CIENCIA
DE DATOS

Estadísticas Descriptivas: datos agrupados

ESTADÍSTICA PARA CIENCIA DE DATOS

PROF. ESTEBAN BALLESTERO

Agrupamiento de datos

- Más que ver el agrupamiento de datos como una técnica estadística, esta puede ser más común de lo que pensamos

- Por ejemplo, tendemos a asociar la como excelente una nota entre 90 y 100, decimos que una persona tiene 18 años cuando en realidad su edad está entre 18 y menos de 19 años, etc.
- ¿Qué razones justificaría el agrupamiento de datos cuantitativos en estadística?

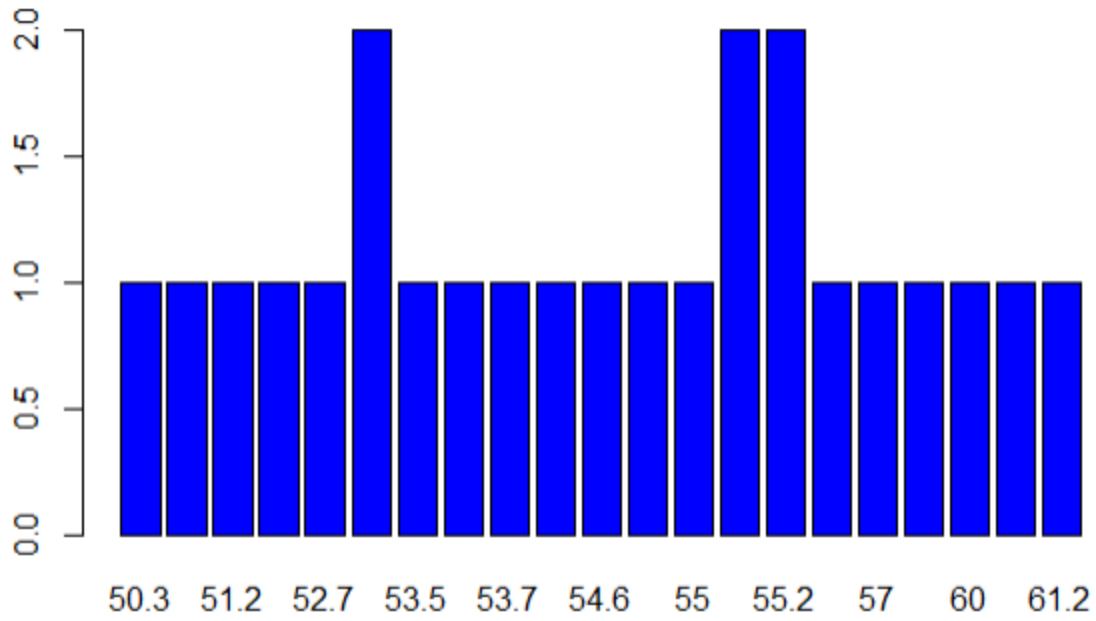
Agrupamiento de datos

Uno de estos motivos puede ser perfectamente que los datos sean muy heterogéneos, en este caso, nos encontraríamos con que las frecuencias de los valores individuales serían todas muy similares abarcando un amplio rango, lo que daría lugar a un diagrama de barras muy difícil de interpretar.

Suponga los siguientes datos para pesos:

