Summary of project subcourse,
Design and analysis of clinical trials

Carl-Fredrik Burman, 29 Febr 2008.

Based on the draft reports from the four groups, I will try to summarise a few learnings from the project course. You're advised to use these notes, in addition to other material, when preparing for the exams.

The idea behind the project course is that you should "learn from experience". By looking at what the other groups have done, you may hopefully get experience from 4 trials "for the price of one". Experience typically means that one has made a lot of errors, and by remembering them one could avoid repeating them. I'm very happy with your work so far. This doesn't mean that I think that all the studies have been perfect. On the contrary, the interesting things are the un-perfections, and I will mainly use those in the text below. You've done a lot of good things also but we will probably learn less from them.

Study protocol

I hope that you have learnt by experience that preplanning is essential. A good study protocol is a great help when conducting, analysing, interpreting and reporting a trial. It's not easy to write a protocol and you may have encountered a number of issues during conduct and analysis, which you didn't think of at the planning stage but would have been good to handle in the protocol.

It is rather common that the study protocol excludes so many patients that it's hard to find study patients. One should of course consider which inclusion criteria are needed and how those chosen would effect the population and number of eligible subjects. It may also be good to consider what to do if some individual is included without fulfilling the criteria. For example, "women of child-bearing potential" are often excluded according to the study protocol from a pharmaceutical trial, as one is afraid of damages on foetuses. If such a woman despite this receive treatment and fulfil the trial, should we use her data in the analysis? One way of arguing is to say that these data are equally useful as the data from anyone else; the reason for the exclusion criterion was only safety, not that this subpopulation should be irrelevant.

The standard deviation indicates that Group D included an individual older than 30 years. This is in violation of the protocol, "*20-30 years old. This narrow age range is chosen since the metabolism tends to change as people get older*". What to do?

In the chocolate study, "One person was asked to come to the table from a corridor nearby. He was an expert in the field, working for Selecta or a similar company. He was later excluded though because he was considered as an outlier." I don't know in what sense he was regarded as an outlier. Should he be excluded?

Another examples of discrepancies between protocol and study conduct: According to the protocol: "Treatment … (approximately 3 to 4 hours after last meal)". In study, one individual had 19 (?) hours since last meal.

Study objectives

What to compare

At several occasions, I don't think that the description in the protocol or report is clear on exactly what should be compared. Ideally, a non-statistician should understand what is compared and the main assumptions. A statistician should be able to follow exactly which test is used, if change from baseline is taken, which covariates are included, etc. It might be useful to give a precise mathematical description of the models (perhaps in an appendix). For example: The responses in individual i=1,…,n are assumed to be $y_i = m + a*X + b*Y + c*Z + e$, where e are i.i.d. N(0,sigma) residuals, X is treatment (1=active, 0=placebo), Y the baseline value of …

Example.
The study protocol was very clear in saying that the primary objective was "To compare the instant effect of oral sugar intake versus placebo on blood glucose". Here it is clear that the two treatment groups (with/without sugar) are to be compared head-to-head. It would be good to be equally clear when presenting the results.

What should be in the report?

The draft reports have clear differences in what aspects of the study that are discussed at length. Report B has a long and illuminating section about the practical procedure, whereas some other report focuses almost only on statistics. The reports are currently more or less early drafts and the final reports will hopefully cover all key aspects.

Anyway, it is worth thinking of what is needed in a report or a scientific publication. For the results to be convincing, all main components (design, conduct, analysis, interpretation and reporting) have to be of sufficiently good quality. If one link is weak, the value of the trial may be greatly diminished.

For example, if the design is inappropriate, the trial may not be able to answer the study question(s). In one of the trials, the pulse (heart rate) decreased from baseline after espresso intake. (The effect had been statistically significant in a 2-sided test without multiplicity corrections. For this example, assume that the decrease was clearly statistical significant, so that it can be ruled out that it was a chance effect.) Does this result show that espresso leads to lower pulse? No, the conclusion is that sitting down, having a double espresso, and waiting a certain time, leads to lower pulse. Perhaps it would have been enough to sit down and rest for a while to decrease the pulse? How could we design the trial to give an unambiguous answer to the question about the effect of espresso? The answer is to have a control group with e.g. no drink or other treatment. The design could be parallel or cross-over. In any case, one should randomise to espresso or no treatment. (I think the group's original design was like this, but I asked them to include exercise to get a clear significance, and exercise took the place of treatment B.)

