MAT-042: Tablas de frecuencia

Felipe Osorio

http://fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

Ejemplo con datos discretos

Ejemplo: Número de cargas familiares

Se consultó el registro de empleados de una fábrica, extrayendo el número de cargas familiares. Se obtuvo los siguientes datos:

1	2	4	2	2	2	3	2	1	1	0	2	2
0	2	2	1	2	2	3	1	2	2	1	2	

La tabla de frecuencias asociada a esta muestra[1] es

Cargas familiares	Número de empleados	Porcentaje de empleados	Num. acumulado de empleados	Porcentaje acumulado
0	2	8%	2	8%
1	6	23%	8	32%
2	14	56%	22	88%
3	2	8%	24	96%
4	1	4%	25	100%
Total	25	100%	–	–

[1] Las últimas 2 columnas carecen de sentido para variables nominales

Tabla de distribución de frecuencias: Datos discretos

Datos discretos: Considere k categorías de una variable x. Entonces,

Variable	Frecuencia Absoluta	Frecuencia Relativa	Frec. Abs. Acumulada	Frec. Rel. Acumulada
x_1	n_1	f_1	N_1	F_1
x_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	f_k	N_k	F_k
Total	n	1	–	–

donde

$$f_i = \frac{n_i}{n}, \qquad N_i = \sum_{j=1}^{i} n_j, \qquad F_i = \sum_{j=1}^{i} f_j,$$

para $i = 1, \ldots, k$. Evidentemente tenemos que

$$\sum_{i=1}^{k} n_i = n, \qquad \sum_{i=1}^{k} f_i = 1, \qquad F_k = 1.$$

Tabla de distribución de frecuencias: Datos discretos

En este contexto, tenemos que la media y varianza muestrales, están dadas por:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} f_i x_i$$

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

Para el ejemplo anterior es fácil notar que

$$\overline{x} = \frac{2\cdot 0 + 6\cdot 1 + 14\cdot 2 + 2\cdot 3 + 1\cdot 4}{25} = \frac{44}{25} = 1.760$$

$$s^2 = \frac{2(0-1.76)^2 + 6(1-1.76)^2 + 14(2-1.76)^2 + 2(3-1.76)^2 + 1(4-1.76)^2}{25-1}$$

$$= \frac{18.56}{24} = 0.773$$

Tabla de distribución de frecuencias: Datos continuos

Datos continuos: Los datos deben ser agrupados en k categorías,

Marca de clase	Frecuencia Absoluta	Frecuencia Relativa	Frec. Abs. Acumulada	Frec. Rel. Acumulada
C_1	n_1	f_1	N_1	F_1
C_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_k	n_k	f_k	N_k	F_k
Total	n	1	–	–

donde

$$C_i = \frac{LI_i + LS_i}{2}, \qquad i = 1, \ldots, k,$$

se denomina marca de clase.

Ejemplo con datos continuos

Ejemplo: Precios de cierre de una acción

Considere los precios de cierre de acciones de una determinada empresa nacional, obteniendo:

179	173	181	170	158	174	172	166	194	185
162	187	198	177	178	165	154	188	166	171
175	182	167	169	172	186	172	176	168	187

La tabla de frecuencias es dada por

Precio de cierre	Marca de clase	Número de días	Porcentaje de días	Num. de días acumulado	Porcentaje acumulado
(150,160]	155	2	7%	2	7%
(160,170]	165	8	27%	10	34%
(170,180]	175	11	36%	21	70%
(180,190]	185	7	23%	28	93%
(190,200]	195	2	7%	30	100%
Total	–	30	100%	–	–

"Algoritmo" para construir una tabla de frecuencias

- Determinar el número de categorías, k usando por ejemplo:

$$k = \sqrt{n},$$
$$k = 1 + 3.3 \log_{10}(n), \qquad \text{(regla de Sturges)}$$
$$k = \lceil 2n^{1/3} \rceil, \qquad \text{(regla de Rice)}$$

- Calcular el rango,

$$R = \max\{x_i\}_{i=1}^{n} - \min\{x_i\}_{i=1}^{n},$$

- Determinar la longitud de los intervalos a_i para $i = 1, \ldots, k$. Usualmente,

$$a_i = a, \qquad a = \frac{R}{k},$$

- General los límites de clases

$$LI_1 = \min\{x_i\}_{i=1}^{n} - \triangle, \qquad LS_1 = LI_1 + a,$$
$$LI_2 = LS_2, \qquad LS_2 = LI_2 + a,$$
$$\vdots$$

