

Análisis y evaluación de resultados.

Análisis y evaluación de resultados.

Caso práctico



Ramón y Claudia ya han vuelto de sus “vacaciones” con la ONG, y han empezado con su vida de cada día en el centro de salud y en la clínica dental. Hoy han quedado con su grupo para **analizar** los datos y la información que han obtenido. Estará con ellos **Maite**, una amiga que trabaja en la universidad y que les va a ayudar.

Comienza la reunión y le recuerdan que estuvieron midiendo el ICAOD y el índice de restauración en niños y niñas de 6 y 12 años, enfermedad periodontal y ausencias en adolescentes y adultos, y que pasaron cuestionarios a la gente para conocer algunos hábitos sobre consumos y cepillado.

—Maite, ¿qué crees que podemos hacer con estos datos?

—Mira Ramón, tenéis que ordenarlos y calcular las medias aritméticas de las variables con sus desviaciones típicas. También es importante, que incluyáis algunos datos para caracterizar la población, como la distribución por edades o por sexo.

Interviene Claudia:

—¡Ya! pero... y después de hacer los cálculos ¿qué hacemos?

—Podéis escribir el informe y preparar algunas tablas y gráficos que recojan los resultados de forma más visual o más gráfica. También sería bueno que relacionarais alguna variable con otra.

—¿Cómo? ¿Qué quieres decir con eso?

—Me dijisteis que os parecía que se tomaba mucho azúcar y que había muy poca costumbre de cepillarse los dientes. Podríais tratar de establecer alguna relación estadística entre estos hábitos y la caries de los niños y niñas. Así podríais demostrar que la relación no es solo una intuición, y sería mucho más fácil obtener dinero para algún programa de educación para la salud.

Están contentos con el trabajo que han hecho. Les salió muy bien y ahora tiene curiosidad por saber exactamente qué nivel de salud hay en la zona y si se confirman o no, las impresiones que tenían mientras estaban allí.

Esto es más o menos lo que vas a aprender a hacer en esta unidad de trabajo. Vas estudiar cómo se analizan los resultados y cómo se hace el análisis estadístico de los datos, para luego poder sacar conclusiones y responder a las primeras preguntas que dieron lugar al trabajo o investigación.

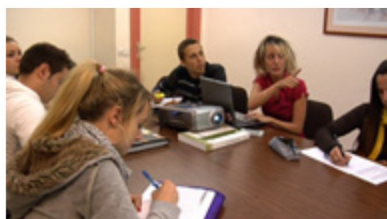


Materiales formativos de FP Online propiedad del Ministerio de Educación y Formación Profesional.

[Aviso Legal](#)

1.- Métodos de análisis de datos.

Caso práctico



Hoy ha quedado el grupo de trabajo y van a hacer una revisión de las fichas que tienen. Se han repartido el trabajo por parejas y, mientras unos empiezan a ordenar fichas, otros irán revisando la calidad de éstas. Van a hacer tres grupos con ellas, el primero con los niños y niñas de 6 años, otro grupo con los de 12 años y un último grupo, con las fichas de los adolescentes y adultos.

Después pasarán a hacer el análisis estadístico de datos, para lo que tienen que meter los datos en un programa informático que les han preparado y tienen mucha curiosidad por saber lo que van a encontrar.

Esta es la última etapa de un trabajo epidemiológico. Ya se ha hecho el trabajo de campo, es decir, se ha recogido la información (mediante exploraciones orales, cuestionarios, consultando bibliografía...), y ahora hay que analizar estos datos y sacar conclusiones.

El objetivo final de esta etapa es confirmar o no, la hipótesis planteada y responder a las preguntas hechas al inicio del trabajo de investigación.



El tipo de **análisis** depende del estudio que se haya hecho. Por ejemplo, en el caso de los estudios transversales, cuyo objetivo es describir una situación, puede bastar con hacer recuentos de datos y algún sencillo cálculo estadístico como la media aritmética y la desviación típica. Pero si se realiza un estudio de casos y controles o de cohortes, en el que se pretende establecer la relación entre un factor y una enfermedad, el análisis se irá haciendo más complejo.

Este proceso se suele comenzar con la **edición de datos**, que consiste en comprobar los datos registrados, revisar las fichas o cuestionarios en los que se han obtenido, ver si son correctos, si están completos, si son precisos... Es una primera visión de la información obtenida y es muy importante dado que en ocasiones, habrá habido distintas personas obteniendo la información que, incluso, no son propiamente los miembros del equipo investigador, por lo que es más fácil que haya confusiones.

Después, se sigue con el [procesamiento de datos](#), es decir, se ordenan y se intentan resumir o presentar de una forma más concisa. Para ello se hacen recuentos, se codifican las respuestas a cuestionarios, se tabulan en tablas de distribución de frecuencia o de contingencia (según el tipo de análisis)...

En la actualidad, prácticamente en la mayoría de los casos, se utilizan **programas informáticos** (como el EPINFO, SPSS...), en los cuales se introducen los datos y el programa realiza el proceso de análisis.

Es frecuente contar en estos análisis, con el asesoramiento de personas expertas que garantizan que el tratamiento de los datos sea el adecuado. Sin embargo, es interesante conocer los fundamentos matemáticos y epidemiológicos en los que estos programas se basan, para saber que se está haciendo y qué es lo que se puede hacer.

Para saber más

A partir de ahora nos vamos a centrar en el análisis estadístico de datos. Antes de empezar con ello, puedes consultar el siguiente enlace (u otros que tú busques), para prepararte o “situarte” un poco. A veces, leer la información proporcionada por distintas personas facilita la comprensión, y en esta unidad de trabajo, puede hacerte falta.

[Apuntes de estadística.](#)

1.1.- Análisis estadístico de datos.



Es posible que te de “pereza” empezar a pensar en “estadística”. Tal vez no se te daban bien las matemáticas, ¡no te preocupes! La estadística que vas a estudiar no es demasiado difícil, y tiene la ventaja de que una vez que aprendes a hacer algo, siempre se hace igual.

El análisis estadístico se hace para caracterizar los datos obtenidos, lo que quiere decir, describir cómo se distribuyen, cómo es su tendencia y posición en cuanto a un valor central y su variabilidad. Esto mismo, dicho de una forma más coloquial, consiste en organizar los datos y construir las tablas de frecuencias y calcular las medidas de centralización, posición y dispersión, que es lo que vamos a hacer a partir de ahora, “poco a poco”.

La **estadística** es la técnica que permite la cuantificación de los hechos o fenómenos.

Según el tipo de análisis que se haga se habla de estadística descriptiva o inferencial.

- [Estadística descriptiva](#) es aquella que se dedica a describir y a representar los datos de una población.
- [Estadística inferencial](#) es la que deduce propiedades (“infiere”) de una población a partir del estudio de una parte de ella o muestra.

Es más fácil verlo con un ejemplo. Piensa en unas elecciones municipales. Cuando se hace un sondeo los días anteriores, ¿qué se está haciendo?... y cuando se hace el recuento, ¿de qué se

trata?, ¿de estadística descriptiva o inferencial? En el primero de los casos, se está intentando averiguar qué va a pasar conociendo lo que unas pocas personas van a votar, pero en el segundo caso, ya se sabe lo que ha votado cada persona, y lo que se hace es el recuento. Por tanto, ¿qué dirías?...

El “sondeo” es estadística inferencial porque estudiando una muestra (o subgrupo), se infiere o se deduce lo que va a hacer todo el grupo. El “recuento” es estadística descriptiva, se conocen ya las votaciones y se describe lo que ha pasado (el 45% ha votado a...; el 35%..., el 10%... y el 10%...).

Autoevaluación

¿Recuerdas lo que era una "variable"? Decíamos que variable es la característica que se va a medir. En el siguiente caso: "Medir en un colegio, el número de personas que tienen manchas en el esmalte dental", ¿cuál sería la variable a estudiar?

- ☐ Número de personas.
- ☐ Colegio.
- ☐ Manchas en el esmalte.

No, te has confundido. Esto será el número de personas que tienen o no, manchas en el esmalte. Es lo que a partir de ahora vamos a empezar a llamar "frecuencia".

Te has confundido. Esto es el lugar en el que se desarrollará el estudio.

Muy bien, esta es la característica que se va a estudiar. Recuerda que una variable es lo que "varía" en el estudio.

Solución

1. Incorrecto
2. Incorrecto
3. Opción correcta

Entre los siguientes casos, ¿cuál de ellos es inferencial?

- En un ambulatorio se registran todos los cánceres orales y se pasa una información a la administración en la que se especifica la histología de la lesión y la edad de la persona afectada.
- En una consulta dental se está haciendo una biopsia de una de cada cinco lesiones precancerosa para hacer un estudio histológico de las lesiones.

No, en este caso se van a describir todos los casos que se ven, por tanto no se infiere nada.

Muy bien, en este caso estudiando uno de cada cinco casos que se presentan, se va a concluir cual es la histología más frecuente.

Solución

1. Incorrecto
2. Opción correcta

El sistema de trabajo que te proponemos en este apartado es ir estudiando los contenidos teóricos (muy pocos) según se van desarrollando unos ejercicios, que te podrán servir como modelo para resolver otros. Recuerda que aparte del cálculo, es importante que vayas comprendiendo para qué sirve lo que estás calculando, y si en algún momento la teoría te parece complicada, busca un ejercicio en el que se aplique lo explicado, porque a veces resulta más sencillo de comprender.

1.2.- Ordenación de datos.



El primer paso para analizar una información suele ser organizarla. Se suele empezar organizando el conjunto de datos, al que llamaremos n (utilizaremos en general la n minúscula, aunque también podría ser N o $\sum f_i$) en una tabla de frecuencias, estableciendo en ella, una relación entre los distintos valores que toma la variable, a la que llamaremos x_i y el número de veces o frecuencia que esto sucede.

Por tanto, lo primero que se hace es abrir una tabla para organizar los datos recogidos. Para ello hay que decidir si los valores de la variable se van a presentar solos (1, 2, 3... n) o agrupados en intervalos (de 0 a 10; de 10 a 20; de 20 a 30...).

Se entiende por [intervalo](#) (también llamado “categoría” o “clase”) cada una de las partes en las que pueden agruparse los datos. Se hacen cuando hay muchos datos o el recorrido de la variable es muy amplio y hay que simplificar el estudio.

En caso de que se opte por presentar los **datos agrupados** en intervalos hay que decidir:

- El número de intervalos que se van a hacer, lo que dependerá de la amplitud que tenga la variable. No es lo mismo una variable que toma dos valores (0 o 1) que una que oscila entre 0 y 100 (por ejemplo un porcentaje de placa). Se puede usar como criterio la raíz cuadrada del número de sujetos seleccionados en la muestra:

$$\sqrt{n}$$

Por ejemplo, si hubiera 900 personas en un estudio, los datos se podrían agrupar en 30 intervalos, aunque por regla general no se suele trabajar con más de 10-12 intervalos. Hay que recordar, que cuanto menos intervalos haya la precisión del estudio será

menor.

- Los **extremos** del intervalo, es decir, el umbral o límite inferior, L_i , y el umbral o límite superior L_s .
- Y la **longitud** o amplitud del intervalo, que es la diferencia entre los extremos: $L_s - L_i$.

Los intervalos pueden ser **abiertos** (...) o **cerrados** [...], lo que significa en el caso de los abiertos, que el valor extremo del intervalo (por ejemplo el 10 en un supuesto intervalo 10-20), no estaría incluido; pero en el caso de que fuera cerrado, si que se incluiría en el recuento de ese intervalo. Generalmente se usan cerrados a la izquierda y abiertos a la derecha (...) y cuando esto no está especificado con ningún signo, como (...) o [...], se supone que es así.

La **marca de clase** es el punto medio de cada intervalo y para su cálculo se usa la formula:

$$\frac{L_s + L_i}{2}$$

Es la cifra que se va a utilizar como representativa del intervalo.

