Lecture 1

Yale grades data [Stats 101-106 homepage ---> datasets ---> Yale grades data]

The data consist of the scores for students on 11 homework sheets, the midterm, and the final exam. The last row gives the scores for a hypothetical perfect student, that is, it gives the maximum score possible. A score of zero always indicates that a student failed to hand in a homework sheet. The combined homeworks (with midterm score added in) counted for 50% of the grade; the final exam counted for the other 50%.

Questions

How would we summarize the distributions of scores on the homework (plus midterm) and on the final exam?

How should the scores from the homework plus midterm be combined with the scores from the final to fairly represent the 50% weightings?

How well, at various stages throughout the course, would the performance on homework predict performance on the final exam, and predict grades on the course?

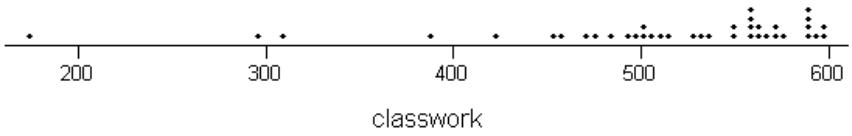
Let me write ***classwork*** for the sum of the scores on sheets 1 through 11 plus the midterm. Here are the classwork scores, sorted into increasing order:

```
176   297   310   390   421   452   457   470   475   483
495   497   500   501   504   509   516   526   531   535
548   548   558   558   558   560   561   561   569   571
573   575   587   587   588   591   592   596   596
```

Note: It is easier to get a feel for the data if you sort the rows into some meaningful order. I will use order of increasing classwork score.

Some pictures of the distribution of classwork scores.

Dotplot:

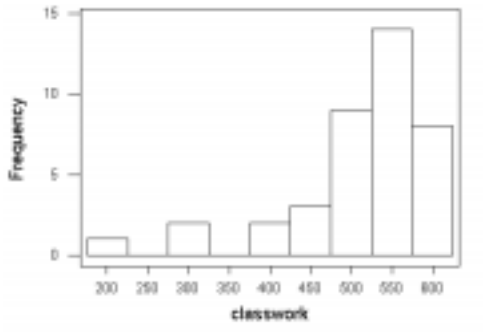


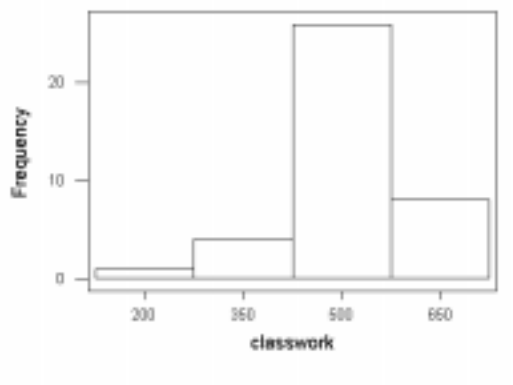
Stemplot:

```
  1    1 7
  1    2
  2    2 9
  3    3 1
  4    3 9
  5    4 2
 12    4 5577899
(10)   5 0000123344
 17    5 55566667778889999
```

Try to match up the leaves (in the stemplot) and the dots (in the dotplot) with the sorted scores.

Histograms:





The height of each bar gives the number of student scores (frequency) in the interval (the bin) specified by the base of the bar. (For example, in the histogram at left , two students had classwork scores in the range 275 to 325.) Notice the effect of increasing the width of the ranges. Compare with the dotplot. Notice the similarity of the stemplot to a histogram tipped on its side.

Percentiles:

There are many slightly different definitions in use. Here is a simple method for N data points.

1. Sort data into increasing order.
2. For the pth percentile, express the fraction (N+1).p/100 as an integer part, k, plus a fractional part, f.
3. Put the pth percentile a fraction f of the distance between the kth and (k+1)st data point.

For example, for the 39 coursework scores:

The 50% percentile is called the median. Calculate (39+1).(50/100) = 20. The median is right at the 20^{th} smallest score, that is, median = 535.

The 25% percentile is called the first quartile (Q1 in the table below). Calculate (39+1)(25/100) = 10. The 25^{th} percentile is at the 10^{th} smallest score, that is, 483.

For the 16^{th} percentile calculate (39+1).(16/100) =6.4. The 16^{th} percentile lies 0.4 of the way between 452 (the 6^{th} smallest score) and 457 (the 7^{th} smallest score)—that is, at
$0.4 \times .452 + 0.6 \times 457 = 455$

You really shouldn't worry too much about the calculation. Minitab will do it for you.

