Chapter 6: Descriptive statistics

Statistics is a science of data. An important aspect of dealing with data is organizing and summarizing the data in ways that facilitate its interpretation and subsequent analysis. This aspect of statistics is called descriptive statistics. In statistics and quantitative research methodology, a data sample is a set of data collected and/or selected from a statistical population by a defined procedure.

1 Measure of center

Suppose we observe n subjects in a study and for each subject, we observe one variable x.

Sample mean: Denote the n observations for a variable in a sample as $x_1, \ldots, x_n$, the sample mean is

$$\bar{x} = \frac{x_1 + x_2 + \ldots x_n}{n} = \sum_{i=1}^{n} x_i/n.$$

Sample Median: The median of a finite list of numbers can be found by arranging all the observations from lowest value to highest value and picking the middle one. If there is an even number of observations, then there is no single middle value; the median is then usually defined to be the mean of the two middle values.

Sample Mode: the most common data point.

Example 1.1. The number of earth quakes of magnitude 7 or greater for years 1980-1999 is 18, 14, 10, 15, 8, 15, 6, 11, 8, 7, 12, 11, 23, 16, 15, 25, 22, 20, 16, 23. For the above data, find the sample mean, median and mode.

2 Measure of spread/variablity

Sample variance and sample standard deviation: if the n observations in a sample are denoted by $x_1, \ldots, x_n$, the sample variance is

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

The sample standard deviation is

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

.

Sample range: if the n observations in a sample are denoted by $x_1, \ldots, x_n$, the sample range is

$$r = \max\{x_i, i = 1, \ldots, n\} - \min\{x_i, i = 1, \ldots, n\}$$

Sample Quartiles: the quartiles of a ranked set of data values are the three points that divide the data set into four equal groups, each group comprising a quarter of the data. The first quartile is often denoted as Q_1 and the third quartile is often denoted as Q_3. Q_1 is also the median of the first half of the data and Q_3 is also the median of the second half of the data. **How to calculate quartiles?**

(a) Arrange the observations in increasing order and locate the median M in the ordered list of observations.

(b) The first quartile Q_1 is the median of the observations whose position in the ordered list is to the left of the location of the overall median.

(c) The third quartile Q_3 is the median of the observations whose position in the ordered list is to the right of the location of the overall median.

Five number summary: min, Q_1, Median, Q_3, max.

Inter quartile range=IQR=Q_3-Q_1

Example 2.1. The number of earth quakes of magnitude 7 or greater for years 1980-1999 is 18, 14, 10, 15, 8, 15, 6, 11, 8, 7, 12, 11, 23, 16, 15, 25, 22, 20, 16, 23. For the above data, find the sample variance, sample range and quartiles.

3 Numerical summaries using R

Example 3.1. The number of earth quakes of magnitude 7 or greater for years 1980-1999 is 18, 14, 10, 15, 8, 15, 6, 11, 8, 7, 12, 11, 23, 16, 15, 25, 22, 20, 16, 23.

R command to input the data:

```
Earthquake <- c(18, 14, 10, 15, 8, 15, 6, 11, 8, 7, 12, 11, 23, 16, 15, 25, 22,
20, 16, 23)
Earthquake
 [1] 18 14 10 15  8 15  6 11  8  7 12 11 23 16 15 25 22 20 16 23
```

```
mean(Earthquake)
[1] 14.75
```

```
median(Earthquake)
[1] 15
```

R command to get the frequencies of data:

```
table(Earthquake) #Mode is then 14.
data
 6  7  8 10 11 12 14 15 16 18 20 22 23 25
 1  1  2  1  2  1  1  3  2  1  1  1  2  1
```

```
var(Earthquake)
32.72368
```

$s = \sqrt{32.72368} = 5.72.$

```
quantile(Earthquake)
 0%    25%    50%    75%   100%
 6.00 10.75 15.00 18.50 25.00
```

4 Graphical displays of data

4.1 Histogram

Quantitative variables often take many values. The distribution tells us what values the variable takes and how often it takes these values. A graph of the distribution is clearer if nearby values are grouped together. The most common graph of the distribution of one quantitative variable is a histogram.

Steps to construct a histogram:

(1) Choose the classes. Divide the range of the data into classes of equal width.

(2) Count the individuals in each class.

