Sample Size, Power, and Precision

Håkon K. Gjessing

Professor/Principal Investigator

Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo

Department of Global Public Health and Primary Care, University of Bergen

Makerere

Wednesday, 7 June 2023

CENTRE FOR FERTILITY AND HEALTH

Testing and Confidence intervals

What do we need? (much simplified! But still OK)

1. Effect estimate $\widehat{M}$ of population value M. Could be "anything", like
   - $\widehat{M} = \overline{X}$, mean value in one group
   - $\widehat{M} = \widehat{p}$, risk in one group
   - $\widehat{M} = \overline{X}_2 - \overline{X}_1$, difference between mean values in two groups
   - $\widehat{M} = \widehat{p}_2 - \widehat{p}_1$, Risk Difference between two groups
   - $\widehat{M} = \log(\widehat{p}_2) - \log(\widehat{p}_1) = \log(\frac{\widehat{p}_2}{\widehat{p}_1}) = \log(RR)$, difference between log risks in two groups, i.e. log of Relative Risk

CENTRE FOR FERTILITY AND HEALTH

Testing and Confidence intervals

What do we need? (much simplified! But still OK)

1. Effect estimate $\widehat{M}$ of population value M. Could be "anything", like
   - $\widehat{M} = \overline{X}$, mean value in one group
   - $\widehat{M} = \widehat{p}$, risk in one group
   - $\widehat{M} = \overline{X}_2 - \overline{X}_1$, difference between mean values in two groups
   - $\widehat{M} = \widehat{p}_2 - \widehat{p}_1$, Risk Difference between two groups
   - $\widehat{M} = \log(\widehat{p}_2) - \log(\widehat{p}_1) = \log(\frac{\widehat{p}_2}{\widehat{p}_1}) = \log(RR)$, difference between log risks in two groups, i.e. log of Relative Risk
2. Standard Error of Estimate $SE(\widehat{M})$.

   How to compute this can depend on
   - Whether you are *testing* or *computing CI*
   - Realistic assumptions about the situation at hand
   - What the statistician fancies, or what the software fancies

   BUT don't fret too much over that....

CENTRE FOR FERTILITY AND HEALTH

Testing and Confidence intervals

What do we need? (much simplified! But still OK)

1. Effect estimate $\widehat{M}$ of population value M. Could be "anything", like
   - $\widehat{M} = \overline{X}$, mean value in one group
   - $\widehat{M} = \widehat{p}$, risk in one group
   - $\widehat{M} = \overline{X}_2 - \overline{X}_1$, difference between mean values in two groups
   - $\widehat{M} = \widehat{p}_2 - \widehat{p}_1$, Risk Difference between two groups
   - $\widehat{M} = \log(\widehat{p}_2) - \log(\widehat{p}_1) = \log(\frac{\widehat{p}_2}{\widehat{p}_1}) = \log(RR)$, difference between log risks in two groups, i.e. log of Relative Risk
2. Standard Error of Estimate $SE(\widehat{M})$.
   How to compute this can depend on
   - Whether you are *testing* or *computing CI*
   - Realistic assumptions about the situation at hand
   - What the statistician fancies, or what the software fancies

   BUT don't fret too much over that....
3. Quantile from the normal distribution, typically $z_{0.025} = 1.96 \approx 2$
   - 0.025 means 5% level two-sided test, or 95% CI
   - Again, other distributions may sometimes be used

   BUT again, don't worry, be happy....

CENTRE FOR FERTILITY AND HEALTH

Confidence Interval for M:

$$\left(\widehat{M} - 1.96 \cdot SE(\widehat{M}), \widehat{M} + 1.96 \cdot SE(\widehat{M})\right)$$

Lower limit: $\widehat{M} - 1.96 \cdot SE(\widehat{M})$ Upper limit: $\widehat{M} + 1.96 \cdot SE(\widehat{M})$

Testing $H_0 : M = M_0$:

$$Z = \frac{\widehat{M} - M_0}{SE(\widehat{M})}$$

Reject H_0 if Z-score is bigger than 1.96 or less than -1.96.

M_0 is the value of M in the null hypothesis H_0.

NOTE: When M is a difference, we often use $M_0 = 0$ as the H_0.

