Categorical and Quantitative variables

Example

Categorical	Quantitative
Type of pet owned (cat, fish, dog)	Numbers of pets owned (2 pets)
Favorite book, song	Numbers of books in the library (100 books)
Gender	Weight in pounds
Model of car	Bank account balance

Gender is a categorical variable but looks like quantitative. Because arithmetic operations doesn't make sense for it.

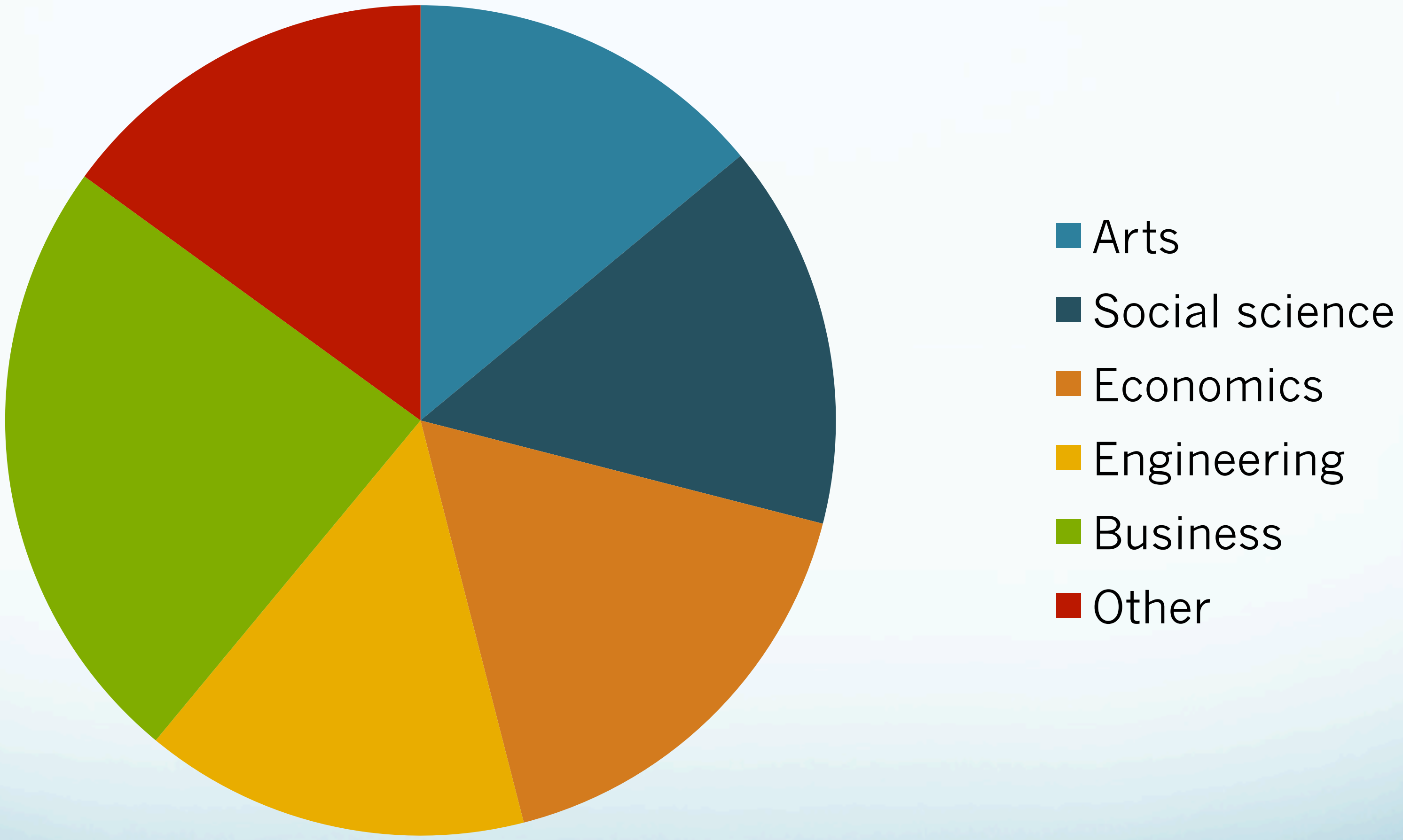
Example 1.3

Here are data on the percents of first-year students who plan to major in several areas:

