TEMA 5

REDUCCION DE DATOS , MEDIDAS DESCRIPTIVAS BASICAS Y REPRESENTACIONES GRAFICAS

1-INTRODUCCION DE LA ORGANIZACION A LA DESCRIPCION DE DATOS.

Una vez que hemos depurado los datos y compro bado que la tabla de dato sesta bien , pasaremos ha analizar esos datos.

Si para la depuracion de datos hemos usado las frecuencias ya estara hecho el primer paso, sino deberemos realizaresta estadistica descriptiva (la distribucion de frecuencais)

La variabilidad es el estudio de la dispersion de los valores, contribuendo ha explicar lo grande que es la muestra y la naturaleza de los valores(o sea las medidas de dispersin nos muestra la variabilidad). Ahora el libro nos ofrece la descripcion de conceptos ya vistos. Cuando usamos la estadistica para constrastar hipotesis buscamos la relation posible entre variables

Variable es lo que puede cambiar de un sujeto a otro

Constante es un valor comun para todos los sujetos algo que no cambia

La estadistica descriptiva es el procedimento de organizar clasificar y resumir el conjunto de datos , o sea crear una tabla donde organizar y clasificar y despues crear frecuencias y graficas donde se resumen. Pero tambien se usaran formulitas que nos ayudan a resumir los datos como las medidas de tendencia central y las medidas de dispersion o variabilidad (medidas centralizadas son la media aritmetica, la mediana y la moda) , (las medidas de dispersion son la desviacion con respecto a la media la desviacion media la varianza y la desviacion tipica) Cuando los datos son muy diferentes los valores , ej 150 valores como en el caso del CI, las reducismo a intervalos. Primero se hacen las medidas centralizadas y luego las de dispersion (logico porque para hacer las de disperison necesitamos la media aritmetica).

5.2 MEDIDAS DE TENDENCIA CENTRAL : MEDIA , MEDIANA Y MODA. USOS HE INTERPRETACION

Busca cual es el valor que mejor representa al grupo .

La media aritmetica se halla sumando todos los valores dados y dividiendoles entre el total. Se representa con el signo $\bar{X}$ y la formula es $\bar{X} = \frac{X_1 + X_2 + X_3 + \ldots + X_n}{N}$ o sea la suma de todos los datos entre la poblacion o en reducido $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{N}$ que quiere decir exactamente lo mismo la suma de todas los valores , pero esta forma seria un poco dificil de llevar a cabo si tenemos muchos valores, imaginar una encuesta de 100 personas sobre el numero de hijos , o sea que si nos pusiesemos a sumar los hijos de cada una de las cien personas nos perderiamos antes o despues para muestra un boton 1111122222223333334444444222224342323131, si no te pierdes sumando todo eso seria impressionante,

Por eso los estadisticos han hecho una trampita es la creacion de una columna nueva que ya sabei visto antes la de Xi x fi en esta multiplicamos el valor (xi) por el numero de veces que aparece (fi) y sumando la multiplicacion de todas los valores ∑ Xi nos daria lo mismo que sumar todos los valores. Ya vereis que lo entendeis mejor con un ejemplo :

xi	fi	Xi * fi
1	5	5
2	2	4
3	3	9
∑	10	18

Hemos preguntado a 10 amigos por los hijos que tienen , y el resultado es el siguiento hay 5 amigos que tienen 1 hijo , 2 amigos que tienen 2 hijos y 3 amigos que tienen 3 hijos . O sea 1,1,1,1,1,2,2,3,3,3, la media la podriamos hacerla sumando todos los datos 1+1+1+1+1+2+2+3+3+3 =18 , pero mirar que coincidencia la suma de la columna Xi*fi es tambien 18. Porque si sumamos 1+1+1+1+1=5 y si multiplicamos 1 numero de hijos por 5(fi) tambien nos da 5 ,y asi sucesibamente. Por eso la suma de esas columnas es igual que la suma entre el total de valores. Pero si os cuesta entenderlo no lo penseis simplemente tomarlo como una formula mas , si sumas todas las columnas de Xi *fi y lo divides entre el total de la muestra nos da la media aritmetica.

Con esta medida podemos comparar los grupos miraremos si es inferior o superior al resultado… pero todo esto como pedagogo no es estadistica. La media solo puede usarse n medidas de intervalo y de razon, o dicotomicas.

La moda es el valor que mas aparece, no necesita ningun calculo . Mira la columna de la frecuencia absoluta (fi) y el numero mas alto nos indicara cual es el valor xi de la moda.

