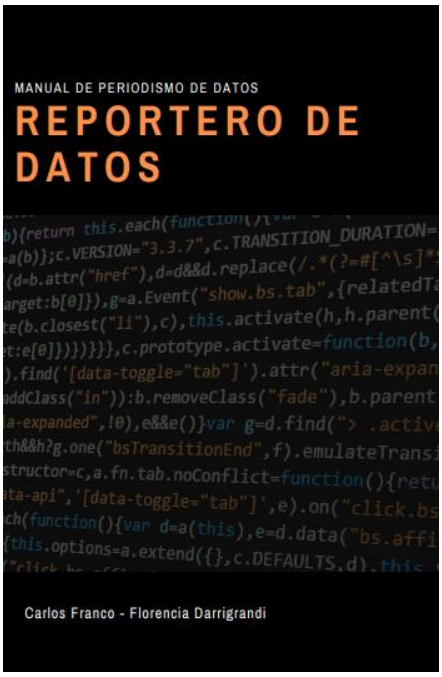


Media, mediana, moda y otras medidas de resumen

Por lo general en estadística se trabaja con grandes volúmenes de datos y para poder entender el comportamiento de ellos y encontrar patrones es necesario sintetizar esta información. Las medidas de resumen como su nombre lo indica resumen en una sola cifra toda la información contenida en una variable.

Las medidas de resumen se dividen entre grupos.

- Medidas de tendencia central
- Medidas de dispersión
- Medidas de posición

Las medias de tendencia central son la media (promedio), mediana y moda. Son llamadas así dado que representan un punto central en torno al cual se encuentran las observaciones.

Las medidas de dispersión cuantifican la variabilidad de los datos. Las más usadas son la varianza, la desviación estándar y el rango

Las medidas de posición reciben ese nombre pues ayudan a comprender, valga la redundancia, cuál es la posición de una observación con respecto al conjunto total de observaciones. Para ello se divide el conjunto total de observaciones en subgrupos con el mismo número de datos.

Las medias de posición más usuales son los percentiles, cuartiles, quintiles y deciles.

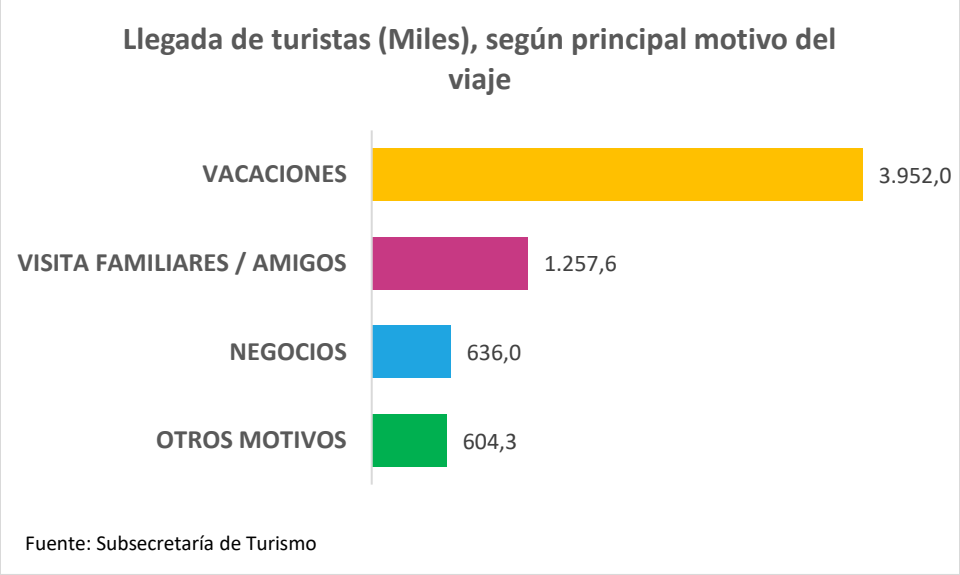
A continuación, discutiremos acera de estas medidas utilizando notas que han aparecido en los medios en este último tiempo.

Medidas de tendencia central

Como ya se mencionó con anterioridad las medidas de tendencia central son la media, la mediana y la moda. Las tres buscan representar donde se encuentran centrados los datos. La discusión acerca de estas medidas la centraremos en torno de cuándo es más recomendable usar una medida u otra.

La moda es la observación que más se repite o en otras palabras la de mayor frecuencia. La moda suele usarse principalmente para variables categóricas y corresponde a la categoría de mayor frecuencia.