(3) Draw the histogram. Mark the scale for the variable whose distribution you are displaying on the horizontal axis.

Example 4.1. What percent of your home state's residents were born outside the United States? The country as a whole has 12.5% foreign-born residents, but the states vary from 1.2% in West Virginia to 27.2% in California. The following table presents the data for all 50 states and the District of Columbia. The individuals in this data set are the states. The variable is the percent of a state's residents who are foreign-born.

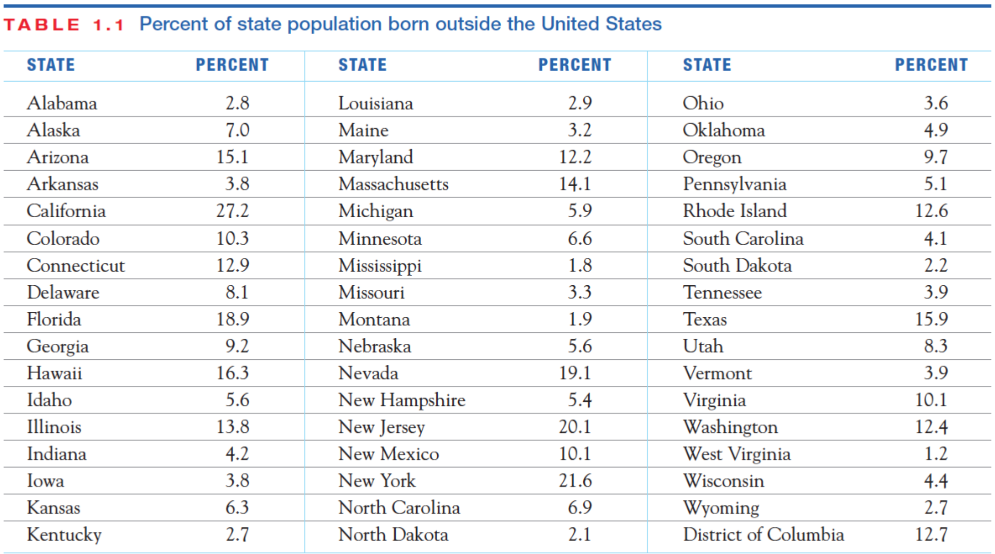
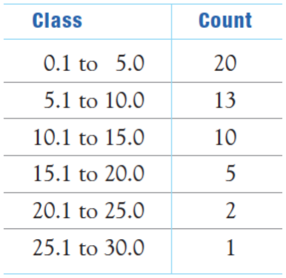
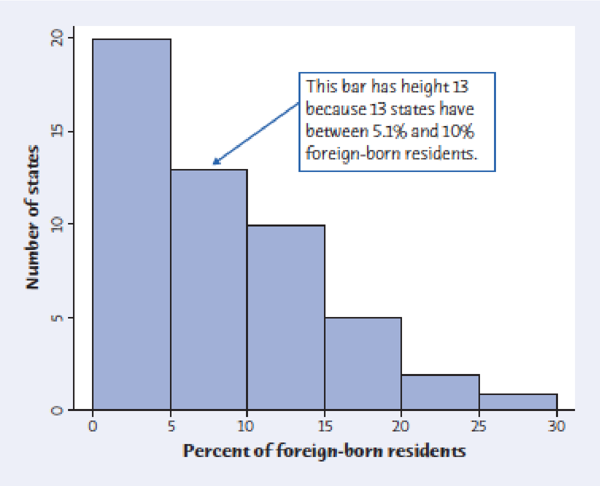


TABLE 1.1 Percent of state population born outside the United States

STATE	PERCENT	STATE	PERCENT	STATE	PERCENT
Alabama	2.8	Louisiana	2.9	Ohio	3.6
Alaska	7.0	Maine	3.2	Oklahoma	4.9
Arizona	15.1	Maryland	12.2	Oregon	9.7
Arkansas	3.8	Massachusetts	14.1	Pennsylvania	5.1
California	27.2	Michigan	5.9	Rhode Island	12.6
Colorado	10.3	Minnesota	6.6	South Carolina	4.1
Connecticut	12.9	Mississippi	1.8	South Dakota	2.2
Delaware	8.1	Missouri	3.3	Tennessee	3.9
Florida	18.9	Montana	1.9	Texas	15.9
Georgia	9.2	Nebraska	5.6	Utah	8.3
Hawaii	16.3	Nevada	19.1	Vermont	3.9
Idaho	5.6	New Hampshire	5.4	Virginia	10.1
Illinois	13.8	New Jersey	20.1	Washington	12.4
Indiana	4.2	New Mexico	10.1	West Virginia	1.2
Iowa	3.8	New York	21.6	Wisconsin	4.4
Kansas	6.3	North Carolina	6.9	Wyoming	2.7
Kentucky	2.7	North Dakota	2.1	District of Columbia	12.7



