# Contents

# Chapter 1

# Chemometrics

## 1.1 Statistical Assumptions

## 1.2 Overview

- Normal probability plot
- Outliers
- dixon test
- Grubbs test

## 1.3 Grubb's Test

Grubb's Test for Detecting Outliers Statisticians have devised several ways to detect outliers. Grubbs' test is particularly easy to follow. This method is also called the ESD method (extreme studentized deviate). The first step is to quantify how far the outlier is from the others. Calculate the ratio Z as the difference between the outlier and the mean divided by the SD. If Z is large, the value is far from the others. Note that you calculate the mean and SD from all values, including the outlier.

Since 5

When analyzing experimental data, you don't know the SD of the population. Instead, you calculate the SD from the data. The presence of an outlier increases the calculated SD. Since the presence of an outlier increases both the numerator (difference between the value and the mean) and denominator (SD of all values), Z does not get very large. In fact, no matter how the data are distributed, Z can not get larger than, where N is the number of values. For example, if N=3, Z cannot be larger than 1.155 for any set of values.

Grubbs and others have tabulated critical values for Z which are tabulated below. The critical value increases with sample size, as expected.

If your calculated value of Z is greater than the critical value in the table, then the P value is less than 0.05. This means that there is less than a 5% chance that you'd encounter an outlier so far from the others (in either direction) by chance alone, if all the data were really sampled from a single Gaussian distribution. Note that the method only works for testing the most extreme value in the sample (if in doubt, calculate Z for all values, but only calculate a P value for Grubbs' test from the largest value of Z.

Once you've identified an outlier, you may choose to exclude that value from your analyses. Or you may choose to keep the outlier, but use robust analysis techniques that do not assume that data are sampled from Gaussian populations.

If you decide to remove the outlier, you then may be tempted to run Grubbs' test again to see if there is a second outlier in your data. If you do this , you cannot use the same table.

### 1.3.1 Critical values for Z

Calculate Z as shown above. Look up the critical value of Z in the table below, where N is the number of values in the group. If your value of Z is higher than the tabulated value, the P value is less than 0.05.

### 1.3.2 Computing an approximate P value

You can also calculate an approximate P value as follows.

N is the number of values in the sample, Z is calculated for the suspected outlier as shown above. Look up the two-tailed P value for the student t distribution with the calculated value of T and N-2 degrees of freedom. Using Excel, the formula is =TDIST(T,DF,2) (the '2' is for a two-tailed P value).

Multiply the P value you obtain in step 2 by N. The result is an approximate P value for the outlier test. This P value is the chance of observing one point so far from the others if the data were all sampled from a Gaussian distribution. If Z is large, this P value will be very accurate. With smaller values of Z, the calculated P value may be too large.

### 1.3.3 Dixon's Q test

In statistics, Dixon's Q test, or simply the Q test, is used for identification and rejection of outliers. This test should be used sparingly and never more than once in a data set. To apply a Q test for bad data, arrange the data in order of increasing values and calculate Q as defined:

$$Q = \frac{\text{Gap}}{\text{Range}} \tag{1.1}$$

Where gap is the absolute difference between the outlier in question and the closest number to it. If $Q_calculated > Q_table$, then reject the questionable point.

## 1.4 Overview of experimental design

Introduction Analysis of variance (ANOVA) is a popular tool that has an applicability and power that we can only start to appreciate in this course. The idea of analysis of variance is to investigate how variation in structured data can be split into pieces

associated with components of that structure. We look only at one-way and two-way classifications, providing tests and confidence intervals that are widely used in practice.

- Two-way ANOVA without interactions.
- Two-way ANOVA with interactions.
- Two-way ANOVA with replicates
- Three-way factorial design.

## 1.5 MA4605: ANOVA

We compute the test statistics $F = 62/3 \sim 20.7$ while the 95% quantile of F distribution with 3 and 8 degrees of freedom is given as

```
>qf(0.95,3,8)
4.066181
```

We clearly see that the test informs us about a significant difference between the means. But which means are different?

The least significant difference method described in Section 3.9.

