PONTIFICIA UNIVERSIDAD CATOLICA DE VALPARAISO
INSTITUTO DE ESTADISTICA

CALCULO DE MEDIDAS DE RESUMEN CON DATOS TABULADOS

Jorge Galbiati Riesco

Si los datos se presentan en tablas de frecuencias por intervalos, se pueden obtener valores aproximados de las medidas de resumen, utilizando las respectivas marcas de clase como valores representativos de todas las observaciones agrupadas en cada uno de los intervalos. Las marcas de clase son los valores promedio entre los límites inferior y superior de los respectivos intervalos. Por esa razón los valores obtenidos no son exactos, pero se aproximan debido a que los errores por defecto tienden a cancelarse con los errores por exceso.

Veremos el caso de la media, la desviación estándar y los percentiles.
Supóngase que los datos se presentan en una tabla de frecuencias con K intervalos, como sigue:

intervalo número (j)	Límite inferior (LI_j)	Límite superior (LS_j)	Marca de clase (x_j)	frecuencias (f_j)	Frecuencias acum. (F_j)
1	LI_1	LS_1	x_1	f_1	F_1
2	LI_2	LS_2	x_2	f_2	F_2
3	LI_3	LS_3	x_3	f_3	F_3
..	..	..	..	..	..
K	LI_K	LS_k	x_K	f_K	F_K

La media.

La media (aproximada) se calcula como el promedio ponderado de las marcas de clase por las respectivas frecuencias.

$$\bar{x} = \frac{\sum_{j=1}^{K} x_j f_j}{n}$$

donde *n* es el total de observaciones, obtenido sumando las frecuencias:

$$n = \sum_{j=1}^{K} f_j$$

No confundir *K*, el número de intervalos, con *n*, el número de observaciones. Este último es mucho más grande que *k*.

La varianza y la desviación estándar.

La varianza se obtiene también a partir de las marcas de clase, ponderando por las frecuencias:

$$s^2 = \frac{\sum_{j=1}^{K} f_j x_j^2 - n \cdot \bar{x}^2}{n-1}$$

La desviación estándar es, igual que siempre, la raíz cuadrada de la varianza.

EJEMPLO 1

La Tabla 1 presenta los **puntajes corregidos** obtenidos por los 153.470 jóvenes que rindieron la Prueba de Selección Universitaria (PSU) Matemática en el año 2003. Estos consisten en el número de preguntas respondidas correctamente menos un cuarto por el número de preguntas respondidas incorrectamente. Las no respondidas no se cuentan.
Como la prueba consta de 70 preguntas, el rango de **puntajes corregidos** tiene un rango entre -17.5 (todas incorrectas) y 70 (todas correctas), variando en 0.25.
La tabla 1 muestra los intervalos, las marcas de clase, frecuencias y frecuencias acumuladas, en ambos casos absolutas y relativas.
El intervalo 3, entre 7.5 y 14.5 puntos, se denomina **intervalo modal**, por ser el que tiene mayor frecuencia.
A partir de esta tabla se calcularán la media, la varianza y la desviación estándar.

Int. num. j	limite inferior li	limite superior ls	marca de clase x	frec f	frec acum. F	frec relativa fr	frec. rel. acum Fr
1	-6,5	0,5	-3	2658	2658	0,0173	0,0173
2	0,5	7,5	4	24630	27288	0,1605	0,1778
3	7,5	14,5	11	31081	58369	0,2025	0,3803
4	14,5	21,5	18	21424	79793	0,1396	0,5199
5	21,5	28,5	25	16798	96591	0,1095	0,6294
6	28,5	35,5	32	14242	110833	0,0928	0,7222
7	35,5	42,5	39	11199	122032	0,0730	0,7952
8	42,5	49,5	46	9454	131486	0,0616	0,8568
9	49,5	56,5	53	8521	140007	0,0555	0,9123
10	56,5	63,5	60	7340	147347	0,0478	0,9601
11	63,5	70,5	67	4910	152257	0,0320	0,9921
12	70,5	77,5	74	1213	153470	0,0079	1,0000
Totales				153470		1,0	

TABLA 1. Distribución de frecuencias de los puntajes corregidos de la PSU Matemática, 2003.

La Figura 1 muestra un histograma de estos datos. Se puede apreciar un sesgo pronunciado hacia los valores grandes.

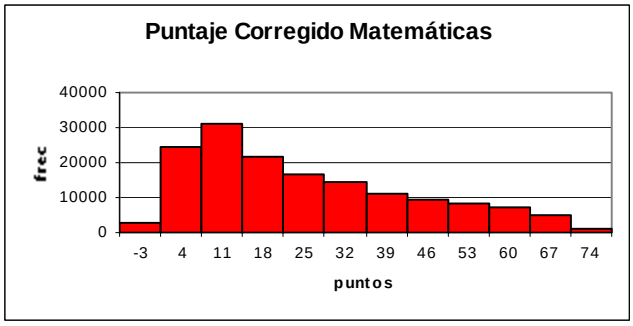

FIGURA 1. Histograma de los puntajes corregidos PSU Matemática 2003.