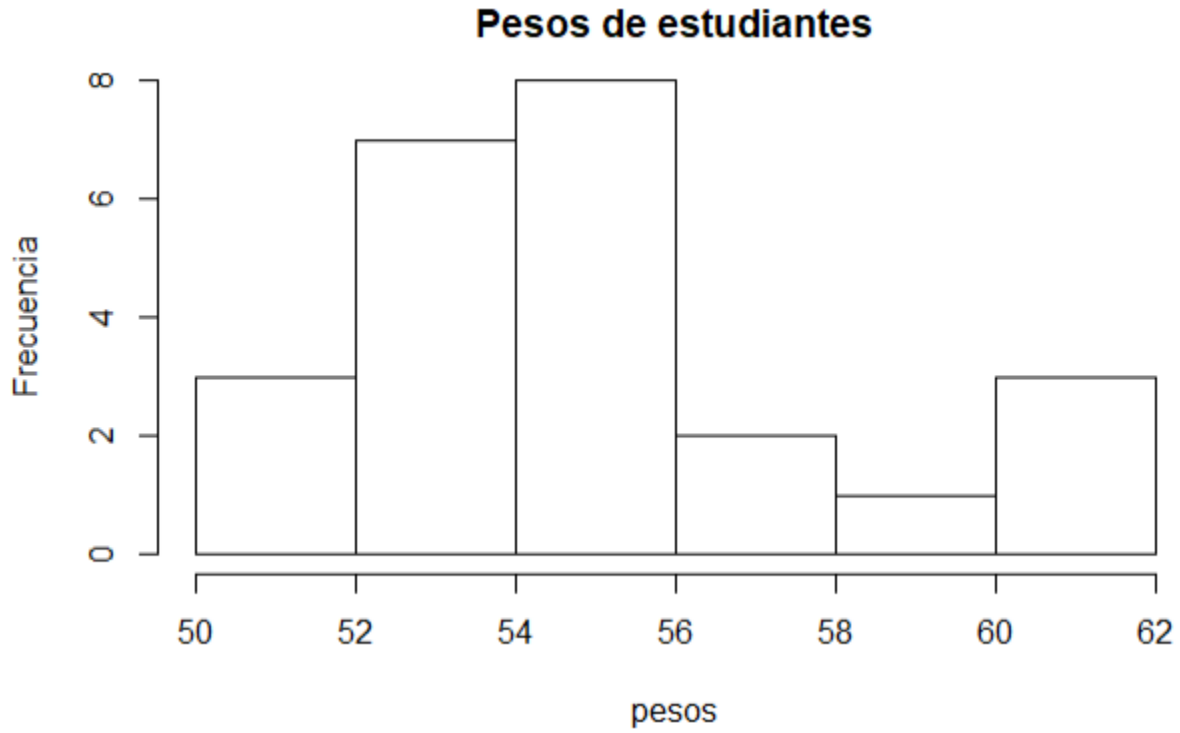
55.2,54.0,55.2,53.7,60.2,53.2,54.6,55.1,51.2,53.2,54.8,
52.3,56.9,57.0,55.0,53.5,50.9,55.1,53.6,61.2,59.5,50.3,
52.7,60.0

Agrupamiento de datos



Agrupamiento de datos

Si se generaran intervalos apropiados, donde la frecuencia correspondería a la cantidad de datos que quedan en el Intervalo, entonces el análisis de los datos cobra sentido



¿Cómo realizar el agrupamiento?

Antes de estudiar unos datos agrupados, hay que, obviamente, agruparlos. Este proceso consta de 4 pasos:

- Decidir el número de intervalos que vamos a utilizar
- Decidir la amplitud de estos intervalos
- Definir los extremos de los intervalos
- Calcular el valor representativo de cada intervalo, lo que se conoce como marca de clase.

Nota: tome en cuenta que no hay una forma mejor que otra para agrupar datos, pero si hay algunos criterios que ayudaría a clasificar como no adecuado cierto tipo de agrupamiento. Cada agrupamiento puede mostrar cosas distintas de los datos.

Paso1: número de clases k

Regla	Requerimientos	Fórmula para k
Raíz cudrada	n	$\sqrt{n}$
Sturges	n	$1 + \log_2(n)$
Scott	Amplitud teórica: $A_s = 3.5 \cdot s \cdot n^{\frac{-1}{3}}$	$\frac{\max(x) - \min(x)}{A_s}$
Freedman-Diaconis	Amplitud teórica: $A_{FD} = 2 \cdot (Q_3 - Q_1) \cdot n^{\frac{-1}{3}}$	$\frac{\max(x) - \min(x)}{A_{FD}}$

Paso 2: Amplitud de clase

- La amplitud de clase podría generar clases de igual tamaño o no, esto según como se desee presentar la información
- La forma más utilizada es usando clases de igual amplitud, y la forma de calcular dicha amplitud sería con la fórmula:

$$\boldsymbol{A = \frac{Rango}{k}}$$

Donde:

$Rango = V_{max} - V_{min}$

$k = número\ de\ clases$

Paso 3: Extremos de los intervalos

Cada intervalo tiene un extremo izquierdo y uno derecho, que puedo no incluirse dentro del mismo. Para este curso se tomarán estos intervalos siempre cerrados por su izquierda y abiertos por la derecha, debido a que esta es la forma en que R los construye y porque es así como se utilizan en Teoría de Probabilidades al definir la distribución de una variable aleatoria discreta.

Notación para los intervalos: $[L_1, L_2), [L_2, L_3), \dots, [L_k, L_{k+1})$

Para la precisión p tome en cuenta que si las medias están dadas en unidades $p = 1$, si están dadas en décimas $p = 0,1$, etc

Paso 3: Extremos de los intervalos

$$L_1 = min(x) - \frac{1}{2} \cdot precisión$$

A partir L_1, el resto de los intervalos se obtienen de forma sucesiva:

$$L_2 = L_1 + A$$

$$L_3 = L_2 + A$$

.

.

