Sample Size Determination

José Sánchez 2020-02-26

- **“The number of subjects in a clinical study should always be large enough to provide a reliable answer to the question(s) addressed.”**

- **“The sample size is usually determined by the primary objective of the trial.”**

- **“ Sample size calculation should be explicitly mentioned in the protocol .”**

(from ICH-E9)

Errors with Hypothesis Testing

	Accept H_0	Reject H_0
		False positives
E = C	**OK**	Type I error
E ≠ C	Type II error False negatives	**OK**

Ho: E=C (ineffective) H1: E≠C (effective)

E: Effect of **E**xperimental Drug

C: Effect of **C**ontrol

Type I Error

Concluding for alternative hypothesis (new drug effective) while null-hypothesis is true (false positive)

- Probability of Type I error = *significance level* or α *level*
- Value needs to be specified in the protocol and should be small
- Explicit guidance from regulatory authorities, e.g. *0.05*
- One-sided or two-sided?
- Multiple testing?

Type II Error

Concluding for null-hypothesis (new drug ineffective) while alternative hypothesis is true (false negative)

- Probability of Type II error = β *level*
- Value to be chosen by the sponsor, and should be small
- No explicit guidance from regulatory authorities, e.g. 0.20, or 0.10
- Power

  =1 - β

= 1 - Type II error rate

=P(reject H_0 when H_1 is true)

=P(concluding that the drug is effective when it is)

- Typical values 0.80 or 0.90

Power and sample size

- Suppose we want to test if a drug is better than a placebo, or if a higher dose is better than a lower dose.

- Sample size: How many patients should we include in our clinical trial, to give ourselves a good chance of detecting any effects of the drug?

- Power: Assuming that the drug has an effect, what is the probability that our clinical trial will give a significant result?

Sample Size and Power

- **Sample size is contingent on design, true effect, underlying variation, analysis method, outcome etc.**
- **With the wrong sample size, you will either**
  - Not be able to make conclusions because the study is "underpowered"
  - Waste time and money because your study is larger than it needed to be to answer the question of interest

Sample Size and Power

- **Sample size ALWAYS requires the investigator to make some assumptions**
  - How much better *do you expect* the experimental therapy group to perform than the standard therapy groups?
  - How much variability *do we expect* in measurements?
  - What *would be* a clinically relevant improvement?
- **The statistician CANNOT tell what these numbers should be**
- **It is the responsibility of the clinical investigators to define these parameters**

Sample Size Calculation

- Following items should be specified
  - a primary variable
  - the statistical test method
  - the null hypothesis; the alternative hypothesis; the study design
  - the Type I error
  - the Type II error
  - variability in the study population
  - way how to deal with treatment withdrawals
  - presumed effect

Three common settings

1. Continuous outcome: e.g., number of units of blood transfused, CD4 cell counts, LDL-C.

2. Binary outcome: e.g., response vs. no response, disease vs. no disease

3. Time-to-event outcome: e.g., time to progression, time to death.

Continuous outcomes

- Easiest to discuss

- Sample size depends on
  - Δ: difference under the null hypothesis
  - α: type 1 error
  - B: type 2 error
  - σ: standard deviation
  - r: ratio of number of patients in the two groups (usually r = 1)

Tests and Confidence Intervals

- You can use a confidence interval (CI) for hypothesis testing. In the typical case, if the CI for an effect does not span 0 then you can reject the null hypothesis.
- But a CI can be used for more. For example, you can make a statement about the range of effects you believe to be likely (the ones in the CI). You can't do that with just a t-test.
- You can also use it to make statements about the null, which you can't do with a t-test. If the t-test doesn't reject the null then you just say that you can't reject the null, which isn't saying much. But if you have a narrow confidence interval around the null then you can suggest that the null, or a value close to it, is likely the true value and suggest the effect of the treatment is too small to be meaningful.

