BCBB Workshop

Introduction to Prism 8

For Visualization and Statistics

Qinlu (Claire) Wang & Jingwen Gu

Bioinformatics and Computational Biosciences Branch (BCBB)
Office of Cyber Infrastructure and Computational Biology (OCICB)
National Institute of Allergy and Infectious Diseases (NIAID)

National Institute of Allergy and Infectious Diseases

NIAID

OCICB Bioinformatics and Computational Biosciences Branch (BCBB)

- Part of NIAID
- Group of ~50
- Software developers
- Computational Biologists
- Project Management & Analysis Professionals
- Biostatistics, Phylogenetics, Genomics, Structural Biology, Proteomics, Programming

National Institute of Allergy and Infectious Diseases

NIAID

How to contact us?

1. Submit Request

   Send emails to bioinformatics@niaid.nih.gov OR fill a request online at Online Request

2. Attend our workshops

   http://www.eventzilla.net/user/NIAID_OCICB_BCBB

3. Tell us what statistical topics you want to learn about

   BCBB Statistical Training – Suggest a class!

4. Join our Slack group!

   The invitation will be sent out after this workshop

What is Prism 8?

GraphPad Prism is a versatile statistics tool for **scientists**.

1. *Structured Data Tables for scientific research*
2. *Statistical analyses and explanation*
3. *Countless ways to customize your graphs*
4. *Gain insights and guidance at every step*
5. *Collaborate with colleagues and share your research with the world.*

Prism Guides

Prism provides three guides:

- *Prism User Guide*
- *Statistics Guide*
- *Curve Fitting (Regression) Guide*

Prism Labs

Lab 0: How to install Prism 8 on Mac

Lab 1: Create and edit data tables

Lab 2: Visualization

Lab 3: Descriptive Statistics

Lab 4: Statistical Testing

Lab 5: Survival Analysis

Lab 6: Categorical Data Analysis

Lab 7: Regression and Curve Fitting

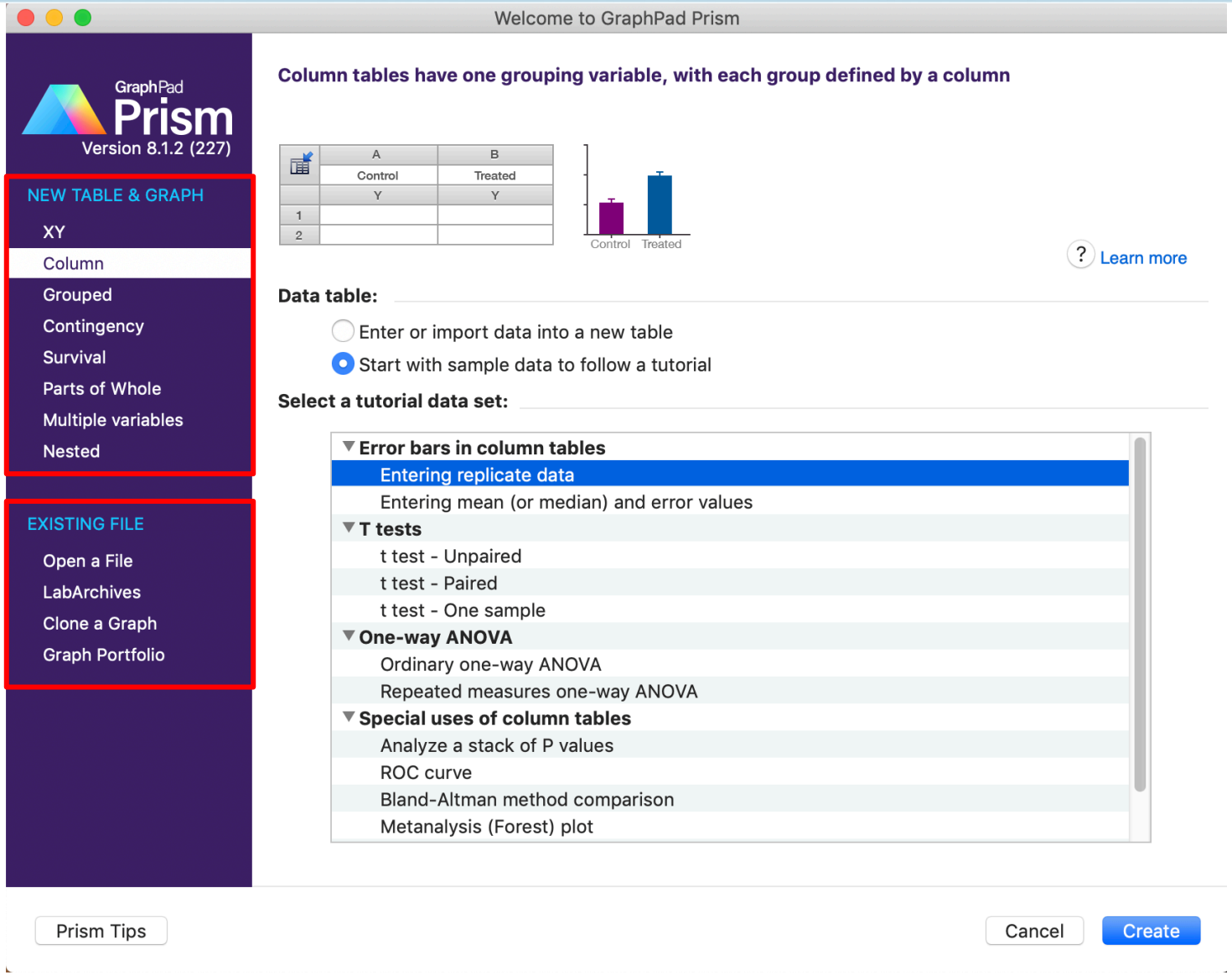
Lab 0 – How to install Prism?

- If install Prism 8 on NIH desktops/laptops, install Prism 8 via *NIH Self Service*. How to Install Prism 8
- If install Prism 8 on personal desktops/laptops, download and install Prism 30-day free trial at Prism Free Trial

Lab 1 – Create and Edit Data Tables

1st Part: 8 Types of Data Tables

2nd Part: Read Existed Files



National Institute of Allergy and Infectious Diseases

NIAID

1st Part: 8 Types of Data Tables

Welcome to GraphPad Prism

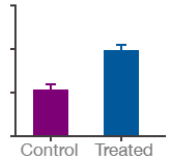
GraphPad Prism
Version 8.1.2 (227)

NEW TABLE & GRAPH
- XY
- Column
- Grouped
- Contingency
- Survival
- Parts of Whole
- Multiple variables
- Nested

EXISTING FILE
- Open a File
- LabArchives
- Clone a Graph
- Graph Portfolio

Column tables have one grouping variable, with each group defined by a column

	A	B
	Control	Treated
	Y	Y
1		
2		



Learn more

Data table:

- ○ Enter or import data into a new table
- ◉ Start with sample data to follow a tutorial

Select a tutorial data set:

- **Error bars in column tables**
  - Entering replicate data
  - Entering mean (or median) and error values
- **T tests**
  - t test - Unpaired
  - t test - Paired
  - t test - One sample
- **One-way ANOVA**
  - Ordinary one-way ANOVA
  - Repeated measures one-way ANOVA
- **Special uses of column tables**
  - Analyze a stack of P values
  - ROC curve
  - Bland-Altman method comparison
  - Metanalysis (Forest) plot

Prism Tips

Cancel | Create

NIH National Institute of Allergy and Infectious Diseases

NIAID

Eight Types of Data Tables

1. *XY Tables*
2. *Column Tables*
3. *Grouped Tables*
4. *Contingency Tables*
5. *Survival Tables*
6. *Parts of Whole Tables*
7. *Multiple Variable Tables*
8. *Nested Tables*

1.XY Tables

An XY table is a graph where every point is defined by both an X and a Y value. This kind of data are often fit with linear or nonlinear regression.

