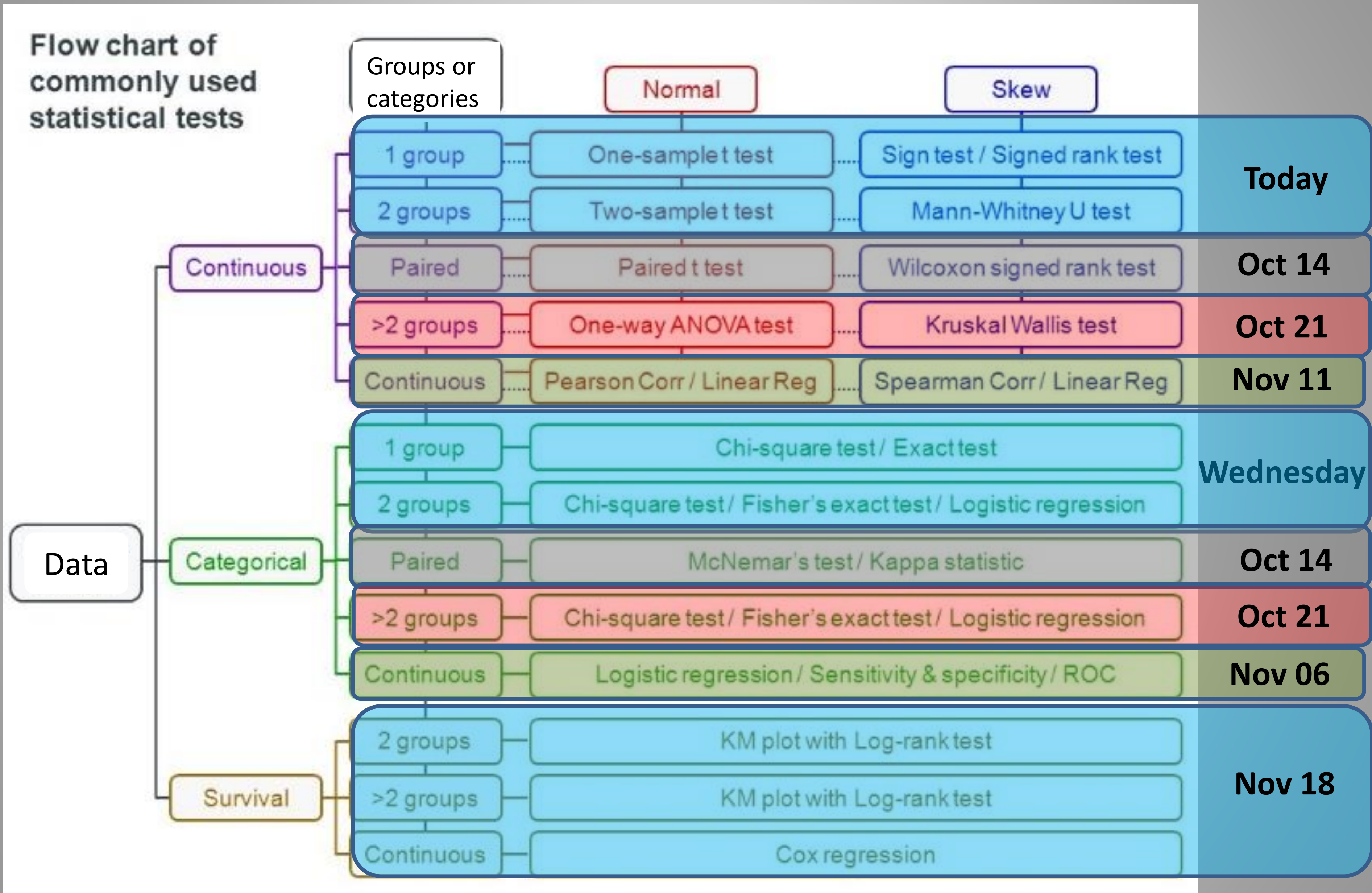
# Comparing one or two independent groups of continuous data:

# One sample and two sample (AKA independent or unpaired) t-tests, Wilcoxon signed rank test, Mann-Whitney U (Wilcoxon rank sum)

Kathleen Torkko

October 07, 2019


Flow chart of commonly used statistical tests
Groups or categories
Normal
Skew
Data
Continuous
1 group
One-sample t test
Sign test / Signed rank test
2 groups
Two-sample t test
Mann-Whitney U test
Today
Paired
Paired t test
Wilcoxon signed rank test
Oct 14
>2 groups
One-way ANOVA test
Kruskal Wallis test
Oct 21
Continuous
Pearson Corr / Linear Reg
Spearman Corr / Linear Reg
Nov 11
Categorical
1 group
Chi-square test / Exact test
2 groups
Chi-square test / Fisher's exact test / Logistic regression
Wednesday
Paired
McNemar's test / Kappa statistic
Oct 14
>2 groups
Chi-square test / Fisher's exact test / Logistic regression
Oct 21
Continuous
Logistic regression / Sensitivity & specificity / ROC
Nov 06
Survival
2 groups
KM plot with Log-rank test
>2 groups
KM plot with Log-rank test
Continuous
Cox regression
Nov 18

# Objectives

Learn to use and interpret parametric and non-parametric tests for testing differences between:

one mean compared to a known mean

one distribution to a known distribution

two independent (unpaired) means

two independent distributions

Learn about the assumptions for each test

# Example

- *Maximum relaxation of bladder muscles in rats:* Frazier, Schneider and Michel (2006) measured how well the neurotransmitter norepinephrine relaxes bladder muscles in rats. Here we will compare the maximal relaxation between old and young rats.
- *Data:*
  - Old: 20.8, 2.8, 50.0, 33.3, 29.4, 38.9, 29.4, 52.6,14.3.
  - Young: 45.5, 55.0, 60.7, 61.5, 61.1, 65.5, 42.9, 37.5.

Frazier EP, Schneider T, Michel MC. Naunyn Schmiedebergs Arch Pharmacol. 2006 Jul;373(4):300-9.

# History of the t-test
# How beer created a statistical test

At Guinness Brewery in Ireland until the early 1900s, hops quality were determined by “looks and fragrance” but they wanted to use a more scientific method: proportion of soft resins to hard.

Problem: small number of hops batches to test. The statistical test available at the time,  the z-statistic, needed a sample size >30

Also, the z-statistic needs a known population mean and SD for comparison

How could they test the consistency of the resin proportion across small batches of hops and when the population mean and SD are not known?

They turned to William Gosset (mathematician, scientist, and experimental brewer; Employed by Guinness Brewery, Dublin, Ireland, from 1899-1935).

# History of the t-test

Guiness asked Gosset to work on the problem because he had studied a bit of Math at Oxford and was less scared of this kind of work than the other brewers.

He developed the t-test (which used a t-distribution) around 1905 (while on sabbatical working in Karl Pearsons' lab) for dealing with small samples in brewing quality control.

Published in 1908 under pseudonym "Student" ("Student's t-test")

THE PROBABLE ERROR OF A MEAN

BY STUDENT

Introduction

Any experiment may he regarded as forming an individual of a "population" of experiments which might he performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a greater number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information

The t-test was designed to work in small sample sizes and when the population mean and SD are unknown (otherwise you would use a Z score)

For continuous data that meet assumptions of normality and homoscedasticity:
t-tests are used to compare ***two*** means

**Independent t-test**

data collected from two independent groups

Prism calls these “unpaired” t-tests

**One-sample t-test**

used to compare one mean to a known or standard mean

**For paired data** used the paired t-test – next week

**For >2 groups**, use the ANOVA - which will be covered later

*t-tests are called t-tests because they use the t-statistic and the t-distribution*

# The t Statistic

Independent t-test

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1^{\ 2}}{N_1} + \frac{S_2^{\ 2}}{N_2}}}$$