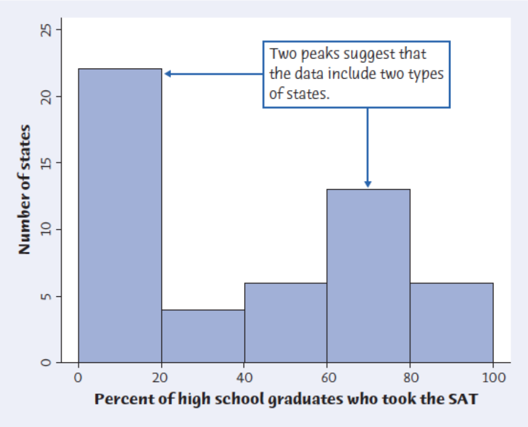
Class	Count
0.1 to 5.0	20
5.1 to 10.0	13
10.1 to 15.0	10
15.1 to 20.0	5
20.1 to 25.0	2
25.1 to 30.0	1



This bar has height 13 because 13 states have between 5.1% and 10% foreign-born residents.

FIGURE 1.5
Histogram of the distribution of the percent of foreign-born residents in the 50 states and the District of Columbia, for Example 1.4.

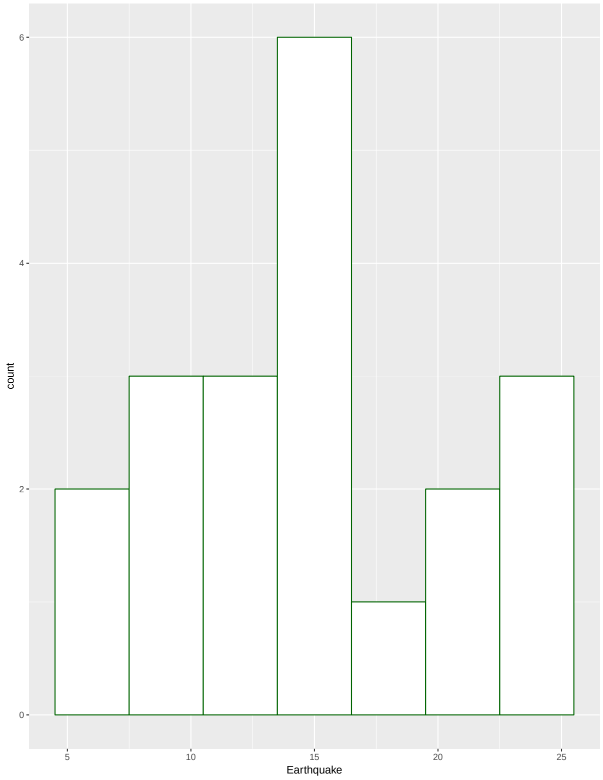
Example 4.2. Who takes the SAT? Depending on where you went to high school, the answer to this question may be "almost everybody" or "almost nobody." The following figure is a histogram of the percent of high school graduates in each state who took the SAT Reasoning test.



Two peaks suggest that the data include two types of states.

FIGURE 1.8
Histogram of the percent of high school graduates in each state who took the SAT Reasoning test, for Example 1.7. The graph shows two groups of states: ACT states (where few students take the SAT) at the left and SAT states at the right.

Figure 1: Histogram for the number of earth quakes of magnitude 7 or greater for years 1980-1999.



4.2 Box plots

A box plot, also called box-and-whisker plots, display the three quartiles, the minimum, and the maximum of the data on a rectangular box. The line extending from each end of the box is called whisker. There are multiple ways to display a box plot.

