BIOL 366

Data Analysis

Stuart R. Borrett

Ecological Detectives

What do our data mean?

http://www.sherlockandwatson.com/Images/sherlock2.gif

Pop Quiz

1. Name two statistical tests we will perform today.
2. What is the formula for the sum of squares (aka sum of squared error, sum of squared deviation)?
3. What is the name of the statistical software we will be using in class today?
4. Distinguish between a parametric and non-parametric statistic.

Learning Objectives

At the end of today's laboratory, you should be able to:

- Describe the purpose of comparative statistics
- Understand and apply the t-test
  - Identify its assumptions
- Understand and apply the chi-square test

Statistical Analyses

- Descriptive Statistics to describe your sample population
  - Central Tendency – mean, median, mode
  - Variability – standard deviation, variance

- Comparative Statistics to test your null hypotheses (aka Hypothesis Testing)
  - Are the ***means*** of two sample distributions different?
    - t-test for continuous data
  - Is your sample normally distributed?
    - Shapiro–Wilk test
  - Are two populations of counts or frequencies different?
    - χ^2 test

Comparative Statistics

Are two populations different?

Symbols and Terms

x_i = single observation or data point

n = sample size (number of data points)

$\sum_{i=1}^{n}$ = sum of $i = 1, 2, \ldots, n$ observations

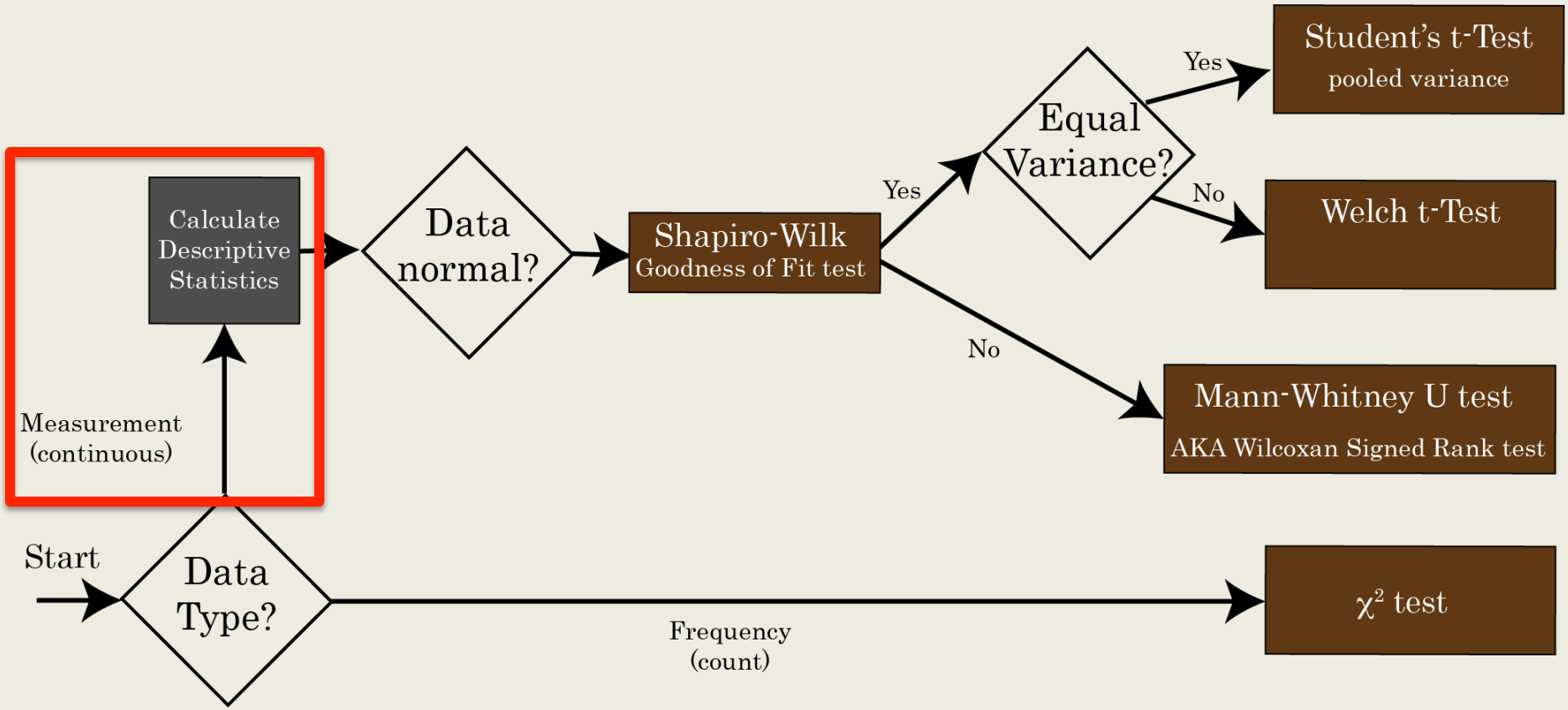
$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$ = mean