One sample t-test

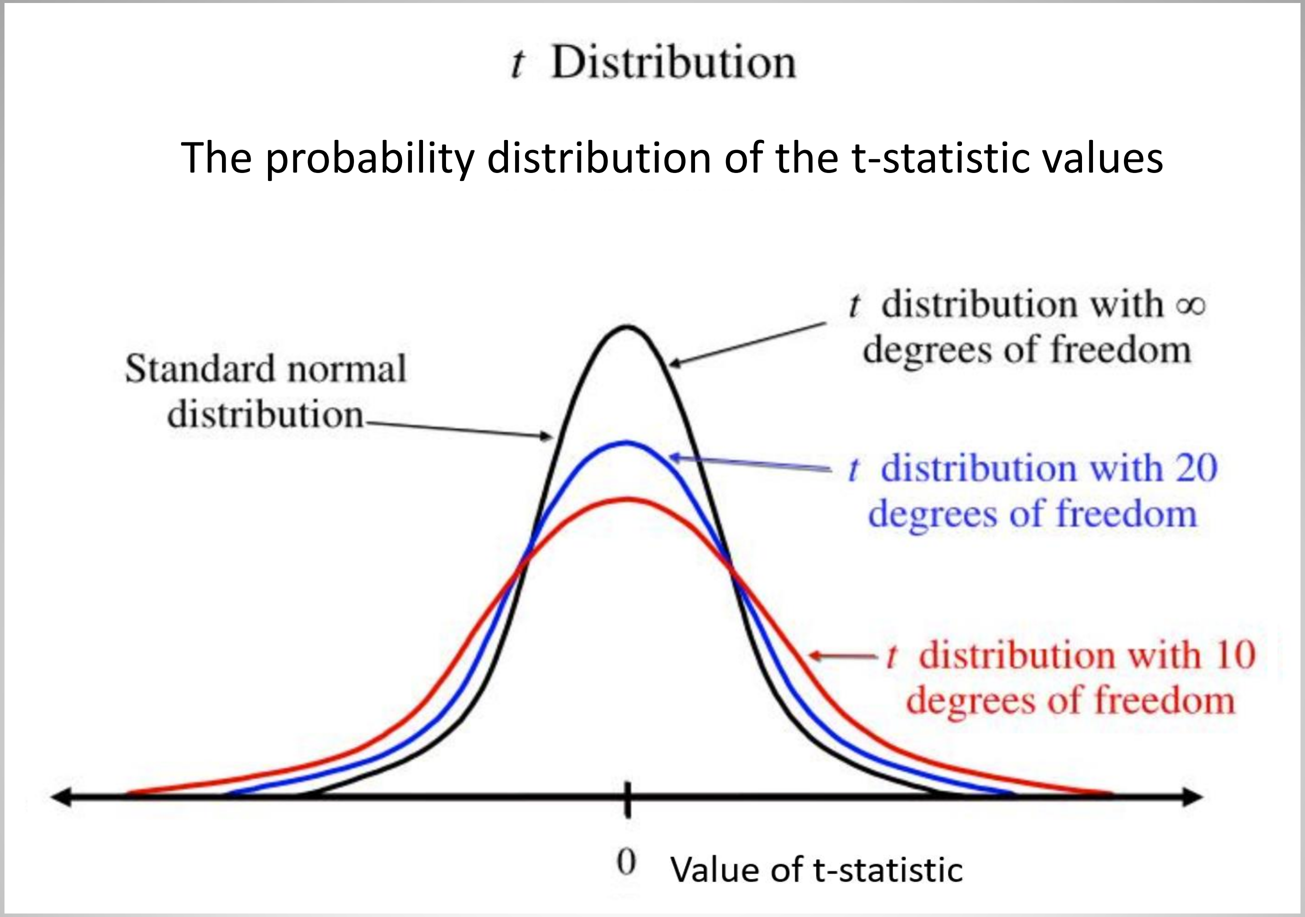
$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

The t statistic allows researchers to use sample data to test hypotheses about an unknown population mean.

The t statistic does not require any knowledge of the population standard deviation and variance.

Uses *degrees of freedom* (df; calculated from sample sizes) to determine *critical values* for the *t-distribution* for a specific value of df.

$$z = \frac{x - \mu_0}{\sigma}$$




The t-distribution has
- slightly less area near the expected central value
- more area in the "tails"  (fatter; helps account for uncertainty with small n
- a larger spread
- a different shape for different sample sizes

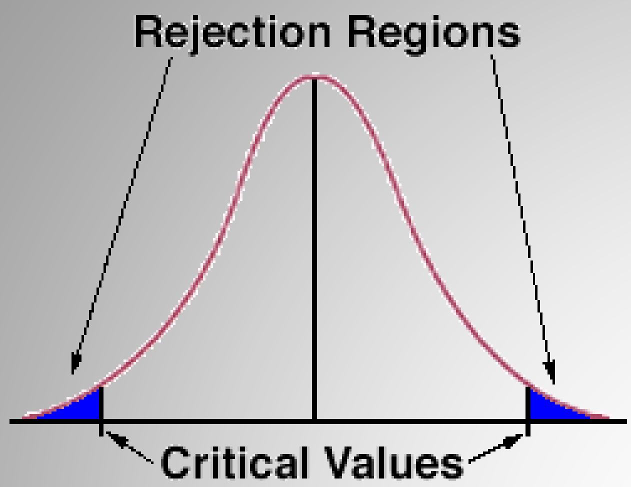
# The t-statistic and the t-distribution

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1{}^2}{N_1} + \frac{S_2{}^2}{N_2}}}$$

Example: A t-test is performed on two data samples and it calculates a t-statistic value of 2.5. What does that mean?

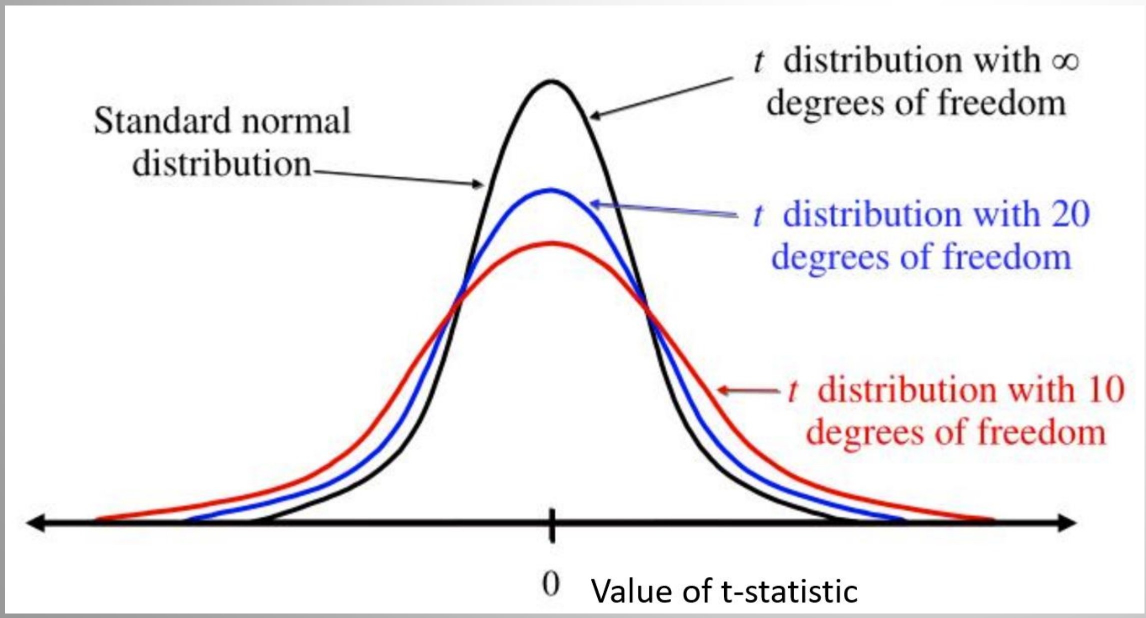
By itself, a t-statistic = 2.5 doesn't mean anything on its own

It needs to be compared to a t-distribution with specific degrees of freedom to determine critical values and *rejection regions*




Rejection regions = reject the null hypothesis

Upper blue area = 2.5% of t-statistic values
Lower blue area = 2.5% of t-statistic values




Critical values at $\alpha$=0.05 for a two-tailed test:

| t with | |
|---|---|
| 2 df | +/- 4.303 |
| 10 df | +/- 2.228 |
| 20 df | +/- 2.086 |
| 120 df | +/- 1.980 |
| ∞ df | +/- 1.960 |

## Critical values of Student's *t* distribution with *v* degrees of freedom

Probability less than the critical value ($t_{1-\alpha,v}$)

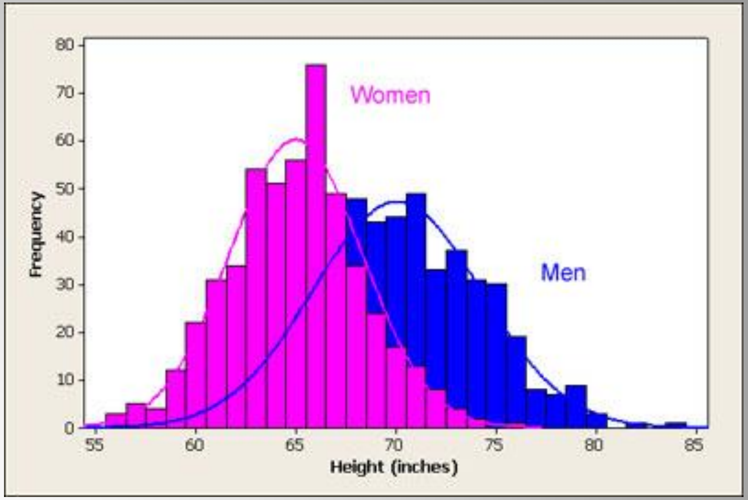
| v | 0.90 | 0.95 | 0.975 | 0.99 | 0.995 | 0.999 |
|---|---|---|---|---|---|---|
| 1. | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 | 318.313 |
| 2. | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 22.327 |
| 3. | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 10.215 |
| 4. | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173 |
| 5. | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.893 |
| 6. | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208 |
| 7. | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.782 |
| 8. | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 4.499 |
| 9. | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.296 |
| 10. | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.143 |
| 11. | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.024 |
| 12. | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.929 |
| 13. | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.852 |
| 14. | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.787 |
| 15. | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.733 |
| 16. | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.686 |
| 17. | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.646 |
| 18. | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.610 |
| 19. | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.579 |
| 20. | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.552 |
| 21. | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.527 |

$v = df = (n_1-1) + (n_2-1)$

Alpha level for a two-tailed test: 0.05/2 = 0.025 or 0.975

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1^{\,2}}{N_1} + \frac{S_2^{\,2}}{N_2}}}$$

## Let's start with the Independent t-test




Continuous variable (not nominal or ordinal)

Can be done only in 2 independent groups

Both samples are random samples from their respective populations
iid: **I**ndependent and **I**dentically **D**istributed random variables

$H_0$: mean height of women is the same as the mean height of men
$H_A$: mean height of women is different from the mean height of men

$H_0$: $\mu_{women} = \mu_{men}$

$H_A$: $\mu_{women} \neq \mu_{men}$

$H_0$: $\mu_{women} - \mu_{men} = 0$

$H_A$: $\mu_{women} - \mu_{men} \neq 0$

# iid: <u>**I**</u>ndependent and <u>**I**</u>dentically <u>**D**</u>istributed

Example: a coin toss to count the number of heads in ten tosses

Each toss is *independent* since every time you flip a coin, the previous result doesn't influence your current result

They are *identically distributed*, since every time you flip a coin, the chances of getting head (or tail) are identical, no matter if its the 1st or the 100th toss (probability distribution is identical over time). If the coin is "fair" the chances are 0.5 for each event (getting head or tail).

