

# Introduction to Applied Statistics for Medical Researchers

Short Course: 12-13th May 2011, AO Research Institute Davos

Universität Bern | Universität Zürich

**vetsuisse-fakultät**

Instructor: F. I. Lewis PhD  
Applied Statistician - Section of Epidemiology  
[fraseriain.lewis@uzh.ch](mailto:fraseriain.lewis@uzh.ch); [www.vetepi.uzh.ch](http://www.vetepi.uzh.ch)



**University of  
Zurich<sup>UZH</sup>**

# 1 Course Outline

## Day 1

1. Introduction to Statistical Analysis
  - Example case study in SPSS
  - Objective of statistical analyses
2. Exploratory Data Analysis
  - Why data exploration is essential in statistical inference
  - Informal descriptions of data - numerical and visual summaries
3. Introduction to Hypothesis Testing
  - Translating scientific questions into statistical hypotheses
  - Formulating and testing one sample hypotheses
  - Interpreting the output of a hypothesis test - p-values and errors
4. Choosing an Appropriate Statistical Test
  - One sample case study using t-test and Wilcoxon signed rank test
  - Parametric or Non-parametric tests
5. Comparing Two Populations
  - Extending one population hypotheses to two populations
  - Two sample t-test and Wilcoxon rank sum test
  - Paired t-test and Wilcoxon signed rank test

## Day 1 - Exercises

- Exercise 1. Using SPSS
- Exercise 2. Importing data into SPSS
- Exercise 3. Numerical and visual data summaries
- Exercise 4. One sample t-test
- Exercise 5. One sample t-test, Wilcoxon signed rank test
- Exercise 6. Which test and why?
- Exercise 7. Two sample t-test or Wilcoxon rank sum test
- Exercise 8. Paired t-test and Wilcoxon signed rank test

## Day 2

### 6. Comparing Multiple Populations - One Way ANOVA

- Validity, parametric and non-parametric
- Case Study - weight gain by diet treatment

### 7. Hypothesis Testing - Linear Modelling

- Extending one way ANOVA
- Adjusting for covariates and nuisance variables
- Case Study - vitamin C and teeth length

### 8. Grouped and Repeated Observations - Introduction

- Adjusting for within-subject variability
- Different approaches in SPSS
- Case Study - longitudinal observations

### 9. Survival Analyses - a Brief Overview

- Kaplan-Meier
- Cox Proportional Hazards
- Case Study - Leukemia survival times

### 10. Other topics - a Brief Mention

## Day 2 - Exercises

- Exercise 9. One Way ANOVA - chicken weights
- Exercise 10. One Way ANOVA - patient data
- Exercise 11. One Way ANOVA - treatment effects
- Exercise 12. Linear Modelling - fixed and covariate effects, tooth data
- Exercise 13. Linear Modelling - fixed and covariate effects, puromycin data
- Exercise 14. Repeated Measures - rat data
- Exercise 15. Repeated Measures - patient data
- Exercise 16. Survival Analyses - Kaplan-Meier + Cox Proportional Hazards

# Introduction to Applied Statistics for Medical Researchers

## Part 1 - Introduction to Statistical Analysis



University of  
Zurich <sup>UZH</sup>

## What is Statistics?

### A first example in SPSS

An experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens. Newly hatched chicks were randomly allocated into six groups, and each group was given a different feed supplement. Their weights in grams after six weeks are given along with feed types. 71 observations on 2 variables.

Research question - do different feeds give rise to different growth rates? E.g. if the experiment was repeated with another sample of similar chickens would we observe qualitatively similar results?

< Demonstration → SPSS >

## What is Statistics?

### One definition

A set of analytical tools designed to quantify uncertainty

- If an experiment or procedure is repeated, how likely is it that the new results will be similar to those already observed?
- What is the likely variation in results if the experiment was repeated?

## What is Statistics?

< Demonstration → SPSS >

- Statistical analyses allows us to **infer** what we are likely to observe on data **we have not yet seen** generated from the same population as the current sample
- SPSS - like an Excel for Statistics, an easy to use point and click interface, allows different types of information to be stored separately for re-use and to allow a structured record of the analyses performed
- .sav - data files; .spv - output/viewer files; .sps stored syntax files.

## Exercise 1. Using SPSS.

- ① Can you re-run the analyses just presented?
- ② Can you re-run the analyses just presented using the Graphical Menus?

## What is Statistics?

## Lies, damn lies and statistics

‘‘An old jest runs to the effect that there are three degrees of comparison among liars. There are liars, there are outrageous liars, and there are scientific experts. This has lately been adapted to throw dirt upon statistics. There are three degrees of comparison, it is said, in lying. There are lies, there are outrageous lies, and there are statistics’’

Robert Giffen in 1892

many similar variations exist circa 1880-1900 - see wikipedia

## Putting Statistics in Context

... and with modern computer software such “lies” are easier to produce than ever before ☺

- Remember - computers are very stupid and only do exactly as they are told, even if this makes no sense

## Putting Statistics in Context

... and with modern computer software such “lies” are easier to produce than ever before ☺

- Remember - computers are very stupid and only do exactly as they are told, even if this makes no sense
- If you ask the computer to do a calculation in most cases it will produce a **numerical answer**. It is up to you to decide whether the analysis is i) appropriate for your data and ii) as optimal as is reasonably possible

## Putting Statistics in Context

- The vast majority of analyses can be done in a straightforward fashion - just remember and always use common sense as a guide - be skeptical!

## Putting Statistics in Context

- The vast majority of analyses can be done in a straightforward fashion - just remember and always use common sense as a guide - be skeptical!
- It is very easy to get “lost” in the statistical software and technical jargon, which differs markedly between different software packages. Terminology can also differ greatly between textbooks...
- Wikipedia is as good a resource as any for finding out about different statistical tests and terminology

## Putting Statistics in Context

- SPSS and other packages can produce copious amounts of output, including many highly specific - and sometimes rather obscure (wiki!) - statistical tests

## Putting Statistics in Context

- SPSS and other packages can produce copious amounts of output, including many highly specific - and sometimes rather obscure (wiki!) - statistical tests
- Do not assume that just because the software provides values as part of the output that this is applicable to your analyses

## Putting Statistics in Context

- SPSS and other packages can produce copious amounts of output, including many highly specific - and sometimes rather obscure (wiki!) - statistical tests
- Do not assume that just because the software provides values as part of the output that this is applicable to your analyses
- If in doubt - keep it small and simple (KISS)

## Putting Statistics in Context

### There is always more than one way to do any analyses

This course will cover many of the basic types of analyses which are useful and widely used in experimental and medical settings - there are many many others which we will not cover but the general concepts are broadly similar.

## Study Design and Planning - Statistical Aspects

- It is crucially important to determine how data from an experiment/study will be statistically analysed **BEFORE** generating/collecting any data!
- Firstly, this ensures that the experiment has a clear objective
- Secondly, data collection is expensive and this ensures that all the appropriate data/measurements are taken and that the study design is sound
- Even the most sophisticated analyses cannot rescue a poorly designed study

## End of Section Milestones

After this short section you should have

- a general idea as to the purpose of statistical analysis in research
- be able to open up SPSS, open up a data set, run analyses provided, and save the output

## Introduction to Applied Statistics for Medical Researchers

### Part 2 - Exploratory Data Analysis



**University of  
Zurich** <sup>UZH</sup>

## Exploratory Data Analysis

- It is crucially important to explore your data fully before considering any “formal” statistical analyses
- What explorations are done depends on the objective of the study - the research question(s)

## Exploratory Data Analysis

- It is crucially important to explore your data fully before considering any “formal” statistical analyses

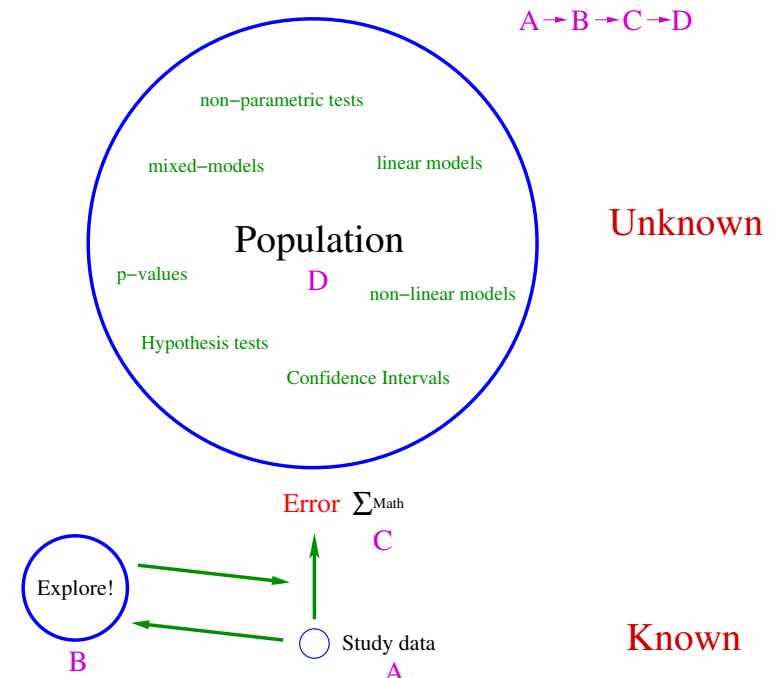
## Exploratory Data Analysis

- It is crucially important to explore your data fully before considering any “formal” statistical analyses
- What explorations are done depends on the objective of the study - the research question(s)
- Helps to decide what kind of formal statistical analyses might be most appropriate for the data available



# Exploratory Data Analysis

- It is crucially important to explore your data fully before considering any “formal” statistical analyses
- What explorations are done depends on the objective of the study - the research question(s)
- Helps to decide what kind of formal statistical analyses might be most appropriate for the data available
- What a simple descriptive analysis **does not** provide is evidence of whether the observed treatment effects are large enough to be notable once sampling variation has been accounted - that is the role of formal analyses, e.g. hypothesis testing



## Numerical Data Summaries

- Mean - a measure of location. Always examine the average value of the response variable(s) for the different “treatment” effects in your data
- Median - a robust single value summary of a set of data (50% quantile point) - most useful in highly skewed data or data with outliers
- Standard deviation - a measure of spread, how variable the data are
- Correlation coefficient - a measure association between continuous variables (common but somewhat limited)

Exercise 2. Importing data into SPSS.