Analyses performed with XY data

- *Nonlinear regression (curve fit)*
- *Linear regression*
- *Area Under Curve*
- *Deming (Model II) Linear regression*
- *Correlation*
- *……*

Table format: XY	X	Group A			Group B		
	Seconds	Control			Treated		
	X	A:Y1	A:Y2	A:Y3	B:Y1	B:Y2	B:Y3
1 Title	1	35	31	42	36	39	25
2 Title	2	43	49				
3 Title	3	50	57	67	87	89	62
4 Title	4						
5 Title	5	77	89	99	102	145	154
6 Title	10	145		121			
7 Title	15				254	269	231
8 Title	20	167	187	145			
9 Title	25				289	296	271
10 Title							

2.Column Tables

Use column tables if your groups are defined by one scheme, perhaps control vs. treated, or placebo vs. low-dose vs. high-dose. Each column defines one group.

Analyses performed with Column data

- *Unpaired / Paired t-test*
- *Mann-Whitney / Kolmogorov-Smirnov test / Wilcoxon test*
- *One-way ANOVA / Kruskal-Wallis test / Friedman test*
- *Descriptive Statistics*
- *Normality and Lognormality Tests*
- *Frequency Distribution*
- *ROC Curve*
- *Identify Outliers*
- *……*

	Group A	Group B
	Placebo	Active Drug
	Y	Y
1	45	67
2	23	46
3	56	113
4	76	79
5	81	123
6	87	
7	99	

3.Grouped Tables

The idea of two-way variables is best understood by example. One grouping variable (male vs. female in the example below) is defined by rows; the other grouping variable (control vs. treated) is defined by columns.

Table format: Grouped		Group A			Group B		
		Control			Treated		
		A:Y1	A:Y2	A:Y3	B:Y1	B:Y2	B:Y3
1	Men	34.5	32.9	43.3	87.5	*321.5**	81.7
2	Women	42.3		45.9	109.4	111.2	115.4

Analyses performed with Grouped data

- *Two-way ANOVA (and mixed model)*
- *Three-way ANOVA (and mixed model)*
- *Row means with SD or SEM*
- *Multiple t tests - one per row*

4.Contingency Tables

Contingency tables are used to tabulate the actual number of subjects (or observations) that fall into the categories defined by the rows and columns of a table.

Table format: Contingency		Outcome A	Outcome B
		Graft Patient	Graft Obstructed
		Y	Y
1	Standard Treatment	45	5
2	Experimental Treatment	49	1

Analyses performed from a contingency table

- *Chi-square and Fisher's exact test (also computes odds ratios and relative risk)*
- *Fraction of total*

5.Survival Tables

Survival tables are used to enter information for each subject. Prism then computes percent survival at each time, plots a Kaplan-Meier survival plot, and compares survival with some tests.

Analyses performed with survival data

- *Kaplan-Meier*
- *Log-rank test*
- *Wilcoxon-Gehan test*

Table format: Survival		X	Group A	Group B	Group C
		Days	Control	Treatment A	Treatment B
		X	Y	Y	Y
1	CO	78	1		
2	NT	34	1		
3	RO	123	0		
4	LT	45	1		
5	RE	234	1		
6	AT	345	1		
7	ME	123	1		
8	NT	211	1		
9	WO	356	1		
10	RK	378	0		
11	PL	88		1	
12	AC	321		1	
13	EB	211		1	
14	OT	111		0	
15	RE	156		0	
16	AT	178		1	
17	ME	236		1	
18	NT	198		1	
19	XX	211		1	
20	XY	234		1	
21	BO	257			1
22	OW	322			1
23	HO	344			1
24	TO	365			1
25	YO	245			0
26	UT	299			1
27	OO	351			1

6.Parts of Whole Tables

A Parts of whole table is used when it makes sense to ask: What **fraction of the total** is each value? This table is often used to make a pie chart.

Analyses performed on parts of whole data

- *Fraction of total*
- *Chi-square goodness of fit (compare observed distribution with theoretical distribution)*

Table format: Parts of whole		A
		Number of Students
		Y
1	A	23
2	B	29
3	C	7
4	D	2
5	E	0

7.Multiple Variable Tables

A multiple variable table is arranged the same way most statistics programs organize data.

Table format: Multiple variables		Variable A	Variable B	Variable C	Variable D	Variable E	Variable F	Variable G	Variable H	Variable I	Variable J
		Glycosylated hemoglobin %	Total cholesterol	Glucose	HDL	Age in years	Male?	Height in inches	Weight in pounds	Waist in inches	Hip in inches
1	Title	4.309999943	203	82	56	46	0	62	121	29	38
2	Title	4.440000057	165	97	24	29	0	64	218	46	48
3	Title	4.639999866	228	92	37	58	0	61	256	49	57
4	Title	4.630000114	78	93	12	67	1	67	119	33	38
5	Title	7.719999790	249	90	28	64	1	68	183	44	41

Analyses performed on multiple variable data

- Correlation matrix
- Multiple linear regression
- Transform and select
- Identify outliers
- Descriptive statistics
- ……

8.Nested Tables

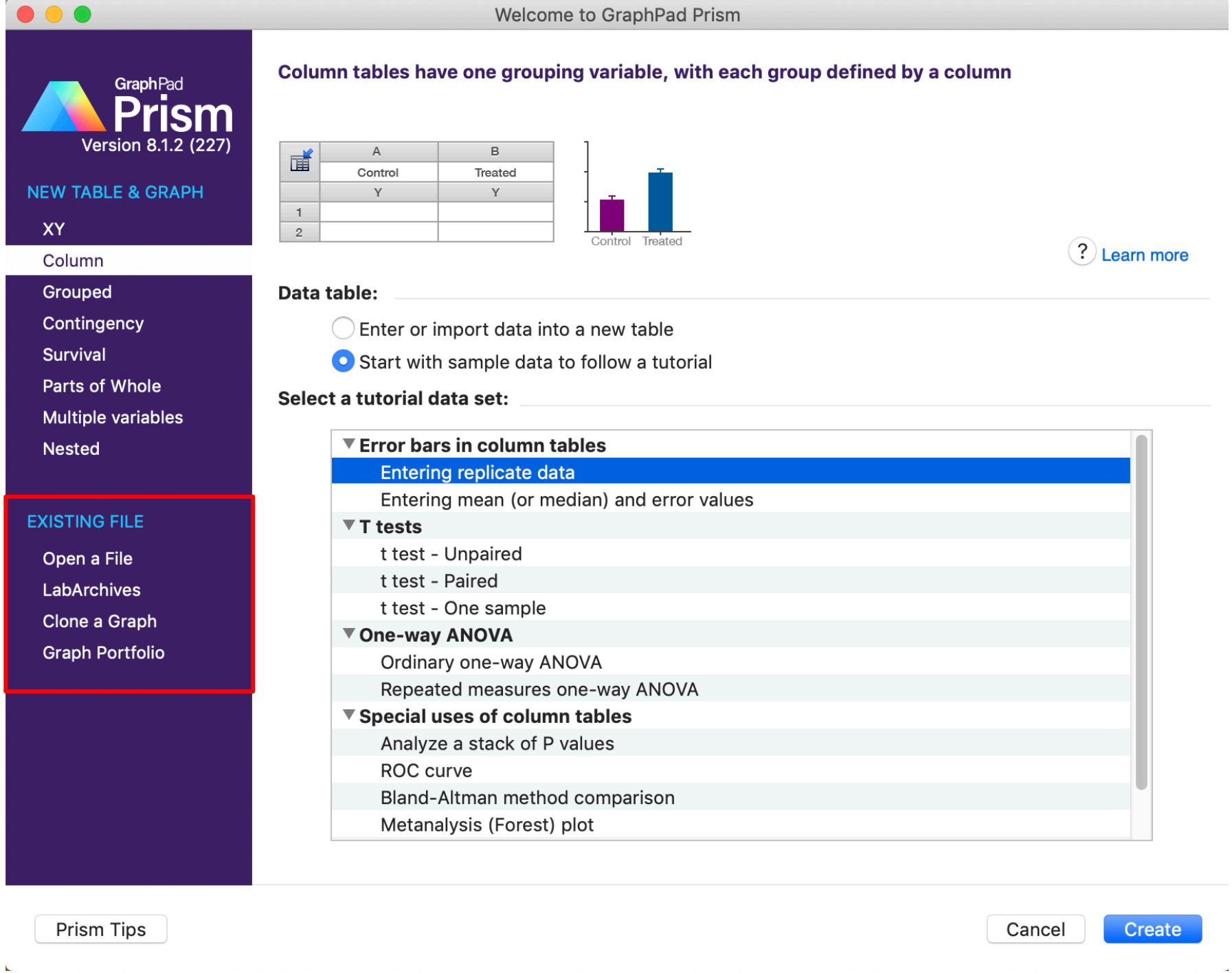
A nested table is used when there are two levels of nested or hierarchical replication.

