Lecture Notes

Analysis : conclusion validity

Conclusion validity - is there relationship between things that u measuring, we are not concerned with causality just weather we can infer relationship from data

Threats to conclusion validity

1. Significant result/no significant result - true positive false positive

Type 1 error

* Replicability crisis- replication of findings

Holm-bonferroni correction

Type 2

* Assumption is incorrect then it is a bias
* Statistical power: porb of no type 2 error

Effect size

**Questions**

* **Research question:**

Does COVID affect post trends from mental illness related subreddits?

* **Hypothesis:**

Mental health-related subreddits exhibit a shift in post trends and the number of engagements based on post texts, number of comments and upvote ratios between periods *pre-covid and post-covid*

Time series - seasonality or stationary for bias

1. **Describe in detail the independent variables, dependent variables, and the relationship that you expect to find between them in your study.**

**Post text, num comments, id, upvote ratio→independent variable in the study**

**dependent variable: post trends (engagement)**

**Relationship: direct relationship**

* Added: Post Length

1. **What statistical or machine learning analysis do you plan to conduct to determine whether this relationship exists or not? Justify this analysis relative to the types of variables (interval, ordinal, nominal, etc.) you are relating**

**Statistical analysis - g power:**

* 1. **t test (compare mean of two groups) → compare mean or median of num of comments, post length, upvote ratio - pre and post covid**
  2. **Chi square test (two variables are independent or related of each other, using categorical variables) → post text (shift of trends pre vs post)**
  3. **regression analysis: evaluating correlation ?**
  4. **Time series analysis: seasonal transition**

**Subgroup analysis: difference between subreddit groups**

**ML: topic modeling**

**Post length - post text (nominal)**

**Pre vs post - created on interval**

**Id - frequency (nominal)**

**Total comments - comparing frequency (nominal)**

**Upvote ratio- interval**

**Score - interval**

1. **Generate and clean pilot data reflecting each of the variables identified above. Display descriptive statistics of these data, including histograms, measures of central tendency and dispersion.**

**Histogram of num of comments, length, upvote ratio - pre vs post & measure of central tendency (mode, mean, median) and dispersion (variance, standard deviation, inter quartile done check stage6 analysis colab**

**null hypothesis: covid affects post trends**

| **null** | **Relationship (true)** | **No relationship (false)** |
| --- | --- | --- |
| **significant** | True positive | False positive type 1 error |
| **Not significant** | False negative type 2 error | True negative |

Type 1 - theres relationship but actually no relationship

Type 2 - theres no relationship but actually theres relationship

Precision - beginning of covid, test and result in positive to covid is very concise

Recall- many false negative later on

**Type 1 error - COVID doesn’t affect post trends**

**Type 2 error -**

1. **For each variable, indicate what, if any, data transformation is necessary in order to conduct your intended analysis. Justify your choice of transformation relative to the distributional assumptions underlying your chosen statistical or ML model.**

**Data transformation**

* **Post length from post text**
* **post text into NLP - topic modeling doc2vec**

**Relative to chosen statistical/ML model:**

* **Post length will be t test**
* **Post text into nlp might be chi square - need to check**

1. **Conduct a preliminary test of the relationships between your variables of interest using a correlational method. Justify your choice of this method given the distribution of the data.**

* **Finish scraping and find correlations of 9 subreddits**

1. **How many analyses do you plan to conduct using your dataset? Make sure you control for multiple comparisons. Choose a Type I error rate using this information.**
2. **Covariates should be included in your analysis if you expect confounds but cannot control for them because of limitations on yourdesign. What covariates do you plan to control for statistically, if any? Justify your selection of these.**

* **Major events during the time periods**
* **Characteristics of users**
* **Demographic**
* **Time-related external factors**

1. **Conduct a power analysis to determine how much data is necessary to detect the hypothesized relationship. You may use off-the-shelf software. Your power analysis should be based on pilot data.**

* **G power thing**

1. **Conduct your intended analysis using your pilot data. Specify and control for any covariates that you have identified. What does your analysis indicate regarding the hypothesized relationship in the data?**
2. **Provide an overall summary of your plan thus far:question,sampling strategy,measurement strategy, design, and analysis plan. Indicate what you expect to find in the analysis, what the design allows you to say given these findings, what the implications are for your theory, how far you can generalize, and to what extent this answers your question.**

