How To Choose The Right Statistic Analysis Method

Yao Zhu

Descriptive Statistics

Let’s Start From the Basic

Different types of data:

- **Continuous**
  -Height, Weight, etc
- **Discrete**
  1. Binomial distribution
     -Infected vs. Non-infected, etc.
  2. Categorical
     -Nominal: Race, Color, etc
     -Ordinal: “On a scale from 1 to 5, how bad is your injury”, etc.
- **Time-event**
  -focusing on Time and Event at the same time
- **Count**
  - “How many times have you been infected with this disease”, etc

Continuous

- ✓ Measures of Central Tendency
  - Mean
  - Median
  - Mode
  - Percentile
  - Range
- ✓ Measures of Differences
  - Variance
  - Standard Deviation
- ✓ Measures of Distribution
  - Symmetry
  - Skewness
  - Kurtosis
  - Usually use boxplot or histogram

Discrete

- Frequency, proportion, percent
- table, pie charts, and bar charts

Statistical Analysis

Before you start your analysis...

1. What do you want to know
   - -Correlation?
   - -Discrepancy?
   - -Independency?

2. What is your data type?
   - -Continuous, discrete, etc.

3. How is your data distributed?
   - -Normal distribution? Non-normal distribution?

4. How many variables do you have?

5. Is there a Dependent variable or an Independent variable?

Here's an example

Assume we have 1000 gene expression samples from normal population, and 1000 gene expression samples from patients who were infected with an unknown disease.

Question: Is there a significant difference between their gene expressions?
What statistical analysis method do we use?

Do we use Two-sample t-test?

Ask yourself these questions first...

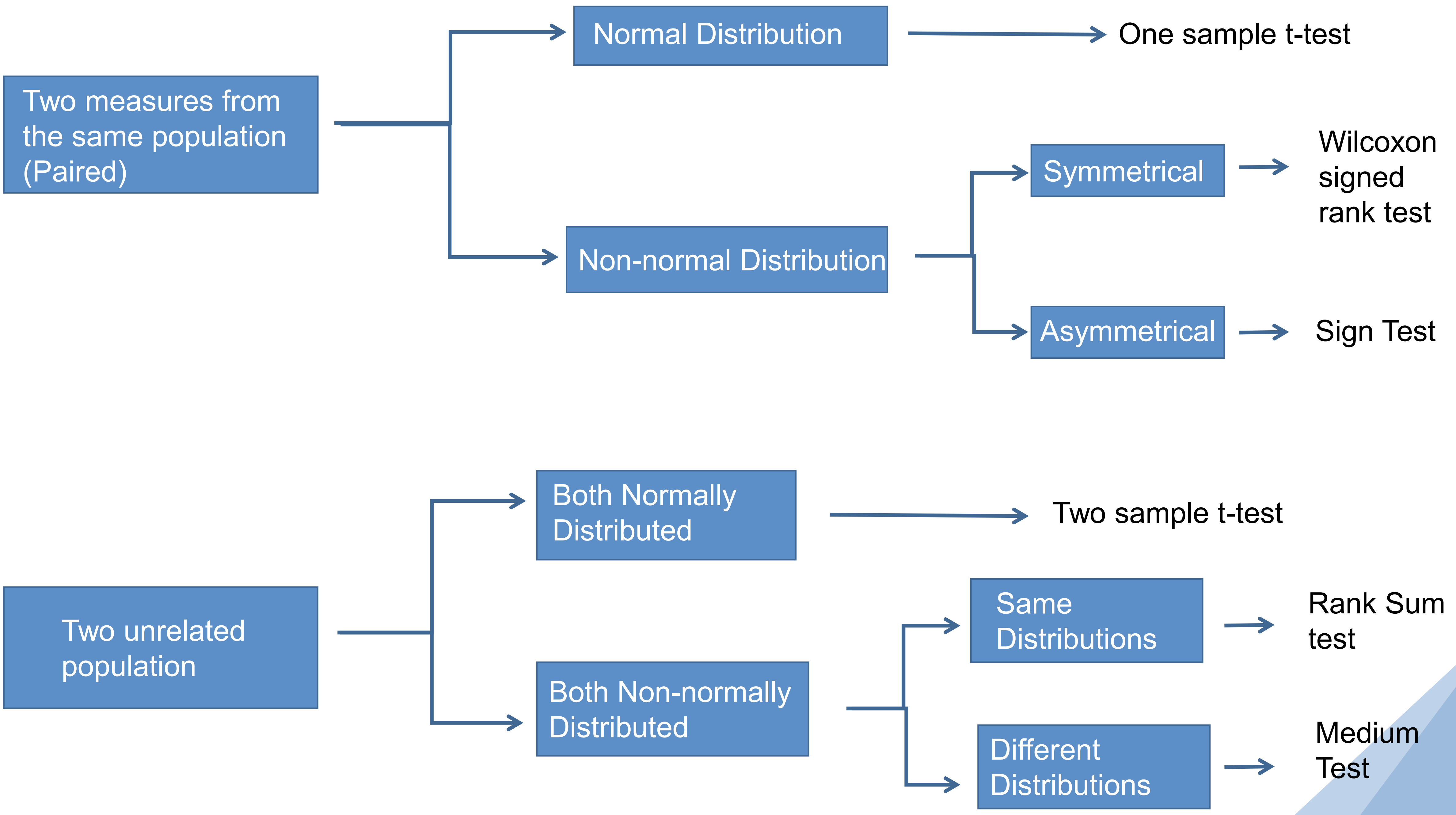
1. Is there any connections between “Normal population” and “Patient population”?
   - Gene expression from the same person? (Measured twice from the same population)
   - Are they related? Family members?

