



Sample Size, Power, and Precision



Håkon K. Gjessing

Professor/Principal Investigator

⇒ Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo
– a Norwegian Research Council Centre of Excellence

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



Kathmandu
Thursday 1 October 2026



Testing and Confidence intervals

What do we *actually* need? (Simplified! But still to the point)

Simplification:

Assume that sample sizes are “reasonably” large, so that the Central Limit Theorem holds

What do we need?

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

Testing and Confidence intervals

What do we *actually* need? (Simplified! But still to the point)

2. **Standard Error** of Estimate $SE(\hat{M})$.

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....

Testing and Confidence intervals

What do we *actually* need? (Simplified! But still to the point)

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 (like the t-distribution) may sometimes be used

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

That is, we need:

1. **Effect estimate $\hat{M}$**
2. **Standard Error $SE(\hat{M})$**
3. **$z_{0.025} = 1.96 \approx 2$**

Confidence Interval for M :

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

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

Testing $H_0 : M = M_0$:

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

M_0 is the value of M in the null hypothesis H_0 .
(When M is a difference, we often use $M_0 = 0$ as the H_0 .)

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

Note that

Rejecting H_0 (two-sided, 5% level)

is equivalent to

M_0 outside the CI (two-sided, 95% confidence)

Important qualification:

- $SE(\hat{M})$ sometimes calculated differently for the two
- In that case the results may be slightly different
- Thus CIs are CIs, tests are tests!
- But the intuition is the same

Testing and Confidence intervals

Example: Comparing two proportions

Two common formulas

- SE for confidence intervals:

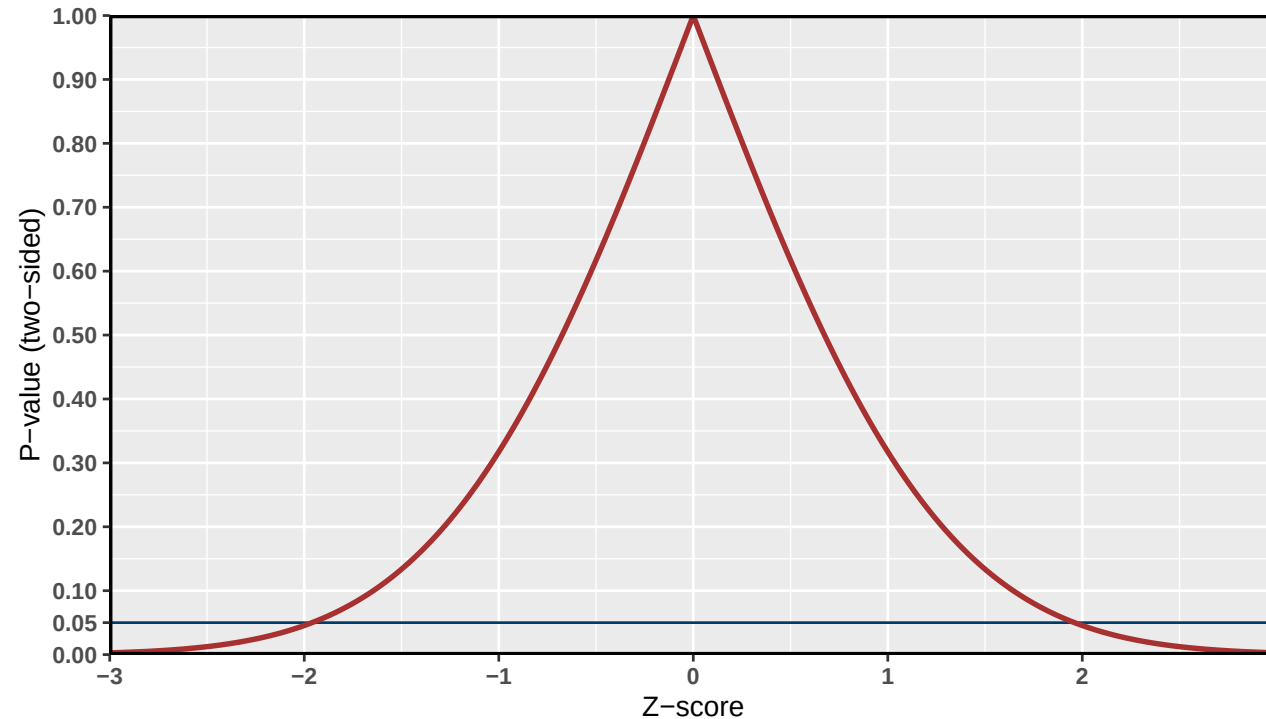
$$SE(RD) = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

