Contents

- Need to choose size of trial to give a reasonable chance of detecting differences of clinical interest
- Unethical to conduct trial with
  - Little chance of reaching conclusion
  - More patients than necessary to reach decision

Reasonable chance?

- 80%? 95%? 70%? 90%? ?????
  - Essentially the choice of the clinician
  - Higher the chance → the greater the size
- **80% common benchmark standard**

- **Difference of clinical interest?**
  - Solely responsibility of clinician
  - Clinicians are over-ambitious
    - Start by saying '***any difference***'
  - Smaller desired detectable difference → the greater the size

- Detecting a difference of size δ:
  - obtaining a significant result when there is indeed a difference δ
    - depends on significance level used

- **Specifications required**
  - main outcome measure
  - method of analysis
    - e.g. two-sample t-test
  - result on standard treatment
    - or pilot results
  - minimum difference required to detect (δ)
  - degree of certainty with which we wish to detect it (i.e. the ***power: 1 – β***)

- **non-significant difference**
  different from
- **no clinically relevant difference**
  - ‘non-significant difference’
    - statistical evidence is not convincing
  - no clinically **relevant** difference
    - if any difference then not medically important

- Types of Error
  - Type I: false positive
    - treatments equivalent but result significant
      - α represents risk of false positive result
  - Type II: false negative
    - treatments different but result non-significant
      - β represents risk of false negative result
      - ***power* = 1 – β**
  - ***c.f. sensitivity & specificity***

- **Sensitivity & Specificity:**

		Disease Status		
		Positive	Negative	Total
Test	Positive	a	b	a+b
	Negative	c	d	c+d
	Total	a+c	b+d	n

Sensitivity= a/(a+c) Specificity=d/(b+d)

- **Sensitivity & Specificity:**
  - Sensitivity:
    - probability test ‘correctly identifies’ null hypothesis (=***1 – α***)
  - Specificity
    - probability test ‘correctly rejects’ null hypothesis (=***1 – β***)

- **General Method:**
  - Determine criterion for declaring a significant result:
  - i.e. value of test statistic > critical value
    - e.g. z-value > 1.96
  - **NB:** P[test statistic > critical value when no difference] = significance level
    - P[test stat>crit val | H_0 true] = α
      - P[z > 1.96 | H_0 true] = 0.025

- Find probability of significant result when difference is δ
  - i.e. the **power** of the test when difference is δ
- This probability depends on δ and n
- Relates **power,** δ and **n** in one equation
- Given 2 of these we can calculate 3rd.

- e.g. H_0: $\mu_1 = \mu_2$ *vs* H_A: $\mu_1 \neq \mu_2$
  - Let $\delta = \mu_1 - \mu_2$
    - then H_0: $\delta = 0$
  - Test statistic z, reject H_0 at level α if $z > c_\alpha$
- Then $P[z > c_\alpha \mid \delta = 0] = \alpha$
- Power when difference is δ is
  $P[z > c_\alpha \mid \delta]$
  - **NB:** the power depends on δ
    - (& on sample size **n**)

- Can calculate n for given power and δ
- Or power for given n and δ
- Or δ for given power and n

- **Often n is limited by practicalities**
  - Often clinician asks about **n**
  - Finds required **n** is unrealistic
  - Tells you what is a realistic **n**
  - You can give a table of obtainable **power** and **δ** for this maximum sample size n

- Testing equality of two binomial proportions θ_1 and θ_2
  - (using Normal approximation)
- General formula

$$n \approx \frac{\theta_2(1-\theta_2)+\theta_1(1-\theta_1)}{(\theta_2-\theta_1)^2}\left\{\Phi^{-1}(\beta)+\Phi^{-1}(\alpha/2)\right\}^2$$

- Can use to find any one of
  n, β or θ_2
  from the other 2

- **N.B. both $\Phi^{-1}(\beta)$ and $\Phi^{-1}(\alpha/2)<0$**
  - Be careful if inverting formula and need to take square roots

