Contents

Chapter 1

Introductory Statistics

1.1 Systematic and random errors

Experimental scientists make a fundamental distinction between ***random***, and ***systematic*** errors. To distinguish between random and systematic errors let us consider a real experiment.

Four students (A-D) each perform an analysis in which exactly 10.00 ml of exactly 0.1 M sodium hydroxide is titrated with exactly 0.1 NI hydrochloric acid. Each student performs five replicate titrations, with the results shown in Table 1.1.

Student	Results	(ml)				Comment
A	10.08	10.11	10.09	10.10	10.12	Precise, biased
B	9.88	10.14	10.02	9.80	10.21	Imprecise unbiased
C	10.19	9.79	9.69	10.05	9.78	Imprecise, biased
D	10.04	9.98	10.02	9.97	10.04	Precise, unbiased

Graphical illustration

The results of experiment represented by dot-plots. (The true value is 10.00). Recall the average for each student $10.0950, 9.9600, 9.9300, 10.0025$ respectively.

Systematic error and bias

Systematic error is a deviation of all measurements in one direction from the true value. It is well represented by the difference between the average value of the determined values and the true value of the measured quantity. This difference is called the bias of measurements.

Random error and precision

Random error is a deviation of a measurement from the average of measured values. It is well represented by the standard deviation of measurements. This value is often called precision of measurements.

Combined error vs. accuracy

Accuracy is in inverse relation to the total deviation of a single measurement from the true value.

1.2 Section 1A : Introductory Statistics

Population and samples Survey sampling Descriptive statistics Tabular displays

1.3 Section 1B : Descriptive Statistics

Mean Variance and standard deviation Median and IQR Quantile statistics Graphical Procedures: Histograms and Barcharts

1.4 Syllabus

- Descriptive Statistics
- Probability Distributions
- Statistical Inference
- Information Theory

On successful completion of this module, students should be able to:

1. Apply probability theory to problem solving.
2. Employ the concepts of random variables and probability distributions to problem solving.
3. Apply information theory to solve problems in data compression and transmission.
4. Analyse rates and proportions.
5. Perform hypothesis tests for a variety of statistical problems.

1.4.1 About this Course

The Book used in this course is:
There will be a mid-term assessment worth 20 of the overall mark.
The end of year exam is worth 80%

1.5 Section 1: Descriptive Statistics

Descriptive statistics

Chapter 2

Sampling Schemes

2.1 Population

As is so often the case in statistics, some words have technical meanings that overlap with their common use but are not the same. Population is one such word. It is often difficult to decide which population should be sampled. For instance, if we wished to sample 500 listeners to an FM radio station specializing in music should the population be of listeners to that radio station in general, or of listeners to that stations classical music programme, or perhaps just the regular listeners, or any one of many other possible populations that you can construct for yourself? In practice the population is often chosen by finding one that is easy to sample from, and that may not be the population of first choice.

In medical trials (which are an important statistical application) the population may be those patients who arrive for treatment at the hospital carrying out the trial, and this may be very different from one hospital to another. If you look at any collection of official statistics (which are most important for state planning) you will be struck by the great attention that is given to defining the population that is surveyed or sampled, and to the definition of terms. For instance, it has proved difficult to get consensus on the meaning of unemployed in recent years but a statistician must be prepared to investigate the population of unemployed. Think about this carefully. It should help you with your Sociology and Marketing and Market Research modules as well as this one:

2.2 Bias

In addition to the common-sense meaning of bias, there is also a more technical meaning for the word in statistics. This will be found both in Chapter 10 of this guide and in the work on estimators in Statistics 2. It seems natural enough to wish to avoid bias, but it is not helpful to be swayed by the value judgements inherited from the use of a word outside the limits of academic discussion.

2.3 Sampling

Explain the difference between sampling error and sampling bias. Explain briefly which is taking place in the following situations, if the population under study consists of all the pupils in a certain school:

i. Your sample is a list of all pupils at the school except those who arrived at the school in the last school year.

ii. You take a random sample of names from the school register of all pupils at the school.

Define quota sampling. In what circumstances would you use it? In what circumstances would you use stratified random sampling? Give two ways in which stratified random sampling differs from quota sampling.

2.4 Sampling

Although it seems sensible to sample a population to avoid the cost of a total enumeration (or census) of that population, it is possible to make a strong argument against the practice. One might well consider that sampling is fundamentally unfair because a sample will not accurately represent the whole population, and it allows the units selected for the sample to have more importance than those not selected. This might be thought undemocratic. Many countries continue to take a full census of their population, even though sampling might be cheaper. It is less obvious, but true, that sampling might well be more accurate, because more time can be spent verifying the information collected for a sample.

2.5 Random Sampling

Random sampling Though it may be clear that random sampling should avoid a sample biased by the prejudices of the sampler, you should think carefully whether that is always a good idea. In other times it was thought that a sampler could produce a better sample by using personal judgement than by randomizing, because all the relevant facts could be taken into consideration. Randomization is popular because this belief seems to be contradicted by experience. Remember that a random sample may (though not very often) come out to look very biased just by chance should one then reject it and try again? If the population has both men and women, would you accept a random sample that just by chance included only men? These difficulties are usually dealt with by some form of restricted randomization. One might, for instance in the example above, use a stratified random sample and select half the sample from the men and half from the women. To carry out a random sample one needs a sample frame. This is the list of all the population units. For instance, if you want to investigate UK secondary schools, the sample frame would be a list of all the secondary schools in the UK. The key advantages of random sampling are that it avoids systematic bias and it allows an assessment of the size of sampling error.

Random sampling is carried out, for many populations, by choosing at random without replacement a subset of the very large number of units in the population. If the sample is a small proportion of the population, then there is no appreciable difference between the inferences possible for sampling with replacement and sampling without replacement.

2.6 Stratified Sampling

In a stratified sample the sampling frame is divided into non-overlapping groups or strata, e.g. geographical areas, age-groups, genders. A sample is taken from each stratum, and when this sample is a simple random sample it is referred to as stratified random sampling.

Chapter 3

Probability

3.1 Introductory Statistics: Basic Concepts

Uniform Distribution: Exercise 24

Use the uniform distribution to simulate 100 throws of two dice. The outcome is the combined values of both dice. Use the appropriate R command to discretize values.

- What is the mean and standard deviation of the outcomes?
- Make a stem-and-leaf plot of the outcomes.
- Make a histogram of the outcomes. (hint: use breaks =seq(1.5,12.5))

Inference Procedures: Exercise 25: Confidence intervals

Seven meas urements 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. Calculate (i) the 95% and (ii) the 99% confidence limits for the true pH utilizing R.

```
> x<-c(5.12, 5.20, 5.15, 5.17, 5.16, 5.19, 5.15)
> n=length(x)
>
> alpha5=0.05
> alpha1=0.01
>
> LB95=mean(x)+qt(alpha5/2,n-1)*sd(x)/sqrt(n)
> UB95=mean(x)+qt(1-alpha5/2,n-1)*sd(x)/sqrt(n)
>
> LB95
[1] 5.137975
> UB95
[1] 5.187739
```

Thus the 95% confidence interval for this problem is approximately [5:14; 5:19], while using the same code only taking α =0.01 will give us [5:13; 5:20].

Exercise 29

Add up all the numbers for 1 to 50

- Using the sum() command.
- using a for loop.

Exam overview

Formulae

Familiarize yourself thoroughly with the Statistical Formulae (See Module Website)

Question 1 : Probability Distribution

Introduction

Consider playing a game in which you are winning when a ***fair die*** is showing 'six' and losing otherwise.

Part 1

If you play three such games in a row, find the probability mass function (pmf) of the number X of times you have won.

- Firstly: what type of probability distribution is this?
- Is this the distribution ***discrete*** or ***continuous***?
- The outcomes are whole numbers - so the answer is discrete.
- So which type of discrete distribution? (We have two to choose from. See first page of formulae)
- **Binomial:** characterizing the number of ***successes*** in a series of n ***independent trials***, with the ***probability of a success*** in each trial being p.

- **Poisson:** characterizing the ***number of occurrences*** in a ***unit space*** (i.e. a unit length, unit area or unit volume, or a unit period in time), where λ is the the number of occurrences per unit space.

- This distribution is **Binomial**
- A success here is throwing a six
- An independent trial is a roll of a die
- There are 3 independent trials ($n = 3$)
- The probability of a success is 1 in 6 ($p = 1/6$)
- (Lets look at the formulae to see what we have to work with)

Probability Mass Function

- This is a tabulation of all the possible outcomes (k), with the probability of that outcome ($P(X = k)$)
- The possible outcomes are
  - No successes (i.e. no sixes) (k=0)
  - One Success (k=1)
  - Two Success (k=2)
  - Three Success (k=3)

Factorials

- $5! = 5 \times 4 \times 3 \times 2 \times 1 (= 120)$
- $5! = 5 \times 4!$

The Choose Operator

$$\binom{n}{k} = \frac{n!}{k! \times (n-k)!}$$

$$\binom{3}{1} = \frac{3!}{1! \times (3-1)!} \frac{3 \times 2!}{1! \times 2!} = \frac{3}{1} = 3$$

- $\binom{3}{0} = 1$
- (Remember 0! is always equal to 1)

- $\binom{3}{1} = 3$
- $\binom{3}{2} = 3$
- $\binom{3}{3} = 1$

$$P(X = 0) = \binom{3}{0}(1/6)^0(5/6)^3 = 125/216 = 0.5787$$

$$P(X = 1) = \binom{3}{1}(1/6)^1(5/6)^2 = 75/216 = 0.3472$$

$$P(X = 2) = \binom{3}{2}(1/6)^2(5/6)^1 = 15/216 = 0.0695$$

$$P(X = 3) = \binom{3}{3}(1/6)^3(5/6)^0 = 1/216 = 0.0046$$

Probability Mass Function

k	0	1	2	3
P(X=k)	0.5787	0.3472	0.0695	0.0046

Part 2 Suppose that you are playing thirty such games and X is the number of times you have won. What are possible values for X?

Answer: The random variable X can take any value between 0 and 30. $k \in \{0, 1, 2, \ldots 30\}$

Part 3 What a well-known distribution describes pmf of X? Identify its parameters. Write the formula for the pmf.

- We actually covered this already. The distribution is the binomial distribution.
- The parameters are similar, except that the number of independent trials is now 30.
- So $n = 30$ and $p = 1/6$
- $X \sim \text{Bin}(30, 1/6)$
- A General Formula for the PMF is

$$P(X = k) = \binom{30}{k} (1/6)^k (5/6)^{30-k}$$

- where $k \in \{0, 1, 2, \ldots 30\}$

Part 5 What is the expected value of the number X of wins when you are playing thirty games?

Part 6 What is the Standard Deviation of X in this case?

- From formulae: The expected value is $E(X) = n \times p = 30 \times (1/6) = 5$
- From formulae: The variance is $Var(X) = n \times p \times (1-p) = 30 \times (1/6) \times (5/6) = 4.1666$
- The standard deviation σ is the square root of the variance : $\sigma = \sqrt{4.1667}$

Question 2

- Ohm's Law: $U = IR$
  - U: Potential (also known as Voltage)
  - I: Current
  - R: Resistance
- Suppose that the resistance R is a random variable having uniform distribution over interval $[1.5, 2.25]$, i.e. $R \sim U(1.5; 2.25)$
- U has normal distribution with the mean 12[V] and standard deviation $1.2[V]$, i.e. $U \sim N(12, 1.2^2)$.
- Assume that R and U are independent and solve the following problems.

Part 1 Find the expected values for R, I. $E[R]$ and $E[I]$

$$\int_a^b \frac{1}{x} dx = log(b) - log(a)$$

- $R \sim U(a, b)$
- From Formulae: $E(R) = \frac{a+b}{2} == \frac{1.5+2.25}{2} = 1.875$
- $E(R) = 1.875[\Omega]$
- We are told that $E(U) = 12[V]$
- $$E[I] = E\left[\frac{U}{R}\right] = E[U] \times E[1/R]$$
- We need to compute $E[1/R]$
- Recall
$$E[R] = \int [rf(r)]\, dr$$
- Recall
$$E[R^2] = \int \left[r^2 f(r)\right] dr$$
- Therefore
$$E[1/R] = \int [1/rf(r)]\, dr$$
- In this case, the density function $f(r)$ is a constant.

$$f(r) = \frac{1}{b-a} = \frac{1}{2.25 - 1.5} = \frac{1}{0.75}$$

- $$E[1/R] = \int [1/rf(fr)]\, dr = f(r) \times \int [1/r]\, dr$$
- Density function is a constant
$$f(r) = \frac{1}{b-a} = \frac{1}{2.25 - 1.5} = \frac{1}{0.75}$$
- Using the hint
$$\int_a^b \frac{1}{r} dr = log(b) - log(a) = log(2.25) - log(1.5)$$
- $$E(1/R) = \frac{log(2.25) - log(1.5)}{0.75} = 0.54063$$
- So the answer is
$$E(I) = 12 \times 0.54063 = 6.4876[A]$$

Part 3 Find the variances for R and I

hint:
$$\int_a^b \frac{1}{x^2} dx = 1/a - 1/b$$

- We use a very similar approach to the last part
- $R \sim U(a, b)$
- From Formulae: $Var(R) = \frac{(b-a)^2}{12} = \frac{(2.25-1.5)^2}{12}$
- $Var(R) = 0.046875$
- We are also told that $Var(U) = 1.2^2 = 1.44$
- Using formulae
$$Var(I) = E(I^2) - E(I)^2$$
- We need to compute $E(I^2)$
$$E(I^2) = E\left[\frac{U^2}{R^2}\right] = E[U^2] \times E\left[\frac{1}{R^2}\right]$$
- We can say: $Var(U) = E[U^2] - E[U]^2$
- We know $Var(U)$ and can compute $E[U]^2$ easily
- $1.44 = E[U^2] - 12^2$
- Necessarily $E[U^2] = 145.44$

- To find $E[1/R^2]$ we use a very similar approach to the last question.

$$E[1/R^2] = \int \left[\frac{1}{r^2} f(r)\right] dr = f(r) \int \frac{1}{r^2} dr$$

$$E[1/R^2] = \frac{(1/1.5) - (1/2.25)}{0.75} = 0.29627$$

- Therefore $E[I^2] = 145.44 \times 0.29627 = 43.09$
- $Var[I] = E[I] - E[I]^2 = 43.09 - (6.4876)^2 = 1.001[A]$

Part 3 Find the standard deviations for R and I?

- Simply compute the square roots of the respective variances.
- $\sigma_R = 0.21651[\Omega]$
- $\sigma_I = 1.0005[A]$

Part 4 If you observe I=7[A], would yo reconsider this a rather unusual value for current in the view of the obtained values for the mean and the standard deviation of I? Justify your answer

- The expected value of I is 6.4876 [A]
- The standard deviation is 1.005
- We expect most values to be within $6.48 \pm 1[A]$
- The observed value of the current seems to be: close to the mean, and within the range of typical values for current I

Part 5 Find the covariance between R and I

- $Cov(R, I) = E(R \times I) - (\mu_R \times \mu_I)$
- Alternatively :$Cov(R, I) = E(R \times I) - (E(R) \times E(I))$
- Using Ohm's law: $E(I \times R) = E(U) = 12$
- $E(I) \times E(R) = 1.875 \times 6.4876 = 12.164$
- $Cov(R, I) = -0.164$

Part 6 Evaluate the correlation coefficients between R and I.

