Clase 1: Estadística descriptiva 1D

Matías Carrasco

29 de septiembre de 2019

Índice

1. Diagramas de tallo y hojas

Diagrama simple

La Tabla 1 muestra las notas finales de un examen de tres grupos de estudiantes de un curso básico de estadística. La nota máxima es 120. Las notas se han tomado de una lista oficial en la cual los estudiantes están ordenados alfabéticamente. En tal disposición, es casi imposible ver qué está pasando. Una mirada rápida nos muestra que hay gran variación en las notas, pero no mucho más.

Escribir las notas de los estudiantes en orden alfabético puede ser ventajoso para propósitos administrativos, pero nos provee de poca ayuda si queremos, por ejemplo, comparar el rendimiento de los tres grupos.

Un artilugio llamado *diagrama de tallo y hojas* es un primer paso muy útil para crear un poco de orden entre tanta confusión.

Tomemos, por ejemplo, la nota 82 del primer estudiante del Grupo 1. Nos referiremos al dígito 8 que caracteriza el nivel de rendimiento del estudiante como el *tallo*, y al dígito 2 que simplemente nos provee de información más detallada como la *hoja* del número 82. Para el estudiante que sacó 115, el tallo es 11 y la hoja es 5.

En un diagrama de tallo y hojas, los tallos se ordenan verticalmente, mientras que las hojas se marcan horizontalmente en el valor del tallo correspondiente.

La Tabla 2 representa el diagrama de tallo y hojas para los 24 estudiantes del Grupo 3. El rendimiento de los estudiantes del Grupo 3 parece haber sido bastante errático: el 50 % de los estudiantes tiene notas bajas (por debajo de 70); el 25 % tiene notas intermedias (70 -

Tabla 1: Notas finales del examen (escala de 0 a 120).

Grupo 1		Grupo 2		Grupo 3	
82	90	99	73	58	70
45	104	87	81	46	84
89	64	72	96	72	96
82	83	81	60	84	63
67	77	88	92	74	90
72	83	82	85	48	
64	78	66	104	116	
89	81	71	98	91	
93	96	88	104	69	
78	62	58	57	53	
87	77	84	25	65	
75	53	68	96	109	
115	113	86	74	91	
57	67	70	74	69	
86	103	88	72	69	
73	39	91	96	86	
86		71	88	45	
85		108	84	48	
82		109	62	61	

Tabla 2: Diagrama de tallo y hojas para el Grupo 3.

4	6858
5	83
6	959913
7	240
8	464
9	1160
10	9
11	6

89); y el 25 % restante tiene notas altas (90 o más). Una mirada a la lista de notas en orden alfabético difícilmente hubiera sugerido esta descripción.

Veamos qué nos revela el diagrama de tallo y hojas para los 97 estudiantes de los tres grupos. Para hacer el diagrama de los tres grupos juntos, haremos un refinamiento más. En el diagrama de la Tabla 2, por simplicidad, hemos introducido los valores de las hojas en el mismo orden en el que aparecían en la Tabla 1. Podemos mejorar la forma de agrupar las hojas ordenando de forma creciente las entradas de cada fila.

El diagrama de tallo y hojas retiene toda la información numérica de la tabla original. Si hubieramos querido, podríamos haber mantenido la información sobre los distintos grupos marcando con diferentes colores los dígitos de cada grupo, por ejemplo, rojo para el Grupo 1, verde para el Grupo 2, y azul para el Grupo 3.

En el diagrama de la Tabla 3, las 97 notas no solo están ordenadas de la más chica a la más grande, si no que también es una clara imagen de cómo están *distribuidas* a lo largo

Tabla 3: Diagrama de tallos y hojas para los tres grupos juntos.

2	5
3	9
4	55688
5	337788
6	012234456778999
7	001122223344457788
8	11122223344445566667888899
9	00111236666689
10	3444899
11	356

del rango 0 a 120. Por ejemplo, podemos responder facilmente a preguntas como ¿cuántos estudiantes hay con notas entre 75 y 85? Vemos a simple vista que más estudiantes tienen notas en los 80's que en cualquier otra categoría de tallos.

Los estadísticos usualmente llaman *moda* a la categoría de tallos que ocurre con mayor frecuencia. Para nuestro ejemplo, la nota 80 representa la moda de la distribución de notas. La moda de una distribución es una forma de caracterizar un valor típico. Hay otras formas de hacerlo que veremos más adelante.

Espalda con espalda

Vimos que el rendimiento de los estudiantes del Grupo 3 fue bastante errático y más bien tirando a ser bajo. ¿Qué se puede decir de los otros dos grupos? Podemos hacer diagramas de tallo y hojas separados para el Grupo 1 y el Grupo 2, y luego compararlos entre ellos y con el Grupo 3. Una forma más efectiva de compararlos es poner los diagramas de tallo y hojas espalda con espalda.

Tabla 4: Diagrama de tallo y hojas espalda c/ espalda para los grupos 2 y 3.

Grupo 3		Grupo 2
	2	5
	3	
8865	4	
83	5	78
999531	6	0268
420	7	01122344
644	8	112445678888
6110	9	1266689
9	10	4489
6	11	

La Tabla 4 muestra tal comparación entre los grupos 2 y 3. El rendimiento de los estudiantes del Grupo 2 es muy diferente al de los estudiantes del Grupo 3. Las notas del Grupo 2 se concentran alrededor de una sola moda en 80, mientras que las del Grupo 3 no se concentran alrededor de la moda 60, sino que presentan dos segundas modas más en 40 y 90.

Datos atípicos

Cuando tenemos un gran número de mediciones, como nuestras notas de examen, a menudo sucede que una, o incluso dos o tres de las mediciones difieren marcadamente de todas las demás mediciones.