- SE for hypothesis testing:

$$SE(RD) = \sqrt{\hat{p}(1 - \hat{p}) \left\{ \frac{1}{n_1} + \frac{1}{n_2} \right\}}$$

-
- You can use the **first** for BOTH if you like
 - But the second assumes specifically that H_0 is true, i.e. $p_1 = p_2 = p = \frac{X_1 + X_2}{n_1 + n_2}$, in the calculation

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



Large (positive or negative) value of $Z \implies$ Small p-value $\implies H_0$ will be rejected.

Important problem:

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

Z can be large (and thus the p-value small) for **two** reasons:

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

The p-value is still valuable, but **it mixes up size and precision** of estimate

Testing and Confidence intervals

... thus

Confidence intervals provide

- Estimate: $\hat{M}$
- Uncertainty/range of estimate: $(\hat{M} - 1.96 \cdot \text{SE}(\hat{M}), \hat{M} + 1.96 \cdot \text{SE}(\hat{M}))$
- ... and **length of CI** is $\approx 4 \cdot \text{SE}(\hat{M})$
- 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*

Sample Size, Power, Precision

Essential observations:

- $SE(\hat{M})$ measures precision of $\hat{M}$. Length of CI $\approx 4 \cdot SE(\hat{M})$
- Smaller $SE(\hat{M})$ makes Z larger and thus easier rejection of H_0
- Smaller $SE(\hat{M})$ gives shorter CIs

and importantly,

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

in almost all conceivable situations...

- Sample size n can vary
- γ is a fixed number for a given situation, design etc. (Often $\gamma = \sigma$, population standard deviation)

Sample Size, Power, Precision

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

$$\text{SE}(\widehat{M}_1) = \frac{\gamma}{\sqrt{n_1}}$$

and

$$\text{SE}(\widehat{M}_2) = \frac{\gamma}{\sqrt{n_2}},$$

respectively.

398 deaths among 2086 followed, $\hat{p} = \frac{398}{2086} = 0.191$.

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

In this case,

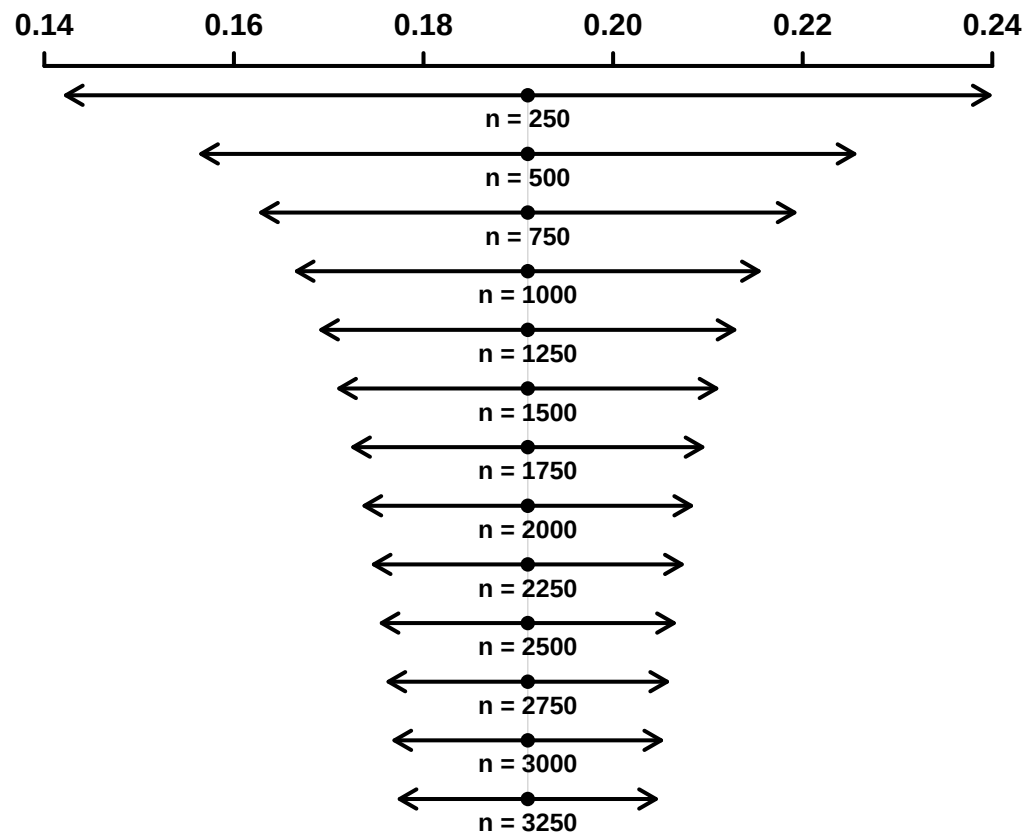
$$\hat{\gamma} = \sqrt{0.191 \cdot (1 - 0.191)} = 0.393$$