CENTRE FOR FERTILITY AND HEALTH

NOTE that

$$\begin{aligned} \frac{\widehat{M} - M_0}{SE(\widehat{M})} &> 1.96 \\ \widehat{M} - M_0 &> 1.96 \cdot SE(\widehat{M}) \\ \widehat{M} - 1.96 \cdot SE(\widehat{M}) &> M_0 \end{aligned}$$

That is, *in this simplified setup*, the following two are equivalent:

- H_0 is rejected because Z is too big
- Lower limit of CI is above M_0

... and vice versa when Z is less than -1.96.

CENTRE FOR FERTILITY AND HEALTH

- $(Z = -1.96)$ or $(Z = 1.96)$ $\Rightarrow$ p-value = 0.05
- $(Z < -1.96)$ or $(Z > 1.96)$ $\Rightarrow$ p-value < 0.05

Thus, H_0 is rejected at the 5% level if p-value < 0.05

The p-value is calculated directly from the Z-score

Thus, a large (positive or negative) value of Z

$\Rightarrow$ a small p-value

$\Rightarrow$ H_0 will be rejected.

CENTRE FOR FERTILITY AND HEALTH

Important issue:

Z can be large (and thus the p-value small)

for **two** reasons:

$$Z = \frac{\widehat{M} - M_0}{SE(\widehat{M})}$$

1. The effect $\widehat{M} - M_0$ is large, i.e. *strong effect*
2. The $SE(\widehat{M})$ is small, i.e. *precise estimate*

The p-value is still valuable,

but it mixes up estimate size and precision

CENTRE FOR FERTILITY AND HEALTH

Confidence intervals provide

- Estimate: $\widehat{\mathrm{M}}$
- Uncertainty/range of estimate: $\left(\widehat{\mathrm{M}} - 1.96 \cdot \mathrm{SE}(\widehat{\mathrm{M}}), \widehat{\mathrm{M}} + 1.96 \cdot \mathrm{SE}(\widehat{\mathrm{M}})\right)$
- ... and **length of CI** is $\approx 4 \cdot \mathrm{SE}(\widehat{\mathrm{M}})$
- M_0 outside of interval is a *good indication* that H_0 will be rejected

p-values provide

- Large p-value means result is probably uninteresting
- Small p-value means result is either *strong* or *precise* or both
- Small p-value means H_0 is rejected

But still, use CI or p-values *as they are intended*

CENTRE FOR FERTILITY AND HEALTH

Sample Size, Power, Precision

Essential observations:

- Standard error measures precision of estimate
- Length of CI $\approx 4 \cdot \mathrm{SE}(\widehat{M})$
- Smaller $\mathrm{SE}(\widehat{M})$ makes Z larger
- I.e. smaller $\mathrm{SE}(\widehat{M})$ gives shorter CIs and easier rejection of H_0

AND

$$\mathrm{SE}(\widehat{M}) = \frac{\gamma}{\sqrt{n}}$$

in almost all conceivable situations....

... where γ is a fixed number for a given situation, design etc.

(often, $\gamma = \sigma$, i.e. the standard deviation in the population)

... but sample size n can vary

CENTRE FOR FERTILITY AND HEALTH

Sample Size, Power, Precision

For instance, in two similar studies with different sample sizes n_1 and n_2, the standard errors are

$$\mathrm{SE}(\widehat{M_1}) = \frac{\gamma}{\sqrt{n_1}}$$

and

$$\mathrm{SE}(\widehat{M_2}) = \frac{\gamma}{\sqrt{n_2}},$$

respectively.

CENTRE FOR FERTILITY AND HEALTH

Mortality among males: 398 deaths among 2086 followed,

$\widehat{p} = 398/2086 = 0.191$.

$$SE(\widehat{p}) \approx \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = \frac{\sqrt{\widehat{p}(1-\widehat{p})}}{\sqrt{n}} = \frac{\widehat{\gamma}}{\sqrt{n}}$$