File> Open > data..

Data Editor screen, variable view tab

## Visual Data Summaries

A picture is worth a thousand words....

- Histograms, Q-Q plots, P-P plots - describes the distribution of data
- Boxplots - allows the distribution of different variables to be easily compared
- Scatter plots - used to examine relationships between variables

## Exploratory Data Summaries - References

Again wikipedia has detailed pages on most of the terms mentioned along with citations and examples.

Textbook: *A Primer in Data Reduction: An Introductory Statistics Textbook* by A. S. C. Ehrenberg is also excellent - highly accessible and aimed at students who have to use statistics as part of there other studies. This text also covers most of the topics we will be discussing later. There are very many other introductory textbooks covering similar topics.

## Exploratory Data Summaries

As we will see later, many statistical analyses rely on the **data having particular properties** - such as variation within different groups being similar. It is also particularly common to need to check that data conform to a particular type of probability distribution, specifically what is known as the Normal (or Gaussian) distribution (details later). One way to visually check whether data are consistent with a particular probability distribution is to use quantile-quantile (Q-Q) plots or probability (P-P) plots.

## Exercise 3. Exploring Data.

```
chickwts_balanced.sav
Tooth.sav
data_A.sav
data_B.sav
See questions in handout.
Analyze> Descriptive Statistics>..
Graphs> Legacy Dialogues>..
```

## Exercise 3. Solutions/Comments

- Not all the output from SPSS will be useful or informative for all data sets
- With small numbers of observations it is very difficult to assess normality (or otherwise)
- It is important to note the observed sizes of treatment effects
  - are they large enough to be clinically relevant?
- An exploratory analysis is important as unlike formal statistical analyses this is not subject to any assumptions or estimations or approximations. **It is the truth** - and while it has its limitations - it is worth referring back to for this reason.

## End of Section 2 Milestones

After this section you should be able to

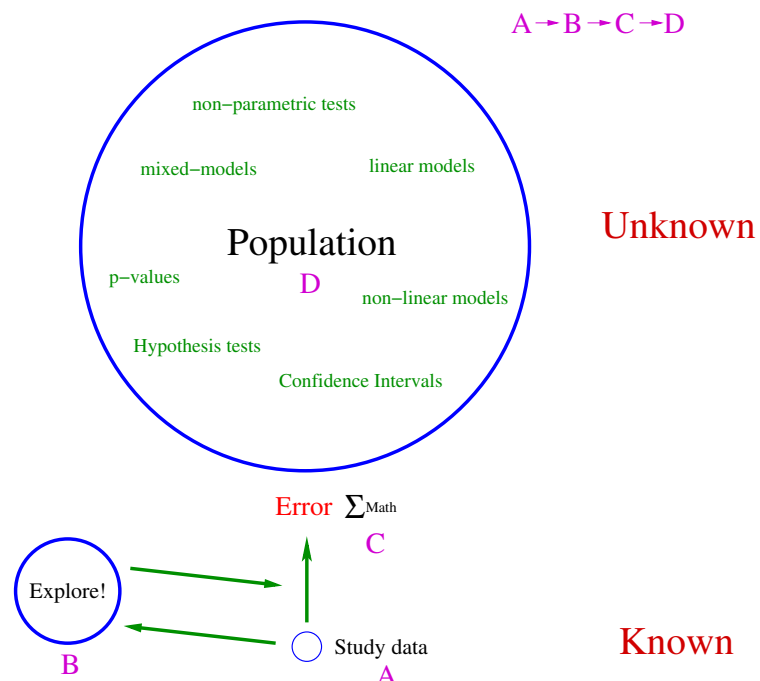
- load data into SPSS and set the appropriate properties
- produce simple numerical summaries which describe the main treatment effects of interest in study data
- produce simple numerical summaries which describe aspects such as the variation within different treatment types
- produce graphical summaries to identify features of the data such its approximate distribution and relationships between variables

# Introduction to Applied Statistics for Medical Researchers

## Part 3 - Introduction to Hypothesis Testing



University of  
Zurich<sup>UZH</sup>



# Introduction to Hypothesis Testing

Study data is collected for a purpose - to answer one or more specific scientific questions. The classical way to perform a formal statistical analyses of these data is to formulate these research questions into statistical **hypothesis tests**.

In this section we will go through a simple example in detail to highlight some of the important concepts - the general approach for more complex analyses is exactly same. *Note: the precise technical details are much less important than the concepts!*

## Simple Example - One Population

**After six weeks will the mean weight of a chicken be more than 250 grams?**

There are 71 observations [chicken weight data set] from which to answer this question. This can be formulated into a statistical hypothesis test. A hypothesis test has two parts, the null hypothesis and the alternative hypothesis. This is typically written as follows:

$$H_0 : \mu \leq 250$$

$$H_A : \mu > 250$$

where  $\mu$  is the mean weight in the **population** of chickens from which the sample of 71 chickens was drawn. Remember - we know the mean weight in the sample of chickens is greater than 250 it is the **population** of chickens which we are interested in.

## Simple Example - One Population

After six weeks will the mean weight of a chicken be at least 250 grams?

$$H_0 : \mu \leq 250$$

$$H_A : \mu > 250$$

The null hypothesis ( $H_0$ ) is the default situation, sometimes called the “state of nature”. In a treatment-control trial,  $H_0$  is typically that the effect of the treatment is no different from the control. In this example our default position is that the mean weight of chickens is  $\leq 250$ . This is called a single-sided hypothesis test.

## Simple Example - One Population

We now analyse the 71 observations to see whether there is evidence to **REJECT** the null hypothesis  $H_0$ , and if the null hypothesis is rejected then we can conclude that the available evidence supports the alternative hypothesis.

Note that hypothesis testing is concerned with finding evidence in support of the null hypothesis  $H_0$  - the default situation - rather than evidence in favour of the alternative hypothesis.

## One Sample t-test

For the chicken weights data an appropriate formal analyses is to use a **one-sample t-test**, why this test is appropriate will be discussed later. This analysis involves calculating a simple summary statistic - called a  $t$ -statistic - which we do entirely from the observed data.

$$T_{obs} = \frac{\bar{x} - \mu}{s/\sqrt{n}}, \quad (1)$$

where  $\bar{x}$  is the sample mean,  $s$  the sample standard deviation and  $\mu$  is the population mean in the null hypothesis which we wish to test for. We then look up the value of  $T_{obs}$  in a set of statistical tables/computer to see what the answer is to our research question.

## Simple Example - One Population

## Important concept - sampling

Why is  $T_{obs} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$  called a  $t$ -statistic?

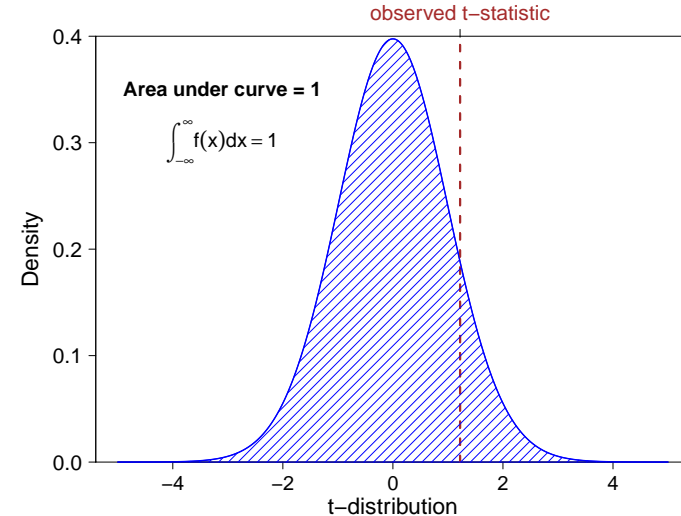
If another sample of 71 chickens from the same population were weighed then the values for  $\bar{x}$  and  $s$  would be different, and hence the value for  $T_{obs}$ . If this was repeated many times and a histogram/Q-Q/P-P plot produced of the values for  $T_{obs}$  then this would follow the shape of a known distribution - **student-t probability distribution**. It is this piece of mathematics - knowing what the sampling distribution of  $T_{obs}$  is - which allows us to infer information about the population of chickens from which our original 71 chickens were sampled - without actually having to collect lots and lots of other samples of chickens! Mathematical theory is used to fill this data gap.