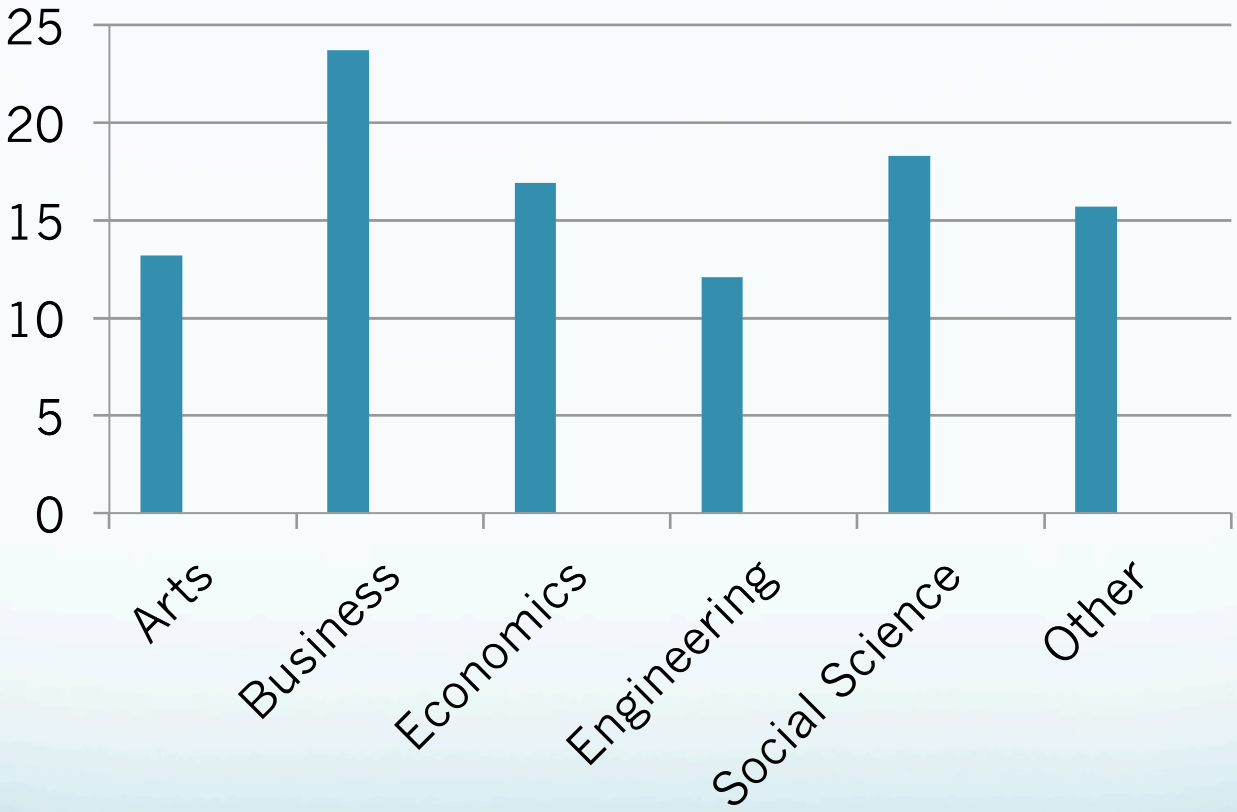
Field of study	Percent of students
Arts	13.2
Social science	18.3
Economics	16.9
Engineering	12.1
Business	23.7
Other majors	15,7
Total	99.9

Why not 100%? The exact percents would add to 100, but each percent is rounded to the nearest tenth. This is roundoff error.