   (If yes, how do we deal with the data? And if not?)

2. How’s the distribution of the data?
   - Normally distributed? Non-normally distributed?

   (How do you deal with them?)

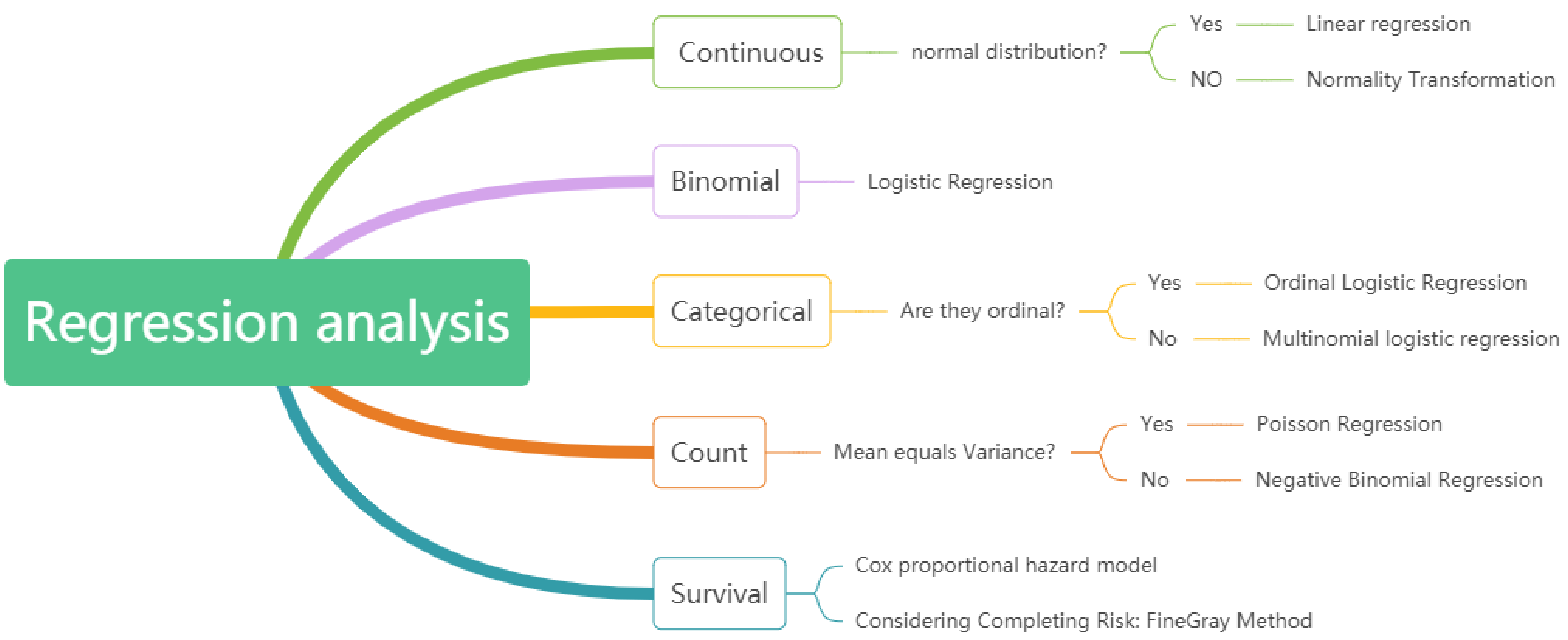
3. Is there only two variables in this question?