.

$$L_{k+1} = L_k + A$$

En resumen: $L_i = L_1 + (i-1)A, con\ i = 2,3, \ldots . k+1$

Paso 4: Marca de clase

El cálculo de medidas de tendencia central para datos agrupados, requiere del uso de las marcas de clases. La marca da clase x_i del intervalo $[L_i, L_{i+1})$ se calcula de la siguiente manera:

$$x_i = \frac{L_i + L_{i+1}}{2}$$

Es decir, es el punto medio de la clase.

Ejemplo 1: pesos de estudiantes

Datos:

55.2,54.0,55.2,53.7,60.2,53.2,54.6,55.1,51.2,53.2,54.8,52.3,56.9,57.0,

55.0,53.5,50.9,55.1,53.6,61.2,59.5,50.3, 52.7,60.0

$n = 24$

$Vmax = 61{,}2$

$Vmin = 50{,}3$

$s = 2.93$

$Q_1 = 53{,}2\ y\ Q_3 = 55{,}62$

Ejemplo 1: pesos de estudiantes

	Sturges	Scott	Freedman-Diaconis
k	$5{,}58 \approx 6$	$3{,}062 \approx 4$	$6{,}48 \approx 7$

A manera de ejemplo, se trabajará con la regla de Scott, así:

$A = \frac{61{,}2-50{,}3}{4} = 2{,}725 \approx 2{,}8$ (se redondea por exceso)

Calculemos el primer extremo:

$$L_1 = 50{,}3 - \frac{1}{2} \cdot 0{,}1 = 50{,}25$$

Ejemplo 1: pesos de estudiantes

Clase	Marca de clase	Frec. Abs	Frec.Abs.Acum	Frec. Rel	Frec.Re.Ac
[50.25, 53.05)	51.65	5	5	0,208	0,208
[53.05, 55.85)	54.45	13	18	0,542	0,75
[55.85, 58.65)	57.25	2	20	0,083	0,833
[58.65, 61.45)	60.05	4	24	0,167	1

Datos ordenados:

50.3 50.9 51.2 52.3 52.7 53.2 53.2 53.5 53.6 53.7 54.0 54.6 54.8 55.0 55.1 55.1 55.2 55.2 56.9 57.0 59.5 60.0 60.2 61.2

Medidas de tendencia central para datos agrupados

El cálculo de medidas de tendencia central es más impreciso, que si se tuviesen los datos crudos, no obstante con las siguientes fórmulas podríamos calcular valores para estas medidas de TC muy similares a los valores reales, lo que facilitaría el análisis práctico de los datos una vez que estos están agrupados

Media para datos agrupados: OPC 1

Para una determinada clase, se toma el valor representativo de esa clase, al cual llamamos marca de clase.

Se puede utilizar la fórmula:

$$\bar{x}_{da} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

donde:

f_i : frecuencia de la i-esima clase.
x_i : marca de la i-esima clase.

Media para datos agrupados: OPC 2

O bien, la fórmula alternativa:

$$\bar{x}_{da} = A + \frac{\sum_{i=1}^{n} f_i d_i}{\sum_{i=1}^{n} f_i}$$

donde:

A : *supuesta media de las marcas de clase. (frecuencia mayor)*

f_i : *frecuencia de la i-ésima clase.*

d_i: $x_i - A$ *(desviación entre la marca de la i-ésima clase y la supuesta media)*

Mediana para datos agrupados:

$$\tilde{x}_{da} = L_{\inf} + \left(\frac{\frac{N}{2} - \left(\sum_{i=1}^{n} f \right)_{\inf}}{f_{mediana}} \right) c$$

donde:

L_{inf} :	Límite real inferior de la clase que contiene a la mediana.
N:	Número total de datos.
c :	Tamaño del intervalo que tiene a la mediana.
$F_{mediana}$:	Frecuencia de la clase de la mediana.
$(\Sigma f)\ \inf$:	Suma de las frecuencias menores a la clase de la mediana.

Moda para datos agrupados:

$$\hat{x}_{da} = L_{\inf} + \frac{d_1}{d_1 + d_2} c$$

donde:

L_{inf} : límite real inferior de la clase que contiene a la moda.

$d1$: diferencia entre la frecuencia de la clase modal y la frecuencia de la clase anterior.

$d2$: diferencia entre la frecuencia de la clase modal y la frecuencia de la clase posterior.

c: tamaño del intervalo que tiene a la moda.

Ejercicio:

- Calcular las medidas de tendencia central para el ejemplo estudiado
- Realizar una tabla de datos agrupados para los datos crabs del paquete MASS:

```
crabdata <- read.csv("http://www.hofroe.net/stat557/data/crab.txt", header=T, sep="\t")
str(crabdata)
cw <- crabdata$Width
cw
```

Sobre los datos:
https://www.rdocumentation.org/packages/MASS/versions/7.3-51.4/topics/crabs