Probabilidad y Estadística (93.24)
Estadística descriptiva

Índice

1. Repaso de algunos conceptos

Población y muestra

Se denomina **población** al conjunto de elementos o individuos cuyas características quieren estudiarse. A veces es imposible (en un población infinita), costoso (una población grande) o ridículo (si para realizar el estudio hay que destruir los elementos estudiados) analizar todo la población. Por estos motivos, comúnmente se estudia un subconjunto denominado **muestra**. Idealmente, se espera que la muestra sea *representativa* de toda la población. Una forma, no siempre sencilla, de obtener una muestra representativa es eligiendo los elementos al azar de entre toda la población.
Sea $\{x_i\}_{i=1}^n$ el conjunto de características relevadas de la población, siendo n el número de elementos o individuos muestrados. Existen dos tipos de muestras, *cualitativas* y *cuantitativas*. Mientras las primeras sólo contienen características cualitativas (por ej., la profesión de un individuo), las segundas contienen características numéricas. Ciertamente, también puede haber muestras que contengan tanto elementos cuantitativos como cualitativos. Nosotros nos concentraremos en muestras cuantitativas.

Cuartiles

Cuando la característica de interés toma valores reales, se definen tres **cuartiles** que separan la muestra en cuatro partes. En particular, sea $\tilde{x}_1 \leq \tilde{x}_2 \leq \cdots \leq \tilde{x}_n$ un ordenamiento de la muestra. Entonces el j-ésimo cuartil q_j es un número tal que $q_j \in [\tilde{x}_k, \tilde{x}_{k+1}]$, donde

$$\frac{k}{n} \leq j \times 0.25 < \frac{k+1}{n}. \tag{1}$$

Intuitivamente, el primero, segundo y tercer cuartil dejan por debajo el 25 %, 50 % y 75 % de los individuos, respectivamente. Se observa, sin embargo, que hay ambigüedad en la definición de los cuartiles. En efecto, diferentes paquetes de software o diferentes libros, usan definiciones ligeramente distintas. Estas diferencias suelen ser insignificantes para muestras grandes.
El segundo cuartil recibe también el nombre de **mediana** y se lo denota con la letra m.
También es habitual hablar de **deciles** y **percentiles** que separan la muestra en 10 y 100 partes, respectivamente.

Medidas de tendencia central

Estas medidas son números que pretenden informar "por dónde andan los valores". Una medida muy común es la **media muestral** (el *promedio*)

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{2}$$

Otra medida es la **mediana**, que ya fue definida más arriba.
Finalmente, la **moda** corresponde a aquel valor que "aparece más frecuentemente". Algunos detalles a tener en cuenta:

1. Puede haber más de una moda. Se habla de **distribuciones multimodales**.
2. En el caso de números reales (un continuo de valores) se procede de la siguiente manera: i) se divide el rango en intervalos de igual longitud; ii) se cuenta cuántos valores se encuentran en cada uno de los intervalos; iii) la moda corresponde al valor medio del intervalo con mayor número de datos. La moda depende del número de intervalos escogido.
3. Se puede utilizar también para datos cuantitativos (por ej., el apellido más común).

Mediana vs. media

La media (el promedio) es una métrica muy conocida. Sin embargo, la mediana tiene algunas ventajas sobre aquella. En particular, se dice que la mediana es *robusta* ante la presencia de *outliers*.
¿Qué son outliers? Son datos que parecen estar más allá de un rango normal, es decir, son muy grandes o muy pequeños. Estos datos pueden ser correctos o pueden deberse a errores de medición.
Supongamos que se miden alturas (en centímetros) de alumnos en y se tiene una muestra $\{150, 160, 170, 180, 190\}$. La media y la mediana coinciden: $\overline{x} = m = 170$ cm. Ahora supongamos que una de las alturas es medida incorrectamente y se obtiene la muestra $\{150, 160, 170, 180, \mathbf{290}\}$. La mediana sigue siendo $m = 170$ cm, pero la media pasa a ser $\overline{x} = 190$ cm como consecuencia del error.
Ahora consideremos otro ejemplo importante. Se toma una muestra de los ingresos diarios (en dólares) de una población, obteniéndose $\{1, 1, 2, 100\}$. El *ingreso medio* U\$S 20.8 ($= \overline{x}$), pero el 50 % de la población gana menos de U\$S 1.5 ($= m$). ¿Cuál de las dos medidas le parece más relevante en este contexto?