(a) A central box spans the quartiles Q_1 and Q_3; a line in the box marks the median; lines extend from the box out to the smallest and largest observations.

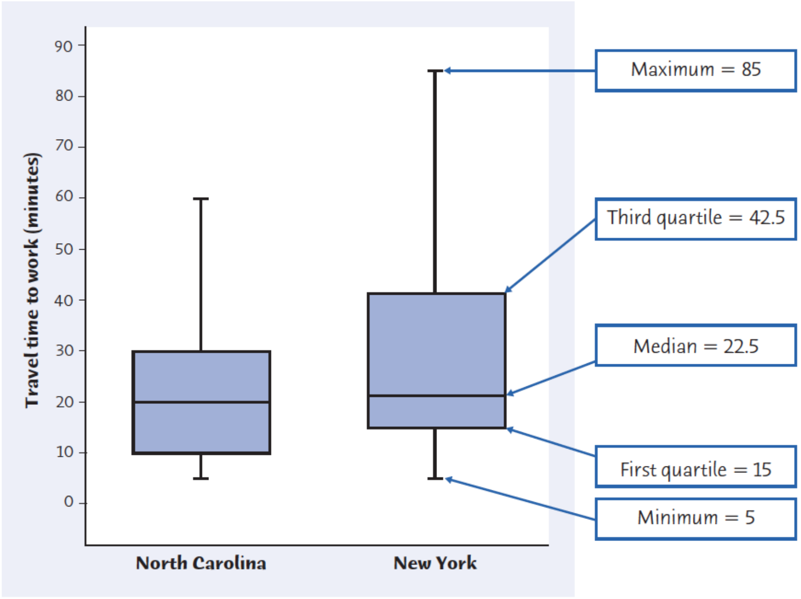
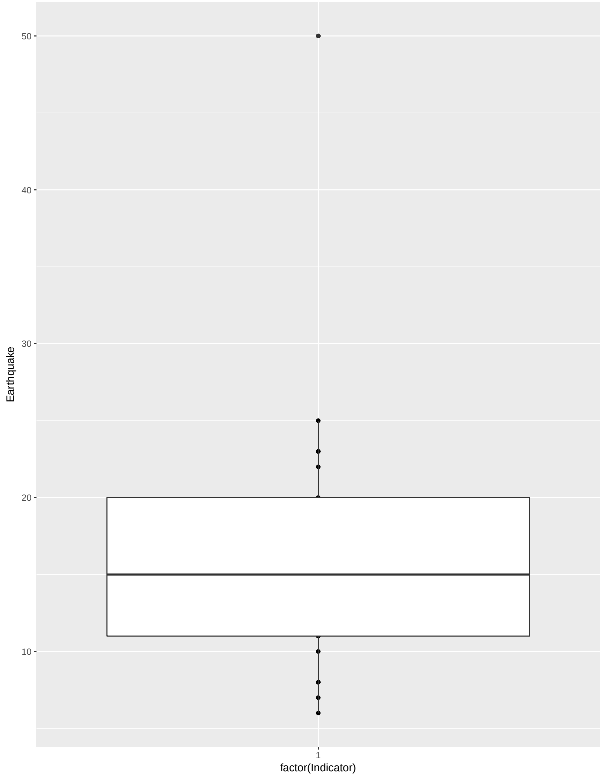


FIGURE 2.1

Boxplots comparing the travel times to work of samples of workers in North Carolina and New York.