* **How do we control covariates and incorporate that into our analysis**

**Statistics notes**

* **T test paired (compare two groups) - numerical paired**
  + **Num of comments**
  + **Post len**
  + **Using one or two tail? What is effect size, sd of two groups**
  + **T value is effect size, anova - f value is effect size**
  + **Two tails - null: there is a post trend change, alt: there is no difference**
  + **Power : 0.8**

**P value is probability of data given the hypothesis**

**ANOVA is used to test for significant differences in means among groups, while ANCOVA is used to test for significant differences in means while controlling for the effects of one or more covariates**

* **ANCOVA**
  + **Mean of post lengths of 9 subreddit**
  + **Factor variable:**
  + **Covariate: pre and post**
  + **Response variable: post len**
* **correlation**

**Columns we are using: pre vs post →**

**post text (post length), user id, num of comments, upvote ratio, month**

**Control group: pre post**

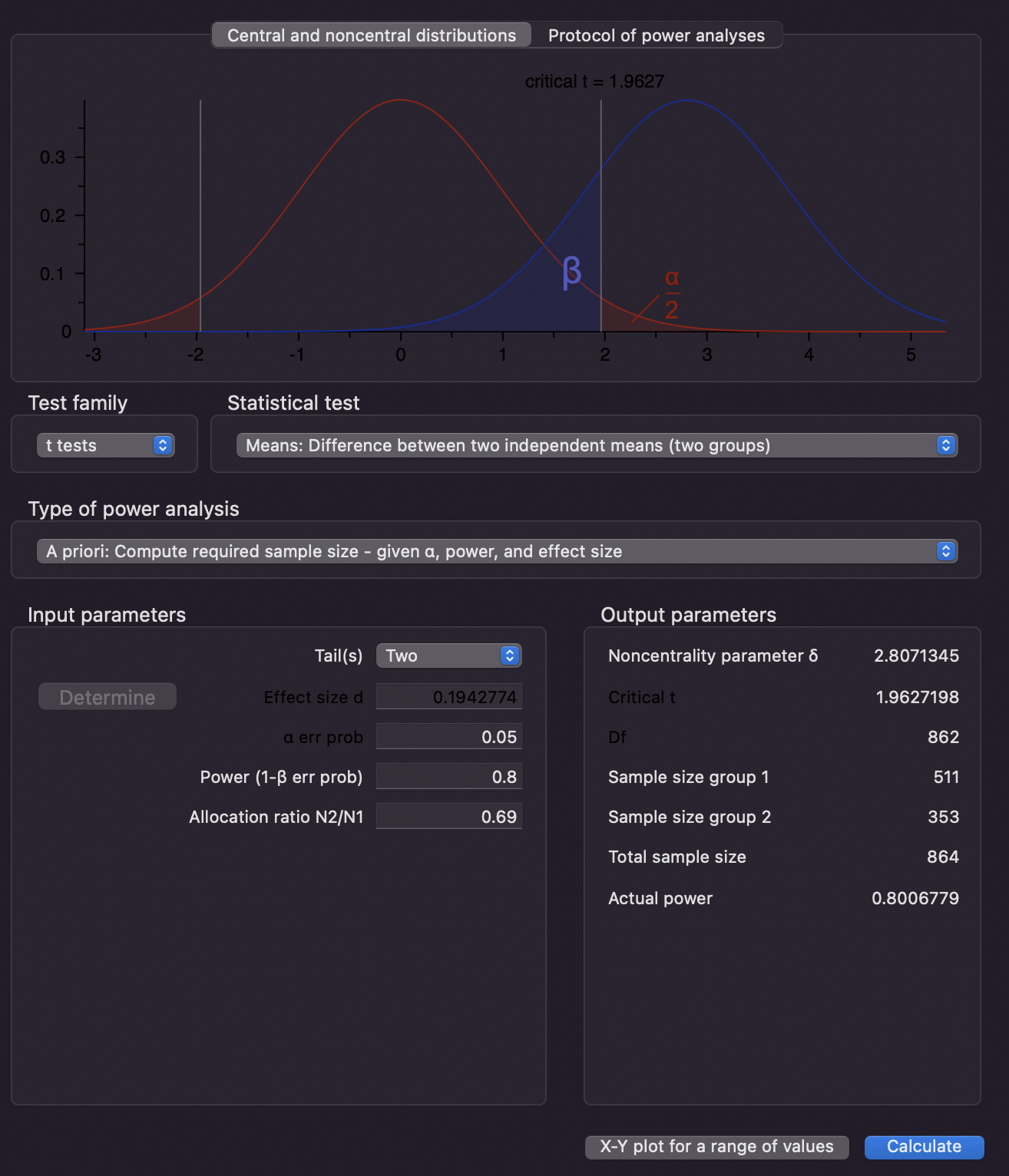
**Treatment group: post post**

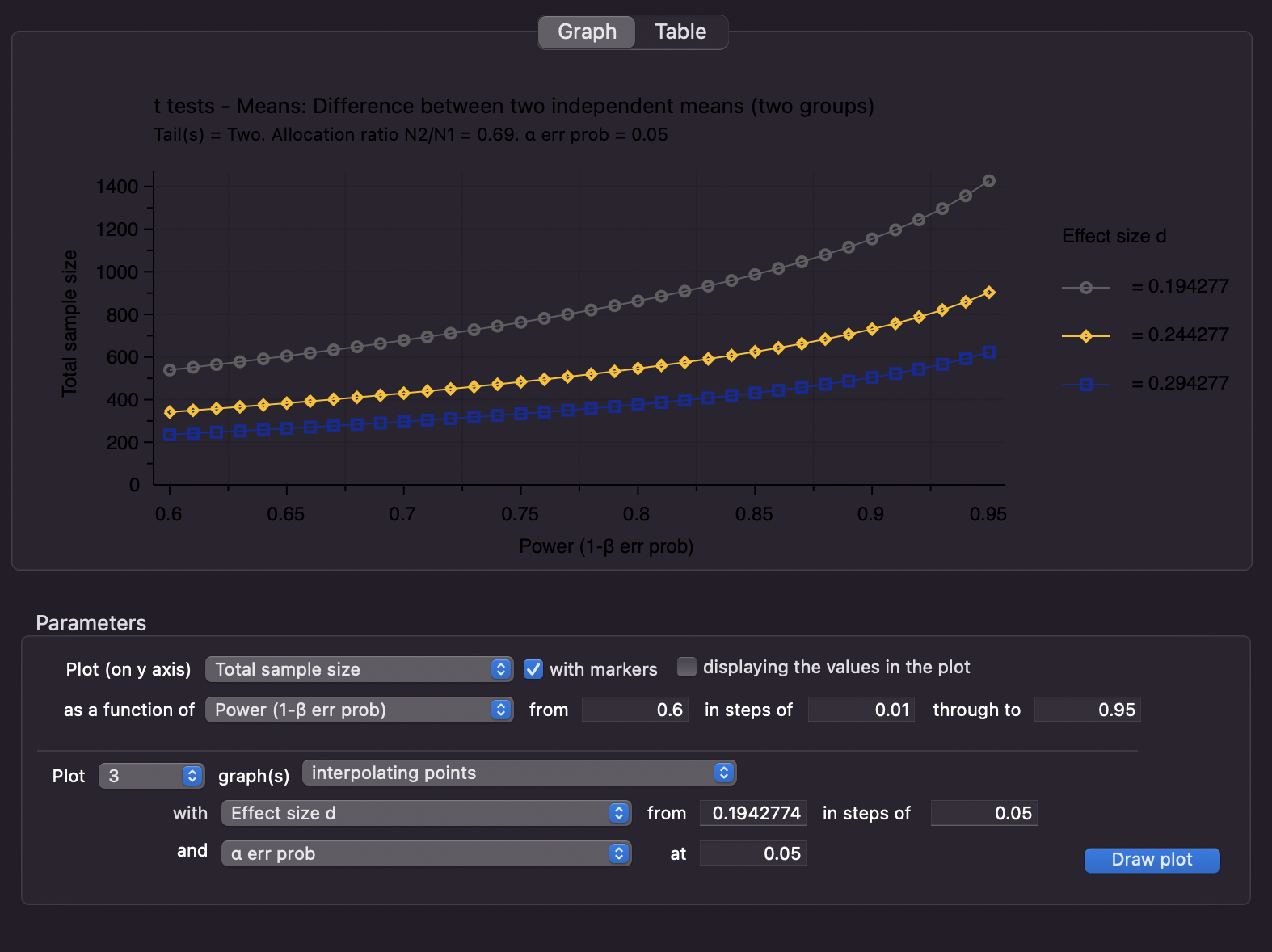
**Compare t test of pre and post of 9 subreddit**