Ejercicio resuelto

Intenta responder a estas preguntas para un estudio en el que se fuera a revisar el número de dientes con obturaciones que presentan un grupo de 400 jóvenes de 25 años, de una ciudad determinada.

1. ¿Qué variable se va a estudiar?
2. ¿Cuál es el número de casos?
3. ¿Cuál es el número de intervalos que habría que hacer?
4. En el primer intervalo, ¿cuál sería el límite inferior y el límite superior? y ¿cuál sería la marca de clase?
5. Si los intervalos son cerrados a la izquierda y abiertos a la derecha, ¿en qué intervalo incluirías el valor 4 para unos intervalos [0-4) y [4 – 8)?

Mostrar retroalimentación

1. La variable a estudiar es el número de dientes con obturaciones.
2. El número de casos es 400 ($n = 400$).
3. La variable puede oscilar entre 0 (para aquellas personas que no tengan ningún diente obturado) hasta 32 (para aquellas personas que tengan todos los dientes obturados). El número podría ser 20, o para trabajar con más comodidad 8. En este apartado hay muchas respuestas válidas, seguramente lo que has pensado estará bien.
4. Un primer intervalo podría ser de 0 a 4. (Recuerda hay más respuestas posibles como de 0 a 3, o de 0 a 6...). El límite inferior sería 0 y el límite superior sería 4. La marca de clase sería 2.
5. El 4 estaría en el intervalo 4-8. El primero $[0-4)$ es abierto a la derecha (por eso lleva un paréntesis y no un corchete) por lo que el cuatro estaría incluido en el $[4 - 8)$, que como es cerrado a la izquierda lo incluye.

1.3.- Preparación de tablas para datos no agrupados en intervalos.

Vamos a hacer un ejemplo de organización de unos datos que no están agrupados en intervalos, y que por tanto en la tabla, aparecerá cada uno de los valores que toma la variable.

Esto puede hacerse en los casos en los que estos valores diferentes no son muchos, porque en los casos contrarios en número de filas en las tablas sería tan grande, que no sería manejable.

Suponemos un caso en el que se ha estudiado a un grupo de personas de 30 a 40 años, que acuden a un gimnasio, y se ha medio el número de piezas afectadas por caries. Se han obtenido los siguientes resultados:

8	2	12	4	6	2	2	9	10	7
5	9	6	1	7	8	5	4	3	2
3	6	5	4	1	10	2	7	0	4
5	6	4	4	8	0	7	3	4	2
5	1	7	9	1	3	1	8	0	5

Vamos a preparar la tabla para ordenar estos datos. La variable a estudiar es el “número de piezas afectadas con caries” y se representa con la letra x_i . El subíndice i significa cada uno de los valores que toma la x , de forma que x_1 es el primer valor de la variable, x_2 el 2º, x_3 el 3º... hasta x_n que sería el último.

Los valores que toma la variable ordenados de menor (el 0 es la cifra menor) a mayor (el 12 es la cifra mayor en este ejercicio del gimnasio) se colocan en la primera columna, tal y como aparece en el ejemplo.

Tabla del ejercicio modelo 1: Estudio de caries en un gimnasio

x_i	
0	

1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	

1.4.- Preparación de tablas para datos agrupados en intervalos.



Vamos a hacer otro ejemplo, pero en este caso los datos están agrupados en cinco intervalos. Piensa en un ejemplo en el que se ha medido el **porcentaje de placa** a un grupo de 25 personas, tras unas sesiones de formación de cepillado, y se han encontrado los datos de la siguiente tabla:

12	23	51	6	92
68	47	24	15	95
74	65	49	72	5
15	46	1	94	36
22	87	66	68	3

En este caso y en otros que veremos más adelante, hemos puesto 25 personas, para facilitar el ejercicio. Normalmente los estudios estadísticos y epidemiológicos se hacen con mucha más gente (1.000, 2.500, 25.000...), pero por no complicar demasiado el cálculo matemático, vamos a ver los ejemplos con pocas personas.

Decidimos agrupar los datos en intervalos, ya que la variable (porcentaje de placa) puede tomar valores de 0 a 100 y una tabla con 100 filas sería incómoda. El valor más pequeño que toma la variable es 1 –no hay nadie que tenga menos placa– y el valor mayor es 95, y podemos hacer cinco intervalos, por ejemplo, de 0 a 20 (¡aunque hay otras opciones

correctas!).

Preparamos la tabla para ordenar estos datos, colocando los valores que puede tomar la variable, en la primera columna. Se colocan en intervalos como plantea el ejercicio, se calcula la marca de clase o el punto medio del intervalo, y la recogemos en la columna x_i . Este es el número que vamos a tomar como referencia y que nos va a servir para operar con los valores de ese intervalo. No podríamos trabajar con un dato de 0 a 20, por lo que utilizaremos el 10 como representativo del resto.

Tabla del ejercicio modelo 2:

Estudio de placa

	x_i
[0-20)	10
[20-40)	30
[40-60)	50
[60-80)	70
[80-100)	90

Ejercicio resuelto

Vamos a repasar algunas de las cuestiones estudiadas hasta ahora, usando el ejemplo de este apartado.

Piensa en las cuestiones y escribe las respuestas en un papel, para poderlas corregir.

1. ¿Cuáles son los límites superiores de cada uno de los intervalos del ejemplo?
2. ¿Cuál es la amplitud de estos intervalos?
3. ¿En qué intervalo incluirías el 60?
4. ¿Recuerdas algún sinónimo de “intervalo”?
5. ¿En cuántos intervalos se ha organizado esta variable?
6. ¿Cuál es el número total de casos?

Mostrar retroalimentación

1. 20, 40, 60, 80 y 100.
2. 20.
3. En el 4º, porque son abiertos a la derecha y cerrados a la izquierda.
4. Clase o categoría.
5. En cinco.
6. $n = 25$ (se va a estudiar un grupo compuesto por 25 personas).

1.5.- Frecuencias.



El segundo paso es contar cuantas veces sucede cada valor de la variable o lo que es lo mismo, calcular los distintos tipos de **frecuencias** que se suelen recoger y representar en una tabla (lo veremos en el siguiente apartado).

Las frecuencias a calcular son las siguientes:

Frecuencia absoluta (f_i o n_i): es decir el número de veces que se repite cada valor que toma la variable. Por ejemplo si estamos estudiando el tipo de colutorio usado por un grupo de 90 pacientes, y una de las opciones es colutorio fluorado, la frecuencia absoluta sería el número de personas de ese grupo, que lo usan, por ejemplo 40.

Frecuencia relativa (f_r o h_i): es el número de veces que se repite un valor en relación al número total de casos. Es el cociente entre frecuencia absoluta y el número de casos.

$$f_r = f_i / n$$

En el caso del ejemplo anterior, la frecuencia relativa sería 40 personas de cada 90, es decir $40 / 90$ o lo que es lo mismo 0,44. Es una medida útil para comparar datos, ya que si te dicen que hay 40 personas (f_i) no puedes valorar si son muchas o pocas, porque no sabes de cuántas, no siendo lo mismo 40 de 90 (que es bastante) que 40 de 90.000 (que sería muy poco).

Porcentaje (% o p_i): es la frecuencia relativa multiplicada por 100. Este es un dato de uso muy habitual, aunque a veces no sepamos de dónde sale. En el caso del ejemplo anterior, sería 0,44 multiplicado por 100, o lo que es lo mismo, hay un 44% de gente que usa colutorio

fluorado según este estudio.

Estos mismos datos se pueden recoger de forma acumulada donde se expresan sus datos y los que son menores que él. Para el cálculo los datos deben de estar ordenados de menor a mayor, por lo tanto en algunas variables cualitativas (por ejemplo “color del esmalte dental”) no tiene sentido su cálculo porque no se puede ordenar.

Se puede calcular la **frecuencia absoluta acumulada (F)**: es la suma de todas las frecuencias absolutas (f_i) de los valores menores o iguales que él; y lo mismo se puede hacer con la frecuencia relativa y porcentajes, calculándose las **frecuencias relativas acumuladas (F_r)** y los **porcentajes acumulados ($\%_a$)**.

Autoevaluación

En el siguiente ejercicio tienes que relacionar cada concepto explicado en las frases con su símbolo. La variable es el porcentaje de placa, o sea que presta atención para no confundir variables y frecuencias (número de personas, número de gente...).

Ejercicio de relacionar

Concepto	Relación Símbolo
Por cada persona, hay 0,9 personas con una placa del 90% o menos.	☐ 1. F
Hay 4 personas que tienen un 60% de placa	☐ 2. f_i
Hay 35 personas que tienen un porcentaje de placa de 90% o menos.	☐ 3. F_r
Hay un 40% de gente que tiene un 60% de placa.	☐ 4. f_r
Por cada persona hay 0,4 con un 60% de placa.	☐ 5. $\%$
Hay un 8% de personas tienen un 60% de placa o menos.	☐ 6. $\%_a$

Enviar

Es importante que te fijas en la letra que representa cada concepto, en si son mayúsculas o minúsculas y en el subíndice que llevan.

1.6.- Cálculo de frecuencias para datos no agrupados en intervalos.



En el apartado anterior hemos visto seis tipos de frecuencias, por tanto vamos a hacer seis tipos de cálculos. Para organizarse bien se suele abrir la tabla y se deja espacio libre a la derecha, de manera que según se va necesitando se van incluyendo las columnas.

Siguiendo con el primer ejemplo, ahora vamos a contar cuantas veces ha salido cada uno de los valores de la variable. Es decir, cuántas personas han tenido 0 caries, cuántas personas 1 caries, etc. Pasamos el recuento a la tabla y completamos la columna **fi**. En ella lo que decimos es que 0 caries han tenido 3 personas, 1 caries tienen 5 personas, 2 caries tienen 6 personas... hasta las 50 personas estudiadas.

Tabla del ejercicio modelo 1: Estudio de la
caries en un gimnasio

x_i f_i

0 3

1 5

2 6

3 4

4 7

5 6

6 4

7 5

x_i	f_i
8	4
9	3
10	2
11	0
12	1
$n=50$	

Después se construye el resto de la tabla con las demás frecuencias. En alguna fila está detallado el cálculo realizado.

Tabla del ejercicio modelo 1: Estudio de caries en un gimnasio

x_i	f_i	f_r	%	F	F_r	% _a
0	3	0,06 (= 3 / 50)	6 (= 3 / 50 • 100)	3	0,06	6
1	5	0,10 (= 5 / 50)	10 (= 0,10 • 100)	8 (= 3 + 5)	0,16 (= 0,06 + 0,10)	16
2	6	0,12	12	14 (= 3 + 5 + 6)	0,28	28
3	4	0,08	8	18 (...)	0,36	36
4	7	0,14	14	25	0,5	50
5	6	0,12	12	31	0,62	62
6	4	0,08	8	35	0,70	70
7	5	0,1	10	40	0,80	80
8	4	0,08	8	44	0,88	88
9	3	0,06	6	47	0,94	94
10	2	0,04	4	49	0,98	98
11	0	0	0	49	0,98	98
12	1	0,02	2	50	1	100
$n = 50$		1	100			

1.7.- Cálculo de frecuencias para datos agrupados en intervalos.



Seguimos con el ejercicio modelo 2, en el que se hacía un recuento de placa en personas a las que se les había dado un curso de cepillado. El recuento es similar al del apartado anterior (donde estudiábamos caries en un gimnasio). Se cuenta cuántas veces ha salido cada uno de los valores de la variable, es decir entre 0 y 20% de placa hay 7 personas, entre 20 y 40 hay 4 personas... hasta las 25 personas sobre las que se ha hecho el estudio.

