APNU1030 Data Handling Exercise

Introduction

In this exercise you will work on a quality control/quality assurance exercise using statistical analysis. Read **all** of this handout carefully before you start work. The exercise represents a full day of work for a professional person. Allowing for the fact that you may be less familiar with the work this exercise may take up to ten hours.

You may be tempted to avoid the computer and try to do the data processing on paper. Please don't, for your own sake, as you will never get it finished and it will be impossible to avoid data entry mistakes on your calculator. If you make one slip with a decimal point the result will be out and you will have to work though all the calculations again. Check your progress with your tutor/demonstrator as you proceed if you wish.

The Scenario

You are supervising a food analysis laboratory. Your client is a margarine manufacturer and their brand of margarine has nutritional information printed on the package. Your job is to take samples from several batches of margarine, measure the amount of unsaturated fat and indicate whether this is consistent with the claims on the packaging. **The packaging states that the fat component of the product is 53% saturated and 47% monounsaturated.**

You will include some data validation in order to convince your client of the reliability of your results. You have four technicians who are trained to make measurements and you will use forty replicate measurements on each of the five batches of margarine. (The client is very determined to get good results.) There are several questions you must answer.

a) Are any of your technicians introducing consistently wrong results?

b) Are the batches significantly different from each other?

c) If the answer to b is 'no', is the 'global' mean from all batches taken together consistent with the manufacturers stated value allowing for your level of measurement error?

In order to answer a) and b) you must arrange for each technician to make some measurements from each of the batches. If you assigned a batch to one technician you wouldn't be able to see if the different means were due to the margarine being genuinely different in different batches or if certain technicians were producing biased results.

So each technician will make ten replicate measurements for each of the five batches of margarine. This means that each technician will make fifty measurements and each batch will be tested forty times. Because you have a lot of data the process would be very time consuming if done on paper with a pocket calculator. You will use a spreadsheet to help you come up with fast and reliable results.

Your technicians will provide the data already entered into the spreadsheet and you will have to add the formulae for calculating the results.

Obtaining the Data

The technicians will send you the data in an Excel spreadsheet file attached to an electronic mail message. (Put in a request for data by mailing your APNU1030 tutor.) To get the file onto your disc wait for a message from the technicians. The message will indicate that there is an attachment to the message. Click on the attachments button of the message window.

You will be shown a list of files attached to the message. There will only be one file in the list. Click on the filename to select it and use the save button. Another box pops up which allows you to choose which disc to put the file on. Unfortunately this box is a bit different to what you are used to. Directories and disc drives are listed together in a box marked 'Directories'. Scroll until you find your disc drive indicated as, for example, [-a-] and double click on it. Now you are on the correct disc drive you can click on the OK button to save the file there.

Calculations

The margarine you are testing contains 98% lipid. This is manufactured using an oil with three identical monounsaturated acyl groups, which has been partially hydrogenated in order to make it solid at room temperature. The percentage of saturated fat is taken to be 100% if all the available double bonds are hydrogenated and 0% if none of the double bonds have been hydrogenated. You may take the molecular weight of the saturated and unsaturated lipids to be both approximately 880 gmol-1 (The addition of two hydrogen atoms makes little difference to the molecular weight.) The values supplied by your technicians give the number of double bonds in mmol g^{-1}. The packaging states the percentage of fat which is monounsaturated.

You can analyse the raw data to answer the first two questions but to compare your mean with the stated value on the packaging you will either convert your calculated mean to a percentage or the stated percentage to millimoles of double bonds per gramme.

You will have to arrange data from the technicians' spreadsheet in different ways for the different calculations so leave that spreadsheet intact and simply copy blocks of data into a separate sheet for each calculation.

The Global Mean

It is quite quick to test your 'grand mean' from all the data against the stated value on the packaging so you can start with this. You can go on to look at the data more closely at a later stage to test the quality assurance side of the project. If you are lucky your team will be up to scratch and this test of the mean will stand up. If not, you may have to repeat it!

