ANALISIS DE DATOS

Dr. Edgar Acuna

http://academic.uprm.edu/eacuna

UNIVERSIDAD DE PUERTO RICO

RECINTO UNIVERSITARIO DE MAYAGUEZ

2. ESTADÍSTICA DESCRIPTIVA

En este capítulo se verán las técnicas que se usan para la organización y presentación de datos en tablas y gráficas, así como el cálculo de medidas estadísticas. Se considerarán solamente datos univariados y bivariados.

EJEMPLO

Row	edad	sexo	escuela	programa	creditos	gpa	familia	hestud	htv
1	21	f	públ	biol	119	3.60	3	35	10
2	18	f	priv	mbio	15	3.60	3	30	10
3	19	f	priv	biot	73	3.61	5	5	7
4	20	f	priv	mbio	*	2.38	3	14	3
5	21	m	públ	pmed	114	3.15	2	25	25
6	20	m	públ	mbio	93	3.17	3	17	6
7	22	m	públ	pmed	120	2.15	5	20	10
8	20	m	priv	pmed	*	3.86	5	15	5
9	20	m	priv	pmed	94	3.19	4	10	2
10	20	f	públ	pmed	130	3.66	6	20	33
11	21	f	priv	mbio	97	3.35	1	15	20
12	20	m	priv	mbio	64	3.17	4	30	2
13	20	f	públ	mbio	*	3.23	2	5	3
14	21	f	publ	mbio	98	3.36	4	15	10
15	21	f	priv	biol	113	2.88	5	15	3
16	21	f	priv	pmed	124	2.80	5	20	10
17	20	f	públ	eagr	*	2.50	4	10	5
18	20	f	priv	mbio	*	3.46	4	18	5
19	22	f	priv	pmed	120	2.74	2	10	15
20	20	f	priv	mbio	95	3.07	3	15	12
21	22	f	priv	biol	125	2.20	3	20	10
22	23	m	públ	eagr	13	2.39	3	10	8
23	21	m	priv	pmed	118	3.05	4	10	10
24	20	f	públ	mbio	118	3.55	5	38	10
25	21	f	públ	mbio	106	3.03	5	36	35
26	20	f	priv	mbio	108	3.61	3	20	10
27	22	f	públ	mbio	130	2.73	5	15	2
28	21	f	priv	pmed	128	3.54	3	18	5

2.1Organización de datos Cuantitativos Discretos

2.1.1 Tablas de Frecuencias: Los datos cuantitativos discretos se organizan en tablas, llamadas ***Tablas de Distribución de frecuencias. Los*** tipos de frecuencias: son

Frecuencia absoluta: Indica el número de veces que se repite un valor de la variable.

Frecuencia relativa: Indica la proporción con que se repite un valor. Se obtiene dividiendo la frecuencia absoluta entre el tamaño de la muestra. Para una mejor interpretación es más conveniente mutiplicarla por 100 para trabajar con una ***Frecuencia relativa porcentual***.

Frecuencia absoluta acumulada: Indica el número de valores que son menores o iguales que el valor dado.

Frecuencia relativa porcentual acumulada: Indica el porcentaje de datos que son menores o iguales que el valor dado.

Tabla de distribucion de frecuencias

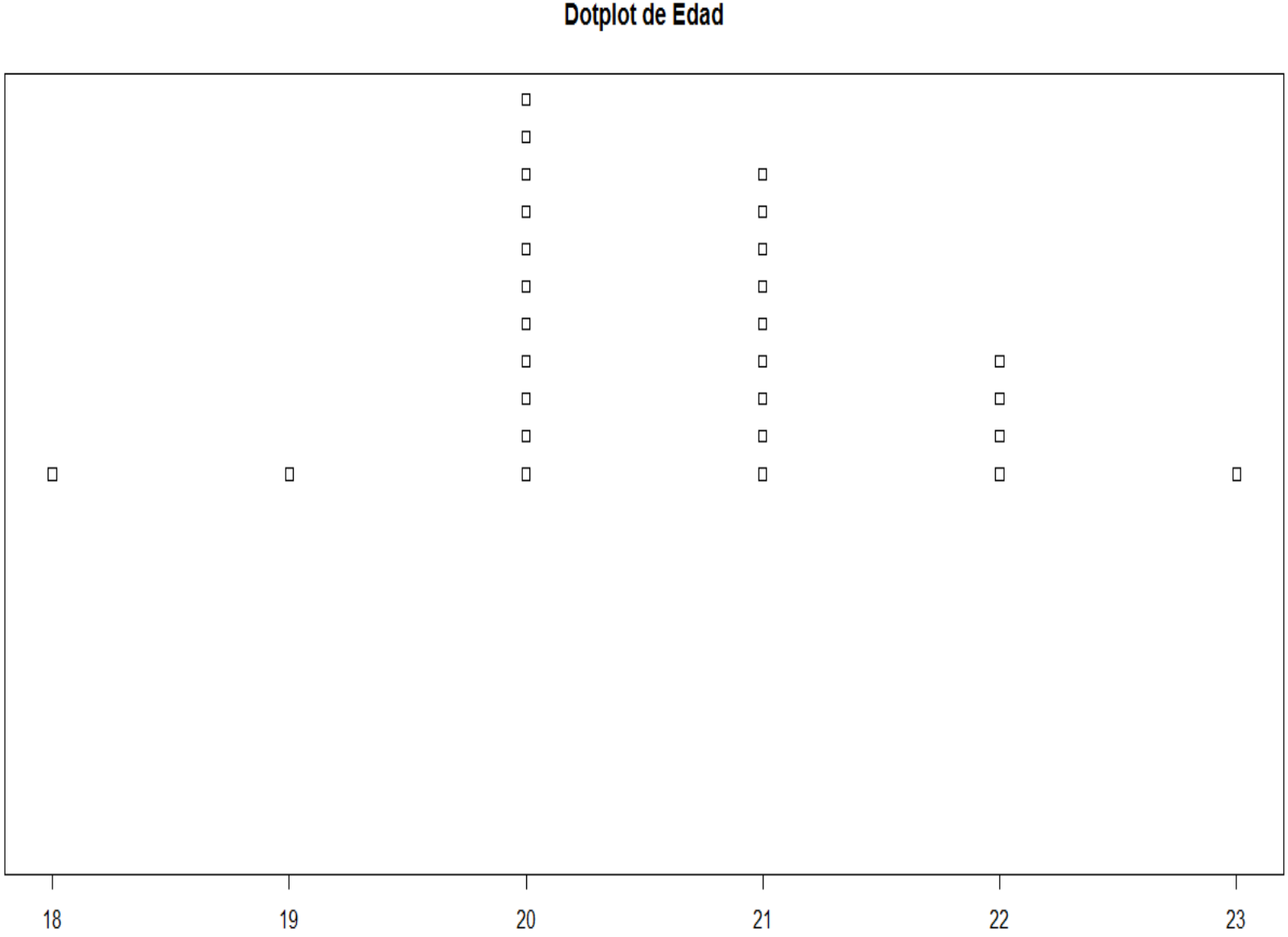
edad	fabs	frelat	fabs.acum	frelat.acum
18	1	3.571429	1	3.571429
19	1	3.571429	2	7.142857
20	12	42.857143	14	50.000000
21	9	32.142857	23	82.142857
22	4	14.285714	27	96.428571
23	1	3.571429	28	100.000000