Group A			Group B			Group C		
Traps + Pour on			Pour on			No vector control		
Herd 1	Herd 2	Herd 3	Herd 4	Herd 5	Herd 6	Herd 7	Herd 8	Herd 9
28	32	27	25	26	25	21	19	18
26	27	25	24	28	26	19	18	20
27	28	29	27	29	24	17	23	19
31	29	27	23	27	23	20	20	18

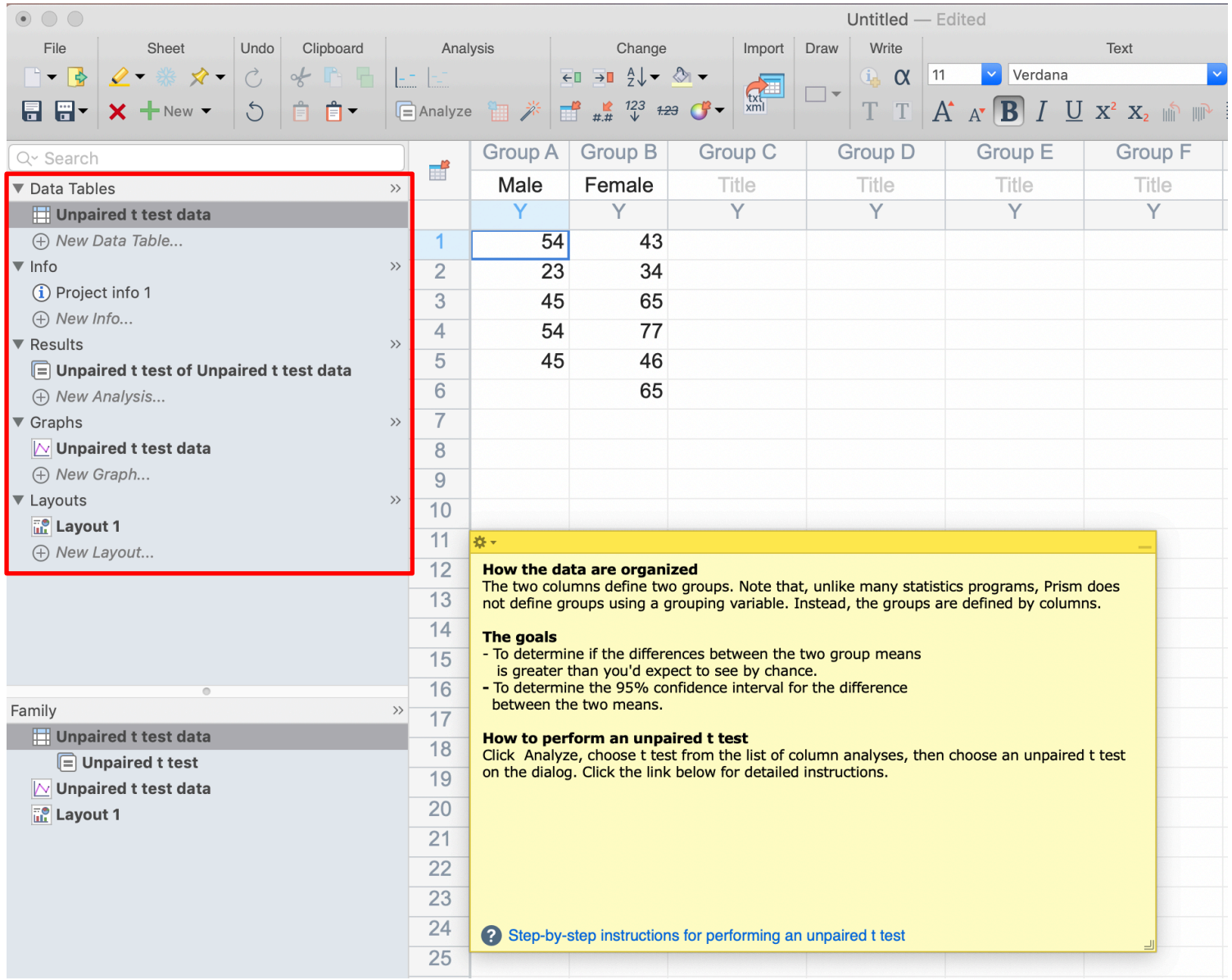
Analyses performed from nested tables

- *Nested t test*
- *Nested one-way ANOVA*
- *Descriptive statistics*
- *Normality and lognormality tests*
- *Outlier tests One-sample t test*

2nd Part: Read Existed Files



The 5 Sections of a Prism Project



National Institute of Allergy and Infectious Diseases

NIAID

Two Formats

When you save a Prism file, you save the entire project into one file. You have a choice of two file formats:

- **PZF format:** This is a binary format that can be opened by Prism 4 or later, but not by other applications.
- **PZFX format:** This is a format that can be opened only by Prism 5 or later. The first part of the file contains all the data tables and info sheets in a plain-text XML format that can be viewed by other programs. After that comes information about results, graphs and layouts in a format that is incomprehensible to any program but Prism.

Lab 2 – Visualization

Exporting images from Prism

Export Formats	PDF	TIFF / TIF
Advantages	File can be stretched to any size with no loss of quality	Tend to be trouble-free when submitted to journals
Disadvantages	1. Few Windows programs import pdf images 2. Few Journals accept them.	You need to choose resolution and size

Lab 3 – Descriptive Statistics

Value	Meaning
Minimum	The smallest value.
25th Percentile	25% of values are lower than this.
Median	Half the values are lower; half are higher.
75th Percentile	75% of values are lower than this.
Maximum	The largest value.
Mean	The average.
Standard Deviation	Quantifies variability or scatter.
Standard Error of Mean	Quantifies how precisely the mean is known.
95% Confidence Interval	Given some assumptions, there is a 95% chance that this range includes the true overall mean.
Coefficient of Variation	The standard deviation divided by the mean.
Geometric Mean	Compute the logarithm of all values, compute the mean of the logarithms, and then take the antilog of that mean. It is a better measure of central tendency when data follow a lognormal distribution (long tail).
Harmonic Mean	Compute the reciprocal of all values, compute the mean of the recoprocals, and then take the reciprocal of that mean.
Quadratic Mean	Compute the square of all values, compute the mean of the squares, and then take the square root of that mean.
Skewness	Quantifies how symmetrical the distribution is. A distribution that is symmetrical has a skewness of 0.
Kurtosis	Quantifies whether the tails of the data distribution matches the Gaussian distribution. A Gaussian distribution has a kurtosis of 0.

