ANÁLISIS DE DATOS EN PSICOLOGÍA 1

- **CONCEPTOS GENERALES**
- **Introducción**

La estadística actual no sólo es un conjunto de técnicas para resumir y transmitir información cuantitativa, sino que sirve también, y fundamentalmente, para hacer inferencias, generalizaciones y extrapolaciones de un conjunto relativamente pequeño de datos a un conjunto mayor.

Estadística es la ciencia que se ocupa de la ordenación y análisis de datos procedentes de muestras, y de la realización de inferencias acerca de las poblaciones de las que éstas proceden.

- **Conceptos generales**

Se llama población estadística al conjunto de todos los elementos que cumplen una o varias características o propiedades.

Una muestra es un subconjunto de los elementos de una población.

Un parámetro es una propiedad descriptiva de una población.

Un estadístico es una propiedad descriptiva de una muestra.

Una característica es una propiedad o cualidad de un individuo.

Una modalidad es cada una de las maneras como se presenta una característica.

- **Medición**

La estadística no realiza sus funciones directamente sobre las modalidades observadas, sino que éstas se representan por números, y la estadística realiza sus funciones sobre esos números.

Se llama medición al proceso de atribuir números a las características.

La asignación de números a las características se hace siguiendo unas reglas, del estudio de los modelos mediante los cuales conocemos las reglas para una correcta atribución de los números se ocupa la Teoría de la Medida.

A partir de una característica se puede establecer un sistema relacional empírico (empírico, porque se refiere a entidades y relaciones reales). El sistema numérico está formado por un conjunto de entidades (números) y unas relaciones entre ellos. Se trata de un sistema relacional numérico.

El objetivo de la medición de una característica es conectar un sistema relacional empírico y un sistema relacional numérico, de tal forma que las relaciones entre las entidades se reflejen en las relaciones entre los números que los simbolizan. Sólo si se consigue este objetivo ocurrirá que de las relaciones entre los números podrán hacerse inferencias válidas acerca de las relaciones entre las entidades.

La medición estudia las condiciones de construcción de representaciones numéricas, y los modelos desarrollados para la medición se llaman escalas.

Tenemos las escalas nominales, ordinales, cuantitativas de intervalo y cuantitativas de razón.

Escalas nominales: la clave de estas escalas de medida es que sólo informan de la igualdad o desigualdad de los individuos en una característica, pero no de posibles ordenaciones, puesto que la característica a la que se refieren no se tiene en mayor o menor medida, sino que simplemente adopta formas cualitativamente distintas.

Un concepto íntimamente ligado al concepto de escala, y que de hecho las caracteriza, es el de transformación admisible, que hace referencia al problema de la unicidad de la medida. La cuestión de la unicidad puede plantearse de la siguiente forma: ¿es la representación numérica que hemos construido la única posible? En general la respuesta será negativa.

Se dice que una transformación de los números asignados en una escala es una transformación admisible si preserva las características que definen a esa escala, es decir, si los números transformados también representan al sistema empírico.

Esta transformación de los valores originales es una transformación admisible porque los valores obtenidos mediante su aplicación siguen cumpliendo las condiciones especificadas anteriormente para toda escala nominal. En términos más técnicos diríamos que en una escala nominal son admisibles todas las transformaciones que supongan aplicaciones inyectivas.

La aplicación de una regla de asignación de números a las diferentes cantidades de tal forma que los números asignados a los objetos reflejen esos distintos grados en los que se presenta la característica. Los números asignados nos permitirán extraer conclusiones acerca de las magnitudes. A veces lo único que esos números nos permiten inferir son relaciones de tipo mayor que o menor que

A aquellas escalas de medida que cumplen estas características se les llama escalas ordinales. También se dice que se está haciendo una medición a nivel ordinal. Los objetos pueden ordenarse, y de ahí el nombre de la escala.

En psicología son muchas las características cuya medición se considera que está a nivel ordinal, pues son muchos los casos en los que lo único que puede decirse es que un individuo es más extravertido que otro, que un niño es más hiperactivo que otro, o que el aprendizaje es más rápido con el método A que con el método B.

Apliquemos de nuevo el concepto de transformación admisible a este tipo de escalas. No todas las transformaciones que eran admisibles en las escalas nominales lo son para las escalas ordinales.

Al igual que en las escalas nominales, las ordinales tienen unas transformaciones admisibles, que lógicamente serán todas aquellas que preserven las características de la escala ordinal. Se puede demostrar que esto ocurre con todas aquellas transformaciones que cumplan la condición de ser transformaciones crecientes.

Se dice que la transformación es creciente si para todo par de objetos a y b se cumple la siguiente condición:

Si n (a) > n (b), entonces t[n (a)] > t[n(b)]

La limitación de las escalas ordinales es que aunque nos informa de que un objeto presenta la característica en cuestión en una mayor magnitud que otro objeto, no nos dice en cuánto más. Para poder extraer conclusiones más precisas, como la de en cuánto más presenta la característica un objeto sobre otro, hay que contar con una unidad de medida, y para ello hay que pasar al siguiente tipo de escala.

Escala de intervalo, la tercera condición añadida a las exigidas para una escala ordinal impone que el número asignado al objeto y que representamos por n(oi), sea una función lineal de la magnitud real que ese objeto representa en la característica en cuestión. Cuenta con una unidad de medida, si se cumple esta tercera

condición podemos extraer consecuencias acerca de la igualdad o desigualdad de diferencias.

Si la diferencia entre los números asignados a dos objetos es igual a la diferencia entre los números asignados a otros dos, también son iguales las diferencias en magnitudes entre estos dos pares. Una mayor diferencia entre los números asignados implica una mayor diferencia entre las magnitudes representadas.

El ejemplo clásico de este tipo de escalas es el de las temperaturas.

