Stats141XP Discussion 1: HW 1 (Team 2)

Team Members: Charles Liu, Joseph Gallegos, Anshu Mahalley, Jacob Samuels, Jordan Tallman

Questions:

1. Charles
2. Charles
3. Jordan
4. Jordan
5. Anshu
6. Anshu
7. Jacob
8. Jacob
9. Joseph
10. Joseph

Table of Contents:

Answers:

Question 1

a. Data science is important for understanding Covid-19 through methods such as development and testing of the Covid-19 vaccine, factors that'll slow the infection rate, understanding how airborne transmission works, reopening schools and when, finding socioeconomic and environmental factors, and tracking mobility to contain the virus. Covid-19 has been filled with lots of uncertainties making it difficult to decide when to reopen, but it was easily called when to close and commence shut down. Data scientists choose to acknowledge the uncertainties and embrace them as compared to leaders, who may choose to reject uncertainty. Data scientists use a range of numbers to determine their findings (such as “Confidence Intervals”), and a true statistician shouldn’t just trust a single number but a range of numbers. Sometimes, there are problems that not even data scientists can solve instantly because data scientists do not have an emergency response team. It can be difficult task for data scientists to have to collect the data and then analyze it. The collection process of data is difficult in itself because there are many details about the data that is left out, and they’ll receive only the results to analyze. Dr. Meng tried to emphasis the point that data science is meant to be targeted for everyone to know and learn about to some degree. Due to Covid-19 pandemic, data science is currently focusing on Covid-19 and racial discrimination data and analysis. It is also important for data scientists to remain ethical when analyzing the data.

b. One problem that was being dealt with during the early stages of the Covid-19 pandemic is that people knew the number of confirmed Covid-19 cases were lower than the number of exposed cases, but they were unsure how much lower. This problem was further exacerbated due to the fact that early stages of Covid-19 testing were slower. A Stanford study showed that 50-85-fold the amount of people were confirmed Covid-19 cases in Santa Clara County. This event had people questioning the study. The rate went from 1.5% to 2.5% confirmed rate. This rate actually came from a sample size rather than the actual population. Sadly, the adjustments made do not account for several factors such as age, ZIP code difficulty, and others. After a while, the Stanford study was shown to have incorrect Confidence Intervals, and it was not the best choice to use classical statistical methods for so many uncertainties. As a statistician, they are meant to focus on the data at hand and whatever assumptions are being made on the data. This is called a “statistical error” if a quantitative claim is not supported by the data and assumptions made. This is not a matter of ethics because statisticians have a chance of testing errors in their research. Once they find a mistake/error, data scientists should do their best to correct their errors, acknowledge their errors, and find out what they did wrong. It is considered an ethical violation if you do not acknowledge someone pointing out a possible error in your research, even if they are an outsider. Data science is difficult to do, and it can be hard to distinguish truth vs. evidence. Overall, as long as data scientists are able to acknowledge their errors and correct their mistakes, data scientists are always able to continue working and improving their research.

1. One study on Covid-19 would be to track which states have the most infections (or deaths by Covid-19) in the U.S., along with their political affiliation with their respective state. This might help statisticians understand if a certain political state and their policies on Covid-19 response might be a possible causation in increased infections.
2. One possible problem they may face is that some states refuse to give accurate data on their Covid-19 infection and death rates. Without an accurate data for statisticians to study, we will be going back to the article “Evidence vs. Truth”, where we question if the actual data is correct or the statisticians study. Another problem is to determine what states are politically affiliated with prior and during the pandemic.
3. One such ethical dilemma that statisticians may face is that the state might attempt to have the statistician alter the data/study to make it seem like their state’s infection/death rates are not that high. As stated in the “Evidence vs. Truth” article, statisticians could also be incorrect in their study of the data and the analysis that follows.
4. The first dilemma is quite simple. Statisticians should NOT alter the data/study in any way. If the state asks such a request, they’ll either refuse the request or continue to publish the study and analysis no matter how good or bad it is reflected on the state. As for the second dilemma, if a statistician’s work is deemed incorrect in some way by an outsider, they should take the time to reevaluate and ensure that their findings are accurate. They might ask other statisticians to do the study to ensure that multiple statisticians’ findings are accurate (or even relatively close) to the findings of the original. If they are incorrect, the statistician should immediately remedy their mistake by stating they made a mistake and should recall their findings.