**Power analysis: 9 subreddit**

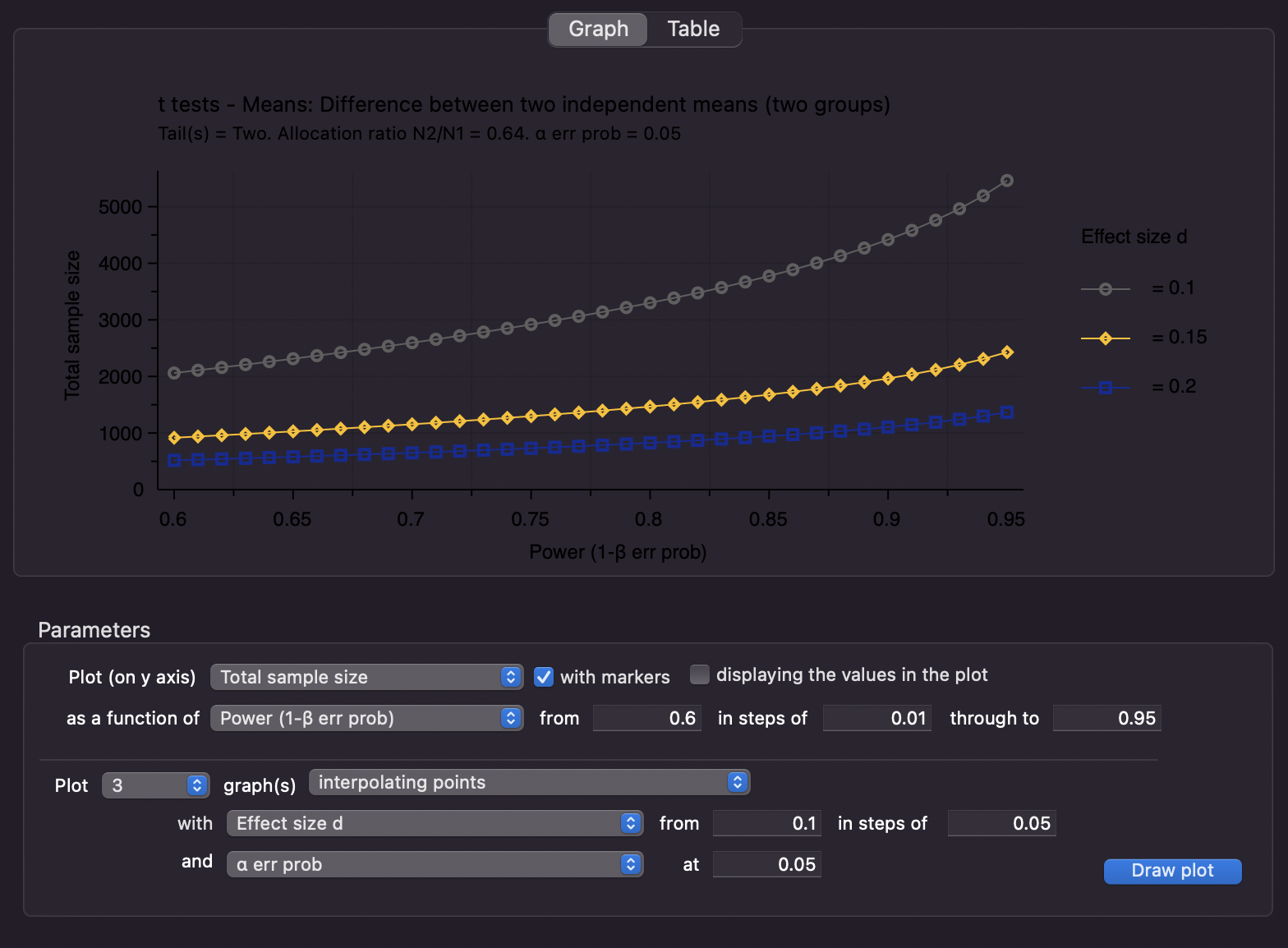
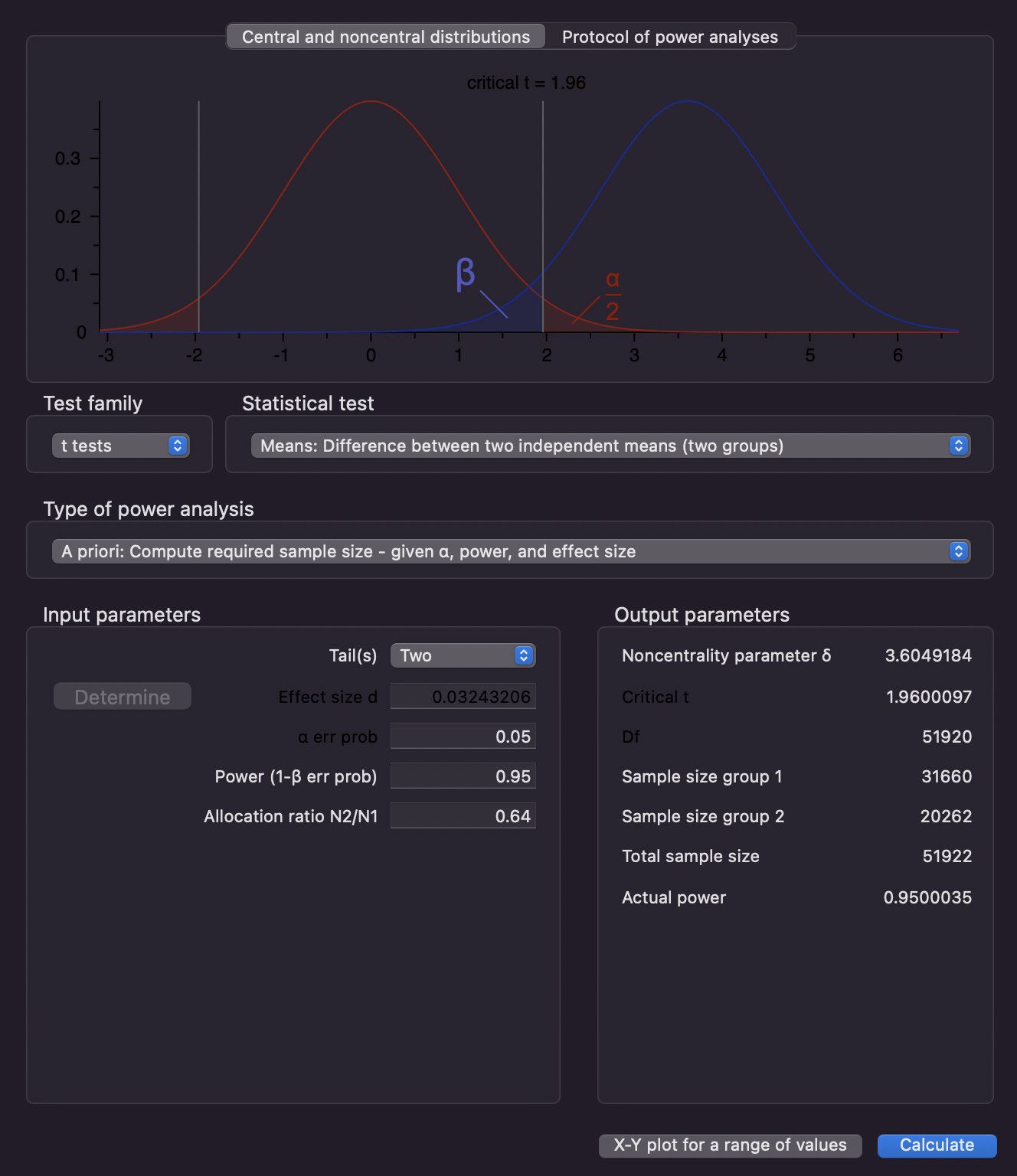
**Depression post length pre vs post - using mean, sd, and correlation of two**

