Power and Sample Size

UCR GradQuant Workshop

11/04/2016

UCR GradSuccess
graduate.ucr.edu/success

Plan of Attack

- Review statistical hypothesis testing, which is the necessary foundation for power analyses
- Define power and the components of power analyses
- Software
- Complete several examples of power analyses for different situations and statistical analysis setups

How much data do we need

- How many subjects should be included in the research.

  Without considering the expenses, the more data the better.

- It is not feasible to collect data on the entire population of interest.

- Consider the collected data as a random sample of the population of interest.

Rules of Thumb

- Feasible in terms of budget and research time frame
- Sufficient data to ensure results to be Accurate, Efficient, and Credible

What is Power Analysis

- Based upon the statistical test for the main research question,
- Power analysis is intended to determine the minimal data (or sample size) required for detecting a significant research finding.

Example: An experimental Study

- We want to determine whether Drug A lowers cholesterol levels in adults. We plan to have one control group and one treatment group whose members will receive the drug. We will measure their cholesterol levels at the end of a 4 week study.

Example: An experimental Study

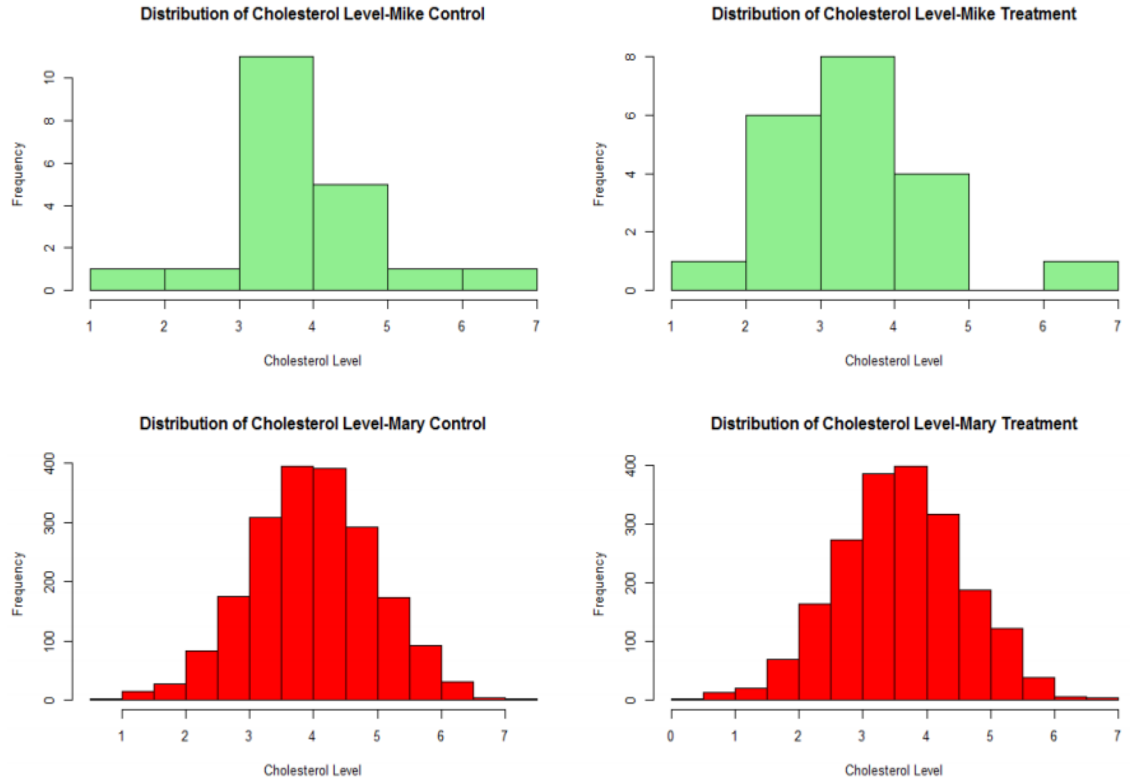
- It’s practically impossible to collect data on an entire population of interest
- Solution: examine data from a random sample to provide support for or against your hypothesis
- How many samples/participants/data points should we collect?
- Which is better? 20 data points or 2000?

Example: An experimental Study

- let’s look at two different research strategies
- Plan 1 is to enroll 20 participants in each treatment group, while Plan 2 is to enroll 2000 participants in each group

Example: An experimental Study

- Results of two plans



Example: An experimental Study

- Plan 1: t = 1.5596, df = 38, p-value = 0.1271
- Plan 2: t = 15.984, df = 3998, p-value < 0.000000000000022
- Conclusion of Plan 1: there is insufficient evidence to support the claim that the use of Drug A results in lower cholesterol levels in adults
- Conclusion of Plan 2: there is sufficient evidence to reject the null hypothesis and conclude that the use of Drug A results in lowers cholesterol levels in adults

Example: An experimental Study

- The Truth: simulated data from both control groups had mean=4, sd=1, while data from both treatment groups had mean=3.5, sd=1
- The population for both sets of data had an effect size of d=0.5, but only Plan 2 had enough participants to observe the difference between groups
- The difference that we were suspected would be present was observed in Plan 2, but huge amounts of time and money was spend to enroll 2000 participants per group

Components of Sample size Calculation

- What test statistic will be employed ? Hypothesis Testing: The null hypothesis vs. The alternative hypothesis
- Alpha Level (or desired accuracy; width of confidence interval)
- Power
- Effect size: expected differences and variation of outcome measures
- Sample size

Types of Statistics

- Means
  - Compare 2 means (t-test)
  - Compare 3 or more means (ANOVA)
- Proportions
  - Compare 2 proportions
- Bivariate relationship – correlation (r)
- Multiple regression – Multiple R2
- Cluster sampling/multi-level

Hypothesis Testing

- The null hypothesis: This hypothesis predicts that there is no effect on the variable of interest
- The alternative hypothesis: This hypothesis predicts that there is an effect on the variable of interest (or a difference between groups).
- Statistical tests look for evidence to reject the null hypothesis and conclude the alternative hypothesis (an effect is existing)
- Sample size calculation: Determine the minimal amount of data required.

Alpha level and Power

Table of error types		Null hypothesis (H_0) is	
		True	False
Judgement of Null Hypothesis (H_0)	Reject	Type I error (False Positive) α	Correct inference (True Positive) (1- β)
	Fail to reject	Correct inference (True Negative) (1- α)	Type II error (False Negative) β
Type-1 = False result but accept it (False Positive) Type-2 =True result but rejected it (False Negative)			

Alpha Level and Power

- Alpha level: Probability of incorrectly concluding (from sample data) a significant effect when it does not really exist in the population (Type-I error).

  -- Alpha level is usually set as .05

- Power: Probability of correctly concluding (from sample data) a significant effect when it really exist in the population.

  -- Power is usually set as .80

Effect size

- Effect sizes - standardized measure of the magnitude of a difference or relationship.
- There are different measures used in different types of analysis
- Larger effect sizes are easier to observe (require a smaller sample size), while smaller effect sizes are more difficult to observe (require more samples)

Computing Effect Size

- Various formulas depend on type of statistic

  e.g., for difference in means (t-test)

$$d = \frac{mean_1 - mean_2}{standard\ deviation}$$

Various labels:

- d for difference in two means
- w for difference in proportions
- r for correlations
- f for difference in many means (e.g., One-way ANOVA)
- η^2 for variance explained
- R^2 for multiple regression

Determining Effect Size

- Based on substantive knowledge
- Based on findings from prior research
- Based on a pilot study
- Estimate required sample size for a range of effect sizes

  -- e.g., small, medium and large effect size defined by Cohen

Magnitude of Effect size

From by Cohen, 1988
The bigger the effect size, the easier the detection.

Statistic	small	medium	Large
Means - d	0.20	0.50	0.80
Association – Chi-square - w	0.10	0.30	0.50
ANOVA - f	0.1	0.25	0.4
ANOVA - $\eta 2$	0.01	0.06	0.14
Correlations - r	0.10	0.30	0.50
Multiple regression - Partial R2	0.02	0.13	0.26

Four types of power analysis

- Determining sample size is an *a priori* analysis
  - How many participants or samples should be in the study? (n)
- Determining achieved power is a *post hoc* power analysis
  - What are the chances I observe the difference that's actually there?
- Determining effect size is a *sensitivity* analysis
  - How large of an effect will the treatment have on our response?
- Determining Type I error rate (alpha) is a *criterion* power analysis
  - What's the probability I will see a false positive?