## Chicken weights t-test

$$T_{obs} = \frac{261.31 - 250}{78.07/\sqrt{71}} = 1.22 \quad (2)$$

Put the values for the sample mean and standard deviation into the t-statistic formula along with the  $\mu = 250$ . We now look up the value of this in a t-distribution reference table. All this calculation will be done for you in SPSS but it is important to understand the general process as this is the same for hypothesis testing in other more complex analyses.

## One Sample, one-sided, t-test



## Important concept - p-values

- The result of a hypothesis test is usually communicated in the form of a **p-value**
- The interpretation of a p-value is of crucial importance - it is the *probability that the test statistic takes values at least as extreme as that observed assuming that  $H_0$  is true*
- Exactly what **at least as extreme as** refers to depends on the alternative hypothesis  $H_A$ .
- This may sound rather abstract but it is usually obvious in practice

## Simple Example - One Population

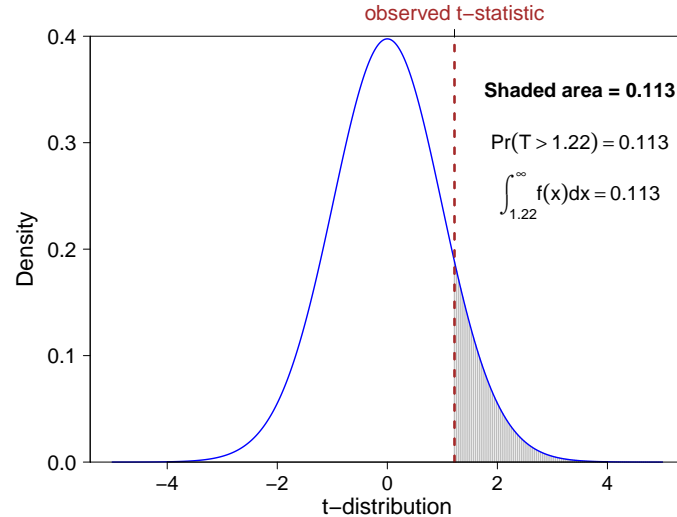
**After six weeks will the mean weight of a chicken be at least 250 grams?**

$$H_0 : \mu \leq 250$$

$$H_A : \mu > 250$$

The alternative hypothesis is  $\mu > 250$  so in this test **at least as extreme as** in the definition of a p-value is the probability of observing a t-statistic which is  $> 1.22$  **assuming that  $H_0$  is true** - this is why 250 was used for  $\mu$  when calculating  $T_{obs}$ .

# One Sample, one-sided, t-test



## Research Question - be pragmatic with p-values

By convention a p-value of less than 0.05 is considered to provide reasonable evidence for rejecting  $H_0$ . A p-value of between 0.05 and 0.1 might be considered as weak evidence against  $H_0$ . Values of less than 0.01 are generally considered as very strong evidence for rejecting  $H_0$ . It is **always** best to provide a p-value in any analyses to let the reviewer/client see the strength of evidence rather than simply claiming statistically significant findings!

## Research Question

The purpose of this hypothesis test analysis is to answer a very specific scientific question:

**After six weeks will the mean weight of a chicken be more than 250 grams?**

So what is our answer?

The p-value for this hypothesis test is 0.113. Based on this value we can either **reject**  $H_0$  and conclude that the mean weight of chickens in the population is likely to be greater than 250 grams or else we can **accept**  $H_0$  where the mean chicken weight is less than 250 grams.

## Communicating Results of Hypothesis Tests

Transparency is essential - the devil can be in the detail - which at the very least should comprise:

- i. what hypothesis was being tested - be clear and precise
- ii. what statistical test was used
- iii. what the p-value is
- iv. what the treatment effect is (more later).

This is particularly crucial if the analyses are to be given to someone *else* to then make a judgment on the scientific significance.

This is the main theory block in the course - we're about half way through - some exercises soon ☺

## Questions?

Later we will look at many different hypothesis tests with multiple populations and other more complex data - these analyses will all follow a similar general approach to that used in the simple chicken weights example.

## Two-sided Tests

A two-sided test is similar to a one-sided test - the key difference is in what is now defined as **at least as extreme** in the definition of the p-value. This time the alternative hypothesis refers to observing a value of **either**  $\bar{x} > 250$  or  $\bar{x} < -250$  **assuming that  $H_0$  is true**, which using the t-test approach is equivalent to the probability of observing  $T_{obs} > 1.22$  or  $T_{obs} < -1.22$  which we can again look up in reference tables.

## Two-sided Tests

### One Population

**After six weeks will the mean weight of a chicken be equal to 250 grams?**

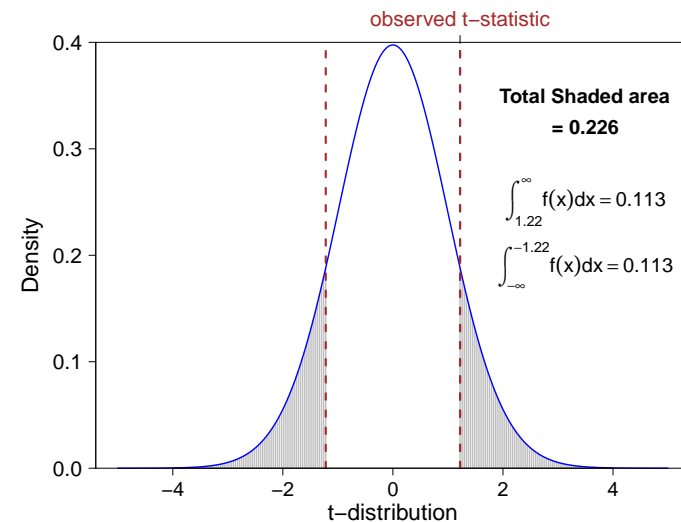
This is now a two sided hypothesis test:

$$H_0 : \mu = 250$$

$$H_A : \mu \neq 250$$

This time the hypothesis test is asking how much evidence is there in our sample data to conclude that in the population of all chickens the mean weight is not equal to 250 grams.

## One Sample, two-sided, t-test





## Two-sided Tests

- The two-sided t-test has a p-value which is exactly double the single sided test
  - Think! - intuitively the p-value should be less for a single sided test as the research question you are asking is much narrower e.g. greater than 250 grams, as opposed to whether the mean chicken weight might be **either** less than 250 grams **or greater** than 250 grams.
- You are using the same amount of information [71 observations] to answer a narrower research question and so all else being equal you should expect a “more powerful” analyses (e.g. a lower p-value all else being equal)

## Exercise 4. One Population - Two sided t-test

Analyse > Compare Means > One sample t-test

## Exercise 4 - comments

This exercise used simulated data so we know what the true values for  $\mu_{x1}$ ,  $\mu_{x2}$ ,  $\mu_{x3}$  and  $\mu_{x1}$  are.

- Did all the hypothesis tests give the correct answers? No!
- We failed to reject some of the  $H_0$ 's even although they were false
- What affected the error? Variation.

With simulated data we know the true answer - with real data we will not

## Hypothesis Testing Errors

- In a hypothesis test we either reject or accept the  $H_0$
- This decision is based on a random sample of data
- Therefore there is a chance that we draw the wrong conclusion!

## Hypothesis Testing Errors

There are two errors associated with a hypothesis test

- ① incorrectly rejecting  $H_0$  even although it is true - **Type I** error, and is usually denoted by  $\alpha$
- ② not rejecting  $H_0$  even although it is false **Type II** error, and is usually denoted by  $\beta$

## Hypothesis Testing Errors

- **Type I** error:  $\alpha = \text{p-value}$  - think about it!
- **Type II** error:  $\beta$  - an error of  $\approx 0.2$  generally considered acceptable
- These errors are NOT independent e.g. we could decide to never reject  $H_0$  and so  $\alpha = 0$  but then  $\beta$  would be very large
- In almost all statistical analyses it is typically  $\alpha$  - the p-value - which is used to determine the answer to the research question

## Hypothesis Testing v Estimation

### What is the likely value of the true population mean?

This information is included in the SPSS output (Exercise 4.) - called a **confidence interval**

Given a sample of data from some population, the best estimate we have of the population mean is the observed sample mean, e.g. we use  $\bar{x}$  as our best guess at  $\mu$ . Using this information and a little math it is possible to estimate a **confidence interval** - this provides an estimate as to the likely value of  $\mu$  in terms of an upper and lower bound, e.g. the observed sample mean  $\pm$  some error, where the error is due to us only having a sample of data from the population so we do not have complete information about  $\mu$ .

## Estimation - Confidence Interval Interpretation

Given a sample of data it is usual to calculate a **95% confidence interval** - we will ignore how this is calculated, SPSS will do it for you - which has the following interpretation: suppose another sample of data were taken from the same population then  $\bar{x}$  and  $s$  would be slightly different and hence the estimated 95% confidence interval for  $\mu$  would also be slightly different. Imagine we were able to keep collecting samples from the same population and each time calculated a 95% confidence interval. On average 95% of these intervals will contain the true value of  $\mu$  - that is why they are called 95% confidence intervals

## Estimation - Confidence Interval Interpretation

### Width of Intervals

A 95% confidence interval will be wider than a 90% confidence interval but narrower than a 99% confidence interval

Why? - think - common sense!

