PSY 201: Statistics in Psychology

Lecture 26

Hypothesis testing for two means

Planning a replication study.

Greg Francis

Purdue University

Fall 2019

TESTING MEANS

- we want to test

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 - \mu_2 \neq 0$$

- but the techniques of last time require $\sigma_1^2 = \sigma_2^2$
- pooled estimate of variance

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

- and use the t-distribution

REVISED TESTING MEANS

- when

$$\sigma_1^2 \neq \sigma_2^2$$

- we must make two changes
  - different estimate of standard error of the difference $s_{\overline{X}_1 - \overline{X}_2}$
  - adjustment of degrees of freedom
- still use the t distribution

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)}$$

- or

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{s_{\overline{X}_1}^2 + s_{\overline{X}_2}^2}$$

DEGREES OF FREEDOM

- when $\sigma_1^2 \neq \sigma_2^2$ we calculate *df* as:

$$df = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$$

- or

$$df = \frac{\left(s_{\overline{X}_1}^2 + s_{\overline{X}_2}^2\right)^2}{\left(s_{\overline{X}_1}^2\right)^2/(n_1-1) + \left(s_{\overline{X}_2}^2\right)^2/(n_2-1)}$$

- looks (and is) messy
- just a matter of plugging in numbers carefully
- still use the *t*-test as before!
- We call it Welch's test

EXAMPLE

- A researcher wants to know if single or married parents are more satisfied with their status. She randomly samples 61 single and 161 married parents. Each parent rates her/his marital status satisfaction, with higher scores indicating greater satisfaction. The researcher wants to know if there is a difference between the population means of single versus married parents.
- data summary

Variable	n	$\overline{X}$	s	$s_{\overline{X}}$
Group 1	61	2.6557	0.602	0.077
Group 2	161	2.7516	0.461	0.036

HYPOTHESES

$$H_0 : \mu_1 - \mu_2 = 0$$

- indicating there is no difference in satisfaction between the two groups

$$H_a : \mu_1 - \mu_2 \neq 0$$

- indicating there is a difference in satisfaction between the two groups
- not an ordered hypothesis because we do not know who might be more satisfied
- level of significance is set at $\alpha = 0.05$

WORRY ABOUT HOMOGENEITY

- We do not know the true values of σ_1 and σ_2, but we notice that $n_1 < n_2$ and that $s_1 > s_2$.
- This makes us worry that maybe our Type I error rate will be off (and maybe too big), so we use Welch's *t*-test
- The online calculator in the textbook uses Welch's test unless $n_1 = n_2$

TEST STATISTIC

- pooled standard error estimate

$$s_{\overline{X}_1-\overline{X}_2} = \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}$$

$$s_{\overline{X}_1-\overline{X}_2} = \sqrt{\left(\frac{(0.602)^2}{61} + \frac{(0.461)^2}{161}\right)}$$

$$s_{\overline{X}_1-\overline{X}_2} = 0.075683$$

TEST STATISTIC

- the formula for the test statistic is

$$\text{Test statistic} = \frac{\text{Statistic} - \text{Parameter}}{\text{Standard Error}}$$

- or

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_{\overline{X}_1 - \overline{X}_2}}$$

- or

$$t = \frac{(2.6557 - 2.75162) - (0)}{0.075683} = -1.267$$

TEST STATISTIC

- adjusted degrees of freedom

$$df = \frac{\left(s^2_{\overline{X}_1} + s^2_{\overline{X}_2}\right)^2}{\left(s^2_{\overline{X}_1}\right)^2/(n_1 - 1) + \left(s^2_{\overline{X}_2}\right)^2/(n_2 - 1)}$$

$$df = \frac{\left((0.077)^2 + (0.036)^2\right)^2}{\left((0.077)^2\right)^2/(61 - 1) + \left((0.036)^2\right)^2/(161 - 1)}$$

$$df = 87.995$$

p VALUE

- From the t-distribution calculator, we find (for a two-tailed test with $df = 87.995$) that

$$p = 0.208455 > \alpha = 0.05$$

- we do not reject H_0
  - there is no evidence that satisfaction with marital status differs for married versus single parents
  - the probability that the observed (or more extreme) difference in means would occur by chance if $\mu_1 - \mu_2 = 0$ is greater than 0.05

ONLINE CALCULATOR

- As always, it is best to use a computer. We can enter the summary statistics.

