American University of Beirut

STAT 210: Elementary Statistics for Sciences

2022–2023 Fall

Siamak Taati

Chapter 2
Organizing and visualizing data
Part 1

Organizing and visualizing data

Raw data

Scenario

A group of zoologists gather data on a sample of size 50 from a species of birds living in Lebanon.

	sex	weight	tail.len	feather.col	beak.col	noct
1	F	98.80	7.50	yellow	red	0
2	M	95.80	7.10	dark blue	pink	0
3	F	92.50	6.80	yellow	pink	0
4	M	86.20	7.30	light blue	pink	0
5	F	87.30	6.40	green	red	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	F	98.00	7.50	yellow	pink	0

This data set corresponds to the raw data as collected by the zoologists (i.e., before processing).

Raw data

Scenario

A group of zoologists gather data on a sample of size 50 from a species of birds living in Lebanon.

	sex	weight	tail.len	feather.col	beak.col	noct
1	F	98.80	7.50	yellow	red	0
2	M	95.80	7.10	dark blue	pink	0
3	F	92.50	6.80	yellow	pink	0
4	M	86.20	7.30	light blue	pink	0
5	F	87.30	6.40	green	red	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	F	98.00	7.50	yellow	pink	0

`sex`	sex of the bird	(categorical)
`weight`	weight in grams	(numerical)
`tail.len`	tail length in centimeters	(numerical)
`feather.col`	feather color	(categorical)
`beak.col`	beak color	(categorical)
`noct`	has nocturnal activity (1: yes, 0: no)	(categorical)

Organizing and visualizing categorical data

Frequency distribution of categorical variables

What is the frequency of birds with each feather color?

feather.col	frequency
dark blue	11
green	8
light blue	12
yellow	19

- The frequency (or count) of a category is the number of cases belonging to that category.

Organizing and visualizing categorical data

Frequency distribution of categorical variables

What is the relative frequency of birds with each feather color?

feather.col	relative frequency
dark blue	0.22
green	0.16
light blue	0.24
yellow	0.38

- The frequency (or count) of a category is the number of cases belonging to that category.
- The relative frequency of a category is the proportion of cases belonging to that category.

Organizing and visualizing categorical data

Frequency distribution of categorical variables

What is the percentage of birds with each feather color?

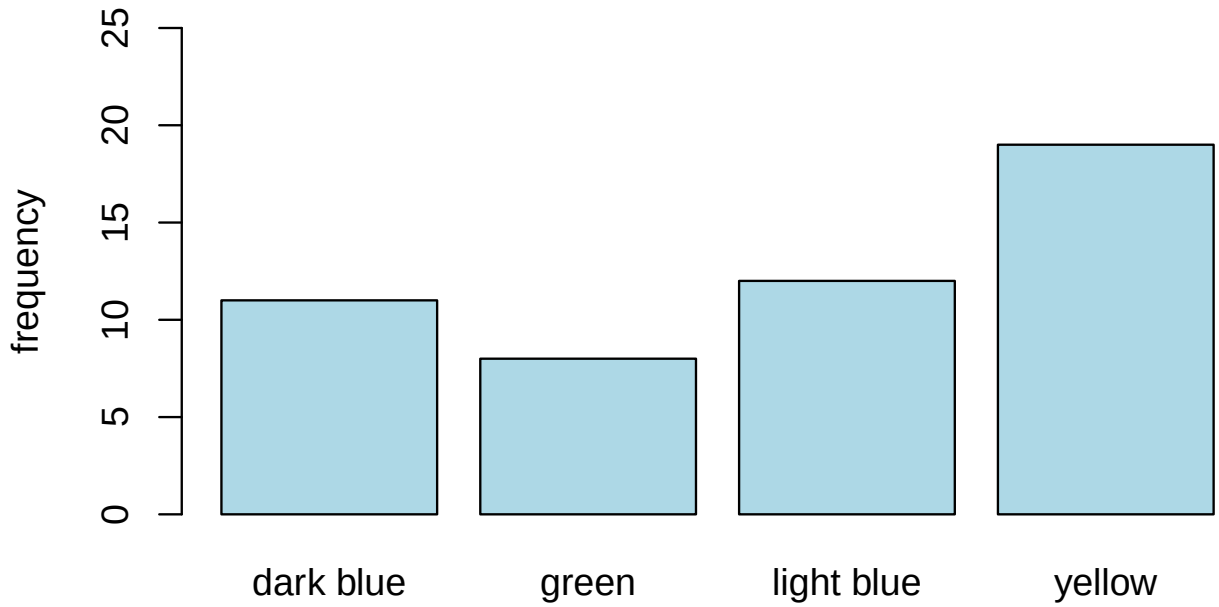
feather.col	percentage
dark blue	22.00
green	16.00
light blue	24.00
yellow	38.00

- The frequency (or count) of a category is the number of cases belonging to that category.
- The relative frequency of a category is the proportion of cases belonging to that category.
- Relative frequencies can also be expressed as percentages.

Organizing and visualizing categorical data

Frequency distribution of categorical variables

The distribution of birds with each feather color can be visualized with a bar chart.