- See Formula
$$\rho = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$
- From before $\sigma_I = 1.005[A]$ and $\sigma_R = 0.21651[\Omega]$
- The correlation coefficient is defined as the ratio between the covariance and the product standard deviation
$$\rho_{R,I} = \frac{-0.164}{(0.21651 \times 1.005)} = -0.7537$$

Part 7 Based on the obtained values of characteristics, if you observe R = 2.1[Ω], then out of two possible values 7.5[A] and 5.2[A] for the current I which you would consider more plausible and why?

- From the correlation coefficient, we see a fairly strong negative relationship
- If we see high values of one variable, we expect lower values of the other.
- In this instance, we have a reasonably high value of resistance 2.1, where the max is 2.25.
- Therefore we expected a low value for current.
- We expect that the value of 5.2 [A] is more likely.
- Exercise 14: question on trigonometric functions omitted.
- Exercise 17: inverse cosine function is acos()

```
> cos(pi)
[1] -1
> acos(-1)
[1] 3.141593
>
```

- Exercise 18: use the standard normal distribution to generate random numbers, unless told otherwise.

```
> rnorm(4)
[1] -1.68732684 -0.62743621  0.01831663  0.70524346
```

- Exercise 18: We have not covered the material for parts 5, 10 and 11 yet.

Exercise 20

Construct a vector of the 0.9, 0.95, 0.975 and 0.99 quantiles of the $t-$ distribution (degrees of freedom = 16). For hep with the $t-$ distribution use the command '?qt'.

```
> perc =c(0.9, 0.95, 0.975, 0.99)
> perc
[1] 0.900 0.950 0.975 0.990
> qt(perc, 16)
[1] 1.336757 1.745884 2.119905 2.583487
```

What is the 0.975 quantile for the $t-$ distribution when the degrees of freedom is 40? (N.B. This exercise is relevant to forthcoming topics, such as confidence intervals.)

```
> qt(0.975, 40)
[1] 1.336757 1.745884 2.119905 2.583487
```

Exercise 21

	Make	Model	Cylinder	Weight	Mileage	Type
1	Honda	Civic	V4	2170.00	33.00	Sporty
2	Chevrolet	Beretta	V4	2655.00	26.00	Compact
3	Ford	Escort	V4	2345.00	33.00	Small
4	Eagle	Summit	V4	2560.00	33.00	Small
5	Volkswagen	Jetta	V4	2330.00	26.00	Small
6	Buick	Le Sabre	V6	3325.00	23.00	Large
7	Mitsbusihi	Galant	V4	2745.00	25.00	Compact
8	Dodge	Grand Caravan	V6	3735.00	18.00	Van
9	Chrysler	New Yorker	V6	3450.00	22.00	Medium
10	Acura	Legend	V6	3265.00	20.00	Medium

Advantages

Stratification will always achieve greater precision provided that the strata have been chosen so that members of the same stratum are as similar as possible in respect of the characteristic of interest. The bigger the differences between the strata, the greater the gain in precision. For example, if you were interested in Internet usage you might stratify by age, whereas if you were interested in smoking you might stratify by gender or social class.

It is often administratively convenient to stratify a sample. Interviewers can be

specifically trained to deal with a particular age-group or ethnic group, or employees in a particular industry. The results from each stratum may be of intrinsic interest and can be analysed separately.

It ensures better coverage of the population than simple random sampling.

3.2 Introduction to Statistics

Sampling Research

3.2.1 Cluster sampling

Cluster sampling is a sampling technique that generates statistics about certain populations. It has a specific format required to obtain an appropriate sample, and though this sampling can help accurately gauge some information, it is not thought as accurate as simple random samples, where all groups of the same size have the same exact chance of being selected. Despite lacking the assurance that comes from using simple random samples or random samples, cluster sampling is used frequently in business and other applications.

The basic procedure for creating cluster sampling is to divide the full population into some sort of meaningful groups. For instance, McDonald's might want a sense of what the most popular item ordered on their menu is. They might create a cluster/group for each McDonald's store. They would then pick some of these clusters and obtain a sample from all people in that cluster. They could keep track of each customer's order and decide which menu item is most popular or survey customers eating, but the company would only survey or track people in the chosen clusters; they'd also try to get all people at selected clusters.

Cluster sampling is very popular on big voting nights. A natural division exists between voter precincts. By choosing some of the precincts and surveying or using exit polls at the chosen ones, there's often a good sense what issues or what elected officials appear to be winning. The results of cluster samples are extrapolated to the entire population, and they're often fairly representative of it.

When people study statistics, they often find it challenging to remember the features of cluster sampling as opposed to the features of stratified sampling. The two have some similarities and key differences that are worth understanding.

3.2.2 Stratified Sampling

In a stratified sample, a population is also divided into groups, though number of groups tends to be smaller. Population could be divided by gender, age, income, and region in which they live, and comparing the result of each group may be part of the reason the stratified sample is performed. The huge and appreciable difference between stratified and cluster sampling is that when the groups are created, some members from each group or strata are selected. With a cluster, when clusters are created, the whole population of some of the clusters are used.

The degree to which cluster sampling works tends to depend on what is being evaluated and how diverse of a population clusters represent. Say a statistician decided to break down voting precincts in a predominantly Republican state and create clusters of some of them to look for predictions about a national election. These results would likely be skewed and not representative of the complete population in the US. On the other hand, cluster sampling with exit polling in a Republican or Democrat state could say a lot about the voting trends in the individual state.

3.3 Research

What is research? What is the aim of research? What is the Frascati document? Distinguish between empirical research and theoretical research. What are the phases of the research process? Distinguish between basic research and applied research. What makes business research different from other types of research? (2 characteristics) Distinguish between market research and marketing research.

- Hypothesis testing - type I and type II error, one and two-tailed tests, oc curves.
- Statistical process control - various charts, mean/range, individuals/moving range, cusum charts.
- Capability studies - capability indices.
- Correlation and Regression - method of least squares, multiple regression, linear and non-linear models, regression analysis, analysis of residuals.
- Importance of plotting data.
- Design of experiments and analysis of variance - one and two way ANOVA, interaction, factorial designs, responses and factors, Plackett-Burman design, response surface methodology.

3.3.1 Recommended Text

1 Statistical Analysis Methods for Chemists (Author : William P Gardiner)

2 simpleR - Using R for Introductory Statistics (Author : John Verzani)

3 An Introduction to R (Authors: The R Project)

3.4 Probability Formulae

- Conditional probability:

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}.$$

- Bayes' Theorem:

$$P(B|A) = \frac{P(A|B) \times P(B)}{P(A)}.$$

Disadvantages

- Difficulty in identifying appropriate strata.
- More complex to organise and analyse results.

3.4.1 Quota sampling

Quota sampling avoids the need for a sample frame. The interviewer seeks out units to ensure that the sample contains given quotas of units meeting specified criteria (known as quota controls). Quota sampling is cheap, but it may be biased systematically by the choices made by the interviewer. For instance, interviewers avoid anyone who looks threatening, or too busy, or too strange. Quota sampling does not allow an accurate assessment of sampling error.

3.4.2 Joint Distributions

Sample Question

The random variables X and Y have the following joint distribution

	Y =0	Y =1	Y=2
X=0	0.2	0	0.2
X=1	0	0.2	0.1
X =2	0.1	0.1	0.1

- Calculate the marginal distributions of X and Y.
- Calculate the coefficient of the correlation between X and Y.
- Are the random variables X and Y independent?

3.4.3 Example

The distribution of (X,Y) is specified in the following table.

X	Y	Probability
1	8	1/6
2	4	1/4
3	6	1/3
5	7	1/4

Find the correlation coefficient of X and Y .

Hypothesis Testing : Worked Example

1. The following are measurements (in mm) of a critical dimension on a sample of engine crankshafts:

224.120	224.017	223.976	223.961
224.089	223.982	223.980	223.989
223.960	223.902	223.987	224.001

(a) Calculate the mean and standard deviation for these data.
(b) The process mean is supposed to be $\mu = 224$mm. Is this the case? Give reasons for your answer.
(c) Construct a 99% confidence interval for these data and interpret.
(d) Check that the normality assumption is valid using 2 suitable plots.

Using `R` for Linear Models

- Simple linear regression
  - The `lm()` command
- Scatter plots
- Pearson's correlation
- Importance of "summary()" and "names()"
- Accessing importance values in data objects

Answers

3.4.4 Quantiles

The quantile function is the inverse of the cumulative distribution function. The p-quantile is the value with the property that there is probability p of getting a value less than or equal to it. The median is by definition the 50% quantile.

Theoretical quantiles are commonly used for the calculation of confidence intervals and for power calculations in connection with designing and dimensioning experiments.

3.4.5 Covariance

The covariance of any two random variables X and Y , denoted by $\mathrm{Cov}(X, Y)$, is defined by $\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$.

Properties of Covariance

For any random variables X, Y,Z and constant c,

1. Cov(X,X) = Var(X),
2. Cov(X,Y) = Cov(Y,X),
3. Cov(kX,Y) = kCov(X, Y),
4. Cov(X, Y +Z) = Cov(X,Y) + Cov(X,Z).

3.4.6 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.

3.4.7 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.

3.5 Overview of experimental design

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

3.6 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

$$SST = \sum_j \sum_j (x-\bar{x})^2 \tag{3.1}$$

$$SSM = \sum_j \sum_j (x-\bar{x})^2 \tag{3.2}$$

$$SSE = \sum_j \sum_j (x-\bar{x})^2 \tag{3.3}$$

$$\tag{3.4}$$

```
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
```

3.7 Normal Distribution: Worked Example

The intelligence quotient (IQ) of 36 randomly chosen students was measured. Their average IQ was 109.9 with a variance of 324. The average IQ of the population as a whole is 100.

1. Calculate the p-value for the test of the hypothesis that on average students are as intelligent as the population as a whole against the alternative that on average students are more intelligent.
2. Can we conclude at a significance level of 1% that students are on average more intelligent than the population as a whole?
3. Calculate a 95% confidence interval for the mean IQ of all students.

$$Z_{Test} = \frac{X - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{109.9 - 100}{\frac{18}{\sqrt{36}}} = \frac{9.9}{3} = 3.3 \tag{3.5}$$

$$p.value = P(Z \geq Z_{Test}) = P(Z \geq 3.3) = 0.00048 \tag{3.6}$$

$$\bar{X} \pm t_{1-\alpha/2,\nu} S.E.(\bar{X}) \tag{3.7}$$

$\nu = 1.96$

$$t_{1-\alpha/2,\nu} = 1.96 \tag{3.8}$$

$$109.9 \pm (1.96 \times 3) = [104.02, 115.79] \tag{3.9}$$

3.8 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.

3.9 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

The Addition Rule for Probability

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$$\binom{8}{2} = \frac{8!}{2!(8-2)!} = \frac{8 \times 7 \times 6!}{2! \times 6!} = \frac{8 \times 7}{2 \times 1} = \frac{56}{2} = 28$$

3.9.1 Independent Events

$$P(AandB) = P(A)P(B)$$

$$P(A)P(B) = 0.98 \times 0.95 = 0.931$$

$P(D) = P(CandD) + P(MandD) + P(LandD)P(D) = P(D|C)P(C) + P(D|M)P(M) + P(D|L$

3.9.2 Example

For a particular Java assembler interface, the operand stack size has the following probabilities:

Stack Size	0	1	2	3	4
Probability	.15	.05	.10	.20	.50

- Calculate the expected stack size.
- Calculate the variance of the stack size.

3.10 Random Variables

Suppose X has the following probability mass function: $p(0) = 0.2$, $p(1) = 0.5$, $p(2) = 0.3$. Calculate $E[X]$ and $E[X^2]$

Chapter 4

Statistical Inference

Confidence Intervals

One sample

$$\begin{aligned} S.E.(\bar{X}) &= \frac{\sigma}{\sqrt{n}}. \\ S.E.(\hat{P}) &= \sqrt{\frac{\hat{p} \times (100 - \hat{p})}{n}}. \end{aligned}$$

Two samples

$$\begin{aligned} S.E.(\bar{X}_1 - \bar{X}_2) &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}. \\ S.E.(\hat{P}_1 - \hat{P}_2) &= \sqrt{\frac{\hat{p}_1 \times (100 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 \times (100 - \hat{p}_2)}{n_2}}. \end{aligned}$$

Hypothesis tests

One sample

$$\begin{aligned} S.E.(\bar{X}) &= \frac{\sigma}{\sqrt{n}}. \\ S.E.(\pi) &= \sqrt{\frac{\pi \times (100 - \pi)}{n}} \end{aligned}$$

Two large independent samples

$$\begin{aligned} S.E.(\bar{X}_1 - \bar{X}_2) &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}. \\ S.E.(\hat{P}_1 - \hat{P}_2) &= \sqrt{(\bar{p} \times (100 - \bar{p})) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}. \end{aligned}$$

Two small independent samples

$$\begin{aligned} S.E.(\bar{X}_1 - \bar{X}_2) &= \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}. \\ s_p^2 &= \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}. \end{aligned}$$

Paired sample

$$S.E.(\bar{d}) = \frac{s_d}{\sqrt{n}}.$$

Standard deviation of case-wise differences

$$s_d = \sqrt{\frac{\sum d_i^2 - n\bar{d}^2}{n - 1}}.$$

4.0.1 Sample Size Estimation

- Recall the formula for margin of error, which we shall denote E.

$$E = Q_{(1-\alpha)} \times \text{Std. Error}$$

- $Q_{(1-\alpha)}$ denotes the quantile that corresponds to a $1-\alpha$ confidence level. (There is quite a bit of variation in notation in this respect.)
- Also recall that the only way to influence the margin of error is to set the sample size accordingly.
- Sample size estimation describes the selection of a sample size n such that the margin of error does not exceed a pre-determined level E.

4.0.2 Sample Size Estimation for the Mean

- The margin of error does not exceed a certain threshold E.

$$E \geq Q_{(1-\alpha)} \times S.E.(\bar{x}),$$

- which can be re-expressed as

$$E \geq Q_{(1-\alpha)} \times \frac{\sigma}{\sqrt{n}}.$$

- Divide both sides by $\sigma \times Q_{(1-\alpha)}$.

$$\frac{E}{\sigma Q_{(1-\alpha)}} \geq \frac{1}{\sqrt{n}}$$

- Square both sides

$$\frac{E^2}{\sigma^2 Q^2_{(1-\alpha)}} \geq \frac{1}{n}$$

4.0.3 Sample Size Estimation for the Mean

- Square both sides

$$\frac{E^2}{\sigma^2 Q^2_{(1-\alpha)}} \geq \frac{1}{n}$$

- Invert both sides, changing the direction of the relational operator.

$$\frac{\sigma^2 Q^2_{(1-\alpha)}}{E^2} \leq n$$

- The sample size we require is the smallest value for n which satisfies this identity.
- The sample standard deviation s may be used as an estimate for σ.
- (This formula would be provided on the exam paper).

4.0.4 SSE for the Mean: Example

- An IT training company has developed a new certification program. The company wishes to estimate the average score of those who complete the program by self-study.
- The standard deviation of the self study group is assumed to be the same as the overall population of candidates, ie. 21.2 points.
- How many people must be tested if the sample mean is to be in error by no more than 3 points, with 95% confidence.

4.0.5 SSE for the Mean: Example

- The sample size we require is the smallest value for n which satisfies this identity.

$$n \geq \frac{\sigma^2 Q^2_{(1-\alpha)}}{E^2}$$

- Remark: $1-\alpha = 0.95$, therefore $Q_{(1-\alpha)} = 1.96$. Also $E = 3$ and $\sigma = 21.2$.

$$n \geq \frac{(21.2)^2 \times (1.96)^2}{3^2}$$

- Solving, the required sample size is the smallest value of n that satisfies

$$n \geq 191.8410$$

- Therefore, the company needs to test 192 self-study candidates.

