ACE2046 Quantitative Techniques
Statistical Computing

M. Farrow
School of Mathematics and Statistics
Newcastle University

Semester 1, 2012-13

2 Inference for Normal Populations: Student's t

2.1 Introduction

Last week we looked at samples of data rather informally. Often we wish to have a more formal assessment of, for example, the strength of evidence against a hypothesis. There are many different cases with different assumptions etc. In this course we will mainly restrict our attention to cases where we assume that the data are drawn from a *normal distribution*.

2.2 Normal distribution

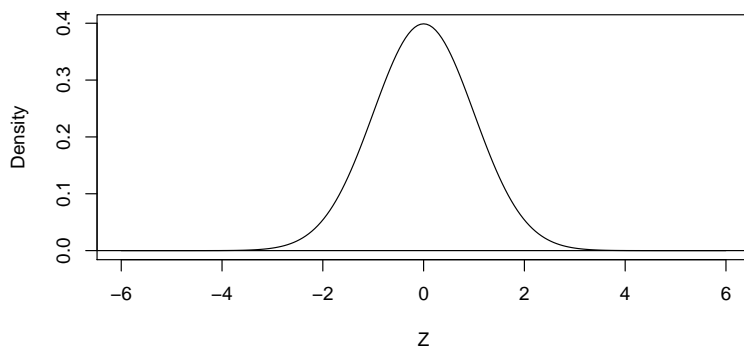Recall that the Normal distribution is symmetric and bell–shaped (see Figure 7). Obviously, the

Figure 7: Probability density function for the (standard) normal distribution

distribution of our data is never going to look as 'perfect' as this but, as long as it *resembles* this shape, we can make inferences on the population mean μ of our distribution by assuming that the sample mean $\overline{X}$ has a normal distribution. By using the mean and variance of our data, we can fit the normal distribution to it. (Imagine drawing a bell-shaped curve over the graphs you have).

The key assumptions are:

- The data are independent;

- the data are **normally distributed**.

We can check these assumptions in `Minitab` .

2.3 The one–sample t–test

Here, from a **single** population we draw a **single** sample, and we estimate the population mean μ with the **sample mean**, $\overline{x}$. We would then like to test how convincing a proposal for the population mean is, based on the information from our sample.

Procedure

- Check the normality assumption using `Minitab` by using a **normal probability plot**;

  ```
  Graph -> Probability Plot -> Single
  ```

- Set up your **hypotheses**: both the *null* and *alternative*. In this test,

  - the null is $H_0 : \mu = \mu_0$, where μ_0 is a proposed value for the population mean. So, for example, we could have the null hypothesis $H_0 : \mu = 10$

– If you have no reason to believe the population mean should be greater, or less than, μ_0, then the alternative is $H_A : \mu \neq \mu_0$ (**two–sided test**).

– In other cases we might want to use a **one-sided** alternative hypothesis which could be, for example, $H_A : \mu > 10$ or $H_A : \mu < 10$. For example, we might be concerned to test the null hypothesis that the mean amount of cheese in nominally 100g packs is 100g against the alternative that it is less than 100g.

- Calculate the **test statistic**, which is

$$t_{n-1} = \frac{\overline{x} - \mu_0}{s/\sqrt{n}},$$

on $n-1$ degrees of freedom. The quantity $s/\sqrt{n}$ is called the *standard error*. So we measure how many standard errors there are between $\bar{x}$ and μ_0. Under the null hypothesis, the test statistic has a Student's t-distribution on $n-1$ degrees of freedom (a t_{n-1} distribution). This distribution is symmetric and bell-shaped but not quite the same as a normal distribution. Its exact shape depends on the number of degrees of freedom $n-1$.

- The properties of the t-distribution are well known. We can calculate a p-value. This is the probability, when the null hypothesis is true, of observing a test statistic at least as extreme (as far from the null hypothesis value) as the observed value. If the p-value is small we say that the result is *significant* and take this as evidence for rejecting the null hypothesis in favour of the alternative.

- Typically we set a cut-off (or p value) of 0.05.

