Estadística y probabilidad

Resumen para el grado séptimo

Carlos Andrés Pérez M.

I.E. Aureliano Flórez Cardona

1 Estadística descriptiva

En términos generales hablamos de **estadística descriptiva** cuando nos remitimos a tomar un conjunto de datos y realizar sobre estos la organización, sistematización y visualización para efectos de describirlos a través de ciertos *parámetros estadísticos* preestablecidos para tal efecto (medidas de tendencia central, de dispersión, de posición, etc.) y presentarlos bien sea en forma de *tablas* o *gráficas* apropiadas.

> NOTA: En este documento solo se plasman conceptos y procedimientos para estadística *univariada* (donde se analiza una sola variable).

1.1 Tablas y gráficas estadísticas

Una vez se tiene un conjunto de datos un procedimiento básico estándar consiste en realizar un conteo de repeticiones para cada uno de los valores que toma la variable analizada. Sin embargo, este conteo y el proceso que se sigue está sujeto al tipo de variable a saber:

- Si la variable es de tipo *cualitativo* (bien sea nominal u ordinal) se considera cada valor de la variable de manera independiente; esto es, se realiza un análisis de datos NO agrupados.
- Si la variable es de tipo *cuantitativo continuo* los distintos valores de la variable se analizan agrupados en intervalos; esto es, un análisis de datos agrupados.
- Si la variable es de tipo *cuantitativo discreto* el análisis de los datos puede hacerse agrupándolos o sin agrupar, dependiendo del caso y la cantidad de valores que toma la variable. Por ejemplo, si se analiza la cantidad de mascotas por hogar en las casas de los estudiantes de grado séptimo, lo más seguro es que no se necesite agrupar datos ya que las posibilidades no son muchas (1 mascota, 2 mascotas, 3 mascotas,…). Sin embargo, si se desea estudiar la cantidad de asistentes a las diferentes funciones en una sala de cine agrupar los datos seguramente sí sea lo más conveniente (de 1 a 10 personas, de 11 a 20, de 21 a 30, etc.) pues en este caso la variedad de datos distintos puede ser enorme.

1.1.1 Tablas para datos no agrupados

Como se mencionó arriba, en este caso simplemente consideramos cada valor de la variable por separado. Vamos a considerar dos ejemplos: uno para variable cualitativa y otro para variable cuantitativa discreta.

1.1.1.1 Ejemplo 1

Supongamos que se consulta a un grupo de 30 estudiantes acerca de su materia favorita y se obtienen los siguientes datos:

```
##  [1] "Lenguaje"    "Matemáticas" "Lenguaje"    "Matemáticas" "Lenguaje"
##  [6] "Lenguaje"    "Historia"    "Lenguaje"    "Lenguaje"    "Matemáticas"
## [11] "Ciencias"    "Historia"    "Matemáticas" "Lenguaje"    "Ciencias"
## [16] "Ciencias"    "Matemáticas" "Lenguaje"    "Lenguaje"    "Ciencias"
## [21] "Lenguaje"    "Ciencias"    "Lenguaje"    "Ciencias"    "Lenguaje"
## [26] "Matemáticas" "Ciencias"    "Historia"    "Lenguaje"    "Matemáticas"
```

Si realizamos un conteo de estos datos tendríamos:

Asignatura	Conteo
Ciencias	7
Historia	3
Lenguaje	13
Matemáticas	7

Dado que se trata de una variable cualitativa nominal entonces el orden de los valores de la variable en la tabla y gráfica puede ser cualquiera. Conservaremos para este ejemplo el orden alfabético.

Asignatura	f_i	F_i	h_i	H_i
Ciencias	7	7	0.2333333	0.2333333
Historia	3	10	0.1000000	0.3333333
Lenguaje	13	23	0.4333333	0.7666667
Matemáticas	7	30	0.2333333	1.0000000

Donde

- f_i corresponde a la *frecuencia absoluta* (conteo de cada valor de la variable entre los datos analizados),
- F_i la *frecuencia absoluta acumulada* (que a la cuenta actual le suma los anteriores f_i: $7 = 7$, $7 + 3 = 10$, $7 + 3 + 13 = 23$, $7 + 3 + 13 + 7 = 30$),
- h_i la *frecuencia relativa* (razón entre f_i y el total de datos; ej. $7/30 = 0.2333333$) y
- H_i que es la *frecuencia relativa acumulada* (que se comporta igual a F_i pero con h_i y por ende, también puede calcularse como F_i dividido entre el total de datos).

Se suele, generalmente, agregar una línea al final para los totales, quedando entonces de la siguiente manera:

Asignatura	f_i	F_i	h_i	H_i
Ciencias	7	7	0.2333333	0.2333333
Historia	3	10	0.1000000	0.3333333
Lenguaje	13	23	0.4333333	0.7666667
Matemáticas	7	30	0.2333333	1.0000000
Total	30		1.0000000	

Un par de consideraciones importantes:

- Las frecuencias relativas también pueden expresarse o bien como *fracción* ($0.233333 = 7/30$) o como *porcentaje* (ej. $0.233333 = 23.3333\%$).
- Advíertase además que el último valor de las frecuencias absolutas acumuladas ha de coincidir con el total de datos (n), así como el último valor de las frecuencias relativas acumuladas ha de coincidir con la unidad (o el 100%, si se expresa como porcentaje).

1.1.1.2 Ejemplo 2

Supongamos que se indaga acerca de la cantidad de asignaturas que perdió un grupo de 40 estudiantes durante el tercer periodo académico. Los resultados fueron los siguientes:

```
##  [1] 2 2 3 3 2 2 0 1 2 5 2 0 1 3 2 1 2 5 0 2 2 3 5 0 3 1 4 2 0 2 0 3 4 0 1
## [36] 1 2 1 2 3
```

Para visualizarlos un poco mejor los ordenaremos:

```
##  [1] 0 0 0 0 0 0 0 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3
## [36] 4 4 5 5 5
```

Un conteo previo nos da:

Materias perdidas	Alumnos
0	7
1	7
2	14
3	7
4	2
5	3

Al realizar la tabla de distribución de frecuencias para datos cuantitativos discretos o cualitativos ordinales el orden debe respetarse. La tabla nos quedaría en este caso:

Perdidas	f_i	F_i	h_i	H_i	$x_i \cdot f_i$
0	7	7	0.175	0.175	0
1	7	14	0.175	0.350	7
2	14	28	0.350	0.700	28
3	7	35	0.175	0.875	21
4	2	37	0.050	0.925	8
5	3	40	0.075	1.000	15
Total	40		1.000		79

Nótese que para esta tabla agregamos una columna al final que corresponde al producto entre el valor de la variable y su respectiva frecuencia absoluta. Esta columna resulta muy útil y más si consideramos su suma, pues permite hallar fácilmente el promedio total de los datos.