**Depression**

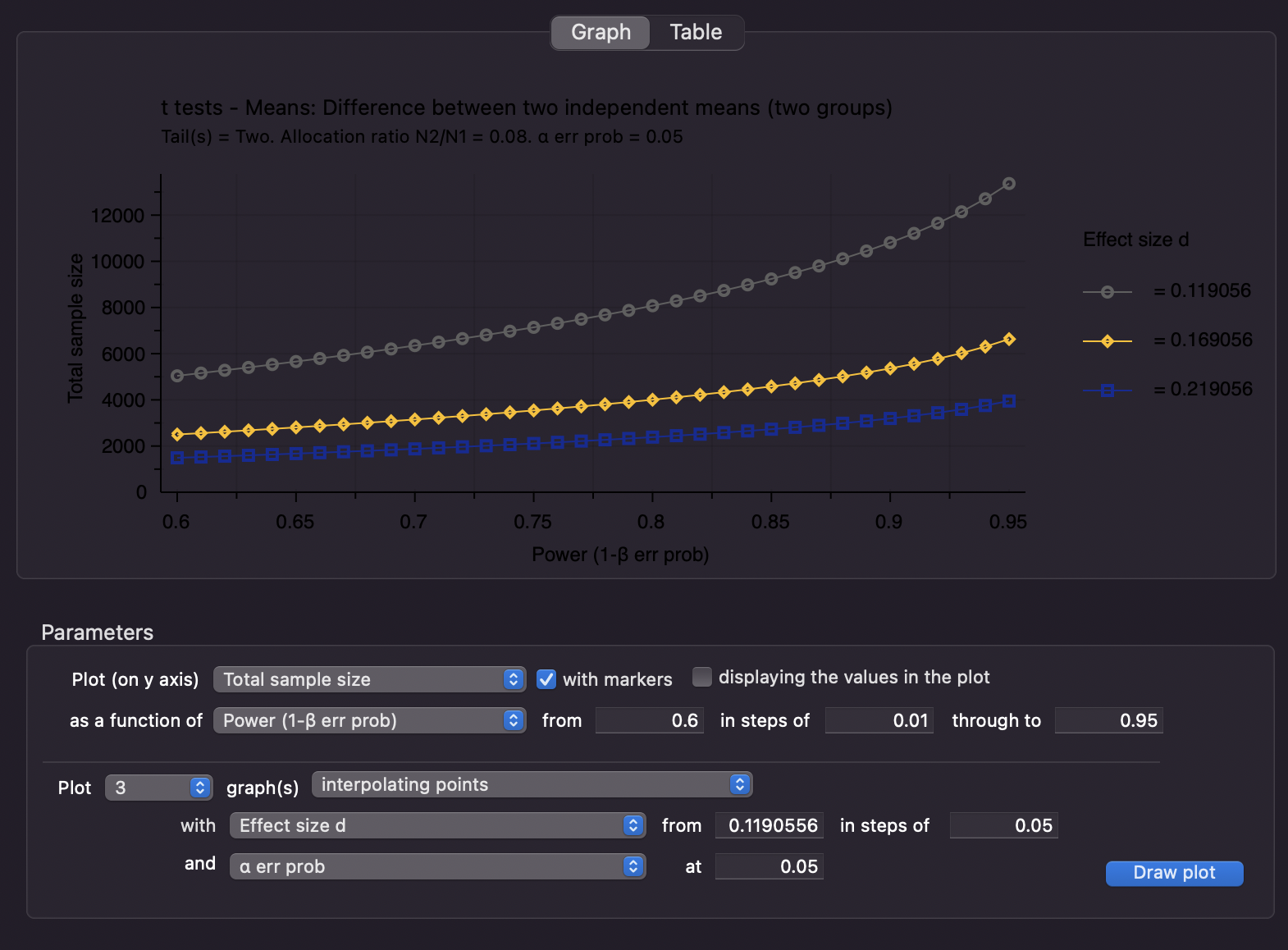
