

Random Sampling and data description

- **Recall:** we are looking at ways to summarize data
 - **Numerical summaries:**
 - measures of center
(mean, median, mode)
 - measures of spread
(sample variance, range, IQR)
 - **Graphical summaries:**
 - Stem and leaf plots
 - Histograms
 - Box Plots

6-1 Numerical Summaries

Definition: Sample Mean

If the n observations in a sample are denoted by x_1, x_2, \dots, x_n , the sample mean is

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n} \quad (6-1)$$

EX: # earthquakes of magnitude 7 or greater for years 1980-1990:

18, 14, 10, 15, 8, 15, 6, 11, 8, 7, 12, 11, 23, 16, 15, 25, 22, 20, 16, 23

$$\bar{x} = \frac{\sum_{i=1}^{20} x_i}{20} = \frac{18 + 14 + \dots + 23}{20} = 14.75$$

Definition: Median

First we need to order the data 6,7 ,8, 8, 10, 11,11,12,14, 15, 15, 15, 16, 16, 18, 20, 22, 23, 23, 25 and then choose the value that divides the data into 2 halves.

If n is even, then

$$\text{Median} = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}$$

If n is odd, then

$$\text{Median} = X_{\frac{n+1}{2}}$$

Ex: $n=20$ is even, so Median is $X_{10} + X_{11} = (15 + 15)/2 = 15$

Definition: Mode

The mode is the value that occurs the most frequently in a data set or a probability distribution

In our example, hence the mode is 15.

Remark:

The sample mean is affected by large values in the observations. Hence, if the data are highly skewed, it might not be the best measure to use. Instead, the median is a more robust measure, because it is always half way the data, no matter the value assumed by our observations.

Measures of spread or variability

Definition: Sample Variance

If x_1, x_2, \dots, x_n is a sample of n observations, the **sample variance** is

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \quad (6-3)$$

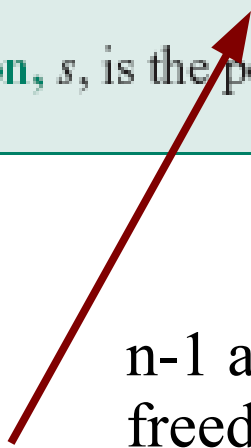
The **sample standard deviation**, s , is the positive square root of the sample variance.

Definition: Sample Variance

If x_1, x_2, \dots, x_n is a sample of n observations, the **sample variance** is

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \quad (6-3)$$

The **sample standard deviation**, s , is the positive square root of the sample variance.



$n-1$ are the degrees of freedom. We lose one degree of freedom for using the sample mean instead of the true mean

Definition: Sample Variance

If x_1, x_2, \dots, x_n is a sample of n observations, the **sample variance** is

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \quad (6-3)$$

The **sample standard deviation**, s , is the positive square root of the sample variance.

In our example, we obtain


$$s^2 = 32.72 \quad s = \sqrt{32.72} = 5.72$$

Definition: Five Number Summary

Min	Q_1	Median	Q_3	Max
6		15		25

Definition: Five Number Summary

Min	Q_1	Median	Q_3	Max
6	10.5	15	19	25



Q_1: First Quartile

is the median of the first $\frac{1}{2}$ of the data

Q_2: Second Quartile

is the median of the 2^{nd} $\frac{1}{2}$ of the data

Range:

$$R = \max - \min = 25 - 6 = 19$$

Interquartile Range:

$$IQR = Q_3 - Q_1 = 19 - 10.5 = 8.5$$

The interquartile range is less sensitive to the extreme values in the sample than is the ordinary sample range