In this case,

$$\widehat{\gamma} = \sqrt{0.191 \cdot (1 - 0.191)} = 0.393$$

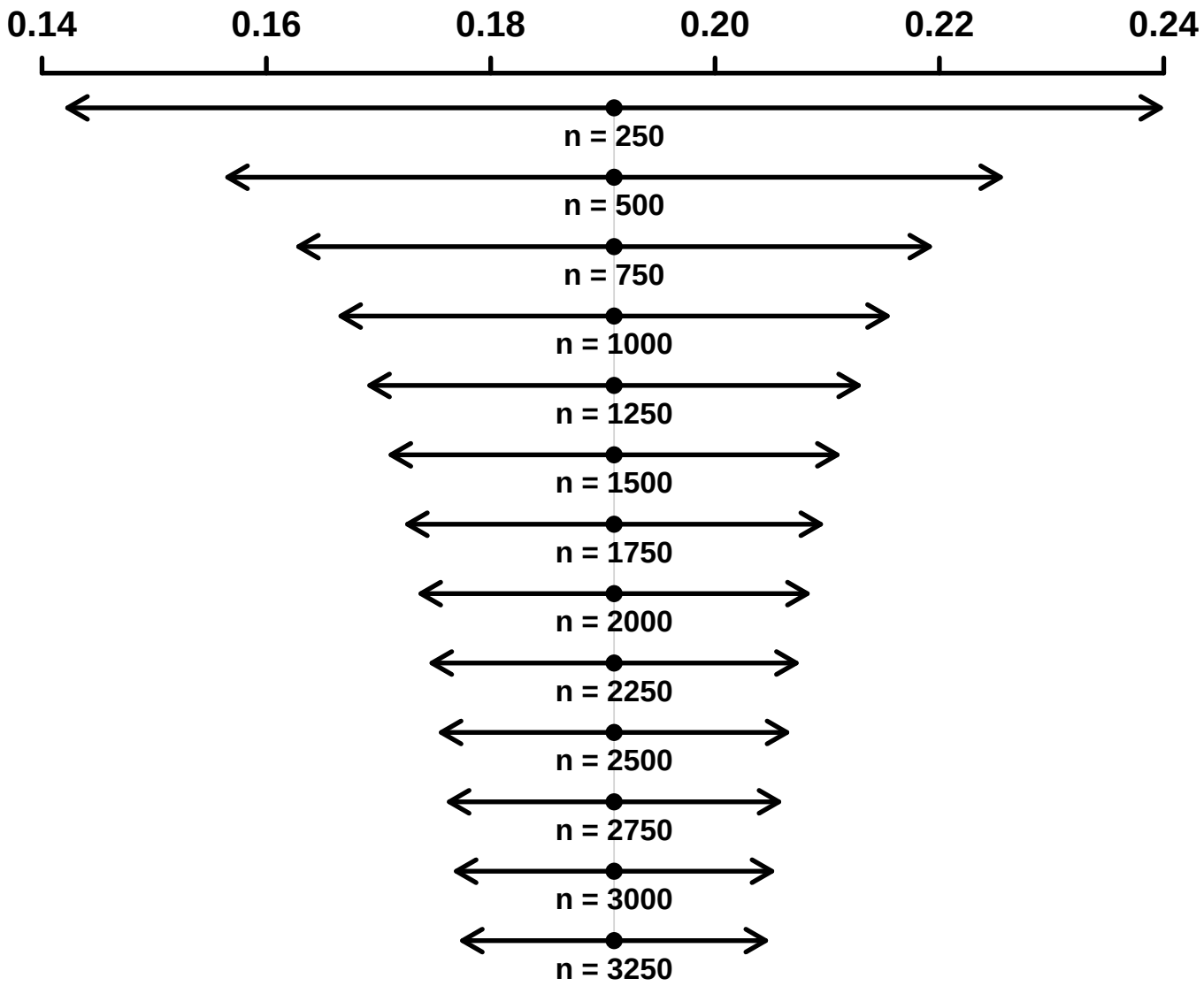
and $n = 2086$.

What happens if n changes (and $\widehat{p}$ stays the same)?

CENTRE FOR FERTILITY AND HEALTH

PRECISION ACCORDING TO SAMPLE SIZE

MALE MORTALITY



CENTRE FOR FERTILITY AND HEALTH

Problem of diminishing returns

4 times sample size gives only double precision:

$$\frac{\gamma}{\sqrt{4n}} = \frac{\gamma}{2\sqrt{n}} = \frac{1}{2} \cdot \frac{\gamma}{\sqrt{n}}.$$

To obtain 10 times the precision you need a hundred times the sample size:

$$\frac{\gamma}{\sqrt{100n}} = \frac{\gamma}{10\sqrt{n}} = \frac{1}{10} \cdot \frac{\gamma}{\sqrt{n}}.$$

... sad fact of life...