Por ejemplo, la nota más baja en el Grupo 2 está 32 puntos más abajo que la siguiente nota más baja del grupo. También es 14 puntos más baja que la nota más baja de los otros dos grupos.

Estos rezagados se conocen como *valores atípicos*. Los valores atípicos pueden ser causados por muchos factores diferentes. En el presente ejemplo, el valor 25 podría deberse a un error administrativo. Quizás la nota debería haber sido 55, pero se tipeó como 25. O podría haber habido un error de cálculo. Es de esperar que esos errores administrativos sean detectados y corregidos. Pero un valor atípico también podría haber sido causado por razones más sustanciales. El estudiante en cuestión podría no haber estado preparado, en tal caso no se requiere corrección.

Como veremos más adelante, los valores atípicos entre las mediciones frecuentemente causan problemas, y necesitamos algunas reglas para identificarlos. Un posible procedimiento lo discutiremos más adelante.

Diagramas de tallo y hojas extendidos

Consideremos la Tabla 5 que muestra las alturas (en pulgadas) de 30 estudiantes de primer año de facultad.

Tabla 5: Alturas (en pulgadas) de 30 estudiantes de primer año de facultad.

64	71	68	69	67	71	70	74	67	73
63	71	70	59	73	66	77	64	68	64
67	67	66	78	72	64	68	61	69	69

Supongamos que queremos construir un diagrama de tallo y hojas. Si hacemos como en el ejemplo de las notas del examen anterior, tendríamos solo 3 valores de tallo: 5, 6 y 7 con 1, 18 y 11 valores de hojas, respectivamente. Tal diagrama no será mucho más informativo que la lista original de 30 mediciones. Ver la Tabla 6.

Tabla 6: Diagrama de tallo y hojas para los datos de la Tabla 5.

5	9
6	134444667777888999
7	00111233478

Podemos obtener más detalles dividiendo un tallo en varios subgrupos, trazando cada subgrupo como una fila separada. Hay dos maneras convenientes de formar subgrupos, uno que involucra dos grupos de 5 valores de hoja cada uno, y el otro, cinco grupos de dos valores de hoja cada uno. Ilustremos ambos métodos.

En el primer caso, mantenemos juntos los valores de hoja de 0 a 4 y los valores de hoja de 5 a 9. La Tabla 7 muestra el diagrama de tallo y hoja para dos subgrupos.

Tabla 7: Diagrama de tallo y hojas para los datos de la Tabla 5 con dos subgrupos.

5	9
6	134444
6	667777888999
7	001112334
7	78

En el segundo caso, formamos subgrupos que constan de los valores de hoja 0 y 1, 2 y 3, 4 y 5, 6 y 7, y 8 y 9 como en la Tabla 8.

Tabla 8: Diagrama de tallo y hojas para los datos de la Tabla 5 con cinco subgrupos.

5	9
6	1
6	3
6	4444
6	667777
6	888999
7	00111
7	233
7	4
7	7
7	8

Para los datos de altura, el segundo método parece ser más informativo.

2. Histogramas

Histograma de puntos

Para muchas investigaciones estadísticas, es suficiente saber cuántas mediciones caen dentro de límites especificados sin conocer su valor preciso. En la Tabla 9, mantenemos la cantidad de hojas en cada tallo del diagrama de tallo y hojas de la Tabla 3, pero ignoramos la información detallada proporcionada por los valores de las hojas.

Tabla 9: Histograma de puntos para los tres grupos juntos (97 estudiantes).

25	*
35	*
45	*****
55	******
65	**********
75	**********
85	**********
95	**********
105	*******
115	***

Al sustituir los valores numéricos precisos por asteriscos, perdemos cierta información, pero la nueva representación es más simple y, a menudo, transmite toda la información que se requiere.

Notar además un cambio adicional respecto de la Tabla 3. Los valores anteriores del tallo que indicaron el nivel de rendimiento han sido reemplazados por los *puntos medios* de los respectivos intervalos del tallo. Dado que el valor del tallo 8 indica un puntaje de examen de al menos 80, pero menos de 90, el punto medio del intervalo es 85, como se indica en la Tabla 9. El nuevo diagrama se denomina con frecuencia *histograma de puntos*.

En realidad, la forma más habitual de un histograma se obtiene rotando la Tabla 9 90 grados, en sentido antihorario, y al mismo tiempo reemplazando las columnas de asteriscos por barras cuyas alturas son iguales al número de mediciones (asteriscos) en un intervalo dado. En la Figura 1, hemos hecho otro cambio. Los límites de los intervalos, en lugar de los puntos medios, están marcados a lo largo del eje horizontal.

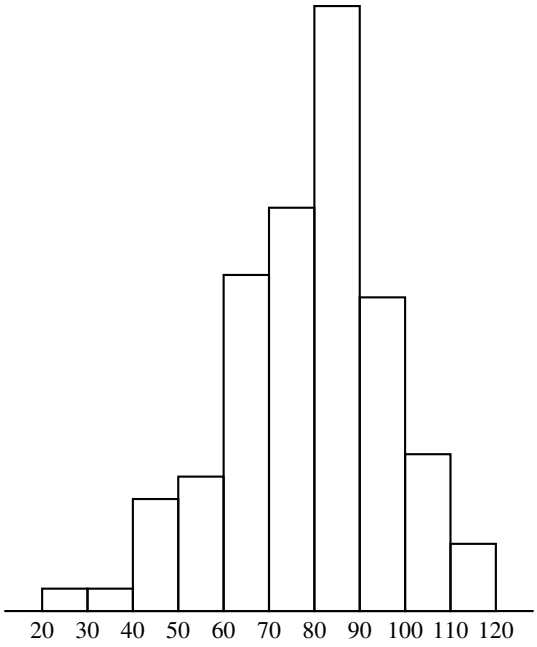

Figura 1: Histograma de barras de las 97 notas.