Tabla de distribución de frecuencias: Datos continuos

De este modo, para calcular la media y varianza muestrales, usamos:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i C_i = \sum_{i=1}^{k} f_i C_i$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i (C_i - \overline{x})^2.$$

En el ejemplo de los precios de cierre, tenemos

$$\overline{x} = \frac{2 \cdot 155 + 8 \cdot 165 + 11 \cdot 175 + 7 \cdot 185 + 2 \cdot 195}{30} = \frac{5240}{30} = 174.667$$

mientras que

$$\sum_{i=1}^{5} n_i (C_i - \overline{x})^2 = 2(155 - 174.667)^2 + 8(165 - 174.667)^2 + 11(175 - 174.667)^2$$
$$+ 7(185 - 174.667)^2 + 2(195 - 174.667)^2 = 3096.667,$$

de este modo $s^2 = 3096.667/(30-1) = 106.782$.

Tabla de distribución de frecuencias (usando los datos crudos)

Note que, usando los datos crudos se obtiene:

```
# precios de cierre (datos a granel)
> x <- c(179, 173, 181, 170, 158, 174, 172, 166, 194, 185,
+   162, 187, 198, 177, 178, 165, 154, 188, 166, 171, 175,
+   182, 167, 169, 172, 186, 172, 176, 168, 187)

# cálculo de media y varianza muestrales
> mean(x)
[1] 175.0667
> var(x)
[1] 105.7195
```

Tabla de distribución de frecuencias

Para calcular la mediana en tablas de distribución de frecuencias, debemos hacer:

$$\text{me} = L_i + \frac{n/2 - N_{i-1}}{n_i} \cdot a_i = L_i + \frac{1/2 - F_{i-1}}{f_i} \cdot a_i.$$

En general, el k-ésimo percentil P_k es

$$P_k = L_i + \frac{n(k/100) - N_{i-1}}{n_i} \cdot a_i = L_i + \frac{k/100 - F_{i-1}}{f_i} \cdot a_i.$$

Mientras que la moda es dada por

$$\text{mo} = L_{\text{m}} + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot a_{\text{m}},$$

con L_{m} el límite inferior de la clase modal, $\Delta_1 = n_{\text{m}} - n_{\text{m}-1}$, $\Delta_2 = n_{\text{m}} - n_{\text{m}+1}$ donde n_{m} la frecuencia absoluta de la clase modal, $n_{\text{m}-1}$ y $n_{\text{m}+1}$ corresponden a las frecuencias absolutas de la clase anterior y posterior a la modal, respectivamente, y a_{m} es la amplitud de la clase modal.

Ejemplo con datos continuos

Ejemplo: Salarios de trabajadores

Los trabajadores de una empresa, cuya tarea es clasificar y envasar fruta, obtuvieron los siguientes salarios semanales (clasificados según sexo).

Ingreso (UM)	Mujeres	Hombres
65 – 75	10	0
75 – 85	15	0
85 – 95	60	5
95 – 105	15	10
105 – 115	10	50
115 – 125	0	25
125 – 135	0	10
Total	110	100

Consideraremos solamente el grupo de mujeres y dejaremos el análisis del grupo de hombres y el total de trabajadores como ejercicio.

Ejemplo con datos continuos

La tabla de frecuencias para el grupo de mujeres adopta la forma:

Ingreso (UM)	C_i	n_i	f_i	N_i	F_i
65 – 75	70	10	0.090	10	0.090
75 – 85	80	15	0.136	25	0.226
85 – 95	90	60	0.548	85	0.774
95 – 105	100	15	0.136	100	0.910
105 – 115	110	10	0.090	110	1.000
Total	–	110	1.000	–	–

En este caso, tenemos $n = 110$ y

$$\sum_{i=1}^{5} n_i C_i = 10 \cdot 70 + 15 \cdot 80 + 60 \cdot 90 + 15 \cdot 100 + 10 \cdot 110 = 9\,900.$$

De este modo, la media aritmética para el grupo de mujeres es:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{5} n_i C_i = \frac{9900}{110} = 90 \text{ (UM)}.$$