CENTRE FOR FERTILITY AND HEALTH

Pilot study:

Uses 100 individuals, observes 17 deaths, i.e. $\widehat{p} = 0.17$.

Our study:

With the same p, what sample size n do we need to get $SE(\widehat{p}) = 0.01$?

(Because we find this to be sufficiently precise.)

$$SE(\widehat{p}) = \frac{\sqrt{0.17 \cdot (1 - 0.17)}}{\sqrt{n}} = \frac{0.376}{\sqrt{n}} = 0.01$$

So we need

$$n = \left(\frac{0.376}{0.01}\right)^2 \approx 1400$$

Note the square term, which is our “big little problem” again.

CENTRE FOR FERTILITY AND HEALTH

Sample size: The easy (and best?) way

More general

Pilot study:

$$\mathrm{SE}(\widehat{M_1}) = \frac{\gamma}{\sqrt{n_1}}$$

Your study:

$$\mathrm{SE}(\widehat{M_2}) = \frac{\gamma}{\sqrt{n_2}}.$$

We have

$$\frac{n_2}{n_1} = \left(\frac{\mathrm{SE}(\widehat{M_2})}{\mathrm{SE}(\widehat{M_1})}\right)^2$$

γ is gone! (Good riddance!)

CENTRE FOR FERTILITY AND HEALTH

Sample size: The easy (and best?) way

More general

$$\frac{n_2}{n_1} = \left(\frac{\mathrm{SE}(\widehat{M_2})}{\mathrm{SE}(\widehat{M_1})} \right)^2$$

That is,

The ratio of the sample sizes is the square of the ratio of precisions

or equivalently

The ratio of the sample sizes is the square of the ratio of CI lengths

CENTRE FOR FERTILITY AND HEALTH

Pilot study:

Uses 100 individuals, observes 17 deaths, i.e. $\widehat{p} = 0.17$.

$$SE(\widehat{p}) = \frac{\sqrt{0.17 \cdot (1 - 0.17)}}{\sqrt{100}} = \frac{0.376}{10} = 0.0376$$

Our study:

With the same p, what sample size n do we need to get $SE(\widehat{p}) = 0.01$?

(Because we find this to be sufficiently precise.)

$$n_2 = 100 \cdot \left(\frac{0.376}{0.01}\right)^2 \approx 1400$$

CENTRE FOR FERTILITY AND HEALTH

Sample size: The easy (and best?) way

A few comments

- In particular, our Norwegian mortality data are population registry data with decades of follow-up, so a power calculation like this is less meaningful. *However,* it can still be a meaningful guide to whether or not to apply for access to data. Or access to specific risk factors that seem interesting but may be under restrictions since they are sensitive.
- Note that for binomial data, we only need an estimate $\widehat{p}$ of the probability, because

$$\widehat{\gamma} = \sqrt{\widehat{p}(1-\widehat{p})}$$

in the binomial case.

CENTRE FOR FERTILITY AND HEALTH

- **In general**, if we only knew, say, that a pilot study showed

$$\mathrm{SE}(\widehat{\mathrm{M}}) = 0.0376$$

with a sample size of $n_1 = 100$, we could find

$$n_2 = n_1 \cdot \left(\frac{\mathrm{SE}(\widehat{\mathrm{M}_2})}{\mathrm{SE}(\widehat{\mathrm{M}_1})} \right)^2 = 100 \cdot \left(\frac{0.0376}{0.01} \right)^2 \approx 1400$$

- NOTE that this **does not** assume binomial data. It works just as well for Normally distributed data, or whenever we can use the Central Limit Theorem.

CENTRE FOR FERTILITY AND HEALTH

Sample size: The easy (and best?) way

Example

A **pilot study** of Birth Weight finds an average of

$$3400\text{g}, \text{ with 95\% CI } (3147\text{g}, 3653\text{g})$$

when looking at $n = 10$ newborns,
i.e. a width of the CI of $3653\text{g} - 3147\text{g} = 506\text{g}$.