We enter the data
using the STAT menu

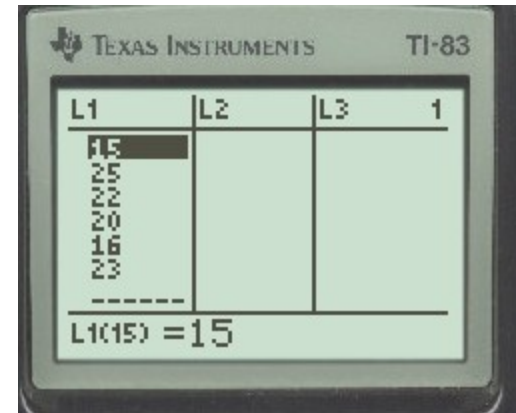
and then we press 1: Edit



We enter the data
using the STAT menu

and then we press 1: Edit

At the end we have
the data all in one
column



**Press STAT and > to display the choices for the
STAT CALC menu**

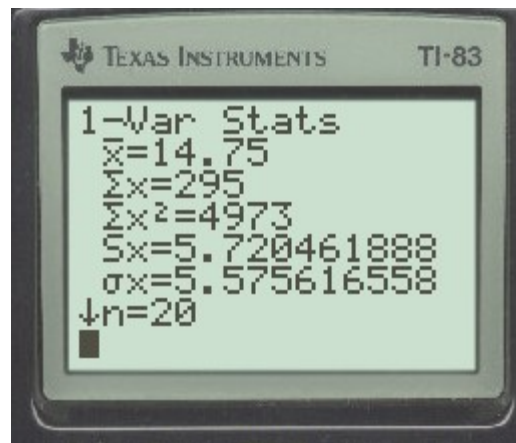


We are going to use **1-Var Stats**



We are going to use **1-Var Stats**

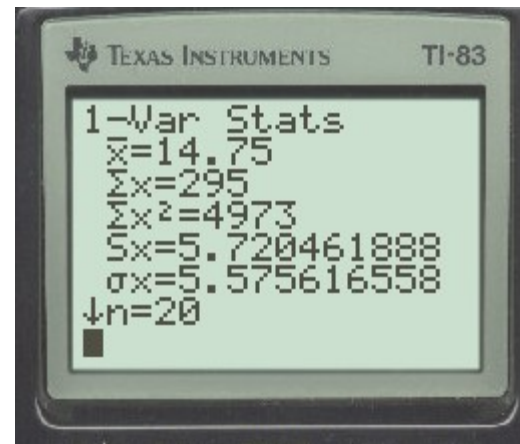
Mean, sum, sum of squares,
sample standard deviation $[\sqrt{n-1}]$,
population standard deviation $[\sqrt{n}]$



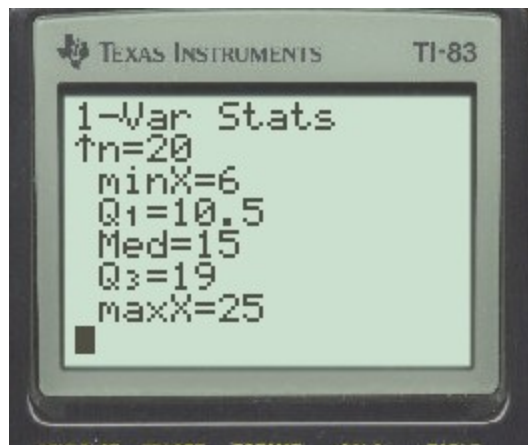


We are going to use **1-Var Stats**

Mean, sum, sum of squares,
sample standard deviation $[\sqrt{n-1}]$,
population standard deviation $[\sqrt{n}]$



Min, Q₁, Med, Q₃, max



Graphical summaries

6-2. Stem-and-Leaf Diagrams

A **stem-and-leaf diagram** is a good way to obtain an informative visual display of a data set x_1, x_2, \dots, x_n , where each number x_i consists of at least two digits. To construct a stem-and-leaf diagram, use the following steps.

Steps for Constructing a Stem-and-Leaf Diagram

- (1) Divide each number x_i into two parts: a **stem**, consisting of one or more of the leading digits and a **leaf**, consisting of the remaining digit.
- (2) List the stem values in a vertical column.
- (3) Record the leaf for each observation beside its stem.
- (4) Write the units for stems and leaves on the display.

Ex: earthquakes in 1980-1999

0		6788
1		01124
1		555668
2		0233
2		5

Ex: earthquakes in 1980-1999

0		6788
1		01124
1		555668
2		0233
2		5

Ex: waiting time in minutes for 20 patients at a public health clinic.

16 45 16 54 15 49 12 54 91 21 33 20
27 53 24 46 39 31 41 27

1		2566
2		01477
3		139
4		1569
5		344
6		
7		
8		
9		1

N.B. It's easy to produce stems of different length by rearranging the leaves around the stems....

```
0 | 2566
2 | 01477139
4 | 1569344
6 |
8 | 1
```

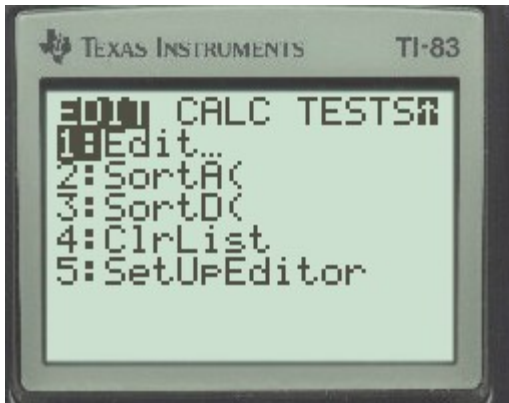
```
1 | 2566
2 | 01477
3 | 139
4 | 1569
5 | 344
6 |
7 |
8 |
9 | 1
```

```
1 | 2
1 | 566
2 | 014
2 | 77
3 | 13
3 | 9
4 | 1
4 | 569
5 | 344
5 |
6 |
6 |
7 |
7 |
8 |
8 |
9 | 1
```

But the overall impression about the distribution might change with it (so, be careful in interpreting these plots)

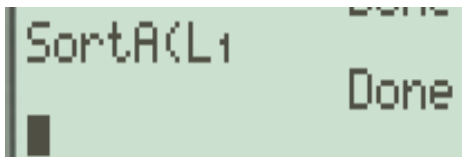
TI 83/84 – how to create stem and leaf plots

We order the data with sortA (sort ascending)



L1	L2	L3	1
6	-----	-----	
7			
8			
8			
10			
11			
11			

L1(1)=6



and then rearrange the data on the basis of the ordered values

Some links:

<http://www.andrews.edu/~calkins/math/webtexts/stat08.htm>

http://wind.cc.whecn.edu/~pwildman/statnew/new_page_13.htm

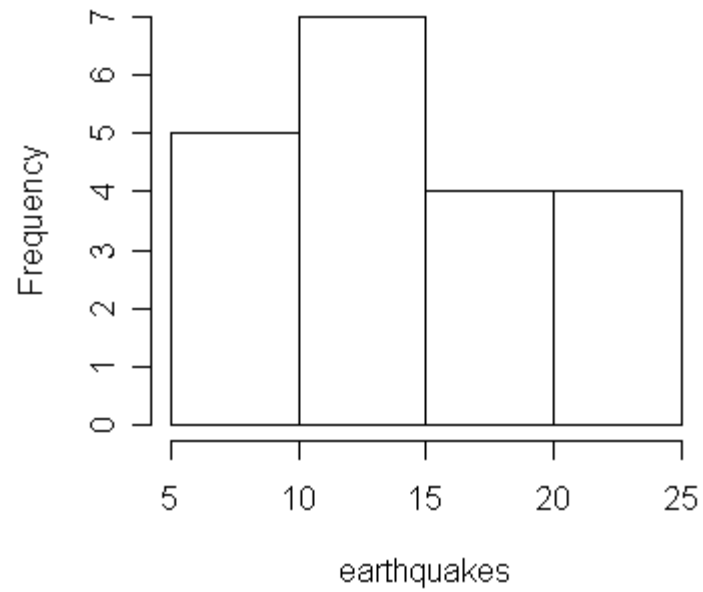
6-3 Frequency Distributions and Histograms

- A **frequency distribution** is a more compact summary of data than a stem-and-leaf diagram.
- To construct a frequency distribution, we must divide the range of the data into intervals, which are usually called **class intervals**, **cells**, or **bins**.

Constructing a Histogram (Equal Bin Widths):

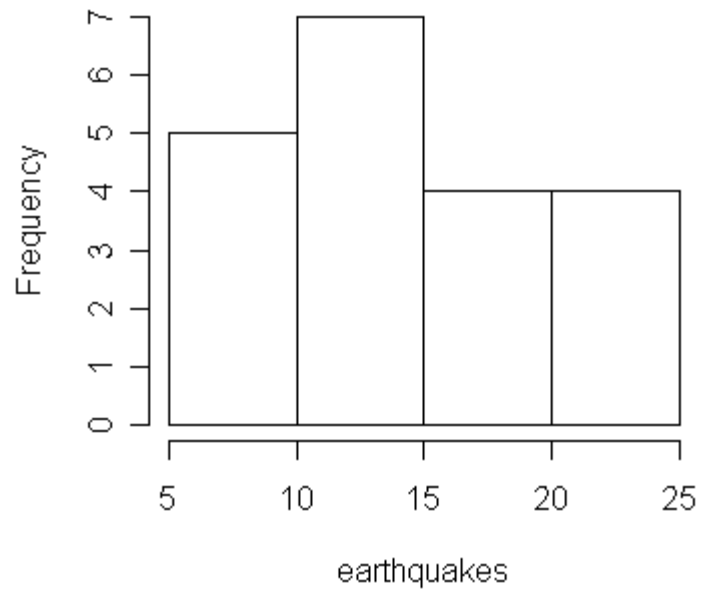
- (1) Label the bin (class interval) boundaries on a horizontal scale.
- (2) Mark and label the vertical scale with the frequencies or the relative frequencies.
- (3) Above each bin, draw a rectangle where height is equal to the frequency (or relative frequency) corresponding to that bin.