Las transformaciones admisibles para las escalas de intervalo no significan más que un cambio en la unidad de medida y en el origen asignado a la escala, valores ambos arbitrarios en este tipo de escalas.

La principal limitación de este tipo de escalas es que no tiene un cero absoluto. El número cero no representa realmente la ausencia de esta característica.

Las escalas de razón

Esta tercera condición cumple la función de preservar el significado del valor cero, de forma que siempre representa la ausencia de esa característica. La consecuencia fundamental de la presencia de un origen absoluto, y no arbitrario, es que además de poder extraer conclusiones acerca de la igualdad o desigualdad de diferencias, también puede hablarse de la igualdad o desigualdad de razones.

NOMINAL

El sexo de los individuos se clasifica simbolizando con un 0 hembra y con un 1 varón. Posteriormente se hace una transformación admisible, 0 ! 5 y 1 ! 3.

ORDINAL

La dureza de los elementos se ordena, asignándoles números que representen esa ordenación. Posteriormente se hace una transformación admisible, es decir, una que respeta esa ordenación.

INTERVALO

Las cantidades de calor, pueden representarse por distintos conjuntos de números, en tanto en cuanto en ellos se mantenga la diferencia de temperatura entre los objetos 1 y 2 sea la misma que la diferencia entre los objetos 3 y 4, y ambas sean mayores que la diferencia entre los objetos 2 y 3. Estas condiciones las cumplen tanto la escala centígrada como la escala Fahrenheit. Además, de cualquiera de ellas puede pasarse a la otra, pues cada una es una transformación admisible para la otra. Cada una tiene su propia unidad de medida y su origen propio.

RAZÓN

Las longitudes, pueden representarse también por distintos conjuntos de números, en tanto en cuanto en ellos se mantenga que le objeto 2 tenga el doble que le objeto 1, y que el cociente entre los números asignados a los objetos 3 y 1 sea mayor que el cociente entre los números asignados a los objetos 2 y 1. Estas condiciones se cumplen tanto al medir en metros como al medir en yardas. Se puede pasar de una a otra, son transformaciones mutuamente admisibles, ya que aunque cada una tiene su unidad de medida, ambas respetan el cero absoluto, que coincide con las dos, y representa la ausencia de esta característica.

Tipo	**Información deducible**	**Transform. admisibles**	**Ejemplos**
Nominal			Sexo, estado civil, diagnóstico clínico.

	Relaciones igual que o distinto que	Aplicaciones inyectivas	
Ordinal	Relaciones mayor que o igual que	Funciones crecientes	Dureza, nivel socioeconómico, grado de asertividad
Intervalo	Igualdad o desigualdad de diferencias	A + b . x (b " 0)	Temperatura, calendario, inteligencia.
Razón	Igualdad o desigualdad de razones	B . x (b " 0)	Longitud, peso.

1.3.1 Las variables: clasificación y notación

- Una variable es una representación numérica de una característica.

Ejemplo	**Tipo de estudio**	**Variables**	**Tipo de escala**
1	Descriptivo	Grado de patrón A	Intervalo
2	Inferencial	Grupo, Nivel cultural, Inteligencia, estrés.	Nominal, Ordinal, Intervalo, Intervalo
3	Inferencial	Tiempo de reacción	Razón
4	Inferencial	Intención de voto	Nominal

Estadística descriptiva con una variable

- **ORGANIZACIÓN Y REPRESENTACIÓN DE DATOS**

- **Distribución de frecuencias**

La distribución de frecuencias es un instrumento diseñado para cumplir tres funciones:

- Proporcionar una reorganización y ordenación racional de los datos recogidos.
- Ofrecer la información necesaria para hacer representaciones gráficas
- Facilitar los cálculos necesarios para obtener los estadísticos muestrales

- Se llama frecuencia absoluta de un valor Xi, y se simboliza por ni, al número de veces que se repite el valor Xi, en la muestra.

- Se llama frecuencia relativa de un valor Xi, y se simboliza por pi, al cociente entre la frecuencia absoluta de ese valor y el tamaño de la muestra. Es decir: pi = ni / n

- Se llama frecuencia absoluta acumulada de un valor Xi, y se simboliza por na, al número de veces que se repite en la muestra ese valor Xi, o cualquier otro valor inferior.

- Se llama frecuencia relativa acumulada de un valor Xi, y se simboliza por pa, al cociente entre su frecuencia absoluta acumulada y el tamaño de la muestra. Es decir: pa = na / n.

Las frecuencias relativas, se expresan en términos porcentuales. Suelen representarse con mayúsculas, para obtenerlas basta con multiplicar por 100 las frecuencias relativas. Para cualquier valor de la variable, Xi tenemos que:

Pi = pi · 100 y Pa = pa · 100

Una distribución de frecuencias se organiza en forma de tabla. En una distribución de frecuencias completa

aparece, una columna con los valores que adopta la variable, creciendo de abajo hacia arriba.

Construimos la distribución de frecuencias siguiendo los pasos descritos:

- Ponemos en la primera columna esos valores, creciendo de abajo hacia arriba
- Para la columna de frecuencias absolutas contamos el número de veces que se repite cada valor, si el número de valores es muy grande conviene ir haciendo marcas por cada valor, para contarlas al final
- Para la columna de frecuencias relativas dividimos cada frecuencia absoluta por n
- Para obtener las frecuencias absolutas acumuladas sumamos para cada valor su frecuencia absoluta más la absoluta acumulada del valor anterior.
- Para las frecuencias relativas acumuladas dividimos cada frecuencia absoluta acumulada por n. La frecuencia relativa acumulada del valor mayor debe ser igual a 1.