We compute the least significant difference $s\sqrt{2/n} \times t$, where $s^2$ is within sample estimate of variance and $t$ is the 97.5% quantile of Student-$t$ distribution with $h(n-1)$ degrees of freedom.

```
>sqrt(mean(s))*sqrt(2/3)*qt(0.975,8)
# 3.261182
>m=apply(x,1,mean)
>m
#[1] 101 102 97 92
```

The associated degrees of freedom: for within-sample $h(n-1)$ (in our example $4 \times 2 = 8$), for between-sample $h-1$ (in our example 3). Total number of degrees freedom $hn-1$ and we see $hn-1 = h(n-1)+h-1$.

But there is more then the relation between degrees of freedom. Namely SST = SSM + SSR; where

WRONG

$$SST = \sum_j \sum_j (x-\bar{x})^2 \tag{1.2}$$

$$SSM = \sum_j \sum_j (x-\bar{x})^2 \tag{1.3}$$

$$SSE = \sum_j \sum_j (x-\bar{x})^2 \tag{1.4}$$

$$\tag{1.5}$$

```
x=c(102,100,101,101,101,104,97,95,99,90,92,94)
factors=c(rep("A",3),rep("B",3),rep("C",3),rep("D",3))
res=aov(xfactors) anova(res)

Analysis of Variance Table Response:

x   Df Sum Sq Mean Sq   F value Pr(>F) factors 3 186 62
20.6670.0004002 *** Residuals 8 24 3
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
```

## 1.6 Weighted Regression

**Homoscedasticity** - the standard deviations of y-observations from the straight line are the same independently of the underlying x-observations.

**Heteroscedasticity** - the standard deviations of y-observations depend on the underlying x-observations.

In the first case, standard regression analysis should be performed, while in the second the weighted regression is more suitable.

```
>Conc=c(0,2,4,6,8,10)
>StDev=c(0.001,0.004,0.010,0.013,0.017,0.022)
>Abs=c(0.009,0.158,0.301,0.472,0.577,0.739)
>n=length(Conc)
>weights=StDev(-2)/mean(StDev(-2))
>wreg=lm(AbsConc,weights=weights)
>reg=lm(AbsConc)
>summary(wreg)
```

It is often convienent to express the regression analysis using ANOVA table. The following equation is the basis for such representation

It is often shortened to SST = SSLR + SSR; where SST is referred

to as the total sum of squares, SSLR is the sum of squares due to linear regression (within regression), SSR is the sum of squares due to residuals (outside regression).

### 1.6.1 R square

$R^2$ is a measure of variation explained by regression.

The following coefficient has a natural interpretation as amount of variability in the data that is explained by the regression fit: $R^2 = SSLR/SST = 1 - SSR/SST$.

A similar interpretation is given to the adjusted coefficient $R^2_{adj}$ which is given by $R^2_{adj} = 1 - MSR/MST$; where MSR is the mean squared error due to residuals, and

MST is the total mean squared error.

The adjusted coefficient is accounting for the degrees of freedom used for each source of variation and is often a more reliable indicator of variability than $R^2$. $R^2_{adj}$ is always smaller than $R^2$.

## 1.7 Testing Normality

An assessment of the normality of data is a prerequisite for many statistical tests as normal data is an underlying assumption in parametric testing. There are two main methods of assessing normality - graphically and numerically.

# Chapter 2

# Linear Models

## 2.1 Multiple Linear Regression

## 2.2 Variable Selection Procedures

| | Coefficients | Std Error | t Stat | P-value |
|---|---|---|---|---|
| Intercept | 0.002107 | 0.004787 | 0.440144 | 0.678209 |
| Conc | 0.025164 | 0.000266 | 94.76047 | 2.48E-09 |

Intercept and Slope estimates are the coefficient.

- Akaike Information Criterion
- Multicollinearity

95% confidence interval for slope

- Estimate of Slope = 0.025164
- Std. Error for slope = 0.000266 from R output
- Quantiles (given) = -2.57 for Lower bound = 2.57 for Upper bound
- Lower bound = 0.025164 + (-2.57)( 0.000266) = 0.0243 ‘

- Upper bound = 0.025164 + (2.57)( 0.000266) = 0.0257
- Confidence Interval = [0.0243, 0.0257]

## 2.3 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test is defined by:
$H_0$: The data follow a specified distribution
$H_1$: The data do not follow the specified distribution

Test Statistic: The Kolmogorov-Smirnov test statistic is defined as

where F is the theoretical cumulative distribution of the distribution being tested which must be a continuous distribution (i.e., no discrete distributions such as the binomial or Poisson), and it must be fully specified