4.0.6 Sample Size Estimation for proportions

We can also compute appropriate sample sizes for studies based on proportions.

- From before;

$$E \geq Q \times S.E.(\hat{p}).$$

(For the sake of brevity, we will just use the notation Q for quantile.)

- Divide both sides by Q.

$$E \geq Q \times \sqrt{\frac{\pi(1-\pi)}{n}}.$$

4.0.7 Sample Size Estimation for proportions

- Remark: E must be expressed in the same form as π, either as a proportion or as a percentage.
- Remark : The standard error is maximized at $\pi = 0.50$,which is to say $\pi(1-\pi)$ can never exceed 0.25 (or 25%). Therefore the standard error is maximized at $\pi = 0.50$. To make the procedure as conservative as possible, we will use 0.25 as our value for $\hat{p}_1 \times (1-\hat{p}_1)$.
- If we use percentages, $\pi \times (100-\pi)$ can not exceed 2500 (i.e $50 \times (100-50) = 2500$).

$$E \geq Q \times \sqrt{\frac{2500}{n}}.$$

4.0.8 Sample Size Estimation for proportions

- Dive both sides by Q, the square both sides:

$$\left(\frac{E}{Q}\right)^2 \geq \frac{2500}{n}.$$

- Invert both sides, changing the direction of the relational operator, and multiply both sides by 2500.

$$\left(\frac{Q}{E}\right)^2 \times 2500 \leq n.$$

- The sample size we require is the smallest value for n which satisfies this identity. (This formula would be provided on the exam paper, but without the maximized standard error).

4.0.9 SSE for proportions: Example

- An IT journal wants to conduct a survey to estimate the true proportion of university students that own laptops.
- The journal has decided to uses a confidence level of 95%, with a margin of error of 2%.
- How many university students must be surveyed?

4.0.10 Sample Size estimation for proportions

- Confidence level = 0.95. Therefore the quantile is $Q_{(1-\alpha)} = 1.96$
- Using the formula:

$$n \geq \left(\frac{1.96}{2}\right)^2 \times 2500$$

- The required sample size is the smallest value for n which satisfies this identity:

$$n \geq 2401$$

- The required sample size is therefore 2401.

4.0.11 Independent Samples (New Section)

- Two samples are referred to as independent if the observations in one sample are not in any way related to the observations in the other.
- This is also used in cases where one randomly assign subjects to two groups, i.e. in give first group treatment A and the second group treatment B and compare the two groups.
- Often we are interested in the difference between the mean value of some parameter for both groups.

4.0.12 Independent Samples

The approach for computing a confidence interval for the difference of the means of two independent samples, described shortly, is valid whenever the following conditions are met:

- Both samples are simple random samples.
- The samples are independent.
- Each population is at least 10 times larger than its respective sample. (Otherwise a different approach is required).
- The sampling distribution of the difference between means is approximately normally distributed

4.0.13 Difference Of Two Means

In order to construct a confidence interval, we are going to make three assumptions:

- The two populations have the same variance. This assumption is called the assumption of homogeneity of variance.
- For the time being, we will use this assumption. Later on in the course, we will discuss the validity of this assumption for two given samples.
- The populations are normally distributed.
- Each value is sampled independently from each other value.

4.0.14 Computing the Confidence Interval

- As always the first step is to compute the point estimate. For the difference of means for groups X and Y, the point estimate is simply the difference between the two means i.e. $\bar{x} - \bar{y}$.
- As we have seen previously, sample size has a bearing in computing both the quantile and the standard error. For two groups, we will use the aggregate sample size $(n_x + n_y)$ to compute the quantile. (For the time being we will assume, the aggregate sample size is large $(n_x + n_y) > 30$.)
- Lastly we must compute the standard error $S.E.(\bar{x} - \bar{y})$. The formula for computing standard error for the difference of two means, depends on whether or not the aggregate sample size is large or not. For the case that the sample size is large, we use the following formula (next slide).

4.0.15 Computing the Confidence Interval

Standard Error for difference of two means (large sample)

$$S.E.(\bar{x} - \bar{y}) = \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

- s_x^2 and s_x^2 is the variance of samples X and Y respectively.
- n_x and n_y is the sample size of both samples.

- For small samples, the degrees of freedom is $df = n_x + n_y - 2$. If the sample size $n \leq 32$, we can find appropriate $t-$quantile, rather than assuming it is a $z-$quantile.

4.0.16 CI for Difference in Two Means

A research company is comparing computers from two different companies, X-Cel and Yellow, on the basis of energy consumption per hour. Given the following data, compute a 95% confidence interval for the difference in energy consumption.

Type	sample size	mean	variance
X-cel	17	5.353	2.743
Yellow	17	3.882	2.985

Remark: It is reasonable to believe that the variances of both groups is the same. Be mindful of this.

- Point estimate : $\bar{x} - \bar{y} = 1.469$
- Standard Error: 0.5805

$$S.E.(\bar{x} - \bar{y}) = \sqrt{\frac{2.743}{17} + \frac{2.985}{17}} = \surd$$

- Quantile : 1.96 (Large sample, with confidence level of 95%.)

$$1.469 \pm (1.96 \times 0.5805) = (0.3321, 2.607)$$

This analysis provides evidence that the mean consumption level per hour for X-cel is higher than the mean consumption level per hour for Yellow, and that the difference between means in the population is likely to be between 0.332 and 2.607 units.

4.0.17 Computing the Confidence Interval

Standard Error for difference of two means (small aggregate sample)

$$S.E.(\bar{x} - \bar{y}) = \sqrt{s_p^2 \left(\frac{1}{n_x} + \frac{1}{n_y} \right)}$$

Pooled Variance s_p^2 is computed as:

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{(n_x - 1) + (n_y - 1)}$$

4.0.18 CI for Difference in Two Means

From the previous example (comparing X-cel and Yellow) lets compute a 95% confidence interval when the sample sizes are $n_x = 10$ and $n_y = 12$ respectively. (Lets assume the other values remain as they are.)

Type	sample size	mean	variance
X-cel	10	5.353	2.743
Yellow	12	3.882	2.985

The point estimate $\bar{x} - \bar{y}$ remains as 1.469. Also we require that both samples have equal variance. As both X and Y have variances at a similar level, we will assume equal variance.

4.0.19 Computing the Confidence Interval

- Pooled variance s_p^2 is computed as:

$$s_p^2 = \frac{(10 - 1)2.743 + (12 - 1)2.985}{(10 - 1) + (12 - 1)} = \frac{57.52}{20} = 2.87$$

- Standard error for difference of two means is therefore

$$S.E.(\bar{x} - \bar{y}) = \sqrt{2.87 \left(\frac{1}{10} + \frac{1}{12} \right)} = 0.726$$

- The aggregate sample size is small i.e. 22. The degrees of freedom is $n_x + n_y - 2 = 20$. From Murdoch Barnes tables 7, the quantile for a 95% confidence interval is 2.086.

- The confidence interval is therefore

$$1.469 \pm (2.086 \times 0.726) = 1.4699 \pm 1.514 = (-0.044, 2.984)$$

4.0.20 Difference in proportions

We can also construct a confidence interval for the difference between two sample proportions, $\pi_1 - \pi_2$. The point estimate is the difference in sample proportions for tho both groups , $\hat{p}_1 - \hat{p}_2$.
Estimation Requirements The approach described in this lesson is valid whenever

the following conditions are met:

- Both samples are simple random samples.
- The samples are independent.
- Each sample includes at least 10 successes and 10 failures.
- The samples comprises less than 10% of their respective populations.

4.0.21 Standard Error for Difference of Proportions

- $\hat{p}_1$ and $\hat{p}_2$ are the sample proportions of groups 1 and 2 respectively.
- n_1 and n_2 are the sample sizes of groups 1 and 2 respectively.

N.B. This formula will be provided in the exam paper. Also, there is no accounting for small samples.

4.0.22 Difference of Proportions : Example

- A study finds that a percentage of 40% of IT users out of a random sample of 400 in a large community preferred one web browser to all others.
- In another large community, 30% of IT users out of a random sample of 300 prefer the same web browser.
- Compute a 95 percent confidence interval for the difference in the proportion of IT users who prefer this particular web browser.

4.0.23 Confidence Interval

Compute the standard Error

$$S.E.(\hat{p}_1 - \hat{p}_2) = \sqrt{\left[\frac{\hat{p}_1 \times (1 - \hat{p}_1)}{n_1}\right] + \left[\frac{\hat{p}_2 \times (1 - \hat{p}_2)}{n_2}\right]}$$

$$S.E.(\hat{p}_1 - \hat{p}_2) = \sqrt{\left[\frac{40 \times 60}{400}\right] + \left[\frac{30 \times 70}{300}\right]} = \sqrt{\left[\frac{2400}{400}\right] + \left[\frac{2100}{300}\right]}$$

$$S.E.(\hat{p}_1 - \hat{p}_2) = \sqrt{6 + 7} = 3.6\%$$

4.0.24 Confidence Interval

- The point estimate is the difference in two proportions i.e. $\hat{p}_1 - \hat{p}_2 = 40\% - 30\% = 10\%$
- We have a large sample, and the confidence level is 95%. Therefore the quantile is 1.96.
- We can now compute the confidence interval for the difference of proportions:

$$10\% \pm (1.96 \times 3.6\%) = \ 10\% \pm 7.05\% = \ (2.95\%, 17.05\%)$$

- SE $= \sqrt{[p_1 \times (1 - p_1)/n_1] + [p_2 \times (1 - p_2)/n_2]}$
- SE $= \sqrt{[0.40 \times 0.60/400] + [0.30 \times 0.70/300]}$
- SE $= \sqrt{[(0.24/400) + (0.21/300)]} = \sqrt{(0.0006 + 0.0007)} = \sqrt{0.0013} = 0.036$

4.0.25 Mean Difference Between Matched Data Pairs

The approach described in this lesson is valid whenever the following conditions are met:

- The data set is a simple random sample of observations from the population of interest.
- Each element of the population includes measurements on two paired variables (e.g., x and y) such that the paired difference between x and y is: d = x - y.
- The sampling distribution of the mean difference between data pairs (d) is approximately normally distributed.

The observed data are from the same subject or from a matched subject and are drawn from a population with a normal distribution does not assume that the variance of both populations are equal

4.0.26 Computing the Case Wise Differences

Student	Before	After	Difference (d_i)	$(d_i - \bar{d})^2$
1	90	95	5	16
2	85	89	4	9
3	76	73	-3	4
4	90	92	2	1
5	91	92	1	0
6	53	53	0	1
7	67	68	1	4
8	88	90	2	9
9	75	78	3	16
10	85	89	4	25

4.0.27 Computing the Case Wise Differences

Compute the mean difference

$$\bar{d} = \frac{\sum d_i}{n} = \frac{3+6}{8}$$

Compute the variance of the differences.

$$s_d^2 = \frac{\sum (d_i - \bar{d})^2}{n-1} = \frac{3+6}{9}$$

4.0.28 How a paired t test works

- The paired t test compares two paired groups.
- It calculates the difference between each set of pairs, and analyzes that list of differences based on the assumption that the differences in the entire population follow a Gaussian distribution.
- First we calculate the difference between each set of pairs, keeping track of sign.
- If the value in column B is larger, then the difference is positive. If the value in column A is larger, then the difference is negative.
- The t ratio for a paired t test is the mean of these differences divided by the standard error of the differences. If the t ratio is large (or is a large negative number), the P value will be small. The number of degrees of freedom equals the number of pairs minus 1. Prism calculates the P value from the t ratio and the number of degrees of freedom.

$$(\bar{X} - \bar{Y}) \pm \left[\text{Quantile } \times S.E(\bar{X} - \bar{Y})\right]$$

- If the combined sample size of X and Y is greater than 30, even if the individual sample sizes are less than 30, then we consider it to be a large sample.
- The quantile is calculated according to the procedure we met in the previous class.

Assume that the mean (μ) and the variance (σ) of the distribution of people taking the drug are 50 and 25 respectively and that the mean (μ) and the variance (σ) of the distribution of people not taking the drug are 40 and 24 respectively.

4.0.29 Difference in Two means

For this calculation, we will assume that the variances in each of the two populations are equal. This assumption is called the assumption of homogeneity of variance.

The first step is to compute the estimate of the standard error of the difference between means ().

$$S.E.(\bar{X} - \bar{Y}) = \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

- s_x^2 and s_x^2 is the variance of both samples.
- n_x and n_y is the sample size of both samples.

The degrees of freedom is $n_x + n_y - 2$.

4.1 What is Statistical Inference?

- Hyptothesis testing
- Confidence Intervals
- Sample size estimation.

4.2 Confidence Interval examples

4.2.1 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 ± 3.83
i.e. $(90.37, 98.03)$

4.2.2 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

4.2.3 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
```

4.2.4 Example 3

Ten replicate analyses of the concentration of mercury in a sample of commercial gas condensate gave the following results (in ng/ml) :

23.3	22.5	21.9	21.5	19.9	21.3	21.7	23.8	22.6	24.7

Compute 99% confidence limits for the mean.

4.3 Hypothesis Tests for Two Means

If the population standard deviations σ_1 and σ_2 are known, the test statistic is of the form:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{4.1}$$

The critical value and p-value are looked up in the normal tables.

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:

$$H_0 : \mu_1 = \mu_2 \qquad (\mu_1 - \mu_2 = 0) \tag{4.2}$$
$$H_a : \mu_1 \neq \mu_2 \qquad (\mu_1 - \mu_2 \neq 0) \tag{4.3}$$

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
```

4.3.1 Basic Probability Questions

Q1a. Two fair dice are thrown. What is the probability of at least one odd number?
Q1b. What is the probability of at least one odd number if four fair dice are thrown?

4.3.2 Hypothesis testing: introduction

The objective of hypothesis testing is to access the validity of a claim against a counterclaim using sample data

- The claim to be proved is the alternative hypothesis(H_1).
- The competing claim is called the null hypothesis(H_0).
- One begins by assuming that H_0 is true.

If the data fails to contradict H_0 beyond a reasonable doubt, then H_0 is not rejected. However, failing to reject H_0 does not mean that we accept it as true. It simply means that H_0 cannot be ruled out as a possible explanation for the observed data. A proof by insufficient data is not a proof at all.

> The process by which we use data to answer questions about parameters is very similar to how juries evaluate evidence about a defendant. from Geoffrey Vining, Statistical Methods for Engineers, Duxbury, 1st edition, 1998.

4.3.3 Hypothesis testing

The standard deviation of the life for a particular brand of ultraviolet tube is known to be $S = 500hr$, and the operating life of the tubes is normally distributed. The manufacturer claims that average tube life is at least 9,000hr. Test this claim at the 5 percent level of significance against the alternative hypothesis that the mean life is less than 9,000 hr, and given that for a sample of $n = 18$ tubes the mean operating life was $\bar{X} = 8,800hr$.

4.3.4 two populations

Two samples drawn from two populations are independent samples if the selection of the sample from population 1 does not affect the selection of the sample from population 2. The following notation will be used for the sample and population measurements:

- p_1 and p_2 = means of populations 1 and 2,
- σ_1 and σ_2 = standard deviations of populations 1 and 2,
- n_l and n_2 = sizes of the samples drawn from populations 1 and 2 ($n_1 > 30$, $n_2 > 30$),
- x_1 and x_2, = means of the samples selected from populations 1 and 2,

- s_1 and s_2 = standard deviations of the samples selected from populations 1 and 2.