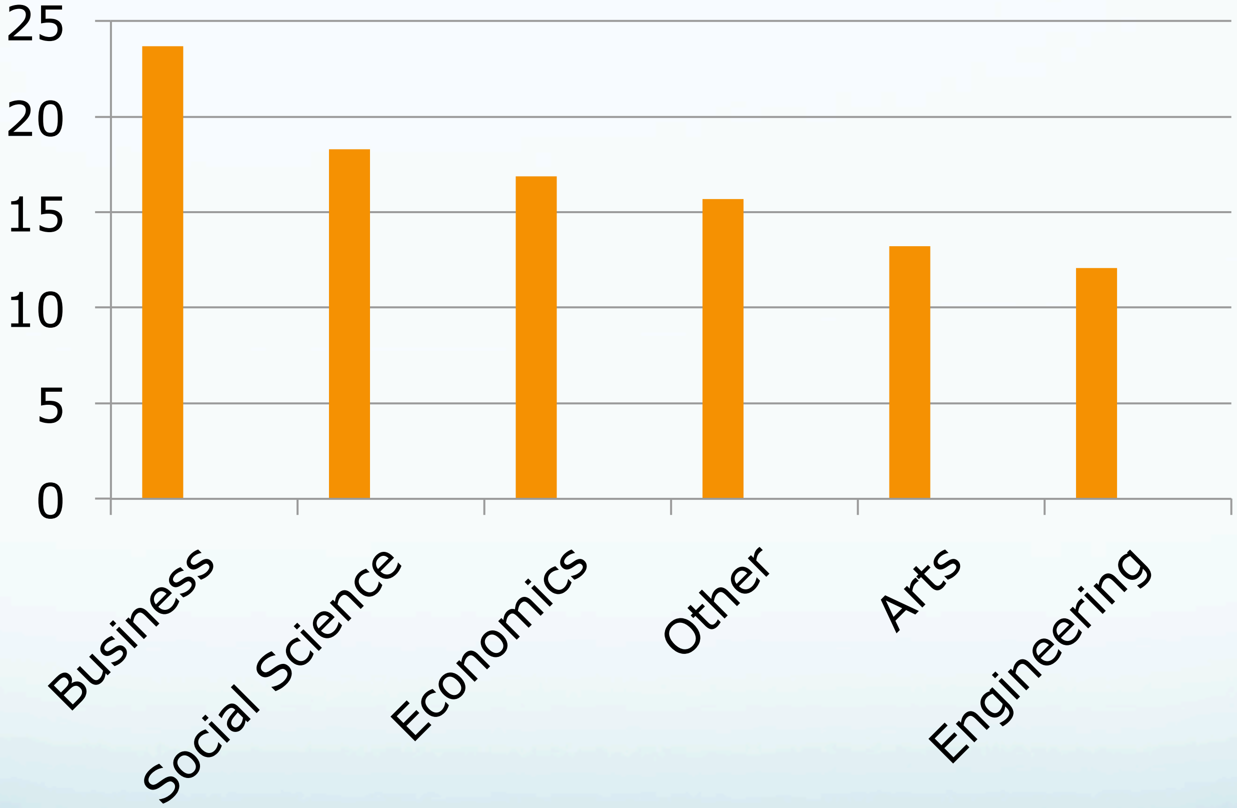
A pie chart must include all the categories that make up a whole



The bar heights show the category counts or percents (the bar in alphabetical order).



In order of height



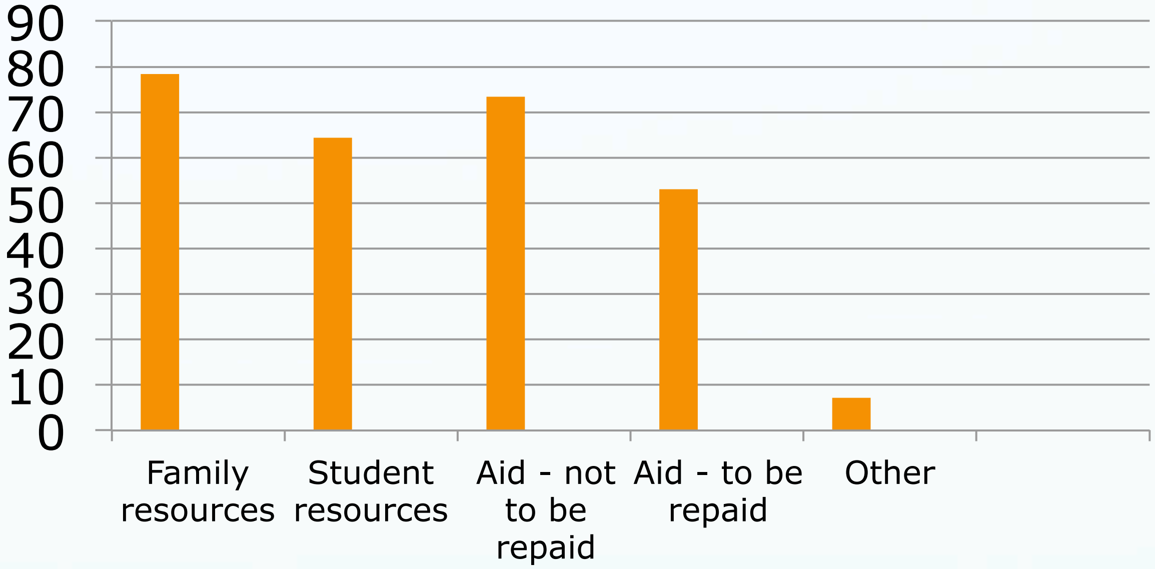
Example 1.4 (homework 1.4 in book)

The Higher Education Research Institute’s Freshman Survey reports the following data on the sources students use to pay for college expenses.

Source for college expenses	Students
Family resources	78,4%
Student resources	64,3%
Aid – not to be repaid	73,4%
Aid – to be repaid	53,1%
Other	7,1%

Why it is not correct to use a pie chart?