Si hay dos valores con la misma frecuencia absoluta mas alta lo llamaremos bimodal, o sea moda de dos valores.(importante la moda siempre es xi o sea el valor a estudiar no se pondria el numero de veces. Por ejemplo la moda de la mitabla anterior sera 1 hijo. Si son tres o mas valores los altos sera, plurimodales. La mediana es el valor que esta en medio , el del centro cuando las puntuaciones estan en tabla o ordenados es facil , es el que ocupa la posicion central , coincide con el percental 50 (o sea percentil viene de porciento y el percentil 50 es el que esta en el 50% o sea la mitad, porque la mitad de 100 es 5'.

Para calcularlo miraremos la columna de frecuencia absoluta acumulada es decir Fi y mararemos n (el numero de gente) y lo dividiremos entre dos y cuando la frecuencia absoluta acumulada pase ese numero de n/2 ese valor o sea xi sera la mediana. Veamos esta tabla de 15 personas entonces n/2=15/2= 7,5

xi	fi	Fi
1	7	7
2	4	11
3	5	15
	15	

Donde estara el valor 7,5 pues como con 1 hijo se queda en 7,7,5, estaran en dos hijos. Cuando la division n/2 da un numero entero la mediana necesitara la media aritmetica de esos dos valores centrales por ejemplo en la tabla 14/2=n/2=7

xi	fi	Fi
1	5	5
2	2	7
3	7	14

Pero como es un numero entero no nos serviria porque la mediana seria un intervalo que va de 7 a 8 por eso su mediana sera dos hijos y tres hijos porque entre ellos esta 7,5.

La medida mas precisa de tendencia centrale s la media aritmetica, pero solo se puede hacer en medidas de intervalo o de razon (es decir las que tienen numeros) Es la suma de todos los valores entre el total de la mustra por eso se hace una columna nueva, donde se moltiplica la fi por xi y al sumarlo nos seria la suma de todos los valores. En esta tabla seria 30/14=2,14

xi	fi	Fi	Fi·xi
1	5	5	5
2	2	7	4
3	7	14	21
∑	14		30

Si hay muchas diferencias entre la media y la moda nos indica un dato muy interesante, nos indica que hay muchos valores bajos , lo entendereis mejor con un ejemplo de las notas de clase la media es 5,2 pero la moda es 7 eso significa que la nota que mas se repite es notable siete pero hay un grupo grande de membro de la clase con notas muy muy bajas que nos bajan la media dos puntos y deberiamos preguntarnos porque (el metodo no se adapta bien a todos los alumnos por ejemplo) por eso la media es la mas imporntante porque al verse afectada por la puntuaciones extremas da una vision mas global no solo mira lo normal sino que ve un poco mas alla .La media se ve arrastarada por las puntuaciones extremas. La distribucion de frecuencias normale s cuando la media aritmetica , la mediana y la moda son el mismo valor

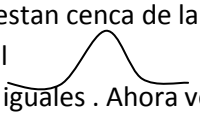
5.3MEDIDAS DE VARIABILIDAD O MEDIDAS DE DISPERSION

Tambien necesitamos saber si hay muchos valores que se alejan de la media aritmetica , o si la mayoria estan cerca de la media para ello se usa la medida de dispersion

5.3.1 LA DESVIACION MEDIA .

Es la media aritmetica de la desviacion del valor con respecto a la medai . Traducido es la media aritmetica de las desviaciones respecto a la media , que son la resta a cada valor de la media aritmetica, se suma todos los resultados de las restas y se dividen entre el numero total de la muestra. Por eso es necesario tener la deviacion respecto a la media para lo que se hara una nueva columnita que sera xi-$\overline{x}$. Es $\overline{X}$ 18/10 = 1,8

xi	fi	Fi	Fi·xi	x-$\overline{x}$
1	5	5	5	1-1,8=0,8
2	2	7	4	2-1.8=0,2
3	3	10	9	3-1,8=1.2
∑	10		18	2,2

Si el valor de esta desviacion media es muy alta significa que la mayoria de los datos sus valores estan muy lejos de la media aritmetica, vamos para pedagogia que hay un problema. Pero si es pequena significa que los valores estan cenca de la media , significa que todo esta bien. Ej grafico todo bien es una campana de guas normal  todo mal  Si la desviacion media es 0 significa que todos los valores son iguales . Ahora veremos como se halla la deviacion media tomaremos los valores sin

importarnos el signo que sea negativo o positivo. Es por eso que una desviacion media 0 significa que todos los valores son iguales porque si tuviesemos estos valores.