### 2.3.1 Characteristics and Limitations of the K-S Test

An attractive feature of this test is that the distribution of the K-S test statistic itself does not depend on the underlying cumulative distribution function being tested. Another advantage is that it is an exact test (the chi-square goodness-of-fit test depends on an adequate sample size for the approximations to be valid). Despite these advantages, the K-S test has several important limitations:

1. It only applies to continuous distributions.
2. It tends to be more sensitive near the center of the distribution than at the tails.
3. Perhaps the most serious limitation is that the distribution must be fully specified. That is, if location, scale, and shape parameters are estimated from the data, the critical region of the K-S test is no longer valid. It typically must be determined by simulation.

Due to limitations 2 and 3 above, many analysts prefer to use the Anderson-Darling goodness-of-fit test.

However, the Anderson-Darling test is only available for a few specific distributions.

## 2.4 The AndersonDarling test

The AndersonDarling test is a statistical test of whether there is evidence that a given sample of data did not arise from a given probability distribution.

In its basic form, the test assumes that there are no parameters to be estimated in the distribution being tested, in which case the test and its set of critical values is distribution-free. However, the test is most often used in contexts where a family of distributions is being tested, in which case the parameters of that family need to be estimated and account must be taken of this in adjusting either the test-statistic or its critical values.

When applied to testing if a normal distribution adequately describes a set of data, it is one of the most powerful statistical tools for detecting most departures from normality.

## 2.5 The Shapiro-Wilk test of normality

Performs the Shapiro-Wilk test of normality.

```
> x<- rnorm(100, mean = 5, sd = 3)
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  rnorm(100, mean = 5, sd = 3)
W = 0.9818, p-value = 0.1834
```

In this case, the p-value is greater than 0.05, so we fail to reject the null hypothesis that the data set is normally distributed.

```
>y <- runif(100, min = 2, max = 4)
> shapiro.test(y)

        Shapiro-Wilk normality test

data:  runif(100, min = 2, max = 4)
W = 0.9499, p-value = 0.0008215
```

In this case, the p-value is less than 0.05, so we reject the null hypothesis that the data set is normally distributed.

## 2.6 Analysis of Two-factor Designs

A two-factor analysis of variance consists of three significance tests: a test of each of the two main effects and a test of the interaction of the variables. An analysis of variance summary table is a convenient way to display the results of the significance tests. A summary table for the hypothetical experiment described in the section on factorial designs and a graph of the means for the experiment are shown below.

| SOURCE | df | Sum of Squares | Mean Square | F | p |
|---|---|---|---|---|---|
| T | 1 | 47125.3333 | 47125.3333 | 384.174 | 0.000 |
| D | 2 | 42.6667 | 21.3333 | 0.174 | 0.841 |
| TD | 2 | 1418.6667 | 709.3333 | 5.783 | 0.006 |
| ERROR | 42 | 5152.0000 | 122.6667 | | |
| TOTAL | 47 | 53738.6667 | | | |

### 2.6.1 Sources of Variation

The summary table shows four sources of variation: (1) Task, (2) Drug dosage, (3) the Task x Drug dosage interaction, and (4) Error.

### 2.6.2 Degrees of Freedom

- The degrees of freedom total is always equal to the total number of numbers in the analysis minus one. The experiment on task and drug dosage had eight subjects in each of the six groups resulting in a total of 48 subjects. Therefore, df total = 48 - 1 = 47.

- The degrees of freedom for the main effect of a factor is always equal to the number of levels of the factor minus one. Therefore, df task = 2 - 1 = 1 since there were two levels of task (simple and complex). Similarly, df dosage = 3 - 1 = 2 since there were three levels of drug dosage (0 mg, 100 mg, and 200 mg).

- The degrees of freedom for an interaction is equal to the product of the degrees of freedom of the variables in the interaction. Thus, the degrees of freedom for the Task x Dosage interaction is the product of the degrees of freedom for task (1) and the degrees of freedom for dosage (2). Therefore, df Task x Dosage = 1 x 2 = 2.

- The degrees of freedom error is equal to the degrees of freedom total minus the degrees of freedom for all the effects. Therefore, df error = 47 - 1 - 2 - 2 = 42.