Distribución de frecuencias construida sobre el ejemplo del número de hijos (texto)

Xi	**ni**	**pi**	**na**	**pa**
4	1	0.05	20	1.00
3	3	0.15	19	0.95
2	7	0.35	16	0.80
1	6	0.30	9	0.45
0	3	0.15	3	0.15
	20	1.00		

Agrupación en intervalos: consisten en formar grupos de valores consecutivos, llamados intervalos, y poner uno de estos grupos en cada fila, en lugar de poner cada valor individual por separado. Cada uno de estos grupos suele indicarse en la distribución de frecuencias poniendo los valores mayor y menos incluidos en él.

A continuación se calculan las frecuencias absolutas conjuntas de los valores incluidos en el intervalo, haciendo lo mismo después con las frecuencias relativas, las absolutas acumuladas y las relativas acumuladas.

- Se llama intervalo a cada uno de los grupos de valores que ocupan una fila en una distribución de frecuencias. En algunos textos se llaman clases.

- Se llaman límites aparentes o informados de un intervalo a los valores mayor y menor que puede adoptar la variable dentro de ese intervalo, según el instrumento de medida utilizado.

- Se llaman límites exactos de un intervalo a los valores máximo y mínimo incluidos en el intervalo y que podrían medirse si se contara con un instrumento de precisión perfecta.

- Se llama punto medio de un intervalo a la suma de sus límites exactos partido por dos. En algunos libros se llama marca de clase.

- Se llama amplitud de un intervalo a la diferencia entre su límite exacto superior y su límite exacto inferior. Suele representarse por la letra I.

Para hacer una distribución de frecuencias:

- El intervalo superior debe incluir al mayor valor observado.
- el intervalo inferior debe incluir al menor valor observado.
- Cada intervalo debe incluir el mismo número de valores.

Dado que el objetivo de una distribución de frecuencias es conseguir una ordenación manejable que ayude a comprender el significado de los datos, no es conveniente que el número de intervalos sea demasiado grande.

Como consecuencia de lo anterior, podemos sentirnos inclinados a reducir al máximo el número de intervalos, pero lo cierto es que esto traería consigo una consecuencia negativa, los intervalos tendrían una excesiva amplitud y acabaríamos teniendo a sujetos con puntuaciones muy distintas en el mismo intervalo.

El número apropiado de intervalos debe ser tal que, con ella se consiga una agrupación operativa y que cumpla los objetivos para los que ha sido diseñada la distribución de frecuencias, pero sin distorsionar excesivamente los valores con el error de agrupamiento.

A veces hay casos en los que hacer un número de intervalos siguiendo las directrices que acabamos de plantear distorsionarían demasiado los datos.

Para evitar eso se utilizan lo que se denomina intervalos abiertos, en los cuales no se pone el límite inferior del intervalo que incluye los valores menores, el límite superior del intervalo que incluye los valores mayores o no se pone ninguno de estos dos. Ej. + de ....

Problema de los bordes

Supongamos que vamos a construir una agrupación en intervalos, siendo los valores mayor y menor observados iguales a 79 y 43, respectivamente. Como el número de valores distintos sería igual a 37, que es un número primo, no pueden hacerse intervalos de amplitud constante tales que el mayor tenga al 79 como límite aparente superior y al 43 como límite aparente inferior. En estos casos suele añadirse al listado de valores distintos observados algunos otros valores no observados en la muestra.

Estos valores tendrán frecuencias absolutas iguales a cero, pero nos permitirán conseguir un número de valores distinto que sea múltiplo del número de intervalos que queremos hacer.

Para no distorsionar demasiado ninguno de los intervalos extremos es preferible repartirlos lo más homogéneamente posible entre los dos.

- **Supuestos de distribución intraintervalo**

Una vez confeccionada una distribución de frecuencias con datos agrupados en intervalos, ésta se puede utilizar para hacer representaciones gráficas y para facilitar los cálculos de estadísticos que iremos explicando.

Dado que de cada puntuación sólo sabemos el intervalo al que pertenece, un procedimiento que a veces resultará útil consiste en asumir el supuesto de concentración en el punto medio. Según este supuesto, trataríamos a esos dos datos como si fueran dos valores iguales. Entonces este es el punto medio de su intervalo.

El supuesto de distribución homogénea, los valores incluidos en un intervalo se reparte con absoluta conformidad en su interior, si en un intervalo hay cinco observaciones, aceptaremos que sus valores son los que tendríamos si partiéramos al intervalo en cinco subintervalos de igual amplitud y asignáramos a cada individuo el punto medio de un subintervalo.

- **Representaciones gráficas**

A partir de las distribuciones de frecuencias se pueden construir representaciones gráficas. La función de éstas es dar informaciones globales mediante un solo golpe de vista.

- **Representaciones gráficas de uso frecuente**

    - Diagrama de rectángulos: Para hacer un diagrama de rectángulos se colocan en el eje de abscisas las modalidades (o los números que las representan) y en el eje de ordenadas las frecuencias. Sobre cada modalidad se levanta un rectángulo cuya altura es la frecuencia correspondiente. Este tipo de representaciones se suele utilizar para variables nominales, pero también se utiliza para variables ordinales, como el nivel cultural.

    - Perfil ortogonal: Se utiliza mucho en informes psicopedagógicos o de rendimiento. Calificaciones obtenidas por un alumno a lo largo de 4 exámenes.

    - Pictograma: Son representaciones en forma de círculos en las que éstos son divididos en secciones cuya superficie es proporcional a la frecuencia de la modalidad correspondiente.

    - Diagrama de barras: Se utiliza para variables cuantitativas discretas. En el eje de abscisas se colocan los distintos valores de la variable y en el eje de ordenadas las frecuencias. Sobre cada valor de la variable se traza una línea o barra perpendicular cuya altura debe ser igual a la frecuencia.

    - Histograma: Se utiliza para variables cuantitativas continuas con datos agrupados en intervalos. En el eje de abscisas se colocan los límites exactos de los intervalos, y en el eje de ordenadas las frecuencias.