⇒ Take an extreme example - we can be 100% confident that every person in the world weighs between zero and 10000 kilograms, e.g. a 100% confidence interval is therefore (0, 10000) kilograms, however, if we restrict this interval to (10, 80) kilograms then we can no longer be 100% confident that this interval will include all possible people. The narrower the interval the less coverage it has - by convention 95% is considered a reasonable trade off between width and coverage.

## Summary

### Key Points from this section

- The purpose of a hypothesis test analysis is to answer a very specific scientific question
- Defining the hypothesis is the first step
- Equally important is to understand exactly what hypothesis is being performed e.g. by SPSS
- A hypothesis test calculates a test statistic based on the observed data e.g. a t-statistic
- It is then determined how extreme or otherwise this value is using statistical tables/software
- A p-value is typically used to summarise the strength of evidence for either rejecting or accepting the null hypothesis - but note that this says nothing of the actual magnitude of the treatment effect

## Hypothesis Testing v Estimation

### Statistical significance **does not** equate to scientific significance

All p-values should be treated with some caution, they are almost always only approximate - sometimes very approximate - and they are commonly abused. More importantly, a p-value can be very small indicating strong statistical significance, but the actual effect of a treatment, e.g. as given by a confidence interval, might be so small as to be of little or no scientific or clinical relevance.

## End of Section Milestones

After this section you should be able to

- translate a scientific question into a precise hypothesis test in a one population situation
- be able to perform one-sample t-tests in SPSS and meaningfully interpret the output

## More Advanced Closely Related Topics

### Other types of hypotheses relevant for medical research

- Non-inferiority hypotheses
- Bioequivalent hypotheses

These types of hypotheses examine whether, for example, a new treatment is **No Worse** than an existing treatment. This is not the same hypothesis testing framework presented, one difference is that it is usually necessary to define some range over which the treatments being compared are allowed to differ. E.g. in a surgical procedure perhaps a difference in mobility of a joint of 5 degrees may be acceptable between different methods.

## Introduction to Applied Statistics for Medical Researchers

### Part 4 - Choosing an Appropriate Statistical Test



University of  
Zurich<sup>UZH</sup>

## Wilcoxon signed rank test

To use the Wilcoxon signed rank test a similar process is used as to the one-sample t-test

- ① Calculate the value of an appropriate test statistic
- ② Examine the alternative hypothesis to see what **as extreme as means**
- ③ Calculate the p-value for the test from appropriate statistical tables/software

The technical details differ - but the process is identical to applying a t-test, or any other hypothesis test. As we have seen the computer will do all of this calculation for you.

## Hypothesis Testing - One Population Alternatives

We have studied the one-sample t-test in some detail. An alternative which is often used in a one-population scenario is the **Wilcoxon signed rank test**. Considering again the chicken weights data, we would use this test with the following - slightly different research question

**After six weeks will the median weight of a chicken be at least 250 grams?**

$$H_0 : \mu_{0.5} = 250$$

$$H_A : \mu_{0.5} \neq 250$$

using  $\mu_{0.5}$  to denote the median weight in the population of chickens from which our 71 observations were drawn.

## Wilcoxon signed rank test

**After six weeks will the median weight of a chicken be at least 250 grams?**

$$H_0 : \mu_{0.5} = 250$$

$$H_A : \mu_{0.5} \neq 250$$

using SPSS to perform the Wilcoxon signed rank test we get a p-value = 0.282. Recall that the t-test had a p-value of 0.226 - and tested a slightly different hypothesis as it considered the mean rather than median.

Exercise 5. One-sample t-test and Wilcoxon signed rank test

```
Analyse> Compare Means > One sample t-test
Analyse> Non parametric> One sample...
```

Do the results differ?

### Choosing an appropriate Hypothesis Test

This is arguably the most important part of any applied statistical analyses. Earlier we examined exploratory analyses which included examining the distribution of the observed data, and in particular assessing whether it was approximately **Normally distributed**.

A great deal of statistical analyses uses this assumption and we will return to this again and again. This is also directly relevant to deciding whether a **one sample t-test** or the **Wilcoxon signed rank test** is more appropriate for a given data set.

### Choosing an appropriate Hypothesis Test

- For a specific hypothesis test to be statistically valid for a given data set certain criteria must be met - at least approximately
- **Every** hypothesis test has some criteria/conditions which must be met
- For a **one sample t-test** to be valid a necessary condition is that the observed data should be approximately **Normally distributed** - if they are not then this test may be unreliable
- The **Wilcoxon signed rank test** does **not** require that the data are **Normally distributed** and so it is generally used as an alternative to the one-sample t-test when the data are not Normally distributed

### As small aside - why is Normality important?

The key calculation in performing a hypothesis test is to estimate the probability of observing a value as extreme as the observed test statistic. This applies to any hypothesis test - the actual test statistic differs from test to test - but this calculation is required in each case. So how it is possible to estimate the **sampling distribution** of a test statistic? E.g. how can we know what

$$\frac{\bar{x} - \mu}{s/\sqrt{n}}$$

has a t-distribution? This mathematical result is only possible by making certain assumptions about the observed data, **x**, for example that **x** comes from a Normal distribution. The criteria relating to a particular hypothesis test are there to ensure that the particular test statistic follows a known distribution - if some of these assumptions are not met then the test statistic for the study data may not then follow a known distribution and the test will be unreliable.

### Parametric or Non-parametric

The t-test is referred to as a **parametric test**, whereas the Wilcoxon signed rank test is referred to as a **non-parametric test**.

The usual defining difference between a parametric test and non-parametric test, is that the former requires that the observed data follow a particular known probability distribution e.g. to use a t-test the data must be approximately **Normally distributed**.

A non-parametric test does not require that the data conform to a particular known probability distribution.

⇒ Does this imply that non-parametric tests are to be preferred since they require less assumptions?

### Parametric or Non-parametric

- It is a **very common misconception** that non-parametric tests can be applied to any data - this is erroneous!
- It is true that for small datasets where it is impossible to determine Normality or not, then non-parametric tests are generally more appropriate
- However, non-parametric tests **do** have conditions of their own
- For example, while the Wilcoxon signed rank test does not require that the observed data are approximately Normally distributed, technically it does assume that the observed data are from a symmetrical distribution (a big assumption!)

## Common sense - Parametric or Non-parametric

- In practice when doing analyses with small sample sizes, it is very difficult, if not impossible, to check for the shape of the distribution e.g. Normality or other properties such as symmetry
- It is therefore usual and generally acceptable to use non-parametric tests for small data sets - there are not many other alternatives!
- It is worth remembering, however, that this is a pragmatic solution and that the p-values produced by non-parametric tests may not always be entirely reliable

Exercise 6. Which test to use?

For each of the data sets - see appendix - decide which test is most appropriate and why?

Variable Y4 has a very non-normal distribution - can we fix this?

## Normality (again)

- If the data are not Normally distributed then it may be possible to transform them to make this assumption more reasonable - a standard approach is to take logs of the data (any number base will do - but usually the natural logarithm is used)
- Generally speaking, if the data (single population) comprise more than 20-30 observations then Normality will be a reasonable assumption - even if the data look rather non-normal - central limit theorem
- If there are a very small number of observations e.g.  $< 10$  then it may be more conservative to choose a non-parametric approach
- Previous experience with similar kinds of data is also relevant

## End of Section Milestones

After this section you should be able to

- Decide whether a one-sample t-test or Wilcoxon signed rank test may be more appropriate for a given data set
- Provide justification as to why one approach may be preferable over the other
- Recognise the role of assumptions in hypothesis testing and the presence of subjectivity

## Small sample sizes - Non-parametric

### A final word of caution about small sample sizes

While it is generally true that non-parametric hypothesis tests are the most appropriate formal analyses for very small sample sizes, and the computer may be able to provide a p-value, even an exact p-value, there are other other implicit scientific assumptions being made in such analyses which require some careful consideration:

- It is implicitly assumed that observed data are a random sample from the population of interest
- What is the precise population to which the hypothesis test refers?
- Is your “small” sample of data really representative of this population? Bias?
- Does it include the full range of variation in “treatment” response which you might expect? Stratification by covariates?



## Introduction to Applied Statistics for Medical Researchers

### Part 5 - Comparing Two Populations



University of  
Zurich<sup>UZH</sup>

#### Exercise 7.

For the Tooth length data set can you formulate an appropriate null and alternative hypothesis for the scientific question?

Can you do a parametric and non-parametric analyses for the appropriate hypotheses in SPSS?

**hint:** the two tests we used for the one-population analyses have closely related versions for comparing two populations. Note also that a Mann Whitney test is also called a **Wilcoxon** rank sum test.

## Hypothesis Testing - Comparing Two Populations

### Tooth Length Case Study

The tooth length dataset comprises of two independent populations. The scientific question of interest is whether animals given vitamin C via orange juice have different teeth lengths from animals given vitamin C via water solution. This analysis involves comparing **two independent populations**.

**Analysing this two population scenario is very similar to the one-population examples we have already discussed.**

### Tooth Length Case Study - i) Two sample t-test

The scientific question of interest is whether animals given vitamin C via orange juice have different teeth lengths from animals given vitamin C via water solution.

**Is the mean tooth length in the “orange juice” population equal to the mean in the “water solution” population?**

$$H_0 : \mu_{OJ} = \mu_{O_2}$$

$$H_A : \mu_{OJ} \neq \mu_{O_2}$$

using  $\mu_{OJ}$  to denote the mean tooth length in the population of guinea pigs which are given vitamin C via orange juice and  $\mu_{O_2}$  similarly but for vitamin C given via water.

