432 Class 13

https://thomaselove.github.io/432-2025

2025-02-25

## Today’s Agenda

* How Big a Sample Size Do I Need?
* Type I and Type II Errors in Hypothesis Testing
* Power and Sample Size Calculations
  + When Comparing Two Means
    - Paired vs. Independent Samples
  + When Comparing Two Proportions
  + The March of Science

## Today’s R Setup

knitr::opts\_chunk$set(comment = NA)  
  
library(pwr)

## How Big A Sample Size Do I need?

1. What is the budget?
2. What are you trying to compare?
3. What is the study design?
4. How big an effect size do you expect (hope) to see?
5. What was that budget again?
6. OK, tell me the maximum allowable rates of Type I and Type II error that you want to control for. Or, if you like, tell me the confidence level and power you want to have.
7. What sort of statistical inference do you want to plan for?

## Errors in Hypothesis Testing

In testing hypotheses, there are two potential decisions and each one brings with it the possibility that a mistake has been made.

| – | is true | is true |
| --- | --- | --- |
| Test rejects | Correct | Type I error (False positive) |
| Test retains | Type II error (False negative) | Correct |

## Errors in Hypothesis Testing

* A Type I error can only be made if is actually true.
* A Type II error can only be made if is actually true.

| – | is true | is true |
| --- | --- | --- |
| Test rejects | Correct | Type I error (False positive) |
| Test retains | Type II error (False negative) | Correct |

## Specifying Error Probabilities (Type I)

If we say we are using 90% confidence, this means:

* we have a 10% significance level
* , the probability of Type I error, is set to 0.10
* In general, confidence level = 100(1-).
* The probability of correctly retaining is designed to be 0.90.

## Specifying Error Probabilities (Type II)

* A Type II error is made if the alternative hypothesis is true, but you fail to choose it.
  + The probability depends on exactly which part of the alternative hypothesis is true, so that computing the probability of making a Type II error is not feasible.
* The **power** of a test is the probability of making the correct decision when the alternative hypothesis is true.
* is defined as the probability of concluding that there was no difference, when in fact there was one (a Type II error).
* A more powerful test has a lower Type II error rate, .

## Significance () and power ().

In many sample size decisions,

* we find that people set , the tolerable rate of Type I error, to be 0.05.
* they then often try to set the sample size and other parameters so that the power (1 - ) is at least 0.80.

We’ll advocate for thinking hard about the relative costs of Type I and Type II errors.

* Assuming that a power of 80% with a significance level of 5% is sufficient for most studies is not a great assumption.

## Power and Sample Size Calculations

A power calculation is likely the most common element of an scientific grant proposal on which a statistician is consulted.

* The tests that have power calculations worked out in intensive detail using R are mostly those with more substantial assumptions.
  + t tests that assume population normality, common population variance and balanced designs in the independent samples setting
  + paired t tests that assume population normality

## Power and Sample Size Calculations

* These power calculations are also usually based on tests rather than confidence intervals.
* This process of doing power and related calculations is far more of an art than a science, which when done properly, involves a lot of iteration.
* Simulation is your friend here when confronted with more complex scenarios.

# Power and Sample Size: Comparing Means

## Paired vs. Independent Samples

If you can afford to obtain n = 400 observations to compare means under exposure A to means under exposure B, and you could either:

1. select a random sample from the population of interest containing 400 people and then randomly assign 200 people to receive exposure A and the remaining 200 people to receive exposure B (thus doing an independent samples study), or
2. select a random sample from the population of interest containing 200 people and then randomly assign 100 of them to get exposure A first, and then, a little later, when the effects have worn off, to then receive exposure B, while the other 100 people are assigned to receive B first, then A (thus doing a paired samples study)

Assuming the effect size is unchanged, which seems as though it would be the more powerful study design?

## How do power calculations work?

Characteristics of a study comparing two means using **paired** samples:

* The sample size (number of pairs of measurements).
* The difference in means we’d like to be able to detect.
* The standard deviation for that difference in means.
* The significance (or confidence) level for our comparison.
* The power of the test.

If we can estimate any four of these values, we can calculate the fifth.