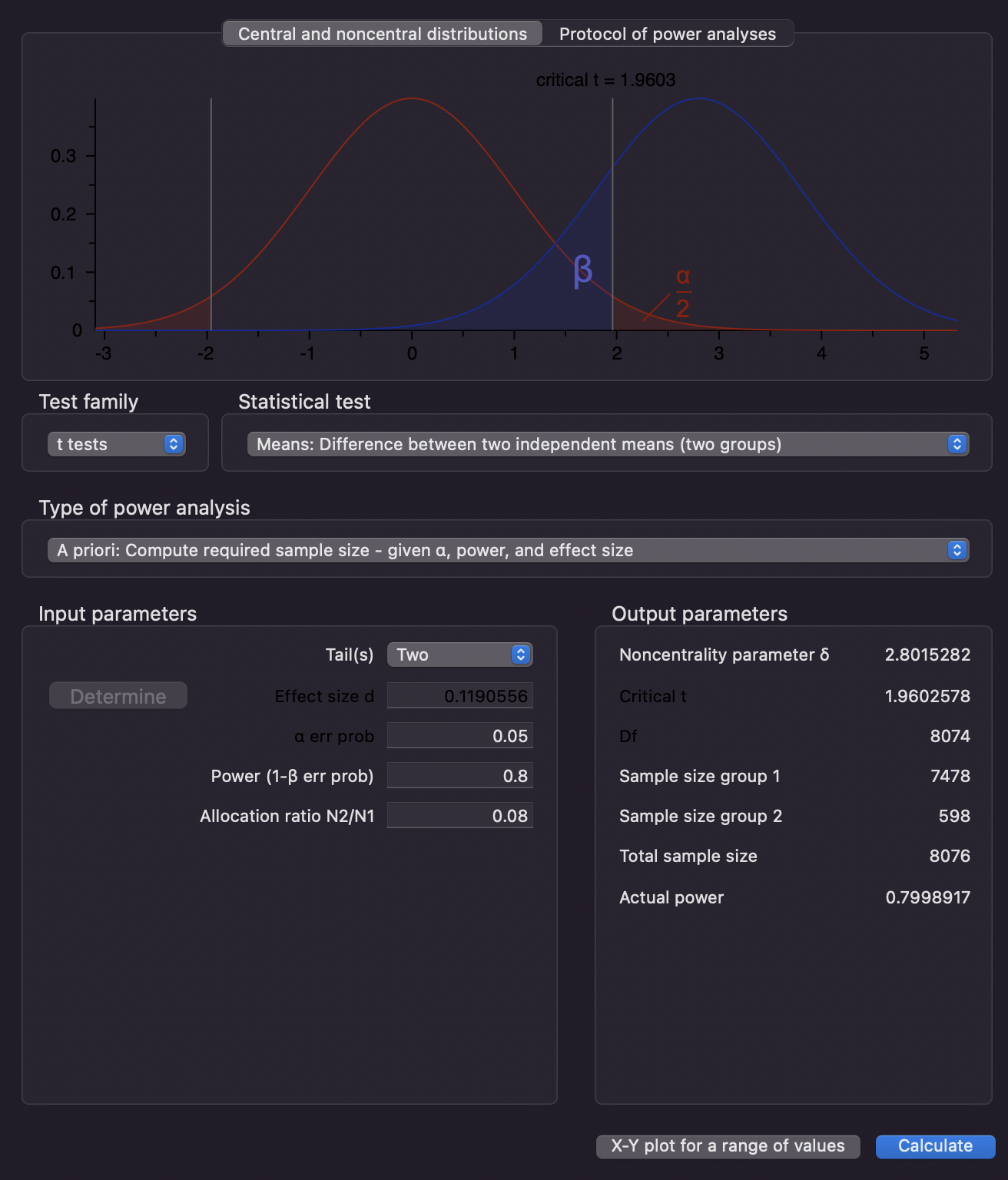
****