The interquartile range (IQR) is the distance between the quartiles: 571– 483 = 88 for the coursework scores.

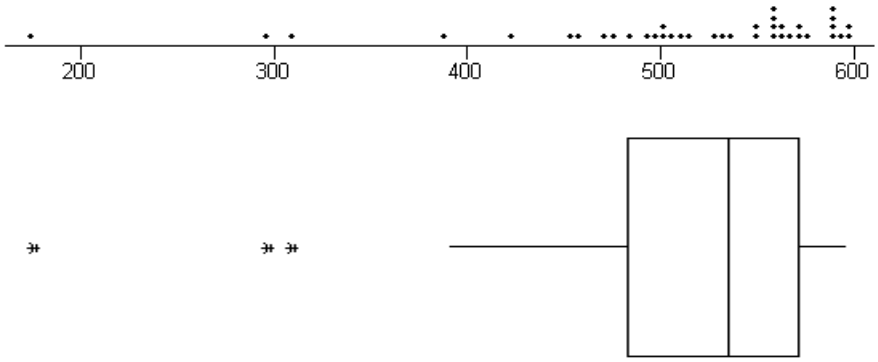
N	Mean	Median	StDev	Minimum	Maximum	Q1	Q3
39	510.8	535.0	90.3	176.0	596.0	483.0	571.0

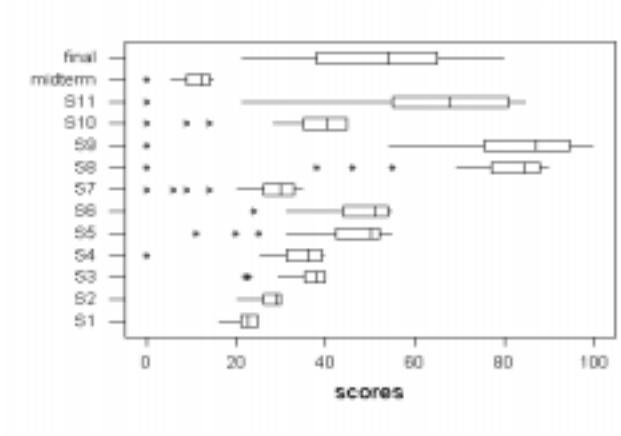
Boxplots:

I usually look at the big box in the middle, which gives the 25^{th}, 50^{th},, and 75^{th} percentiles at a glance.

For the lower whisker, stick pins into the data points, stretch a rubber band 1.5 $\times$ IQR (see normal distribution, below) from 25^{th} percentile, let it snap. For the coursework, the rubber band stretches from

483 to 483-1.5×88 = 351, then it snaps back to the score at 390. All scores below 390 get marked with a star (*), to flag possible outliers. The upper whisker is constructed similarly. (See M&M page 48.)





Boxplots are very useful for comparison between many similar distributions..

Why are there so many stars down the left edge?

What do you learn about how the homeworks changed during the semester?

Was it difficult to do well on the homeworks?

Should a score of 80 on the homeworks carry the same weight as a score of 80 on the final exam, if coursework and final are each to carry 50% weight towards the grade?

Location:
What is the 'average' score on the coursework?

mean = (176 + 297 + … + 596)/39 = 510.8
median = 50^{th} percentile = 535

In general, for values $x_1,x_2,\ldots,x_N$, the mean equals = $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

Spread:
standard deviation = square root of variance

variance = $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$

Why squares? (Convenience? Tradition?) Why N-1? (See M&M page 53.)
Why subtract the mean? (Spread not affected by change of location.)
Why square root? (Double the values, double the spread.)

Standardize:
For data $\mathbf{x_1,x_2,\ldots,x_N}$, define standardized data $z_i = \frac{x_i - \bar{x}}{s}$

The standardized data have 0 mean and standard deviation 1

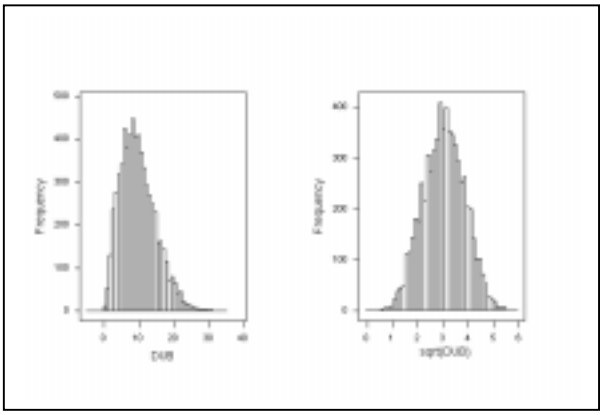
Why not "standardize" to get zero median and unit IQR?