Question 2

Coefficients	Model 1	Model 2	Model 3	Model 4
Intercept	45.148	24.10712	12.49668	11.58254

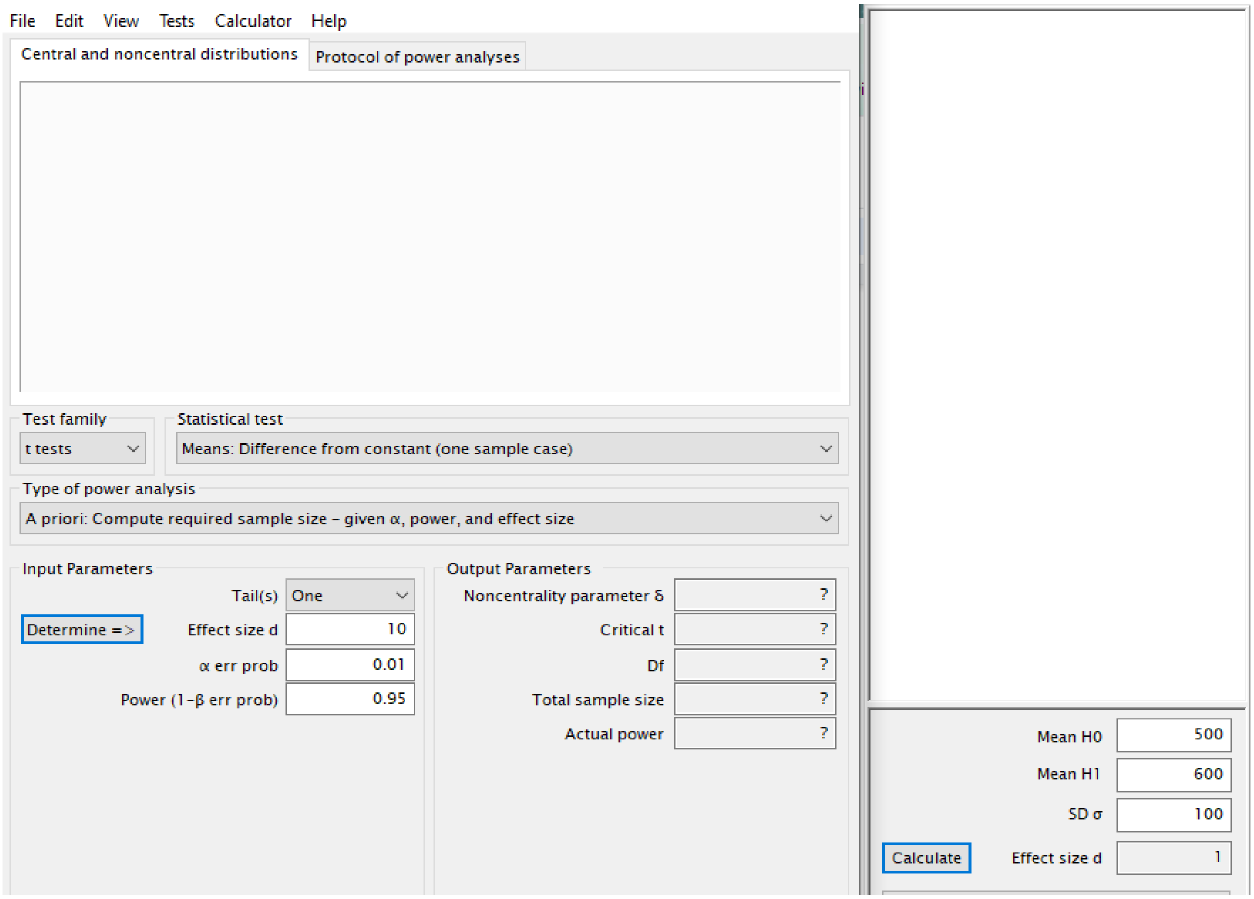
Algeb1	N/A	0.63726	0.22789	0.19576
Arith1	N/A	N/A	0.59954	0.54230
Geom1	N/A	N/A	N/A	0.14189
5 Years vs. 2 Years	-13.919	-10.08922	-4.44090	-4.39818
> 8 Years vs. 2 Years	9.307	4.40583	2.81335	2.62025
SST	340765	340764	340764	340765
SS due to Covariates (SSR)	300756	211440	159559	157662
SS due to Years More than Covariates (SSE)	40009	129324	181205	183103
R-Squared due to Covariates	0.1174	0.3795	0.5318	0.5373
R-Squared due to Years More than Covariates	0.1144	0.3763	0.5285	0.5333

a. “completed in the table above”

b. The covariates are continuous predictor variables that are complementary to the response variable. They are related in some way to the dependent variable. As for what it is, the covariate is a possible explanatory variable of the dependent variable. The covariates help with reducing error variance and reducing bias.
c. We can see that as we add more covariates into our model, the coefficient for “5 years vs. 2 years” increases and the coefficient for “> 8 years vs. 2 years” decreases. The intercept coefficient (“5 to 8 years”) decreases as more covariates are added. The reason the coefficients decrease is because as we add more covariates to the model, this will reduce the variance of the residuals.
d. As we increase the number of covariates we use, the statistical significance of “years more education” decreases. The p-value of “yearsmoreeducation” can be seen increasing, but it isn’t enough to change the fact that it is still a statistically significant variable to our model.
e. Mathematically, the SST is pretty much the same. Model 1 and Model 4 have an SST of 340765, and Model 2 and Model 3 have an SST of 340764. The SST is the dispersion of observed dependent variables around the mean. Conceptually, this tells us the total amount of variation in our model. We can see that there is slightly more variation for Model 1 and Model 4.
f. One assumption that is being made is that the adjustment of the within-groups Sum of Squares is that the within-groups regression coefficients are all estimates of the same common population regression coefficient. The size of the error variance is determined by our dispersion of the “Conditional Distributions”. The conditionals being “algeb1”, “algeb2”, and the different type of “yearsmoreeducation”. The higher they are correlated with each other, the narrower our scattered points become, and the higher our reduction in the error variance becomes due to ANCOVA.
g. The main difference between one-way ANOVA and one-way ANCOVA is that the Yij(adjusted) takes into account the variation of dependent scores that is not associated with the linear regression of Y on X. The adjustment made is made to remove the linear effect on the covariate(s). This will reduce the Sum of Squares for Y if the slope of the regression does not equal zero. The covariates and their differences affect the observed difference between Yij and Yij(adjusted). The terms from Yij can be rearranged to get the “adjusted score”. The thing about the “adjusted score” is that it is free from the effects of the covariates, and it also provides an estimate of the ANOVA terms of the model equation for a completely randomized design. One difference is that the ANCOVA error term will be usually smaller than the error term in ANOVA.