Testing the Mean

Copy all the data onto a separate new spreadsheet as a block. It doesn't matter what order it goes in as long each datum is pasted in only once. Give the block of data the name 'data'. (See your course manual.) Use some convenient cells on the spreadsheet to enter formulae that will give you the number of datum points, the mean and standard deviation.

=COUNT(data)

=AVERAGE(data)

=STDEV(data)

Name the cells with these results in 'N', 'mean', and 's' respectively. You will either have to convert these figures to percentage saturation using more cells on the spreadsheet or use some cells to convert the manufacturer's stated value into mmol g^{-1}. This will be necessary in order to make the comparison.

You now have a mean calculated from a large number of measurements which you can compare with the stated value on the margarine's packaging. How do you do it? First you need to estimate your confidence in the measured mean. This is the standard error of the mean. This standardised measure of the experimental error depends on the spread of the

measurements but also the number of measurements used to estimate the mean.

Divide your standard deviation by the square root of the number of measurements used to estimate the mean. This is your standard error. You can see that as you make more measurements the standard error decreases and you are therefore more confident about your mean value.

Standard error is a measure of the range of likely values for the **actual** saturated fat content of the margarine. There is an 84% chance (worked out from the normal distribution) that the **actual** value lies within the range of the **estimated** mean plus or minus the standard error. If you compare your mean value with the expected you can test the hypothesis that the measured value is consistent with the expected value. In other words, what is the probability that your measured mean could have resulted from measurements of saturated fat content, with normal experimental error, of margarine of the stated saturated fat content.

If the probability is very low then you will have satisfactorily proved that the level in the supplied samples is not as stated on the packaging.

Calculate the deviation of your measured mean from the expected (hypothetical) mean (a simple subtraction). You must scale this in terms of the standard error. Simply divide by the standard error. **You now have a measure of how far the expected value is from the measured mean scaled according to your confidence in the measurement. A deviation between two means (measured or hypothetical) scaled in this way is usually called 't'.** Now you need to know the probability.

When you look at the deviation of a **measurement** from a mean the deviation is scaled to **standard deviation.** When you look at the deviation between two means the deviation is scaled to **standard error.** In either case it is the normal distribution itself which relates the scaled deviation to a probability of reaching such a deviation by chance alone. However, because the normal distribution is described partly by the standard deviation and because this is only estimated by your data, if you have small numbers of datum points the distribution is adjusted. That is why the t distribution is used. The t distribution is an adjusted version of the normal distribution for different numbers of data points. The deviation you calculate is referred to as t.

The calculation of p from d is complex to perform on paper and the calculation of p from t is more complex because you must involve the number of measurements as well, but the spreadsheet program you are using has a simple function for calculating either. *(If you were doing this on paper you would probably accept that, with the large amount of data you have, you have such a good estimate of standard deviation that a normal distribution would be acceptable - the deviation is called d instead of t.)*

On a spreadsheet enter a formula like this;

=TDIST(t, count-1, TRUE)

The TRUE in the function makes it perform a two tailed test. This is needed because you are concerned about the measurement being too low as well as too high. If the probability is less than 0.05 you can confidently reject the value stated on the packaging.

Comparing Means

Assume initially that the various batches of margarine are all the same at first. It is safe to look at the four mean values each of your technicians produced (using all the batches) and compare them. Are they consistent with each other? You may have already met t tests as a way of comparing two means. Analysis of variance is an alternative way of comparing means. It allows you to compare two means like a t test but can be extended to compare any number of means. Analysis of variance is also a little easier to understand than the various t tests.

Variation

Variation between individual values in our situation could be due to unavoidable measurement error or due to consistent error introduced by the technician (we are ignoring any possible variation due to batches being different for the moment.) So if you take all 240 measurements and calculate the variance these two forms of variation will both be contributing to it. If you could separate the two contributions and compare them it will show you if consistent error, due to the technician, is real and significant

That is what analysis of variance is about. You calculate the two variance values which contribute to the total variance. If variation introduced by the technician is significant then you will have demonstrated that the different means calculated by different technicians are really at odds with each other. The technique is called one factor analysis of variance because, apart from the error in the measurement process you are looking at how one factor (technician bias) affects the measured means.