Two measures from the same population (Paired)
Normal Distribution
One sample t-test
Non-normal Distribution
Symmetrical
Wilcoxon signed rank test
Asymmetrical
Sign Test
Two unrelated population
Both Normally Distributed
Two sample t-test
Both Non-normally Distributed
Same Distributions
Rank Sum test
Different Distributions
Medium Test

1 \ 2	Continuous Normal distribution	Continuous Non-normal distribution	Binomial distribution	Nominal distribution	Ordinal distribution	Any type
Continuous Normal distribution	Pearson Correlation	Spearman correlation	t-test	Analysis of variance	Linear Regression	Linear Regression
Continuous Non-normal distribution		Spearman correlation	Wilcoxon rank test	Kruskal-Wallis test	Jonckheere-Terpstra test	Linear Regression
Binomial distribution			Chi-squared test	Chi-squared test	Cochran-Mantel-Haenszel test	Logistic Regression
Nominal distribution				Chi-squared test	Cochran-Mantel-Haenszel test	Multinominal Logistic Regression
Ordinal distribution					Cochran-Mantel-Haenszel test	Ordinal Logistic Regression
Time-event data						Kaplan-Meier Curve, Log-rank test

If there're multiple variables...



Sample Size

What should we know about sample size calculation

- Things that will effect sample size:

  1. Significant level. (We usually pick 0.05)
  2. Power. (power = 1-β)
  3. Effect size

- Common effect size: Risk Ratio(RR), Odds Ratio(OR), Hazard Ratio(HR)

- What if we do not have any of those, and still needs an effect size?
  - 1. Use effect size from previous related research (such as meta analysis)
  - 2. Perform a small sample size pre-research/pre-experiment to determine your effect size
  - 3. In 1988, Cohen mentioned in one of his research that effect size can be 0.2, 0.5, or 0.8 when you do not know the effect size for your experiment. 0.2, 0.5, and 0.8 represents small, medium, and large for your effect size. (Usually we choose 0.5)