But we can build a bar graph for these data



Histograms

- Appropriate for quantitative variables that take many values and/or large datasets.
- Divide the possible values into classes (equal widths).
- Count how many observations fall into each interval (may change to percents).
- Draw picture representing the distribution—bar heights are equivalent to the number (percent) of observations in each interval.

Interpreting Histograms

EXAMINING A HISTOGRAM

- In any graph of data, look for the overall pattern and for striking deviations from that pattern.
- You can describe the overall pattern by its shape, center, and variability. You will sometimes see variability referred to as spread.
- An important kind of deviation is an outlier, an individual that falls outside the overall pattern.

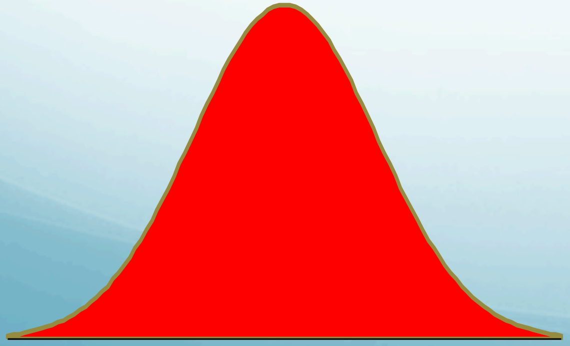
Describing Distributions

- A distribution is symmetric if the right and left sides of the graph are approximately mirror images of each other.

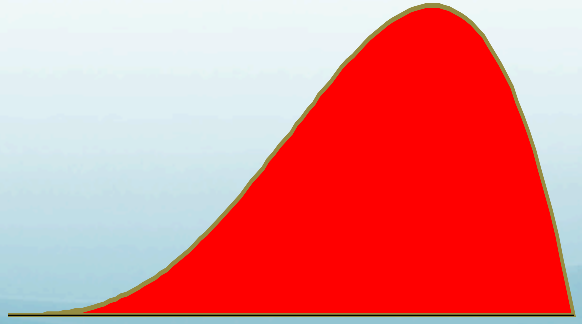
- A distribution is skewed to the right (right-skewed) if the right side of the graph (containing the half of the observations with larger values) is much longer than the left side.

- It is skewed to the left (left-skewed) if the left side of the graph is much longer than the right side.