Question 3

a. Effect Size = 1


File
Edit
View
Tests
Calculator
Help
Central and noncentral distributions
Protocol of power analyses
Determine =>
Calculate

Test family: t tests

Statistical test: Means: Difference from constant (one sample case)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Tail(s): One

Effect size d: 10

α err prob: 0.01

Power (1−β err prob): 0.95

Output Parameters

Noncentrality parameter δ: ?

Critical t: ?

Df: ?

Total sample size: ?

Actual power: ?

Mean H0: 500

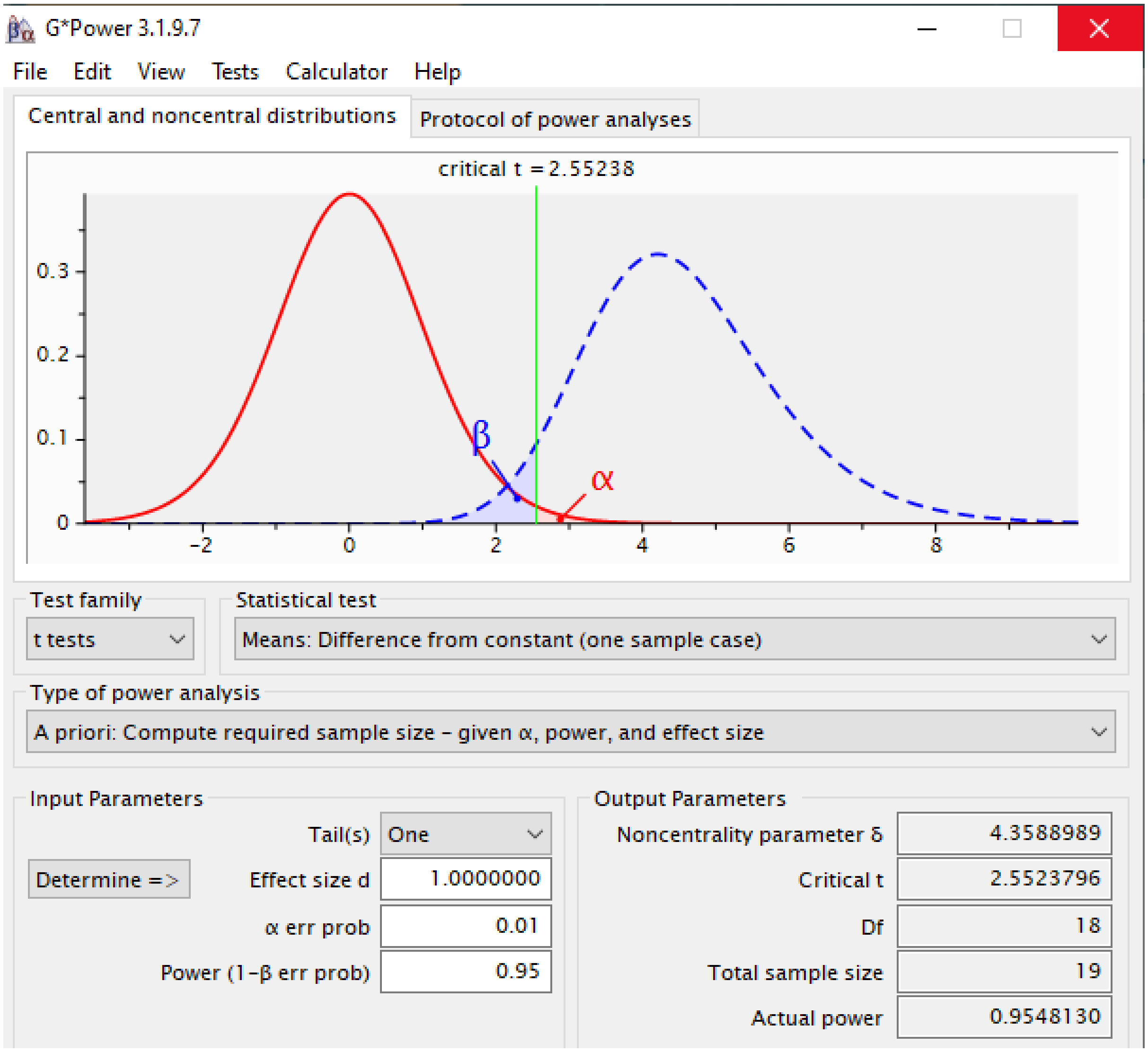
Mean H1: 600

SD σ: 100

Effect size d: 1

(t-test, difference from constant – one sample case)

b. Sample Size = 19



Test family: t tests

Statistical test: Means: Difference from constant (one sample case)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Tail(s): One