4.3.5 Standard Error

$$S.E(\bar{X}_1 - \bar{X}_2) = \sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})} \tag{4.4}$$

4.3.6 Two Sample t-test : Example

The mean height of adult males is 69 inches and the standard deviation is 2.5 inches. The mean height of adult females is 65 inches and the standard deviation is 2.5 inches. Let population 1 be the population of male heights, and population 2 the population of female heights. Suppose samples of 50 each are selected from both populations.

4.3.7 One Sample confidence Interval: 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 ± 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.

4.3.8 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 $5.2 - 2.26(7.41)/\sqrt{10}, 5.2 + 2.26(7.41)/\sqrt{10} = (.0.1, 10.5)$.

4.3.9 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:

4.3.10 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.

4.3.11 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.

4.4 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

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{1/m + 1/n}} \tag{4.5}$$

where S_p is the pooled standard deviation.

$$S_p = \sqrt{\frac{(m-1)S_x^2 + (n-1)S_y^2}{m+n-2}} \tag{4.6}$$

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, σ_x and σ_y, are equal

.

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

If the underlying assumptions of

4.4.1 Example using R

Finding confidence intervals for the mean for the nitrate ion concentrations in Example 2.7.1.

```
#Typing data in
x=c(102,97,99,98,101,106)
mean(x)
sd(x)
n=length(x)
#setting the confidence level
CL=0.95
#computing confidence interval
pm=sd(x)*c(qt(0.025,n-1),qt(0.975,n-1))/sqrt(n)
CI=mean(x)+pm
```

4.5 Independent one-sample t-test

In testing the null hypothesis that the population mean is equal to a specified value μ_0, one uses the statistic

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \tag{4.7}$$

where s is the sample standard deviation and n is the sample size. The degrees of freedom used in this test is $n-1$.

4.5.1 Confidence intervals

Most studies will only sample part of a population and then the result is used to interpret the null hypothesis in the context of the whole population. Any estimates obtained from the sample only approximate the population value. Confidence intervals allow statisticians to express how closely the sample estimate matches the true value in the whole population. Often they are expressed as 95% confidence intervals. Formally, a 95% confidence interval of a procedure is any range such that the interval covers the true population value 95% of the time given repeated sampling under the same conditions.

If these intervals span a value (such as zero) where the null hypothesis would be confirmed then this can indicate that any observed value has been seen by chance. For example a drug that gives a mean increase in heart rate of 2 beats per minute but has 95% confidence intervals of -5 to 9 for its increase may well have no effect whatsoever.

The 95% confidence interval is often misinterpreted as the probability that the true value lies between the upper and lower limits given the observed sample. However this quantity is more a credible interval available only from Bayesian statistics.

Confidence intervals - example

A researcher was investigating computer usage among students at a particular university. Three hundred undergraduates and one hundred postgraduates were chosen at random and asked if they owned a laptop. It was found that 150 of the undergraduates and 80 of the postgraduates owned a laptop.

Find a 95% confidence interval for the difference in the proportion of undergraduates and postgraduates who own a laptop. On the basis of this interval, do you believe that postgraduates and undergraduates are equally likely to own a laptop?

4.6 F-test of equality of variances

The test statistic is

$$F = \frac{S_X^2}{S_Y^2} \tag{4.8}$$

has an F-distribution with $n-1$ and $m-1$ degrees of freedom if the null hypothesis of equality of variances is true.

Chapter 5

Probability

Section 3 : Probability

How to Compute Probability: Equally Likely Outcomes Sometimes, a statistical experiment can have n possible outcomes, each of which is equally likely. Suppose a subset of r outcomes are classified as "successful" outcomes.

The probability that the experiment results in a successful outcome (S) is:

P(S) = (Number of successful outcomes) / (Total number of equally likely outcomes) = r / n

Consider the following experiment. An urn has 10 marbles. Two marbles are red, three are green, and five are blue. If an experimenter randomly selects 1 marble from the urn, what is the probability that it will be green?

In this experiment, there are 10 equally likely outcomes, three of which are green marbles. Therefore, the probability of choosing a green marble is 3/10 or 0.30.

- Conditional probability
- Independent events
- Repeated independent events

Chapter 6

Introduction to R

6.1 The R Project for Statistical Computing

R is a language and environment for statistical computing and graphics. Rprovides a wide variety of statistical and graphical techniques, and is highly extensible. Among its tools one can find implemented

- linear and nonlinear modelling,
- classical statistical tests,
- time-series analysis,
- classification,
- clustering,
- ...and many more.

One of R's strengths is the ease with which well-designed publication quality plots can be produced. including mathematical symbols and formulae where needed.

- R is a computing software for statistical analysis
- The package is available for all popular operating systems: Windows, Mac or Linux.
- It is free!
- Everyone (knowledgeable enough) can contribute to the software by writing a package.
- Packages are available for download through a convenient facility
- It is fairly well documented and the documentation is available either from the program help menu or from the web-site.

- It is the top choice of statistical software among academic statisticians but also very popular in industry specially among biostatisticians and medical researchers (mostly due to the huge package called Bioconductor that is built on the top of R).
- It is a powerful tool not only for doing statistics but also all kind of scientific programming.

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent. integrated collection of intermediate tools for data analysis, graphical facilities for data analysis and display either on-screen or on hard-copy, and
- a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

6.2 Downloading and Installing R

- R can be downloaded from the CRAN website: http://cran.r-project.org/
- You may choose versions for windows, mac and linux.
- As per the instructions on the respective pages, you require the "base" distribution.
- Now you can download the installer for latest version of R , version 2.17.
- Select the default settings. Once you finish, the R icon should appear on your desktop.
- Clicking on this icon will start up the program.

6.3 Statistical Tables using R

The following is a fragment of the tables of the values of $F(x)$ for the standard normal ('Z') cumulative distribution function from page 254 of the main textbook.

```
# Segment 1A-1
# Preceding line with the symbol # makes it a comment in R
# The following line produce a single value of the standard normal cumulative
# function. It is the value corresponding to the first value in the table

pnorm(-3.4)

#[1] 0.0003369293
#Then the first row of the table

z=seq(-3.4,-3.31,by=0.01)
pnorm(z)

# [1] 0.0003369293 0.0003494631 0.0003624291 0.0003758409 0.0003897124
# [6] 0.0004040578 0.0004188919 0.0004342299 0.0004500872 0.0004664799
# And all values from the table

z=seq(-3.4,3.4,by=0.01)
pnorm(z)

#  [1] 0.0003369293 0.0003494631 0.0003624291 0.0003758409 0.0003897124
#  [6] 0.0004040578 0.0004188919 0.0004342299 0.0004500872 0.0004664799
# [11] 0.0004834241 0.0005009369 0.0005190354 0.0005377374 0.0005570611
# [16] 0.0005770250 0.0005976485 0.0006189511 0.0006409530 0.0006636749
# [21] 0.0006871379 0.0007113640 0.0007363753 0.0007621947 0.0007888457
# [26] 0.0008163523 0.0008447392 0.0008740315 0.0009042552 0.0009354367
# [31] 0.0009676032 0.0010007825 0.0010350030 0.0010702939 0.0011066850
```

There is more than meets the eye in the table. It is not only the table values that can be explored for the standard normal distribution using R. Recall that the normal distribution is defined by the density function:

$$f(z) = \frac{1}{\sqrt{(2\pi)}} e^{-Z^2/2}.$$

The density represents distribution of probability for a random variable associated with it. The area under the density represents the probability so the that the total area under it is equal to one. The area accumulated up to certain value z_o represents probability that a corresponding random variable takes value smaller than z and this probability defines the cumulative distribution function $F(z)$ which is tabularized.

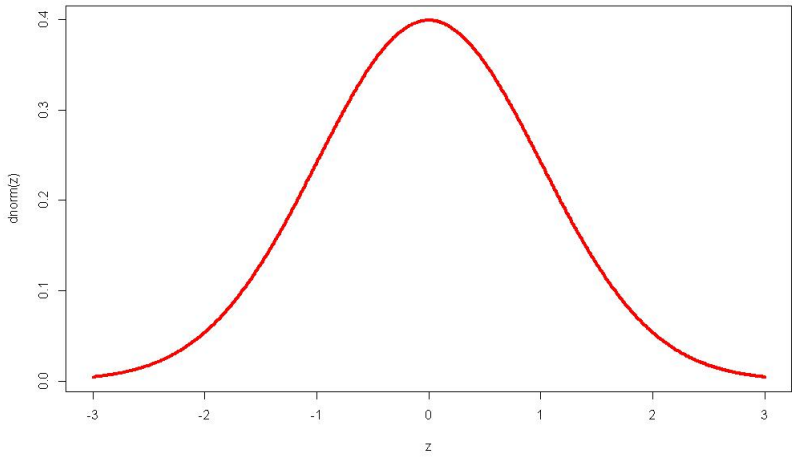
All this can be seen in R. The following code explores various aspects of the standard normal distribution:

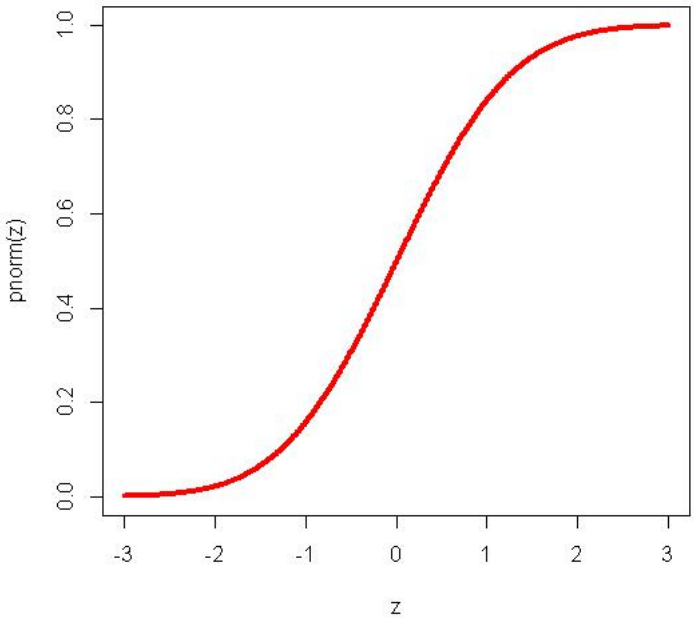
```
#Plotting the density function of the standard normal variable
z=seq(-3,3,by=0.01)
plot(z,dnorm(z),type='l',col="red",lwd=4)

#Plotting the cumulative distribution function (that one from the table)
plot(z,pnorm(z),type='l',col="red",lwd=4)
```

The R code results in the following plots.

- The probability density function.
- The cumulative density function.





6.4 Data Analysis with R

Data from Table 1.1 of the textbook

Table 1.1 Random and systematic errors

Student	Results	(ml)				Comment
A	10.08	10.11	10.09	10.10	10.12	Precise, unbiased
B	9.88	10.14	10.02	9.80	10.21	Imprecise unbiased
C	10.19	9.79	9.69	10.05	9.78	Imprecise, biased
D	10.04	9.98	10.02	9.97	10.04	Precise, unbiased

This is also given in the text file Table1 − 1.txt, the contents of which is given below:

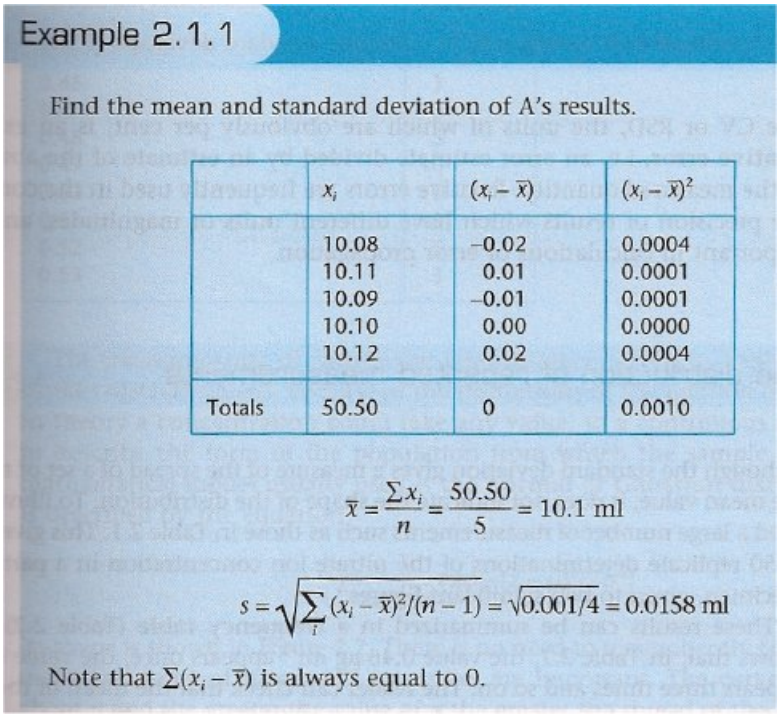
```
A 10.08 10.11 10.09 10.10
B 9.88 10.14 10.02 9.80
C 10.19 9.79 9.69 10.05
D 10.04 9.98 10.02 9.97
```

Reading data from a file to R:

```
#Reading the data from
Titra=read.table("Table1-1.txt", row.names = 1)
Titra
# V2 V3 V4 V5
#A 10.08 10.11 10.09 10.10
#B 9.88 10.14 10.02 9.80
#C 10.19 9.79 9.69 10.05
#D 10.04 9.98 10.02 9.97
#Listing the first row
Titra[1,]
#and the fourth column
Titra[,4]
```

Means and standard deviations

Find the mean and standard deviation of A's results.



Example 2.1.1

Find the mean and standard deviation of A's results.

	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
	10.08	−0.02	0.0004
	10.11	0.01	0.0001
	10.09	−0.01	0.0001
	10.10	0.00	0.0000
	10.12	0.02	0.0004
Totals	50.50	0	0.0010

$$\bar{x} = \frac{\sum x_i}{n} = \frac{50.50}{5} = 10.1 \text{ ml}$$

$$s = \sqrt{\sum_i (x_i - \bar{x})^2/(n-1)} = \sqrt{0.001/4} = 0.0158 \text{ ml}$$

Note that $\sum(x_i - \bar{x})$ is always equal to 0.

Means and standard deviations much faster and better

```
#Comuting means
rowMeans(Titra)
# A B C D
#10.0950 9.9600 9.9300 10.0025
#and standard deviation
apply(Titra,1,sd)
# A B C D
#0.01290994 0.15055453 0.23036203 0.03304038
```

Nitrate ion concentration from Table 2.1

Table 2.1 Results of 50 determinations of nitrate ion concentration, in μg ml^{-1} (Also in the file Table $2-1$.txt.)

0.51	0.51	0.51	0.50	0.51	0.49	0.52	0.53	0.50	0.47
0.51	0.52	0.53	0.48	0.49	0.50	0.52	0.49	0.49	0.50
0.49	0.48	0.46	0.49	0.49	0.48	0.49	0.49	0.51	0.47
0.51	0.51	0.51	0.48	0.50	0.47	0.50	0.51	0.49	0.48
0.51	0.50	0.50	0.53	0.52	0.51	0.50	0.50	0.51	0.51

```
0.51 0.51 0.51 0.50 0.51 0.49 0.52 0.53 0.50 0.47
0.51 0.52 0.53 0.48 0.49 0.50 0.52 0.49 0.49 0.50
0.49 0.48 0.46 0.49 0.49 0.48 0.49 0.49 0.51 0.47
0.51 0.51 0.51 0.48 0.50 0.47 0.50 0.51 0.49 0.48
0.51 0.50 0.50 0.53 0.52 0.51 0.50 0.50 0.51 0.51
```

Compute the mean concentration, and the standard deviation:

```
#Getting data in a vector
x=scan('Table2_1.txt')
mean(x)
#[1] 0.4998
sd(x)
#[1] 0.01647385
```

6.4.1 Bootstrap Methods

If we would repeat our experiment of collecting 50 samples of nitrate concentrations many times we would see the range of error. But it would be a waste of resources and not a viable method.
Instead we re-sample 'new' data from our data and use so obtained new samples for assessment of the error. The following R code does the job.

```
#Getting data in a vector
m=mean(x)
bootstrap=vector('numeric',500)
for(i in 1:500)
 {
 bootstrap[i]=mean(sample(x,replace=T))-mean(x)
 }
#The distribution of estimation error
hist(bootstrap)
```

The conclusion of this procedure is that the nitrate concentration is 4999 ± 0.005. We are specifically interested in how R was easily able to implement a solution for this.