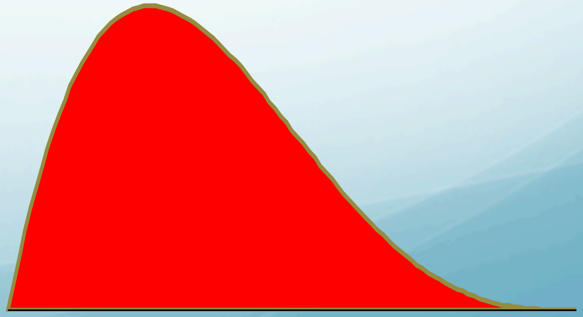
Symmetric

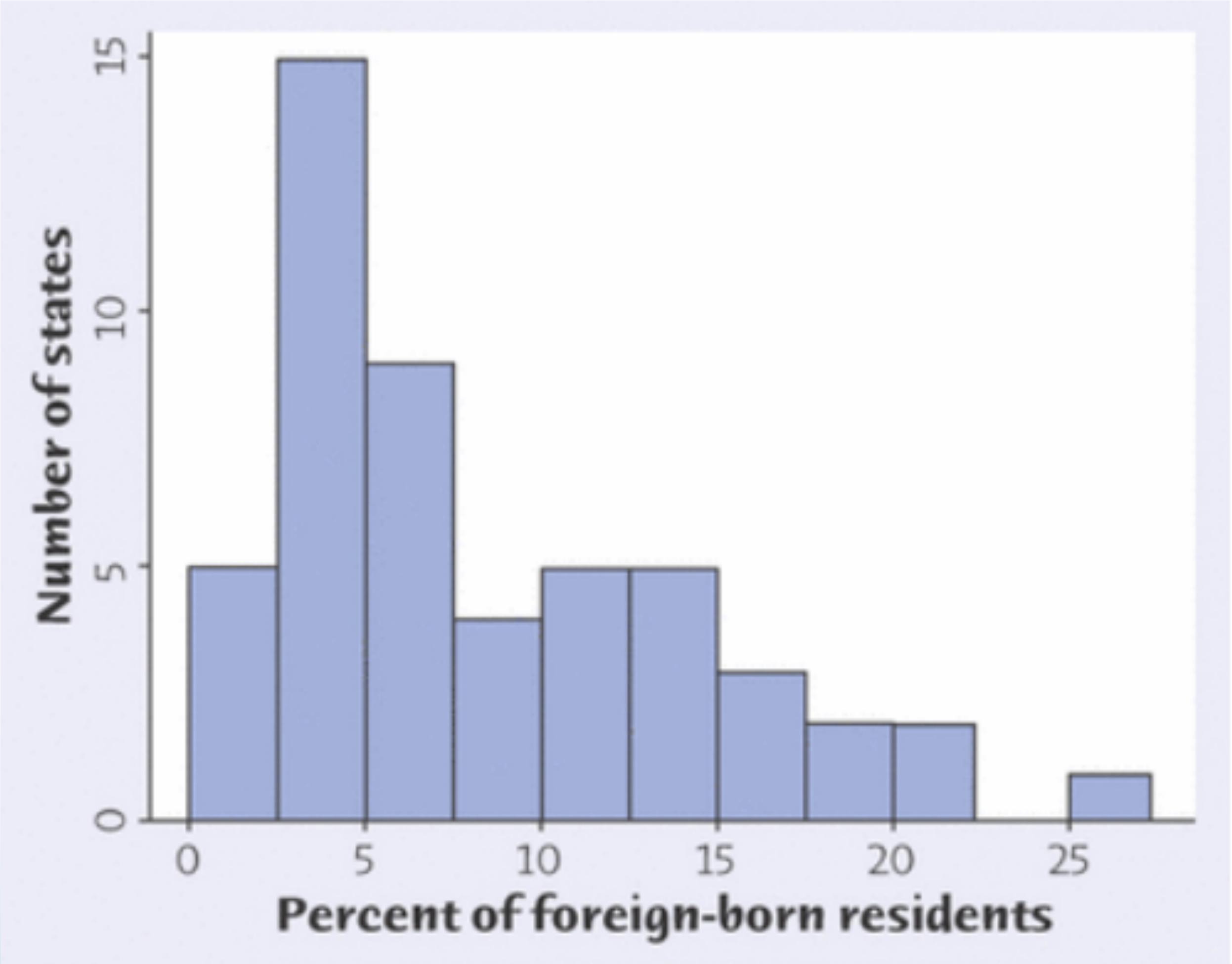


Skewed-left



Skewed-right




Number of states
15
10
5
0
0
5
10
15
20
25
Percent of foreign-born residents

Stemplots (Stem-and-Leaf Plots)

STEMPLOT

- To make a stemplot:

1. Separate each observation into a stem, consisting of all but the final (rightmost) digit, and a leaf, the final digit. Stems may have as many digits as needed, but each leaf contains only a single digit.
2. Write the stems in a vertical column with the smallest at the top, and draw a vertical line at the right of this column. Be sure to include all the stems needed to span the data, even when some stems will have no leaves.
3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.

Stemplots (Stem-and-Leaf Plots)

- *If there are very few stems* (when the data cover only a very small range of values), then we may want to create more stems by splitting the original stems.

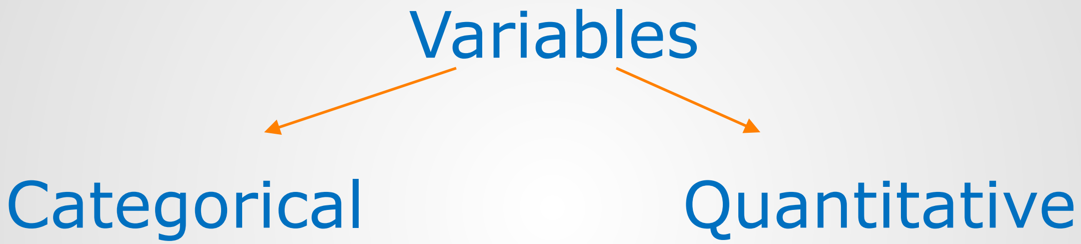
- Example: If all of the data values were between 150 and 179, then we may choose to use the following stems:

```
15 |
15 |
16 |
16 |
17 |
17 |
```

Leaves 0-4 would go on each upper stem (first “15”), and leaves 5-9 would go on each lower stem (second “15”).

Individuals – the objects described by a set of data (people, animals, things).

Variables – characteristics of an individual (it can take different values for different individuals).