Si se usan barras en lugar de asteriscos, se lo llama simplemente *histograma*.

Histogramas de área uno

Un histograma es una de las herramientas estadísticas más simples y útiles para representar un conjunto de medidas tales como calificaciones, ingresos o edad.

En un histograma, como el de la Figura 1, las *frecuencias absolutas* representan las alturas de las barras rectangulares.

Como todos los intervalos son iguales en longitud, las áreas de las barras también son proporcionales a las frecuencias, en donde la constante de proporcionalidad depende de la longitud común de los intervalos. De este modo, el área total bajo el histograma depende del número de mediciones en el conjunto de datos.

Esto puede complicar la comparación de dos conjuntos de datos de diferente tamaño y, por lo tanto, de área bajo el histograma diferente.

Podemos evitar esta dificultad haciendo que las áreas de las barras sean iguales a las *frecuencias relativas* en lugar de ser proporcionales a las frecuencias absolutas.

Las frecuencias relativas se obtienen dividiendo las frecuencias absolutas por el número de mediciones en el conjunto de datos.

En un histograma de este tipo, el área de cada barra indica qué proporción del conjunto total de datos cae en el intervalo correspondiente. Las frecuencias relativas siempre suman uno, independientemente del tamaño del conjunto de datos. Por lo que el área debajo del histograma es igual a uno.

Tabla 10: Puntajes de 1000 estudiantes en una prueba SAT para ingresar a una universidad de Estados Unidos.

Intervalo		Frec. Asb.	Frec. Rel.	Densidad
Bordes	Punto medio			
475-525	500	2	.002	.00004
525-575	550	0	.000	.00000
575-625	600	7	.007	.00014
625-675	650	5	.005	.00010
675-725	700	13	.013	.00026
725-775	750	17	.017	.00034
775-825	800	39	.039	.00078
825-875	850	62	.062	.00124
875-925	900	75	.075	.00150
925-975	950	102	.102	.00204
975-1025	1000	113	.113	.00226
1025-1075	1050	126	.126	.00252
1075-1125	1100	113	.113	.00226
1125-1175	1150	104	.104	.00208
1175-1225	1200	70	.070	.00140
1225-1275	1250	63	.063	.00126
1275-1325	1300	43	.043	.00086
1325-1375	1350	20	.020	.00040
1375-1425	1400	12	.012	.00024
1425-1475	1450	9	.009	.00018
1475-1525	1500	5	.005	.00010
	Total	1000	1	.02

Cuando el histograma tiene área uno, la escala vertical ya no representa frecuencias (ni absolutas ni relativas). En estos casos, la altura h de la barra sobre un intervalo de longitud ℓ verifica

$$h \times \ell = \text{área de la barra} = f,$$

en donde f es la frecuencia relativa de mediciones en el intervalo correspondiente.

Dicho de otro modo

$$\text{altura de la barra} = h = \frac{f}{\ell}.$$

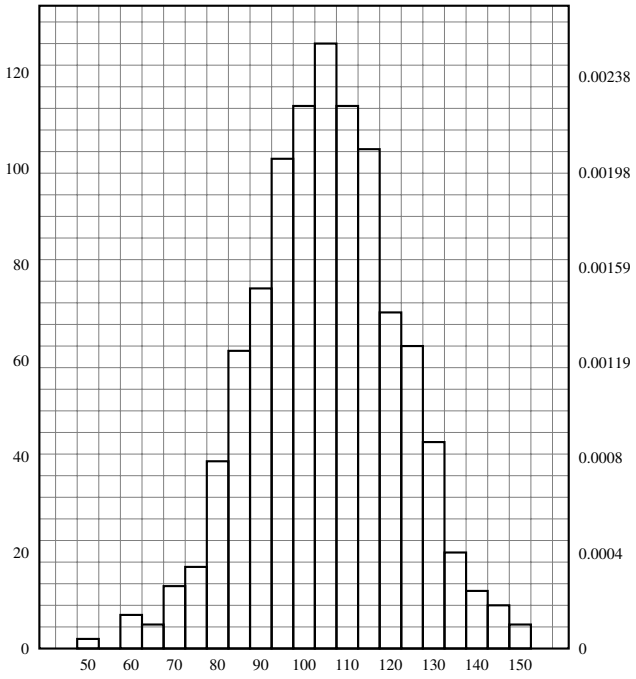

Figura 2: Histograma de los 1000 puntajes del SAT.

¿En qué unidades debe estar la escala vertical? Depende de las unidades de los datos observados. Por ejemplo, en el caso de las notas, las unidades son puntos. Las frecuencias relativas no tienen unidades, pero podemos pensar que las unidades son porcentajes sobre 100. Por lo que la altura de la barra tendrá en este caso unidades de $\%/(100 \text{ puntos})$. En general

$$\text{la escala vertical es } \%/(100 \text{ unidades de los datos}).$$

Esta escala se llama la *escala de densidad*.

El nombre se debe a la analogía con las densidades de masa: si graficamos la densidad lineal de masa de un objeto como una barra rígida que pesa 1 kg, el área debajo de la curva de densidad es igual a 1kg, de modo que la unidad de densidad es kg/cm. Para las densidades de probabilidad se debe cambiar la unidad de masa kg por la unidad de probabilidad $\%/100$.

Esta analogía permite interpretar mejor las lecturas hechas en un histograma de área 1. Por ejemplo, supongamos que una barra rígida con densidad de masa no uniforme está hecha de un mismo material. La densidad en un punto de la barra simplemente indica cuántas partículas de ese material están en la vecindad del punto. Así, si la densidad es más alta de un lado de la barra que del otro, es porque las partículas están más *amontonadas* en ese lado que en el otro. Es decir, la densidad mide el *amontonamiento* de partículas en la vecindad de un punto.

