

Research Education Seminar Series

Lecture 2



Alex Stephens, PhD, Director of Research

Basic outline

1. *Research Overview*

- Overarching aims of the research office
- Why do we engage in research?
- Simplified research framework
- Research translation framework
- Evidence-based practice pyramid

2. *Robust study design*

- Epidemiological methods – study designs
- Key elements of study validity and critical appraisal
- Measures of association
- Determining sample size – power analysis

3. *Analysis*

- Basic biostatistical methods and analysis

4. *Writing for research*

- Writing grants, papers and scientific presentations

5. *Ethics, governance and software*

- Research ethics and governance
- Research software

Epidemiological methods – study designs

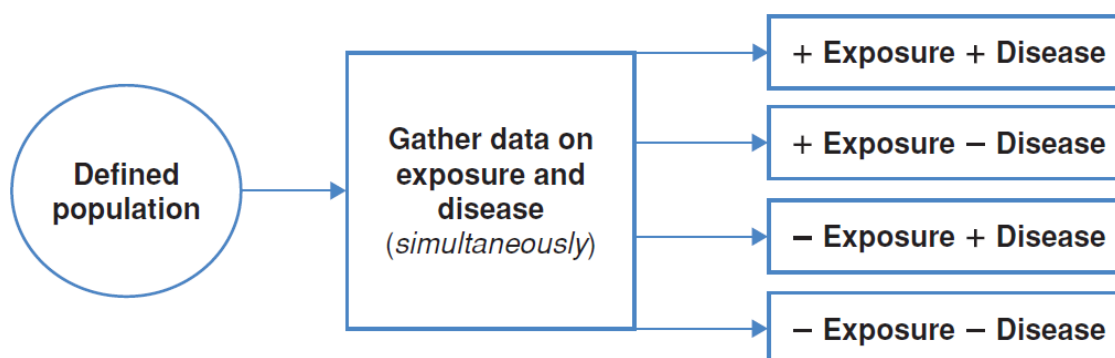
Main study designs:

- Case studies/series
- Cross-sectional (point in time) studies
- Case-control studies
- Observational (cohort) studies
- Randomised Controlled Trials
- Systematic reviews (and meta-analysis)

Case study/series

- The collection of information and detailed presentation of findings about a particular patient/participant or a small group of patients/participants
- These are descriptive studies in nature (i.e. are not analytical) and typically will not involve a hypothesis. However, they may be hypothesis generating.