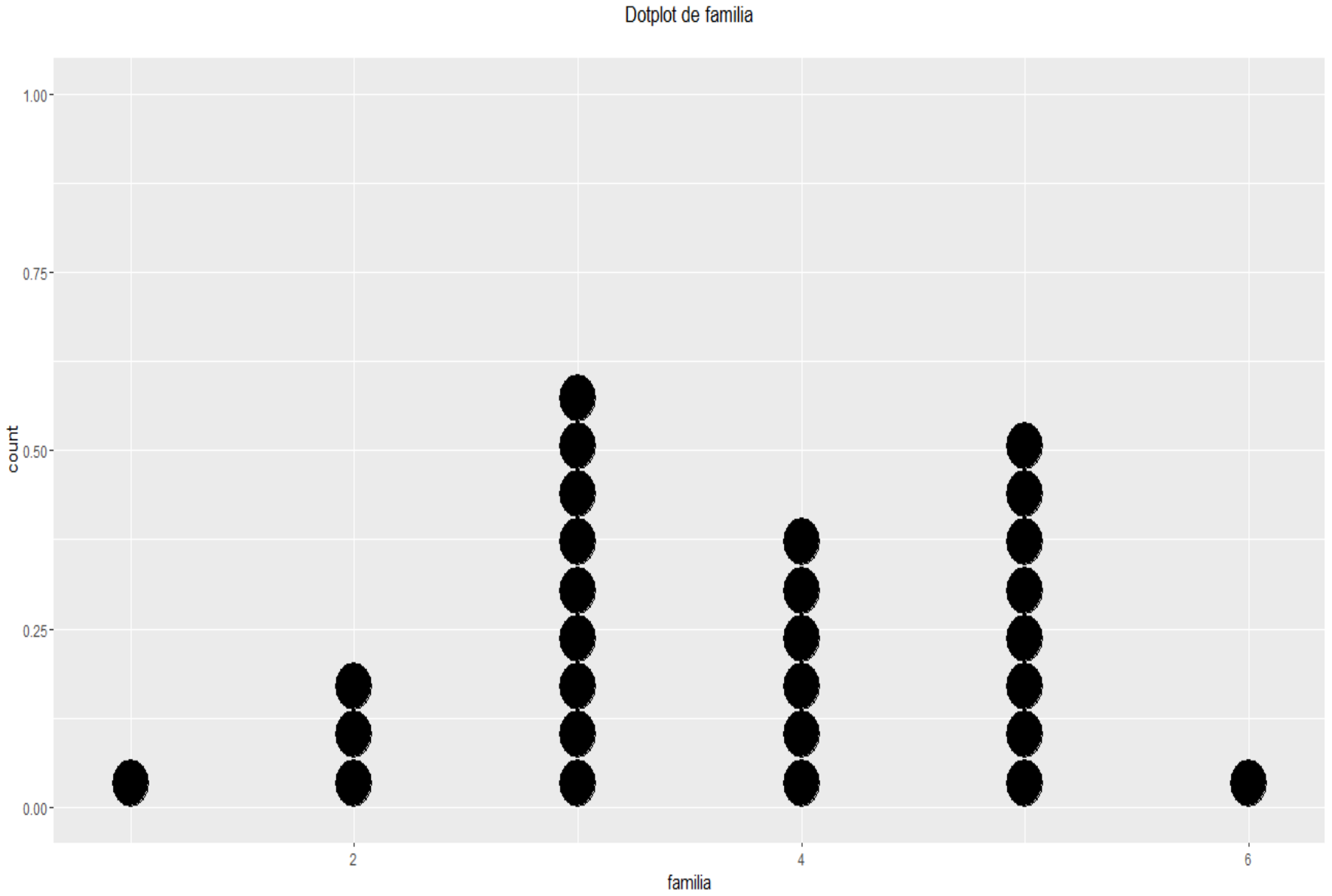
Laboratorio 2

Lab 2: Leyendo datos y tabla de frecuencias datos discretos

2.1.2 El plot de puntos (“Dotplot”)

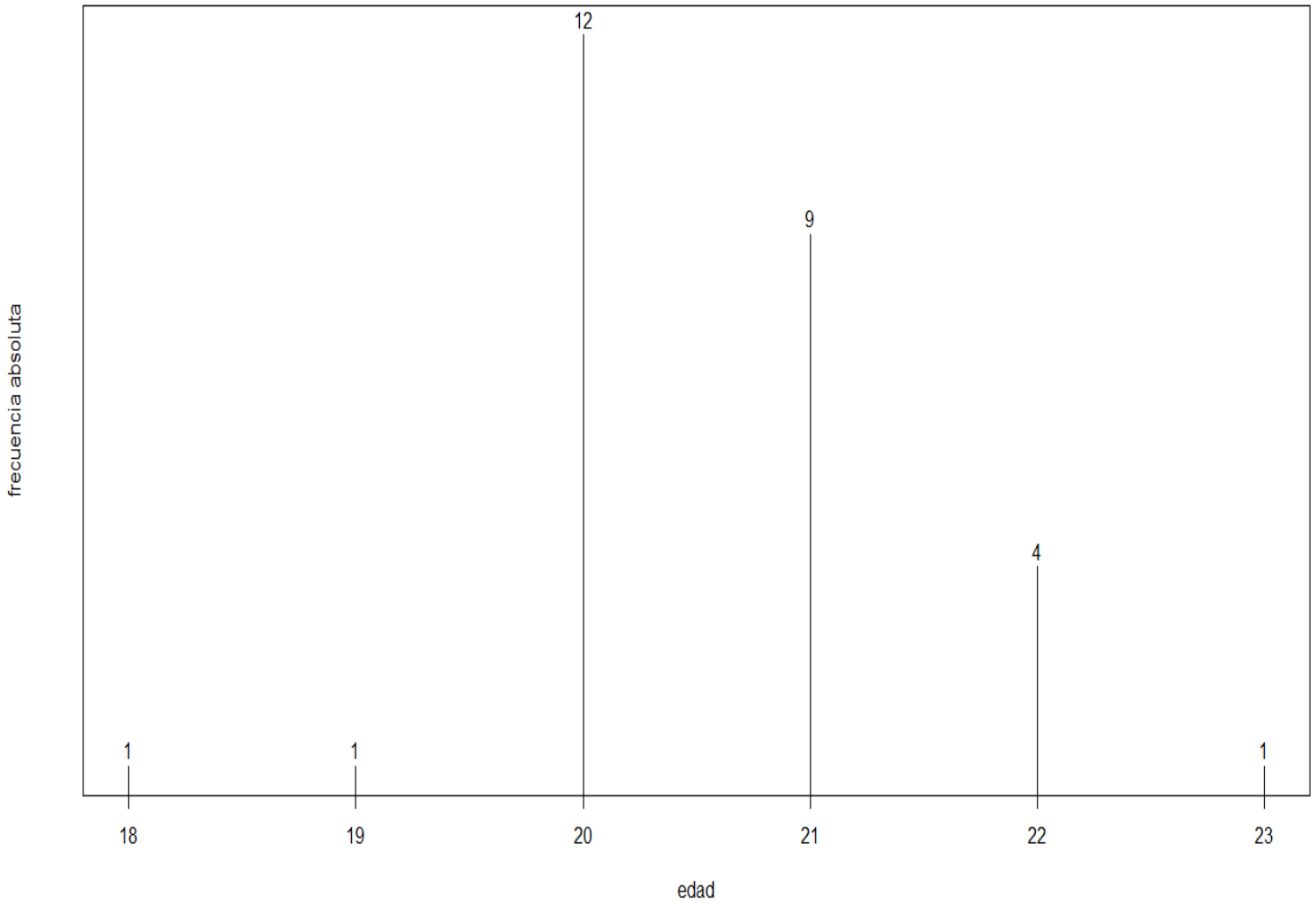
La gráfica más elemental para la distribución de fercuencias de datos discretos es el plot de puntos (“Dotplot”) que consiste en colocar un punto cada vez que se repite un valor. Esta gráfica permite explorar la simetría y el grado de variabilidad de la distribución de los datos con respecto al centro, el grado de concentración o dispersión de los datos con respecto al valor central y permite detectar la presencia de valores anormales (“outliers”).