## Paired Samples t test

* Goal: detect a paired mean difference as small as 1/4 of a standard deviation, using 200 pairs of observations, and a 90% confidence level.

power.t.test(n = 200, delta = 0.25, sd = 1, sig.level = 0.1, type = "paired")

Paired t test power calculation   
  
 n = 200  
 delta = 0.25  
 sd = 1  
 sig.level = 0.1  
 power = 0.9698521  
 alternative = two.sided  
  
NOTE: n is number of \*pairs\*, sd is std.dev. of \*differences\* within pairs

## What sample size do we need?

How many pairs of observations would we need to maintain 80% power in our paired samples study?

power.t.test(delta = 0.25, sd = 1, sig.level = 0.10,   
 power = 0.80, type = "paired")

Paired t test power calculation   
  
 n = 100.2877  
 delta = 0.25  
 sd = 1  
 sig.level = 0.1  
 power = 0.8  
 alternative = two.sided  
  
NOTE: n is number of \*pairs\*, sd is std.dev. of \*differences\* within pairs

* Note that we’d need 101 pairs of measurements.

## How do power calculations work?

Characteristics of a study comparing two means with **independent** samples:

* The sample size in each group (balanced design)?
* The difference in means we’d like to be able to detect.
* The standard deviation for that difference in means.
* The significance (or confidence) level for our comparison.
* The power of the test.

If we estimate any four of these, R will calculate the fifth.

## Independent Samples t test

* Goal: detect a difference in group means as small as 1/4 of a standard deviation, using 200 observations in each group, and a 90% confidence level.

power.t.test(n = 200, delta = 0.25, sd = 1, sig.level = 0.1)

Two-sample t test power calculation   
  
 n = 200  
 delta = 0.25  
 sd = 1  
 sig.level = 0.1  
 power = 0.8025858  
 alternative = two.sided  
  
NOTE: n is number in \*each\* group

## What sample size do we need?

How many observations would we need in each group to reach 90% power in our independent samples study?

power.t.test(delta = 0.25, sd = 1, sig.level = 0.10, power = 0.90)

Two-sample t test power calculation   
  
 n = 274.7222  
 delta = 0.25  
 sd = 1  
 sig.level = 0.1  
 power = 0.9  
 alternative = two.sided  
  
NOTE: n is number in \*each\* group

* So, we’d need 275 measurements in each group, or 550 overall.

## Changing assumptions?

In our independent-samples test, we chose

* n = 200 (per group)
* delta = 0.25 (the minimum clinically important difference in means that we want to detect)
* sd = 1 (assumed population standard deviation in each group)
* sig.level = 0.10 (since we want 90% confidence)

## Which direction will power move in?

Original Setup yielded power = 0.802

* If we change from 200 to 400, leaving everything else untouched, do you think the power will increase or decrease?
* If we change from 200 to 400, power = 0.970
* What if we change from 200 to 100?
* If we change from 200 to 100, power = 0.546

## Changing other parameters

Original power = 0.802. Which changes **increase** the power?

| New Setup | Resulting Power |
| --- | --- |
| a. from 0.25 to 0.5 | Higher or Lower than 0.802? |
| b. from 0.25 to 0.1 | ? |
| c. sd from 1 to 2 | ? |
| d. sd from 1 to 0.5 | ? |
| e. from 0.1 to 0.05 | ? |
| f.  from 0.1 to 0.2 | ? |

## Changes a, d, f increase power

Original Setup yielded power = 0.802

| Change | Resulting power |
| --- | --- |
| a. from 0.25 to 0.5 | 0.9996 |
| b. from 0.25 to 0.1 | 0.259 |
| c. sd from 1 to 2 | 0.345 |
| d. sd from 1 to 0.5 | 0.9996 |
| e. from 0.1 to 0.05 | 0.703 |
| f.  from 0.1 to 0.2 | 0.888 |

## Balanced vs. Unbalanced design?

The most efficient design for an independent samples comparison will be balanced.

* Our original setup for , sd and , (with n = 200 in each group yielded power = 0.802). Now, we place:
  + 150 subjects into one exposure group, and
  + plan to recruit X (larger than 150) for the other.
* How many people would we have to recruit into the second exposure group to yield the same power as our original 200 in each group result?

## pwr.t2n.test from the pwr package

* Note the use here of d = /sd.

pwr.t2n.test(n1 = 150, d = 0.25/1, sig.level = 0.10, power = 0.802)

t test power calculation   
  
 n1 = 150  
 n2 = 298.1132  
 d = 0.25  
 sig.level = 0.1  
 power = 0.802  
 alternative = two.sided

So we can either have 200 and 200, or we can have 150 and 299 to maintain the same power.

## Assessing Unbalanced Designs

The power is always stronger for a balanced design than for an unbalanced design with the same overall sample size.

### One-Sided or Two-Sided Comparisons

Note that I used a two-sided test to establish my power calculation - in general, this is the most conservative and defensible approach for any such calculation, unless there is a strong and specific reason to use a one-sided approach in building a power calculation, don’t.