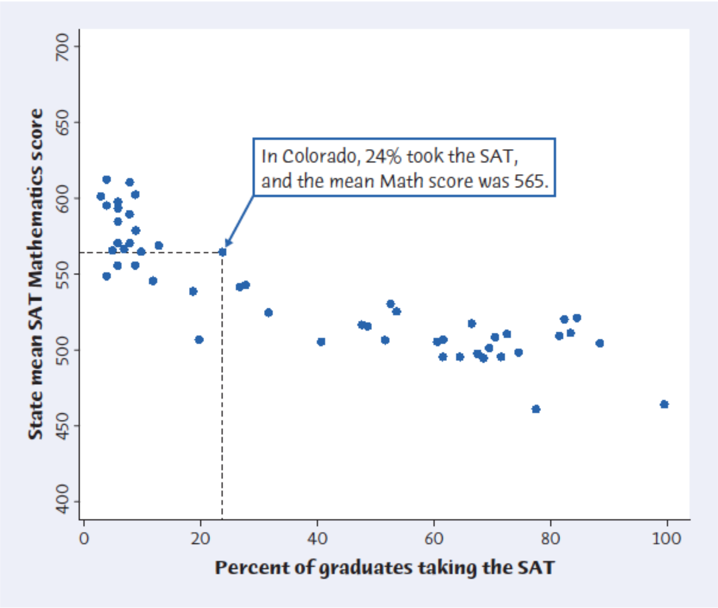
(b) A central box spans the quartiles Q_1 and Q_3; a line in the box marks the median; line extend from the bottom to smallest observation greater than or equal to $Q_1 - 1.5(Q_3 - Q_1)$; line extend from the top to largest observation smaller than or equal to $Q_3 + 1.5(Q_3 - Q_1)$. A point beyond the whisker is called an outlier.



4.3 Scatter plots

To understand a statistical relationship between two variables, we measure both variables on the same individuals. Each observation consist of measurements of two variables, i.e (x, y). Therefore we observe $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ for n subjects. **A response variable** measures an outcome of a study. An **explanatory variable** may explain or influence changes in a response variable. A scatter plot is used to graphically display the potential relationship between the response and the explanatory variables of the observations. Linear (straight-line) relations are particularly important because a straight line is a simple pattern that is quite common. A linear relation is strong if the points lie close to a straight line, and weak if they are widely scattered about a line.

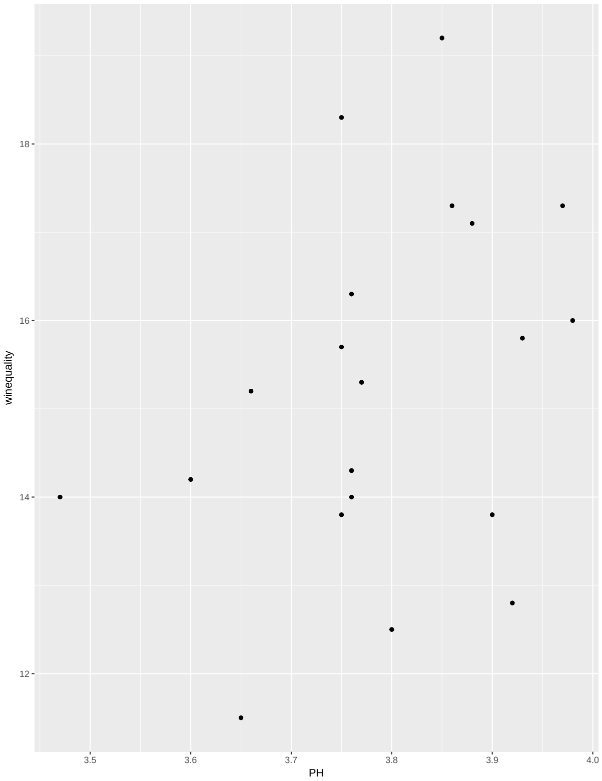
Example 4.3. The percent of high school students who take the SAT varies from state to state. Does this fact help explain differences among the states in average SAT Mathematics score?



In Colorado, 24% took the SAT, and the mean Math score was 565.

FIGURE 4.1

Scatterplot of the mean SAT Mathematics score in each state against the percent of that state's high school graduates who take the SAT, for Example 4.3. The dotted lines intersect at the point (24, 565), the data for Colorado.



4.4 Probability plots

A probability plot is a graphical method for determining whether sample data conform to a hypothesized distribution based on a subjective visual examination of the data. Consider n observations for a variable in a sample $x_1, \ldots, x_n$. Construct the observations from the smallest to the largest. The arranged sample is $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. A normal probability plot can be constructed on ordinary axes by plotting the standardized Normal scores z_j where $\Phi(z_j) = \frac{j-0.5}{n}$ against $x_{(j)}$; $\Phi(\cdot)$ function returns the probability of a z-score for a Normal distribution. If data comes from a Normal distribution, then pairs of (z_j, x_j) would scatter around a straight line closely.

Figure 2: Normal probability plot for wine quality scores of 20 wine bottles.