Effect size d: 1.0000000

α err prob: 0.01

Power (1-β err prob): 0.95

Output Parameters

Noncentrality parameter δ: 4.3588989

Critical t: 2.5523796

Df: 18

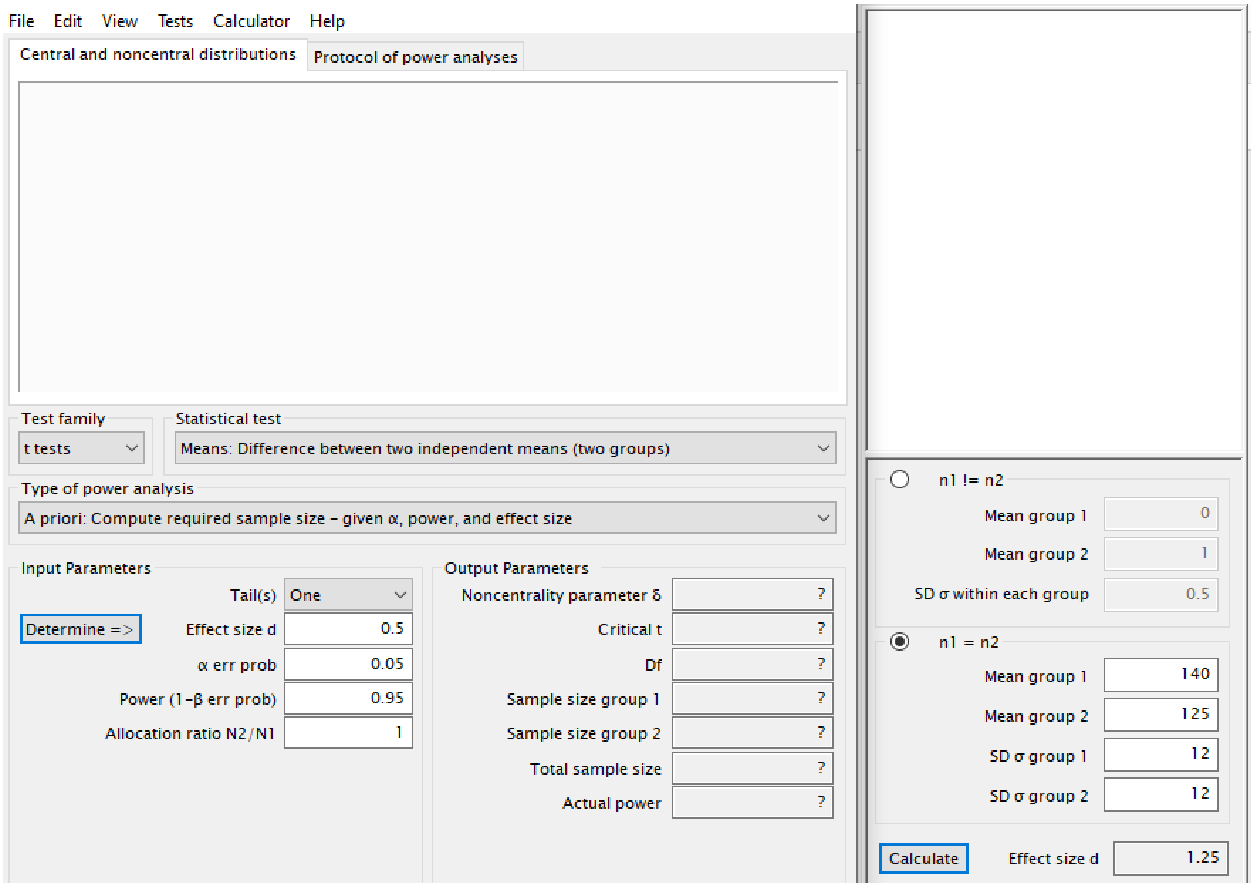
Total sample size: 19

Actual power: 0.9548130

(t-test, difference from constant – one sample case)

Question 4

a. Effect Size = 1.25


File
Edit
View
Tests
Calculator
Help
Central and noncentral distributions
Protocol of power analyses
Determine =>
Calculate

Test family: t tests

Statistical test: Means: Difference between two independent means (two groups)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Tail(s)	One
Effect size d	0.5
α err prob	0.05
Power (1−β err prob)	0.95
Allocation ratio N2/N1	1

Output Parameters

Noncentrality parameter δ	?
Critical t	?
Df	?
Sample size group 1	?
Sample size group 2	?
Total sample size	?
Actual power	?