Recuerda que los intervalos son cerrados a la izquierda y abiertos a la derecha, es decir si hubiera un valor de 20, se contaría en la 2ª fila, 40 en la 3ª, 60 en la 4ª y 80 en la última, y que cuando esto no está especificado con ningún signo (...) o [...], se supone que es así.

En este caso, como los datos están en intervalos, se necesita la marca de clase o punto medio del intervalo, como referencia de él.

Tabla del ejercicio modelo 2:

Estudio de placa

	x_i	f_i
[0-20)	10	7
[20-40)	30	4
[40-60)	50	4
[60-80)	70	6
[80-100)	90	4
	25	

Después se construye el resto de la tabla con las demás frecuencias, haciendo los cálculos igual que en el intervalo anterior.

Tabla del ejercicio modelo 2: Estudio
de placa

	x_i	f_i	f_r	%	F	F_r	% _a
0-20	10	7	0,28	28	7	0,28	28
20-40	30	4	0,16	16	11	0,44	44
40-60	50	4	0,16	16	15	0,60	60
60-80	70	6	0,24	24	21	0,84	84
80-100	90	4	0,16	16	25	1	100
		25	1	100			

Autoevaluación

Para repasar un poco las tablas, piensa qué significan las siguientes “cifras” en la tabla del ejercicio modelo 2. En los casos en los que la cifra aparece varias veces en la tabla, elige uno de los “significados” aportados.

Ejercicio de relacionar

Cifra Relación Significado

- | | | |
|----|--------------------------|---|
| 7 | ☐ | 1. Porcentaje de gente que tiene entre 80 y 100% de placa. |
| 16 | ☐ | 2. Porcentaje de gente que tiene entre 40 y 60% de placa o menos. |
| 60 | ☐ | 3. Número de personas que tienen entre 40 y 60% de placa o menos. |
| 15 | ☐ | 4. Marca de clase del último intervalo. |
| 90 | ☐ | 5. Número de personas que tienen entre 0 y 20% de placa. |

Enviar

Es importante que comprendas lo que estás calculando.

1.8.- Gráficos para la representación de datos.



Los gráficos son representaciones para expresar la frecuencia de los datos que permiten verlos de forma clara, rápida y fácil. Hay distintos tipos y hay que seleccionar el más conveniente a cada caso.

Diagrama de barras o bastones: son gráficos que asignan a cada valor de la variable, una barra o rectángulo proporcional a la frecuencia. Se representan los datos en un [sistema de ejes cartesianos](#), de manera que en el [eje de las abscisas](#) se indica el valor de la variable estadística (x_i) y en el [eje de las ordenadas](#) se señala su frecuencia absoluta (f_i).

Estos gráficos se usan en representación de caracteres cualitativos y cuantitativos discretos.

Fíjate en la imagen, en el eje de las abscisas se ha colocado los valores de la variable, que en este caso es el número de dientes con caries que tienen unas personas, y en el eje de las ordenadas las frecuencias correspondientes a cada caso.

Histogramas: Son gráficos para variables continuas que se presentan agrupadas en intervalos. Asocian a cada intervalo de valores, un rectángulo proporcional a la frecuencia. Dado que la variable es continua, y por tanto todos los valores son posibles, los rectángulos están adosados uno a otros. Veremos un ejemplo en el apartado siguiente, pero estos gráficos los viste en la primera unidad de trabajo, cuando estudiaste las pirámides de población.

Polígono de frecuencias: Es la línea que une los puntos correspondientes a las frecuencias de cada valor o de los extremos de las barras. En el caso de los datos agrupados en intervalos, une las marcas de clase.

Pictogramas: Son gráficos que utilizan imágenes alusivas a la variable, y generalmente a

través de su tamaño, se puede intuir la frecuencia. Por ejemplo si estamos estudiando prevalencia de caries en distintas ciudades, usaríamos la imagen de una muela cariada, de distintos tamaños según la prevalencia.

Diagrama de sectores: Se representa mediante sectores circulares proporcionales a las frecuencias.

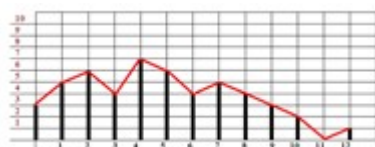


Es importante seleccionar bien el gráfico en función de lo que se pretende con él (presentar los datos bien organizados, informar de forma rápida e inmediata de las frecuencias, complementar textos escritos...).

1.9.- Representación de datos de los ejemplos en gráficos.

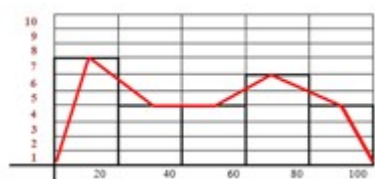
Acabas de estudiar la representación gráfica de los datos. Los ejemplos que has visto son imágenes obtenidas con un programa estadístico y ahora vamos a hacer un ejercicio de representación de datos para los dos ejemplos de esta cuarta unidad de trabajo de una forma más detallada.

En el ejercicio modelo 1, el que trataba sobre la caries, vamos a representar los datos en un diagrama de barras o de bastones y vamos a hacer su polígono de frecuencias. Se colocan los valores de la variable en el eje de abscisas y los valores de la frecuencia en el eje de ordenadas. Se dibuja una barra o bastón para cada valor de x_i que llegue hasta el valor de f_i que le corresponde. Por ejemplo, el valor 0 caries lo han tenido 3 personas, 1 caries: 5 personas, el valor 2: 6 personas, y así sucesivamente.



La **línea roja** representa el polígono de frecuencias, que se puede presentar solo o asociado a otro gráfico. Fíjate en el valor de 11 caries, nadie lo tenía, o sea la frecuencia es 0 y eso se refleja porque no hay barra y el polígono baja hasta el 0.

En el segundo ejemplo, el del recuento de placa, el tipo de gráfico que corresponde es un histograma porque los datos se presentan agrupados en intervalos o para variables continuas. Vamos a representarlos y también el polígono.



El desarrollo es igual que en el caso anterior, pero en este caso la frecuencia es de todo el intervalo. Por ejemplo, en el primero de 0 a 20, el rectángulo que representa esto sube hasta el valor 7, en el de 20 a 40 hasta 4, en el de 40 a 60 lo mismo, etc. Al igual que en el caso anterior la línea que une los puntos medios representa el polígono de frecuencias. Este ejemplo es el mismo que el gráfico de sectores del subapartado anterior, pero en este caso es más fácil averiguar cuánta gente tiene las distintas prevalencias de placa.

1.10.- Ejercicios de repaso de cálculo de frecuencias.

Vamos a hacer un “parón” y a repasar lo que has estudiado hasta ahora. Puedes hacer estos dos ejercicios y comprobar la solución en el enlace que te proponemos al final del apartado.

Ejercicio resuelto

EJERCICIO 1.- Se han contado el número de piezas ausentes de un grupo de 40 pacientes en un ambulatorio, y se han obtenido los siguientes resultados:

6, 1, 5, 2, 5, 1, 3, 6, 1, 5, 4, 2, 3

6, 6, 5, 3, 4, 3, 1, 2, 2, 5, 3, 2, 2

6, 5, 3, 4, 2, 2, 3, 5, 2, 1, 4, 2, 1

4,

- Prepara la tabla.
- Calcula la frecuencia absoluta, frecuencia relativa, porcentaje.
- Calcula los mismos datos acumulados.

Mostrar retroalimentación

Tabla con los datos de piezas
ausentes

x_i	f_i	f_r	%	F	F_r	% _a
1	6	0,15	15	6	0,15	15
2	10	0,25	25	16	0,4	40
3	7	0,175	17,5	23	0,575	57,5
4	5	0,125	12,5	28	0,7	70
5	7	0,125	17,5	35	0,875	87,5
6	5	0,125	12,5	40	1	100

x_i	f_i	f_r	%	F	F_r	% _a
40	1		100			

Ejercicio resuelto

EJERCICIO 2.- Los datos siguientes corresponden a la prevalencia de maloclusiones en un grupo de países de Europa, en chicos y chicas de 12 años:



8%, 12%, 5%, 4%, 7%, 13%, 14%, 16%, 17%, 9%, 10%, 11%, 11%, 16%, 18%, 4%, 5%, 7%, 8%, 15%, 18%, 18%, 17%, 15%, 17%, 11%, 10%, 9%, 15%, 14%, 14%, 16%, 17%, 17%, 16%, 14%, 13%, 15%, 9%, 17%

- ¿Cuántas veces ha salido el porcentaje 11 de maloclusiones?
- ¿Cuántas veces ha salido el 11% de maloclusiones o menos?
- ¿Cuál es la tasa de maloclusiones más alta que se ha encontrado?
- ¿Hay algún país con un porcentaje cero de maloclusiones? ¿Qué tasa de maloclusiones es la más baja que se ha encontrado?

Mostrar retroalimentación

En este ejercicio la tabla es similar a la de otros ejercicios, pero se presenta de otra manera. Los datos de las frecuencias están las filas y las cabeceras para los cálculos, en la primera columna.

Tabla para el ejercicio del segundo
ejercicio resuelto de cálculo de
frecuencias

4 5 7 8 9 10 11 12 13 14 15 16 17 18

f_i 2 2 2 2 3 2 3 1 2 4 4 4 6 3 40

F 2 4 6 8 11 13 16 17 19 23 27 31 37 40

a. 3 veces

b. 16 veces

c. 18%

d. No. 4%

2.- Cálculo de medidas de centralización, de posición y de dispersión.

Caso práctico



El grupo de trabajo ya ha ordenado los datos que ha obtenido de las distintas personas exploradas y ahora tienen que empezar a hacer los cálculos correspondientes para poder sacar algunas conclusiones.

En el apartado anterior se ha tratado cómo se ordenan y cómo se representan los datos, que es la primera parte o parte previa al análisis. A partir de ahora, vas a calcular medidas estadísticas que pretenden resumir la información de grupos grandes de población.

Estas medidas son útiles, precisamente por ello, por que al ser los grupos tan numerosos (10.000 personas, 100.000, etc.), de otra manera los datos no serían manejables. Con su cálculo se obtienen unas “cifras” o “valores” que nos dan información sobre la serie de datos que se está manejando.

Vamos a estudiar algunos de estos parámetros, que tal vez recuerdes de algunos estudios anteriores. De todas formas, si no te acuerdas o es nuevo para ti, no te preocupes, porque es más fácil “hacer” que “entender” los términos.

2.1.- Medidas de centralización. Media aritmética.

Estas medidas son usadas habitualmente y seguro que las conoces. Por ejemplo, imagina que has hecho dos exámenes en un módulo y has tenido un 7 y un 9. ¿Qué nota crees que te corresponde? Para responder a esto es posible que hayas calculado la “media” y digas que tienes un 8. O tal vez alguien te pregunte cuál ha sido la nota que más gente ha sacado, en este caso le responderás usando la “moda”.

Las medidas de centralización o promedios son parámetros o números que miden la tendencia central de los datos y expresan los valores centrales hacia los que los datos parecen agruparse. Son la media, la mediana y la moda.

La **media aritmética** ($\bar{x}$) es la suma de todos los valores de una variable dividida por el número de valores estudiados. La media es el parámetro más utilizado ya que tiene en cuenta todos los datos y el cálculo resulta sencillo.

Es única y no sirve para variables cualitativas ni para datos con intervalos abiertos.

Para su cálculo se utiliza la siguiente fórmula:

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 \dots + f_nx_n}{f_1 + f_2 + f_3 \dots + f_n} = \frac{\sum f_i x_i}{\sum f_i}$$

En los casos sencillos es fácil entender como se hace. Piensa por ejemplo que estamos estudiando el número de ausencias de piezas dentales en un grupo pequeño de pacientes (10 personas). Hemos obtenido los siguientes datos: 1, 3, 1, 4, 2, 5, 3, 3, 1, 5.