    - Polígono de frecuencias: Para variables discretas, el polígono de frecuencias es la figura que resulta de unir los extremos superiores de las que hubieran sido las barras. Si se trata de las bases superiores de los rectángulos correspondientes a un hipotético histograma construido con esos mismos datos.

    - Diagrama de barras acumulativo: Se utiliza en variables discretas. En el eje de abscisas se colocan los valores de la variable, y en el de ordenadas las frecuencias acumuladas, ya sean absolutas o relativas. Sobre cada valor se traza una perpendicular cuya longitud sea igual a la frecuencia acumulada. Desde el extremo superior de cada una de estas barras se traza una línea horizontal que se une con la barra situada a su derecha.

    - Polígono de frecuencias acumuladas: Se utilizan en variables continuas. El eje de abscisas se construye igual que en los histogramas, pero en el de ordenadas se incluyen las frecuencias acumuladas, ya sean absolutas o relativas. Sobre cada límite se levanta una perpendicular cuya longitud sea idéntica a la frecuencia acumulada y se unen los extremos superiores de dichas perpendiculares.

- **Convenciones sobre las representaciones gráficas**
- En el eje de abscisas colocamos los valores de la variable, y en el de ordenadas las frecuencias (absolutas o relativas, simples o acumuladas).
- La intersección de los dos ejes es el origen , de modo que en el eje de abscisas las puntuaciones más bajas estarán a la izquierda, y las más altas a la derecha; en el de ordenadas los valores los valores pequeños estarán abajo y los altos arriba.
- Si el valor mínimo del eje de abscisas fuera excesivamente grande, se debe cortar la línea
- Conviene incluir en cada gráfico toda la información posible para evitar ambigüedades y facilitar su interpretación por otras personas o por nosotros mismos al cabo del tiempo.
- Cuando un mismo gráfico se representan dos o más grupos simultáneamente y éstos son de tamaños considerablemente distintos se deben utilizar frecuencias relativas.

Las representaciones sirven para comunicar información de un solo golpe de vista, y por ello en su construcción debe tenerse en cuenta el público al que va dirigida, sus necesidades de informaciones más bien

globales y generales o específicas y precisas, y cualquier otra consideración que pueda mejorar la transmisión de información ágil y precisa.

- **Tendenciosidad en las representaciones gráficas**

Un primer método consiste en recortar el eje de ordenadas, eliminando los menores valores de frecuencias con la excusa de que no hay ninguna observación que las adopte. Esto tiene como consecuencia que pequeñas diferencias parezcan mayores.

Un segundo tipo de distorsión se produce cuando se utilizan figuras representativas de aquello que se está midiendo. Suelen hacerse proporcionando sus alturas a las frecuencias correspondientes, el incremento en la altura conlleva también un incremento en la anchura. Como consecuencia de ello, la superficie de las figuras no guarda relación con las frecuencias observadas, dando la impresión de que la diferencia es mayor que la realmente registrada.

- **Propiedades de las distribuciones de frecuencias**

Los polígonos de frecuencias dependen demasiado de la unidad de medida utilizada, de la agrupación en intervalos hecha y de las fluctuaciones particulares esperables en una muestra concreta. Las curvas suavizadas suelen ser representaciones más apropiadas que los polígonos de frecuencias simples. Son cuatro las propiedades con las que describiremos las distribuciones de frecuencias:

- Tendencia central: Una primera propiedad es la que se refiere a la magnitud general de las observaciones hechas. Esta magnitud general puede cuantificarse mediante unos índices conocidos como índices de tendencia central o promedios, y que reciben ese nombre porque pretenden ser síntesis de los valores de la variable.
- Variabilidad: Grado de concentración de las observaciones en torno al promedio. Una distribución de frecuencias será homogéneo o poco variable si los datos difieren poco entre sí, y por tanto, se agolpan en torno a su promedio. Sería heterogénea o muy variable si los datos se dispersan mucho con respecto al promedio.
- Asimetría o sesgo: Esta propiedad se refiere al grado en que los datos tienden a concentrarse en los valores centrales, en los valores inferiores al promedio, o en los valores supriores a éste. Existe simetría perfecta cuando en caso de doblar la representación gráfica por una vertical trazada sobre la media, las dos mitades se superponen perfectamente. Las distribuciones con asimetría negativa son propias de las pruebas, tareas o tests fáciles, en las que la mayoría de los sujetos puntúan alto. Las distribuciones asimétricas positivas son típicas de pruebas, tareas o tests difíciles en las que la mayoría de los sujetos puntúan bajo.
- Curtosi: Se refiere al grado de apuntamiento de la distribución de frecuencias. Si es muy apuntada se llama leptocúrtica y si es muy aplastada, se llama platicúrtica.
- **Diagrama de tallo y hojas**

Las distribuciones de frecuencias no son el único medio para resumir y exponer conjuntos de datos; una alternativa a ellas son los llamados diagramas de tallo y hojas.

Su obtención requiere separar cada puntuación en dos partes. El primer o primeros dígitos, que reciben el nombre de tallo, y el dígito o dígitos restantes, que reciben el nombre de hojas; por ejemplo, X = 56 se puede separar en 5 (tallo) y 6 hoja. Estos diagramas tienen la suficiente flexibilidad como para admitir otras posibilidades.

- Se identifican los valores máximo y mínimo observados.
- Se toma una decisión acerca del número más apropiado de tallos distintos.
- Se listan todos los tallos distintos en una columna, ordenados de forma creciente de arriba abajo.
- Se escribe cada hoja junto al tallo que le corresponda, preferiblemente ordenados según su valor.