Analysis Checklist

NIAD

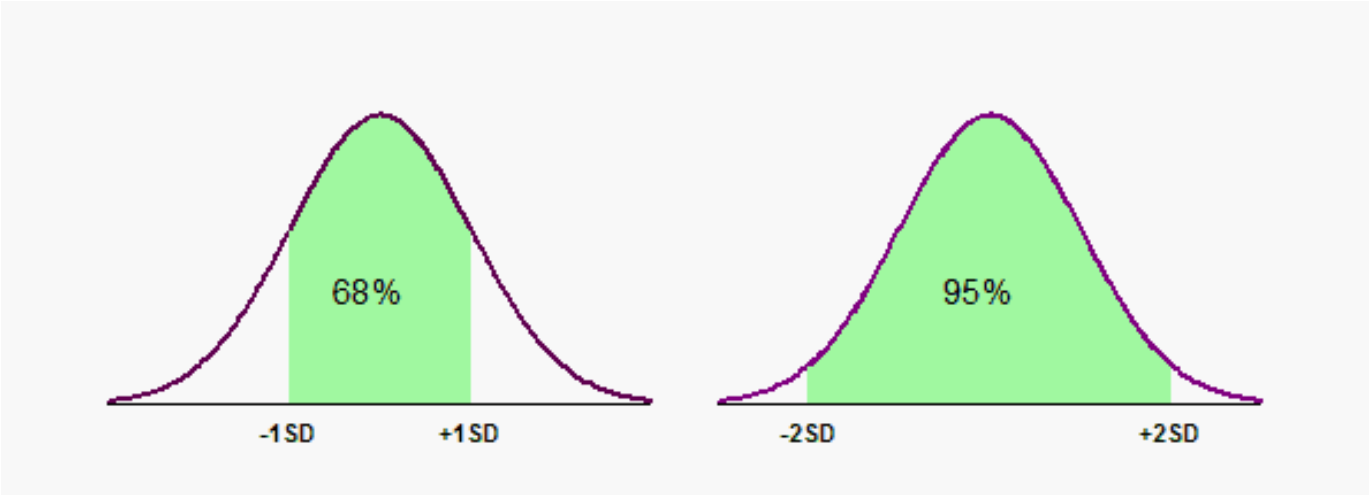
Standard Deviation (SD)

1. What is SD?

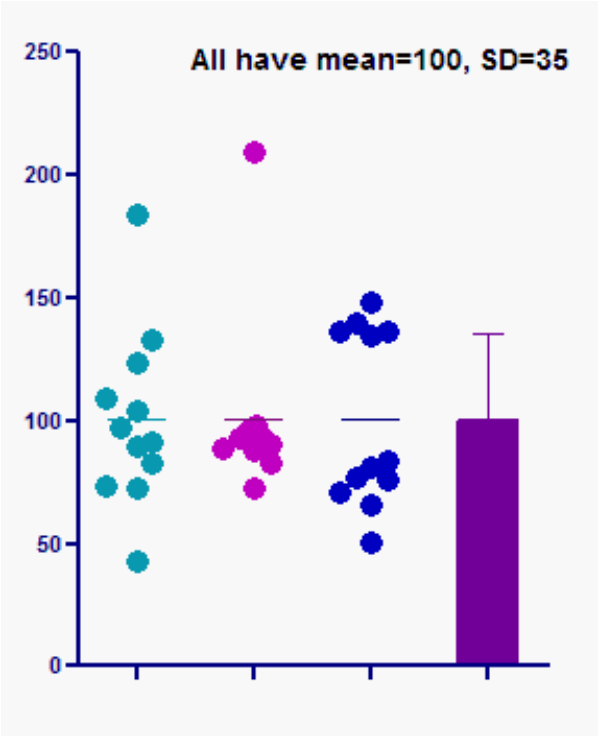
The standard deviation (SD) quantifies variability or scatter, and it is expressed in the same units as your data.

2. How to interpret the SD?

When the data are Gaussian:



When the data are not Gaussian:



3. How to report SD?

Mean (SD = xx)

Standard Error of the Mean (SEM/SE)

1. What is SEM?

The standard error of the mean (SEM) quantifies the precision of the mean. It is a measure of how far your sample mean is likely to be from the true population mean. It is expressed in the same units as the data.

2. How to interpret the SEM?

Although scientists often present data as mean and SEM, interpreting what the SEM means is not straightforward. It is much easier to interpret the 95% confidence interval, which is calculated from the SEM.

Comparison between SD and SEM

1. Calculation

$$\text{standard deviation } \sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

$$\text{variance} = \sigma^2$$

$$\text{standard error } (\sigma_{\bar{x}}) = \frac{\sigma}{\sqrt{n}}$$

where:

$\bar{x}$ = the sample's mean

n = the sample size

2. The SD quantifies scatter — how much the values vary from one another. The SEM quantifies how precisely you know the true mean of the population. It takes into account both the value of the SD and the sample size.
3. If you want to show the variation in your data – Plot SD

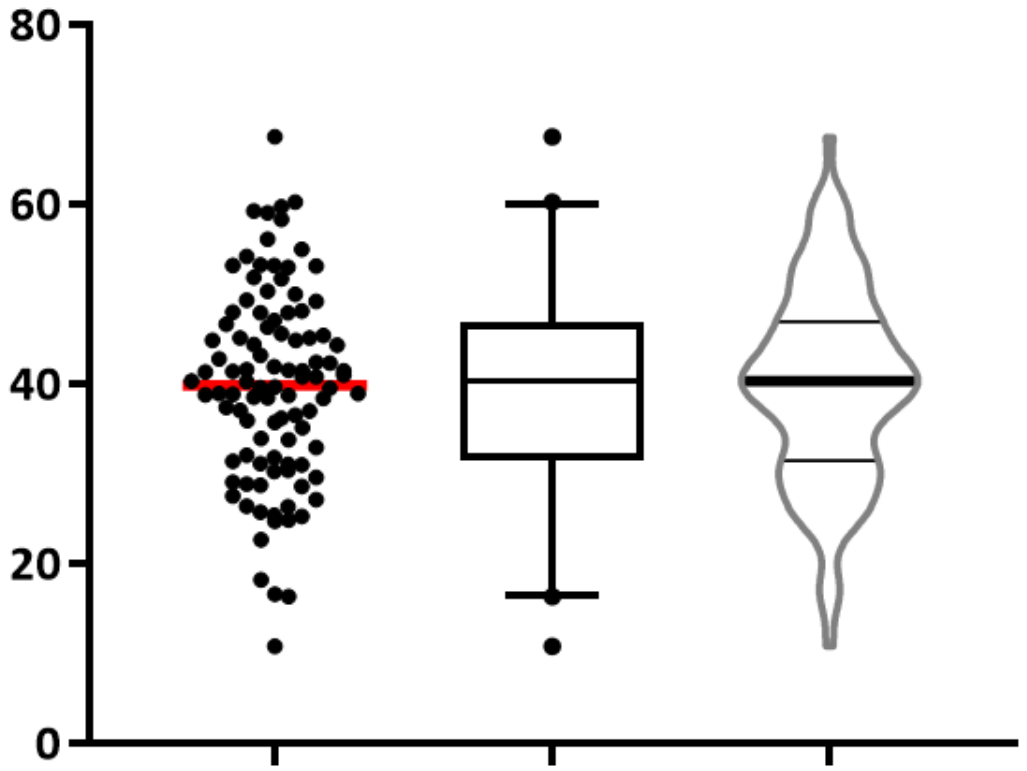
   If you want to show how precisely you have determined the mean – Plot SEM

Viewing Data Distributions

Left: Scatter Plot

Middle: Box-and-Whiskers Graph

Right: Violin Plot



Normal distribution?

Prism offers four normality tests. In this case, the null hypothesis is that all the values were sampled from a population that follows a Normal distribution.