- Approximation requires

$$\frac{\sqrt{2\theta_1(1-\theta_1)}}{\sqrt{\theta_2(1-\theta_2)+\theta_1(1-\theta_1)}} \approx 1$$

- Here = 1.14, so reasonable
  - Packages may use exact formula

- Term $\left\{\Phi^{-1}(\beta)+\Phi^{-1}(\alpha/2)\right\}^2$
  - is the same for many different tests
  - only multiplier differs

- If we can really justify a 1-sided test (e.g. from a pilot study)
- then put $\Phi^{-1}(\alpha/2) \rightarrow \Phi^{-1}(\alpha)$.
- However, this MUST be specified in the protocol
  - Cannot decide after trial has begun
  - Definitely not after collecting data

- 1–sided testing
  - Reduces the required sample size
  - Increase the power
  - Reduces the CRD

- e.g. a p-value of 8% on a 2-sided test becomes 4% if test is converted to 1-sided
- – cannot decide to use 1-sided just because a 2-sided p-value is 8%

- For given α and β, n depends [inversely] on $(\theta_2 - \theta_1)^2$
- **half** the CRD needs
  - **C**linically **R**elevant **D**ifference
- **fourfold** increase in trial size.

- Increase in power from 50% to 95% needs sample size $\times$ 3
- NB In practice need to allow for a drop-out rate so need to inflate sample size to allow this:
  - e.g. if 10% drop-out rate expected then increase sample size by 11.1% [=1/(1-10%)]

- Only possible to do sample size calculations in simplified situations
  - Cannot take into account more complex statistical tests that are actually likely to be used.
    - e.g cannot do sample size calculations for linear regression or ANOVA
    - Need to use calculations from a simplified situation as a guide

- Typically sample size is limited by practical situations
- More useful to find the power of a study for a range of sample sizes and CRDs

- Continuous data
  - Same general method
  - Need to know standard deviation σ
    - Pilot study or published data
    - ¼ × likely range
      - Guess at maximum and minimum possible

- Testing equality of two normal means μ_1 and μ_2
  - (using Normal approx to t-distribution or known variance)
- General formula

$$n = \frac{2\sigma^2}{(\mu_2 - \mu_1)^2}\left\{\Phi^{-1}(\beta) + \Phi^{-1}(\alpha/2)\right\}^2$$

- Same term as before
- Same dependence on CRD

- **Computational Note**
  - Sample size facilities in **R,** S-PLUS, Minitab & StatsDirect
  - These have different default choices & may give slightly different answers
  - Tables by Machin & Campbell (1997) provide a disk with SampSize (V2.0)
  - nQuery Advisor is industry standard
  - Programme POWER.EXE available through course web page

- **Implementation in R**
- Three functions
  - `power.t.test()`
  - `power.prop.test`
  - `power.anova.test()`
  - Each will find any of power, sample size or CRD from the other 2

Example:

```
> power.prop.test(p1=0.9,p2=0.95,n=581,sig.level=0.05)

     Two-sample comparison of proportions power
calculation

              n = 581
             p1 = 0.9
             p2 = 0.95
      sig.level = 0.05
          power = 0.8999597
    alternative = two.sided

 NOTE: n is number in *each* group
```