○ n1 != n2

Mean group 1	0
Mean group 2	1
SD σ within each group	0.5

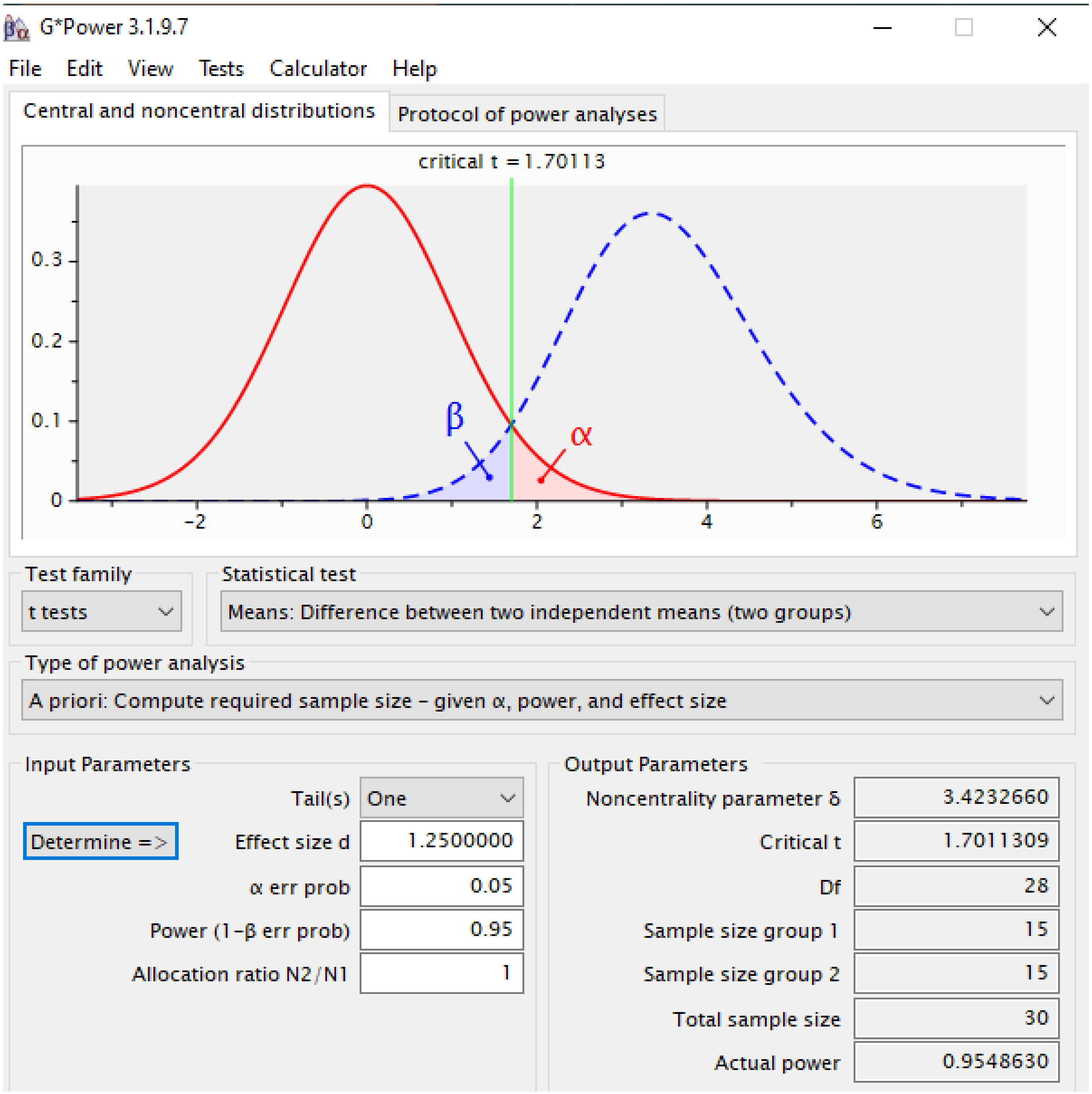
● n1 = n2

Mean group 1	140
Mean group 2	125
SD σ group 1	12
SD σ group 2	12

Effect size d 1.25

(t-test, difference between two independent means)

b. Sample Size = 30



Test family: t tests

Statistical test: Means: Difference between two independent means (two groups)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters		Output Parameters	
Tail(s)	One	Noncentrality parameter δ	3.4232660
Effect size d	1.2500000	Critical t	1.7011309
α err prob	0.05	Df	28
Power (1-β err prob)	0.95	Sample size group 1	15
Allocation ratio N2/N1	1	Sample size group 2	15
		Total sample size	30
		Actual power	0.9548630

(t-test, difference between two independent means)