****

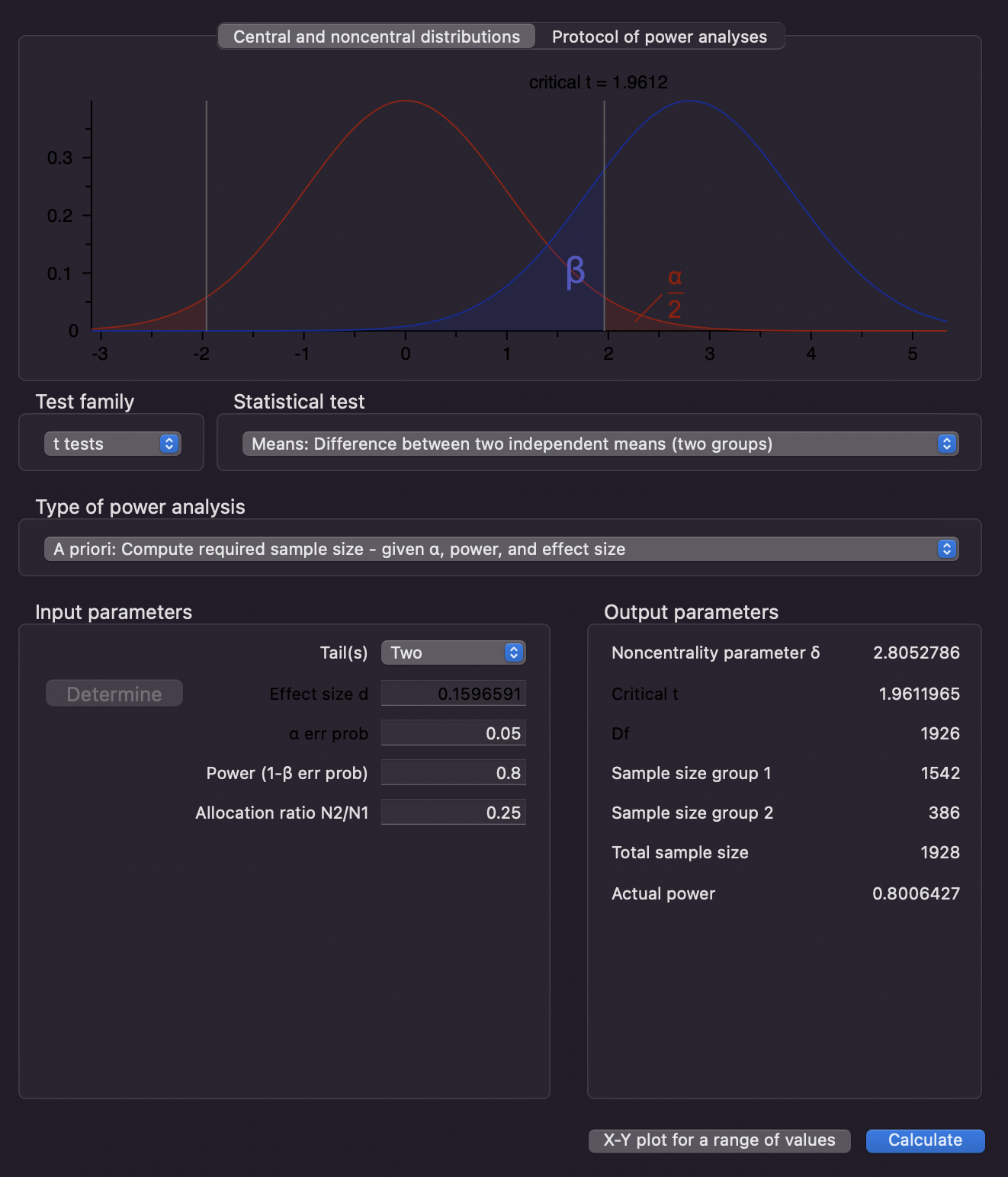
**anxiety**

****

**Adhd**

****

**ptsd**

**ed**