### 2.6.3 Mean Squares

As in the case of a one-factor design, each mean square is equal to the sum of squares divided by the degrees of freedom. For instance, Mean square dosage = 42.6667/2 = 21.333 where the sum of squares dosage is 42.6667 and the degrees of freedom dosage is 2.

### 2.6.4 F Ratios

The F ratio for an effect is computed by dividing the mean square for the effect by the mean square error. For example, the F ratio for the Task x Dosage interaction is computed by dividing the mean square for the interaction ( 709.3333) by the mean square error (122.6667). The resulting F ratio is: F = 709.3333/122.6667 = 5.783

### 2.6.5 Probability Values

To compute a probability value for an F ratio, you must know the degrees of freedom for the F ratio. The degrees of freedom numerator is equal to the degrees of freedom for the effect. The degrees of freedom denominator is equal to the degrees of freedom error. Therefore, the degrees of freedom for the F ratio for the main effect of task are 1 and 42, the degrees of freedom for the F ratio for the main effect of drug dosage are 2 and 42, and the degrees of freedom for the F for the Task x Dosage interaction are 2 and 42.

An F distribution calculator can be used to find the probability values. For the interaction, the probability value associated with an F of 5.783 with 2 and 42 df is 0.006.

### 2.6.6 Drawing Conclusions

When a main effect is significant, the null hypothesis that there is no main effect in the population can be rejected. In this example, the effect of task was significant. Therefore it can be concluded that, in the population, the mean time to complete the complex task is greater than the mean time to complete the simple task (hardly surprising). The effect of dosage was not significant. Therefore, there is no convincing evidence that the mean time to complete a task (in the population) is different for the three dosage levels

The significant Task x Dosage interaction indicates that the effect of dosage (in the population) differs depending on the level of task. Specifically, increasing the

dosage slows down performance on the complex task and speeds up performance on the simple task. The effect of increasing the dosage therefore depends on whether the task is complex of simple.

There will always be some interaction in the sample data. The significance test of the interaction lets you know whether you can infer that there is an interaction in the population.

## 2.7 Regression

Unweighted regression requires that the variability of the residuals is constant over the measured range of values. (This is called homoskedasticity).

Weighted regression does not have this requirement. There may be differing variability over the range of values. (This is called heteroskedasticity).

Weighted regression requires extra information on the standard deviations of the responses so as to compute the weights.

Unweighted regression doesnt need or use any information on the response standard deviations.

Weighted regression is preferable if heteroskedasticity evident in the data

(If there is not constant variance for the residuals over the range of values)

### 2.7.1 R square

The model with the highest R2 and adjusted R2 is the preferable of all candidate models The quadratic model is the preferable model in that case.

## 2.8 Example: Poisson

A computer server breaks down on average once every three months.

- What is the probability that the server breaks down three times in a quarter?

- What is the probability that a server breaks down exactly five times in one year?

## 2.9 Example

An accounting firm wishes to test the claim that no more than 1% of a large number of transactions contains errors. In order to test this claim, they examine a random sample of 144 transactions and find that exactly 3 of these are in error.

An accounting firm wishes to test the claim that no more than 5number of transactions contains errors. In order to test this claim, they examine a random sample of 225 transactions and find that exactly 20 of these are in error.

## 2.10 Example

In the past, 18% of shoppers have bought a particular brand of breakfast cereal. After an advertising campaign, a random sample of 220 shoppers is taken and 55 of the sample have bought this brand of cereal.

Write down the null and the alternative hypothesis for this problem, and state whether it is a one tailed or two tailed test

The conventional treatment for a disease has been shown to be effective in 80% of all cases. A new drug is being promoted by a pharmaceutical company; the Department of Health wishes to test whether the new treatment is more effective than the conventional treatment.

Write down the null and the alternative hypothesis for this problem, and state whether it is a one tailed or two tailed test

## 2.11 Sample size Estimation

For a certain variable, the standard deviation in a large population is equal to 12.5. How big a sample is needed to be 95% sure that the sample mean is within 1.5 units

of the population mean?

For a certain variable, the standard deviation in a large population is equal to 8.5. How big a sample is needed to be 90% sure that the sample mean is within 1.5 units of the population mean?