x	fi	xi. fi	Xi -$\bar{x}$
80	2	160	-24
128	2	256	24

$\bar{X}$= 416/4 =104

Ahora haremos la desviacion media que es una media aritmetica de la desviacion respecto a la media , y ccomo lo que es la desviacion media es una media , lo que debemos hacer es sumarla desviacion respecto a la media de todos los valores que han dado los encuestados . Por eso como ocurre al hacer la media aritmetica debemos hacer una nueva columna que sera $|xi-\bar{x}| \cdot fi$

xi	fi	Xi·fi	xi-$\bar{x}$	$\|Xi - \bar{X}\| \cdot fi$
1	5	5	-0,8	0,8·5 =4
2	2	4	0,2	0,2·2=0,04
3	3	9	1,2	1,2·3=3,6
∑	10	18		7,64

$\bar{X}$ = 1,8

Otra cosita si os fijais en la desviacion respecto a la media del valor 1 tendria que restar a 1- 1.8 y por tanto me da un producto negativo, pero en estadistica para la elaboracion de tablas no vamos a tener en cuenta el signo negativo(ese es el significado de esa especie de parentesis en forma de lineas que tienen algunos valores). Una vez hecha esta nueva columna nos iremos a la fila de los sumatorios y sumaremos todos sus valores y lo dividiremos entre el total de la muestra diez en este caso y esta sera la desviacion media

Dm =7,64/10 = 0,76

5.3.2 La desviacion tipica

Otro indice conocido es la desviacion tipica representada por una letra S para estadistico(basado en la muestra), y б para parametro(para poblacion). Ademas esta la varianza que se representa con una σ^2 para parametros y una s al cuadrado para muestra. Se realiza elevando al cuadrado cada desviacion respecto a la media de todos los valores o sea tendriamos que poner al cuadrado (multiplicar por si mismo) las desviaciones respecto a la media de todos y cada uno de los sujetos , por tanto haremos una nueva columna que multipli caria la frecuencia absoluta por el cuadrado de la desviacion respecto a la media. La suma de estos cuadrados se ha de fi y asi no tendremos que sumarlo uno a uno, ademas debemos una vez hecha la suma dividirlo entre el numero de la muestra .

Vamos paso a paso

1) Restamos la media aritmetica a los valores
2) Ponemos al cuadrado la desviacion respecto a la media (paso 1)
3) Multiplicamos por el numero de veces que aparece cada valor fi.
4) Sumamos el resultado de toda esta nueva columna
5) Dividimos el resultado de la suma entre el total de la muestr

6) Al resultado le racemo su raiz cuadrada

Aqui da igual que utilicemos el numero de la desviacion respecto a la media en negativo o en positivo porque al ponerlo al cuadrado el negativo desapareceria porque menos por menos es mas y al hacer el cuadrado desaparece por tanto el negativo

xi	fi	Xi·fi	xi-$\widetilde{x}$	|Xi − $\overline{X}$ | al cuadrado	|Xi − $\overline{X}$ | al cuadrado por fi
1	5	5	-0,8	0,64	3,2
2	2	4	0,2	0,04	0,08
3	3	9	1,2	1,44	4,32
∑	10		8		7,6

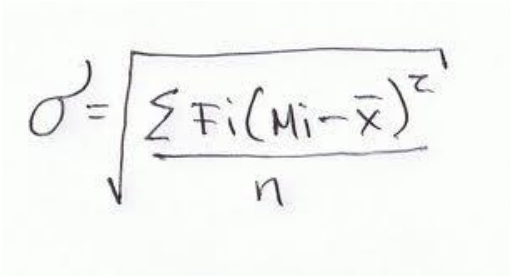
Media aritmetica es = a 1,8

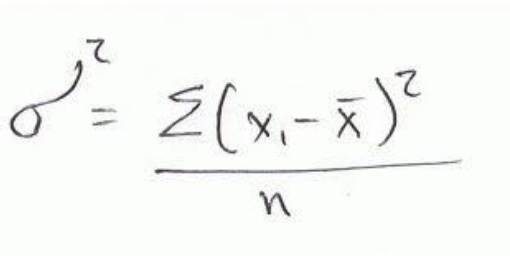
Desviacion media 0,76

Varianza = 0,76

La desviacion tipica es = 0,87

La formula es





La desviacion tipica es dificil de interpretar y se utiliza para comprobar la dispersion entre gupos distintos. Y por tanto eso a veces se hace la misma desviacion tipica y la maxima

