
Techniques and tools for data analysis

Preview

Introduction

In chapter 3 of *Statistics In A Day* different combinations of numbers and types of variables are presented. We go through these sections one by one here. The aim is primarily to illustrate the different analyses that occur. There is only a little new ‘theory’ and only the occasional new idea is introduced—most of basic ideas were covered in chapter 2 of *Statistics In A Day* and discussed in topic 11.

READING **Read section 3.1 of *Statistics In A Day*. This simply provides a summary of the ideas covered in chapter 2.**

Analyses for one variable

Analysis for one numerical variable

The ‘Yield’ example in chapter 2 of *Statistics In A Day* involved one numerical variable, so there is not a lot of new material in section 3.2 of *Statistics In A Day*. Example 3.3 involves almost identical work to that done in the ‘Yield’ example, in fact.

EXERCISE 12.1 **Work through section 3.2 of *Statistics In A Day*. Reproduce the work for examples 3.3 and 3.4 on your own.**

Example 3.4 has a couple of new ideas. The first is that you’ve always got to look out for *paired data*—working with the difference between the variables reduces the number of variables by one (KISS principle). Second, the data turns out to be quite badly skewed. So, instead of working with the mean (which, as you know from our earlier work, can be unreliable in the presence of skewness), the median is used. The corresponding hypothesis test appears almost identical to the test for the mean but it doesn’t assume the data is bell-shaped (it’s a so-called non-parametric test).

There are other analyses for a single numerical variable of which you should be aware. For example, there are a number of variations of the usual frequency histogram. On the Summaries sheet in the XLStatistics workbook 1Num, the frequency chart can show relative frequencies (percentages in each class), for example. Or sometimes you might want to label the actual classes, at other times endpoints.

Also on the Summaries sheet in 1Num is a boxplot of the data. This useful plot showing the min., first quartile, median, third quartile, and max. was briefly discussed in chapter 2 of *Statistics In A Day*. It is particularly useful for detecting outliers. Try it on the data for example 3.4, for example.

The Mean plot on the Summaries sheet is a useful summary chart although it is really more useful when there are a number of groups (see analysis for one numerical and one categorical

variable below). The bar shows the (sample) mean; the error bars (the little lines that extend up and down) can show the range or standard deviation (to show the spread of the data about the mean) or a confidence interval (to give an idea of how accurately we know the true mean).

There are extra analyses on the Tests sheet in 1Num. The Wilcoxon Signed Rank test is rarely used in business circles—it is a nonparametric (doesn't assume the data is bell-shaped) alternative to the more common t-test that is usually applied to differences from paired data. The Chi-square test on the variance can be applied if your question is about the variability of your data (the variance is the square of the standard deviation).

Grouped data

It is quite common that data comes already grouped in classes, as for a frequency chart. For example, in example 3.3 of *Statistics In A Day* we might have just been given the following frequency table (table 12.1):

TABLE 12.1

Class	Frequency
10–14	7
15–19	32
20–24	37
25–29	12
30–34	2

It is not quite possible to recover the original data from this, but we can come close. We know, for example, that there are seven numbers between 10 and 14—pretend they are all 12 (halfway between 10 and 14). We know there are 32 numbers between 15 and 19—pretend they are all 17. Now analyse this resulting data. This is effectively what is done in the Grouped Data workbook on the Extra Tools sheet in 1Num. Try it with the above grouped data. Enter it as displayed in figure 12.1:

FIGURE 12.1

Data		Variable X	
Lower	Upper	Frequency	
10	15	7	
15	20	32	
20	25	37	
25	30	12	
30	35	2	

—you'll see that you'll approximately recover the results obtained with the original data in example 3.3 of *Statistics In A Day*.

Analysis for one categorical variable

The analyses appropriate for single-categorical-variable data almost all involve proportions of the various categories that comprise the variable. These proportions are often pictured in a pie chart but simple tables are good, or bar charts.

EXERCISE 12.2

Work through section 3.3 of *Statistics In A Day*. Reproduce the work for example 3.9 on your own.

Often you will need confidence intervals for the various proportions, or hypothesis tests. These are on the Tests sheet in 1Cat, for example. The analysis for determining sample sizes is similar to that for a single numerical variable. The ‘usual’ analyses assume that the sample sizes are ‘large’ but there are small-sample tests that can be used if necessary.

Analyses for two variables

Analysis for one numerical variable and one categorical variable

Problems involving one numerical variable and one categorical variable are everywhere—it’s probably the most commonly encountered combination of variables. The analysis almost always involves splitting the data up into the various categories that comprise the categorical variable, and then comparing them.

