PSY 201: Statistics in Psychology

Lecture 34
Power for Analysis of Variance
Keep it simple!

Greg Francis

Purdue University

Fall 2019

HYPOTHESES

- The null for an ANOVA is an *omnibus* hypothesis. It suppose no difference between any population means

$$H_0 : \mu_i = \mu_j \; \forall \; i, j$$

- the alternative is the complement

$$H_0 : \mu_i \neq \mu_j \;\; \text{for some } i, j$$

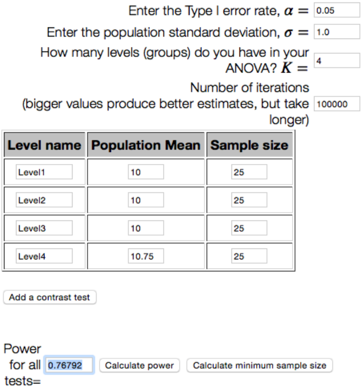
- To compute power, we have to provide the standard deviation, α, n's, and specific values for the means

POWER CALCULATOR

- For other power calculators, it was kind of easy to identify how power is affected by the specific alternative:
- bigger differences (between population means, proportions, or correlations) leads to more power
- That is also true for ANOVA, but it can be more complicated because there are multiple means

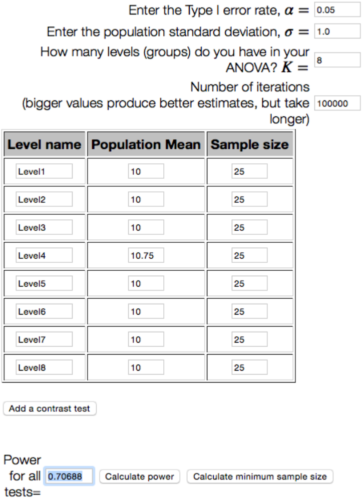
POWER CALCULATOR

- Consider a situation with $K = 4$ means (one different from the others):
- We estimate the power to be 0.76792



POWER CALCULATOR

- Consider a situation with $K = 8$ means (one different from the others):
- We estimate the power to be 0.70688.
- Power is affected by the ratio of the variability between group means and the variability within each group ($\sigma = 1$). If just one mean is different from the others, this ratio decreases as K gets bigger


Enter the Type I error rate, $\alpha =$ 0.05

Enter the population standard deviation, $\sigma =$ 1.0

How many levels (groups) do you have in your ANOVA? $K =$ 8

Number of iterations (bigger values produce better estimates, but take longer) 100000

Level name	Population Mean	Sample size
Level1	10	25
Level2	10	25
Level3	10	25
Level4	10.75	25
Level5	10	25
Level6	10	25
Level7	10	25
Level8	10	25

Add a contrast test

Power for all tests= 0.70688 Calculate power Calculate minimum sample size

POWER CALCULATOR

- Consider a situation with $K = 8$ means (four different from the others):
- We estimate the power to be 0.98324

Enter the Type I error rate, $\alpha =$ 0.05

Enter the population standard deviation, $\sigma =$ 1.0

How many levels (groups) do you have in your ANOVA? $K =$ 8

Number of iterations (bigger values produce better estimates, but take longer) 100000

Level name	Population Mean	Sample size
Level1	10	25
Level2	10	25
Level3	10	25
Level4	10	25
Level5	10.75	25
Level6	10.75	25
Level7	10.75	25
Level8	10.75	25

Add a contrast test

Power for all tests= 0.98324 Calculate power Calculate minimum sample size

POWER CALCULATOR

- Consider a situation with $K = 4$ means (two different from the others):
- We estimate the power to be 0.88392
- Thus, it is *not* just that power decreases as K increases. It depends on the values of the means

Enter the Type I error rate, $\alpha =$ 0.05

Enter the population standard deviation, $\sigma =$ 1.0

How many levels (groups) do you have in your ANOVA? $K =$ 4

Number of iterations (bigger values produce better estimates, but take longer) 100000