Medidas de dispersión

Estas medidas son números que pretenden informar "qué tan dispersos" están los valores.

1. **Rango**: $R = |\text{máx}_i\, x_i - \text{mín}_i\, x_i|$.
2. **Rango intercuartil**: $\text{IQR} = q_3 - q_1$. Es la longitud de los datos que ocupan el 50 % central.
3. **Varianza muestral**:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \,. \tag{3}$$

   Un problema de la varianza muestral es que tiene las unidades al cuadrado.

4. **Desvío muestral**: $s = \sqrt{s^2}$.
5. **Media del desvío absoluto**:

$$w = \frac{1}{n} \sum_{i=1}^{n} |x_i - \overline{x}| \,. \tag{4}$$

   Parece más natural que la varianza. Sin embargo, desde el punto de vista analítico, resulta más sencillo trabajar con la varianza (entre otras cosas, x^2 es derivable en todos lados, no así $|x|$).

6. **Mediana del desvío absoluto**:

$$\text{MAD} = \text{mediana}\left\{|x_i - \overline{x}|\right\} . \tag{5}$$

   Es una medida robusta, utilizada en muchos contextos.

Parámetros de forma

1. El **coeficiente de simetría** es una medida de la simetría de los datos respecto de la media:

$$\gamma = \frac{1}{n s^3} \sum_{i=1}^{n} (x_i - \overline{x})^3 \qquad \text{sin corrección por muestra pequeña.} \tag{6}$$

2. El **coeficiente de curtosis** se refiere a la concentración en torno a la media (positivo si es alta la concentración):

$$\kappa = \frac{1}{n s^4} \sum_{i=1}^{n} (x_i - \overline{x})^4 - 3 \qquad \text{sin corrección por muestra pequeña.} \tag{7}$$

Frecuencia de ocurrencia

Es habitual dividir el rango de valores en intervalos. De esta forma, se consideran los números $a_1 < a_2 < a_3 < \cdots < a_{N+1}$, donde N es el número de intervalos, $a_1 < \min_i x_i$, y $a_{N+1} \geq \max_i x_i$. Es habitual que los intervalos sean de la misma longitud, esto es, $|a_{k+1} - a_k| =$ constante. A cada intervalo se lo suele denominar **bin**.
La **frecuencia absoluta de ocurrencia** n_k cuenta la cantidad de elementos en el k-ésimo intervalo:

$$n_k = |\{x_i : x_i \in (a_k, a_{k+1}]\}| . \tag{8}$$

Por su parte, la **frecuencia relativa de ocurrencia** f_k viene dada por

$$f_k = \frac{n_k}{n}. \tag{9}$$

Tanto n_k como f_k suelen representarse con gráficos de barras denominados **histogramas**. Las siguientes igualdades se cumplen:

$$\sum_{k=1}^{N} n_k = n, \qquad \sum_{k=1}^{N} f_k = 1. \tag{10}$$

Se define también la **función de frecuencia relativa acumulada** como

$$F(\alpha) = \frac{|\{x_i : x_i \leq \alpha\}|}{n}. \tag{11}$$

Siempre $F(\alpha) \in [0, 1]$.