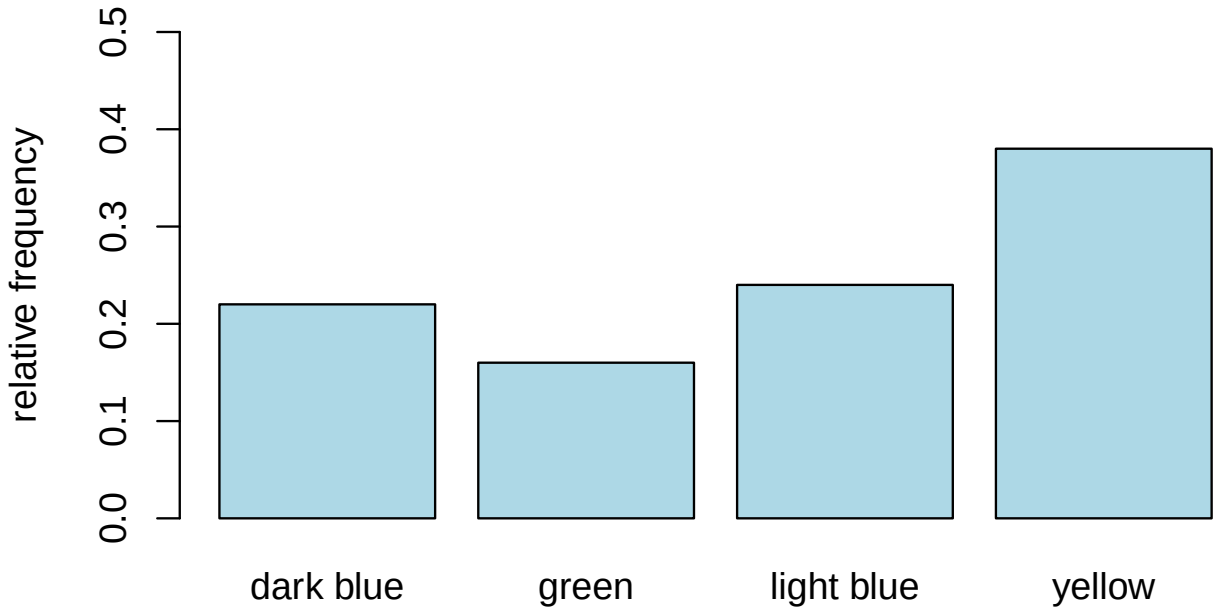


A bar chart (frequency type)

Organizing and visualizing categorical data

Frequency distribution of categorical variables

The distribution of birds with each feather color can be visualized with a bar chart.

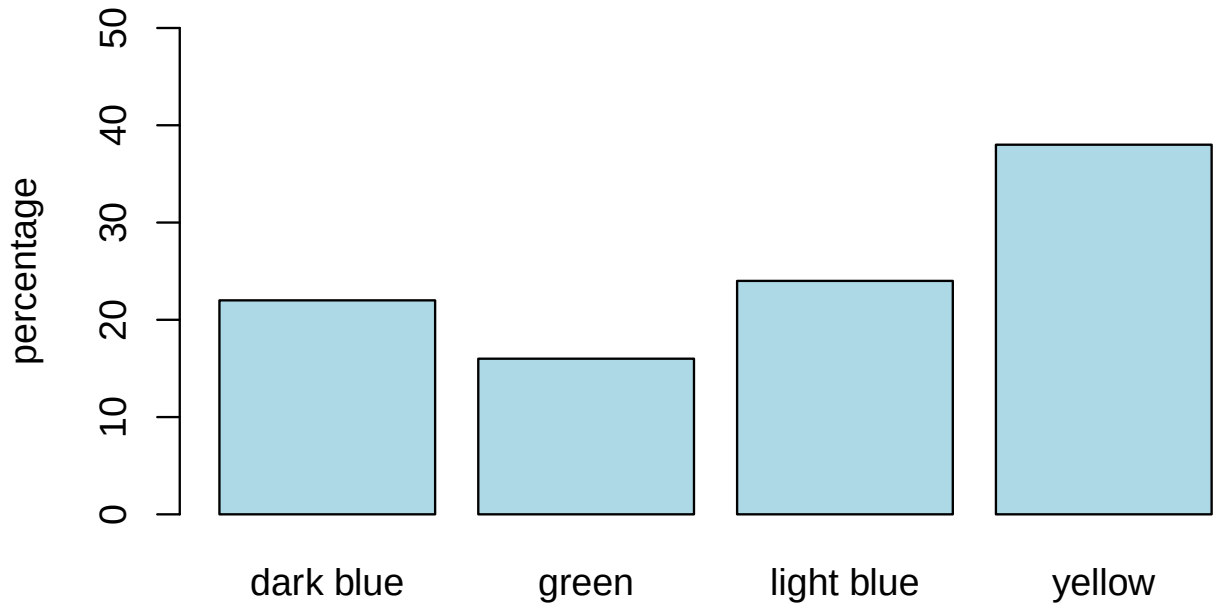


A bar chart (relative frequency type)

Organizing and visualizing categorical data

Frequency distribution of categorical variables

The distribution of birds with each feather color can be visualized with a bar chart.