Enter data:

Sample size for group 1 n_1 = 61

Sample mean for group 1 $\overline{X}_1$ = 2.6557

Sample standard deviation for group 1 s_1 = 0.602

Sample size for group 2 n_2 = 161

Sample mean for group 2 $\overline{X}_2$ = 2.7516

Sample standard deviation for group 2 s_2 = 0.461

Specify hypotheses:

$H_0 : \mu_1 - \mu_2 =$ 0

H_a : Two-tails

$\alpha =$ 0.05

Run Test

ONLINE CALCULATOR

- You need to understand how to pull out the information you want

Test summary	
Type of test	Welch's Test
Null hypothesis	$H_0 : \mu_1 - \mu_2 = 0$
Alternative hypothesis	$H_a : \mu_1 - \mu_2 \neq 0$
Type I error rate	$\alpha = 0.05$
Label for group 1	Group 1
Sample size 1	$n_1 = 61$
Sample mean 1	$\overline{X}_1 = 2.6557$
Sample standard deviation 1	$s_1 = 0.602000$
Label for group 2	Group 2
Sample size 2	$n_2 = 161$
Sample mean 2	$\overline{X}_2 = 2.7516$
Sample standard deviation 2	$s_2 = 0.461000$
Pooled standard deviation	$s =$NA
Sample standard error	$s_{\overline{X}_1 - \overline{X}_2} = 0.075683$
Test statistic	$t = -1.267121$
Degrees of freedom	$df = 87.99504946388605$
p value	$p = 0.208455$
Decision	Do not the reject null hypothesis
Confidence interval critical value	$t_{cv} = 1.987291$
Confidence interval	$CI_{95} = (-0.246305, 0.054505)$

CONFIDENCE INTERVAL

- Basic formula for all confidence intervals:

$$CI = \text{statistic} \pm (\text{critical value})(\text{standard error})$$

- for a difference of sample means

$$CI = (\overline{X}_1 - \overline{X}_2) \pm t_{cv} s_{\overline{X}_1 - \overline{X}_2}$$

- We already have most of the terms (we get t_{cv} from the Inverse t-distribution calculator, so

$$CI_{95} = (2.6557 - 2.7516) \pm (1.987291)(0.075683)$$

$$CI_{95} = (-0.246305, 0.054505)$$

POWER

- Power is treated much the same as for the one-sample case
- We just have to keep track of whether we are using the standard t-test or Welch's test
- The on-line calculator of our textbook does this for you automatically
- Power is very important when designing an experiment

REPLICATION

- An important characteristic of science is *replication*
- Show that the same methods and measures produce the same results
- "Hard" sciences are very good at this (e.g., physics, chemistry)
- Sciences that depend on statistics face challenges
- We *always* face a risk of making a Type I or a Type II error
- Thus, successful replication is not expected even for real effects
- You can mitigate these problems by designing good replication studies that use the same methods, but have high power

INTERESTING STUDY

- Consider a study on how nonconformity can induce higher status in certain environments
- Participants were 52 shop assistants working in downtown Milan, Italy boutiques (Armani, Burberry, Christian Dior, La Perla, Les Copains, and Valentino)
- Two groups of 26 each read a vignette:
- Imagine that a woman is entering a luxury boutique in downtown Milan during summer. She looks approximately 35 years old.
- Nonconforming condition (Group 1): She is wearing plastic flip-flops and she has a Swatch on her wrist.
- Conforming condition (Group 2): She is wearing sandals with heels and she has a Rolex on her wrist.
- Rate the status of the woman on a scale of 1–7 (bigger means higher status)

