36-309 / 36-749 Homework 6: Power

Due Wednesday, October 30, 11:59pm on Gradescope

Emily Sands

Question 1: Computing Error Rates across Many Experiments (28pts)

Professor T. Test performs 1000 treatment-vs-control experiments in her lifetime. She always uses a significance level of alpha = 0.05. *She does not know this*, but because I am omniscient I can tell you the following information:

1) In 160 of her studies the null hypothesis is true.
2) In 200 of her studies the null hypothesis is false with 20% power for the true effect size.
3) In 360 of her studies the null hypothesis is false with 55% power for the true effect size.
4) In 280 of her studies the null hypothesis is false with 80% power for the true effect size.

Given this information, answer the following questions.

PART A (16pts)

First we will consider "positive results" and "negative results." For this part, answer the following two questions.

- (8pts) Out of all 1000 studies, how many "positive results" (i.e., reject the null hypothesis results) do we expect, and how many "negative results" (i.e., fail to reject the null hypothesis results) do we expect? Please show your work for how you arrived at your answer.

Positive Results: $8 + 462 = 470$ **total**

- **False Positive:** $0.05 * 160 = 8$
- **True Positive:** $0.2 * 200 + 0.55 * 360 + 0.8 * 280 = 462$

Negative Results: $378 + 152 = 530$ **total**

- **False Negative:** $(1 - 0.2) * 200 + (1 - 0.55) * 360 + (1 - 0.8) * 280 = 378$
- **True Negative:** $0.95 * 160 = 152$
- (8pts) What *percentage* of the *positive results* are expected to be correct? Furthermore, what *percentage* of the *negative results* are expected to be correct? Please show your work for how you arrived at your answer. (**Hint**: This question is asking about percentages *among positive results specifically* and *among negative results specifically*. For example, if we expected 100 positive results, and we expected 20 of those positive results to be correct, then the answer would be 20%.)

% Positive Results Correct: $462/470 = 0.9829787 \implies 98.30$

% Negative Results Correct: $152/530 = 0.2867925 \implies 28.68$

PART B (12pts)

Now we'll consider making "errors" when hypothesis testing. For this part, answer the following two questions.

- (6pts) Above we communicated four kinds of studies, labeled (1) through (4). For which of these studies is it possible to make a *Type 1 error*, and for which of these studies is it possible to make a *Type 2 error*? Please explain in 1-3 sentences, and please use the (1) through (4) labels to refer to these studies in your answer.

Possible Type 1 Error: Study (1)

- **Type 1 Error is a false positive (i.e. reject the null hypothesis when it's actually true), which means only study 1 fits as it's the only study with a true null hypothesis.**

Possible Type 2 Error: Study (2), (3), (4)

- **Type 2 Error is a false negative (i.e. fail to reject the null hypothesis when it's actually false), which means only studies 2, 3, and 4 fits as they are the only ones with a false null hypothesis.**
- (6pts) Out of all 1000 studies, what is the expected total number of **incorrect** hypothesis testing conclusions? Please show your work for how you arrived at your answer.

Incorrect Hypothesis Testing Conclusions: $378 + 8 = 386$

- **Incorrect hypothesis testing conclusions are when it is "False" from above (i.e. reject the null when it's true, or fail to reject the null when it's false).**

Question 2: Type 1 and 2 errors and their relationship with alpha (12pts)

We often use alpha = 0.05 when doing statistical hypothesis testing; alpha = 0.05 corresponds to rejecting the null hypothesis when the p-value is below 0.05. **For this problem, consider raising to alpha = 0.1 instead of 0.05.** Given this, answer the following questions.

- (4pts) When we raise alpha to 0.1 instead of 0.05, does the **Type 1 error rate** increase, decrease, or stay the same? Discuss your reasoning in 1-2 sentences.

When we raise alpha to 0.1 instead of 0.05, the Type 1 error rate increases, because alpha represents the probability of rejecting the null hypothesis when it is true (i.e. a false positive) which is exactly what type 1 error is.

- (4pts) When we raise alpha to 0.1 instead of 0.05, does the **Type 2 error rate** increase, decrease, or stay the same? Discuss your reasoning in 1-2 sentences.

When we raise alpha to 0.1 instead of 0.05, the Type 2 error rate decreases, because raising alpha makes it more likely we'll reject the null hypothesis, but type 2 error corresponds to when we fail to reject the null hypothesis when it's actually false.

- (4pts) When we raise alpha to 0.1 instead of 0.05, does the **power** increase, decrease, or stay the same? Discuss your reasoning in 1-2 sentences.