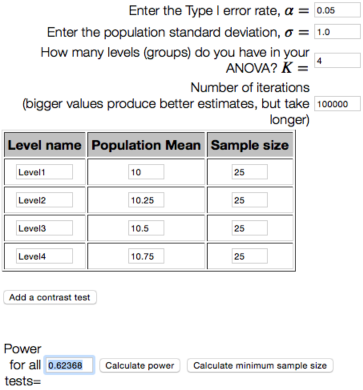
Level name	Population Mean	Sample size
Level1	10	25
Level2	10	25
Level3	10.75	25
Level4	10.75	25

Add a contrast test

Power for all tests= 0.88392 Calculate power Calculate minimum sample size

POWER CALCULATOR

- Consider a situation with $K = 4$ means (every mean is different from the others):
- We estimate the power to be 0.62368
- The biggest and smallest means differ by 0.75, just like previous cases, but that alone does not determine power



TRUST THE MATH

- With sufficient experience, you can learn to recognize what types of situations produce large (or small) power
- Until you get that experience, rely on the calculator (even after you get the experience you need the calculator to do the actual computations)
- It is still the case that larger samples lead to higher power.

EXAMPLE

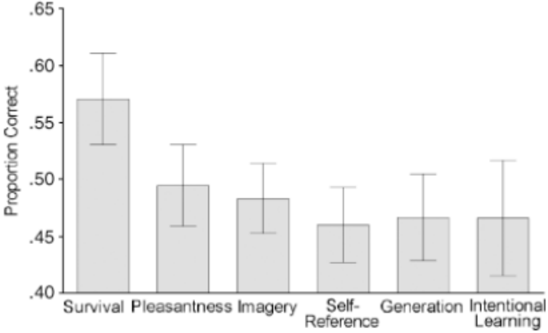
- There are lots of mnemonic tricks to try to improve your memory. They really do work!
- To compare these tricks we can use a standard memory test (Nairne, Pandeirada & Thompson, 2008):
- A subject is shown a word and asked to do some kind of task. This is repeated for 30 words.
- At the end of the experiment, the subject is asked to recall as many words as possible. Usually, this is a surprise memory task.
- For each subject, we compute the proportion of recalled words.
- We are interested in the mean value of the proportion across subjects.
- We can compare how well different tasks influence memory.

TASKS

- *Pleasantness*: Rate the pleasantness of the word on a scale from 1 to 5.
- *Imagery*: Rate how easy it is to form a mental image of the word on a scale from 1 to 5.
- *Self-reference*: Rate how easily the word brings to mind an important personal experience on a scale from 1 to 5.
- *Generation*: Words are partially scrambled; unscramble and then rate the pleasantness of the word on a scale from 1 to 5. (e.g., "iktten")
- *Survival*: Rate the relevance of the word for survival if you are stranded in the grasslands of a foreign land, on a scale from 1 to 5.
- *Intentional learning*: Try to remember the words for a future memory test.
- Different subjects are assigned to different conditions

ORIGINAL RESULTS

- Nairne, Pandeirada & Thompson (2008) found a big advantage for survival processing compared to the other methods. $n_i = 50$ for each group



- $F_{5,294} = 4.41$, $p = 0.00178$, $MS_W = 0.019$

NEW METHOD

- Suppose that you want to further explore these kinds of memory tricks. You think that the survival processing method does well because it gets subjects to be really be engaged in thinking about the word. You come up with a new method
- *Vacation*: Rate the relevance of the word for enjoyment while on vacation at a fancy resort, on a scale from 1 to 5.
- You expect that the vacation task will do about the same as the survival task
- You worry that other details of the experiment may change the overall level of performance for all tasks, so you decide to repeat the full study, with the addition of your new, Vacation, task. So there will be seven groups.
- How do you plan an appropriate sample size?