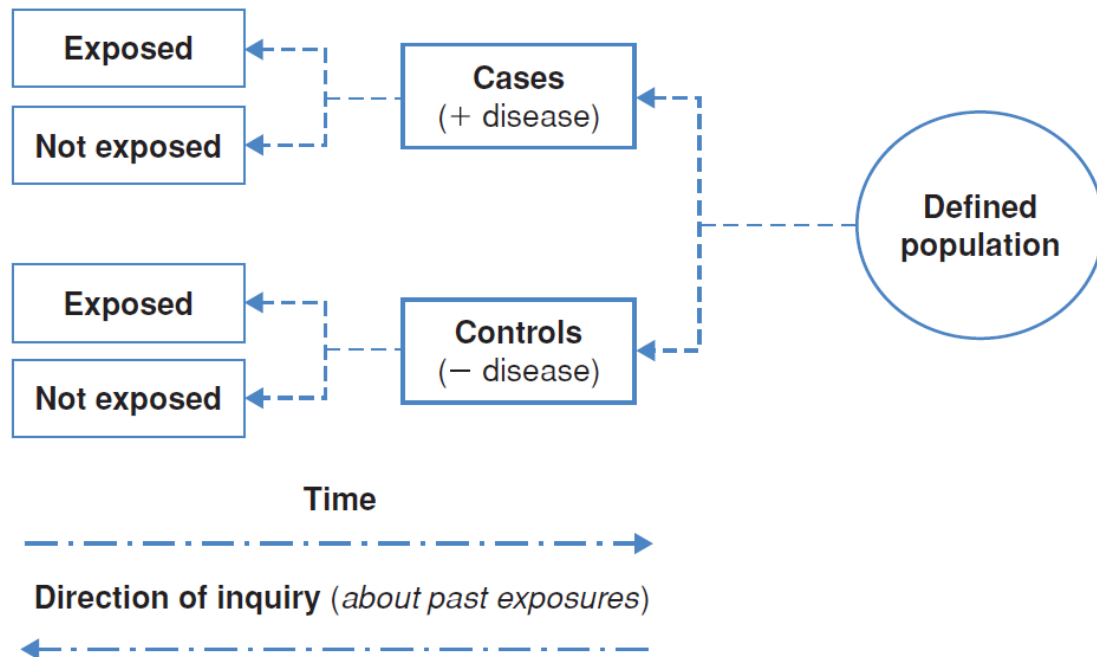
Cross-sectional studies



Adapted from Figure 4.1 *Essential Epidemiology, An Introduction for Students and Health Professionals*, Cambridge University Press, 2011

- Information is gathered from a defined population, or typically a representative sample, where data on exposure and disease/outcome is collected at the same time (i.e. simultaneously).
- Cross-sectional studies can be descriptive – providing information on the prevalence and magnitude of disease/health outcomes – and they can also be analytical, where simple or sophisticated statistical techniques can be applied to estimate associations between (potential) predictive factors and outcomes (in the form of odds ratios, relative risks, rate ratios, mean differences...).
- Surveys are a good example of cross-sectional studies (e.g. NHANES in the US)

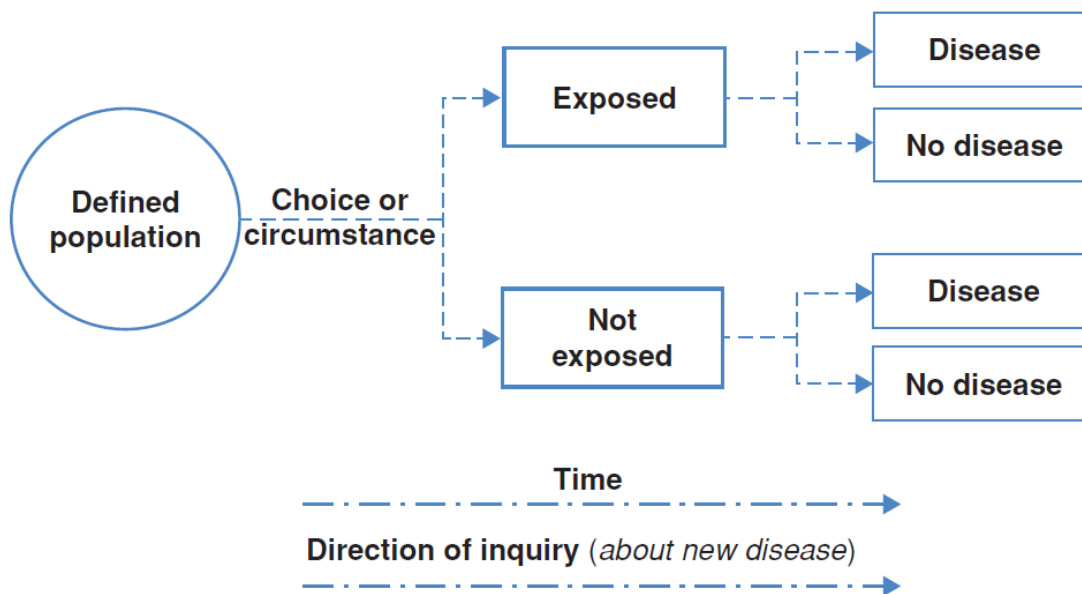
Case-control studies



Adapted from Figure 4.3 *Essential Epidemiology, An Introduction for Students and Health Professionals*, Cambridge University Press, 2011

- As the name suggests, participants are selected who have developed the disease/outcome of interest (the cases) and a representative sample from the same population as the cases (but which do not have the disease/outcome) are selected as the controls.
- For example, a register of disease (e.g. cancer) could be used as the source of cases, and a sample of the electoral role for the same catchment area (e.g. the same state as the cancer register) could be used to obtain a sample of controls without disease.
- Once cases and controls are selected, we work backwards to determine which exposures may be different between the two groups. For example, in a case-control study of cancer, we would ask participants about previous exposure to toxins/carcinogens and compare exposure between the cases and controls.

