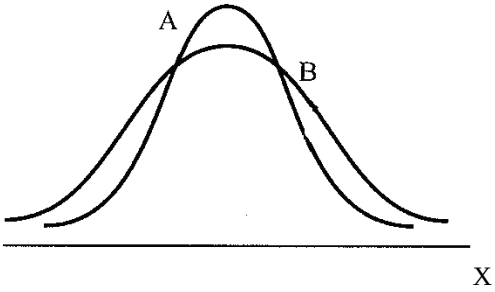
3.2 – Caracterización de grupos: Estadísticos de dispersión

1. **Variables categóricas: el índice de variación cualitativa.**
2. **Variables ordinales: el rango y el rango intercuartil.**
3. **Variables cuantitativas: la varianza, la desviación típica y el coeficiente de variación.**

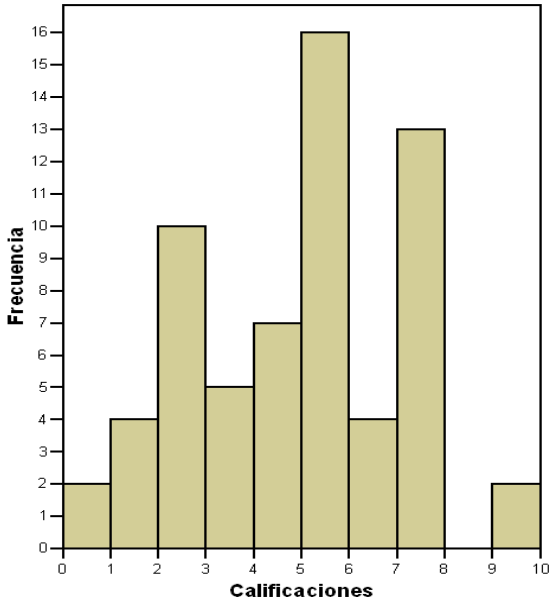
• En diversos textos de Estadística se hace referencia a la dispersión o variabilidad como la razón de ser de esta disciplina; por ejemplo, de Veaux, Bock y Velleman (2003) afirman de forma rotunda en su manual *Intro Stats* lo siguiente: "Statistics is about variation". En efecto, si no existiese heterogeneidad o dispersión en las variables que estudiamos sería muy obvio resumir la información de las mismas, no haciendo ninguna falta los métodos estadísticos: con el dato de un caso, conoceríamos lo que ocurre para el resto de los casos –en vez de variables, tendríamos, en realidad, constantes.

• El concepto de dispersión hace referencia al grado en que los datos de una variable son más homogéneos (=> menor dispersión o variabilidad) o más heterogéneos (=> mayor dispersión o variabilidad). Es un concepto que tal vez pueda captarse de un modo más intuitivo a nivel gráfico: los siguientes polígonos de frecuencias suavizados muestran gráficamente la distribución de una misma variable (*X*) en dos grupos distintos de sujetos (A y B); observándolos ¿cuál de los dos grupos diríamos que tiene mayor dispersión en dicha variable?

BY NC SA



• Origen de la variabilidad: la dispersión en los valores de los sujetos en una variable puede deberse a diferentes causas, a las cuales se suele hacer referencia como fuentes de variación de los datos. Por ejemplo, la variabilidad en las calificaciones de *Estadística* de los estudiantes del grupo E de un curso académico reciente (ver histograma), ¿a qué puede deberse? En este caso, una fuente de variabilidad fundamental será el conocimiento y dominio de la materia. Es de esperar que diferencias individuales en este aspecto sean la principal causa de la dispersión existente en las calificaciones de la asignatura.



Ahora bien, supóngase el caso ficticio en que todos los alumnos hubieran tenido el mismo dominio y nivel de conocimientos de la asignatura, ¿es de esperar que la nota hubiese sido las misma para todos ellos? –Es más que probable que éste no sea el caso. No es difícil pensar en otras posibles fuentes de variación que tendrán su influencia en las diferencias existentes en las calificaciones, por ejemplo, lo bien que se haya dormido la noche antes del examen, la capacidad para afrontar situaciones estresantes, la habilidad para responder al tipo de preguntas planteadas en el examen (objetivas, abiertas…), la fiabilidad y validez del instrumento de medida (el examen), cómo nos haya sentado el desayuno (o la comida) previa al examen, etc.

BY NC SA

• A continuación se presentan una serie de índices estadísticos y representaciones gráficas orientados a describir la dispersión de una variable. Estos aparecen diferenciados en 3 apartados que se corresponden con la distinción planteada para las variables en función de su escala de medida.

1. Variables categóricas: el índice de variación cualitativa

• El índice de variación cualitativa (*IVC*) se obtiene a través de la siguiente fórmula, donde *k* es el número de modalidades de la variable y p_i la frecuencia relativa asociada a cada una de ellas:

$$IVC = \frac{1-\sum p_i^2}{(k-1)/k}$$

• El *IVC* expresa el grado en que los casos están dispersos en las diferentes modalidades de la variable, alcanzando su máximo (*IVC* = 1) en el caso en que las frecuencias relativas sean iguales para todas las modalidades de la variable (caso que se corresponde al de una variable con una distribución uniforme, esto es, de máxima dispersión). El *IVC* sería igual a 0 cuando la frecuencia relativa de una modalidad de la variable fuese igual a 1, esto es, el caso en que todos los casos tuviesen el mismo valor observado en la variable (dispersión nula).

Ejemplo: Sea la variable "Religión que se profesa" [Codificación: 0: Católica; 1: Protestante; 2: Otra; 3: Ninguna] de la que se han obtenido datos para una muestra de 50 personas, cuya distribución de frecuencias se muestra a continuación:

X_i	*Frec. absoluta* (n_i)	*Frec. relativa* (p_i)
0	12	0,24
1	10	0,2
2	10	0,2
3	18	0,36
	50	1,00

El valor del *IVC* será igual a:

$$IVC = \frac{1-(0.24^2+0.2^2+0.2^2+0.36^2)}{(4-1)/4} = 0.98$$

Ejercicio 1: Obtener el *IVC* en la distribución de frecuencias de la variable "Estado civil" que se presentó en los dos temas previos y que aparece a continuación:

X_i	Frec. absoluta (n_i)	Frec. relativa (p_i)	Porcentaje ($\%_i$)
soltero/a	15	0,3	30
casado/a	20	0,4	40
separado/a	11	0,22	22

CC BY NC SA

viudo/a	4	0,08	8
	50	1,00	100

Ejercicio 2: Inventa dos distribuciones de frecuencias para la variable "Estado civil" en que el *IVC* sea, respectivamente, tan bajo y tan alto como sea posible.