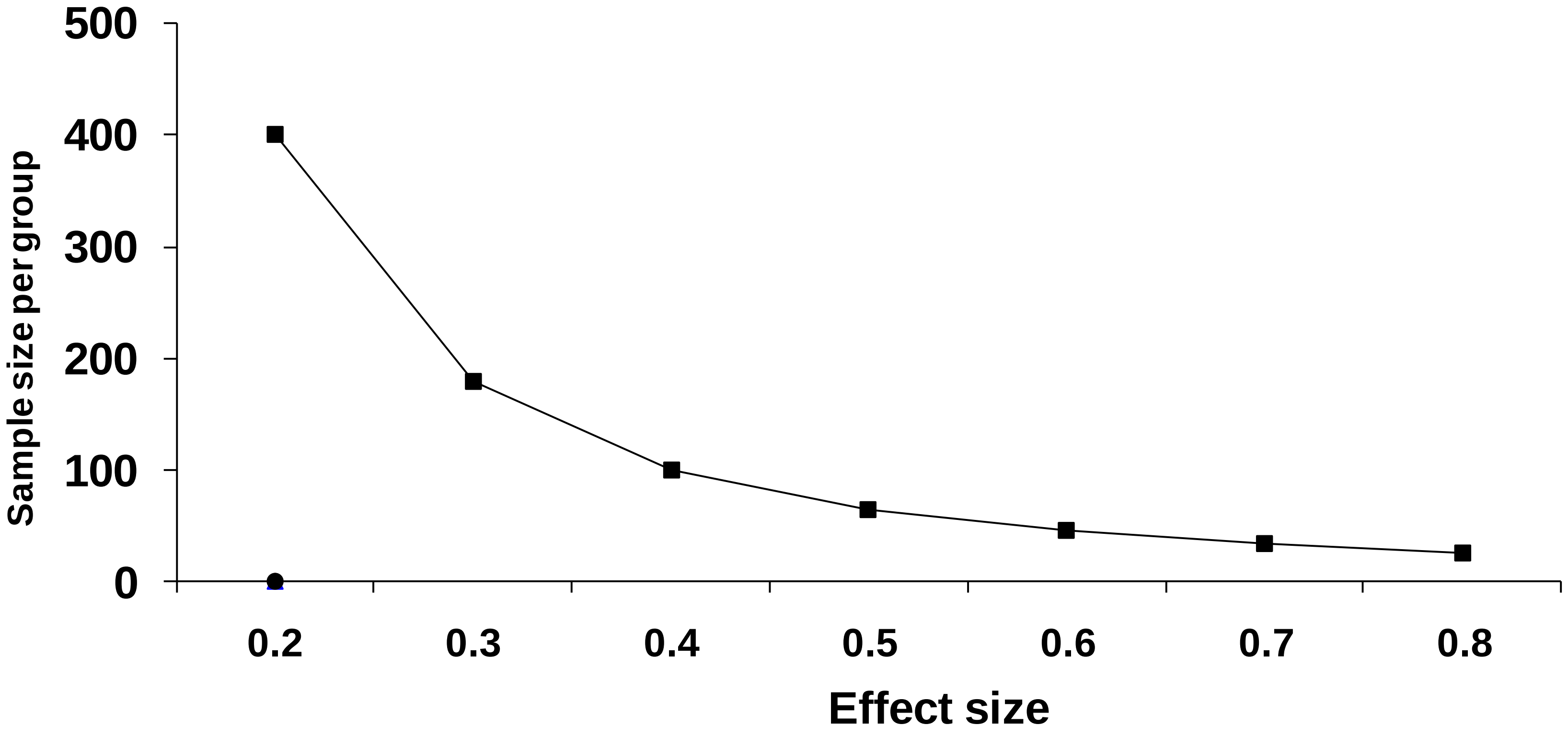
Steps for Sample Size Determination

- Decide types of outcome statistics (e.g., mean, proportion, correlations,…)
- Specify 1- or 2-tailed tests
- Specify desired alpha level and power
- Specify the desired effect size (from literature, pilot study, or best guess)

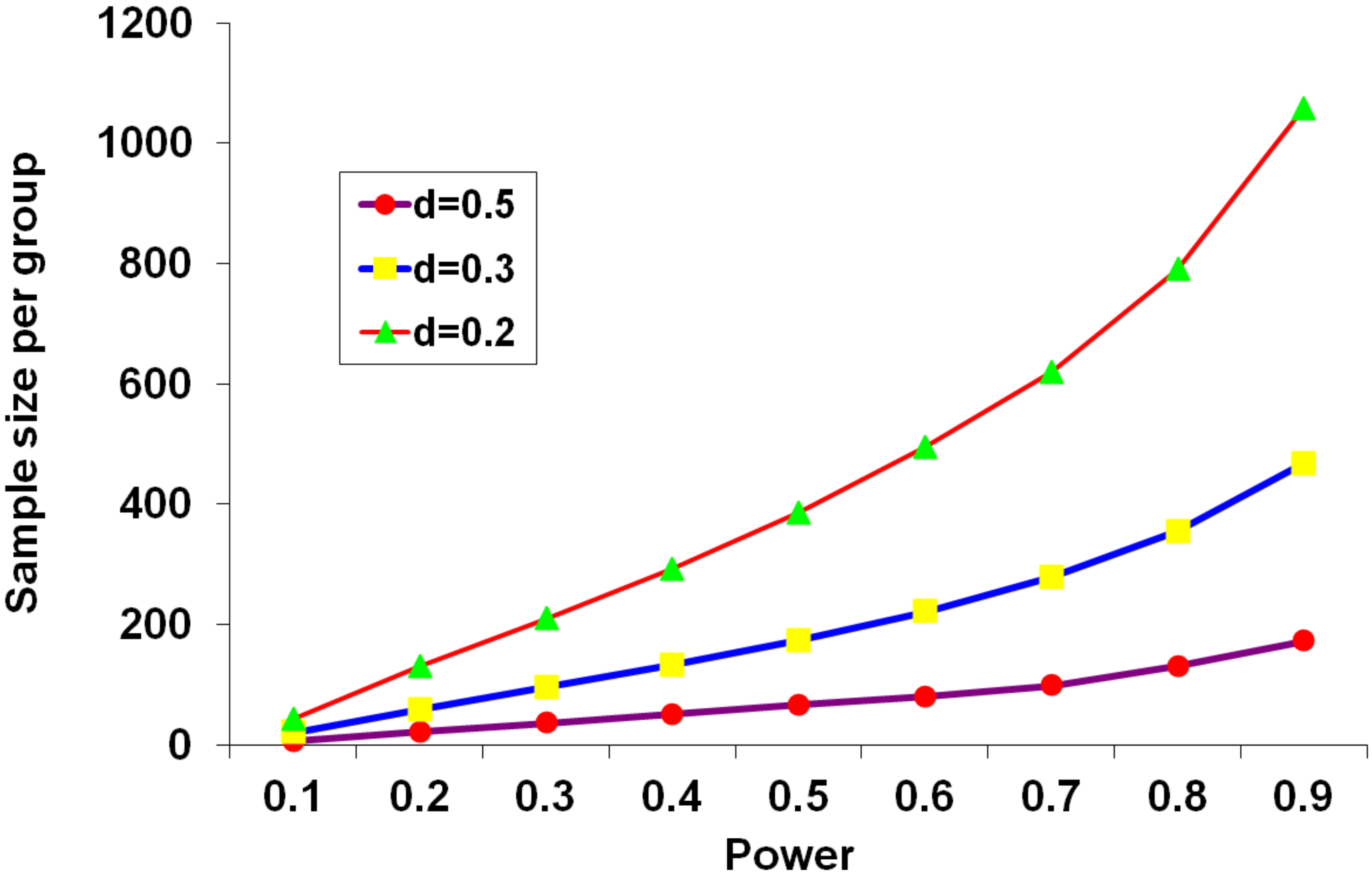
General Rules: Required Sample Size

- Detecting small effect sizes --> larger N
- Smaller alpha or greater power --> larger N
- 2-tailed test --> larger N than 1-tailed test
- Addition of covariates (e.g., ANCOVA) → reduce error variance, then increase effect size and decrease N

Effect Size vs. Number of Subjects per Group for two-tailed t-test with α=.05, power=.80



Functions of Power vs. Number of Subjects per Group for two-tailed t-test with α=.05



General Rules: Required Sample Size

- Cluster sampling/multi-level data structure:

  -- Larger N as intra-class correlation increase

- Follow-up with repeated measures:

  -- More repeated measures, smaller N per group

Software

- SamplePower. SPSS
  - SPSS module for computing power/sample size.
- Proc POWER and Proc GLMPOWER. SAS
  - SAS procedures for computing power/sample size.
- R pwr package
- G*Power - a free tool to compute statistical power analyses. It supports many designs (t-test, ANOVA, ANCOVA, repeated measures, correlations, regression, logistic, proportions, Chi- square, nonparametric equivalents).