# Assumptions of independent t-tests
# Every test has them

Textbooks usually list the following assumptions:

1. The data must be randomly sampled from normally distributed populations
   - Or, the sample data for each group must be normally (symmetrically) distributed
2. The two samples must have equal variances (homoscedasticity).
3. Each observation must be independent from the others.

The third of these is the most important assumption
- Determines use of an independent or a paired t-test

Samples from real data rarely meets the first two assumptions
- Real populations are rarely truly normally distributed

But the t-test is a *robust* test concerning normality and homoscedasticity violations

## There is no formal definition of a "robust statistical test"

Robust tests are any statistical test that yield good performance when data are mostly unaffected by outliers or have only small departures from model assumptions in a given dataset.

*t*-tests are *robust* to violations of the assumptions of normality and equal variances, particularly in certain circumstances

larger samples sizes (*i.e.*, >30)
numbers are equal in each group (known as a balanced design)
the data distribution of the  groups are similar

# *Testing* t-test Assumptions

**Normality:** Plot frequency histograms and box plots

for each group of independent samples

Statistical tests (e.g., Shapiro-Wilk) for normality in terms of the null hypothesis (but remember the issues with normality tests)

**$H_0$**: Sample is drawn from a **normal** population

**$H_A$**: Sample is drawn from a **non-normal** population

**Homoscedasticity:** Scatter plots, compare sample standard deviations between the 2 groups

As a rough estimate, one SD should be no more than twice the other

Do an F-test to formally test for differences (remember problems)

**Independence:** Determine from study design

# The F test for Equal Variances Used by Prism

Divide the larger sample SD by the smaller

$$F = s_1^2 / s_2^2$$

S= the standard deviation of the sample

Degrees of freedom = n-1, calculated for both the numerator and denominator

$H_0$: variance of sample1 = variance of sample2
$H_A$: variance of sample1 ≠ variance of sample2

<u>EXAMPLE two-tailed test</u>
SD sample 1 = 9.6, n=25
SD sample 2 = 10.9, n=21

$F = 10.9^2 / 9.6^2 = 118.81/92.16 = 1.289$

<u>To find critical value</u>
Calculate degrees of freedom for the numerator = 21-1 = 20
Calculate degrees of freedom for the denominator = 25-1 = 24
Set Alpha level for a two-tailed test: 0.05/2 = 0.025

## IV. Percentage Points of the F Distribution (*continued*)

$F_{0.025,\nu_1,\nu_2}$

$\alpha = 0.025$

Degrees of Freedom for the Numerator ($\nu_1$); Degrees of Freedom for the Denominator ($\nu_2$)

| $\nu_2$ \ $\nu_1$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | ∞ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 647.8 | 799.5 | 864.2 | 899.6 | 921.8 | 937.1 | 948.2 | 956.7 | 963.3 | 968.6 | 976.7 | 984.9 | 993.1 | 997.2 | 1001 | 1006 | 1010 | 1014 | 1018 |
| 2 | 38.51 | 39.00 | 39.17 | 39.25 | 39.30 | 39.33 | 39.36 | 39.37 | 39.39 | 39.40 | 39.41 | 39.43 | 39.45 | 39.46 | 39.46 | 39.47 | 39.48 | 39.49 | 39.50 |
| 3 | 17.44 | 16.04 | 15.44 | 15.10 | 14.88 | 14.73 | 14.62 | 14.54 | 14.47 | 14.42 | 14.34 | 14.25 | 14.17 | 14.12 | 14.08 | 14.04 | 13.99 | 13.95 | 13.90 |
| 4 | 12.22 | 10.65 | 9.98 | 9.60 | 9.36 | 9.20 | 9.07 | 8.98 | 8.90 | 8.84 | 8.75 | 8.66 | 8.56 | 8.51 | 8.46 | 8.41 | 8.36 | 8.31 | 8.26 |
| 5 | 10.01 | 8.43 | 7.76 | 7.39 | 7.15 | 6.98 | 6.85 | 6.76 | 6.68 | 6.62 | 6.52 | 6.43 | 6.33 | 6.28 | 6.23 | 6.18 | 6.12 | 6.07 | 6.02 |
| 6 | 8.81 | 7.26 | 6.60 | 6.23 | 5.99 | 5.82 | 5.70 | 5.60 | 5.52 | 5.46 | 5.37 | 5.27 | 5.17 | 5.12 | 5.07 | 5.01 | 4.96 | 4.90 | 4.85 |
| 7 | 8.07 | 6.54 | 5.89 | 5.52 | 5.29 | 5.12 | 4.99 | 4.90 | 4.82 | 4.76 | 4.67 | 4.57 | 4.47 | 4.42 | 4.36 | 4.31 | 4.25 | 4.20 | 4.14 |
| 8 | 7.57 | 6.06 | 5.42 | 5.05 | 4.82 | 4.65 | 4.53 | 4.43 | 4.36 | 4.30 | 4.20 | 4.10 | 4.00 | 3.95 | 3.89 | 3.84 | 3.78 | 3.73 | 3.67 |
| 9 | 7.21 | 5.71 | 5.08 | 4.72 | 4.48 | 4.32 | 4.20 | 4.10 | 4.03 | 3.96 | 3.87 | 3.77 | 3.67 | 3.61 | 3.56 | 3.51 | 3.45 | 3.39 | 3.33 |
| 10 | 6.94 | 5.46 | 4.83 | 4.47 | 4.24 | 4.07 | 3.95 | 3.85 | 3.78 | 3.72 | 3.62 | 3.52 | 3.42 | 3.37 | 3.31 | 3.26 | 3.20 | 3.14 | 3.08 |
| 11 | 6.72 | 5.26 | 4.63 | 4.28 | 4.04 | 3.88 | 3.76 | 3.66 | 3.59 | 3.53 | 3.43 | 3.33 | 3.23 | 3.17 | 3.12 | 3.06 | 3.00 | 2.94 | 2.88 |
| 12 | 6.55 | 5.10 | 4.47 | 4.12 | 3.89 | 3.73 | 3.61 | 3.51 | 3.44 | 3.37 | 3.28 | 3.18 | 3.07 | 3.02 | 2.96 | 2.91 | 2.85 | 2.79 | 2.72 |
| 13 | 6.41 | 4.97 | 4.35 | 4.00 | 3.77 | 3.60 | 3.48 | 3.39 | 3.31 | 3.25 | 3.15 | 3.05 | 2.95 | 2.89 | 2.84 | 2.78 | 2.72 | 2.66 | 2.60 |
| 14 | 6.30 | 4.86 | 4.24 | 3.89 | 3.66 | 3.50 | 3.38 | 3.29 | 3.21 | 3.15 | 3.05 | 2.95 | 2.84 | 2.79 | 2.73 | 2.67 | 2.61 | 2.55 | 2.49 |
| 15 | 6.20 | 4.77 | 4.15 | 3.80 | 3.58 | 3.41 | 3.29 | 3.20 | 3.12 | 3.06 | 2.96 | 2.86 | 2.76 | 2.70 | 2.64 | 2.59 | 2.52 | 2.46 | 2.40 |
| 16 | 6.12 | 4.69 | 4.08 | 3.73 | 3.50 | 3.34 | 3.22 | 3.12 | 3.05 | 2.99 | 2.89 | 2.79 | 2.68 | 2.63 | 2.57 | 2.51 | 2.45 | 2.38 | 2.32 |
| 17 | 6.04 | 4.62 | 4.01 | 3.66 | 3.44 | 3.28 | 3.16 | 3.06 | 2.98 | 2.92 | 2.82 | 2.72 | 2.62 | 2.56 | 2.50 | 2.44 | 2.38 | 2.32 | 2.25 |
| 18 | 5.98 | 4.56 | 3.95 | 3.61 | 3.38 | 3.22 | 3.10 | 3.01 | 2.93 | 2.87 | 2.77 | 2.67 | 2.56 | 2.50 | 2.44 | 2.38 | 2.32 | 2.26 | 2.19 |
| 19 | 5.92 | 4.51 | 3.90 | 3.56 | 3.33 | 3.17 | 3.05 | 2.96 | 2.88 | 2.82 | 2.72 | 2.62 | 2.51 | 2.45 | 2.39 | 2.33 | 2.27 | 2.20 | 2.13 |
| 20 | 5.87 | 4.46 | 3.86 | 3.51 | 3.29 | 3.13 | 3.01 | 2.91 | 2.84 | 2.77 | 2.68 | 2.57 | 2.46 | 2.41 | 2.35 | 2.29 | 2.22 | 2.16 | 2.09 |
| 21 | 5.83 | 4.42 | 3.82 | 3.48 | 3.25 | 3.09 | 2.97 | 2.87 | 2.80 | 2.73 | 2.64 | 2.53 | 2.42 | 2.37 | 2.31 | 2.25 | 2.18 | 2.11 | 2.04 |
| 22 | 5.79 | 4.38 | 3.78 | 3.44 | 3.22 | 3.05 | 2.93 | 2.84 | 2.76 | 2.70 | 2.60 | 2.50 | 2.39 | 2.33 | 2.27 | 2.21 | 2.14 | 2.08 | 2.00 |
| 23 | 5.75 | 4.35 | 3.75 | 3.41 | 3.18 | 3.02 | 2.90 | 2.81 | 2.73 | 2.67 | 2.57 | 2.47 | 2.36 | 2.30 | 2.24 | 2.18 | 2.11 | 2.04 | 1.97 |
| 24 | 5.72 | 4.32 | 3.72 | 3.38 | 3.15 | 2.99 | 2.87 | 2.78 | 2.70 | 2.64 | 2.54 | 2.44 | 2.33 | 2.27 | 2.21 | 2.15 | 2.08 | 2.01 | 1.94 |
| 25 | 5.69 | 4.29 | 3.69 | 3.35 | 3.13 | 2.97 | 2.85 | 2.75 | 2.68 | 2.61 | 2.51 | 2.41 | 2.30 | 2.24 | 2.18 | 2.12 | 2.05 | 1.98 | 1.91 |
| 26 | 5.66 | 4.27 | 3.67 | 3.33 | 3.10 | 2.94 | 2.82 | 2.73 | 2.65 | 2.59 | 2.49 | 2.39 | 2.28 | 2.22 | 2.16 | 2.09 | 2.03 | 1.95 | 1.88 |
| 27 | 5.63 | 4.24 | 3.65 | 3.31 | 3.08 | 2.92 | 2.80 | 2.71 | 2.63 | 2.57 | 2.47 | 2.36 | 2.25 | 2.19 | 2.13 | 2.07 | 2.00 | 1.93 | 1.85 |
| 28 | 5.61 | 4.22 | 3.63 | 3.29 | 3.06 | 2.90 | 2.78 | 2.69 | 2.61 | 2.55 | 2.45 | 2.34 | 2.23 | 2.17 | 2.11 | 2.05 | 1.98 | 1.91 | 1.83 |
| 29 | 5.59 | 4.20 | 1.61 | 3.27 | 3.04 | 2.88 | 2.76 | 2.67 | 2.59 | 2.53 | 2.43 | 2.32 | 2.21 | 2.15 | 2.09 | 2.03 | 1.96 | 1.89 | 1.81 |
| 30 | 5.57 | 4.18 | 3.59 | 3.25 | 3.03 | 2.87 | 2.75 | 2.65 | 2.57 | 2.51 | 2.41 | 2.31 | 2.20 | 2.14 | 2.07 | 2.01 | 1.94 | 1.87 | 1.79 |
| 40 | 5.42 | 4.05 | 3.46 | 3.13 | 2.90 | 2.74 | 2.62 | 2.53 | 2.45 | 2.39 | 2.29 | 2.18 | 2.07 | 2.01 | 1.94 | 1.88 | 1.80 | 1.72 | 1.64 |
| 60 | 5.29 | 3.93 | 3.34 | 3.01 | 2.79 | 2.63 | 2.51 | 2.41 | 2.33 | 2.27 | 2.17 | 2.06 | 1.94 | 1.88 | 1.82 | 1.74 | 1.67 | 1.58 | 1.48 |
| 120 | 5.15 | 3.80 | 3.23 | 2.89 | 2.67 | 2.52 | 2.39 | 2.30 | 2.22 | 2.16 | 2.05 | 1.94 | 1.82 | 1.76 | 1.69 | 1.61 | 1.53 | 1.43 | 1.31 |
| ∞ | 5.02 | 3.69 | 3.12 | 2.79 | 2.57 | 2.41 | 2.29 | 2.19 | 2.11 | 2.05 | 1.94 | 1.83 | 1.71 | 1.64 | 1.57 | 1.48 | 1.39 | 1.27 | 1.00 |

