Medical Statistics

MATH 38071

Notes

(Part II)

7. The Analysis of Equivalence and Non-Inferiority Trials

7.1 Equivalence Trials

Usually, the aim of a clinical trial is to test whether a new treatment is better than the existing standard treatment or a placebo. Such trials are sometimes called *Superiority Trials* as their purpose is to test whether one treatment is superior to another.

Some trials are designed to establish that two treatments are equally effective. A new drug may have fewer side-effects, be cheaper, or be more convenient than the current standard treatment. In such circumstances one may wish to establish that the new treatment has the same ability to treat the condition. For example some pain relief medications have side effects such as causing gastric bleeding or ulceration. A replacement therapy might have fewer side effects, but one may want to check that the two drugs have the same ability to relieve pain. Studies testing this type of hypothesis are called *Equivalence Trials*.

An equivalence trial design may be relevant to testing non-drug interventions. For example a trial to test whether say nurses were as effective as a general practitioner at delivering a particular type of care might wish to demonstrate that outcomes are the same.

Hypotheses for Equivalence Trials

In a *superiority* trial the hypotheses used for testing for a treatment effect are $H_0 : \tau = 0$ vs. $H_1 : \tau \neq 0$. It is sometime suggested that failure to reject $H_0 : \tau = 0$ mean that treatments are the same. This is incorrect as a small study that is underpowered would lead to this conclusion by default, whereas a larger trial might detect a small difference. The issue can be summed up by the statement "*An Absence of Evidence is not Evidence of Absence*"**.** Use of a statistical test with the hypotheses above is therefore inappropriate for demonstrating equivalence.

This suggests the hypothesis for equivalence trials could be

$$H_0 : \tau \neq 0 \text{ vs. } H_1 : \tau = 0.$$

Unfortunately, it is never possible to show that two treatments are identical. Instead, one tests whether the treatment effect falls in range, say $\left(-\tau_E, \tau_E\right)$, called the *range of equivalence*. This should be defined by considering what range could be considered clinically equivalent. The hypotheses are now

$$H_0 : \left|\tau\right| \geq \tau_E \text{ vs. } H_1 : \left|\tau\right| < \tau_E$$

Rather than using formal significance testing, statistical analysis of equivalence trials is often based on the confidence interval of the difference between treatments. Equivalence is established by demonstrating that the confidence interval of the difference between treatment lies in the range of equivalence$\left(-\tau_E, \tau_E\right)$.

7.2 Analysis of Equivalence Trials for a Continuous Outcome Measures

Notation

Suppose n_T and n_C patients have been randomly allocated to groups *T* and *C* and suppose outcome measure *Y* is continuous and normally distributed with a mean μ_T for the new treatment and mean μ_C for the control so that the treatment effect $\tau = \mu_T - \mu_C$. Let $\bar{y}_T$ and $\bar{y}_C$ are the sample means and $\hat{\tau} = \bar{y}_T - \bar{y}_C$. As previously

$$SE\left[\hat{\tau}_C\right] = SE\left[\bar{y}_T - \bar{y}_C\right] = s\lambda$$

where $\lambda = \sqrt{1/n_T + 1/n_C}$ and the pooled sample standard deviation *s* is estimated by

$$s = \sqrt{\frac{\left(n_T - 1\right)s_T^2 + \left(n_C - 1\right)s_C^2}{n_T + n_C - 2}}$$

with s_T and s_C being the sample standard deviations for the two treatment groups. $t_\alpha(\nu)$ is the value of the t-distribution with $\nu = n_T + n_C - 2$ degrees of freedom having a cumulative probability equal $(1-\alpha)$.

Rejection of the null hypothesis that $H_0 : |\tau| \geq \tau_E$ vs. $H_1 : |\tau| < \tau_E$, when the $(1-2\alpha)$ confidence interval $\left(\hat{\tau} - t_\alpha(\nu)SE\left[\hat{\tau}\right], \hat{\tau} + t_\alpha(\nu)SE\left[\hat{\tau}\right]\right)$ is within the interval $(-\tau_E, \tau_E)$, has a Type I error less than α.

Proof

To determine the Type 1 error we needs to estimate $\Pr\left[\text{Reject } H_0\right]$ under H_0. This probability depends on the value of τ. Since H_0 is a range of values of τ, $\Pr\left[\text{Reject } H_0\right]$ will take a range of values also.

$$\Pr\left[\text{Reject } H_0 \middle| \tau\right] = \Pr\left[\left(\hat{\tau} - t_\alpha(\nu) SE\left[\hat{\tau}\right], \hat{\tau} + t_\alpha(\nu) SE\left[\hat{\tau}\right]\right) \subseteq \left(-\tau_E, \tau_E\right)\right]$$

The distribution of $\hat{\tau}/SE\left[\hat{\tau}\right]$ has a non-central t-distribution, if $\tau \neq 0$.

To simplify the proof we will assume the variance is known, and equal to say σ, so $\hat{\tau}$ has a distribution $N\left[\tau, \sigma^2\lambda^2\right]$, so that we can replace $t_\alpha(\nu)$ with z_α. Hence

$$\begin{aligned}
\Pr\left[\text{Reject } H_0 \middle| \tau\right] &\cong \Pr\left[\left(\hat{\tau} - z_\alpha\sigma\lambda, \hat{\tau} + z_\alpha\sigma\lambda\right) \subseteq \left(-\tau_E, \tau_E\right)\right] \\
&= \Pr\left[\left(\hat{\tau} - z_\alpha\sigma\lambda > -\tau_E\right) \bigcap \left(\hat{\tau} + z_\alpha\sigma\lambda < \tau_E\right)\right] \\
&= \Pr\left[\left(\hat{\tau} > -\tau_E + z_\alpha\sigma\lambda\right) \bigcap \left(\hat{\tau} < \tau_E - z_\alpha\sigma\lambda\right)\right]
\end{aligned}$$

If $\tau_E \leq z_\alpha\sigma\lambda$, $\left(\hat{\tau} > -\tau_E + z_\alpha\sigma\lambda\right) \bigcap \left(\hat{\tau} < \tau_E - z_\alpha\sigma\lambda\right)$ is the null set, hence $\Pr\left[\text{Reject } H_0 \middle| \tau\right] = 0$.

If $\tau_E > z_\alpha\sigma\lambda$,

$$\begin{aligned}
\Pr\left[\text{Reject } H_0 \middle| \tau\right] &= \Pr\left[\hat{\tau} < \tau_E - z_\alpha\sigma\lambda\right] - \Pr\left[\hat{\tau} < -\tau_E + z_\alpha\sigma\lambda\right] \\
&= \Phi\left(\frac{\tau_E - z_\alpha\sigma\lambda - \tau}{\sigma\lambda}\right) - \Phi\left(\frac{-\tau_E + z_\alpha\sigma\lambda - \tau}{\sigma\lambda}\right)
\end{aligned}$$

The next step is to find τ that maximizes $\Pr\left[\text{Reject } H_0 \middle| \tau\right]$ under H_0.

Differentiation with respect to τ gives

$$\frac{d}{d\tau}\Pr\left[\text{Reject } H_0\right] = -\frac{1}{\sigma\lambda}\phi\left(\frac{\tau_E - \sigma\lambda z_\alpha - \tau}{\sigma\lambda}\right) + \frac{1}{\sigma\lambda}\phi\left(\frac{-\tau_E + \sigma\lambda z_\alpha - \tau}{\sigma\lambda}\right)$$

where ϕ is the density $N\left[0,1\right]$.

Since $\phi(z)=\phi(-z)$, it follows that

$$\phi\left(\frac{-\tau_E+\sigma\lambda z_\alpha-\tau}{\sigma\lambda}\right)=\phi\left(\frac{\tau_E-\sigma\lambda z_\alpha+\tau}{\sigma\lambda}\right).$$

Hence

$$\frac{d}{d\tau}\Pr\left[\text{Reject } H_0\right]=-\frac{1}{\sigma\lambda}\phi\left(\frac{\tau_E-\sigma\lambda z_\alpha-\tau}{\sigma\lambda}\right)+\frac{1}{\sigma\lambda}\phi\left(\frac{\tau_E-\sigma\lambda z_\alpha+\tau}{\sigma\lambda}\right),$$

which is zero when $\phi\left(\frac{\tau_E-\sigma\lambda z_\alpha-\tau}{\sigma\lambda}\right)=\phi\left(\frac{\tau_E-\sigma\lambda z_\alpha+\tau}{\sigma\lambda}\right)$.

Since $\tau_E>\sigma\lambda z_\alpha$, it follows that $\tau=0$

Since $\Pr\left[\text{Reject } H_0\right]$tend to zero as τ tends to $\pm\infty$, $\tau=0$ must be maximum.

Hence $\Pr\left[\text{Reject } H_0\right]$is monotone increasing for $\tau<0$ and monotone decreasing for $\tau>0$. Maximum of the Type 1 error are therefore the boundary values $\tau=-\tau_E$ and $\tau=\tau_E$.

When $\tau=\tau_E$,

$$\Pr\left[\text{Reject } H_0\right]=\Phi\left(\frac{\tau_E-\sigma\lambda z_\alpha-\tau_E}{\sigma\lambda}\right)-\Phi\left(\frac{-\tau_E+\sigma\lambda z_\alpha-\tau_E}{\sigma\lambda}\right)$$

$$=\Phi(-z_\alpha)-\Phi(-2\tau_E/\sigma\lambda+z_\alpha)=\alpha-\Phi(z_\alpha-2\tau_E/\sigma\lambda)<\alpha.$$

Similarly, when $\tau=-\tau_E$, $\Pr\left[\text{Reject } H_0\right]<\alpha$.

Therefore under the null hypothesis $H_0:|\tau|\geq\tau_E$, $\Pr\left[\text{Reject } H_0\right]<\alpha$ as required ■

Ex 7.1. An equivalence trial is carried out to test the pain relief of a new medication thought to have fewer side effects than the current standard treatment. It was felt that mean pain score should not differ by more than 5 to demonstrate equivalent pain relief with higher pain scores representing greater pain. Fifty subjects were randomized to receive each treatment. Mean pain for the standard treatment is 45.1 (s.d.=20.6) and 46.3 (s.d.=19.4) for the new treatment. Test the null hypothesis of non-equivalence using a 5% significance level.

By default most computer packages give a 95% confidence intervals, but they can give confidence intervals with different levels of coverage if specified.

Figure 7.1 Stata output for a two sample t-test

(i) 95% confidence interval

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      50        46.3     2.91328        20.6    40.44554    52.15446
       y |      50        45.1    2.743574        19.4    39.58658    50.61342
---------+--------------------------------------------------------------------
    diff |                 1.2      4.0018               -6.741441    9.141441
------------------------------------------------------------------------------
    diff = mean(x) - mean(y)                                      t =   0.2999
Ho: diff = 0                                     degrees of freedom =       98

     Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.6175         Pr(|T| > |t|) = 0.7649          Pr(T > t) = 0.3825
```

(ii) 90% confidence interval

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [90% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      50        46.3     2.91328        20.6    41.41574    51.18426
       y |      50        45.1    2.743574        19.4    40.50026    49.69974
---------+--------------------------------------------------------------------
    diff |                 1.2      4.0018               -5.445193    7.845193
------------------------------------------------------------------------------
    diff = mean(x) - mean(y)                                      t =   0.2999
Ho: diff = 0                                     degrees of freedom =       98

     Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.6175         Pr(|T| > |t|) = 0.7649          Pr(T > t) = 0.3825
```

7.3 Sample Size for Equivalence Trials

Consider a continuous and normally distributed outcome measure *Y* with means μ_T and μ_C for the new and control treatment, and suppose the range of equivalence is $(-\tau_E, \tau_E)$. Assuming $\tau = 0$ under the alternative hypothesis $H_1 : |\tau| < \tau_E$, the sample size required to reject $H_0 : |\tau| \geq \tau_E$ using a $(1-2\alpha)$ confidence interval with power $(1-\beta)$ is $n = \frac{2\sigma^2}{\tau_E^2}\left(z_\alpha + z_{\beta/2}\right)^2$ per treatment group assuming two equal size groups.

