%
1.24	1	12	3,3%	40,0%
1.25	2	14	6,6%	46,6%
1.26	3	17	10,0%	56,6%
1.27	3	20	10,0%	66,6%
1.28	4	24	13,3%	80,0%
1.29	3	27	10,0%	90,0%

1.30	3	30	10,0%	100,0%

1º cuartil: es el valor 1.22 cm, ya que por debajo de él se sitúa el 25% de la frecuencia (tal como se puede ver en la columna de la frecuencia relativa acumulada).

2º cuartil: es el valor 1.26 cm, ya que entre este valor y el 1ʃ cuartil se sitúa otro 25% de la frecuencia.

3º cuartil: es el valor 1.28 cm, ya que entre este valor y el 2ʃ cuartil se sitúa otro 25% de la frecuencia. Además, por encima de él queda el restante 25% de la frecuencia.

Atención: cuando un cuartil recae en un valor que se ha repetido más de una vez (como ocurre en el ejemplo en los tres cuartiles) la medida de posición no central sería realmente una de las repeticiones.

6 Medidas de dispersión

Las medidas de dispersión estudian la distribución de los valores de la serie, analizando si estos se encuentran más o menos concentrados, o más o menos dispersos.

Existen diversas **medidas de dispersión**, entre las más utilizadas podemos destacar las siguientes:

1.- Rango: mide la amplitud de los valores de la muestra y se calcula por diferencia entre el valor más elevado y el valor más bajo. Es muy sensible a los valores extremos. Por ejemplo de la serie 2,**1**,4,3,**8**,4. Los valores extremos son 1 y 8, por lo tanto el rango es 8-1 = 7.

2.- Varianza: Es la media aritmética de los cuadrados de las desviaciones de cada valor con respecto a su media. Mide la distancia existente entre los valores de la serie y la media. Se calcula como sumatorio de las diferencias al cuadrado entre cada valor y la media, multiplicadas por el número de veces que se ha repetido cada valor. El sumatorio obtenido se divide por el tamaño de la muestra, como se muestra en la siguiente formula.

$$S_x^2 = \frac{\sum (x_i - x_m)^2 * n_i}{n}$$

La varianza siempre será mayor que cero. Mientras más se aproxima a cero, más concentrados están los valores de la serie alrededor de la media. Por el contrario, mientras mayor sea la varianza, más dispersos están.

3.- Desviación típica: A cada medida de centralización podemos asociarle una medida de la variabilidad de los datos respecto a ella. A la media se le asocia la desviación típica y se calcula como raíz cuadrada de la varianza (para datos sin agrupar).

$$\sigma = (S_x^2)^{(1/2)} \qquad \sigma = \sqrt{\frac{\sum (x_i - x_m)^2 * n_i}{n}}$$

ó

4.- Coeficiente de variación: El coeficiente de variación es una medida relativa de "variabilidad". Se calcula como cociente entre la desviación típica y la media. No debe usarse cuando la variable presenta valores negativos o donde el valor 0 sea una cantidad fijada arbitrariamente. Es frecuente mostrarla en porcentajes. Por ejemplo si la media es 80 y la desviación típica 20 entonces CV=20/80=0,25=25% (variabilidad relativa).

Cv = σ / media

Ejemplo: utilizando la serie de datos de la estatura de los alumnos de una clase, vamos a calcular sus medidas de dispersión.

Variable	Frecuencias absolutas		Frecuencias relativas	
(Valor)	Simple	Acumulada	Simple	Acumulada
1.20	1	1	3,3%	3,3%
1.21	4	5	13,3%	16,6%
1.22	4	9	13,3%	30,0%
1.23	2	11	6,6%	36,6%
1.24	1	12	3,3%	40,0%
1.25	2	14	6,6%	46,6%
1.26	3	17	10,0%	56,6%
1.27	3	20	10,0%	66,6%
1.28	4	24	13,3%	80,0%
1.29	3	27	10,0%	90,0%
1.30	3	30	10,0%	100,0%

1.- Rango: Diferencia entre el mayor valor de la muestra (1.30) y el menor valor (1.20). Luego el rango de esta muestra es 10 cm.

2.- Varianza: recordemos que la media de esta muestra es 1.253. Luego, aplicamos la fórmula:

$$S^2_x = \frac{((1.20-1.253)^2 * 1) + ((1.21-1.253)^2 * 4) + ((1.22-1.253)^2 * 4) + \ldots + ((1.30-1.253)^2 * 3)}{30}$$

Por lo tanto, la varianza es 0,0010

3.- Desviación típica: es la raíz cuadrada de la varianza.

$$\sigma = (S_x^2)^{(1/2)}$$

Luego:

$$\sigma = (\ 0{,}010\ )^{(1/2)} = 0{,}0320$$

4.- Coeficiente de variación de Pearson: se calcula como cociente entre la desviación típica y la media de la muestra.

Cv = 0,0320 / 1.253

Luego,

$$Cv = 0{,}0255$$

El interés del coeficiente de variación es que al ser un porcentaje permite comparar el nivel de dispersión de dos muestras. Esto no ocurre con la desviación típica, ya que viene expresada en las mismas unidas que los datos de la serie.

Por ejemplo, para comparar el nivel de dispersión de una serie de datos de la altura de los alumnos de una clase y otra serie con el peso de dichos alumnos, no se puede utilizar las desviaciones típicas (una viene expresada en cm y la otra en kg.). En cambio, sus coeficientes de variación son ambos porcentajes, por lo que sí se pueden comparar.