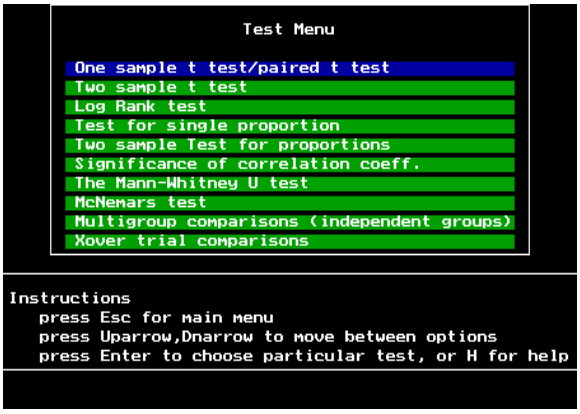
Find power

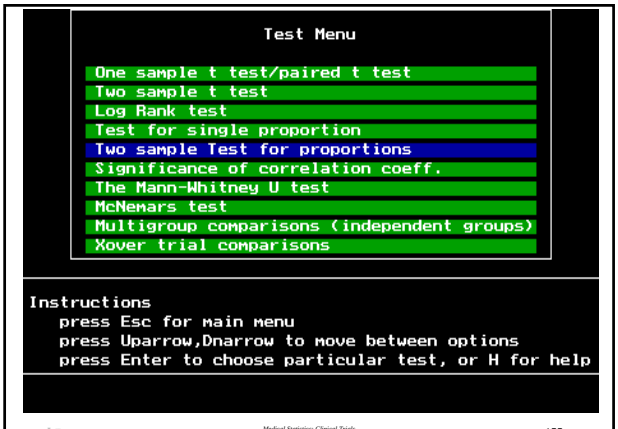
Example:

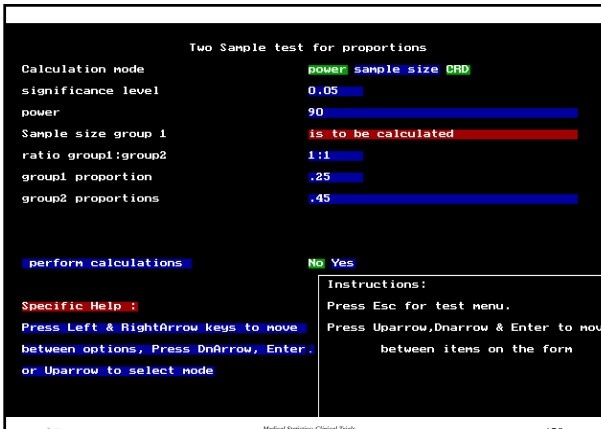
```
>
power.t.test(delta=1.6577,sd=3.6,power=0.8,sig.level=0.05)

     Two-sample t test power calculation

              n = 75.00648
          delta = 1.6577
             sd = 3.6
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

 NOTE: n is number in *each* group
```

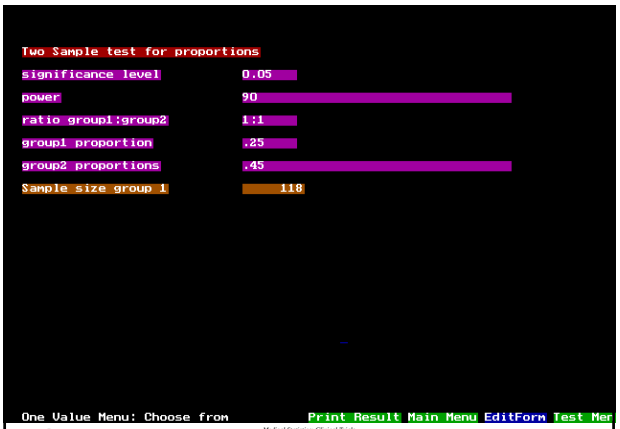
Find n

Programme POWER.EXE available from course web page









- **Summary and Conclusions**
  - Samples too small
    ⇒ little chance of conclusion
    ⇒ exposing patients to risk with no benefit
  - Samples too large
    ⇒ may expose too many subjects to inferior treatment

- For sample size calculation need:
  - outcome measure
  - method of analysis
    - including significance levels
  - CRD
  - power
  - results on standard treatment
    - including likely variability

- Practical Limitations:
  - Drop-outs
    - Inflate sizes to allow for this
  - Use simplified version of statistical design as a guide to sample sizes
  - Consider effects on power and CRD for practical range of sample sizes