# Results of Our Example

$F = 10.9^2 / 9.6^2 = 118.81/92.16 = 1.29$

Critical value from the table = 2.41

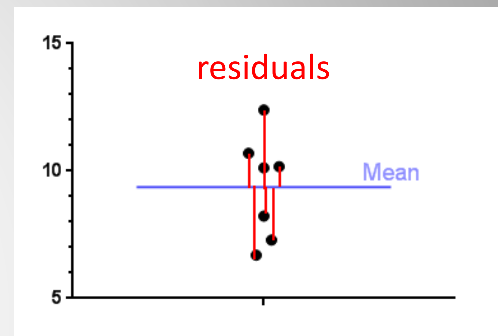
Because $1.29 < 2.41$, then $p > 0.05$ (if greater than critical value, $p \leq 0.05$)

Fail to reject the null and accept that the variances are the same

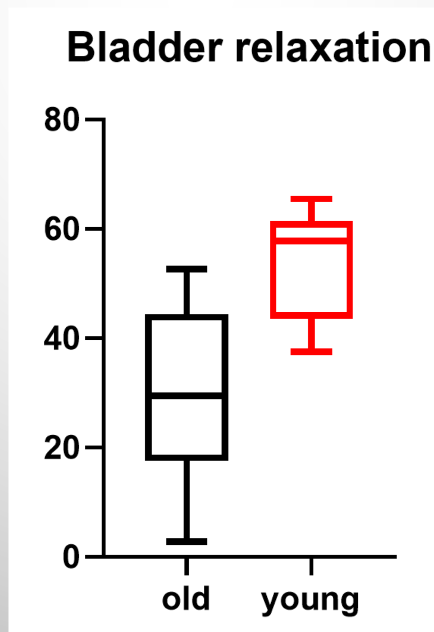
# Example using bladder relaxation data in old and young rats

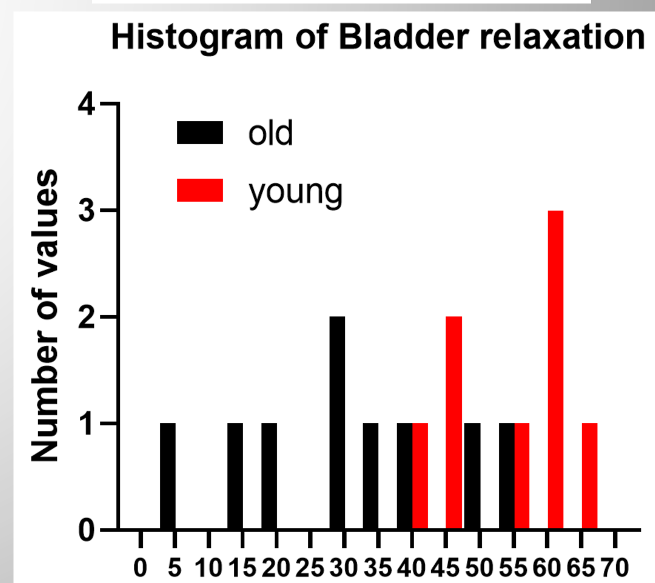
$H_0$: the means are equal
$H_A$: the means are not equal

Normality?
Homoscedasticity?
Independent?