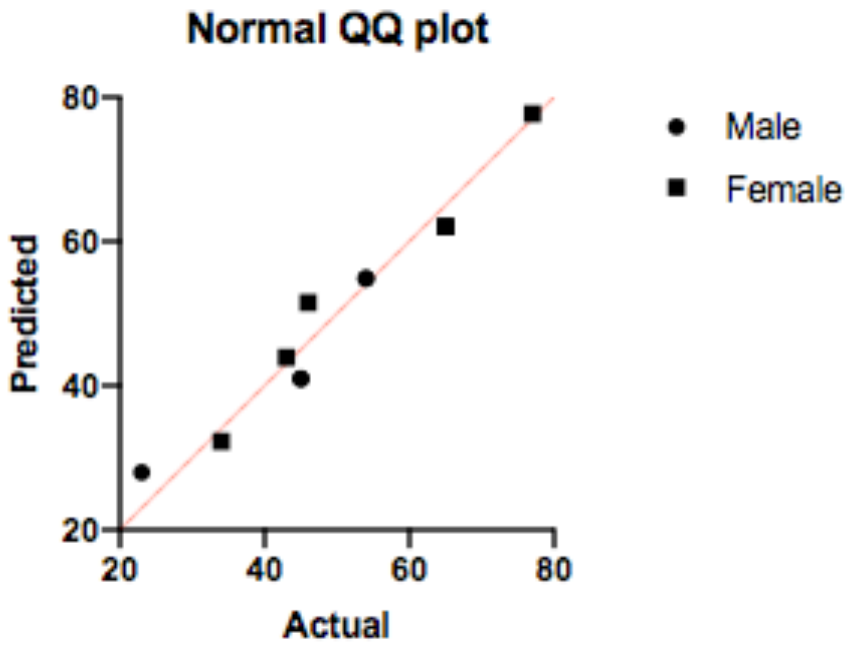
- When p-value is high: The data are not inconsistent with a Normal distribution
- When p-value is low: Reject that null hypothesis that the data are not sampled from a Gaussian population.

[The difference between different normality tests]

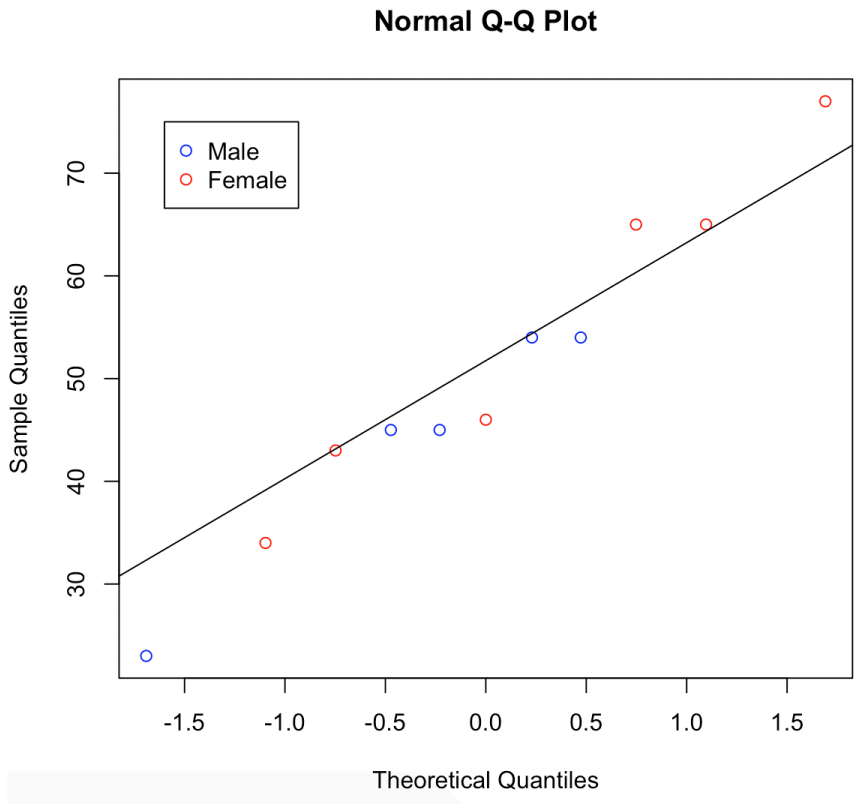
	Normality and Lognormality Tests Tabular results	A Male	B Female
		Y	Y
1	**Test for normal distribution**		
2	**Anderson-Darling test**		
3	A2*	N too small	N too small
4	P value		
5	Passed normality test (alpha=0.05)?		
6	P value summary		
7			
8	**D'Agostino & Pearson test**		
9	K2	N too small	N too small
10	P value		
11	Passed normality test (alpha=0.05)?		
12	P value summary		
13			
14	**Shapiro-Wilk test**		
15	W	0.8070	0.9335
16	P value	0.0923	0.6076
17	Passed normality test (alpha=0.05)?	Yes	Yes
18	P value summary	ns	ns
19			
20	**Kolmogorov-Smirnov test**		
21	KS distance	0.3252	0.2286
22	P value	0.0906	>0.1000
23	Passed normality test (alpha=0.05)?	Yes	Yes
24	P value summary	ns	ns
25			
26	**Number of values**	5	6

QQ Normality Plots

QQ plot in Prism



QQ plot in R



Lab 4 – Statistical Testing

Useful References for choosing appropriate tests

1. Prism 8 Statistics Guide (link)

   - This guide examines general principles of statistical analysis, looks at how to conduct those analyses in Prism, and how to interpret results of these analyses.

2. Choosing the Correct Statistical Test in SAS, Stata, SPSS and R (link)

   - A flow chart to choose the best statistical test and the related codes with SAS, Stata, SPSS and R

3. Summary and Analysis of Extension Program Evaluation in R (link)

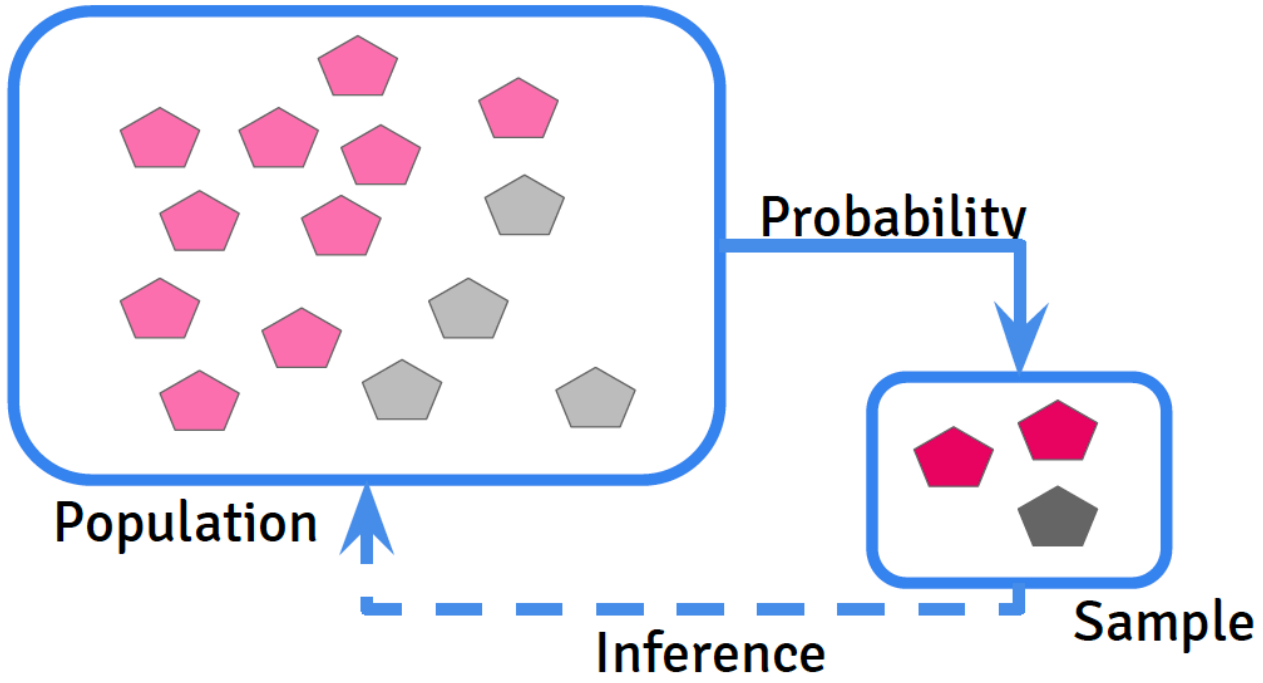
   - A dictionary of statistical testing methods and the related R codes