Para la minima desviacion tipica todos los valores deben sere l mismo vaolor y para la maxima , la mitad de lso sujetos deberian obtener la puntuacion maxima de la escala y la otra mitad la minima. Esto solo se ve en variablaes dicotomicas o dicotomizadas y cuanto mas grande sea la distancia entre valores mas improbabile es que se de esta situation…Si una desviacion tipica es mas baja en un grupo que en otro significa que los sujetos estan mas igualados entres si. La desviacion tipica es un apoblacion sera mas elevada que en una muestra, al haber mas jujetos , seran mas probabile encontrar mayores diferencia s interindividuales.

Por eso se ve que cuanto mas baja , los valores son mas iguales. Estas que hemos visto son desviaciones tipica sesgada que acentua el efecto de las grandes desviaciones y es siempre superior a la desviacion media.

La desviacion tipica y varianza insesgada mide lo mismo que la sesgada aunqaue tiene diferentes propiedades , que es en vez de parala muestra para la poblacion . Y es mas elevada que en una muestra (al haber mas sujetos esmas probable encontrar mas diferencias por la que ponemos el 1 para disminuir el denominador , el resultado de la desviacion sera por tanto mayor . La ionsesgada se codifica para muchos б-1 y la sesgadaб

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

5.3.3 La amplitud o recorrido

Otra medida de dispersion son la amplitud o recorrido de una variable que es la puntuacion mayor menos la menor mas 1.

A=X1 – Xultimo +1

Y se usa para organizar intervalos y para los graficos. Y afecta a la dispersion y las otras medidas porque no es lo mismo una varianza de 5 para 10 personas que para 1000. Y se utiliza como medida de dispersion unica cuando no se pueden usar otras, o para complementar la moda. Solo se hace con dos puntuaciones por lo que puede dar resultados que confundan sino se toma con cui dado porque puede haber valores extremos (outliers)

5.3.4 La desviacion semi intercuartilica

Se denomina desviacion semiintercuartilica Q que muestra la dispersion en el 50% central de la distribucion. Util para nivel de medida ordinal y como complemento a la mediana . Asi se eliminan las puntuaciones extremas al prescindir del 25% superior e inferior de las puntuaciones. Para hallarla se ha de saber el 1° cuartil y el 3°cuartil recuerda 1Q =1·N÷4. El 3Q = 3·N÷4

$$\text{desv.int cuartil.} = \frac{Q_3 - Q_1}{2}$$

5.3.5 El coeficiente de variacion

Se representa con V o CV. Permite comparar variables (valores) de medida de datos diferentes, por ejemplo el coefficiente intelectual de amplitud 50 y actitud hacia el estudio de amplitud 10.

$$C.V. = \frac{S}{\overline{X}} * 100$$

Este indice tiene limites fijos dependiendo de valor de la media. Suele ser menos que 1. Se interpreta en porcentaje depende de la desviacion tipica o estandar. Vamos es la relation entre esta desviacion tipica y su media.

5.4 Media y desviacion tipica para variables dicotomicas

La variables dicotomicas es un tipo de variable Xi que solo tiene dos posibles valores . Valores a los que se puede hacer la media aritmetica para cada uno . .La suma de ambas medias debe dar siempre 1

P+q=1

La varianza tambien es muy facil de calcular

σ^2 = p·q

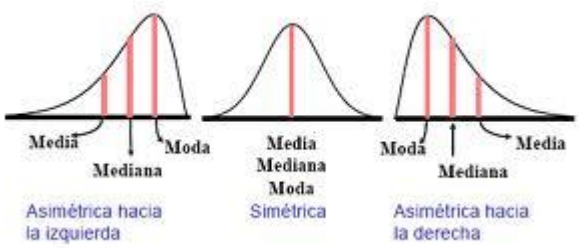
S=√P·Q

5.5 Asimetria y apuntamiento : relation con la curva normal

La representacion grafica de las variables , sobre todo cuando trabajamos con maestra grandes, tienden a ser curvas que por su grado de asimetria, pueden asemejarse a una de estas tres .

Una curva normal carece de asimetria , es simetrica o tiene indice de asimetria igual a 0

Por su grado de asimetria puede asemejarse a una de estas tres



La asimetrai positiva indica que la mayoria de los sujetos esta en la parte baja de la puntuacion de la distribucion de frecuencia. La cola de la distribucion esta a la derecha . Lo cual no indica que los sujetos tengan puntuaciones bajas , todo depende donde se encuentra el grupo mas grande de puntuaciones que puede ser mas alta.