Our study wants a CI with width of about 100g:

$$n_2 = 10 \cdot \left(\frac{500\text{g}}{100\text{g}}\right)^2 = 256$$

CENTRE FOR FERTILITY AND HEALTH

Sample size: The easy (and best?) way

- This calculation is easy to understand
- It is easy to perform
- It is relevant
- If SE is not mentioned in the pilot study,
  it can often be computed from the length of a CI

But keep in mind:

Sample size calculations always need a somewhat subjective subject-matter decision:

What precision is “needed”, i.e., is biologically relevant?

This is not a statistical decision as such

CENTRE FOR FERTILITY AND HEALTH

Sample size: The hard way

Testing situation:

$H_0 : p = 0.15$ versus $H_1 : p \neq 0.15$

Types of errors:

- Type 1 error: H_0 is correct, but you *reject* it by mistake
  - The probability of Type 1 error is the **level** of the test, e.g. 5%, as specified
- Type 2 error: H_0 is *not* correct, but you *keep* it by mistake
  - The **power** of a test is the probability of *rejecting* H_0 when it is wrong

That is, the **power** is the "ability" of your test to detect a true p which is away from $p_0 = 0.15$.

CENTRE FOR FERTILITY AND HEALTH

Sample size: The hard way

Remember:

- Large (or small) Z-score $\Rightarrow$ low p-value $\Rightarrow$ likely rejection