Ejemplo con datos continuos

Para calcular la mediana, debemos ubicar el intervalo mediano. Es decir, la primera frecuencia relativa acumulada (F_i) que supere 0.5. Así, el intervalo mediano es $(85, 95]$. Además, $a_i = 10$ para todos los intervalos. Luego,

$$\text{me} = L_i + \frac{1/2 - F_{i-1}}{f_i}\, a_i,$$

donde $L_i = 85$, $F_{i-1} = 0.226$, $f_i = 0.548$ y $a_i = 10$. De este modo,

$$\text{me} = 85 + \frac{0.500 - 0.226}{0.548} \cdot 10 = 85 + 0.5 \cdot 10 = 90 \text{ (UM)}.$$

Ejemplo con datos continuos

Note que,

$$s^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{5} n_i C_i^2 - n\overline{x}^2\Big).$$

En nuestro caso,

$$\sum_{i=1}^{5} n_i C_i^2 = 902\,000, \qquad \overline{x}^2 = 8\,100.$$

Así,

$$s^2 = \frac{1}{110-1}(902\,000 - 110 \cdot 8\,100) = \frac{1}{109}(902\,000 - 891\,000)$$
$$= \frac{11\,000}{109} = 100.9174 \ (\text{UM})^2.$$

Además, tenemos que $s = \sqrt{11\,000/109} = 10.0458$ (UM). Lo que lleva a

$$\text{CV} = \frac{10.0458}{90} = 0.1116.$$

Ejemplo con datos continuos

Podemos evaluar la simetría usando el coeficiente de Galton. Ahora, Q_1 y Q_3, son dados por:

$$Q_1 = 85 + \frac{0.250 - 0.226}{0.548} \cdot 10 = 85.438$$
$$Q_3 = 85 + \frac{0.750 - 0.226}{0.548} \cdot 10 = 94.562,$$

de este modo $IQR = 9.1240$, y

$$b_{\mathsf{G}} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$$
$$= \frac{(94.562 - 90) - (90 - 85.438)}{9.124} = \frac{4.562 - 4.562}{9.124} = 0.000$$

Es decir, la distribución de los datos es simétrica.

Ejemplo: Datos de falla de resortes

En un experimento industrial se desea determinar la confiabilidad de cierto tipo de resortes cuando son sometidos a repetidos ciclos de esfuerzo hasta que estos fallen.

Los tiempos de falla, en unidades de 10^3 ciclos de esfuerzo para 60 resortes fueron divididos en grupos de 10 considerando 6 diferentes niveles de presión.

Note que conforme la presión decrece existe un rápido aumento de número promedio de ciclos hasta la falla.

Además, existe un patrón lineal entre $\log \overline{x}$ y $\log s^2$, sugiriendo que la varianza es proporcional al promedio al cuadrado.

Datos de falla de resortes

Tiempos de falla (en unidades de 10^3 ciclos) de resortes sometidos a ciclos de presión bajo un esfuerzo dado.

	Stress (N/mm^2)					
	950	900	850	800	750	700
	225	216	324	627	3402	12510
	171	162	321	1051	9417	12505
	198	153	432	1434	1802	3027
	189	216	252	2020	4326	12505
	189	225	279	525	11520	6253
	135	216	414	402	7152	8011
	162	306	396	463	2969	7795
	135	225	379	431	3012	11604
	117	243	351	365	1550	11604
	162	189	333	715	11211	12470
$\overline{x}$	168.3	215.1	348.1	803.3	5636.1	9828.4
s	33.1	42.9	57.9	544.0	3864.3	3354.7

Boxplot

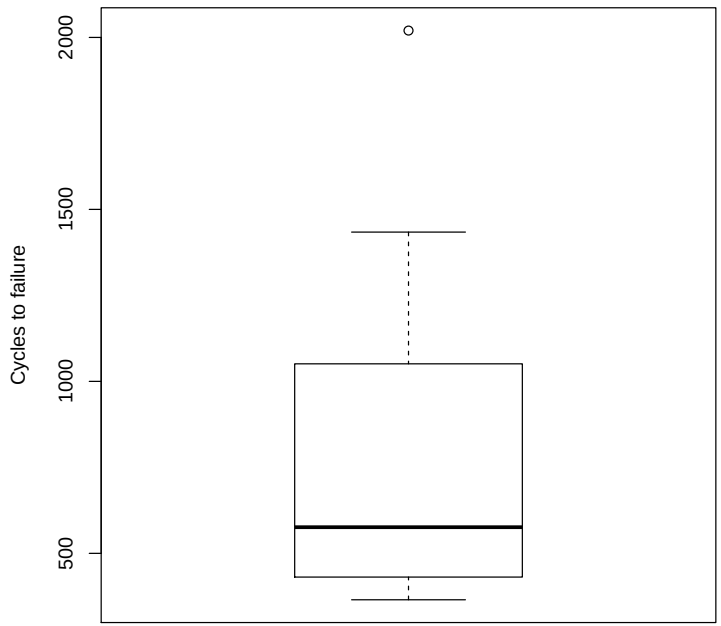
Boxplot: Representación gráfica de 5 estadísticas de resumen.