En general, un número de tallos superior a cinco y que no pase de 20 suele ser apropiado. Aparte de ser más fácil de construir, el diagrama de tallo y hojas tiene varias ventajas sobre la distribución de frecuencias, y también algún inconveniente:

- Ventaja: permite identificar cada puntuación individual. En las distribuciones tradicionales sólo conocemos la frecuencia del intervalo y nos obliga a tratar los datos de ciertas maneras distorsionantes. La ventaja de retener cada valor individual viene acompañada del inconveniente de que le diagrama de tallo y hojas no facilita, como la distribución de frecuencias clásica, el cálculo de los estadísticos que estudiaremos más adelante.
- Ofrece simultáneamente tanto un listado de las puntuaciones como un dibujo de distribución, si tumbamos el diagrama obtenemos una especie de histograma.
- Al contener los valores de cada observación, es más fácil de modificar para obtener un dibujo con un nivel de detalle distinto, mayor o menor, de la distribución.
- Pueden presentarse dos conjuntos de datos simultáneamente en el mismo diagrama, con lo que se facilita la comparación.

- **MEDIDAS DE POSICIÓN**

- **Centiles o percentiles**

Son 99 valores de la variable que dividen a la distribución en 100 secciones, cada una conteniendo a la centésima parte de las observaciones. Se pueden representar por la inicial de cada uno de los dos términos que los designan más el subíndice correspondiente, Ck o Pk (k = 1,2,...99). Se simboliza por C28 a aquella puntuación que deja por debajo de sí al 28 por 100 de las observaciones y que es superada por el 72 por 100.

Aunque por definición son sólo 99 valores, por extensión a veces se utilizan posiciones intermedias, como, por ejemplo, el centil 88,5 o C88,5, que sería aquel valor de la variable por debajo del cual se encuentra el 88,5 por 100 de las observaciones.

Dado que los valores correspondientes a los centiles se determinan en función de los porcentajes de observaciones, normalmente las distancias entre ellos, en términos de puntuación, no serán constantes. Generalmente las distancias entre los centiles intermedios serán menores que las distancias entre centiles extremos.

Las puntuaciones correspondientes a los centiles 55 y 56 serán más cercanas entre sí que las puntuaciones correspondientes a los centiles 98 y 99, o las de los centiles 2 y 3. Esto se dará, en distribuciones simétricas, mientras que a medida que las distribuciones se van haciendo más asimétricas esta relación hay que matizarla algo más.

Los centiles no suelen calcularse con cantidades pequeñas de datos y cuando es necesario hacerlo se obtienen sencillamente ordenando las puntuaciones y calculando la proporción de éstas que superan al valor que se quiere comparar. Normalmente los centiles se obtienen sobre datos agrupados en intervalos, y en su cálculo se asume el supuesto de distribución homogénea intraintervalo.

Estamos buscando la puntuación que deja por debajo de sí a 140, que es una cantidad intermedia entre 90 y 150. El procedimiento que se utiliza para calcular ese valor, y que se recoge en la fórmula que veremos a continuación, consiste en adoptar como representación del C70 un valor perteneciente al intervalo 11 − 14 que mantenga una relación de proporcionalidad con respecto a la frecuencia buscada. Se trata de buscar una puntuación de ese intervalo que divida a las observaciones pertenecientes a él en dos grupos, uno que incluya a las 50 observaciones inferiores y otro que incluya a las 10 restantes. De esta forma, ese valor dejará por debajo de sí a 50 observaciones pertenecientes al intervalo, más las 90 que quedaban por debajo de su límite exacto inferior, totalizando las 140 buscadas. Dado que esa puntuación debe dejar a 50 por debajo y 10 por

encima, debe ser una puntuación más cercana al límite superior del intervalo que al límite inferior. El procedimiento se resume en la siguiente fórmula:

$$Ck = Li + I \cdot k \cdot n - na$$

ni 100

Ck es la puntuación correspondiente al centil k

Li es el límite exacto inferior del intervalo crítico

I es la amplitud de los intervalos

ni es la frecuencia absoluta del intervalo crítico.

k es el porcentaje de observaciones inferiores a Ck

n es el número de observaciones hechas

na es la frecuencia absoluta acumulada hasta Li

- **Otros cuantiles**
- **Deciles**

Son nueve puntuaciones que dividen a la distribución en 10 partes, cada una conteniendo al 10 por 100 de las observaciones. Se representan por Dk, donde k indica el número del decil al que se refiere.

- **Cuartiles**

Son tres puntuaciones que dividen a la distribución en cuatro partes, cada una conteniendo al 25 por 100 de las observaciones. Se representan por Qk, donde k indica el número del cuartil al que se refiere.

- **MEDIDAS DE TENDENCIA CENTRAL**

- **La media aritmética**

El índice de tendencia central más utilizado es la media. Se define como l asuma de los valores observados, dividida por el número de ellas. Se representa con la misma letra que representa la variable, en mayúsculas , con una barra horizontal encima. Por tanto, si recogemos n observaciones de la variable X, entonces la medida de los valores observados es:

$$øX = Xi / n$$

De donde se deduce que:

$$Xi = n \cdot øX$$

- **Cálculo en una distribución de frecuencias**

Cuando se tiene un conjunto grande de observaciones, éstas tradicionalmente se han agrupado en distribuciones de frecuencias, para luego hacer los cálculos sobre la distribución.

Vamos a describir el procedimiento para hacer los cálculos de la media con datos agrupados en una distribución de frecuencias.

Para hacer los cálculos se asume el supuesto de concentración en el punto medio del intervalo.

Para hallar la media se asume el supuesto de concentración en el punto medio del intervalo.

$$\text{ø}X = \frac{ni \cdot Xi}{n}$$

Se diferencia de la fórmula anterior en que: el sumatorio no tiene n sumandos, como en la fórmula anterior sino tantos como intervalos tenga la distribución y las Xi no son los datos directos, sino los puntos medios de los intervalos.

- **Propiedades de la media aritmética**

Puntuaciones directas: valores brutos y los representamos por la letra de la variable en mayúsculas.