Software

- The steps involved in conducting a power analysis are as follows:
  - Select the type of power analysis desired (a priori, post-hoc, criterion, sensitivity)
  - Select the expected study design that reflects your hypotheses of interest (e.g. t-test, ANOVA, etc.)
  - Select a power analysis tool that supports your design
  - Provide 3 of the 4 parameters (usually alpha=.05, power = .80, expected effect size, preferably supported by pilot data or the literature)
  - Solve for the remaining parameter, usually sample size (N)

Examples

T-test: Difference between two independent means (two groups)

- The null and alternate hypothesis of this t test are:

$$H0\colon \mu_1 - \mu_2 = 0$$
$$H1\colon \mu_1 - \mu_2 \neq 0$$

- The two-sided test ("two tails") should be used if there is no restriction on the sign of the deviation assumed in the alternate hypothesis. Otherwise use the one-sided test ("one tail").
- The effect size index d is defined as: $d = (\mu_1 - \mu_2)/\sigma$
- conventional values for d: small d = 0.2, medium d = 0.5, large d = 0.8

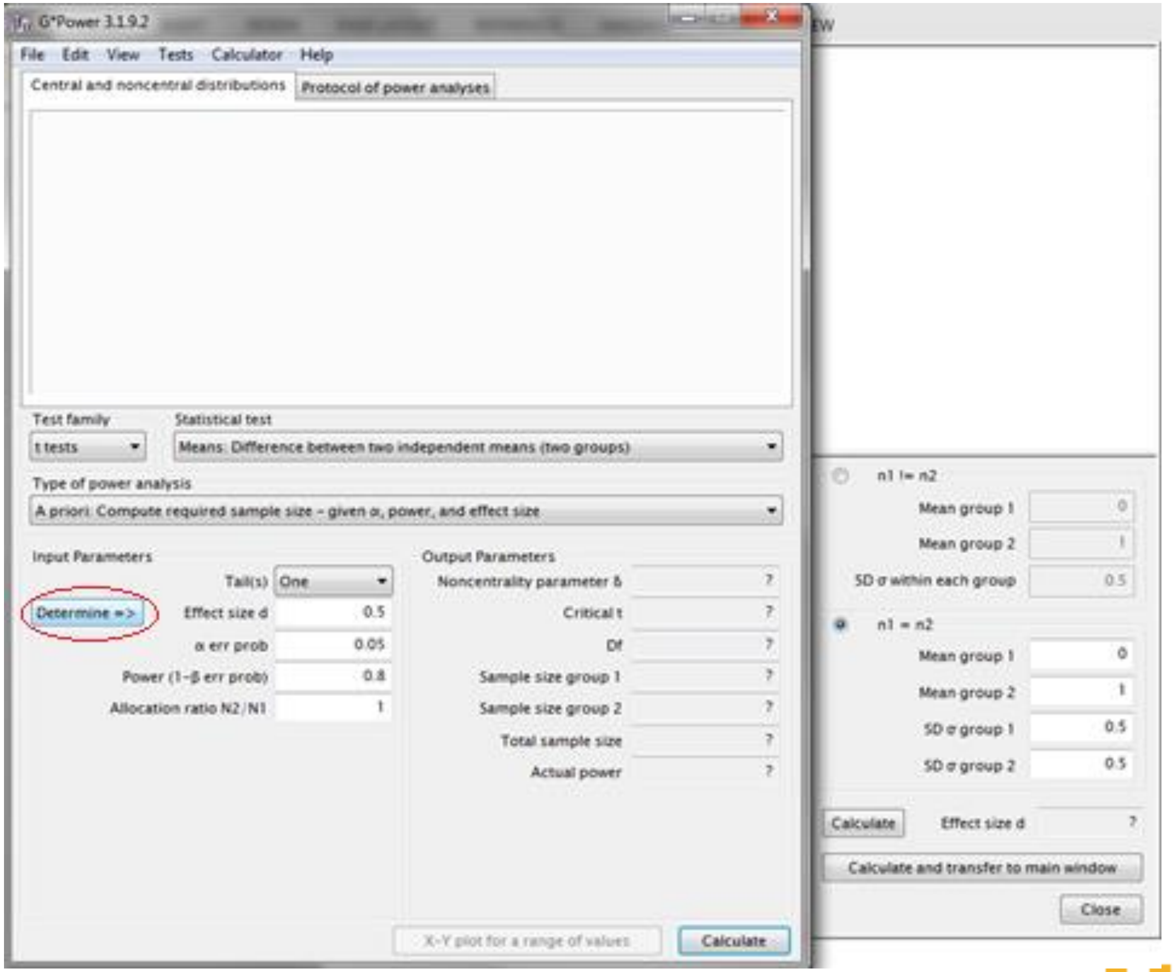
T-test: Difference between two independent means (two groups)

- Remember the cholesterol study example?
  - Goal: determine whether the intervention group has a lower cholesterol level on average at the end of the study
  - Previous studies have indicated that drugs like Drug A have had a medium effect when it comes to lowering cholesterol (convention: d = 0.5)
  - How can we figure out how many participants to enroll to achieve 80% power?

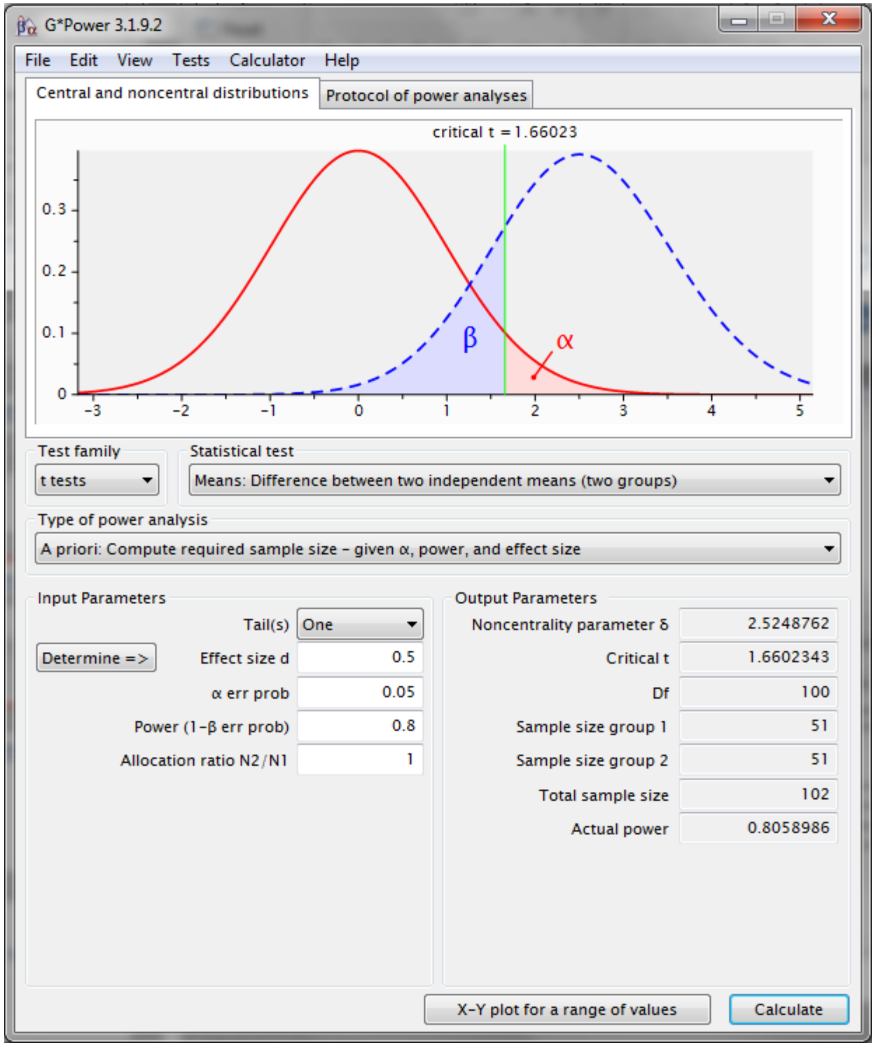
T-test: Difference between two independent means (two groups)