Datos agrupados

A veces, los datos son resumidos en una tabla de frecuencias (absolutas o relativas) de ocurrencia para un cierto número de intervalos adecuadamente escogidos. Esta forma de presentación se denomina de **datos agrupados**.
Es posible aproximar las medidas presentadas más arriba usando datos agrupados. Por ejemplo:

1. Media, varianza, simetría y curtosis muestrales:

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{N} m_k \, n_k, \qquad s^2 = \frac{1}{n-1} \sum_{k=1}^{N} (m_k - \overline{x})^2 \, n_k, \tag{12}$$

$$\gamma = \frac{1}{n s^3} \sum_{k=1}^{N} (m_k - \overline{x})^3 \; n_k, \qquad \kappa = \left[\frac{1}{n s^4} \sum_{k=1}^{N} (m_k - \overline{x})^4 \; n_k \right] - 3, \tag{13}$$

   donde $m_k = (a_k + a_{k+1})/2$ es la **marca de clase**.

2. Función de frecuencias acumuladas ("una escalera"):

$$F(\alpha) = \begin{cases} 0 & \alpha < a_1, \\ \sum_{k=1}^{r} f_k & \alpha \in [a_r, a_{r+1}), \\ 1 & \alpha \geq a_{N+1}. \end{cases} \tag{14}$$

3. El j-ésimo cuartil: $q_j \in [a_k, a_{k+1}]$ con k tal que $F(a_k) \leq j \times 0.25 \leq F(a_{k+1})$. Habitualmente, se usa una interpolación lineal (regla de tres).

2. Guía de ejercicios

Los enunciados que siguen incluyen preguntas en el estilo de las que encontrarán en el EAC 1. Se recomienda una lectura cuidadosa del capítulo 1 de Devore.

1. La tabla que sigue corresponde a la vida en horas de las lámparas incandescentes de la marca MARSO para una muestra de 10 lámparas elegidas al azar.

1067	919	1196	785	1126	936	918	1156	920	948.

   Determine la media, mediana, rango y desviación estándar de la muestra considerada.

2. En un negocio se registró la demanda semanal de un repuesto durante 90 semanas obteniéndose los datos que se muestran en la siguiente tabla:

9	8	3	18	4	5	6	7	7	6	7	5	4	3	15
3	8	6	11	10	9	8	7	13	3	4	5	5	6	4
3	6	7	9	8	7	4	5	6	7	8	10	11	3	2
1	7	6	17	7	9	8	6	11	0	20	1	4	5	12
2	2	1	4	5	6	7	8	10	9	8	7	7	6	5
2	7	7	10	6	6	14	2	4	5	12	10	9	8	7

   a) Agrupe estos datos en una tabla de frecuencias, y haga el diagrama de barras (es interesante observar que en este ejemplo la variable aleatoria que se muestrea es discreta).

   b) Determine la media, varianza, desviación estándar y moda muestrales. Los tres primeros calcúlelos con los datos originales y también con los datos agrupados de la tabla de frecuencias.

3. Los siguientes datos son mediciones de la resistencia a la ruptura (en onzas) de una muestra de 60 hilos de cáñamo:

32.5	15.2	35.4	21.3	28.4	26.9	34.6	29.3	24.5	31.0
21.2	28.3	27.1	25.0	32.7	29.5	30.2	23.9	23.0	26.4
27.3	33.7	29.4	21.9	29.3	17.3	29.0	36.8	29.2	23.5
20.6	29.5	21.8	37.5	33.5	29.6	26.8	28.7	34.8	18.6
25.4	34.1	27.5	29.6	22.2	22.7	31.3	33.2	37.0	28.3
36.9	24.6	28.9	24.8	28.1	25.4	34.5	23.6	38.4	24.0

   a) Agrupe los datos en una tabla de frecuencias considerando las clases 15.0 - 19.9, 20.0 - 24.9, ..., 35.0 - 39.9. Represente el histograma y el polígono de frecuencias absolutas.

   b) Realice un histograma y el polígono de frecuencias acumuladas.

   c) Determine media, mediana y dispersión o desvío estándar usando los datos agrupados.

   d) Determine los cuartiles y el rango intercuartil usando todos los datos. Realice un diagrama boxplot o de caja

   e) Determine, en base a todas las observaciones muestrales la proporción de la muestra para la cual la resistencia a la ruptura toma valores en un intervalo con centro en la media muestral y de semiamplitud igual al doble de la dispersión muestral. Realice el cálculo nuevamente con los datos agrupados.