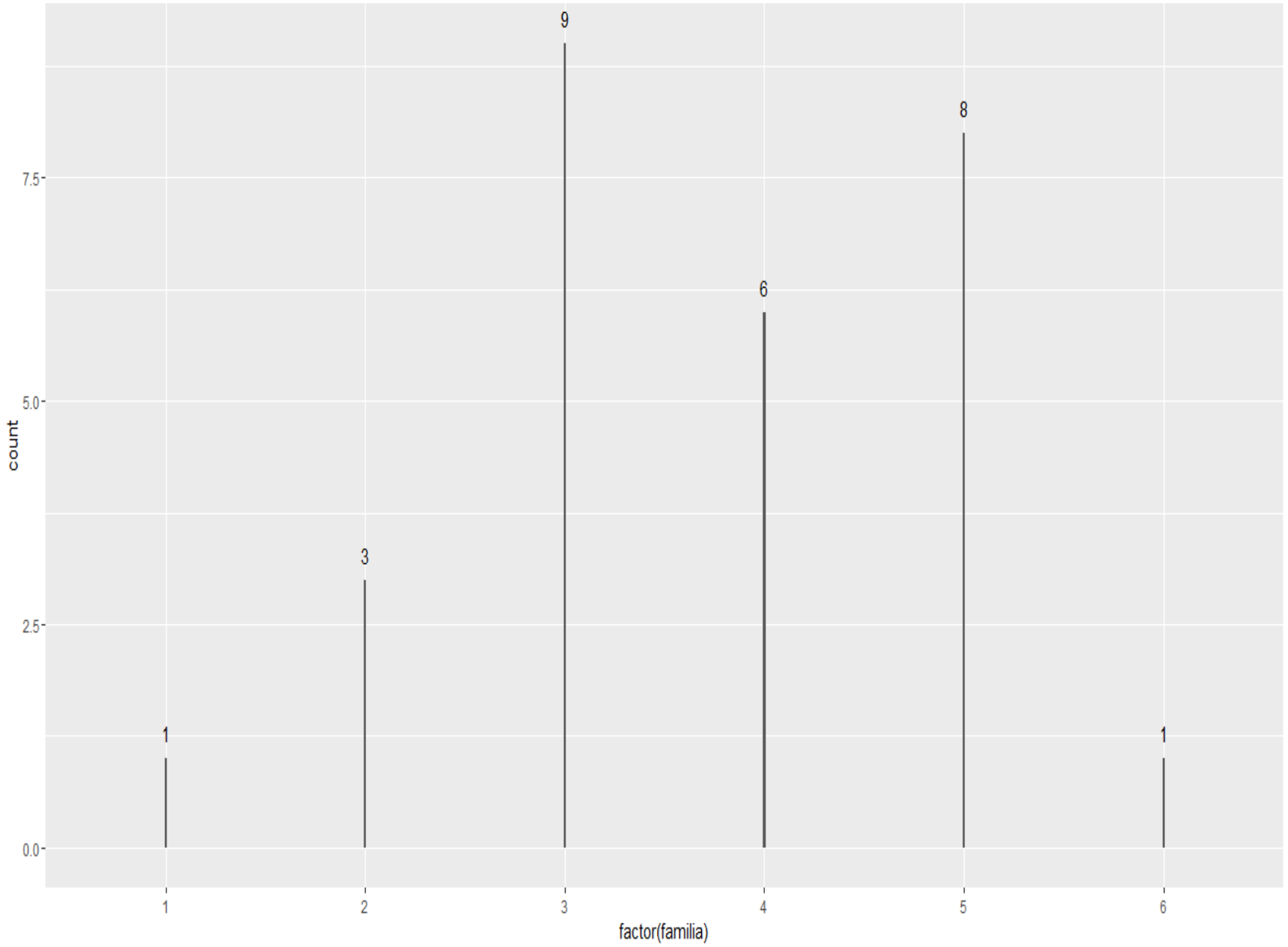

Dotplot de Edad
18
19
20
21
22
23


Dotplot de familia
1.00
0.75
0.50
0.25
0.00
count
2
4
6
familia

2.1.3 Gráfica de Línea

La gráfica de línea es una alternativa a la gráfica de puntos. Por cada valor de la variable se traza una linea vertical de altura proporcional a la frecuencia absoluta del valor de la variable.


12
9
4
1
1
1
frecuencia absoluta
18
19
20
21
22
23
edad


9
8
7.5
6
5.0
count
3
2.5
1
1
0.0
1
2
3
4
5
6
factor(familia)

2.2 Organización de datos Cuantitativos Continuos

Cuando los datos son de una variable continua o de una variable discreta que asume muchos valores distintos, ellos se agrupan en clases que son representadas por intervalos y luego se construye una tabla de frecuencias, cada frecuencia absoluta (relativa porcentual) representa el número (porcentaje) de datos que caen en cada intervalo.

Recomendaciones acerca del número de intervalos de clases:

- El número de intervalos de clases debe variar entre 5 y 20.
- Se debe evitar que hayan muchas clases con frecuencia baja o cero, de ocurrir ello es recomendable reducir el número de clases.
- A un mayor número de datos le corresponde un mayor número de clases.

2.2 Organización de datos Cuantitativos Continuos (cont)

Una regla bien usada es que el número de clases debe ser aproximadamente igual a la raíz cuadrada del número de datos. También está la regla de Sturges,en donde el número de clases está dado por 1+3.3*log(número de datos)

Una vez que se determina el número de clases se determina la amplitud de cada clase usando la siguiente fórmula:

Amplitud del intervalo de clase ≈ (Mayor Valor-Menor Valor)/numero de intervalos

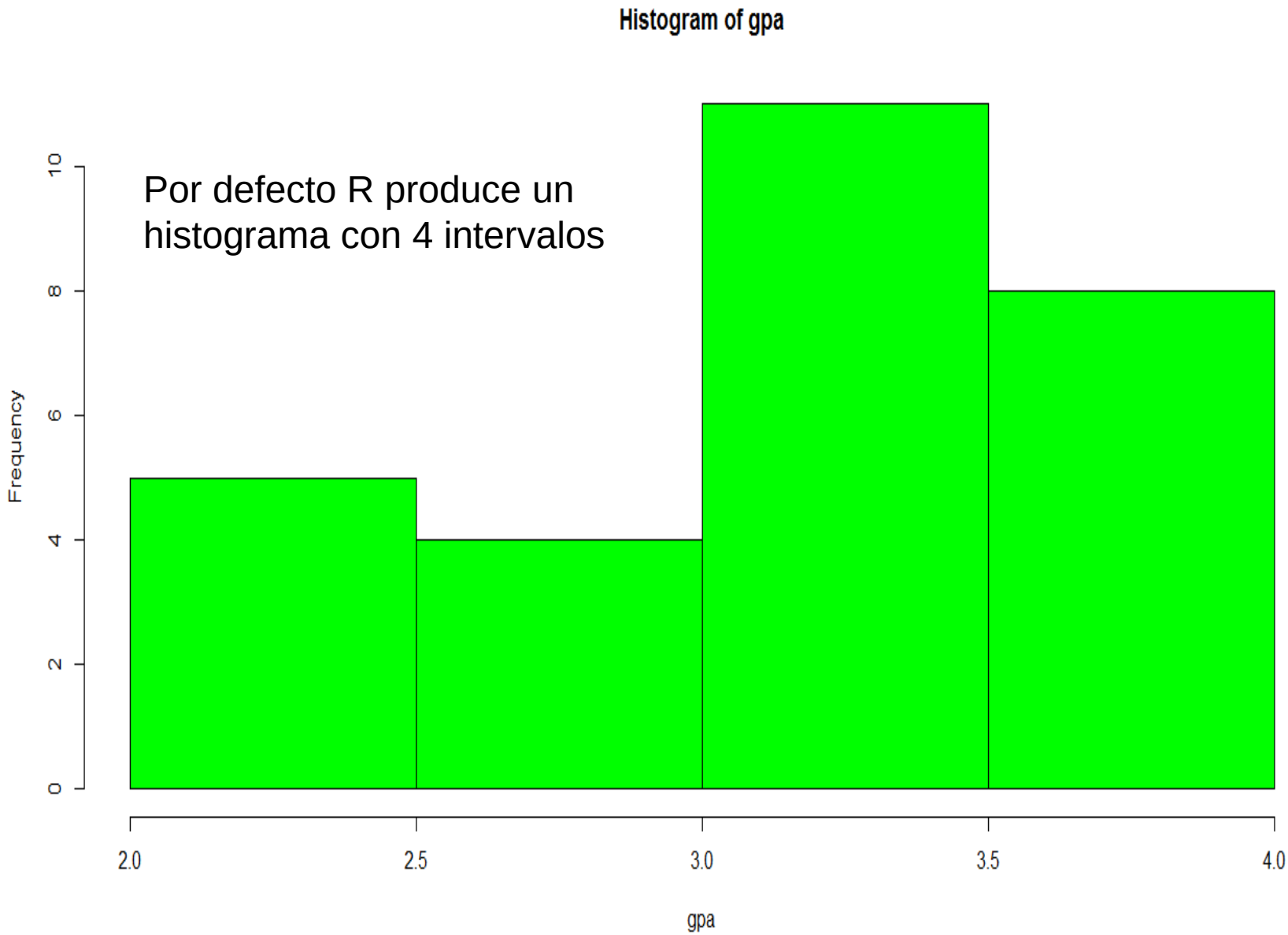
Usualmente la amplitud se redondea a un número cómodo de usar.

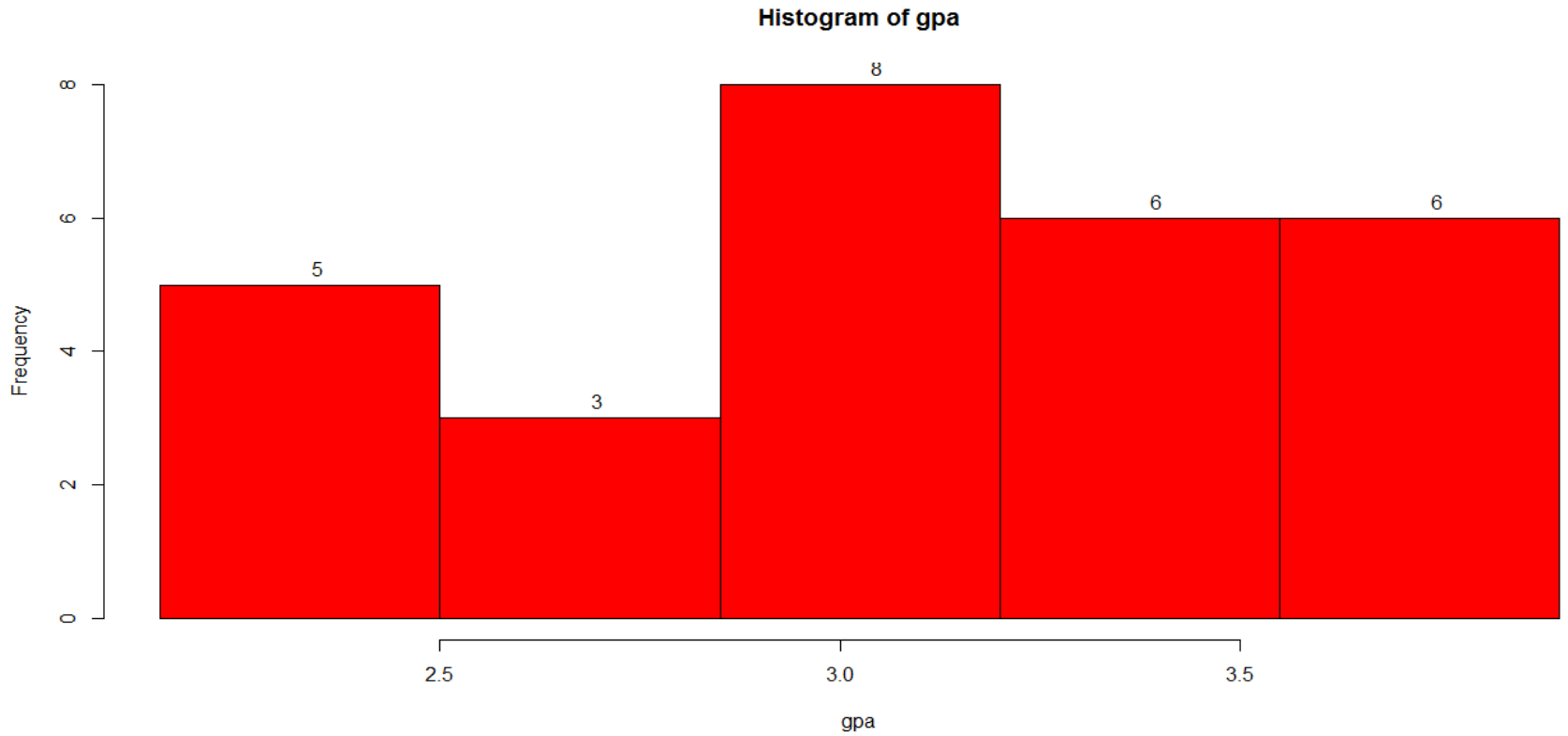
3.2.2 Histograma

Es la gráfica de la tabla de distribución de frecuencias para datos agrupados, consiste de barras cuyas bases son los intervalos de clases y cuyas alturas son proporcionales a las frecuencias absolutas (o relativas) de los correspondientes intervalos.