6.5 Introduction - systematic vs. random errors

6.6 Statistics of Repeated Measures

6.6.1 Titration experiment

Recall the titration experiment from the last class. 4 Students performing the same experiment five times, hence each yield 5 results.(Table 1.1 random and systematic errors).

Student	Results	(ml)				Comment
A	10.08	10.11	10.09	10.10	10.12	Precise, biased
B	9.88	10.14	10.02	9.80	10.21	Imprecise unbiased
C	10.19	9.79	9.69	10.05	9.78	Imprecise, biased
D	10.04	9.98	10.02	9.97	10.04	Precise, unbiased

Two criteria were used to compare these results, the average value (technically know as a measure of location and the degree of spread (or dispersion). The average value used was the arithmetic mean (usually abbreviated to *the mean*), which is the sum of all the measurements divided by the number of measurements.

The mean,$\bar{X}$, of n measurements is given by

$$\bar{X} = \frac{\sum x}{n}$$

In Chapter 1 the spread was measured by the difference between the highest and lowest values (i.e. the range). A more useful measure, which utilizes all the values, is the sample standard deviation, s, which is defined as follows:

The standard deviation, s, of n measurements is given by

$$s = \sqrt{\frac{\sum (x - \bar{X})^2}{n-1}} (2.2)$$

Example 1 A fair die is thrown. The number shown on the die is the random variable X. Tabulate the possible outcomes. Solution X takes the six possible outcomes 1, 2, 3, 4, 5, 6 which each have probability 1/6 (i.e. one sixth).

r 1 2 3 4 5 6 P(X = r) 1/6 1/6 1/6 1/6 1/6 1/6 Example 2 Two unbiased spinners, one numbered 1, 3, 5, 7 and the other numbered 1, 2, 3 are spun. The random variable X is the sum of the two results. Find the probability distribution for X.

Solution Listing all the possible outcomes is best done in a table.

Chapter 7

Stochastic Processes

7.1 Probability

Probability generating function

$$G_N(s) = \sum_{n=0}^{\infty} p_n s^n \tag{7.1}$$

$$\frac{dG(s)}{ds} = G'(s) = \sum_{n=1}^{\infty} n p_n s^{n-1} \tag{7.2}$$

$$\frac{d^2G(s)}{ds^2} = G''(s) = \sum_{n=2}^{\infty} n(n-1) p_n s^{n-2} \tag{7.3}$$

$$G_N(1) = \sum_{n=0}^{\infty} p_n \tag{7.4}$$

$$G'_N(1) = \sum_{n=1}^{\infty} n p_n = E(N) = \mu \tag{7.5}$$

The random variable N has a binomial distribution with paramtrics terms m and p. Its probability function is given by

$$p(n) = p_n = P(N = n) = \binom{m}{n} p^n q^{m-n} \tag{7.6}$$

Find its probability generating function.

$$G(s) = \sum_{n=0}^{m} s^n \binom{m}{n} p^n q^{m-n} \tag{7.7}$$

$$G(s) = \sum_{n=0}^{m} \binom{m}{n} (ps)^n q^{m-n} \tag{7.8}$$

$$G(s) = (q + ps)^m \tag{7.9}$$

$$(x+y)^k = [\binom{k}{0}x^k y^0]+[\binom{k}{1}x^{k-1}y^1]+\cdots+[\binom{k}{n}x^{k-n}y^n]+\ldots[\binom{k}{k-1}x^1 y^{k-1}]+[\binom{k}{k}x^0 y^k]$$

$$(x+y)^k = \sum_{n=0}^{k}\binom{k}{n}x^{k-n}y^n$$

$$(x+y)^3 = \left[\binom{3}{0}x^{3-0}y^0\right] + \left[\binom{3}{1}x^{3-1}y^1\right] + \left[\binom{3}{2}x^{3-2}y^2\right] + \left[\binom{3}{3}x^{3-3}y^3\right]$$

$$(x+y)^3 = x^3 + 3x^2y + 3xy^2 + y^3$$

$$\sum_{n=0}^{k}\binom{k}{n}p^{k-n}q^n = (p+q)^k$$

This result is not interesting because $p+q=1$

$$(p+q)^k = 1^k = 1$$

However when using terms ps and q we can say:

$$\sum_{n=0}^{k}\binom{k}{n}ps^{k-n}q^n = (ps+q)^k$$

$$\binom{m}{n} = \frac{m!}{n!\times(m-n)!}$$

7.2 Poisson

$$p(x) = \frac{e^{\alpha}\alpha^x}{x!} \qquad x = 0, 1, 2\ldots$$

variance and mean are both equal to α.

7.3 PGF

$$G(s) = \frac{1-\alpha(1-s)}{1+\alpha(1+s)}$$

Chapter 8

Markov Chains

8.1 Markov Chains

$$T = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Find the eigenvalues of the following matrix.

$$T = \begin{bmatrix} 0.25 & 0.5 & 0.25 \\ 0.50 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.50 \end{bmatrix}$$

$$Det(T - I_3\lambda) = \begin{vmatrix} 0.25 - \lambda & 0.5 & 0.25 \\ 0.50 & 0.25 - \lambda & 0.25 \\ 0.25 & 0.25 & 0.50 - \lambda \end{vmatrix}$$

8.2 Markov Chains

$$T = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{bmatrix}$$

8.2.1 Classification of States

1. Absorbing state
2. Periodic state
3. Persistent state
4. Transient state
5. Ergodic state

8.2.2 Absorbing states

Absorbing states are characterized in Markov chains by a value of 1 in the diagonal element of the matrix ($p_{ii} = 1$). Once entered, there is no escaping the absorbing states.

8.2.3 Classification of Chain

1. Irreducible sets
2. Closed sets
3. Ergodic chains

8.2.4 Irreducible Chain

An irreducible chain is a chain in which every state is accessible from any other state in a finite number of steps.

8.2.5 Closed Sets

A closed set is a subset of states that can not be escaped once entered. An absorbing state is a closed set composed of one state.

Closed Sets

$$T = \begin{bmatrix} 0.2 & 0.2 & 0.3 & 0.3 \\ 0.1 & 0.2 & 0.3 & 0.4 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 0.5 \end{bmatrix}$$

Once either states 3 or 4 are entered, it is not possible to revert to states 1 or 2. States 3 and 4 will be the only states visited.

8.2.6 Ergodic Chains

For Ergodic Chains, the invariant distribution is the vector of the mean recurrence time reciprocals.

Chapter 9

Birth and Death Processes

9.1 Birth processes

Generating function equations

$$\frac{ds}{dz} = \lambda s(s-1)$$

This is a first order separable equation.

$$\int \frac{s(s-1)}{ds} = \int -\lambda dz = -\lambda z$$

9.2 Pure death processes

In a pure death process, the population numbers decline by dying off, with no replacement births.

$$\int \frac{ds}{s(1-s)} = \int \mu dz = \mu z$$

$$\int \frac{ds}{s(1-s)} = \int \frac{1}{s} ds 1 \int \frac{1}{s-1} ds = ln\left[\frac{s}{1-s}\right]$$

$$\ln\left[\frac{s}{1-s}\right] = -\lambda z$$

$$\left[\frac{s}{1-s}\right] = exp(-\lambda z)$$

$$s = \left[\frac{1}{1+exp(-\lambda z)}\right]$$

Calculations

1. $\frac{1}{s(1-s)} = \frac{1}{s} + \frac{1}{1-s} = \frac{1}{s} - \frac{1}{s-1}$
2. $\ln(a-b) = \ln(\frac{a}{b})$

9.3 Combined birth and death processes

birth rate λ and death rate μ.

Chapter 10

Reliability Theory

10.1 Notation

1. F(t) : Failure function. Probability that a component fails before time 't'.
2. R(t) : Reliability function (Survivor function) Probability that a component has survived to time 't'.
3. r(t) : Failure rate function (hazard function)

10.2 Reliability Theory

1. Exponential distribution and reliability
2. Mean time to failure
3. Reliability of series and parrallel systems
4. Renewal processes

Chapter 11

Chemometrics

11.1 Chemometrics

1. Calibration
2. Paired T test
3. Confidence intervals

11.2 Calibration

11.3 Blank Signals

11.4 Chemometrics

- Chemometrics is the science of extracting information from chemical systems by statistical means.
- An analyte is a substance or chemical constituent that is determined in an analytical procedure, such as a titration.
- the detection limit, lower limit of detection, or LOD (limit of detection), is the lowest quantity of a substance that can be distinguished from the absence of that substance (a blank value) within a stated confidence limit (generally 1%).
- The detection limit is estimated from the mean of the blank, the standard deviation of the blank and some confidence factor. Another consideration that affects the detection limit is the accuracy of the model used to predict concentration from the raw analytical signal.

11.4.1 Geometric notation

Chapter 12

Statistical Inference

12.1 Introduction to Inference

This chapter is concerned with data based decision-making. It is about making a decision which involves a population. The population is made up of a set of individual items. This could be, for example, a set of individuals or companies which constitute the market for your product. It could consist of the items being manufactured from a production line.
The sort of information needed for a decision may be a mean value, (e.g. How many items does an individual purchase per year, on average?) or a proportion (What proportion of items manufactured have a fault?). The associated decision may range from setting up extra capacity to cope with estimated demand, to stopping the production line for readjustment.
In most cases it is impossible to gather information about the whole population, so one has to collect information about a sample from the population and infer the required information about the population.

Statistical inference is the process of making inferences about populations from information provided by samples.

12.1.1 *Sampling

Random sampling is a sampling procedure by which each memmber of a population has an equal chance of being selected for a sample.

what does statistical inference refer to? Estimation and hypothesis testing

what are the names of the descriptive characteristics of populations and samples? paramaters and statistics respecgtively. In statistics, inferences are made about parameters by analysing their corresponding statistics.

How can representative samples be obtained? by random sampling.

The central limit theorem allows statisticians to use sample statistics to make inferences about the population parameters without knowing about the distribution of the parent population .

An interval estimate is the range of values used to estimate an unknown parameter together with the probability, known as the *confidence level*, that the unknown population parameters in within that range. Confidence intervals are conventionally centered around the point estimate.

The two numbers defining confidence interval are the lower confidence limit and the upper confidence limit, collectively as the confidence limits. A confidence interval expresses the degree of accuracy or confidence we have in an estimate.

12.2 ANOVA

ANOVA is a general technique that can be used to test the hypothesis that the means among two or more groups are equal, under the assumption that the sampled populations are normally distributed.

12.3 Normal Distribution: Worked Examples

MA4413 Computer Maths 3 January 2007

Q5. a) Assume that the amount of wine poured into a bottle has a normal distribution with a mean of 750ml and a variance of 144ml^2.

1. Calculate the probability that a bottle contains more than 765ml. (2 marks) 2. Calculate the probability that a bottle contains between 744ml and 759ml. (3 marks)

MA4403 Computer Maths 3 January 2006

A machine fills bags with animal feed. The nominal weight of a bag is 50kg. Because random variations the weight of a filled bag is normally distributed $N(\mu, \sigma^2)$. The variance (σ^2) is known to be 0.01kg^2 and μ is set by the operator to a particular value.

(i) If ? = 50kg calculate the probability of a bag containing less than 49.95kg? (ii) Calculate the value of "?" such that only 2% of the output are under the nominal weight?

MA4004 Engineering Statistics SPRING 2008

a) The amount of beer in a bottle has a normal distribution with mean 500ml and variance 25ml2. i) Calculate the probability that the amount of beer in the bottle is between 498ml and 504ml. ii) What volume is exceeded by 20% of the bottles? (6 marks)

12.3.1 The Standard Normal Distribution

12.3.2 Standardisation Formula

$$Z = (X - \mu)/\sigma \tag{12.1}$$

12.3.3 Example 2

A machine produces components whose thicknesses are normally distributed with a mean of 0.40 cm and a standard deviation of 0.02 cm. Components are rejected if they have a thickness outside the range 0.38 cm to 0.41 cm. (i) What is the probability that a component will have a thickness exceeding 0.41 cm? (4 marks) (ii) What is the probability that a component will have a thickness between 0.38 cm and 0.41 cm? (4 marks) (iii) What is the thickness below which 25% of the components will be? (4 marks)

12.3.4 Example 3

A charity believes that when it puts out an appeal for charitable donations the donations it receives will normally distributed with a mean 50 and standard deviation 6, and it is assumed that donations will be independent of each other.

- Find the probability that the first donation it receives will be greater than 40.
- Find the probability that it will be between 55 and 60.
- Find the value x such that 5% of donations are more than x.

Chapter 13

Bivariate Analysis: Linear Regression and Correlation

Chapter 14

Linear Regression

14.1 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:

$$Y = \alpha + \beta X + \epsilon \tag{14.1}$$

where α and β are constants and ϵ $N(0, \sigma^2)$, a random error term. The coefficients α and β 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.

14.1.1 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.

14.1.2 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

14.1.3 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 tdistribution 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].

14.1.4 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?

14.2 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.

14.2.1 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.

14.2.2 Regression

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

14.3 Inference for Regression

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

$$b \pm t_{1-\alpha/2,n-2} S.E.(b) \tag{14.2}$$

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

$$(0.0251643) \pm (0.0002656)(2.570582) = [0.0245, 0.0258] \tag{14.3}$$

14.3.1 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
```

14.4 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)

14.4.1 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.

Ansombe's Quartet

Set 1		Set 2		Set 3		Set 4	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Exercise 26: Correlation test

Assessment of tuna quality: We compare the Hunter L measure of lightness to the averages of consumer panel scores (recoded as integer values from 1 to 6 and averaged over 80 such values) in 9 lots of canned tuna. (Hollander & Wolfe (1973), p. 187f.)

```
x <- c(44.4, 45.9, 41.9, 53.3, 44.7, 44.1, 50.7, 45.2, 60.1)
y <- c( 2.6,  3.1,  2.5,  5.0,  3.6,  4.0,  5.2,  2.8,  3.8)
```

- Test the hypothesis that the correlation coefficient is not zero.
- Test the hypothesis that the correlation coefficient is positive.
- What is the test statistics in both cases?
- What is the p-value in both cases?
- Interpret the p-values.

The alternative hypothesis of interest is there is a correlation between the Hunter L value and the panel score.

```
> cor.test(x, y , alternative = "two.sided")

        Pearson's product-moment correlation

data:  x and y
t = 1.8411, df = 7, p-value = 0.1082
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1497426  0.8955795
sample estimates:
      cor
0.5711816
```

The alternative hypothesis of interest is that the Hunter L value is positively associated with the panel score.

