

Review Notes

Hypothesis Testing, P values, Power and Sample Size

- with a clinical trial the conclusions of the trial are usually expressed as either the experimental therapy being better or not better than usual care, and the results are either statistically significant or not

P value

- the letter P stands for the probability of obtaining the observed difference by chance, if in reality the null hypothesis is true and there is no difference between the groups
- another way of thinking about the P value is 'if there is no difference between the experimental and control groups and the trial was repeated over and over again, what proportion of the trials would conclude that the difference between the two groups was at least as large as that found in the study?'
- *where did $p < 0.05$ come from? Sir Ronald Fisher proposed it 'if the probability of such an event (falsely rejecting the null hypothesis) were sufficiently small – say 1 chance in 20, then one might regard the result as significant'*
- *so our traditional P value is really a historic (somewhat arbitrary) number that we continue to use*
- *an example of probability: if you were to toss a coin, about 1 in 16 times you would get 4 tails in a row, and about 1 in 32 times you would get 5 tails in a row (and this would be a $p < .05$ and considered 'statistically significant' even though it occurred by chance; the exact p value for tossing 5 tails in a row is 0.031)*

P values: some important points

- statistically significant does not necessarily mean clinically significant
- P values don't tell us anything about the size of the effect of the therapy
- the smaller the P value the stronger the evidence (the results are less likely to have occurred by chance)
- results of research should not be interpreted simply as 'significant' or 'non-significant' but should be interpreted in the context of the type of study, and possible problems with the validity of the study (Is the Study Valid?)

Hypothesis Testing

- hypotheses are a prediction about what the data will show
- in a clinical trial we start with a **null hypothesis (H_0)**: this assumes there is no real underlying difference between treatments, and any observed difference in the results has occurred by chance
- we continue to support the null hypothesis until the data makes it untenable, then we reject the null hypothesis and accept the alternative hypothesis
- the **alternative hypothesis (H_1)** states that a difference between the two groups exists
- what is untenable? when the probability of the observed result occurring by chance is less than a prespecified threshold (or alpha)
- this threshold has traditionally been set at a chance of less than or equal to 1 in 20 or $p < 0.05$

A 2 x 2 Table for Hypothesis Testing

		Truth	
		There <u>is</u> a difference	There is <u>no</u> difference
Study	Positive Result Null hypothesis(Ho) rejected	Correct Conclusion Power = 1 - beta	Type I error (alpha)
	Negative Study (No difference found) Null hypothesis (Ho) accepted	Type II error (beta)	Correct Conclusion

- there are two types of error that may occur with a clinical trial (i.e. two ways of coming to the wrong conclusion)
- these are known as Type I and Type II errors

Type I error

- this is falsely concluding that there is a difference between the groups, when in reality there is not
- this can also be thought of as a false positive study result
- the likelihood of having a Type I error is known as **alpha**
- by custom the level of alpha is usually set at $p = 0.05$ ('statistically significant'); this means that the investigator is willing to run a 5% chance (but no higher) of being in error and falsely concluding that there is a difference between the experimental and control groups (rejecting the null hypothesis)
- the P value is a quantitative estimate of the probability that the difference found in the study could have occurred by chance alone that is determined statistically when the study is completed
- *remember that we must have first decided that the study was valid and bias minimized before we will accept the results*

Type II error

- this is falsely concluding that there is no difference between the experimental and control groups, when in fact there is
- this can also be thought of as a false negative study result (a Type II error only applies to negative trials)
- the possibility of a Type II error must be considered when reviewing negative trials
- the likelihood of having a Type II error is known as **beta**
- traditionally beta is set at 0.20; this is the maximum probability of failing to find a statistically significant difference when a true difference exists, or the maximum probability of making a Type II error

Power

- power is defined as $1 - \beta$; the power of a study represents the ability of the study to detect a difference when it exists
- for example if β is 0.20 then the power is 0.80, or the study has an 80% probability of being able to demonstrate a statistically significant difference between the two groups if it actually exists

What does power depend on?

1. **sample size**
 - the smaller the sample size the harder it will be to find statistical significance
2. **effect size**
 - before the trial is started the effect size is estimated as the difference between the two groups that will be considered clinically important
 - it is easier to detect a large effect size (for example a decrease in mortality of 50%) than it is to detect a small effect size (a decrease in mortality of 2%)
3. **level of significance**
 - traditionally α is set at 0.05
 - if α were to be set at 0.001 then we would need to increase sample size
4. *standard deviation of the data (don't worry about this)*
 - *if data sets have a very large 'spread' or a large standard deviation, the power of the study will be lower*

Sample Size Calculations

- investigators have to decide **before** the study begins how many people will be required
- studies should describe how the sample size was determined

Components of a sample size calculation: (just background information)

1. power:

- *investigators must decide on the power of the study (before it starts); that is the ability to detect a true difference in outcome between groups*
- *if power is set at 80% then there is a 20% (β) likelihood of missing a difference when it exists (a false-negative result or Type II error)*

2. level of significance:

- *the significance level is the α level, or the likelihood of making a Type I error*
- *significance level is traditionally set a 5% (or $p = 0.05$)*

3. population event rate:

- *this is the best estimate of how often an event (outcome) will occur*
- *information from other studies is used to predict this*

4. size of treatment effect:

- *the effect of treatment that will be considered to be clinically significant must be decided before the trial since this will impact the sample size required*
- *this can be a challenge, and an implausibly large treatment effect (for example a 50% decrease in mortality) means that the trial may miss a smaller, but still important effect of the treatment*

Some examples of sample size requirements:

Alpha	Beta (Type II Error)	Power	Mortality (Control Group)	Mortality (Experimental)	Sample Size Required (for each group)
.05	.20	.80	40%	20%	64
.01	.20	.80	40%	20%	105
.05	.10	.90	40%	20%	89
.05	.20	.80	20%	10%	157
.05	.20	.80	20%	15%	714
.05	.20	.80	4%	2%	899
.05	.20	.80	2%	1%	1826

- *this shows that if the outcome is uncommon (ie. mortality of 2%) you need a very large study to detect a decrease in mortality to 1%*
- *if you accept a Type II error of 20% (the risk of a false negative study) you don't need as many people as you would if you decided on a Type II error or beta of 10%*
- *if the effect size is large you don't need as many people (for example if the effect you are trying to find is a decrease in mortality from 40% to 20% you don't need a very large study), however most therapies do not have this effect size!*
- *when calculating the required sample size assumptions must be made – for example the expected mortality in the control group and a decrease in mortality in the experimental group that would be considered clinically important*

Here is what it you should look for when reading a trial. The authors have planned their sample size, the power of the study (91%), and the effect size they can detect (a 20% reduction). Alpha is set at 0.05.

“Our planned study sample size of 5000 patients was based on the assumptions of a 6% annual primary event rate in the placebo group, recruitment of patients over 18 months, and a total trial duration of 4 years. A time-to-event analysis was planned, and thus the study had 91% power to detect a 20% reduction in the hazard with a type I error of 0.05. To maintain this power, all patients had to be followed-up until at least 760 patients had one endpoint event or more.” (from a study in Lancet)

Negative studies

- if a clinical study is ‘negative’, that is it concludes that there is no difference between the groups (accepts the null hypothesis) you need to review the methods of the study to try to ensure that the study was adequately ‘powered’ – i.e. it was large enough to find a difference if there truly is a difference