Si tuvieras que hacer la media de las ausencias, sumarías todos los valores y los dividirías entre los 10 pacientes estudiados, ¿no? Harías: $1 + 3 + 1 + 4 \dots$ o sea $28 / 10$. Es decir, la media de ausencias en este grupo es de 2,8.

Usando una fórmula matemática y teniendo en cuenta que hay valores que se repiten, esto se podría expresar así:

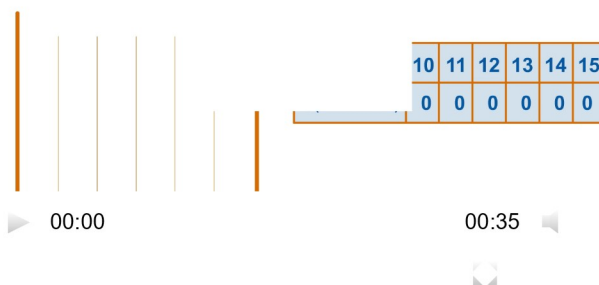
$$\bar{x} = \frac{(1 \cdot 3) + 2 + (3 \cdot 3) + 4 + (5 \cdot 2)}{10} = 2,8$$

En este caso es sencillo identificar los cuatro valores que toma la variable y los que se repiten porque son 10 personas, pero ¿si fueran 1.000 o 100.000? No sería posible, por lo que cuando hay muchos datos se necesita ordenarlos y una tabla en la que hacerlo.

Debes conocer

La siguiente animación te puede ayudar a comprender mejor la media aritmética. Fíjate en cómo se colocan las bolas agrupadas según su valor y cómo va cambiando la frecuencia. En el cálculo de la media es importante cuántas veces sucede cada valor de la variable.

$$\text{media} = \bar{x} = \frac{\sum x_i \cdot f_i}{\sum f_i} =$$



[Resumen textual alternativo](#)

Ejercicio resuelto

Un grupo de 4 personas tienen 6, 7, 3 y 3 caries respectivamente. ¿Qué valor sería en este caso $\sum f_i$ de la fórmula de la media?

Mostrar retroalimentación

4. El $\sum f_i$ es lo mismo que en otras ocasiones hemos denominado como “n” o número de casos. Es la suma de las frecuencias obtenidas por cada valor de la

variable.

2.2.- Cálculo de la media aritmética.

Para calcular la media aritmética ($\bar{x}$) cuando hay muchos datos es imprescindible organizarlos en una tabla. Vamos a seguir con los mismos ejemplos de apartados anteriores como modelo, y añadiendo columnas en la tabla según se van necesitando.

- a. Para datos sin agrupar en intervalos: Si observas la fórmula, para este cálculo se necesita la columna del producto de cada valor de la variable (x_i) por el número de veces que pasa (f_i), y es lo que hemos calculado en la última columna de la tabla siguiente. Después se suman todas las celdas de la fila para calcular el Σ , y es lo que hemos colocado en la última celda, y por último se divide por el número de casos (n o Σf_i).

Tabla del ejercicio modelo 1: Estudio de
caries en un gimnasio

x_i	f_i	f_r	%	F	F_r	% _a	$x_i f_i$
0	3	0,06	6	3	0,06	6	0 = (0 • 3)
1	5	0,10	10	8	0,16	16	5 = (1 • 5)
2	6	0,12	12	14	0,28	28	12 = (2 • 6)
3	4	0,08	8	18	0,36	36	12
4	7	0,14	14	25	0,5	50	28
5	6	0,12	12	31	0,62	62	30
6	4	0,08	8	35	0,70	70	24
7	5	0,1	10	40	0,80	80	35
8	4	0,08	8	44	0,88	88	32
9	3	0,06	6	47	0,94	94	27
10	2	0,04	4	49	0,98	98	20
11	0	0	0	49	0,98	98	0
12	1	0,02	2	50	1	100	12
n = 50		1	100				$\Sigma f_i x_i = 237$

Aplicando la fórmula para hacer los cálculos:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{237}{50} = 4,7 \text{ piezas de media afectadas por caries}$$

- b. Para datos agrupados en intervalos: Hacemos el mismo proceso y añadimos en la tabla la columna que se necesita. En este caso para operar usamos la marca de clase (x_i).

Tabla del ejercicio modelo 2: Estudio de placa

Placa	x_i	f_i	f_r	%	F	F_r	% _a	$x_i f_i$
0-20	10	7	0,28	28	7	0,28	28	70 = (10 · 7)
20-40	30	4	0,16	16	11	0,44	44	120 = (30 · 4)
40-60	50	4	0,16	16	15	0,60	60	200
60-80	70	6	0,24	24	21	0,84	84	420
80-100	90	4	0,16	16	25	1	100	360
		$\sum f_i = n = 25$	1	100				$\sum f_i x_i = 1170$

Aplicando la fórmula para hacer los cálculos:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{1170}{25} = 23,4\% \text{ de placa}$$

Es decir la media de placa en este grupo de gente es el 23,4%.

2.3.- Parámetros de centralización: moda y mediana.

La **moda** (x_0 o m_0): es el valor que más se repite en un conjunto de datos. Representa el valor dominante (el más frecuente, “lo que más se lleva”...). En su cálculo no intervienen todos los valores y tiene la ventaja de que se puede usar para valores cualitativos.

Para calcular la moda se siguen estos pasos:

Se observa la columna de la frecuencia absoluta (f_i), y aquel valor de la variable cuya f_i sea la mayor será la moda.

Cuando hay más de un valor con la misma frecuencia se dice que la distribución es bimodal, trimodal... y si todos los valores de la variable tuvieran la misma frecuencia absoluta, se dice que no hay moda.

En el caso de que los datos se presenten agrupados en intervalos, el mayor valor de f_i se corresponde con un intervalo o clase modal (en el que estará la moda), y para calcularla se usa la siguiente fórmula:

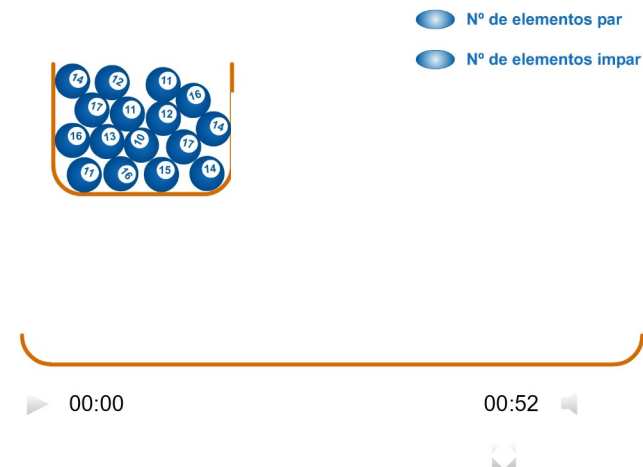
$$x_0 = L_i + c \frac{D_1}{D_1 + D_2}$$

En esta fórmula L_i es el límite inferior de la clase o intervalo modal; c la amplitud del intervalo; D_1 es la diferencia entre la frecuencia absoluta de la clase modal y de la clase inmediatamente anterior; y D_2 es la diferencia entre la frecuencia absoluta de la clase modal y de la clase siguiente.

La **mediana** es un valor de la variable, tal que el número de observaciones menores o iguales que él, es igual al número de observaciones mayores o iguales que él. O dicho de otra manera, es aquel valor, que ordenados todos de menor a mayor, se encuentra en la “mitad” o en el centro de la distribución.

Debes conocer

En la siguiente animación puedes observar cómo se colocan las bolas ordenadas de menor a mayor, y cómo se destaca la bola o las dos bolas del medio para el cálculo de la mediana. Fíjate en el caso de que los datos sean pares e impares.



[Resumen textual alternativo](#)

Es una medida única y su información complementa a la media. Es un parámetro útil para distribuciones con algún intervalo abierto o cuando hay valores muy extremos que afectan a la media.

Cuando los datos son impares hay un valor central pero cuando son pares hay dos. Fíjate en los siguientes casos, los números que están en negrita.

0, 1, 1, 3, 4, 6, 7

1, 2, 2, 2, 4, 6.

Para calcular la mediana se siguen estos pasos:

- Se ordenan los datos de menor a mayor.
- Para datos que **no** se presentan en intervalos: la mediana es el primer valor de la variable cuya frecuencia absoluta acumulada (F) es mayor de la mitad del número de datos ($n/2$). Se usa $n/2$ porque es el valor que está en la mitad.
- Para datos que **si** se presentan en **intervalos**: se determina primero **el intervalo o la clase mediana** (aquel en el que F supera a $n/2$, y por tanto contendrá a la mediana) y después se calcula el valor en concreto, mediante la fórmula:

$$Md = L_i + c \frac{\frac{n}{2} - F_{i-1}}{f_i}$$

En esta fórmula:

L_i es el límite inferior del intervalo mediano; c es la amplitud; n , el número total de datos; F_{i-1} la frecuencia absoluta acumulada de la clase anterior (por eso se escribe $i-1$) al intervalo mediano; y f_i la frecuencia absoluta de la clase mediana.

2.4.- Cálculo de la moda.

Para calcular la moda cuando los datos están sin agrupar en intervalos, vamos a usar el ejercicio modelo 1 de los apartados anteriores, en el cual los datos no están agrupados. En este caso, se busca el valor que más se repite, observando la columna de las frecuencias absolutas f_i , que ya la tenemos hecha de ejercicios anteriores, por lo que no “hacemos” nada nuevo.

La mayor cifra (en la columna f_i) es 7, que corresponde al valor de la variable 4, o sea que se puede decir, que lo más frecuente en esta distribución, es tener 4 piezas afectadas con caries. Es importante que no confundas el 7 con la moda, ya que 7 es el número de veces que ha pasado el 4.

Tabla del ejercicio modelo 1: Estudio de

caries en un gimnasio

x_i	f_i	f_r	%	F	F_r	% _a	$x_i f_i$
0	3	0,06	6	3	0,06	6	0
1	5	0,10	10	8	0,16	16	5
2	6	0,12	12	14	0,28	28	12
3	4	0,08	8	18	0,36	36	12
4	7	0,14	14	25	0,5	50	28
5	6	0,12	12	31	0,62	62	30
6	4	0,08	8	35	0,70	70	24
7	5	0,1	10	40	0,80	80	35
8	4	0,08	8	44	0,88	88	32
9	3	0,06	6	47	0,94	94	27
10	2	0,04	4	49	0,98	98	20
11	0	0	0	49	0,98	98	0
12	1	0,02	2	50	1	100	12
n = 50		1	100	$\sum f_i x_i = 237$			



En el caso de **datos agrupados en intervalos**, y usando para ello el ejemplo 2, en el que estábamos estudiando porcentaje de placa dental, de nuevo hay que buscar el valor que más se repite observando la columna de las frecuencias absolutas (f_i).

Tabla del ejercicio modelo 2: Estudio de placa

Placa	x_i	f_i	f_r	%	F	F_r	% _a	$x_i f_i$
0-20	10	7	0,28	28	7	0,28	28	70
20-40	30	4	0,16	16	11	0,44	44	120
40-60	50	4	0,16	16	15	0,60	60	200
60-80	70	6	0,24	24	21	0,84	84	420
80-100	90	4	0,16	16	25	1	100	360
		$\sum f_i = n = 25$	1	100				
						$\sum f_i x_i = 1170$		

La cifra mayor es 7, y está en el intervalo [0-20). Aplicando la fórmula para datos agrupados en intervalos:

$$L_i = 0$$

$$c = 20$$

$$D_1 = 7$$

$$D_2 = 3$$

Es decir en este caso, el valor más frecuente es un 14% de placa. Fíjate qué está comprendido en el intervalo modal. Si no estuviera es que hemos hecho mal los cálculos.

2.5.- Cálculo de la mediana para datos no agrupados en intervalos.



Para el **cálculo de la mediana** seguimos usando los mismos ejercicios modelo 1 y 2 de los apartados anteriores. En ambos casos ya tenemos los datos ordenados de menor a mayor.