Flaws in the study conduct may spoil a trial by introducing bias. Cf. the discussions from Group B in the next section. Double-blinding, where possible, could clearly improve the credibility. It is also important to describe the study procedures well in the protocol and clearly state in the report how the procedures were followed.

The main purpose of the report is to convey the scientific results. Which statistical analysis was used is for example not the most important thing, but it is essential to

describe the analysis in order to convince the reader that the results are trustworthy. We would like to build a solid case for our conclusions. For example, a trial of caffeine's effect on pulse may be reported with this abstract (details have to be provided in the rest of the report):
The primary objective was to investigate whether caffeine intake leads to increased pulse. The design, study conduct, and statistically analysis were all prespecified in the protocol. A total of 100 healthy volunteers (age 18-70 years, 55% females) were randomised into two equally sized groups, to receive either a tablet containing X mg of caffeine or a placebo tablet, similar in appearance, taste and smell. The primary variable was change from baseline to 15 minutes after tablet intake in pulse. A 2-sample 2-sided t-test was statistically significant ($p<0.001$) and the corresponding 95% confidence interval for the placebo-controlled effect of caffeine on pulse was 1.9 ± 1.1 beat/min. Due to the randomisation, the only explanations for the result are a) that caffeine increases the pulse, or b) that the effect is due to chance. The small p-value speaks against the latter explanation. Our conclusion is therefore that caffeine increases the pulse. (The generalisation of the results to other populations or other times after intake does not follow from pure logic. However, using medical and pharmacological knowledge, it is possible to argue for such generalisations.)

Study conduct

Group B has already produced a report with a lot of text and a number of interesting observations, especially regarding how the trial was carried out. They give a number of details (no milk or sugar was offered with the drinks) that would help an independent research team to replicate the study. This may also help e.g. a referee to assess whether the study was well conducted or may have flaws.

The group reports that the questionnaire (which one could have expected to be fool-proof) was not understood and filled in correctly:
"We were interested in using a grading scale on preference for each chocolate, to be able to better estimate the magnitude of difference in preference. However, our results indicate that people have misinterpreted the scale used for the grading. Some seem to have understood that 1 was the best and 4 the worst, whilst others obviously have inverted the scale. This underscores the importance of using an easily understandable and comprehensible scoring system." The group later continues:
The participants were asked both to rank the taste of each chocolate on a scale from 1 to 4, but also to decide which chocolate they did prefer. Even if participants gave higher rank to chocolate "A" he/she could have marked "B" as the perefered one. We clearly stated that "1" was the highest grade and "4" the lowest but it seemed like the participants didn't really read the instructions for the scales. We don't think the instruction could be much clearer than it was but maybe we should really ask each participant to read the form carefully before answering. One way of solving this would be to use another type of scale. We did think about two options:

1. *Terrible – Bad – Good – Delicious*
2. *Some visual representation like smiling faces. "*

Two participants filled out the drink question wrong. They did actually drink tea but marked their questionnaire with coffee. These were afterwards substituted by two new participants.
Should these individuals been kept and the data on drink corrected? You may think of pros and cons. What could have been done to eliminate the risk of false reporting?

"Everyone except two people was participating directly. These two people draw questionnaires but then disappeared. They wanted to come back later but they didn't come before the experiment was over."

Blinding

Group B writes that the chocolates A and B were cut in 3*3 cm and 1.5*1.5 cm squares, respectively. They were put in two different pots marked A and B. This is no perfect blinding, and you may think about possible risks of bias. Preferably, the treatment should look, taste, spell equally; side effects should not be able to reveal the treatment; there should be no pattern in the coding (different random numbers for all medicine packs e.g., not label A and B), and all involved should be blinded. Of course, the ideal is often not possible to achieve.

"There were quite many people participating in the experiment at the same time and we couldn't actually stop them talking to each other. One person was trying to convince her friend about which chocolate was the best one." This is clearly an indication that observations may not be independent.

Randomisation

I have the impression that many students have not fully grasped the importance and significance of randomisation. I would recommend that you carefully read what the lecture notes and the book have to say about this topic.