## Tooth Length Case Study - ii) Wilcoxon rank sum test

Do the teeth lengths in the orange juice population have the same distribution as those in the water solution population?

It is possible to write this out formally as an  $H_0$  and  $H_A$  hypothesis but it is rather more technical than the two-sample t-test since this test is comparing not just **single statistics**, e.g. means or medians, but **distributions**. The alternative hypothesis ( $H_A$ ) is that at least one distribution differs by a *location shift*.

The key point is straightforward - this is a useful test for comparing two populations, but it is worth noting that this test is more general than simply comparing medians (there is also a test for equality of medians in SPSS - try it!).

It is **always** a good idea to find exactly what hypothesis the test you are using is actually doing.

## Two Populations - Parametric v Non-parametric

Reasons for preferring either the **two-sample t-test** or the **Wilcoxon rank sum test** are identical to those in one-population case, e.g. are the data approximately Normal. In addition:

- With two populations we have a decision about assuming either constant or differing variances in each of the two populations - see SPSS output
- The Wilcoxon rank sum is the default choice for comparing two populations of ordinal - not interval scaled - data

## Two Populations - Paired Observations

In the two population scenario **paired data** are common. For example this could be the result of *before* and *after* observations on the same subjects.

- It is possible to analyse paired data by treating it as two independent populations as in the previous examples, and use either the two-sample t-test or else Wilcoxon rank sum test
- However, this is a **very poor method of analysis** as this does not allow for control of within patient variability
- A much more powerful analysis is to work with a new dataset comprising the **differences** for each patient, e.g. the difference in blood pressure before and after treatment

## Exercise 8.

For the Hamilton depression data - does giving the patients a tranquilizer influence their depression score? Parametric or Non-parametric? Are the results different?

## End of Section Milestones

After this section you should be able to

- Decide whether a **two-sample t-test** or **Wilcoxon rank sum test** is more appropriate for a given two population data set
- Decide whether to use a paired or independent samples analysis

## Introduction to Applied Statistics for Medical Researchers

### Part 6 - Comparing Multiple Populations - One Way ANOVA



University of  
Zurich <sup>UZH</sup>

## Hypothesis Testing - One Way ANOVA

We have seen how to perform hypothesis tests when comparing two populations using the two sample t-test and Wilcoxon rank sum tests. In many analyses we may have multiple populations - for example suppose we have a treatment which has a number of different levels high/medium/low/placebo, or equivalently a number of different treatments. What then is the hypothesis we wish to test?

**Is there a difference in the effect of the treatment?**

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

$$H_A : \text{at least one pair of } \mu_1, \dots, \mu_k \text{ are different}$$

where  $\mu_1, \dots, \mu_k$  denote the mean effect of treatment levels 1 through  $k$ .

### Chicken Weights - Differences between Multiple Treatments

The scientific question of interest in the chicken weight data [60 observations] is

**Is there a difference in the growth rate of chickens fed on either: casein, horsebean, linseed, meatmeal, soybean, sunflower?**

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_6$$

$$H_A : \text{at least one pair of } \mu_1, \dots, \mu_6 \text{ are different}$$

using  $\mu_1$  to denote the mean weight of chickens fed on casein,  $\mu_2$  to denote the mean weight of chickens fed on horsebean etc. This is similar to comparing two populations but a slightly difference approach (and test) is needed.

### Chicken Weights - Differences between Multiple Treatments

Exercise 9 - we will work through this together since there are number of different ways to do this in SPSS and many different options.

- Load the dataset `chickwts_balanced.sav`
- First remind ourselves of the what the data looks like - means, boxplots, P-P plots etc
- Then `Analyze...> One Way ANOVA` with only minimal outputs

SPSS provides the output as an **analysis of variance table**. The hypothesis test being used here is called an **F-test**, it can be used for many kinds of analyses. In the output we have  $F=13.401$  which gives a p-value  $< 0.001$ .

### Chicken Weights - Differences between Multiple Treatments

Exercise 9 continued - is an F-test valid for this data?

- an F-test needs Normally distributed data
- and it requires that the variance within each group is approximately constant
- Next lets turn on some more diagnostics  
Analyze...> One Way ANOVA with homogeneity of variance test

### Chicken Weights - Differences between Multiple Treatments

Exercise 9 continued - A slightly different way

- Analyze...> General linear Model...> Univariate
- Save unstandardized residuals
- Save Cooks distance

The usual way to check for the validity of an ANOVA is to plot the residuals - these should be randomly scattered and Normally distributed. It is also important to check for the sensitivity of results to individual observations, so-called leverage points and the Cooks distance is one way to assess this

Exercise 10. i) considering data\_A.sav (ignoring, for the moment, the patient variable)

- 1 Is there a statistically significant effect of treatment?
- 2 Is the F-test valid here?
- 3 Can you provide an annotated SPSS viewer file which supports your analyses?

ii) now consider data\_B.sav (ignoring the specimen variable)

- 4 Is there a statistically significant effect of Group with either i) Reponse1? or ii) Response2?
- 5 Is the F-test valid here? If not can you make it valid? hint: transformation or perhaps dropping a group?
- 6 Can you provide an annotated SPSS viewer file which supports your analyses?

### Comments on Exercise 10

There is quite a lot to think about - and it is to some degree subjective

- Check for constant variance
- Check that the residuals are random and normal
- No high leverage points
- Removing observations is generally a very dubious practice - change the analyses not the data!
- If the removal of some data is unavoidable then it is better to remove a whole treatment group than discarding individual observations simply because they do not fit the particular analytical method chosen

## Patient dataset B - Non-parametric

From Exercise 10 it appears that an F-test is not appropriate for the examining the hypothesis that there is a difference in the mean value of Response 2 between the six populations for which we have sample data.

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_6$$

$$H_A : \text{at least one pair of } \mu_1, \dots, \mu_6 \text{ are different}$$

where  $\mu_1, \dots, \mu_6$  denote the mean value of Response 2 with each of the six different treatment groups.

As with the previous one population and two population cases there is a non-parametric test we can use as an alternative. The **Kruskal - Wallis test** is an extension of the **Wilcoxon rank sum** to three or more populations.

## Patient dataset B - Kruskal - Wallis test

This examines the hypothesis that the distribution is the same for Response 2 across each of the six treatment groups in the study, with the alternative hypothesis ( $H_A$ ) that at least one distribution differs by a location shift.

Does not require that the data are normally distributed.

## Kruskal - Wallis Test

As with the other non-parametric tests we have examined, while normality is not required, the Kruskal - Wallis Test has its own conditions:

- **to interpret this test as a difference in location the distribution of values within each group should be identically shaped**

Does this assumption appear reasonable for Patient dataset B?

Perhaps not - look at SPSS output such as boxplots - they look rather different. Also even after log transforming the data the equality of variance test still suggests that the variance within each group is different, this was also still true even after removing Group 2. Although this variance test assumes Normality so it is perhaps not entirely reliable.

## Patient dataset B: Kruskal - Wallis Test

- This is just real life!
- It generates data which sometimes does not fit nicely into some pre-defined category
- You simply have to make the best you can with the tools available 😊

⇒ There are only four observations available per treatment group so it is always going to be very difficult to tell whether the assumption about similar distributional shapes is reasonable. For this data set/research question then the Kruskal - Wallis does look more reasonable than the F-test. I would also take a careful look at specimen 2 (in group 2). If this is a true observation then perhaps more observations are needed?

## Post Hoc Analyses and Estimating Effects

### Patient dataset B - Summary

Using SPSS we get a p-value of 0.014 using the Kruskal - Wallis test which is strong evidence that the distribution of Response 2 does differ across treatment groups. We can use this non-parametric approach or else either drop treatment group 2 from the analyses and use the F-test analyses (which obviously uses a slightly different hypothesis) which has a p-value of  $\ll 0.001$ . Which is more appropriate is less of a statistical question, and more about whether the person in-charge of the study is happy to have stronger evidence but for a reduced hypothesis.

So far, we have only compared populations to identify evidence of differences existing between them. As mentioned earlier, just as important is estimating the size of differences, as if these are very small then the p-values/statistical significance may be irrelevant.

Consider again the Chicken weights data [Exercise 1] i) what are the size of treatment effects; ii) and which effects are different?

Such analyses are generally called post-hoc analyses - we explore the differences in treatment effects which resulted in the hypothesis being rejected

Bonferroni Correction (wikipedia) - making multiple comparisons has a statistical cost - Type I error increases

## Comments on Exercise 11

Exercise 11. i) considering data\_A.sav (ignoring, for the moment, the patient variable)

- 1 what are the sizes of treatment effects and which treatments are statistically different?
- ii) now consider data\_B.sav (ignoring the specimen variable)
  - 2 what are the sizes of treatment effects and which treatments are statistically different?

There are many different post-hoc analyses possible - the key is to remember not to over-interpret results, and if too many comparisons are made the chance of finding a spurious result increases substantially. Also, the size of an effect may again be more relevant than its significance

What about post-hoc analyses for Response 2 in Patient data B? In SPSS you need to double-click into the table to see the post-hoc results.

A potential drawback of a non-parametric approach is that it is not always possible to readily estimate size effects, and if this is available then the interpretation might not be readily apparent.