SPECIFIC MEANS

- As the values for the specific means, we can use the sample means found in the original study
- We get them from the figure
- For the Vacation task, we expect performance to be the same as the Survival task
- For the standard deviation, we can use the square root of MS_W

$$\sigma = \sqrt{MS_W} = \sqrt{0.019} = 0.1378$$

POWER FOR ANOVA

- Power is quite high (0.999) if we use $n_i = 50$, as in the original study

Enter the Type I error rate, $\alpha =$ 0.05

Enter the population standard deviation, $\sigma =$ 0.1378

How many levels (groups) do you have in your ANOVA? $K =$ 7

Number of iterations (bigger values produce better estimates, but take longer) 5000

Level name	Population Mean	Sample size
Survival	.57	50
Pleasantnes:	0.49	50
Imagery	.48	50
SelfReferenc	.46	50
Generation	.47	50
IntentionalLe	.47	50
Vacation	.57	50

Add a contrast test

Power for all tests= 0.999 Calculate power Calculate minimum sample size

POWER FOR ANOVA

- If we accept power of 0.9, $n = 25$ subjects in each sample is sufficient

Enter the Type I error rate, $\alpha =$ 0.05

Enter the population standard deviation, $\sigma =$ 0.1378

How many levels (groups) do you have in your ANOVA? $K =$ 7

Number of iterations (bigger values produce better estimates, but take longer) 5000

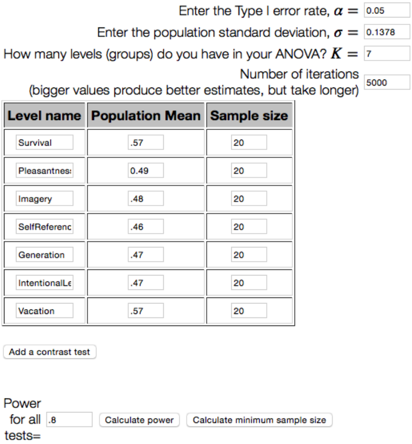
Level name	Population Mean	Sample size
Survival	.57	25
Pleasantnes:	0.49	25
Imagery	.48	25
SelfReferenc	.46	25
Generation	.47	25
IntentionalLe	.47	25
Vacation	.57	25

Add a contrast test

Power for all tests= .9 Calculate power Calculate minimum sample size

POWER FOR ANOVA

- If we accept power of 0.8, $n = 20$ subjects in each sample is sufficient



Enter the Type I error rate, $\alpha =$ 0.05

Enter the population standard deviation, $\sigma =$ 0.1378

How many levels (groups) do you have in your ANOVA? $K =$ 7

Number of iterations (bigger values produce better estimates, but take longer) 5000

Level name	Population Mean	Sample size
Survival	.57	20
Pleasantnes:	0.49	20
Imagery	.48	20
SelfReferenc	.46	20
Generation	.47	20
IntentionalLε	.47	20
Vacation	.57	20

Add a contrast test

Power for all tests= .8 Calculate power Calculate minimum sample size

CONTRASTS

- However, just a significant ANOVA is not enough for what we are studying
- We want to show that the Vacation task is better than most of the other tasks (not including the Survival task)
- We also want to show that the Survival task is better than most of the other tasks (not including the Vacation task)
- For our hypothesis test, we will set up two contrasts to test Vacation and Survival against the other tasks:
- We need to include those contrasts in the power analysis (more subjects)

CONTRASTS

Enter the Type I error rate, $\alpha =$ 0.05

Enter the population standard deviation, $\sigma =$ 0.1378

How many levels (groups) do you have in your ANOVA? $K =$ 7

Number of iterations
(bigger values produce better estimates, but take longer) 5000

Level name	Population Mean	Sample size
Survival	.57	27
Pleasantnes:	0.49	27
Imagery	.48	27
SelfReferenc	.46	27
Generation	.47	27
IntentionalLe	.47	27
Vacation	.57	27

Add a contrast test

Specify hypotheses for Contrast1

H0: 0 $\mu_{Survival}$ + 1 $\mu_{Pleasantness}$ + 1 $\mu_{Imagery}$ + 1 $\mu_{SelfReference}$ + 1 $\mu_{Generation}$ + 1 $\mu_{IntentionalLearning}$ + -5 $\mu_{Vacation} = 0$