1.1.2 Tablas para datos agrupados

El análisis de datos agrupados es un poco más dispendioso y, de hecho, también más caprichoso.

Lo primero que debe advertirse es que lo más importante es determinar en cuántos intervalos (m) o clases se han de agrupar los datos. No existe una regla universal y lo más seguro es que la naturaleza misma de los datos y/o los requerimientos puntuales de quien los analiza sugiera cómo ha de realizarse el particionado.

No es necesario que los intervalos sean del mismo tamaño, aunque este es el procedimiento más acostumbrado. Ahora, si el usuario no sabe cuántos intervalos tomar, existen algunas reglas que se suelen usar en estos casos:

- *Regla de la raíz*: Si n es la cantidad de datos, entonces un buen número de intervalos a tomar sería $m = \sqrt{n}$.
- *Regla de Sturges*: Si n es la cantidad de datos, entonces un buen número de intervalos a tomar sería $m = 1 + 3.322 \log_{10} n$.

A título de ejemplo, si contamos con 150 datos, entonces por la regla de la raíz deberíamos usar $\sqrt{150} = 12.2474 \approx 12$ intervalos; mientras que por la regla de Sturges serían $1 + 3.322 \log_{10}(150) = 8.229 \approx 8$ intervalos.

En realidad no son reglas como tal, sino sugerencias. Así que es en últimas quien analiza los datos el que debe decidir cuál de los dos valores usar (o por el contrario usar un tercero cercano a estos pero más conveniente de acuerdo a su criterio).

Una vez definida la cantidad de intervalos se realiza la partición de los mismos, para la cual se considera como primera medida los valores límite de los datos (mínimo y máximo) y el rango entre estos (ver ejemplo 3). Seguida-

mente se realiza el conteo de los datos considerando los intervalos escogidos y lo que se sigue es similar que para los datos no agrupados.

1.1.2.1 Ejemplo 3

Se tiene un grupo de 80 estudiantes a los cuales se les mide su estatura (en metros) para obtener los siguientes datos:

```
##  [1] 1.60 1.61 1.62 1.63 1.60 1.59 1.55 1.59 1.61 1.67 1.60 1.56 1.58 1.63
## [15] 1.51 1.57 1.60 1.67 1.55 1.60 1.61 1.63 1.68 1.55 1.62 1.53 1.66 1.60
## [29] 1.55 1.61 1.55 1.64 1.64 1.56 1.53 1.58 1.60 1.59 1.61 1.62 1.65 1.62
## [43] 1.59 1.59 1.58 1.60 1.51 1.61 1.57 1.59 1.62 1.54 1.56 1.62 1.60 1.67
## [57] 1.55 1.60 1.66 1.60 1.62 1.58 1.75 1.60 1.54 1.60 1.62 1.61 1.62 1.57
## [71] 1.60 1.60 1.64 1.69 1.60 1.64 1.66 1.60 1.67 1.68
```

Organizamos los datos en forma ascendente y obtenemos:

```
##  [1] 1.51 1.51 1.53 1.53 1.54 1.54 1.55 1.55 1.55 1.55 1.55 1.55 1.56 1.56
## [15] 1.56 1.57 1.57 1.57 1.58 1.58 1.58 1.58 1.59 1.59 1.59 1.59 1.59 1.59
## [29] 1.60 1.60 1.60 1.60 1.60 1.60 1.60 1.60 1.60 1.60 1.60 1.60 1.60 1.60
## [43] 1.60 1.60 1.60 1.61 1.61 1.61 1.61 1.61 1.61 1.61 1.62 1.62 1.62 1.62
## [57] 1.62 1.62 1.62 1.62 1.62 1.63 1.63 1.63 1.64 1.64 1.64 1.64 1.65 1.66
## [71] 1.66 1.66 1.67 1.67 1.67 1.67 1.68 1.68 1.69 1.75
```

Es claro que el valor mínimo es 1.51 mientras que el máximo es 1.75. Por tanto, el rango es $R = 1.75 - 1.51 = 0.24$ metros.

Definiremos ahora la cantidad de intervalos a usar considerando las dos reglas vistas:

- $m = \sqrt{80} = 8.9443 \approx 9$ intervalos (regla de la raíz)
- $m = 1 + 3.322 \log_{10}(80) = 7.3221 \approx 7$ intervalos (regla de Sturges)

Usaremos entonces, de manera conveniente 8 intervalos básicamente por dos razones: 1) es un valor intermedio a los dos sugeridos, y por tanto muy razonable, y 2) es evidente que 8 divide a 0.24 (el rango) sin dejar exceso de decimales.

NOTA: Si la cantidad de intervalos no divide de manera cómoda al rango entonces puede optarse o por ampliar el rango a ambos extremos de los límites hasta acomodar las particiones o forzar las divisiones sin importar los decimales.