El grafico que se muestra a continuación muestra información del estudio Turismo Receptivo 2017 del la Subsecretaría de Turismo (Fuente: http://www.subturismo.gob.cl/wp-content/uploads/2015/10/TURISMO-RECEPTIVO-ANUAL-2017.pdf). Motivo del viaje corresponde a una variable categórica donde las categorías son vacaciones, visita familiares /amigos, negocios y otros motivos. Observemos que la categoría de mayor frecuencia es vacaciones con 3.952.000 observaciones. Es decir la moda o categoría modal es Vacaciones.



No suele usarse para variables cuantitativas ya que en estos casos puede que haya varias modas o que no haya ninguna. Para ejemplificar esta situación consideremos el siguiente conjunto de observaciones que corresponden a las edades de una muestra de 25 personas que asistieron a una conferencia.

47	50	55	31	39
45	34	48	42	53
35	52	50	46	40
37	48	30	29	44
52	28	49	40	41

Observe que el 40, 48, 50 y 52 se repiten dos veces. El resto de las edades no se repiten por lo que 49, 48, 50 y 52 serían las modas.

Consideremos ahora otro conjunto de observaciones que también corresponden a las edades de una muestra de 25 personas que asistieron a una conferencia.

47	50	55	31	39
45	34	48	42	53
35	36	33	46	40
37	32	30	29	44
52	28	49	27	41

Observe que en este caso no se repite ningún valor por lo que no hay moda.

Cuando la variable es cuantitativa se recomienda usar la media o la mediana como medida de tendencia central.

La media es lo mismo que el promedio y suele denotarse por $\bar{x}$. Se calcula, primeramente sumando todos los datos y luego dividiendo por la cantidad de observaciones. La mediana es un valor tal que bajo ella se encuentran al meno el 50% de las observaciones. Con el fin de ilustrar este concepto calcularemos la mediana del siguiente conjunto de datos:

47	50	55	31	39
45	34	48	42	53
35	36	33	46	40
37	32	30	29	44
52	28	49	27	41

Para calcular la mediana lo primero que debemos hacer ordenar los datos de menor a mayor:

27 28 29 30 31 32 33 34 35 36 37 39 **40** 41 42 44 45 46 47 48 49 50 52 53 55

Como hay un número impar de observaciones, la mediana es justo el valor que se encuentra en la mitad de los datos. Dicho de otra forma, la mediana es el valor que se encuentra en la posición $\frac{n+1}{2}$ donde n es la cantidad de observaciones.

¿Cómo se interpreta? Esto quiere decir que el 50% de los asistentes tenía a lo más 40 años.

Si hubiera un número par de observaciones, la mediana sería el promedio de las dos observaciones centrales.

27 28 29 30 31 32 33 34 35 36 37 39 **40** **41** 42 44 45 46 47 48 49 50 52 53 55 57

En este caso sería $\frac{40+41}{2} = 40{,}5$. Esto quiere decir que el 50% de los asistentes tenía a los más 41,5 años.

El cálculo aquí realizado solo fue hecho con fin ilustrativo ya que hoy en día la mediana, al igual que la media, se puede calcular usando software estadístico o inclusive Excel.

Hay casos en que es más recomendable usar la media y otros la mediana.

Consideremos la siguiente nota publicada el 6 de mayo del 2019 en el portal del Diario Concepción titulada "*Fundación Sol: mediana de sueldos en la Región del Bío Bío sólo alcanza a los $300.000*" (Fuente: https://www.diarioconcepcion.cl/economia-y-negocios/2019/05/05/fundacion-sol-mediana-de-sueldos-en-la-region-del-bio-bio-solo-alcanza-a-los-300-000.html). En ella se menciona lo siguiente. "*"La mediana, es decir, el umbral de ingresos/salarios para el 50% de los trabajadores de Chile, es $350.000 líquidos, lo que equivale a sólo dos tercios del ingreso promedio y da cuenta de que en países como Chile que presentan altos niveles de desigualdad, el promedio no es un valor representativo", explica Kremerman*."

Lo que está detrás de la explicación de Kremerman es que la media es sensible a la presencia de datos atípicos o extremos en tanto que la mediana no. Como dato atípico o extremo entenderemos a aquel que se aleja, ya sea por que es mucho más grande o mucho más chico, del resto de los datos.

Por ejemplo, si 21, 22; 22; 23 y 23 son las edades de 5 alumnos de un curso de periodismo, en este caso la mediana es 22 y el promedio 22,2.