Ex: earthquakes in 1980-1999

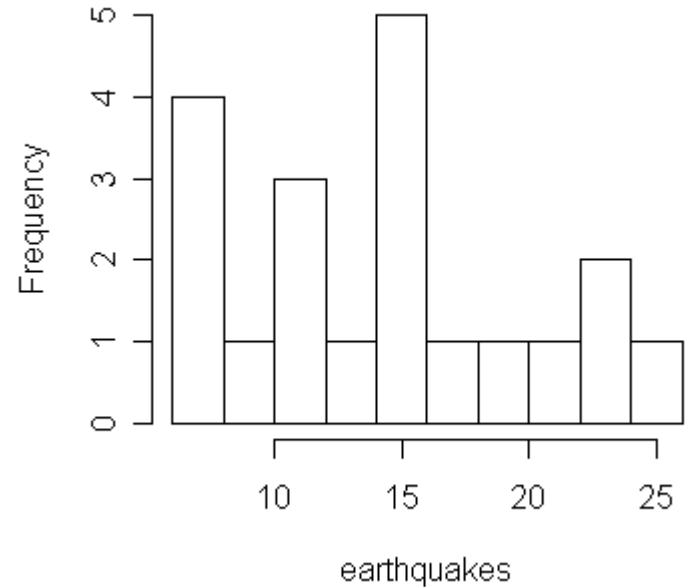


with 5 classes

Ex: earthquakes in 1980-1999

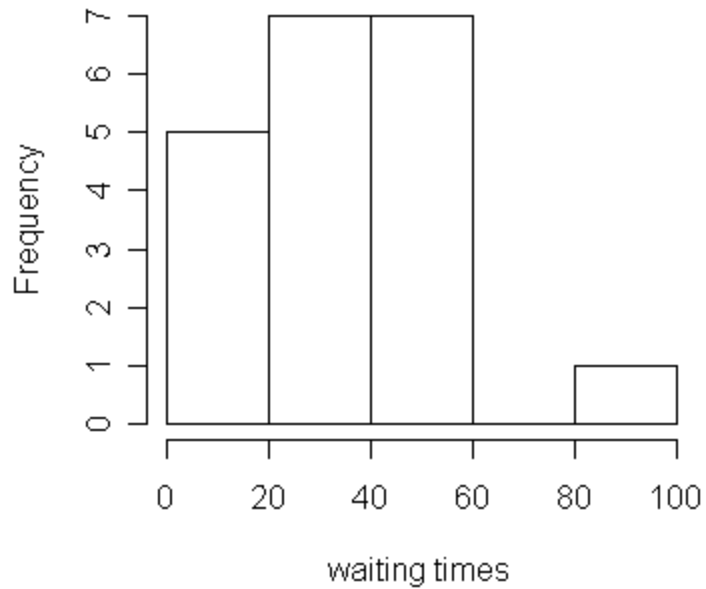


with 5 classes

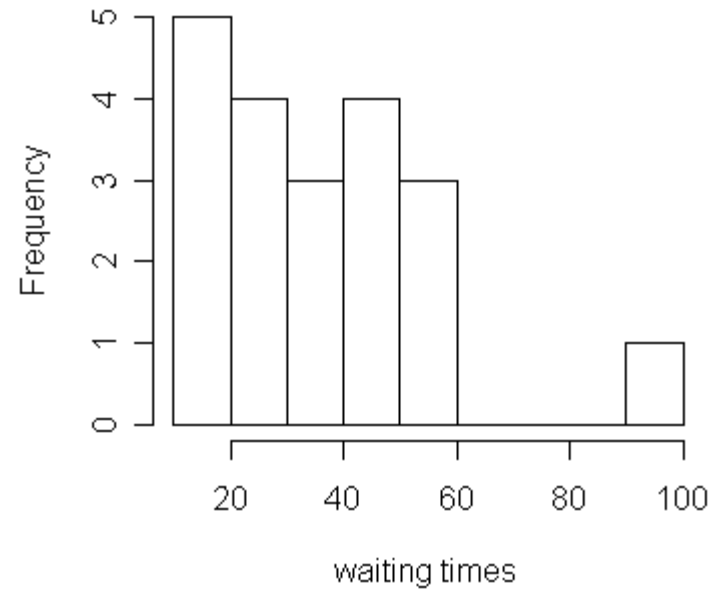


with 10 classes

Ex: waiting time in minutes for 20 patients at a public health clinic.

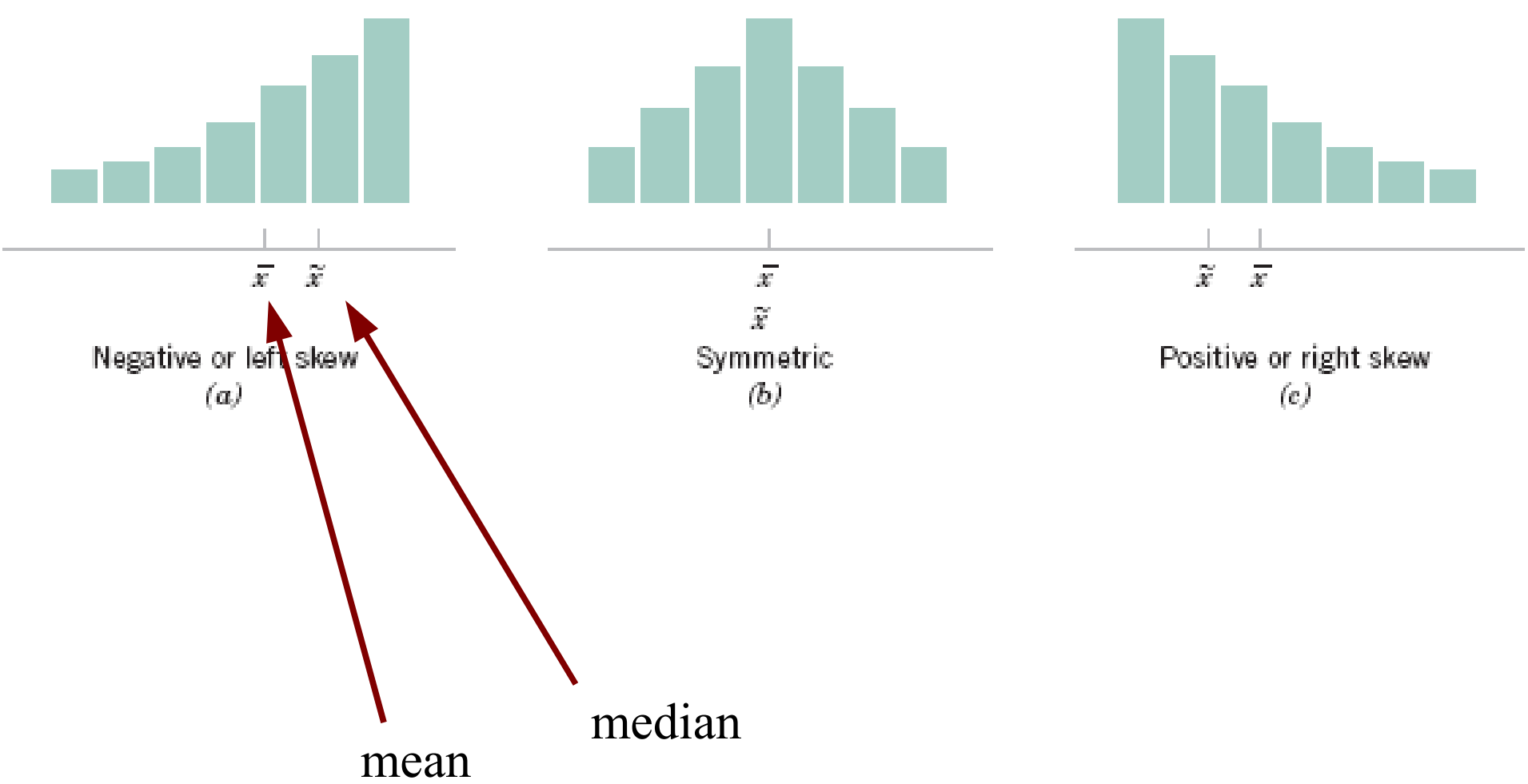


with 5 classes



with 10 classes

Figure 6-11 Histograms for symmetric and skewed distributions.