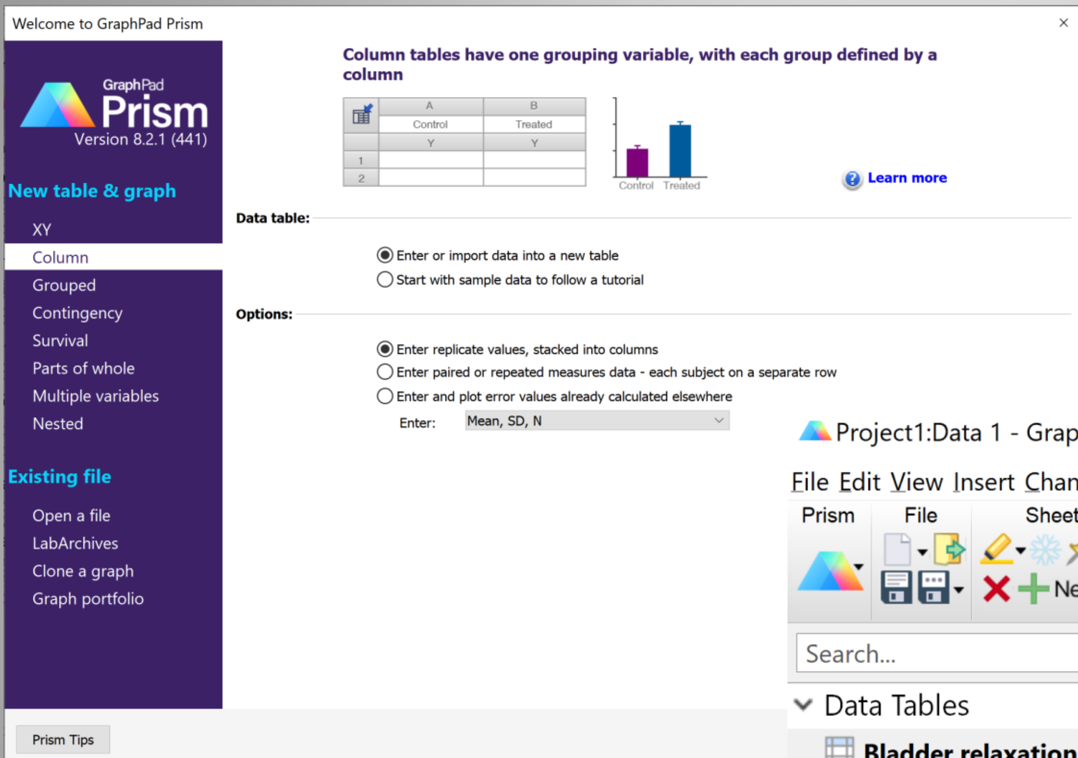


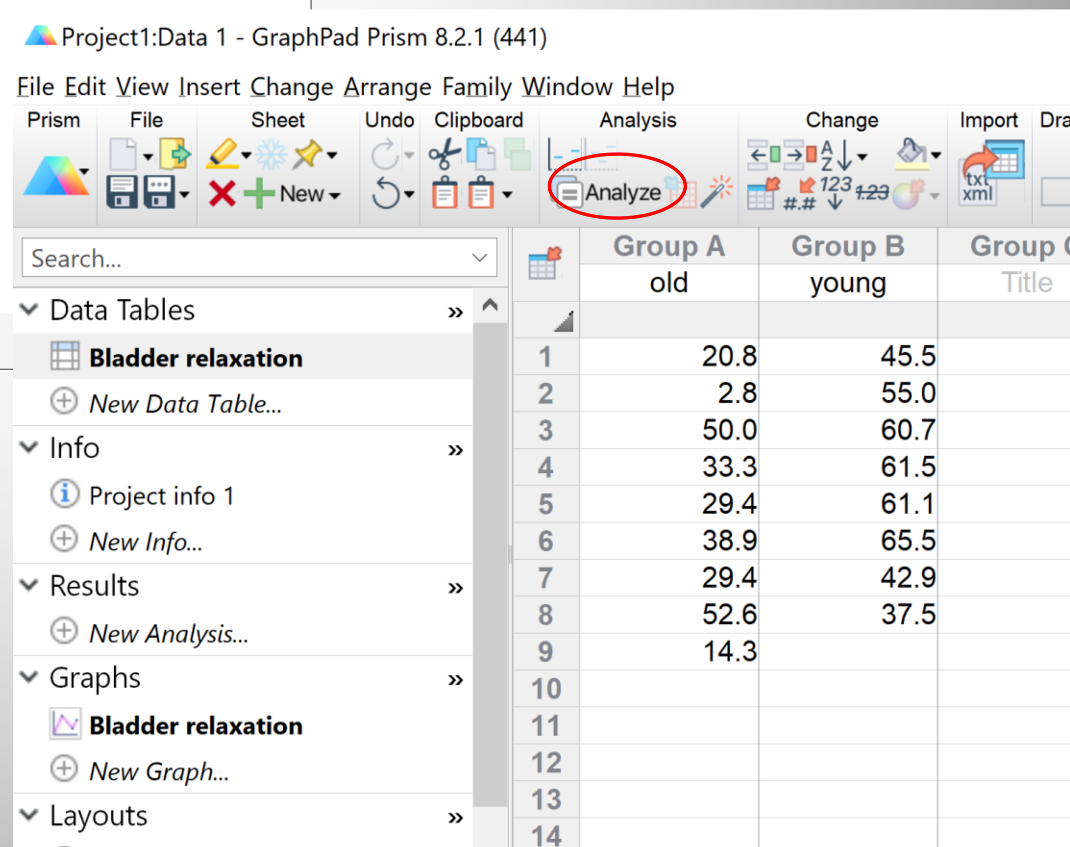

| | | A | B |
|---|---|---|---|
| | | old | young |
| 1 | Number of values | 9 | 8 |
| 2 | | | |
| 3 | Minimum | 2.800 | 37.50 |
| 4 | 25% Percentile | 17.55 | 43.55 |
| 5 | Median | 29.40 | 57.85 |
| 6 | 75% Percentile | 44.45 | 61.40 |
| 7 | Maximum | 52.60 | 65.50 |
| 8 | Range | 49.80 | 28.00 |
| 9 | | | |
| 10 | Mean | 30.17 | 53.71 |
| 11 | Std. Deviation | 16.09 | 10.36 |
| 12 | Std. Error of Mean | 5.365 | 3.664 |
| 13 | | | |
| 14 | Skewness | -0.2130 | -0.5515 |
| 15 | Kurtosis | -0.3964 | -1.463 |



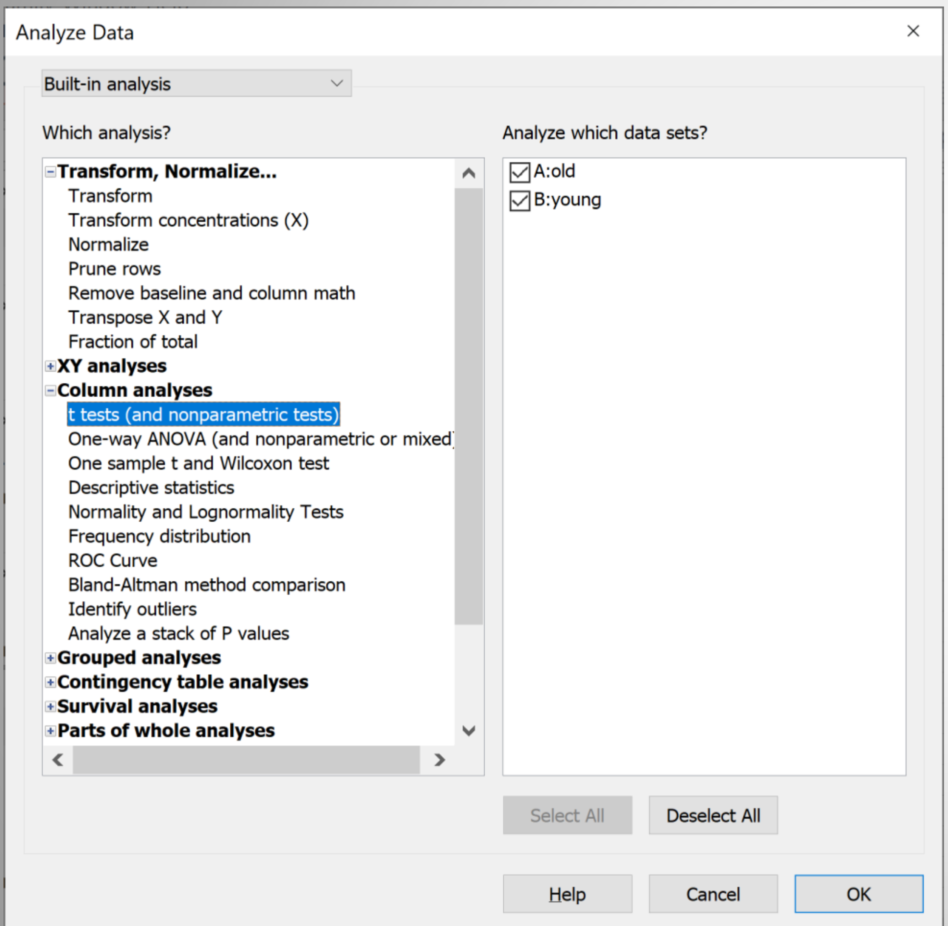
# Example using bladder relaxation data in old and young rats