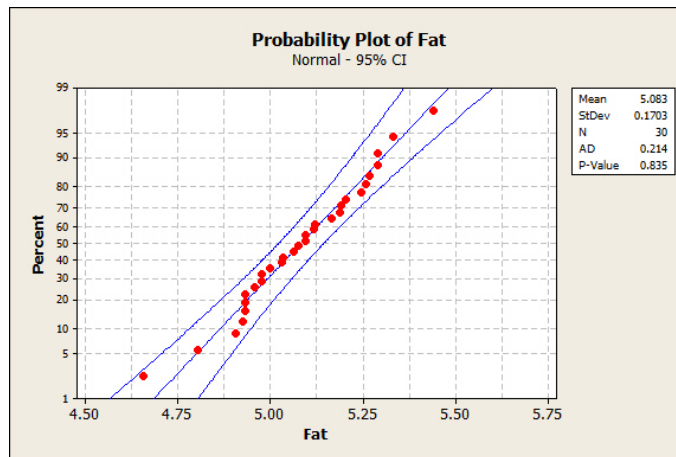We can do this by hand; but why? `Minitab` can do it for us!

Figure 8: Normal probability plot for % fat data.

Example

Thirty specimens of a processed food product are analysed and the percentage fat determined.

Figure 8 shows the normal probability plot. We see that there is no reason to suppose that we should not use a t-test here.

Suppose first that it is claimed that the product contains "5% fat" and we simply wish to test whether there is evidence that this claim is false.

In Minitab we use

```
Stat -> Basic Stats -> 1-sample t
```

We tick the box for `Perform hypothesis test` and enter 5 for the `Hypothesized mean`.

```
One-Sample T: Fat

Test of mu = 5 vs not = 5


Variable   N    Mean   StDev  SE Mean       95% CI          T      P
Fat       30  5.0830  0.1703   0.0311  (5.0194, 5.1466)  2.67  0.012
```

Notice that `Minitab` has calculated all the necessary statistic for us, including $p = 0.012$. To reject the null hypothesis, it is usually thought that we need a p–value less then 5%, or 0.05. Since $p = 0.012$ which is less than 0.05, we reject our null hypothesis. The data suggest that the mean fat% is not 5.00. In fact, since the sample mean is greater than 5 we can conclude that the data suggest that the population mean μ is greater than 5.

Recall that **confidence intervals** for μ can also be calculated. Again, we do not need to calculate this by hand – `Minitab` gives us this interval in the output from the test. The default is a 95% confidence interval, though others can be set. This means that if we were to take 100 independent samples and calculate a 95% confidence interval, about 95 of these intervals would contain the actual population mean μ. From the `Minitab` output, we can see the 95% confidence interval is (5.0194, 5.1466), which supports our decision to reject the null hypothesis, since this interval does not contain $\mu_0 = 5$.

Suppose now that it is claimed that the mean fat % is "not more than 5." In this case we might want to test the null hypothesis that $\mu = 5$ against the one-sided alternative that $\mu > 5$. In this case we need to click on `Options` and select the approrpiate alternative: `Greater than`. Here is the result.

```
One-Sample T: Fat

Test of mu = 5 vs > 5


                                      95% Lower
Variable   N    Mean   StDev  SE Mean    Bound     T      P
Fat       30  5.0830  0.1703   0.0311   5.0302  2.67  0.006
```

Again we can reject the null hypothesis. In fact the p-value is half as big as before since we are not considereing the possibility of a sample mean less than the null value.

2.4   The two–sample t–test

Often we wish to compare the means of two normal populations. As before, we assume

- Normality

- Independence

However, for the two–sample t–test we have the further requirement of

- equal variance

between the two populations. (Note that `Minitab` provides functions to deal with unequal variances as well).

Procedure

- Check the normality assumption for both populations – i.e. produce a Normal probability plots for each dataset. Check also the assumption of equal variances by looking at the sample variances for each sample taken. They do not have to be exactly equal, of course. we would not expect this.

- Set up your hypotheses. Usually, $H_0 : \mu_1 = \mu_2$. If you have reason to believe one population mean should be greater than the other, then the alternative could be $H_A : \mu_1 > \mu_2$, for example. Otherwise, $H_A : \mu_1 \neq \mu_2$;

- Calculate the test statistic, which is

$$t_{n_1+n_2-2} = \frac{\overline{x}_1 - \overline{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

  on $n_1 + n_2 - 2$ degrees of freedom;

- Compare this value to those from tables and obtain a p–value;

- Retain or reject the null hypothesis.

Again, `Minitab` can calculate all the necessary statistics for us!