Puntuaciones diferenciales: diferencias de cada sujeto con respecto a la media grupal, las representamos por la letra minúscula.

$$Xi = Xi - \text{ø}X$$

- La suma de las diferencias de n puntuaciones con respecto a su media, o puntuaciones diferenciales, es igual a cero: $xi = 0$

La razón por la que la suma de las diferenciales es igual a cero es que unas son positivas y otras negativas y se compensan unas con otras.

- La suma de los cuadrados de las desviaciones de unas puntuaciones con respecto a su media es menor que con respecto a cualquier otro valor.

- Si sumamos una constante a un conjunto de puntuaciones, la media aritmética quedará aumentada en esa misma constante.

- Si multiplicamos por una constante a un conjunto de puntuaciones, la media aritmética quedará multiplicada por esa misma constante.

- La media total de un grupo de puntuaciones, cuando se conocen los tamaños y las medias de varios subgrupos hechos a partir del grupo total, mutuamente exclusivos y exhaustivos, puede obtenerse ponderando las medias parciales a partir de los tamaños de los subgrupos en que han sido calculadas.

$$\text{ø}XT = \frac{n1 \cdot \text{ø}X1 + n2 \cdot \text{ø}X2 + \ + nk \cdot \text{ø}Xk}{n1 + n2 + n3 + \ + nk}$$

- Una variable definida como la combinación lineal de otras variables tiene como media la misma combinación lineal de las medias de las variables intervinientes en su definición

- **La mediana**

Mediana: aquella puntuación que fuera superada por la mitad de las observaciones, pero no por la otra mitad, se suele representar por Mdn.

Para su cálculo podemos encontrarnos en dos casos generales, aquel en el que contamos con un número impar de observaciones y aquel que nos encontramos con un número par de ellas. En el primero se toma como mediana el valor central: en el segundo se da la circunstancia de que cualquier valor comprendido entre los dos centrales cumple con la definición de la mediana. Fechner propuso tomar la media aritmética de los dos valores centrales.

- **La moda**

Una tercera vía para representar la tendencia central de un conjunto de valores consiste en informar del valor más frecuentemente observado. Se presenta por Mo, y se define sencillamente como el valor de la variable con mayor frecuencia absoluta.

Como norma, para obtener la moda ordenaremos los valores de menor a mayor para así facilitar la identificación del de mayor frecuencia.

- Cuando todos los valores tienen la misma frecuencia, es un caso en el que la moda no se puede calcular, decimos que es una distribución amodal.
- Cuando hay dos valores con la misma y máxima frecuencia en este caso se dice que la distribución tiene dos modas o que es una distribución bimodal.

- Cuando disponemos de una distribución de frecuencias, se toma como moda el punto medio del intervalo con mayor frecuencia. También en distribuciones de frecuencias pueden darse los casos anteriores. En ellos se utilizarían las mismas reglas que acabamos de exponer, pero aplicadas a los puntos medios de sus intervalos.

- **Comparación entre medidas de tendencia central**

Si no hay ningún argumento de peso en contra, se preferirá siempre la media. Hay dos razones para apoyar esta norma general. La primera es que en ella se basan otros estadísticos y la segunda es que es mejor estimador de su parámetro que la mediana y la moda.

Hay al menos 3 situaciones en las que se preferirá la mediana a la media:

- Cuando la variable esté medida en escala ordinal
- Cuando haya valores extremos que distorsionen la interpretación de la media
- Cuando haya intervalos abiertos, situaciones en las que el intervalo superior carece de límite superior, el intervalo inferior carece de límite inferior o ambos.

La media es extremadamente sensible a las puntuaciones y un cambio en sólo una de ellas supone un cambio en la media aritmética, mientras que la mediana sólo se vería alterada por cambios en los valores centrales.

La mediana será la segunda candidata para representar la tendencia central y si no hay argumentos de peso en contra se preferirá la mediana a la moda:

- Cuando se trate de una variable medida en escala nominal
- Cuando haya intervalos abiertos y la mediana pertenezca a uno de ellos.

- **MEDIDAS DE VARIACIÓN**

- **Medidas de variación**
- **Varianza y desviación típica**

Una idea que se ha demostrado útil a la hora de cuantificar la variabilidad es la de trabajar con las distancias desde los valores hasta algún poste centra, que podría ser la media aritmética y basar la medición de la dispersión en algún tipo de separación promedio hasta ese poste

Una solución al problema de que las distancias con respecto a la media sumen cero consiste en elevar al cuadrado esas distancias antes de hallar su promedio, dado que los cuadrados son siempre positivos. El índice basado en esta idea se llama varianza y se representa por la expresión S2x, donde el subíndice recoge la letra con la que se representa la variable. Al cálculo del promedio de las desviaciones cuadráticas con respecto a la media:

$$S2x = \frac{(X - øX)2}{n}$$

$$S2x = \frac{x2}{n}$$

La cuestión que puede surgir es la de cómo valorar el grado de dispersión cuantificado mediante este índice. En realidad no tiene mucho sentido hablar de niveles altos o bajos de dispersión en términos absolutos, sino, en todo caso, en términos comparativos.

La varianza sirve sobre todo para comparar el grado de dispersión de dos o más conjuntos de valores en una misma variable, llegando a conclusiones como la siguiente: La población de hombres presenta una mayor variabilidad en su estatura que la población de mujeres, que son más homogéneas en esa característica

El valor 27,2 no parece un número claramente relacionado con lo que se pretendía medir. Las mayores distancias que presentan esos valores con respecto a la media son de 8 puntos y parece que una representación numérica de la magnitud general de esas distancias estaría bastante alejada de 27,2. La razón de esta discrepancia es que las distancias no se han tratado como tales, sino que para evitar el problema de que las diferenciales sumen cero se han elevado éstas al cuadrado. Por ello es frecuente que, con objeto de retornar las unidades originales de esas distancias, se calcule la raíz cuadrada de la cantidad obtenida. Al índice así hallado se le llama desviación típica, se representa por Sx y se define sencillamente como la raíz cuadrada de la varianza:

$$Sx = \frac{(Xi - øX)2}{n}$$

La desviación típica es un mejor descriptor de la variabilidad, aunque la varianza tenga algunas notables propiedades que la hacen idónea para basar en ella los análisis estadísticos complejos.