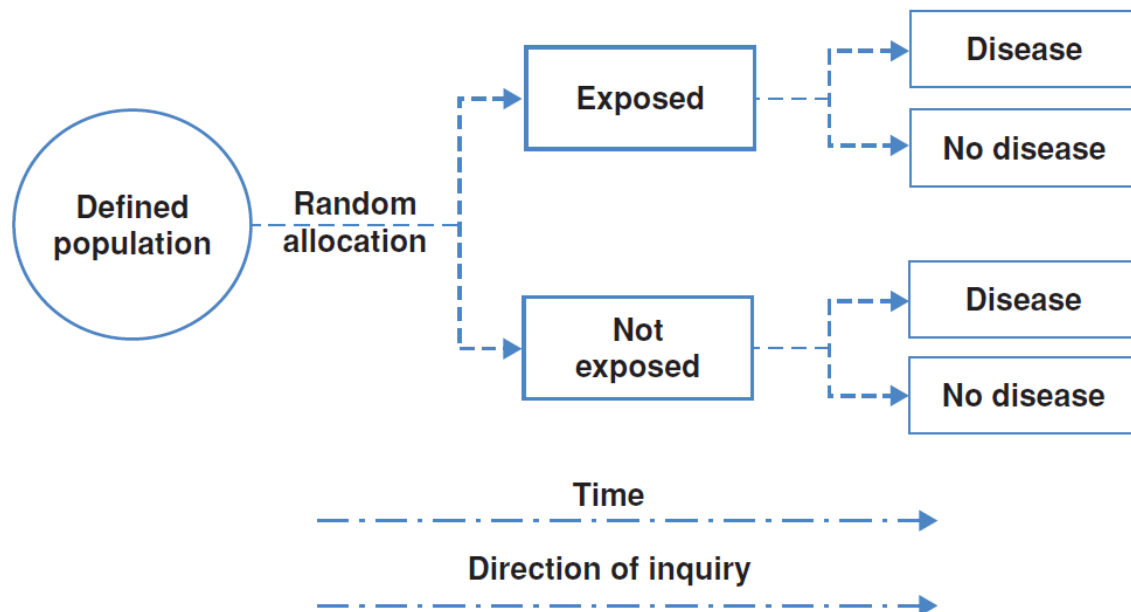
Cohort studies



Adapted from Figure 4.2 *Essential Epidemiology, An Introduction for Students and Health Professionals*, Cambridge University Press, 2011

- Cohort studies are also referred to as prospective or longitudinal studies. In these studies, we start with a defined study population and observe them over time to see what happens to them. Whether a participant is exposed or not is dependent on choice or circumstance (and is not imposed). Analyses are conducted to determine if any associations exist between exposures and outcomes.
- The Framingham Heart Study is a good example of an early and very long running cohort study. Modern cohort studies have become massive and can include hundreds of thousands of participants (e.g. European Investigation into Cancer – EPIC – includes more than 500,000 individuals).

Randomised Controlled Trials



Adapted from Figure 4.4 *Essential Epidemiology, An Introduction for Students and Health Professionals*, Cambridge University Press, 2011

- Randomised controlled trials (RCTs) are the classic, “gold standard” for generating new evidence regarding an intervention.
- In a RCT, a group of patients that meet the inclusion criteria are randomly allocated to receive either the intervention or placebo, and are followed up to evaluate outcomes.
- As treatment allocation is random, on average, the distribution of important covariates and confounding factors should be equal/similar between intervention and placebo groups – i.e. the design ensures that groups are as similar as possible to start with, and any important differences, apart from the effect of treatment, occur by chance.
- The quality of evidence generated from RCTs is highly dependent on appropriate selection (inclusion criteria), randomisation and blinding.

Measures of frequency and association

Measures of frequency

The occurrence of an outcome or disease can be measured several ways. Two common ways are **incidence** and **prevalence**. Although related, these two measures are distinguished by time of onset and the nature of the outcome (some outcomes are incident in nature – e.g. heart attack – and others are prevalent in nature – e.g. osteoarthritis).