¿Qué pasa si se integra un nuevo alumno mucho mayor, digamos de 55 años? La mediana ahora sería 22,5 (observe que casi no varió con respecto al valor anterior), pero el promedio ahora sería 27,67. La media, en relación a la mediana, sufrió una mayor variación, pero además, ¿podemos decir que 27,67 representa la edad de los alumnos de periodismo? La respuesta es no, ya que la mayoría de las edades están entre los 21 y 23 años. Por lo que en este caso la mediana es la medida de tendencia central más apropiada para representar este conjunto de datos.

Cuando hay presencia de datos extremos se recomienda utilizar la mediana como medida de tendencia central pues esta será más representativa que la media.

Medidas de dispersión

Las medida de dispersión por si solas no representan en su totalidad el comportamiento de un conjunto de datos.

Consideremos la siguiente nota publicada por el Diario Financiero el 12 de marzo del 2019 (Fuente: **https://www.df.cl/noticias/mercados/pensiones/ciedess-fondos-de-pensiones-le-ganan-a-la-bolsa-chilena-exponiendose-a/2019-03-12/121819.html)**

Ciedess: Fondos de pensiones le ganan a la bolsa chilena exponiéndose a un menor riesgo

¿Qué significará que los fondos de pensiones se expongan a menor riesgo? ¿Cómo se evalúa el riesgo de un fondo?

El riesgo de un fondo está asociado a la variabilidad de su rendimiento. Mientras más variable sea, más riesgoso será.

La nota del Diario Financiero señala lo siguiente:

"*Muchas veces nos preguntamos si los fondos de pensiones están rentando bien o si conviene refugiarse en un fondo conservador cuando la volatilidad se toma los mercados. Bueno, un estudio de Ciedess que analizó el riesgo y el retorno de los fondos entre 2002 y 2018, afirma que como alternativa de inversión a largo plazo, "los multifondos destacan por sus buenos resultados y la posibilidad de elegir la alternativa de acuerdo a las preferencias de cada afiliado".*

Se suma el hecho de que el riesgo asumido para alcanzar dichos retornos es menor que al que uno se expone invirtiendo directamente en los índices. Como referencia toma el desempeño del IGPA, indicador que agrupa a gran parte de las acciones transadas en la Bolsa de Comercio de Santiago. "Todos los fondos poseen rentabilidades positivas y un riesgo menor al del mercado"".

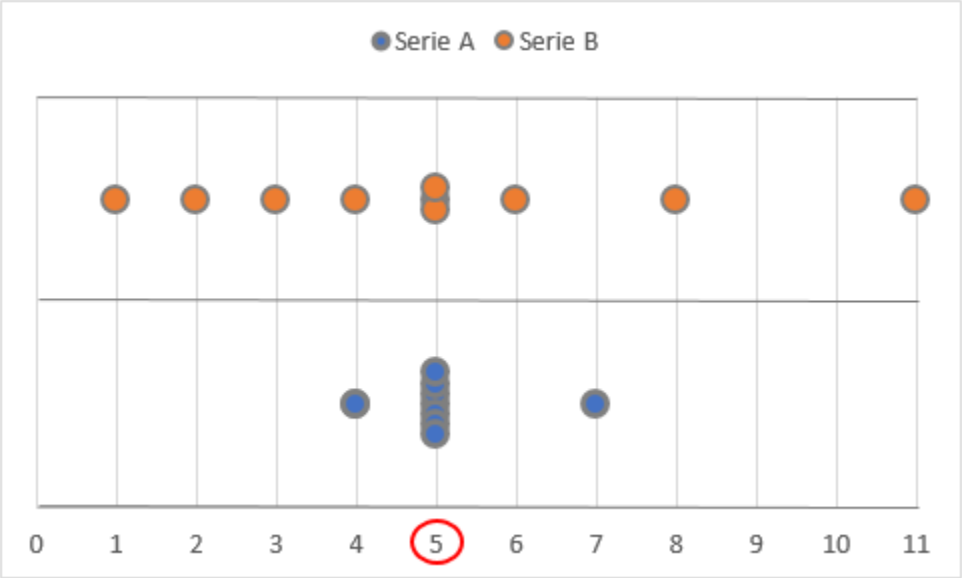
Como mencionamos en un inicio las medias más usadas de dispersión son el rango, la varianza y la desviación estándar. El rango es la diferencia entre el valor máximo observado y el valor mínimo observado. La varianza y la desviación estándar miden que tan dispersos están los datos en torno a la media.