One-sample

- Assume that we wish to study a hypothesis related to a sample of independent and identically distributed normal random variables with mean μ and standard deviation σ. Assuming σ is known, a (1-α)100% confidence interval for μ is given by

$$\overline{Y} \pm Z_{\left(\frac{\alpha}{2}\right)} \frac{\sigma}{\sqrt{n}}$$

- The maximum error in estimating μ is $E = Z_{\left(\frac{\alpha}{2}\right)} \frac{\sigma}{\sqrt{n}}$ which is a function of n. If we pre-specify E, then n can be chosen according to

$$n = \frac{Z_{\left(\frac{\alpha}{2}\right)}^{2} \sigma^2}{E^2}$$

- This takes care of the problem of determining the sample size required to attain a certain precision in estimating μ using a confidence interval at a certain confidence level.
- Since a confidence interval is equivalent to a test, the above procedure can also be used to test a null hypothesis. Suppose thus that we wish to test the hypothesis
  - H_0: $\mu_1 = \mu_0$
  - H_a: $\mu_a > \mu_0$
- with a significance level α. For a specific alternative

H_a: $\mu_a = \mu_0 + \Delta$ where $\Delta > 0$, the power of the test is given by

$$1-\beta = P\{\text{reject } H_0 | H_a \text{ is true}\}$$

$$= P\left\{\frac{\bar{Y} - (\mu_0+\Delta)}{\sigma/\sqrt{n}} > Z(\alpha) - \frac{\Delta}{\sigma/\sqrt{n}} \,\middle|\, \mu = \mu_0 + \Delta\right\}.$$

Under the alternative hypothesis that $\mu = \mu_0 + \Delta$, the test statistic

$$\frac{\bar{Y} - (\mu_0+\Delta)}{\sigma/\sqrt{n}}$$

follows a standard normal variable. Therefore

$$1-\beta = P\left\{Z > Z(\alpha) - \frac{\Delta\sqrt{n}}{\sigma}\right\},$$

from which we conclude that

$$-Z(\beta) = Z(\alpha) - \frac{\Delta\sqrt{n}}{\sigma}$$

and hence

$$n = \frac{\sigma^2[Z(\alpha)+Z(\beta)]^2}{\Delta^2}.$$

Two samples

- H_0: $\mu_1 - \mu_2 = 0$
- H_a: $\mu_1 - \mu_2 \neq 0$
- With known variances σ_1 and σ_2, the power against a specific alternative $\mu_1 = \mu_2 + \Delta$ is given by

$$1 - \beta = P\left\{\left|\frac{\bar{Y}_1 - \bar{Y}_2}{\sigma_d}\right| > Z(\alpha/2) | \mu_1 = \mu_2 + \Delta\right\},$$

where

$$\sigma_d = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Therefore

$$\beta = P\left\{-Z(\alpha/2) - \frac{\Delta}{\sigma_d} < \frac{(\bar{Y}_1 - \bar{Y}_2) - \Delta}{\sigma_d} < Z(\alpha/2) - \frac{\Delta}{\sigma_d} | \mu_1 = \mu_2 + \Delta\right\}.$$

Under the alternative hypothesis, the statistic

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - \Delta}{\sigma_d}$$

- is normal so

$$\beta = P\left[-Z(\frac{\alpha}{2}) - \frac{\Delta}{\sigma_d} < Z < -Z(\frac{\alpha}{2}) - \frac{\Delta}{\sigma_d}\right]$$

$$\Rightarrow -Z(\beta) = Z(\frac{\alpha}{2}) - \frac{\Delta}{\sigma_d}$$

$$\Rightarrow n = \frac{\left[Z(\frac{\alpha}{2}) + Z(\beta)\right]^2 ({\sigma_1}^2 + {\sigma_2}^2)}{\Delta^2}$$

- In the one sided case we have

$$n = \frac{[Z(\alpha) + Z(\beta)]^2 ({\sigma_1}^2 + {\sigma_2}^2)}{\Delta^2}$$

- In the case of equal variances, ${\sigma_1}^2 = {\sigma_2}^2 = \sigma^2$, we have

$$n = 2\frac{[Z(\alpha) + Z(\beta)]^2 \sigma^2}{\Delta^2}$$

One proportion

We want to test

- H_0: $P = P_0$
- H_a: $P > P_1$

with a certain power (1-β) and with a certain significance level (α). The required sample of size is

$$n = \frac{\left(Z(\alpha)\sqrt{P_0(1-P_0)} + Z(\beta)\sqrt{P_1(1-P_1)}\right)^2}{(P_2 - P_1)^2}$$