- What information do we need?
  - Type of analysis →two-sample t-test
  - Sample size → unknown
  - Effect size → medium (convention: d=0.5)
  - Type I error rate (alpha) → fixed at 0.05
  - Power → 80%

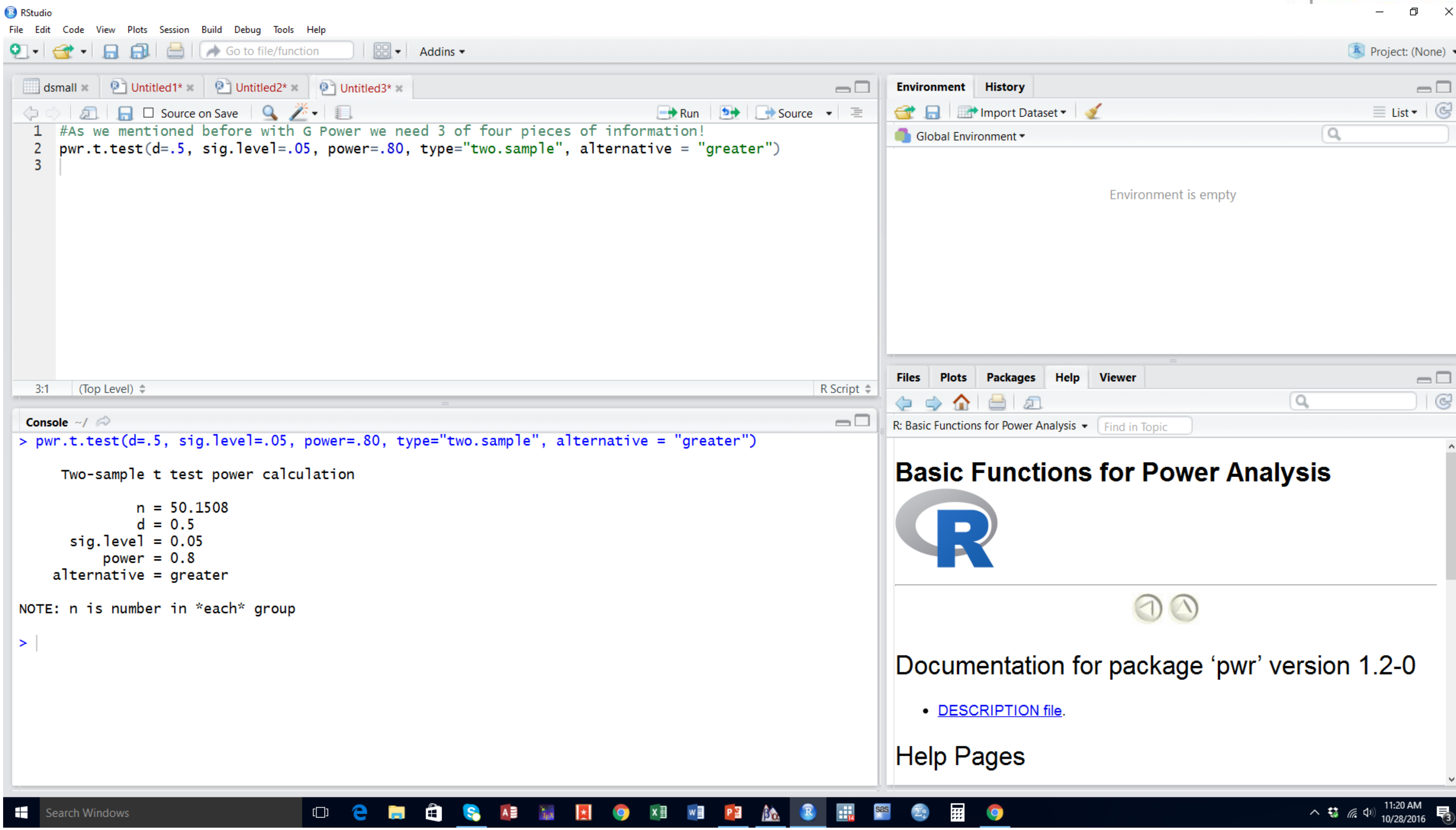
T-test: Difference between two independent means (two groups)



T-test: Difference between two independent means (two groups)

