**Slide 3**

If you learn that the mean or median salary is $100,000 this tells you the typical salary but it tells you nothing about the variability of the salaries.

The range is a fairly crude measure of variability. It is sensitive to the extremes.

Eg if one players salary increased to 40 million the range would increase by 10 million just because of one player. More frequently used is the variance and std dev.

**Slide 4**

Sample variance (VAR function) and population variance (VARP function)

**Slide 5**

Squared dollars don’t mean much..

**Slide 6**

This example shows why variability is important.

Both suppliers have an average diameter of 100 for the product, however the variance shown by supplier 2 is much larger.

**Slide 7**

Fortunately many variables in real world data are approximately normally distributed, so these rules apply.

If supplier 1’s diameters are normally distributed then the rules apply as follows:

68% of suppliers parts will have diameters from 97 to 103 (+- 3 for one standard dev)

95% of parts will be 94 to 106

99% will be 91 to 109

For supplier 2 it is much worse, with the std dev larger than 25 the variability is large.

**Slide 8**

Std dev of salaries was 4.535 million (the variance is shown but because it is in squared dollars it is huge and has no meaningful interpretation)

You can always try to apply the emipirical rules but if the salaries are not at least approximately normally distributed the rules won’t be very accurate. Because of the skewness in this data we can assume it is not normally distributed.

In this example the standard dev is larger than the mean!, thus all three lower end points are negative

The percentages given at the end show the difference from the rules of 68%, 95% and 99%

**Slide 9**

Skewness occurs where there is a lack of symmetry. A few players have really large salaries and no players have really small salaries. The largest salaries are much further to the right of the mean than the smallest salaries are to the left of the mean. This is apparent when you draw a histogram.

These salaries are skewed to the right or positively skewed.

The value doesn’t make much difference, positive is skewed right, negative is skewed left and zero is no skew.

**Slide 13**

We are particularly interested in whether the distribution is symmetric or is skewed in one direction.

Each of the summary measures discussed on numerical variables (mean etc) describe only one aspect of a numerical variable. In contrast a histogram provides the complete picture. It indicates the centre, the variability, the skewness all in one chart.

**Slide 16**

A box plot is not as popular as the histogram but side by side box plots are useful for comparing distributions such as salaries for men versus salaries for women. They show you at a glance some of the key features of the distribution

**Slide 17**

The box extends left to right from the 1st quartile to the 3rd quartile. The box contains the middle half of the data, the line in the box is the median, the line to the right is a whisker that extends to the outliers. It shows again that it is a right skewed distribution. Box plots and histograms are complemntary ways of displying the distribution of a numerical variable, histograms are more popular and intuitive.

**Slide 18**

When we analyse time series data summary measures such as means and std dev and charts often don’t make much sense.

If you are interested in daily closing prices of a stock, if you create a histogram you will get counts for each bin range of closing value but you won’t know when they occurred, if you report the mean of stock prices over 40 years it really isn’t useful or relevant.

**Slide 23**

As mentioned, traditional summaries such as means, medians and std dev are not often very meaningful. It is useful to find differences or percentage changes in the data from period to period and then report traditional summaries of these.

**Slide 24**

The percentage changes in the DOW have been calculated