This formula assumes that sample size is sufficient such that the normal approximation for the t-distribution is valid.

The power is $\Pr[\text{Reject } H_0]$ under the alternative hypothesis H_1.

From the derivation above

$$\Pr[\text{Reject } H_0] = \Phi\left(\frac{(\tau_E - \sigma\lambda z_\alpha) - \tau}{\sigma\lambda}\right) - \Phi\left(\frac{(-\tau_E + \sigma\lambda z_\alpha) - \tau}{\sigma\lambda}\right)$$

Since $\tau = 0$ is assumed for the alternative hypothesis

$$Power = 1-\beta = \Phi\left(\tau_E/\sigma\lambda - z_\alpha\right) - \Phi\left(-\tau_E/\sigma\lambda + z_\alpha\right)$$

Since $\Phi(x) = 1 - \Phi(-x)$, it follows that the second term

$$\Phi\left(-\tau_E/\sigma\lambda + z_\alpha\right) = 1 - \Phi\left(\tau_E/\sigma\lambda - z_\alpha\right).$$

Hence

$$1-\beta = 2\Phi\left(\tau_E/\sigma\lambda - z_\alpha\right) - 1.$$

Rearrangement gives

$$1-\beta/2=\Phi\left(\tau_E/\sigma\lambda-z_\alpha\right).$$

Since $\Phi^{-1}\left(1-\beta/2\right)=z_{\beta/2}$, it follows that $z_{\beta/2}=\dfrac{\tau_E}{\sigma\lambda}-z_\alpha$.

Hence $\dfrac{\tau_E}{\sigma\lambda}=z_\alpha+z_{\beta/2}$

Assuming equal sample size $n_T=n_C=n$, then $\lambda=\sqrt{\dfrac{2}{n}}$.

Therefore $\sqrt{\dfrac{n}{2}}=\dfrac{\sigma}{\tau_E}\left(z_\alpha+z_{\beta/2}\right)$.

Squaring and rearrangement gives $n=\dfrac{2\sigma^2}{\tau_E^2}\left(z_\alpha+z_{\beta/2}\right)^2$ as required■

Ex 7.2 Pain relief trial continued. From the data above we have an estimate of $\sigma\cong 20$. Taking the range of equivalence as (-5,5) that is τ_E =5, estimate the sample size per arm required to have 80% power to reject the null hypothesis of non-equivalence using a 90% confidence interval (i.e. 5% level test) assuming two equal size groups.

7.4 Non-Inferiority Trials

In the trial comparing conservative treatment with suturing for small lacerations of the hand (Critical Appraisal 2), the objective was to establish that conservative management gave as good aesthetic outcome after 3 months. If conservative treatment gave a better outcome, one would not be concerned and would want to reject the null hypothesis as the main concern is check that conservative treatment did no worse.

If one is only concerned to demonstrate that a new treatment is as good or better, rather than equivalent to an existing treatment, only one bound is needed. This design is referred to as *non-inferiority* trials. Analysis of a non-inferiority trial can be based on a single-sided confidence interval. Note a (1-α) single-sided confidence interval is define by one or other limits of the usual two sided confidence interval, but with coverage (1-2α). So the upper and lower 95% single-sided confidence intervals are the upper and lower limits of a 90% confidence interval.

The hypotheses for a non-inferiority trial are therefore

$$H_0 : \tau \leq -\tau_N \text{ vs. } H_1 : \tau > -\tau_N$$

if τ>0 represents benefit to the patient or

$$H_0 : \tau \geq \tau_N \text{ vs. } H_1 : \tau < \tau_N$$

if τ < 0 represents benefit.

Suppose Y is continuous and normally distributed outcome measure with a mean μ_T for the new treatment and mean μ_C for the control treatment so the treatment effect $\tau = \mu_T - \mu_C$. Suppose also that $\bar{y}_T$ and $\bar{y}_C$ are the sample means and $\hat{\tau} = \bar{y}_T - \bar{y}_C$.

(i) (Higher score for a better outcome) Rejection of the null hypothesis $H_0 : \tau \le -\tau_N$ vs. $H_1 : \tau > -\tau_N$ if the $(1-\alpha)$%single-sided confidence interval, given by $\hat{\tau} - t_\alpha(\nu) SE[\hat{\tau}]$, is greater than or equal to $-\tau_N$ will have a *Type 1* error ≤ *α*.

(ii) (Lower score for a better outcome) Rejection of the null hypothesis $H_0 : \tau \ge \tau_N$ vs. $H_1 : \tau < \tau_N$, if the $(1-\alpha)$% single-sided confidence interval, given by $\hat{\tau} + t_\alpha(\nu) SE[\hat{\tau}]$, is less than or equal to τ_N will have a *Type 1* error ≤ *α*.

(i) Higher score for a better outcome

Assuming a normal approximation to the t-distribution and a known standard deviation, σ, the $(1-\alpha)$% single sided lower confidence interval for $\hat{\tau}$ is given by $\hat{\tau} - z_\alpha \sigma\lambda$. H_0 will be rejected provided $\hat{\tau} - z_\alpha \sigma\lambda > -\tau_N$.

Therefore

$$\Pr\left[\text{Reject } H_0 \middle| \tau\right] = \Pr\left[\hat{\tau} - z_\alpha \sigma\lambda > -\tau_N\right] = \Pr\left[\hat{\tau} > -\tau_N + \sigma\lambda z_\alpha\right].$$

Since $\hat{\tau}$ is $N\left[\tau, \sigma^2\lambda^2\right]$,

$$\Pr\left[\text{Reject } H_0 \middle| \tau\right] = 1 - \Phi\left(\frac{\left(-\tau_N + \sigma\lambda z_\alpha\right) - \tau}{\sigma\lambda}\right) = \Phi\left(\frac{\tau_N - \sigma\lambda z_\alpha + \tau}{\sigma\lambda}\right)$$

The maximum of this can be obtained by differentiation w.r.t. τ . The derivative is

$$\frac{d}{d\tau}\Pr\left[\text{Reject } H_0 \middle| \tau\right] = \frac{1}{\sigma\lambda}\phi\left(\frac{\tau_N - \sigma\lambda z_\alpha + \tau}{\sigma\lambda}\right) \quad (*)$$

where ϕ is the standard normal density function.

Since ϕ >0 for finite values, it follows that $\frac{d}{d\tau}\Pr\left[\text{Reject } H_0 \middle| \tau\right]$is positive and so $\Pr\left[\text{Reject } H_0 \middle| \tau\right]$is monotone increasing with τ.

Hence, the type 1 error rate has a maximum when $\tau = -\tau_N$.

Setting $\tau = -\tau_N$, in (*)

$$\Pr\left[\text{Reject } H_0 \middle| \tau\right] = \Phi\left(\frac{\tau_N - \sigma\lambda z_\alpha - \tau_N}{\sigma\lambda}\right) = \Phi\left(-z_\alpha\right) = \alpha$$

Hence, the type 1 error must be less than or equal to α ■

Result (ii) is left as an exercise.

Ex 7.3 Pain relief example continued. Assuming that higher pain scores represent more pain we would require the upper confidence interval to be less than 5. Are we able to show non-inferiority?

7.5 Sample Size for Parallel Group Non-Inferiority Trials

Suppose higher scores represent a better outcome for the patient.

For a continuous and normally distributed with means μ_T and μ_C for the new treatment and control groups. If one considers treatment to be non-inferior provided the (1-α) % one-sided confidence interval for $\hat{\tau} = \bar{y}_T - \bar{y}_C$ is greater than $-\tau_N$, the sample size required to demonstrate non-inferiority with a power (1-β) is

$$n = \frac{2\sigma^2}{\tau_N^2}\left(z_\alpha + z_\beta\right)^2$$

per treatment group assuming $\tau = 0$ under the alternative hypothesis.

The derivation of the sample size formula for non-inferiority trials is similar to that for equivalence trials. Again the derivation will assume that sample size is sufficient for a normal approximation to the t-distribution is reasonable so that $\hat{\tau} \sim N\left[\tau, \sigma^2\lambda^2\right]$.

With higher score being a better outcome, one is testing $H_0 : \tau \le -\tau_N$ vs. $H_1 : \tau > -\tau_N$. From above

$$\Pr\left[\text{Reject } \mathrm{H}_0 \middle| \tau\right] = \Phi\left(\frac{\tau_N - \sigma\lambda z_\alpha + \tau}{\sigma\lambda}\right)$$

Under the alternative hypothesis, $\tau = 0$. Therefore, the power

$$1 - \beta = \Phi\left(\frac{\tau_N - z_\alpha\sigma\lambda}{\sigma\lambda}\right)$$

Since $\Phi^{-1}\left(1-\beta\right) = z_\beta$, by taking inverses $z_\beta = \frac{\tau_N}{\sigma\lambda} - z_\alpha$.

Hence $\frac{\tau_N}{\sigma\lambda} = z_\alpha + z_\beta$.

Assuming equal sample sizes $n_T = n_C = n$ then $\lambda = \sqrt{\frac{2}{n}}$.

Substitution gives the $\sqrt{\frac{n}{2}} = \frac{\sigma}{\tau_N}\left(z_\alpha + z_\beta\right)$ giving

$n = \frac{2\sigma^2}{\tau_N^2}\left(z_\alpha + z_\beta\right)^2$ as required ■

The derivation assuming lower scores are better is left as an exercise.

Ex 7.4 Pain relief trial continued. Calculate the sample size required to give 80% power to reject the null hypothesis of non-inferiority using a single-sided 95% confidence interval.

7.6 Limitations of Equivalence and Non-inferiority Trials

One problem with equivalence and non-inferiority trials is that poor design and sloppy implementation reduce the differences between treatment groups biasing the study towards the alternative hypothesis of equivalence. It is important therefore that patients adhere to their treatment in this type of trial. We will return to this point when we consider intention to treat analyses in the next section.

Equivalence trials show that two treatments may give the same average outcome but difference patients may benefit from a particular treatment. Equivalence trials do not demonstrate bio-equivalence, that is patients have the same outcome which ever treatment they receive.

Comparison of Sample Size Formulae for Parallel Group Trials and a Continuous Outcome

Sample size per group for a continuous outcome measure for difference hypotheses:

Superiority $$n = \frac{2\sigma^2}{\tau^2}\left(z_{\alpha/2} + z_{\beta}\right)^2$$ – two sided

$$n = \frac{2\sigma^2}{\tau^2}\left(z_{\alpha} + z_{\beta}\right)^2$$ – one sided

Equivalence $$n = \frac{2\sigma^2}{\tau_E^2}\left(z_{\alpha} + z_{\beta/2}\right)^2$$

Non-inferiority $$n = \frac{2\sigma^2}{\tau_N^2}\left(z_{\alpha} + z_{\beta}\right)^2$$

8. Analysis with Treatment Protocol Deviations

Sometime patients in a randomised controlled trial do not receive the treatment allocated. After consenting they or their care provider may change their mind, perhaps due to the change in the patient's health. Patients may decide not to take the tablets. A patient may start a treatment but then default or change to another before receiving an adequate dose. In these situations the patient may be said to be non-compliant or non-adherent. These changes from the randomly allocated treatment are sometimes referred to as *Treatment Protocol Deviations.*

If patients do not adhere to their randomly allocated treatment, should they be included in the statistical analyses, and if so how?