En el primer caso (modelo 1), los datos **no están en intervalos**, seguimos por tanto los pasos antes citados:

Se observa la columna de las frecuencias absolutas acumuladas (F), y se busca el valor en el que F es mayor por primera vez que $n/2$. El número de casos total es 50, por tanto ha de ser F mayor que 25 ($50/2$). A esta norma hay alguna excepción que no vamos a estudiar, solo hay que recordar que ha de ser mayor, no igual ni menor).

Eso sucede en el valor $F = 31$, por que 25 no es mayor.

El valor de la variable que tiene una frecuencia absoluta acumulada de 31 es 5. Esto significa que por debajo de 5 caries, está la mitad de la gente, o sea 25 personas, y por encima, están otras 25, que tienen más de 5 caries.

Tabla del ejercicio modelo 1: Estudio de
caries en un gimnasio

x_i	f_i	f_r	%	F	F_r	% _a	$x_i f_i$
0	3	0,06	6	3	0,06	6	0
1	5	0,10	10	8	0,16	16	5
2	6	0,12	12	14	0,28	28	12
3	4	0,08	8	18	0,36	36	12
4	7	0,14	14	25	0,5	50	28

x_i	f_i	f_r	%	F	F_r	% _a	$x_i f_i$
5	6	0,12	12	31	0,62	62	30
6	4	0,08	8	35	0,70	70	24
7	5	0,1	10	40	0,80	80	35
8	4	0,08	8	44	0,88	88	32
9	3	0,06	6	47	0,94	94	27
10	2	0,04	4	49	0,98	98	20
11	0	0	0	49	0,98	98	0
12	1	0,02	2	50	1	100	12
$n = 50$		1	100	$\sum f_i x_i = 237$			

Para saber más

En el siguiente enlace puedes repasar algunos de los contenidos vistos hasta ahora. Hemos seleccionado el apartado de la mediana, pero puedes revisar otros que ya has estudiado.

[Mediana.](#)

Autoevaluación

Vamos a repasar la información que viene en la tabla del ejercicio modelo 1. Escribe las respuestas en los huecos.

- ¿Qué variable se está estudiando en este ejercicio?
- ¿Qué valores puede tomar esta variable? Desde hasta
- ¿A cuántas personas se ha estudiado?
- ¿Qué porcentaje de gente ha tenido 3 caries? %
- ¿Cuántas personas no tienen ninguna caries? ¿Qué porcentaje representan? % del grupo.

Enviar

2.6.- Cálculo de la mediana para datos agrupados en intervalos.

El cálculo de la mediana para datos agrupados en intervalos, lo vamos a ver con esta tabla del ejercicio modelo 2, en el que estábamos estudiando el porcentaje de placa.

Tabla del ejercicio modelo 2: Estudio de
placa

Placa	x_i	f_i	f_r	%	F	F_r	% _a	$x_i f_i$
0-20	10	7	0,28	28	7	0,28	28	70
20-40	30	4	0,16	16	11	0,44	44	120
40-60	50	4	0,16	16	15	0,60	60	200
60-80	70	6	0,24	24	21	0,84	84	420
80-100	90	4	0,16	16	25	1	100	360
		25	1	100				1170

- Lo primero es ordenar los datos de menor a mayor que ya está hecho en la tabla.
- En segundo lugar, hay que comprobar para que valor F es mayor que $n/2 = 12,5$. Eso sucede en el 16, que corresponde al intervalo 40-60, en el cual estará la mediana.
- Aplicando la fórmula para calcular el valor exacto:

$$Md = Li + c \cdot \frac{\frac{n}{2} - F_{i-1}}{f_i} = 40 + 20 \cdot \frac{25/2 - 11}{16} = 41,875$$

- Es decir, la mitad de la gente tiene menos de 41,875% de placa y hay otras tantas personas, que tienen por encima de este porcentaje.

Debes conocer

En la siguiente animación puedes observar de forma muy esquemática y sencilla los parámetros que hemos visto hasta ahora. Pincha en cada una de las tres palabras y fíjate en cómo son conceptos diferentes.

media**mediana****moda**

1 4
2 8 2

▶ 00:00

00:37 ◀

[Resumen textual alternativo](#)

Autoevaluación

Si tuvieras que relacionar cada frase con un parámetro de centralización, ¿cómo lo harías?

Tabla para relacionar un parámetro estadístico y una frase

Frase	Relación Parámetro	
Se ha visto que por debajo de 6 puntos está la mitad de la población.	☐	1. Mediana
En una serie de datos el valor más frecuente es 5 puntos.	☐	2. Media
La mayoría de los datos se aproximan a 5,4 puntos.	☐	3. Moda

Enviar

Las medidas de centralización son parecidas, pero cada una aporta una información diferente.

2.7.- Medidas de posición y percentiles.

Mientras que las medidas o parámetros de centralización señalan los valores del centro hacia los que tiende la distribución, estas medidas de posición permiten caracterizar las series o distribuciones de datos según otras características diferentes a los valores centrales.

Los más usados son los **cuartiles**, Q_1 , Q_2 y Q_3 , que son tres puntos, que dividen la distribución en cuatro partes iguales (¡fíjate en la imagen!); **deciles** o nueve puntos que dividen la distribución en 10 partes iguales (d_1 , d_2 ... d_9) y **percentiles** (p_1 , p_2 , p_3 ..., p_{99}), que son 99 puntos o parámetros que dividen la distribución en cien partes iguales.



En el caso de temas de salud, se usan mucho los percentiles y son los únicos que vamos a estudiar en este apartado.

Los **percentiles** son 99 puntos que dividen una distribución de datos en 100 partes iguales.

Veamos lo primero qué significa este parámetro estadístico antes de aprender a calcularlo. Por ejemplo, si una persona está en el percentil 25 en cuanto al peso, significa que hay un 25% de gente que pesa menos que ella y un 75% que pesa más que ella. Si alguien está el percentil 80 en cuanto a talla, quiere decir que hay un 80% de gente que está por debajo y un 20% que está por encima, o sea es una persona bastante alta en su grupo de población.

Es decir, cuando calculamos el valor de estos parámetros, sabemos qué parte (porcentaje en el caso de los percentiles) de los datos están por encima y qué parte está por debajo (es mayor o menor). Es la misma idea que para la mediana (“en la mitad”), por lo que se calculan de modo similar. Siempre se representan los menores valores de la variable a la izquierda y los valores más altos a la derecha.



Autoevaluación

Piensa para los siguientes casos qué percentil habría que calcular. Igual te conviene hacer un "dibujo" similar a la imagen del apartado.



Se quiere seleccionar al 80% de personas que tienen más placa dental de un grupo de gente que ha estado incluida en una investigación de salud oral, para darles próximamente, un curso de cepillado. ¿Qué percentil habría que calcular?

- ☐ P_{80} .
- ☐ P_{20} .

Muy bien. Has entendido bien el cocepto.

Te has confundido. Este parámetro habría que calcularlo si se cogiera a los que menos placa tienen. Igual te conviene hacer un dibujo y ver de dónde se empieza a seleccionar a la gente para coger a los de más placa.

Solución

1. Opción correcta
2. Incorrecto

Se quiere seleccionar el 30% de los niños de un grupo con un índice de restauración más alto. ¿Qué parámetro hay que calcular?

- ☐ P_{30} .

- P70.

No es éste, ya que éste sería en el caso de que quieras seleccionar el 30% de niños y niñas con el índice de restauración más bajo. Para seleccionar al grupo con IR más alto empiezo a coger desde la derecha.

Muy bien.

Solución

1. Incorrecto
2. Opción correcta

2.8.- Cálculo de percentiles para datos no agrupados en intervalos.

Para el cálculo en el caso de datos que **no están en intervalos**, el percentil que se busca, será el primer valor de la variable cuya frecuencia absoluta acumulada (F) es mayor que un valor que depende del percentil que queramos calcular. Por ejemplo:

- Para el P_{25} : $(n/100) \cdot 25$
- Para el P_{75} : $(n/100) \cdot 75$

Si te fijas, esto quiere decir que se divide el número de casos (n) entre 100 porque son percentiles (100 trozos) y se multiplica por el número de partes que queremos coger (En los dos casos de antes, P_{25} : 25 partes y P_{75} : 75 partes; si fuera P_{30} : 30 partes...).

Vamos a suponer que tenemos la serie de datos sobre la caries del ejercicio del modelo 1. Imagina que queremos seleccionar al 30% de la gente que menos caries tiene, ¿qué parámetro tenemos que calcular?



Tenemos que buscar un valor que deje por debajo (o a la izquierda en la representación) al 30% de la gente y por tanto encima (o sea a la derecha en la representación) quedaría un 70%, o sea el P_{30} .

Para el **cálculo del percentil 30** seguimos los pasos señalados en el apartado anterior:

- F debe ser mayor de:

$$\frac{50}{100} \cdot 30 = 15$$

- Observando en la tabla la columna F, es 18 y eso corresponde al valor 3.
- Es decir, hay un 30% de personas que tienen menos de 3 caries y hay un 70% que tienen más, y respondiendo a la pregunta de una forma más precisa.

Tabla del ejercicio modelo 1: Estudio
de caries en un gimnasio

x_i	f_i	f_r	%	F	F_r	% _a	$x_i f_i$
0	3	0,06	6	3	0,06	6	0
1	5	0,10	10	8	0,16	16	5
2	6	0,12	12	14	0,28	28	12
3	4	0,08	8	18	0,36	36	12
4	7	0,14	14	25	0,5	50	28
5	6	0,12	12	31	0,62	62	30
6	4	0,08	8	35	0,70	70	24
7	5	0,1	10	40	0,80	80	35
8	4	0,08	8	44	0,88	88	32
9	3	0,06	6	47	0,94	94	27
10	2	0,04	4	49	0,98	98	20
11	0	0	0	49	0,98	98	0
12	1	0,02	2	50	1	100	12
n = 50		1	100				237

2.9.- Cálculo de percentiles para datos agrupados en intervalos.

Para su cálculo en el caso de datos que **están en intervalos**, se determina primero la clase o intervalo en el que se encuentra el percentil que se busca, de la misma forma que en el caso anterior (aquel en el que F supera a...) y después se calcula el valor en concreto, mediante la fórmula:

$$P_{75} = L_i + \frac{\frac{n}{100} \cdot 75 - F_{i-1}}{f_i} \times c$$

En esta fórmula L_i es el Límite inferior de la clase o intervalo; c la [amplitud del intervalo](#); N o n es el número total de datos; $F_{(i-1)}$ la frecuencia absoluta acumulada de la clase anterior a la clase; f_i la frecuencia absoluta de la clase.



Supón ahora que vamos a iniciar unas sesiones de algún tratamiento especial para la placa, y en este caso queremos al 20% de gente que peor tiene la boca, ¿qué parámetro necesitamos calcular? Tenemos que buscar un valor que deje por debajo o a la izquierda al 80% de los datos y el resto, 20%, por encima.

Para el **cálculo del percentil 80**, procedemos de la misma manera.

- F debe ser mayor que:

$$\frac{25}{100} \cdot 80 = 20$$

- Observando en la tabla la columna F, eso sucede con el número 21, que corresponde al intervalo 60-80.
- Aplicando la fórmula para calcular el valor: $L_i = 60$ $c = 20$ $F_{i-1} = 15$, $f_i = 6$, obteniendo 76.6 de resultado.

Es decir, seleccionaremos para el curso a la gente con más del 76,6% de placa.