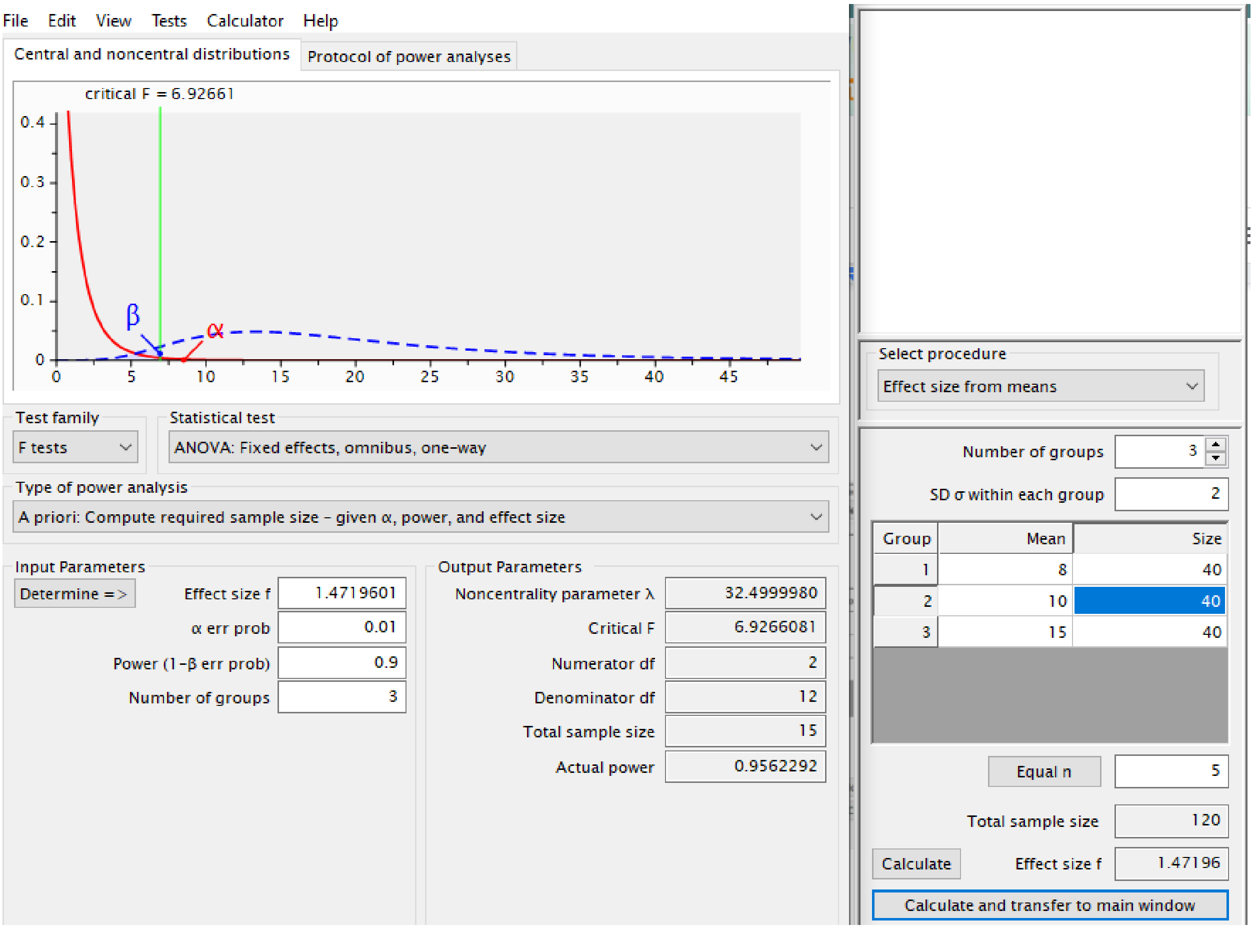
Question 5

1. Number of groups = 3
   Total sample size = 120

   (Used A Priori Analysis in GPower to calculate effect size for below table)

Table Two

Effect Size	Power and alpha	Estimated power
1.4719601	0.80, 0.05	0.8605124
1.4719601	0.80, 0.01	0.8312605
1.4719601	0.90, 0.05	0.9743308
1.4719601	0.90, 0.01	0.9562292



Test family: F tests

Statistical test: ANOVA: Fixed effects, omnibus, one-way

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Effect size f	1.4719601
α err prob	0.01
Power (1-β err prob)	0.9
Number of groups	3

Output Parameters

Noncentrality parameter λ	32.4999980
Critical F	6.9266081
Numerator df	2
Denominator df	12
Total sample size	15
Actual power	0.9562292