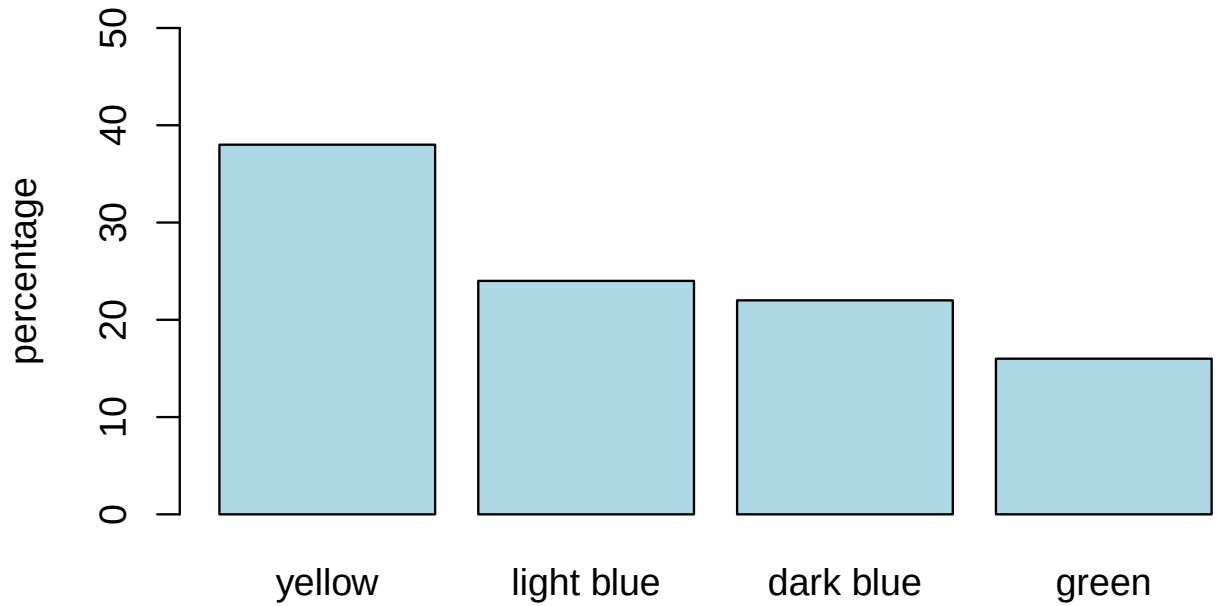


A bar chart (percentage type)

Organizing and visualizing categorical data

Frequency distribution of categorical variables

The distribution of birds with each feather color can be visualized with a bar chart.

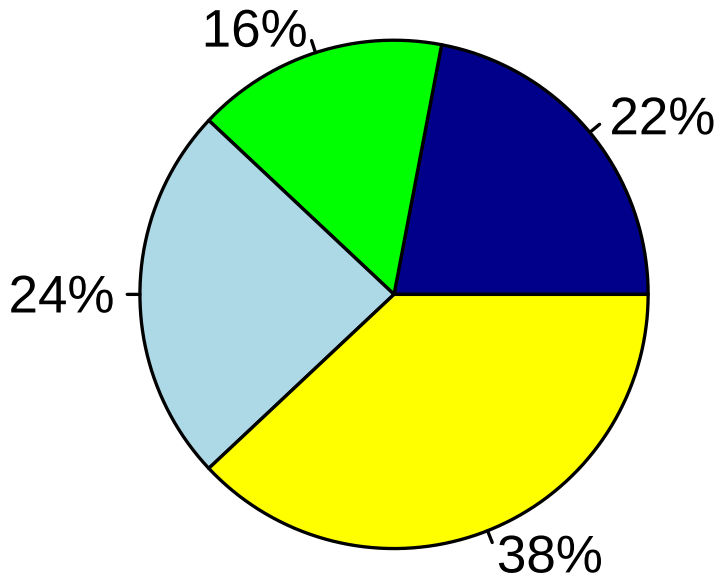


A Pareto chart (percentage type)

Organizing and visualizing categorical data

Frequency distribution of categorical variables

The distribution of birds with each feather color can also be visualized with a pie chart.



A pie chart

Organizing and visualizing numerical data

Frequency distribution of numerical variables

What is the distribution of weight among the sampled birds?

weight	frequency
75 to <80	1
80 to <85	1
85 to <90	6
90 to <95	14
95 to <100	20
100 to <105	7
105 to <110	1

The logic of the table:

- We have divided the range of values for `weight` into classes (or bins, or intervals). The table contains the frequency of birds within each class.
- The width of each class is chosen to be 5 grams.
- The lowest value is 79.7 grams, the highest value is 108.0 grams. We have rounded these to 75 and 110 to make the classes nicer.

Organizing and visualizing numerical data

Frequency distribution of numerical variables

What is the distribution of weight among the sampled birds?

weight	frequency
75 to <80	1
80 to <85	1
85 to <90	6
90 to <95	14
95 to <100	20
100 to <105	7
105 to <110	1

Grouped data (frequencies)

Organizing and visualizing numerical data

Frequency distribution of numerical variables

What is the distribution of weight among the sampled birds?

weight	relative frequency
75 to <80	0.02
80 to <85	0.02
85 to <90	0.12
90 to <95	0.28
95 to <100	0.40
100 to <105	0.14
105 to <110	0.02

Grouped data (relative frequencies)

Organizing and visualizing numerical data

Frequency distribution of numerical variables

What is the distribution of weight among the sampled birds?