Example: Cholesterol and diet

Here, the two–sample t–test is demonstrated by comparing the cholesterol level of people who have 6 servings per day of grain, against those who have 0 servings per day. First we should do a probability plot:

    Graph -> Probability Plot -> Multiple

Insert the Graph and Grouping columns to get Figure 9 Again, all points on the plot lie within
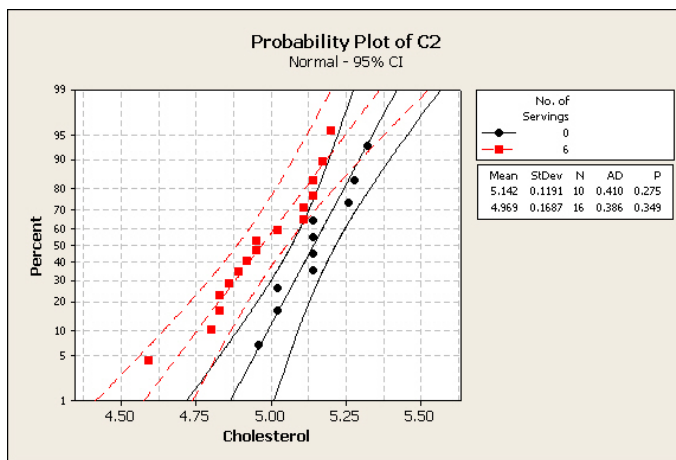

Figure 9: Normal probability plot for cholesterol data

the 95% confidence bands, and so we can safely assume the data are normally distributed. The other assumption of constant variance is easy to check in `Minitab` , which calculates the variances in each group for us.

    Stat -> Basic Statistics -> Display Desc Stats

To get

Variable	C1	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
C2	0	10	0	5.1420	0.0376	0.1191	4.9600	5.0200	5.1400	5.2650
	6	16	0	4.9694	0.0422	0.1687	4.5900	4.8375	4.9500	5.1325

We have $s_0 = 0.1191$ and $s_6 = 0.1687$, where 0 and 6 denote the number of servings. Thus, the variances are roughly the same and so the assumption of constant variance holds.

Without any expert knowledge about how grain affects the cholesterol level, we set up our hypotheses:

$$H_0 : \mu_6 = \mu_0 \quad H_1 : \mu_6 \neq \mu_0.$$

We can set this up in `Minitab` in either of two ways. We can either

16

1. have all of the cholesterol values in one column and a "group" variable (0 or 6 in this case) in another column or

2. have the two samples in separate columns.

Method 1 has been used here. In either case we use

```
Stat -> Basic Statistics -> 2-sample t
```

The output from the two–sample t–test in `Minitab` is shown below:

```
Two-sample T for C2

C1   N   Mean  StDev  SE Mean

0   10  5.142  0.119    0.038
6   16  4.969  0.169    0.042

Difference = mu (0) - mu (6)
Estimate for difference:  0.1726
95% CI for difference:  (0.0462, 0.2991)
T-Test of difference = 0 (vs not =): T-Value = 2.82  P-Value = 0.010  DF = 24
Both use Pooled StDev = 0.1520
```

Since we have obtained a p–value of 0.01, which is *less* than 0.05, we accept the alternative hypothesis. So it appears that indeed $\mu_6 \neq \mu_0$, i.e. there is a difference between the population means. So we conclude that, on average, having 6 portions of grain lowers your cholesterol. The 95% confidence interval given by `Minitab` is (0.0462, 0.2991).

2.5 Paired comparisons t–test

The paired comparisons t–test is used when we have drawn a single sample from the population, but have made two observations on each element within the sample. The assumptions and procedure are exactly the same as that for the one–sample t–test (we have, after all, only taken one sample from one population). Essentially, we are actually performing a one–sample t–test, where we are testing that the *differences* are zero.

Example: Treatment of anorexia

Seventeen girls suffering from anorexia nervosa are given a particular treatment. The measurements are their weights in kg before and after a period of treatment. A normal probability plot applied to the differences does not suggest any problem.