Two proportions

We want to test

- H_0: $P_1 = P_2$
- H_a: $P_1 \neq P_2$

with a certain power (1-β) and with a certain significance level (α). The required sample of size is

$$n = \frac{\left(Z(\alpha/2)\sqrt{P(1-P)} + Z(\beta)\sqrt{P_1(1-P_1) + P_2(1-P_2)}\right)^2}{(P_2 - P_1)^2},$$

$$P = \frac{(P_1 + P_2)}{2}$$

Example 11.3.1 Suppose that a pharmaceutical company is interested in conducting a clinical trial to compare two cholesterol lowering agents for treatment of hypercholesterolemic patients. The primary efficacy parameter is a low-density lipidprotein cholesterol (LDL-C). Suppose that a difference of 8% in the percent change of LDL-C is considered a clinically meaningful difference and that the standard deviation is assumed to be 15%. Then, by (11.3.8), at $\alpha = 0.05$, the required sample size for having an 80% power can be obtained as follows:

$$n = \frac{2\sigma^2[Z(\alpha/2) + Z(\beta)]^2}{\Delta^2}$$

$$= \frac{2(15)^2[1.96 + 0.842]^2}{(8)^2}$$

$$= 55.2 \approx 56.$$

Therefore a sample size of 56 patients per arm is required to obtain an 80% power for detection of an 8% difference in percent change of LDL-C for the intended clinical study.

Using SAS

```
proc power;
    twosamplemeans test=diff
    meandiff = 0.08
    stddev = 0.15
    power = 0.8
    npergroup = . ;
 run;
```

The POWER Procedure
Two-Sample t Test for Mean Difference

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Mean Difference	0.08
Standard Deviation	0.15
Nominal Power	0.8
Number of Sides	2
Null Difference	0
Alpha	0.05

Computed N per Group	
Actual Power	N per Group
0.806	57

A sample size of 57 in each group will have 80% power to detect a difference in means of 0,08 assuming that the common standard deviation is 0,15 using a two group t-test with a 0,05 two-sided significance level.

Example 11.3.2 A pharmaceutical company is interested in examining the effect of an antidepressant agent in patients with generalized anxiety disorder. A double-blind, two-arm parallel, placebo-controlled randomized trial is planned. To determine the required sample size for achieving an 80% power, the HAM-A scores is considered as the primary efficacy variable. It is believed that a difference of 4 in the HAM-A scores between the antidepressant and the placebo is of clinical importance. Assuming that the standard deviation is 7.0 obtained from previous studies, the required sample size can be obtained based on (11.3.8), which is given by

$$n = \frac{2\sigma^2[Z(\alpha/2) + Z(\beta)]^2}{\Delta^2}$$

$$= \frac{2(7)^2[1.96 + 0.842]^2}{(4)^2}$$

$$= 48.1 \approx 49.$$

Using SAS

```
proc power;
   twosamplemeans test=diff
   meandiff = 4
   stddev = 7
   power = 0.8
   npergroup = . ;
run;
```

The POWER Procedure
Two-Sample t Test for Mean Difference

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Mean Difference	4
Standard Deviation	7
Nominal Power	0.8
Number of Sides	2
Null Difference	0
Alpha	0.05

Computed N per Group	
Actual Power	N per Group
0.808	50

A sample size of 50 in each group will have 80% power to detect a difference in means of 4 assuming that the common standard deviation is 7 using a two group t-test with a 0,05 two-sided significance level.