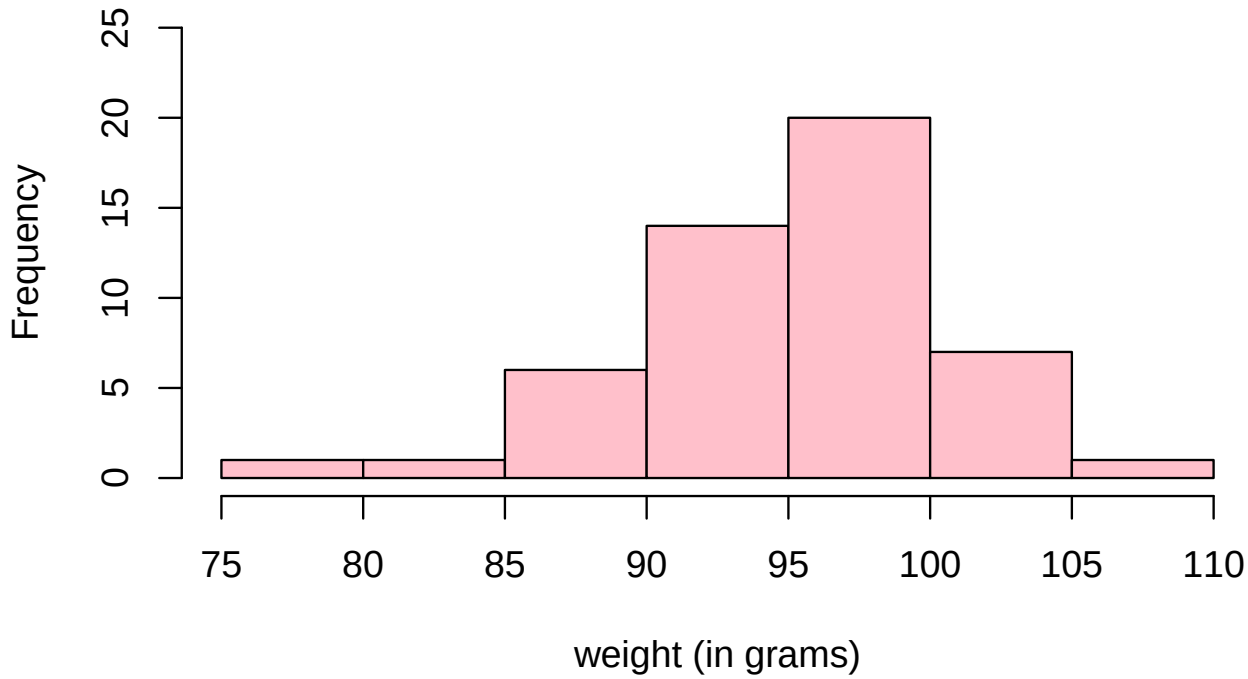
weight	percentage
75 to <80	2.00
80 to <85	2.00
85 to <90	12.00
90 to <95	28.00
95 to <100	40.00
100 to <105	14.00
105 to <110	2.00

Grouped data (percentages)

Organizing and visualizing numerical data

Frequency distribution of numerical variables

The distribution of weight among the sampled birds can be visualized with a histogram.

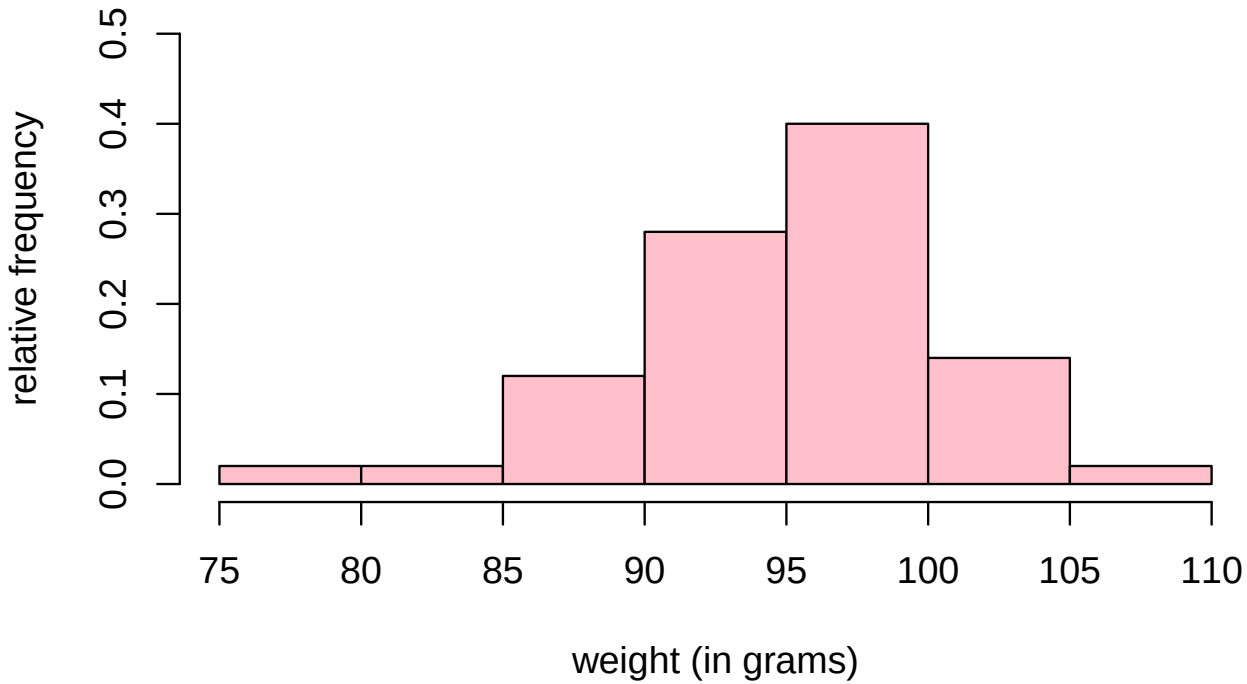


A histogram (frequency type)

Organizing and visualizing numerical data

Frequency distribution of numerical variables

The distribution of weight among the sampled birds can be visualized with a histogram.