2. Variables ordinales: el rango y el rango intercuartil

2.1. El rango

• También denominado como amplitud, consiste en obtener la diferencia entre el mayor y el menor valor observado de la variable:

$$Rango = Máximo - Mínimo$$

Ejemplo de obtención del *rango* para la variable con los datos recogidos con la pregunta "Ansiedad que siente cuando se encuentra con mucha gente alrededor" de un test orientado a medir la ansiedad (escala de respuesta: 1: Nada; 2: Algo; 3: Bastante; 4: Mucha.).

X_i	n_i	$\%_i$	n_a	$\%_a$
1	23	19,0	23	19
2	36	29,7	59	48,7
3	47	38,9	106	87,6
4	15	12,4	121	100
	121	100		

$Rango = 4 - 1 = 3$

• La principal desventaja del rango es que al basarse su cálculo en los valores mínimo y máximo, si la distribución tiene valores atípicos, su cálculo se verá muy influido por los mismos. En estos casos, el rango puede proporcionar valores que no sean buenos indicadores de la verdadera dispersión de los datos –por ejemplo, en la variable X :{8, 8, 9, 10, 10, 12, 50}, el rango es igual a 42 cuando, en realidad, todos los datos, salvo uno, son bastante homogéneos.

Ejercicio 3: Obtener el *Rango* de la variable obtenida a partir de los datos recogidos con la pregunta: "Se valora en los empleados la creatividad y la capacidad de creación" de un test de cultura organizacional en empresas que se aplicó a una muestra de 200 empleados de diferentes empresas [1: Muy en desacuerdo; 2: Bastante en desacuerdo; 3: Algo en desacuerdo; 4: Ni en desacuerdo ni de acuerdo; 5: Algo de acuerdo; 6: Bastante de acuerdo; 7: Muy de acuerdo]:

CC BY NC SA

X_i	n_i	$\%_i$	$\%_a$
2	21	10,5	10,5
3	31	15,5	26
4	36	18	44
5	47	23,5	67,5
6	38	19	86,5
7	27	13,5	100
	200	100	

• En lo que respecta a la interpretación del rango, tanto éste como el resto de índices de variabilidad que se van a tratar en las siguientes secciones (exceptuando, parcialmente, el coeficiente de variación) ofrecen resultados que no tienen una interpretación directa en términos absolutos. Así, ¿qué significa un rango de 4 o un rango de 10, mucha o poca dispersión?

– El único caso en que la interpretación de estos índices es inequívoca es cuando dan igual a 0, indicando la ausencia de variabilidad en los datos de la variable –caso por otra parte bastante excepcional. Valores mayores que 0 indicarán dispersión en los datos, tanto mayor cuanto mayor sea ese valor, pero sin existir un techo que nos permita establecer interpretaciones en términos absolutos.

– La interpretación de estos índices depende de la naturaleza de la variable considerada y de la escala utilizada al ser medida –por ejemplo, un rango de 10 en la variable *Peso* (kg) en una muestra de personas adultas sí que nos da una idea de la dispersión de esa variable: se trata de una variable con muy poca dispersión dado que cabría esperar que, en una muestra de personas adultas, la diferencia entre el valor máximo y el mínimo de peso fuese bastante mayor que 10. Sin embargo, en otros muchos ejemplos la interpretación podría resultar mucho más incierta, por ejemplo, un rango de 840 milisegundos en la variable tiempo de reacción para reconocer un determinado estímulo visual, ¿indica mucha o poca dispersión? Tal vez para alguien con experiencia en experimentos de tiempo de reacción con estímulos visuales, ese valor de rango sí que le permita interpretar la variabilidad de los datos asociada a ese resultado pero, en caso de no contar con esa formación, puede resultar más que aventurado realizar una interpretación al respecto.

– Ahora bien, sí que será siempre posible con los resultados de cualquiera de los índices de dispersión realizar interpretaciones en términos relativos, por ejemplo, establecer en dos muestras de las que se tiene datos en una misma variable, cuál de los dos tiene una mayor dispersión en sus datos o, también, comparar la dispersión de los datos de una misma variable medida en dos momentos temporales distintos. No olvidar que no tendrá sentido comparar estos

CC BY NC SA

índices de dispersión cuando se obtengan para variables diferentes –tan solo una salvedad a esta última afirmación: cuando se trate de variables que estén expresadas en las mismas unidades y que tenga sentido comparar (por ejemplo, las variables ingresos y gastos mensuales para una muestra de consumidores).

2.2. El rango intercuartil

• El rango o amplitud intercuartil (*RIC*) se obtiene como diferencia entre los cuartiles 3º y 1º:

$$RIC = Q_3 - Q_1$$

Una variante del mismo es el conocido como amplitud o rango semi-intercuartil:

$$RSIC = (Q_3 - Q_1)/2$$

• Ambos índices tienen como ventaja respecto al *Rango* que no se ven afectados por la existencia de valores atípicos en la variable, pues no se obtienen a partir de los dos valores más extremos de la variable sino a partir de dos valores más centrados como son el Q_3 y el Q_1.

Ejemplo de obtención del *RIC* y del *RSIC* para la variable "Ansiedad que siente cuando se encuentra con mucha gente alrededor" (ver distribución de frecuencias más arriba).

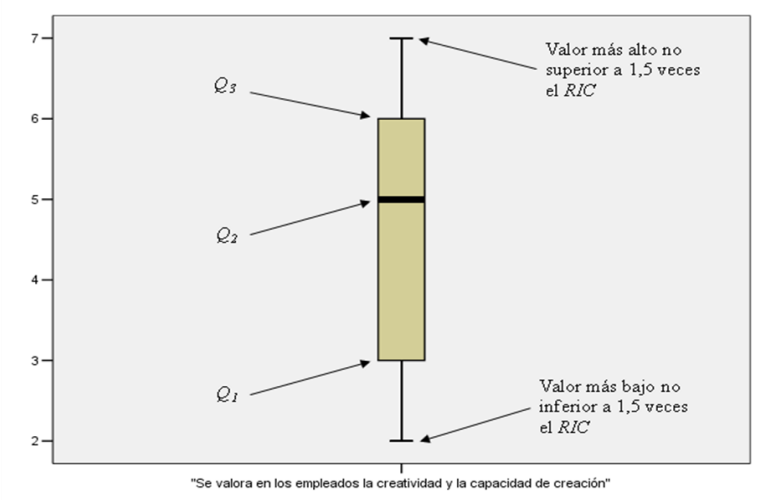
$$RIC = 3 - 2 = 1 \qquad RSIC = (3 - 2)/2 = 0{,}5$$

Ejercicio 4: Obtener el *RIC* y del *RSIC* de la variable "Se valora en los empleados la creatividad y la capacidad de creación" (ver distribución de frecuencias más arriba).