La amplitud de cada intervalo sería entonces $c = 0.24/8 = 0.03$ metros. (En otras palabras, se harán grupos de estudiantes en orden creciente cada 3 centímetros)

Las particiones que vamos a buscar son entonces:

```
## [1] 1.51 1.54 1.57 1.60 1.63 1.66 1.69 1.72 1.75
```

Un conteo de las estaturas para los intervalos planteados nos da:

Intervalo de estatura	Estudiantes
[1.51,1.54]	6
(1.54,1.57]	12
(1.57,1.60]	27
(1.60,1.63]	19
(1.63,1.66]	8
(1.66,1.69]	7
(1.69,1.72]	0
(1.72,1.75]	1

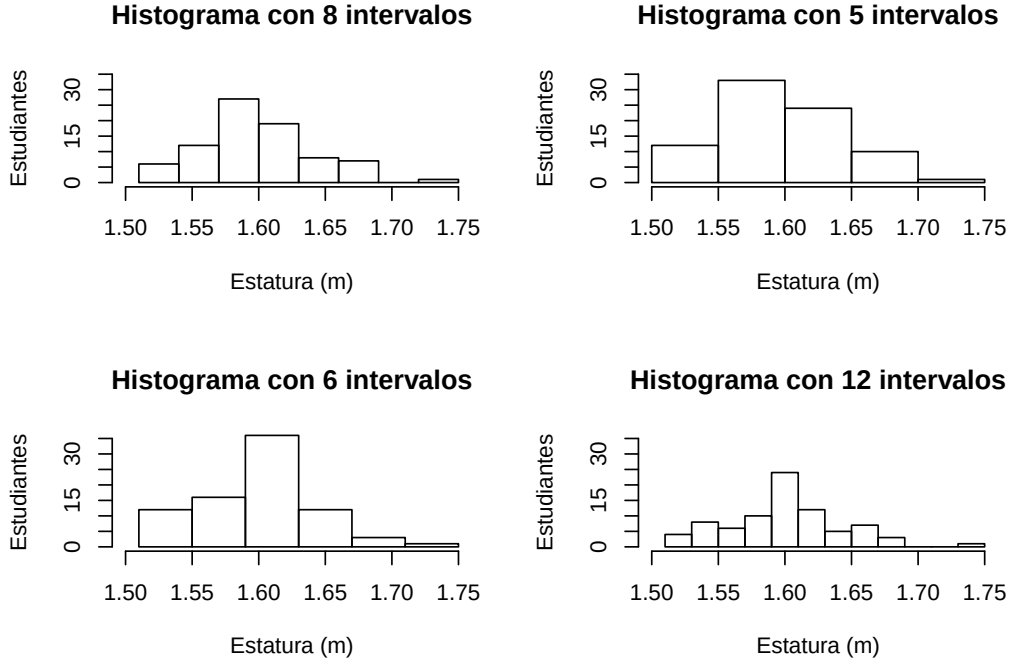
El paréntesis cuadrado indica que el extremo forma parte del intervalo y el paréntesis curvo que no. A manera de ejemplo, si tenemos una estatura de 1.63 metros, esta formaría parte del intervalo $(1.60, 1.63]$ y no del intervalo $(1.63, 1.66]$.

Estatura	X_i	f_i	F_i	h_i	H_i	$X_i \cdot f_i$
[1.51,1.54]	1.525	6	6	0.0750	0.0750	9.150
(1.54,1.57]	1.555	12	18	0.1500	0.2250	18.660
(1.57,1.60]	1.585	27	45	0.3375	0.5625	42.795
(1.60,1.63]	1.615	19	64	0.2375	0.8000	30.685
(1.63,1.66]	1.645	8	72	0.1000	0.9000	13.160
(1.66,1.69]	1.675	7	79	0.0875	0.9875	11.725
(1.69,1.72]	1.705	0	79	0.0000	0.9875	0.000
(1.72,1.75]	1.735	1	80	0.0125	1.0000	1.735
Total		80		1.0000		127.910

En este caso aparece una nueva columna llamada X_i; es la columna de las *marcas de clase* las cuales no son otra cosa que los valores intermedios de cada intervalo y que para efectos prácticos se considerará como el valor representativo del intervalo en cuestión. A manera de ejemplo: $\frac{1}{2}(1.51 + 1.54) = 1.525$.

Adicionalmente, la última columna se vio modificada en virtud de lo anteriormente dicho.

Es importante saber que la cantidad de intervalos escogidos afecta la forma en que se van a visualizar los datos (y los cálculos realizados a partir de los valores de la tabla). A continuación se muestran 4 histogramas (diagramas especiales para datos agrupados que se explicarán con más detalle más adelante) para el mismo conjunto de datos pero con diferente cantidad de intervalos y particiones:



1.1.3 Principales gráficos estadísticos

Son muchos los tipos de gráficos estadísticos existentes debido primero en parte al tipo de datos manejados y segundo a lo que se quiere comunicar con ellos.

Entre los principales tenemos los diagramas *circulares*, los *diagramas de barras* y los *histogramas*.

IMPORTANTE: Aunque parezca tentador hacerlos, los gráficos estadísticos con elementos decorativos en tres dimensiones deben evitarse (barras 3D, pasteles 3D, etc). Estos no aportan sino problemas a la correcta visualización e interpretación.

1.1.3.1 Diagramas circulares

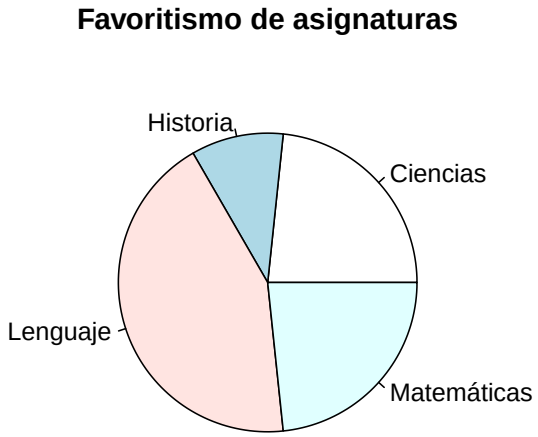
También conocidos como diagramas tipo *pastel*. Son usados generalmente para representar porcentajes para variables de datos no agrupados.