When we raise alpha to 0.1 instead of 0.05, power increases because power = 1 - (Type 2 error rate), so if raising alpha decreases type 2 error rate, then increasing alpha will also increase power by the equation above.

Question 3: Power and Sample Size Calculations (60pts)

Let's say we're working at a tech company, and the company is considering updating their app to a new design. Their software engineering and human-computer interaction teams have developed two possible new designs (which we'll call Design A and Design B). The company has decided it would be beneficial to run an experiment, where they enroll a set of current app users, and then randomize them to one of three treatment groups: an app with Design A, an app with Design B, or an app with the Old Design. After this happens, the company will measure how much time each user spends on the app, which will be the outcome for the experiment.

There are three treatment groups and a quantitative outcome. Thus, the most appropriate statistical analysis will be one-way ANOVA. However, the company would like to get some more information before implementing the experiment. The company has assigned us to run a power analysis to better understand how the experiment should be run. We'll explore various aspects of power analyses for this example.

PART A (11pts)

For simplicity, the company is first considering enrolling 65 subjects in each of the three treatment groups. Given this information, answer the following three questions.

- (3pts) For this hypothetical experiment, what is the between-groups degrees of freedom and within-groups degrees of freedom? Explain in one sentence.

Between-groups degrees of freedom: num groups $- 1 = 3 - 1 = 2$

Within-groups degrees of freedom: total num subjects $-$ num groups $= 3 * 65 - 3 = 192$

- (4pts) Let's say we ran this hypothetical experiment and observed an F statistic equal to 1. What would be the p-value in that case? Explain how you arrived at your answer, and provide any code you used to arrive at your answer.

```
1 - pf(1, df1 = 2, df2 = 192)
```

```
## [1] 0.3697872
```

Answer: 0.37

The p-value is the proportion of the null distribution that's greater than the observed F-statistic (in this case 1). The above code calculates this metric, given between group df of 2 and within group df of 192.

- (4pts) Let's say that we would reject the null hypothesis only if the p-value from the one-way ANOVA was less than 0.05. What would be the critical F value in this case? Explain how you arrived at your answer, and provide any code necessary to arrive at your answer.

```r
f.crit = qf(1 - 0.05, df1 = 2, df2 = 192)
f.crit
```

```
## [1] 3.042964
```

Answer: 3.04

The critical value is $1 - \alpha = 1 - 0.05$ and the critical value is a quantile, so we need to compute the quantile of a null distribution

PART B (10pts)

In the remaining parts of this homework, we'll conduct some power analyses for some hypothetical experiments. But first, we'll take a step back and consider power in theory. In class, we discussed how to approximate the expected value of the F statistic under the alternative hypothesis. What are three ways that we can *increase* power, according to that approximation? In your answer, **do not** explicitly mention the terms σ^2, σ_A^2, and n from lecture; instead, please discuss these terms qualitatively within the context of our tech experiment.

1. **Increase Sample Size: By having more participants in each of the three groups, we can make more precise inference (e.g. narrower confidence intervals) which means we'll be more likely to reject the null hypothesis if it's false, boosting power.**
2. **Increase between group variation: Increasing differences in the three group means (i.e. the time each group spends on the app) increases power because bigger differences between each groups' means are easier to detect to a statistically significant degree.**
3. **Decrease within group variability: Smaller variance in time spent within each group means inference is more precise (e.g. smaller confidence intervals), so we'll be more likely to reject the null hypothesis if it's actually false.**

PART C (15pts)

Now we'll compute statistical power for one possible proposed experiment. For this part, let's say the company is still considering enrolling 65 subjects in each of the three treatment groups. Answer the following questions.

- (4pts) This actually isn't the first experiment this company has run. Looking back at some reports, the company provides the following one-way ANOVA table that was produced after a similar experiment (see the PDF for easier readability):

	Df	Sum Sq	Mean Sq	F value	p-value
design	3	5640	1880	1.88	0.137
Residuals	116	117160	1010		

Note that it isn't necessary that this past experiment had the same number of treatment groups or subjects as the new proposed experiment. Using the above one-way ANOVA table, what was the estimate of sigma-squared based on this old experiment? Explain in one sentence, and define this estimate as `sigma2` (template code given below; be sure to uncomment your code such that `sigma2` is defined).

```
sigma2 = 1010
```

σ^2 is represented as Mean Square of the residuals in the anova table

- (5pts) Recently, users have spent about 55 minutes on the app, on average. Meanwhile, the company would be very happy if either of the new designs (Design A or Design B) changed users' time on the app by 15 minutes (i.e., from 55 to 70). It also does not necessarily anticipate that either Design A or Design B will be better. Thus, the company proposes the following treatment group means to consider:

```
groupmeans.better = c(55, 70, 70)
groupmeans.worse = c(55, 40, 40)
```

The first vector, `groupmeans.better`, is the case where Design A and Design B perform better than the Old Design, and `groupmeans.worse` is the case where they perform worse.