Laboratorio 3

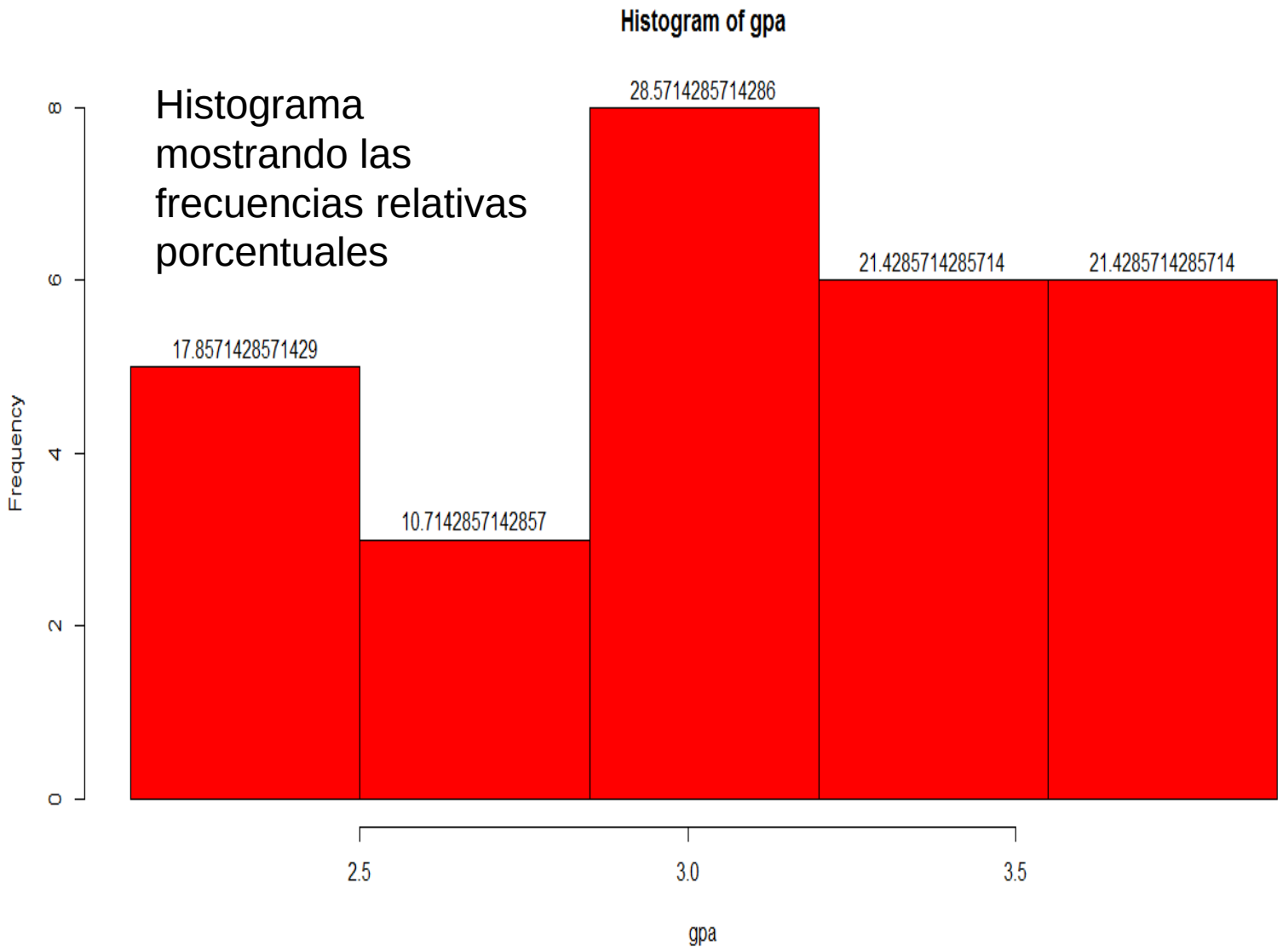
Tabla de frecuencias para datos continuos e histograma
Tabla de frecuencias para datos categoricos