## End of Section Milestones

After this section you should be able to

- Understand the hypotheses being tested by an **F-test** or the **Kruskal - Wallis test** and recognise data sets in which these may be appropriate
- Understand more about the statistical decision making and judgments required when dealing with a single response variable and single categorical explanatory variable



## Introduction to Applied Statistics for Medical Researchers

### Part 7 - Hypothesis Testing - Linear Modelling



University of  
Zurich<sup>UZH</sup>

## Hypothesis Testing - Linear Modelling

We have so far covered enough statistical methods and techniques for you to start to analyse more complex data: containing both **multiple populations** and other **nuisance variables** (covariates and fixed effects in SPSS terminology).

The basic approach is almost identical to our previous analyses in SPSS.

In this section we will focus on an in-depth exercise in SPSS to provide experience in the problem solving aspects of analysing more realistic data

## Hypothesis Testing - Linear Modelling

We have seen how to perform hypothesis tests when comparing multiple populations by fitting a linear model and performing F-tests - equivalent to one way ANOVA.

Using a linear statistical model provides an extremely flexible method of analysing study data, in particular it allows us to adjust for other covariates, e.g. age and sex, and thus provide much more powerful hypotheses tests for the treatment effects - which is what we are primarily interested in.

## Tooth growth in Guinea Pigs

### The Effect of Vitamin C on Tooth Growth in Guinea Pigs

The response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid). Variable names: length, treatment, dose.

Source: C. I. Bliss (1952), *The Statistics of Bioassay*. Academic Press.

## Linear Modelling Exercise - Guinea Pigs

### Exercise 12

The Guinea pig data is an excellent case study as it provides almost all the common statistical aspects you are likely to meet in much more complex data.

The scientific question of interest is in determining whether there is **a difference in teeth length when vitamin C is given via orange juice or via water solution.**

- The Guinea Pig data set is in `tooth.sav` - load this into SPSS
- `General Linear Model > Univariate ...`

### Exercise 12 - hints (ctd)

- Always check the analyses is valid - how?
- Could/should dose be treated as a number rather than category?
- Is the effect of treatment the same at different doses? This might be practically important.

### Exercise 12 - hints

- As always first explore the data
- Do the analyses step by step - KISS - start off with very simple analyses and compare subsequent results
- First use the statistical tools we have previously examined e.g. t-test, one way anova etc
- If dose is related to tooth length then a more powerful analyses might be to include this in the model as it may explain some additional variation?

## Additional Exercise - Enzyme reaction with/without Puromycin

### Exercise 13.

Similar structure to Guinea pig data - continuous response variable (rate), one treatment (state - treated or untreated with Puromycin) and one covariate (concentration of Puromycin - continuous). Scientific question of interest is whether the state has an effect on rate. Data are in `Puromycin.sav`.

hint: for the results of any model/hypotheses test to be valid the residuals must be random and approximately normal - or a different model is needed

## End of Section Milestones

After this section you should

- have a basic understanding of how to perform analyses - hypothesis tests - using simple linear models to estimate the effect of treatments after adjusting for variation due to other “nuisance” covariates present in the study design.
- recognise that each data set is different and even if it is of a similar form the appropriate statistical model may be very different
- be able to analyse a wide range of different study data

## Introduction to Applied Statistics for Medical Researchers

### Part 8 - Grouped and Repeated Observations - Introduction



University of  
Zurich <sup>UZH</sup>

## Grouped and Repeated Observations

### Example 1 - Multi-clinic/centre/site studies

In the Guinea pig tooth length data the scientific question of interest was whether there existed a difference between giving vitamin C via orange juice or in water. We also corrected for the impact of different dosages of vitamin C, which helped explain a lot more of the variation in teeth length, and allowed a much more accurate assessment of the treatment effect.

**Question:** Suppose now that this experiment was duplicated at a number of different clinics. This would provide very much more data for analyses. Can the data be analysed in the same way?

## Grouped and Repeated Observations - Introduction

So far we have considered how to determine whether statistically significant differences exist **between** different “treatment” groups, including adjusting for other “nuisance” variables to explain more variation and therefore improve the power of any analyses.

This general approach is applicable to many, if not all, study designs. What we have not yet considered, however, are complications due to **grouping** effects.

A special case of this are repeated observations - **repeated measures** - on the same subject.

### Example 2 - Repeated (longitudinal) Observations

The body weights of 12 rats, in grams, are measured on day 1 and every seven days thereafter until day 64, with an extra measurement on day 44 [the experiment started several weeks before day 1].

There are three groups of rats, each on a different diet. A total of 132 observations.

The dataset is also **balanced** - this is of particular importance in such analyses as it affects what is the best method to use (see later).

### Example 2 - Rat body weight data

**Scientific question: Is there is a difference in the growth rate of rats between the different diets?**

*Potential complication:* Observations within one rat may be more similar to each other than to observations from another rat even although they are both in the same treatment group

⇒ model residuals may not be randomly distributed - a key requirement for any analyses

### Example 2 - Rat body weight data

**Scientific question: Is there is a difference in the growth rate of rats between the different diets?**

Using the same methods as we have looked at previously - a **linear model containing Diet and Time as covariates** - we find an F-statistic for the Diet of  $\approx 513$  giving a p-value  $\ll 0.0001$ , we also find that there is no evidence (p-value 0.2) of changes over Time.

The residuals are clearly NOT normally distributed so these results are highly questionable!

see `exercise14.supplementary.spv`

## Using SPSS for Repeated Measures

## Using SPSS for Repeated Measures

Fitting linear statistical models which comprise both **fixed effects**, e.g. all the analyses we have considered to far, and also **random effects** - which is what is typically needed for analyses of grouped data - is an area which can be extremely confusing in terms of understanding exactly what analyses the computer may be doing.

There are many different terms used e.g. **subject, repeated, correlated, fixed, random, covariance structure**, which are not mutually exclusive and can sometimes mean slightly different things and different software (and books) specify models differently.

In principle, the actual analyses for repeated measures are similar to that which we have worked through previously. The class of statistical models used to perform these types of analyses is, however, very general, and vastly more sophisticated than the other statistical models we have so far used.

As always keep it small and simple, use common sense and try to avoid assuming that the software is doing what you want - if all else fails check the manual 😊

### Rat Body Weights - a detailed worked example in SPSS

- As always begin by exploring the data
- With longitudinal data plotting the observations for each subject is a sensible idea as this can highlight issues of unusual variation
- There are several ways to fit an appropriate model in SPSS - each provides slightly different output
- General Linear Model > Repeated Measures...
- Mixed Models > Linear..
- General Linear Model > Variance Components...

Using a mixed model is generally preferred, and if the data are unbalanced then definitely more appropriate. We will **not** consider unbalanced data.

### Rat Body Weights - a detailed worked example in SPSS

**The scientific question is whether there is a difference in growth rate between the different diets**

We can proceed as previously, the only difference is that now the model contains an adjustment to allow for correlation within the observed data.

- A key output from this analyses is the result of the F-test which tests the hypothesis that all three diets have the same mean - the main result
- We are also interested in the effect of Time - several outputs relevant to this - typically F-tests
- A new output of relevance are the **variance components** - this provides estimates of variation due to subjects - intra-class covariance which is commonly used to provide the intra-class correlation coefficient

### Rat Body Weights - a detailed worked example in SPSS

In this example it may make some sense to treat time as a continuous variable rather than a set of categories e.g. the number of days elapsed from the start of monitoring.

The simplest way is to create a new variable in SPSS

Following on from this, it also makes some sense to consider whether there is evidence of different slopes - e.g. evidence of different growth rates after accounting for within subject covariance

Exercise 14. Can you repeat in SPSS the analyses of the rat data we have just discussed? Start off small and simple - exploratory, then first without the random effect aspect, and then proceed step by step. Do not be too concerned if you find this initially confusing - there are many different ways to do this in SPSS!

## Grouped data

- Analysing repeated measures is a special case of analysing data which has within group correlation
- Analyses would proceed in a very similar way for, for example, trials repeated at different clinics
- The key decision is whether to treat a variable as either a fixed or random effect - this is obvious in repeated measures
- In short, a factor should probably be treated as random if it is a sample from a larger population, e.g. people, animals, doctor, and therefore the effect of a given subject is not of interest.
- With groups such as clinics or hospitals the choice is more ambiguous and will depend on the purpose of the analyses

Exercise 15. Patient data A - data\_A.sav

This data set is can be analysed using General Linear Model with random effects and Variance Components OR Linear mixed model - try them all - differences?

## End of Section Milestones

After this section you should

- Recognise when it might be appropriate to use analyses which adjust for grouping effects
- Have some familiarity with how such analyses might be conducted in SPSS

## Introduction to Applied Statistics for Medical Researchers

### Part 9 - Survival Analyses - a Brief Overview



University of  
Zurich <sup>UZH</sup>

## Survival Analysis - a Brief Overview

Of key interest in survival analysis is estimating the **survivor function**  $S(t)$ , the probability of surviving until time  $t$ ,  $P(T > t)$ .

E.g. determining whether  $S(t)$  differs between treatments.

Other important - and related quantities - are the **hazard function**  $h(t)$  which denotes the instantaneous “death” rate, and the **cumulative hazard function**  $H(t)$ . We will focus our attention on  $S(t)$ .

