

## Graphical methods of exploring distributions

### Quantitative Methods in Geography

Geo 340

William D. McCoy

Department of Geosciences  
University of Massachusetts, Amherst

Autumn, 2007

- ▶ stem-and-leaf display
- ▶ histogram
- ▶ probability density plot
- ▶ empirical cumulative distribution plot
- ▶ boxplot
- ▶ quantile plot
- ▶ quantile-quantile plot

### Numerical summaries of data

- ▶ Extremes
  - ▶ maximum
  - ▶ minimum
- ▶ Measures of central tendency
  - ▶ mean
  - ▶ median
  - ▶ mode
- ▶ Measures of spread or dispersion
  - ▶ variance
  - ▶ standard deviation
  - ▶ interquartile range
  - ▶ range

### Example dataset

**R** has many built-in datasets to use to try out functions. Data can be retrieved from **R** and placed in an object in your workspace using `data()`:

```
> data(cars)
```

This retrieves the dataset `cars` and places it in an object of the same name in your workspace.

## Extremes

We can easily find the extreme values in our data. To find the maximum value in a vector, use `max()`, and to find the minimum, use `min()`. Now you can find the extreme values in each of the two vectors in the `cars` dataset: `speed` and `dist`. Here are the extremes of `speed`:

```
> max(cars$speed)
```

```
[1] 25
```

```
> min(cars$speed)
```

```
[1] 4
```

You can try the same for `dist`. Note the use of the '\$' to select the speed component of the `cars` dataset.

## Measures of central tendency: mean

You can use the function `mean()` to find the average of a vector or any set of numbers given as arguments.

```
> mean(cars$dist)
```

```
[1] 42.98
```

The mean is defined as the sum of the elements of the vector divided by the number of elements:

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

The mean can be a misleading measure of central tendency for an asymmetric dataset.

## Properties of the mean

The mean has some special properties that are very important in data analysis. One such property is that the sum of the deviations of each value from the mean is zero:

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

Also, the sum of the squared deviations from the mean is a minimum, *i.e.* it is less than the sum of the squared deviations from any other number.

$$\sum_{i=1}^n (x_i - \bar{x})^2 < \sum_{i=1}^n (x_i - y)^2$$

where  $y \neq \bar{x}$ .

## Measures of central tendency: median

The median is the middle value of the sorted data. Use `median()` to find the median of your data.

```
> median(cars$dist)
```

```
[1] 36
```

The median is a very resistant measure of central tendency. That is to say, unlike the mean, the median is not strongly influenced by a few extreme values.

## Properties of the median

The median has the special property that the sum of the absolute values of the deviations is minimized by the median. That is to say the expression,

$$\sum_{i=1}^n |x_i - M|$$

is a minimum when  $M$  is the median.

## Measures of central tendency: mode

The mode is not well defined for continuous data. There is no **R** function that gives a single-valued result for the mode of a dataset. We will look at estimates of the mode when we discuss graphical methods.

## Measures of spread: variance

The variance is an important measure of spread or dispersion around the mean. It is defined as:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

The dimension of the variance is the square of the units of the data values, *e.g.* if the data values are in inches (*in*), the variance is square inches (*in*<sup>2</sup>). Use the function `var()` to find the variance of a vector.

```
> var(cars$dist)
[1] 664.0608
```

## Measures of spread: standard deviation

The standard deviation ( $s$ ) is the square root of the variance. Therefore, the standard deviation has the same dimensions as the original data values. Use `sd()` in **R** to find the standard deviation.

```
> sd(cars$dist)
[1] 25.76938
```

## Measures of spread: interquartile range

The interquartile range is the difference between the first and third quartiles. One-half (50%) of the data values lie within this range. That is to say, the interquartile range contains the middle half of the data values.

```
> IQR(cars$dist)
```

```
[1] 30
```

To see that this is the difference between the first and third quartiles, we can use the `summary()` function.

```
> summary(cars$dist)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.00	26.00	36.00	42.98	56.00	120.00

## Graphical methods: stem-and-leaf plot

The stem-and-leaf plot was developed by John Tukey. He developed many methods of exploratory data analysis (EDA) and wrote a book with that title. The stem-and-leaf display is a clever, semi-graphical display that has the advantage of showing the distribution of the data along with every data value.

```
> stem(cars$dist)
```

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

```
0 | 24004678
2 | 002466668822244466
4 | 002668024466
6 | 046806
8 | 04523
10 |
12 | 0
```

## Measures of spread: range

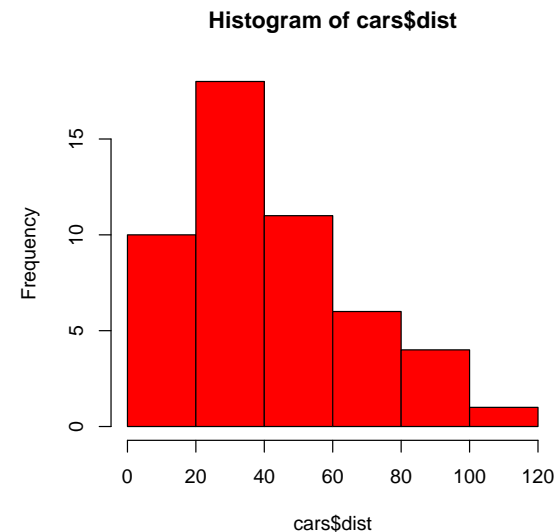
The range is a simple measure of spread. The range is the maximum data value minus the minimum data value,  $x_{max} - x_{min}$ . The range, of course, includes 100% of the data values. In **R**, the `range()` function returns the minimum and maximum data values. So if you take the difference of those values you have the range.