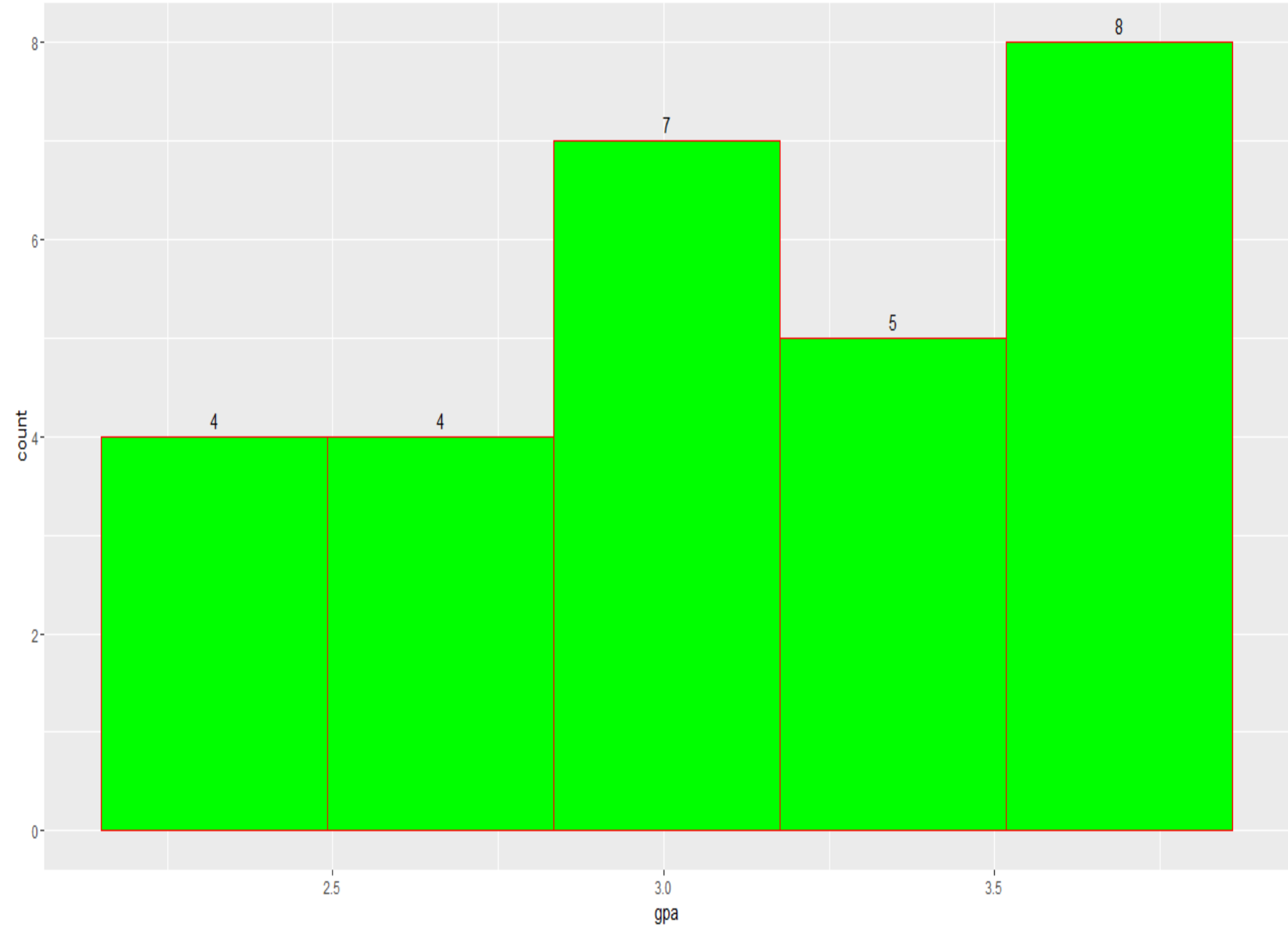

Histogram of gpa
Por defecto R produce un histograma con 4 intervalos
Frequency
0
2
4
6
8
10
2.0
2.5
3.0
3.5
4.0
gpa



Histograma asimetrico hacia la izquierda, la frecuencia es baja para valores bajos de GPAy la frecuencia es alta para valores altos de GPA

Histograma mostrando las frecuencias relativas porcentuales




8
6
4
2
0
count
4
4
7
5
8
2.5
3.0
3.5
gpa

Interpretacion del histograma

Un histograma es simetrico con respecto al centro si los intervalos a la derecha e izquierda del centro tienen similar frecuencia.