• Una representación gráfica de una variable basada en los Q_3 y Q_1 (y también en la mediana), cuya utilización está cada vez más extendida, es el conocido como gráfico de caja y bigotes, el cual ofrece información simultanea sobre la posición y variabilidad de la distribución de frecuencias de una variable. Como veremos más adelante, también ofrece información sobre la asimetría de la distribución, así como sobre la posible existencia de valores atípicos en los datos de la variable. Además es un gráfico muy utilizado con la finalidad de comparar grupos.

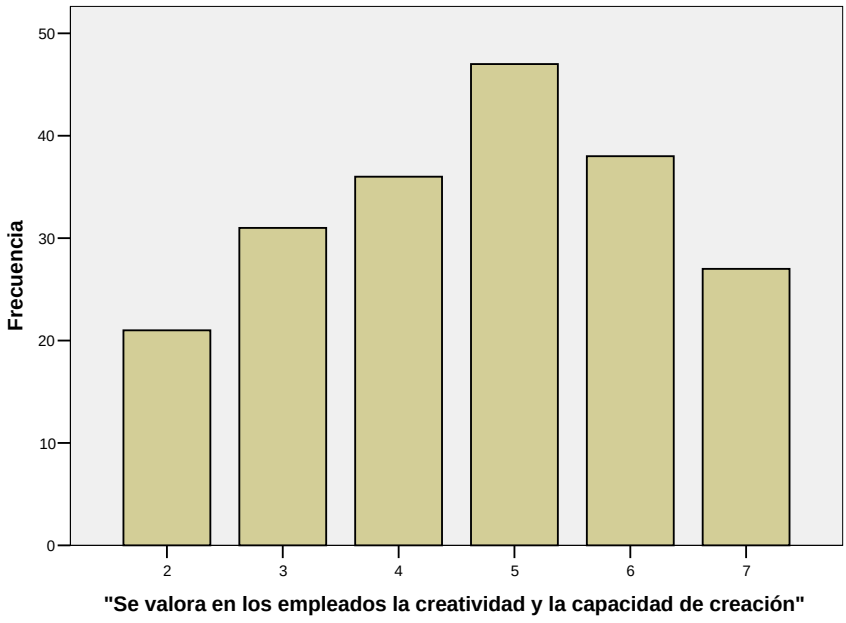
Como **ejemplo**, el gráfico de caja y bigotes de la variable "Se valora en los empleados la creatividad y la capacidad de creación" obtenida a partir de una muestra de 200 empleados:

BY NC SA



El mismo se construye situando la escala de respuesta de la variable en el eje vertical y dibujando una *caja* delimitada por la mediana y los cuartiles 1º y 3º (la distancia entre ambos es, precisamente, el rango intercuartil), y unos *bigotes* que se extienden hasta los valores más extremos de la variable que se encuentren dentro de 1,5 veces la longitud de la caja medida desde los lados de la misma. Los valores más allá de 1,5 veces la longitud de la caja, si existen, se representan por puntos y suelen indicar valores anómalos (atípicos o extremos) por lo raro de los mismos en relación al grueso de los datos.

Se muestra a continuación el gráfico de barras de la misma variable a fin de que pueda compararse con el correspondiente gráfico de caja y bigotes:



Ejercicio 5: Realizar la representación gráfica de la variable "Ansiedad que siente cuando se encuentra con mucha gente alrededor" con un gráfico de caja y bigotes.

3. Variables cuantitativas: la varianza, la desviación estándar y el coeficiente de variación

3.1. La varianza y la desviación estándar

• La distancia de los valores de una variable respecto a su media aritmética ofrece, de forma intuitiva, el fundamento para la propuesta de un índice de dispersión. Cuanto mayor sean esas distancias, más dispersos serán los datos; cuanto menor, más homogéneos resultarán ser.

Esa distancia de un dato (X_i) respecto a la media es más conocida en estadística como desviación o puntuación diferencial (d_i) correspondiente a ese dato ($d_i = X_i - \bar{X}$). Intuitivamente, el índice de dispersión más sencillo basado en este concepto consistiría en la obtención del promedio de las desviaciones ($\bar{d}_i$):

$$\bar{d}_i = \frac{\sum d_i}{n} = \frac{\sum (X_i - \bar{X})}{n}$$

Ejemplo de cálculo para la variable *X*: {6, 7, 4, 2, 5, 6}:

$$\frac{\sum (X_i - \bar{X})}{n} = \frac{(6-5)+(7-5)+(4-5)+(2-5)+(5-5)+(6-5)}{6} = \frac{0}{6} = 0$$

El resultado obtenido (0) resulta un tanto contradictorio, en cuanto que la simple observación de los datos nos dice que la dispersión de esa variable es cualquier cosa menos nula.