Se trata de dividir un círculo en regiones proporcionales a la frecuencia hallada para la variable estudiada. Dado que el círculo barre un ángulo total de 360° entonces el ángulo que le corresponde a cada sector para su respectivo valor de variable se calcula mediante cualquiera de las siguientes expresiones:

$$\theta_i = 360^\circ \left(\frac{f_i}{n}\right); \quad \theta_i = 360^\circ h_i; \quad \theta_i = 360^\circ \left(\frac{h_i}{100\%}\right)$$

(esta última expresión si la frecuencia relativa está expresada como porcentaje)

Para el caso del ejemplo 1 tendríamos:



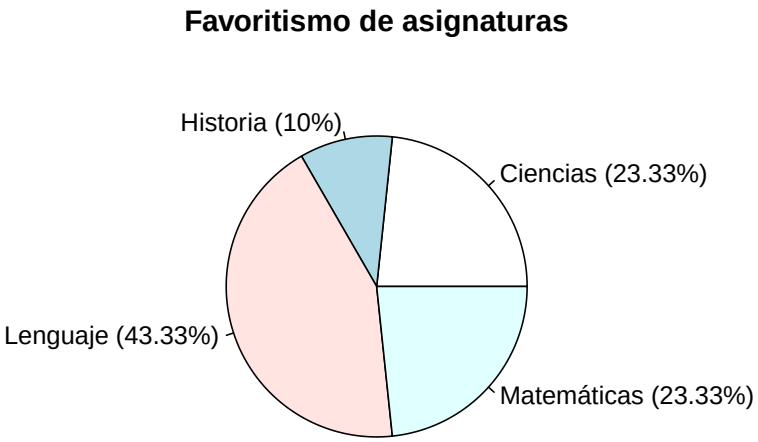
donde el ángulo del sector correspondiente para "Ciencias" es

$$\theta_1 = 360^\circ \left(\frac{7}{30}\right) = 84^\circ$$

(y así con los demás).

Este tipo de diagrama, aunque fácil de entender, no siempre es el mejor en virtud de que su misma representación a veces no permite comparar tan fácilmente entre las diferentes regiones circulares. Por tanto, se sugiere agregar información adicional para clarificar y ayudar en la lectura (preferiblemente los porcentajes respectivos a cada sector).

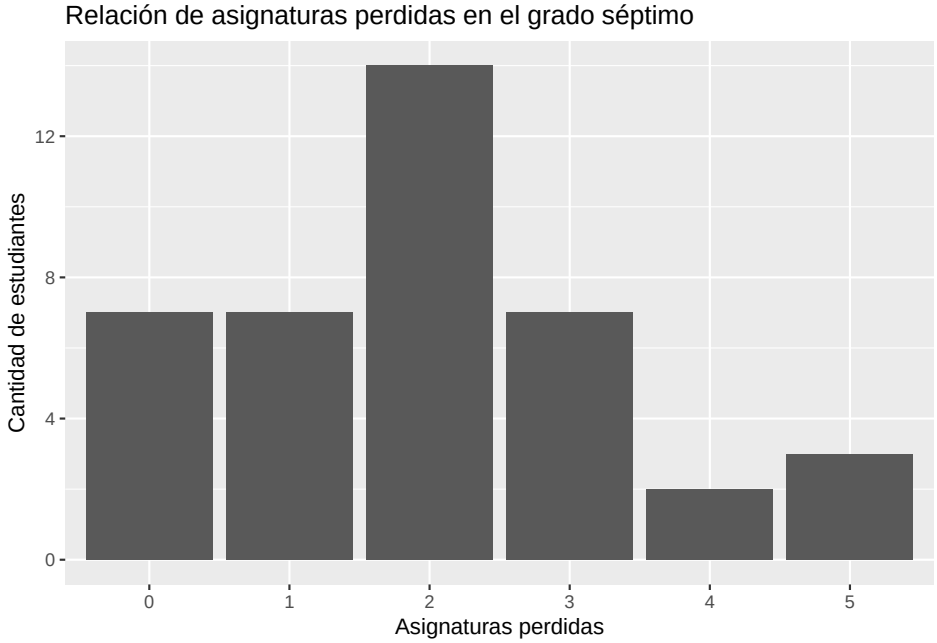
A continuación el mismo gráfico circular pero con porcentajes en sus etiquetas:



1.1.3.2 Diagramas de barras

Este es quizá el tipo de diagrama más común y fácil de leer. Relaciona la frecuencia relativa con barras de altura igual a dichas frecuencias.

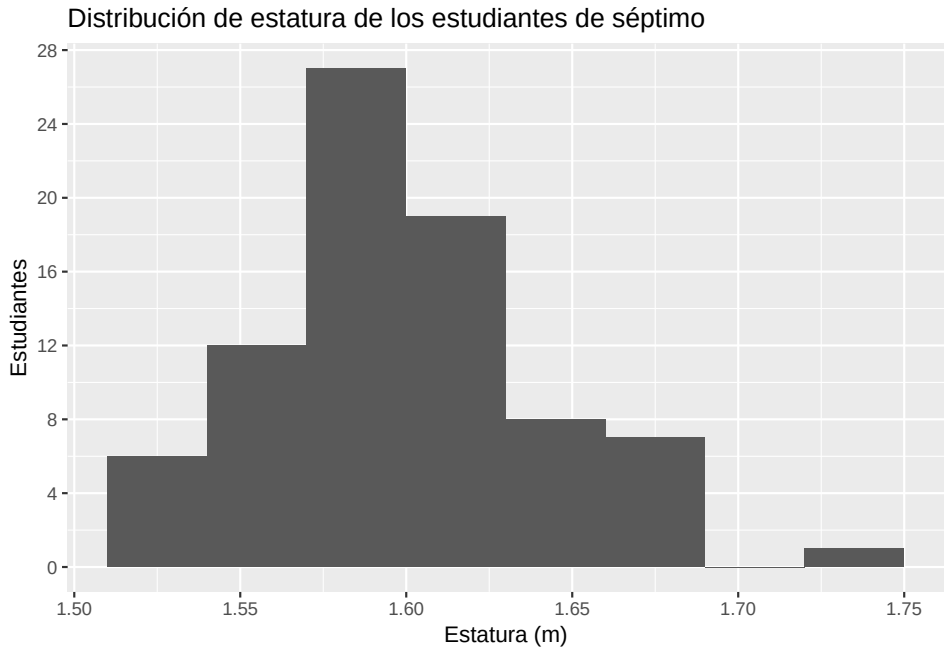
Si consideramos los datos y tabla respectiva del ejemplo 2 tendríamos el siguiente diagrama de barras:



Vale la pena mencionarse que las barras también pueden plasmarse de manera horizontal si el usuario así lo desea. Además, no solo se usan para representar frecuencias absolutas sino también porcentajes si la situación lo amerita.