```
> diff(range(cars$dist))
```

```
[1] 118
```

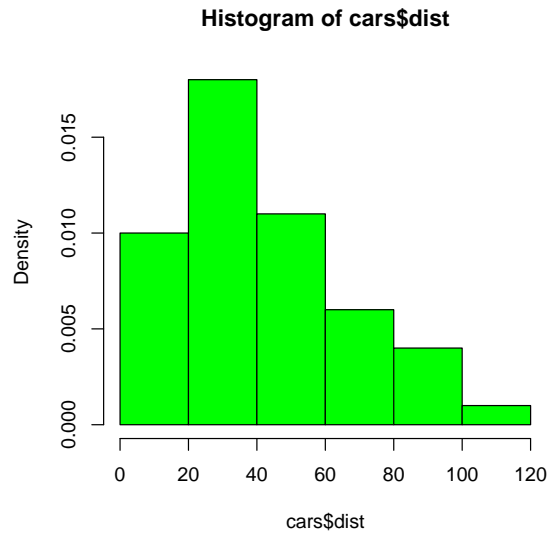
## Graphical methods: histogram

```
> hist(cars$dist, col = "red")
```



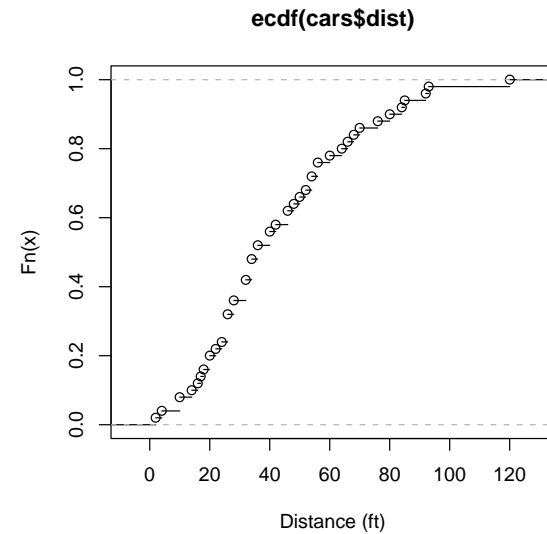
## Graphical methods: histogram

```
> hist(cars$dist, prob = TRUE, col = "green")
```



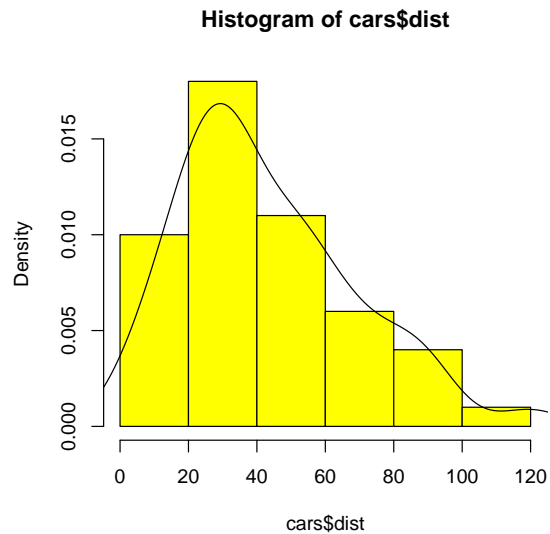
## Graphical methods: empirical cumulative distribution plot

```
> plot(ecdf(cars$dist), xlab = "Distance (ft)")
```



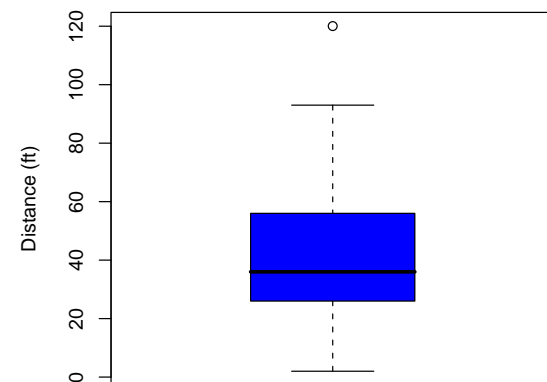
## Graphical methods: probability density

```
> hist(cars$dist, prob = TRUE, col = "yellow")  
> lines(density(cars$dist))
```



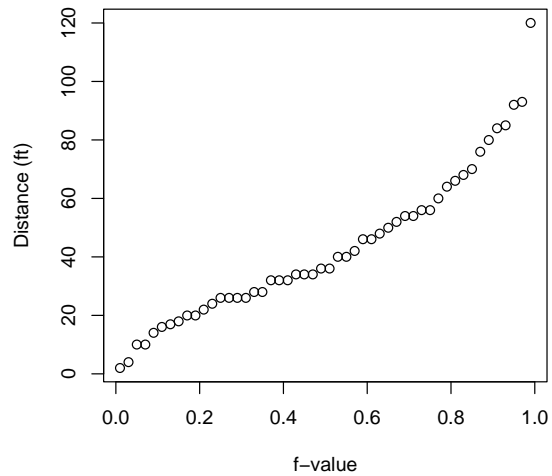
## Graphical methods: boxplot

```
> boxplot(cars$dist, ylab = "Distance (ft)",  
+ col = "blue")
```



## Graphical methods: quantile plot

```
> plot(ppoints(cars$dist), sort(cars$dist),  
+      ylab = "Distance (ft)", xlab = "f-value")
```



## Graphical methods: quantile-quantile (q-q) plot

```
> qqnorm(cars$dist)  
> qqline(cars$dist)
```