Variables

Categorical

(places an indiv. into categories)

Quantitative

(takes numerical value, arithmetic operations make sense)

Two principles to organize our exploration of a set of data:

1. Exploring data (examining each var by itself, study the relationships among the var-s).

2. Distribution of a variable (what values it takes and how often it takes these values).

Variables

Categorical

Quantitative

Pie charts

(slices are sized by the counts or percents for the categories)
Use to emphasize each category’s relation to the whole

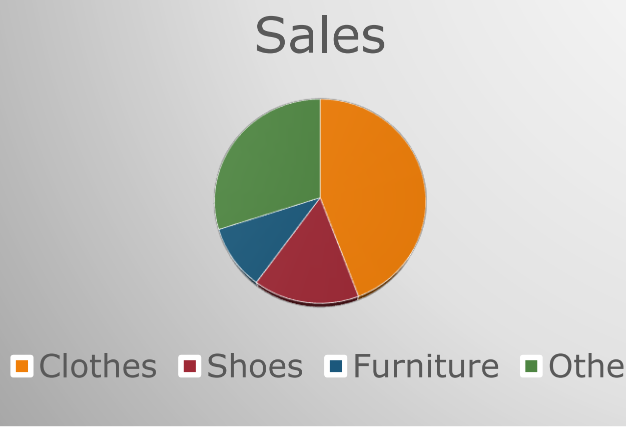
Bar graphs

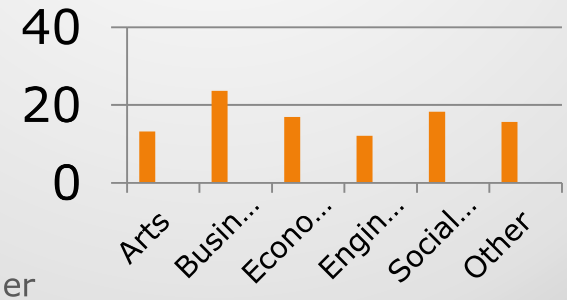
(represent each category as a bar. The bar heights show the category counts or percents)