4. Analysis Data Model (ADaM) Examples in Commonly Used Statistical Analysis Methods (link)

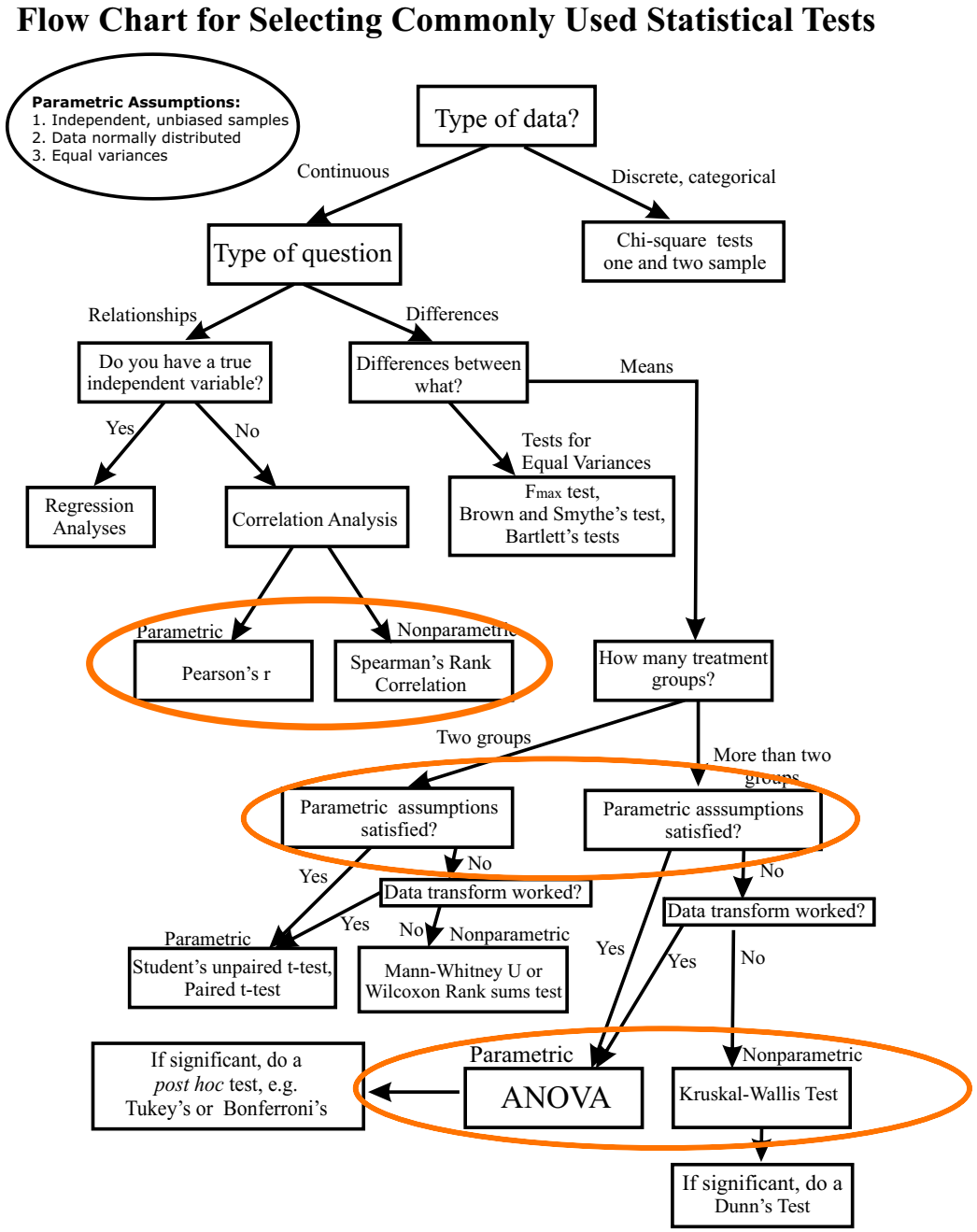
   - Commonly used statistical analysis methods in clinical trials (proved by FDA)

Statistical Testing Process

The best way to determine whether a statistical hypothesis is true would be to examine the entire population. Since that is often impractical, researchers typically examine a random sample from the population. If sample data are not consistent with the statistical hypothesis, the hypothesis is rejected.



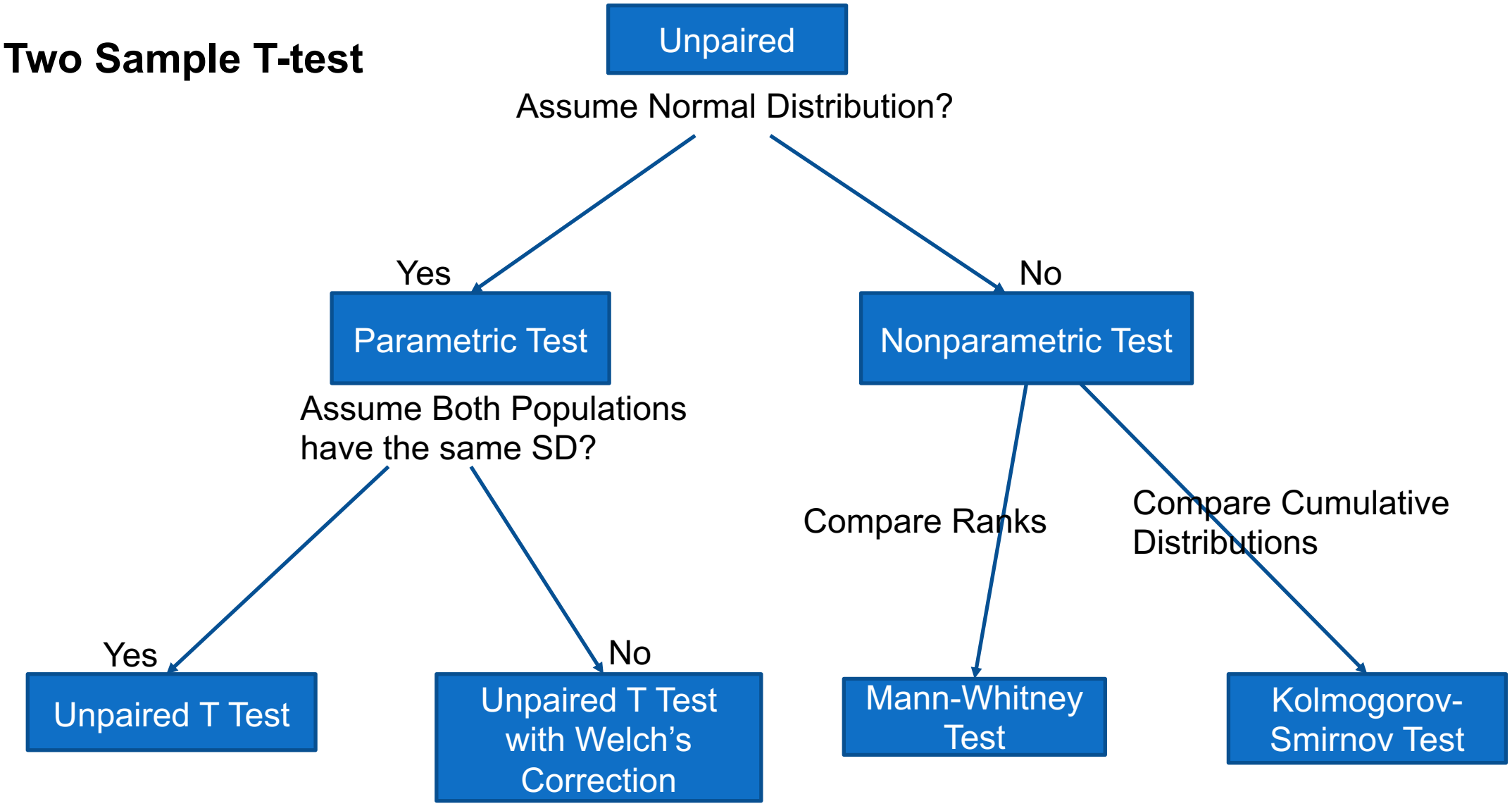
How to choose appropriate statistical test?