• En efecto, la anterior fórmula nos plantearía una contrariedad importante si la utilizáramos como índice de dispersión: siempre va a dar 0, sea cual sea el conjunto de datos que consideremos. Así, otras variantes de esa expresión han sido propuestas a fin de superar este inconveniente. Una muy utilizada es la varianza (S_X^2 *o* σ_X^2):

$$S_X^2 = \frac{\sum d_i^2}{n} = \frac{\sum (X_i - \bar{X})^2}{n}$$

A la expresión del numerador de esta fórmula se la conoce en la literatura estadística como *suma de cuadrados* (*SC*), por lo que la anterior fórmula puede quedar expresada como:

$$S_X^2 = SC_X / n$$

CC BY NC SA

Ejemplo de cálculo para la variable *X*: {6; 7; 4; 2; 5; 6}:

$$S_X^2 = \frac{\sum (X_i - \bar{X})^2}{n} = \frac{(6-5)^2 + (7-5)^2 + (4-5)^2 + (2-5)^2 + (5-5)^2 + (6-5)^2}{6} = \frac{16}{6} = 2,67$$

En el caso en que la varianza se obtenga a partir de una distribución de frecuencias:

$$S_X^2 = \frac{\sum n_i \cdot (X_i - \bar{X})^2}{n}$$

Ejemplo de cálculo de la varianza para la variable "Tiempo empleado en completar un laberinto" por una muestra de 20 ratas (n = 20):

Tiempo (seg)	n_i	p_i
9	3	0,15
10	8	0,4
11	6	0,3
12	2	0,1
13	1	0,05

$$\bar{X} = \frac{9 \cdot 3 + 10 \cdot 8 + 11 \cdot 6 + 12 \cdot 2 + 13 \cdot 1}{20} = 10,5\, seg$$

$$s_x^2 = \frac{3 \cdot (9-10,5)^2 + 8 \cdot (10-10,5)^2 + 6 \cdot (11-10,5)^2 + 2 \cdot (12-10,5)^2 + 1 \cdot (13-10,5)^2}{20} = 1,05\, seg^2$$

Una fórmula alternativa en el cálculo de la varianza a partir la información de una distribución de frecuencias consiste en sumar el producto de cada desviación al cuadrado por su frec. relativa:

$$S_X^2 = \sum p_i \cdot (X_i - \bar{X})^2$$

Ejemplo para la variable "Tiempo empleado en completar un laberinto":

$$S_x^2 = 0,15 \cdot (9-10,5)^2 + 0,4 \cdot (10-10,5)^2 + 0,3 \cdot (11-10,5)^2 + 0,1 \cdot (12-10,5)^2 + 0,05 \cdot (13-10,5)^2 = 1,05\, seg^2$$

• Al calcular la varianza de una variable, las unidades del valor resultante son el cuadrado de la unidad de medida de la variable en cuestión, lo cual complica la interpretación del mismo. La desviación típica o estándar (S_X o σ_X), al obtenerse como raíz cuadrada de la varianza, ya no tiene este inconveniente pues la unidad en que se exprese será la misma que la de la variable a partir de la que se haya obtenido.

$$S_X = \sqrt{S_X^2}$$

CC BY NC SA

Ejemplo de cálculo de la desviación estándar para la variable "Tiempo empleado en completar un laberinto":

$$S_X = \sqrt{1,05} = 1,02\,seg$$

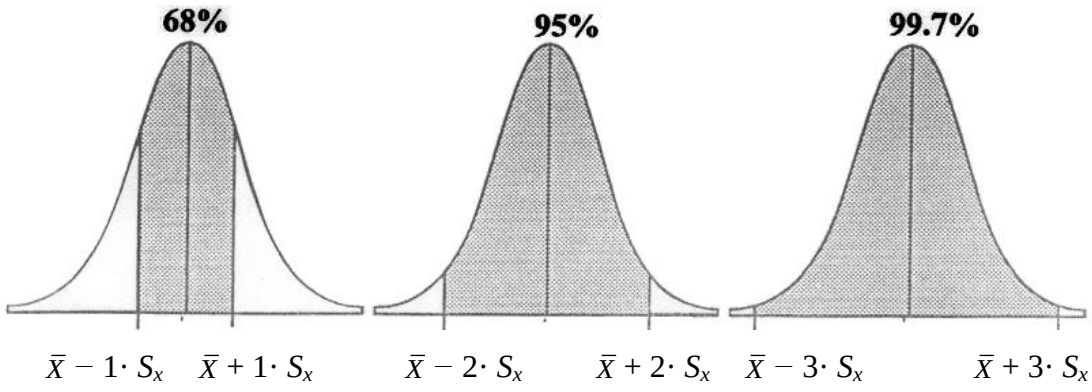
Ejercicio 6: Obtener la S_X^2 y S_X de una variable cuantitativa *X* para la que se han obtenido los siguientes datos para un grupo reducido de 7 sujetos: *X:* {6; 7; 4; 3; 5; 4; 6}

Ejercicio 7: Inventar 2 conjuntos de 6 datos (valores enteros entre 0 y 10, se pueden repetir) cada uno con $\bar{X}$ = 7 pero diferente s_x.

Ejercicio 8: Inventar 5 datos (valores enteros entre 0 y 10, se pueden repetir), que tengan S_x mínima (diferente de 0).

Ejercicio 9: Inventar 6 datos (enteros entre 0 y 10, se pueden repetir), que tengan S_x máxima.

• Una particularidad de la desviación estándar es que si tenemos una variable con una distribución de frecuencias que se ajuste a la curva normal (campana de Gauss), entonces resulta ya conocido el porcentaje de casos cuyos valores observados quedan entre los valores $\bar{X} \pm k$ veces la S_x. Por ejemplo, si k = 1 (es decir, la media ± una vez el valor de la desviación típica), podemos afirmar que el 68% de los sujetos tendrán sus valores en esa variable entre los valores $\bar{X} \pm 1 \cdot s_x$. Gráficamente, para k = 1, 2 y 3 en una variable *X* distribuida normalmente:



Ejercicio 10: Tras haber recogido datos de estatura para un grupo de 500 sujetos, se ha obtenido que la media es igual a 170 cm y la varianza igual a 81 cm^2. Sabiendo que la distribución de la variable se ajusta a la curva normal: (1) ¿entre qué valores de estatura están el 68% central de los sujetos?; (2) el 99,7 % central de los sujetos mide entre ... y ... ; (3) ¿cuántos sujetos miden entre 161 y 179 cm?

3.2. El coeficiente de variación

• La varianza o la desviación típica nos permiten comparar la dispersión de diferentes distribuciones de frecuencias obtenidas para una misma variable en diferentes grupos de sujetos. Por ejemplo, dados dos grupos de personas que vamos a llamar *G1* y *G2*, las desviaciones típicas de la variable

BY NC SA

Peso en ambos grupos, S_{Peso_G1} = 4,18 y S_{Peso_G2} = 14,55, evidencian la diferente dispersión de la variable *Peso* en los dos grupos en que ha sido medida, algo que podemos contrastar en este caso a simple vista si observamos más abajo los datos de ambas variables.