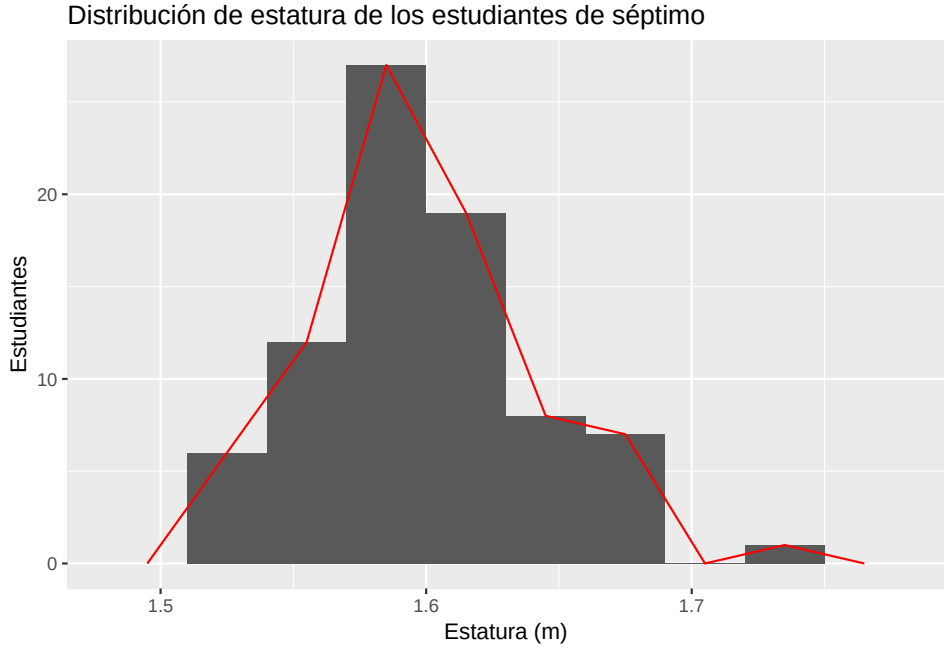
1.1.3.3 Histogramas

Los histogramas son diagramas de barras pero para datos agrupados, esto es, para datos que han sido distribuidos por intervalos; ahora veremos un histograma para los datos del ejemplo 3 (también con 8 intervalos):



Nótese que hay continuidad en las barras para los valores del eje horizontal (es decir, no hay separación como en el diagrama de barras) lo cual se debe a que los histogramas llevan la secuencia completa de valores durante todo el rango de datos.

En algunos casos es útil agregar un *polígono de frecuencias* a los datos, el cual relaciona las marcas de clase de cada intervalo con su respectiva frecuencia absoluta. Para combinar estos dos gráficos hacemos:



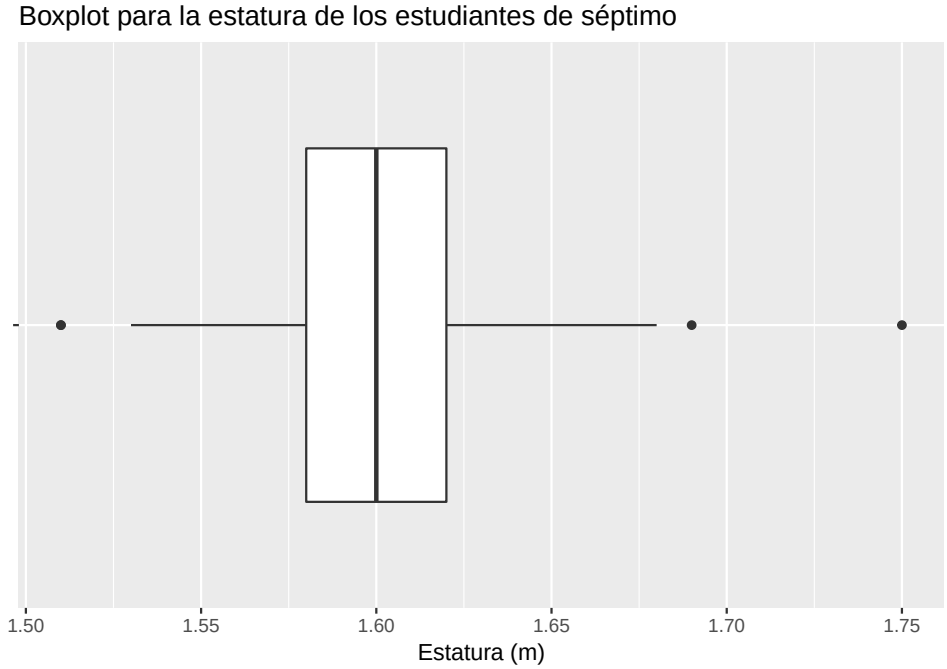
1.1.3.4 Otros tipos de diagramas (para datos univariados)

A continuación se mencionan otros tipos útiles de gráficos estadísticos que, aunque su construcción no se detalla por escaparse del alcance de este documento, sí se mencionan para contar con una necesaria familiarización.

Ya se mencionó que el problema de los histogramas es que la cantidad de intervalos puede modificar un poco la

percepción que se tiene de la distribución de los datos. Este problema es resuelto por otro tipo de diagrama muy útil: el *gráfico de densidad*, el cual juega con las probabilidades.

Otro tipo de diagrama muy usado es el "*boxplot*" o *diagrama de caja y bigotes* el cual permite visualizar la distribución de datos a partir de otra perspectiva: los cuartiles. Cada una de las cuatro regiones corresponde a un cuartil distinto.



Una ventaja interesante de este tipo de gráfico es que permite identificar muy fácilmente valores atípicos y, además, es excelente para comparar distribuciones de datos que comparten alguna relación.

1.2 Medidas de tendencia central

Este tipo de parámetros estadísticos nos permiten determinar a qué valor tienden los datos a centrarse o acumularse. Son muchos, pero para efectos prácticos y para el alcance de este documento consideraremos los tres más importantes: moda, mediana y media.

La **moda** (*Mo*) es el *valor que más se repite dentro de un conjunto de datos*. Si miramos los datos del ejemplo 1 es claro que este valor corresponde a "Lenguaje", y para el ejemplo 2 es "2 materias perdidas".