Analysis of variance

We want to to test

- H_0: $\tau_1 = \tau_2 = \ldots = \tau_k$
- H_a: at least one τ_i is not zero

Take:

$$\nu_1 = k - 1, \nu_2 = k(n-1) = N - k,$$

$$F_A^* = F(\alpha, \nu_1, \nu_2), \delta^2 = \frac{n\sum \tau_i^2}{2\sigma^2},$$

$$Z(\beta) = \frac{\left(\sqrt{\nu_2[2(\nu_1 + \delta^2)^2 - (\nu_1 + 2\delta^2)^2]} - \sqrt{\nu_1[(\nu_1 + \delta^2)(2\nu_2 - 1)F_A^*]}\right)}{\sqrt{\nu_1[(\nu_1 + \delta^2)F_A^* + \nu_2(2\nu_1 + 2\delta^2)]}}$$

The last equation is difficult to solve but there are tables.

Example 11.4.1 To illustrate the use of (11.4.4) for the sample size determination in comparing more than two treatments, consider the following example: Suppose that we are interested in conducting a four-arm parallel group, double-blind, randomized clinical trial to compare four treatments. The comparison will be made based on an F test with a significance level of $\alpha = 0.05$. Assume that the standard error within each group is expected to be $\sigma = 3.5$ and that the clinically important differences for the four treatment groups are given by

$$\tau_1 = -0.75, \quad \tau_2 = 3.0, \quad \tau_3 = -0.5, \quad \text{and} \quad \tau_4 = -1.75$$

Thus we have

$$\delta^2 = \frac{n\sum_{i=1}^{4}\tau_i^2}{\sigma^2}$$

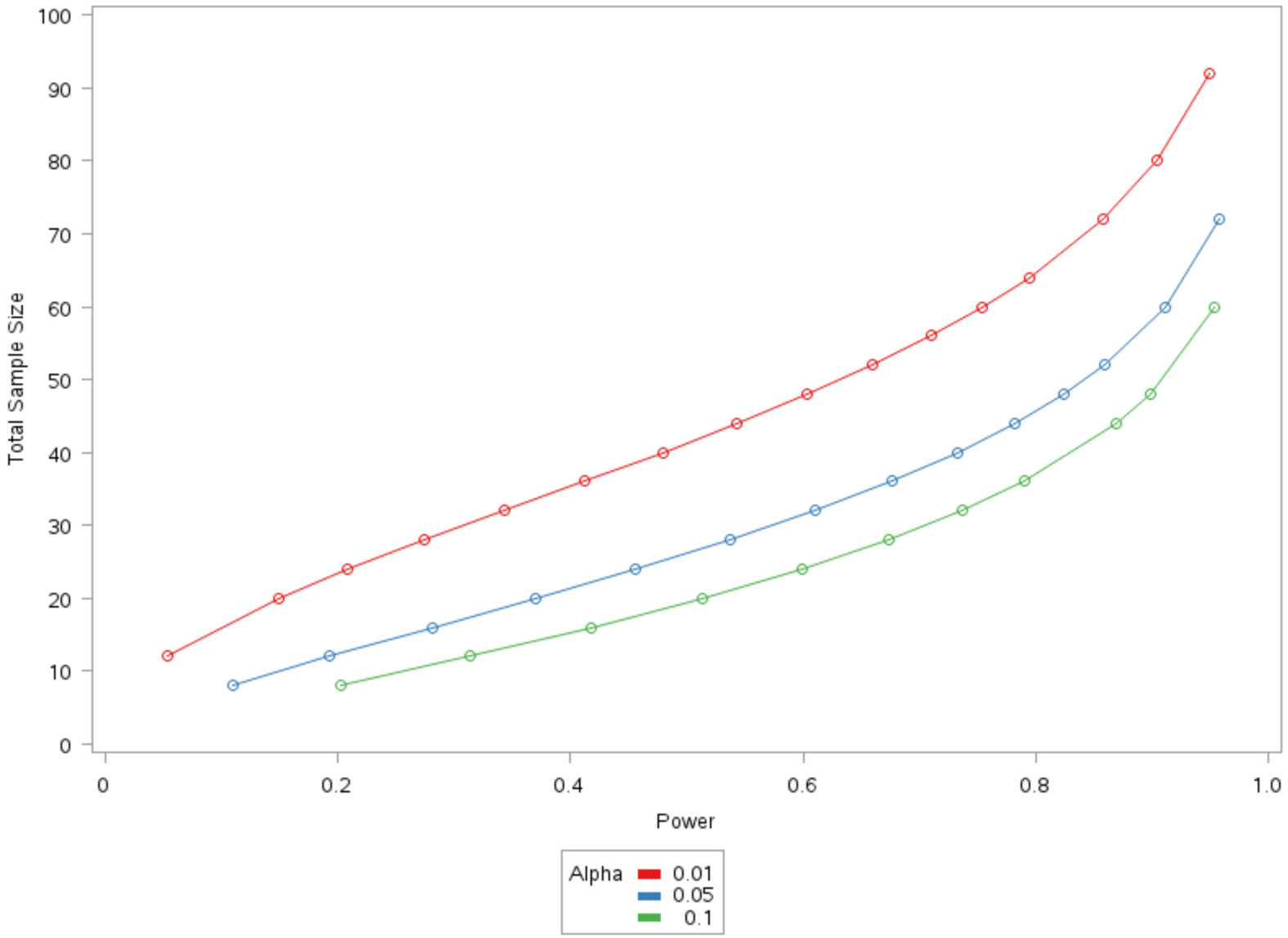
$$= \frac{n[(0.75)^2 + (3.0)^2 + (0.5)^2 + (1.75)^2]}{(3.5)^2}$$