• Esa misma diferencia en variabilidad también se puede observar en los datos de los dos grupos de personas, *G5* y *G6*, en que fue medida la variable *Altura* (S_{Altura_G5} = 0,036 y S_{Altura_G6} = 0,227), poniéndose de manifiesto como los valores de la desviación estándar están intrínsecamente vinculados a la escala de medida de la variable considerada. Así, para la variable *Altura* son aparentemente bajos los valores de S_x, en comparación con los obtenidos para la variable *Peso*, aun cuando en el grupo *G6* existe una dispersión considerable en los valores observados de *Altura*, tal como se pone de manifiesto si observamos los datos originales de esta variable para ese grupo. Parece obvio que no resulta coherente comparar la dispersión de variables de diferente naturaleza con coeficientes que se expresan en las mismas unidades que las de las variables.

Nombre variable	n	Mínimo	Máximo	Rango	Media	Desv. típ.	CV
Peso_G1	5	70	81	11	75,00	4,18	5,57
Peso_G2	5	59	94	35	75,20	14,55	19,35
Peso_G3	5	4800	5100	300	4960,00	119,37	2,40
Peso_G4	5	4200	6800	2600	5180,00	1028,1	19,85
Altura_G5	5	1,68	1,77	0,09	1,72	,036	2,12
Altura_G6	5	1,45	1,98	0,53	1,74	,227	13,04

Peso_G1 (kg): {73; 77; 81; 74; 70}
Peso_G2 (kg): {65; 94; 86; 72; 59}
Peso_G3 (kg): {4800; 4950; 5100; 4900; 5050}
Peso_G4 (kg): {4200; 5500; 6800; 4500; 4900}
Altura_G5 (m): {1,70; 1,72; 1,77; 1,75; 1,68}
Altura_G6 (m): {1,45; 1,56; 1,98; 1,91; 1,80}

• Incluso la comparación de la variabilidad para diferentes subgrupos en una misma variable puede resultar desacertada en algunos casos al hacerla con la desviación estándar, en concreto, cuando se trate de subgrupos con medias bastante distintas en la variable en cuestión. Ello es debido a que suele haber en las variables una asociación entre la posición de los datos y su dispersión, de modo que datos de mayor magnitud están intrínsecamente vinculados a una mayor variabilidad. A modo de ejemplo, si miramos en la tabla las desviaciones típicas para la variable *Peso* medida en dos grupos de elefantes *G3* y *G4* (S_{Peso_G3} = 119,4 y S_{Peso_G4} = 1028,1), se observa como son valores muy elevados, por lo menos en comparación con los obtenidos con los dos grupos de personas para la variable *Peso*. Sin embargo, si nos fijamos en los datos de la variable *Peso_G3* se pone de manifiesto como, en realidad, se trata de un conjunto de datos muy homogéneo para lo que sería de esperar para una muestra de elefantes. Conclusión, si comparáramos las desviaciones típicas

correspondientes a las variables *Peso_G3* y *Peso_G2* podríamos llegar a conclusiones totalmente equívocas.

• Este problema de la comparación de la variabilidad de subgrupos con medias bien distintas puede soslayarse a través de un índice propuesto por K. Pearson, el coeficiente de variación (CV_X), el cual relativiza el peso de la desviación típica dividiéndola por la media (en consecuencia, no tiene unidades):

$$CV_X = \frac{S_X}{\bar{X}} \cdot 100$$

• En la práctica, el *CV* puede tomar cualquier valor por encima de 0, ahora bien, tal como señalan Solanas et al. (2005), es lo habitual que no tome valores superiores a 100: valores por encima pondrían de manifiesto una dispersión excepcionalmente alta en los datos. En ese caso, se aconseja indagar las fuentes de variabilidad de los datos, pues podría existir algún tipo de error o sesgo en la recogida de los datos que diera lugar a una dispersión tan elevada. Es más, Pardo, Ruiz y San Martín (2009) señalan que valores del *CV* superiores a 50 ya son indicativos de mucha dispersión.

Como se puede observar en la tabla de estadísticos para nuestro ejemplo, los valores del *CV* son más acordes a la realidad de los datos, por lo que se pone de manifiesto su conveniencia a la hora de comparar la variabilidad de subgrupos con medias diferenciadas. Es más, al tratarse de un coeficientes adimensional, puede resultar también útil para comparar la dispersión de variables distintas –cuando ello tenga sentido–, como podría ser el caso de las variables de Altura respecto a las de Peso en nuestro ejemplo.

Ejercicio 11: Obtener todas las medidas de dispersión presentadas en este tema para la variable "Nº de hijos" a partir de la distribución de frecuencias de los datos:

X_i	n_i
0	40
1	80
2	60
3	20

Ejercicio 12: Tenemos datos sobre el gasto anual en nuevas tecnologías en los colegios públicos de 2 ciudades ¿En cuál de las 2 ciudades presenta más dispersión esta variable? (Aplíquese el índice más apropiado para este caso)

Ciudad A	*Ciudad B*
$\bar{X}$ = 24000 € *Sx* = 3300 €	$\bar{X}$ = 15000 € *Sx* = 2900 €

BY NC SA

3.3. Algunas anotaciones sobre los índices de dispersión orientados a variables cuantitativas

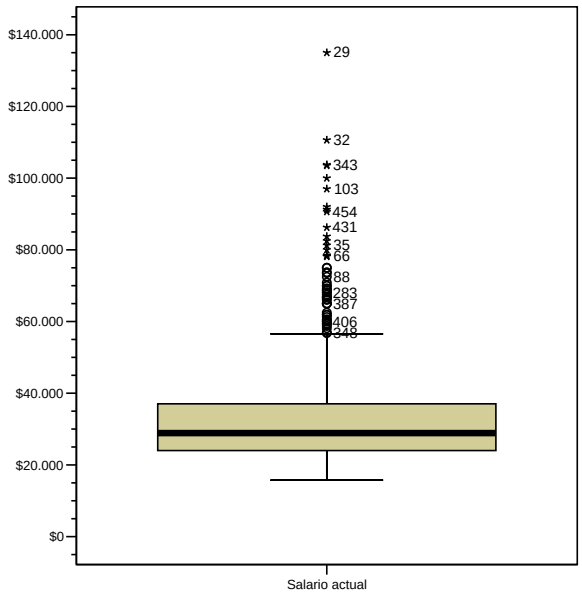
• Al estar basados en la media, se hace extensible a los mismos lo que se comentó al tratar este índice, en concreto, su sensibilidad a valores anómalos o atípicos, valores que se apartan en exceso del grueso de los valores (=> distribuciones de frecuencias muy asimétricas), por lo que se recomienda en estos casos aplicar el rango intercuartil.