Using your estimate of sigma-squared `sigma2` from the previous part, what is the *effect size* (which we've defined as "eta" in class and lab) for these two different treatment group means? Remember that the company has proposed to enroll 65 subjects for each of the three treatment groups. Template code is provided below, which you must fill in. Your code should successfully print out the two effect sizes, `eta.better` and `eta.worse`. (**Hint**: If you did everything correctly, you should find that `eta.better` and `eta.worse` are equal. It's worth thinking about why they're equal.)

```
eta.better = sqrt( sum((65/195)*(groupmeans.better - mean(groupmeans.better))^2) / sigma2 )
eta.better
```

```
## [1] 0.2224971
```

```
eta.worse = sqrt( sum((65/195)*(groupmeans.worse - mean(groupmeans.worse))^2) / sigma2 )
eta.worse
```

```
## [1] 0.2224971
```

- (6pts) Now, compute the *statistical power* we would achieve if we ran an experiment where we had 65 subjects in each of the three treatment groups, the sigma-squared was equal to your `sigma2` above, and the group means were equal to `groupmeans.better` above, and we used alpha = 0.05 when testing the null hypothesis. After computing the statistical power, write your interpretation of that power value within the context of this experiment in 1-2 sentences. (**Hint**: Use `power.anova.test()`.)

```
power.anova.test(groups = 3, n = 65, between.var = var(groupmeans.better), within.var = sigma2)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##           groups = 3
##                n = 65
##     between.var = 75
##      within.var = 1010
##       sig.level = 0.05
##           power = 0.7942126
##
## NOTE: n is number in each group
```

The above code shows that the power is approximately 79.42% which means that if the three group means are (55, 70, 70), $\sigma^2 = 1010$, and there are 65 subjects per group, then we have a 79.42% chance of correctly rejecting the null hypothesis.

PART D (10pts)

In Part C, you should have computed the statistical power for one particular setting. For this part, we'll still assume that sigma-squared is equal to your `sigma2` in Part C, the group means are equal to `groupmeans.better` in Part C, and we use alpha = 0.05. However, now we'll consider what happens to power when we change the sample size.

(For the sake of this part, it's okay if your `sigma2` is incorrect—we'll grade according to the `sigma2` you computed.)

The company would like to consider the following sample sizes for each treatment group:

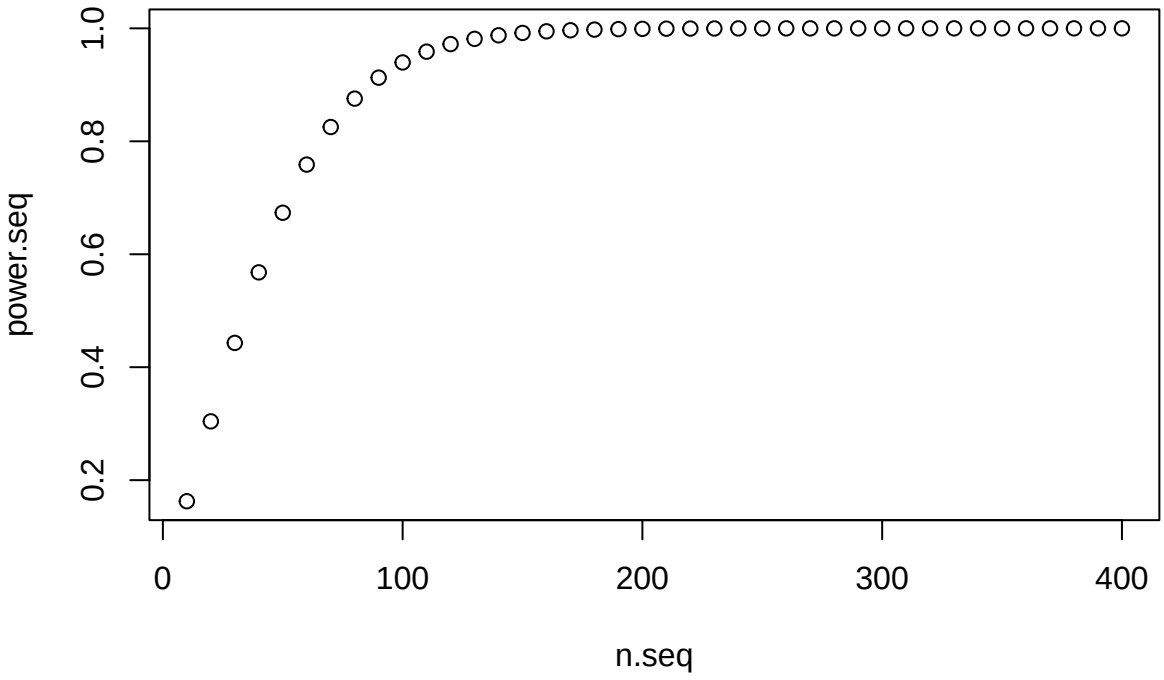
```
n.seq = seq(10, 400, by = 10)
```

In the above code, the variable `n.seq` is defined as a vector of numbers between 10 and 400 in increments of 10 (i.e., 10, 20, 30, etc.)

In this part, your task is to compute the statistical power for each sample size in `n.seq`. Then, make a *scatterplot*, with `n.seq` on the x-axis, and power on the y-axis. Finally, write a 1-2 sentence interpretation of the plot, where you discuss the relationship between sample size and power according to the plot. Template code is provided below, which you must fill in. (**Hint**: The "for loop" below simply runs `power.anova.test()` for each element in `n.seq`. It's okay if you aren't familiar with for loops, especially because we have not covered them in this class; all you need to do is specify `groups`, `between.var`, and `within.var` as you've done previously.)

```
#compute power for different sample sizes
power.seq = vector(length = length(n.seq))
#for each element in n.seq, compute power
for(j in 1:length(n.seq)){
 power.seq[j] = power.anova.test(
 groups = 3,
 n = n.seq[j],
 between.var = var(groupmeans.better),
 within.var = sigma2)$power
}

#Now make a scatterplot, with
# n.seq on the x-axis, and power.seq on the y-axis.
plot(n.seq, power.seq)
```



As sample size increases, the power also increases — specifically this relationship is similar to a log relationship in growth but with a limit of 1. Thus, as sample size increases we're more likely to reject the null hypothesis that all times on the apps are the null hypothesis being false is indeed correct.

PART E (14pts)

Now we'll consider computing the minimum sample sizes and minimum effect sizes to achieve 80% power. For this part, we'll still assume that sigma-squared is equal to your `sigma2` in Part C, the group means are equal to `groupmeans.better` in Part C, and we use alpha = 0.05. Answer the following two questions.

(For the sake of this part, it's okay if your `sigma2` is incorrect—we'll grade according to the `sigma2` you computed.)

- (6pts) Given `sigma2` and `groupmeans.better`, what is the *minimum sample size* we could assign to each treatment group and still obtain at least 80% power? Please report your answer as a whole number. Explain how you arrived at your answer, and provide any code you used to arrive at your answer.

```
power.anova.test(groups = 3, between.var = var(groupmeans.better), within.var = sigma2, power = 0.8)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##           groups = 3
```

```
##               n = 65.88247
##     between.var = 75
##      within.var = 1010
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: n is number in each group
```

I arrived at my answer by specifying the group size, between.var, within.var, and the power within a power.anova.test which will return the minimum sample size needed to meet these parameters, that sample size being 66 subjects (round up because you can't have fractional people).

- (8pts) After looking at your answer above, the company's CEO said, "L'anno scorso mi sono dato un grosso bonus, quindi non c'è modo che potremmo permetterci un esperimento così grande. Il numero massimo di soggetti che sarei disposto ad assegnare a ciascuno dei tre gruppi di trattamento è di 75 soggetti."[1] Again, luckily you haven taken 82-216 Intermediate Italian I and you know that this translates to: "I gave myself a big bonus last year, so there's no way that we could afford an experiment this big. The most subjects I'd be willing to assign to each of the three treatment groups is 75 subjects." Given this and `sigma2`, what is the *minimum variance in treatment group means* you could successfully detect with at least 80% power? After finding this minimum, answer the following: Is this bigger or smaller than the variance of `groupmeans.better`? Explain how you arrived at your answer, and provide any code you used to arrive at your answer.

```
power.anova.test(groups = 3, n = 75, within.var = sigma2, power = 0.8)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##          groups = 3
##               n = 75
##     between.var = 65.757
##      within.var = 1010
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: n is number in each group
```

```
var(groupmeans.better)
```

```
## [1] 75
```

The minimum variance in treatment group means you could successfully detect with 80% power with the above parameters is between.var = 65.757. This is smaller than the variance of groupmeans.better which is 75.