La asimetria negativa indica lo contrario , que los sujetos tienden a agruparse en torno a las puntuaciones altas de la distribucion , lo que como antes no significa necessariamente que los sujetos tengan puntuaciones altas.

La asimetria positiva a negativa se debe solo al signo del calculo del indice

El indice Perarson es la forma de calcolarlo mas facil

El coeficiente de correlación de Pearson, pensado para variables cuantitativas (escala

mínima de intervalo), es un índice que mide el grado de covariación entre distintas

variables relacionadas. El coeficiente de correlación de Pearson es un índice de fácil ejecución e, igualmente, de

fácil interpretación. Digamos, en primera instancia, que sus valores absolutos oscilan

entre 0 y 1. Hemos especificado los términos "valores absolutos" ya que en realidad si se contempla

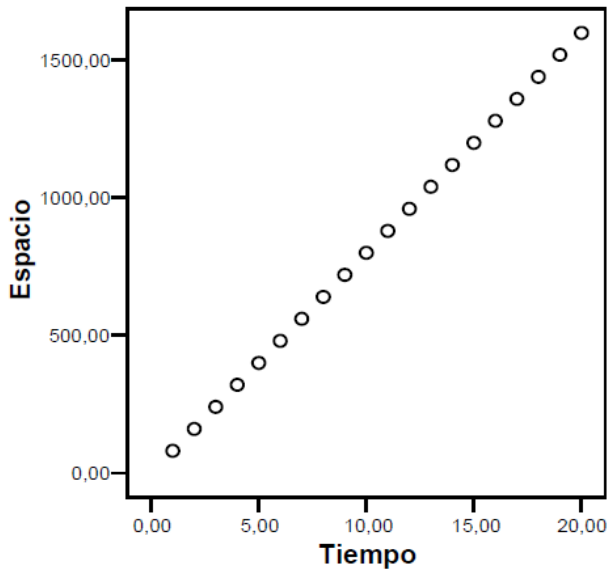
el signo el coeficiente de correlación de Pearson oscila entre –1 y +1.

En este sentido, tan fuerte

es una relación de +1 como de -1. En el primer caso la relación es perfecta positiva y en el segundo perfecta negativa.

Decimos que la correlación entre dos variables X e Y es perfecta positiva cuando exactamente en la medida que aumenta una de ellas aumenta la otra. Esto sucede cuando la relación entre ambas variables es funcionalmente exacta.

Por ejemplo, la relación entre espacio y tiempo para un móvil que se desplaza a velocidad constante. Gráficamente la relación ser del tipo:



Vamos resumiendo cuanto más intensa sea la concordancia (en sentido directo o inverso) de las posiciones relativas de los datos en las dos variables, el producto del numerador toma mayor valor (en sentido absoluto). Si la concordancia es exacta, el numerador es igual a N (o a -N), y el índice toma un valor igual a 1 o -1.

La formula es

$$A = \frac{\overline{X} - Mo}{S}$$

Esa si porque la media , a diferencia de la moda y la mediana se ve atraido por portuaciones extremas, luego su posicion relativa se vera desplazada hacia puntuaciones extremas.
El apuntamiento o la curiosi indica el grado en que la distribucion es mas o menos picudo , es decir si la distribucion es mas abierta o dispersa respecto a la media.
Si las puntuaciones estan poco concentradas respecto a la media y por tanto mas chata o aplastada (platicurtica) o por el contrario mas apuntada y por tanto mas estrecha o con las distribuciones mas concentradas en torno a la media (leptocurtica). Una curiosi igual a cero viene representada por la distribucion normal. Curtosis superior a cero nos indica una distribucion leptocurtica maestra que si es inferior a cero , esta sera platicurtica. Par calcular la curiosi puede utilizarse la siguiente formula.