You should remember that variance is calculated as µ § divided by the degrees of freedom (usually N-1). Because of this division the component variance values are not *directly additive* but the component µ § values **are** additive. For this reason you work with µ § initially and then work out the different variance values.

Sum of Squares of Deviations from the Mean

µ §

This value, when calculated by hand or pocket calculator, is easier to work out in a different way;

µ § = µ §

This second method gives an identical result but doesn't require you to work out the mean in advance and then subtract each datum from it. **If you use a computer to work out the value using a built in function you need not worry about how it is done.** For this reason the first formula is used throughout this text as it describes the statistic better, i.e. it shows how much your data varies from its own mean. The square makes sure that each individual deviation is positive so low values contribute as much as high values.

Finding µ § is one step away from the variance (divide by degrees freedom) and two steps away from the standard deviation (find square root of variance).

In some text books this value is called 'sums of squares' for short. Don't be confused by this.

Comparing Technicians' Means

You need to get four columns of data together on your spreadsheet each containing the sixty measurements made by each technician. There is no need to keep different batches separate for this calculation so simply cut and paste to produce four long columns. Leave a space of ten or more rows at the top of the sheet to allow space for the statistical results.

Graphical Method

Before you do the statistical analysis (one factor ANOVA) you can use a graphical method to compare technicians. Produce a histogram for each technician's set of data. This involves

counting data points that fall in certain ranges and then graphing the frequency totals in a column chart. Fortunately the spreadsheet program you are using has a facility for automating this. What you will have to do though is set up a table of 'bins'. A bin is a category for counting. It is defined by giving the lowest acceptable value for the bin. A datum between that value and the value for the next bin up will be counted. A data point *equal* to the bin level will also be counted.

So, to the side of your data set up a column of figures starting with 0 and increasing in 0.2 steps to beyond the highest measurement you have. Now, select the column of data for your first technician and choose the analysis tools command on the options menu. You will see that the input range box has already been filled in with your data selection. You must now click on the bin range box and enter the range of cells where your column of bin values will be found. (In excel dialogue boxes, instead of typing a cell reference you can click on the spreadsheet to indicate a region - this speeds things up.) Then click on the output box and enter the cell where you want the results to be placed. There are three check boxes on the dialogue box - Pareto and Cumulative Percentage should be switched off but the Chart Output option can be switched on. Click on the O.K. button to obtain the results. You may have to change the title of the chart before you save it to disc.

Repeat this for each technician. You now have a graphical way of comparing data sets; can you pick out any set of data as being significantly biased compared to the others? To really be certain you will have to work out the statistics.

Numerical Method

It will be useful to give names to areas of data before you start. Select all four columns of data and use the appropriate command to name this 'data'. Name each column 'datax', 'datay', 'dataz' and 'dataw'.

To start with you want to know the global degrees freedom, mean, sums of squares of deviation from the mean and variance. Use a space to the right of the data to enter formulae for these and label them using neighbouring cells.

=COUNT(data)-1

=AVERAGE(data)

=DEVSQ(data)

For variance you must enter a formula which refers to two cells above and does a division.

Next you must start looking at the separate means. Above each column enter a formula of the type;

=AVERAGE(datax)

Have a look at the means. They will certainly be each different to the overall mean but is this due to the normal error associated with making the measurements or due to the technicians introducing consistent error? This is where the analysis of variance comes in. Imagine that a technician always makes a measurement 2.5 units too high. The normal error will give measurements spread out either side of a mean shifted 2.5 units up instead of either side of the real mean. So if you look at deviation from the technicians own mean you will see the variance due only to the normal measurement error without the extra deviation due to the 2.5 unit shift. So this is how to calculate the error variance;

In cells above the four data columns calculate individual sums of squares of deviations from the individual mean, e.g.;

=DEVSQ(datax)

Somewhere on your spreadsheet add up the four separate values. You now have the

contribution to µ § from measurement error alone.