Incidence is defined as the number of new cases of an outcome (within a defined time interval and population). For example, the number of new cases of Hepatitis C in a given year for the NSW North Coast population → in epidemiology, this is termed the **cumulative incidence**.

$$\text{Incidence} = \frac{\text{Number of new cases in a specified period}}{\text{Total population at risk at start of specified period}}$$

Prevalence is the total number of existing and new cases of a disease/outcome in a given population. For example, total number of individuals with active Hepatitis C infection for the NSW North Coast population (typically at a defined date/year).

$$\text{Prevalence} = \frac{\text{Number with disease (or outcome) at specific point in time}}{\text{Total population at risk at specific point in time}}$$

Measures of incidence and prevalence come in several forms and include:

- **Count:** The number of individuals that meet the case definition
- **Proportion:** The number of individuals that meet the case definition expressed as a proportion of the total population
- **Rate:** The number of individuals that meet the case definition for every X number of the population; e.g. per 1,000 or per 100,000 population

Measures of association

The main objective of epidemiology is to identify causes of disease. This concept can be extended to other non-disease outcomes (e.g. recovery times from surgery, patient satisfaction with a health service). Often, we cannot infer causation from an analysis, but we can determine associations and their strength. Associations between outcomes and exposures (study factors) can be represented a number of ways.

Ratio-scale measures

These are relative measures which compare the incidence or prevalence of an outcome between two (exposure) groups.

The **risk ratio** (often called the **relative risk**) is defined as the ratio of the cumulative incidence (CI) of an outcome in the exposed (or intervention) group divided by the cumulative incidence of an outcome in the unexposed (or control) group.

$$\text{Relative risk (RR)} = \frac{CI_{\text{exposed}}}{CI_{\text{unexposed}}} = \frac{CI_e}{CI_o}$$

Example – cumulative incidence of congenital conditions in newborn infants by exposure to environmental toxins

Table: Number of newborns with and without congenital defects by exposure to environmental hazard

	Defects	No defects	Total
Exposed	75	75	150
Unexposed	50	400	450
Total	125	475	600

$$CI_o = \frac{50}{450} = 0.11$$

$$CI_e = \frac{75}{150} = 0.50$$

$$RR = \frac{0.5}{0.11} = 4.5$$

The risk of congenital conditions in newborns is **4.5 times** greater in mothers living in areas exposed to environmental hazards compared to those living in areas not exposed to environmental hazards.

A **rate ratio** is analogous to a risk ratio except that rates (e.g. number of new cases of myocardial infarction per 1,000 person-years of follow-up) are compared.

The **prevalence ratio** is given by the prevalence of disease (or outcome) in the exposed population divided by the prevalence of disease (or outcome) in the unexposed population.

$$\text{Prevalence ratio} = \frac{\text{Prevalence}_e}{\text{Prevalence}_o}$$

A measure closely related to the relative risk is the **odds ratio**. The odds ratio is commonly used in case-control studies, and is a ratio of the odds of disease between exposed and unexposed groups.

$$\text{Odds ratio (OR)} = \frac{\text{Odds}_e}{\text{Odds}_o}$$

Where

$$\text{Odds} = \frac{\# \text{ Cases}}{\# \text{ Non - cases}}$$

When applied to our **congenital conditions example**

Unexposed group

$$\text{Odds}_o = \frac{50}{400} = 0.125$$

Exposed group

$$\text{Odds}_e = \frac{75}{75} = 1$$

Odds ratio

$$\text{OR} = \frac{1}{0.125} = 8$$

Tip: When the proportion of an outcome is low ($\sim < 0.1$), the *OR* and *RR* will give similar results.

Difference or attributable scale measures

A limitation of ratio scale measures is it tells us little about the actual burden or amount of disease (or outcome). This is where difference and attributable measures become useful.

The **risk (or rate) difference** is simply the difference in the risk between exposed and unexposed groups.

$$\text{Risk difference} = \text{risk}_e - \text{risk}_o$$

The risk difference is sometimes referred to as the **attributable risk (AR)**, and is interpreted as the excess risk of disease that occurs in the exposed group.