and $n = 2086$.

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

Precision according to sample size

Example: Mortality among males



$\hat{p} = 0.191$ stays the same. CIs get narrower.

A 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...

Sample size: The easy (and best?) way

Example

Pilot study:

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

Our study:

With the same p , what sample size n do we need to get $SE(\hat{p}) = 0.01$?
(Because we find this to be sufficiently precise.)

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

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.

Sample size: The easy (and best?) way

More general

Pilot study:

$$\text{SE}(\widehat{M}_1) = \frac{\gamma}{\sqrt{n_1}}$$

Your study:

$$\text{SE}(\widehat{M}_2) = \frac{\gamma}{\sqrt{n_2}}.$$

We have

$$\frac{n_2}{n_1} = \left(\frac{\text{SE}(\widehat{M}_1)}{\text{SE}(\widehat{M}_2)} \right)^2$$

γ is gone! (Good riddance!)

Sample size: The easy (and best?) way

More general

$$\frac{n_2}{n_1} = \left(\frac{\text{SE}(\widehat{M}_1)}{\text{SE}(\widehat{M}_2)} \right)^2$$

That is,

The ratio of the sample sizes is the square of the (inverse) ratio of precisions

or equivalently

The ratio of the sample sizes is the square of the (inverse) ratio of CI lengths

Sample size: The easy (and best?) way

Example

Pilot study:

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

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

Our study:

With the same p , what sample size n do we need to get $SE(\hat{p}) = 0.01$?
(Because we find this to be sufficiently precise.)

$$n_2 = 100 \cdot \left(\frac{0.0376}{0.01} \right)^2 \approx 1400$$

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 $\hat{p}$ of the probability, because

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

in the binomial case.

Sample size: The easy (and best?) way

A few comments

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

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

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

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

- 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.

Sample size: The easy (and best?) way

Example

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

3400g, with 95% CI (3150g, 3650g)

when looking at $n = 10$ newborns,
i.e. a width of the CI of $3650\text{g} - 3150\text{g} = 500\text{g}$.

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

$$n_2 = 10 \cdot \left(\frac{500\text{g}}{100\text{g}} \right)^2 = 250$$

Thus, we need a sample of 250 to obtain the required precision

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

Sample size: The hard way

... which is, unfortunately, the most common

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 $p_0 = 0.15$.

Sample size: The hard way

Remember:

- Large (or small) Z-score $\implies$ low p-value $\implies$ likely rejection of H_0