Ha: Two-tails

α 0.05

Specify hypotheses for Contrast2

H0: -5 $\mu_{Survival}$ + 1 $\mu_{Pleasantness}$ + 1 $\mu_{Imagery}$ + 1 $\mu_{SelfReference}$ + 1 $\mu_{Generation}$ + 1 $\mu_{IntentionalLearning}$ + 0 $\mu_{Vacation} = 0$

Ha: Two-tails

α 0.05

Power for all tests= .8 Calculate power Calculate minimum sample size

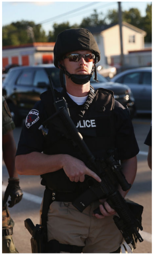
NULL

- You might also want to demonstrate that memory performance is the same for the Survival and Vacation tasks (after all, your idea is that both tasks are engaging, so they should have similar performance)
- Unfortunately, hypothesis testing cannot show that two groups have equal means (that would be *proving* the null hypothesis)
- Thus, we cannot set a sample size so that we are sure the Survival and Vacation tasks are equally effective for improving memory

ANOTHER EXAMPLE

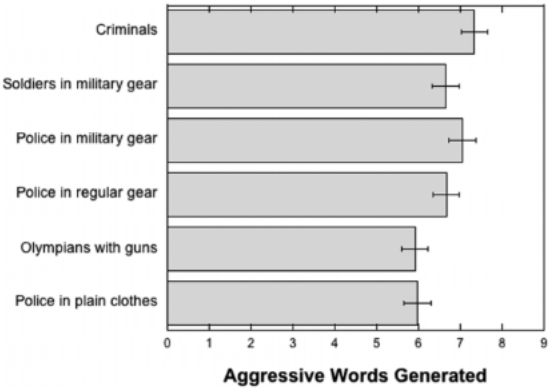
- Bushman (2018) investigated the "weapons effect": the mere presence of weapons can increase aggression
- Subjects were assigned to view a set of images of one type:
- Criminals, Soldiers, Police in military gear, Police in regular gear, Olympians with guns, Police in plain clothes
- Afterwards, complete a word fragment task:
    - C H O __ E
    - K I __ __
    - M U __ __ E R
    - C __ T
- each fragment can be completed to form an aggressive or non-aggressive word
- Count how many aggressive words are formed: measure of aggressive thoughts

EXAMPLE IMAGES



DATA

- Roughly $n = 100$ for each image set



MANY TESTS

- Conclusions are based on many contrasts
  - Significant ANOVA (some difference across image types)
  - Contrast between people with guns vs. plainclothes police (no guns): Weapon is important
  - Contrast between Olympians vs. Others: Person must intend to hurt others
  - Contrast between people with guns vs. Olympians: Weapon must be to hurt people
- Conclusion: only guns intended to shoot human targets prime aggressive thoughts

REPLICATION STUDY

- Suppose you want to replicate this study. To estimate power you use the means and standard deviation of the original finding. You want to see what happens if you use a similar sample size as the original study, $n = 100$, for each sample
- We enter the information in the ANOVA Power calculator

Enter the Type I error rate, $\alpha =$ 0.05

Enter the population standard deviation, $\sigma =$ 3.4

How many levels (groups) do you have in your ANOVA? $K =$ 6

Number of iterations (bigger values produce better estimates, but take longer) 5000

Level name	Population Mean	Sample size
Criminal	7.1	100
Soldiers	6.65	100
PoliceMilitar	6.8	100
PoliceReguli	6.7	100
Olympians	5.9	100
PolicePlainC	5.95	100

REPLICATION STUDY

- We set up each of the contrast tests in the ANOVA Power calculator:

Specify hypotheses for Contrast1

H_0: 1 $\mu_{Criminal}$ + 1 $\mu_{Soldiers}$ + 1 $\mu_{PoliceMilitaryGear}$ + 1 $\mu_{PoliceRegularGear}$ + 0 $\mu_{Olympians}$ + -4 $\mu_{PolicePlainClothes}$ = 0