We use

```
Stat -> Basic statistics -> Paired t

Paired T-Test and CI: After, Before

Paired T for After - Before

              N    Mean   StDev   SE Mean
After        17   90.49    8.48      2.06
Before       17   83.23    5.02      1.22
Difference   17    7.26    7.16      1.74


95% CI for mean difference: (3.58, 10.94)
T-Test of mean difference = 0 (vs not = 0): T-Value = 4.18  P-Value = 0.001
```

We see that the result is significant and that there is therefore strong evidence that the mean weight after treatment is different from (in fact, greater than) the mean weight before treatment.

17

Student Reference Numbers

In the practicals for this week and later weeks, in at least some of the exercises, each student will have different data. In order to identify which data each student should have, each student is given a refernce number according to the table below. Please make a note of your reference number because we will use the same numbers in later weeks.

Agapiou	Kristy Marie	1
Au	Hoi Ching	2
Begin	Abigail Ruth	3
Breininger	Stella Panagio	4
Cepelis	Aivaras	5
Choo	Siew Li	6
Clarke	Timothy Oliver	7
Crawley	Michael	8
Darwin	Darwin	9
Drayton	Chloe Roseanne	10
Faleti	Hannah Ifedolapo	11
Golding	Carolina De Arouca	12
Grant	Georgia Kate	13
Hui	Angeline Kosim	14
Jenkins	Grace Evangeline	15
Linn	Sophia Campbell	16
Lo	Kin Wa	17
MacKinlay	Brittney Lauren	18
Obertelli	Georgia May	19
Rathod	Neel Harish	20
Rowland	Maisie Katharine	21
Saberioon	Ghazaleh	22
Song	Anqi	23
Tam	Wai Tsun Rachel	24
Trikerioti Chatziioannou	Danai	25
Wright	Jessica Laura Kate	26
Zawarska	Santana Anna	27

Practical 2

Instructions

1. You have an individual reference number (see previous page). You should use the data assigned to your reference number. Please write your reference number on your report as well as your name.

2. Please also mark on your report my name (Dr Malcolm Farrow) and "School of Mathematics & Statistics".

3. Answer both questions.

4. This assignment is to be submitted to the AFRD Office in the Agriculture Building, by no later than 2.00pm on Thursday 18th October.

5. Write each solution in the form of a (brief) report. This should have an introduction to the problem, a description of the analysis and a clear statement of conclusions, illustrated, where appropriate, with graphs and tables.

6. Graphs should be properly labelled with appropriate axis labels etc.

7. In your report explain (briefly) how you obtained your results using Minitab so that, if anything has gone wrong, I might be able to see where it is.

Questions

1. As part of a large study carried out in California, a number of middle-aged men were categorised according to "Behaviour type", A or B. Blood cholesterol measurements (mg per 100 ml) were made on these men. You are provided with the cholesterol measurements and the behaviour type (1 for A, 2 for B).

    The data file `behavedat.txt` can be downloaded from the module Web page (via Blackboard or otherwise) or it can be read directly into `Minitab` using

    ```
    File -> Other files -> Import special text
    ```

    entering `c1-c31` in the box marked `Store data in column(s)` and entering

    ```
    http://www.mas.ncl.ac.uk/~nmf16/teaching/ace2046/behavedat.txt
    ```

    for the file name.

    The behaviour types are in column 31. Name this column `Type` or some other suitable name.

    Your cholesterol measurements are in the column corresponding to your reference number. For example, if your reference number is 10 then your data are in column `c10`. Name your column `Cholesterol` (or some other suitable name).

    (a) Check that it is reasonable to assume normality. Use

    ```
    Graph -> Probability Plot -> Multiple
    ```

    The `Graph variable` is `Cholesterol` and the `Categorical variable for grouping` is `Type`.

    (b) Does it seem reasonable to assume equal variances?

    (c) Use a two-sample t-test to test the null hypothesis that there is no difference between the mean cholesterol values for the two behaviour types against the two-sided alternative. Use

    ```
    Stat -> Basic statistics -> 2-sample t -> Samples in one column
    ```

The `Samples` are in your `Cholesterol` column and the `Subscripts` are given by `Type` . Select `Assume equal variances` .

Report your conclusions clearly.

(d) Write down a 95% confidence interval for the true difference between the means.

2. Twenty patients were admitted to hospital for treatment for anorexia. For each patient, three BMI values were recorded, as follows.

- B_1: the patient's BMI on admission to hospital,
- B_T: the patient's target BMI,
- B_2: the patients' BMI at follow-up after discharge from hospital.