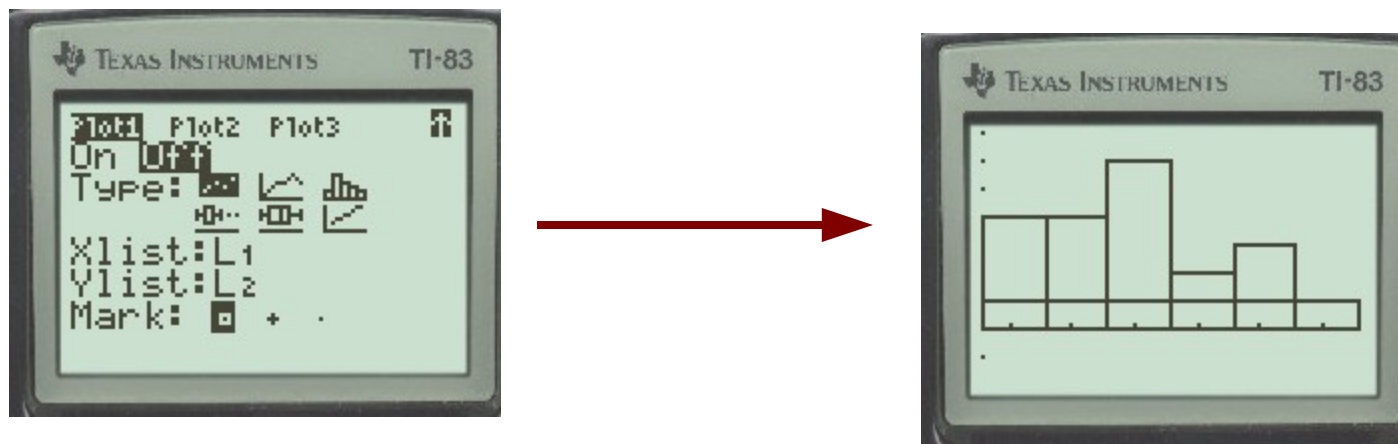


6-4 Box Plots

- The **box plot** is a graphical display that **simultaneously** describes several important features of a data set, such as center, spread, departure from symmetry, and identification of observations that lie unusually far from the bulk of the data.
- **Whisker**
- **Outlier**
- **Extreme outlier**

TI 83/84 – how to create histograms

There is the possibility to create histograms in the Stat plot screen: to access the screen hit 2nd function Y (stat plot), then hit 1 for Plot 1



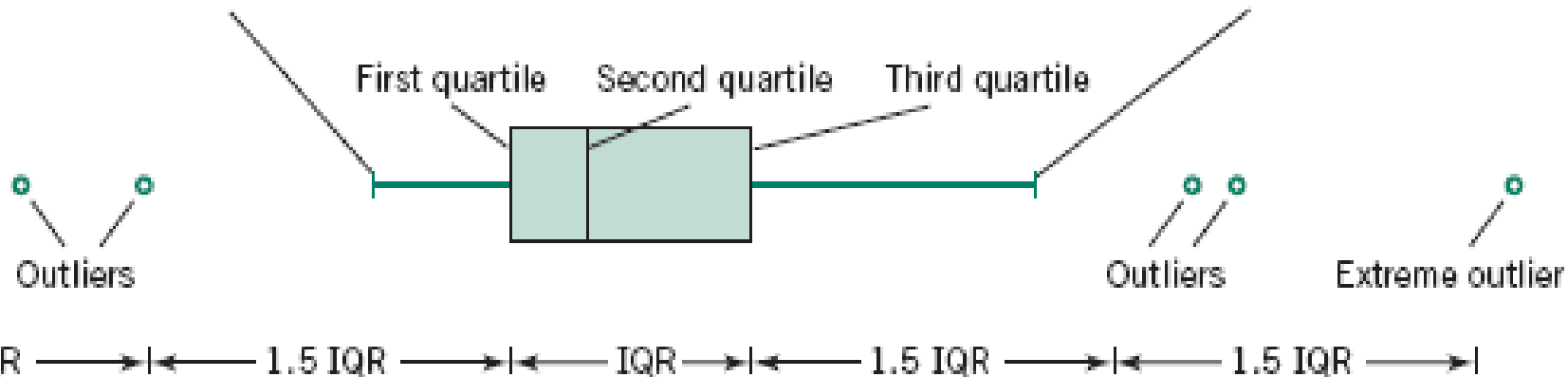
See complete description, for example, at