$$Z = \frac{\hat{p} - p_0}{SE(\hat{p})}$$

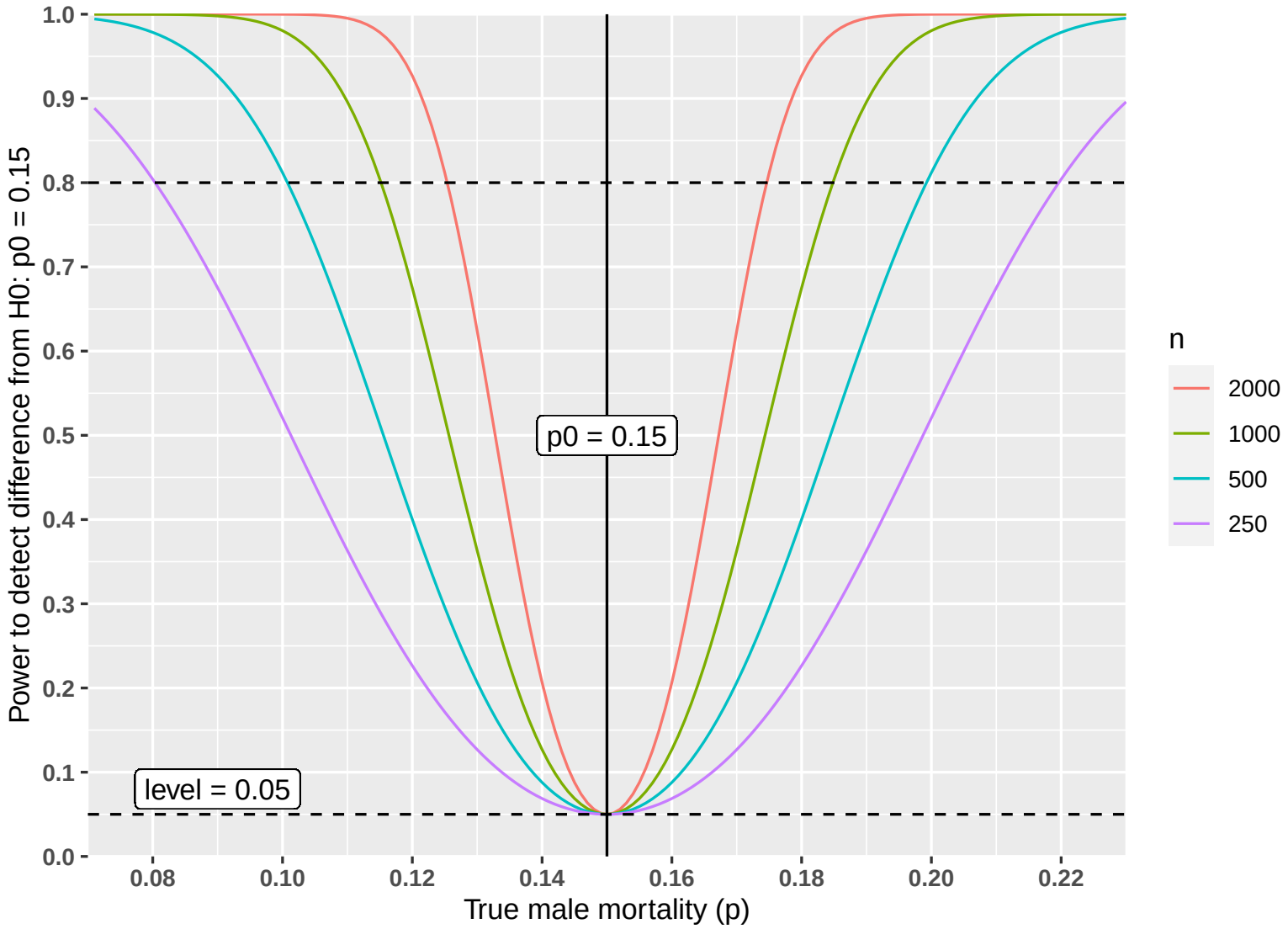
So the power *increases* when

- p is far away from $p_0 = 0.15$
- $SE(\hat{p})$ gets smaller, i.e. when n gets bigger.

CENTRE FOR FERTILITY AND HEALTH

Sample size: The hard way

Power curves



CENTRE FOR FERTILITY AND HEALTH

Sample size: The hard way

Typcial approach:

- Choose level, typically 5%
- Choose the power you want, typically 80%
- Decide (subjectively) how far away from $p_0 = 0.15$ is meaningful
- Choose n large enough to guarantee the 80% power at this difference

But harder with power, so software should be used.

(Well, I admit.... using good software is *always* useful...
easy to make mistakes.)

CENTRE FOR FERTILITY AND HEALTH

Special case: Comparing two independent groups

Standard error for difference:

$$SE(\overline{X}_2 - \overline{X}_1) = \sqrt{SE(X_1)^2 + SE(X_2)^2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Equal variances, and assume $n_1 = n, n_2 = k \cdot n$:

$$SE(\overline{X}_2 - \overline{X}_1) = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{k \cdot n}} = \sqrt{\frac{k+1}{k}} \cdot \frac{\sigma}{\sqrt{n}}$$

That is, the SE of the difference is $\sqrt{\frac{k+1}{k}}$ times larger than for a single group. k is the ratio of group 2 to group 1.

CENTRE FOR FERTILITY AND HEALTH

Equal variances, equal sizes:

The SE of the difference is about 40% larger than each SE.

$$\sqrt{\frac{k+1}{k}} = \sqrt{2} \approx 1.4$$

CENTRE FOR FERTILITY AND HEALTH

Another case of diminishing returns

Case-control study:

Cases: $n_1 = n$ is kept fixed (no available extra cases)

Controls: $n_2 = k \cdot n$, where k can be increased.

$$k = 1 \quad \Rightarrow \quad \sqrt{\frac{k+1}{k}} = \sqrt{2} \approx 1.4 \quad \text{(40\% more than a single group)}$$

$$k = 2 \quad \Rightarrow \quad \sqrt{\frac{k+1}{k}} = \sqrt{3/2} \approx 1.22 \quad \text{(22\% more than a single group)}$$

$$k = 3 \quad \Rightarrow \quad \sqrt{\frac{k+1}{k}} = \sqrt{4/3} \approx 1.15 \quad \text{(15\% more than a single group)}$$

$$k = 4 \quad \Rightarrow \quad \sqrt{\frac{k+1}{k}} = \sqrt{5/4} \approx 1.12 \quad \text{(12\% more than a single group)}$$