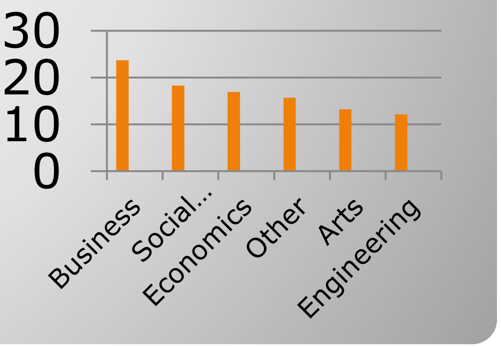
order

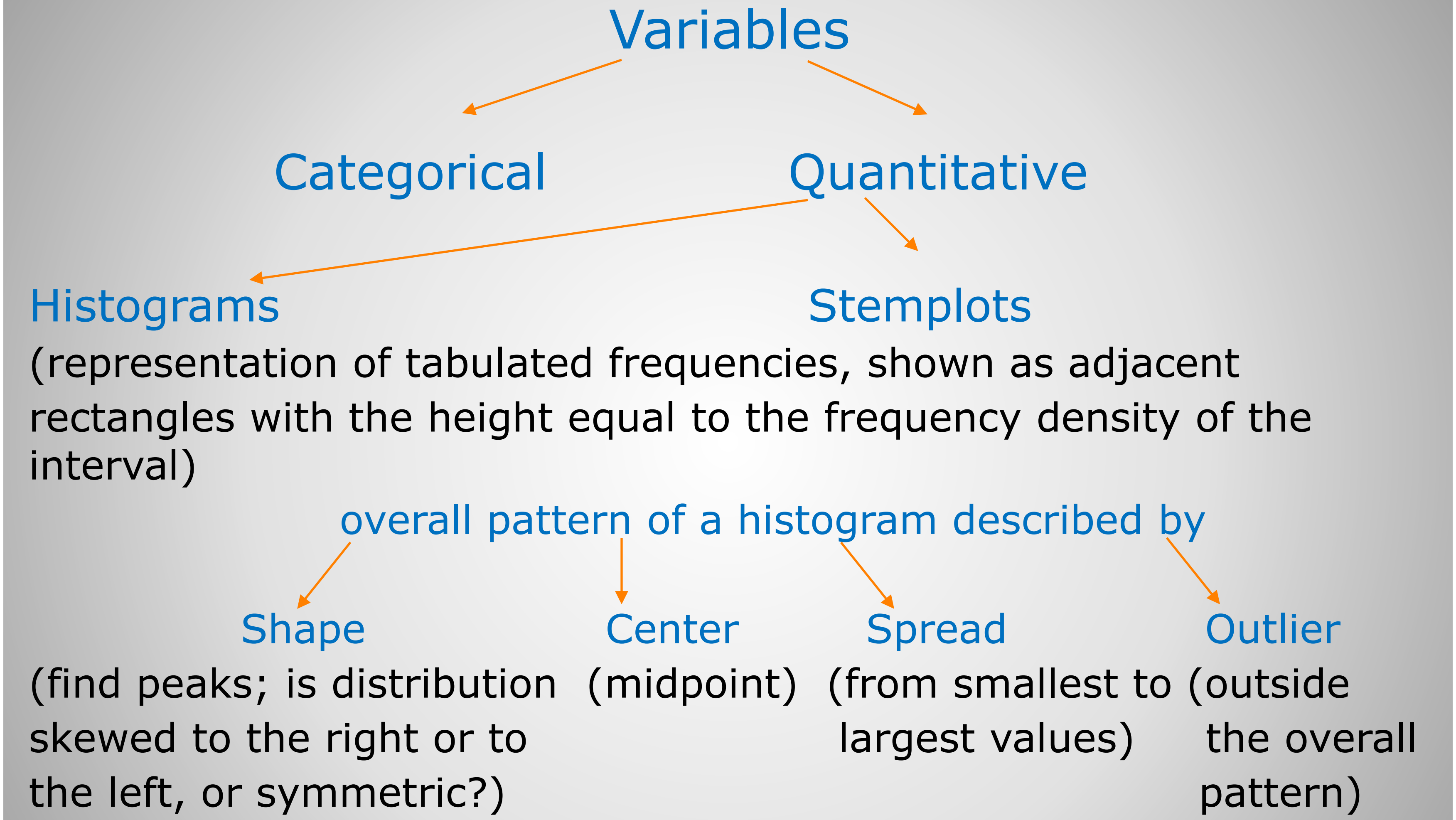
alphabetically

by height









Variables

Categorical

Quantitative

Histograms

(representation of tabulated frequencies, shown as adjacent rectangles with the height equal to the frequency density of the interval)

Stemplots

overall pattern of a histogram described by

Shape

(find peaks; is distribution skewed to the right or to the left, or symmetric?)

Center

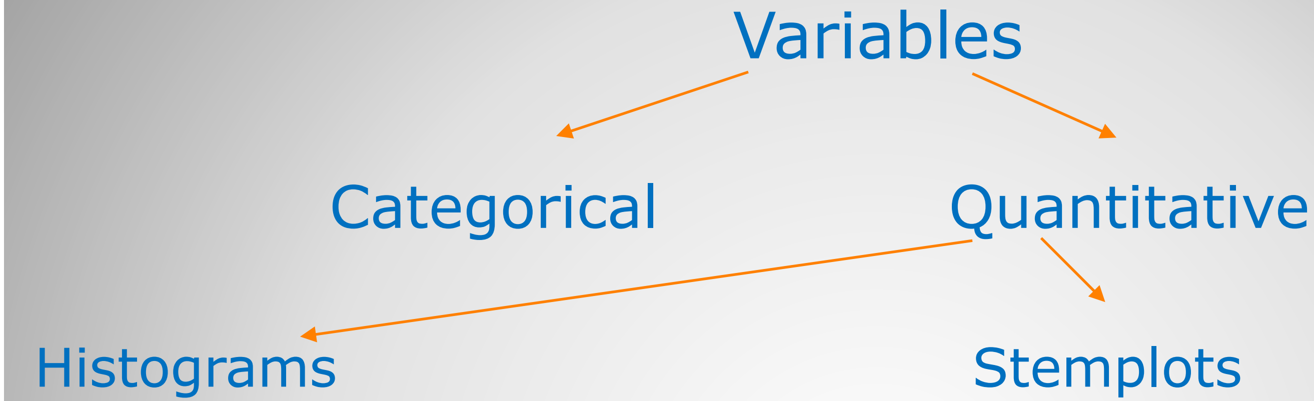
(midpoint)

Spread

(from smallest to largest values)

Outlier

(outside the overall pattern)