Tabla del ejercicio modelo 2: Estudio de placa

Placa	x_i	f_i	f_r	%	F	F_r	% _a	$x_i f_i$
0-20	10	7	0,28	28	7	0,28	28	70
20-40	30	4	0,16	16	11	0,44	44	120
40-60	50	4	0,16	16	15	0,60	60	200
60-80	70	6	0,24	24	21	0,84	84	420
80-100	90	4	0,16	16	25	1	100	360
		$\sum f_i = n = 25$	1	100				$\sum f_i x_i = 1170$

Autoevaluación

Se quiere seleccionar a determinados grupos de personas que han participado en estudios epidemiológicos. Relaciona en la tabla para cada caso el parámetro que se usaría.

Ejercicio de relacionar

Caso	Relación	Parámetro
Se ha hecho un curso de cepillado y se va coger al 30% de personas con la boca más limpia para reforzar el comportamiento.	☐	1. P_{25}
Se sabe que en una población hay 1000 habitantes y se quiere coger al 25% con más caries para hacer un tratamiento experimental.	☐	2. P_{30}
Para seleccionar de un grupo, el 25% de gente que menos se cepilla.	☐	3. P_{85}
Para seleccionar de un grupo, el 15% que más usan colutorio tras el cepillado.	☐	4. P_{75}

Enviar

Es importante que te fijas en el texto, y si no te coincide la respuesta haz un dibujo y representa lo que te dice.

2.10.- Medidas de dispersión y coeficiente de variación.

Estos parámetros son los que informan de cómo se concentran o cómo se separan los datos de los valores centrales o parámetros de centralización.

Estos parámetros que expresan la mayor o menor concentración de los datos con respecto a las medidas de centralización, son medidas fundamentales para caracterizar una serie de datos, Imagínate, que has hecho dos exámenes y has tenido una media de 5. Ha podido ser un 10 y un 0 o ha podido ser un 6 y un 4, ¿sería lo mismo?

Reflexiona

Las medidas de centralización siempre han de ir acompañadas de las medidas de dispersión. Decir solo la media aritmética de una distribución es decir “muy poco”.



Hay bastantes parámetros pero sólo vamos a estudiar el recorrido o rango y la desviación típica que se usan bastante. La varianza sólo la presentamos como un paso intermedio para la obtención de la desviación típica, y vamos a intentar usar solo una de las distintas fórmulas de cálculo.

Amplitud, rango o recorrido (R): Es la diferencia entre el mayor y el menor valor que toma la variable estadística (observaciones extremas). Sirve para valorar entre que valores se mueve la distribución.

Varianza (S^2 , σ^2 o sigma^2): Es un parámetro que mide la distancia existente entre los valores de la serie y la media. Es la media aritmética de los cuadrados de las desviaciones respecto a la media. La varianza siempre será mayor que cero. Cuanto más se aproxima a cero, más

concentrados están los valores de la serie alrededor de la media, y por el contrario, mientras mayor sea la varianza, más dispersos están.

$$varianza = s^2 = \frac{\sum f_i x_i^2}{\sum f_i} - \bar{x}^2$$

Desviación típica o estándar (s; S o sigma): Es la raíz cuadrada positiva de la varianza. Es una medida que expresa, en las mismas unidades de la variable (puntos, porcentajes, kilos, metros...) lo que se separan los datos de los valores centrales. Al igual que en el caso de la varianza, cuando más pequeña es más próximos están los datos al valor de la media. Para su cálculo vamos a utilizar la fórmula más sencilla.

$$Desviacion\ típica = s = \sqrt{s^2} = \sqrt{\frac{\sum f_i x_i^2}{\sum f_i} - \bar{x}^2}$$

El **coeficiente de variación (CV)**, que se define como la relación existente entre la desviación típica y la media. Expresa la variabilidad de los datos de una distribución y se dice que cuanto más pequeño sea, más útil será la distribución para trabajar con ella, e incluso cuando es mayor de un 33%, se considera que la distribución es poco homogénea, poco uniforme, muy dispersa y poco útil.

$$CV = \frac{s}{\bar{x}} k = \dots \%$$

Autoevaluación

Señala de estas dos afirmaciones la que consideras que es más cierta.

- ☐ Las medidas de centralización son mucho más útiles que las medidas de dispersión.
- ☐ Las medidas de dispersión y las medidas de centralización se deben utilizar conjuntamente.

No es cierta, aunque con frecuencia se utilizan solas.

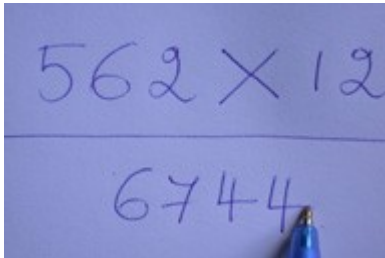
Recuerda el ejemplo de la nota de dos exámenes del apartado. Es importante usar

ambas para saber si la media que manejamos es representativa o no.

Solución

1. Incorrecto
2. Opción correcta

2.11.- Cálculo de medidas de dispersión y del coeficiente de variación.



Vamos a calcular estos parámetros estadísticos en el caso del ejemplo del cálculo del porcentaje de placa en las 25 personas.

Usamos la tabla que ya tenemos hecha, aunque algunas de las columnas ahora no se necesiten:

Tabla del ejercicio modelo 2: Estudio de placa

Placa	x_i	f_i	f_r	%	F	F_r	% _a	$x_i f_i$	$f_i x_i^2$
0-20	10	7	0,28	28	7	0,28	28	70	700
20-40	30	4	0,16	16	11	0,44	44	120	3600
40-60	50	4	0,16	16	15	0,60	60	200	10000
60-80	70	6	0,24	24	21	0,84	84	420	29400
80-100	90	4	0,16	16	25	1	100	360	32400
		25	1	100				$\sum f_i x_i = 1170$	$\sum f_i x_i^2 = 76.100$

La **amplitud o rango** es la diferencia entre el menor valor de la variable y el mayor, por tanto en este caso, y tomando la marca de clase como referencia:

Amplitud o rango = $90 - 10 = 80$ puntos.

Para el cálculo de la desviación típica se calcula primero la varianza, para lo que se aplica la fórmula:

$$Varianza = S^2 = \frac{f_i \cdot x_i^2}{f_i} - \bar{x}^2 = \frac{76100}{25} - 23,4^2 = 3044 - 547,56 = 2.496,44$$

$$desviación\ típica = s = \sqrt{2.496,44} = \pm 49.96$$

Para el cálculo del coeficiente, buscamos la media que ya la teníamos calculada en el apartado 2.2.

$$Cv = \frac{s}{\bar{x}} \cdot k = \frac{49,96}{23,4} \cdot 100 = 2,13\%$$

Es decir es una distribución uniforme, con los datos poco dispersos, con la que se puede trabajar.

Recomendación

Es importante poner siempre la unidad en el resultado hallado para saber de qué se está hablando. En los casos como la media, mediana, percentiles... fíjate en la variable y según esté esta expresada (% , kg, puntos...) puedes hacerlo.

Otros casos, como el coeficiente de variación, casi siempre se expresa como un porcentaje.

Por último, cuando la pregunta hace referencia a número de personas, como en el caso de las frecuencias, no te confundas porque en este caso son personas.

2.12.- Utilización conjunta de la desviación típica y de la media aritmética.

Hay que recordar que la media aritmética sin la desviación típica no informa de nada, y que es importante siempre saber cómo están distribuidos los datos para poder sacar conclusiones.

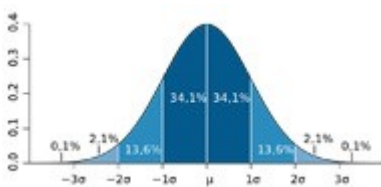
Por ejemplo, si a una persona le faltan 0 dientes y a otra le faltan 10, la media de ausencias es 5, y en otro caso, si a una le faltan 6 y a la otra persona le faltan 4, la media también es 5. Sin embargo, en este segundo caso, el 5 representa mejor a las dos personas que en el primer caso, que realmente no representa a ninguna.

Por ello se utilizan las dos medidas conjuntamente para dar la información. Se suele usar una expresión como las siguientes:

- En un pueblo A, la media aritmética del índice periodontal de la comunidad (CPI) es del 4.85 con una desviación típica de $\pm 0,12$.
- En otro pueblo B, la media del índice es índice periodontal de la comunidad (CPI) del 4,85 pero con una desviación típica de $\pm 2,02$.

Esto nos indica que en el primero de los casos, en el pueblo A, el CPI calculado representa mejor a la mayoría de la gente, y por tanto se puede sacar mejores conclusiones de ese dato. (Es decir, si calculamos el coeficiente de variación en el segundo caso es mayor).

Pero además de esto, el uso conjunto de las dos medidas también sirve para saber cómo se distribuyen los datos en la distribución, ya que la mayoría de las series de datos, y en especial las constantes biológicas, tienden a tener un patrón muy parecido, que consiste en pocos datos en los extremos y más en el centro -distribución normal o curva de Gauss-.



Fíjate en el gráfico que representa una distribución normal. En él se puede ver como la mayoría de los datos están en el centro, y como hay menos en los valores extremos. En

concreto, se sabe que el intervalo comprendido entre:

- $\bar{x}-1s$ y $\bar{x}+1s$: es probable que estén aproximadamente el 68% de los datos.
- $\bar{x}-2s$ y $\bar{x}+2s$: es probable que estén aproximadamente el 95% de los datos.
- $\bar{x}-3s$ y $\bar{x}+3s$: es probable que estén aproximadamente el 99% de los datos (casi el 100%).

Autoevaluación

Supón una población en la que se sabe que la media aritmética del porcentaje de superficies con placa es del 20% con una desviación típica de ± 5 . Cumplimenta el siguiente párrafo con lo que has aprendido en el último apartado.

El 68% de los datos estaría entre y % de placa. El 95% de los datos estaría entre y % de placa y el 99% de los datos estaría entre y % de placa.

Enviar

2.13.- Distribuciones bidimensionales.

Estas series de datos son diferentes de las que hemos estudiado hasta ahora. No vamos a trabajar mucho con ellas, pero como están muy relacionadas con los estudios ecológicos que estudiamos en la primera unidad de trabajo, vamos a presentar algo sobre ellas.

Imagina un grupo de grupo de personas de las que se conoce el peso y la talla de cada una de ellas, o unas personas de las que se sabe el índice de enfermedad periodontal y la frecuencia diaria de cepillados. En ambos casos hay personas (n) y dos variables para cada una (x_i e y_j). Sobre cada una de ellas se podría hacer un análisis estadístico separado, y así, estudiar la variable “índice de enfermedad periodontal” y también estudiar la variable “frecuencia de cepillado”, pero en este caso de las distribuciones bidimensionales, sabemos además también qué índice de enfermedad periodontal coincide en cada individuo con qué frecuencia de cepillado.

Distribuciones bidimensionales son aquellas distribuciones en las que intervienen dos conjuntos de datos diferentes (x_i e y_j), variando ambos a la vez. En ellas, cada elemento de la distribución (individuos, muestras, personas...) está definido por dos valores.



A veces entre estas dos variables (por ejemplo tasa de caries y tasa de consumo de azúcar) se establece un tipo de relación, de forma que una de ellas –variable dependiente– depende o cambia con la otra –variable independiente– (es probable que a más azúcar, más caries). Esto se denomina **CORRELACIÓN** y permite estudiar, en términos de probabilidad, lo que sucede entre ambas variables.

Se puede estudiar varios aspectos, y aunque no vamos a hacer ningún cálculo, vamos a concretarlos en tres cuestiones para facilitar las cosas:

1. El **tipo de relación** que se da entre ellas, es decir si a aumentos de una de las variables corresponden aumentos o disminuciones de la otra.

Para medir esto se suele calcular la **covarianza** y el **coeficiente de correlación (o de Pearson)** que mide la fuerza de asociación entre las dos variables.