Analysis Strategies Where There is Non-Compliance

- ***Intention-To-Treat analysis (ITT):*** Patients are analysed according to the group to which they were randomized, irrespective of whether they received the intended intervention. Also called ***As-Randomized***
- ***Per-Protocol (PP):*** Patients are analysed within the intervention group to which they were randomized after exclusion of non-compliant patients.
- ***As-Treated (AT):*** Patients are analysed according to the treatment they actually received irrespective of the random allocation.

Table 8.1 Comparison of *intention-to-treat* with *as treated* and *per-protocol* analysis

Survival at 2 years	Randomization			
	Medical treatment		Surgery	
	Received medicine	Switch to surgery	Received Surgery	Switched to medicine
	[1]	[2]	[3]	[4]
Died	27	2	15	6
Alive	296	48	354	20
Total	323	50	369	26

Two-year mortality in the coronary bypass surgery trial published by the European Coronary Study Group (1979) from E Marubini M.G. Valsecchi Analysing survival Data from Clinical Trials and Observational Studies, p22 Wiley 1996.

Table 8.2 Summary of Mortality Rates for each Analysis Method

Analysis	Medical P_M	Surgical P_S	Treat. Effect $P_M - P_S$
Intention-to-treat			
Per-protocol			
As-treated			

Table 8.3 Summary of Inferential Analysis using a z-test for Proportions

Analysis	**Treat. Effect P_M- P_S**	**95% c.i.**	**p-value**	**Signif. In a 5% level test**
Intention-to-treat	2.45%	-1.05% to 5.96%	0.168	
Per-protocol	4.29%	0.66% to 7.92%	0.018	
As-treated	5.40%	1.79% to 9.00%	0.003	

In this trial the patients that changed from medical treatment to surgery appear to be different from those patients who changed from surgery to medical treatment. Only 2/50 (4%) of those that switched from medicine to surgery died, whereas 6/26 (23%) of those that switched from surgery to medicine died, a difference in mortality of 19%. This suggests that the prognosis of these two patient sub-groups were very different.

8.1 Comparison of ITT, PP And AT Analyses

When testing $H_0: \tau = 0$ vs $H_1: \tau \neq 0$, where there is non-compliance the *Intention-to-Treat* estimate $\hat{\tau}_{ITT}$ is biased toward H_o whereas the *Per-Protocol* estimate $\hat{\tau}_{PP}$ and the *As-Treated* estimate $\hat{\tau}_{AT}$ may be biased either towards or away from H_o.

A simple mathematical model can be constructed to illustrate the difference between the three estimates of treatment effect. We suppose that the patient population can be divided into three sub-groups as follows:

Group A - who comply with the allocated treatment (Compliers – Always as randomised)
Group B - who will always receive control treatment regardless of allocation (Always Control Treatment)
Group C- who will receive the new treatment regardless of allocation. (Always New Treatment)
It is assumed that there are no **defyers**, that is patients who will always receive the opposite of the treatment to which they are randomly allocated.

As patients enter the trial the sub-group membership of a patient is not known or “latent”. Patients in each of the three compliance sub-groups or “latent classes” are likely to have a different prognosis. Considering example 7.1 if surgery was the New Treatment and medical was the Control, group B is patient that would always receive

medical treatment. We saw that those patients appear to have a worse prognosis.

As the patients are randomly allocated the expected proportions of patients in each of the three latent classes is the same in each arm of the trial. For simplicity of presentation it will be assumed that the treatment effect compared to the control treatment is τ, in all three "latent classes". The quantity τ is the causal effect of treatment, sometimes called by the *Compliance Average Causal Effect (CACE)*, which is the average treatment effect in patients that comply with the New treatment.

Table 8.4 Model of expected mean outcome for each treatment and latent sub-group

	Latent Class	**Control Treatment Group**	**New Treatment Group**	**Probability in Latent Class**
As Randomized	**A**	μ	$\mu + \tau$	θ_A
Always Control	**B**	$\mu + \gamma_B$	$\mu + \gamma_B$	θ_B
Always New Treatment	**C**	$\mu + \gamma_C + \tau$	$\mu + \gamma_C + \tau$	θ_C $(=1 - \theta_A - \theta_B)$

Assumptions of Model

- No *defyers*, that is patients who always receive the opposite of the allocated treatment.
- Proportion and characteristics of the three latent classes *compliers*, *always control*, *always new treatment* is the same in both arms. This is justified by randomization.
- Randomization only effects outcome through treatment.

From the table of expected means the *Intention-to-Treat* estimate is

$$\tau_{ITT} = \left[\theta_A(\mu+\tau)+\theta_B(\mu+\gamma_B)+\theta_C(\mu+\gamma_C+\tau)\right] - \left[\theta_A\mu+\theta_B(\mu+\gamma_B)+\theta_C(\mu+\gamma_C+\tau)\right] = \theta_A\tau$$

as second and third terms in each bracket cancel.

Hence $|\tau_{ITT}| \le \tau$ which means τ_{ITT} is biased towards zero if θ_A <1 i.e. if some patients do not comply with treatment. Hence $E\left[\hat{\tau}_{ITT}\right] \le \tau$

The *Per-Protocol* estimate is

$$\begin{aligned}
\tau_{pp} &= \left[\frac{\theta_A(\mu+\tau)+\theta_C(\mu+\gamma_C+\tau)}{\theta_A+\theta_C}\right] - \left[\frac{\theta_A\mu+\theta_B(\mu+\gamma_B)}{\theta_A+\theta_B}\right] \\
&= \left[\frac{(\theta_A+\theta_C)\mu+\theta_C\gamma_C+(\theta_A+\theta_C)\tau}{\theta_A+\theta_C}\right] - \left[\frac{(\theta_A+\theta_B)\mu+\theta_B\gamma_B}{\theta_A+\theta_B}\right] \\
&= \tau+\mu+\left[\frac{\theta_C\gamma_C}{\theta_A+\theta_C}\right]-\mu-\left[\frac{\theta_B\gamma_B}{\theta_A+\theta_B}\right] \\
&= \tau+\left[\frac{\theta_C\gamma_C}{1-\theta_B-\theta_C+\theta_C}\right]-\left[\frac{\theta_B\gamma_B}{1-\theta_B-\theta_C+\theta_B}\right] \\
&= \tau+\left[\frac{\theta_C\gamma_C}{1-\theta_B}\right]-\left[\frac{\theta_B\gamma_B}{1-\theta_C}\right]
\end{aligned}$$

τ_{PP} is biased by terms involving γ_B and γ_C. Since γ_B and γ_C can be either positive or negative $\hat{\tau}_{PP}$ may be biased either towards or away from zero.

A similar expression can be derived for the As-Treated estimate that also shows that it can also be biased towards or away from zero depending on the magnitude of γ_B and γ_C.

Advantages of Intention-to-Treat

The Intention-to-Treat analysis is always biased towards zero so that the efficacy of the treatment is being under-estimated. In a superiority trial, use of *intention-to-treat* biases the statistical analysis towards the null hypothesis. If one rejects the null hypothesis H_o: $\tau = 0$ based on an *intention-to-treat* analysis, one can feel confident that the treatment effect is larger in patients that actually take the treatment. An analysis based on *intention-to-treat* is therefore said to be conservative. This is not true for *per-protocol* and *as-treated* analyses as both can be biased either towards or away from the null hypothesis.

Another advantage of intention-to treat analysis is that randomization clearly defines the groups being compared so there is no ambiguity as to how the patients should be included in the analysis. In contrast, the groups being compared in per-protocol or as-treated analyses may be less well defined. Whether a particular patient completes treatment is often difficult to obtain. Even if one is able to collect reliable data on the treatment, the researchers needs to agree how many tablets or therapy sessions a patient has to receive before they can be considered to have complied with treatment, which is an issue for which there may be no consensus. For this reason an ITT analysis may therefore be easier to implement than Per-protocol or As-treatment analyses.

It is important that all patients are followed-up, not just those that receive treatment, for ITT analysis to be carried out.

8.2 Efficacy and Effectiveness

Efficacy and *Effectiveness* are two terms used to describe the ability to produce an effect such cure a specific illness. In clinical trials a distinction is drawn between efficacy (also known as ideal use) and effectiveness (also known as typical use). We have already seen that where there is non-compliance, Intention-to-treat underestimates the efficacy of a treatment.

Intention to Treat Analyses and Effectiveness

Researchers may not be just interested in whether treatment works in patients who receive a treatment. They may want to know the overall effect of offering a treatment. This is particularly true for health policy makers. For example in a trial of exercise for the treatment of back-pain some patients may not comply. If only a small proportion of patient take the treatment, the average benefit of offering the treatment may be small, even if it is beneficial in patients that comply. It may be important to know the effectiveness, which is the effect taking account of non-compliance, as there are likely to be “costs” associated with offering the treatment to patients that do not comply. It can be argued that the intention-to-treat (ITT) analysis gives an estimate of treatment effect taking account of non-compliance. For this reason ITT is sometimes said to give an estimate of the effectiveness of treatments. This interpretation of ITT assumes that the proportion of patients that comply in the trial is the same as in normal care, which may not be true.

8.3 Estimating Efficacy and the CACE estimate

Suppose instead the researcher is interested in efficacy. Provided the assumptions below table 7.5 hold, the compliance average causal effect (CACE) estimate can be obtained.

From above the ITT estimate $\tau_{ITT} = \tau\theta_A$. Hence, the *Compliance Average Causal Effect* is

$$\tau = \tau_{ITT} / \theta_A .$$

$\hat{\tau}_{ITT} = \bar{y}_T - \bar{y}_C$ for continuous data and

$\hat{\tau}_{ITT} = p_T - p_C$ for binary data.

One needs an estimator of θ_A , the proportion of patients who comply with randomization. This can be obtained as follows:

- Suppose the observed proportions who receive the new treatment in the treatment and control groups are respectively q_T and q_C.
- Considering the control group, $q_c = \hat{\theta}_C$
- Considering the new treatment group, $q_T = \hat{\theta}_A + \hat{\theta}_C$
- Hence θ_A can be estimated by $\hat{\theta}_A = q_T - q_C$.

Hence the *Compliance Average Causal Effect* can be estimated by

$$\hat{\tau} = \frac{\hat{\tau}_{ITT}}{q_T - q_C}.$$

It should be noted that this estimate assumes that there are only two treatments that patients can switch between. This method does not work where one is comparing two active treatments and some patients default to a third option such as no treatment.

Ex 8.2 For the bypass surgery example above the Intention-to-Treat estimate of the treatment was 2.45% , 4.29% for a Per-Protocol analysis and 5.40% for an As-treated analysis. Estimate the Compliance Average Causal Effect , τ.

$q_T =$

$q_C =$

Hence $\hat{\tau} =$

The causal effect of treatment is 3.1% , which is smaller than the Per-Protocol (4.3%) and As-Treat estimates (5.4%). Under the assumption we have made, one can see that the Per-Protocol and As-Treat estimates are both biased away from the null.

The test that the compliance average causal effect (CACE) is zero is equivalence to the test that the intention to treat effect (ITT) is zero, that is $H_0 : \tau = 0$ is equivalent to $H_0 : \tau_{ITT} = 0$.

Intention-to-Treat and Equivalence and Non-inferiority Trials

Application of the intention-to-treat (ITT) analysis in an equivalence trial has problems, as it is biased towards the alternative hypothesis of no difference between treatments. An ITT analysis may therefore increase the probability of accepting the alternative hypothesis. Good compliance with treatment is therefore very important in both equivalence and non-inferiority trials.