Gets better, but *always* stuck with uncertainty in the case group

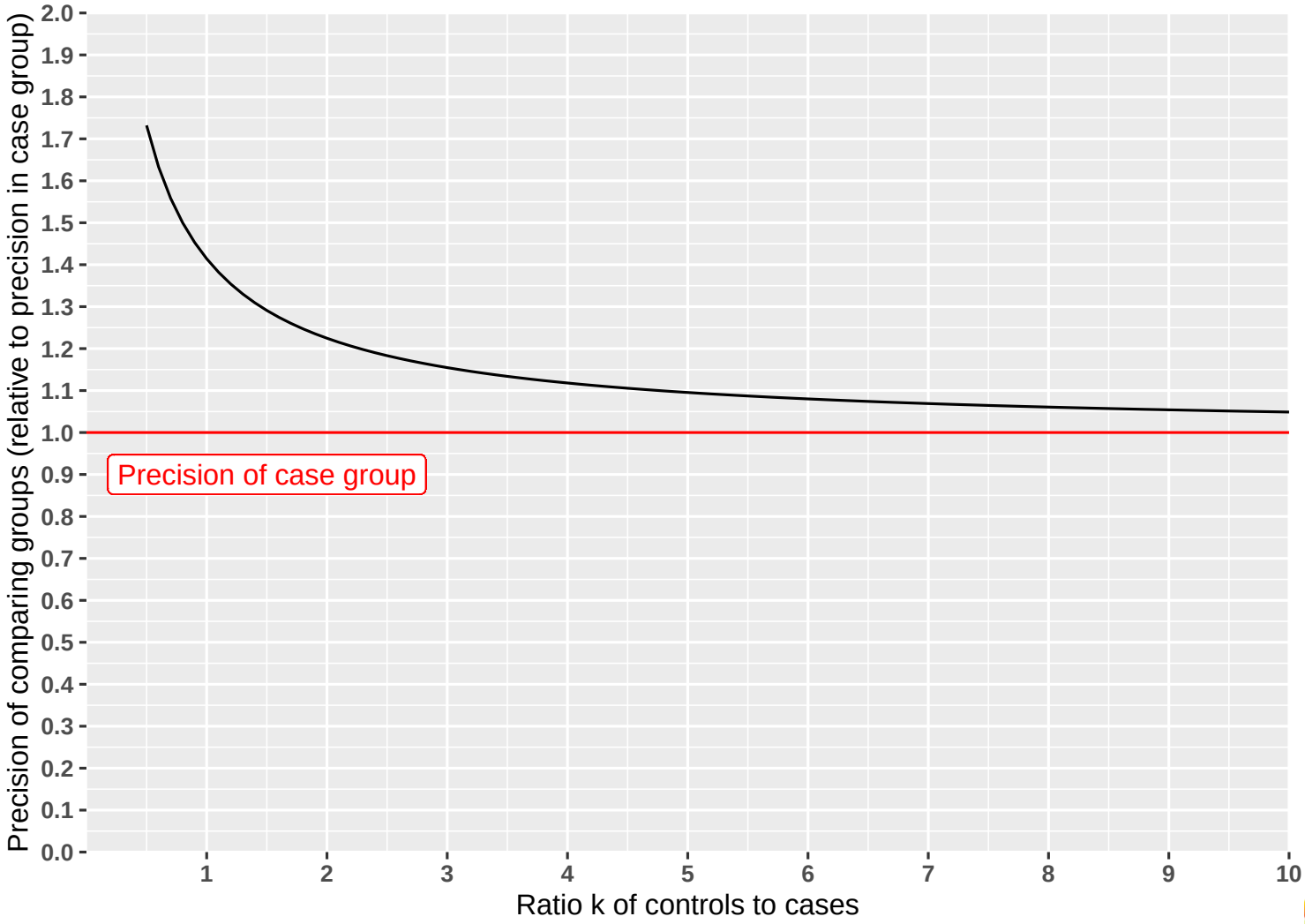
$$SE(\overline{X}_2 - \overline{X}_1) = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{k \cdot n}} = \sqrt{\frac{k+1}{k}} \cdot \frac{\sigma}{\sqrt{n}}$$

CENTRE FOR FERTILITY AND HEALTH

Special case: Comparing two independent groups

For instance, Case-control studies

Precision of difference between groups, relativ to single group:



CENTRE FOR FERTILITY AND HEALTH

Additional thoughts

- Remember to allow for dropouts!
- With time-to-event studies, things may be a bit harder
- With cluster-randomized trials, the design factor is often useful
- Clever designs *can* sometimes help precision and power

  ... but there is a limit
- Post-hoc power calculations are not very helpful.... CI is better
- With extensive multiple testing, special methods may be employed, e.g. False Discovery Rates (FDR)

CENTRE FOR FERTILITY AND HEALTH