- Si es positivo la relación es **directa**, es decir, cuando aumenta una de ellas (independiente) aumenta la otra (dependiente).
 - Si es negativo la relación es **inversa**, o sea, cuando aumenta la variable independiente, disminuye la variable dependiente.
2. El **nivel de dependencia** o cuánto depende una de otra, lo que es muy útil para programas de educación para la salud, porque tiene más sentido desarrollarlos sobre la dependencia fuerte.
 - Si el valor del coeficiente es cercano a $[-1]$ y $[+1]$ se dice que la correlación es **fuerte**, o sea depende mucho una de la otra.
 - Si el valor del coeficiente es cercano a 0, la correlación es **débil** o depende poco.
 3. El **cálculo del valor de la variable dependiente** en función de un valor de la variable independiente. Por ejemplo, sabiendo la tasa de consumo de azúcar en un país, se puede calcular la tasa probable de caries que le corresponde. En este caso, cuanto más fuerte es la correlación (más depende una variable de la otra), el cálculo realizado se acercará más a la realidad (es más probable acertar).

Autoevaluación

Intenta establecer el tipo de correlación que se da entre estas variables.

Correlación entre el consumo de azúcar y la tasa de caries

- Directa.
- Inversa.

Muy bien, La tasa de caries aumenta cuando aumentan las tasas de consumo de azúcar.

No, te has confundido.

Solución

1. Opción correcta
2. Incorrecto

Correlación entre consumo de flúor y caries.

- ☐ Directa.
- ☐ Inversa.

No, la caries no aumenta cuando aumenta el uso de flúor.

Muy bien, la caries disminuye con el consumo de flúor. Por cierto, algunas personas dicen que esta asociación es muy fuerte, y que la caries, depende mucho de esta variable.

Solución

1. Incorrecto
2. Opción correcta

Correlación entre tabaco y cáncer oral.

- ☐ Directa.
- ☐ Inversa.

Muy bien. El cáncer oral depende del tabaco, de forma que cuanto más tabaco se consume, aumenta la incidencia de cáncer oral. De hecho, la incidencia de cáncer

oral está aumentando en las mujeres, debido al aumento del hábito tabáquico.

¿Seguro? Repasa los apuntes.

Solución

1. Opción correcta
2. Incorrecto

2.14.- Ejercicios de repaso de cálculo de parámetros.

Para terminar este apartado de estadística, puedes resolver los siguientes ejercicios y consultar el resultado en el enlace propuesto.

EJERCICIO 1.- En un pueblo se ha medido el **índice de restauración –IR–** en una muestra de 80 personas y se han obtenido los resultados siguientes:

0%-10%: 0; 10%-20%: 2; 20%-30%: 5;
30%-40%: 9; 40%-50%: 13; 50%-60%:
19;
60%-70%: 11; 70%-80%: 13; 80%-90%:
5;
90%-100%: 3



Ordena los datos en una tabla y calcula las frecuencias.

- ¿Cuánta gente ha tenido un índice de restauración entre 80 y 90%?
- ¿Cuánta gente ha tenido un IR entre 80 y 90% o menos?
- ¿Qué porcentaje de personas han tenido un IR de entre 0 y 10%?
- ¿Qué porcentaje de gente ha tenido un IR de 50% o menos?
- Representa los datos en un gráfico.
- ¿Qué media de índice de restauración corresponde a este pueblo?
- ¿Qué valor ha sido el más frecuente? Calcúlalo.
- Se va a seleccionar al 40% de la gente que menos tratada tiene la boca para dar un curso de prevención. ¿Con que parámetro se haría? Calcúlalo.
- Se quiere seleccionar también a un 10% de las personas con el IR más alto, para hacer una revisión y un seguimiento durante unos años. ¿Qué parámetro se usaría?
- Calcula la desviación típica y relaciónala con la media aritmética.
- ¿Es esta una distribución homogénea?
- ¿Entre qué dos valores se encontrarán probablemente el 65 % de los datos?

EJERCICIO 2.- En una comunidad autónoma se ha hecho un estudio ecológico y se ha

estudiado la **correlación** entre el índice de caries y el consumo de bebidas azucaradas en distintos pueblos. Se ha visto que el coeficiente de correlación entre ambas variables es de 0,97. ¿Cómo es esta correlación? ¿Directa o inversa, fuerte o débil...? ¿Qué significa esto?

En el caso de que el coeficiente de correlación entre el índice de caries y la ausencia de cepillado fuera de 0,17, ¿qué significaría? Analiza los dos resultados y compáralos sobre la variable dependiente.

Debes conocer

Puedes consultar la solución a estos ejercicios en el Anexo I de este documento.

Para saber más

Puedes repasar estos contenidos en una de las páginas relacionadas con estadística, como la que a continuación te proponemos.

[Estadística.](#)

3.- Interpretación de resultados.

Caso práctico



El equipo de **Claudia** y **Ramón** ya ha terminado de hacer el análisis estadístico de los datos, han calculado medias aritméticas, desviaciones típicas, rangos o recorridos, porcentajes... y ahora tienen que interpretar los datos que han obtenido.

Es el momento de pensar en lo que estos “números” quieren decir, y concretar el índice de caries, de enfermedad periodontal o de ausencias que hay en la población estudiada, la distribución de estos datos según edades o sexo, comparar estos datos con otros y sacar conclusiones sobre si son adecuados o no, relacionar el consumo de azúcares y la caries, etc.



Imagínate que en el cálculo del ICAOD han obtenido una media de 6,4 con una desviación típica de $\pm 0,5$. La mayoría de la gente no sabría que pensar. El dato o la cifra en sí no dice nada, no se sabe si es mucha caries o poca, si está bien o mal...

Es el momento de valorar los números obtenidos, y desde luego, hace falta tener formación

específica sobre el tema, para poder analizar otros datos anteriores de la zona y deducir si está aumentando o disminuyendo, revisar datos de otras zonas o poblaciones similares y poder valorar las cifras como convenientes o excesivas, intuir posibles factores (o “variables”) asociados...

Se suele decir que “los datos no hablan solos”, hacen falta personas que los interpreten. Algunos dicen que es el momento de “disfrute” del equipo investigador, porque van a encontrar la respuesta a la pregunta planteada, y en caso de no encontrarla, deberán plantear una nueva pregunta para reiniciar el proceso (acuérdate de la segunda unidad de trabajo, en ella se hablaba de las fases de un trabajo epidemiológico).

La interpretación de los resultados pretende responder a las preguntas planteadas en el inicio de la investigación a través del análisis y estudio de los datos obtenidos y procesados.

En este momento, los datos ya están ordenados y se han realizado los cálculos necesarios, hay por tanto, que sacar conclusiones como las siguientes:

- Describir las características de la población estudiada.
- Expresar la medición de algunos índices o de algunos hábitos.
- Establecer causas para algunos fenómenos.
- Determinar la efectividad de tratamientos y medidas de prevención.
- Relacionar un factor de protección y una enfermedad.
- Medir el incremento de riesgo que aporta un factor específico.
- Etc.

Así, si se quería ver la fuerza de asociación entre el tabaco y la enfermedad periodontal, habremos hallado un número que habrá que interpretar; o si se trataba de describir la prevalencia de bolsas periodontales en mujeres embarazadas, habrá que dar una tasa y unas características de esta distribución... hay muchas respuestas posibles, pero siempre en relación con el objetivo que se hubiera planteado.

3.1.- Errores y sesgos.



Una de las “pesadillas” de los equipos investigadores, son los errores y sesgos en la investigación, o dicho de manera más coloquial, que alguien demuestre que su trabajo y sus conclusiones no se sustentan en planteamientos correctos, que no han acertado o que se han “confundido”.

Por esto presentamos aquí este apartado, que aunque no tienes nada que estudiar en él, debes reconocer estos términos que todos los investigadores e investigadoras, incluyen en sus trabajos de forma que demuestran que han cuantificado en lo posible, el margen de error y sus planteamientos son correctos.

Al realizar un estudio es posible cometer distintos errores, y aunque no hay métodos que permitan evitar por completo estos errores, si se preparan bien las cosas es posible reducirlos al mínimo. No hay que olvidar, que en estos trabajos de epidemiología se trabaja en base a probabilidades, por tanto es “probable” (o bastante posible), que los datos obtenidos sean esos pero no siempre se sabe, ya que cuando se trabaja con muestras de grupos se está “infiriendo” datos.

Las causas de los errores pueden ser múltiples:

- Una muestra mal seleccionada por que es poco representativa, o porque es demasiado pequeña...
- Pérdidas de personas por abandonos o fallecimientos durante la investigación...
- Información de poca calidad por que se han recogido mal los datos o por mal entrenamiento de las personas que los toman...
- Etc.

La ventaja de trabajar con un método científico es que el nivel de algunos errores se puede

minimizar o al menos, se puede cuantificar a través de distintos sistemas (como la determinación del error típico o estándar, las pruebas de significación estadística como la “T” de Student, el cálculo del Chi cuadrado, el cálculo de intervalos de confianza, las pruebas de hipótesis...) lo que supone, de alguna manera, paliar su consecuencias.

Para saber más

Los contenidos que has estudiado en esta cuarta unidad de trabajo, se usan en los trabajos epidemiológicos que se hacen, y te permitirán en el futuro, comprender mejor y ser más crítico con la información que manejes cuando estés trabajando.

Busca la encuesta de salud oral del año 2005 que has ido leyendo en las unidades de trabajo anteriores. Lee el apartado 4.5, en el que describe el análisis estadístico. Fíjate en los programas informáticos que menciona, en los parámetros y test estadísticos calculados, ¿te suena alguno?

[Encuesta de Salud Oral en España 2005.](#) (0.87 MB)

Lee ahora la encuesta de salud oral del 2015 que también has leído en unidades anteriores, y observa la similitud entre ambas.

[Encuesta de Salud Oral en España 2015](#)

Anexo I.- Ejercicios resueltos de cálculo de parámetros.

En este documento puedes revisar lo que has hecho. Hemos optado por poner la tabla al principio y hemos ido añadiendo las columnas según se han ido necesitando. Acuérdate de poner las unidades, que en este caso del IR son porcentajes (%), no lo confundas con la columna de porcentajes e intenta explicar los resultados.

1. Ordena los datos en una tabla y calcula las frecuencias.

- Serían las seis columnas: x_i , f_i , f_r %, F , F_r y %_a de la tabla.

Índice de restauración

	x_i	f_i	f_r	%	F	F_r	% _a	$x_i f_i$	$f_i x_i^2$
0-10	5	0	0	0	0	0	0	0	0
10-20	15	2	0,025	2,5	2	0,025	2,5	30	450
20-30	25	5	0,062	6,2	7	0,087	8,7	125	3.125
30-40	35	9	0,112	11,2	16	0,199	19,9	315	11.025
40-50	45	13	0,162	16,2	29	0,361	36,1	585	26.325
50-60	55	19	0,237	23,7	48	0,598	59,8	1.045	57.475
60-70	65	11	0,137	13,7	59	0,735	73,5	715	46.475
70-80	75	13	0,162	16,2	72	0,897	89,7	975	73.125
80-90	85	5	0,062	6,2	77	0,959	95,9	425	36.125
90-100	95	3	0,037	3,7	80	0,996 (aprox = 1)	99,6 (aprox = 100)	285	27.075
	80	1		100					

2. ¿Cuánta gente ha tenido un IR entre 80 y 90%?

- 5 personas.

3. ¿Cuánta gente ha tenido un IR entre 80 y 90% o menos?

- 77 personas.

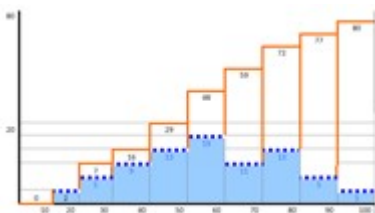
4. ¿Qué porcentaje de personas han tenido un IR de entre 0 y 10%?

- Nadie, 0%.

5. ¿Qué porcentaje de gente ha tenido un IR de 50% o menos?

- 36,1%.