9. Crossover Trials

9.1 Motivation for Crossover Trials

When we first consider statistical methods for clinical trial the notion of potential outcome was introduced (See section 2.1) A patient had two *potential outcomes*, say $Y_i(T)$ and $Y_i(C)$. For the i^{th} patient the treatment effect was defined as $\tau_i = Y_i(T) - Y_i(C)$. The expected treatment effect is therefore, $\tau = E\left[\tau_i\right] = E\left[Y_i(T) - Y_i(C)\right]$. For many conditions a single course of treatment may cure some if not all patients, so they would no longer be eligible for the comparator treatments. This applies to most treatments for acute (short-term) conditions such as antibiotic treatment for an infection, trauma, or surgery. For these conditions, it is only possible to measure one of the two potential outcomes in the same patient - one outcome is said to be counter factual. As previously discussed, the expected treatment effect is therefore estimated as the difference in average outcome between patients receiving the intervention (T) and those receiving the control treatment (C). The treatment effect is estimated between groups and so the precision of the estimate of the average treatment effects depends on in the variance of the outcome $Var\left[Y_i\right]$.

In some for chronic (long-term) diseases such as arthritis, asthma, diabetes, or high blood pressure, the condition is not cured by treatment. Instead symptoms may be reduced or the disease progression slowed by continuing treatment. It may be possible to measure both potential outcomes in the same patient, and so the

treatment effect τ_i can be estimated in each patient. The average treatment effect could then be estimated with greater precision for a given sample size of patients. Trial size would then be reduced making clinical trials easier to conduct.

Suppose one is able to compare two treatments, say A and B, in the same patient. One option would be to give treatment A followed by treatment B. Such a design is potentially biased, because the patient's condition may deteriorate or improve over time irrespective of treatment. To overcome this, patients should be randomly allocated to receive treatments in either order. This trial design is the called the AB-BA Crossover Design. In such a design:

- The trial is divided into two *periods* with one treatment given during each.
- Patients are randomly allocated to two groups, one receiving treatment sequence *A then B* and the other receiving *B then A*.

Random allocation is important for two reasons:

- To prevent bias due to possible change in the patient over time.
- To maintain concealment prior to treatment allocation and post treatment allocation where the trial is double blind.

Use of a Crossover design

A crossover trial designs may be suitable where:

- The condition being treated is a chronic disease (e.g. chronic diabetes).
- The condition is stable and so unlikely to change greatly from one period to the next.
- The intervention has a rapid effect.

A crossover designs are not suitable where:

- The conditions may resolve quickly, that is an acute condition, making the second period treatment unnecessary. (e.g. infections, trauma, and rehabilitation).
- Patients are likely to withdraw from treatment or be lost to follow-up.
- The effect of the first treatment could plausibly contaminate the effect of second. Note for some treatments contamination can be prevented by having a “wash-out” period between treatments.

9.2 Analysis of an *AB-BA* Crossover Design

Suppose patients are randomly allocated to either treatment *A*, followed by treatment B (Sequence *AB*) or treatment *B* then treatment *A* (Sequence *BA*). Suppose that there are n_{AB} and n_{BA} patients in each sequence with total sample size $N = n_{AB} + n_{BA}$. Even though we have stated that patients should be stable, a statistical model for a crossover design needs to include parameters called a period effect, which is the difference between period 2 and period 1 irrespective of treatment order as patients health will change over time. Two sources of variation can be identified, variation *between-patients*, and variation *within-patient*.

If Y_{ij} is the response for the i^{th} subject during period $(j = 1, 2)$, a model for an *AB-BA* crossover trial can be define as follows:

$y_{ij} = \mu + \xi_i + \varepsilon_{ij}$ Sequence *AB* Period 1

$y_{ij} = \mu + \tau + \phi + \xi_i + \varepsilon_{ij}$ Sequence *AB* Period 2

$y_{ij} = \mu + \tau + \xi_i + \varepsilon_{ij}$ Sequence *BA* Period 1

$y_{ij} = \mu + \phi + \xi_i + \varepsilon_{ij}$ Sequence *BA* Period 2

μ = mean in period 1 for the Sequence *AB.*

τ = treatment effects of *B* compared to *A*.

ϕ = period effect.

ξ_i = random variable for patient i with $E[\xi_i] = 0$ and variance σ_B^2.

ε_{ij} = random variable for patient i in period j with $E[\varepsilon_{ij}] = 0$ and variance σ_ε^2 assumed to be normally distributed, $N[0, \sigma_\varepsilon^2]$.

σ_B^2 is called the between-patient variance and σ_ε^2 is called the within-patient variance.

Using a Single Sample t-Test to Analyse Crossover Trials

Sometime crossover trials are analysed by using a single sample t-test applied to the difference between the two treatments, also called the paired t-test.

Define $c_i = y_{i2} - y_{i1}$ for $i \in AB$ and $c_i = y_{i1} - y_{i2}$ for $i \in BA$, which are the difference in outcome between when a patient receives treatment B and treatment A. The treatment effect is then estimated by $\hat{\tau}_C = \bar{c} = \frac{\sum c_i}{N}$ with $N = n_{AB} + n_{BA}$. The hypothesis $H_0 : \tau = 0$ is then tested using $T_C = \frac{\bar{c}}{SE[\bar{c}]}$, where $SE[\bar{c}] = \sqrt{\frac{s_C^2}{N}}$ with s_C^2 being the sample standard deviation of the differences, c_i. If c_i can be assumed to be normally distributed, the test statistic T_C has a t-distribution with $\nu = N - 1$ degrees of freedom.

Unfortunately, $\hat{\tau}_C$ can give a biased estimate of τ.

Suppose the data generation model for an AB-BA crossover trial defined above applies. For the treatment effect estimator $\hat{\tau}_C = \bar{c}$

$E\left[\hat{\tau}_C\right] = \tau + \dfrac{\left(n_{AB} - n_{BA}\right)\phi}{N}$ where $N = n_{AB} + n_{BA}$. If the period effect $\phi \neq 0$ and $n_{AB} \neq n_{BA}$, $\hat{\tau}_C$ will be a biased estimator of the treatment effect, τ.

Substitution from the model above gives

$$E\left[c_i \middle| i \in AB\right] = E\left[\left(\mu + \tau + \phi + \xi_i + \varepsilon_{ij}\right) - \left(\mu + \xi_i + \varepsilon_{ij}\right)\right].$$

$$= E\left[\left(\tau + \phi + \varepsilon_{ij}\right) - \left(\varepsilon_{ij}\right)\right].$$

Since $E\left[\varepsilon_{ij}\right] = 0$, $E\left[c_i \middle| i \in AB\right] = \tau + \phi$.

Similarly, $E\left[c_i \middle| i \in BA\right] = E\left[\left(\tau + \varepsilon_{i1}\right) - \left(\phi + \varepsilon_{i2}\right)\right] = \tau - \phi$.

Now $E\left[\hat{\tau}_C\right] = E\left[\bar{c}\right] = E\left[\dfrac{\sum c_i}{N}\right] = \dfrac{\sum E\left[c_i\right]}{N}$.

Hence,

$$E\left[\hat{\tau}_C\right] = \frac{n_{AB}\left(\tau + \phi\right) + n_{BA}\left(\tau - \phi\right)}{N} = \tau + \frac{\left(n_{AB} - n_{BA}\right)\phi}{N} \bullet$$

It is rarely possible to rule-out a period effect $\left(\phi\right)$ completely. Even where n_{AB} and n_{AB} are planned to be equal, the may differ due to imbalance arising from randomisation, such as incomplete blocks in block randomisation, or patients dropout from the trial. The treatment effect estimate $\hat{\tau}_C$ may therefore be biased and is therefore not recommended.

An Unbiased Estimator of the Treatment Effect

Notation

For subject *i* , define $d_i = y_{i2} - y_{i1}$, that is the difference between period 2 and period 1 irrespective of treatment order. Let $\bar{d}_{AB} = \dfrac{\sum_{i \in AB} d_i}{n_{AB}}$ and $\bar{d}_{BA} = \dfrac{\sum_{i \in BA} d_i}{n_{BA}}$ be the sample means for sequences AB and BA respectively.

Suppose the data generating model for an AB-BA crossover trial defined above applies, then $\hat{\tau} = \dfrac{\bar{d}_{AB} - \bar{d}_{BA}}{2}$ is an unbiased estimator of τ, that is $E\left[\hat{\tau}\right] = E\left[\dfrac{\bar{d}_{AB} - \bar{d}_{BA}}{2}\right] = \tau$.

Substitution from the model above gives

$$E\left[d_i \middle| i \in AB\right] = E\left[\left(\mu + \tau + \phi + \xi_i + \varepsilon_{ij}\right) - \left(\mu + \xi_i + \varepsilon_{ij}\right)\right]$$

The terms ξ_i on the RHS cancel so

$$E\left[d_i \middle| i \in AB\right] = E\left[\left(\tau + \phi + \varepsilon_{ij}\right) - \left(\varepsilon_{ij}\right)\right].$$

Since $E\left[\varepsilon_{ij}\right] = 0$, $E\left[d_i \middle| i \in AB\right] = \tau + \phi$.

Hence $E\left[\bar{d}_{AB}\right] = E\left[\dfrac{\sum_{i \in AB} d_i}{n_{AB}}\right] = \dfrac{n_{AB} E\left[d_i \middle| i \in AB\right]}{n_{AB}} = \phi + \tau$.

Similarly $E\left[d_i \middle| i \in BA\right] = E\left[\left(\phi + \varepsilon_{ij}\right) - \left(\tau + \varepsilon_{ij}\right)\right] = \phi - \tau$.

and $E\left[\bar{d}_{BA}\right] = E\left[\dfrac{\sum_{i\in BA} d_i}{n_{BA}}\right] = \phi - \tau$.

Hence $E\left[\dfrac{\bar{d}_{AB} - \bar{d}_{BA}}{2}\right] = \dfrac{E\left[\bar{d}_{AB}\right] - E\left[\bar{d}_{BA}\right]}{2} = \dfrac{(\tau+\phi)-(\phi-\tau)}{2} = \tau$ ●

Being the difference of two sample means of the two sequences, statistical inference on the estimator $\hat{\tau} = \dfrac{\bar{d}_{AB} - \bar{d}_{BA}}{2}$ can be based on a two sample t-test provided the assumptions of the test are satisfied. Since the patients in sequences AB and BA are independent, this assumption it follows that this assumption is satisfied by the design.

The Variance of the Difference d_i

Define $Var\left[d_i \middle| i \in AB\right] = \sigma^2_{d_{AB}}$ and $Var\left[d_i \middle| i \in BA\right] = \sigma^2_{d_{BA}}$.

From the model for a cross-over trial (in 9.1), for sequence AB

$$\sigma^2_{d_{AB}} = Var\left[d_i\right] = Var\left[y_{2i} - y_{1i}\right] = Var\left[\left(\mu + \phi + \xi_i + \varepsilon_{i2}\right) - \left(\mu + \tau + \xi_i + \varepsilon_{i1}\right)\right]$$

$$= Var\left[\varepsilon_{i2} - \varepsilon_{i1}\right] = Var\left[\varepsilon_{i1}\right] + Var\left[\varepsilon_{i2}\right] = 2\sigma^2_{\varepsilon} \text{ , since } \operatorname{cov}\left[\varepsilon_{i1}, \varepsilon_{i2}\right] = 0.$$

Similarly, for sequence BA, $\sigma^2_{d_{BA}} = 2\sigma^2_{\varepsilon}$.