$SSD = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$ = sum of squared deviations

df = degrees of freedom, often $n - 1$

$s^2 = SSD/df$ = variance

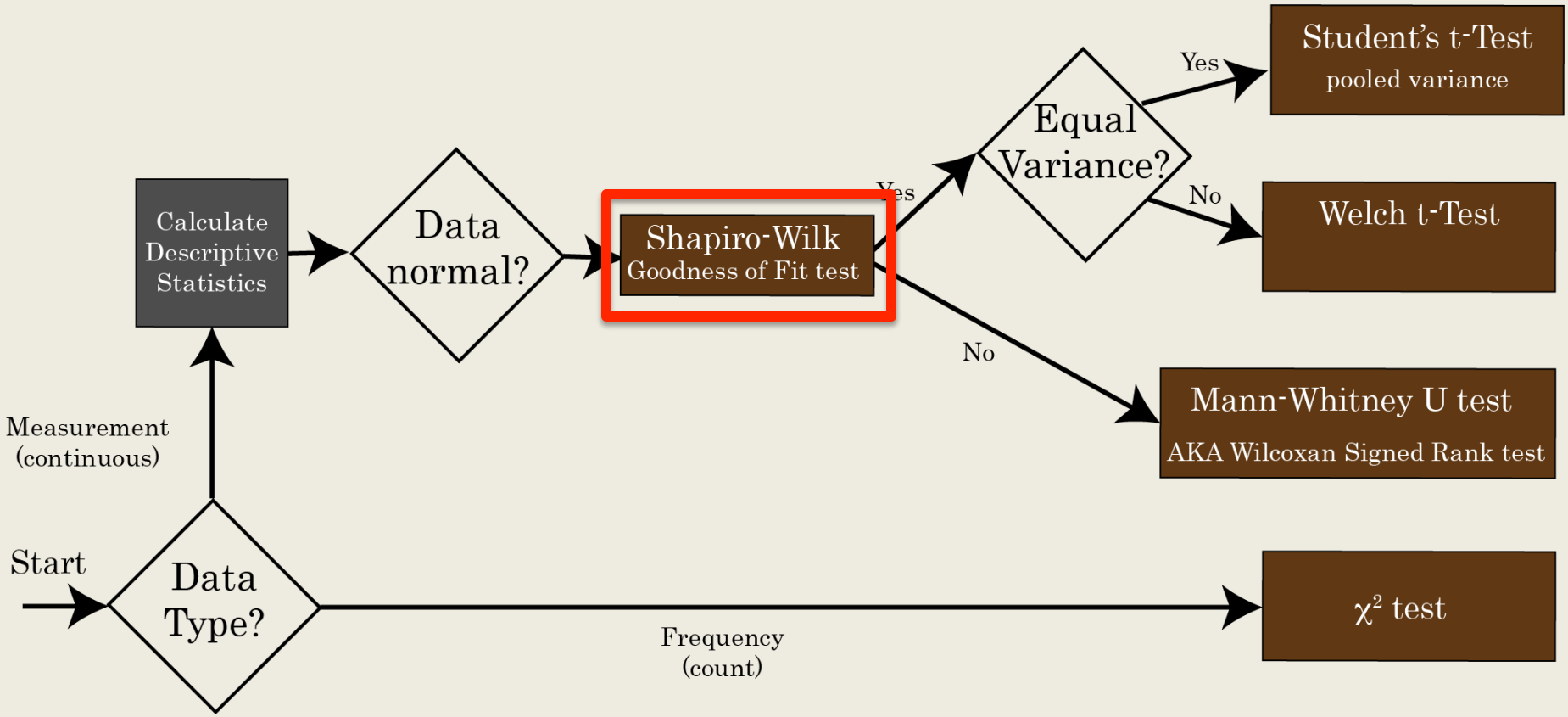
Analysis Flow Chart



Enter Data Into JMP

- Open JMP from http://tealware.uncw.edu
  - May try to install citrix client on your machine – allow
- Open new data table
- Enter data in columns
  - Show example
  - **Data Type** is essential – how JMP determines which tests are available/appropriate
- Calculate descriptive statistics
  - Analyze → Distribution → {y=circumference, by = type}