Consideremos las siguientes series de datos

											Promedio $(\bar{x})$	**Des. Estándar**
Serie A	5	4	5	5	4	5	5	5	5	7	5	0,82
Serie B	1	5	2	4	8	6	3	5	11	5	5	2,91

Ambas series tienen el mismo promedio, sin embargo la segunda es mucho más variable que la primera pues su desviación estándar es mayor.

Gráficamente se observa lo siguiente:



Observe que en el caso de la serie A, que es la que tiene menor desviación estándar la mayoría de las observaciones están muy cerca de la media, de hecho, en este caso la mayoría coinciden con la media. En cambio, para la serie B, hay datos que coinciden con la media, pero hay otros que están más alejados.

La desviación estándar es la raíz cuadrada de la varianza. Solo con el fines educativos mostraremos aquí como se calcula la varianza, ya que al igual que en el caso de la media y la mediana la varianza se puede calcular usando software estadístico o inclusive Excel.

Calculo de la varianza de la serie B

1. Primero debemos calcular el promedio.

$$\bar{x} = \frac{1+5+2+4+8+6+3+5+11+5}{10} = \frac{50}{10} = 5$$

2. Luego, como se muestra en la imagen, a cada observación (x_i) le restamos la media y elevamos el resultado al cuadrado.

3. Por último se suman todos los valores de la tercera columna, es decir los, $(x_i - \bar{x})^2$ y se divide por el número de observaciones n o por $n-1$. Se divide por n cuando se trabaja con todos los datos de la población y se divide por $n-1$ cuando se trabaja con un subconjunto o muestra de la población. En este caso consideraremos que la serie de datos es una muestra por lo que dividiremos por $n-1$, es decir, 9.

Serie B	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	$1-5=-4$	16
5	$5-5=0$	0
2	$2-5=-3$	9
4	$4-5=-1$	1

8	$8-5=3$	9
6	$6-5=1$	1
3	$5-3=2$	4
5	$5-5=0$	0
11	$11-5=6$	36
5	$5-5=0$	0
	Suma =	**76**

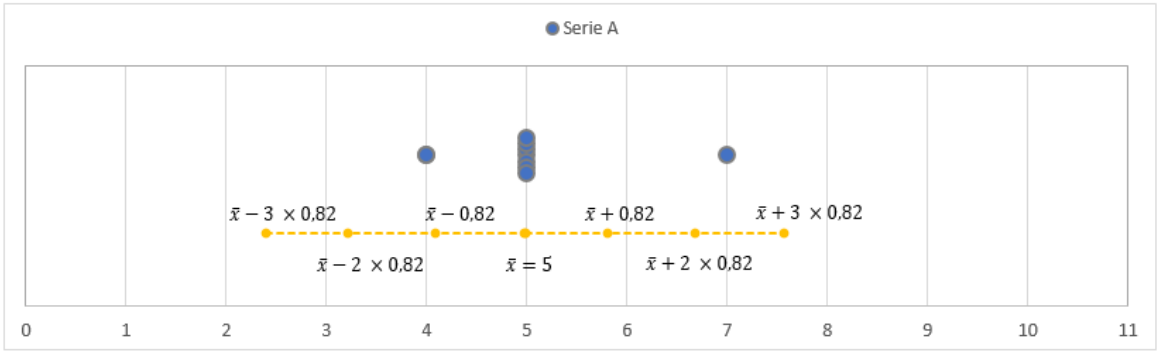
Finalmente la varianza será $82/9 = 8{,}44$ y la desviación estándar la raíz cuadrada de $\sqrt{9{,}11} = 2{,}91$.

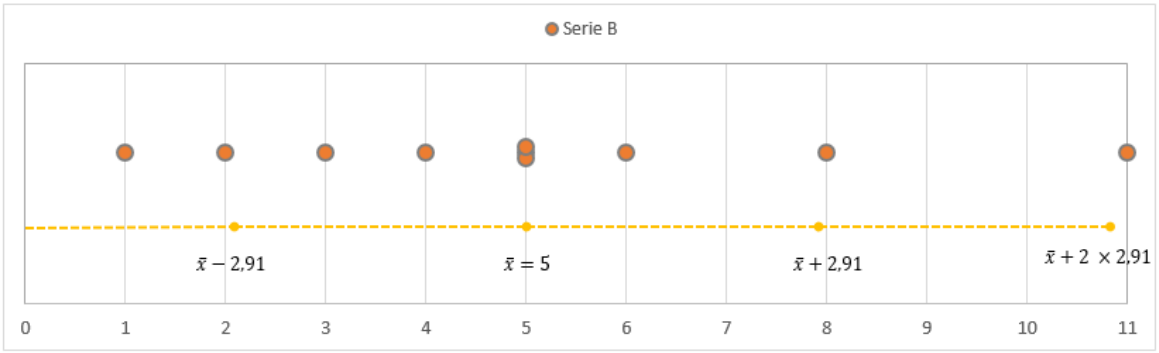
En general se trabaja más con la desviación estándar que con la varianza y la razón es muy simple. Supongamos que las series anteriores representan el tiempo en minutos que los clientes de dos bancos, A y B, deben esperar para ser atendidos. Esto implicaría que la unidad de medición de la varianza es min^2 lo que es poco tangible y resulta difícil de comprender. Como la desviación estándar es la raíz de la varianza, la unidad de medición es minutos y esta se puede incluso graficar.