## Survival Analysis - a Brief Overview

Survival analysis is concerned with the distribution of lifetimes e.g. given a treatment how long will it be until a subject reaches a given state - death/remission/recovery. In statistical terms this topic is similar to all those previously discussed: a treatment is applied to a set of subjects, if a different set of subjects were taken from the same population then the results would be slightly different. We wish to quantify this variability and determine which factors are statistically supported as being influential - using hypothesis tests/confidence intervals - which is exactly what we have been doing previously. Same ideas just different technical details.

## Survival Analysis - a Brief Overview

A distinguishing feature of survival analysis is **censoring** - this is where incomplete information is known about some subject in the study. For example, it is extremely common for studies to be of limited (or fixed) duration and therefore some subjects may yet to have reached the state of interest (e.g. death/remission) by the end of the study. Some information is known - that the subject survived at least until the end of the study duration - this is called censoring and it is important for survival models to be able to deal with this aspect in order to make best use of the data available.



### Leukemia survival times

Survival times are given for 33 patients who died from acute myelogenous leukemia. Also measured was the patient's white blood cell count at the time of diagnosis. The patients were also factored into 2 groups according to the presence or absence of a morphologic characteristic of white blood cells. Patients termed AG positive were identified by the presence of Auer rods and/or significant granulation of the leukemic cells in the bone marrow at the time of diagnosis.

### Leukemia survival times

Scientific question of interest: Do the survival times differ between patients who were AG positive from those who were AG negative?

In SPSS the Kaplan-Meier analyses estimate survival functions and perform the **Mantel-Cox log-rank** test where the null hypothesis is that the survival function is the same across the groups being compared, e.g that the lifetime distributions do not differ

We have two independent populations - and no censoring - could a two sample t-test or linear model be used to analyse this data instead?

### Leukemia survival times

Using the Kaplan-Meier analyses in SPSS we find very strong evidence against the null hypothesis that the survival function is the same in each treatment group - a p-value of 0.004 (and there is also a clear difference in means).

Exercise 16: Can you repeat similar analyses for the data set in `gehan.sav` - see handout for details. Note - this data contain **censoring**

Extra exercise: Can you repeat similar analyses for the data set in `leuk.sav`?

### Leukemia survival times

Suppose now that we also wish to examine the impact of the covariate `wbc` - white blood cell - on the survival time. We can do this by fitting a **Cox Proportional Hazards Model** - think of it as just another type of model similar to the linear statistical models we have already seen. In SPSS a Wald test is used to test whether this term can be set to zero. Note that in this example we use the log of `wbc` (why?)

### Leukemia survival times

Now let's fit a Cox Proportional Hazards Model to investigate the effect of both treatment and also adjusting for the wbc covariate - again this is similar to how we would do a normal linear modelling analysis. We find that both wbc and treatment are strongly supported. We also find that the "effect" of treatment is that `treatment=present` decreases the hazard rate relative to `treatment=absent` by a factor of 0.361. This is the usual way of quantifying the relative effects of variables.

### Leukemia survival times

In the previous Cox Proportional Hazards Model it was assumed that the hazard function for each treatment level only differed by a multiplicative factor - a single baseline hazard function was assumed. A more flexible model is to allow each treatment level to have its own baseline hazard function. This is achieved using the **stratum** command in SPSS. This allows the effect of the covariate wbc to be examined after allowing for more flexibility in the distribution of survival times in each treatment cohort.

One additional analysis which may be relevant for this data would be to examine whether the effect of wbc was different within each treatment group - e.g. by including an interaction between treatment and wbc.

## End of Section Milestones

### Gehan Data - Exercise

Exercise 16: Can you perform Cox Proportional Hazards analyses with the data in `gehan.sav` - see handout for details. What is the effect of the treatment?

Extra exercise: Can you repeat similar analyses for the data set in `leuk.sav`?

After this section you should

- Have some familiarity with how survival data might be undertaken in SPSS
- Be able to perform simple analyses using Kaplan-Meier and Cox Proportional Hazards Models

## Introduction to Applied Statistics for Medical Researchers

### Part 10 - Other topics - a brief mention



University of  
Zurich <sup>UZH</sup>

## Normal Data

So far we have only considered analyses where the response variable is a **continuous** measurement

We now have a brief look at three other common types of **response** variables

The methods we use are largely similar to what we have looked at previously

## Other topics - a brief mention

We have covered in brief some of the basic topics necessary for helping you to start to analyse your own study data.

Many things have not been included - some are worth mentioning as they are particularly common in some areas of medical research

## Other types of data

- **Bounded counts** - e.g. proportions such as disease prevalence,  $r$  out of  $n$  subjects test positive for disease
- **Unbounded counts** - e.g. incidence rates, faecal egg counts,  $r$  cases per year/sample
- **Category/Ordinal** - e.g. questionnaire responses Yes/No/Don't Know, or scoring systems 0,1,2,3

These types of data can all be analysed using a form of linear statistical model

## Other types of data

Some common tests we have not discussed

- **Fisher Exact** - test for dependence in 2x2 tables
- **Chi-squared** ( $\chi^2$ ) another test for dependence in 2x2 tables
- **Binomial Exact** - tests where the proportions are the same in two groups e.g. is there an equal proportion of men and women with diabetes
- **Test of Proportions** - similar to Binomial exact but for comparing multiple populations, e.g. is the proportion of animals with disease the same in all clinics/farms

## Bounded Counts - Example

### BVDV in beef cattle

235 farms were visited and on each between 7 and 14 animals were tested for antibodies to BVDV

Each farmer also completed a questionnaire  $\approx$  50 questions - we wish to determine whether there exist any variables which are related to disease exposure

## i) Bounded Counts - Proportions

- Referred to often as Binomial observations
- **GLM** - generalised linear model is used to analyse this kind of data
- Sometimes called **logistic regression**

## ii) Unbounded Counts

- Referred to often as Poisson observations
- **GLM** - generalised linear model is used to analyse this kind of data
- Sometimes called **loglinear regression**

## Unbounded Counts - Example

### Tick Abundance - Lyme Disease

800 blanket drags were performed across different pastures. For each drag the climate and habitat conditions were recorded.

We wish to identify any factors that affect tick abundance and hence risk of exposure to Lyme disease.

## iii) Categorical/Ordinal Response

- **GLM** - generalised linear model is used to analyse this kind of data
- Sometimes called **proportional odds regression**

## Ordinal - Example

### Soup Taste Test

1847 observations where respondents were asked to say how sure they were that a soup was a particular brand, 1 - sure it was the brand, through to 6 - sure it was NOT the brand. 13 other variables were measured and we wish to identify which variables are associated with the brand preference.

## Some things to remember

- Using generalised linear models is very similar to the previous models we have seen for normal data - things like residuals need to be checked for randomness, although this is more difficult with counts. Also the residuals being normally distributed is not relevant to these kinds of models.
- As always explore the data first visually and with tables before doing any formal analyses.
- keep it small and simple

# Appendices

## A Exercises in SPSS

### Exercise 1. Using SPSS

Relevant files are `chickwts_balanced.sav`, `exercise_1.sps`, `exercise_1.spv`.

### Exercise 2. Importing data into SPSS

A range of different data sets will be used throughout the course. Some of these are currently in Excel or in plain text files, e.g. comma separated. Can you read/import this data into SPSS and provide appropriate variable labels and categories? Things to remember: check for blank rows and/or columns and odd formatting; set the correct measure type; always add labels to nominal/ordinal variables. Give variables helpful labels or names.

Relevant files are `data_chickwts.xls`, `data_tooth.csv`, `data_A.csv`, `data_B.xls`.

Remember - if the data contain errors at this stage then all subsequent analyses may be incorrect so it is important to check the data thoroughly.

### Exercise 3. Numerical and visual data summaries

#### Chicken Weights

How many observations are there per treatment type?

What is the effect of the treatment?

Is the variation within treatments similar?

What is the distribution of weights within treatments?

Outliers?

#### Teeth Growth

How many observations are there per treatment type?

How many observations are there per dose?

How many observations are there per dose and per treatment type?

What the effect of the treatment?

What is the effect of dose?

What is the distribution of length within treatment or dose? Are they different?

Is the treatment effect the same at different dose levels?

Outliers?

#### Patient Experiment A

How many observations are there per treatment type?

How many observations are there per patient?

How many observations are there per patient and per treatment type?

What is the effect of treatment?

What is the effect of patient?

What is the distribution of outcome within treatment or patient? Are they different?  
Outliers?

### Patient Experiment B

How many observations are there per treatment (group) type?

What is the effect of the treatment on response1?

What is the effect of the treatment on response2?

What is the distribution of response1 within treatment?

Outliers?

Is there a relationship between Response1 and Response2?

The scales are very different - can we transform the data to see better?

### Exercise 4. One sample t-test

Load in the SPSS data file `onesampleT.sav`. What is the answer to the following hypotheses tests?

1.

$$H_0 : \mu_{x1} = 100$$

$$H_A : \mu_{x1} \neq 100$$

2.

$$H_0 : \mu_{x2} = 300$$

$$H_A : \mu_{x2} \neq 300$$

3.

$$H_0 : \mu_{x3} = 1$$

$$H_A : \mu_{x3} \neq 1$$

4.

$$H_0 : \mu_{x4} = 100$$

$$H_A : \mu_{x4} \neq 100$$

Now what about these hypothesis tests

1.

$$H_0 : \mu_{x1} = 105$$

$$H_A : \mu_{x1} \neq 105$$

2.

$$H_0 : \mu_{x2} = 305$$

$$H_A : \mu_{x2} \neq 305$$

3.

$$\begin{aligned}H_0 : \mu_{x4} &= 105 \\ H_A : \mu_{x4} &\neq 105\end{aligned}$$

Which results are correct? What are the true values for  $\mu_{x1}, \mu_{x2}, \mu_{x3}$ ?