La variabilidad de los datos está reflejando el hecho incuestionable de las diferencias individuales y éstas son uno de los objetos de estudio primordiales de la psicología.

Cuasivarianza: dividiendo por n − 1, representamos por S'2x

$$n \cdot S2x = (n - 1) \cdot S'2x$$

- **Cálculo y propiedades de la varianza**

    - La varianza y la desviación típica como medidas de dispersión, son valores esencialmente positivos.

    - Si sumamos una constante a un conjunto de puntuaciones, su varianza no se altera

    - Si multiplicamos por una constante a un conjunto de puntuaciones, la varianza quedará multiplicada por el cuadrado de la constante, y la desviación típica por el valor absoluto de esa constante.

    - La varianza total de un grupo de puntuaciones, cuando se conocen los tamaños (ni), las medidas (øXi) y las varianzas (S2i) de varios subgrupos hechos a partir del grupo total, mutuamente exclusivos y exhaustivos, puede obtenerse sumando la media (ponderada) de las varianzas y la varianza (ponderada) de las medias:

S2T = ni · S2i + ni (øXi − øXT)2

ni ni

    - La desigualdad de Tchebychev recoge el hecho de que las distancias menores hasta la media son más frecuentes que las distancias mayores. Así, entre las puntuaciones correspondientes a la media +− una desviación típica se encontrarán menos observaciones que entre las puntuaciones correspondientes a la media +− una dos desviaciones típicas. Según la desigualdad de Tchebychev, el porcentaje de puntuaciones que quedan entre las correspondientes a la media +− k desviaciones típicas es, como mínimo, el:

(1 − 1 por 100 de las observaciones

k2

- **Otras medidas de variación**

Una forma muy sencilla de indicar el grado de dispersión consiste en calcular la distancia entre el mayor y el menor de los valores observados. Este índice se llama amplitud total, rango o recorrido, y se obtiene sencillamente hallando la diferencia entre los valores extremos:

AT = Xmáx − Xmin

La principal desventaja de este índice es que es muy sensible a los valores extremos, y nada sensible a los intermedios, pudiendo carecer de toda representatividad.

Otro inconveniente de este índice es que está ligado al tamaño de la muestra utilizada, Si se quiere comparar la variabilidad de la dispersión de dos conjuntos de datos de tamaño marcadamente distinto, es probable que la muestra de mayor tamaño presente una mayor amplitud aunque las poblaciones de referencia tengan la misma variabilidad.

Desviación media: tomar las desviaciones con respecto a la media, o puntuaciones diferenciales, en valor absoluto. Se representa por sus iniciales (DM):

DM o ST = | Xi − øX|

n

La desviación media representa un promedio de distancias tomadas en valor absoluto y representa bien el concepto de dispersión y su cuantificación, aunque no es muy utilizado en psicología debido a la dificultad que supone el trabajo con valores absolutos, y que hace que no haya muchas técnicas de análisis estadístico basadas en ella.

Cuando en las puntuaciones hay algún valor extremo que pudiera distorsionar la representatividad de la varianza se puede utilizar otro índice, basado sólo en las puntuaciones correspondientes a los cuartiles primero y tercero. Se denomina amplitud semi − intercuartil, se representa por la letra Q

Q = Q3 − Q1

2

Esta medida de variabilidad elimina del cómputo las puntuaciones extremas que no le afectan.

Coeficiente de variación: Comparar la variabilidad de grupos cuya media es claramente distinta, relativizar la desviación típica con respecto a la media, expresado como un porcentaje, se representa por CV

CV = Sx · 100

øX

Cuanto mayor es el coeficiente de variación, menos representativa es la media.

- **Representación gráfica de la variabilidad**

Box and whiuskers, que significa literalmente caja y bigotes. Para su construcción se marcan señales de tal forma que las distancias entre ellas sean proporcionales a las distancias entre la puntuación máxima y mínima y los 3 cuartiles. Con los 3 cuartiles se forma una especie de ficha de dominó, mientras que las puntuaciones máxima y mínima se unen mediante líneas rectas a los bordes de esta forma geométrica. Se puede comparar la variabilidad de dos distribuciones haciendo representaciones paralelas de caja y bigotes.

En otros casos se quiere representar la evolución de los valores medios, se pueden unir mediante un trazo los puntos correspondientes y añadir unos bigotes verticales que indiquen los valores correspondientes a una desviación típica.

- **PUNTUACIONES TÍPICAS Y ESCALAS DERIVADAS**

6.1 Puntuaciones típicas

Puntuación diferencial ! la distancia o diferencia, entre esa puntuación y la media del grupo de puntuaciones.

Las puntuaciones diferenciales son más informativas e interesantes que las directas, pues al menos nos indican si la puntuación es superior o inferior a la media o si coincide con ella. Esta información es insuficiente para comparar puntuaciones de sujetos pertenecientes a distintos grupos o a distintas variables.

Variabilidad del grupo de referencia: se trataría de indicar cómo de grande es una distancia en términos de las distancias observadas en general en esas puntuaciones. Esa distancia general es la desviación típica. Las puntuaciones así conseguidas se denominan, puntuaciones típicas, se representan por letras z minúsculas y su fórmula es:

zi = Xi − øX

Sx

Es idéntica al cociente entre la puntuación diferencial y la desviación típica:

zi = xi

Sx

Al proceso de obtención de las puntuaciones típicas se le llama tipificación. La definición de las puntuaciones típicas puede basarse en esta idea y expresarse como sigue:

- La puntuación típica de una observación indica el número de desviaciones típicas que esa observación se separa de la media del grupo de observaciones.