Many studies within medicine and other fields are *observational*. The treatments are not randomised; they are spontaneously chosen by the individuals. As an example, a recent study, mentioned in the radio news (but I didn't here the reporting myself and don't know the details), found that the risk of overweight in children was correlated to their consumption of salt. The researchers concluded that if you eat more salt, you will drink more (sugar-containing) soda (läsk), and that would lead to overweight. We could ask a number of questions: Does salt intake lead to soda consumption? Or does soda cause salt intake? Or is everything a matter of social class: less educated families allow (or has less control) more salt, more soda, and children in these families exercise less than in other families. Thus, social class may be the cause behind less exercise leading to overweight and behind increased salt consumption. In short, the causal structure may be very unclear. Furthermore, just observing what people do cannot clearly resolve all questions.

Randomisation is what makes clinical trials the golden standard. Observational trials may be needed (for example, for ethical reasons) but clinical trials have strong advantages when they can be performed. If it was possible to randomise children to receive more or less salt, one would be able to find out if salt causes over-weight (possible through e.g. soda consumption).

Group A included lavatory visits in the main analysis. This may remove some of the logical advantages of the randomisation. It is also preferred to ask the participants to choose water consumption (0.5, 0.75, or 1 l) before randomising the order of the treatments. The participant answering 0.75 l would than have to take 0.75 l, either the same day (period 1) or in period 2.

The usual rule is to use only covariates that were measures before randomisation (or, as lab variables, measured completely independent of the randomisation code). Exploratory analyses may use post-randomisation covariates but the results need careful interpretation.

Instruments

How to measure a treatment effect can be critical. Besides ensuring that the instruments are unbiased, decreasing variability will transfer to greater trial efficiency. A recommendation is to analyse from where the variability comes and how much one

would gain (in terms of efficiency, possibility of decreasing the sample size) if certain measured could be done to improve the precision. My impression is that it's common that teams work hard to reduce variabilities that are so small that they could be ignored. On the other hand, simple ways of decreasing variability may be missed. One simple idea is to use duplicate measurements. In a long-term weight loss trial, for example, it may be worthwhile to measure body weight on two different days at the end of the treatment period.

Group A had two different scales (my fault) and used the average in the main analysis. One of the scales ("Opera") seems to be considerably more precise than the other ("Caffe"). Would it have been better to ignore the readings from the latter scale? It should be possible to address this question by analysing the data! The group has started to look into the performance of the scales. I would suggest looking directly at the residuals. For example, after no treatment (no water intake), the body masses are usually relatively constant. Does one scales show larger variations in the four measurements. One should of course pool data from all participants, and may adjust for an estimated decrease.

Sample size calculation

This should reflect the primary objective and the statistical analysis planned to answer the primary question. In case a larger sample size is needed for an important secondary variable, this sample size may be chosen instead.

It is hard to predict how large the effect will be in a trial. It's usually easier to assess the standard deviation but even this estimate may be far from what is later observed in the trial. If you haven't done so already, please compare your guesses before the trial with the estimates from the trial.

Group B assumed that 75% would prefer the high-cocoa chocolate. The results were that this chocolate was significantly *less* preferred than the control. Through history, this is certainly not the first time when the results of a study are contrary to what was anticipated.

Group A had four primary objectives and dimensioned the trial based on the primary objective for which they thought that it would be hardest to show an effect. Assumed effect of coffee (compared with baseline) was 5 mmHg. Observed effect was 0.75 if I understand the results (not significant). Furthermore, the estimated effect of coffee on pulse was *negative* (p=0.02 in 2-sided test) although the group had predicted that they could demonstrate a significant positive effect. The results for the other two variables were more in line with expectations.

Group D assumed an effect of 0.5 mmol/L, observed 0.233 +/- 0.613 (?), and got a non-significant result. Looking at the length of the confidence interval, we can conclude that the result would not have been significant even if the observed effect had been the same as the assumed, 0.5. Thus, it seems as if the variability was larger than expected. We don't know what the true effect is, but note that if the effect is 0.25 instead of 0.5, assuming the same standard deviation, four times more patients are needed to give the same power.