An extension of the attributable risk is to determine the percentage or proportion of the total risk in the exposed group that is associated with (or a consequence of the) exposure i.e. the excess risk of disease in the exposed group (compared to unexposed) divided by the total risk of disease in the exposed group. This is called the attributable fraction or attributable percent.

For our **congenital conditions example**

The cumulative incidence or proportion of infants born with congenital conditions in the unexposed group was 0.11 (11%). The proportion of infants with congenital conditions in the exposed group was 0.50 (50%). Therefore, the excess risk of congenital conditions in the exposed group is $0.50 - 0.11 = 0.39$. **The attributable risk percent is $0.39/0.50 = 0.78$** . Therefore, this suggests that 78% of the risk of congenital conditions is associated with exposure to environmental toxins.

Population attributable risk (PAR) is an extension of the AR and places the excess risk associated with exposure in the context of the total burden of disease/outcome in the population at risk.

For our **congenital conditions example**

The excess risk of congenital conditions in the exposed group was 0.39. Therefore, the excess number of infants born with congenital conditions in the exposed group (above that expected based on non-exposed levels) is $0.39 \times 150 = 58.33$. The PAR is the proportion of congenital conditions in the entire population associated with the excess amount due to exposure, and is equal to: $58.33 / (75 + 50) = 0.467$. Therefore, 46.7% of all congenital conditions are attributable to exposure to environmental hazards.

Simple formula for PAR using relative risk:

$$PAR(\%) = \frac{F(RR - 1)}{1 + F(RR - 1)} \times 100$$

Where F is the fraction of the total population exposed

For our example:

$$F = 150/600 = 0.25$$
$$PAR(\%) = \frac{0.25(4.5 - 1)}{1 + 0.25(4.5 - 1)} \times 100 = 46.7$$

Key elements of study validity and critical appraisal

These elements are what is used to assess (appraise) the quality of research, and determine where the research sits within the “quality of evidence” pyramid

For individual research studies:

- There is a well-defined research question – the study **population, intervention, comparator group and outcome** (PICO) are explicitly stated → PICO also informs the study hypotheses
- The study design is appropriate – minimises **selection bias** (considers representativeness of study population, limits non-response that may be associated with exposure and outcome, and ensures that cases and controls are sampled from the same underlying population) and **loss to follow-up** (that is associated with exposure and outcome)
- Robust and appropriate data and biostatistical analysis – is the analysis appropriate for the outcome (e.g. continuous, binary/categorical, count data), have assumptions been met, has potential confounding been accounted for or controlled (or at the very least considered in the discussion)

For systematic reviews:

- There is a well-built question – again follows PICO
- Well developed and explicitly stated inclusion criteria for studies found in search
- A wide and all-inclusive search – want to try to capture all studies that potentially fit the inclusion criteria → electronic databases/search engines, reference lists of published studies, study registers, research networks
- Consider publication bias (e.g. studies that showed positive outcomes were more likely to be published) and duplication bias (pretty much the same study published multiple times)
- Internal validity (quality) of constituent studies appraised → consider risk of bias for randomised controlled trials – Cochrane’s risk of bias tool (randomisation, allocation concealment, performance bias, detection bias, attrition bias, reporting bias and other biases – the use of allocation concealment and blinding of all involved should address performance bias, detection bias and reporting bias).
- Heterogeneity of results – forest plots and measures of heterogeneity → Cochrane’s Q statistic, I^2

Cochrane's Q statistic & I²

$$Q = \sum_{i=1}^k W_i(T_i - \bar{T})^2$$

W_i = weight for the i^{th} study

T_i = mean effect for the i^{th} study

\bar{T} = Pooled/combined effect for all studies

$$I^2 = 100 \times \frac{(Q - df)}{Q}$$

df = degrees of freedom, given by number of studies (k) - 1

I^2 Describes the percentage of variation across studies that is due to heterogeneity rather than chance

Determining sample size – power analysis

Is a common requirement for project proposals – e.g. in study protocols for ethics applications – and a must for grant applications.