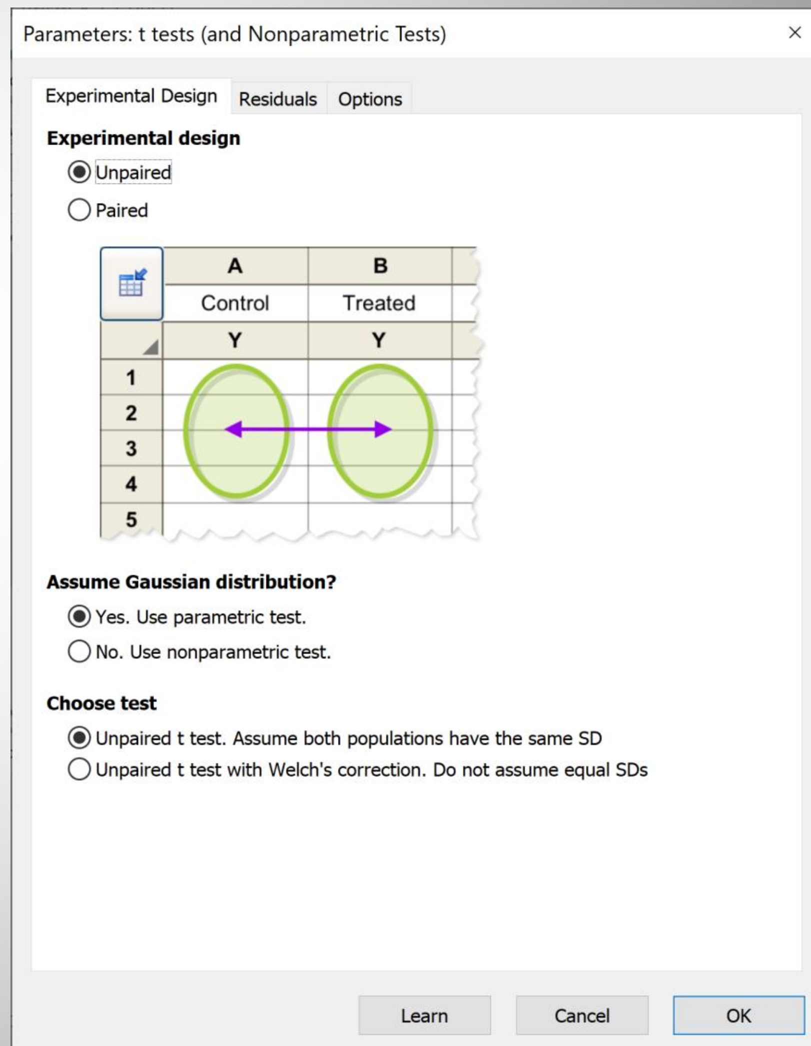


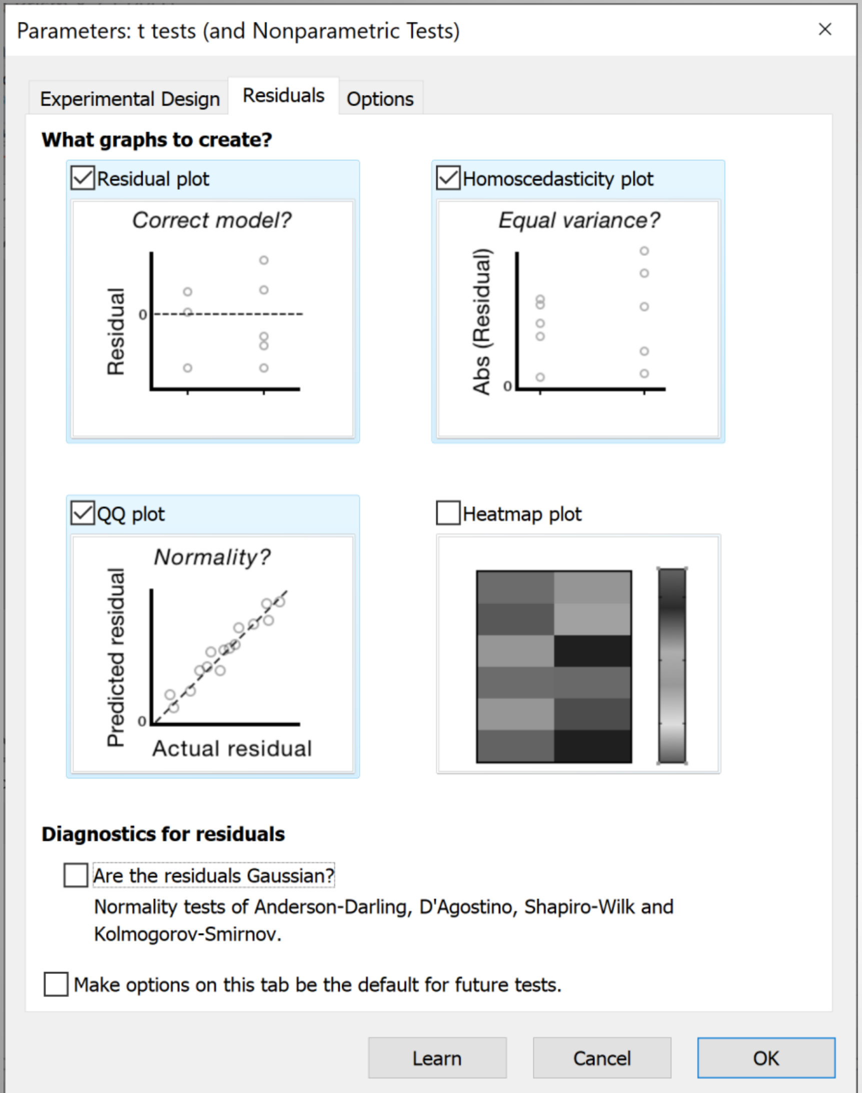



# T-test for differences in independent groups of continuous data With F Test for Equal Variances (Homoscedasticity)

## Unpaired (independent) t-test





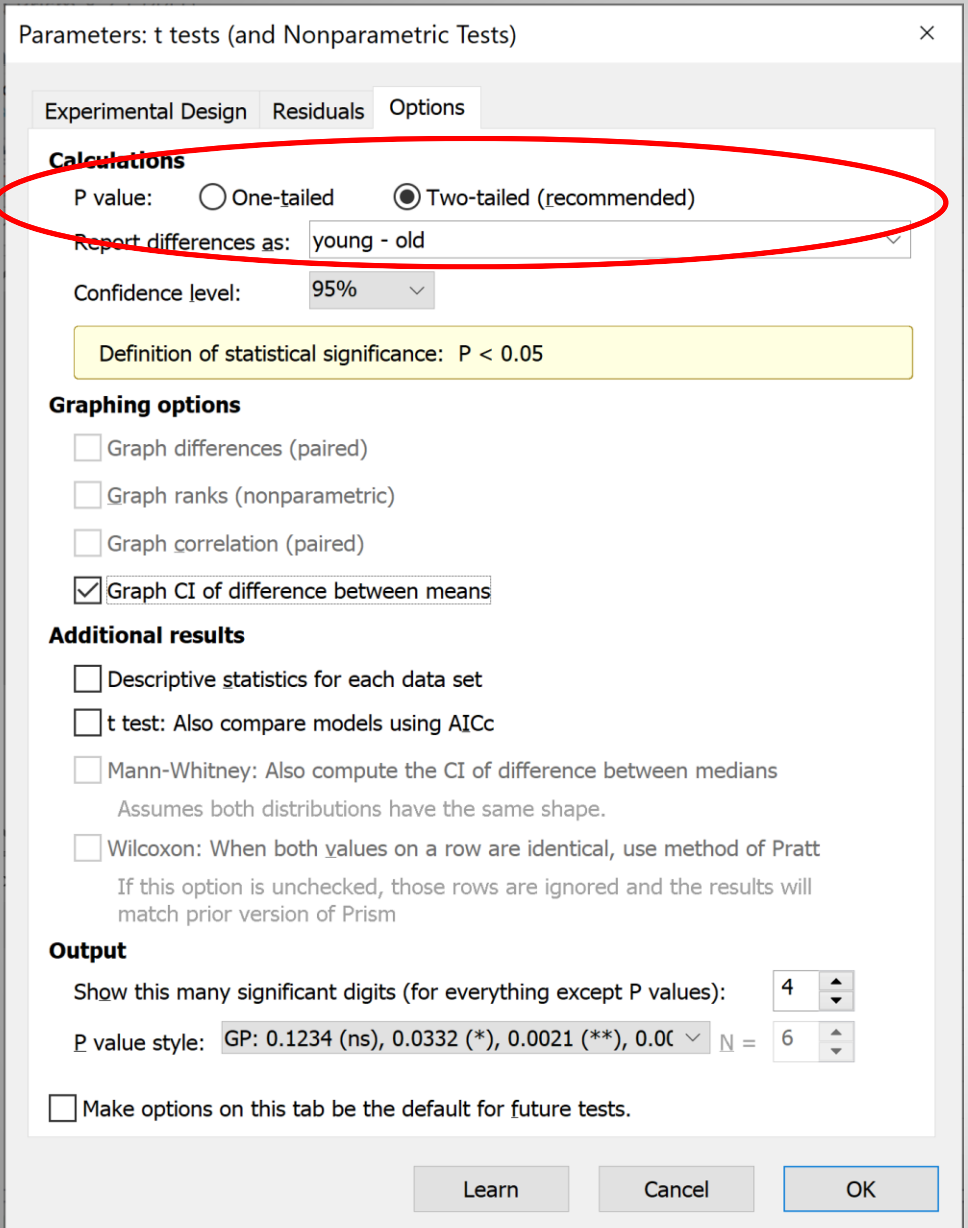

One of the assumptions of t-tests is that the *residuals* from a Gaussian distribution.

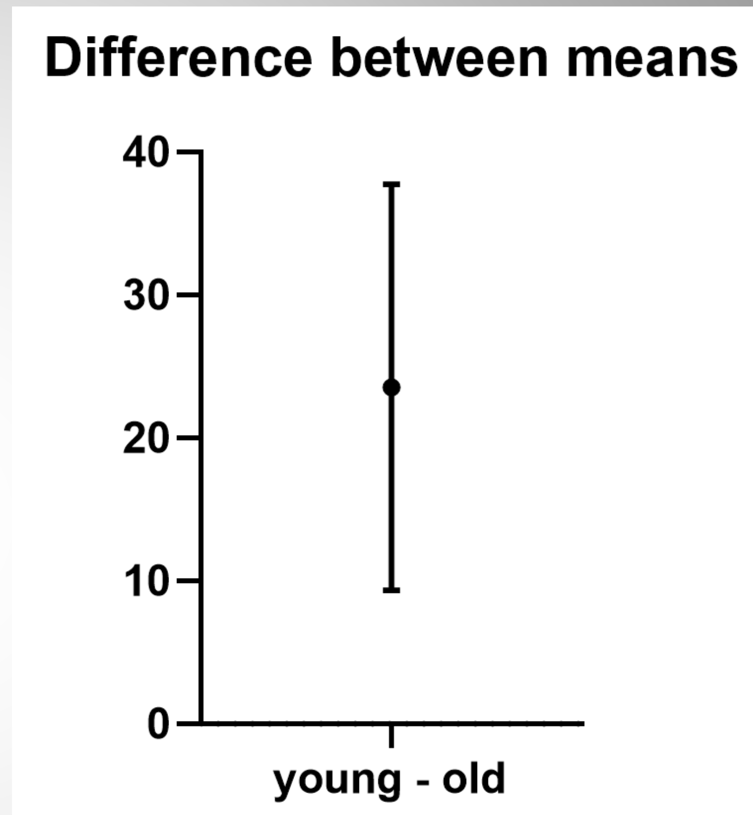
## Which residual graph to use?

**Residual plot**. The X axis is the actual value of the value (unpaired tests) or difference (paired test). The Y axis is the residual. This lets you spot residuals that are much larger or smaller than the rest.

**Homoscedasticity plot**. The X axis is the actual value of the value (unpaired tests) or difference (paired test). The Y axis is the absolute value of the residual.

