431 Class 20

github.com/THOMASELOVE/2019-431

2019-11-07

Today's Setup (No Data so no Tidyverse!)

```r
library(pwr) # new today
library(here)

source(here("R", "Love-boost.R"))
```

Today's Agenda (Notes Chapters 25-27)

- Power and Sample Size Considerations Comparing 2 Means
  - With `power.t.test` for balanced designs (independent & paired)
  - With `pwr` for unbalanced designs
- Power and Sample Size for ANOVA comparisons of 3+ Means
  - With `power.anova.test` for balanced designs
- Power and Sample Size when comparing Proportions
  - With `power.prop.test` for balanced designs using independent samples
  - With `pwr` for unbalanced designs
- What We're NOT talking about today

Types of Errors

In a hypothesis testing structure, we have a null hypothesis H_0 and an alternative hypothesis H_A.

- α is the probability of rejecting H_0 when H_0 is true.
    - So $1 - \alpha$, the confidence level, is the probability of retaining H_0 when that's the right thing to do.
    - Rejection of H_0 when H_0 is true is referred to as a **Type I error**.
    - So α is the probability of Type I error, associated with our scenario.
- β is the probability of retaining H_0 when H_A is true.
    - So $1 - \beta$, the power, is the probability of rejecting H_0 when that's the right thing to do.
    - Retaining H_0 when H_A is actually true is referred to as a **Type II error**.
    - So β is the probability of Type II error, associated with our scenario.

Error Types, Confidence, Power, α and β

- α is the probability of rejecting H_0 when H_0 is true.
    - So $1 - \alpha$, the confidence level, is the probability of retaining H_0 when that's the right thing to do.
- β is the probability of retaining H_0 when H_A is true.
    - So $1 - \beta$, the power, is the probability of rejecting H_0 when that's the right thing to do.

–	H_A is True	H_0 is True
Test Rejects H_0	Correct Decision (1 - β)	Type I Error (α)
Test Retains H_0	Type II Error (β)	Correct Decision (1 - α)

Most common (not necessarily sensible) approach

- Pre-specify the significance level α to be 0.05.
- Pre-specify the power (1 - β) to be 0.80.

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. And what sort of statistical inference do you want to plan for?

Section 1

Comparing Two Means (with Paired or Independent Samples)

Using `power.t.test`

Measure	Paired Samples	Independent Samples
`type =`	`"paired"`	`"two.sample"`
n	# of paired diffs	# in each sample
δ	true mean of diffs	true diff in means
$s =$ sd	true SD of diffs	true SD, either group[1]
$\alpha =$ sig.level	max. Type I error rate	Same as paired.
$1 - \beta =$ power	power to detect effect δ	Same as paired.

Specify `alt = "greater"` or `alt = "less"` for a 1-sided comparison.

Sample Size & Power: Pooled t Test

For an independent-samples t test, with a balanced design $(n_1 = n_2)$ then R can estimate any one of the following elements, given the other four, using the `power.t.test` function, for a one-sided or two-sided t test.

- $n = n_1 = n_2 =$ the sample size in each of the two groups being compared
- $\delta =$ delta $=$ the true difference in means between the two groups
- s $=$ sd $=$ the true standard deviation of the individual values in each group (assumed to be constant, since we assume equal population variances)
- $\alpha =$ sig.level $=$ the significance level for the comparison (maximum acceptable risk of Type I error)
- 1 - $\beta =$ power $=$ the power of the t test to detect the effect of size δ

If you want a two-sample power calculation for an unbalanced design, you will need to use a different package and function in R.

A Small Example: Studying Satiety

- I want to compare people eating meal A to people eating meal B in terms of impact on satiety, as measured on a 0-100 scale. I'm interested in a two-sided test, since I don't know which will be better in advance.
- I can afford to enroll 160 people (or, more specifically, prepare 160 meals) in the study.
- I expect that a difference that's important will be about 10 points on the satiety scale.
  - Perhaps this is because I saw a **17** point difference in a pilot study.
- I don't know the standard deviation, but the whole range (0-100) gets used, so I'll estimate the SD conservatively with (range/4) or 25.

The Key Questions

- How many should eat meal A and how many meal B to maximize my power to detect such a difference?
- And how much power will I have if I use a 90% confidence level?

Satiety Example: Power

- n = the sample size in each of the two groups being compared
- δ = delta = the true difference in means between the two groups
- s = sd = the true standard deviation of the individual values in each group (assumed to be constant, since we assume equal population variances)
- α = sig.level = the significance level for the comparison (maximum acceptable risk of Type I error)
- 1 - β = power = the power of the t test to detect the effect of size δ

What do I know?

