Graphs and Describing Distributions

Section 1.5

Cathy Poliak, Ph.D.
cathy@math.uh.edu

Department of Mathematics
University of Houston

January 26, 2016

Outline

1. Graphs for Categorical Variables

2. Graphs for Quantitative Variables

3. Describing Distributions

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Data of basketball shoes

We will be demonstrating these graphs with this dataset called **shoes.**

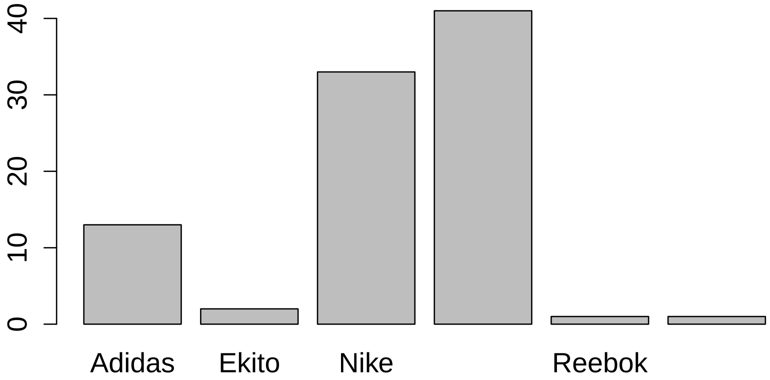
Name	Brand	Price
adiPower Howard 2	Adidas	75
adiZero Crazy Light	Adidas	90
adiZero Crazy Light 2	Adidas	140
⋮		
1 Flight	Nike Jordan	100
1 Flight Low	Nike Jordan	95
⋮		
Air Max CB34	Nike	110
Air Max Dominate	Nike	75
⋮		

Graphs of a Categorical Variable

- **Bar graphs**: Each individual bar represents a category and the height of each of the bars are either represented by the count or percent.
- **Pie charts**: Helps us see what part of the whole each group forms.

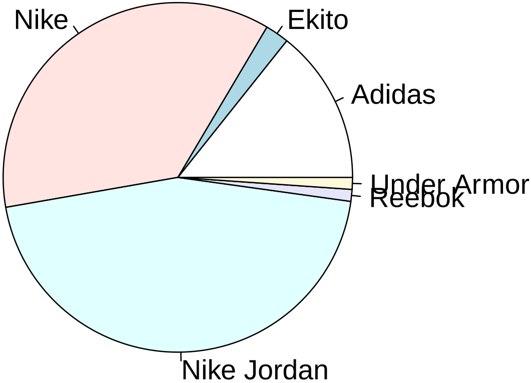
UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Bar graph



```
plot(shoes$Brand)
```

Pie Chart



R code

- For bar graph: plot(**datasetname$variablename**)
- For pie chart:

```
> counts<-table(shoes$Brand)
> pie(counts)
```

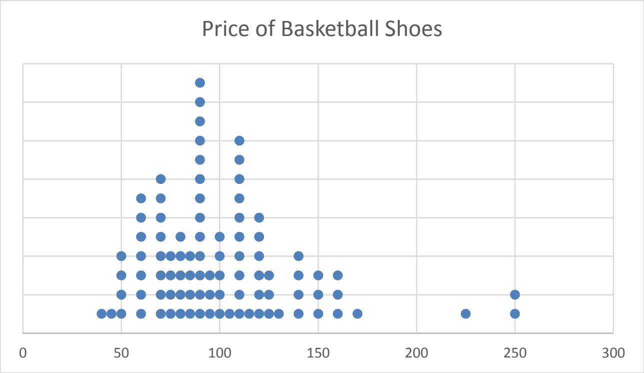
Graphs for quantitative variables

- Dotplot
- Stemplot
- Histogram
- Boxplot

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Dot plots

A **dot plot** is made by putting dots above the values listed on a number line.



Stem - and - leaf plot

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.
3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.