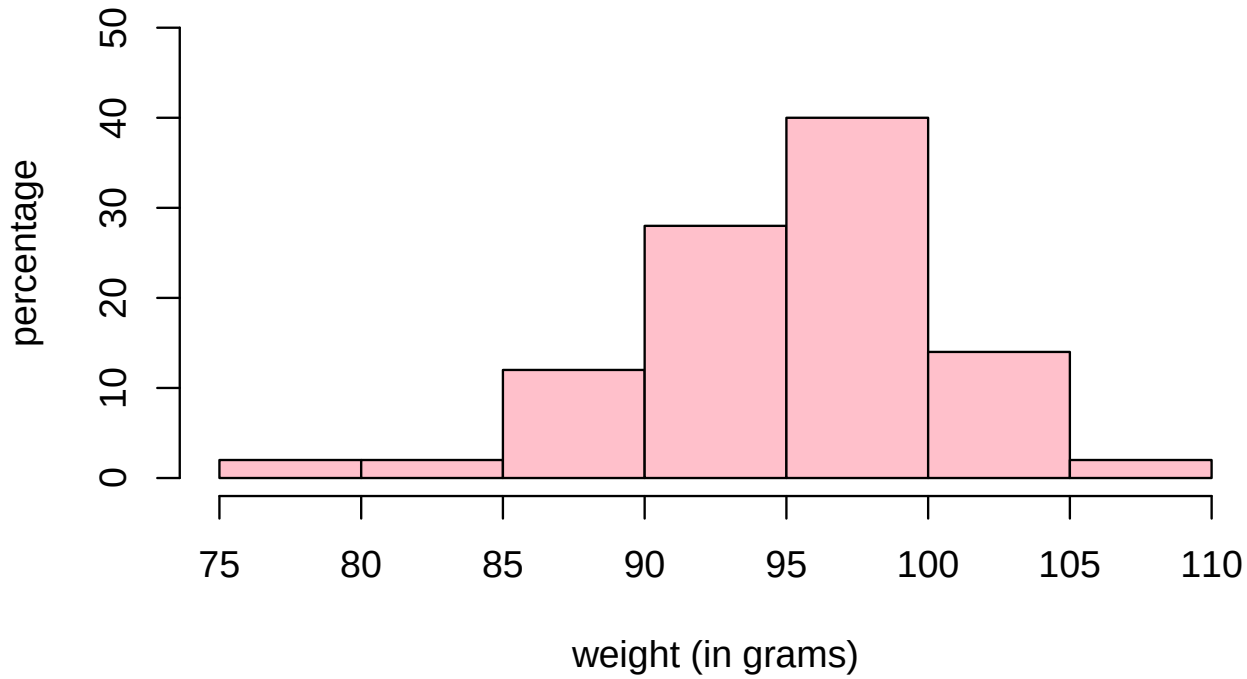


A histogram (relative frequency type)

Organizing and visualizing numerical data

Frequency distribution of numerical variables

The distribution of weight among the sampled birds can be visualized with a histogram.