Satiety: Assumption Set 1

Let's go with:

- Independent Samples, so test type = "`two.sample`"
- Total sample size I can afford = 160, so a balanced design gives 80 in each group.
- δ = smallest difference I want to be sure I detect = 10 points.
- standard deviation of satiety scores is unknown, but we'll guess 25 (since range/4 = 25).
- we'll use alpha = .10 (90% confidence) and a two-sided test.

Satiety Example Calculation

```r
power.t.test(n = 80, delta = 10, sd = 25,
             sig.level = 0.10, alt = "two.sided",
             type = "two.sample")
```

```

     Two-sample t test power calculation

              n = 80
          delta = 10
             sd = 25
      sig.level = 0.1
          power = 0.8089716
    alternative = two.sided

NOTE: n is number in *each* group
```

What happens if you change something else?

If we start with $\texttt{n} = 80$, $\delta = 10$, sd $= 25$, $\alpha = 0.10$, that yields power 0.809.

To increase the power to 0.90, what can we do?

1. Increase the sample size, *n* to 108 in each group
    - If $\texttt{n} = 108$ instead, power is now 0.90
2. *or* Increase the minimum detectable effect size δ to 11.7
3. *or* Reduce the standard deviation to 21.5
4. *or* Increase α (willingness to tolerate Type I error) to 0.215
    - So confidence level would 79.5% instead of 90%.
5. *or* Switch from independent to matched/paired samples.
    - Even if we leave `sd` alone, 55 matched differences yields 90% power

What if we use a paired samples design instead?

- Paired samples, so test `type = "paired"`
- Total = 160 meals, so that's `n = 80` paired differences (each subject eats A and B, in a random order)
- standard deviation of *differences* in satiety will be smaller than 25; let's be very conservative and say 20.
- δ remains 10 points, two-sided test with 90% confidence.

So, we want...

```r
power.t.test(n = 80, delta = 10, sd = 20,
             sig.level = 0.10, alt = "two.sided",
             type = "paired")
```

Any guesses?

Satiety: Paired Samples Estimated Power (Results)

```r
power.t.test(n = 80, delta = 10, sd = 20,
             sig.level = 0.10, alt = "two.sided",
             type = "paired")
```

```
     Paired t test power calculation 

              n = 80
          delta = 10
             sd = 20
      sig.level = 0.1
          power = 0.9973528
    alternative = two.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences*
```

Note: If sd $= 25$ with paired samples, power $= 0.971$.

What if 32 people ate both meals (different times?)

Impact on standard deviation? Let's say $\sigma_d = 15$...

```r
power.t.test(delta = 10, sd = 15, sig.level = 0.10,
        n = 32, alt = "two.sided", type = "paired")
```

```

     Paired t test power calculation 

              n = 32
          delta = 10
             sd = 15
      sig.level = 0.1
          power = 0.979437
    alternative = two.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences*
```

Power for an unbalanced design

- If you have independent samples, the most powerful design for a given total sample size will always be a balanced design.
- If you must use an unbalanced design in setting up a sample size calculation, you typically have meaningful information about the cost of gathering samples in each group, and this may help you estimate the impact of Type I and Type II errors so you can trade them off appropriately.

The `pwr` package function to do this is `pwr.t2n.test`.

- Must specify both n_1 and n_2
- Instead of specifying δ and sd separately, specify their ratio, d.

Satiety Example with an Unbalanced Design?

If we can only get 40 people in the tougher group to fill, how many people would we need in the easier group to get at least 80% power to detect a difference of 10 points, assuming a standard deviation of 25, and using 90% confidence.

Remember that we met this standard with 80 people in each group using a balanced design. . .

So we have $n_1 = 40$, d = 10/25 (δ / sd), sig.level $\alpha = 0.1$ and power (1 - β) = 0.8

- What's your guess, before I show you the answer, as to the number of people I'll need in the easier group?

Satiety Example, Unbalanced Design

```
pwr::pwr.t2n.test(n1 = 40, d = 10/25, sig.level = .1,
             power = .80, alt="two.sided")
```

```
     t test power calculation

             n1 = 40
             n2 = 1174.101
              d = 0.4
      sig.level = 0.1
          power = 0.8
    alternative = two.sided
```

- OK. So use balanced samples when planning a study.
- What's next?

Section 2

Designing an ANOVA?

Power/Sample Size for designing an ANOVA study

Is there a `power.anova.test` approach in R?