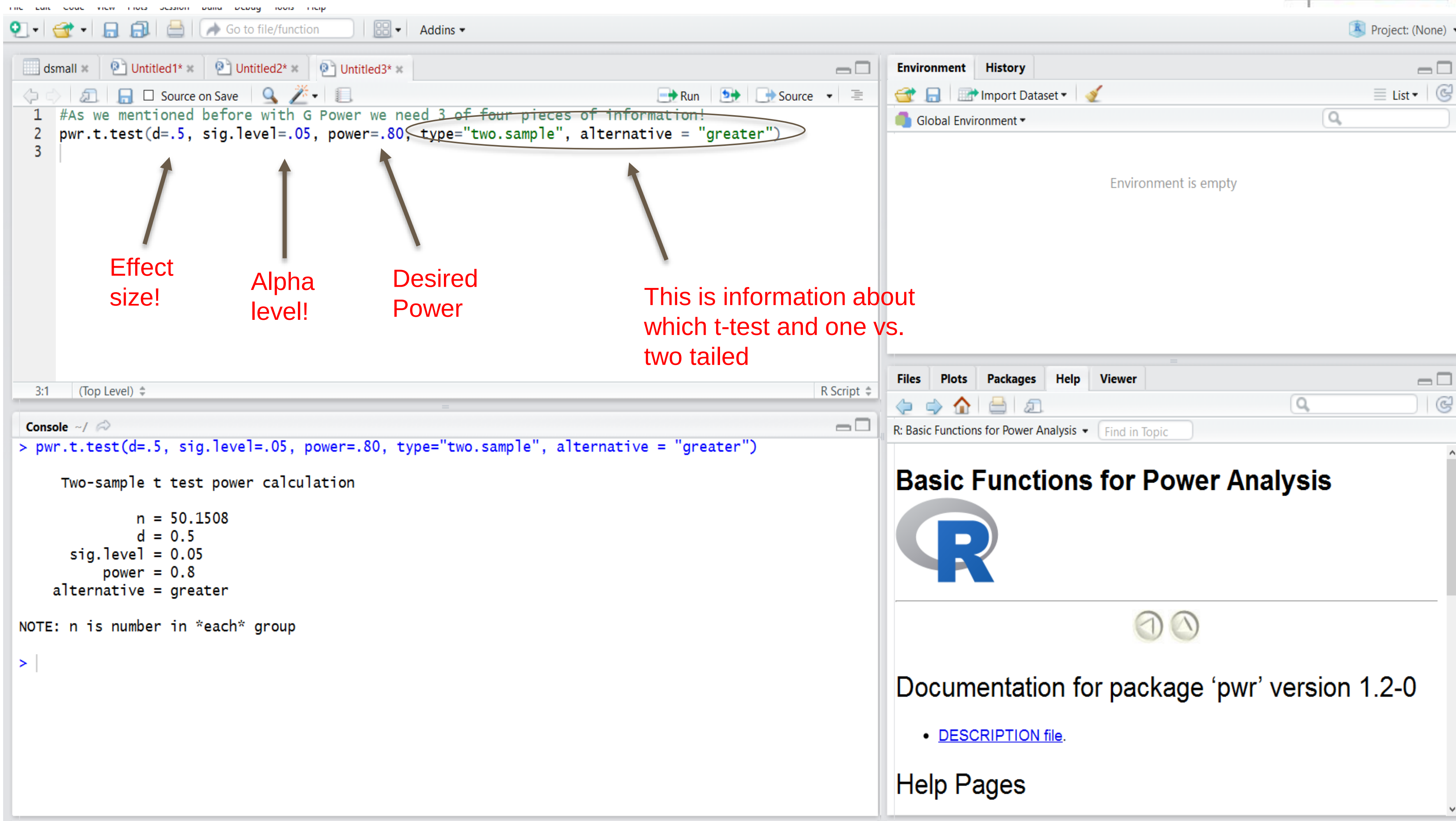


Let's see Power Analysis in R now

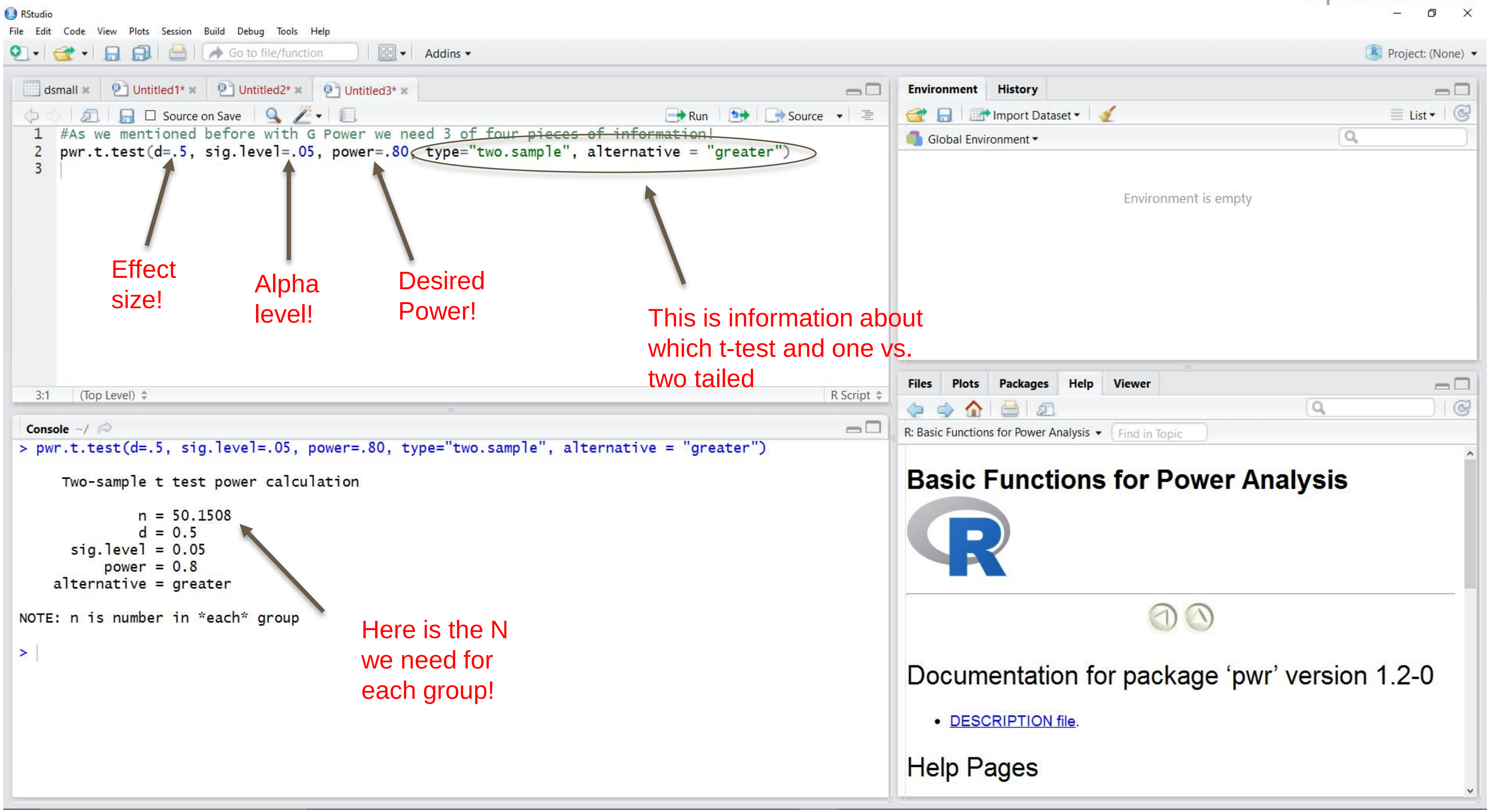


UCR GradSuccess graduate.ucr.edu/success

Let’s see Power Analysis in R now



Let’s see Power Analysis in R now



UCR GradSuccess
graduate.ucr.edu/success

Comparison of proportions between two independent groups

- Use Fisher's exact test

	Group1	Group2	Total
Success	x1	x2	m
Failure	n1 – x1	n2 – x2	N - m
Total	n1	n2	N

- The null and alternate hypothesis are:

$$H0:\ \pi_1 - \pi_2 = 0$$
$$H1:\ \pi_1 - \pi_2 \neq 0$$

where π_1 and π_2 are the probability of success in two groups respectively.

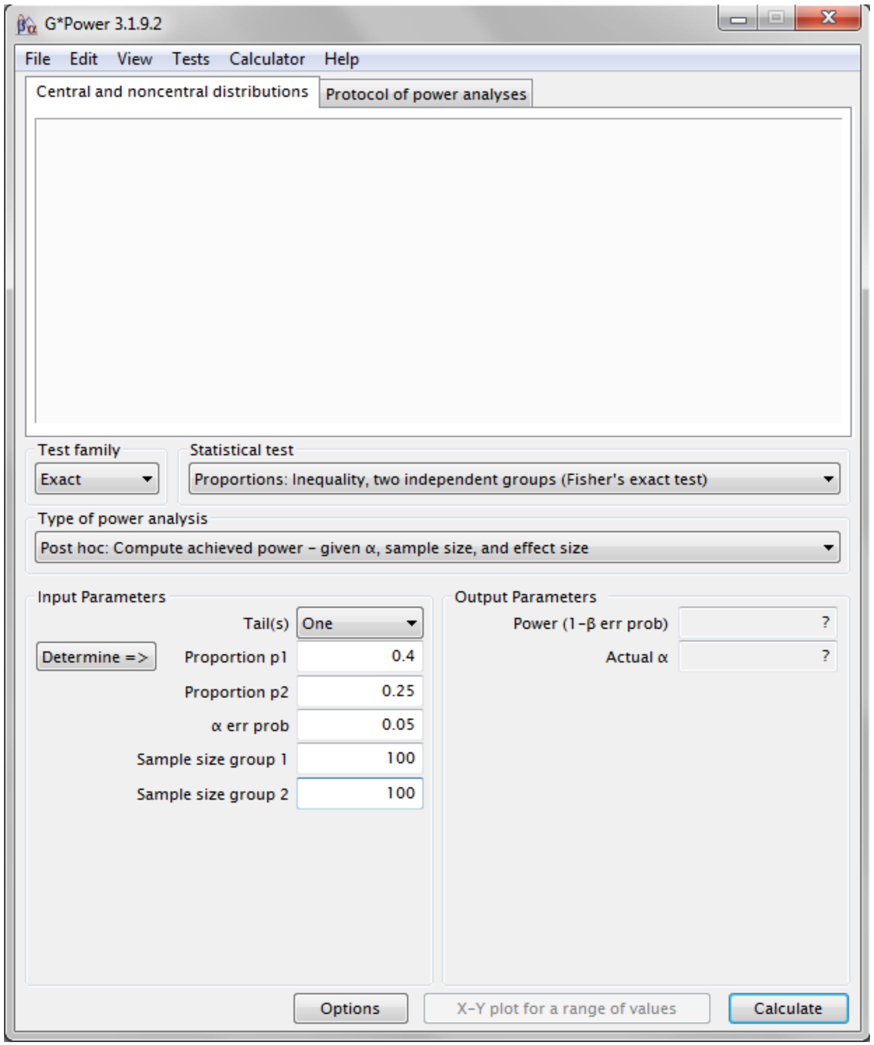
Comparison of proportions between two independent groups

- **Example**: We want to compare the occurrence of a plant disease between a control group and a group treated with Spray Z, an anti-fungal plant spray. Our outcome is binary: 0 if the plant doesn't get the disease, and 1 if it does. These plants are rare, and Spray Z is expensive. Your department will only let you order 200 plants. Before you do an experiment, you want to make sure you can achieve a satisfactory level of statistical power. This species of plant has a 40% disease rate when left untreated, and Spray Z has been shown to lower the occurrence of the disease to 15-20%.