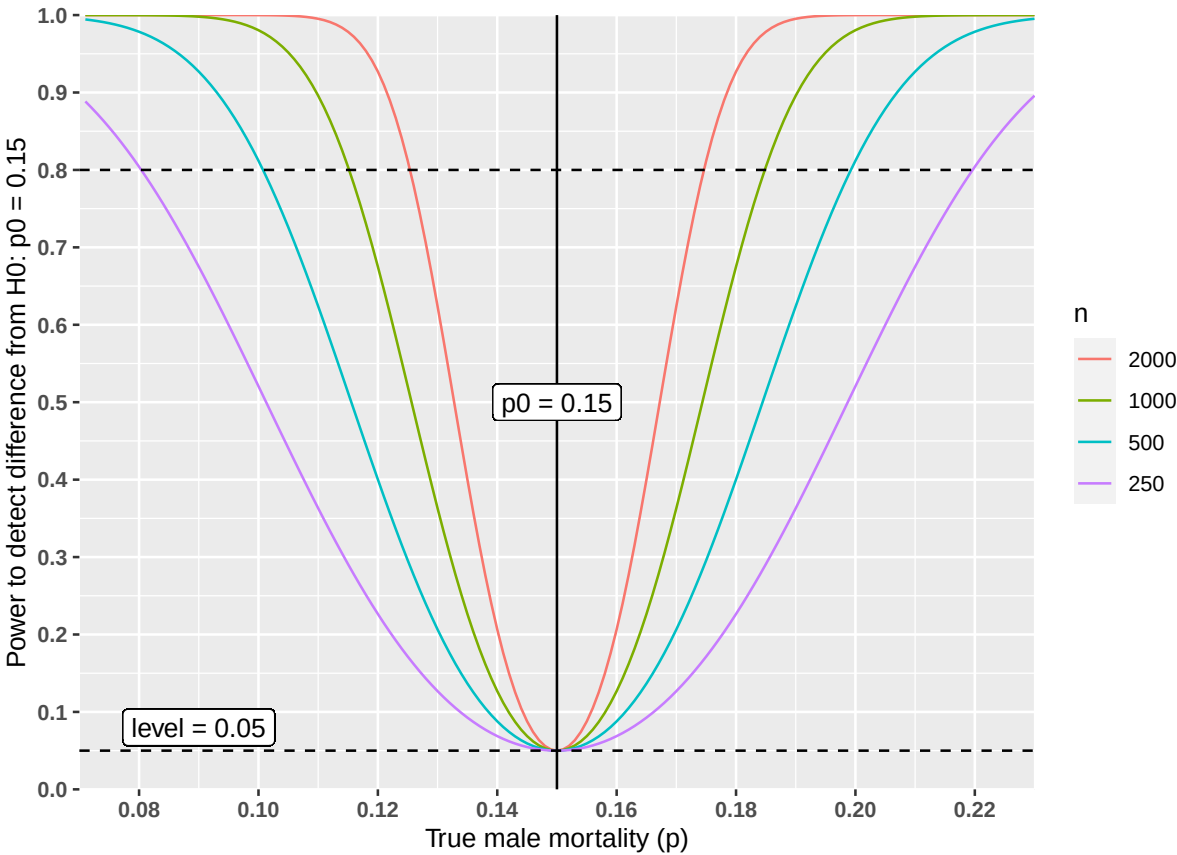
$$Z = \frac{\hat{p} - p_0}{\text{SE}(\hat{p})}$$

So the power *increases* when

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

Sample size: The hard way

Power curves



Sample size: The hard way

Typical 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... it's easy to make mistakes.)

Special case: Comparing two independent groups

Standard error for difference:

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

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

$$\text{SE}(\bar{X}_2 - \bar{X}_1) = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{k \cdot n_1}} = \sqrt{\frac{k+1}{k}} \cdot \frac{\sigma}{\sqrt{n_1}}$$

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.

Special case: Comparing two independent groups

Equal variances, equal sizes:

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

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

Special case: Comparing two independent groups

For instance, Case-Control studies: Another case of diminishing returns

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

Controls: $n_2 = k \cdot n_1$, where k can be increased.

$$k = 1 \implies \sqrt{\frac{k+1}{k}} = \sqrt{2} \approx 1.4 \quad (\text{SE 40\% higher than for a single group})$$

$$k = 2 \implies \sqrt{\frac{k+1}{k}} = \sqrt{\frac{3}{2}} \approx 1.22 \quad (\text{SE 22\% higher than for a single group})$$

$$k = 3 \implies \sqrt{\frac{k+1}{k}} = \sqrt{\frac{4}{3}} \approx 1.15 \quad (\text{SE 15\% higher than for a single group})$$

$$k = 4 \implies \sqrt{\frac{k+1}{k}} = \sqrt{\frac{5}{4}} \approx 1.12 \quad (\text{SE 12\% higher than for a single group})$$