Un histograma es asimetrico hacia la derecha si hay mayor concentracion de datos a la izquierda que a la derecha del centro. O sea para valores altos de las variablesle corresponde mayor frecuencia que valores bajos de la variable.

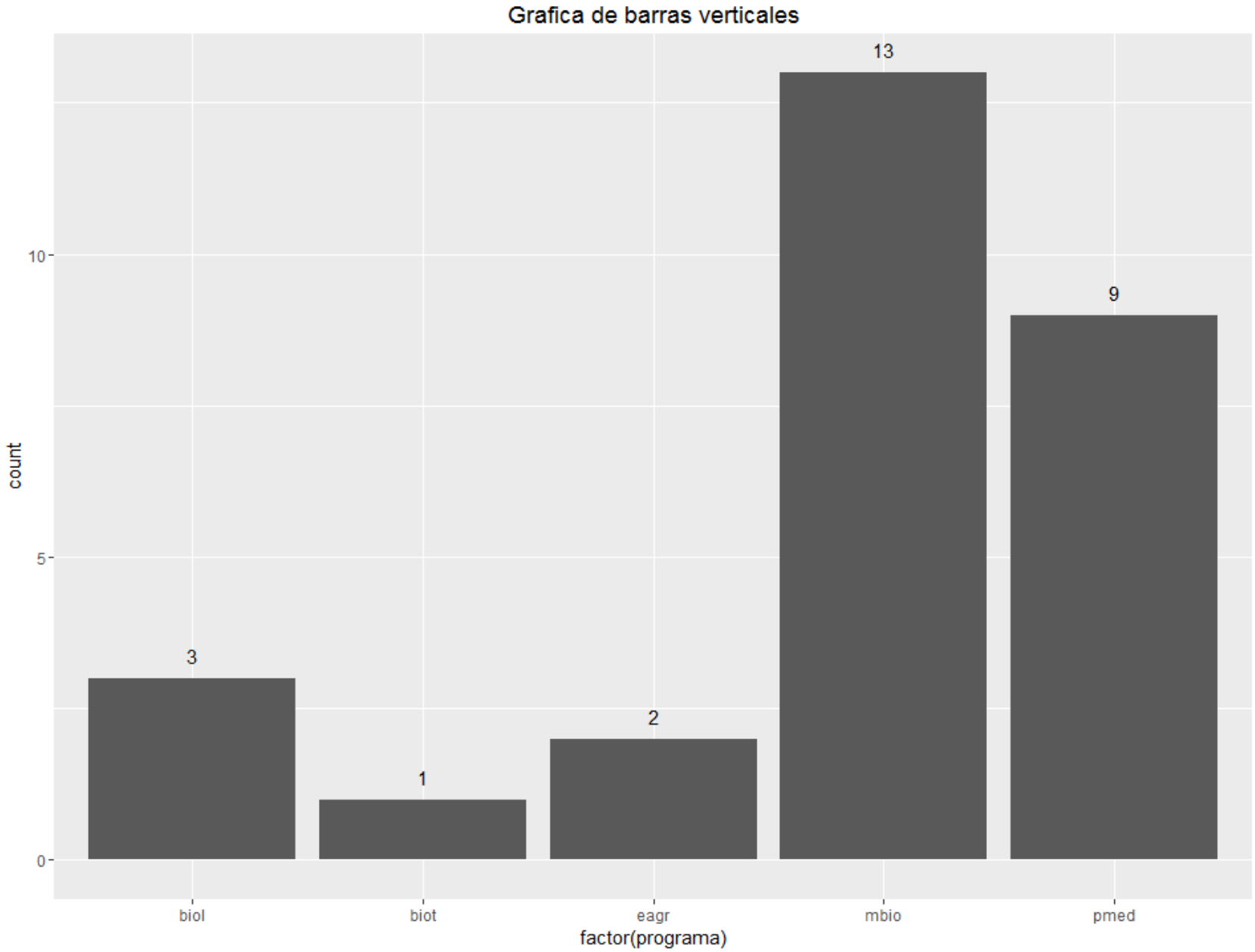
Un histograma es asimetrico a la izquierda si hay mayor concentracion de datos a la derecha que a la izquierda del centro. O sea que a valores bajos de la variable le corresponde mayor frecuencia que a valores altos de la variable.