Sure.

- `groups` = number of groups
- `n` = number of observations per group
- `between.var` = between-group variance
- `within.var` = within-group variance
- `sig.level` = α (significance level)
- `power` = $1 - \beta$

Specify five, and the computer will calculate the sixth. This does require a **balanced design**.

So, what do we use for `between.var` and `within.var`?

Determining `between.var` and `within.var`

- If you have prior knowledge of what you expect the true group means to be, then you can take their variance to get the `between.var` value.
- The `within.var` value is the within-group variance. To get that, realize that ANOVA assumes that each group will have the same standard deviation of outcome values. Square that "within-group standard deviation" you estimate to obtain the within-group variance.

Powering an ANOVA study, Setup

PI wants to plan a study:

- to compare four groups, and she wants to be sure she can detect a difference if the means turn out to be any more different than they would be if they were 560, 585, 610 and 625.
- using a balanced design, 90% power and $\alpha = 0.05$
- where she thinks that the standard deviation in each group of the scores will be 80.

`power.anova.test` needs five of these six things:

- `groups` = number of groups
- `n` = number of observations per group
- `between.var` = between-group variance
- `within.var` = within-group variance
- `sig.level` = α (significance level)
- `power` = $1 - \beta$

Powering an ANOVA study, Results

```r
groupmeans <- c(560, 585, 610, 625)
power.anova.test(groups = 4,
                 between.var = var(groupmeans),
                 within.var = 80^2, power = 0.90)
```

```
     Balanced one-way analysis of variance power calculation 

         groups = 4
              n = 38.01195
    between.var = 816.6667
     within.var = 6400
      sig.level = 0.05
          power = 0.9

NOTE: n is number in each group
```

Is there a power and sample size approach for unbalanced ANOVA?

Not a straightforward one, and not within the `pwr` package, no.

- This is a good time for us to mention that the number one place where Ph.D. statisticians get involved in grant proposals is when people realize they need a non-standard power calculation.

Section 3

Comparing Proportions (using Independent Samples)

Tuberculosis Prevalence Among IV Drug Users

Suppose we are investigating factors affecting tuberculosis prevalence among intravenous drug users.

We collect the following information:

- Among 97 individuals who admit to sharing needles,
  - 24 (24.7%) had a positive tuberculin skin test result.
- Among 161 drug users who deny sharing needles,
  - 28 (17.4%) had a positive test result.

What does the 2x2 table look like?

Tuberculosis Prevalence Among IV Drug Users

Among 97 individuals who admit to sharing needles, 24 (24.7%) had a positive tuberculin skin test result; among 161 drug users who deny sharing needles, 28 (17.4%) had a positive test result.

The 2x2 Table is. . .

```
         TB+   TB-
share     24    73
don't     28   133
```

- rows describe needle sharing, columns describe TB test result
- row 1 people who share needles: 24 TB+, and 97-24 = 73 TB-
- row 2 people who don't share: 28 TB+ and 161-28 = 133 TB-

twobytwo (with Bayesian Augmentation)

To start, we'll test the null hypothesis that the population proportions of intravenous drug users who have a positive tuberculin skin test result are identical for those who share needles and those who do not.

$$H_0 : \pi_{share} = \pi_{donotshare}$$

$$H_A : \pi_{share} \neq \pi_{donotshare}$$

We'll use the Bayesian augmentation.

```r
twobytwo(24+1, 73+1, 28+1, 133+1,
         "Sharing", "Not Sharing",
         "TB test+", "TB test-")
```

Two-by-Two Table Result

```
Outcome   : TB test+
Comparing : Sharing vs. Not Sharing

          TB test+ TB test- P(TB test+) 95% conf. int.
Sharing         25       74      0.2525  0.1767 0.3471
Not Sharing     29      134      0.1779  0.1265 0.2443

                                    95% conf. interval
             Relative Risk: 1.4194    0.8844   2.2779
         Sample Odds Ratio: 1.5610    0.8520   2.8603
Conditional MLE Odds Ratio: 1.5582    0.8105   2.9844
    Probability difference: 0.0746   -0.0254   0.1814

Exact P-value: 0.1588      Asymptotic P-value: 0.1495
```

What conclusions should we draw?

Designing a New TB Study

PI:

- OK. That's a nice pilot.
- We saw $p_{nonshare} = 0.18$ and $p_{share} = 0.25$ after your augmentation.
- Help me design a new study.
    - This time, let's have as many needle-sharers as non-sharers.
    - We should have 90% power to detect a difference as large as what we saw in the pilot, or larger, so a difference of 7 percentage points.
    - We'll use a two-sided test, and $\alpha = 0.05$, of course.