Statistical analysis

My main reflection is that you tend to use so complicated statistical methods: GLM, stepwise regression, etc. Many sophisticated methods are better (e.g. give more answers, or are more powerful) but it's often useful to consider the basics, the

simplest method that would do the job. Do you really always understand how the methods you apply work and how the results should be interpreted?

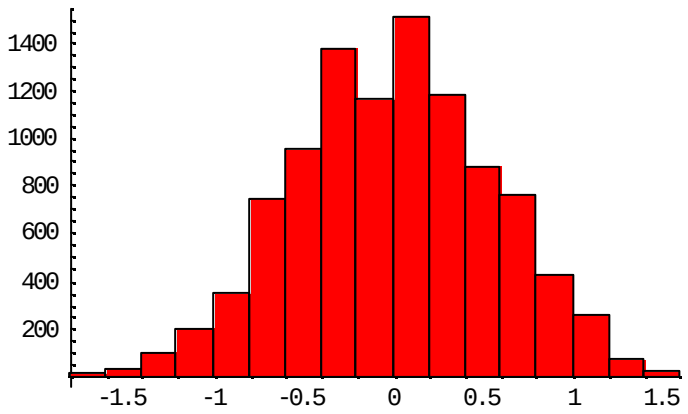
Randomisation test

Say that we are to test the null hypothesis:
H_0: The response variable does not depend on the treatment, versus the alternative
H_A: The response tends to be higher after treatment A than after treatment B.
We plan to conduct a parallel group trial, where participants are randomised to receive either treatment A or treatment B.
Before running the study, we define any reasonable test statistic T, which is expected to get a higher value if the alternative is true than if the null hypothesis holds. The statistic T may, for example, be the difference (between treatment group B and A) in average response; difference in median; estimated treatment difference using a certain linear model with covariates; the rank sum for treatment A; or (if variables have to be positive) the product of all responses in group A. After collecting the data, we calculate the observed value T_{obs} of the test statistic we chosed before the trial.

The p-value is the probability that one by chance (that is, assuming the null hypothesis) would observe a greater value of T than T_{obs}. How could one calculate a p-value? Usually we are basing this calculation on distributional assumptions. If we assume independent normal variables, we may e.g. get a t-test. The randomisation test approach is less common, and is usually not needed, but I think that if the approach is understood it may help one's thinking in non-standard situations.

The randomisation test starts with all the observations, ignoring from which treatment group they stem. Say that from the study we have the following 14 responses: 4.9; 4.9; 5.2; 5.4; 5.7; 5.7; 5.9; 6.1; 6.4, 6.8; 6.8; 7.4; 7.6; 7.8. (These are the data from the blood glucose experiment with a small modification to get a clearer treatment effect).

If we know that 9 of the participants received treatment A and 5 received B, what values for T = (the difference in average response) would we expect? If the null hypothesis is true, the two treatments give the same effect (or has the same distribution for the response). Thus, assuming H_0 we have no idea about which 9 of the 14 responses that come from treatment group A. We would expect these 9 values to appear as drawn by random from the 14. The histogram below displays the difference in average between the two treatment groups, when the nine A observations are drawn at random. The graph is based on 10,000 simulations. (In this case, one could even calculate the exact randomisation distribution. One may choose 9 out of 14 observations in "14 over 9" = 2002 different ways, so there is no match for a computer to check all possibilities.)



Now it's time to reveal which observations that belonged to the different groups in the study.

A	B	B	B	A	B	B	A	A	A	A	A	A	A
4.9	4.9	5.2	5.4	5.7	5.7	5.9	6.1	6.4	6.8	6.8	7.4	7.6	7.8.

The observed statistic is T_{obs}=1.19. The tail area in the histogram above this value (1.19) is 1.2%. The p-value is thus 0.012 and there is a statistically significant difference between the groups.

In a cross-over trial, one can base the randomisation test on the null hypothesis assumption that treatment A and B are exchangeable *within* subject. This translates into simulate the signs (+ or -) for the treatment difference for each individual. The randomisation tests can be made more powerful by including covariates, as in an ANCOVA or logistic model e.g. The randomisation test usually gives very similar results as a corresponding distribution-based test. The randomisation t-test has e.g. the same asymptotic efficiency as the usual t-test.

Group C rightly states "All zero effect (can be) ignored" in the Wilcoxon sign-rank test.