Del mismo modo, la densidad de probabilidad mide el amontonamiento de mediciones que caen en un determinado intervalo. Si la densidad es más alta en un lugar que en otro, quiere decir que las mediciones se amontonan más en ese lugar.

Nosotros usaremos siempre histogramas de área 1. En estos histogramas, la escala vertical es la densidad. Así, la altura de un bloque representa el amontonamiento de los datos en el intervalo correspondiente, y sus unidades son $\%$ cada 100 unidades de medición. Las áreas representan frecuencias relativas.

Veamos otro ejemplo para clarificar estas ideas.

El examen SAT es una prueba de acceso a la universidad en Estados Unidos. En el SAT no se evalúan los conocimientos adquiridos durante el liceo, sino la capacidad de aprendizaje futura del estudiante, es por tanto un examen de razonamiento.

En la Tabla 10 se muestran los puntajes SAT de 1000 estudiantes que han solicitado la admisión a una universidad. Los puntajes están desglosados en intervalos de longitud 50, y se dan las frecuencias absolutas, las frecuencias relativas, y la densidad de cada intervalo.

La Tabla 10 proporciona toda la información relevante sobre la distribución de las 1000 puntuaciones SAT, pero el histograma correspondiente de la Figura 2 proporciona una imagen mucho más clara de lo que está sucediendo. En la Figura 10, las frecuencias absolutas se leen en la escala vertical a la izquierda, las densidades en la escala vertical a la derecha en $\%/(100 \text{ puntos})$.

3. Resumen de los cinco números

Los diagramas de tallo y hojas y, en menor grado, los histogramas brindan información detallada y útil sobre un conjunto de mediciones. Pero para muchos propósitos prácticos, considerablemente menos detalles pueden ser suficientes e incluso preferibles. Los estadísticos han propuesto muchas formas de resumir un conjunto de datos. Un resumen de cinco números es simple y al mismo tiempo altamente informativo al indicar dónde se centran las mediciones y qué tan dispersas están.

La mediana

Un diagrama de tallo y hojas ordena las mediciones desde la más pequeña hasta la más grande. Por lo tanto, es natural observar la medición en el medio de este arreglo como su "centro". El término técnico para este centro es *mediana*, y la denotamos por m. Antes de describir cómo encontrar la mediana de un diagrama de tallo y hojas, veamos un ejemplo más simple basado en las primeras 7 notas del Grupo 1 de la Tabla 1:

$$82\ 45\ 89\ 82\ 67\ 72\ 64,$$

o, reorganizado, del más pequeño al más grande,

$$45\ 64\ 67\ 72\ 82\ 82\ 89.$$

La mediana es 72, la cuarta más chica y también la cuarta más grande de las 7 mediciones. Por lo tanto $m = 72$.

La determinación de la mediana es menos obvia cuando hay un número par de mediciones, por ejemplo, las primeras 8 notas del Grupo 3 de la Tabla 1. Reorganizados según el tamaño, son:

$$46\ 48\ 58\ 72\ 74\ 84\ 91\ 116.$$

Esta vez, no hay puntaje que esté igualmente distante de cualquier extremo. La cuarta medición más chica es 72; la cuarta más grande, 74. Así que definimos la mediana como la más chica de esas dos posibilidades, es decir $m = 72$ también.

Cuando hay un número impar de mediciones, hay exactamente una medición central, que es la mediana. Para un número par de mediciones, hay dos medidas centrales, y tomamos la más chica como la mediana.

Siempre que no tengamos demasiadas medidas, un método conveniente para determinar la mediana m que ni siquiera requiere reorganizar las medidas según el tamaño es el siguiente. Tachamos la medición más grande y la más chica y continuamos tachando mediciones en pares hasta que queden una o dos. En cada paso, se tachan las mediciones más grande y más chica que no se eliminaron previamente. Si queda una sola medición en el último paso, esta es m. Si quedan dos mediciones, m se elige como la más chica de las dos. Cualquier medición que se repita más de una vez, debe usarse tantas veces como ocurra en el proceso de tachado.

Por ejemplo, si las mediciones son

$$405,\ 280,\ 73,\ 440,\ 179 \text{ y } 405,$$

comenzamos tachando 440 y 73. Esto deja 405, 280, 179 y 405. Ahora tachamos 405 y 179, dejando 280 y 405. Así $m = 280$.

Para la determinación de la mediana (y las otras cuatro cantidades que se analizarán en esta sección) de un arreglo ordenado de mediciones, es útil asignar a cada medición una cantidad llamada *profundidad* y que denotaremos por p.

La profundidad de una medición nos dice "qué tan lejos" está del extremo izquierdo en el arreglo ordenado de todas las mediciones (es decir, la medición más chica). Para el ejemplo anterior de las 7 notas, tenemos:

mediciones :	45	64	67	72	82	82	89
profundidad p :	1	2	3	4	5	6	7

por lo que la profundidad de la mediana es 4, $p(m) = 4$. Para el ejemplo de las 8 notas, tenemos

mediciones :	46	48	58	72	74	84	91	116
profundidad p :	1	2	3	4	5	6	7	8

En este caso también $p(m) = 4$.

Notar que la profundidad de una medición indica la cantidad de mediciones que están a su izquierda (incluyendose a sí misma).

Regla general

Para un conjunto de n mediciones, la profundidad de la mediana m es igual a

$$p(m) = \left\lfloor \frac{n+1}{2} \right\rfloor,$$

tanto si n es par o impar.