What sample size would be required to accomplish these aims?

How `power.prop.test` works

`power.prop.test` works much like the `power.t.test` we saw for means.

Again, 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 (α)
- The power (1 - β)

The big weakness with the `power.prop.test` tool is that it doesn't allow you to work with unbalanced designs.

Using `power.prop.test` for Balanced Designs

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

- we will use a two-sided test, with $\alpha = .05$, and power $= .90$,
- we estimate that non-sharers have probability .18 of positive tests,
- and we will try to detect a difference between this group and the needle sharers, who we estimate will have a probability of .25

Finding the required sample size in R

```
power.prop.test(p1 = .18, p2  = .25,
                alternative = "two.sided",
                sig.level = 0.05, power = 0.90)
```

Any guess as to needed sample size?

Results: `power.prop.test` for Balanced Design

```r
power.prop.test(p1 = .18, p2  = .25,
                alternative = "two.sided",
                sig.level = 0.05, power = 0.90)
```

```
Two-sample comparison of proportions power calculation
n = 721.7534
p1 = 0.18, p2 = 0.25
sig.level = 0.05, power = 0.9, alternative = two.sided
NOTE: n is number in *each* group
```

So, we'd need at least 722 non-sharing subjects, and 722 more who share needles to accomplish the aims of the study, or a total of 1444 subjects.

Another Scenario

Suppose we can get 400 sharing and 400 non-sharing subjects. How much power would we have to detect a difference in the proportion of positive skin test results between the two groups that was identical to the data above or larger, using a *one-sided* test, with $\alpha = .10$?

```r
power.prop.test(n=400, p1=.18, p2=.25, sig.level = 0.10,
                alternative="one.sided")
```

```
Two-sample comparison of proportions power calculation
n = 400, p1 = 0.18, p2 = 0.25
sig.level = 0.1, power = 0.8712338
alternative = one.sided
NOTE: n is number in *each* group
```

We would have just over 87% power to detect such an effect.

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 n_1 is not equal to n_2.

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

- `n1` = The sample size in group 1
- `n2` = The sample size in group 2
- `sig.level` = The significance level (α)
- `power` = The power (1 - β)
- `h` = the effect size h, which can be calculated separately in R based on the two proportions being compared: p_1 and p_2.

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, in our comparison, we have the following effect size.

```r
ES.h(p1 = .18, p2 = .25)
```

```
[1] -0.1708995
```

Using `pwr.2p2n.test` in R

Suppose we can have 700 samples in group 1 (the not sharing group) but only 400 in group 2 (the group of users who share needles).

How much power would we have to detect this same difference (p1 = .18, p2 = .25) with a 5% significance level in a two-sided test?

R Command to find the resulting power

```
pwr::pwr.2p2n.test(h = ES.h(p1 = .18, p2 = .25),
                   n1 = 700, n2 = 400, sig.level = 0.05)
```

Results of using pwr.2p2n.test

```
pwr::pwr.2p2n.test(h = ES.h(p1 = .18, p2 = .25),
                   n1 = 700, n2 = 400, sig.level = 0.05)
```

```
difference of proportion power calculation
for binomial distribution (arcsine transformation)

h = 0.1708995, n1 = 700, n2 = 400
sig.level = 0.05, power = 0.7783562
alternative = two.sided
NOTE: different sample sizes
```

We will have just under 78% power under these circumstances.

Comparison to Balanced Design

How does this compare to the results with a balanced design using 1100 drug users in total, i.e. with 550 patients in each group?

```
pwr::pwr.2p2n.test(h = ES.h(p1 = .18, p2 = .25),
                   n1 = 550, n2 = 550, sig.level = 0.05)
```

which yields a power estimate of 0.809. Or we could instead have used…

```
power.prop.test(p1 = .18, p2 = .25, sig.level = 0.05,
                n = 550)
```

which yields an estimated power of 0.808.

Each approach uses approximations, and slightly different ones, so it's not surprising that the answers are similar, but not identical.

What haven't I included here?

1. Some people will drop out.
2. What am I going to do about missing data?
3. What if I want to do my comparison while adjusting for covariates?
4. And what if I want to compare proportions using paired samples?

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. And what sort of statistical inference do you want to plan for?

Coming Soon

- Working with Larger Contingency Tables (Chi-Square Tests of Independence)
- Mantel-Haenszel Procedures for Three-Way Tables