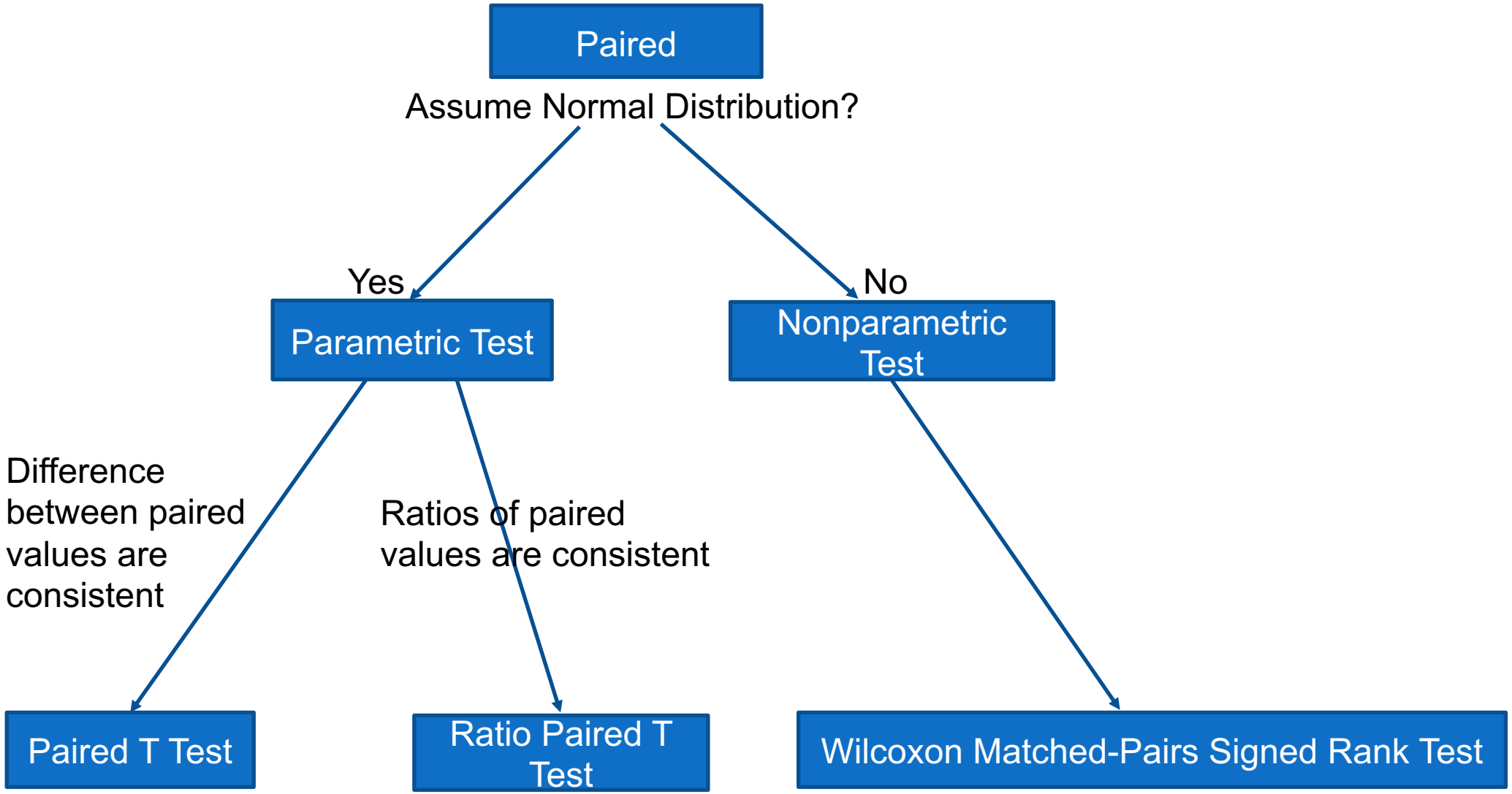


National Institute of Allergy and Infectious Diseases

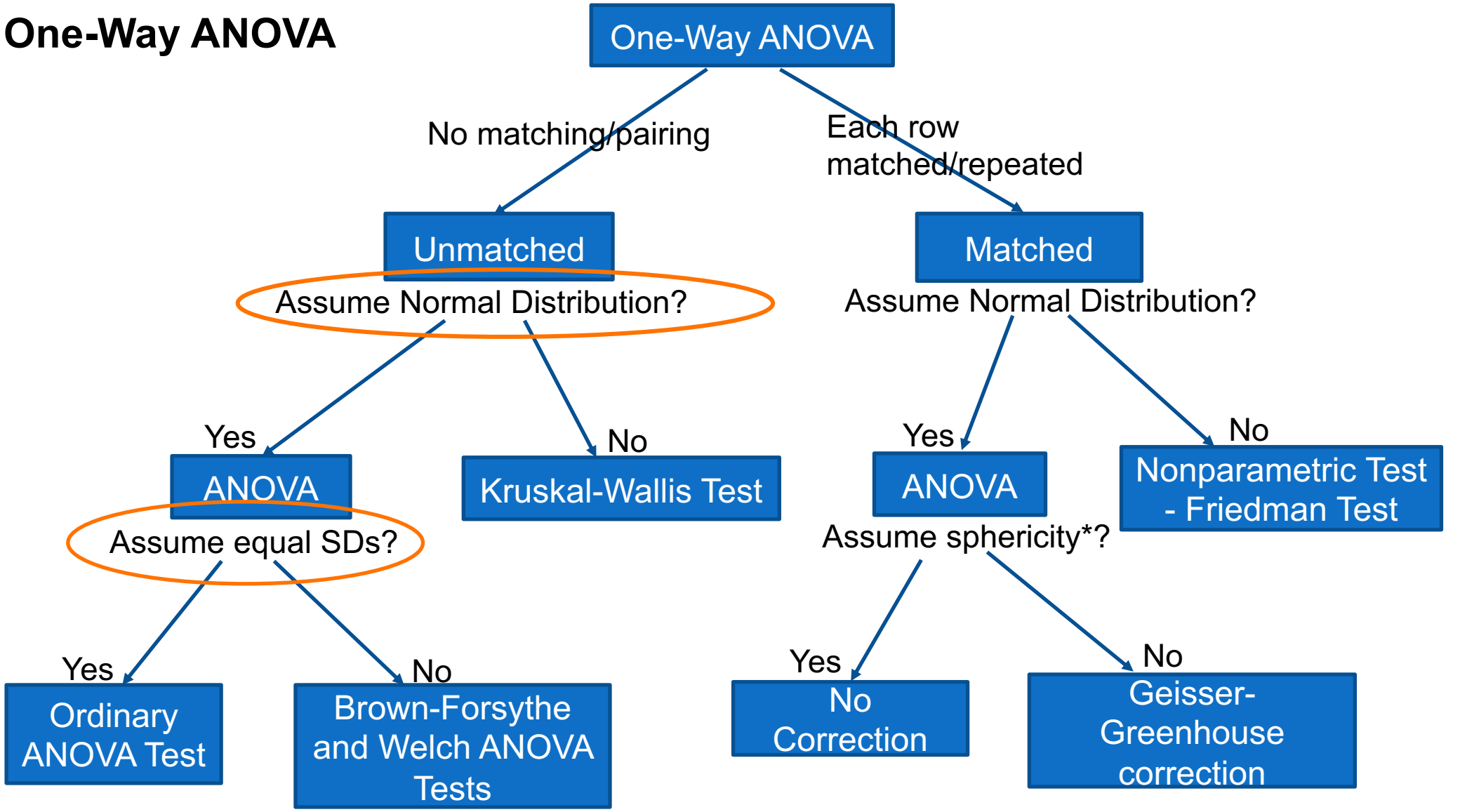
NIAID

Two Sample T-test




Paired
Assume Normal Distribution?
Yes
No
Parametric Test
Nonparametric Test
Difference between paired values are consistent
Ratios of paired values are consistent
Paired T Test
Ratio Paired T Test
Wilcoxon Matched-Pairs Signed Rank Test

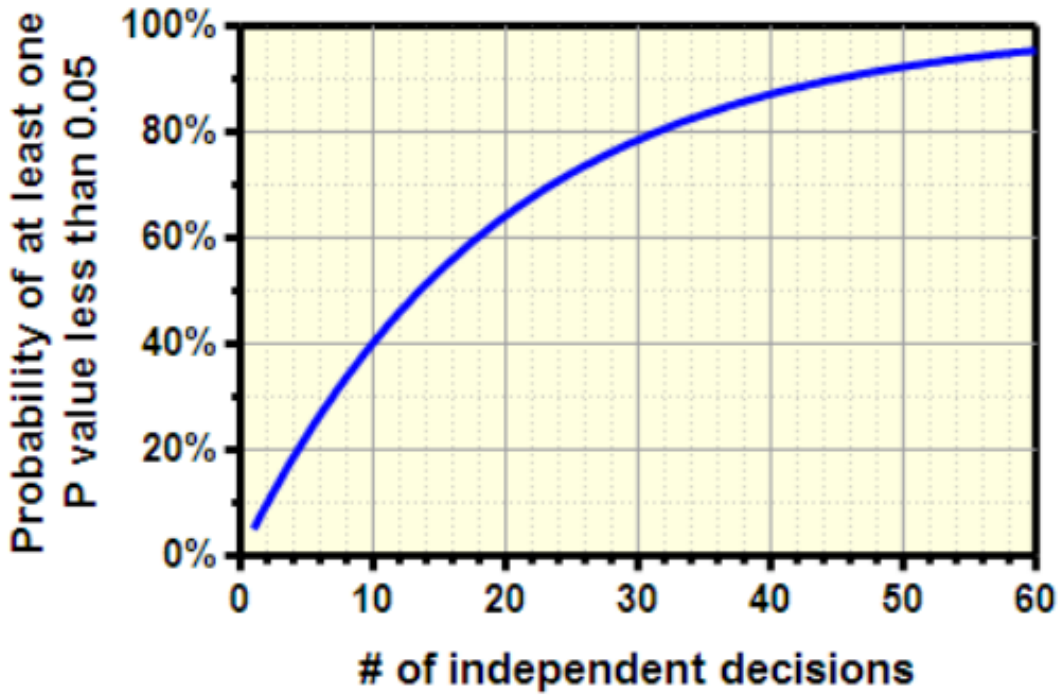
One-Way ANOVA



*Sphericity: Equal variability of differences

Multiple Comparison

Interpreting multiple P values is difficult. If you test several independent null hypotheses and leave the threshold at 0.05 for each comparison, the chance of obtaining at least one "statistically significant" result is greater than 5% (even if all null hypotheses are true).



Three Approaches to Multiple Comparisons

1. Don’t correct for multiple comparisons
2. Control the Type I error rate for the family of comparisons
3. Control the False Discovery Rate (FDR)

Lab 5 – Survival Analysis

In many clinical and animal studies, the outcome is survival time. The goal of the study is to determine whether a treatment changes survival. Prism creates survival curves, using the product limit method of Kaplan and Meier, and compares survival curves using both the **log-rank test** and the **Gehan-Wilcoxon test**.

Examples of **events** and **censor**:

Events	Censor
1. Death 2. Targeted Events	1. Still alive at the end of the study 2. Drop out of the study

Log-rank Test	Gehan-Breslow-Wilcoxon test
• Gives equal weights to all time points • More standard • More powerful if the assumptions of proportional hazards** is true.	• Gives more weight to deaths at early time points • The results can be misleading when a large fraction of patients is censored at early time points. • Doesn't require a consistent hazard ratio, but does require that one group consistently have a higher risk than the other.

** Proportional hazards mean that the ratio of hazard functions (deaths per time) is the same at all time points.

Lab 6 – Categorical data analysis

What is categorical data?

- Categorical data arises when individuals are categorized into one of two or more mutually exclusive groups.
- Continuous data could be transformed to categorical data with respect to some predefined criteria.

Why contingency table?

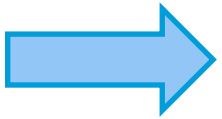