Las puntuaciones típicas permiten hacer comparaciones entre unidades de distintos grupos, entre variables medidas de distintas formas o incluso entre variables diferentes. Siempre nos indican el número de desviaciones típicas que se separan de la media y si esa desviación es por encima o por debajo de la media.

Las típicas no son más que una transformación lineal que consiste en multiplicar las directas por una constante (el inverso de la desviación típica) y luego sumar a esos productos otra constante (el cociente entre la media y la desviación típica, con signo negativo)

Zi = Xi − øX = 1 . Xi + − øX

Sx Sx Sx

Estas características de las puntuaciones típicas son universales, no dependen del tipo de puntuaciones, ni de su dispersión, ni de su número.

- La media de las puntuaciones típicas es cero, mientras que su varianza y desviación típica son iguales a uno.

Las puntuaciones típicas reflejan las relaciones esenciales entre las puntuaciones con independencia de la unidad de medida que se haya utilizado en la medición.

6.2 Escalas derivadas

Inconveniente ! La medida de las típicas es cero y su desviación típica uno, buena parte de las puntuaciones suelen ser negativas y casi todas decimales.

Un procedimiento consiste en transformar las puntuaciones típicas en otras que retengan todas las relaciones que manifiestan las puntuaciones originales, que sean puntuaciones equivalentes pero evitando la dificultad operativa y que constituyen lo que se denomina una escala derivada.

Las puntuaciones transformadas tienen como media y desviación típica las constantes utilizadas para la transformación, podemos conseguir que las puntuaciones en una escala derivada tengan las características que nos resulten más cómodas, sencillamente haciendo la transformación con las constantes que deseamos como media y desviación típica.

- Si transformamos linealmente las puntuaciones típicas, multiplicándolas por una constante a, y sumando una constante b, entonces las puntuaciones transformadas tendrán como media la constante sumada, b, como desviación típica el valor absoluto de la constante multiplicada (a) y como varianza

el cuadrado de esta constante, a2

La construcción de una escala derivada parte de unas puntuaciones directas, éstas se tipifican y después se transforman linealmente en otras puntuaciones.

Esquema de la transformación en una escala derivada:

Tipificación

Transformación en escala derivada:

Ti = a . zi + b

La cuestión fundamental de las escalas derivadas consiste en transformar las puntuaciones originales, Xi, en otras puntuaciones transformadas, Ti, tales que sean más cómodas de tratar e interpretar, pero que a la vez retengan las relaciones esenciales entre los valores, que sean puntuaciones equivalentes.

Cualquier transformación lineal en la que la constante multiplicadora sea positiva da lugar a unas puntuaciones equivalentes.

- **MEDIDAS DE ASIMETRÍA Y CURTOSI**

- **Índices de asimetría**

El grado de asimetría de una distribución hace referencia al grado en que los datos se reparten equilibradamente por encima y por debajo de la tendencia central. Hay diferentes índices con los que cuantificar esta propiedad:

El primero de ellos se basa en la relación entre la media y la moda, y se define como la distancia entre la media y la moda, medida en desviaciones típicas.

La media es inferior a la moda, y por tanto este índice dará un valor negativo, mientras que en la figura c la media es superior y el índice dará positivo. En la distribución de la figura b coinciden los dos índices de tendencia central, y por tanto el índice de asimetría dará cero.

Las distribuciones del tipo de la figura a se dice que tienen asimetría negativa y el índice da valores menores que cero. Las del tipo de la figura c se dice que tienen asimetría positiva y este índice da valores mayores que cero. Las del tipo de la figura b se dice que son distribuciones simétricas, puesto que no están inclinadas hacia ningún lado; este índice da en ellas valores en torno a cero y si la simetría es perfecta entonces da exactamente cero. Este índice tiene la dificultad de que sólo se puede calcular en distribuciones unimodales.

Índice de asimetría de Pearson, es igual al promedio de las puntuaciones típicas elevadas al cubo:

As = (øz3) = [(øXi − øX)3/ Sx3] = 1 . (Xi − øX)3

n n. S3x

Los valores menores que cero indican asimetría negativa, los mayores de cero asimetría positiva y los valores en torno a cero indican distribuciones aproximadamente simétricas. Es el índice de asimetría más utilizado.

El índice de asimetría intercuartílico se basa, en los cuartiles. Su fórmula es:

As = (Q3 − Q2) − (Q2 − Q1)

Q3 − Q1

Los valores mayores de cero indican asimetría positiva, los menores indican asimetría negativa y los valores en torno a cero reflejan distribuciones aproximadamente simétricas. Tiene una ventaja sobre los índices anteriores y es que tiene un valor máximo y mínimo con lo que se facilita su interpretación en términos relativos.

- **Índice de curtosis**

Sólo vamos a estudiar el que se basa en el promedio de las típicas elevadas a la cuarta potencia. Su fórmula es:

Cr = (øz4) = [(Xi − øX)4 / S4x] − 3 = 1 . . (Xi − øX)4 − 3

n n. S4x

Al restar un tres al índice lo que se consigue es utilizar ese modelo como patrón de comparación. Una distribución en la que el índice sea igual a cero tiene un grado de curtosis similar al de la distribución normal y, siguiendo la terminología propuesta por Pearson, se dice que es mesocúrtica, mientras que si es positivo su grado de apuntamiento es mayor que el de la distribución normal y se dice que es una distribución leptocúrtica y si es negativo su apuntamiento es menor que el de la distribución normal y se dice que es platicúrtica.

23

Puntuaciones transformadas (Ti)

Media: b

Varianza: a2

Puntuaciones típicas (zi)

Media: 0

Varianza: 1

Puntuaciones directas (Xi)

Media: øX

Varianza: S2x