<http://wind.cc.whcn.edu/~pwildman/stat/ti83/tihistogram/TIHistograms.html>

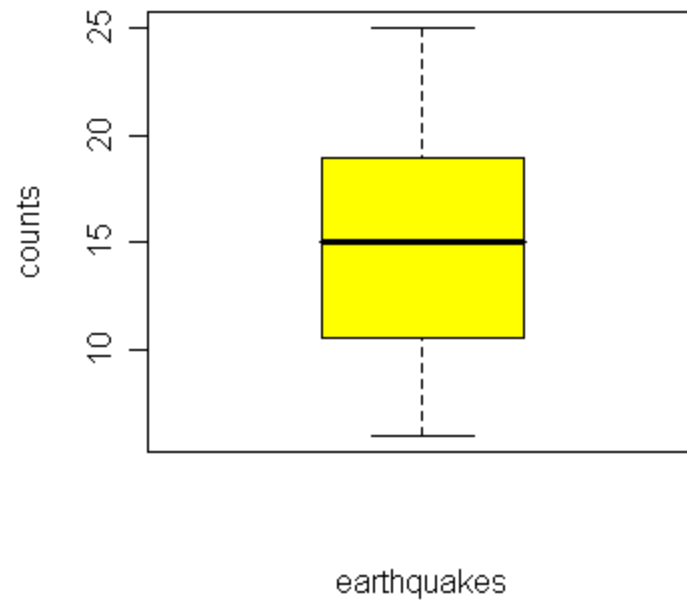
Box Plots

Whisker extends to smallest data point within 1.5 interquartile ranges from first quartile

Whisker extends to largest data point within 1.5 interquartile ranges from third quartile



Ex: earthquakes in 1980-1999



```

earthquake<-c(18,14,10,15,8,15,6,11,8,7,12,11,23,16,
15,25,22,20,16,23)
> sort(earthquake) #sort data from smallest to largest
[1]  6  7  8  8 10 11 11 12 14 15 15 15 16 16 18
   20 22 23 23 25
> fivenum(earthquake)
[1]  6.0 10.5 15.0 19.0 25.0

```

$$Q_1 = 10.5, Q_2 = 15, Q_3 = 19$$

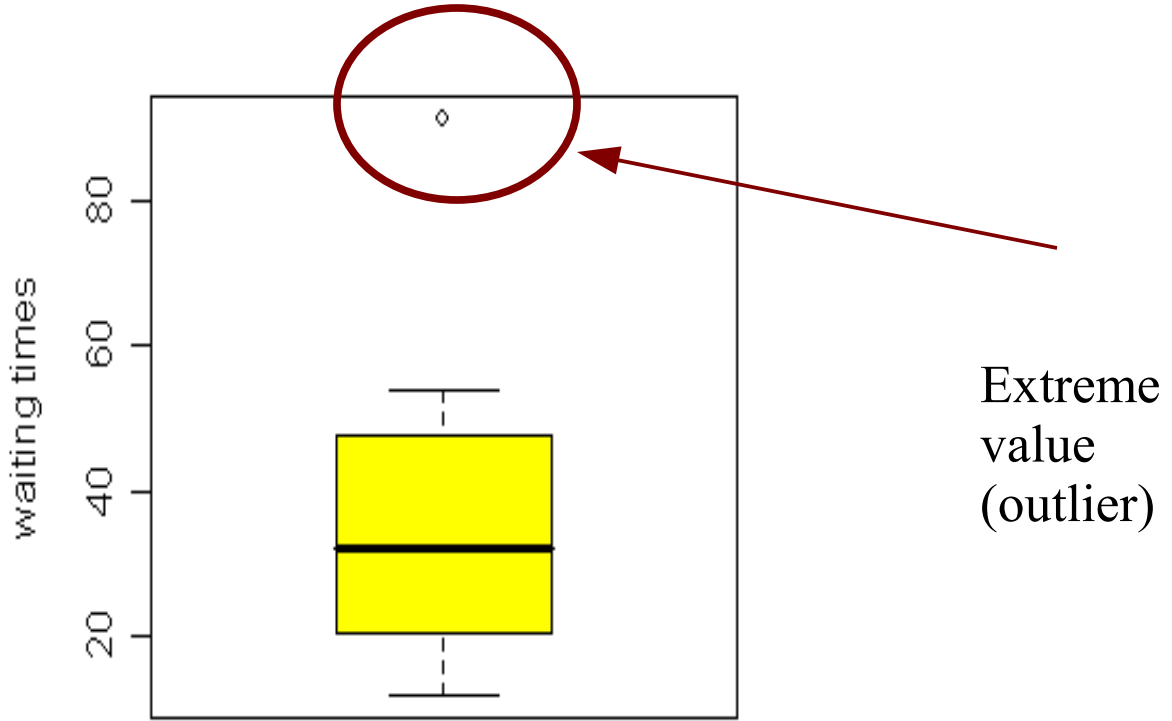
$$IQR = 19 - 10.5 = 8.5$$

$$Q_1 - 1.5 * IQR = 10.5 - 1.5 * 8.5 = -2.25$$

$$Q_3 + 1.5 * IQR = 19 + 1.5 * 8.5 = 31.75$$

No data point is less than -2.25 or greater than 31.75, therefore, no outlier is detected.