For t-tests, the effect size is assessed as

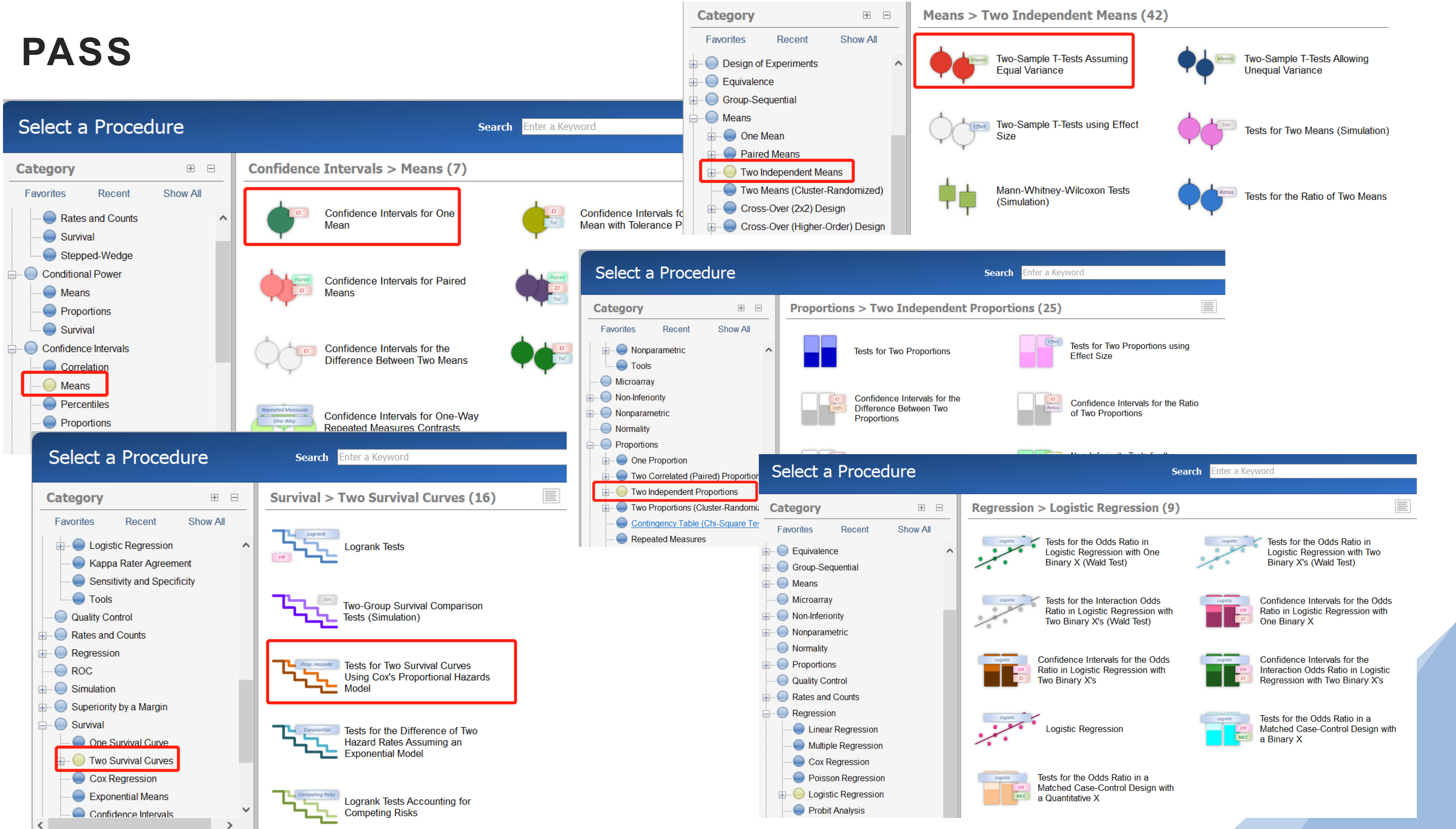
$$d = \frac{|\mu_1 - \mu_2|}{\sigma}$$

where μ_1 = mean of group 1
μ_2 = mean of group 2