Analysis Flow Chart



Shapiro-Wilk Goodness of Fit

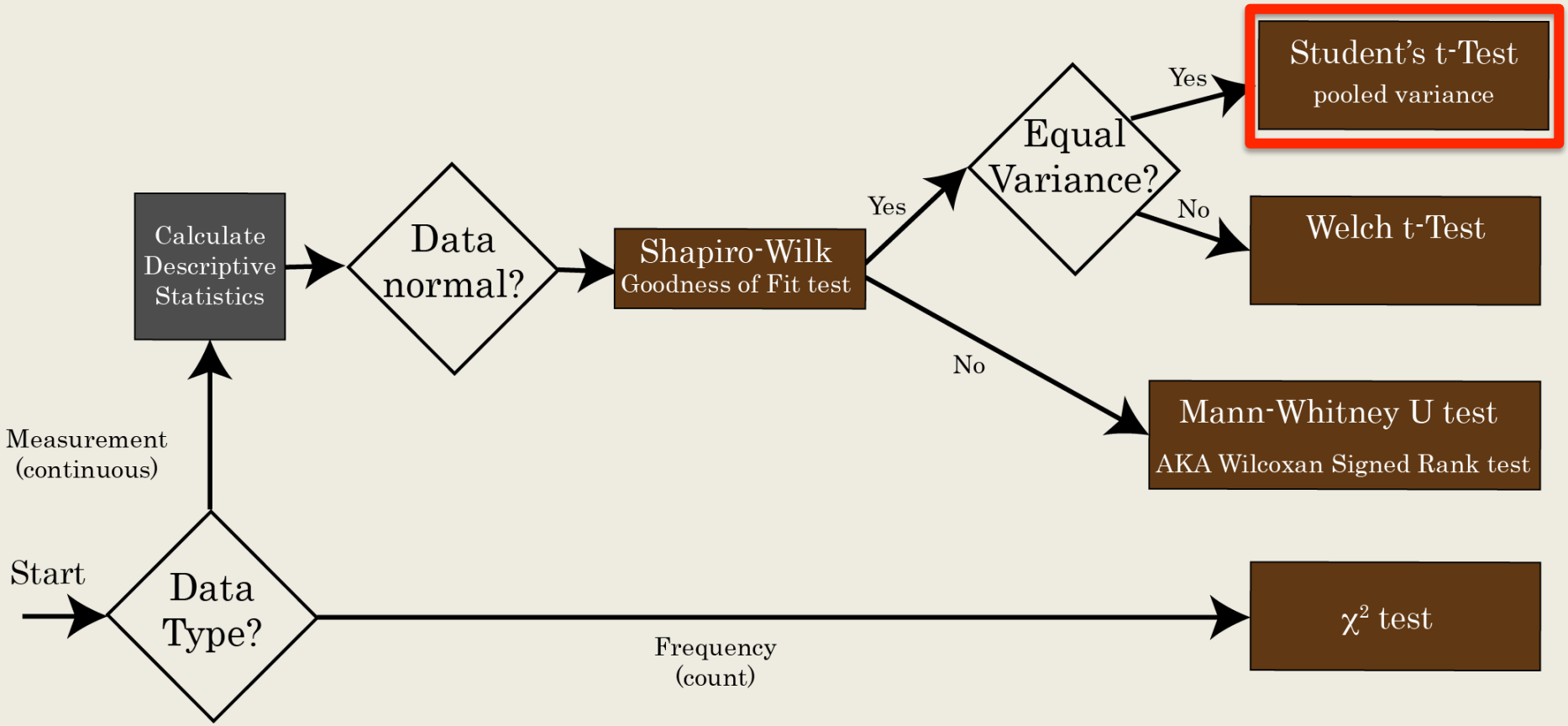
- Compares two distributions

  H_o: There is no difference between the distribution of our data and a normal distribution

- Use JMP to calculate (~~as in laboratory manual~~)
  - Circumference → Fit Distribution → Normal
  - Fitted Normal → Goodness of Fit → (W, p-value)

- Test each population separately
  - i.e. Forest A and B

Analysis Flow Chart



Student's t-test

- Compares the mean value of two samples

  H_o: There is no difference between the means

- Assumes that the variable is
  - continuous
  - has a normal distribution
- Example
- Use JMP to calculate (as in laboratory manual)
  - Analyze → Fit Y by X → (y = circumference, by = type)
  - Means/ANOVA/Pooled t → {t Ratio, DF, p-value}

Student's t-test statistic

$$t = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

Comparing the means, but considering the variation.