Para el caso del ejemplo 3 tenemos un problema: para datos continuos y agrupados no siempre es posible hablar con propiedad de una moda como tal en virtud de que potencialmente los valores que puede tomar la variable son infinitos. En este caso es posible hablar de un *intervalo modal* o clase modal que en el caso del ejemplo 3 sería $(1.57, 1.60]$.

Si definitivamente se desea tener un valor que, a manera de moda, represente de manera aproximada los datos agrupados una vez se tiene la tabla de distribución de frecuencias para estos, usamos la relación

$$Mo = L_m + \frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} c_m$$

donde L_m hace referencia al límite inferior de la clase o intervalo modal, f_m, f_{m-1} y f_{m+1} a las frecuencias de la clase modal, la anterior y la posterior a esta, respectivamente, y finalmente c_m la amplitud de la clase modal.

Como ejemplo recordemos la tabla para los datos del ejemplo 3:

Estatura	X_i	f_i	F_i	h_i	H_i	$X_i \cdot f_i$
[1.51,1.54]	1.525	6	6	0.0750	0.0750	9.150
(1.54,1.57]	1.555	12	18	0.1500	0.2250	18.660
(1.57,1.60]	1.585	27	45	0.3375	0.5625	42.795
(1.60,1.63]	1.615	19	64	0.2375	0.8000	30.685
(1.63,1.66]	1.645	8	72	0.1000	0.9000	13.160
(1.66,1.69]	1.675	7	79	0.0875	0.9875	11.725
(1.69,1.72]	1.705	0	79	0.0000	0.9875	0.000
(1.72,1.75]	1.735	1	80	0.0125	1.0000	1.735
Total		80		1.0000		127.910

En este caso se tendría que

$$Mo = 1.57 + \frac{27 - 12}{(27 - 12) + (27 - 19)}(0.03) = 1.589565 \approx 1.59$$

La **mediana** (*Me*) representa el *dato que divide al conjunto total de datos en dos partes iguales*. Dicho de otra manera, la mediana es el valor para el cual el 50% de los datos son menores que este (o mayores, según como quiera verse).

Consideremos estos dos conjuntos de datos a los que llamaremos respectivamente *Datos x* y *Datos y*:

```
##  [1] 3 5 6 8 3 5 4 6 3 5 4 2 7
```

```
##  [1] 2 4 5 2 1 2 4 5 2 1 3 1 4 5 3 2
```

Para calcular la mediana primero ordenamos los datos en orden ascendente:

```
##  [1] 2 3 3 3 4 4 5 5 5 6 6 7 8
```

```
##  [1] 1 1 1 2 2 2 2 2 3 3 4 4 4 5 5 5
```

Ahora, si la cantidad de datos es *impar* (como ocurre con los datos *x* donde $n = 13$) entonces usamos la relación $Me = X_{(n+1)/2}$ donde X es el conjunto de datos en cuestión ya ordenado ascendentemente.

Por tanto, tendríamos en este caso que $Me = X_{(13+1)/2} = X_7 = 5$ que según como se ve, podríamos afirmar que "el 50% de los datos de x son menores que 5" (o "mayores o iguales que 5").

Si por el contrario la cantidad de datos analizados es *par* (como ocurre con los datos y donde $n = 16$), entonces la mediana equivale al promedio de los dos valores intermedios del conjunto de datos ordenado: $Me = \frac{1}{2}\left(X_{n/2} + X_{n/2+1}\right)$.

De esta manera se tendría que $Me = \frac{1}{2}\left(X_{16/2} + X_{16/2+1}\right) = Me = \frac{1}{2}\left(X_8 + X_9\right) = Me = \frac{1}{2}\left(2+3\right) = 2.5$; esto es, "el 50% de los datos del conjunto y son menores que 2.5" (o "mayores que 2.5").

Si los datos ya se encuentran tabulados, el criterio se aplica igual pero considerando que dicho valor ha de buscarse en la columna de frecuencias absolutas acumuladas. Por ejemplo, si miramos la tabla del ejemplo 2:

Perdidas	f_i	F_i	h_i	H_i	$x_i \cdot f_i$
0	7	7	0.175	0.175	0
1	7	14	0.175	0.350	7
2	14	28	0.350	0.700	28
3	7	35	0.175	0.875	21
4	2	37	0.050	0.925	8
5	3	40	0.075	1.000	15
Total	40		1.000		79

tendríamos que al haber 40 datos en total, los datos que ocupen los lugares 20 y 21 en la lista ordenada determinarían la mediana. De acuerdo a la columna F_i, estos datos deben ser ambos 2 y 2, por lo que la mediana es, en efecto, 2.

> NOTA: La mediana NO se aplica a datos de variables cualitativas, solo cuantitativas, a diferencia de la moda que sí puede aplicarse a variables cualitativas.

Para el caso de datos agrupados, la relación que permite determinar una aproximación razonable para la mediana es

$$Me = L_m + \frac{n/2 - F_{m-1}}{f_m} c_m$$

donde L_m es el límite de la clase donde se encuentre la mediana (de acuerdo a columna de frecuencias absolutas acumuladas), n es la cantidad de datos, F_{m-1} es el valor de la frecuencia absoluta acumulada de la clase anterior a la clase de la mediana, f_m es la frecuencia absoluta de la clase de la mediana, y c_m la amplitud de la clase de la mediana.

Recordemos nuevamente los datos de la tabla 3:

Estatura	X_i	f_i	F_i	h_i	H_i	$X_i \cdot f_i$
[1.51,1.54]	1.525	6	6	0.0750	0.0750	9.150
(1.54,1.57]	1.555	12	18	0.1500	0.2250	18.660
(1.57,1.60]	1.585	27	45	0.3375	0.5625	42.795
(1.60,1.63]	1.615	19	64	0.2375	0.8000	30.685
(1.63,1.66]	1.645	8	72	0.1000	0.9000	13.160
(1.66,1.69]	1.675	7	79	0.0875	0.9875	11.725
(1.69,1.72]	1.705	0	79	0.0000	0.9875	0.000
(1.72,1.75]	1.735	1	80	0.0125	1.0000	1.735
Total		80		1.0000		127.910

Nótese que al haber 80 datos, la mediana estaría determinada por los valores que ocupen los puestos 40 y 41 de la lista ordenada de datos. Estos valores estarían (de acuerdo a la columna F_i) en el intervalo $(1.57, 1.60]$ (el cual llamaríamos como *intervalo mediano*).