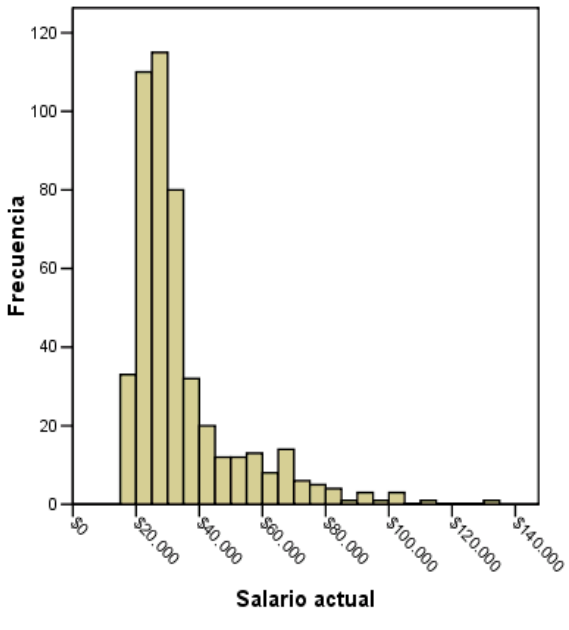
• No se debe olvidar, como ocurría en el tema precedente y ocurrirá en otros sucesivos, que los índices presentados para un determinado tipo de variable, también son aplicables para variables de orden superior –por ejemplo, los índices presentados para las variables categóricas se pueden aplicar a las variables ordinales y a las cuantitativas.

3.4. Visualización gráfica de la dispersión con variables cuantitativas

• Al igual que con las variables ordinales, el gráfico de caja y bigotes resulta también adecuado como representación gráfica de la posición y dispersión de variables cuantitativas. En variables cuya distribución contiene valores atípicos, el gráfico de caja y bigotes ofrece una identificación inequívoca de los mismos (los puntos que van más allá de los bigotes, esto es, los casos cuyo valor en la variable es superior a 1,5 veces el *RIC* más el Q_3, o bien, los casos cuyo valor es inferior a 1,5 veces el *RIC* menos el Q_1).

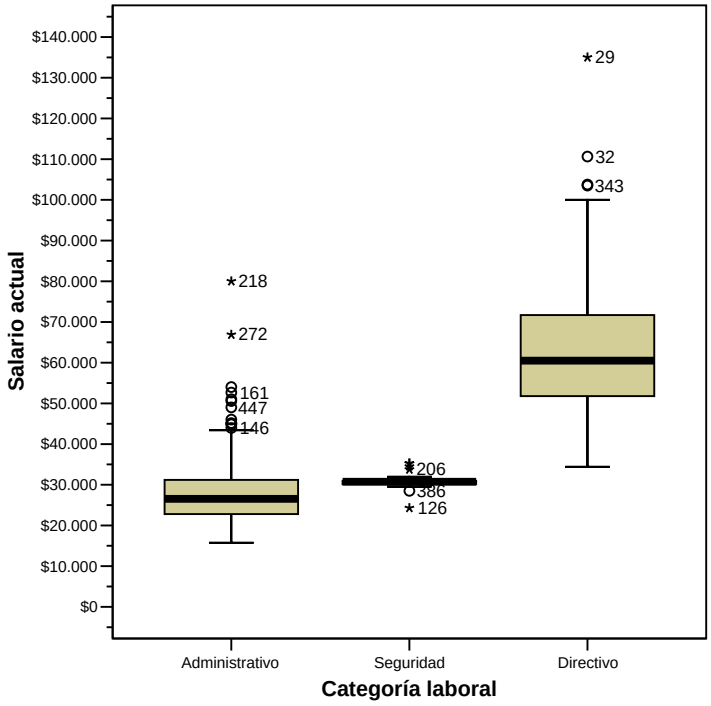
Ejemplo de gráfico de caja y bigotes con una distribución de frecuencias con valores atípicos (variable "Salario actual" para los 474 empleados de una empresa de servicios). Se muestra también el histograma de la misma variable a fin de que pueda compararse con el correspondiente gráfico de caja y bigotes





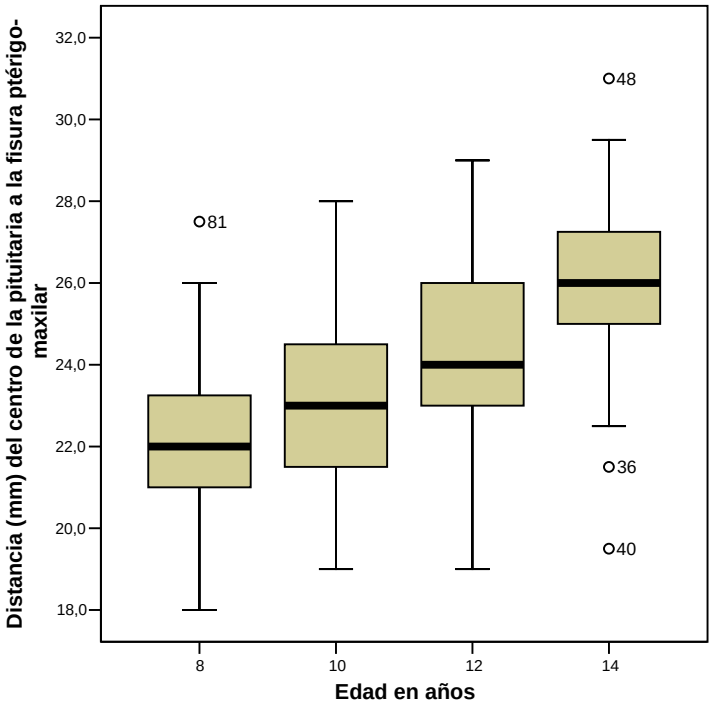
Nótese que, dada la presencia de valores tan atípicos en la distribución de esta variable, no sería adecuado en este caso describir la dispersión de la misma a partir de los índices de dispersión orientados a variables cuantitativas.