Comparison of proportions between two independent groups

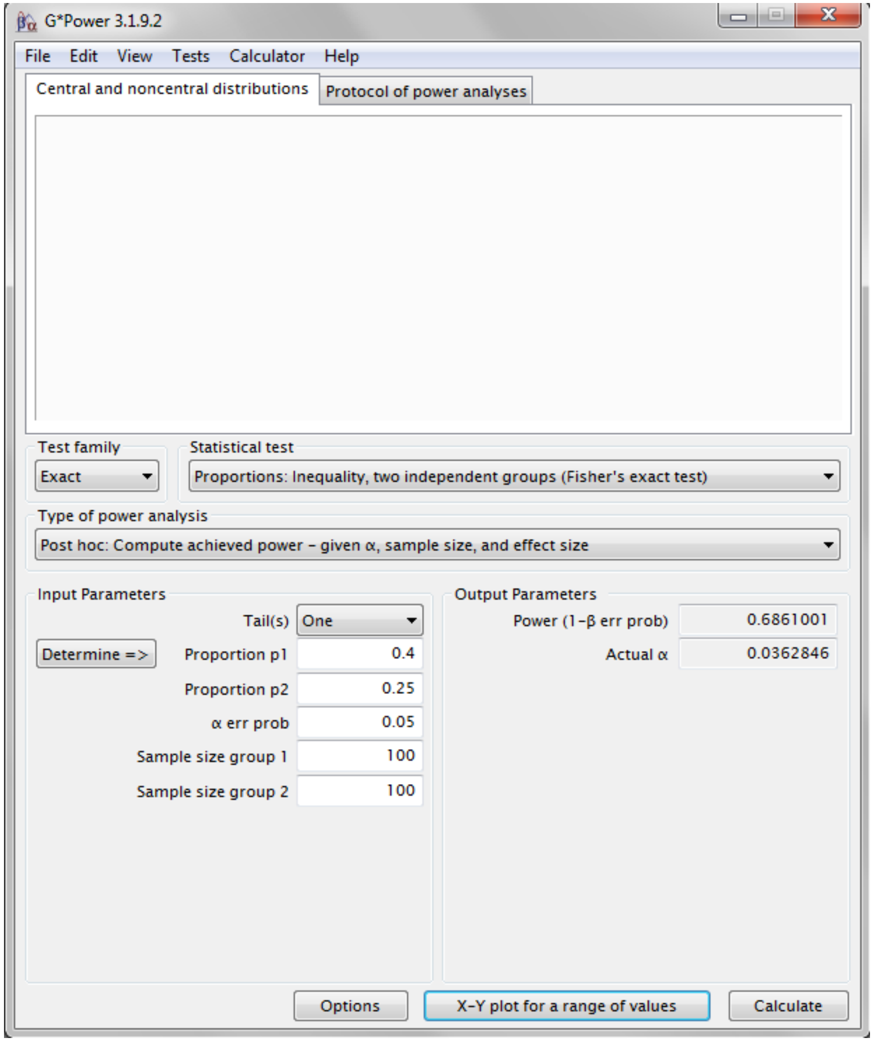
- What information do we need?
  - Type of analysis → two-sample comparison of proportions
  - Sample size → 200 total (100 per group)
  - Effect size → P1=0.40, P2=0.15-0.20
  - Type I error rate (alpha) → fixed at 0.05
  - Power → Unknown

Comparison of proportions between two independent groups



Post hoc power analysis

Comparison of proportions between two independent groups



ANOVA – Compare Two or More Group Means

- The null hypothesis is that all k means are identical $H0$: μ_1=$\mu_2 = \cdots = \mu_k$. The alternative hypothesis states that at least two of the k means differ. $H0$: $\mu_i \neq \mu_j$, for at least one pair i, j with $1 \leq i, j \leq k$.
- Effect size index $f = \sigma_m / \sigma$, where σ_m is the standard deviation of the group means μ_i and σ the common standard deviation within each of the k groups.
  - Small f = 0.10
  - Medium f = 0.25
  - Large f = 0.40

ANOVA – Compare Two or More Group Means

- We want to compare common dieting strategies to see which best helps adults lose weight. Participants will be randomly assigned to one of four groups
  - Low carb diet
  - Low fat diet
  - Low calorie diet
  - Control (placebo effect of being in a weight loss study)
- Outcome of interest: weight loss (baseline weight-final weight)

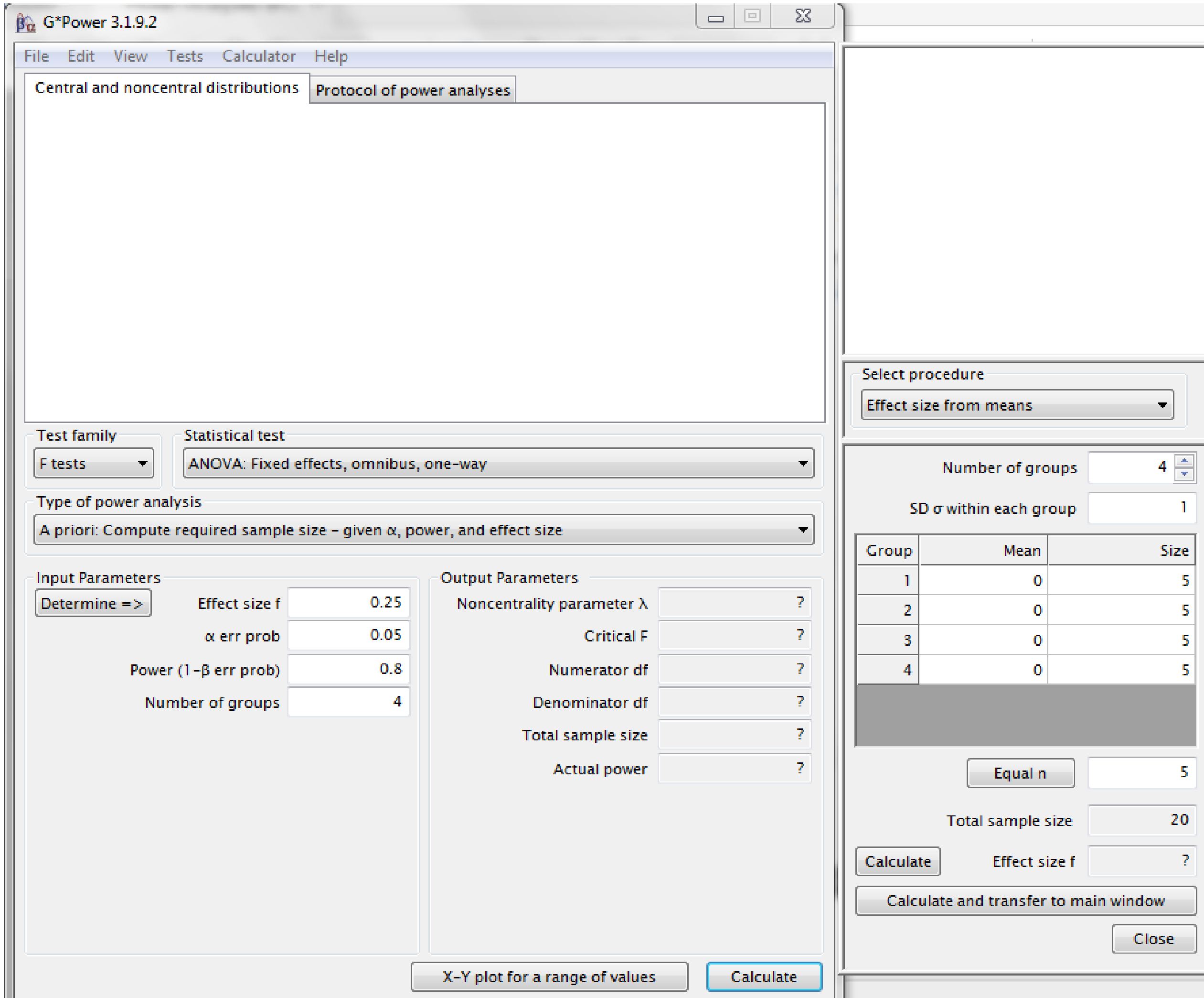
ANOVA – Compare Two or More Group Means

- Let's pretend this has never been studied…
  - We have no idea how large of an effect each of these diets will have on weight loss. What can we do?
  - We have conventions for small, medium, and large effects
  - How will each of these effect sizes change required sample size?
  - Let's aim for 80% power to detect at least one difference in weight loss among the 4 groups