$$= 1.092n.$$

The sample size can be determined using (11.4.4). To obtain the required sample size, we apply various n to (11.4.4). The results are summarized in Table 11.4.2. From Table 11.4.2, it can be seen that for $n = 11$ and $F^* = F(0.05, 3, 40) = 2.8387$, equation (11.4.4) yields $z(\beta) \approx 0.853$ which is the closest to $z(0.2)=0.842$. Therefore $n = 11$ is the required sample size per treatment group. Note that the required sample size $n = 11$ can also be obtained from Table 11.4.1 by specifying λ and ϕ.

Using SAS

```
proc power;
   onewayanova
      groupmeans = -0.75 | 3.0 | -0.5 | -1.75
      stddev = 3.5
      groupweights = (1 1 1 1)
      alpha = 0.01 0.05 0.1
      ntotal = .
      power = 0.8;
   plot x=power min=0 max=1
     vary (color);
run;
```


Total Sample Size
100
90
80
70
60
50
40
30
20
10
0
0
0.2
0.4
0.6
0.8
1.0
Power
Alpha
0.01
0.05
0.1

Survival analysis

- Assume we plan a RCT aiming at comparing a new treatment (drug) with an old one (control). Such a comparison can be made in terms of the hazard ratio or some function of it. Let *d* stand for the drug and *c* for control.
- In the sequel we describe the sample size required to demonstrate that the drug is better than the control treatment. This will be made under some assumptions in two steps:
    - Step 1: specify the number of events needed.
    - Step 2: specify the number of patients needed to obtain the number of events from step 1 (this is affected by the recruitment and follow-up times).

Step 1

- In what follows we will use the notation below to give a formula specifying the number of events needed to have a certain asymptotic power in many different cases like.

$p =$ The proportion of individual receiveing the new drug
$q = 1 - p =$ The proportion of individual receiveing the old drug
$d =$ the required number of events
$\mathrm{Z}(v) =$ The upper v quantile of the standard normal distribution

- Using the 2-sided log-rank test for the hazard ratio, the number of events required at significance $\alpha/2$ and power v is given by:

$$d = \frac{\left(Z\left(\frac{\alpha}{2}\right) + Z(v)\right)^2}{pq\beta^2}$$

- It is not clear how β should estimated. The difficulty lies in the fact that this depends on the exact situation at hand. We illustrate that through an example.