EXERCISE 12.3 **Work through section 3.4 of *Statistics In A Day*. Reproduce the work for examples 3.14–3.17 on your own.**

Examples 3.14–3.17 take you through four common types of analysis:

- Example 3.14 involves comparing means (and uses a technique called Analysis of Variance).
- Example 3.15 involves a common situation where the categorical variable has only two levels. The ‘extra’ analysis that you can do in this case is to directly compare the magnitude of the means, or medians, and so on.
- Example 3.16 is a non-parametric test—the data is quite badly skewed so we compare medians instead.
- Example 3.17 involves paired data again—as before, the ‘trick’ is to work with the differences. This example is also interesting in that it illustrates how ‘including important sources of variation’ can make a huge difference to the conclusion.

The tests for the above examples are all carried out using the XLStatistics workbook 1Num1Cat. There are other analyses in this workbook that are worthy of a mention, too.

- The plot of means on the Summaries sheet shows the means together with error bars showing variability (range or standard deviation, for example), or confidence intervals for the true means. There are a number of useful alternatives. The one with the bars is good if the levels of the categorical variable don’t have any special order. The ones with the lines are good if the levels comprising the categorical variable can be ordered in some way (like months, for example).
- The side-by-side boxplots on the Summaries sheet show separate boxplots for each level of the categorical variable, but they are on the same set of axes so they can be easily compared.
- The frequency chart on the Summaries sheet has the same sort of options as the one in 1Num. All are used at various times.
- On the Tests-2 Levels sheet there is a link to analysis for sample-size determination. This is similar to the analysis found in 1Num, for example. It is actually more often used than the analysis in 1Num. ‘What sized sample do I need?’ is probably the most common question asked of statisticians, and two-level data is probably the most common type of situation where it is asked!

Analysis for two categorical variables

There are lots of problems involving two categorical variables. As with single-categorical data, the analysis usually involves proportions.

EXERCISE 12.4 **Work through section 3.5 of *Statistics In A Day*. Reproduce the work for examples 3.23–3.24 on your own.**

Examples 3.23 and 3.24 take you through two common situations:

- Example 3.23 is similar to example 3.15—it involves a common situation where one of the variables has only two levels. The ‘extra’ analysis that you can do in this case is to directly compare the proportions.
- Example 3.24 involves a so-called Chi-square test. This is a test for detecting a relationship between two categorical variables.

The tests for these examples are all carried out using the XLStatistics workbook 2Cat. There are other analyses in this workbook that are worthy of a mention, too.

- There are various tables on the Summaries sheet, showing, for example, percentages calculated in various ways. All are useful from time to time. In general these are called pivot tables. The charts based on the tables are not as popular as the tables themselves (as you can imagine, they get pretty complicated pretty quickly).
- There are special analyses for data where each variable has only two levels (on the Extra Tools sheet).

Analysis for two numerical variables

Analyses for two-numerical-variable data commonly involve fitting curves of various types to the data points on a scatterplot.

EXERCISE 12.5 **Work through section 3.6 of *Statistics In A Day*. Reproduce the work for example 3.31 on your own.**

The results on the Linear Regression sheet in the XLStatistics workbook 2Num deal with fitting straight lines. On the Extra Tools sheet are analyses for fitting other functions (power functions, quadratics, ‘s’-shaped curves, etc., etc.). Also on the Extra Tools sheet are analyses for fitting curves that do not have a pre-specified form (one is used in example 3.31, for example). All are useful in various situations—have a look around in the workbook.

Another version of the KISS principle applies when fitting curves to two-numerical-variable data: (as a rule) fit the simplest curve you can. So, don’t try fitting a quadratic if a straight line will do just as well. There are analyses that will tell you if the quadratic terms are important (a confidence interval for the coefficient, for example), but it’s best to try and avoid the complication in the first place, if possible.

Analyses for three variables

Analysis for two numerical variables and one categorical variable

Analysis of data with this combination of variables often involves splitting the data up into the various categories comprising the categorical variable and doing analysis of the resulting

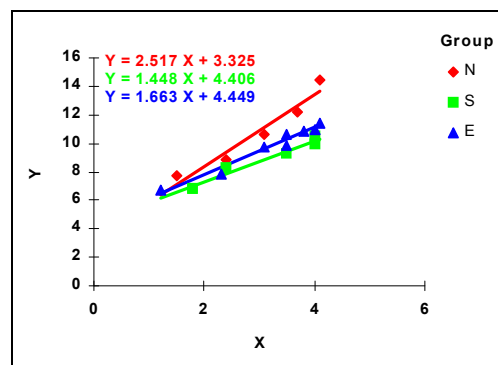
two numerical-variable data. So you typically get analyses involving fitting curves to the data, then, maybe, comparing the curves somehow.

EXERCISE 12.6 **Work through section 3.7 of *Statistics In A Day*. Reproduce the work for example 3.36 on your own.**

Example 3.36 illustrates this idea with fitted straight lines—the most common analysis (because it’s the simplest). There are many variations on a theme. Even fitting straight lines you have many choices—you can fit the straight lines so that their slopes are the same, or you might want to set intercepts (constant terms) to be the same so that all the lines start out from the same point on the vertical axis.