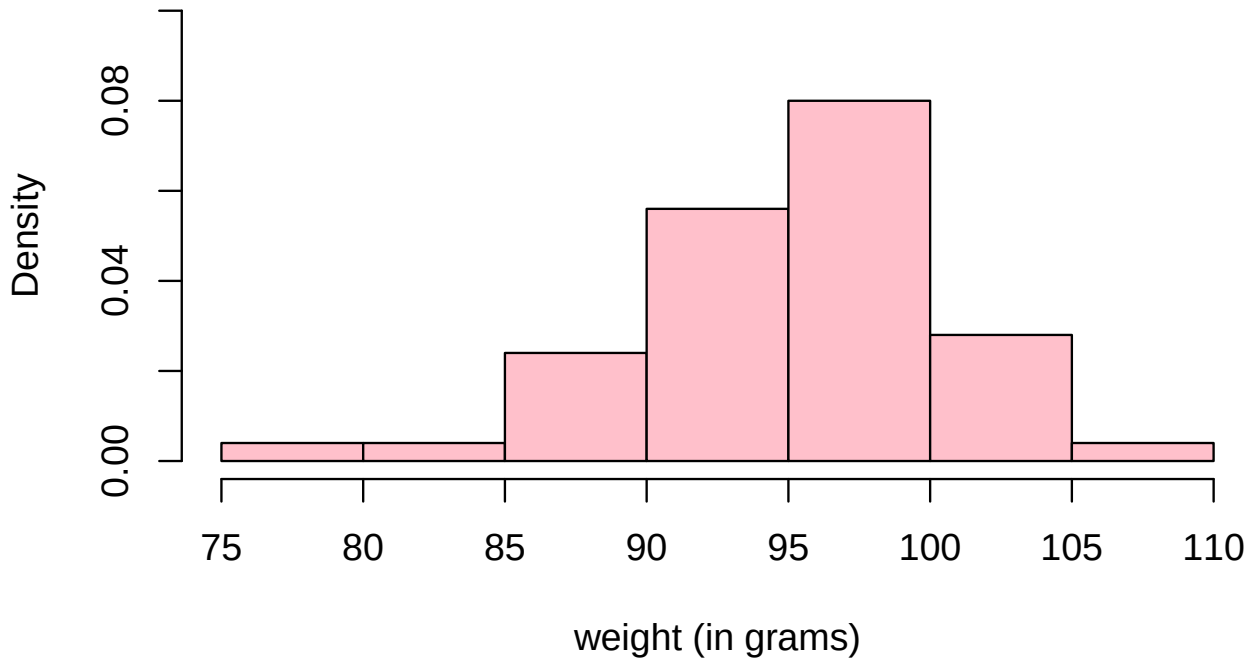


A histogram (percentage type)

Organizing and visualizing numerical data

Frequency distribution of numerical variables

The distribution of weight among the sampled birds can be visualized with a histogram.



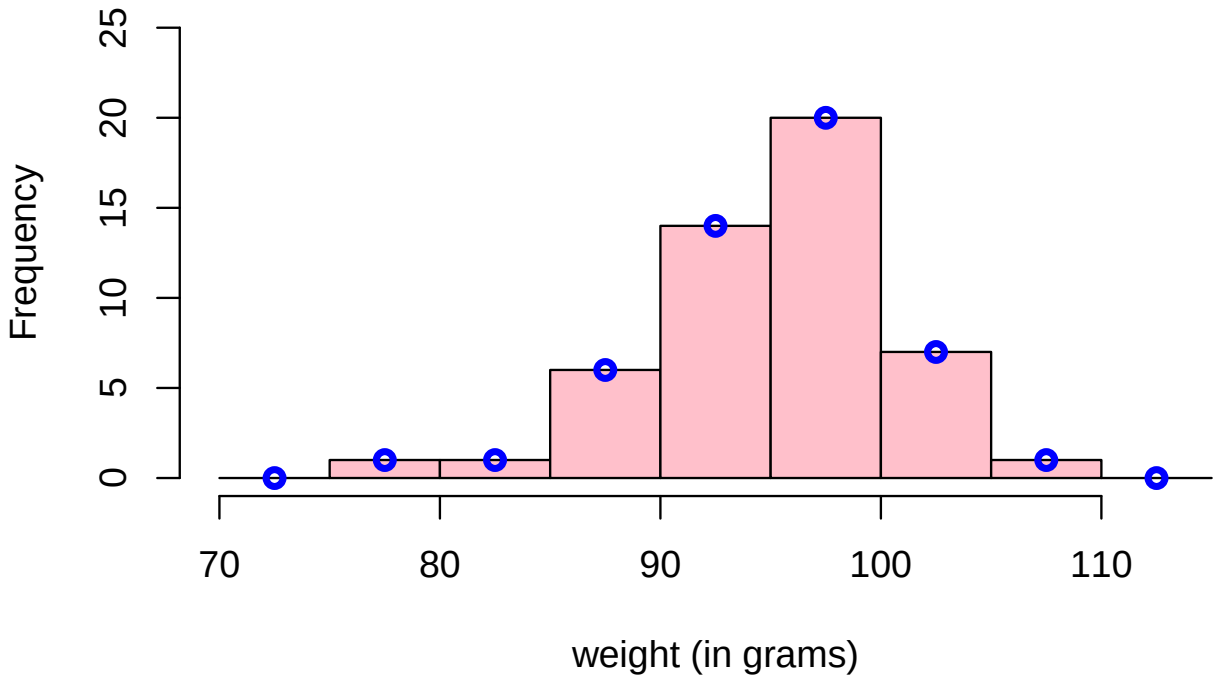
A histogram (density* type)

* Not required in this course.

Organizing and visualizing numerical data

Frequency distribution of numerical variables

The distribution of weight among the sampled birds can also be visualized with a polygon.