Para el conjunto completo de notas de los 97 estudiantes de la Tabla 1, $n = 97$ y

$$p(m) = \left\lfloor \frac{97+1}{2} \right\rfloor = 49.$$

Para encontrar la nota mediana, tenemos que movernos 49 mediciones desde el extremo más chico del diagrama de tallo y hojas. Esta operación se simplifica considerablemente si agregamos una columna de profundidad al diagrama de tallo y hojas. Para hacerlo, definimos la profundidad de una fila (o profundidad del tallo) como igual al valor de profundidad más grande de cualquier hoja en la fila.

Por ejemplo, la profundidad de la fila con tallo 5 es igual a $6+5+1+1=13$, ya que hay 6 hojas en esa fila y $5+1+1$ más en las filas de más arriba. Del mismo modo, la profundidad de la fila con tallo 9 es igual a 87, ya que son $7+3=10$ hojas en las filas de más abajo. La Tabla 11 representa el diagrama de tallo y hojas con los valores de profundidad agregados para todas las filas.

Tabla 11: Diagrama de tallos y hojas para los tres grupos juntos (97 estudiantes) con los valores de profundidad de los tallos.

1	2	5
2	3	9
7	4	55688
13	5	337788
28	6	012234456778999
46	7	001122223344457788
73	8	11122223344445566667788899
87	9	00111236666689
94	10	3444899
97	11	356

La determinación de la mediana ahora procede de la siguiente manera. Hemos visto que $p(m)=49$. De acuerdo con la columna de profundidad, contando desde la parte superior de la tabla (donde se registran las mediciones chicas), hay 46 mediciones en la fila con tallo 7. La mediana corresponde a la tercera hoja más chica en la fila siguiente, representada simbólicamente como 8|1 u 81. La nota mediana del examen es 81.

Dadas n mediciones $x_1, x_2, \ldots, x_n$ la mediana m es por definición

$$m = x^*_{\lfloor \frac{n+1}{2} \rfloor},$$

en donde

$$x^*_1 \leq x^*_2 \leq \cdots \leq x^*_n$$

es la lista ordenada de mediciones.

Notar que la mediana divide en dos parte "iguales" el conjunto de mediciones en el siguiente sentido: es la medición más chica que deja a su izquierda (incluyéndose a sí misma) a al menos el 50 % del total.

El promedio

Hay muchas otras maneras de medir el centro de una distribución de mediciones. Una de las medidas más frecuentemente utilizadas es el *promedio*, comúnmente conocido también como la *media*.

Para encontrar el promedio de un conjunto de mediciones, simplemente calculamos la suma de todas ellas, y luego dividimos esta suma por la cantidad total que hay.

Para las 97 notas en nuestro ejemplo de la Tabla 1,

$$\text{Nota promedio} = \frac{82 + \cdots + 90}{97} = \frac{7622}{97} = 78.6.$$

En general:

Dadas n mediciones $x_1, x_2, \ldots, x_n$ el promedio $\bar{x}$ es por definición

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Promedio vs. mediana

¿Cuál es la mejor medida de centralidad: la mediana o el promedio? Antes de poder responder a esta pregunta, debemos decidir qué queremos decir con "mejor". Y eso no es fácil.

La mediana es conceptualmente más simple, pues es el “punto medio” de la lista ordenada de mediciones. Conceptualmente, el promedio es una cantidad más oscura. Sin embargo, el promedio es mucho más fácil de calcular que la mediana.

Una ventaja de la mediana respecto al promedio es su *robustez*. Esto quiere decir que la mediana es poco sensible a datos atípicos. Esto no ocurre con el promedio.

Recordar, que al igual que la esperanza para variables aleatorias discretas, el promedio representa el lugar en donde debemos colocar un punto de apoyo para que el histograma de las mediciones se mantenga en equilibrio.

Veamos un ejemplo sencillo para entender la diferencia. Supongamos que nuestra lista de datos es

$$1, 2, 2, 3.$$

En este caso, el histograma es simétrico respecto a la mediana que es $m = 2$, y el promedio es también $\bar{x} = 2$. Ver la Figura 3.

¿Qué pasa cuando el valor 3 de la lista se cambia por 5, o 7? Como se muestra en la Figura 3, el bloque sobre dicho valor se mueve hacia la derecha, destruyendo así la simetría del histograma. El promedio $\bar{x}$ se desplaza también hacia la derecha, como si estuviera siguiendo, con menor velocidad, al bloque. Sin embargo, la mediana m es igual a dos en todos los casos.

Notar que cuando cambiamos el valor 3 por 5 o 7, el 75 % de las mediciones quedan a la izquierda del promedio. Cuando un histograma no es aproximadamente simétrico, el promedio no es un valor representativo de los datos. Ver la Figura 4.

Medidas de dispersión

La mediana y el promedio son medidas útiles que caracterizan el centro de un conjunto de mediciones. Pero, por lo general, también queremos saber algo sobre su dispersión.

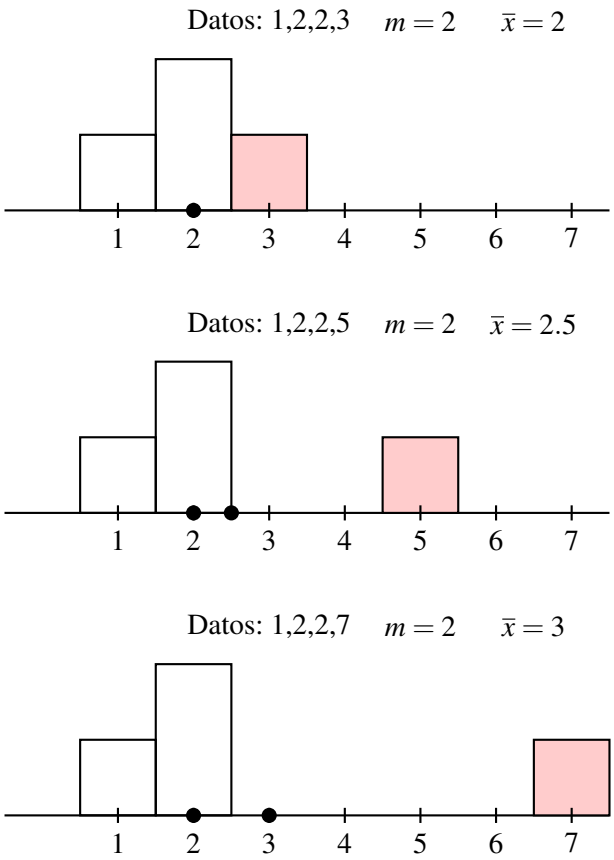

Figura 3: La figura muestra tres histogramas. El promedio y la mediana están indicados arriba a la derecha de cada histograma.