Considerando lo anterior, para estos datos la mediana sería:

$$Me = 1.57 + \frac{80/2 - 18}{27}(0.03) = 1.59444444 \approx 1.59$$

La mediana real de este conjunto de datos (usando el paquete estadístico R) es de 1.6.

Finalmente tenemos que la **media** ($\bar{x}$) de un conjunto de datos corresponde al valor que se obtendría si todos los valores registrados se recogieran para luego ser redistribuidos de manera equitativa. Matemáticamente se obtiene al dividir la suma de todos los datos entre el total de los mismos:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

En realidad existen diferentes tipos de media; la descrita anteriormente es la *media aritmética* (o *promedio*). Otros tipos de media son la geométrica, la armónica, la podada, la winsorizada, entre otras, las cuales se escapan del alcance de este documento.

Si consideramos el siguiente conjunto de datos:

su media aritmética corresponde a

$$\bar{z} = \frac{3 + 5 + 6 + 7 + 7 + 2 + 5 + 40}{8} = \frac{39}{8} = 4.875$$

Si los datos ya se hallan tabulados en una tabla de datos no agrupados (como en el ejemplo 2), entonces la media se calcula rápidamente usando la relación

$$\bar{x} = \frac{\sum (x_i \cdot f_i)}{n} = \frac{79}{40} = 1.975$$

Si hablamos de datos agrupados ya tabulados y se desea tener una estimación válida para la media, esta se calcula usando la relación

$$\bar{x} = \frac{1}{n}\sum (X_i \cdot f_i)$$

que para el caso de los datos del ejemplo 3 sería:

$$\bar{x} = \frac{127.910}{80} = 1.598875$$

Calculada con el paquete estadístico R se tiene que la media exacta de los datos es de 1.604.

> NOTA: En la práctica, el cálculo de los parámetros estadísticos no se realiza a partir de las tablas de frecuencias sino sobre la data misma. Esto no solo resulta más cómodo ya que se cuenta con herramientas de software especializadas (como R) sino que también se garantiza más precisión.

2 Probabilidad

Hablar de probabilidad es hablar del grado de certeza y/o incertidumbre que, en mayor o menor medida, permea a los eventos que nos rodean; ie. la probabilidad de que esta tarde llueva, que mi equipo llegue la final, que un nuevo procedimiento médico tenga éxito, etc.

La **teoría de la probabilidad** es la rama de las matemáticas que nos ayuda a comprender y analizar el aparente caos que existe detrás de eso que llamamos *azar* y que, como ya se mencionó, imprime algún grado de incertidumbre a todos los fenómenos del Universo.

Aquellos eventos que se rigen enteramente por el azar se conocen como *experimentos aleatorios*. La clave para entender cómo funciona la probabiliad en dichos eventos es asumir que aunque no es posible conocer a priori el resultado final, sí se conocen todos los resultados que pueden darse. A manera de ejemplo, si vamos a lanzar un dado legal de seis caras es imposible estar seguros de cuál cara caerá, y sin embargo sí estar preparados para que sea un 1, 2, 3, 4, 5 ó 6.

2.1 Principios básicos de conteo

Para poder determinar la probabilidad de un evento lo primero que debemos conocer es todo el abanico de posibilidades que puede abarcar los resultados para dicho evento; este se conoce matemáticamente como *espacio muestral*.

Algunos ejemplos de espacio muestral son:

- Para el lanzamiento de una moneda es $\{C, S\}$ (donde C es cara y S es sello),
- pero para el lanzamiento de dos monedas sería $\{(C,C),(C,S),(S,C),(S,S)\}$;
- para un dado justo es $\{1,2,3,4,5,6\}$;
- para una rifa que juega con las tres últimas cifras de una lotería es $\{000, 001, 002, 003, \ldots, 998, 999\}$

Determinar el tamaño del espacio muestral en ocasiones es un poco más complejo (por ejemplo determinar todas las posibles combinaciones para el Baloto) y es precisamente en este caso que hacemos uso de algunos principios útiles como los que se presentan a continuación.

2.1.1 Principio de multiplicación

Se usa cuando se tienen dos o más *eventos independientes* (la ocurrencia de uno no afecta la del otro). En este caso, el evento combinado posee un espacio muestral cuyo tamaño es igual al producto entre todos los tamaños de los espacios muestrales de los eventos independientes.

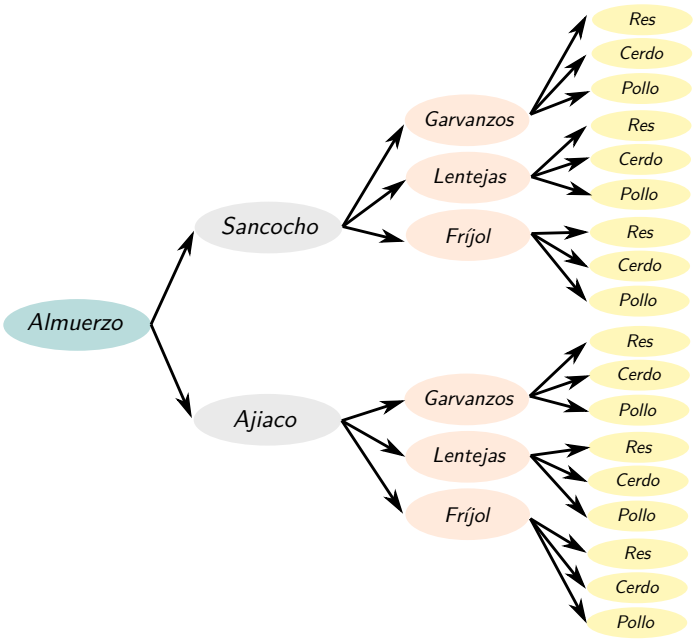
Por ejemplo, el tamaño del espacio muestral para el lanzamiento de una moneda es 2 y el de un dado es 6; pero si lanzamos una moneda y un dado a la vez (los cuales son eventos independientes entre sí), la cantidad total de posibles resultados es 12 ($2 \times 6 = 12$).