La tabla 2 muestra los cálculos intermedios para la obtención de la media y varianza.

Int. num. j	marca de clase x	frec f	f*x	f*x2
1	-3	2658	-7974	23922
2	4	24630	98520	394080
3	11	31081	341891	3760801
4	18	21424	385632	6941376
5	25	16798	419950	10498750
6	32	14242	455744	14583808
7	39	11199	436761	17033679
8	46	9454	434884	20004664
9	53	8521	451613	23935489
10	60	7340	440400	26424000
11	67	4910	328970	22040990
12	74	1213	89762	6642388
Totales	--	153470	3876153	152283947

TABLA 2. Puntajes corregidos de la PSU Matemática 2003, Cálculos intermedios.

Con los totales, mostrados en la última fila de la tabla, se pueden calcular las medidas de resumen aproximadas.

La media es $\bar{x} = \frac{3876153}{153470} = 25.257$

La varianza es $s^2 = \frac{152283947 - 153470 \times 25.257^2}{153470 - 1} = 354.371$

y la desviación estándar es $s = \sqrt{354.3708} = 18.825$

EJEMPLO 2

La Tabla 3 presenta los **puntajes de la PSU** obtenidos por los mismos 153.470 jóvenes de la Tabla 1. Estos puntajes son el resultado de aplicar una transformación no lineal a los puntajes corregidos.

El rango de **puntajes PSU** va de 200 a 800 puntos.

El intervalo modal es el intervalo 9, entre 500 y 549 puntos.

Con esta tabla se calcularán la media, la varianza y la desviación estándar.

Int. núm. j	limite inferior li	limite superior ls	marca de clase x	frec f	frec acum. F	frec relativa fr	frec. rel. acum Fr
1	100	149	125	193	193	0,0013	0,0013
2	150	199	175	251	444	0,0016	0,0029
3	200	249	225	1156	1600	0,0075	0,0104
4	250	299	275	2747	4347	0,0179	0,0283
5	300	349	325	9152	13499	0,0596	0,0880
6	350	399	375	10718	24217	0,0698	0,1578
7	400	449	425	24176	48393	0,1575	0,3153
8	450	499	475	27609	76002	0,1799	0,4952
9	500	549	525	28480	104482	0,1856	0,6808
10	550	599	575	22830	127312	0,1488	0,8296
11	600	649	625	14183	141495	0,0924	0,9220
12	650	699	675	6223	147718	0,0405	0,9625
13	700	749	725	2721	150439	0,0177	0,9803
14	750	799	775	1822	152261	0,0119	0,9921
15	800	849	825	1209	153470	0,0079	1,0000
	Totales	--	--	153470	--	1,0000	--

TABLA 3. Distribución de frecuencias de los puntajes PSU Matemática, 2003.

La Figura 2 muestra el histograma de estos datos. Se puede observar que es bastante simétrico, lo que se logró por medio de la transformación no lineal que se aplicó a los puntajes corregidos. En el eje horizontal aparecen las marcas de clase de los intervalos de orden impar.

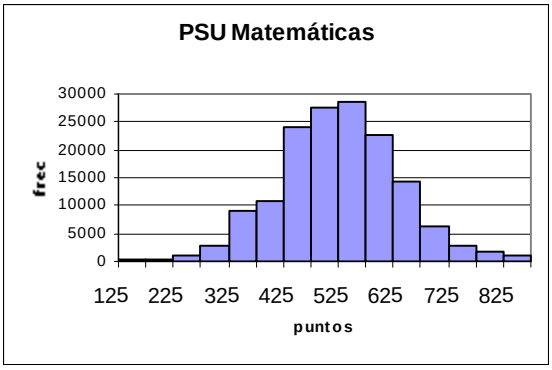

FIGURA 2. Histograma de los puntajes PSU Matemática 2003.

En la tabla 4 se muestran los cálculos intermedios para la obtención de la media y varianza.