```
> cor.test(x, y, alternative = "greater")

        Pearson's product-moment correlation

data:  x and y
t = 1.8411, df = 7, p-value = 0.05409
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
 -0.02223023  1.00000000
sample estimates:
      cor
0.5711816
```

```
> ct2 <- cor.test(x, y, alternative = "greater")
> names(ct2)
[1] "statistic"   "parameter"   "p.value"   "estimate"  "null.value"  "alternativ
[9] "conf.int"
>
> ct2$p.value
[1] 0.05408653
>
> ct2$statistic
       t
1.841083
```

Exercise 27: Kolmogorov-Smirnov test

```
x <- rnorm(50)
y <- runif(30)
# Do x and y come from the same distribution?
ks.test(x, y)

        Two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.46, p-value = 0.0004387
alternative hypothesis: two-sided
```

14.5 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).

14.5.1 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.

14.6 Advanced Regression

1. Multiple Linear regression
2. Deming regression
3. Weighted Linear Regression

14.7 Multiple Linear regression

In multiple linear regression, there are p explanatory variables, and the relationship between the dependent variable and the explanatory variables is represented by the following equation:

$$y = \tag{14.4}$$

Examples where multiple linear regression may be used include:

- Trying to predict an individuals income given several socio-economic characteristics.
- Trying to predict the overall examination performance of pupils in A levels, given the values of a set of exam scores at age 16.
- Trying to estimate systolic or diastolic blood pressure, given a variety of socioeconomic and behavioural characteristics (occupation, drinking smoking, age etc).

14.8 Deming regression

Whereas the ordinary linear regression method assumes that only the Y measurements are associated with random measurement errors, the Deming method takes measurement errors for both methods into account.

14.9 Weighted regression

14.10 scatterplots

The first part of the question will require the drawing of a scatter plot. When doing do, remember to label the axes, and to put in the relevant units. (i.e. Metres, Degrees, Hours etc)

The Explanatory variable is on the X-axis and the Response variable is on the Y Axis.

A Trend line will be useful in demonstrating what type of relationship exists between the response variable and the explanatory variable.

There are five possible plot types

- Strong positive linear relationship
- Weak positive linear relationship
- Strong negative linear relationship
- Weak negative linear relationship
- No Relationship

In the strong case - the points of the graph correspond to the trend line quite closely, whereas in the weak case they don't. In the positive case the response values Y increase as the explanatory values X increases. In the negative case the response values Y decrease as the explanatory values X increases.

Part 2 Correlation This requires a simple calculation based in values given and the relevant formula.

Strength of a linear relationship between X and Y

```
M=1000
CorrData=numeric(M)
for (i in 1:M)
{
CorrData[i] = cor(rnorm(10),rnorm(10))
}
```

14.11 Regression: R-Square

Any model is only as good as it is able to predict the actual outcome with accuracy. R-Square is a measure of how well the model is able to predict the changes in the actual data. R-Square ranges between 0 and 1, with values over 0.5 indicating a good fit between the predictions and actual data.

14.12 Regression: Multi-collinearity

Multi-collinearity is a condition when independent variables included in a model are correlated with each other. Multi-collinearity may result in redundant variables being included in the model, which in itself is not such a terrible thing. The real damage caused by multi-collinearity is that it causes large standard errors in estimating the coefficients. In simpler terms it causes the estimated t-statistics for correlated or multi-collinear variables to be insignificant, thus resulting in significant variables to appear to be insignificant. Multi-collinearity can be identified by the **Variance Inflation factor (VIF)**, which is a statistic calculated for each variable in a model. A VIF greater than 5 may suggest that the concerned variable is multi-collinear with others in the model and may need to be dropped. VIF cannot

be calculated with the Excel Regression package, it needs to be calculated using more sophisticated packages like SAS, SPSS or S-Plus.

Multicollinearity can be controlled by shrinkage techniques like Ridge Regressions, but a better strategy is to combine collinear variables using techniques like Factor Analysis or Principal Components Analysis.

Chapter 15

Chi Square Goodness of Fit tests

15.1 Contingency Tables

In the case of goodness of fit tests, there is only one categorical variable, such as the screen size of TV sets that have been sold, and what is tested is a hypothesis concerning the pattern of frequencies, or the distribution, of the variable.

The observed frequencies can be listed as a single row, or as a single column, of categories.

Tests for independence involve (at least) two categorical variables, and what is tested is the assumption that the variables are statistically independent.

Independence implies that knowledge of the category in which an observation is classified with respect to one variable has no affect on the probability of the other variable being in one of the several categories.

When two variables are involved, the observed frequencies are entered in a two-way classification table, or contingency table.

The dimensions of such tables are defined by $r \times c$, in which r indicates the number of rows and c indicates the number of columns.

If the null hypothesis of independence is rejected for classified data such as in Table 12.3, this indicates that the two variables are dependent and that there is a relationship between them. For Table 12.3, for instance, this would indicate that there is a relationship between age and the sex of stereo shop customers.

Given the hypothesis of independence of the two variables, the expected frequency associated with each cell of a contingency table should be proportionate to the total observed frequencies included in the column and in the row in which the cell is located as related to the total sample size.

Where fr is the total frequency in a given row and fc is the total frequency in a given column, a convenient formula for determining the expected frequency for the cell of the contingency table that is located in that row and column is

The general formula for the degrees of freedom associated with a test for independence is $df = (r - 1)(c - 1)$.

15.2 Chi Square

p_i = Expected proportion for digit i.

For this test we used the chi-squared test statistic which is given by:

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{15.1}$$

The Chi Square test tests a null hypothesis stating that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution. The events considered must be mutually exclusive and have total probability 1. A common case for this is where the events each cover an outcome of a categorical variable.

- X^2 = the test statistic that asymptotically approaches a χ^2 distribution.
- O_i = an observed frequency;
- E_i = an expected (theoretical) frequency, asserted by the null hypothesis;
- n = the number of possible outcomes of each event.

The chi-square statistic can then be used to calculate ap-value by comparing the value of the statistic to a chi-square distribution. The number of degrees of freedom is equal to the number of cells "n", minus the reduction in degrees of freedom, "p".

15.2.1 Chi Square example

In reading a burette to 0.01ml the final figure has to be estimated. The following frequency table gives the final figures of 40 such readings.

Digit	0	1	2	3	4	5	6	7
Frequency	1	6	4	5	3	11	2	8

The null hypothesis is that each digit has equal chances of occurring. Since we have ten digits this implies that each digit should have a 12.5% chance of occurring.

Mathematically we can express the null hypothesis as:

$H_0 : p_0 = p_1 = p_1 = \cdots = p_7 = 0.125$

$$X^2 = \frac{(1-5)^2}{5} + \frac{(6-5)^2}{5} + \frac{(4-5)^2}{5} + \frac{(5-5)^2}{5} + \frac{(3-5)^2}{5} + \frac{(11-5)^2}{5} + \frac{(2-5)^2}{5} + \frac{(8-5)^2}{5} \tag{15.2}$$

15.2.2 Goodness of fit example

Pressure readings are taken regularly from a meter. It transpires that, in a random sample of 100 such readings, 45 are less than 1, 35 are between 1 and 2, and 20 are between 2 and 3.

Perform a χ^2 goodness of fit test of the model that states that the readings are independent observations of a random variable that is uniformly distributed on (0, 3).

15.2.3 Chi Square contingency tables

In a survey the samples from five factories were examined for the number of skilled or unskilled workers.

Factory	skilled workers	unskilled workers
A	80	184
B	58	147
C	114	276
D	55	196
E	83	229

Does the population proportion of skilled and unskilled workers vary with the factory?

15.3 Chi Square

The table below shows the relationship between gender and party identification in a US state.

Democrat Independent Republican Total Male 279 73 225 577 Female 165 47 191 403 Total 444 120 416 980

Test for association between gender and party affiliation at two appropriate levels and comment on your results.

Set out the null hypothesis that there is no association between method of computation and gender against the alternative, that there is. Be careful to get these the correct way round!

H0: There is no association. H1: There is an association.

Work out the expected values. For example, you should work out the expected value for the number of males who use no aids from the following: (95/195) 22 = 10.7.

15.4 Chi Square Example

In a large country each district is permitted to have its own policy on the death penalty. Some districts choose to have it, others choose not to. The table below shows the relationship between having the death penalty (No, Yes) and the crime rate (Low, High) for a sample of 200 districts.

Death penalty	Low crime	High crime	Total
No	30	70	100
Yes	60	40	100
Total	90	110	200

Calculate the value of chi-squared for the table and say what you would conclude.

Chapter 16

Advanced Inference Procedures

16.1 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% of the values in a Gaussian population are more than 1.96 standard deviations from the mean, your first thought might be to conclude that the outlier comes from a different population if Z is greater than 1.96. This approach only works if you know the population mean and SD from other data. Although this is rarely the case in experimental science, it is often the case in quality control. You know the overall mean and SD from historical data, and want to know whether the latest value matches the others. This is the basis for quality control charts.

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.

16.2 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:

$$Q = \frac{\text{Gap}}{\text{Range}} \tag{16.1}$$

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.

16.3 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

16.3.1 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.

16.4 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.

16.5 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.

16.5.1 Shapiro-Wilk test: Example

```
> x<-rnorm(100, mean = 5, sd = 3)
> y<-runif(100, min = 2, max = 4)
> shapiro.test(x)