Most often, does not need to be overly complicated and sample size based on detecting a desired difference in means or proportions will suffice.

Sample size for continuous variables

Comparing Two Means

Equation is:

$$n_{\text{per group}} = \frac{2(Z_{\alpha/2} + Z_{\beta})^2}{\left(\frac{\mu_0 - \mu_1}{\sigma}\right)^2}$$

Assumes:

- The sampling distribution of the mean differences ($\mu_0 - \mu_1$) follows a normal distribution
- The variances (σ) are equal for each group (homogeneity of variances)
- α , which defines the type I error rate (incorrectly reject null), is two-sided, and β , which defines the type II error rate (fail to reject null), is one sided.
- Common values for α and β are 0.05 (95% confidence) and 0.20 (80% power – remembering that power is given by $1 - \beta$)
- And the quantiles (Z values) are 1.96 for a 0.05 two-sided α ($Z_{\alpha/2}$), and 0.84 for a 0.20 one-sided β (Z_{β})

Stat's corner

Derivation of sample size formula from t -test equation

$$t = \frac{\mu_0 - \mu_1}{\sqrt{\frac{2\sigma^2}{n}}}$$

For 80% power and 95% confidence, t must be 1.96 + 0.84 standard errors (SEs) away from null (0)

$$(1.96 + 0.84) = \frac{\mu_0 - \mu_1}{\sqrt{\frac{2\sigma^2}{n}}}$$

Solve for n

$$\sqrt{\frac{2\sigma^2}{n}} = \frac{\mu_0 - \mu_1}{(1.96 + 0.84)}$$

Square both sides

$$\begin{aligned}\frac{2\sigma^2}{n} &= \frac{(\mu_0 - \mu_1)^2}{(1.96 + 0.84)^2} \\ n &= \frac{2\sigma^2(1.96 + 0.84)^2}{(\mu_0 - \mu_1)^2} \\ n &= \frac{2(1.96 + 0.84)^2}{\left(\frac{\mu_0 - \mu_1}{\sigma}\right)^2}\end{aligned}$$

Substituting common values for $Z_{\alpha/2}$ and Z_{β} into the sample size equation leads to a basic formula for sample size:

$$n_{per\ group} = \frac{2(1.96 + 0.84)^2}{\left(\frac{\mu_0 - \mu_1}{\sigma}\right)^2}$$

Which, when rounding up, leads to the basic formula called Lehr's equation:

$$n_{per\ group} = \frac{16}{\Delta^2}$$

where

$$2(1.96 + 0.84)^2 \sim 16$$

and

$$\Delta = \frac{\mu_0 - \mu_1}{\sigma}$$

which is the difference in means expressed in units of standard deviations – the standardised difference

Example corner

A revolutionary wonder pill is being tested for decreasing fat-mass in overweight and obese individuals. The outcome of interest is weight (in kg), standardised to height (in m²) in the form of body mass index (BMI) (kg/m²). Population based studies suggest the current standard deviation of BMI is approximately 5 kg/m². How many patients per group are required to detect a 3-point change in BMI at a 95% confidence level and 80% power?

$$Z_{\alpha/2} = 1.96 \mid Z_{\beta} = 0.84 \mid \mu_0 - \mu_1 = 3 \mid \sigma = 5$$

$$n = \frac{2(1.96 + 0.84)^2}{\left(\frac{3}{5}\right)^2}$$

$$n = 43.56 \sim 44 \text{ patients per group}$$

Using Lehr's equation

$$n_{per\ group} = \frac{16}{\left(\frac{3}{5}\right)^2}$$

$$n = 44.44 \sim 45 \text{ patients per group}$$

Or use an online calculator (e.g. <https://www.stat.ubc.ca/~rollin/stats/ssize/n2.html>)

Sample size for binary variables

Comparing Two Proportions

Equation is:

$$n_{per\ group} = \frac{\left[Z_{\alpha/2} \sqrt{2\bar{p}\bar{q}} + Z_{\beta} \sqrt{p_0(1-p_0) + p_1(1-p_1)} \right]^2}{(p_0 - p_1)^2}$$

where

p_0 = proportion in group 0

p_1 = proportion in group 1

$$\bar{p} = \frac{p_0 + p_1}{2}$$

$$\bar{q} = 1 - \bar{p}$$

Alternative equation:

$$n_{per\ group} = \frac{\left(Z_{\alpha/2} + Z_{\beta} \right)^2 \times (p_0(1-p_0) + p_1(1-p_1))}{(p_0 - p_1)^2}$$