Int. núm. j	marca de clase x	frec f	f*x	f*x2
1	125	193	24125	3015625
2	175	251	43925	7686875
3	225	1156	260100	58522500
4	275	2747	755425	207741875
5	325	9152	2974400	966680000
6	375	10718	4019250	1507218750
7	425	24176	10274800	4366790000
8	475	27609	13114275	6229280625
9	525	28480	14952000	7849800000
10	575	22830	13127250	7548168750
11	625	14183	8864375	5540234375
12	675	6223	4200525	2835354375
13	725	2721	1972725	1430225625
14	775	1822	1412050	1094338750
15	825	1209	997425	822875625
	--	153470	76992650	40467933750

Se obtienen las medidas de resumen con los totales:

La media es $\bar{x} = \frac{76992650}{153470} = 501.68$

La varianza es $s^2 = \frac{40467933750 - 153470 \times 501.68^2}{153470 - 1} = 12004.7$

y la desviación estándar es $s = \sqrt{12004.7} = 109.6$

Percentiles

La forma de calcular los percentiles con datos tabulados es algo más complicada. Supongamos que buscamos el percentil P_q , en que *q* es un número entero entre 1 y 99. Se siguen los cuatro pasos siguientes:

1.- Primero se debe calcular la posición del percentil. La posición del percentil es $R=q*n/100$ aproximado al entero más cercano, con *n* el número total de observaciones.

2.- Luego se debe buscar en qué intervalo está el percentil, para lo cual se observa la columna de frecuencias acumuladas de la tabla. En esa columna se busca el menor valor que sea mayor o igual a R.
El percentil está en el intervalo correspondiente a ese valor, intervalo que designaremos con el índice k.

3.- El percentil es el dato que ocupa la posición R. La frecuencia acumulada hasta el intervalo anterior al del percentil es F_{k-1}, menor o igual que R. Entonces debemos avanzar F_{k-1}-R lugares para alcanzar el percentil dentro del intervalo.

4. - Sea f_k la frecuencia absoluta (no acumulada) del intervalo La fracción de intervalos que falta para completar el número buscado *R*, es $(F_{k-1}-R)/f_k$.

5.- Vamos a asumir que las f_k observaciones están a igual distancia en el intervalo, de longitud *long*. Este supuesto no se cumple necesariamente, por lo que el valor del percentil que resulta es aproximado, al igual que todos los cálculos efectuados con tablas de frecuencia de datos agrupados por intervalos.

6.- Bajo este supuesto, la distancia que se debe recorrer desde el límite inferior para llegar al percentil es la fracción $(F_{k-1}-R)/f_k$ por la longitud de los intervalo, *long.*

7.- El percentil buscado es el límite inferior del intervalo k, $Linf_k$ más $(F_{k-1}-R)/f_k$ veces la longitud del intervalo, *long*. La fórmula que da un valor aproximado del percentil, es

$$P_q = L\inf_k + \frac{(R - F_{k-1})}{f_k} \times long$$

en que

$Linf_k$ es el límite inferior del intervalo que contiene al percentil,

f_k es la frecuencia del intervalo,

F_{k-1} la frecuencia acumulada del intervalo inmediatamente anterior,

long es la longitud de los intervalos,

R es *q*n*/100 aproximado al entero,

n el número de observaciones.

EJEMPLO 3

Obtendremos los cuartiles 1 y 3 y la mediana, para los datos tabulados del Ejemplo 1, correspondientes a los puntajes corregidos.

Cuartil 1:

Lugar que ocupa el cuartil 1: $R = q*n/100 = 0.25 \times 153470 = 38368$

Observando la Tabla 1, vemos que la frecuencia acumulada del intervalo 2 llega hasta F_2 = 27288, menos que R, y del intervalo 3 llega a F_4 = 58369. Por lo tanto el cuartil 1 está en el intervalo 3, entre 7.5 y 14.5 puntos corregidos.

Su frecuencia absoluta es f_3 = 31081.

La frecuencia acumulada hasta el intervalo anterior es F_2 = 27288, luego lo que le falta para llegar al cuartil 1 es $R\text{-}F_3$ = 38368 - 27288 = 11080

Luego el valor de la fracción es $\frac{(R-F_3)}{f_4} = \frac{11080}{31080} = 0.3565$

El límite inferior del intervalo 3 es $Linf_3$ = 7.5

La longitud de los intervalos es *long*=7

Por lo tanto el cuartil 1 es $P_{25} = L\inf_3 + \frac{(R-F_2)}{f_3} \times long = 7.5 + 0.3565 \times 7 = 10.0$

Entonces Q3=10.0 puntos corregidos.

Mediana:

Lugar que ocupa la mediana: $R = 0.50 \times 153470 = 76735$

De la Tabla 1, la frecuencia acumulada del intervalo 3 llega hasta F_3 = 58369, y del intervalo 4 llega a F_4 = 79793. Por lo tanto la mediana está en el intervalo 4, entre 14.5 y 21.5 puntos corregidos.

Su frecuencia absoluta es f_4 = 21424

Lo que falta para llegar a la mediana es $R\text{-}F_3$ = 76735 - 58369 = 18366

El valor de la fracción es $\frac{(R-F_3)}{f_4}=\frac{18366}{21424}=0.8573$

El límite inferior del intervalo 4 es $Linf_4$ = 14.5

Por lo tanto la mediana es $P_{50}=L\inf_4+\frac{(R-F_3)}{f_4}\times long=$ 14.5 + 0.8573 x 7 = 20.5

Entonces Mn=20.5 puntos corregidos.

