Slide 2

Most books and examples use cleaned data, the data you will encounter in business will not be clean. Outliers and missing data will be problems you encounter.

Alex Rodriguez’s salary of 32 million is an outlier.

Outliers are extreme values, you can decide for a given data set how extreme a value needs to be to qualify as an outlier.

Sometimes an outlier is easy to detect and deal with, sometimes they are identified as errors. You can go an check the observation. Sometimes combinations of values are noticed, like someone aged 10 with a height of72 inches.

It is not appropriate to remove outliers just to produce nicer results, there has to be a legitimate reason for removing outliers, e.g. if you want to analyse typical manager salaries then it makes sense to remove CEO salary.

Slide 3

There are no missing values in the baseball salaries data, all 843 observations have values for each of the four variables. For real data sets this is probably the exception rather than the rule.

They could be missing because a person didn’t want to provide the info or that it didn’t exist for that observation, or because some values are not known. Missing values are obvious in excel by finding blank cells. Some are coded such as 9999 or -9999 or – or \* if you know the code you can do a search and replace with blanks.

What happens if you do an average on a column with missing values? It divides my the number of nonmissing values

Slide 5

Apply these tools to the new file type .xlsx

Slide 6

The data set is now designated as a table, it is formatted, and a dropdown arrow appears next to each variable name, a new table tools design ribbon is now available. It is called Table1 by default (on ribbon), you can change it to a more descriptive name. variable names remain visible even when you scroll down, this is only when the active cell is within the table.

Slide 7

The two main advantages of filtering on a table as opposed to the three options here are the nice formatting (banded rows etc) provided by tables and more importantly the total row. If this total row is showing it totals only those visible records not the hidden ones.

Slide 8

On continuous fields the dropdown arrow gives you toption like all salaries greater than 75,000 choose from number filters, other number filters include top10, above average, below average etc. The top 10 filter can be used to select the top n items.

For text values you can enter choice like begins with, ends with, contains and others.

Dates show as year with a plus sign you can drill down to months, days etc. in the date filters option you can access yesterday, next week, last month etc.

Coloured cells are often used for size of number value.

If nothing else works you can use a custom filter.

When you click the down arrow you always have sorting options also

Slide 11

Smoking and drinking, non smoker, occasional smoker, heavy smoker, same for drinking.

Slide 13

=countifs($b$2:$B$8762,F$3,$C$2:$C$8762,$E4)

The countif only lets you specify on a single criteria, now there are two criteria.

So the above formula in entered in F4 and the copied across from F4 to H6

The first two arguments are for the condition on smoking the last two on drinking you then also get the totals across categories.

There are big differences between the totals (many more non smokers that heavy smokers) any relationship is difficult to detect. It is useful to express these raw counts as percentages.

The second tables show a relationship clearly,

You can create column charts of the percentage crosstabs, highlight E10:H13 and insert a column chart

Slide 14

The list could go on and on, exploratory methods are used to investigate if there are differences across the subpops on the numerical variable of interest. Generalisation of the relationships are then tested by confidence intervals and hypothesis testing (beyond scope of this module)