Practical issues in study conduct

Sample size estimation

Database design EPIDATA

Data analysis – simple descriptive EPICALC2000

Epicalc 2000

- Freeware program
- Useful for quickly checking data in the literature
- Great for simple sample size calculation

- http://www.brixtonhealth.com/epicalc.html

Sample Size Estimation

- Design stage
- Ensure that proposed number of subjects recruited appropriate to answer main objectives of the study
  - Required for grant proposal
  - Required by ethics committee

- Too few patients
  - may fail to detect important effects (clinically important difference)
  - or estimate them too imprecisely no matter how good the design of the study in other respects
- Too many patients
  - Waste resources
  - Therapy may have risks

- All else being equal a larger sample size increases precision in estimates
- Law of large numbers
  - The average of results obtained from a large number of trials should be close to the expected value
  - What is the expected value of the roll of a fair dice?

How big a study? $\Delta\sigma\alpha\beta$

- Δ = difference between two groups that is clinically significant
    - The bigger the difference the smaller the n!
- σ = variability of the estimate
    - This comes from prior studies/pilot studies
    - The smaller the variability the smaller the n!
- α = 0.05 unless you are desperate!
    - The higher the type I error the smaller the n!
- β = 0.2, (power of 80%, power is the likelihood that the study will yield a statistically significant result)
    - The higher the type II error the smaller the n!

Example – sample size of a mean or mean difference

- Precision = Sample error = half the width of the 95% CI
- Precision = 5
- Standard deviation = 15

Sample - Precision - Single mean

SD	:	15.00
Precision	:	5.00
Confidence level	:	95%
Sample size	:	34

Describe - Mean

Sample size	:	34	
SD	:	15.00	
Mean [95% CI]	:	100.00	[94.77, 105.23]

Describe - Mean

Sample size	:	34	
SD	:	15.00	
Mean [95% CI]	:	10.00	[4.77, 15.23]

Sample - Precision - Single mean

SD	:	30.00
Precision	:	5.00
Confidence level	:	95%
Sample size	:	138

Sample - Precision - Single mean

SD	:	15.00
Precision	:	1.00
Confidence level	:	95%
Sample size	:	864

Means and mean differences

- It doesn't matter what the mean is – you are simply interested in describing the mean with a given margin of error
- Same result if you are calculating a mean difference on paired samples
- For independent samples and a mean difference simply use the pooled estimate of the standard deviation

Estimating a single proportion

- Calculate a 95%CI for a proportion with a margin of error of x
- Proportion is 50%
- Margin of error is 10%
- Note, here the proportion being estimated does influence the sample size – the largest sample size is required for a proportion of 50%, so this is recommended as the most ‘conservative’ estimate of sample size

Sample - Precision - Single proportion

Proportion	:	50.00%
Precision	:	10.00%
Confidence level	:	95%
Sample size	:	96

Sample - Precision - Single proportion

Proportion	:	50.00%
Precision	:	5.00%
Confidence level	:	95%
Sample size	:	384

Sample - Precision - Single proportion

Proportion	:	40.00%
Precision	:	5.00%
Confidence level	:	95%
Sample size	:	368

Testing two proportions based on a cohort or cross-sectional study

- Incidence of 10% in one group compared with incidence of 5% in other group

OR

- Prevalence of 10% in one group compared with prevalence of 5% in another group

Sample - Size - Two proportions

Proportion 1	:	10.00%
Proportion 2	:	5.00%
Significance	:	0.05
Power	:	80%
Sample size (each group)	:	433
Sample size (overall)	:	866

Example - to assess the value of a new cancer treatment

- 200 women with newly diagnosed breast cancer were randomly allocated to receive either the new or the standard treatment
- All patients followed up for 1-year or until death occurred
- The outcome of interest was the proportion of women still alive at the end of the trial

RCT new breast cancer drug

		New treatment	Standard treatment	Total
Alive after 1 year after entry into trial	Yes	80	70	150
	No	20	30	50
Total		100	100	200

P1=80/100*100 = 80%
P2 = 70%
Risk difference = 10%

Compare - Two proportions - Counts and sample sizes

Sample 1

Count	:	80
Proportion	:	80.00%
Sample size	:	100

Sample 2

Count	:	70
Proportion	:	70.00%
Sample size	:	100

Difference

Difference	:	10.00	[-2.92, 22.92]
Z	:	1.47	
One-sided p-value	:	0.070822	
Two-sided p-value	:	0.141645	

Sample - Size - Two proportions

Proportion 1	:	80.00%
Proportion 2	:	70.00%
Significance	:	0.05
Power	:	80%
Sample size (each group)	:	292
Sample size (overall)	:	584