Cuartil 3:

Lugar que ocupa el cuartil 3: R = 0.75 x 153470 = 115103

De la Tabla 1, el cuartil 3 está en el intervalo 6, entre 28.5 y 35.5 puntos corregidos.

f_6 = 14242

$R\text{-}F_7$ = 115103 - 96591 = 18512

$\frac{(R-F_5)}{f_6}=\frac{18512}{14242}=1.2998$

El límite inferior del intervalo 6 es $Linf_6$ = 28.5

Por lo tanto la mediana es $P_{75}=L\inf_6+\frac{(R-F_5)}{f_6}\times long=$ 28.5 + 1.2998 x 7 = 37.6

Entonces el cuartil 3 es 37.6 puntos corregidos.

Resumiendo, el cuartil 1, la mediana y el cuartil 3 son, respectivamente, 10.0, 20.5 y 37.6. Esto se interpreta de la siguiente manera: El 25% de los jóvenes obtuvo 10.0 puntos corregidos o menos. Otro 25% estuvo entre 10.0 y 20.5 puntos. Otro 25% obtuvo entre 20.5 y 37.6 puntos. Y finalmente, un 25% obtuvo más de 37.6 puntos.
La distancia entre el cuartil 1 y la mediana es 10.5, menor que la distancia entre la mediana y el cuartil 3, que es 17.1. Esto es por el sesgo a la derecha de los datos.

EJEMPLO 4

Obtendremos los cuartiles 1 y 3 y la mediana, para los datos tabulados del Ejemplo 2, correspondientes a los puntajes PSU.

Cuartil 1:

Lugar que ocupa el cuartil 1: R = 0.25 x 153470 = 38368

De la Tabla 3, el cuartil 1 está en el intervalo 7, entre 400 y 449 puntos PSU.

f_7 = 24176 y F_6 = 24217.

$R\text{-}F_6$ = 38368 - 24217 = 14151

$$\frac{(R-F_6)}{f_7}=\frac{24217}{24176}=0.5853$$

El límite inferior del intervalo 8 es $Linf_8 = 400$

La longitud de los intervalos es *long*=50

Por lo tanto la el cuartil 1 es $P_{25}=L\inf_7+\frac{(R-F_6)}{f_7}\times long = 400 + 0.5853 \text{ x } 50 = 429.3$

Entonces el cuartil 1 es 429.3 puntos PSU.

Mediana:

Lugar que ocupa la mediana: $R = 0.50 \text{ x } 153470 = 76735$

De la Tabla 3, la mediana está en el intervalo 9, entre 500 y 549 puntos PSU.

$f_9 = 28480$ y $F_8 = 76002$

$R\text{-}F_8 = 76735 - 76002 = 733$

$$\frac{(R-F_8)}{f_9}=\frac{733}{28480}=0.0257$$

El límite inferior del intervalo 9 es $Linf_9 = 500$

Por lo tanto la mediana es $P_{50}=L\inf_9+\frac{(R-F_8)}{f_9}\times long = 500 + 0.0257 \text{ x } 50 = 501.3$

Entonces la mediana es 501.3 puntos PSU.

Cuartil 3:

Lugar que ocupa el cuartil 3: $R = 0.75 \text{ x } 153470 = 115103$

De la Tabla 3, el cuartil 3 está en el intervalo 10, entre 549 y 599 puntos PSU.

$f_{10} = 22830$ y $F_9 = 104482$

$R\text{-}F_9 = 115103 - 104482 = 10621$

$$\frac{(R-F_9)}{f_{10}}=\frac{10621}{22830}=0.4652$$

El límite inferior del intervalo 10 es $Linf_{10} = 550$

Por lo tanto la mediana es $P_{75}=L\inf_{10}+\frac{(R-F_9)}{f_{10}}\times long = 550 + 0.4652 \text{ x } 50 = 573.3$

Entonces el cuartil 3 es 573.3 puntos PSU.

En resumen, el cuartil 1, la mediana y el cuartil 3 son, respectivamente, 429.3, 501.3 y 573.3 puntos PSU. Es decir: El 25% de los jóvenes obtuvo 429.3 puntos o menos. Otro 25% estuvo entre 429.3 y 501.3 puntos. Otro 25% obtuvo entre 501.3 y 573.3 puntos. Y

finalmente, un 25% obtuvo más de 573.3 puntos. En este caso la diferencia entre el cuartil 1 y la mediana es igual a la diferencia entre la mediana y el cuartil 3, 72.0 puntos PSU. Esto refleja la simetría de la distribución de frecuencia de los datos.