$\bar{X}_i$ Is the mean of sample i

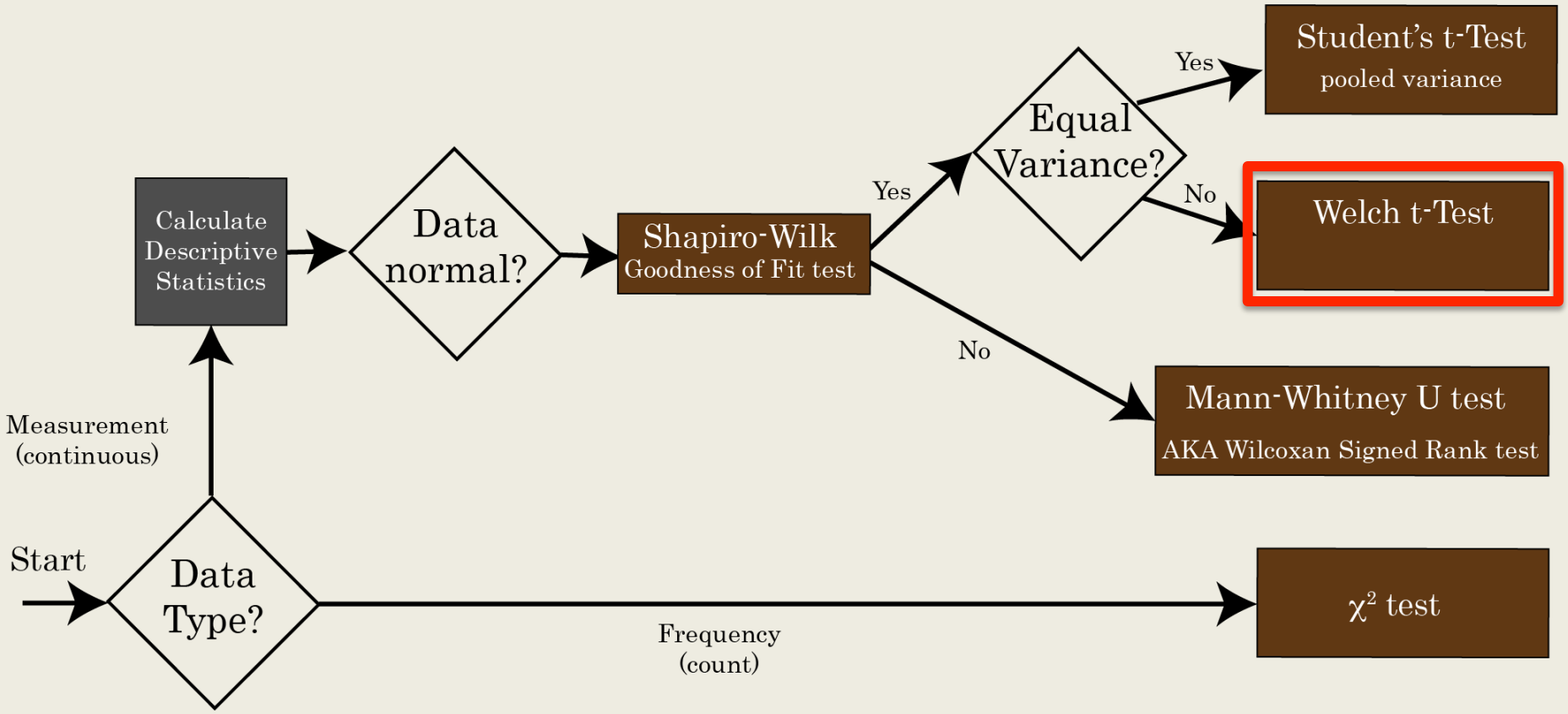
n_i Is the number of observations in sample i

$$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2}$$

Is an estimate of the standard error of the mean.

Assumes equal variances

Analysis Flow Chart



Welch's t-test

The student's t-test described in the laboratory manual assumes that the variances of the two sample populations being compared are equal.

Termed "pooled variance".

Easier to calculate by hand, but may not be true.