H_a: Two-tails

α 0.05

Specify hypotheses for Contrast2

H_0: 1 $\mu_{Criminal}$ + 1 $\mu_{Soldiers}$ + 1 $\mu_{PoliceMilitaryGear}$ + 1 $\mu_{PoliceRegularGear}$ + -5 $\mu_{Olympians}$ + 1 $\mu_{PolicePlainClothes}$ = 0

H_a: Two-tails

α 0.05

Specify hypotheses for Contrast3

H_0: 1 $\mu_{Criminal}$ + 1 $\mu_{Soldiers}$ + 1 $\mu_{PoliceMilitaryGear}$ + 1 $\mu_{PoliceRegularGear}$ + -4 $\mu_{Olympians}$ + 0 $\mu_{PolicePlainClothes}$ = 0

H_a: Two-tails

α 0.05

REPLICATION STUDY

- When we hit the "Calculate power" button, we get:

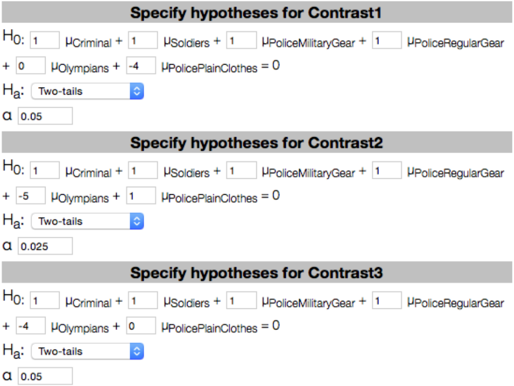
Power for all tests= 0.3936 | Calculate power | Calculate minimum sample size

Test	Estimated Power
ANOVA	0.6828
Contrast1	0.625
Contrast2	0.6456
Contrast3	0.666

- Each test has around a 65% chance of rejecting its H_0, but the probability of **all** tests rejecting the H_0 for one set of samples is only around 40%.

ADJUSTING α

- Bushman (2018) was concerned about multiple tests increasing Type I error, so he set $\alpha = 0.025$ for the second contrast



Specify hypotheses for Contrast1

H0: 1 μCriminal + 1 μSoldiers + 1 μPoliceMilitaryGear + 1 μPoliceRegularGear + 0 μOlympians + -4 μPolicePlainClothes = 0

Ha: Two-tails

α 0.05

Specify hypotheses for Contrast2

H0: 1 μCriminal + 1 μSoldiers + 1 μPoliceMilitaryGear + 1 μPoliceRegularGear + -5 μOlympians + 1 μPolicePlainClothes = 0

Ha: Two-tails

α 0.025

Specify hypotheses for Contrast3

H0: 1 μCriminal + 1 μSoldiers + 1 μPoliceMilitaryGear + 1 μPoliceRegularGear + -4 μOlympians + 0 μPolicePlainClothes = 0

Ha: Two-tails

α 0.05

ADJUSTING α

- When we hit the "Calculate power" button, we get:

Power for all tests= 0.3304 Calculate power Calculate minimum sample size

Test	Estimated Power
ANOVA	0.6776
Contrast1	0.6196
Contrast2	0.5598
Contrast3	0.6724

- The power of the second contrast drops a bit. The other power estimates change, but that is just a side effect of the calculations. We could increase the number of iterations to avoid these changes.
- The power for all tests drops from 40% to around 33%

SAMPLE SIZE

- How big a sample size do we need to have 80% power?
- $n = 223$, which means a total of $6 \times 223 = 1338$ subjects
- The power values would be distributed across the tests as:

Power for all tests= 0.8072 Calculate power Calculate minimum sample size

Test	Estimated Power
ANOVA	0.9712
Contrast1	0.9266
Contrast2	0.8752
Contrast3	0.95

SIMPLE IS BETTER

- If your conclusion depends on many hypothesis tests producing significant results, you should design your study to take into account all of those tests
- Adding tests always lowers power
- Complicated experiments require much larger samples than simple experiments
- Lots of studies that are published are woefully underpowered because they do not consider these details of experimental design

CONCLUSIONS

- power for ANOVA
- power for contrasts
- simple is better

NEXT TIME

- Dependent ANOVA
- Contrasts

Leverage relationships