Example - to increase survival by 5%

- New cancer treatment: 90%
- Standard cancer treatment: 85%
- Wish to have a 80% chance of finding this difference

Sample - Size - Two proportions

Proportion 1	:	90.00%
Proportion 2	:	85.00%
Significance	:	0.05
Power	:	80%
Sample size (each group)	:	684
Sample size (overall)	:	1368

Testing an Odds Ratio from a Case-Control Study

- Alternative hypothesis is OR = 1.0
- Significance level – default is 0.05
- Required power – default is 80%
- 1:1 ratio of cases to controls
- Odds ratio worth detecting = 3
- Prevalence of exposure (in the controls) = 25%

Sample - Size - Case-control study

Ratio of cases to controls	:	1	
OR to detect	:	3.00	
Proportion (%) controls exposed	:	25.00%	
Significance	:	0.05	
Power	:	80%	
Sample size	:	57	(cases)
	:	57	(controls)
	:	114	(overall)

Sample - Size - Case-control study

Ratio of cases to controls	:	4	
OR to detect	:	3.00	
Proportion (%) controls exposed	:	25.00%	
Significance	:	0.05	
Power	:	80%	
Sample size	:	36	(cases)
	:	144	(controls)
	:	180	(overall)

Case-Control Study

		Lung cancer	No lung cancer	Total
Smoking	Yes	200	100	300
	No	200	300	500
Total		400	400	800

Odds of exposure in diseased 200/200 = 1
Odds of exposure in not diseased 100/300 = 0.33
Odds ratio 1/0.33 = 3

Tables - 2-by-2 unstratified

	+	-	Total
+	50	25	75
-	50	75	125
Total	100	100	200

Tests of significance

Fisher exact test (one tailed)	:	0.000210
Fisher exact test (two tailed)	:	0.000420
Uncorrected chi-square	:	13.33
p-value	:	0.000001
Yates corrected Chi-square	:	52.27
p-value	:	0.000261

Measures of exposure effect [95% CI]

Risk ratio	:	1.67	[1.28, 2.18]
Odds ratio	:	3.00	[1.65, 5.46]

Describe - Proportion - Percentage

HIV prevalence among 20-24 year olds in 2008 HSRC survey

Sample size	:	910	
Proportion [95% CI]	:	25.20%	[22.43, 28.18]

Describe - Proportion - Count and sample size

Count	:	22	
Sample size	:	100	
Proportion [95% CI]	:	22.00	[14.58, 31.61]

Other freeware software

- OpenEpi
- PS: Power and Sample Size

Database design

- Junk in = junk out
- Carefully designed questionnaire

EpiData

- Freeware
- EpiData Analysis
- http://www.epidata.dk

- Variable names
  - 10 characters, start with a letter
  - Lower case – for analysis in STATA
    - Two methods
      - Meaningful words (“age”, sex”, “bloodtype”)
      - Numbered (q1,q2,q3 or a1,a2,a3,b1,b2,b3 etc)
  - Keep a coding sheet – will demonstrate later

Epi Data

- Confidentiality/anonymity
  - Database should not contain identifying information such as names and addresses (should be kept in separate document)

Epi Data

- Categorical variables
  - Fixed set of responses with only one correct response
    - Sex: 1=male, 2=female
    - ABO group: 1=O, 2=A, 3=B, 4=AB
    - Do not use Sex: m=male, f=female – will have to be recoded before analysis

Epi Data

- Categorical variables
  - If multiple selections allowed, each option is treated as a separate binary variable
    - Which health care practitioners are involved in managing your diabetes? (tick as many as apply)
      - Endocrinologist
      - GP
      - Nurse
      - Dietician
      - Other (please specify)

Epi Data

- Continuous and discrete variables
  - Contain numbers only
    - Height or weight (continuous)
    - Number of people in household (discrete)
  - Often have reasonable limits that can be specified when setting up database to limit errors in data entry
  - Be consistent in use of units (cm or m)

Epi Data

- Date variables
  - Vary around the world so use either
    - day-month-year or
    - year-month-day
  - Be consistent

Epi Data

- String variables
  - Can contain text and numbers
  - Often used for “open-ended” questions, consequently usually need to be coded for analysis
  - Maximum length of 80 characters, but can use more than one variable to represent the string
    - commment1, comment2, comment3 etc

Epi Data

- Missing data
  - Is represented by special codes to indicate something is missing so we can tell the difference between a response that was really missing from the questionnaire and one that has been accidentally skipped during data entry
    - Often use “9” or “99” or “9999”
    - Must not be a valid response
  - All fields should have data in them unless they have been skipped because of a conditional jump