Existe una regla llamada Regla Empírica que señala:

- Aproximadamente el 68% de las datos estén entre la media y una desviación estándar
- Aproximadamente el 95% de las datos estén entre la media y dos desviaciones estándar
- Aproximadamente el 100% de las datos estén entre la media y tres desviaciones estándar

Para las series de datos que estábamos analizando podemos identificar los tramos determinados por la regla empírica:





Hay que tener en cuenta que la regla empírica es solo una aproximación y va a depender de la forma de la distribución de los datos que tanto se ajusten estos a la regla empírica. Por ejemplo, en la Serie A un 90% de los datos estan entre la media más/menos una desviación estándar en tanto que para la serie B, un 80% de los datos están entre la media más/menos una desviación estándar.

Ahora que ya tenemos claro que es la varianza y la desviación estándar, volvamos a lo nota del Diario Financiero donde dice "*Se suma el hecho de que el riesgo asumido para alcanzar dichos retornos es menor que al que uno se expone invirtiendo directamente en los índices…*"

Además se anexa la siguiente información:

ESTADÍSTICA DESCRIPTIVA MENSUAL DE LOS RETORNOS NOMINALES DE LOS FONDOS DE PENSIONES, PERÍODO 2002-2018

MEDIDA	IGPA	A	B	C	D	E	RF
Promedio	0,96%	0,80%	0,70%	0,66%	0,62%	0,57%	0,26%
Mediana	0,69%	0,99%	0,84%	0,69%	0,62%	0,61%	0,25%
Des. Estándar	0,040	0,034	0,025	0,016	0,010	0,008	0,001
Varianza	0,002	0,001	0,001	0,000	0,000	0,000	0
Máximo	14,97%	9,46%	6,65%	4,20%	3,08%	3,52%	0,43%
Mínimo	-10,21%	-20,48%	-13,34%	-7,01%	-2,98%	-2,34%	0,12%
Beta	1	0,53	0,4	0,25	0,12	-0,02	

FUENTE: SUPERINTENDENCIA DE PENSIONES Y BOLSA DE COMERCIO DE SANTIAGO. ELABORACIÓN CIEDESS.

Analizando la desviación estándar podemos ver que todos los fondos tiene una desviación estándar menor a la del IGPA., lo cual implica que son más estables y por lo tanto menos riesgosos. Además, recordemos que el fondo A es el más riesgoso y que el E es el menos riesgoso. Esto también se ve reflejado en la desviación estándar de los distintos fondos, la que va decreciendo a medida que se va de un fondo a otro.

Una pregunta recurrente que me hacen los alumnos cuando enseño esta materia es "¿Qué es mejor? ¿Una varianza grande o una varianza pequeña? La respuesta es depende y les doy el siguiente ejemplo. Supongamos que la media de la prueba fue 3,0, ¿qué me gustaría a mí como profesora? Que hubiera una alta varianza ya que esto sería reflejo de que hay alumnos que tuvieron notas muy superiores a 3,0 (aunque probablemente también inferiores). Así habría indicios de que, al menos, algunos entendieron los contenidos que fueron evaluados. Por el contrario, si la media de la prueba fue 6,0 me gustaría que la varianza fuera pequeña pues este sería un indicador de que a todos les fue muy bien.

La varianza y la desviación estándar se relacionan también con los conceptos de homogeneidad y heterogeneidad de un grupo con respecto a una variable. Por ejemplo, los alumnos de periodismo de tercer año de la Universidad Adolfo Ibañez son homogéneos con respecto a su edad, ya que sus edades son relativamente similares. Pero, estos mismos alumnos son heterogéneos con respecto a la distancia a la que viven de la Universidad.