σ^2 = common error variance

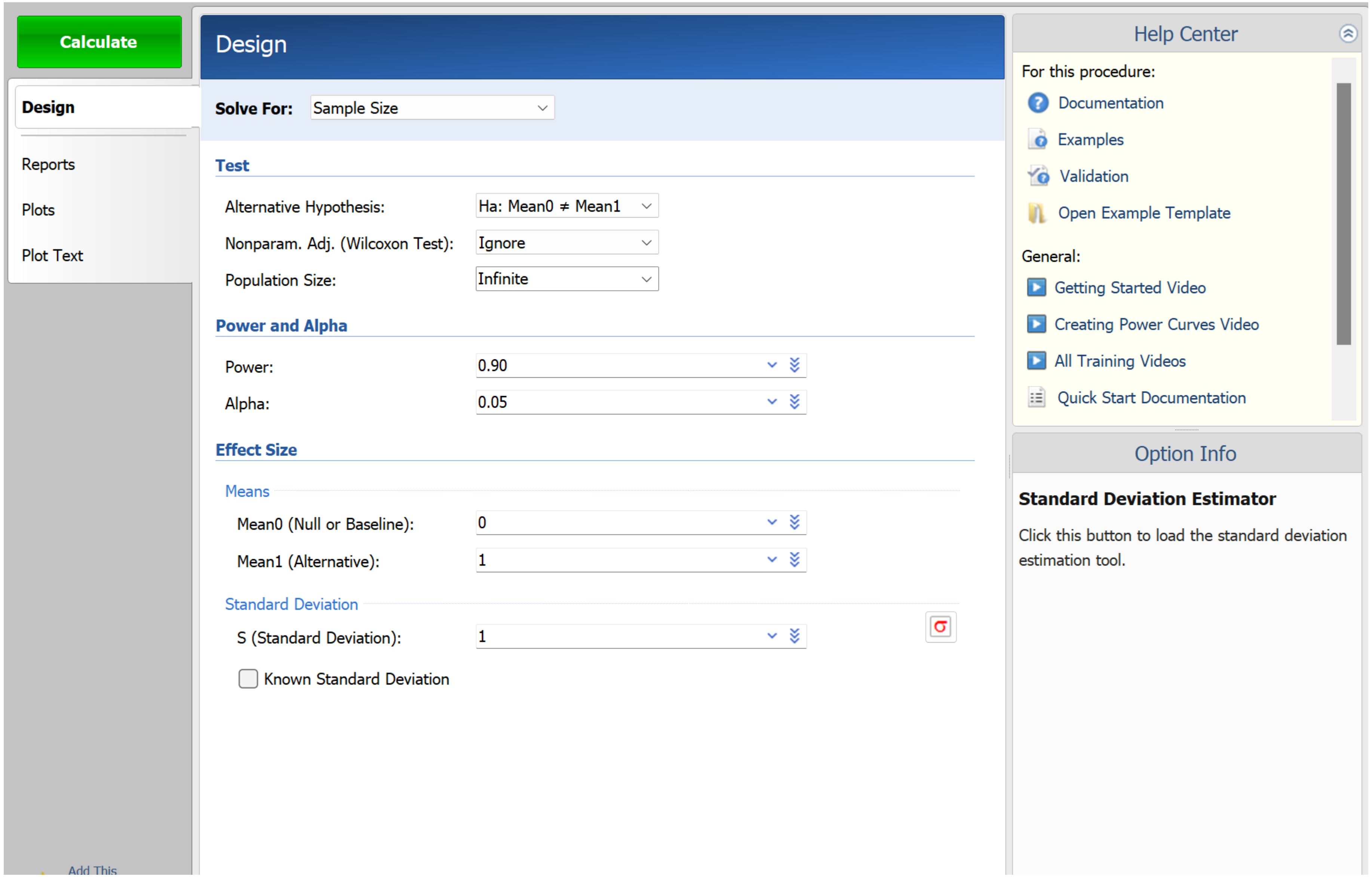
Cohen suggests that d values of 0.2, 0.5, and 0.8 represent small, medium, and large effect sizes respectively.