Hence, the variances of the two sequences are the same for this model, that is $\sigma^2_{d_{AB}} = \sigma^2_{d_{BA}} = \sigma^2_d = 2\sigma^2_{\varepsilon}$.

Analysis of an *AB-BA* Crossover Trial using a Two Sample t-test

Hypothesis Testing

The hypothesis $H_0 : \tau = 0$ vs $H_1 : \tau \neq 0$ can be tested using a two-sample t-test of the means of the differences. The test statistic *T* is defined as

$T = \dfrac{\bar{d}_{AB} - \bar{d}_{BA}}{\hat{SE}\left[\bar{d}_{AB} - \bar{d}_{BA}\right]}$ where $\hat{SE}\left[\bar{d}_{AB} - \bar{d}_{BA}\right] = s_d \sqrt{\dfrac{1}{n_{AB}} + \dfrac{1}{n_{BA}}}$ and

$s_d = \sqrt{\dfrac{\left(n_{AB} - 1\right) s_{d_{AB}}^2 + \left(n_{BA} - 1\right) s_{d_{BA}}^2}{n_{AB} + n_{BA} - 2}}$ with $s_{d_{AB}}^2, s_{d_{BA}}^2$ be the sample variances of the differences for the two sequences. Under assumptions of normality and equality of variance $\left(\sigma_{d_{AB}}^2 = \sigma_{d_{BA}}^2 = \sigma_d^2\right)$, the test statistic *T* will have a t-distribution with $n_{AB} + n_{BA} - 2$ degrees of freedom.

Confidence Interval of the Treatment Effect

A $\left(1 - \alpha\right)$ size confidence interval for the treatment effect τ is defined by

$$\frac{1}{2}\left(\bar{d}_{AB} - \bar{d}_{BA}\right) \pm \frac{1}{2} t_{\alpha/2\left(n_{AB} + n_{BA} - 2\right)} \hat{SE}\left[\bar{d}_{AB} - \bar{d}_{BA}\right].$$

Example Bronchodilators Crossover Trial. The data in the table below is from a two-period *AB-BA* randomised crossover trial in which patients are treated with two bronchodilators, salbutamol (S) and formoterol (F). The outcome is the peak expiratory flow (PEF). Patients were randomised to receive F then S or S then F. Increased PEF is a benefit to patients.

Patient	*PEF period 1*	*PEF period 2*	Y_2-Y_1
F then S	Drug F	Drug S	
1	310	270	-40
4	310	260	-50
6	370	300	-70
7	410	390	-20
9	250	210	-40
10	380	350	-30
13	330	365	35
N_{FS}	7	$\bar{d}_{FS}$	-30.7
		$s_{d_{FS}}$	33.0
S then F	Drug S	Drug F	
2	370	385	15
3	310	400	90
5	380	410	30
8	290	320	30
11	260	340	80
12	90	220	130
N_{SF}	6	$\bar{d}_{SF}$	62.5
		$s_{d_{SF}}$	44.7

Ex 9.1 Test the null hypothesis $H_0 : \tau = 0$ vs $H_1 : \tau \neq 0$

$\bar{d}_{FS} - \bar{d}_{SF}$ =

$$s_d = \sqrt{\frac{(n_{FS}-1)s_{d_{FS}}^2 + (n_{SF}-1)s_{d_{SF}}^2}{n_{FS}+n_{SF}-2}} =$$

$$\hat{SE}\left[\bar{d}_{FS} - \bar{d}_{SF}\right] = s_d\sqrt{\frac{1}{n_{FS}} + \frac{1}{n_{SF}}} =$$

$$T = \frac{\bar{d}_{FS} - \bar{d}_{SF}}{\hat{SE}\left[\bar{d}_{FS} - \bar{d}_{SF}\right]} =$$

Degrees of freedom $= n_{FS} + n_{SF} - 2$

For a 5% level test $t_{\alpha/2}\left(n_{FS}+n_{SF}-2\right) =$

Ex 9.2 Calculate the point estimate and 95% confidence interval of the treatment effect.

The point estimate of the treatment effect of salbutamol (S) compared to formoterol (F) is $\tau =$

The 95% confidence interval of τ is

$$\frac{1}{2}\left(\bar{d}_{FS} - \bar{d}_{SF}\right) \pm \frac{1}{2}t_{0.025}\left(n_1+n_2-2\right)\hat{SE}\left[\bar{d}_{FS} - \bar{d}_{SF}\right]$$

Figure 9.1 Stata Output for Bronchodilators Crossover Trial

```
Summary statistics: mean, sd, N
  by categories of: Sequence

Sequence |      PEF1      PEF2
---------+--------------------
F then S |  337.1429  306.4286
         |  53.76315  64.72469
         |         7         7
---------+--------------------
S then F |  283.3333  345.8333
         |  105.3882  70.88136
         |         6         6
---------+--------------------
```

Two Sample t-test Analysis Applied to the Differences between Period (d_i)

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
F then S |       7   -30.71429    12.46082    32.96824   -61.20482   -.2237503
S then F |       6        62.5      18.246     44.6934    15.59715    109.4028
---------+--------------------------------------------------------------------
    diff |           -93.21429    21.55312               -140.6524   -45.77619
------------------------------------------------------------------------------
    diff = mean(F then S) - mean(S then F)                        t =  -4.3249
Ho: diff = 0                                     degrees of freedom =       11

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0006         Pr(|T| > |t|) = 0.0012          Pr(T > t) = 0.9994
```

The treatment effect and its confidence interval are obtained by halving the values in the printout. From the print-out

`diff` equals `mean(F then S) - mean(S then F)`

The treatment effect for salbutamol compared to formoterol is

$$\hat{\tau} =$$

Conclusion There is evidence that formoterol gives improved outcome compared to salbutamol with an increase in PEF equal to 46.6 (95% c.i.22.9 to 70.3 , p=0.0012).

Note that the p-values should NOT be halved!!!

Figure 9.2 Stata Output for the Single Sample t-test Applied to the Differences between treatments (c_i)

```
One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
SALB-FORM|      13   -45.38462    11.25835    40.59257   -69.91446   -20.85477
------------------------------------------------------------------------------
    mean = mean(SALB-FORM)                                        t =  -4.0312
Ho: mean = 0                                     degrees of freedom =       12

    Ha: mean < 0                 Ha: mean != 0                 Ha: mean > 0
 Pr(T < t) = 0.0008         Pr(|T| > |t|) = 0.0017          Pr(T > t) = 0.9992
```

By this method effect for salbutamol compared to formoterol

$$\hat{\tau}_C =$$

This has a slight bias compared to the unbiased method ($\hat{\tau}$ =-46.6).

It is worth noting also that the standard error for the single sample method is slightly larger. For then unbiased analysis

$$SE\left[\hat{\tau}\right] = SE\left[\frac{\bar{d}_{FS} - \bar{d}_{SF}}{2}\right] = \frac{21.55312}{2} = 10.77656$$

whereas $SE\left[\hat{\tau}_C\right] = 11.25835$, so the unbiased method is also appears to be more precise.

It can be shown that the period effect increases the standard error of the single sample methods compared to the two-sample method.

9.3 Sample Size for *AB-BA* Cross-over Trials

Given that the analysis of an *AB-BA* crossover trial should be based on a two sample t-test of the differences, we can estimate sample size for a crossover trial using the formula previously derived in the notes for sample size for a parallel group trial for a continuous outcome (see section 4.3). Consider the variance of the difference, σ_d^2.

For an *AB-BA* crossover trial with equal numbers randomised to each sequence, the total sample size *N* required to have power $(1-\beta)$ to detect a treatment effect τ using a two-sided α-size test of the hypothesis of superiority, $H_0 : \tau = 0$ vs $H_1 : \tau \neq 0$ is $N = \frac{\sigma_d^2}{\tau^2}\left(z_{\alpha/2} + z_\beta\right)^2$, where σ_d^2 is the variance of the differences.

If analysis of an AB-BA crossover trial is based on a two sample t-test, the total sample size assuming two equal size groups can be estimated from the formula in section 4.3 for sample size per group.

Total sample size is $N = \frac{4\sigma^2}{\tau^2}\left(z_{\alpha/2} + z_\beta\right)^2$.

If σ^2 is replaced σ_d^2 in the above formula, τ should be replace by 2τ, because $\bar{d}_{AB} - \bar{d}_{BA}$ estimates 2τ.

9.4 Analysis of the Period Effect ϕ

A question of secondary interest is ' Is there a period effect?'

How does one test $H_0 : \phi = 0$ vs $H_1 : \phi \neq 0$?

Consider the differences between treatment *B* and *A* for each subject define by $c_i = y_{i2} - y_{i1}$ for $i \in AB$ and $c_i = y_{i1} - y_{i2}$ for $i \in BA$. These are the difference used in the biased single sample method. It can be shown that $E\left[\frac{\bar{c}_{AB} - \bar{c}_{BA}}{2}\right] = \phi$ where $\bar{c}_{AB} = \frac{\sum_{i \in AB} c_i}{n_{AB}} = \bar{d}_{AB}$ and $\bar{c}_{BA} = \frac{\sum_{i \in BA} c_i}{n_{BA}} = -\bar{d}_{BA}$, suggesting $\hat{\phi} = \frac{\bar{c}_{AB} - \bar{c}_{BA}}{2}$. The details are an exercise. A test of the hypothesis $H_0 : \phi = 0$ vs $H_1 : \phi \neq 0$ can be carried out as a two sample t-test between the two sequences.

Figure 8.3 Stata Output for Bronchodilators Crossover Trial
Two Sample t-test Analysis Applied to the Differences between treatments (c_i)

```
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
F then S |       7   -30.71429    12.46082    32.96824   -61.20482   -.2237503
S then F |       6       -62.5      18.246     44.6934   -109.4028   -15.59715
---------+--------------------------------------------------------------------
    diff |            31.78571    21.55312               -15.65238    79.22381
------------------------------------------------------------------------------
    diff = mean(F then S) - mean(S then F)                        t =   1.4748
Ho: diff = 0                                     degrees of freedom =       11