Randomisation tests are described briefly on page 270-272 in van Belle et al. "Biostatistics: A methodology for the health sciences", 2nd edition, Wiley.

Continuous variables

Classic distribution-based tests are the one-sample and two-sample t-tests. If the sample size is rather large, normal distribution tests are very good approximations. A 2-treatment parallel group study can lead to a two-sample test. A 2-treatment cross-over design would rather give a one-sample test, where the difference within individual between the two treatment responses is the variable used in the test.

By using a linear model with covariates, the efficiency can often be improved. If 10% of the variability (residual variance) is explained by covariates, the normal distribution test will have asymptotic efficiency 1-0.1=0.9 compared with the ANCOVA. Thus, the sample size may be decreased 10% when using the more sophisticated test. The linear model is also useful to study whether different covariates affect the response. A simple way of getting a grip of a covariate's effect is to plot the response (or residual without the covariate in the model) as function of the covariate.

Even if you are using a sophisticated procedure in your statistical software, it may be useful to check a simple t-test (normal test) so that you get similar results. It may take some time to get a good intuition of what the software really is producing.

Some groups have been looking at data to determine whether the normality assumption is valid. It's always a good idea to plot some data and residuals. The normal distribution plot is one possible tool. However, it is not possible to prove that the distribution is exactly normal. In your studies, the number of data points is often so few that we don't learn much about the distribution. On the other hand, standard tests are often reasonably robust to moderate deviations from the distribution assumptions.

0-1 variables

In the chocolate-tasting example, the primary variable was the preference (chocolate A or B) reported by the participants. This is a zero-one variable and the simple test is

based on the binomial distribution. Under the null hypothesis, the distribution is BinDistr(n,0.5).

Two participants didn't answer the preference question. What should one do with such missing data? The decision on how to handle this should preferably be made before study start. When this was not done, we may introduce some bias. One possible solution would be excluding the individual from the analysis, another solution to check if the individual had a better score (on the scale 1-4) for one of the chocolates. One could then argue that this difference in score should be interpreted as a corresponding preference. There is a risk that such decisions are affected by the knowledge of which chocolate got the highest score. A double-blind study would have the advantage that such issues could be resolved more objectively, without the risk of being affected by treatment knowledge.

Multiplicity

Only one of the groups, I think, applied multiplicity corrections. (Group C used Hochberg's correction for the four primary objectives.) The other groups, after you have had a lecture on multiple inference, may see some multiplicity problems. Could you trust a significant result for a secondary variable, or a significant result in one alternative analysis of the primary variable when another analysis is non-significant?

Multiplicity is a complicated area. The mathematics is sometimes hard, but the philosophy is much worse. Should we be allowed a total type I error (family-wise error rate) of exact 5% for each trial? Why 5%? Why should we get 5 new % if we stop the trial and start a new similar one?

The basic advice is to preplan. Specify everything in the protocol. Define how you will correct for multiplicity. By thinking carefully, you can often reduce the number of key comparisons to a rather small number, and may proceed to defining how to control the type I error level within that small family. (Weighted) Bonferroni tests, Holm's procedure and sequential testing (if hypothesis 1 is significant, go on to test hypothesis 2, etc) are useful and relatively simple. There are often a lot of other comparisons that could be of interest, and you may choose to test (or just calculate) confidence intervals) while treating them as exploratory. In general, the report should clearly state for which family the error level is controlled, and stress the exploratory nature when discussing the results of exploratory analyses (you don't have to do this for the lab reports).

The frequentist hypothesis testing paradigm is not always useful (some would say that it's never useful!). Bayesian and decision-theoretic approaches are sometimes better fit for your purpose. Confidence intervals are often more useful than p-values.

In many cases, there is a structure in the data that can be utilised. In case there are several doses of the same drug in a trial (different treatment groups), one should consider using a model for dose-response when analysing the data. In the water intake experiment, individuals could choose between three different amounts of water (not randomised). One might have pooled the data from the three groups, for example by estimating weight gain as fraction of water consumption.

Another common structure is that several measurements are made over time. It may then be possible to model time-effect. The power in the primary analysis can often be increased by utilising more than one time point. In the water experiment, one could try a parametric time-response model.

Interpretation of results

The tables that the computer software spits out are not always intelligible to a layman or someone who don't know how variables have been coded.