La mediana divide las mediciones en dos partes iguales, el 50 % inferior y el 50 % superior. Una forma útil de caracterizar la dispersión es determinar dos números que contengan el 50 % central.

Llamaremos al más chico el *cuartil inferior*, q_i, y al más grande *cuartil superior*, q_s.

Para aclarar ideas, volvamos al ejemplo anterior que involucra las primeras 8 notas del Grupo 3 de la Tabla 1:

$$46\ 48\ 58\ 72\ 74\ 84\ 91\ 116.$$

Como ya hemos visto, la mediana es $m = 72$ que corresponde a una profundidad igual a la parte entera de $(8+1)/2$, es decir 4. ¿Cuáles son las mediciones que dejan a su izquierda (incluyéndose a sí mismas) al menos el 75 % del total? El 75 % de 8 es 6, por lo que buscamos las mediciones que dejan a su izquierda al menos 6 mediciones. Estas son

$$84\ 91\ 116.$$

Definimos q_s como la más chica de éstas. Es decir, $q_s = 84$. Análogamente, las mediciones que dejan a su izquierda (incluyéndose a sí mismas) al menos el 25 % del total, son en este caso

$$48\ 58\ 72\ 74\ 84\ 91\ 116.$$

Definimos q_i como la más chica, es decir $q_i = 48$.

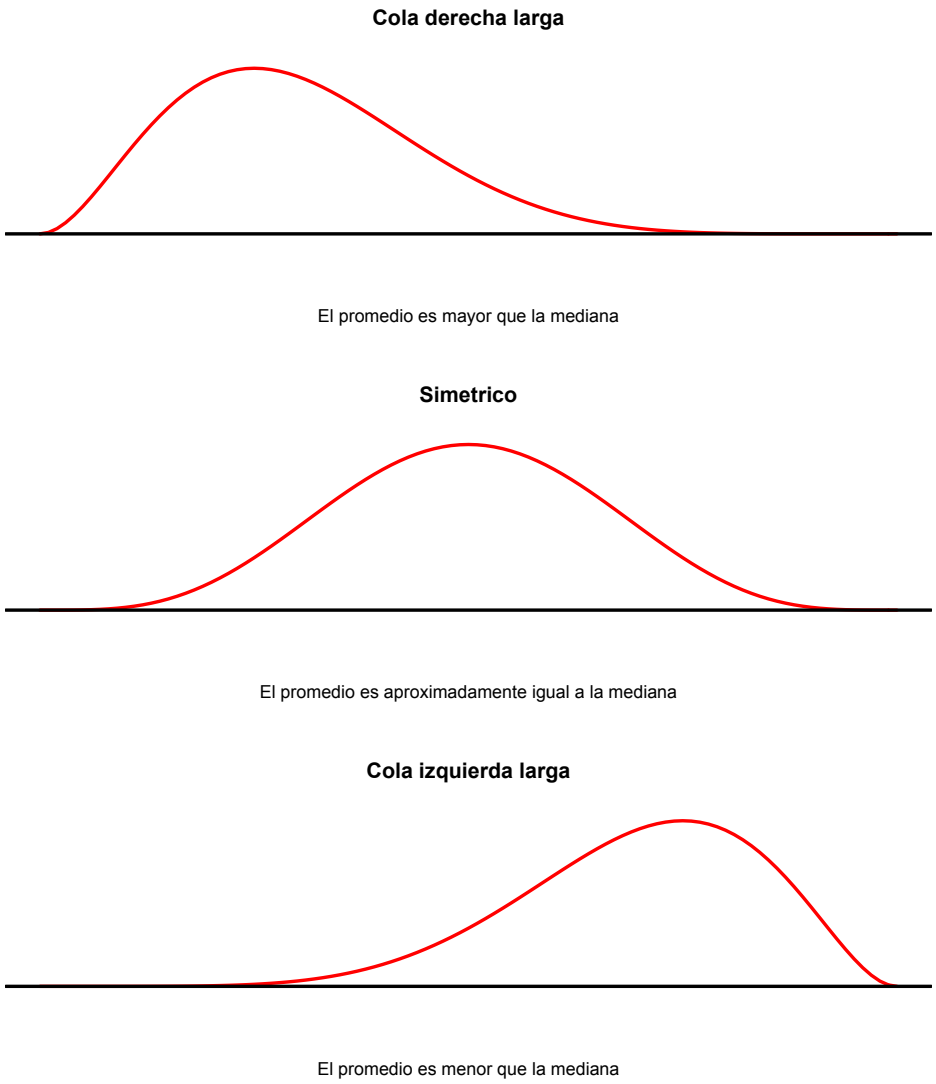

Figura 4: Relación entre la simetría de un histograma con el promedio y la mediana.

Es útil complementar la información proporcionada por la mediana y los cuartiles especificando también los valores extremos del conjunto de mediciones. En nuestro ejemplo, estos son 46 y 116. Podemos resumir la información de la siguiente manera:

mín	q_i	m	q_s	máx
46	48	72	84	116

Estos cinco números constituyen el *resumen de cinco números* para los datos dados. Entonces, un resumen de cinco números consta de la mediana, los dos cuartiles y los dos extremos.