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9158         Pr(|T| > |t|) = 0.1683          Pr(T > t) = 0.0842
```

Ex 9.3 Estimate the period effect from the print-out
The period effect $\phi =$

9.5 The Carry-Over Effect

Above, we considered how to test for a period effect. This assumed the same change ϕ for each sequence. There is the possibility that the change, after accounting for the treatment effect, differ between the two sequences *AB* and *BA*. This can occur if one treatment has greater persistent into the second period than the other and is called the *carry-over effect*. One can incorporate such a differential effect into the model of an *AB-BA* trial by adding a term γ to represent the difference in the period effect for the two sequences, *BA* and *AB*. For each group and period the models are now:

$$y_{ij} = \mu + \xi_i + \varepsilon_{ij} \qquad \text{Sequence } AB \text{ Period 1}$$

$$y_{ij} = \mu + \tau + \phi + \xi_i + \varepsilon_{ij} \qquad \text{Sequence } AB \text{ Period 2}$$

$$y_{ij} = \mu + \tau + \xi_i + \varepsilon_{ij} \qquad \text{Sequence } BA \text{ Period 1}$$

$$y_{ij} = \mu + \phi + \gamma + \xi_i + \varepsilon_{ij} \qquad \text{Sequence } BA \text{ Period 2}$$

With $d_i = y_{i2} - y_{i1}$, it follows that

$$E\left[\bar{d}_{AB}\right] = E\left[\frac{\sum d_i}{n_{AB}}\right] = \tau + \phi \qquad \text{for } AB \text{ and}$$

$$E\left[\bar{d}_{BA}\right] = E\left[\frac{\sum d_i}{n_{BA}}\right] = \phi - \tau + \gamma \qquad \text{for BA.}$$

Hence $E\left[\hat{\tau}\right] = \left[\frac{\bar{d}_{AB} - \bar{d}_{BA}}{2}\right] = \tau + \frac{\gamma}{2}$ instead of τ.

We have shown that the treatment effect will be biased, if there is a carry-over effect.

A Flawed Statistical Analysis involving Carry-Over Effect

To deal with carry-over effect in crossover trials the following analysis procedure has been suggested.

1. Test whether there is a carry-over effect i.e. $H_0 : \gamma = 0$ vs $H_1 : \gamma \neq 0$
2. If $H_0 : \gamma = 0$ is not rejected, the analysis proceeds as described above.
3. If $H_0 : \gamma = 0$ is rejected, the analysis should just be based on the period 1 data reducing the study to a parallel group trial and the period 2 data is then discarded.

The following test of for a carry-over effect has been proposed.

Define $a_i = y_{i2} + y_{i1}$ so that

$$a_i = 2\mu + \phi + \tau + 2\xi_i + \varepsilon_{i2} + \varepsilon_{i1} \qquad i \in AB$$

$$a_i = 2\mu + \phi + \tau + \gamma + 2\xi_i + \varepsilon_{i2} + \varepsilon_{i1} \qquad i \in BA$$

Defining the sample mean for each sequence as $\bar{a}_{AB}$ and $\bar{a}_{BA}$, with corresponding population means μ^a_{AB} and μ^a_{BA}

$E\left[\bar{a}_{AB}\right] = 2\mu + \tau + \phi$ and $E\left[\bar{a}_{BA}\right] = 2\mu + \tau + \phi + \gamma$

Therefore $E\left[\bar{a}_{BA} - \bar{a}_{AB}\right] = \gamma$.

This suggests the hypothesis H_0:γ=0 is equivalent to H_0: $\mu^a_{AB} = \mu^a_{BA}$, which could be tested by $T_a = \dfrac{\bar{a}_{BA} - \bar{a}_{AB}}{\hat{SE}\left[\bar{a}_{BA} - \bar{a}_{AB}\right]}$ using a two sample t-test.

Consider now

$$Var\left[a_i\right] = Var\left[2\xi_i + \varepsilon_{i2} + \varepsilon_{i1}\right] = 4Var\left[\xi_i\right] + Var\left[\varepsilon_{i2}\right] + Var\left[\varepsilon_{i1}\right] = 4\sigma_B^2 + 2\sigma_\varepsilon^2$$

Therefore $Var\left[\bar{a}_{AB}\right] = \dfrac{4\sigma_B^2 + 2\sigma_\varepsilon^2}{n_{AB}}$ and $Var\left[\bar{a}_{BA}\right] = \dfrac{4\sigma_B^2 + 2\sigma_\varepsilon^2}{n_{BA}}$.

Hence $Var\left[\bar{a}_{BA} - \bar{a}_{AB}\right] = \left(4\sigma_B^2 + 2\sigma_\varepsilon^2\right)\left(\dfrac{1}{n_{BA}} + \dfrac{1}{n_{AB}}\right)$,

giving $SE\left[\bar{a}_{BA} - \bar{a}_{AB}\right] = \sqrt{\left(4\sigma_B^2 + 2\sigma_\varepsilon^2\right)\left(\dfrac{1}{n_{BA}} + \dfrac{1}{n_{AB}}\right)}$.

This contains both *within-patient* and *between-patient* variation. We can therefore draw the following conclusion:

Unless σ_B^2 is small relative to σ_ε^2 the test statistic T_a will have low power to reject the hypothesis of no carry-over effect $H_0 : \gamma = 0$ compared to a test of the treatment effect τ.

The advantage of crossover trials is that they remove *between-patient* variance σ_B^2, and so there is only going to be power to detect a carry-over effect where there is little benefit in using a crossover design. To increase the sample size so that a carry-over effect could be detected would remove the main advantage of the crossover design. Note, also that the inferential reasoning of the test is faulty as in practice one would want to show that γ is zero to justify the crossover analysis. As with an equivalence trial one would need to have $H_0 : \gamma \neq 0$ and $H_1 : \gamma = 0$. For these reasons the procedure for testing for carry-over effect in a crossover trial is no longer recommended.

Implications of Carry-Over Effect and Crossover Trials

When planning a crossover trial, it is important to consider whether a carry-over effect may occur, as estimation is only unbiased where it is absent. This assumption has to be based on scientific arguments regarding the way treatments work rather than statistical tests. For example, in a particular situation carry-over may not be plausible so it can be safely ignored. Alternatively, it may be eliminated by a washout period between the two treatments. In such circumstances a crossover trial may be legitimately conducted and the advantages of crossover trial exploited.

Unfortunately, in many situations a carry-over effect may be plausible and lengthy washouts periods may not be feasible as it would be unethical for a patients to go without treatment. There are therefore many circumstances where a crossover trial is not feasible, when testing treatments for chronic conditions.

This is an example of where clinical science is important for determining the choice of design. In summary, if there are good scientific reasons to believe a carry-over effect may occur, the crossover design is not recommended and a parallel group design should be used instead.

9.6 Summary: Comparison of Parallel and Crossover Trials

Parallel Group Design

- Comparison of treatments is between groups of patients.
- Power and sample size depends on between-subject variation.

Crossover Design

- Comparison of treatments is within patients so that each patient acts as their own control.

Advantage of a Crossover Design compared to a Parallel Group Design

- Within patient estimation of treatment effects means that variation between patients is removed from the analysis, hence sample size may be substantially smaller.

Disadvantages of a Crossover Design compared to a Parallel Group Design

- Only applicable to certain types of condition such as stable diseases.
- More complicated to organize.
- Patients withdrawing during the second period mean that their data cannot be included in the statistical analysis.
- Requires the assumption of no Carry-over effect to give unbiased estimates of the treatment effect.

10. Systematic Review and Meta-Analysis

10.1 Systematic Review

If a clinical trial has been properly conducted, it should provide information regarding the efficacy or effectiveness of a new therapy. Once a trial has been published, it might be thought that it would be unethical to undertake another trial making the same comparison. In practice, the decision is rarely so simple. Trial results generally need to be replicated before new treatment can be widely adopted. Clinical trials of new treatments are often repeated within different global regions (e.g. Europe, Americas, or Asia) to assess their generalisability. Modifications to the trial design may be made to remove perceived biases in the design of earlier studies or test the effect of treatment on other outcome measures.

Where several trials have been carried out to compare the same treatments, the traditional method for assessing the evidence involved selecting from the readily available trial reports, appraising each, before drawing conclusions in a narrative discussion. This type of review can be highly subjective and open to selection bias. An alternative is a *systematic review*, in which studies are identified systematically in attempt to find all before combining the results by an overall statistical summary. Systematic review is now an important component of the evaluation of new treatments and diagnostic test procedures, and is also used to combine evidence from epidemiological studies. *Meta-analysis* is the statistical methodology for combining data from several studies of the same question to produce an overall summary.