# Power and Sample Size: Comparing Proportions

## Designing a New Study

(PI): In our pilot, we saw = 0.18 and = 0.24. Help me design a new study using a two-sided test with . This time, let’s have as many in groupA as in groupB. We should have 90% power to detect a difference almost as large as what we saw in the pilot, or larger, so a difference of 6 percentage points.

What sample size is required to achieve these aims?

## How power.prop.test works

We specify 4 of the following 5 elements of the comparison, and R calculates the fifth.

* The sample size (interpreted as the # in each group, so half the total sample size)
* The true probability in group 1
* The true probability in group 2
* The significance level ()
* The power (1 - )

Requires you to work with balanced designs.

## Using power.prop.test

To find the sample size for a two-sample comparison of proportions using a balanced design:

* we will use a two-sided test, with = .05, and power = .90,
* we estimate that groupA has probability .18 of our outcome,
* and we will try to detect a difference between this group and groupB, who we estimate will have a probability of .24 of our outcome.

Any guess as to needed sample size?

## Finding the required sample size

power.prop.test(p1 = .18, p2=.24, alternative = "two.sided",  
 sig.level = 0.05, power = 0.90)

Two-sample comparison of proportions power calculation   
  
 n = 966.3554  
 p1 = 0.18  
 p2 = 0.24  
 sig.level = 0.05  
 power = 0.9  
 alternative = two.sided  
  
NOTE: n is number in \*each\* group

* We’d need at least 967 groupA subjects, and 967 more in groupB to accomplish these goals: a total of 1934 subjects.

## A One-Sided Test?

Suppose we get 400 groupA and 400 groupB subjects. With what power could we detect the difference between 0.18 in one group and 0.24 in the other, in a *one-sided* = .10 test?

power.prop.test(n=400, p1=.18, p2=.24, sig.level = 0.10,  
 alternative="one.sided")

Two-sample comparison of proportions power calculation   
  
 n = 400  
 p1 = 0.18  
 p2 = 0.24  
 sig.level = 0.1  
 power = 0.7892697  
 alternative = one.sided  
  
NOTE: n is number in \*each\* group

## Using the pwr package to assess sample size for Unbalanced Designs

The pwr.2p2n.test function in the pwr package can help assess the power of a test to determine a particular effect size using an unbalanced design, where is not equal to .

As before, we specify four of the following five elements of the comparison, and R calculates the fifth.

## Now the five elements are…

* n1 = The sample size in group 1
* n2 = The sample size in group 2
* sig.level = The significance level ()
* power = The power (1 - )
* h = the effect size, which can be calculated separately in R based on the two proportions being compared: and .

## Calculating the Effect Size h

To calculate the effect size for a given set of proportions, use ES.h(p1, p2) which is available in the pwr package.

For instance, comparing .18 to .24, we have the following effect size.

ES.h(p1 = .18, p2 = .24)

[1] -0.1476473

Suppose we can recruit 800 subjects in groupA but only 400 in groupB.

## What power would we have?

How much power would we have to detect the distinction between p1 = .18 and p2 = .24 with a 90% confidence level (two-sided) with 800 subjects in groupA and 400 in groupB?

pwr::pwr.2p2n.test(h = ES.h(p1 = .18, p2 = .24), n1 = 800, n2 = 400,   
 sig.level = 0.1)

difference of proportion power calculation for binomial distribution (arcsine transformation)   
  
 h = 0.1476473  
 n1 = 800  
 n2 = 400  
 sig.level = 0.1  
 power = 0.7782513  
 alternative = two.sided  
  
NOTE: different sample sizes

## Comparison to Balanced Design

How does this compare to the results with a balanced design using 1200 subjects in total, i.e. 600 per group?

pwr::pwr.2p2n.test(h = ES.h(p1 = .18, p2 = .24), n1 = 600, n2 = 600,   
 sig.level = 0.10)

difference of proportion power calculation for binomial distribution (arcsine transformation)   
  
 h = 0.1476473  
 n1 = 600  
 n2 = 600  
 sig.level = 0.1  
 power = 0.8192533  
 alternative = two.sided  
  
NOTE: different sample sizes

## We could instead have used…

power.prop.test(p1 = .18, p2 = .24, sig.level = 0.10, n = 600)

Two-sample comparison of proportions power calculation   
  
 n = 600  
 p1 = 0.18  
 p2 = 0.24  
 sig.level = 0.1  
 power = 0.8183448  
 alternative = two.sided  
  
NOTE: n is number in \*each\* group

Each approach we’ve used includes approximations, and slightly different ones, so it’s not surprising that the answers are similar, but not identical.

## Some people will drop out

Suppose my calculation says I need 1000 subjects in my study. How many subjects do I need to recruit?