        Shapiro-Wilk normality test
data:  x
W = 0.9851, p-value = 0.3249
>
> shapiro.test(y)
        Shapiro-Wilk normality test
data:  y
W = 0.9585, p-value = 0.003151
```

Chapter 17

ANOVA and Experimental Design

17.1 Designed Experiments

- In observational studies you try to examine and measure things as they naturally occur. In experimental studies you impose some manipulation and measure its effects. It is easier to generalize from observational studies, but it is easier to determine causality from experimental studies.
- The objects on which an experiment is performed are called the experimental units. Human experimental units used to be called subjects," but we now call them research participants."
- The explanatory variables in an experiment are referred to as factors, while the values each factor can have are called its levels. A specifoc experimental condition applied to a set of units is called a treatment.
- The effectiveness of a treatment is generally measured by comparing the results of units that underwent the treatment to the results of units in a control condition. Units in a control condition go through the experiment but never receive the treatment.
- It is generally not a good idea to determine the eect of a treatment by comparing the units before the treatment to the same units after the treatment. Just being in an experiment can sometimes cause changes, even if the treatment itself doesn't really do anything. This is called the placebo effect.
- The basic idea of an experiment is that if you control for everything in the environment except for the factor that you manipulate, then any differences you observe between different groups in your experiment must be caused by the factor. It is therefore very important that the groups of experimental units you have in each level of your factor are basically the same. If there is some difference between the types of units in your groups, then the results might be caused by those differences instead of your treatment.
- Sometimes people use matching to make their groups similar. In this case you assign people to the different levels of your factor in such a way that for every person in one factor level you have a similar person in each of the other levels.

- The preferred way to make your groups similar is by randomization. In this case you just randomly pick people to be in each of your groups, relying on the fact that on average the differences between the groups will average out. Not only does this require less effort, but you don't have to know ahead of time what are the important things that you need to match on.

- There are many different ways to randomly assign your experimental units to conditions. You can use a table of random numbers, a mathematical calculator that produces random digits, or you can use statistical software.

- Randomization is one of the basic ideas behind statistics. When we randomly assign units to be in either the treatment or control conditions, there will be some differences between the groups. However, this is not a problem because mathematicians know a lot about how random variables work. When we see a difference between the groups we can figure out the probability that the difference is just due to the way that we assigned the units to the groups. If the probability is too low for us to believe that the results were just due to chance we say that the difference is statistically significant, and we conclude that our treatment had an effect.

- In conclusion, when designing an experiment you want to:

    1. Control the effects of irrelevant variables.
    2. Randomize the assignment of experimental units to treatment conditions
    3. Replicate your experiment across a number of experimental units.

17.1.1 2^2 Design

The simplest type of 2^k design is the 2^2, i.e. two factors, A nd B, each with two levels. We usually think of these levels as the low and high levels of the factor.

17.2 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			

17.2.1 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.

17.2.2 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.

17.2.3 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.

17.2.4 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

17.2.5 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.

17.2.6 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.

17.3 Fractional factorial design

(d) Define the following terms used in fractional factorial design; Defining relation, Generator, Confounding, Resolution. Which design resolution is considered optimal?

17.4 Factorial Design

Factorial experiments permit researchers to study behavior under conditions in which independent variables, called in this context factors, are varied simultaneously.

Thus, researchers can investigate the joint effect of two or more factors on a dependent variable. The factorial design also facilitates the study of interactions, illuminating the effects of different conditions of the experiment on the identifiable subgroups of subjects participating in the experiment.

A full factorial experiment is an experiment whose design consists of two or more factors, each with discrete possible values or "levels", and whose experimental units take on all possible combinations of these levels across all such factors. A full factorial design may also be called a fully-crossed design. Such an experiment allows studying the effect of each factor on the response variable, as well as the effects of interactions between factors on the response variable.

For the vast majority of factorial experiments, each factor has only two levels. For example, with two factors each taking two levels, a factorial experiment would have four treatment combinations in total, and is usually called a 2×2 factorial design.

17.5 Completely Randomized Design

17.5.1 Questions

- Give the principal features of a balanced completely randomised design, and explain the role of replication in such a design.
- State the statistical model for this design, define the terms in the model and state the standard assumptions made about the error term.
- Briefly explain the principles of randomisation and replication, in the context of a completely randomised experimental design. Write down the model equation for a completely randomised design having equal numbers of replicates in all treatment groups, defining all the symbols that you use.

17.6 Orthogonal Array

Orthogonal array testing is a systematic, statistical way of testing. Orthogonal arrays can be applied in user interface testing, system testing, regression testing, configuration testing and performance testing. All orthogonal vectors exhibit orthogonality. Orthogonal vectors exhibit the following properties: Each of the vectors conveys information different from that of any other vector in the sequence, i.e., each vector conveys unique information therefore avoiding redundancy. On a linear addition, the signals may be separated easily. Each of the vectors is statistically independent of the others. When linearly added, the resultant is the arithmetic sum of the individual components.

17.6.1 Two Factor Interaction

The effects of interest in the 2^2 design are the main effects A and B and the two-factor interaction AB. It is easy to estimate the effects of these factors.

$$A = \frac{a + ab}{2n} - \frac{b + (1)}{2n} \tag{17.1}$$

$$B = \frac{b + ab}{2n} - \frac{a + (1)}{2n} \tag{17.2}$$

$$AB = \frac{ab + (1)}{2n} - \frac{a + b}{2n} \tag{17.3}$$

$$\tag{17.4}$$

The Sums of Squares formulae are

$$SS_A = \frac{[(a+ab)-(b+(1))]^2}{4n^2} \tag{17.5}$$

$$SS_B = \frac{[(b+ab)-(a+(1))]^2}{4n^2} \tag{17.6}$$

$$SS_{AB} = \frac{[(ab+(1))-(a+b)]^2}{4n^2} \tag{17.7}$$

$$\tag{17.8}$$

Chapter 18

Advanced Distribution Theory

18.1 Mixed Joint Probability Distribution

So far we've looked pairs of random variables where both variables are either discrete or continuous. A joint pair of random variables can also be composed of one discrete and one continuous random variable. This gives rise to what is known as a mixed joint probability distribution. The density function for a mixed probability distribution is given by

18.2 Conditional Probability Distribution

Conditional Probability Distributions arise from joint probability distributions where by we need to know that probability of one event given that the other event has happened, and the random variables behind these events are joint. Conditional probability distributions can be discrete or continuous, but the follow the same notation i.e.

18.3 Joint Distribution Functions

Thus far, we have concerned ourselves with the probability distribution of a single random variable. However, we are often interested in probability statements concerning two or more random variables. To deal with such probabilities, we define, for any two random variables X and Y , the joint cumulative probability distribution function of X and Y by

$$F(a, b) = PX < a, Y < b, \qquad (-\infty < a, b < \infty) \tag{18.1}$$

Chapter 19

Statistics for Chemists

19.1 Quantitative nature of analytical chemistry

Modern analytical chemistry is overwhelmingly a quantitative science. A quantitative answer is much more valuable than a qualitative one. It may be useful for an analyst to claim to have detected some boron in a distilled water sample, but it is much more useful to be able to say how much boron is present.

Often it is only a quantitative result that has any value at all.For example, almost all samples of (human) blood serum contain albumin; the only question is, how much ? Even where a qualitative answer is required, quantitative methods are used to obtain it.

Quantitative approaches might be used to compare two soil samples. For example, they might be subjected to a particle size analysis, in which the proportions of the soil particles falling within a number say 10, of particle-size ranges are determined. Each sample would then be characterized by these 10 pieces of data, which could then be used to provide a quantitative assessment of their similarity.

19.1.1 Errors in quantitative analysis

Since quantitative studies play a dominant role in any analytical laboratory, it must be accepted that the errors that occur in such studies are of supreme importance. No quantitative results are of any value unless they are accompanied by some estimate of the errors inherent in them!

Example 1 - detecting a new analytical reagent

- A chemist synthesizes an analytical reagent that is believed to be entirely new.
- The compound is studied using a spectrometric method and gives a value of 104.
- The chemist finds that no compound previously discovered has yielded a value of more than 100.
- Has the chemist really discovered a new compound?
- The answer lies in the degree of reliance to experimental value of 104.
- If the result is correct to within 2 (arbitrary) units, i.e. the true value probably lies in the range 102 ± 2, then a new material has probably been discovered.
- If, however, investigations show that the error may amount to 10 units i.e. 104 ± 10, then it is quite likely that the true value is actually less than 100, in which case a new discovery is tar from certain.
- A knowledge of the experimental errors is crucial!!

Example 2 - replicates in a titrimetric experiment

- Analysts commonly perform several replicate determinations in the course of a single experiment.
- An analyst performs a titrimetric experiment four times and obtains values of 24.69,24.73,24.77 and 25.39 ml.
- All four values are different, because of the variations inherent in the measurements
- The fourth value (25.39 ml) is substantially different from the other three.
- Can it be safely rejected, so that (for example) the mean titre is reported as 24.73 ml, the average of the other three readings?

19.2 Comapring Methods of Measurement

19.2.1 The Bland Altman plot

The Bland Altman plot (Bland & Altman, 1986 and 1999) is a statistical method to compare two measurements techniques. In this graphical method the differences (or alternatively the ratios) between the two techniques are plotted against the averages of the two techniques. Horizontal lines are drawn at the mean difference, and at the limits of agreement, which are defined as the mean difference plus and minus 1.96 times the standard deviation of the differences.

Chapter 20

Information Theory and Data Compression

20.1 Data Compression

1. Explain what an optimal code is, in the context of data compression.
2. Are Huffman Codes Optimal?

The frequency of 0 as an input to a binary channel is 0.6. If O is the input, then 0 is the output with probability 0.8. If l is the input, then l is the output with probability 0.9.

a. (4 marks) Calculate the information per bit contained in the input.

b. (2 marks)Calculate the probability that the output is 0.

c. (2 marks) Calculate the probability that the output is l,

d. (2 marks) Calculate the probability that the input is 0 given that the output is O.

e. (2 marks) Calculate the probability that the input is l given that the output is 1,

f. (2 marks) Calculate the probability that the input is l given that the output is O.

g. (2 marks) Calculate the probability that the input is 0 given that the output is l.

h. (6 marks) Calculate the amount of information transmitted by the channel.

i. (3 marks) Derive the globally optimal reconstruction rule.

Information Theory

- $I(p) = -log_2(p) = log_2(1/p)$
- $I(pq) = I(p) + I(q)$
- $H = -\sum_{i=1}^{m} p_i \; log_2(p_i)$
- $E(L) = \sum_{i=1}^{m} l_i p_i$
- Efficiency $= H/E(L)$
- $I(X;Y) = H(X) + H(Y) - H(X,Y)$
- $I(X;Y) = H(X) - H(X|Y)$
- $P(C[r]) = \sum_{j=1}^{m} P(C[r]|Y = d_j)P(Y = d_j)$

An inspector of computer parts selects a random sample of components from a large batch to decide whether or not to audit the full batch. (i) If 20% or more of the sample is defective, the entire batch is inspected, Calculate the probability of this happening if it is thought that the population contains 4% defective components and a sample of 20 is selected. (ii) If 10% or more of the sample is defective, the entire batch is inspected. Calculate the probability of this happening if it is thought that the population contains 4% defective components and a sample of 50 is selected. (10 marks) (d) A model of an online computer system gives a mean time to retrieve a record from a direct access storage system device of 200 milliseconds with a standard deviation of 58 milliseconds. If it can be assumed that the data are normally distributed: (i) What proportion of retrieval times will be greater than 75 milliseconds? (ii) What proportion of retrieval times will be between 150 milliseconds and 250 milliseconds? (iii) What is the retrieval time below which 10% of retrieval times will be? (9 marks)

$$
\begin{aligned}
r &= \frac{S_{XY}}{\sqrt{S_X^2} \times \sqrt{S_Y^2}}. \\
S_{XY} &= \sum xy - \frac{\sum x \times \sum y}{n}. \\
S_X^2 &= \sum x^2 - \frac{(\sum x)^2}{n}. \\
S_Y^2 &= \sum y^2 - \frac{(\sum y)^2}{n}.
\end{aligned}
$$

Regression coefficients

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{S_{XY}}{S_X^2}. \\
\hat{\beta}_0 &= \bar{y} - (\hat{\beta}_1 \times \bar{x}).
\end{aligned}
$$

1. Given four symbols A, B, C and D, the symbols occur with an equal probability. What is the entropy of the distribution?

2. Suppose they occur with probabilities 0.5, 0.25, 0.125 and 0.125 respectively. What is the entropy associated with the event (experiment)?

3. Suppose the probabilities are 1,0,0,0. What is the entropy?

Chapter 21

Assorted Topics

21.1 Testing Normality

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

21.2 Mallow's Cp

Mallow's Cp coefficient is a metric used in model selection to dissuade the use of over-fitted models.

$$Cp = \frac{RSS}{\hat{\sigma}^2} - (n - 2p) \tag{21.1}$$

This coefficient should be minimized over p.

1. Multicollinearity
2. Biometrics
3. Variance Inflation Factor

   (a) The heights for a group of forty rowing club members are tabulated as follows;

141	148	149	149	155	156	167	169	169	170
171	173	175	176	177	179	182	182	183	183
183	184	184	184	185	185	185	186	186	189
191	191	191	191	192	192	192	193	194	199

   i. (6 marks) Summarize the data in the above table using a frequency table. Use 6 class intervals, with 140 as the lower limit of the first interval.

ii. (6 marks) Draw a histogram for the above data.

iii. (4 marks) Comment on the shape of the histogram. Based on the shape of the histogram, what is the best measure of centrality and variability?

iv. (12 marks) Construct a box plot for the above data. Clearly demonstrate how all of the necessary values were computed.

(b) Data on the construction durations (measured in months) were collected for a random sample of similar infrastructure projects in two neighbouring countries: A and B.
The durations for country A were collected and tabulated as follows;

41	44	44	43	37	37	34

i. (2 marks) Calculate the mean of the durations for country A.

ii. (4 marks) Calculate the variance for country A.

iii. (2 marks) Calculate the standard deviation for country A.

iv. (2 marks) Calculate the coefficient of variation for country A.

v. (2 marks) For the sample in country B, the mean of the durations was found to be 36 weeks, with a standard deviation of 6 weeks. In which country do the durations show a more dispersed distribution?

21.3 Useful formulae

21.3.1 Mathematics

1. Logarithms: If $N = b^n$, then $\log_b N = n$.

2. Compound interest:

$$P_t = P_0 (1+i)^t, \qquad P_t = P_0 \left(1 + \frac{i}{m}\right)^{mt}, \qquad P_t = P_0 \mathrm{e}^{it}.$$

3. Matrices:

   (a) Inverse of a 2*2 matrix:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}.$$

   (b) Deteminant of a 2*2 matrix:

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$

   (c) Cramer's Rule: If

$$\begin{aligned} a_1 x + b_1 y &= d_1, \\ a_2 x + b_2 y &= d_2, \end{aligned}$$

   then

$$x = \frac{\begin{vmatrix} d_1 & b_1 \\ d_2 & b_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}}, \qquad y = \frac{\begin{vmatrix} a_1 & d_1 \\ a_2 & d_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}}.$$

21.3.2 Statistics

1. Sample mean

$$\bar{x} = \frac{\sum x_i}{n}.$$

2. Sample standard deviation

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}.$$

3. Conditional probability:

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}.$$

4. Binomial probability function

$$f(x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} \qquad \text{where} \qquad \left(\begin{array}{c} n \\ x \end{array} \right) = \frac{n!}{x!\,(n-x)!}.$$

5. Poisson probability function

$$f(x) = \frac{m^x \mathrm{e}^{-m}}{x!}.$$

6. Exponential probability distribution

$$P(X \leq k) = 1 - e^{-k/\mu}$$

Question 3:

- There are 100 passengers booked for a flight.
- It was found that the weight of a passenger and its luggage is random with average value of 75[kg] and standard deviation of 15[kg].
- The plane is overloaded if the total weight of its fuel, passengers and their luggage exceeds 9500[kg].
- The plane has 1750[kg] of fuel.
- ***Remark:*** Assume that the plane starts with a tank of 1750 kg of fuel. If the plane is overloaded, fuel is pumped out of the tanks accordingly.

Part 1: Introduce appropriate random variables and express the event that the plane is overloaded by means of these random variables.

Solution:

- Let X_i be the weight of the i-th passenger and his/her luggage.
- Then the variable Y defined as

$$Y = \sum_{i=1}^{100} X_i$$

  is the total weight of passengers and their luggage.
- The plane is overloaded if $Y + 1750 > 9500$, i.e. $Y > 7750$.

Part 2: Use the Central Limit Theorem to approximate the probability that the plane will be overloaded? In the solution you can use the following

```
normcdf(7750,7500,sqrt(22500))
ans = 0.95221
```

Solution:

- By the Central Limit Theorem, $Y = 100X$

$$Y \sim N(100 \times 75, 100 \times 15^2)$$

$$Y \sim N(7500, 22500)$$

- ***Remark*** *we do not need to know the distribution of X. The CLT states that sample statistics (such as sample sums or sample means) are normally distributed, even if the underlying variable is not normally distributed.*

- From the conditions of the problem we need to find

  $$P(Y > 7500) = 1 - P(Y \leq 7750) = 1 - F(7750)$$

  where F is the cdf of $N(7500, 22500)$.

- This can be found using the provided Matlab code.

  $$F(7750) = P(Y \leq 7750) = 0.95221$$

- It follows that there is about $1 - 0.95221 = 0.04779$ of chance for the plane to be overloaded.

Part 3: How much fuel the plane can take in order for the chances of overload to be 1%?

In the solution one can use the following output from Matlab:

```
norminv(0.99,7500,sqrt(22500))
ans = 7849.0
```

- If Y exceeds 7750 [kg], fuel is pumped out of the tanks accordingly, reducing the fuel level to F

- We can say

  $$Y + F = 9500$$

  As Y increases, F must decrease accordingly.

- *If Y does not exceed 7750, we don't need to pump any out. We are not interested in this.*

- The question can be phrased a different way: Find the fuel level x that there is a 1% probability that the fuel level must not exceed in order for the plane not to be overloaded.
- Using the MatLab code, we know that the probability of Y being less than or equal to 7849.0 (99%)
$$P(Y \leq 7849) = 0.99$$
- Necessarily
$$P(Y \geq 7849) = 0.01$$
- There is a 1% chance that the total weight of passengers will be greater than 7849 kg.
- If the aggregate value for Y exceeds 7849, the fuel level must be reduced to 1651 [kg] or less.
- So there is a 1% chance that the fuel level will have to be reduced to 1651 [kg] at least [Answer].

Part 1: Examine data graphically and check if some linear relation can be involved.

[See Graph]

Part 2: Find the LSE of the linear regression line and present it graphically together with the data.

We need to compute

- The Intercept Estimate $\hat{\alpha}$

- The Slope Estimate $\hat{\beta}$
- See Formulae: The predicted value $\hat{y}$ of y given a value of the explanatory variable X

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

- $\bar{x} = 8.266667$, $\bar{y} = 1.566667$, $n = 6$
- Compute the slope estimate

$$\hat{\beta} = \frac{\sum(xy) - n\bar{x}\bar{y}}{\sum(x^2) - n\bar{x}^2}$$

$$= \frac{62.06 - (6 \times 8.266 \times 1.566)}{538.58 - (6 \times 1.566^2)}$$

$$= \frac{-15.64666}{128.55333} = -0.12171$$

- Now compute the intercept estimate

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 1.566 - (-0.12171 \times 8.266) = 2.5728$$

- The regression line is therefore

$$\hat{y} = 2.5728 - 0.1217x$$

Part 3: What would be the prediction of the mercury content at 2[m] from the polarograph?

$$\hat{y} = 2.573 - (0.127 \times 2) = 2.33\text{ng/g}$$

1. Multicollinearity
2. Biometrics
3. Variance Inflation Factor
   (a) The heights for a group of forty rowing club members are tabulated as follows;

141	148	149	149	155	156	167	169	169	170
171	173	175	176	177	179	182	182	183	183
183	184	184	184	185	185	185	186	186	189
191	191	191	191	192	192	192	193	194	199

      i. (6 marks) Summarize the data in the above table using a frequency table. Use 6 class intervals, with 140 as the lower limit of the first interval.
      ii. (6 marks) Draw a histogram for the above data.
      iii. (4 marks) Comment on the shape of the histogram. Based on the shape of the histogram, what is the best measure of centrality and variability?
      iv. (12 marks) Construct a box plot for the above data. Clearly demonstrate how all of the necessary values were computed.

   (b) Data on the construction durations (measured in months) were collected for a random sample of similar infrastructure projects in two neighbouring countries: A and B.

The durations for country A were collected and tabulated as follows;

41	44	44	43	37	37	34

i. (2 marks) Calculate the mean of the durations for country A.

ii. (4 marks) Calculate the variance for country A.

iii. (2 marks) Calculate the standard deviation for country A.

iv. (2 marks) Calculate the coefficient of variation for country A.

v. (2 marks) For the sample in country B, the mean of the durations was found to be 36 weeks, with a standard deviation of 6 weeks. In which country do the durations show a more dispersed distribution?

4. (a) An electronics assembly subcontractor receives resistors from two suppliers: Deltatech provides 70% of the subcontractors's resistors while another company, Echelon, supplies the remainder. 1% of the resistors provided by Deltatech fail the quality control test, while 2% of the chips from Echelon also fail the quality control test.

   i. (5 marks)What is the probability that a resistor will fail the quality control test?

   ii. (4 marks)What is the probability that a resistor that fails the quality control test was supplied by Echelon?

   (b) It is estimated by a particular bank that 25% of credit card customers pay only the minimum amount due on their monthly credit card bill and do not pay the total amount due. 50 credit card customers are randomly selected.

   i. (3 marks) What is the probability that 9 or more of the selected customers pay only the minimum amount due?

   ii. (3 marks) What is the probability that less than 6 of the selected customers pay only the minimum amount due?

   iii. (3 marks) What is the probability that more

than 5 but less than 10 of the selected customers pay only the minimum amount due?

(c) The average lifespan of a PC monitor is 6 years. You may assume that the lifespan of monitors follows an exponential probability distribution.

i. (3 marks) What is the probability that the lifespan of the monitor will be at least 5 years?

ii. (3 marks) What is the probability that the lifespan of the monitor will not exceed 4 years?

iii. (3 marks) What is the probability of the lifespan being between 5 years and 7 years?

(d) A machine is used to package bags of potato chips. Records of the packaging machine indicate that its fill weights are normally distributed with a mean of 455 grams per bag and a standard deviation of 10 grams.

i. (5 marks) What proportion of bags filled by this machine will contain more than 470 grams in the long run?

ii. (5 marks) What proportion of bags filled by this machine will contain less than 445

grams in the long run?

iii. (3 marks) What proportion of bags filled by this machine will be between 465 grams and 475 grams in the long run?

21.4 Useful formulae

21.4.1 Mathematics

1. Logarithms: If $N = b^n$, then $\log_b N = n$.
2. Compound interest:

$$P_t = P_0 (1+i)^t, \qquad P_t = P_0 \left(1 + \frac{i}{m}\right)^{mt}, \qquad P_t = P_0$$

3. Matrices:
   (a) Inverse of a 2*2 matrix:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}.$$

   (b) Deteminant of a 2*2 matrix:

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$

   (c) Cramer's Rule: If

$$\begin{aligned} a_1 x + b_1 y &= d_1, \\ a_2 x + b_2 y &= d_2, \end{aligned}$$

   then

$$x = \frac{\begin{vmatrix} d_1 & b_1 \\ d_2 & b_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}}, \qquad y = \frac{\begin{vmatrix} a_1 & d_1 \\ a_2 & d_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}}.$$

21.4.2 Statistics

1. Sample mean
$$\bar{x} = \frac{\sum x_i}{n}.$$
2. Sample standard deviation
$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}.$$
3. Conditional probability:
$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}.$$

Question 3:

- There are 100 passengers booked for a flight.
- It was found that the weight of a passenger and its luggage is random with average value of 75[kg] and standard deviation of 15[kg].
- The plane is overloaded if the total weight of its fuel, passengers and their luggage exceeds 9500[kg].
- The plane has 1750[kg] of fuel.
- ***Remark:*** Assume that the plane starts with a tank of 1750 kg of fuel. If the plane is overloaded, fuel is pumped out of the tanks accordingly.

Part 1: Introduce appropriate random variables and express the event that the plane is overloaded by means of these random variables.

Solution:

- Let X_i be the weight of the i-th passenger and his/her luggage.
- Then the variable Y defined as

$$Y = \sum_{i=1}^{100} X_i$$

  is the total weight of passengers and their luggage.
- The plane is overloaded if $Y + 1750 > 9500$, i.e. $Y > 7750$.

Part 2: Use the Central Limit Theorem to approximate the probability that the plane will be overloaded? In the solution you can use the following