**Exercise 5. One sample t-test, Wilcoxon signed rank test**

Load in the SPSS data file `onesampleTW.sav`. What is the answer to the following hypotheses tests? ( $\mu_{0.5}$  is used to denote the median)

1.

$$\begin{aligned}H_0 : \mu_{x1} &= 100 \\ H_A : \mu_{x1} &\neq 100\end{aligned}$$

$$\begin{aligned}H_0 : \mu_{0.5,x1} &= 100 \\ H_A : \mu_{0.5,x1} &\neq 100\end{aligned}$$

2.

$$\begin{aligned}H_0 : \mu_{x2} &= 300 \\ H_A : \mu_{x2} &\neq 300\end{aligned}$$

$$\begin{aligned}H_0 : \mu_{0.5,x2} &= 300 \\ H_A : \mu_{0.5,x2} &\neq 300\end{aligned}$$

3.

$$\begin{aligned}H_0 : \mu_{x3} &= 1 \\ H_A : \mu_{x3} &\neq 1\end{aligned}$$

$$\begin{aligned}H_0 : \mu_{0.5,x3} &= 1 \\ H_A : \mu_{0.5,x3} &\neq 1\end{aligned}$$

4.

$$\begin{aligned}H_0 : \mu_{x4} &= 100 \\ H_A : \mu_{x4} &\neq 100\end{aligned}$$

$$\begin{aligned}H_0 : \mu_{0.5,x4} &= 100 \\ H_A : \mu_{0.5,x4} &\neq 100\end{aligned}$$

Now what about these (two-sided) hypothesis tests



1.

$$\begin{aligned}H_0 : \mu_{x1} &= 105 \\ H_A : \mu_{x1} &\neq 105\end{aligned}$$

$$\begin{aligned}H_0 : \mu_{0.5,x1} &= 105 \\ H_A : \mu_{0.5,x1} &\neq 105\end{aligned}$$

2.

$$\begin{aligned}H_0 : \mu_{x2} &= 305 \\ H_A : \mu_{x2} &\neq 305\end{aligned}$$

$$\begin{aligned}H_0 : \mu_{0.5,x2} &= 305 \\ H_A : \mu_{0.5,x2} &\neq 305\end{aligned}$$

3.

$$\begin{aligned}H_0 : \mu_{x4} &= 105 \\ H_A : \mu_{x4} &\neq 105\end{aligned}$$

$$\begin{aligned}H_0 : \mu_{0.5,x4} &= 105 \\ H_A : \mu_{0.5,x4} &\neq 105\end{aligned}$$

Are the results similar?

### Exercise 6. Which test and why?

Load in the SPSS data file `onesampleT_final.sav`. For each of the seven data sets decide which test is most appropriate - give a reason.

### Exercise 7. Two sample t-test or Wilcoxon rank sum test

Load in the SPSS data file `Tooth.sav`. Which test is best? Do they differ? What about medians?

### Exercise 8. Paired t-test and Wilcoxon signed rank test

Load in the SPSS data file `twosample.sav`. Which is the most appropriate test? Are they both appropriate/reliable?

### Exercise 9. One Way ANOVA chicken weights

- Load the dataset `chickwts_balanced.sav`
- First remind ourselves of the what the data looks like - means, boxplots, P-P plots etc
- Then `Analyze...> One Way ANOVA` with only minimal outputs

### Exercise 10. One Way ANOVA patient data

i) considering `data.A.sav` (ignoring, for the moment, the patient variable)

- 1 Is there a statistically significant effect of treatment?
- 2 Is the F-test valid here?
- 3 Can you provide an SPSS viewer file which supports your analyses?

ii) now consider `data.B.sav` (ignoring the specimen variable)

- 4 Is there a statistically significant effect of Group with either of the two response variables?
- 5 Is the F-test valid here? If not can you make it valid? hint: transformation or perhaps dropping a group?
- 6 Can you provide an annotated SPSS viewer file which supports your analyses?

### Exercise 11. One Way ANOVA sizes effects

i) considering `data.A.sav` (ignoring, for the moment, the patient variable)

- 1 what are the sizes of treatment effects and which treatments are statistically different?

ii) now consider `data.B.sav` (ignoring the specimen variable)

- 2 what are the sizes of treatment effects and which treatments are statistically different?

### Exercise 12. Linear Modelling - fixed and covariate effects, Tooth data

The Guinea pig data is an excellent case study as it provides almost all the common statistical aspects you are likely to meet in much more complex data.

The scientific question of interest is in determining whether there is **a difference in teeth length when vitamin C is given via orange juice or via water solution.**

- The Guinea Pig data set is in `tooth.sav` - load this into SPSS
- General Linear Model > Univariate ...

### Exercise 13. Linear Modelling - fixed and covariate effects, Puromycin data

Similar structure to Guinea pig data - continuous response variable (rate), one treatment (state - treated or untreated with Puromycin) and one covariate (concentration of Puromycin). Scientific question of interest is whether the state has an effect on rate. Data are in `Puromycin.sav`.

hint: for results of any model - hypotheses tests to be valid the residuals must be random and approximately normal

#### **Exercise 14. Repeated Measures - rat data**

Can you repeat in SPSS the analyses of the rat data `rats_longformat.sav`, `rats_shortformat.sav` we have just discussed? Start off small and simple - even without the random effect aspect and then proceed step by step.

#### **Exercise 15. Repeated Measures - patient data**

Patient data A - `data_A.sav`

This data set is can be analysed using GLM with random effects OR Linear mixed model - try them both - differences?

## **Exercise 16. Survival Analyses - Kaplan-Meier + Cox Proportional Hazards**

Can you perform a Kaplan-Meier analyses with the data in `gehan.sav`. What is the effect of the treatment? Can you perform Cox Proportional Hazards analyses with the data in `gehan.sav`. What is the effect of the treatment?

## **B Selected Data sets for Exercises**

### **B.1 Teeth Growth in Guinea Pigs**

`data.Tooth.csv`. Experimental data to explore the impact on teeth growth of giving Vitamin C to Guinea pigs. Comprises Three variables: length; Treatment; and Dose. Treatment is categorical where 1 - denotes vitamin C given in “water” and 2 - via “orange juice”. Dose is numerical and in milligrams.

### **B.2 Chicken Weights**

`data_chickwts.xls`. An experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens. Two variables: Weight and feed type. Feed type has six categories 1 - “casein”, 2 - “horsebean”, 3 - “linseed”, 4 - “meatmeal”, 5 - “soybean”, 6 - “sunflower”.

### **B.3 Hamilton Depression data**

`twosample.sav`. Hamilton depression scale factor measurements in 9 patients with mixed anxiety and depression, taken at the first and second visit after initiation of a therapy (administration of a tranquilizer). Taken from Myles Hollander and Douglas A. Wolfe (1973), Nonparametric Statistical Methods. New York: John Wiley & Sons. Pages 2733 (one-sample), 6875 (two-sample). Or second edition (1999).

### **B.4 Patient Experiment A**

`data_A.csv`. 20 observations across three variables. Patient is categorical, Treatment (type) is categorical, Outcome is continuous (scale)

### **B.5 Patient Experiment B**

`data_B.csv`. Six groups of four specimens where each has two different observations made on it. The observations are continuous, specimen and group are categorical.

### **B.6 Rat Body Weights**

`rats_longformat.sav`, `rats_shortformat.sav`. Hand and Crowder (1996) describe data on the body weights of rats measured over 64 days. The body weights of the rats (in grams) are measured on day 1 and every seven days thereafter until day 64, with an extra measurement on day 44. The experiment started several weeks before “day 1.” There are three groups of rats, each on a different diet.

## B.7 Puromycin Enzyme Reaction

`puromycin.sav`. The “Puromycin” data set has 23 rows and 3 columns of the reaction velocity versus substrate concentration in an enzymatic reaction involving untreated cells or cells treated with Puromycin. Data on the velocity of an enzymatic reaction were obtained by Treloar (1974). The number of counts per minute of radioactive product from the reaction was measured as a function of substrate concentration in parts per million (ppm) and from these counts the initial rate (or velocity) of the reaction was calculated (counts/min/min). The experiment was conducted once with the enzyme treated with Puromycin, and once with the enzyme untreated.

## B.8 Survival Times and White Blood Counts for Leukemia Patients

`leuk.sav`. Survival times are given for 33 patients who died from acute myelogenous leukemia. Also measured was the patient’s white blood cell count at the time of diagnosis. The patients were also factored into 2 groups according to the presence or absence of a morphologic characteristic of white blood cells. Patients termed AG positive were identified by the presence of Auer rods and/or significant granulation of the leukemic cells in the bone marrow at the time of diagnosis.

## B.9 Remission Times of Leukemia Patients

`gehan.sav`. A data frame from a trial of 42 leukemia patients. Some were treated with the drug 6-mercaptopurine and the rest are controls. The trial was designed as matched pairs, both withdrawn from the trial when either came out of remission.