BMI is "Body Mass Index". It is the weight in kg divided by the square of the height in m.

We are interested in whether the treatment has an effect. The change in BMI is likely to be related to how far from the target the patient is. Suppose that we adopt the following model. We work in terms of $Y = \ln(B/B_T)$ where B is a BMI value and $\ln()$ denotes the natural logarithm. We propose to use a paired sample t-test to compare $Y_1 = \ln(B_1/B_T)$ with $Y_2 = \ln(B_2/B_T)$. However

$$
\begin{aligned}
Y_2 - Y_1 &= \ln(B_2/B_T) - \ln(B_1/B_T) = \ln(B_2) - \ln(B_T) - \ln(B_1) + \ln(B_T) \\
&= \ln(B_2) - \ln(B_1)
\end{aligned}
$$

so the target BMI cancels out.

Carry out a paired sample t-test and comment on your results as follows.

The data file `bodymass.txt` can be downloaded from the module Web page (via Blackboard or otherwise) or it can be read directly into `Minitab` using

```
File -> Other files -> Import special text
```

entering `c1-c31` in the box marked `Store data in column(s)` and entering

```
http://www.mas.ncl.ac.uk/~nmf16/teaching/ace2046/bodymass.txt
```

for the file name.

Your BMI measurements are in the column corresponding to your reference number. For example, if your reference number is 10 then your data are in column `c10`. The first twenty values are B_1, the BMI on admission. The last twenty values are B_2, the BMI after discharge, in the same order. Name your column `BMI` (or some other suitable name).

Column `c31` contains 1 or 2 according to whether the BMI value is "before" or "after".

(a) Calculate the logarithms of the BMI values. You can use

```
Calc -> Calculator
```

Select a column to store the logarithms. Type the expression " `ln(` " then double click on the name of your BMI column then complete the expression by adding " `)` ". Click OK. Name the resulting column `Y` .

(b) Separate the "before" and "after" values. Use

```
Data -> Unstack Columns
```

Your data are in your column `Y` . The subscripts are in column `c31` .

If you leave `In new worksheet` selected then your two new columns will appear in a new worksheet. This is actually quite convenient.

(c) Carry out a paired-sample t-test. Select

```
Stat -> Basic Statistics -> Paired t
```

20

For your first sample, select `Y_2`. For your second sample, select `Y_1`. (Doing it this way round means that the differences are "after minus before"). Select `Graphs` and tick all three boxes. Click `OK` and `OK` again.

(d) What can you conclude about the effect of the treatment?

(e) Do you think this was a good way to analyse these data? Can you think of a possible alternative or improvement? Should we be using the target BMI values in some way? Comment.

Model Solution

As a guide to how to report results, here is an example question and report.

Question

1. An experiment is conducted in a bakery. Twenty loaves are baked under standard conditions. A further twenty loaves are baked with reduced yeast and a shorter proof time. The specific volume of each loaf is measured. Interest lies in whether the treatment given to the second batch has an effect on specific volume compared to the standard conditions. The data are as follows.

Sample 1 Standard conditions					Sample 2 Reduced yeast and proof time				
521	535	525	527	535	499	469	461	466	504
537	553	540	509	544	465	465	527	462	479
542	552	461	550	533	525	464	514	451	478
523	538	532	543	528	460	484	449	459	438

(a) Use boxplots to display the data. Comment on the comparison between the samples.

(b) Check that it is reasonable to assume normality. Use

        Graph -> Probability Plot -> Multiple

(c) Does it seem reasonable to assume equal variances?

(d) Use a two-sample t-test to test the null hypothesis that there is no difference between the mean specific volume values for the two treatments against the two-sided alternative. Use

        Stat -> Basic statistics -> 2-sample t

Report your conclusions clearly.

(e) Write down a 95% confidence interval for the true difference between the means.

(f) Do you think a two-sided test was appropriate? What happens if you use a one-sided test?

Report

Introduction

The data were collected in an experiment in a bakery. Two batches of twenty loaves each were baked and the specific volume of each loaf was determined. The first batch, "Sample 1", was baked under standard conditions. The second batch, "Sample 2", was baked with reduced yeast and a shorter proof time. Interest lies in whether the treatment given to the second batch has an effect on specific volume compared to the standard conditions. The data are given in Table 2.