Select procedure: Effect size from means

Number of groups: 3

SD σ within each group: 2

Group	Mean	Size
1	8	40
2	10	40
3	15	40

Equal n: 5

Total sample size: 120

Effect size f: 1.47196

Table Three

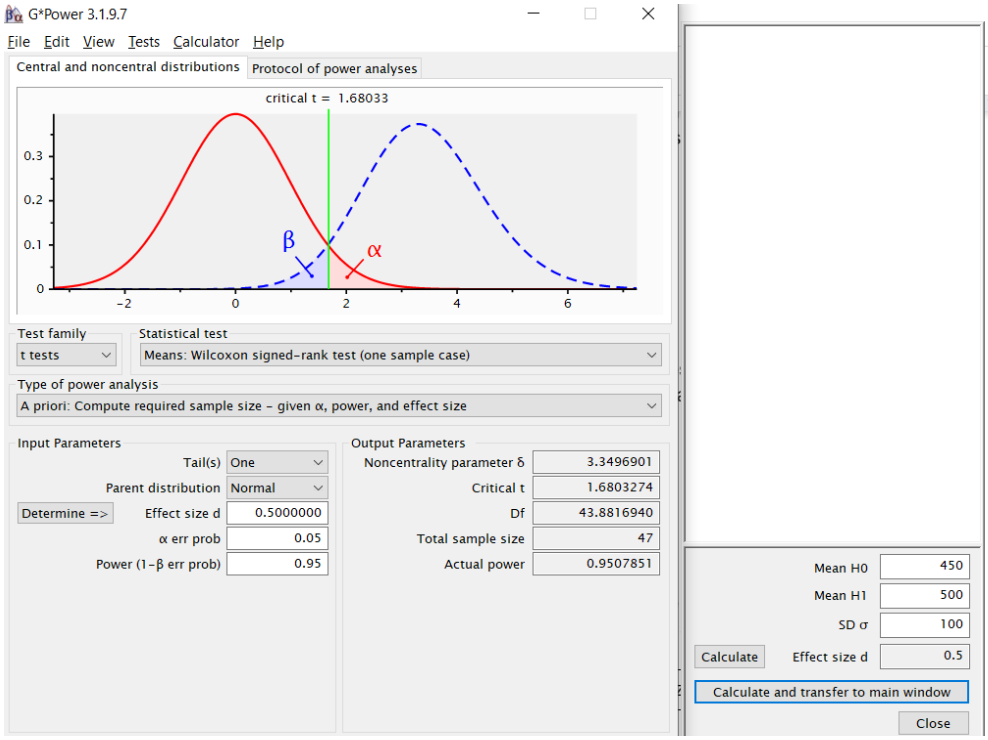
(Used A Priori Analysis to calculate sample size)

Effect Size	Power and alpha	Sample Size
1.4719601	0.80, 0.05	9
1.4719601	0.80, 0.01	12
1.4719601	0.90, 0.05	12
1.4719601	0.90, 0.01	15

2. Effect size, sample size and the significance criterion influence the achieved power in power analysis. A larger sample size yields higher power. The greater the error variance, the smaller the power. You have less power with smaller alpha (keeping effect size and power constant).

Question 6

1. Effect size d = 0.5



G*Power 3.1.9.7
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
critical t = 1.68033
0.3
0.2
0.1
0
β
α
−2
0
2
4
6
Determine =>
Calculate
Calculate and transfer to main window
Close

Test family: t tests

Statistical test: Means: Wilcoxon signed−rank test (one sample case)

Type of power analysis: A priori: Compute required sample size − given α, power, and effect size

Input Parameters

Tail(s): One
Parent distribution: Normal
Effect size d: 0.5000000
α err prob: 0.05
Power (1−β err prob): 0.95