• Una faceta del análisis estadístico en que los gráficos de caja y bigotes resultan especialmente convenientes es para comparar la posición y variabilidad, bien de una misma variable medida en diferentes subgrupos de casos, bien de una misma variable medida en diferentes momentos temporales. A continuación se muestra un **ejemplo** del primer caso, en concreto se trata de la variable "Salario actual" para cada una de las tres categorías laborales diferenciadas en una empresa de servicios:



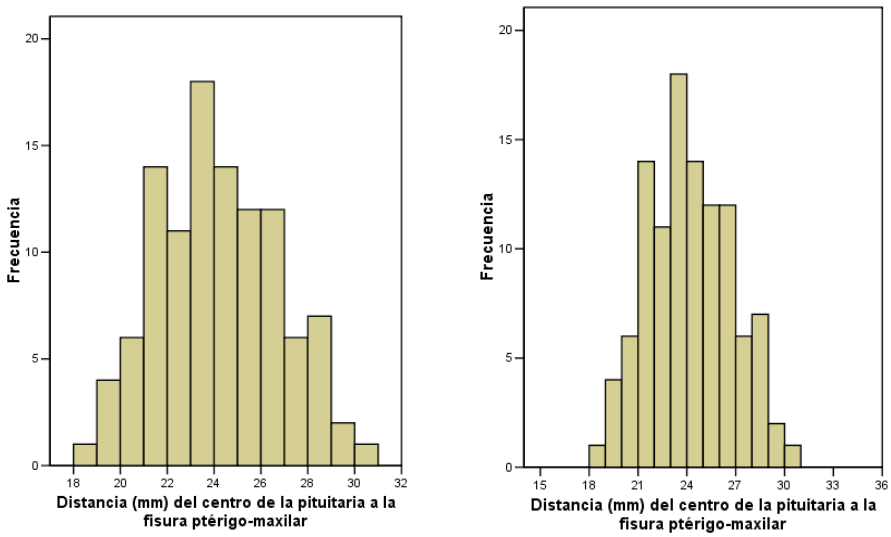
Otro **ejemplo** en que se comparan 4 subgrupos de sujetos definidos en función de la edad (8, 10, 12 y 14 años) en la variable "Distancia en mm del centro de la pituitaria a la fisura ptérigo-maxilar" por medio de un gráfico de caja y bigotes:

BY NC SA



Sería posible representar cada uno de los 4 grupos de sujetos mediante un histograma y comparar los mismos, sin embargo, el gráfico de caja y bigotes ofrece más ventajas en lo que al aprovechamiento del espacio gráfico se refiere. Además, se evita el problema de que cada subgrupo pueda estar representado en una escala diferente, pudiéndose provocar la percepción de diferencias no existentes

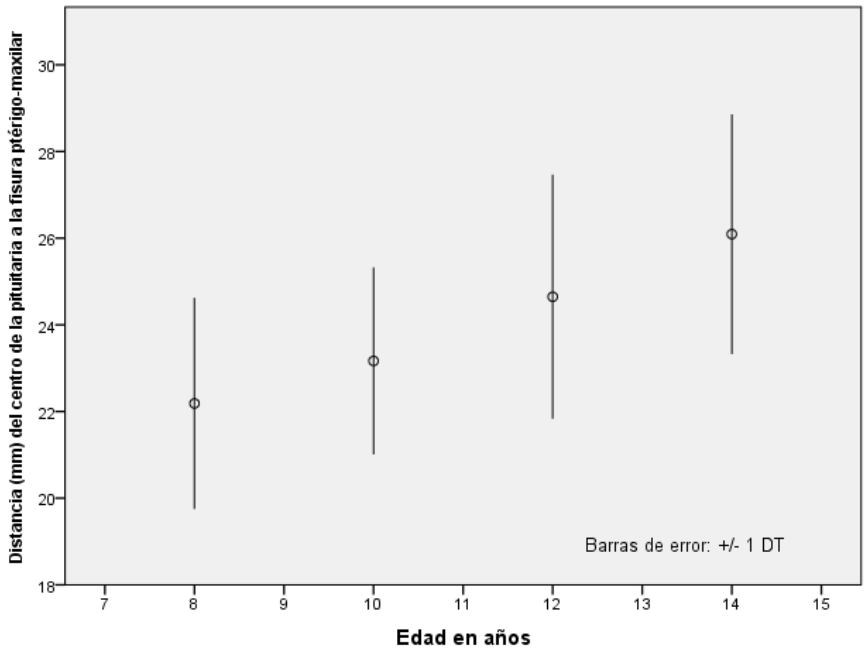
Ejemplo en que se comparan 2 subgrupos de sujetos en una misma variable (”Distancia en mm. del centro ...”) por medio de sendos histogramas:



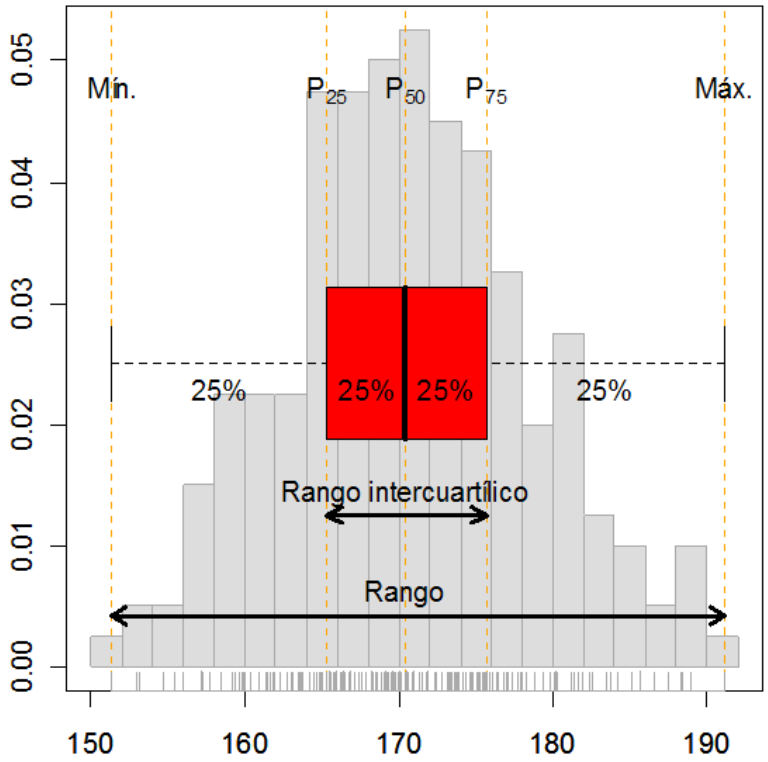
Si nos fijamos bien no se trata en realidad de 2 subgrupos sino del mismo, lo único que cambia de un histograma al otro es la escala del eje horizontal. Ahora bien, una primera impresión rápida podría habernos conducido a concluir erróneamente que los dos subgrupos tienen una posición grupal similar, siendo el segundo menos disperso en sus valores.

