MATH 124 Spring 2005

Lecture: 5
Date: Feb 22, 2005

Skills you should acquire from this lecture:

- Understand and be able to compute Variance, Standard deviation, IQR from a list of data
- Be able to identify outliers using the 1.5IQR rule
- Be able to draw and interpret a modified boxplot

Related readings in the textbook:

-Section 1.2
-Section 2.1

Standard deviation and variance

The *standard deviation*, and a closely related quantity, the *variance* are the most commonly used measures of the variability of data. In words the variance is the *average of the sum of the squares of the deviations from the mean*. Mathematically, the variance is

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

the standard deviation is the square root of the variance. ie $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$. Why do we square the deviations? Because we would get 0 if we summed the deviations. Why is the denominator $n-1$? Because it makes the estimate *unbiased* (we will talk about this more later in the class). Why is the standard deviation used more than the variance? Because it is in the same units of measurement as the data, whereas the variance will be in the squared units.

An alternative formula for the variance is

$$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

The standard deviation is always positive. It only equals zero when there is no spread in the data. The standard deviation is not resistant to outliers.

Range

The *range* is given by the distance between the largest data value and the smallest data value.

IQR

The *interquartile range* is a more robust measure of variability. It is given by the difference between the upper and lower quartile. ie

$$IQR = UQ - LQ$$

Modified Boxplot

A common modification is to use the IQR to establish a criterion for outliers. In particular, any observation that is more than 1.5 times the IQR above the UQ or more than 1.5 times the IQR below the LQ. A modified boxplot is to then drawn such that the whiskers (lines) are only extended out as far as the observations that are not outliers. Outliers are marked as individual points.

An example

Suppose we have the data

`0.8 0.9 1.4 0.6 0.2 2.0 0.9 3.9 2.4 1.0`

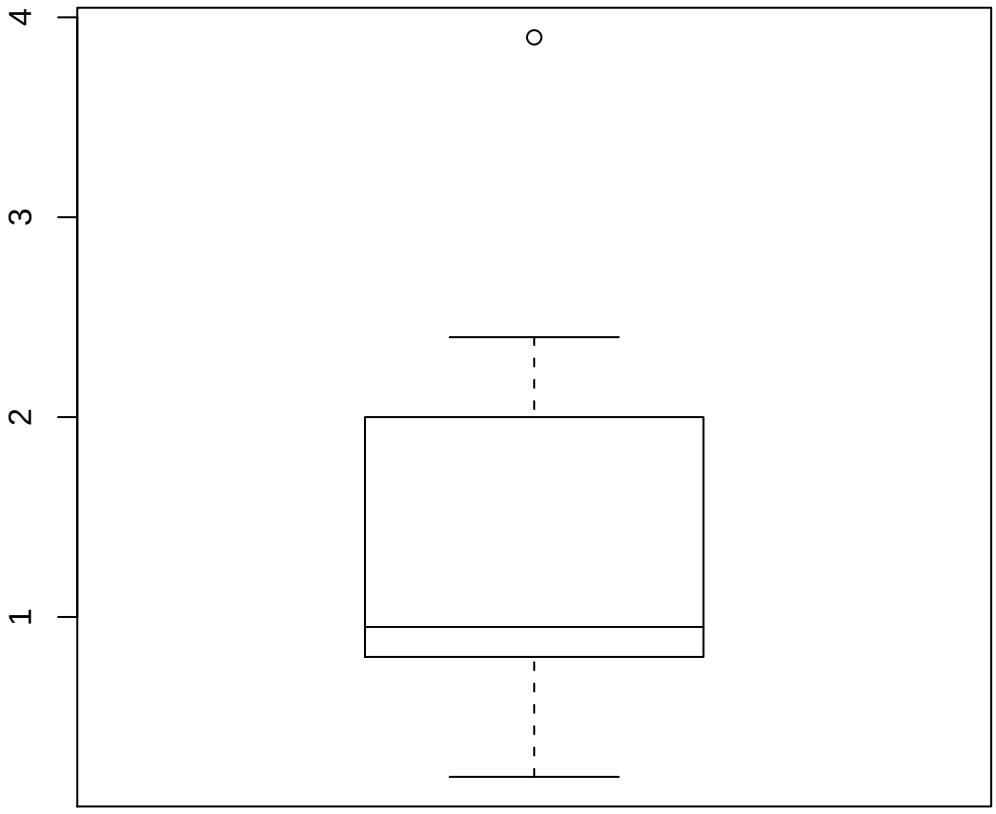
sorting the data gives us

`0.2 0.6 0.8 0.9 0.9 1.0 1.4 2.0 2.4 3.9`

The median is

`(0.9 + 1.0)/2 = 0.95`

the LQ is given by `0.8` and the upper quartile is `2.0`. The IQR is `2.0 - 0.8 = 1.2`. 1.5 times the IQR above the UQ is `3.8` and 1.5 times the IQR below the LQ is `-1.0`.



Relationships between two variables

We have talked about how to inspect the distribution of a single variable, both graphically and using summary statistics. However, often we measure more than one variable for each individual eg height and weight, age and number of car accidents in the last 5 years, cholesterol and heart disease. If is useful to look at the relationship between variables in many contexts. We say that two variables, each measured on the same set of individuals, are *associated* if some values of one variable are more likely to be paired with some values of the other variable. eg higher values of variable 1 are more likely to be recorded with lower values of variable 2. Note that an association is only a tendency. It is not an ironclad rule (ie there may be perfectly sensible exceptions), and it does not prove that there is a causal relationship between the two variables that are associated.

We often describe variables as being either:

1. *Response variable*: a variable which measures the outcome of a study.

2. *Explanatory variable*: a variable that causes or explains the changes in the response variable

When we carryout an experiment where we control one varibale it is easy to say which is the explanatory variable (the one which we can set) and which is the response (the one that we measure). eg consider an experiment using a blood pressure medicine. The experimenter can control the level of the blood pressure medicine (the expalantory variable) and then measures the initial and final blood pressure levels of the patients after some period of time to get the change in blood presure (the response variable).

However, many times you only have observational data and therefore it is not clear as to what is the explantory variable and what is the response variable. In cases like this we generally choose based on what we are going to do with the data. For instance maybe we want to use one variable to predict values of the other variable. Sometimes we are helped by the temporal order in which things are measured (ie one variable happens before the other). For example suppose we measure students high school GPA and college GPA. While a high school GPA does not cause a college GPA, it may well be useful in helping predict it. Since high school is before college, we would treat the high school GPA as the explanatory variable and the college GPA as the response variable.

It is often the case that an unmeasured variable (we call this a "*lurking variable*") often explains or is the vause of both the observed variables. For example the hours studied per week could help explain both high school and college GPA.