2.3 Presentación de datos cualitativos

En este caso los datos también se pueden organizar en tablas de frecuencias, pero las frecuencias acumuladas no tienen mucho significado, excepto cuando la variable es ordinal.

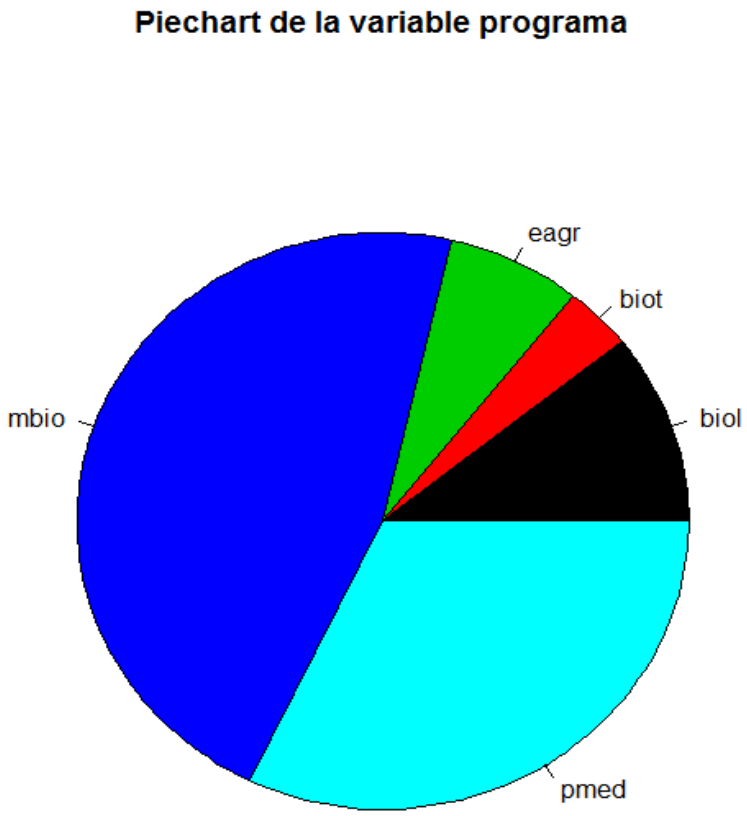
2.3.1 Gráficas de Barras

Las gráficas de barras pueden ser verticales u horizontales.


Grafica de barras verticales
13
10
9
count
5
3
2
1
0
biol
biot
eagr
mbio
pmed
factor(programa)

2.3.2 Gráficas Circulares

Este tipo de gráfica se usa cuando se quiere tener una idea de la contribución de cada valor de la variable al total. Aunque es usada más para variables cualitativas, también podría usarse para variables cuantitativas discretas siempre que la variable no asuma muchos valores distintos.


Piechart de la variable programa
eagr
biot
biol
mbio
pmed

Laboratorio 4

Graficas interactivas en Python

Hay varias librerias para hacer graficas interactivas en Python . Las principales son:

Ggplot
Plolty
Seaborn
Bokeh
Pygal

2.4 Gráfica de tallo y hojas (“Stem-and-Leaf ”)

Es una gráfica usada para datos cuantitativos. Es la gráfica más básica de un conjunto de técnicas conocido con el nombre de Análisis Exploratorio de Datos (EDA) introducida por John Tukey a mediados de los años 70.

La idea es considerar los primeros dígitos del dato como una rama del tallo (“stem”) y el último dígito como una hoja (“leaf”) de dicha rama. Las ramas son ordenadas en forma creciente.

Ejemplo 2.4. Los siguientes datos representan pesos de una muestra de 15 varones adultos.

165 178 185 169 152 180 175 189 195 200 183 191 197 208 179

Hacer su gráfica de “Stem-and Leaf”.

Solución: En este caso las ramas la forman los primeros dos dígitos de los datos, y las hojas serán dadas por los últimos dígitos de los datos.

continuación: ...

Ejemplo 2.4.

Luego el “stem-and leaf “ será de la siguiente manera:

15	2
16	59
17	598
18	0935
19	517
20	08

Interpretación: *El uso del “stem-and-leaf” es exactamente igual al del Histograma, la única diferencia está en que del “stem-and-leaf” se pueden recuperar los datos muestrales, pero de un histograma no se puede hacer. En este ejemplo el “stem-and-leaf” es asimétrico a la izquierda, no tiene mucha variabilidad ni “outliers”.*