A systematic review followed by a meta-analysis can bring together the results from several inconclusive or conflicting studies to give a single conclusive result. It also gives greater power and precision to answer more refined research questions. For example, individual trials are usually designed to answer the question “does a treatment work on average”. They rarely have sufficient power to investigate differences in the treatment effect for specific types of patient, but this may be possible by combining data from several studies in a meta-analysis. A systematic review may also enable one to investigate rarer outcomes, such as serious adverse events, that may not be possible in a single trial. For example by combining trials of treatments for depression it has been possible to show that some drug treatments increased the risk of suicidal behaviour, a result that could not be demonstrated in individual trials due to lack of power for this outcome measure.

Steps in a Systematic Review

A *systematic review* is similar to a clinical trial. It involves several steps.

1. Define precise objectives for the review.
2. Set inclusion and exclusion criteria for trials.
3. Search for trials satisfying the inclusion criteria.
4. Assess methodological quality of studies identified, possibly discarding methodologically poor studies.
5. Extract statistical summary data or obtain raw data for each study.
6. Estimate the overall treatment effect by a meta-analysis.

10.2 Meta-analysis

Meta-analysis of generally address the following questions

1. Are the effects in the studies homogeneous? This is needed to justify estimating an overall treatment effect.
2. What is the overall treatment effect?
3. Do study size, study characteristics or methodological quality correlate with the magnitude of the treatment effect?

The best way to carry out a meta-analysis is to combine the raw data from individual studies into a single large dataset, and then carry out an analysis of all the data to estimate the overall effect. This is method called *individual patient data* meta-analysis. Whilst this is similar to analysing a single large study, analysis should take account of data coming from several studies.

Individual patient data meta-analysis is often not possible, because the original raw data are no longer available for all studies, particularly where some may be many years old. For this reason, most meta-analyses use summary statistics extracted from published reports. This method is called *summary measures* meta-analysis and is a special set of methods.

Fixed or Random Effect Meta-analysis

Suppose there are *k* studies and the treatment effect estimate for the i^{th} study is $\hat{\theta}_i$. Suppose the overall treatment effect is θ. There are two main approaches to estimation of θ.

In the first, we assume that $\hat{\theta}_i$ each trial is estimating the a common effect of treatment θ. Any departure of $\hat{\theta}_i$ from θ is assumed to be simply due to sampling variation. This is called Fixed *Effects* estimation.

The second approach is called *Random Effects* estimation. This assumes that the studies are sampled from a larger population of studies. The treatment effect θ_i is then a random variable with mean equal to the overall effect θ and variance *v*.

If $\hat{\theta}$ is the overall estimate, $Var\left[\hat{\theta}\right]$ will be larger if estimated by random effect estimation than fixed effect estimation due to the additional variance term *v*.

In this module we will just describe methods of analysis for fixed effects estimation.

10.3 Summary Measures Estimation of the Overall Effect

Suppose there are *k* trials comparing two treatments. Let $\hat{\theta}_i$ be the estimate of the treatment effect for the j^{th} trial and let $v_i = Var\left[\hat{\theta}_i\right]$.

For a continuous outcome measure *y* , define $\bar{y}_{ij}$, s_{ij}^2 and n_{ij} $\left(i = 1,...,k; j = 1,2\right)$ to be sample mean and variance, and the sample size respectively of the j^{th} treatment in the i^{th} trial. The treatment effect of the i^{th} trial can be the mean difference, $\hat{\theta}_i = \bar{y}_{i2} - \bar{y}_{i1}$ with

$$\hat{v}_i = \hat{Var}\left[\hat{\theta}_i\right] = \frac{s_{i1}^2}{n_{i1}} + \frac{s_{i2}^2}{n_{i2}}.$$

If *Y* is binary, one could use the rate difference (*RD*), as the summary statistic for each trial. If the observed number of events is r_{ij}, the observed proportions $p_{ij} = r_{ij}/n_{ij}$ $\left(i = 1,...,k; j = 1,2\right)$. One can define

$$\hat{\theta}_i = P_{i2} - P_{i1} \text{ with } \hat{Var}\left[\hat{\theta}_i\right] = \frac{p_{i1}\left(1-p_{i1}\right)}{n_{i1}} + \frac{p_{i2}\left(1-p_{i2}\right)}{n_{i2}} = \hat{v}_i \ .$$

Alternatively, one might want to estimate an overall odds ratio (*OR*) or rate ratio (*RR*). These are generally estimated by taking $\hat{\theta}_i$ equal to the $\log_e\left[OR\right]$ or $\log_e\left[RR\right]$.

For $\log_e\left[OR\right]$, $\hat{v}_i = \frac{1}{r_{i1}} + \frac{1}{n_{i1}-r_{i1}} + \frac{1}{r_{i2}} + \frac{1}{n_{i2}-r_{i2}}$ demonstrated in section 4.

For $\log_e\left[RR\right]$, $\hat{v}_i =$.

The overall estimate for odd ratio and rate ratio are obtained by taking the exponent of the overall $\log_e[OR]$ and $\log_e[RR]$ estimates.

Summary Measures Estimate of the Overall Effect

Whichever type of summary measure is used (mean difference, *RD*, log [*OR*] or log[*RR*]), an overall estimate of θ can be estimated by a weighted mean defined as

$$\hat{\theta} = \sum_{i=1}^{k} w_i \hat{\theta}_i \Big/ \sum_{i=1}^{k} w_i \, .$$

Consider now the variance of the estimate for $\hat{\theta}$.

$$Var\left[\hat{\theta}\right] = \frac{1}{\left(\sum_{i=1}^{k} w_i\right)^2} Var\left[\sum_{i=1}^{k} w_i . \hat{\theta}_i\right]$$

Since the studies are independent, $Cov\left[\hat{\theta}_i, \hat{\theta}_j\right] = 0$

Therefore, $Var\left[\hat{\theta}\right] = \dfrac{\sum_{i=1}^{k} w_i^2 Var\left[\hat{\theta}_i\right]}{\left(\sum_{i=1}^{k} w_i\right)^2}$.

Choice of Weights and the Minimum Variance Estimate

Different weights will give different estimates of θ and $Var\left[\hat{\theta}\right]$. We could weight studies equally by setting $w_i = 1, i = 1, ..., k$, but this is rarely done as the size of studies generally varies greatly. It can be

shown that taking $w_i \propto \frac{1}{Var\left[\hat{\theta}_i\right]}$ for each *i* gives an estimator with minimum variance, that is with greater precision. For this reason, *inverse variance weights* are often used in meta-analysis.

The weighted variance $Var\left[\hat{\theta}\right] = \frac{\sum_{i=1}^{k} w_i^2 . Var\left[\hat{\theta}_i\right]}{\left(\sum_{i=1}^{k} w_i\right)^2}$ will have a minimum when if $w_i \propto 1/Var[\theta_i]$.

The proof uses the Lagrange Multiplier method for obtaining maxima or minima subject to a constraint.

Let $Var\left[\hat{\theta}\right] = F\left(w_1, w_2, ..., w_k\right) = \frac{\sum_{i=1}^{k} w_i^2 . Var\left[\hat{\theta}_i\right]}{\left(\sum_{i=1}^{k} w_i\right)^2}$

Without loss of generality one can apply the constraint $\sum_{i=1}^{k} w_i = 1$.

Define $G\left(w_1, w_2, ..., w_k\right) = \sum_{i=1}^{k} w_i - 1$.

Applying the Lagrange Multiplier Method one defines

$$H\left(w_1, w_2, ..., w_k, \lambda\right) = F\left(w_1, w_2, ..., w_k\right) + \lambda G\left(w_1, w_2, ..., w_k\right)$$

The minimum of *F* subject to the constraint *G* is found by equating the partial derivatives of $H\left(w_1, w_2, ..., w_k, \lambda\right)$ with respect to each w_i to zero. Considering the j^{th} study

$$\frac{\partial H}{\partial w_j}(w_1, w_2, ..., w_k, \lambda) = \frac{\partial}{\partial w_j}\sum_{i=1}^{k} w_i^2 . Var\left[\hat{\theta}_i\right] + \lambda \frac{\partial}{\partial w_j}\left(\sum_{i=1}^{k} w_i - 1\right)$$

Hence

$$\frac{\partial H}{\partial w_j}(w_1, w_2, ..., w_k) = 2w_j Var\left[\hat{\theta}_j\right] + \lambda = 0 \text{ giving } w_i = -\lambda \Big/ \left(2Var\left[\hat{\theta}_i\right]\right)$$

The second derivatives of *H* are positive so this must be a minimum.

Hence $w_i \propto 1 \Big/ Var\left[\hat{\theta}_i\right]$ gives the estimate with minimum variance●

If $\hat{\theta}_{MV}$ is the minimum variance estimate then

$$Var\left[\hat{\theta}_{MV}\right] = \frac{1}{\sum_{i=1}^{k} \frac{1}{Var\left[\hat{\theta}_i\right]}} = \frac{1}{\sum_{i=1}^{k} w_i}$$

Substitutes $\frac{1}{Var\left[\hat{\theta}_i\right]}$ for w_i , into

$$Var\left[\hat{\theta}\right] = \frac{\sum_{i=1}^{k} w_i^2 Var\left[\hat{\theta}_i\right]}{\left(\sum_{i=1}^{k} w_i\right)^2}$$

gives