# Chapter 3

# Statistical Inference

## 3.1 Confidence Interval examples

### 3.1.1 Example

A random sample of 15 observations is taken from a normally distributed population of values. The sample mean is 94.2 and the sample variance is 24.86. Calculate a 99% confidence interval for the population mean.

**Solution**

$t_(14, 0.005) = 2.977$ 99% CI is $94.2 \pm 2.977\sqrt{24.86/15}$

i.e. $94.2 \pm 3.83$

i.e. $(90.37, 98.03)$

### 3.1.2 Example 1: paired T test

| | | | | | | |
|---|---|---|---|---|---|---|
| X | 5.20 | 5.15 | 5.17 | 5.16 | 5.19 | 5.15 |
| Y | 5.20 | 5.15 | 5.17 | 5.16 | 5.19 | 5.15 |

### 3.1.3 Example 2

Seven measurements of the pH of a buffer solution gave the following results:

| | | | | | | |
|---|---|---|---|---|---|---|
| 5.12 | 5.20 | 5.15 | 5.17 | 5.16 | 5.19 | 5.15 |

Task 1: Calculate the 95% confidence limits for the true pH utilizing $R$.

Solution. We are using Student t distribution with six degrees of freedom and the following code gives us the confidence interval for this problem.

```
>x <- c(5.12, 5.20, 5.15, 5.17, 5.16, 5.19, 5.15)
>n =length(x)
>alpha =0.05
>stderr =sd(x)/sqrt(n)
>LB=mean(x)+qt(alpha/2,6)* stderr
>UB=mean(x)+qt(1-alpha/2,6)* stderr
>LB
#[1] 5.137975
>UB
#[1]5.187739
```

**example**

A survey of study habits wishes to determine whether the mean study hours completed by women at a particular college is higher than for men at the same college. A sample of $n_1 = 10$ women and $n_2 = 12$ men were taken, with mean hours of study $\bar{x}_1 = 120$ and $\bar{x}_2 = 105$ respectively. The standard deviations were known to be $\sigma_1 = 28$ and $\sigma_2 = 35$.

The hypothesis being tested is:

$$H_0 : \mu_1 = \mu_2 \qquad (\mu_1 - \mu_2 = 0) \tag{3.1}$$

$$H_a : \mu_1 \neq \mu_2 \qquad (\mu_1 - \mu_2 \neq 0) \tag{3.2}$$

In $R$, the test statistic is calculated using:

```
xbar1 <- 120
```

```
xbar2 <- 105
sd1 <- 28
sd2 <- 35
n1 <- 10
n2 <-12
TS <- ( (xbar1 - xbar2) - (0) )/sqrt( (sd1^2/n1) + (sd2^2/n2) )
TS
[1] 1.116536
```

Now need to calculate the critical value or the p-value.

The critical value can be looked up using qnorm. Since this is a one-tailed test and there is a ¿ sign in $H_1$:

```
qnorm(0.95)
[1] 1.644854
```

Since the test statistic is less than the critical value ( 1.116536 ¡ 1:645 )there is not enough evidence to reject $H_0$ and conclude that the population mean hours study for women is not higher than the population mean hours study for men.

The p-value is determined using pnorm.

Careful! Remember pnorm gives the probability of getting a value LESS than the value specified. We want the probability of getting a value greater than the test statistic.

```
1-pnorm(1.116536) # OR pnorm(1.116536, lower.tail=FALSE)
[1] 0.1320964
```

### 3.1.4 Example

Suppose that 9 bags of salt granules are selected from the supermarket shelf at random and weighed. The weights in grams are 812.0, 786.7, 794.1, 791.6, 811.1, 797.4, 797.8, 800.8 and 793.2. Give a 95% confidence interval for the mean of all the bags on the shelf. Assume the population is normal.

Here we have a random sample of size n = 9. The mean is 798.30. The sample variance is $s^2 = 72.76$, which gives a sample standard deviation $s = 8.53$.

The upper 2.5% point of the Student's $t$ distribution with n-1 (= 9-1 = 8) degrees of freedom is 2.306.

The 95% confidence interval is therefore from

$(798.30 - 2.306 \times (8.53/\sqrt{9}), 798.30 + 2.306 \times (8.53/\sqrt{9})$

which is

$(798.30 - 6.56, 798.30 + 6.56) = (791.74, 804.86)$

It is sometimes more useful to write this as $798.30 \pm 6.56$.