En general, definimos el cuartil inferior q_i como la medición más chica que deja a su izquierda (incluyéndose a sí misma) al menos el 25 % del total. Analogamente, definimos el cuartil superior q_s como la medición más chica que deja a su izquierda al menos el 75 % del total.

Recordar que la profundidad de una medición nos dice exactamente cuántas mediciones

están a su izquierda. Por lo que podemos escribir la definición de los cuartiles de la siguiente forma.

Dadas n mediciones $x_1, x_2, \ldots, x_n$, definimos los cuartiles

$$q_i = \text{mín}\left\{x_i : p(x_i) \geq \frac{n}{4}\right\},$$
$$q_s = \text{mín}\left\{x_i : p(x_i) \geq \frac{3n}{4}\right\}.$$

En el ejemplo de las notas de los $n = 97$ estudiantes de la Tabla 1, tenemos que $n/4 = 97/4 = 24.25$. Luego, debemos mirar la medición más chica con $p(x_i) \geq 24.25$. Como la profundidad es un entero, la definición es equivalente a $p(q_i) = 25$, es decir $q_i = 68$. Análogamente, $3n/4 = 72.75$, por lo que buscamos $p(q_s) = 73$. Esto es $q_s = 89$.

Entonces, el resumen de cinco número queda

mín	q_i	m	q_s	máx
25	68	81	89	116

La diferencia entre los cuartiles se llama *rango intercuartílico*, lo denotamos por R y es una forma de caracterizar las dispersión de las mediciones correspondientes. En el ejemplo de las notas, tenemos $R = 89 - 68 = 21$.

Dadas n mediciones $x_1, x_2, \ldots, x_n$, definimos el rango intercuartílico como

$$R = q_s - q_i.$$

Es una forma de medir la dispersión de las mediciones. Notar que el intervalo $[q_i, q_s]$ contiene aproximadamente al 50 % de los datos.

4. Identificando posibles datos atípicos

Ahora estamos prontos para establecer una regla para identificar los posibles datos atípicos en un conjunto de mediciones. Hemos ya definido el rango intercuartílico

$$R = q_s - q_i.$$

Cualquier medición que esté a más de $1.5 \times R$ de distancia del cuartil más cercano se clasifica como sospechosa de ser atípica. Para el ejemplo de las notas de la Tabla 1, teníamos $R = 89 - 68 = 21$, por lo que $1.5 \times R = 31.5$. Cualquier nota que sea más chica que $68 - 31.5 = 36.5$ o más grande que $89 + 31.5 = 120.5$ es sospecha de ser inusual de una manera u otra. Por lo tanto, el puntaje más bajo de 25 es definitivamente sospechoso. Como la nota máxima posible es 120, no tenemos que preocuparnos de notas inusualmente altas.

Diagrama de caja

Un resumen de cinco números proporciona información numérica útil. Pero para la mayoría de las personas, la información numérica no es tan efectiva como una representación visual.

Un *diagrama de caja* proporciona una traducción visual de un resumen de cinco números. Un diagrama de caja consiste en una caja central que se extiende de q_i a q_u con la mediana m marcada con un segmento vertical.

A esta caja central, agregamos dos líneas horizontales que se extienden hacia los extremos como en la Figura 5. Por lo tanto, la caja representa el 50 por ciento central de las mediciones con la mediana que indica la línea divisoria entre el 25 por ciento central inferior y el superior.

Para hacer las líneas horizontales (a veces llamadas bigotes), debemos calcular las cantidades

$$\ell_i = q_i - 1.5R \text{ y } \ell_s = q_s + 1.5R.$$

El dato más grande que aún no es sospechoso de ser atípico es por lo tanto

$$x_+ = \text{máx}\{x_i : x_i \leq \ell_s\},$$

analogamante el más chico es

$$x_- = \text{mín}\{x_i : x_i \geq \ell_i\}.$$

La línea horizontal a la izquierda se extiende entonces hasta x_-; la línea horizontal a la derecha, hasta x_+. Las mediciones que caen fuera de los límites de estas lineas, son justamente los datos considerados sospechosos de ser atípicos.

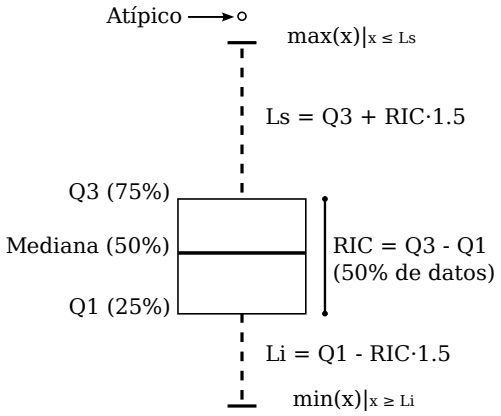

Figura 5: Definición general de un diagrama de caja.

Esta representación gráfica proporciona una imagen mucho más vívida de la distribución de las mediciones que el resumen de cinco números. Por ejemplo, para el ejemplo de las notas del examen, en la Figura 6, vemos que éstas se concentran alrededor del centro y se extienden hacia los extremos. También vemos que las notas no se extienden simétricamente desde la mediana.

5. Desvío estándar

Otra cantidad muy usada para caracterizar la dispersión de un conjunto de mediciones es el *desvío estándar*. Así como el promedio es análogo a la esperanza, el desvío estandar es análogo a la raíz de la varianza.

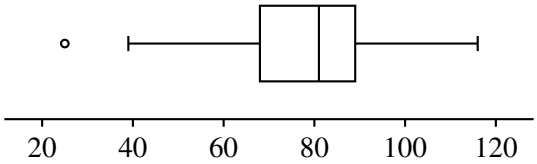

Figura 6: Diagrama de caja para las mediciones de la Tabla 1. Notar que hay un dato sospechosamente atípico.