Welch's t-test does not assume the variances are equal, which is safer.

$$t = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Welch t-test

- Compares the mean value of two samples

  H_o: There is no difference between the means

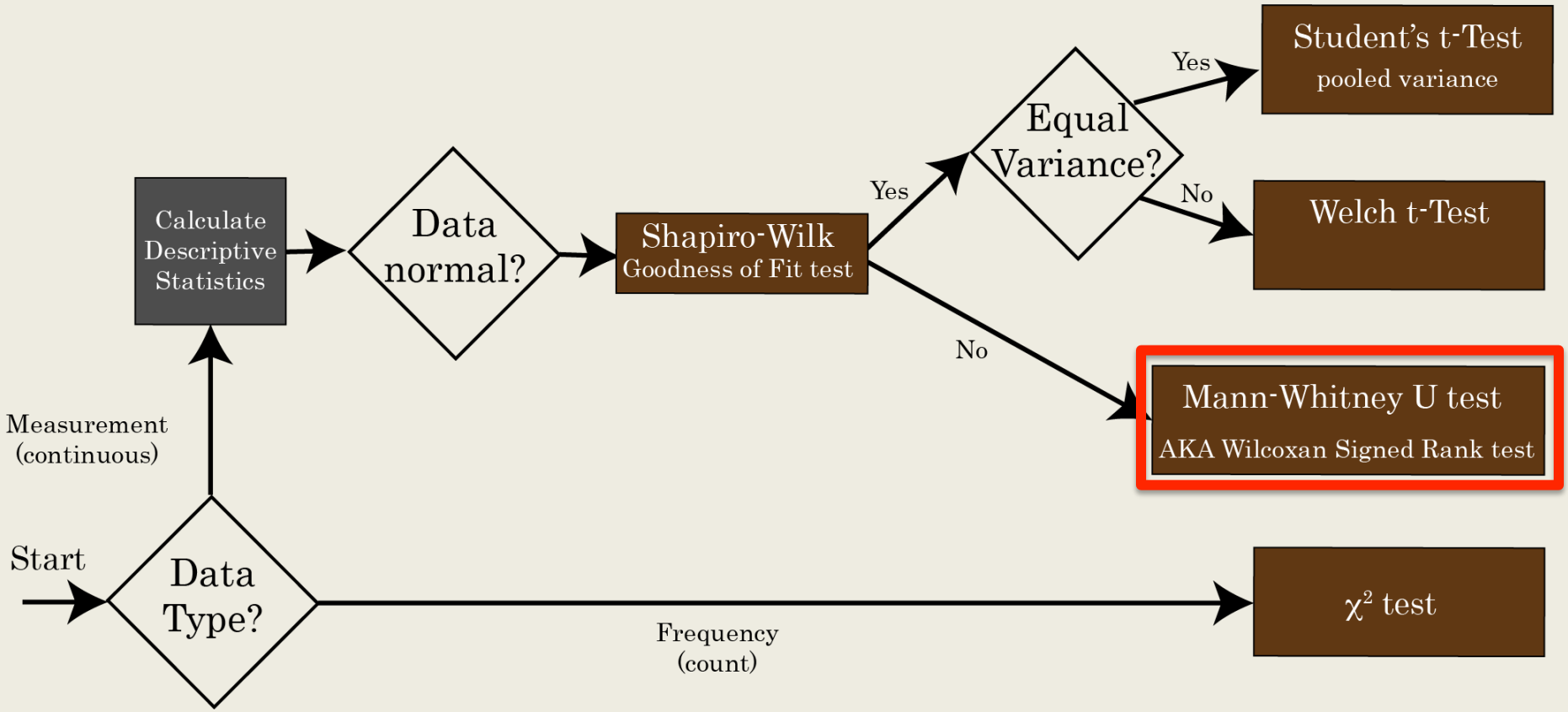
- Assumes that the variable is
  - Continuous, has a normal distribution
  - Variances not equal
- Use JMP to calculate
  - Analyze → Fit Y by X → (y = circumference, x = type)
  - t-Test → {t Ratio, DF, p-value}

What are the key assumptions of the t-test?

The t-test is a parametric statistic, and it assumes

(1) the data are independent and
(2) n*ormally distributed*.