ANOVA – Compare Two or More Group Means

- What information do we need?
  - Type of analysis → ANOVA with 4 groups
  - Sample size → Unknown
  - Effect size → Unknown
  - Type I error rate (alpha) → fixed at 0.05
  - Power → 80%

ANOVA – Compare Two or More Group Means



Correlation

- The null hypothesis is that in the population the true correlation ρ between two bivariate normally distributed random variables has the fixed value ρ_0. The (two-sided) alternative hypothesis is that the correlation coefficient has a different value: $\rho \neq \rho_0$.
- The proper effect size is the difference between ρ and ρ_0: $\rho - \rho_0$. For the special case ρ_0 = 0, effect size conventions was defined
  - For the special case r0 = 0, Cohen (1969, p.76) defines the following effect size conventions: small ρ = 0.1, medium ρ = 0.3 and large ρ = 0.5

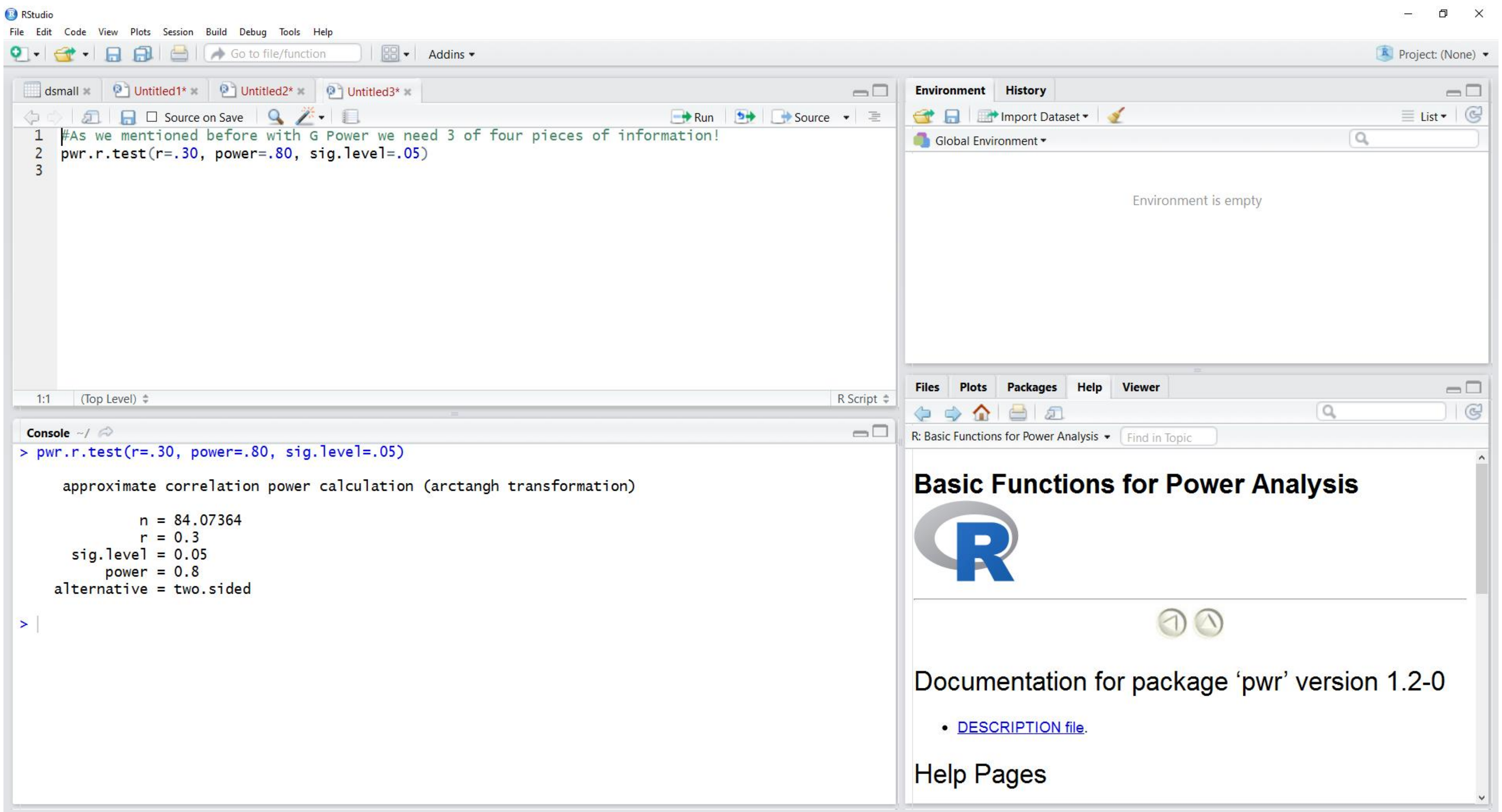
Correlation

- Let's say we are studying the potential relationship between systolic blood pressure and several other variables. One relationship we are particularly interested in exploring is between systolic blood pressure and body mass index (BMI). We hypothesize that these two variables have a positive correlation, but we need to know how many participants we will need in order to detect the correlation. We want to be able to detect a correlation of $r = 0.3$ or greater when our null hypothesis is that the correlation between BMI and blood pressure is zero.

Correlation

- What information do we need?
- Type of analysis → Correlation: Bivariate normal model
- Sample size → Unknown
- Effect size → r = 0.3
- Type I error rate (alpha) → fixed at 0.05
- Power → 80%

Using R for correlation power analysis



UCR GradSuccess
graduate.ucr.edu/success

Multiple Regression

- There are a couple ways of writing the null hypothesis in multiple regression
- We can say that all coefficients equal zero

That is: H_0: $\beta_1 = \beta_2 = \ldots \beta_k = 0$

Where k = Number of predictors

- The alternative is lengthy and clumsy to right out, but it basically says that AT LEAST one coefficient is not zero.