4. La siguiente tabla da la distribución del tiempo de duración en segundos de 1000 llamadas telefónicas :

Duración (marca de clase)	30	60	90	120	150	180	210	240	270	300
Frecuencia	6	28	88	180	247	260	133	42	11	5

a) Represente el histograma y polígono de frecuencias.

b) Obtenga la tabla y represente el polígono de frecuencias acumuladas.

c) Calcule la media, mediana, modo y desvío estándar muestral s.

d) Obtenga los porcentajes de llamadas cuya duración pertenece a los intervalos centrados en la media y de semiamplitud 1) s 2) $2\,s$ 3) $3\,s$.

e) Calcule el porcentaje de llamadas cuya duración supera los 3 minutos.

Nota: La marca de clase es el punto medio del intervalo de clase (que en este ejemplo tiene una amplitud de 30 seg). Para responder c) y d) es conveniente realizar un gráfico "linealizado" de las frecuencias acumuladas (en el que la frecuencia acumulada en cada intervalo crece linealmente).

Resolución aquí.

5. Se toma una muestra de 100 recién nacidos y se los pesa. A partir de esas mediciones, se forma la siguiente tabla de datos agrupados:

Marca [kg]	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5
Frecuencia	2	3	14	10	15	18	16	14	4	4

a) Grafique el histograma.

b) Grafique la poligonal de frecuencias acumuladas.

c) Calcule la media y el desvío muestrales. Detalle los cálculos.

d) Calcule la mediana muestral. Detalle los cálculos.

Resolución aquí.

6. Se toma una muestra de 100 recién nacidos y se los mide. A partir de esas mediciones, se forma la siguiente tabla de datos agrupados:

Marca [cm]	32.325	32.975	33.625	34.275	34.925
Frecuencia	1	4	11	20	24
Marca [cm]	35.575	36.225	36.875	37.525	38.175
Frecuencia	18	13	6	1	2

a) Grafique el histograma.

b) Grafique la poligonal de frecuencias acumuladas.

c) Calcule la media y el desvío muestrales. Detalle los cálculos.

d) Calcule la mediana muestral. Detalle los cálculos.

7. La siguiente tabla da la distribución del tiempo de duración en segundos de 400 llamadas telefónicas:

Intervalos de clase	Frecuencia absoluta
$(0, 10]$	20
$(10, 20]$	60
$(20, 30]$	120
$(30, 40]$	100
$(40, 50]$	60
$(50, 60]$	40

a) Realizar el histograma correspondiente indicando claramente en el gráfico las marcas de clase.

b) Graficar junto al histograma el polígono de frecuencias.

c) Armar una tabla con los valores de las frecuencias relativas acumuladas.

d) Hallar el valor del primer cuartil y de la media.

3. Ejercicios resueltos

Ejercicio 4

La siguiente tabla da la distribución del tiempo de duración en segundos de 1000 llamadas telefónicas:

Duración (marca de clase)	30	60	90	120	150	180	210	240	270	300
Frecuencia	6	28	88	180	247	260	133	42	11	5

1. Representar el histograma y polígono de frecuencias.
2. Obtener la tabla representar el polígono de frecuencias acumuladas.
3. Calcular la media , mediana, modo y desvío standard muestral s.
4. los porcentajes de llamadas cuya duración pertenece a los intervalos centrados en la media y de semiamplitud 1) s 2) 2 s 3) 3 s.
5. Calcular el porcentaje de llamadas cuya duración supera los 3 minutos (Rta: 32 %)

Nota: La marca de clase es el punto medio del intervalo de clase (que en este ejemplo tiene una amplitud de 30 seg). Para responder c) y d) es conveniente realizar un gráfico "linealizado" de las frecuencias acumuladas (en el que la frecuencia acumulada en cada intervalo crece linealmente).