Como segundo ejemplo supóngase que en el menú de un restaurante se ofrece hoy como opciones para el almuerzo dos tipos de sopa (sancocho y ajiaco), tres tipos de principio (garvanzos, lentejas y fríjol), y tres tipos de carne (res, cerdo y pollo). La cantidad posible de almuerzos distintos de los que una persona puede disponer es $2 \times 3 \times 3 = 18$.

Un recurso muy utilizado para visualizar espacios muestrales y cómo funciona en ellos el principio de multiplicación son los *diagramas de árbol*[1]. En estos diagramas, cada nivel corresponde a un evento que se ramifica considerando las posibilidades para el evento siguiente.

El ejemplo de los posibles almuerzos, visualizado en forma de árbol quedaría como:

[1] En realidad este tipo de diagramas es muy versátil y existen muchas variantes usadas en estadística, probabilidad e inteligencia artificial, entre otras.



2.1.2 Factorial

El factorial de un número entero no negativo se define como $n! = n(n-1)(n-2)\cdots 1$; a manera de ejemplo $7! = 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5040$.

Se trata de una operación que aparece de manera muy seguida en algunas ramas de la matemática, en especial en *combinatoria*.

Sus propiedades más importantes (dentro del alcance de este documento) son:

- $0! = 1$
- $1! = 1$
- $n! = n \cdot (n-1)!$

El factorial puede ser usado para calcular el tamaño de espacios muestrales como el de los siguientes ejemplos:

- En la casa de Ana, Daniel y Mateo, solo hay un baño, por lo que el orden en que lo usan para asearse para ir al colegio depende de quién se levante más temprano. ¿De cuántas maneras puede ocurrir el orden de uso del baño un día cualquiera? **Respuesta**: De $3! = 3 \cdot 2 \cdot 1 = 6$ maneras diferentes.
- En la I.E. Aureliano Flórez hay cuatro séptimos los cuales compiten cada periodo por tener la mejor nota promedio del grupo. ¿De cuántas maneras podrían quedar organizados los cuatro grupos de acuerdo a su nota? **Respuesta**: De $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$ maneras diferentes.

2.1.3 Permutaciones y combinaciones

En el caso anterior se tomaban todos los elementos de un conjunto y se replanteaba su orden (sin repetir); sin embargo, en ocasiones es necesario tomar solo una parte de este conjunto (sin llegar a repetir elementos) y después ya determinar si el orden es relevante o no.

Si el orden importa tenemos una *permutación*, y si no importa una *combinación*.

Permutar r elementos de un total de n elementos (que se suele leer como "n permutado r") implica seleccionar r de estos de un conjunto de n totales *considerando el orden en que sean seleccionados*. Esto se calcula como

$$ {}_nP_r = \frac{n!}{(n-r)!} $$

El caso de la combinación ("n combinado r") es idéntico al de la permutación pero *sin coniderar el orden de la selección*, y se calcula como

$$ {}_nC_r = \frac{n!}{(n-r)!r!} $$

A título de ejemplo a la final de 100 metros planos llegaron 8 atletas; si se desea saber de cuántas maneras se puede formar el podio (oro, plata y bronce) haríamos la permutación (ya que el orden sí importa en este caso)

$$ {}_8P_3 = \frac{8!}{(8-3)!} = \frac{8 \cdot 7 \cdot 6 \cdot 5!}{5!} = 336 \text{ podios distintos.} $$

En el caso del modelo antiguo del Baloto, donde se sacaban al azar 6 balotas sin repetición de un total de 45 (numeradas del 1 al 45), se tenía una combinación ya que, a diferencia del ejemplo anterior, el orden de salida de las balotas no importa. La cantidad posible de combinaciones del Baloto mediante ese sistema era de

$$ {}_{45}C_6 = \frac{45!}{(45-6)!6!} = \frac{45 \cdot 44 \cdot 43 \cdot 42 \cdot 41 \cdot 40 \cdot 39!}{39! \cdot 6!} = \frac{45 \cdot 44 \cdot 43 \cdot 42 \cdot 41 \cdot 40}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 8145060. $$

2.2 Probabilidad de ocurrencia de un evento

La probabilidad de ocurrencia de un evento X se define como la razón entre la cantidad de posibles casos favorables a X y la cantidad de eventos totales, esto es:

$$ P(X) = \frac{\text{Cantidad de eventos favorables}}{\text{Cantidad de eventos totales}} $$

De esta manera, si se tiene un grupo de 30 estudiantes de los cuales 21 son mujeres, la probabilidad de escoger una mujer al azar de entre dicho grupo corresponde a $P(M) = 21/30 = 7/10 = 0.7$.

Alternativamente una probabilidad puede expresarse como porcentaje quedando, para el caso anterior, que $P(M) = 0.7 \times 100\% = 70\%$.

El valor de la probabilidad de un evento será siempre un valor entre cero y uno, es decir, $0 \leq P(X) \leq 1$, siendo $P(X) = 0$ un evento *imposible* (como sacar un 8 de un dado estándar) y $P(X) = 1$ un evento *seguro* (como sacar un número menor que 10 cuando se lanza un dado estándar).

La relación entre la ocurrencia de un evento $P(X)$ y su no ocurrencia $P(\neg X)$ es

$$ P(X) + P(\neg X) = 1. $$

A saber, si consideramos el ejemplo anterior, la probabilidad de que el estudiante escogido al azar no sea mujer es $P(\neg M) = 1 - 0.7 = 0.3$ (o del 30%).

Es posible deducir expresiones relacionadas con esta propiedad para casos más específicos en donde un experimento aleatorio cuenta con un número finito de posibilidades. Por ejemplo para el caso del lanzamiento de una moneda sería $P(\text{cara}) + P(\text{sello}) = 1$; para un dado de seis caras sería que $P(1)+P(2)+P(3)+P(4)+P(5)+P(6) = 1$; para el resultado de un partido $P(\text{ganar}) + P(\text{perder}) + P(\text{empatar}) = 1$; etc.