Ex: waiting time in minutes for 20 patients at a public health clinic.



```

> wt<-c(16,45,16,54,15,49,12,54,91,21,33,20,27,53,24,46,
39,31,41,27)
> sort(wt)
 [1] 12 15 16 16 20 21 24 27 27 31 33 39 41 45
    46 49 53 54 54 91
> fivenum(wt)
 [1] 12.0 20.5 32.0 47.5 91.0

```

$$Q_1 = 20.5, Q_2 = 32, Q_3 = 47.5$$

$$IQR = 47.5 - 20.5 = 27$$

$$Q_1 - 1.5 * IQR = 20.5 - 1.5 * 27 = -20$$

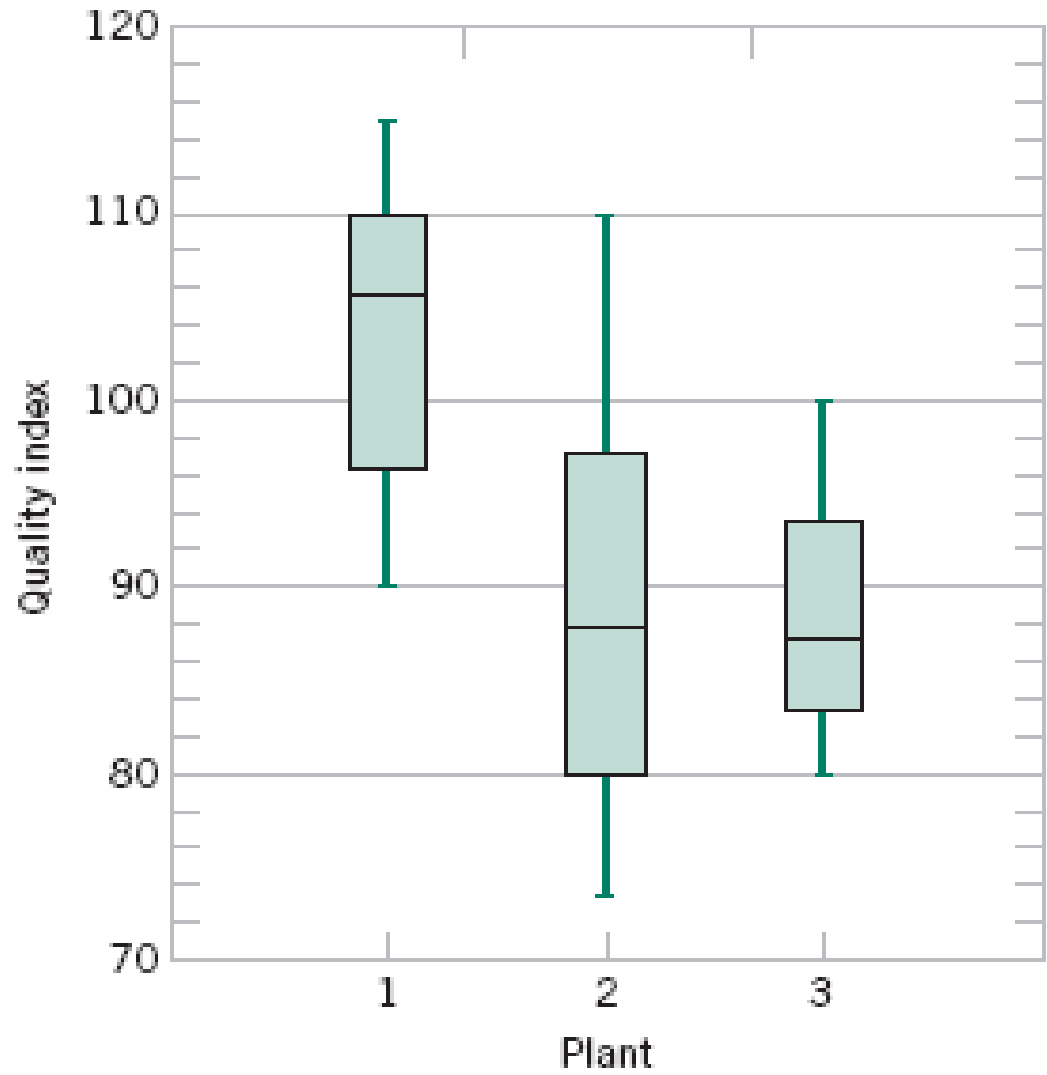
$$Q_3 + 1.5 * IQR = 47.5 + 1.5 * 27 = 88$$

$$Q_3 + 3 * IQR = 47.5 + 3 * 27 = 128.5$$

- ▶ No data point is less than -20.
- ▶ Data point 91 is greater than 88 but less than 128.5. Therefore, it is considered as an outlier but not an extreme outlier.

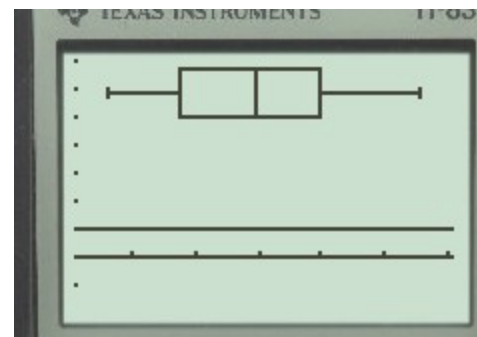
Figure 6-15

Comparative box plots of a quality index at three plants.



TI 83/84 – how to create box plots

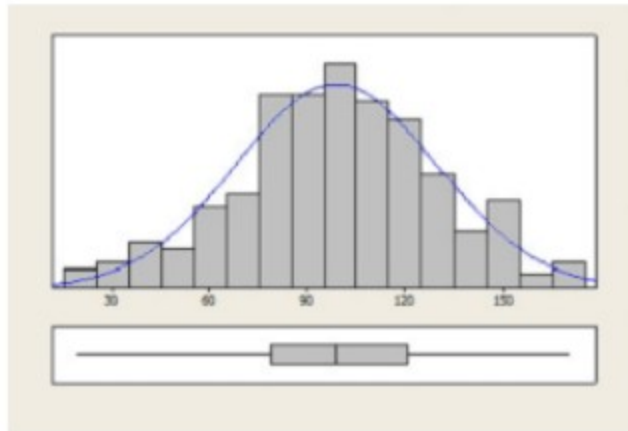
We use again the STAT PLOTS menu.