Rcode: stem(**dataset name**$**variable name**)

Example of Stem-and-leaf Plot

```
> stem(shoes$Price)

  The decimal point is 1 digit(s) to the right of the |

   4 | 050000
   6 | 0000000000000005555
   8 | 0000055550000000000000555
  10 | 00000500000000005
  12 | 0000005550
  14 | 0000000
  16 | 0000
  18 |
  20 |
  22 | 5
  24 | 00
```

Histograms

- Bar graph for quantitative variables.
- Values of the variable are grouped together.
- Bars are touching.
- The width of the bar represents an interval of values (range of numbers) for that variable.
- The height of the bar represents the number of cases within that range of values.

To create a histogram

1. Divide the range of data into classes of equal width. For example the price of the basketballs shoes are from \$40 to \$250 dollars. We can use a width of \$20 for the classes. Thus the classes are:

$$\begin{aligned} 40 \leq & \quad \text{price} & < 60 \\ 60 \leq & \quad \text{price} & < 80 \\ & \quad \vdots & \\ 240 \leq & \quad \text{price} & < 260 \end{aligned}$$

   Be sure to specify the classes precisely so that each individual price falls into exactly one class and all of the prices are counted.

2. Count the number of shoe prices in each class.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Counts of the classes

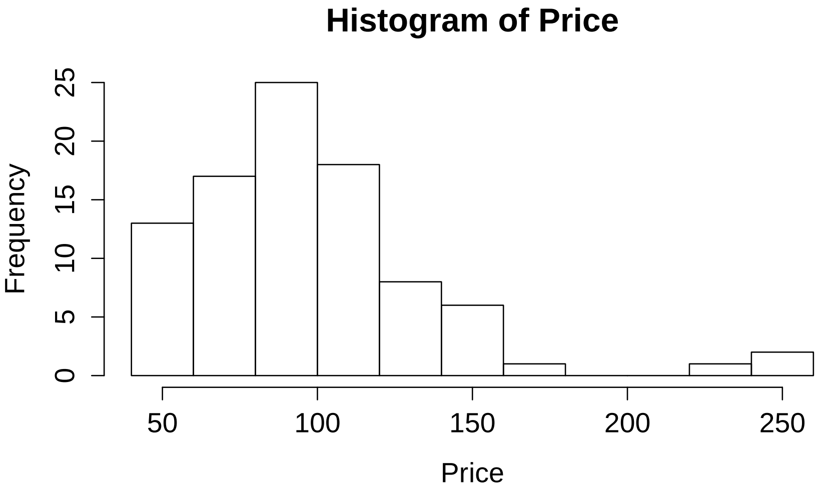
Price (Classes)	Count
$40 \leq \text{price} < 60$	6
$60 \leq \text{price} < 80$	19
$80 \leq \text{price} < 100$	25
$100 \leq \text{price} < 120$	17
$120 \leq \text{price} < 140$	10
$140 \leq \text{price} < 160$	7
$160 \leq \text{price} < 180$	4
$180 \leq \text{price} < 200$	0
$200 \leq \text{price} < 220$	0
$220 \leq \text{price} < 240$	1
$240 \leq \text{price} < 260$	2

Draw the histogram

1. Mark on the horizontal axis the scale for the variable whose distribution you are displaying.
2. The vertical axis contains the scale of the counts.
3. Each bar represents a class. The base of the bar covers the width of the classes, and the bar height is the class count. There is no horizontal space between bars unless a class is empty, so that its bar has height zero.

Rcode: hist(**dataset name**$**variable name**)