> If we wish to compare the 'shapes" of two distributions that have different scales, or centers of location, it is a good idea to perform some form of standardization.

Normal distributions and densities

Look at the wind data on the web site [Stats 101-106 homepage ---> datasets ---> wind data], which give daily average wind speeds at 12 meteorological stations in Ireland , for a period of 18 years.

I have drawn a histogram for both the DUBLIN wind speeds and the square root of the DUBLIN wind speeds. Notice the rough "bell shape" for the square root data.

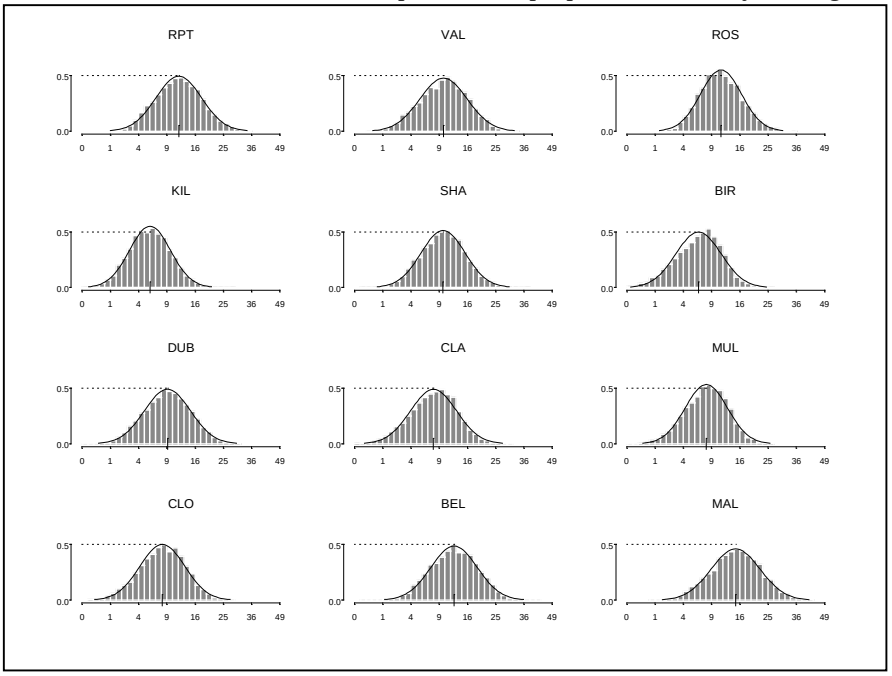


[Little puzzle: How can the maximum heights for the two histograms be different?]

Many data sets, possibly after some suitable transformation, give similar shaped histograms.

If we standardize the data before constructing the histogram, we get a shape known as the *standard normal*. (M&M page 74)

Here are the corresponding histograms for the square roots of daily average wind speeds at all 12 stations. I have superimposed an approximating normal curve on each histogram. The vertical scales have been reduced so that the area of each bar represents the *proportion* of all days falling in the corresponding interval. (For convenience of interpretation I have labeled the horizontal axes with actual wind speeds, rather than their square roots.)



Notice that the approximating curves differ only in the location and scale.

To get the same approximating curve in each case I could have standardized all the data before constructing the histograms.

You could think of the approximating smooth density curve as a histogram with extremely narrow bin widths, an idealized model for the distribution of data (M&M page 69). M&M write μ for the mean of this idealized distribution and σ for its standard deviation.

If the data are standardized, to have zero mean and unit standard deviation, the approximating normal curve has $\mu=0$ and $\sigma=1$. The normal curve is standard normal.

Look at M&M pages 72-73 for the general shape of the normal, and the proportions of the area within various specific ranges.

Normal (probability/quantile) plots:

(Compare with the footnote on M&M page 80 with the captions to the figures on pages 82-83.)

Not all "bell shapes" are normal. Plot percentiles for the data against the corresponding percentiles for a standard normal distribution. If the plot is approximately a straight line then the data histogram has approximately a normal shape.