Resolución

La información disponible corresponde a la tabla de frecuencias construida a partir de la duración en segundos de 1000 llamadas telefónicas. En este caso el rango de la duración de las llamadas fue subdividido en 10 intervalos de igual longitud (30 segundos). El primer intervalo corresponde a llamadas cuya duración es de 15 a 45 (y son 6 llamadas), el segundo intervalo es (45, 75] y se registraron 28 llamadas y así sucesivamente.

El siguiente código en *Octave/Matlab* genera el histograma (Fig. 1) y el polígono de frecuencias (Fig. 2):

```
d=[30 60 90 120 150 180 210 240 270 300]; % marcas de clase
f=[6 28 88 180 247 260 133 42 11 5]; % frecuencias absolutas
bar(d,f,1) % genera el histograma con barras contiguas
% el vector dd tiene los extremos de los intervalos de clase
dd =[ 0 30 60 90 120 150 180 210 240 270 300 330];
% el vector ff tiene las ordenadas para el poligono de frecuencias
ff=[0 6 28 88 180 247 260 133 42 11 5 0];
e =[ 15 15 45 45 75 75 105 105 135 135 165 165];
e=[e 195 195 225 225 255 255 285 285 315 315];
fff =[ 0 6 6 28 28 88 88 180 180 247 247 260 260 133 133 42 42 11 11 5
    5 0];
% los vectores e y fff tienen las coordenadas de la poligonal
% que limita el histograma
% en el grafico que sigue se superponen el histograma y
% el poligono de frecuencias
plot(dd,ff,'-o',e,fff,'-')
```

La tabla con la información necesaria para construir el polígono de frecuencias acumuladas es la siguiente:

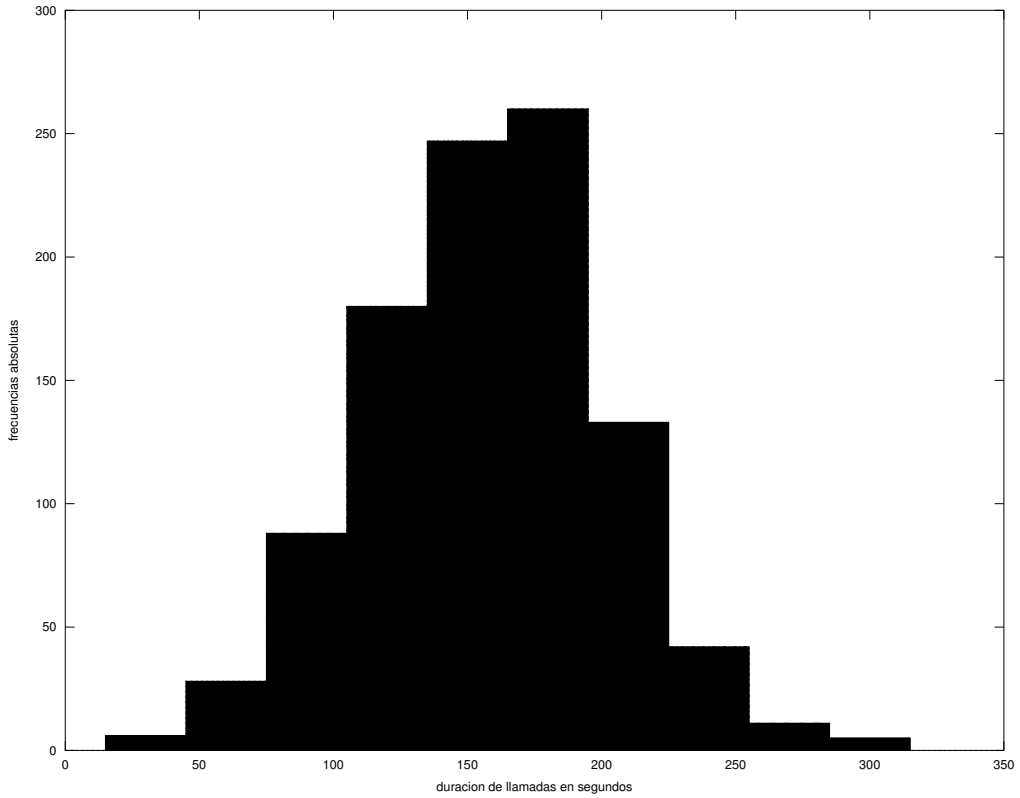

Figura 1: Histograma de frecuencias

Duración (marca de clase)	15	45	75	105	135	165	195	225	255	285	315
Frecuencia	0	6	34	122	302	549	809	942	984	995	1000

El siguiente código en *Octave/Matlab* genera el histograma de frecuencias acumuladas y el polígono de estas frecuencias:

```
x=[15 15 45 45 75 75 105 105  135 135 165 165  195 195 225 225 255 255 285
    285 315];
F=[0 6 6 34 34 122 122 302 302 549 549 809 809 942 942 984 984 995 995 1000
    1000];
xx=[15 45 75 105 135 165 195 225 255 285 315 ];
FF=[0 6 34 122 302 549 809 942 984 995 1000];
plot(x,F,'-',xx,FF,'-o')
ylabel('frecuencias acumuladas absolutas')
xlabel('duracion de llamadas en segundos')
title('Histograma y poligono de frecuencias acumuladas')
```

Con los datos de esta tabla:

$$\overline{x} = 157.71 \text{ seg}, s \approx 45.37 \text{ seg}. \tag{15}$$

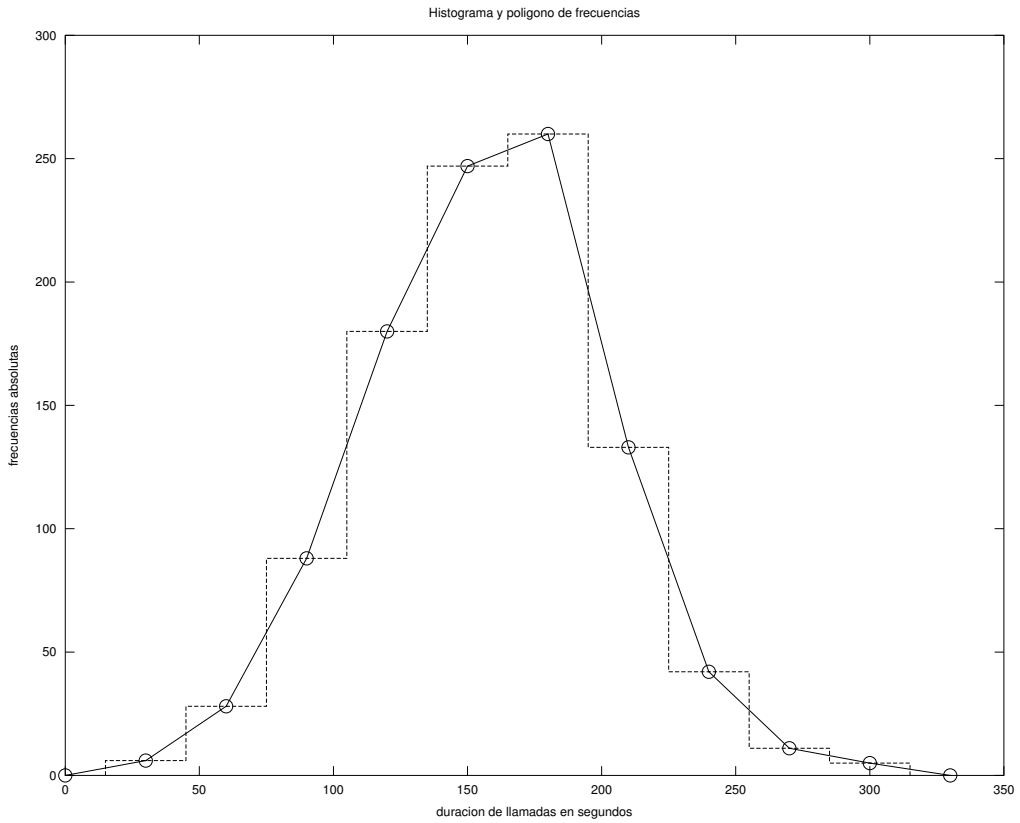

Figura 2: Histograma y polígono de frecuencias

En este caso la distribución de frecuencias tiene un único modo. Una opción es asignarle al modo el valor del punto medio del intervalo al que corresponde la máxima frecuencia (el sexto intervalo) o sea la abscisa del modo del polígono de frecuencias. Así resulta que $M = 180$ seg, el punto medio de ese intervalo. Otra opción utilizada para determinar el modo M tiene en cuenta las frecuencias de los dos intervalos contiguos (por izquierda f_I y por derecha f_D) y realizar una asignación proporcional de la forma $(M - L_I)(f_M - f_D) = (L_D - M)(f_M - f_I)$, donde L_I y L_D son los extremos izquierdo y derecho del intervalo al que pertenece M. En este caso (y en segundos) son $L_I = 165$ y $L_D = 195$ en tanto que $f_I = 247$, $f_D = 133$ y $f_M = 260$. Reemplazando, resulta, $M \approx 167.79$ seg.

Para determinar un valor de la mediana se puede realizar una interpolación lineal sobre el polígono de frecuencias acumuladas para obtener la abscisa a la que corresponde la frecuencia acumulada $n/2$ (en este caso 500). Esa frecuencia acumulada pertenece al intervalo (135, 165) y la frecuencia acumulada a cada extremos es 302 y 549, respectivamente. Entonces la mediana m_e resulta $m_e = 135 + 30\,\frac{500-302}{549-302} \approx 159.05$ seg.

Por interpolación lineal en la tabla de frecuencias acumuladas se puede responder a preguntas relacionadas con la proporción de llamadas en determinado intervalo de valores de la duración de ellas. Los intervalos centrados en la media y de semiamplitud s, 2 s y 3 s son los intervalos (112.34 203.08), (66.97 248.45) y (21.60 293.82) respectivamente. Si $\mathrm{P}(x)$ es la función interpoladora li-