INTERESTING STUDY

- The results are:
- Nonconforming

$$\overline{X}_1 = 4.8$$

- Conforming

$$\overline{X}_2 = 4.2$$

$$t(50) = 2.1$$

$$p = 0.0408$$

REPLICATION

- You want to repeat the study, but it is not easy to get shop assistants from high end stores (you might have to go to Chicago for your subjects)
- The online power calculator requires you to enter estimates of:

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_0 : \mu_{a1} - \mu_{a2} \approx \overline{X}_1 - \overline{X}_2 = 0.6$$

$$\sigma_1, \sigma_2$$

REPLICATION

- for the standard deviations, we use some algebra. We know that for the reported t-test:

$$2.1 = t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}} = \frac{0.6}{s_{\overline{X}_1 - \overline{X}_2}}$$

- so

$$s_{\overline{X}_1 - \overline{X}_2} = \frac{0.6}{2.1} = 0.28571$$

- We can assume the standard t-test was used, so $\sigma_1 = \sigma_2$. Thus

$$s_{\overline{X}_1 - \overline{X}_2} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = s\sqrt{\frac{1}{26} + \frac{1}{26}} = s(0.27735)$$

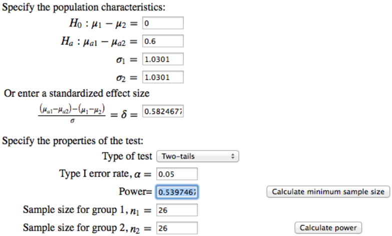
- so

$$s = \frac{0.28571}{0.27735} = 1.030157$$

- which we can use for both σ_1 and σ_2

REPLICATION

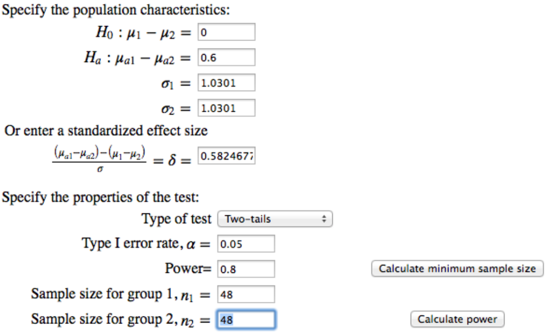
- Oftentimes researchers just use the same sample size as a previous study. After all, that study worked, so it must be an appropriate sample size, right?
- No, if we use $n_1 = n_2 = 26$, the on-line power calculator gives power=0.5397



- this should make sense because the $p = 0.04$ in the original study is just below the $\alpha = 0.05$ criterion
- if we take a different random sample, we will get a different p value, almost half the time it will be bigger than α

REPLICATION

- Suppose you want 80% power
- The calculator tells you that you need $n_1 = n_2 = 48$ participants. Nearly twice as big as the original study!



- if you do the replication correctly, you typically run a *better* study than the original
- That is common in science, where new experiments are better than old experiments

EFFECT SIZE

- The power calculator computes a term called δ. This is an estimate of d' between the null and specific alternative distributions. Bigger values of δ mean it is easier to notice a difference. It can be computed from the means and standard deviation estimates that you provide to the power calculator.
- An estimate, d, can also be computed from the t value and sample sizes

$$d = t\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (2.1)\sqrt{\frac{1}{26} + \frac{1}{26}} = 0.5824$$

- Often called Cohen's d

EFFECT SIZE

- You might worry that the effect size of the original study is an overestimate
- After all, if the researchers had not found a significant difference, they might not have published their paper (publication bias)
- A conservative approach is to divide the estimated effect size half, and do the power calculation from that new effect size.
- Thus, we can directly enter:

$$\delta = \frac{d}{2} = \frac{0.5824}{2} = 0.2912$$

- The power calculator now tells us that to have 80% power, we need $n_1 = n_2 = 187$ subjects
- This could be a very difficult experiment to run

CONCLUSIONS

- Welch's test
- Power
- Replication

NEXT TIME

- hypothesis testing for dependent samples
- sampling distribution
- standard error

Within is better than between.