Analysis

Boxplots of the data are shown in Figure 10. From the boxplots it appears that the loaves baked under standard conditions tend to have greater specific volume, although the distributions do overlap. The observed spread of values in Sample 2 seems to be greater than that in Sample 1 but, in Sample 1 there is a single outlier with an unusually small specific volume.

The boxplots were produced by selecting

```
Graph -> Boxplot -> One Y with Groups
```

in Minitab.

To check the normality of the data the normal probability plots shown in Figure 11 were produced using the slection

```
Graph -> Probability Plot -> Multiple
```

in Minitab.

In the plot for Sample 1 (black dots) the outlier can be seen and its effect shows in that the gradient of the remaining points does not match that of the plotted theoretical mean and 95% bounds. Nevertheless, with the exception of the outlier, the points follows a straight line fairly closely and there seems to be little reason to reject the assumption of normality. In the case of Sample 2 (red squares) the alignment is not as straight but no points lie outside the 95% bounds. Again it seems reasonable to assume normality.

From the box plots and the sample standard deviations given with the normal probability plots, it can be seen that the sample variance is greater for Sample 2 than for Sample 1 but the ratio is not great and may be just due to sampling variation.

There is no indication that the loaves can not be treated as independent (but see Conclusions below). therefore we are prepared to assume normality, independence and equal population variances as required for the t-test which follows.

A two-sample t-test was carried out using the selection

```
Stat -> Basic statistics -> 2-sample t
```

in Minitab. The option to assume equal variances was used. The resulting output is shown in Figure 12. The T-value was 7.71 on 38 dgrees of freedom, giving $p < 0.001$ for a 2-sided test of the null hypothesis that $\mu_1 = \mu_2$ against the two-sided alternative that $\mu_1 \neq \mu_2$, where μ_1 and μ_2 are the population mean specific volumes under the Treatment 1 (standard treatment) and under Treatment 2 (reduced-yeast, shorter-proof-time) respectively. We therefore conclude that the result is highly significant so we reject the null hypothesis and conclude that $\mu_1 \neq \mu_2$. A 95% confidence interval for the difference between the population mean for Treatment 1 and that for Treatment 2 is

$$40.89 < \mu_1 - \mu_2 < 70.01.$$

It seems therefore that Treatment 2 tends to reduce the specific volume compared to Treatment 1.

It might well be felt that reducing the yeast and the proof time would not be expected to increase the specific volume and the only possibilities would be a reduction or no change. Therefore a one-sided test against the alternative $\mu_1 > \mu_2$ might be more appropriate that a two sided test. The results of such a one sided test, obtained by chaning the `Alternative` to `Greater Than` in

23

Sample 1					Sample 2				
Standard conditions					Reduced yeast and proof time				
521	535	525	527	535	499	469	461	466	504
537	553	540	509	544	465	465	527	462	479
542	552	461	550	533	525	464	514	451	478
523	538	532	543	528	460	484	449	459	438

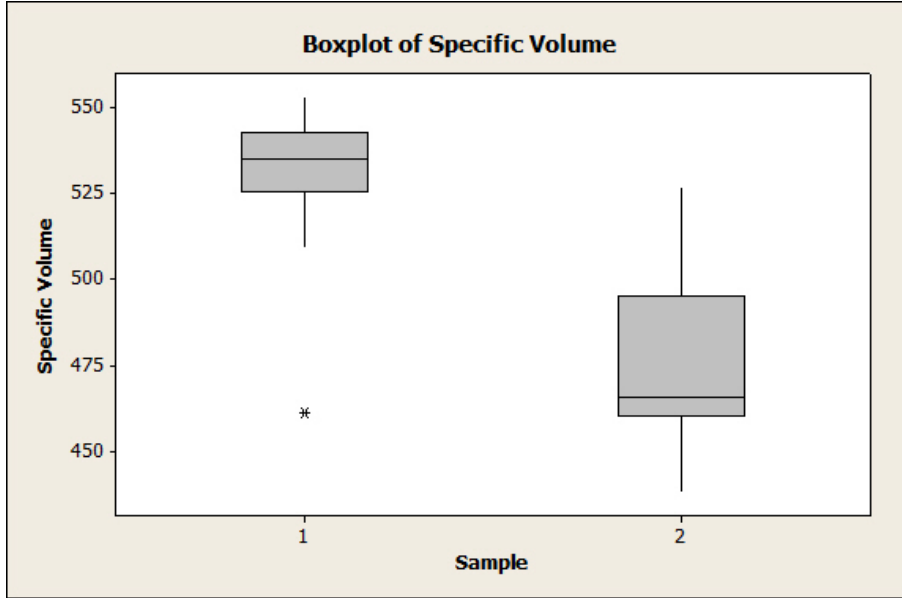Table 2: Specific volumes of loaves from bakery experiment