Note that even if we do not assume the population is normal (that assumption is never really true) the Central Limit Theorem might suggest that the confidence interval is nearly right. A larger confidence would increase the length of the interval, so we trade off increased certainty of coverage against a longer interval.

### 3.1.5 Example

Ten soldiers visit the rifle range on two different weeks. The first week their scores are: 67 24 57 55 63 54 56 68 33 43 The second week they score 70 38 58 58 56 67 68 77 42 38 Give a 95% confidence interval for the improvement in scores from week one to week two.

#### Answer

This is a case of paired samples, for the scores are repeated observations for each soldier, and there is good reason to think that the soldiers will differ from each other in

their shooting skill. So we work with the individual differences between the scores. We shall have to assume that the pairwise differences are a random sample from a normal distribution.

The differences are:

3 14 1 3 -7 13 12 9 9 -5

Effectively we now have a single sample of size 10, and want a 95% confidence interval for the mean of the population from which these differences are drawn. For this we use a Student's $t$ interval. The sample mean of the differences is 5.2, and $s^2 = 54.84$. So $s = 7.41$, and the 95% $t$ interval for the difference in the means is $5.2 - 2.26(7.41)/\sqrt{10}, 5.2 + 2.26(7.41)/\sqrt{10} = (.0.1, 10.5)$.

### 3.1.6 Example

A sample of 50 households in one community shows that 10 of them are watching a TV special on the national economy. In a second community, 15 of a random sample of 50 households are watching the TV special. We test the hypothesis that the overall proportion of viewers in the two communities does not differ, using the 1 percent level of significance, as follows:

### 3.1.7 2 sided test

A two-sided test is used when we are concerned about a possible deviation in either direction from the hypothesized value of the mean. The formula used to establish the critical values of the sample mean is similar to the formula for determining confidence limits for estimating the population mean, except that the hypothesized value of the population mean m0 is the reference point rather than the sample mean.

### 3.1.8 The $t$ distribution

TESTING A HYPOTHESIS CONCERNING THE MEAN BY USE OF THE t DISTRIBUTION:

The $t$ distribution is the appropriate basis for determining the standardized test statistic when the sampling distribution of the mean is normally distributed but $s$ is not known. The sampling distribution can be assumed to be normal either because the population is normal or because the sample is large enough to invoke the central limit theorem. The $t$ distribution is required when the sample is small ($n < 30$). For larger samples, normal approximation can be used. For the critical value approach, the procedure is identical to that described in Section 10.3 for the normal distribution, except for the use of $t$ instead of z as the test statistic.

## 3.2 Confidence Interval

CONFIDENCE INTERVALS FOR THE MEAN

suppose that you wish to estimate the mean sales amount per retail outlet for a particular consumer product during the past year. The number of retail outlets is large. Determine the 95 percent confidence interval given that the sales amounts are assumed to be normally distributed, $\bar{X}$ =3,425, s = 200 , and $n = 25$.

Ans. $3; 346 : 60to3;503{:}40$

8.24. Referring to Problem 8.23, determine the 95 percent confidence interval given that the population is assumed to be normally distributed, $\bar{X}$ =3,425, s = 200 , and $n = 25$.

Ans. $3; 342 : 44to3;507{:}56$

## 3.3 Two sample test

Suppose one has two independent samples, x1, ..., xm and y1, ..., yn, and wishes to test the hypothesis that the mean of the x population is equal to the mean of the y population:

$H0 : \mu_x = \mu_y$.

Alternatively this can be formulated as $H0 : \mu_x - \mu_y = 0$.

Let $\bar{X}$ and $\bar{Y}$ denote the sample means of the xs and ys and let $S_x$ and $S_y$ denote the respective standard deviations. The standard test of this hypothesis $H_0$ is based on the t statistic

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{1/m + 1/n}} \tag{3.3}$$

where $S_p$ is the pooled standard deviation.

$$S_p = \sqrt{\frac{(m-1)S_x^2 + (n-1)S_y^2}{m+n-2}} \tag{3.4}$$

Under the hypothesis $H_0$, the test statistic T has a t distribution with $m + n - 2$ degrees of freedom when

- both the xs and ys are independent random samples from normal distributions
- the standard deviations of the x and y populations, $\sigma_x$ and $\sigma_y$, are equal

.