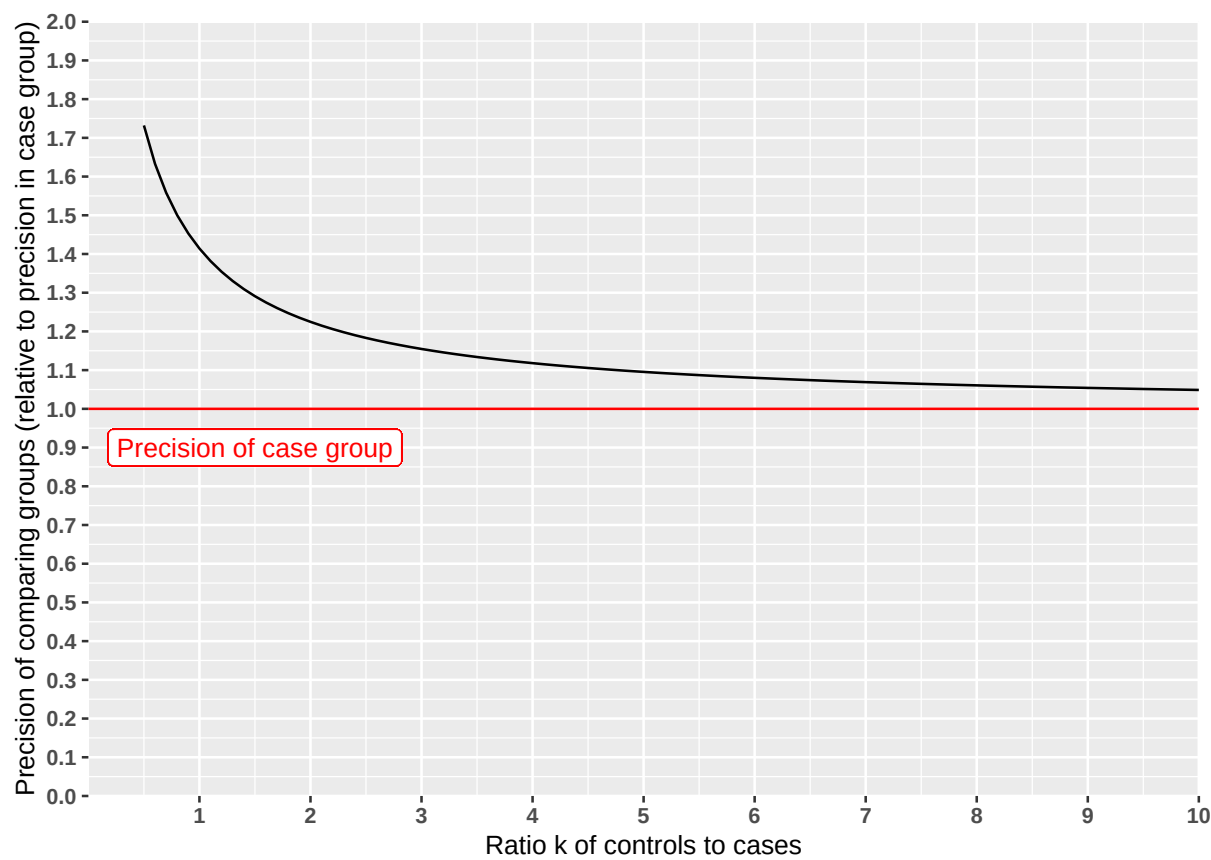
Gets better as k grows larger, but *always* stuck with uncertainty in the case group

$$\text{SE}(\bar{X}_2 - \bar{X}_1) = \sqrt{\frac{k+1}{k}} \cdot \frac{\sigma}{\sqrt{n_1}}$$

Special case: Comparing two independent groups

For instance, Case-Control studies

Precision of difference between groups, relative to single group:



Additional thoughts

- Remember to allow for dropouts!
- With cluster-randomized trials, the **design factor** is often useful (dependent data)
- Clever designs *can* sometimes help precision and power
... but there is a limit
- Post-hoc power calculations are not very helpful.... CI is better post-hoc
- With extensive multiple testing, special methods may be employed, e.g. False Discovery Rates (FDR)