- Contingency tables summarize results where you compared two or more groups and the outcome is a categorical variable (such as disease vs. no disease, pass vs. fail, artery open vs. artery obstructed).

** You must enter data in the form of a contingency table. Prism cannot cross tabulate raw data to create a contingency table.

How to display data?

- For data from prospective and experimental studies, the top row usually represents exposure to a risk factor or treatment, and the bottom row is for controls. The left column usually tabulates the number of individuals with disease; the right column is for those without the disease.
- In case-control retrospective studies, the left column is for cases; the right column is for controls. The top row tabulates the number of individuals exposed to the risk factor; the bottom row is for those not exposed.

Smoking	Lung Cancer	Count
Yes	Case	688
Yes	Control	650
No	Case	21
No	Control	59



Contingency table

	Lung Cancer	
Smoking	Case	Control
Yes	688	650
No	21	59

Reference

CDA Outline

- ❑ Create a contingency table
- ❑ Calculation - sensitivity and specificity *(relative risk, difference between proportion, odds ratio are available from the same option box)*
- ❑ Perform statistical testing for contingency table

Link to the practice manual

How to report statistical results

Overall:

- Every statistical paper should report all methods completely enough so someone else could reproduce the work exactly.
- Every figure and table should present the data clearly
- All the results should be reported completely enough that no one wonders what you actually did.

Statistical Methods:

- State the full name of the test.
- Identify the program of the program that did the calculations
- State all options you selected. Repeated measures? Report enough detail so anyone could start with your data and get precisely the same results you got.

Graphing Data:

- Present data clearly. Focus on letting the reader see the data, and not only your conclusions.
- When possible, graph the individual data, not a summary of the data.

Conclusions

- The materials of this seminar will be updated on: Github - Prism
- If you have any further statistical problem, please send email to bioinformatics@niaid.nih.gov
- Check our training schedule: http://www.eventzilla.net/user/NIAID_OCICB_BCBB
- Take the survey and tell us what statistical topics you want to learn about next: BCBB Statistical Training – Suggest a class!
- Ask question on slack group!

  The invitation to the slack group will be sent after the workshop

Thank You!