Suppose the level of significance of the test is set at $\alpha$. Then one will reject H when $|T| < tn + m.2, \alpha/2$, where $tdf, \alpha$ is the $(1 - \alpha)$ quantile of a t random variable with df degrees of freedom.

If the underlying assumptions of

### 3.3.1 Paired T test

The mean and standard deviation of the sample d values are obtained by use of the basic formulas in Chapters 3 and 4, except that d is substituted for X.

The mean difference for a set of differences between paired observations is $\bar{d} = \frac{\sum d_i}{n}$.

The deviations formula and the computational formula for the standard deviation of the differences between paired observations are, respectively,

$$S_d = \sqrt{\frac{\sum(d_i - \bar{d})^2}{n-1}} \tag{3.5}$$

$$S_d = \sqrt{\frac{\sum(d^2) - n(\bar{d}^2)}{n-1}} \tag{3.6}$$

$$\tag{3.7}$$

The standard error of the mean difference between paired observations is obtained for the standard error of the mean.

**Hypotheses**

$$H_0 : \mu_d = 0$$

$$H_1 : \mu_d \neq 0$$

# Chapter 4

# Linear Regression

## 4.1 Simple Linear Regression

We start with a scatter diagram between two variables as before. This time, we want to know what line will best fit the data.

The theory we learn here assumes we are going to use a straight line, and not a curve of any kind, though in some disciplines (physics or finance, for example) a curve would be more appropriate.

We have to find the line of best fit. Before we can do this, we must assume that one variable is dependent on the other. By convention we call the dependent variable y and the independent variable x. We have to work out the slope of the line, and the point at which it cuts the y axis.

Again, by convention, we call these values and respectively for the population.

The basic model is therefore given as follows.

The model of a random variable Y, the dependent variable, which is related to random variable X, the independent (or predictor or explanatory) variable by the equation:

$$Y = \alpha + \beta X + \epsilon \tag{4.1}$$

where $\alpha$ and $\beta$ are constants and $\epsilon$ $N(0, \sigma^2)$ , a random error term. The coefficients $\alpha$ and $\beta$ are theoretical values and can only be estimated from sample data.

The estimates are generally written as $a$ and $b$.

Given a sample of bivariate data, $(x_1, y_1), (x_2, y_2), ....., (x_n, y_n)$, $a$ and $b$ can be estimated. To fit a line to some data as in this case, an objective must be chosen to define which straight line best describes the data. The most common objective is to minimise the sum of the squared distance between the observed value of $y_i$ and the corresponding predicted value $\hat{y}_i$.

The estimated least squares regression line is written as: $y = a + b \times x$ We can derive the formulae for $b$ and $a$.

The line is called the sample regression line of y on x.

The following example demonstrates the calculation of a and b and the use of the resultant equation to estimate y for a given x.

### 4.1.1 Ordinary least squares

Ordinary least squares (OLS) is a technique for estimating the unknown parameters in a linear regression model. This method minimizes the sum of squared distances between the observed responses in a set of data, and the fitted responses from the regression model.

### 4.1.2 Regression example

A study was made by a retailer to determine the relation between weekly advertising expenditure and sales (in thousands of pounds). Find the equation of a regression line to predict weekly sales from advertising. Estimate weekly sales when advertising costs are 35,000.

Adv. Costs(in 000) 40 20 25 20 30 50 40 20 50 40 25 50

Sales (in 000) 385 400 395 365 475 440 490 420 560 525 480 510

### 4.1.3 example

Concentration (ng/ml) 0 5 10 15 20 25 30 Absorbance 0.003 0.127 0.251 0.390 0.498 0.625 0.763

```
# DO A FULL LINEAR REGRESSION ANALYSIS ON THE DATA

>concentration=c(0,5,10,15,20,25,30)
>absorbance=c(0.03,0.127,0.251,0.390,0.498,0.625,0.763)
>regr=lm(absorbance~concentration)
# READ AS; ABSORBANCE DEPENDENT ON CONCENTRATION
>summary(regr)
```

This output from this code is as follows:

```
Call:
lm(formula = absorbance ~ concentration)

Residuals:
        1         2         3         4         5         6         7
 0.015357 -0.010571 -0.009500  0.006571 -0.008357 -0.004286  0.010786

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.0146429  0.0079787   1.835    0.126
concentration 0.0245857  0.0004426  55.551 3.58e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.01171 on 5 degrees of freedom
Multiple R-squared: 0.9984,     Adjusted R-squared: 0.9981
```

```
F-statistic:  3086 on 1 and 5 DF,  p-value: 3.576e-08
```