Epi Data

- Minimising errors
  - Set limits
    - Continuous variables
  - Categorical variables
    - Specify valid values and restrict database to accept only those
      - E.g. Sex valid codes are 1, 2 and 9
  - Double entry

Epi Data

- Types of error
  - Transposition, e.g. 39 becomes 93
  - Copying errors e.g. 1 as 7 and O as 0
  - Coding errors (Corrected by questionnaire design to include coding rather than doing this after data has been collected)
  - Routing errors – questions asked in wrong order (Corrected by questionnaire design and training interviewers)
  - Consistency errors – two or more responses are contradictory – program consistency checks
  - Range errors - program possible/probable ranges

Epi Data

- Questionnaire file
- Check file
- Record file

Example using a simple questionnaire

My little study

1. Unique ID number
2. Hospital number
3. Date of birth / /

   dd/mm/yyyy
4. Height (in cm)
5. Sex

   1 Male, GO TO Question 5

   2 Female

6. Have you ever been pregnant

1 Yes

0 No

7. Blood Group

1 O

2 A

3 B

4 AB

8 Don’t know

9 Missing

8. Which health practitioners are involved in managing your diabetes (tick as many as apply)

Endocrinologist

GP

Nurse

Dietician

Other (please specify) _____

Epi Data

- Avoid using the following characters

  @

  <,>

  “-”

```
BEFORE FILE
  HELP "Instructions for data entry. \n
Please enter the following data as it
appears on the completed questionnaire."
  HELP "Most variables require just a
numeric code, some require free text."
  HELP "Numeric codes are shown in a
drop-down box, \n then either clicking the
correct code will enter it into the field, \n or
you can enter the number yourself."
  HELP "Free text can be entered in lower
case - it will automatically be recorded as
all capitals. "
  HELP "After entry of data you must press
ENTER for the cursor to jump to the next
field, \n in some cases the cursor is
programmed to jump over questions, \n to
follow skip patterns in the questionnaire."
  CONFIRM
END
```

TYPE COMMENT
COMMENT LEGAL USE LABEL_v5 SHOW

Epi Data

- Prepare for double entry

THE END

Simple formulas

- Two means (continous dependent variable): $n=16(\sigma^2/\Delta^2)$
- Two proportions: $n=4\sqrt{p}(1-p)/(P_p-P_d)$
- Difference among many means (ANOVA):usually choose one comparison between means we care most about and use original formula
- Relation between two continuous variables:

  $n=4 + (8/r)$
- Relation between many variables: rule of thumb 5-10x number of variables!

Example

- Objective: To see if feeding milk to 5-yr olds enhances growth
- Study design: Randomised trial with 2 arms (standard milk diet and extra milk diet)
- Outcome: height (cm)
- Clinically significant difference: 0.5cm

Beta of the test

- Define degree of certainty of finding this difference (beta(ß) or type II error
  - Probability of NOT detecting a significant difference when there is one
  - Risk of a false-negative finding
  - Set at <=20%

Power of test

- Probability of detecting a clinically significant difference
- Power=(1-ß)=1-20%=80%

Define significance level

- Alpha (α) or type I error
- The probability of detecting a significant difference when the treatments are really equally effective
- Risk of a false-positive finding
- 5%

For the milk study

- Type I error (α) = 0.05
- Type II error (β) = 0.20
- Power = (1- β) = 0.80
- Clinically significant difference (Δ) = 0.5cm
- Measure of variation (SD) = 2.0cm
  - Exists in literature/guesstimate/pilot data

$$N=\frac{2(SD)^2}{\Delta^2} \times f(\alpha,\beta)$$

$$N = \frac{2(SD)^2}{\Delta^2} \times f(\alpha,\beta)$$

- Sample size:
- Directly proportional to the variation in measures
- Inversely proportional to the size of the clinically relevant difference

Sample - Size - Two means

```
Mean 1
:       0.50
Mean 2
:       0.00
SD
:       2.00
Significance
:       0.05
Power
:      80%

Sample size
:     250          (each group)

:     500          (overall)
```

Assumptions

- Random sample
- Independent observations
- Variability of the population known (standard deviation)

- As n becomes very large, the variability decreases
- Leads to more sensitive hypothesis tests, greater statistical power, and smaller confidence intervals
- Power=1-ß

Methods

- Web based
  - Online – not recommended as website might ‘disappear’ and you will not be able to reproduce your sample size estimate
  - Freeware that can be downloaded is recommended
    - EpiCalc 2000
    - PS Power and Sample Size (more sophisticated)
- Specialised software
- Tables/Nomograms
- Formulae

What can we do to reduce the sample size?

- Use more complicated sampling techniques
  - Stratified sampling works by reducing the variance of the sample estimates