Comments about Quartiles

Descriptive measures of the relationship of a measurement to the rest of the data are called **measures of relative standing**. One such measure is a measurements **percentile ranking.**

According to our text, for any set of n measurements (arranged in ascending or descending order), the pth **percentile** is a number such that p% of the measurements fall below that number and (100 – p)% fall above it. Furthermore, our text states that the **lower quartile** Q_L is the 25th percentile of a data set; the 50th percentile or **middle quartile M** is the median; the upper **quartile** Q_U is the 75th percentile.

For large data sets and very large data sets, those definitions seem to cause no particular problems. However we have noted that in the case of the EPA Mileage Ratings on 100 Cars we found that the median (50th percentile) was 37 but only 49% of the MPG ratings actually fell below 37. So, in cases where the data set does not contain a very large amount of measurements, it is probably better to say the following:

> The pth **percentile** is a number such that **at most** p% of the measurements fall below that number.

Using the revised wording we can say the median for the EPA Mileage Ratings on 100 Cars is 37.0 which means that at most 50% of the measurements fall below 37.0 and at least 50% of the measurements fall at or above 37.0.
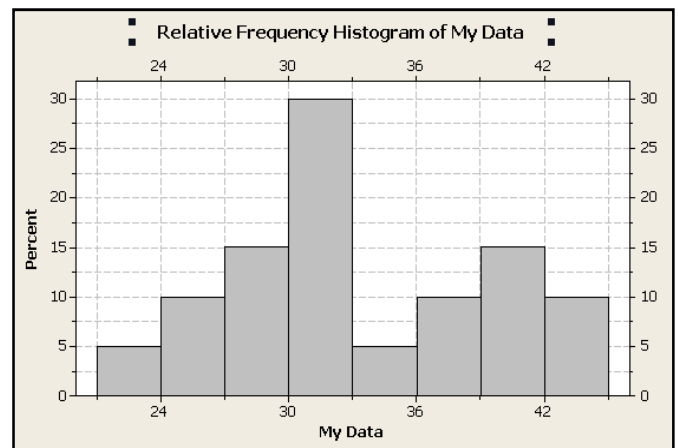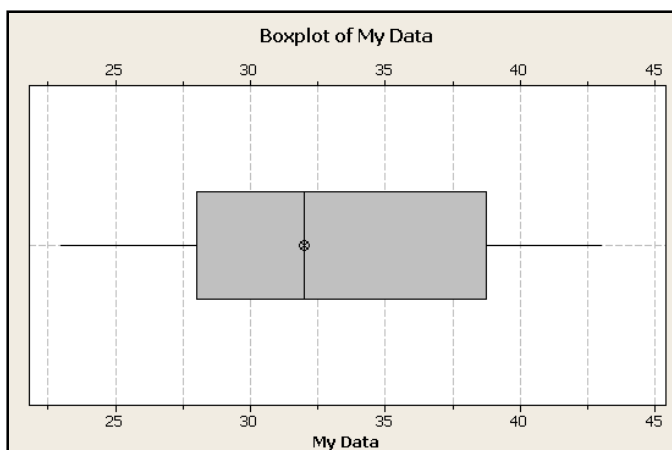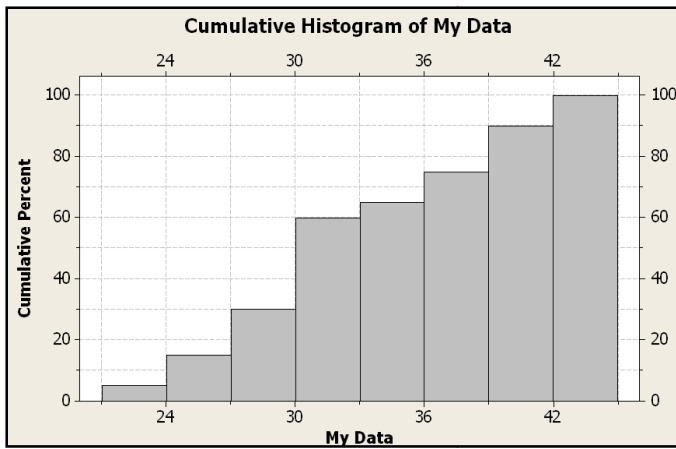
Example:

Data Set: 23, 24, 26, 28, 28, 28, 30, 30, 32, 32, 32, 32, 34, 36, 38, 39, 40, 40, 42, 43

Minitab's Descriptive Statistics for the Data Set:

```
Variable   N  N*    Mean  SE Mean  StDev  Minimum     Q1  Median     Q3  Maximum
C2        20   0   32.85     1.33   5.97    23.00  28.00   32.00  38.75    43.00
```

In this case it might be appropriate for us to say *at most* 25% of the measurements are less than Q_L = 28 and *at most* 50% of the measurements are less than M = 32 and *at most* 75% of the measurements are less than Q_U = 38.75.

Boxplot of My Data

Relative Frequency Histogram of My Data

Cumulative Histogram of My Data

Consider the cumulative distribution graph to the left and complete the following statements.
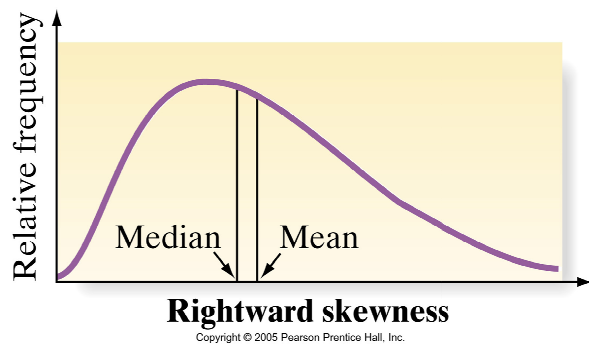
At most about _____% of the measurements are less than 24.

At most about _____% of the measurements are less than 33.

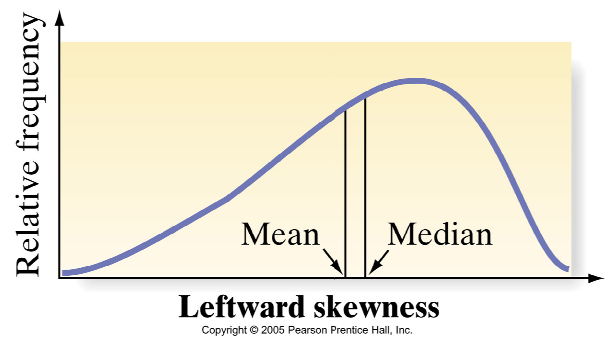At most about 90% of the measurements are less than _____.

At most about _____% of the measurements are less than _____.

Comments about Skewness

Rightward skewness

Copyright © 2005 Pearson Prentice Hall, Inc.

Use of mean overstates average

Leftward skewness

Copyright © 2005 Pearson Prentice Hall, Inc.

Use of mean understates average

Comments about Outliers

Observations more than three standard deviations from the mean are generally considered **outliers**. For some highly skewed data sets, observations more than 2 standard deviations from the mean may be outliers.

Values less than $Q_L - 1.5(IQR)$ or greater than $Q_U + 1.5(IQR)$ are identified as **potential outliers** or **suspect outliers** because they are extreme values that represent relatively rare measures. Those **potential outliers** are identified by asterisks (*) in Minitab. Values less than $Q_L - 3(IQR)$ or greater than $Q_U + 3(IQR)$ are called **outliers** or **highly suspect outliers** and are also identified by asterisks (*) in Minitab.