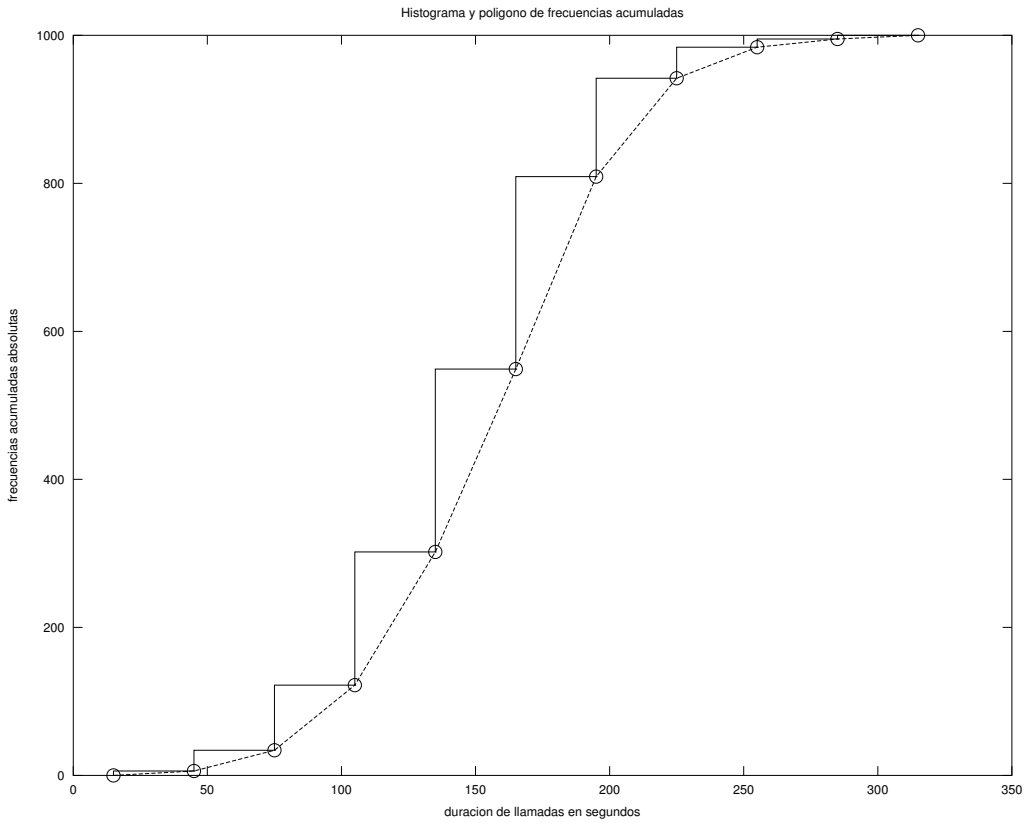

Figura 3: Histograma y polígono de frecuencias acumuladas

neal por tramos de las frecuencias acumuladas entonces la proporción de llamadas cuya duración pertenece al intervalo (a, b) viene dada por $(\mathrm{P}\,(b) - \mathrm{P}\,(a))/n$. El gráfico de P es el polígono de frecuencias acumuladas, la interpolación lineal supone que la densidad de datos es constante que es la suposición acorde a disponer de datos agrupados.

La frecuencia acumulada interpolada para 112.34 se obtiene considerando los pares (105, 122) y (135, 302) y generando $\mathrm{P}\,(112.34) = 122 + (302 - 122)(112.34 - 105)/30 \approx 166.04$. Por otro lado la frecuencia acumulada interpolada para 203.08 se obtiene considerando los pares (195, 849) y (225, 942) y generando $\mathrm{P}\,(203.8) = 809 + (942 - 809)(203.08 - 195)/30 \approx 844.82$. Asi, entonces, la proporción de llamadas cuya duración pertenece al intervalo (112.34 203.08) es $(844.82 - 166.04)/1000 \approx 0.68$ o sea 68 %. En igual forma se pueden obtener los porcentajes de datos en cada uno de los otros dos intervalos.

El porcentaje de llamadas cuya duración supera los 3 minutos corresponde a evaluar $(1 - \mathrm{P}\,(180))/1000$. Es sencillo verificar que $\mathrm{P}\,(180) = (809 + 549)/2 = 679$ de donde el porcentaje de llamadas cuya duración supera los 3 minutos es 32.1 %.

Ejercicio 5

Se toma una muestra de 100 recién nacidos y se los pesa. A partir de esas mediciones, se forma la siguiente tabla de datos agrupados:

Marca [kg]	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5
Frecuencia	2	3	14	10	15	18	16	14	4	4

1. Grafique el histograma.
2. Grafique la poligonal de frecuencias acumuladas.
3. Calcule la media y el desvío muestrales. Detalle los cálculos.
4. Calcule la mediana muestral. Detalle los cálculos.

Resolución

Las marcas $\overline{x}_i$ y las frecuencias f_i son las que están en la tabla. $n = \sum_i f_i$ es la cantidad de datos.

$$\overline{x} = \frac{1}{n}\sum \overline{x}_i f_i = 3.6460, \; s = \sqrt{\frac{1}{n-1}\sum (\overline{x}_i - \overline{x})^2 f_i} = 0.4234.$$

Los datos no se acumulan "hasta" las marcas, sino "hasta" los extremos derechos de los intervalos de clase (*bines*). Por lo tanto, la mediana (que divide a la mitad los datos) tiene que estar en el intervalo $(3.6, 3.8]$. Usando interpolación lineal:

$$\text{mediana} = \frac{50 - 44}{62 - 44} \cdot (3.8 - 3.6) + 3.6 = 3.6667.$$