$$Var\left[\hat{\theta}_{MV}\right] = \frac{\sum_{i=1}^{k} w_i^2 . Var\left[\hat{\theta}_i\right]}{\left(\sum_{i=1}^{k} w_i\right)^2} = \frac{\sum_{i=1}^{k} \frac{1}{Var\left[\hat{\theta}_i\right]}}{\left(\sum_{i=1}^{k} \frac{1}{Var\left[\hat{\theta}_i\right]}\right)^2} = \frac{1}{\sum_{i=1}^{k} \frac{1}{Var\left[\hat{\theta}_i\right]}} = \frac{1}{\sum_{i=1}^{k} w_i}$$

as required ●

Summary Measures Inference

Confidence Intervals

Even if the source data are not normally distributed, it is plausible that the individual study level estimates $\hat{\theta}_i$ are by the central limit theorem. Since $\hat{\theta}_{MV}$ is a linear function of $\hat{\theta}_i$s that are plausibly normally distributed, we can assume that $\hat{\theta}_{MV}$ is also normal. A $(1-\alpha)$ level confidence interval of $\hat{\theta}_{MV}$ can therefore be given by

$$\hat{\theta}_{MV} \pm z_{\alpha/2} \frac{1}{\sqrt{\sum_{i=1}^{k} w_i}}$$

Hypothesize Tests

To test the null hypothesis $H_0 : \theta = 0$, the following test statistic can be used

$$T = \frac{\hat{\theta}_{MV}}{SE\left[\hat{\theta}_{MV}\right]} = \hat{\theta}_{MV} \sqrt{\sum_{i=1}^{k} w_i} \ ,$$

which can be assumed to have a standardised normal distribution under the null hypothesis.

Ex 10.1 Systematic review of the effect of maternal steroid therapy on neonatal mortality.

The table over-page summarizes the results for 12 trial identified by a systematic review of trials testing maternal steroid therapy. The outcome measure is the number of neonatal deaths, which is death within the first 28 days of life. Note that in one study, there are no deaths in both arms and so this study cannot contribute to the meta-analysis and has to be excluded from the analysis.

(i) Estimate the minimum variance estimate and its 95% confidence interval.

(ii) Test the null hypothesis of no overall treatment effect.

Some of the computation is carried out on the table above summarizing the raw data.

(iii) Display the data graphically.

The standard method of graphical display of a meta-analysis is a forest plot illustrated below.

Fixed Effects Meta-Analysis of the Effect of Maternal Steroid Therapy on Neonatal Mortality (Crowley et al,1990)

Trial No.	**Steroid Therapy**			**Control**			$\hat{\theta}_i = P_S - P_C$	$v_i = Var[P_S - P_C]$	$w_i = 1/v_i$	$w_i\hat{\theta}_i$
	$Died\ (r_S)$	P_S	n_S	$Died\ (r_C)$	P_C	n_C				
Liggins	36	0.068	532	60	0.112	538	-0.044	0.000303		
Block	1	0.014	69	5	0.082	61	-0.067	0.001441		
Scliutte	3	0.047	64	12	0.207	58	-0.160	0.003527	283.5	-45.37
Taeush	5	0.089	56	7	0.099	71	-0.009	0.002704	369.9	-3.44
Doran	2	0.025	81	10	0.159	63	-0.134	0.002417	413.8	-55.46
Teranin	0	0.000	38	0	0.000	42	0.000	-	-	-
Gamsu	14	0.107	131	20	0.146	137	-0.039	0.001639	610.3	-23.87
Collab. Grp.	36	0.097	371	37	0.099	372	-0.002	0.000477	2096.7	-5.09
Morales	7	0.058	121	13	0.105	124	-0.047	0.001207	828.3	-38.92
Papagecrgio	1	0.014	71	5	0.067	75	-0.053	0.001025	975.4	-51.29
Morrison	2	0.030	67	7	0.119	59	-0.089	0.002205	453.6	-40.28
Schmidt	5	0.147	34	5	0.161	31	-0.014	0.008053	124.2	-1.77
								Σ	10152.6	-457.2

(i) Estimate the minimum variance fixed effect estimate and its 95% confidence interval.

The fixed effect estimate

$$\hat{\theta}_{MV} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i} =$$

$$\hat{SE}\left[\theta_{MV}\right] = \sqrt{\frac{1}{\sum_{i=1}^{k} w_i}} =$$

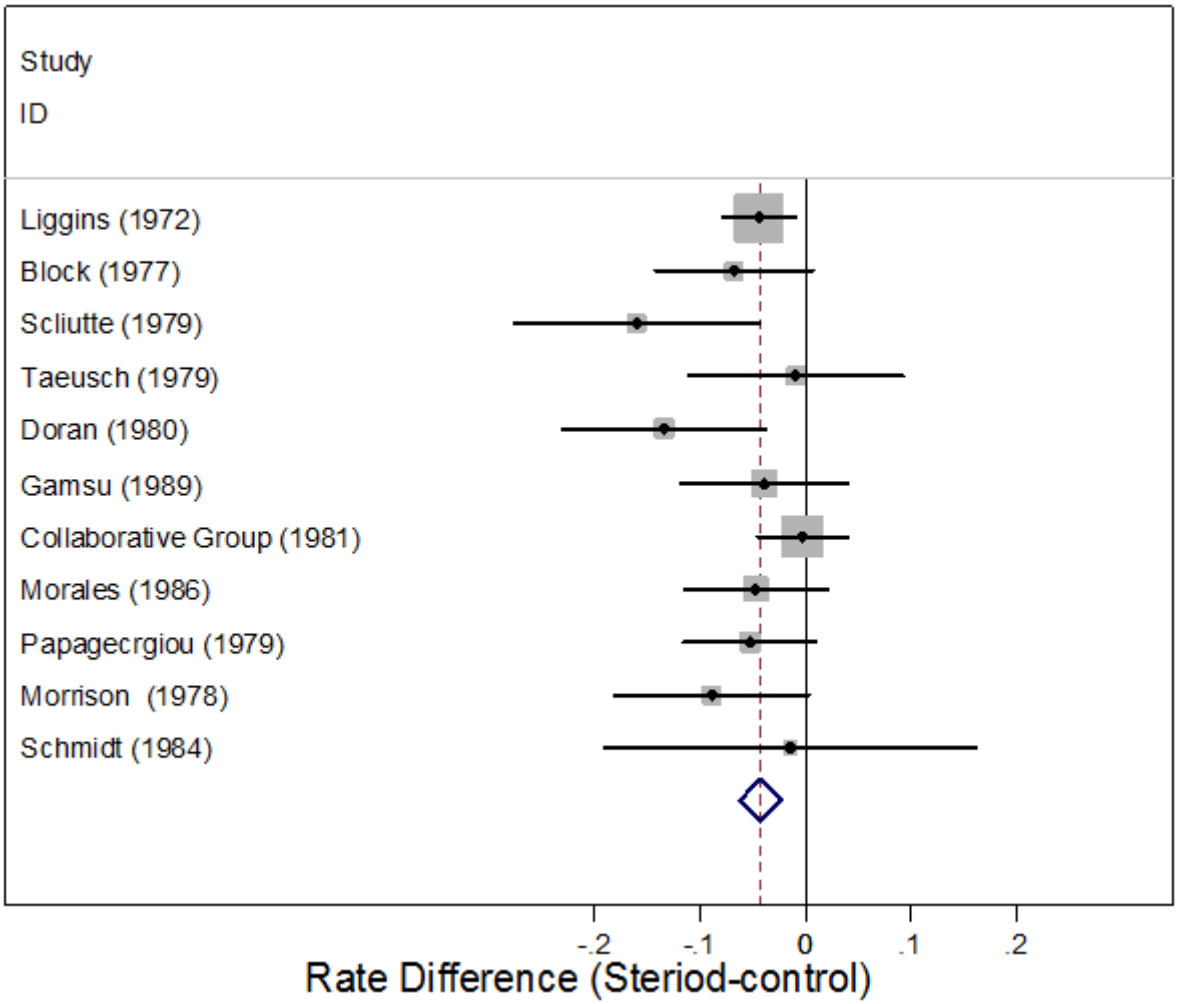
95% C.I. is $\hat{\theta}_{MV} \pm z_{\alpha/2} \hat{SE}\left[\theta_{MV}\right]$,

(ii) Test the null hypothesis of no overall treatment effect

$$T = \frac{\hat{\theta}_{MV}}{SE\left[\hat{\theta}_{MV}\right]} =$$

(iii) **Graphical Display of Meta-Analyses**

Forest Plot of Data from Crowley et al.



The treatment effect in each study is represented by a square with bars represent the 95% confidence interval of the treatment effect. The combined treatment effect and its confidence interval are shown at the bottom of the figure as a diamond. The area of the block representing the point estimate for each study has been made inversely proportional to the variance. Since larger studies will have smaller variance, larger studies will be represented by a large block. This is added otherwise the eye would tend to be attracted towards the studies that have wider confidence intervals which are smaller.

10.4 Investigation of Biases

It is well known that studies that fail to find a statistically significant treatment effect are less likely to be published than those that do. This mean that a meta-analysis based only on published studies may be biased. The term used for this phenomenon is *Publication Bias*.

Possible Causes of Publication Bias

- Selective publications: Studies in which an intervention is found to be ineffective are sometimes never published. Sponsors of research, such as pharmaceutical companies or the innovator of the treatment, have been known to discourage or prevent publication of unfavourable results. If the results are negative, a clinical researcher may be less motivated to get a trial published as they are conscious that they may be considered less interesting to journal editors and so much more difficult to get accepted published.

- Identification: Studies in which results are statistically significant are likely to be published in more prestigious, and hence easily accessible, journals. As an illustration of this, it has been shown that trials carried out in non-English speaking countries are more likely to be published in English where the study result is statistically significant. Hence, a meta-analysis restricted to English language journals may overestimate the treatment effect as studies in other languages will tend to have a smaller effect.

- Selective reporting: Where studies have multiple outcomes measured, statistically significant results may be emphasized in reports whereas non-significant results may be given less prominence or even left out. Glaring examples of this are trial reports that fail to give the primary outcome measure previously specified in the trial protocol, but publish other measures that have been found to be statistically significant.

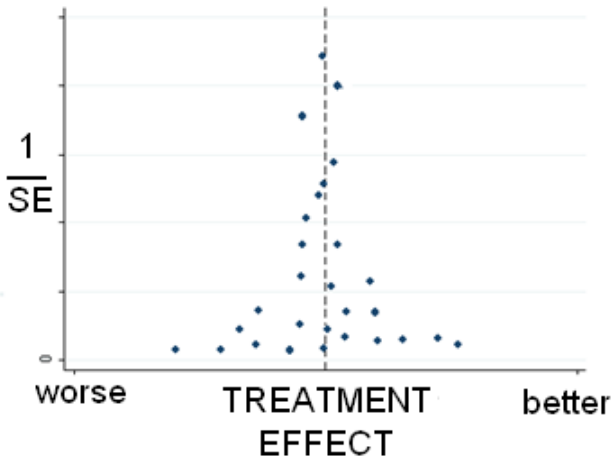
The Funnel Plots

One way to investigate publication bias is a funnel plot. This plots

$$\text{Precison}\left[\hat{\theta}\right] = \frac{1}{\text{SE}\left[\hat{\theta}\right]}$$

against the treatment effect for each trial $\hat{\theta}$. Assuming all studies in the meta-analysis are a random sample of all possible studies of the same treatment, the distribution of points should resemble an inverted funnel shape widening as the precision decreases. This is because studies with larger standard errors (i.e. less precision) will have wider confidence intervals and so estimates of the treatment effect will be more widely dispersed.

Figure 10.2 A Funnel Plots

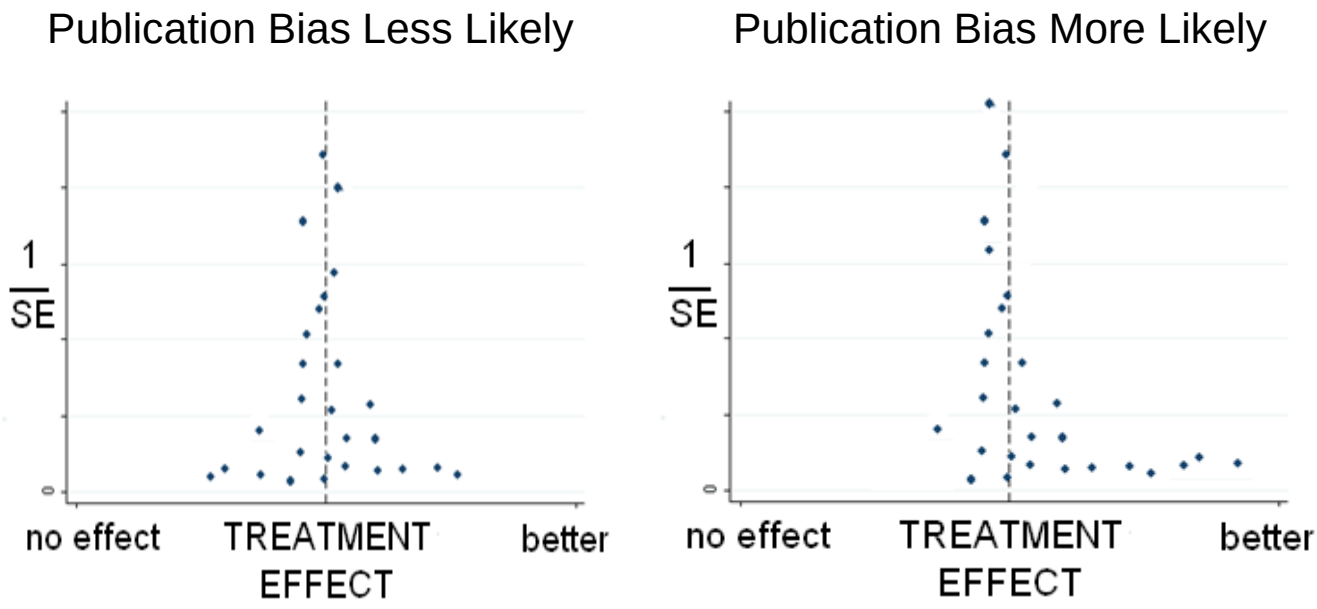


Funnel plots can also constructed by plotting total sample size against the treatment effect and give a similar shaped figure, as precision is related to the square root of the sample size.

Funnel Plots Asymmetry

Studies with greater precision have larger sample size and so tend to get published irrespective of statistically significance. In contrast, studies with less precision are less likely to be published, if they are not statistically significant. Hence, smaller studies showing a smaller treatment are more likely to be missed by a systematic review and so left out of a meta-analysis. This is illustrated in figure 10.3.

Figure 10.3 Illustration of publication bias

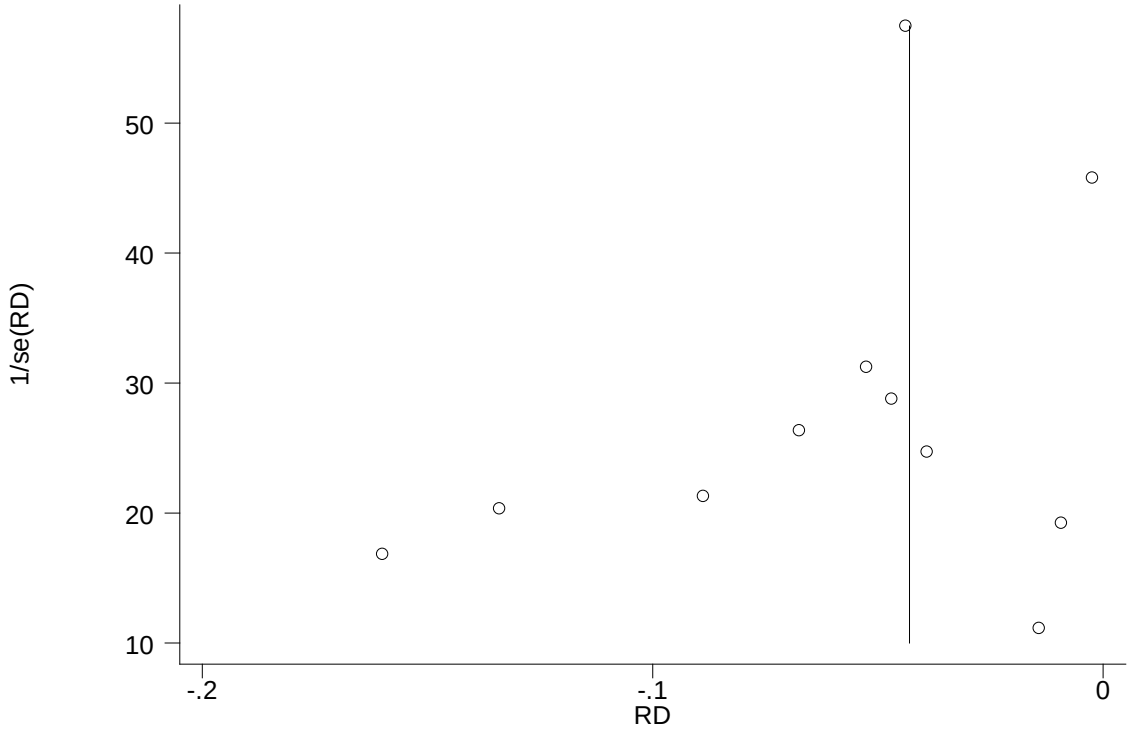


As well as publication bias, lack of symmetry in funnel plots may indicate:

- True Heterogeneity – the treatment in smaller studies may be more intensive than in larger studies or patients in smaller studies may differ systematically from those in larger studies.
- Outcome may be measured in different ways depending on trial size.
- Smaller trials may be more poorly conducted than larger studies and so more likely to be biased.

The possibility of publication bias means that it is particularly important that meta-analyses are based on all relevant studies and not just those that are conveniently available. Researchers carrying out systematic reviews are encouraged to identify trials that have not been published or are reported in more obscure journals. To aid this, an international directory of clinical trials (ISRCTN) has been established with which all new randomised trials should registers.

Figure 10.4 Funnel Plot of Crowley et al.



There is some evidence in the funnel plot above that smaller studies showed a larger effect. This could be due to publication bias.

If there is concern that there may be publication bias, one option would be to carry out a sensitivity analysis excluding smaller studies.

END OF MODULE