Analysis Flow Chart



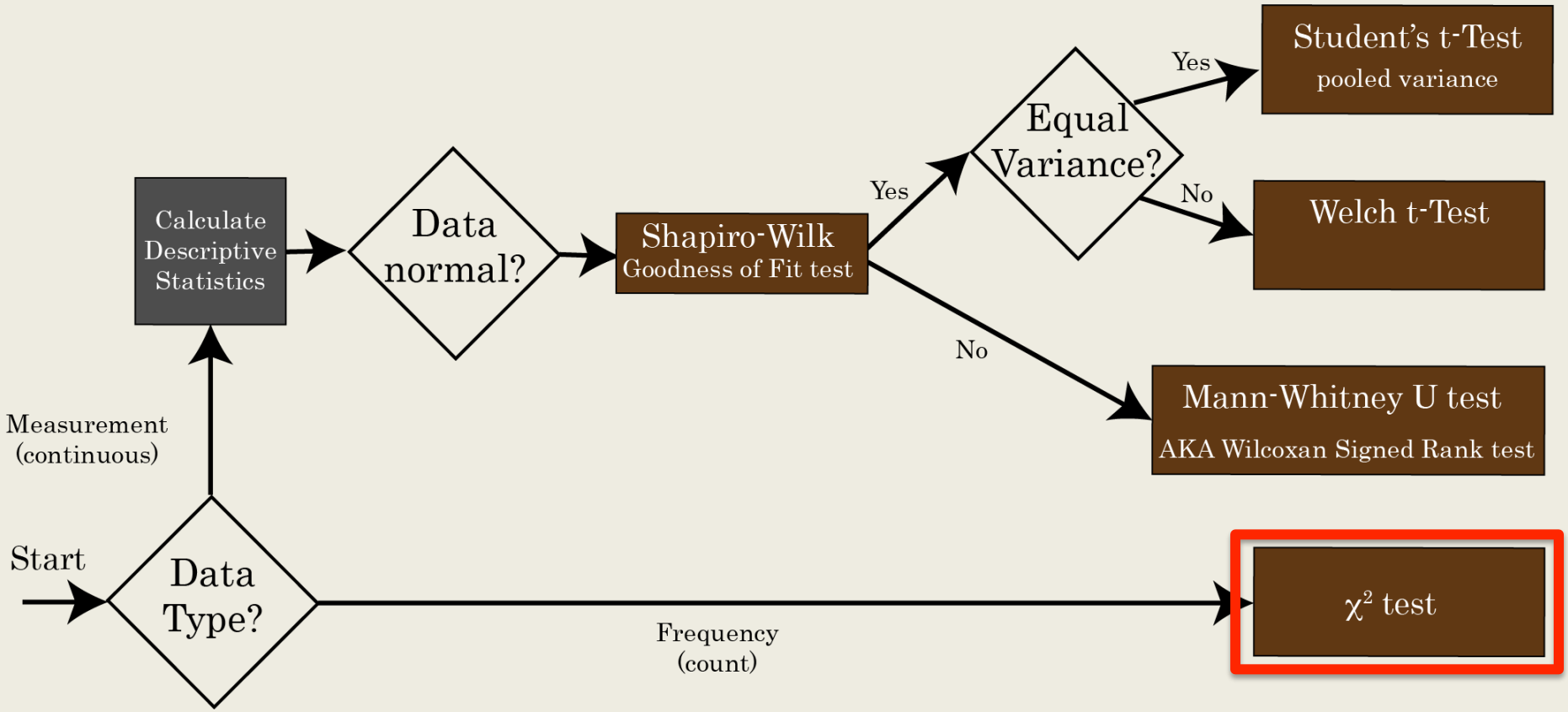
Mann-Whitney U / Wilcoxon Sign Rank

- Compares the mean value of two samples

  H_o: There is no difference between the means

- Non-parametric test
  - Does not assume data distribution
- Use JMP to calculate
  - Analyze → Fit Y by X → (y = circumference, by = type)
  - Non-parametric → Wicoxon → {S, p-value}

Analysis Flow Chart



Chi-Square Test

- Compares totals, counts or frequencies
- Are two values significantly different?

  H_o: There is no difference between the means

- Formula

$$\chi^2 = \sum_{i=1}^{k} \frac{(observed_i - expected_i)^2}{expected_i}$$

- Work example on board.
- We will calculate by hand or in Excel

Type I vs. Type II Errors

Type I error

- Probability of rejecting null hypothesis when it is correct

Type II error

- Probability of accepting the null hypothesis when it is in fact incorrect.

Lab Activity

- Work in teams to analyze your data using the appropriate methods
  - Descriptive Statistics
  - Comparative Statistics; hypothesis testing
- Write a *draft* of your results section
  - Build any necessary figures/tables
  - Determine the evidence for your hypotheses
  - Consider the questions on pages 32 of the lab manual.

Full Draft Report Due Tonight by Midnight

Analysis Flow Chart