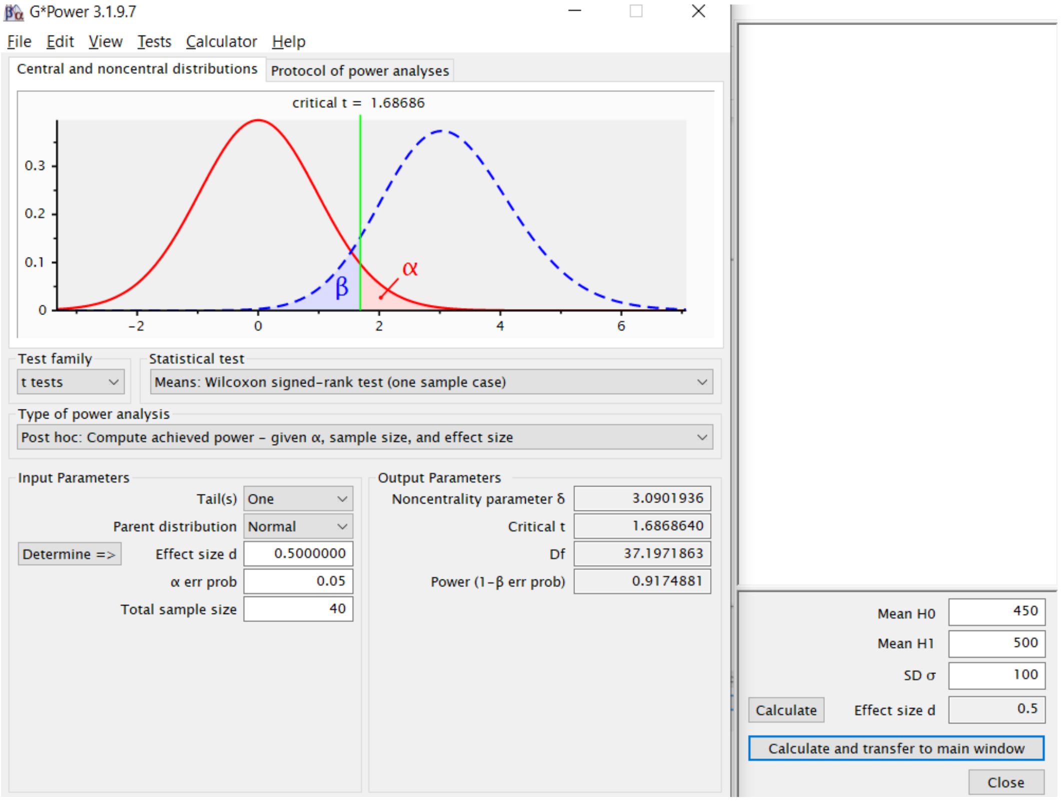
Output Parameters

Noncentrality parameter δ: 3.3496901
Critical t: 1.6803274
Df: 43.8816940
Total sample size: 47
Actual power: 0.9507851

Mean H0: 450
Mean H1: 500
SD σ: 100
Effect size d: 0.5

Figure 1: Calculation of effect size

Power = 0.9174481



G*Power 3.1.9.7
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
critical t = 1.68686
0.3
0.2
0.1
0
β
α
−2
0
2
4
6
Determine =>
Calculate
Calculate and transfer to main window
Close

Test family: t tests

Statistical test: Means: Wilcoxon signed−rank test (one sample case)

Type of power analysis: Post hoc: Compute achieved power − given α, sample size, and effect size

Input Parameters

Tail(s): One
Parent distribution: Normal
Effect size d: 0.5000000
α err prob: 0.05
Total sample size: 40

Output Parameters

Noncentrality parameter δ: 3.0901936
Critical t: 1.6868640
Df: 37.1971863
Power (1−β err prob): 0.9174881

Mean H0: 450
Mean H1: 500
SD σ: 100
Effect size d: 0.5

Figure 2: Calculation of achieved power using Post hoc analysis

2. Compared to question five, the data provided in this question is for only one group compared to three groups in the previous question. The test we perform here is one sample T-test of the mean compared to a F-test with fixed effects in question 5.

Question 7

1. The type of design is a within subject design.
2. The outcome variable is a numerical measure of the knowledge of French taken multiple times across specific intervals for the same subjects each time.
3. TABLE FIVE:

Effects size, power, and alpha	Sample size
0.35, 0.95, 0.05	20
0.25, 0.95, 0.05	36
0.15, 0.95, 0.05	98
0.12, 0.95, 0.05	152

4. Since I kept power and alpha constant, it is clear that as effect size decreases, the required sample size for the experiment increases.
5. I would recommend a sample size of 36, because it is the largest sample size that will allow the experiment to be run in just one French class. That way, you can guarantee that all subjects are getting the same method of instruction and evaluation.

Question 8

1. The design is a Factorial ANOVA with two categorical predictors and a numerical outcome variable.
2. The independent variables are method of teaching (formula or formula plus visuals) and the subject’s major (sociology, science, or humanities).
3. TABLE SEVEN:

Givens	Recommended sample size based on major (df numerator = 2)	Recommended sample size based on method of teaching (df numerator = 1)	Recommended sample size based on interaction effect	Sample size you recommended

Effect size (0.10)
alpha = 0.05
power = 0.80
	967	787	967	161 participants in each group

Effect size (0.10)
alpha = 0.05
power = 0.90
	1269	1053	1269	212 participants in each group

Effect size (0.10)
alpha = 0.01
power = 0.80
	1393	1172	1393	232 participants in each group

Effect size (0.10)
alpha = 0.01
power = 0.90
	1748	1492	1748	291 participants in each group

4. It appears that sample size required gets larger as power gets larger and as alpha gets smaller. With the alpha getting smaller, it means we want a smaller margin for error, which means it will take a larger sample size to achieve that necessary accuracy level. Effect size is held constant in this question, so no conclusions can be drawn on it.

Question 9

1. Between factor would be the treatment level with two factors (Calcium Only & Calcium plus Vitamin D)
2. The within Factor would be the measure of bone over time with three levels (Bone density: initial, after 6 months, after 12 months).
3. THe test I would use would be a repeated measure since the outcome is numerical with three measures over time with the same subjects, which would be in the f-test family.
4. Total sample size would be 116, making it 58 per group.

Question 10

1. Chi-square analysis
2. Logistic Regression
3. Ordinal Regression
4. Two-Sample test of Mean
5. One-way ANCOVA
6. Logistic Regression
7. MANCOVA
8. Simple Linear Regression
9. One-way ANOVA
10. Mixed Design
11. Multinomial Regression
12. Multiple Linear Regression
13. Multiple Linear Regression
14. Repeated Measure
15. Mixed Design
16. Multiple Linear Regression
17. Chi-Squared
18. One-way ANOVA
19. Repeated Measure
20. Multinomial Regression
21. Logistic Regression
22. Mixed Design