Ejemplo: Considere los datos de falla de resortes a una presión de 800 N/mm^2, i.e.,

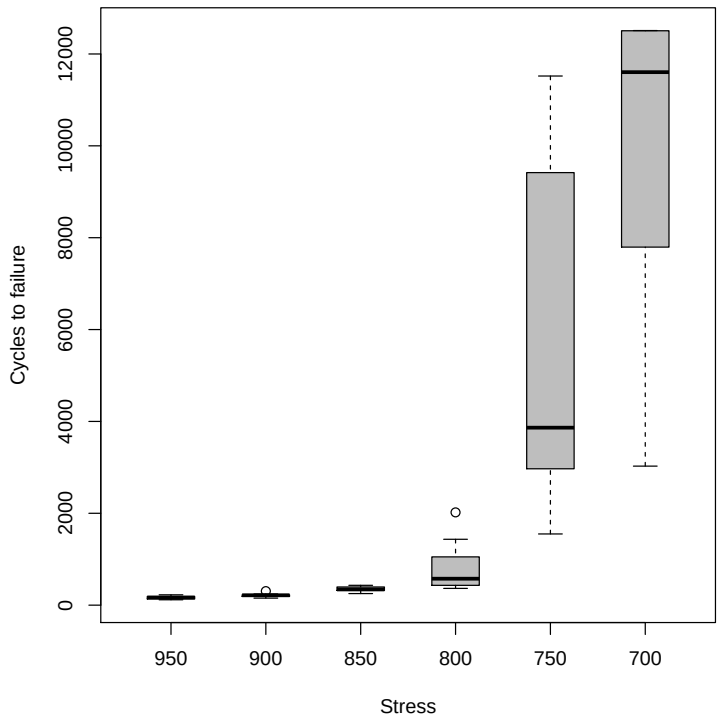
627	1051	1434	2020	525
402	463	431	365	715

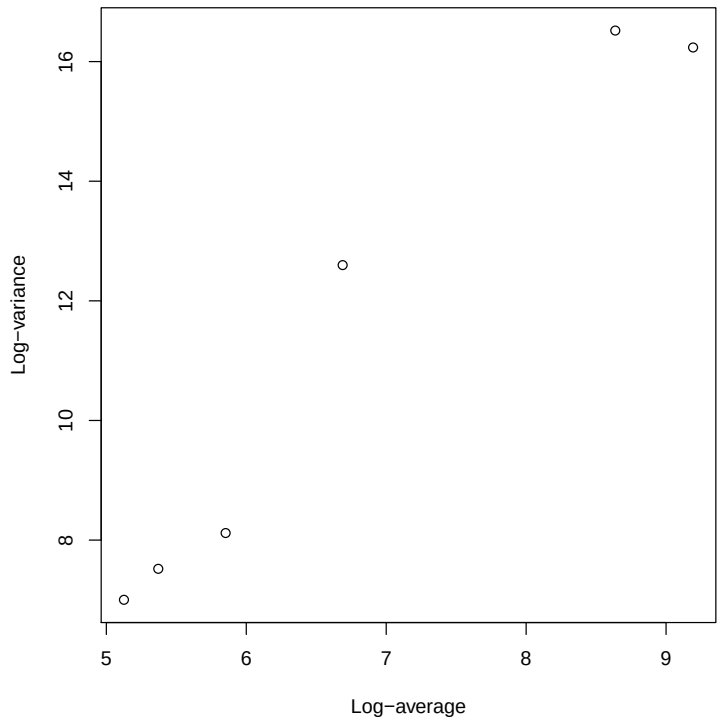
Usando la función `summary` de **R**:

$x_{(1)}$	Q_1	me	$\overline{x}$	Q_3	$x_{(10)}$
305	439	576	803.3	967	2020



Boxplot: datos de resortes





QQ-plot: datos de resortes