- Estimation of Slope = 0.0251643
- Standard error of the estimation of the slope is 0.0002656
- Estimation of Intercept = 0.0021071
- Standard error of the estimation of the intercept is 0.0047874
- Degrees-of-Freedom1 = 7-2 = 5 The critical values for testing is are -2.57 and 2.57 since the area under the Students t distribution curve with 5 degrees-of-freedom outside this range is 5
- p-value for intercept is 67.8% implying that it is not significantly different from zero
- The p-value for the intercept is the area under the Students t-distribution curve with 5 degrees-of-freedom outside the range of [-0.44,0.44].
- p-value for slope is less than 5% implying that it is significantly different from zero
- The p-value for the slope is the area under the Students t-distribution curve with 5 degrees of- freedom outside the range of [-94.76,94.76].

### 4.1.4 Regression example

A survey was conducted in 9 areas of the USA to investigate the relationship between divorce rate (y) and residential mobility (x). Divorce rates in the annual number per 1000 in the population and the residential mobility is measured by the percentage of the population that moved house in the last five years.

| Area | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| x | 40 | 38 | 46 | 49 | 47 | 43 | 51 | 57 | 55 |
| y | 3.9 | 3.4 | 5.2 | 4.8 | 5.6 | 5.8 | 6.6 | 7.6 | 5.8 |

- Check that the following statements are correct.
  - sum of x data = 426
  - sum of squares of x data = 20494
  - sum of y data = 48.7
  - sum of squares of y data = 276.81
  - sum of products of x and y data = 2361
- Derive the estimates for the slope and intercept of the regression line.
- Estimate the divorce rate for areas that has a residential mobility of 39 and 60 respectively.
- Which of these estimates is likely to be more accurate? Why?

## 4.2 Regression

The argument to lm is a model formula in which the tilde symbol ( ) should be read as "described by.

This was seen several times earlier, both in connection with boxplots and stripcharts and with the t and Wilcoxon tests.

### 4.2.1 Multiple Linear Regression

The `lm()` function handles much more complicated models than simple linear regression. There can be many other things besides a dependent and a descriptive variable in a model formula.

A multiple linear regression analysis (which we discuss in Chapter 11) of, for example, y on x1, x2, and x3 is specified as $y\ x1 + x2 + x3$.

This is an F test for the hypothesis that the regression coefficient is zero. This test is not really interesting in a simple linear regression analysis since it just duplicates information already givenit becomes more interesting when there is more than one explanatory variable.

### 4.2.2 Regression

```
> lm(short.velocity~blood.glucose)
```

## 4.3 Inference for Regression

To determine the confidence interval for the slope we use the following equation:

$$b \pm t_{1-\alpha/2,n-2} S.E.(b) \tag{4.2}$$

- b = Estimation of Slope (0.0251643)
- S.E.(b) = Standard Error of Slope(0.0002656)
- n = Sample Size (7)
- $\alpha$ = Alpha Value (5%)
- $t_{1-\alpha/2,n-2}$ = Quantile Value from Students t-distribution (2.570582)

$$(0.0251643) \pm (0.0002656)(2.570582) = [0.0245, 0.0258] \tag{4.3}$$

### 4.3.1 Regression example

In a medical experiment concerning 12 patients with a certain type of ear condition, the following measurements were made for blood flow (y) and auricular pressure (x):

```
x<-c(8.5, 9.8, 10.8, 11.5, 11.2, 9.6, 10.1, 13.5, 14.2, 11.8, 8.7, 6.8)
y<-c(3 ,12, 10, 14, 8 ,7 ,9 ,13, 17, 10, 5 ,5)
```

(Sx =126.5 Sxx =1,381.85 Sy =113 Syy =1251 Sxy =1272.2)

- Calculate the equation of the least-squares fitted regression line of blood flow on auricular pressure.
- Confirm the following values: Sx =126.5, Sxx =1381.85, Sy =113, Syy =1251, Sxy =1272.2.
- Calculate the correlation coefficient.

```
> cor(x,y)
[1] 0.8521414
```