- Assume that our observations follow exp(λ) under the old drug (control) and exp(λ*exp*(β)) under the new drug. Then the ratio of medians of the two distributions will be

$$\frac{M_d}{M_c} = \frac{\lambda}{\lambda e^{\beta}} \frac{\ln(2)}{\ln(2)} = \frac{1}{e^{\beta}}$$

- To be able to discover a 50% increase in the median we take

$$\frac{M_d}{M_c} = \frac{1}{e^{\beta}} = \frac{3}{2} \Rightarrow e^{\beta} = \frac{2}{3}$$

1. It is possible to dress a table specifying β as a function of the remaing parameters when α=0.05 and $p=q$ and the test is two sided.
2. Observe that e^{β} and $e^{-\beta}$ lead to the same number of events.
3. The largest power is obtained when *p*=*q*.
4. When *p* and *q* are not equal divide the table value by *4pq*.
5. For a one sided test, $Z\left(\frac{\alpha}{2}\right)$ needs to be changed to $Z(\alpha)$

Using SAS

```
proc power;
     twosamplesurvival test=logrank
          groupmedsurvts=(1 1.15)
          accrualtime = 0
          followuptime = 1
          groupweights = (1 1)
          alpha=0.05
          power=0.5 0.7 0.8 0.90
          eventstotal=.;
run;
```

The POWER Procedure
Log-Rank Test for Two Survival Curves

Fixed Scenario Elements	
Method	Lakatos normal approximation
Form of Survival Curve 1	Exponential
Form of Survival Curve 2	Exponential
Accrual Time	0
Follow-up Time	1
Alpha	0.05
Group 1 Median Survival Time	1
Group 2 Median Survival Time	1.15
Group 1 Weight	1
Group 2 Weight	1
Number of Sides	2
Number of Time Sub-Intervals	12
Group 1 Loss Exponential Hazard	0
Group 2 Loss Exponential Hazard	0

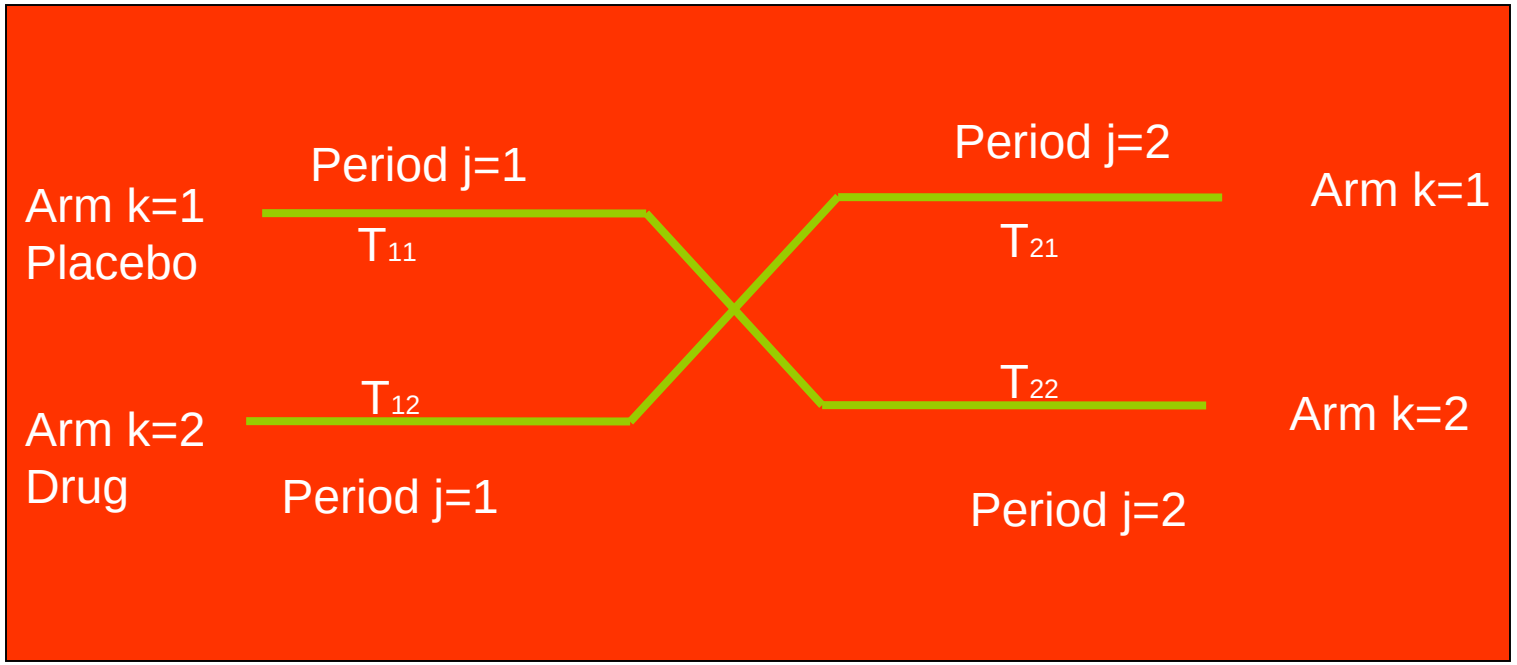
Computed Ceiling Event Total				
Index	Nominal Power	Fractional Event Total	Actual Power	Ceiling Event Total
1	0.5	772.186675	0.500	773
2	0.7	1240.799299	0.700	1241
3	0.8	1577.909004	0.800	1578
4	0.9	2112.376904	0.900	2113