**QQ plot**. The X axis is the actual residual. The Y axis is the predicted residual, computed from the percentile of the residual (among all residuals) and assuming sampling from a Gaussian distribution. Most useful for t-tests


Parameters: t tests (and Nonparametric Tests)
Experimental Design
Residuals
Options
Calculations
P value:
One-tailed
Two-tailed (recommended)
Report differences as:
young - old
Confidence level:
95%
Definition of statistical significance: P < 0.05
Graphing options
Graph differences (paired)
Graph ranks (nonparametric)
Graph correlation (paired)
Graph CI of difference between means
Additional results
Descriptive statistics for each data set
t test: Also compare models using AICc
Mann-Whitney: Also compute the CI of difference between medians
Assumes both distributions have the same shape.
Wilcoxon: When both values on a row are identical, use method of Pratt
If this option is unchecked, those rows are ignored and the results will match prior version of Prism
Output
Show this many significant digits (for everything except P values):
4
P value style:
GP: 0.1234 (ns), 0.0332 (*), 0.0021 (**), 0.0(
N =
6
Make options on this tab be the default for future tests.
Learn
Cancel
OK


Difference between means
40
30
20
10
0
young - old

**Unpaired t test**
Tabular results

| | | |
|---|---|---|
| 2 | | |
| 3 | Column B | young |
| 4 | vs. | vs. |
| 5 | Column A | old |
| 6 | | |
| 7 | **Unpaired t test** | |
| 8 | P value | 0.0030 |
| 9 | P value summary | ** |
| 10 | Significantly different (P < 0.05)? | Yes |
| 11 | One- or two-tailed P value? | Two-tailed |
| 12 | t, df | t=3.531, df=15 |
| 13 | | |
| 14 | **How big is the difference?** | |
| 15 | Mean of column A | 30.17 |
| 16 | Mean of column B | 53.71 |
| 17 | Difference between means (B - A) ± SEM | 23.55 ± 6.667 |
| 18 | 95% confidence interval | 9.335 to 37.76 |
| 19 | R squared (eta squared) | 0.4540 |
| 20 | | |
| 21 | **F test to compare variances** | |
| 22 | F, DFn, Dfd | 2.412, 8, 7 |
| 23 | P value | 0.2631 |
| 24 | P value summary | ns |
| 25 | Significantly different (P < 0.05)? | No |

Prism, unlike most statistics programs, reports a $R^2$ value. It quantifies the fraction of all the variation in the samples that is accounted for by a difference between the group means. If $R^2$=0.36, that means that 36% of all the variation among values is attributed to differences between the two group means, leaving 64% of the variation that comes from scatter among values within the groups.

## Critical values of Student's *t* distribution with *v* degrees of freedom

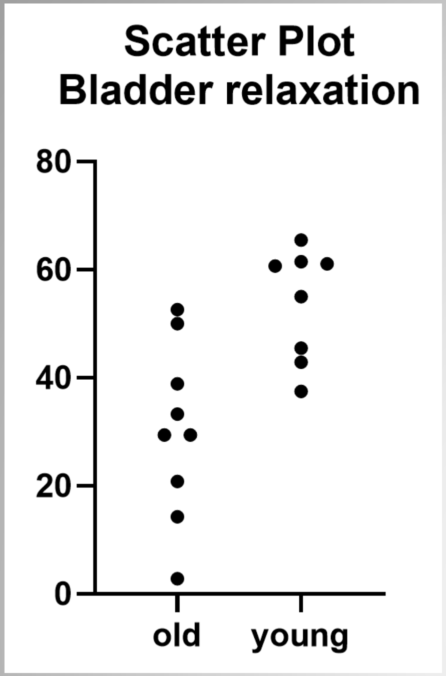
Probability less than the critical value ($t_{1-\alpha,v}$)

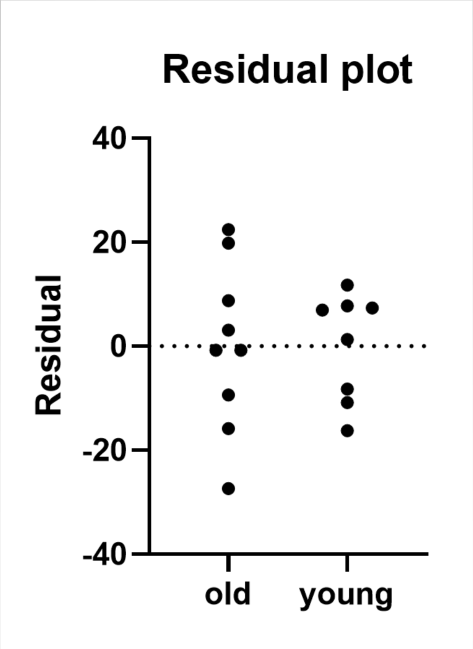
| v | 0.90 | 0.95 | 0.975 | 0.99 | 0.995 | 0.999 |
|---|---|---|---|---|---|---|
| 1. | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 | 318.313 |
| 2. | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 22.327 |
| 3. | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 10.215 |
| 4. | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173 |
| 5. | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.893 |
| 6. | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208 |
| 7. | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.782 |
| 8. | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 4.499 |
| 9. | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.296 |
| 10. | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.143 |
| 11. | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.024 |
| 12. | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.929 |
| 13. | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.852 |
| 14. | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.787 |
| 15. | 1.341 | 1.753 | $\boxed{2.131}$ | 2.602 | 2.947 | 3.733 |
| 16. | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.686 |
| 17. | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.646 |
| 18. | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.610 |
| 19. | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.579 |
| 20. | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.552 |
| 21. | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.527 |

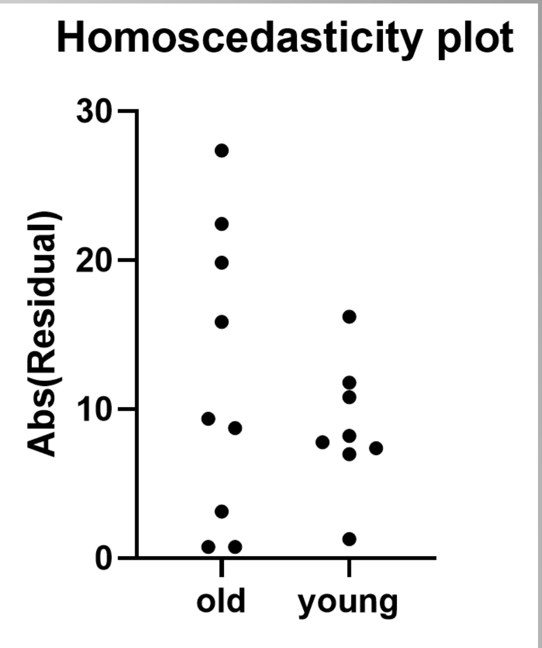
$v = df = (n_1-1) + (n_2-1)$

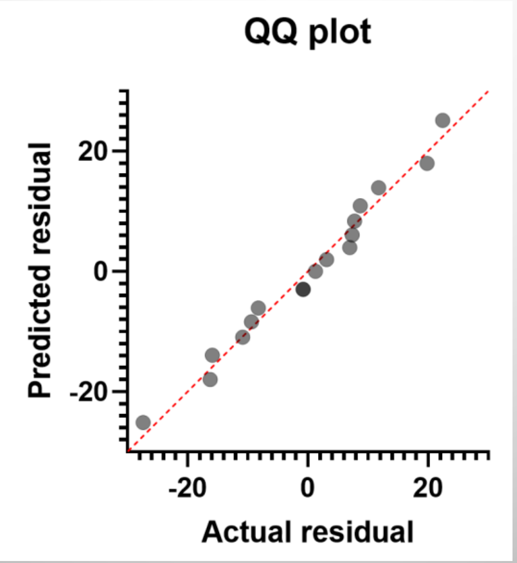
Alpha level for a two-tailed test: 0.05/2 = 0.025 or 0.975

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1^{\,2}}{N_1} + \frac{S_2^{\,2}}{N_2}}}$$