6. Representa los datos de la frecuencia absoluta y de la frecuencia absoluta acumulada en un gráfico.



7. ¿Qué media de índice de restauración corresponde a este pueblo?

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{4.500}{80} = 56,25$$

- Es decir, poniendo la unidad e interpretando la cifra obtenida, la media del índice de restauración es del 56,25% en este pueblo.

8. ¿Qué valor ha sido el más frecuente? Calcúlalo.

- La frecuencia absoluta más alta es 19, que corresponde al intervalo 50-60.

$$X_o = Li + c \frac{D_1}{D_1 + D_2} = 50 + 10 \frac{19 - 13}{(19 - 13) + (19 - 11)} = 50 + 10 \frac{6}{14} = 54,28$$

- Es decir, poniendo la unidad e interpretando la cifra obtenida, el valor del índice de restauración que más se repite, es el 54,28%.

9. Se va a seleccionar al 40% de la gente que menos tratada tiene la boca para dar un curso de prevención. ¿Con qué parámetro se haría? Calcúlalo.

- Se trata de seleccionar a un 40% de gente que tiene la boca menos tratada, por tanto con un IR más bajo. Por ello, se empieza a coger desde la izquierda-los valores más bajos en un eje de representación se colocan a la izquierda y los más altos a la derecha-. El parámetro a calcular es por tanto el P₄₀.
- Para el cálculo del P₄₀: F debe ser mayor de:

$$\frac{80}{100} \cdot 40 = 32$$

- Observando en la tabla la columna F, eso sucede con el número 48, que corresponde al intervalo 50-60.

- Aplicando la fórmula para calcular el valor:

$$P_{40} = 50 + 10 \cdot \frac{\frac{80}{100} \cdot 40 - 29}{19} = 51,27$$

- Es decir, seleccionaremos para el curso a las personas que tengan un índice de restauración menor del 51,57% (eran las personas con la boca menos tratada).

10. Se quiere seleccionar también a un 10% de las personas con el IR más alto, para hacer una revisión y un seguimiento durante unos años. ¿Qué parámetro se usaría?
- En este caso se trata de seleccionar al 10% con la boca más tratada, por lo que se empieza a coger desde la derecha -los valores más bajos en un eje de representación se colocan a la izquierda y los más altos a la derecha-.
 - Este 10% corresponde al P_{90} . Fíjate en la imagen que se adjunta, puede ayudarte.



11. Calcula la desviación típica y relaciónala con la media aritmética.

$$s = \sqrt{s^2} = \sqrt{\frac{\sum f_i \cdot x_i^2}{\sum f_i} - \bar{x}^2}$$

Sustituyendo los valores en la fórmula:

$$s = \sqrt{\frac{281.200}{80} - 56,25^2} = \sqrt{3.515 - 3.164,06} = \sqrt{350,94} = 18,73$$

12. ¿Es esta una distribución homogénea?

$$CV = \frac{s}{\bar{x}} \cdot 100 = \frac{18,73}{56,25} \cdot 100 = 33,29\%$$

- La variabilidad de esta distribución de datos es de un 33,29%. ¿Es mucho o poco? Se puede decir que está en el límite que es del 33%.

13. ¿Entre que dos valores se encontrarán probablemente el 65 % de los datos?

- Se sabe que en la mayoría de estas series de datos estos tienden a adoptar una forma determinada, que se le suele llamar campana de Gauss. Por ello se sabe que es probable que el 68% de los dato estén entre $\bar{x}-1s$ y $\bar{x}+1s$.
- En este caso 74,98 y 37,52% del IR.

EJERCICIO RESUELTO SOBRE CORRELACIÓN

Índice de correlación de 0,97:

- Es **positivo**, por tanto la correlación es directa. Esto significa que a aumentos de una de las variables, corresponden aumentos de la otra. En este caso, que cuando aumenta la tasa del consumo de bebidas azucaradas en estos pueblos, aumenta la tasa de caries.
- Es un valor muy próximo a [1], por tanto es **fuerte**, o lo que es lo mismo, una de las variables depende mucho de la otra. En este caso la tasa de caries depende mucho del consumo de estas bebidas.

Índice de correlación de 0,17:

- En este caso también es **positivo**, por tanto cuanto más ausencia de cepillado (lo que en un lenguaje coloquial diríamos a menos cepillado) más caries.
- En un valor próximo a 0, por tanto la correlación es **débil**, por lo que la caries en estos pueblos no depende mucho de la ausencia de cepillado.

Ambos índices:

- Según este estudio, la correlación entre caries y consumo de bebidas azucaradas es más fuerte que entre caries y ausencia de cepillado, por que el valor obtenido está próximo a [1], mientras que la otra el valor es 0,17, más próximo a 0.
- Es decir la variable caries depende más del consumo de bebidas azucaradas.
- Aunque no lo pregunta en el ejercicio, se podría añadir, que en caso de un programa de educación para la salud, para controlar la caries, se debería hacer sobre el consumo de bebidas azucaradas, y no sobre el cepillado, porque según este estudio la caries depende más de ello.

Anexo.- Licencias de recursos.

Licencias

Recursos

Descripción del recurso

Autor: Ministerio de Educación y Formación Profesional.



Licencia: Uso Educativo no comercial.

Procedencia: Elaboración propia.



Autoría: Ministerio de Educación y Formación Profesional. Fotos casos prácticos. Gestor de proyectos.

Licencia: Uso educativo no comercial para plataforma FPaD

Procedencia: Elaboración propia

Autoría: Centers for Disease Control and Prevention.



Licencia: Copyright (cita)

Procedencia: Imagen extraída de la página de descarga del programa. <http://wwwn.cdc.gov/epiinfo/>



Autoría: ITE - Felix Valles Calvo

Licencia: CC by-nc-sa

Procedencia: ideITE=152556



Autoría: Stockbyte.

Licencia: Uso educativo no comercial para plataformas públicas de Formación Profesional a distancia.

Procedencia: CD-DVD Num. V43



Autoría: ITE - *Luana Fischer Ferreira*

Licencia: CC by-nc-sa

Procedencia: ideITE=181013



Autoría: Stockbyte.

Licencia: Uso educativo no comercial para plataformas públicas de Formación Profesional a distancia.

Procedencia: CD-DVD Num. SS36



Autoría: ITE - Pilar Acero López

Licencia: CC by-nc-sa

Procedencia: ideITE = 145775



Autoría: ITE - Félix Valles Calvo

Licencia: CC by-nc-sa

Procedencia: ideITE = 152295



Autoría: Inés Araico Galdos

Licencia: Uso educativo no comercial para plataformas de FPaD

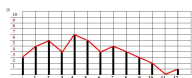
Procedencia: Elaboración propia



Autoría: Inés Araico Galdos

Licencia: Uso educativo no comercial para plataformas de FPaD

Procedencia: Elaboración propia

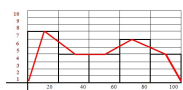


Autoría: Inés Araico Galdos

Licencia: Uso educativo no comercial para plataformas de FPaD

Procedencia: Elaboración propia

Autoría: Inés Araico Galdos



Licencia: Uso educativo no comercial para plataformas de FPaD

Procedencia: Elaboración propia



Autoría: ITE - Antonio Ortega Moreno.

Licencia: CC by-nc-sa

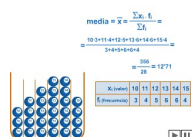
Procedencia: ideITE= 175286



Autoría: Stockbyte.

Licencia: Uso educativo no comercial para plataformas públicas de Formación Profesional a distancia.

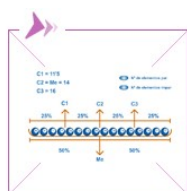
Procedencia: CD-DVD Num. EP006



Autoría: ITE - Isabel Martín Rojo

Licencia: CC by-nc-sa

Procedencia: ideITE = 185460

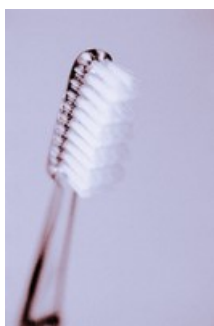


Cálculo de la mediana

Autoría: ITE - Isabel Martín Rojo

Licencia: CC by-nc-sa

Procedencia: ideITE = 183231



Autoría: Stockbyte.

Licencia: Uso educativo no comercial para plataformas públicas de Formación Profesional a distancia.

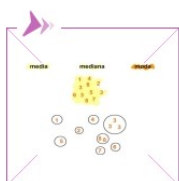
Procedencia: CD-DVD Num. SS36

Autoría: ITE



Licencia: CC by-nc-sa

Procedencia: ideITE = 160013



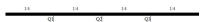
Parámetros de centralización

Autoría: ITE - Jorge Sánchez

Licencia: CC by-nc-sa

Procedencia: ideITE = 195395

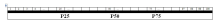
Autoría: Inés Araico Galdos



Licencia: Uso educativo no comercial para plataformas de FPaD

Procedencia: Elaboración propia

Autoría: Inés Araico Galdos



Licencia: Uso educativo no comercial para plataformas de FPaD

Procedencia: Elaboración propia

Autoría: Inés Araico Galdos



Licencia: Uso educativo no comercial para plataformas de FPaD

Procedencia: Elaboración propia

Autoría: Inés Araico Galdos



Licencia: Uso educativo no comercial para plataformas de FPaD

Procedencia: Elaboración propia

Autoría: Inés Araico Galdos



Licencia: Uso educativo no comercial para plataformas de FPaD

Procedencia: Elaboración propia

Autoría: Ministerio de Educación y Formación Profesional.

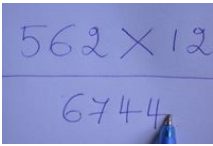


Licencia: Uso educativo no comercial para plataformas FPaD

Procedencia: Elaboración propia

Autoría: ITE

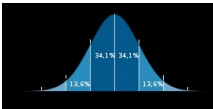
Licencia: CC by-nc-sa



Procedencia: ideITE = 109938

Autoría: Nusha.

Licencia: Documentación libre GNU



Procedencia: <http://commons.wikimedia.org>

[/wiki/File:Standard_deviation_diagram_\(decimal_comma\).svg?uselang=es](/wiki/File:Standard_deviation_diagram_(decimal_comma).svg?uselang=es)

Autoría: ITE



Licencia: CC by-nc-sa

Procedencia: ideITE= 120790

Autoría: ITE



Licencia: CC by-nc-sa

Procedencia: ideITE = 170970



Autoría: Silvia Portero Cano

Licencia: Uso educativo no comercial para plataforma FPaD

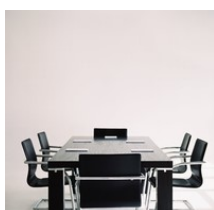
Procedencia: Elaboración propia



Autoría: Silvia Portero Cano

Licencia: Uso educativo no comercial para plataforma FPaD

Procedencia: Elaboración propia



Autoría: Stockbyte.

Licencia: Uso educativo no comercial para plataformas públicas de Formación Profesional a distancia.

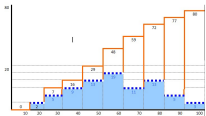
Procedencia: CD-DVD Num. CD165



Autoría: ITE

Licencia: CC by-nc-sa

Procedencia: ideITE = 110282



Autoría: Inés Araico Galdos

Licencia: Uso educativo no comercial para plataformas de FPaD

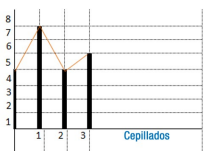
Procedencia: Elaboración propia



Autoría: Inés Araico Galdos

Licencia: Uso educativo no comercial para plataformas de FPaD

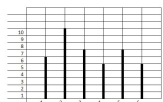
Procedencia: Elaboración propia



Autoría: Inés Araico Galdos

Licencia: Uso educativo no comercial

Procedencia: Elaboración propia



Autoría: Inés Araico Galdos

Licencia: Uso educativo no comercial

Procedencia: Elaboración propia