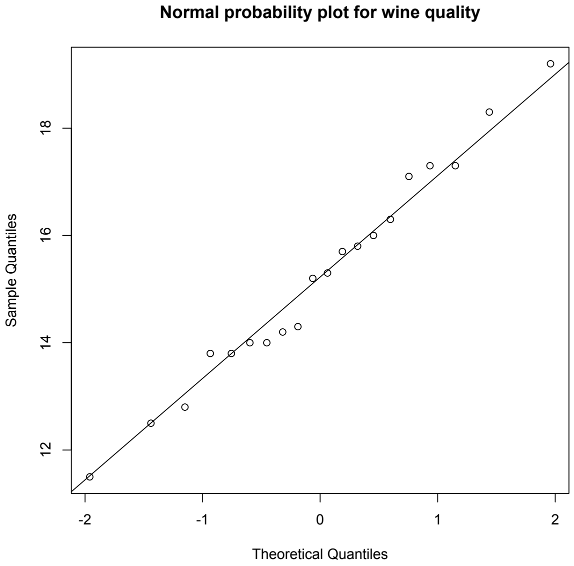
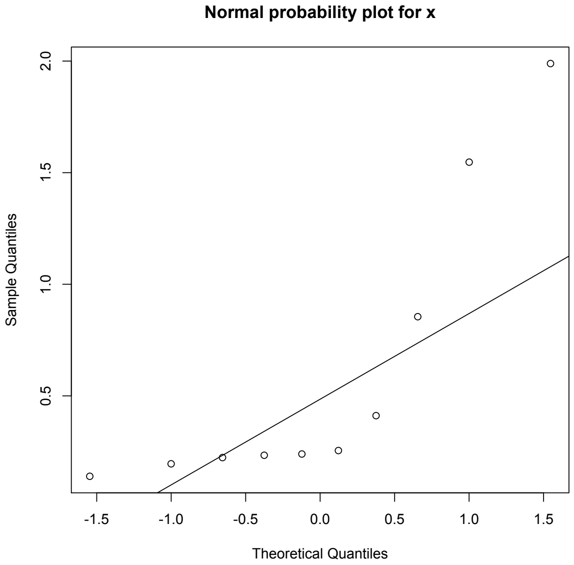


Figure 3: Normal probability for 10 simulated values from Exponential distribution.



4.5 Graphical display using R

```
#Type in data and store it in a vector named "Earthquake".
Earthquake <- c(18, 14, 10, 15, 8, 15, 6, 11, 8, 7, 12, 11, 23, 16, 15, 25,22, 20,
                16, 23)
```

```
#For Homework 5, problem 3
Exer<-c(425, 223, 389, 139, 213, 324, 209, 287, 244, 408, 158, 436)
NoExer<-c(485, 160, 478, 249, 236, 253, 407, 335)

#Wine quality data
winequality <- c(19.2, 18.3, 17.1, 15.2, 14.0, 13.8, 12.8, 17.3, 16.3, 16.0,
                 15.7, 15.3, 14.3, 14.0,13.8, 12.5, 11.5, 14.2,17.3,15.8)
PH<-c(3.85,3.75,3.88,3.66,3.47,3.75,3.92,3.97,3.76,3.98,3.75,3.77,3.76,
      3.76,3.90, 3.80,3.65,3.60,3.86,3.93)

########################################################################################
#Histogram
hist(Earthquake)

par(mfrow=c(1,2))
hist(Exer,main="Exercise group",breaks=5)
hist(NoExer,main="No Exercise group",breaks=5)

########################################################################################
#Boxplot
#Website reference: http://docs.ggplot2.org/0.9.3.1/geom_boxplot.html
Earthquake <- c(18, 14, 10, 15, 8, 15, 6, 11, 8, 7, 12, 11, 23, 16, 15, 25,22, 20,
                16, 23,50)
boxplot(Earthquake)


########################################################################################

########################################################################################
#Scatter plot
#Website: http://docs.ggplot2.org/0.9.3/geom_point.html
plot(PH,winequality,xlab="PH",ylab="Winequality")


########################################################################################
```

```
#Normal probability plot
qqnorm(winequality,main="Normal probability plot for wine quality")
qqline(winequality)
x=rexp(10,1)
qqnorm(x,main="Normal probability plot for a random sample from
       Exponential distribution")
qqline(x)
```