Consecuencia de ello, cuando se plantee utilizar histogramas a la hora de comparar grupos en una misma variable, se debe tener cuidado de que éstos sean representados con la misma escala en el eje de la variable.

• Una variante del gráfico de caja y bigotes es el conocido como gráfico de barras de error, en el cual se representa con un punto la media de la variable y, a partir de ese punto, se extienden dos líneas rectas verticales del mismo tamaño cuya longitud puede representar diferentes elementos de información estadística, por ejemplo, la desviación típica de la variable. También es habitual que en gráficos de este tipo se represente el error estándar o el intervalo de confianza, dos conceptos que se tratarán cuando se profundice en la Estadística Inferencial. En el gráfico del **ejemplo** que se muestra a continuación se muestran 4 barras de error para una variable ya representada anteriormente, la "Distancia en mm del centro de la pituitaria a la fisura ptérigo-maxilar", representando cada barra a uno de 4 subgrupos de sujetos de edades de 8, 10, 12 y 14 años, respectivamente.



• Tanto los gráficos de caja y bigotes como los gráficos de barras los podemos encontrar representados horizontalmente, como es el caso del que se muestra a continuación –el cual además aparece superpuesto sobre un histograma de la misma variable. Ambos muestran gráficamente la distribución de frecuencias de la variable "Altura (cm)" para una muestra de sujetos adultos. Destacar que el gráfico de caja y bigotes que aparece en esta figura es una versión simplificada de la versión original propuesta por John W. Tukey que puede encontrarse en algunos manuales de análisis de datos. Se diferencia del original en que los bigotes se extienden hasta el valor mínimo y máximo de la distribución, independientemente de lo alejados que estén de los lados de la caja.



Ejercicio 13: A partir del gráfico anterior, decir cuáles son los valores de los siguientes índices estadísticos: el mínimo y el máximo, el Q_1, la mediana, el P_{75}, la moda, el rango y el *RIC*.

Ejercicio 14: A continuación se muestra la distribución de frecuencias de la variable "Nº de visitas al servicio de urgencias hospitalario durante el pasado año", obtenida para una muestra de 150 sujetos diagnosticados con hipocondría. A partir de ésta: 1) dibujar el gráfico de barras y el de caja y bigotes para esta distribución de frecuencias. 2) decidir cuáles serían los índices de dispersión más adecuados para este caso y calcularlos.

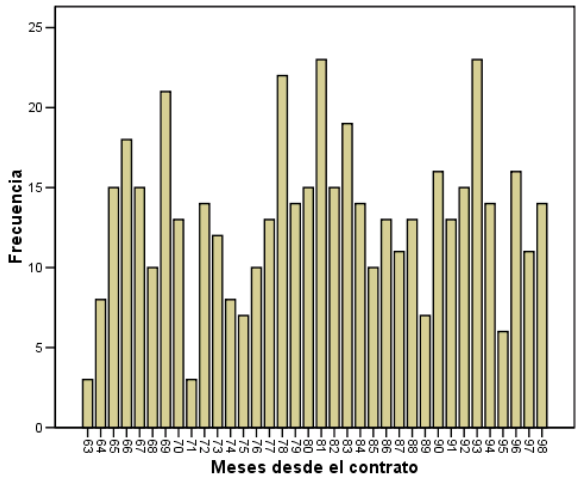
X_i	n_i	%	$\%_a$
0	11	7,33	7,33
1	30	20	27,33
2	41	27,33	54,66
3	27	18	72,66
4	19	12,67	85,33
5	14	9,33	94,66
6	5	3,33	98
7	2	1,33	99,33
10	1	0,67	100
	150	100	

Ejercicio 15: A continuación se muestra la distribución de frecuencias de la variable "Antigüedad en la empresa", medida a partir del "Nº de meses desde el contrato" para los 474 empleados de una empresa de servicios. Además se muestra el gráfico de barras y algunos estadísticos descriptivos

obtenidos con *SPSS*. A partir de esta información: 1) Obtener el gráfico de caja y bigotes. 2) Decidir cuáles serían los índices de dispersión más adecuados.

Meses desde el contrato

N	Válidos	474
	Perdidos	0
Media		81,11
Mediana		81,00
Moda		81[a]
Rango		35
Mínimo		63
Máximo		98
Percentiles	25	72,00
	50	81,00
	75	90,00



Meses desde el contrato

	Frecuencia	Porcentaje	Porcentaje acumulado
63	3	,6	,6
64	8	1,7	2,3
65	15	3,2	5,5
66	18	3,8	9,3
67	15	3,2	12,4
68	10	2,1	14,6
69	21	4,4	19,0
70	13	2,7	21,7
71	3	,6	22,4
72	14	3,0	25,3
73	12	2,5	27,8
74	8	1,7	29,5
75	7	1,5	31,0
76	10	2,1	33,1
77	13	2,7	35,9
78	22	4,6	40,5
79	14	3,0	43,5
80	15	3,2	46,6
81	23	4,9	51,5
82	15	3,2	54,6
83	19	4,0	58,6
84	14	3,0	61,6
85	10	2,1	63,7
86	13	2,7	66,5
87	11	2,3	68,8
88	13	2,7	71,5
89	7	1,5	73,0
90	16	3,4	76,4
91	13	2,7	79,1
92	15	3,2	82,3
93	23	4,9	87,1
94	14	3,0	90,1
95	6	1,3	91,4
96	16	3,4	94,7
97	11	2,3	97,0
98	14	3,0	100,0
Total	474	100,0	

Referencias

De Veaux, R. D., Bock, D. E. y Velleman, P. (2003). Intro Stats. Boston: Addison–Wesley.

Pardo, A., Ruiz, M. A. y San Martín, R. (2009). Análisis de datos en ciencias sociales y de la salud I. Madrid: Síntesis.

BY NC SA

Solanas, A., Salafranca, L., Fauquet, J. y Núñez, M. I. (2005). Estadística descriptiva en Ciencias del Comportamiento. Madrid: Thompson.

BY NC SA