Variables

Categorical

Quantitative

Histograms

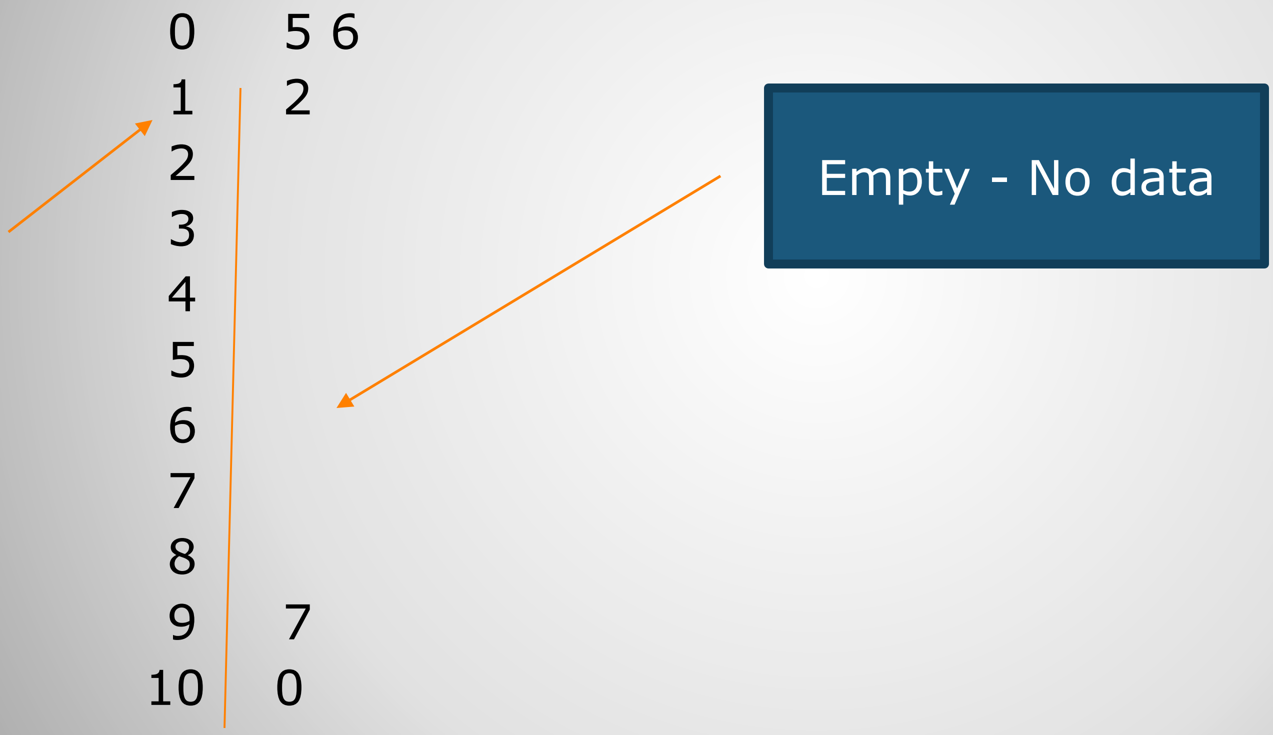
Stemplots

(Looks like a histogram turned on end but preserves the actual value of each observation. Arrange the leaves in ascending order. Don’t use stemplots for large data sets)

12	2
13	1 3
14	2 5 8
15	0 2 5 7 9
16	1 3 5 7
17	1 2

Stemplot

For the data: 6, 5, 12, 100, 97, 3213



0	5 6
1	2
2	
3	
4	
5	
6	
7	
8	
9	7
10	0

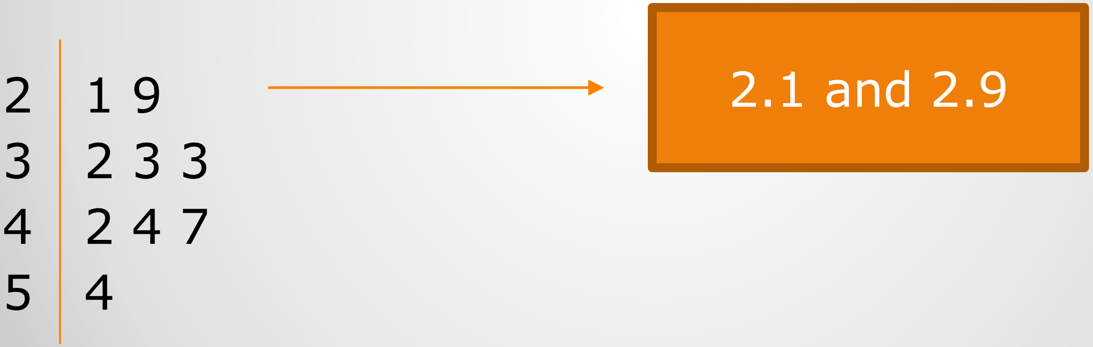
Empty - No data

Stemplot

For the data: 3.231 3.256 4.159 4.422 3.259 2.127 2.908 4.71 5.41

1) Round the data to the nearest hundred and then drop two 0s from the end of each rounded number (3.256 -> 3.300 -> 3.3)

3.2 3.3 4.2 4.4 3.3 2.1 2.9 4.7 5.4



```
2 | 1 9
3 | 2 3 3
4 | 2 4 7
5 | 4
```

2.1 and 2.9