```
normcdf(7750,7500,sqrt(22500))
ans = 0.95221
```

Solution:

- By the Central Limit Theorem, $Y = 100X$

$$Y \sim N(100 \times 75, 100 \times 15^2)$$

$$Y \sim N(7500, 22500)$$

- ***Remark*** *we do not need to know the distribution of X. The CLT states that sample statistics (such as sample sums or sample means) are normally distributed, even if the underlying variable is not normally distributed.*
- From the conditions of the problem we need to find

  $$P(Y > 7500) = 1 - P(Y \leq 7750) = 1 - F(7750)$$

  where F is the cdf of $N(7500, 22500)$.
- This can be found using the provided Matlab code.

  $$F(7750) = P(Y \leq 7750) = 0.95221$$

- It follows that there is about $1 - 0.95221 = 0.04779$ of chance for the plane to be overloaded.

Part 3: How much fuel the plane can take in order for the chances of overload to be 1%?

In the solution one can use the following output from Matlab:

```
norminv(0.99,7500,sqrt(22500))
ans = 7849.0
```

- If Y exceeds 7750 [kg], fuel is pumped out of the tanks accordingly, reducing the fuel level to F

- We can say
$$Y + F = 9500$$
As Y increases, F must decrease accordingly.
- *If Y does not exceed 7750, we don't need to pump any out. We are not interested in this.*
- The question can be phrased a different way: Find the fuel level x that there is a 1% probability that the fuel level must not exceed in order for the plane not to be overloaded.
- Using the MatLab code, we know that the probability of Y being less than or equal to 7849.0 (99%)
$$P(Y \leq 7849) = 0.99$$
- Necessarily
$$P(Y \geq 7849) = 0.01$$
- There is a 1% chance that the total weight of passengers will be greater than 7849 kg.
- If the aggregate value for Y exceeds 7849, the fuel level must be reduced to 1651 [kg] or less.
- So there is a 1% chance that the fuel level will have to be reduced to 1651 [kg] at least [Answer].

Part 1: Examine data graphically and check if some linear relation can be involved.

[See Graph]

Part 2: Find the LSE of the linear regression line and present it graphically together with the data.

We need to compute

- The Intercept Estimate $\hat{\alpha}$
- The Slope Estimate $\hat{\beta}$
- See Formulae: The predicted value $\hat{y}$ of y given a value of the explanatory variable X

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

- $\bar{x} = 8.266667$, $\bar{y} = 1.566667$, $n = 6$
- Compute the slope estimate

$$\hat{\beta} = \frac{\sum(xy) - n\bar{x}\bar{y}}{\sum(x^2) - n\bar{x}^2}$$

$$= \frac{62.06 - (6 \times 8.266 \times 1.566)}{538.58 - (6 \times 1.566^2)}$$

$$= \frac{-15.64666}{128.55333} = -0.12171$$

- Now compute the intercept estimate

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 1.566 - (-0.12171 \times 8.266) = 2.5728$$

- The regression line is therefore

$$\hat{y} = 2.5728 - 0.1217x$$

Part 3: What would be the prediction of the mercury content at 2[m] from the polarograph?

$$\hat{y} = 2.573 - (0.127 \times 2) = 2.33\text{ng/g}$$

21.5 The binomial distribution

The binomial distribution is a discrete probability distribution that is applicable as a model for decisionmaking situations in which a sampling process can be assumed to conform to a Bernoulli process. A Bernoulli process is a sampling process in which (1) Only two mutually exclusive possible outcomes are possible in each trial, or observation. For convenience these are called success and failure. (2) The outcomes in the series of trials, or observations, constitute independent events. (3) The probability of success in each trial, denoted by p, remains constant from trial to trial. That is, the process is stationary. The binomial distribution can be used to determine the probability of obtaining a designated number of successes in a Bernoulli process. Three values are required: the designated number of successes (X); the number of trials, or observations (n); and the probability of success in each trial (p). Where $q = (1 - p)$, the formula for determining the probability of a specific number of successes X

for a binomial distribution is

Formula

21.6 The hypergeometric distribution

When sampling is done without replacement of each sampled item taken from a finite population of items, the Bernoulli process does not apply because there is a systematic change in the probability of success as items are removed from the population. When sampling without replacement is used in a situation that would otherwise qualify as a Bernoulli process, the hypergeometric distribution is the appropriate discrete probability distribution. Given that X is the designated number of successes, N is the total number of items in the population, T is the total number of successes included in the population, and n is the number of items in the sample, the formula for determining hypergeometric probabilities is

21.7 Quantiles

The quantile (this term was first used by Kendall, 1940) of a distribution of values is a number xp such that a proportion p of the population values are less than or equal to xp. For example, the .25 quantile

(also referred to as the 25th percentile or lower quartile) of a variable is a value (xp) such that 25% (p) of the values of the variable fall below that value.

Similarly, the 0.75 quantile (also referred to as the 75th percentile or upper quartile) is a value such that 75% of the values of the variable fall below that value and is calculated accordingly.

See

21.8 Autocorrelation

Autocorrelation can be detected using correlograms.

21.9 Cause and Effect Diagrams

The Ishikawa diagram (also known as a fishbone diagram) is used to associate multiple causes with a single effect.

21.10 Bonferroni Test

A type of multiple comparison test used in statistical analysis. When an experimenter performs enough tests, he or she will eventually end up with a result that shows statistical significance, even if there is none. If a particular test yields correct results 99% of the time, running 100 tests could lead to a false

result somewhere in the mix. The Bonferroni test attempts to prevent data from incorrectly appearing to be statistically significant by lowering the alpha value.

The Bonferroni test, also known as the "Bonferroni correction" or "Bonferroni adjustment" suggests that the "p" value for each test must be equal to alpha divided by the number of tests.

21.11 Control Charts for Attributes

Control charts could also be prepared for ***attributes***, e.g. the proportion showing the proportion that is defective in some way. The Central line would be set at the average proportion defective expected, and the actual amount defective would be plotted on the chart.

21.12 Cronbach's Alpha

Cronbach's α is defined as

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K} \sigma_{Y_i}^2}{\sigma_X^2}\right)$$

21.13 Data Types

- Categorical Data
- Nominal Data
- Ordinal Data
- Interval Data
- Ratio Data

21.14 Dendrograms

Dendrograms are also known as "Tree Diagrams".

21.15 Durbin Watson Statistic

A number that tests for autocorrelation in the residuals from a statistical regression analysis. The Durbin-Watson statistic is always between 0 and 4. A value of 2 means that there is no autocorrelation in the sample. Values approaching 0 indicate positive autocorrelation and values toward 4 indicate negative autocorrelation.

$$d = \frac{n}{n} \tag{21.2}$$

21.16 Experimentally Weighted Moving Average

21.17 Exponential Smoothing

New Forecast = Old Forecast + α(Latest Observation - Old Forecast).

- Greater weight is given to more recent data.
- All past data is incorporated, and there is no cut-off point as with moving averages.

21.18 Finite Population Correction Factor

Where the sample size exceeds 5% of the population, the Finite Population Correction Factor should be applied.

$$\sqrt{\frac{N-n}{N-1}}$$

21.19 Huffman Codes: Characteristics

Huffman codes are prefix-free binary code trees, therefore all substantial considerations apply accordingly.

Codes generated by the Huffman algorithm achieve the ideal code length up to the bit boundary. The maximum deviation is less than 1 bit.

21.20 Integration in Probability

$$\mathrm{E}(X) = \int_l^u x f(X) dx$$

$$P(X \leq A) = \int_l^A f(X) dx$$

$$\mathrm{Var}(X) = \int_b^a x f(X) dx$$

$$\mu = \bar{x} \pm t_{(\nu,\sigma/2)} \frac{s}{\sqrt{n}}$$

21.21 Monte Carlo Simulation

A problem solving technique used to approximate the probability of certain outcomes by running multiple trial runs, called simulations, using random variables.

21.22 Moving Averages : Characteristics

- The different moving averages produce different results.
- The greater the number of periods in the moving average, the greater the smoothing effect.

21.23 Non-Sampling Error

A statistical error caused by human error to which a specific statistical analysis is exposed. These errors can include, but are not limited to, data entry errors, biased questions in a questionnaire, biased processing/decision making, inappropriate analysis conclusions and false information provided by respondents.

21.24 Probability Distribution

A statistical function that describes all the possible values and likelihoods that a random variable can take within a given range. This range will be between the minimum and maximum statistically possible values, but where the possible value is likely to be plotted on the probability distribution depends on a number of factors, including the distributions mean, standard deviation, skewness and kurtosis.

21.25 Process Capability Analysis

Process Capability Indices for a characteristic of interest from a continuous process can be obtained using the command process.capability.

Lower and Upper Specification Limits must be specified.

$$c_p = \frac{\text{USL} - \text{LSL}}{6 \times s} = \frac{12}{6 \times 1.956} = 1.02$$

$$c_{pk}(\text{upper}) = \frac{\text{USL} - \bar{x}}{3 \times s} = \frac{506 - 500.38}{3 \times 1.956} = 0.96$$

$$c_{pk}(\text{lower}) = \frac{\bar{x} - \text{USL}}{3 \times s} = \frac{500.38 - 494}{3 \times 1.956} = 104$$

21.26 Properties of Good Estimators

Unbiased discuss

Consistency discuss

Efficiency discuss

Sufficiency discuss

21.27 Seasonality

A characteristic of a time series in which the data experiences regular and predictable changes which recur every calendar year. Any predictable change or pattern in a time series that recurs or repeats over a one-year period can be said to be seasonal.

Note that seasonal effects are different from cyclical effects, as seasonal cycles are contained within one calendar year, while cyclical effects (such as boosted

sales due to low unemployment rates) can span time periods shorter or longer than one calendar year

21.28 Shannon-Fano Coding

At about 1960 Claude E. Shannon (MIT) and Robert M. Fano (Bell Laboratories) had developed a coding procedure to generate a binary code tree. The procedure evaluates the symbol's probability and assigns code words with a corresponding code length.

Compared to other methods the Shannon-Fano coding is easy to implement. In practical operation Shannon-Fano coding is not of larger importance. This is especially caused by the lower code efficiency in comparison to Huffman coding as demonstrated later.

21.29 Statistical Process Control

-
- Statistical Process Control is, in effect, continuous hypothesis testing.

21.30 Survivorship Function

The survivorship function (commonly denoted as R(t)) is the complement to the cumulative distri-

bution function (i.e., R(t)=1-F(t)); the survivorship function is also referred to as the reliability or survival function (since it describes the probability of not failing or of surviving until a certain time t

21.31 Time Series

A sequence of numerical data points in successive order, usually occurring in uniform intervals. In plain English, a time series is simply a sequence of numbers collected at regular intervals over a period of time.

21.32 Trimmed Means

For certain data sets, an option is available to trim the extreme values from the distribution of values of a variable. For example, we can trim (i.e., remove) the lowest 5% and the highest 5% from the distribution of values. The mean of the trimmed distribution of values is referred to as a "trimmed mean".

21.33 Tukey HSD

This post hoc test (or multiple comparison test) can be used to determine the significant differences between group means in an analysis of variance set-

ting. The Tukey HSD is generally more conservative than the Fisher LSD test but less conservative than Scheffe's test

21.34 Wilcoxon Test

The Wilcoxon test is a nonparametric alternative to t-test for dependent samples. It is designed to test a hypothesis about the location (median) of a population distribution. It often involves the use of matched pairs, for example, "before" and "after" data, in which case it tests for a median difference of zero.

This procedure assumes that the variables under consideration were measured on a scale that allows the rank ordering of observations based on each variable (i.e., ordinal scale) and that allows rank ordering of the differences between variables

21.35 The Inverse Value Problem

21.36 Limits of Detection (Chemistry)

21.37 Multicollinearity

21.38 Mutually Exclusive Events

Mutually exclusive events are events that cannot happen at the same time.

$$P(AandB) = P(A) + P(B)$$

21.39 OC function

Type II error : probability of accepting a process as being in control, when in fact it is not. Based on the following OC

21.40 Skewness: Pearson Coefficient of Skewness

$$S_k = \frac{3(\text{Mean} - \text{Median})}{\sigma}$$

21.41 Spearman Rank Correlation

$$1 - \frac{6\left(\sum d^2 + \frac{t^3 - t}{12}\right)}{n(n^2 - 1)}$$

The adjustment for tied values $\frac{t^3-t}{12}$, where t is the number of tied values

21.42 The Stepping Stone Method (Transportation)

- Start at a cell that has no allocation. (This cell will be a "plus" cell)
- Choose a cell that has received an allocation (This cell will be a "minus" cell)
- Right Angle Turn -
- Keep going until you have returned to the origin cell.

21.43 Variance Inflation Factor

Multicollineaity

21.44 Weibull Distribution

Probability distribution