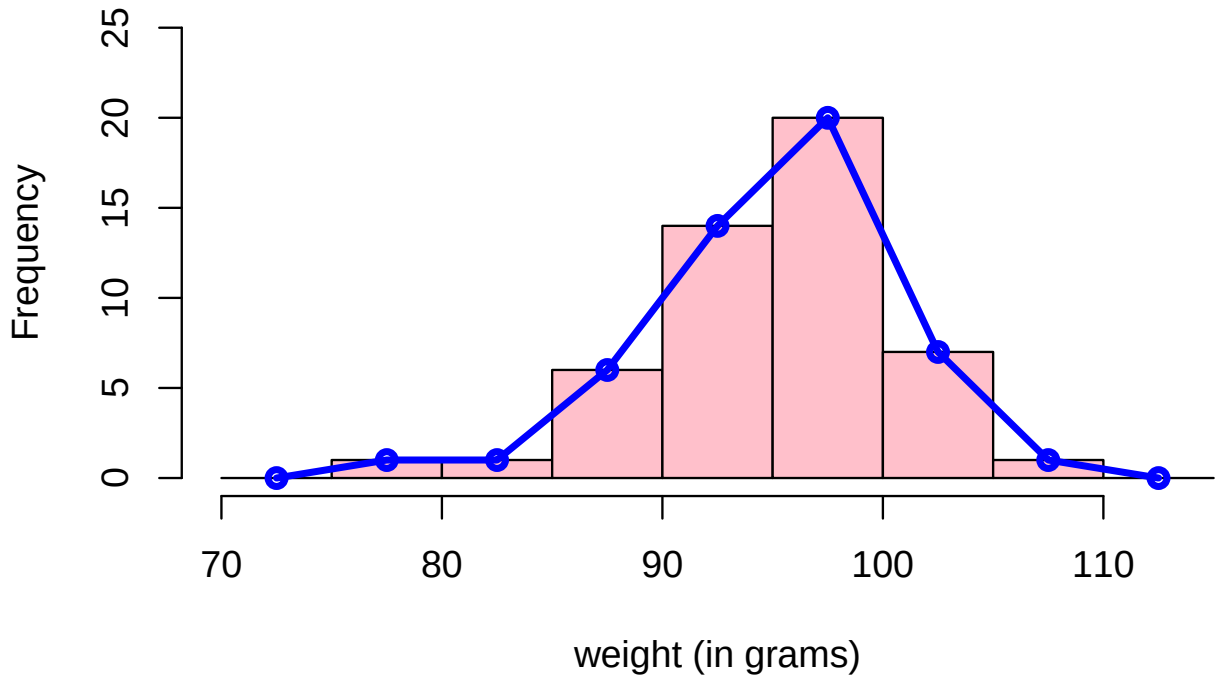


A histogram (frequency type)

Organizing and visualizing numerical data

Frequency distribution of numerical variables

The distribution of weight among the sampled birds can also be visualized with a polygon.



A polygon (frequency type)

Organizing and visualizing numerical data

Frequency distribution of (discrete) numerical variables

For discrete numerical variables (specially when there are only a few distinct values), we may not need to group the values into classes.

Scenario

Researchers from a public policy institution, study the demographics of a country based on a sample of 500 families.

	no.children	monthly.income
1	0	1538.26
2	1	6050.31
3	1	1759.54
4	0	1440.23
5	3	1654.33
⋮	⋮	⋮
500	0	1613.53

no.children	number of children in family
monthly.income	average monthly income in dollars

Organizing and visualizing numerical data

Frequency distribution of (discrete) numerical variables

For discrete numerical variables (specially when there are only a few distinct values), we may not need to group the values into classes.

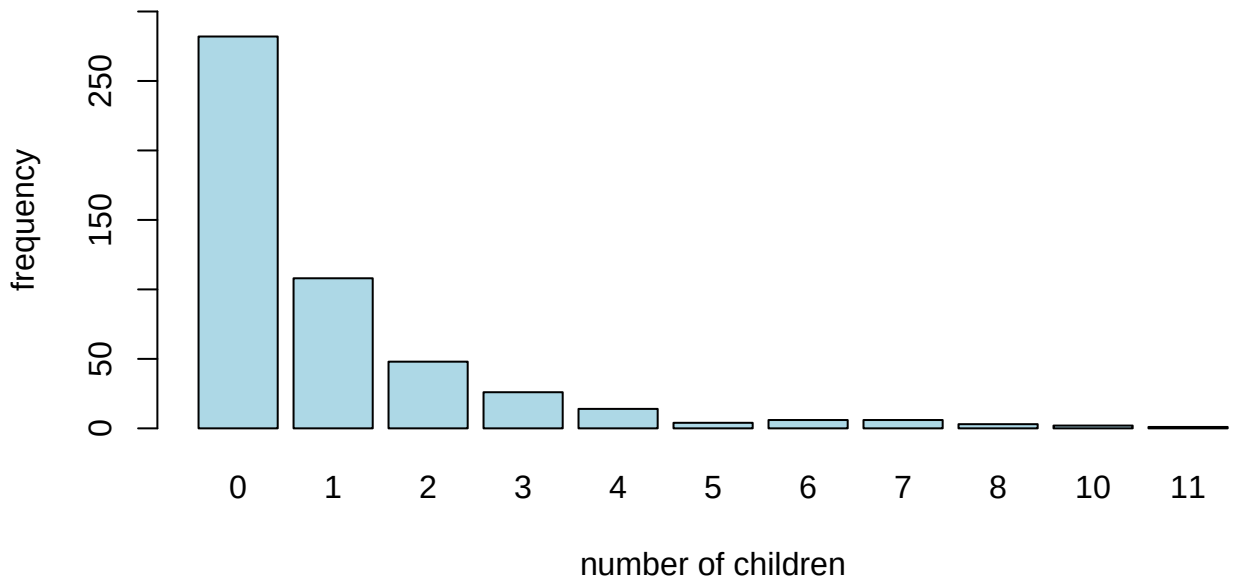
no.children	frequency	relative frequency	percentage
0	282	0.56	56.40
1	108	0.22	21.60
2	48	0.10	9.60
3	26	0.05	5.20
4	14	0.03	2.80
5	4	0.01	0.80
6	6	0.01	1.20
7	6	0.01	1.20
8	3	0.01	0.60
10	2	0.00	0.40
11	1	0.00	0.20

Distribution of the number of children

Organizing and visualizing numerical data

Frequency distribution of (discrete) numerical variables

For discrete numerical variables (specially when there are only a few distinct values), we may not need to group the values into classes.