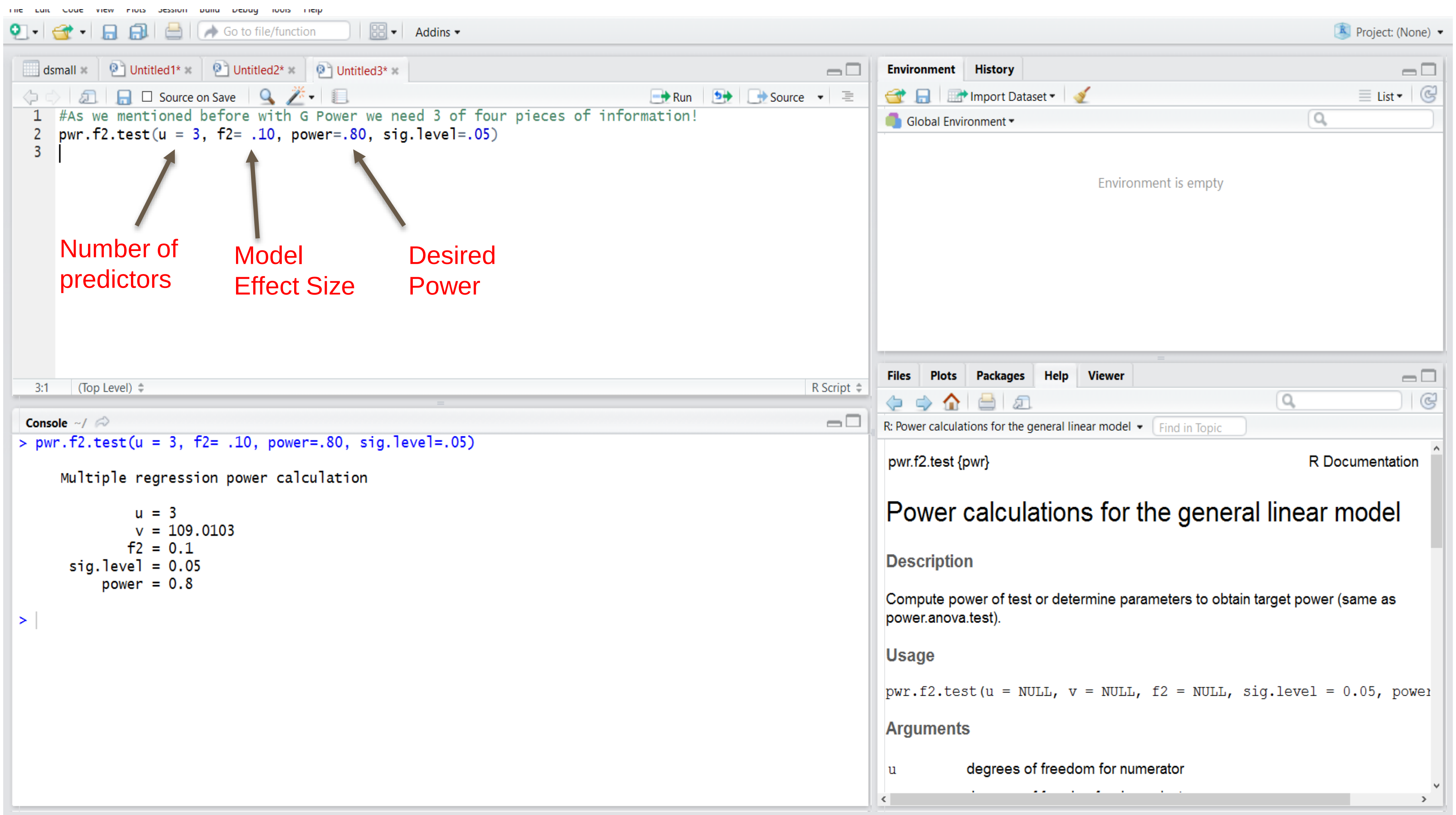
Multiple Regression

- We can also think of the null in terms of the combined effect of all variables in the model
- That is: H_0: R^2 = .00
- Where our predictors are accounting for 0% of the variation in our predictor
- The alternative would be H_1 : $R^2 \neq$ .00
- Where the variance accounted for in the model is not zero

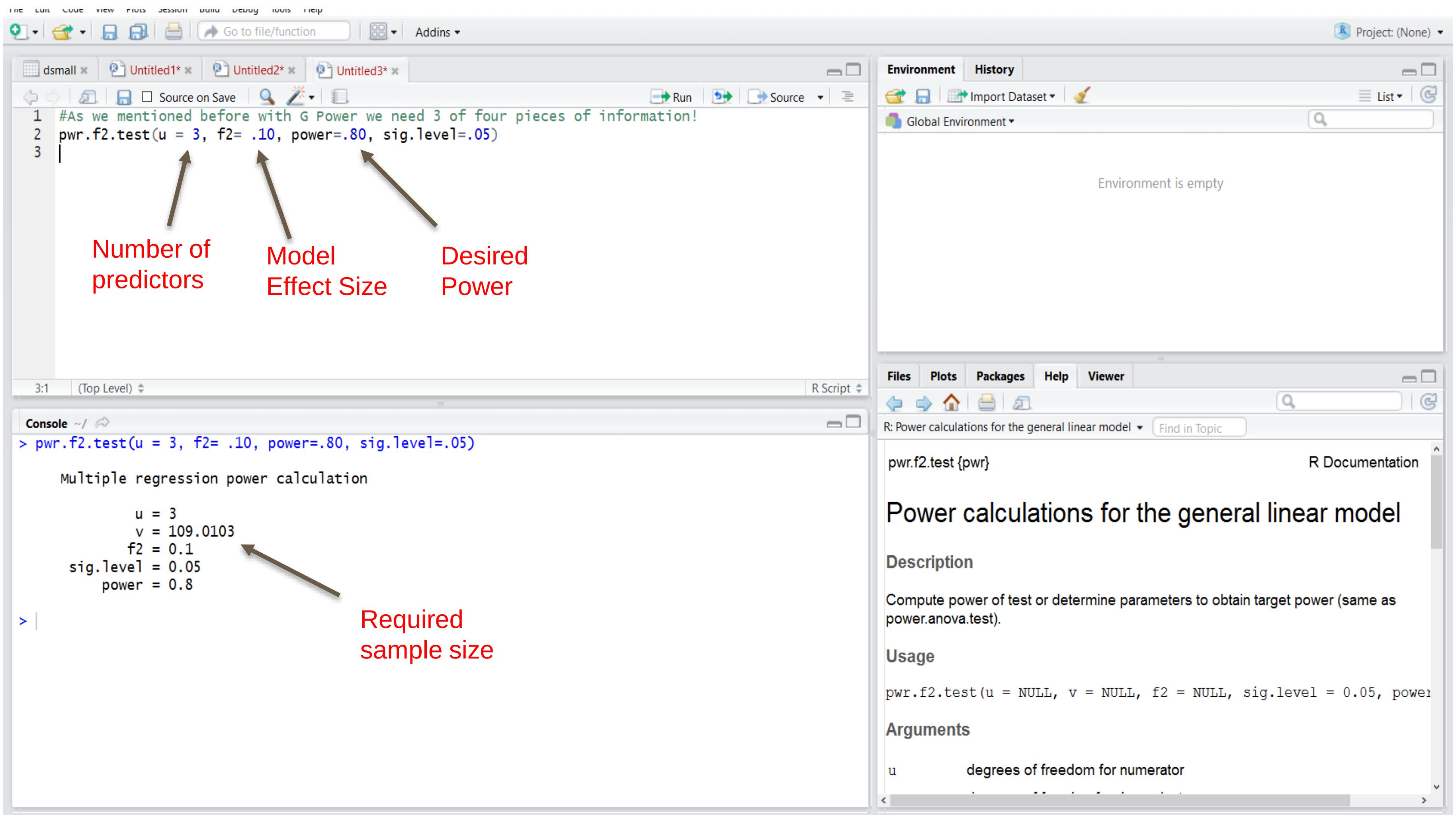
Multiple Regression Example

- Let's say that we want to predict how many apples are going to grow on a tree. To predict the number of apples, we are using the average temperature, amount of fertilizer, and the number of pesticides used. We think that these three variables will account for 10% of the variation in apple number. To get a power equal to .80, how many trees do we need?

Multiple Regression power using R!



Multiple Regression Power using R!



Power: A Cautionary Tale

- Remember, the concept of power still relies on rejecting the null!
- This reliance on dichotomous decision making with research
- We can use to try and figure out if anything fishy is going on
- For example…

How likely?

- Researcher decide to conduct six studies (replication, right?)
- They found a correlation of $r = .30$ pretty much every time
- The *N's* for the six studies are: 31, 50, 40, 50, 118, and 50
- What is the probability they would have found a significant result *every time*?

“Compound Power”

- Power for each individual study was:
- .38, .57, .48, .57, .91, and .57.
- Only the fifth study seems to be adequately powered, but what are the odds altogether?
- .38*.57*.48*.57*.91*.57 = .031
- The researchers only had a 3.1% chance of finding a significant result every time!
- This could be a hint that some unsavory practices were occurring

An Alternative look at Sample Size

- Remember the equation for a confidence interval (for a t-test):
- $LL = t - t_{critical}$ (S.E.)
- $UL = t + t_{critical}$ (S.E.)

- This interval will give us a range of values that has a 95% chance of containing the population value

More on Confidence Intervals

- $t_{critical}$ is fixed depending on our confidence level (typically 95%)

- Imagine if our standard error was zero

- The more observations we have, the lower our standard error and the more precise will be our parameters

Advantages of using CI's

- Does not rely on probability like power does (i.e. we know how wide the CI will be)
- May be how researchers think of sample size anyway
- Can discourage reliance on just p-values to make decisions
- Best to use both!

Thanks!