[1] Did I mention that we're still in Italy?

Question 4 (5pts; ONLY REQUIRED FOR 36-749 STUDENTS; BONUS QUESTION FOR 36-309 STUDENTS)

In Question 3D you should have created what's known as a "power curve," where you visualized how power changes with different sample sizes. In general, you can create power curves for different experimental design specifications. For this part, you'll create another power curve, this time varying sigma-squared.

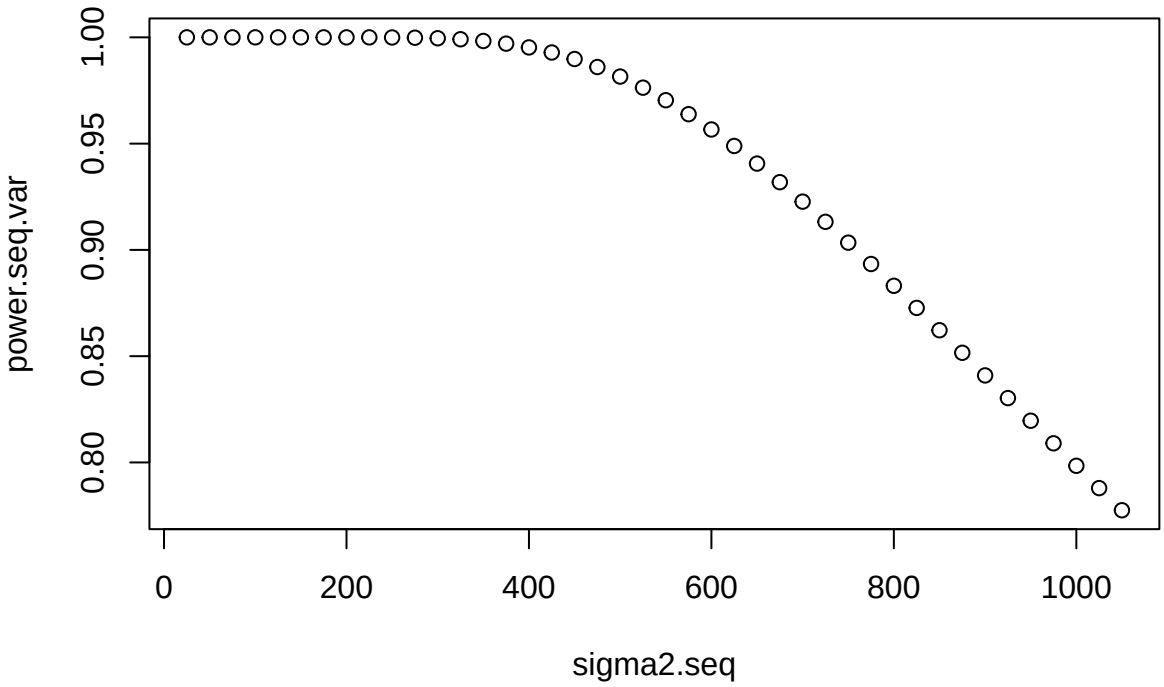
For this question, still consider the case where there are three treatment groups, with means equal to `groupmeans.better` used throughout this homework. Furthermore, assume that we'll assign 65 subjects to each of the three treatment groups.

Given this, define a vector of variances, `sigma2.seq`, equal to a vector of numbers between 25 and 1050 in increments of 25 (i.e., 25, 50, 75, etc.) Then, for each variance in `sigma2.seq`, compute the statistical power. Call the vector of power values `power.seq.var`. Finally, make a scatterplot, with `sigma2.seq` on the x-axis and `power.seq.var` on the y-axis.

After making the plot, write a 1-2 sentence interpretation of the plot, where you discuss the relationship between variance and power according to the plot.

```
sigma2.seq = seq(25, 1050, by = 25)
power.seq.var = vector(length = length(sigma2.seq))
for(j in 1:length(sigma2.seq)){
  power.seq.var[j] = power.anova.test(
    groups = 3,
    n = 65,
    between.var = var(groupmeans.better),
    within.var = sigma2.seq[j])$power
}

plot(sigma2.seq, power.seq.var)
```



As within group variance increases, the statistic power decreases (slowly at first and then more rapidly as within variance gets larger). This means, higher variance within treatment groups makes statistical inference less precise, meaning we'll be less likely to reject the null hypothesis if it's false.