Hemos comentado que el concepto de asimetría se refiere a si la curva que forman los valores de la serie presenta la misma forma a izquierda y derecha de un valor central (media aritemética)

$$g_2 = \frac{1}{N} \cdot \frac{\sum_{i=1}^{n} (X_i - \bar{X})^4 \cdot f_i}{\sigma^4} - 3$$

5.6 Representaciones graficas

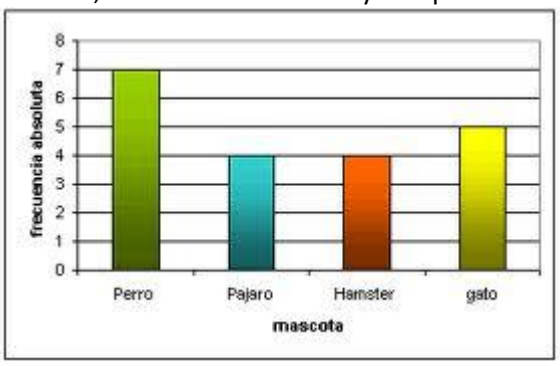
La representacion grafica de los datos como histogramas , poligono de frecuencias, ciclo grama… son buen complemento de los indices numeriocos y ayudan a comprender rapidamente la information descriptiva. A partir de una distribucion de frecuencias es muy facil realizar una representacion grafica, como hemos visto en el capitulo anterior. Y es recomendable adaptar el tipo de grafico al nivel de medida de la variables

5.6.1 Grafica de sectores

Tb llamado ciclo grama o grafico de torta representa variables con nivel de medida nominal. Tiene forma circular y dicidido en porciones cada porcion representa la presencia proporcional de cada una de los niveles de la variable.

5.6.2 Grafico de barras

Es comparativo , compara porcentajes fi , en una muestra. Suele usarse para variables con nivel de medida ordinal , o cuando es nominal y comparaciones de variables clasificatorias o categoricas.



5.6.3 Histograma

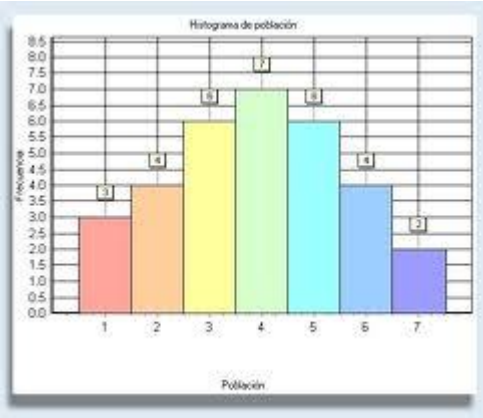
Es parecido al diagrama de barras, pero se usa en variables cuantitativas continuas con nivel de medida de intervalo o de razon . Las barras estan juntas. No aparecen todos los valores de la variable continua , son muchos los valores que pueden obtener los sujetos y seria muy farragoso con un grafico con muchas barras. Por eso se agrupan las puntuaciones en intervalo. Asi cada puntuacion tiene un numero determinado de puntuaciones, denomina valor intervalo .

Cuanto menor es el numero de intervalo ,, mas suavizados aparece la grafica. Los intervalos suelen usarse .

Estimar por tanteo el numero de intervalos que se desea, segun lo que acabamos de indicar

Iniciar el primer intervalo que incluya a puntuacion directa mas baja y qeu sea multiplo del valor del intervalo

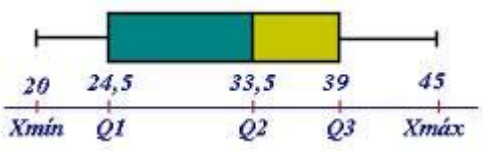
Situar la marca de clase (punto central del intervalo) en el eje de abscisas en el centro de cada barra



Hasta hace unas decadas , la agrupacion por intervalos se utilizaba tambien para el calculo de medidas de tendencia y variabilidad. Porque no se usaban programas informaticos . El poligono de frecuencias es cuando hay una linea que une el punto superior central de cada barra , abriendo y cerrando el eje de abscisas.

5.6.4 Grafico de caja

Tambien llamado como caja y pastillas o caja y bigotes Es practico porque permite hacerse una idea rapida de la distribucion de las puntuaciones en la zona central desde el primer cuartil o percentil 25 hasta el tercer cuartil o percentil 75 y en los extremos. Las pastillas representan las puntuaciones hasta los extremos de la distribucion o sea el ultimo y el primer valor



5.6.5 Grafica de tallo y hojas

El grafico de tallo y hojas combina la representacion numerica y grafica . Es una especie de histograma horizontal , construida con los numeros correspondientes a las puntuaciones . Las hojas es el ultimo digito de la puntuacion y el tallo , es el resto de digitos

15,16,21,23,23,26,26,30,32,41

Tallo	Hoja
1	5 6
2	1 3 3 6 6
3	0 2
4	1