You also need to know the contribution due to technician bias. There is more than one way to find this out. Imagine that each of the measurements a technician made were set to the technicians own mean value, i.e. that each technician had made fifty identical measurements. If this were done you would have eliminated variation due to normal error and the variation left would be down to the technicians. You might then repeat the µ § calculation for this block of data. However there is a simpler way to do it without making up a big table of figures. If you calculate µ § for **the row of four means** you can simply multiply by sixty to get the full figure.

So calculate

=DEVSQ(*insert an appropriate reference to the four means here*)*60

You now have three values for µ § You should see that the two contributing values (for error and for technician bias) can be added together to make the total µ § for all the data. So in reality you could choose to work out any two values and find the other by subtraction or addition as appropriate. On the other hand, by calculating all three values you have got a way of checking your calculations.

What next?

You must find the variance for error and variance for technician. To do this you need to divide by degrees freedom. You have to be careful here. Degrees freedom for the error is (???) , degrees freedom for technician bias is (???).

Now is the time to compare the two contributory variance values. Divide error variance by technician variance. This value is called f. You have to obtain from this the probability that variation between technicians' means is due to chance alone. **Do not look up f in a statistics table.** Statistical tables are provided only because the necessary calculation is too complex and time consuming to perform on paper using a pocket calculator. You have a computer to help you and the spreadsheet program is capable of the doing the proper calculation for you! The function must account for both degrees of freedom so it will take the form;

=FDIST(*reference to your f value*, *reference to error d.f.*, *reference to tech. d.f.*)

This gives you p exactly - it is not an estimate. If the value is low then the null hypothesis (that variation between technicians' means is due to chance) is unlikely. If it is lower than 0.05 then you can confidently discount the null hypothesis. If your value is greater than 0.05 then you must assume that without any further evidence to the contrary your technicians are working reliably.

If there is a significant discrepancy between the means what can you do to satisfy the client? Look at the four means and the histograms you produced. Can you spot a single technician who is at odds with rest? If you can't, you will have to carry on regardless, but if there is one technician who's mean is obviously deviating from the other three you can choose to disregard the data from that person. Repeat the ANOVA using the remaining three technicians' data. Are the three means still at odds with each other? Repeat the test of the mean against the stated value on the packaging using an overall mean taken from the three technicians' data.

Comparing Batches of Margarine

You can use exactly the same technique (one way analysis of variance) to find out if batches of margarine have mean values which are significantly different. Of course you will have to arrange the data in columns differently, so start with a fresh spreadsheet, and remember that

the degrees freedom will be different. If you have chosen to discount the data of one of your technicians you will be using a smaller data set for each batch.

If the batches are significantly different you will have to compare each batch against the value on the packaging separately instead of the overall mean.

Routine use of ANOVA

The spreadsheet program you are using has a built in facility for ANOVA. You have set the calculations up manually so that the client could see all the intermediate values but using the facility provided you can get instant results.

You must still arrange the data into columns as before, but this time select the entire block and choose the Analysis Tools command on the Options menu. A dialogue box will ask you which analysis you wish to use - choose one factor ANOVA. A second dialogue box will ask you questions about your data and how you want the test done. Use the help button on this box to find out what the options do.

Two factor ANOVA

In some studies you may choose to treat the two factors affecting the mean together. In other words look at the technician bias at the same time as batch variation. There will now be more than two contributors to variation in the measurements. There is measurement error, technician bias and batch variation. In addition there may also be an interactive variation if a technician introduces a *different* bias on *different* batches! This seems very unlikely for our study but is a distinct possibility in less clear cut studies.

There is no need to do this type of analysis for your client but you should be aware of the possibility of using it in the future.

The Report

You need to word process a report for your client. Tell them whether the margarine fits in with the value stated on the packaging. If you had to do this with batches of margarine separately explain why. If you had to ignore data from a technician explain why you did this - your data set will be smaller than that paid for by the client so your reasoning must be well thought out. At all points state the statistical test used and the key statistical results.

Quality Assurance Audit

Your company has a policy for running quality assurance audits. Your study has been picked out for audit. You must provide all the documents connected with the project, every spreadsheet file and a copy of the report, before sending off to the client. (Using Email may be a bit of strain on the computer system with all these files so check with your course tutor about this first.)