* Suppose I expect at most 20% of recruits to drop out before the study ends.
* If I expect at most 20% of recruits to drop out, then perhaps I should recruit enough that if 20% leave, I still have 1000.
* If I expect at most 20% of recruits to drop out, that means I would retain at least 80% of the subjects.
* So I recruit R = 1000 / 0.8 = 1250 people. If I retain 80% of my 1250 recruits, that’s 0.8 (1250) = 1000 subjects.

## Power Calculations

Most common scenario: identify desired significance level and desired power (1 - ) and the details of the plan in terms of the comparison to be made, and how the data are to be collected.

Then calculate a minimum necessary sample size to achieve these desires, with and commonly used.

* Neither 95% confidence nor 80% power is a magical choice.
* Anything below 80% power will be hard to justify in real work.

## A useful metaphor?

Sometimes I like to think of science as a march towards a destination.

Actually, I suppose it’s an infinitely long march towards an ever-receding destination, but let’s leave the philosophy out of it for a moment.

Suppose, for example, that we’re trying to make a meaningful change in the world, perhaps to treat an infection.

What we’re trying to do is related to where we are in the March of Science.

## Early vs. Late in the March of Science

In **early** work, we’re focused more on discovery than making final decisions.

* We don’t have a lot of past experience, so we bring little relevant data to the table.
* We’re (often) most concerned about discovering new possibilities, and we don’t have a very clear sense of where to go next.
* We’re (often) less concerned about false starts than we are about missed opportunities.

In **late** work, we’re focused more on making a decision about how to treat.

* We have a fair amount of relevant history to draw on, sometimes quite detailed.
* We’re more concerned about testing the limits of our current knowledge than we are about missing opportunities to consider a new pathway.
* We’re often concerned about doing harm if we implement the strategy that looks most promising.

## Power and “significance”?

If we treat our sample size and study design as fixed strategies, then we must choose between:

* reducing , the rate of Type I error (increasing our confidence) and
* reducing , the rate of Type II error (increasing our power)

Suppose we are testing a new treatment for some condition.

* A Type I error means we conclude this treatment is helpful, when it actually isn’t.
* A Type II error means we conclude this treatment is not helpful, when it actually is.

## Early Work: Power and Sample Size

In early work, we are searching for treatments of promise, and our initial study will inevitably not be the last word on the subject, but rather will be followed up by confirmatory studies. In such a setting, it is often the case that:

* We’re not so concerned about getting results that cause us to continue to explore a treatment that doesn’t actually do what we need it to do.
* We’re really concerned about ruling out a treatment that is promising before we should.

This implies we should prioritize reducing Type II error rates (we want more power to detect small but real effects, even if this means we will occasionally identify something as promising when it isn’t.)

This means setting lower confidence levels and higher power levels, potentially, than the standard 95% confidence and 80% power.

## Late Work: Power and Sample Size

In late work, we have already identified promising treatments, and we are trying to confirm those results. The current study may actually be the last word on the subject, and we want to be sure we do no harm.

* We’re very concerned about getting results that cause us to continue to explore a treatment that doesn’t actually do what we need it to do.
* We’re less concerned about ruling out a treatment that is promising but doesn’t actually work.

This implies we should prioritize reducing Type I error rates (we want greater confidence, even at the expense of power, that the effect we claim based on past data holds up.)

This kind of confirmatory work is usually well suited to studies set up with higher confidence (perhaps 95% or 99% or more) and lower power (80% is the minimum I would recommend) against reasonable alternatives.

## Conclusions

1. If you’re early in the March of Science (perhaps just one pilot study has been done) then I would emphasize Type II error (power) more than usual, perhaps pushing required power to 90%, at least for a reasonably substantial “minimum scientifically important difference” .
2. If you’re late in the March of Science (perhaps confirming the results of multiple prior studies) then I would be happier with 80% power and higher levels of confidence.
3. If you’re in the middle, trading off Type I and Type II error is worth some thought, but I’d never recommend being under 80% power for an effect that matters.
4. If it’s feasible to run a study large enough to have strong performance on both and , that’s obviously ideal. Typically that doesn’t happen in early work.

## So … How Big A Sample Do I Need?

1. What is your research question?
2. What is the budget?

## More Examples?

See the **old** (pre-2024) 431 class notes:

* <https://thomaselove.github.io/431-notes/22-samplesize.html>
* <https://thomaselove.github.io/431-notes/27-power_for_rates.html>
* <https://easystats.github.io/effectsize/articles/statistical_power.html> for interesting initial thoughts[[1]](#footnote-69) on power and sample size from the easystats framework.

1. I keep waiting for this reference to be updated, but it hasn’t been in a while. [↑](#footnote-ref-69)