Conceptualmente el desvío estándar es un poco oscuro. La *media cuadrática* x_{MC} de una lista de números, $x_1, x_2, \ldots, x_n$, es la raíz del promedio de sus cuadrados:

$$x_{MC} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}.$$

La media cuadrática es una forma de caracterizar el "tamaño" promedio de los números x_i. No es una medida de centro, pues no tiene en cuenta el signo de las mediciones.

Por ejemplo, el promedio de la lista

$$1,\ -3,\ 5,\ -6,\ 3$$

es $\bar{x} = 0$, pero la media cuadrática es

$$x_{MC} = \sqrt{\frac{1+9+25+36+9}{5}} \approx 8.94.$$

Volviendo al caso general de una lista de mediciones $x_1, \ldots, x_n$, consideremos los *desvíos del promedio* (con signo)

$$d_i = x_i - \bar{x},\ \ i = 1, \ldots, n.$$

El desvío estándar d_{MC} es la media cuadrática de los desvíos del promedio. Es más común usar $\hat{\sigma}$ para denotar el desvío estándar.

Dada una lista de mediciones $x_1, \ldots, x_n$, el desvío estándar se define como

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

Por ejemplo, consideremos las mediciones 48, 51, 49, 52, 47, 52, 46, 51, 53, 51. Para calcular el desvío estándar es es práctico hacer una tabla como la siguiente:

	x_i	d_i	d_i^2
	48	-2	4
	51	1	1
	49	-1	1
	52	2	4
	47	-3	9
	52	2	4
	46	-4	16
	51	1	1
	53	3	9
	51	1	1
Suma	500	0	50

Notar que la suma de los desvíos del promedio es siempre cero. Así que el desvío estándar es

$$\hat{\sigma} = \sqrt{\frac{50}{10}} = \sqrt{5} \approx 2.236.$$

En el ejemplo de la notas de la Tabla 1, el cáclulo es bastante mas tedioso, pero el desvío es igual a $\hat{\sigma} = 17.6$. Podemos comparar el desvío estándar con R el rango intercuartílico, que en este caso es $R = 21$. Ambas son medidas útiles de la dispersión.

Al igual que ocurre con la comparación entre el promedio y la mediana, el rango intercuartílico es más robusto que el desvío estándar, es decir, es menos sensible a datos atípicos.

	Robusta	**No Robusta**
Medida de centro	Mediana	Promedio
Medida de dispersión	Rango Intercuartílico	Desvío estándar

El $n-1$ en la definidicón de desvío estándar

En muchos libros de texto se define el desvío estándar dividiendo la suma de los desvíos al cuadrado entre $n-1$ en lugar de n. Para distinguirlo de $\hat{\sigma}$ se lo denota por la letra s:

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

Esto hace que la definición de desvío estándar sea aún más misteriosa, pues no solo elevamos los desvíos al cuadrado, ¡si no que dividimos entre $n-1$!

La razón más pertinente de hacer esto es para obtener medidas *insesgadas* de la dispersión, pero eso es un tema que veremos más adelante.

Nosotros usaremos tanto s como $\hat{\sigma}$, dependiendo del contexto. El uso de s se justifica cuando queremos hacer estudios más precisos, en los que disponemos de relativamente pocos datos. Notar que si n es grande, dividir entre $n-1$ o n es casi lo mismo.

Sin embargo, se puede justificar el uso de $n-1$ razonando de la siguiente manera. Una forma natural de medir la dispersión de una lista de mediciones es considerar todas las diferencias entre ellas. Es decir, en lugar de medir los desvíos al promedio d_i, medimos las diferencias

$$d_{ij} = x_i - x_j.$$

Imaginemos ahora que ponemos un resorte entre cada par de datos que va de x_i a x_j, que hace fuerza para acercarlos y por lo tanto quedar en una posición de equilibrio en la cual los dos puntos están juntos. Como se acordaran de física, la energía potencial de un resorte estirado una cantidad x es igual

$$\text{energía potencial del resorte} = \frac{1}{2}kx^2,$$

en donde k es la constante del resorte. Supongamos que nuestros resortes tienen constante $k = 1$, de modo que la energía potencial del par $\{i, j\}$ es

$$E_{ij} = \frac{1}{2}(x_i - x_j)^2 = \frac{1}{2}d_{ij}^2.$$

Cuanto más dispersos estén los datos, más estirados estarán los resortes, y más energía potencial tendrá la colección de mediciones. Usemos esta analogía para definir una energía potencial promedio:

$$E = \frac{1}{\binom{n}{2}} \sum_{\{i,j\}} E_{ij},$$

en donde la suma se extiende sobre todos los pares $\{i, j\}$ posibles. En total hay

$$\binom{n}{2} = \frac{n(n-1)}{2}$$

pares (resortes) distintos. Notar que la suma de los cuadrados de las diferencias es igual a

$$\begin{aligned}\sum_{\{i,j\}} E_{ij} &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} E_{ij} = \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(x_i^2 + x_j^2 - 2x_i x_j\right) \\ &= \frac{n}{2}\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 = \frac{n^2}{2}\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right] = \frac{n^2}{2}\hat{\sigma}^2.\end{aligned}$$

El último paso se debe a que

$$\hat{\sigma}^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2,$$

que se puede ver por cálculo directo. Entonces, si dividimos por la cantidad de pares distintos que hay, obtenemos

$$E = \frac{1}{\binom{n}{2}} \sum_{\{i,j\}} E_{ij} = \frac{n^2}{n(n-1)}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s^2.$$

Es decir, $s^2 = E$ es la energía potencial promedio de los resortes.