And now that we have more than two variables another feature can arise—interaction between the independent variables. This was discussed briefly in chapter 2 of *Statistics In A Day*. In the context of data with two numerical variables and one categorical variable it often manifests itself as non-parallel lines as in figure 12.2.

FIGURE 12.2



Interaction is usually not difficult to handle with three-variable data, but as the number of variables increases it becomes increasingly more difficult to detect and to deal with effectively. This is one of the main reasons that data with many variables becomes hard to analyse—hence the KISS principle.

Not all analyses for two numerical variables and one categorical variable are to do with curve fitting. Another common analysis involves treating the two numerical variables separately and seeing if the relationship with the categorical variable is any different. Multi-axis plots like the one pictured at the end of 3.38 in chapter 3 in *Statistics In A Day* are quite common and quite useful here.

Analysis for one numerical variable and two categorical variables

Analysis of data with one numerical and two categorical variables often involves the use of tables. Pivot-tables, such as in the analysis of two categorical variable data, can be used but this time summaries of the numerical variable can appear in the cells in the table (rather than just counts or proportions). So in the Description box in the XLStatistics workbook 1Num2Cat you’ll find tables. On the Summaries sheet are charts based, typically, on the table of means—these are similar to the plots of means for one numerical and one categorical variable.

EXERCISE 12.7 **Work through section 3.8 of *Statistics In A Day*. Reproduce the work for examples 3.41 and 3.42 on your own.**

The most common test for comparing means is called a two-way analysis of variance. The details can get quite complicated and the XLStatistics workbook 1Num2Cat can only handle so-called balanced data, where all the values in the Counts table on the Data and Description sheet are the same. The relatively simple analysis carried out in the workbook is reasonably straightforward—the hypothesis tests are spelled out fairly clearly. Some details are mentioned in 3.45 in *Statistics In A Day*.

Analysis for three numerical variables

Analysis of data with three numerical variables is mostly a special case of analysis with ‘n’ numerical variables (see below) except that there are a couple of special pictures you can draw, as discussed in section 3.9 of *Statistics In A Day*.

EXERCISE 12.8 **Work through section 3.9 of *Statistics In A Day*.**

Analyses for ‘n’ variables

Analysis for ‘n’ categorical variables

As discussed previously, analysis of data involving ‘n’ variables simultaneously (be they categorical or numerical) becomes increasingly harder as ‘n’ increases.

EXERCISE 12.9 **Work through section 3.11 of *Statistics In A Day*. Reproduce the work for example 3.50 on your own.**

In doing example 3.50 you can already see what happens with ‘n’ categorical variables—the pivot tables (which are the standard tool for summarising ‘n’-categorical-variable data) can quickly become unmanageable. For three variables (as in example 3.50) it’s not too bad, but it becomes more and more complicated as the number of variables increases. I’ve seen tables with five variables in them, but that’s probably about the upper limit.

There are tests that can be applied to ‘n’ categorical data, but they will not be discussed in this Study Guide.

Analysis for one numerical and ‘n’ categorical variables

Analysis for this combination of variables uses pivot tables, like for ‘n’ categorical variables, except that summary measures (like the mean) of the numerical variable can appear in the cells of the table.

EXERCISE 12.10 **Work through section 3.12 of *Statistics In A Day*. Reproduce the work for example 3.55 on your own.**

Analysis for ‘n’ numerical variables

If kept simple, the analysis of ‘n’ numerical variables is actually a straightforward generalisation of the analysis for two numerical variables. For two numerical variables the simplest analysis usually involves fitting a straight line to data in a scatterplot. The same idea extends to fitting (linear) equations to ‘n’-numerical-variable data. The analysis is then called multiple regression. Simple results to do with this can be found on the Multiple Regression sheet in the XLStatistics workbook nNum.

Work through section 3.13 of *Statistics In A Day*. Reproduce the work for example 3.59 on your own.

Of course, it needn't stop there and there are many other types of analyses that can be carried out for 'n'-numerical-variable data. In example 3.59, for example, we looked for groups in the data. It was only done graphically in that example, but there are many types of 'cluster' analyses that could be used to extend the work. There are also many techniques for combining variables. There are special analyses for time-ordered data (where one of the variables is time). There are special analyses that handle non-linear problems. We will not discuss these techniques here.

Review

Summary

In the discussion and examples presented in this topic you saw many different types of data analysis and I suppose it's possible that you are now confused by the huge variety. Don't be! Remember, the analysis appropriate for a given problem is dictated to a very large extent by the number and type of variables involved. This fact doesn't necessarily make data analysis easy, but it does make it manageable. It should help eliminate one of the most common questions students (or anyone wanting to do some analysis for that matter) ask: 'What analysis do I use here?'. If you follow the guide and use a little common sense you cannot go too far wrong.