Another alternative equation based on Lehr's equation (simplest to use):

$$n_{per\ group} = \frac{16\bar{p}\bar{q}}{(p_0 - p_1)^2}$$

Example corner

Dr Neura has received approval to test a lentiviral-based gene therapy for augmenting IQ. The research team believe they can increase the proportion of superior intelligence individuals (i.e. IQ > 130) from 2% to at least 10%. How many patients per group are required to detect this change at 95% confidence and 80% power?

$$n_{per\ group} = \frac{\left[Z_{\alpha/2} \sqrt{2\bar{p}\bar{q}} + Z_{\beta} \sqrt{p_0(1-p_0) + p_1(1-p_1)} \right]^2}{(p_0 - p_1)^2}$$

$$n_{per\ group} = \frac{\left[1.96\sqrt{2 \times 0.06 \times 0.94} + 0.84\sqrt{0.02(1-0.02) + 0.1(1-0.1)} \right]^2}{(0.02 - 0.1)^2}$$

$$n_{per\ group} = 136.99 = 137\ per\ group$$

Example corner continued...

Using alternative equation:

$$n_{per\ group} = \frac{(Z_{\alpha/2} + Z_{\beta})^2 \times (p_0(1 - p_0) + p_1(1 - p_1))}{(p_0 - p_1)^2}$$

$$n_{per\ group} = \frac{(1.96 + 0.84)^2 \times (0.02(1 - 0.02) + 0.1(1 - 0.1))}{(0.02 - 0.1)^2}$$

$$n_{per\ group} = 134.26 = 135\ per\ group$$

Using Lehr's equation

$$n_{per\ group} = \frac{16\bar{p}\bar{q}}{(p_0 - p_1)^2}$$

$$n_{per\ group} = \frac{16 \times 0.06 \times 0.94}{(0.02 - 0.1)^2}$$

$$n_{per\ group} = 141$$

Or use an online calculator (e.g. <https://select-statistics.co.uk/calculators/sample-size-calculator-two-proportions/>)

Sample size for Poisson distributed variables (count data)

Comparing Two Poisson Means

- The Poisson distribution is useful for analysing count data (number of counts per unit time observation)
- It leverages from the fact that the square root of a Poisson variable is approximately normal – *normal approximation*
- For a Poisson distributed variable

$$Y_i \sim Poi(\lambda)$$

$$\sqrt{Y_i} \sim N(\sqrt{\lambda}, 0.25)$$

- Substituting into Lehr's equation

$$n_{per\ group} = \frac{16}{\left[\frac{(\sqrt{\lambda_0} - \sqrt{\lambda_1})^2}{\sigma^2} \right]}$$

$$n_{per\ group} = \frac{16\sigma^2}{(\sqrt{\lambda_0} - \sqrt{\lambda_1})^2}$$

- we know $\sigma^2 \sim 0.25$

$$n_{per\ group} = \frac{16 \times 0.25}{(\sqrt{\lambda_0} - \sqrt{\lambda_1})^2}$$

Equation is:

$$n_{per\ group} = \frac{4}{(\sqrt{\lambda_0} - \sqrt{\lambda_1})^2}$$

Example corner

A de-prescribing protocol is being investigated for its potential to reduce polypharmacy in adults aged 75 years or more with multiple chronic conditions. The research team is wishing to reduce the number of medications prescribed from an average of 16 to 13 (or less) per patient. How many patients per group are required to detect this change at 95% confidence and 80% power?

$$n_{per\ group} = \frac{4}{(\sqrt{\lambda_0} - \sqrt{\lambda_1})^2}$$

$$n_{per\ group} = \frac{4}{(\sqrt{16} - \sqrt{13})^2}$$

$$n_{per\ group} = 25.71 \sim 26\ per\ group$$