Histogram



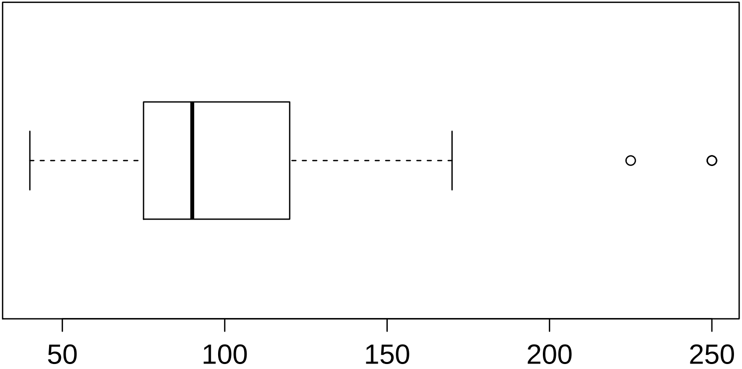
```
hist(shoes$Price, main = "Histogram of Price", xlab = "Price")
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Boxplot

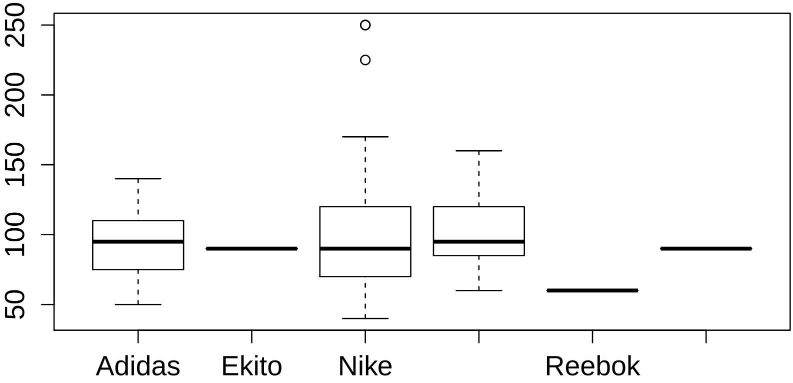
- A graph of the five-number summary.
  - A central box spans the quartiles.
  - A line inside the box marks the median.
  - Lines extend from the box out to the smallest and largest observations.
  - Asterisks represents any values that are considered to be outliers.
- Boxplots are most useful for side-by-side comparison of several distributions.
- Rcode: boxplot(**dataset name**$**variable name**)

Boxplot of Prices



```
boxplot(shoes$Price,horizontal = T)
```

Boxplot of Prices by Brand

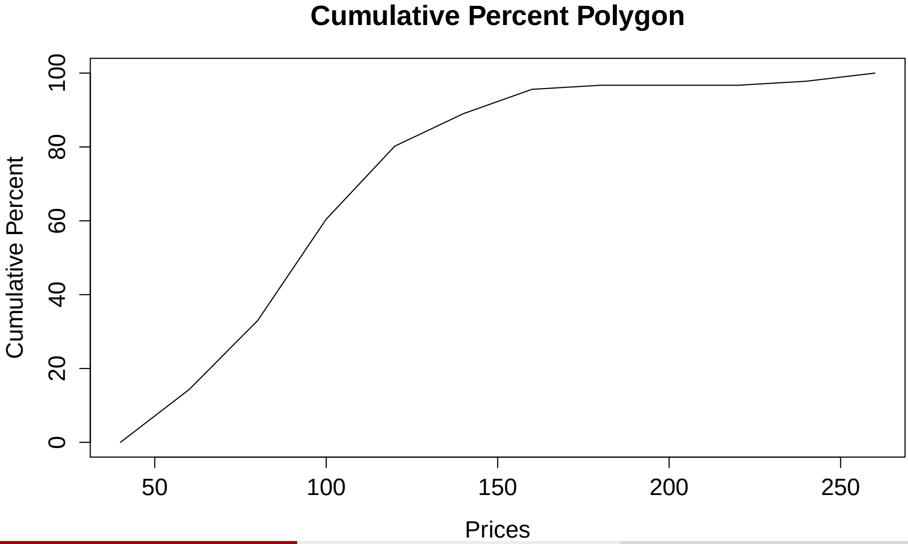


