PSY 201: Statistics in Psychology

Lecture 31
Multiple testing
Error is sneaky.

Greg Francis

Purdue University

Fall 2023

HYPOTHESIS TESTING

- we know how to test the difference of two means

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

- by using the t distribution and estimates of standard error
- what if you have more populations and what to know if they are all equal?

MULTIPLE t- TESTS

- if we have $K = 5$ population means, we might want to compare each mean to all the others
- requires
$$c = \frac{K(K-1)}{2} = 10$$
- different t-tests
- suppose each test is with $\alpha = 0.05$
- What is the Type I error?

MULTIPLE t- TESTS

- we have a risk of making a type I error for *each* t test
- since we have $c = 10$ different t-tests, with $\alpha = 0.05$, the Type I error rate becomes approximately

$$1 - (1 - \alpha)^c = 0.40$$

- bigger risk of error than you might expect!
- to be sure we do not make *any* Type I errors, we would need to set α much smaller to insure that Type I error rate is below 0.05!

ADJUST α

- To a first approximation, to make sure the Type I error rate for $c = 10$ tests is less than 0.05, you could set the α criterion for each t-test to be
$$\alpha = \frac{0.05}{c} = \frac{0.05}{10} = 0.005$$
- Then, the probability of any given test producing a Type I error is 0.005, and the probability than any of the 10 tests produces a Type I error is 0.05
- This is called the Bonferroni correction
- But decision making always involves trade offs.

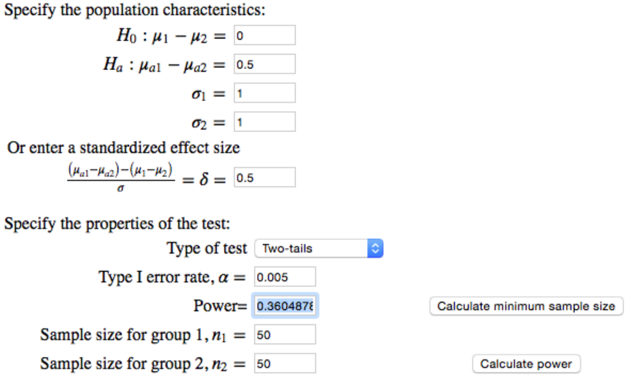
ADJUST α

- What kind of power do we have?
- Suppose $\sigma = 1$ and we take samples of size $n = 50$ for each condition
- If you use $\alpha = 0.005$, and one of the means, $\mu_1 = 0.5$, is *truly* different from the other four means, $\mu_2 = \mu_3 = \mu_4 = \mu_5 = 0$. What is the probability you will reject H_0?
- For the six tests that do not involve μ_1, the probability of any of them producing a Type I error is

$$1 - (1 - 0.005)^6 = 0.029$$

ADJUST α

- For the four tests involving μ_1, we can estimate power of each test, by using the on-line power calculator:


Specify the population characteristics:

$H_0 : \mu_1 - \mu_2 =$ 0

$H_a : \mu_{a1} - \mu_{a2} =$ 0.5

$\sigma_1 =$ 1

$\sigma_2 =$ 1

Or enter a standardized effect size

$\frac{(\mu_{a1}-\mu_{a2})-(\mu_1-\mu_2)}{\sigma} = \delta =$ 0.5

Specify the properties of the test:

Type of test: Two-tails

Type I error rate, $\alpha =$ 0.005

Power= 0.360487

Calculate minimum sample size

Sample size for group 1, $n_1 =$ 50

Sample size for group 2, $n_2 =$ 50

Calculate power

POWER

- We have four chances for one of the tests involving μ_1 to be significant, so the probably of at least one being significant is

$$1 - (1 - 0.360)^4 = 0.832$$

- On the other hand, the probably that each of those four experiments will reject H_0 is

$$(0.360)^4 = 0.0168$$

- So, you are almost surely going to draw *some* wrong conclusions

POWER

- If you want to have a 0.9 probability that all four tests involving μ_1 reject H_0, each test needs a power of

$$(0.9)^{1/4} = 0.974$$

- We can identify the required sample size for *each* condition

Specify the population characteristics:

$H_0 : \mu_1 - \mu_2 =$ 0

$H_a : \mu_{a1} - \mu_{a2} =$ 0.5

$\sigma_1 =$ 1

$\sigma_2 =$ 1

Or enter a standardized effect size

$\frac{(\mu_{a1}-\mu_{a2})-(\mu_1-\mu_2)}{\sigma} = \delta =$ 0.5

Specify the properties of the test:

Type of test: Two-tails

Type I error rate, $\alpha =$ 0.005

Power= 0.974 [Calculate minimum sample size]

Sample size for group 1, $n_1 =$ 183

Sample size for group 2, $n_2 =$ 183 [Calculate power]

- this is an approximation because the tests are not independent

POWER

- Trying to control the error probabilities becomes complicated when you have multiple comparisons
- The probability of making at least one Type I error increases (power for detecting *something* increases)
- The probability of making at least one Type II error increases (power for the full set of differences decreases)
- This will always be true, but there are steps we can take to partially deal with the problem

DEMONSTRATION

- Open your five packages and count the number of *green* M&M's in each package. You have five numbers, $n = 5$, that make a sample
- within each sample, compute:

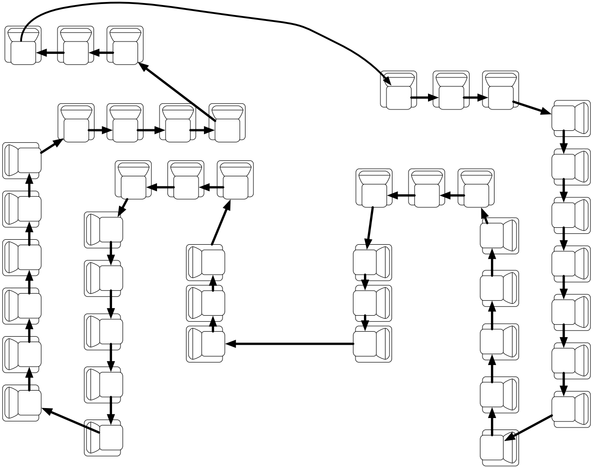
$$\overline{X}_k = \frac{\sum_i X_{ki}}{5}$$

$$s_k = \sqrt{\frac{\sum_i X_{ki}^2 - [(\sum_i X_{ki})^2/5]}{4}}$$

- Use the on-line calculator for *Descriptive Statistics*, if you want (use your phone). No need to log in.

DEMONSTRATION

- now, share your $\overline{X}_j$ and s_j with your "neighbor" by following the arrows below
- get $\overline{X}_j$ and s_j from your "neighbor"



DEMONSTRATION

- Run a hypothesis test to compare your mean to the mean of your neighbor
  - we'll assume homogeneity of variance
- (1) State the hypothesis:

$$H_0 : \mu_k = \mu_j$$

$$H_a : \mu_k \neq \mu_j$$

- and set the criterion
  $\alpha = 0.05$

DEMONSTRATION

- (2) Compute test statistic:

$$s^2 = \frac{(n_k - 1)s_k^2 + (n_j - 1)s_j^2}{n_k + n_j - 2} = \frac{(4)s_k^2 + (4)s_j^2}{8} =$$

$$s_{\overline{X}_k - \overline{X}_j} = \sqrt{s^2 \left(\frac{1}{n_k} + \frac{1}{n_j} \right)} = \sqrt{s^2 \left(\frac{1}{5} + \frac{1}{5} \right)} =$$

$$t = \frac{(\overline{X}_k - \overline{X}_j) - (0)}{s_{\overline{X}_k - \overline{X}_j}} =$$

$$df = n_k + n_j - 2 = 8$$

- (3) Compute the p-value using the t-distribution calculator
  - Instead we will identify the t_{cv} that corresponds to $p = 0.05$. It is $t_{cv} = 2.306$
- (4) Make a decision:

$$t = \quad <?> \quad = \pm 2.306$$

DEMONSTRATION

- We know from the outset that H_0 is actually true here.

$$\mu_k = \mu_j$$

  - because all the samples are actually from the very same population (M&M packages from the same factory have a fixed ratio of colors)
- Still, just due to sampling errors, we expect to have some tests reject H_0. The probability of at least one is around:

$$1 - (1 - \alpha)^c =$$

- where c is the number tests (number of students in the class)

WHAT DO WE MAKE OF THIS?

- Not only is it a pain to compute multiple comparisons of means
- but it tends to lead to more Type I error than α indicates
- we could decrease α to a smaller value so that the overall Type I error is how we want it
- which will decrease power

CONCLUSIONS

- testing multiple means
- loss of control of Type I error

NEXT TIME

- there is a better method
- ANOVA
- two measures of variance

Measure twice, cut once.