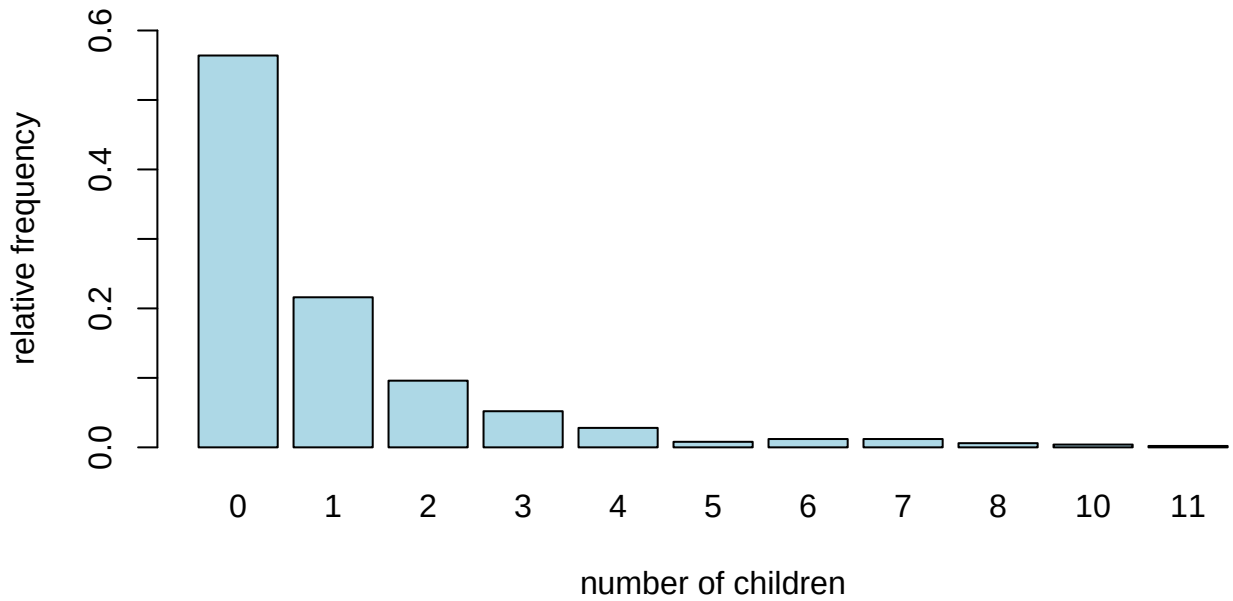


Bar chart for the distribution of number of children

Organizing and visualizing numerical data

Frequency distribution of (discrete) numerical variables

For discrete numerical variables (specially when there are only a few distinct values), we may not need to group the values into classes.

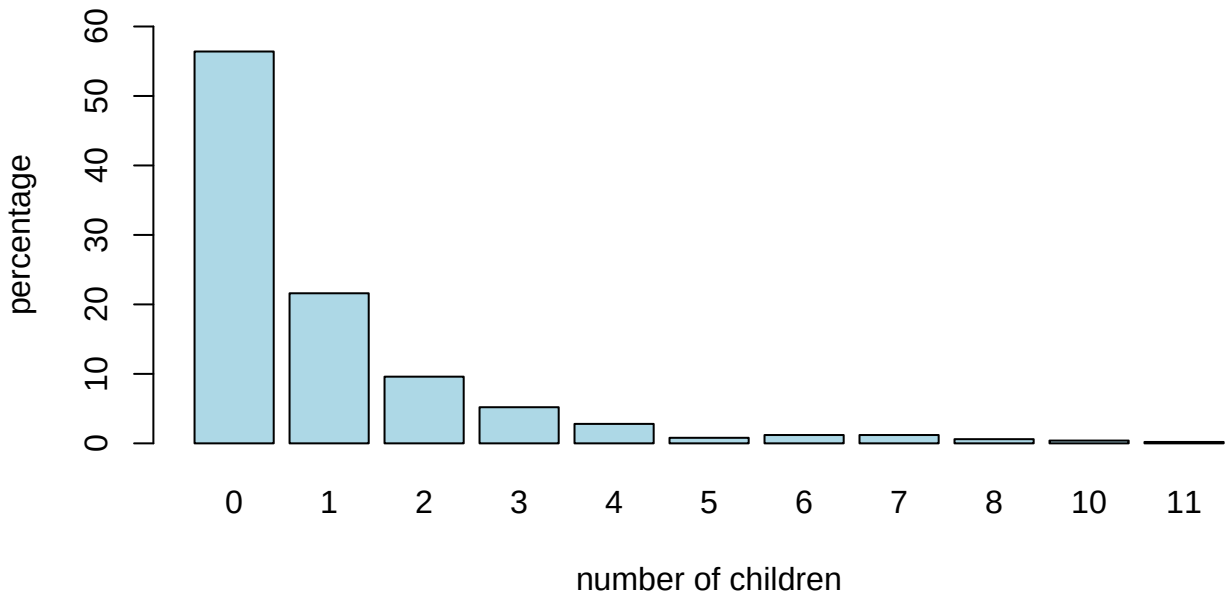


Bar chart for the distribution of number of children

Organizing and visualizing numerical data

Frequency distribution of (discrete) numerical variables

For discrete numerical variables (specially when there are only a few distinct values), we may not need to group the values into classes.



Bar chart for the distribution of number of children