Sample size was estimated using data from previous tDCS studies,[17,18] antidepressant and rTMS meta-analyses,[5,38,39] and rTMS studies in which antidepressant drugs were combined.[40,41] With these data, we estimated a 3-point difference effect (effect size of Cohen *d* = 0.5) for both tDCS only and sertraline only vs placebo and a combined additive effect in the combined treatment group (ie, 6-point difference, with an effect size of Cohen *d* = 1.0), which, considering probabilities of 5% for type I error and 20% for type II error, resulted in an estimated sample size of 30 patients per arm for a total of 120 participants (for an extensive discussion regarding our power analysis, see the articles by Brunoni et al[29,42]). Further, we considered a difference smaller than an effect size of 0.5 or a 3-point between-group difference not to be clinically relevant per the National Institute for Clinical Excellence guidelines.[43]

PASS



Select a Procedure
Search
Enter a Keyword
Category
Favorites
Recent
Show All
Rates and Counts
Survival
Stepped-Wedge
Conditional Power
Means
Proportions
Survival
Confidence Intervals
Correlation
Means
Percentiles
Proportions
Confidence Intervals > Means (7)
Confidence Intervals for One Mean
Confidence Intervals for Paired Means
Confidence Intervals for the Difference Between Two Means
Confidence Intervals for One-Way Repeated Measures Contrasts
Category
Favorites
Recent
Show All
Design of Experiments
Equivalence
Group-Sequential
Means
One Mean
Paired Means
Two Independent Means
Two Means (Cluster-Randomized)
Cross-Over (2x2) Design
Cross-Over (Higher-Order) Design
Means > Two Independent Means (42)
Two-Sample T-Tests Assuming Equal Variance
Two-Sample T-Tests Allowing Unequal Variance
Two-Sample T-Tests using Effect Size
Tests for Two Means (Simulation)
Mann-Whitney-Wilcoxon Tests (Simulation)
Tests for the Ratio of Two Means
Select a Procedure
Search
Enter a Keyword
Category
Favorites
Recent
Show All
Nonparametric
Tools
Microarray
Non-Inferiority
Nonparametric
Normality
Proportions
One Proportion
Two Correlated (Paired) Proportion
Two Independent Proportions
Contingency Table (Chi-Square Tes
Repeated Measures
Proportions > Two Independent Proportions (25)
Tests for Two Proportions
Tests for Two Proportions using Effect Size
Confidence Intervals for the Difference Between Two Proportions
Confidence Intervals for the Ratio of Two Proportions
Select a Procedure
Search
Enter a Keyword
Category
Favorites
Recent
Show All
Logistic Regression
Kappa Rater Agreement
Sensitivity and Specificity
Tools
Quality Control
Rates and Counts
Regression
ROC
Simulation
Superiority by a Margin
Survival
One Survival Curve
Two Survival Curves
Cox Regression
Exponential Means
Confidence Intervals
Survival > Two Survival Curves (16)
Logrank Tests
Two-Group Survival Comparison Tests (Simulation)
Tests for Two Survival Curves Using Cox's Proportional Hazards Model
Tests for the Difference of Two Hazard Rates Assuming an Exponential Model
Logrank Tests Accounting for Competing Risks
Select a Procedure
Search
Enter a Keyword
Category
Favorites
Recent
Show All
Equivalence
Group-Sequential
Means
Microarray
Non-Inferiority
Nonparametric
Normality
Proportions
Quality Control
Rates and Counts
Regression
Linear Regression
Multiple Regression
Cox Regression
Poisson Regression
Logistic Regression
Probit Analysis
Regression > Logistic Regression (9)
Tests for the Odds Ratio in Logistic Regression with One Binary X (Wald Test)
Tests for the Odds Ratio in Logistic Regression with Two Binary X's (Wald Test)
Tests for the Interaction Odds Ratio in Logistic Regression with Two Binary X's (Wald Test)
Confidence Intervals for the Odds Ratio in Logistic Regression with One Binary X
Confidence Intervals for the Odds Ratio in Logistic Regression with Two Binary X's
Confidence Intervals for the Interaction Odds Ratio in Logistic Regression with Two Binary X's
Logistic Regression
Tests for the Odds Ratio in a Matched Case-Control Design with a Binary X
Tests for the Odds Ratio in a Matched Case-Control Design with a Quantitative X


Calculate
Design
Reports
Plots
Plot Text
Add This
Design
Solve For: Sample Size
Test
Alternative Hypothesis: Ha: Mean0 ≠ Mean1
Nonparam. Adj. (Wilcoxon Test): Ignore
Population Size: Infinite
Power and Alpha
Power: 0.90
Alpha: 0.05
Effect Size
Means
Mean0 (Null or Baseline): 0
Mean1 (Alternative): 1
Standard Deviation
S (Standard Deviation): 1
Known Standard Deviation
Help Center
For this procedure:
Documentation
Examples
Validation
Open Example Template
General:
Getting Started Video
Creating Power Curves Video
All Training Videos
Quick Start Documentation
Option Info
Standard Deviation Estimator
Click this button to load the standard deviation estimation tool.

Navigat

- Tests for One Mean
  - Numeric Results
  - References
  - Report Definitions
  - Summary Stateme
  - Dropout-Inflated S
  - Procedure Input S

Tests for One Mean

Numeric Results for One-Sample T-Test

Null Hypothesis: Mean0 = Mean1 Alternative Hypothesis: Mean0 ≠ Mean1
Unknown standard deviation.

Power	N	Alpha	Beta	Mean0	Mean1	S	Effect Size
0.91071	13	0.05000	0.08929	0.0	1.0	1.0	1.000

References

Machin, D., Campbell, M., Fayers, P., and Pinol, A. 1997. Sample Size Tables for Clinical Studies, 2nd Edition. Blackwell Science. Malden, MA.

Zar, Jerrold H. 1984. Biostatistical Analysis (Second Edition). Prentice-Hall. Englewood Cliffs, New Jersey.

Report Definitions

Power is the probability of rejecting a false null hypothesis. It should be close to one.
N is the size of the sample drawn from the population. To conserve resources, it should be small.
Alpha is the probability of rejecting a true null hypothesis. It should be small.
Beta is the probability of accepting a false null hypothesis. It should be small.
Mean0 is the value of the population mean under the null hypothesis. It is arbitrary.
Mean1 is the value of the population mean under the alternative hypothesis. It is relative to Mean0.
Sigma is the standard deviation of the population. It measures the variability in the population.
Effect Size, |Mean0-Mean1|/Sigma, is the relative magnitude of the effect under the alternative.

Summary Statements

A sample size of 13 achieves 91% power to detect a difference of -1.0 between the null hypothesis mean of 0.0 and the alternative hypothesis mean of 1.0 with an estimated standard deviation of 1.0 and with a significance level (alpha) of 0.05000 using a two-sided one-sample t-test.

Thank you