Figure 10: Box plot of specific volumes of loaves from bakery experiment

Minitab, are shown in Figure 13. This makes no important difference to the outcome of the significance test, since the result was already highly significant, but it does change the 95% confidence interval to the one-sided interval

$$\mu_1 - \mu_2 > 43.32.$$

Conclusion

We conclude that the evidence strongly suggests that reducing the yeast and the proof time cause a decrease in the mean specific volume of loaves. A 95% lower bound for the decrease in the mean specific volume is 43.32.

We felt that the assumptions of normality and equal variance could safely be used. For the t-test which was used we also require an assumption of independence bwteen the observations. Considering the nature of bread baking there may be a case for saying that loaves within a batch may not, in fact, be independent and it may be advisable to carry out a larger experiment involving more than one batch with each treatment.
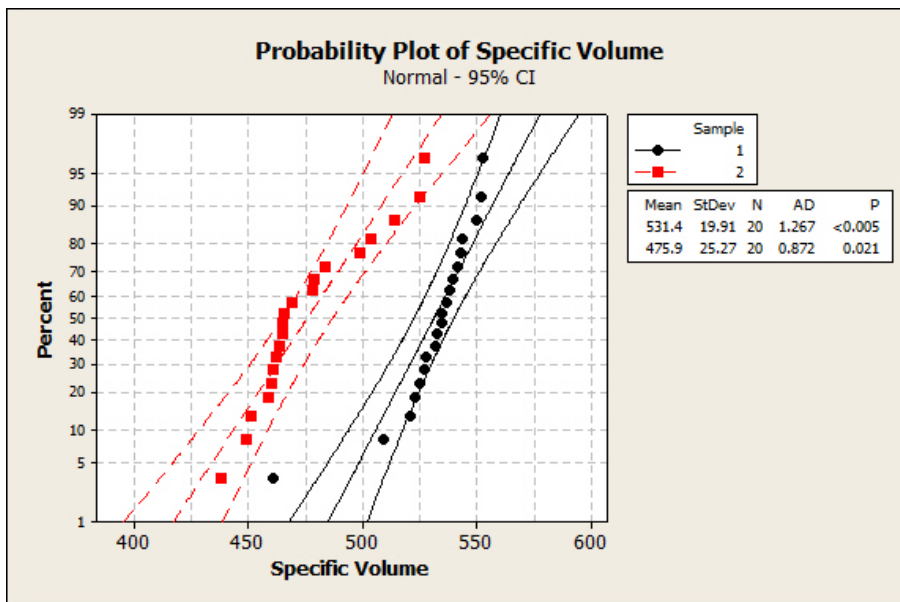
Figure 11: Box plot of specific volumes of loaves from bakery experiment

```
Two-Sample T-Test and CI: Specific Volume, Sample

Two-sample T for Specific Volume

Sample   N    Mean   StDev   SE Mean
1       20   531.4   19.9       4.5
2       20   475.9   25.3       5.7


Difference = mu (1) - mu (2)
Estimate for difference:  55.45
95% CI for difference:  (40.89, 70.01)
T-Test of difference = 0 (vs not =): T-Value = 7.71   P-Value = 0.000   DF = 38
Both use Pooled StDev = 22.7479
```

Figure 12: Minitab output for two-sample t-test in bakery experiment

```
Two-Sample T-Test and CI: Specific Volume, Sample

Two-sample T for Specific Volume

Sample   N    Mean   StDev   SE Mean
1       20   531.4   19.9       4.5
2       20   475.9   25.3       5.7


Difference = mu (1) - mu (2)
Estimate for difference:  55.45
95% lower bound for difference:  43.32
T-Test of difference = 0 (vs >): T-Value = 7.71   P-Value = 0.000   DF = 38
Both use Pooled StDev = 22.7479
```

Figure 13: Minitab output for two-sample t-test in bakery experiment, one-sided alternative