Number of events needed for a certain power under equal allocation

		e^{β}				
		1.15	1.25	1.50	1.75	2.0
Power	0.5	773	304	93	50	33
	0.6	985	388	119	63	42
	0.7	1241	488	150	80	53
	0.8	1578	621	190	101	67
	0.9	2113	831	254	135	90

Crossover Designs

- In this type of studies we are interested in testing a null hypothesis of the form
  - H_0: $\mu_T = \mu_P$ against
  - H_a: $\mu_T \neq \mu_P$.



Crossover design

Paired t-test

- Assume Y_{ij} are the individual responses in groups *j=T,P*.
- For each subject calculate the treatment difference $d_i = Y_{iT} - Y_{iP}$
- Then $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} Y_{iT} - Y_{iP} = \overline{Y_T} - \overline{Y_P}$

$$var(\bar{d}) = var(\overline{Y_T} - \overline{Y_P}) = \sigma_T^2/n + \sigma_P^2/n - 2\sigma_T\sigma_P\,\rho/n$$

$$= \frac{2}{n}\sigma^2 - \frac{2}{n}\sigma^2\rho$$ where $\rho = corr(Y_{iT}, Y_{iP})$ and $\sigma_T^2 = \sigma_P^2$

$$\sigma_d = \sigma\sqrt{\frac{2}{n}(1-\rho)}$$

Then we can construct the test statistic

$$T_d = \frac{\overline{Y_T} - \overline{Y_P}}{\sigma_d} \sim T(2n-2)$$

Assuming equal sample sizes in both arms, and that σ_d can be estimated from old studies we have

$$n \geq \frac{2\sigma_d^2[t(\alpha/2, 2n-2) + t(\beta, 2n-2)]^2}{\Delta^2}$$

where Δ stands for the clinically meaningful difference.

Using SAS

```
proc power;
   pairedmeans test=diff
   meandiff = 2
   stddev = 3
   corr = 0.25 0.5 0.75 0.95
   power = 0.8
   npergroup = . ;
 run;
```

The POWER Procedure
Paired t Test for Mean Difference

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Mean Difference	2
Standard Deviation	3
Nominal Power	0.8
Number of Sides	2
Null Difference	0
Alpha	0.05

Computed N Pairs			
Index	Corr	Actual Power	N Pairs
1	0.25	0.808	29
2	0.50	0.807	20
3	0.75	0.804	11
4	0.95	0.933	5

When the sample size in each sequence group is 11, (a total sample size of 22) a 2 x 2 crossover design will have 80% power to detect a difference in means of 2, assuming an equal standard deviation of 3 in each group and a correlation of 0.75

Software

- For common situations, software is available
- Software available for purchase
  - NQuery
  - StudySize
  - PASS
  - Power and Precision
  - Etc…..
- Online free software available on web search
- http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize
- http://calculators.stat.ucla.edu
- http://hedwig.mgh.harvard.edu/sample_size/size.html