Para cerrar con el tema de la varianza les dejó el link a este artículo en donde me pereció que se usó de manera muy creativa la desviación estándar (denotada aquí por SD). "*Éstas son las 50 películas que más han dividido a la crítica*" https://www.espinof.com/otros/estas-son-las-50-peliculas-que-mas-han-dividido-a-la-critica 24 noviembre 2017

Medidas de posición

Consideremos la siguiente nota publicada en el sitio Pulso de La Tercera el 29 de mayo de 2019 titulada "*Hogares de mayores ingresos gastan 11 veces más en peajes que los del quintil más bajo"* (Fuente: https://www.latercera.com/pulso/noticia/hogares-mayores-ingresos-gastan-11-veces-mas-peajes-los-del-quintil-mas/676282/)

¿A qué se refiere con el quintil más bajo?

De acuerdo con el Ministerio de Desarrollo Social y familia lo quintiles se definen como:

Quintil del ingreso autónomo per cápita del hogar nacional: Quinta parte o 20% de los hogares nacionales ordenados en forma ascendente de acuerdo al ingreso autónomo per cápita del hogar, donde el primer (Quintil I) representa el 20% más pobre de los hogares del país y el quinto quintil (Quintil V) el 20% más rico de estos hogares.

Quintil de ingreso autónomo per cápita del hogar regional**:** Quinta parte o 20% de los hogares de una región ordenados en forma ascendente de acuerdo al ingreso autónomo per cápita del hogar, donde el primer (Quintil I) representa el 20% más pobre de los hogares de la región y el quinto quintil (Quintil V) el 20% más rico de estos hogares.

Fuente: http://observatorio.ministeriodesarrollosocial.gob.cl/casen/casen_def_ingresos.php

El 19 de mayo de 2019 en el portal radiopolar.com se publicó la nota titulada: "*SEREMI DE EDUCACIÓN DESTACÓ MEJORAMIENTO DE LA EDUCACIÓN EN TODOS SUS NIVELES EN CUENTA PÚBLICA"* En el cuerpo de este artículo se menciona: "*En materia de educación técnico profesional de enseñanza media, el Ministerio dispuso el Plan de modernización de la educación Técnico Profesional, que permitirá que los 9 establecimientos educacionales que imparten educación técnico profesional en Magallanes, extiendan la gratuidad al 7mo decil, lo que beneficiará a más de 1.525 alumnos en nuestra región."*

¿Qué es el 7mo decil?

Los percentiles dividen un conjunto de observaciones ordenadas de menor a mayor en 100 partes iguales. Así hasta el primer percentil, p_{10}, hay un 10% de las observaciones, hasta el segundo percentil, p_{20}, hay un 20% de las observaciones ya así sucesivamente. Los quintiles y deciles son casos particulares de los percentiles. Por ejemplo, el percentil 20 es equivalente al primer quintil, primer decil es equivalente al percentil 10 y la mediana es lo mismo que el percentil 50. El *7mo décil*, al que hace referencia la nota, es lo mismo que percentil 70.

En los contextos mencionados se usan los deciles y quintiles para caracterizar a los hogares de acuerdo con su nivel de ingresos. Que un hogar este en el séptimo decil, quiere decir que su ingreso per cápita esta entre los $193.105 a $250.663. Por otra parte esto significa que el 70% de los hogares tienen como máximo un ingreso per cápita del $250.663. Si un hogar esta en el primer quintil quiere decir que su ingreso per cápita es como máximo $74.969, pero también podemos afirmar que el 20% de los hogares en Chile tiene un ingreso per cápita de a lo sumo $74.969. (Fuente: https://www.emol.com/noticias/Nacional/2017/10/23/880299/Conoce-a-que-decil-perteneces-para-postular-a-la-gratuidad-y-becas-de-la-educacion-superior.html)

Otra manera que se usa con frecuencia para dividir un conjunto de observaciones son los cuartiles. Como su nombre lo indica en este caso se dividen las observaciones en cuatro grupos tal que cada uno de ellos contenga el 25% de las observaciones. Bajo el primer cuartil, Q_1 se encuentra el 25% de las observaciones, bajo el cuartil 2, Q_2, se encuentran el 50% de las observaciones y así sucesivamente. Note que el cuartil 2 corresponde a la mediana.

No nos detendremos en la forma de cálculo de estas medidas pues al igual que en el caso de las medidas de tendencia central y de dispersión estas pueden calculares con distintos softwares estadísticos y planillas de cálculo.