It is sometimes unclear in the result tables in what directions the results go. (This might be clearer in the final reports when more text is added.)

Example 1:
Does the significant p-value for "Baseline Blood Glucose" mean that participants with high baseline glucose levels performed better or worse on the word count?

General Linear Model		
Source of Variation	SS	p-value
Treatment	5.556	0.396
BMI	0.651	0.041
Baseline Blood Glucose	27.494	0.003
Error	30.299	
Total	64.000	

Table 7. General Linear Model for Word Count

Example 2:
Does the significant result for "Age" mean that older individuals were comparably fonder of chocolate A or B?

Table 6. Odds ratios and 95% CI for covariates on primary variable

Covariate	Odds ratio	95% CI	p-value
Age	0.947	0.902-0.994	0.028
Drink (Tea vs Coffee)	1.750	0.593-5.162	0.311
Years of work	0.975	0.910-1.045	0.473
Hours of work	0.671	0.420-1.071	0.094
Occupation (employee vs student)*	0.384	0.108-1.370	0.140

Group D checked whether the "instruments" used to measure cognitive ability were calibrated. That is, they checked if the scores were similar for the two different word puzzles. The conclusion was "we could prove that the puzzles have the desired property of having the same difficulty". Mathematically, "the same" is a point and there is not possible to show this by observing random variables. Statistical inference could only give a limited confidence interval (compare equivalence and noninferiority studies). Thus, one extra step is needed in the reasoning. For example: 1) the confidence intervals show that the difference in mean score is no more than 3; 2) differences smaller than 3 are regarded as acceptable (or irrelevant) as …. 3) Thus, we conclude that the puzzles are sufficiently similar in difficulty for our purposes.

The generalisability of results is important to discuss. One group states that *"We had a very specific population (mostly females aged 19-25). It might be interesting to compare our results with a similar study made in the harbour, at "Volvo Torslanda", in a shopping hall, in a church, at the central station, in the middle of the street (hopefully in the summer) etcetera. We are not sure that our results can be extrapolated for instance to the whole population of Gothenburg, Västra Götalands Län, Sweden, Europe or the entire universe."*
Survey theory has ambitions of being able to generalise results based on taking a random sample from a larger population. However, clinical trials are usually not possible to generalise in a narrow statistical sense. What is needed is a

medical/scientific reasoning, which may be supported by statistical analysis based on population characteristics.

In an earlier version of the report, group B had tested whether the chocolates were equally preferred among tea drinkers (B significantly better), and among coffee drinkers (no significance). From this, one *cannot* draw the conclusion that tea drinkers preferred B comparably more than coffee drinkers did. What is needed to answer this question is to consider interaction in the 2*2-table (coffee/tea, prefer A/B). The test can be based on the estimated difference of p2-p1, or using Fisher's exact test, or using a chi2 test. All these approaches will give similar results (for reasonable sample sizes). The current report states no significance.

Group C writes "The table … shows that treatment A has … a small negative effect on the pulse". The estimated effect, -3.63 beats/min, is rather large, I think. More interesting, however, is that the reported (unadjusted) p-value was 0.99. Should p=0.99 be interpreted as "showing" an effect? The test was performed one-sided, with the alternative hypothesis stating that the pulse would *increase* (not decrease). If the test would have been 2-sided, the p-value had been 0.02, indicating a *decrease*. As the results of many real clinical trials have gone in the opposite direction to what was believed at the design stage (e.g. new drug performs worse than active control in trial), the default is to make all tests 2-sided.

Group A concludes that BMI had a significant effect but stated that the effect was very small. The estimated effect was an extra 0.0427 kg body mass increase for each BMI unit (kg/m2). A difference in BMI of 5 kg/m2 would therefore correspond to a difference in weight gain of 0.21 kg --- this seams to be non-ignorable. However, the p-values were close 5%, and I suspect that the estimate by chance was much larger than the true expected value.

Ethics

Luckily, we haven't had any severe ethical mistakes. Real clinical trials will give harder challenges. It's worth noting that the confidentiality of data is important. Group D stated in the protocol "… recorded nameless and kept confidential". Group A had names in the first excel file I got but they had spontaneously improved the confidentiality in the next version.

Please remember to study the course material about ethical aspects of clinical trials.