For more details, see for example:

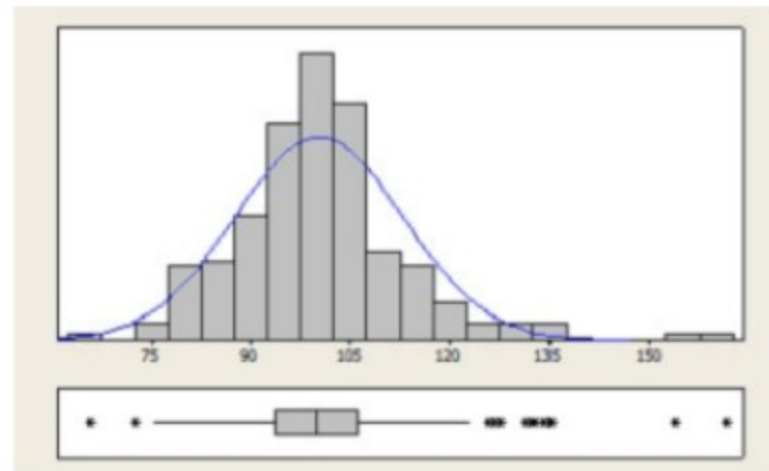
http://faculty.kutztown.edu/schaeffe/Tutorials/Statistics/Box_Plot/Box-Plot.html

Distributional shapes

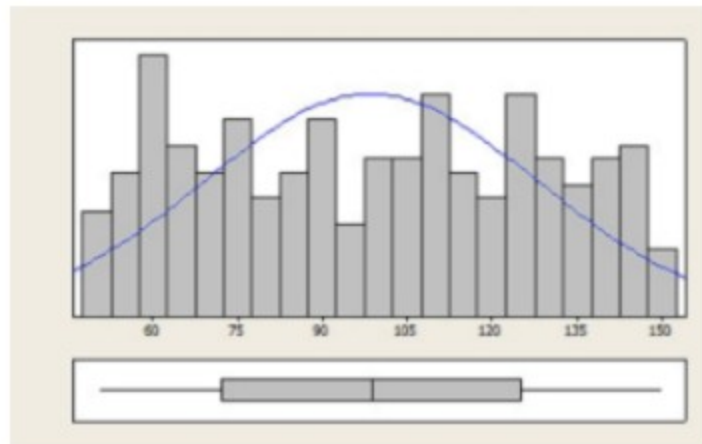


The distribution is **unimodal, symmetric and bell-shaped** (“normal”).

The distribution is **unimodal, symmetric and heavy-tailed**



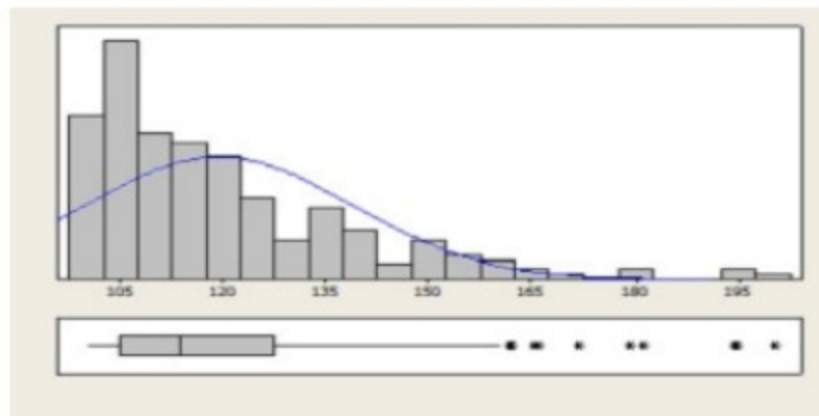
Distributional shapes



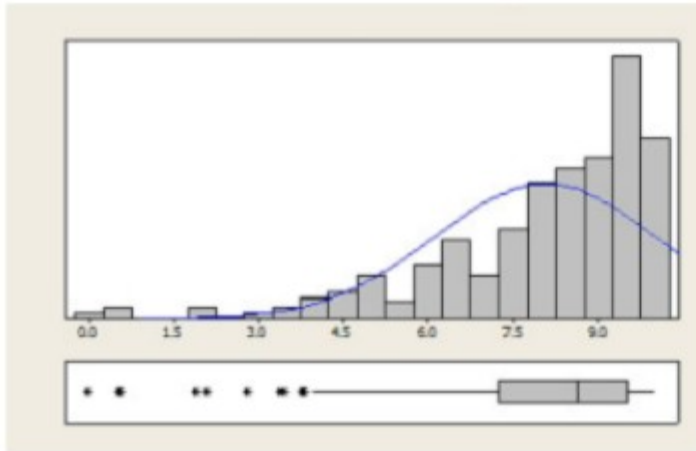
The distribution is **symmetric**,
but not bell-shaped.

The boxplot shows symmetry, but the tails
of the distribution are shorter (lighter)
than in the normal distribution.

The distribution is **skewed to the
right**, because the right tail is
much longer than the left tail.



Distributional shapes



The distribution is **skewed to the left**, because the left tail is much longer than the right tail.

The distribution is **bimodal**.