```
boxplot(shoes$Price~shoes$Brand)
```

Cumulative Frequency Polygon

- Plot a point above each upper class boundary at a height equal to the cumulative frequency of the class.
- Connect the plotted points with line segments.
- A similar graph can be used with the cumulative percents.

Cumulative Percent Polygon



Question about the Graphs

Given the first type of plot indicated in each pair, which of the second plots could not always be generated from it?

a) dot plot, histogram

b) stem and leaf, dot plot

c) histogram, stem and leaf

d) dot plot, box plot

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Distributions

- When observing a data set, one of the first things we want to know is how each variable is *distributed*.
- The **distribution** of a variable tells us what values it takes and how often it takes these values based on the individuals.
- The distribution of a variable can be shown through tables, graphs, and numerical summaries.

Distributions for categorical variables

- Lists the categories and gives either the count or the percent of cases that fall in each category.
- One way is a **frequency table** that displays the different categories then the count or percent of cases that fall in each category.
- Then we look at the graphs (bar or pie) to determine the distribution of a categorical variable.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Describing distributions of quantitative variables

- The **distribution** of a variable tells us what values it takes and how often it takes these values.
- There are four main characteristics to describe a distribution:
  1. Shape
  2. Center
  3. Spread
  4. Outliers

Describing distributions

- An initial view of the distribution and the characteristics can be shown through the graphs.
- Then we use numerical descriptions to get a better understanding of the distributions characteristics.

Describing a distribution

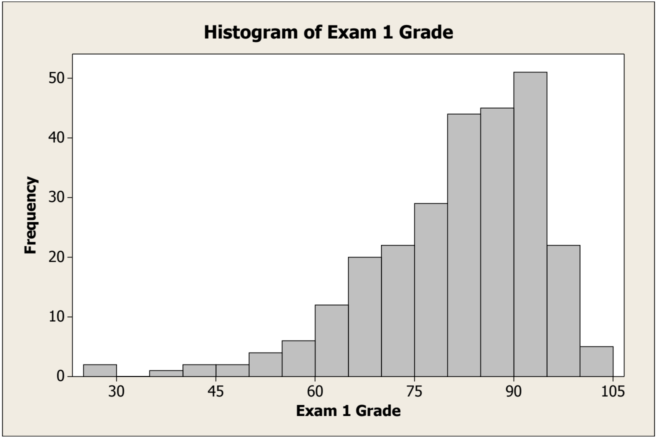
- Shape
  - A distribution is **symmetric** if the right and left sides of the graph are approximately mirror images of each other.
  - A distribution is **skewed to the right** if the right side (higher values) of the graph extends much farther out than the left side.
  - A distribution is **skewed to the left** if the left side (lower values) of the graph extends much farther out than the right side.
  - A distribution is **uniform** if the graph is at the same height (frequency) from lowest to highest value of the variable.
- Center - the values with roughly half the observations taking smaller values and half taking larger values.
- Spread -from the graphs we describe the spread of a distribution by giving *smallest and largest values*.
- Outliers - individual values that falls outside the overall pattern.

Distribution of the price of basketball shoes

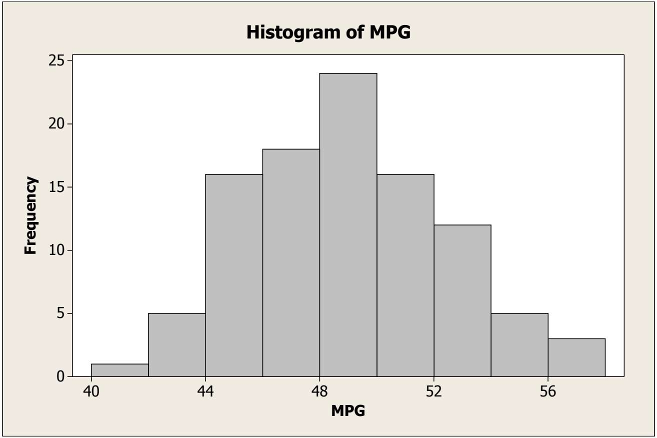
- Shape - longer tail on the upper values (right) **skewed right**.
- Center - Approximately \$90. Looking at the boxplot and from the histogram \$90 is at approximately half of the **area** of the graphs.
- Spread - The interval of values is from \$40 to \$250.
- From the three graphs we can see three distinct outliers.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Distribution of exam 1 grades



Distribution of prius hybrid MPG


Histogram of MPG
Frequency
MPG