Scatter Plot
Bladder relaxation
80
60
40
20
0
old
young


Residual plot
Residual
40
20
0
-20
-40
old
young

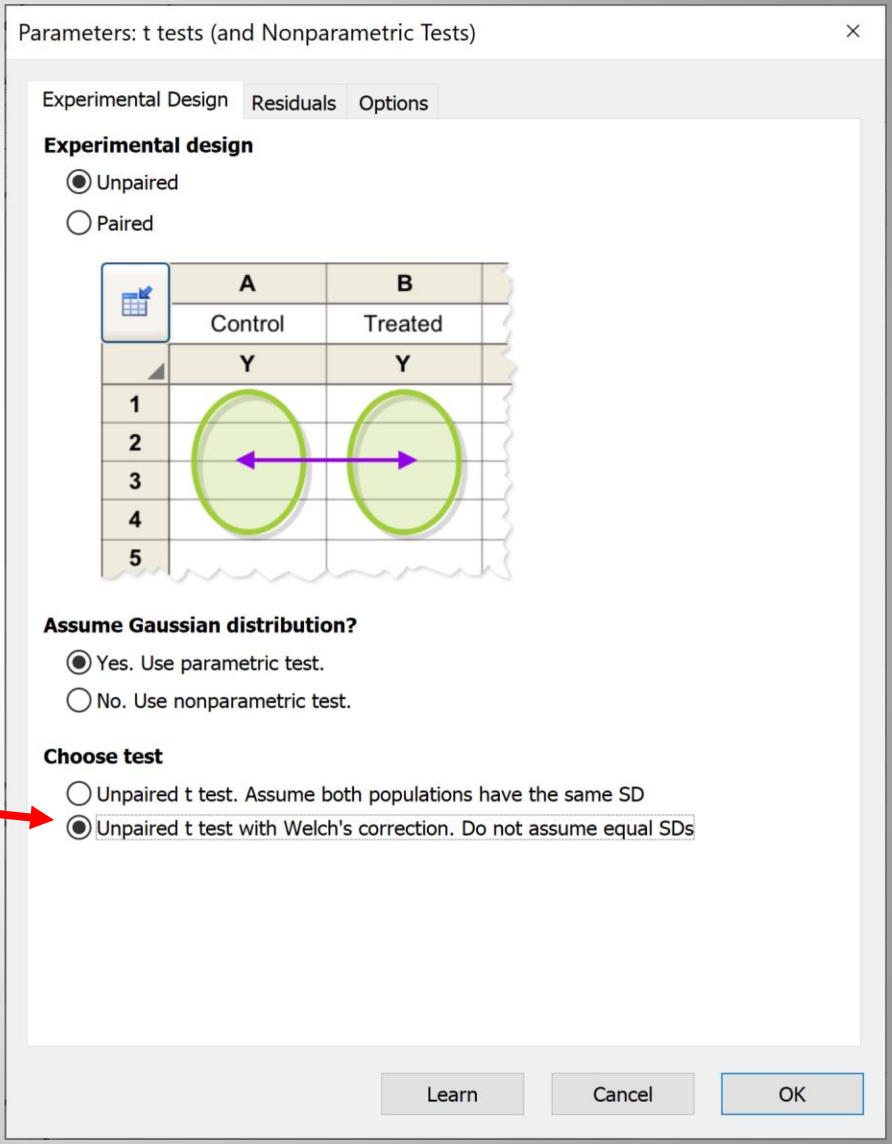

Homoscedasticity plot
Abs(Residual)
30
20
10
0
old
young


QQ plot
Predicted residual
20
0
-20
-20
0
20
Actual residual

# For unequal variances (heteroscedascity)

If the variances are too different, i.e. there is heteroscedasticity

Use “unequal variances” version of the t-test



**Unpaired t test**
Tabular results

| | | |
|---|---|---|
| 2 | | |
| 3 | Column B | young |
| 4 | vs. | vs. |
| 5 | Column A | old |
| 6 | | |
| 7 | **Unpaired t test** | |
| 8 | P value | 0.0030 |
| 9 | P value summary | ** |
| 10 | Significantly different (P < 0.05)? | Yes |
| 11 | One- or two-tailed P value? | Two-tailed |
| 12 | t, df | t=3.531, df=15 |
| 13 | | |
| 14 | **How big is the difference?** | |
| 15 | Mean of column A | 30.17 |
| 16 | Mean of column B | 53.71 |
| 17 | Difference between means (B - A) ± SEM | 23.55 ± 6.667 |
| 18 | 95% confidence interval | 9.335 to 37.76 |
| 19 | R squared (eta squared) | 0.4540 |
| 20 | | |
| 21 | **F test to compare variances** | |
| 22 | F, DFn, Dfd | 2.412, 8, 7 |
| 23 | P value | 0.2631 |
| 24 | P value summary | ns |
| 25 | Significantly different (P < 0.05)? | No |

| | | |
|---|---|---|
| 1 | Table Analyzed | Bladder relaxation |
| 2 | | |
| 3 | Column B | young |
| 4 | vs. | vs. |
| 5 | Column A | old |
| 6 | | |
| 7 | **Unpaired t test with Welch's correction** | |
| 8 | P value | 0.0028 |
| 9 | P value summary | ** |
| 10 | Significantly different (P < 0.05)? | Yes |
| 11 | One- or two-tailed P value? | Two-tailed |
| 12 | Welch-corrected t, df | t=3.624, df=13.78 |
| 13 | | |
| 14 | **How big is the difference?** | |
| 15 | Mean of column A | 30.17 |
| 16 | Mean of column B | 53.71 |
| 17 | Difference between means (B - A) ± SEM | 23.55 ± 6.497 |
| 18 | 95% confidence interval | 9.591 to 37.50 |
| 19 | R squared (eta squared) | 0.4881 |
| 20 | | |
| 21 | **F test to compare variances** | |
| 22 | F, DFn, Dfd | 2.412, 8, 7 |
| 23 | P value | 0.2631 |
| 24 | P value summary | ns |
| 25 | Significantly different (P < 0.05)? | No |

# What happens if your data violate the independent t test assumptions? (hint: biased results)

Assumption of independence for the sample values is violated

- The independent (unpaired) t-test is not appropriate.

Assumption of normality is violated, or outliers are present

- The t-test may not be the most powerful test available
- A non-parametric test or employing a transformation may result in a less biased answer

Assumption of homoscedasticity is violated

- The Welch-Satterthwaite t-test provides means of performing a t-test adjusted for the inequality of the variances
- Or transform the data or use a non-parametric test

# Example: The Independent t-test by hand
# Is there a difference in Test X by Gender?

| Males | Females |
|---|---|
| $\bar{X}_1 = 22.5$ | $\bar{X}_2 = 25.6$ |
| $SD_1 = 2.5$ | $SD_2 = 3.0$ |
| $N_1 = 9$ | $N_2 = 9$ |

For the purposes of this example, we have already determined that all test assumptions have been met.

$H_0$: the means are equal
$H_A$: the means are not equal

# Determining Critical Value

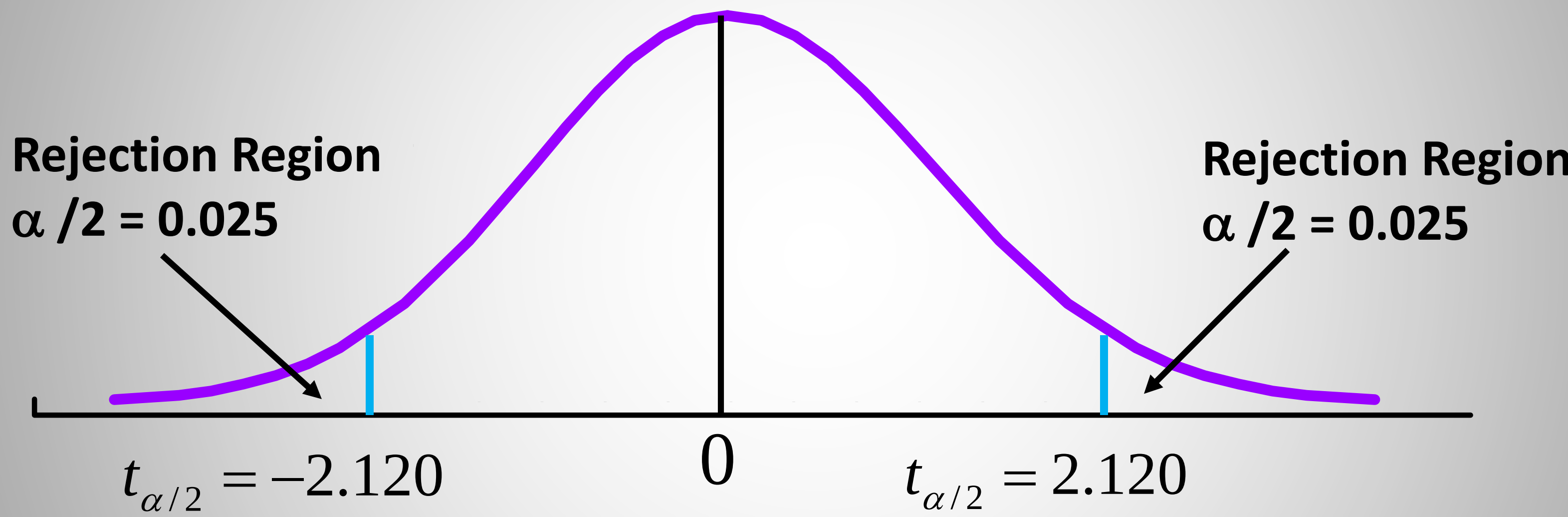
$$df = (n_1 - 1) + (n_2 - 1)$$

$$df = (9 - 1) + (9 - 1) = 16$$

| 1 tail α = | | 0.1 | 0.05 | 0.025 | 0.01 | 0.005 |
|---|---|---|---|---|---|---|
| 2 tails α = | | 0.2 | 0.1 | 0.05 | 0.02 | 0.01 |
| | df =1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.656 |
| | 2 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 |
| | 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 |
| | 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 |
| | 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 |
| | 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 |
| | 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 |
| | 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 |
| | 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 |
| | 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 |
| | 11 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 |
| | 12 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 |
| | 13 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 |
| | 14 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 |
| | 15 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 |
| | 16 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 |
| | 17 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 |
| | 18 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 |
| | 19 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 |
| | 20 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 |

# Hypothesis Tests for Two Means (two-tailed test, df=16)

$H_0$: the means are equal
$H_A$: the means are not equal




If $-2.120 \leq t \leq 2.120$, fail to reject $H_0$

# Example: The Independent t-test by hand
# Is there a difference in Test X by Gender?

| Males | Females |
|---|---|
| $\bar{X}_1 = 22.5$ | $\bar{X}_2 = 25.6$ |
| $SD_1 = 2.5$ | $SD_2 = 3.0$ |
| $N_1 = 9$ | $N_2 = 9$ |

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1^{\ 2}}{N_1} + \frac{S_2^{\ 2}}{N_2}}}$$

$\bar{x}_1 - \bar{x}_2 = 22.5 - 25.6 = -3.1$

$S_1^{\ 2} = 2.5^2 = 6.25$ $\quad$ $S_2^{\ 2} = 3.0^2 = 9$

$6.25/9 = 0.69$ $\quad$ $9/9 = 1$

$\sqrt{0.69 + 1} = \sqrt{1.69} = 1.3$

$-3.1/1.3 = -2.38$ = t statistic

df = (9-1) + (9-1) = 16

Critical value of t(16) = (+/-) 2.120 (two-tailed test)

-2.38 < - 2.120, then $p<0.05$ and we reject the null hypothesis of no differences between groups.

# What about confidence intervals?

95% confidence intervals for the independent (unpaired) t-test
(Confidence intervals for the difference in means)

$$Lower\ bound: (\bar{X}_1 - \bar{X}_2) - t_{\alpha/2} * \sqrt{\frac{s_1^2}{n_1} + \frac{s_{12}^2}{n_2}}$$

$$Upper\ bound: (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2} * \sqrt{\frac{s_1^2}{n_1} + \frac{s_{12}^2}{n_2}}$$

# Independent t-test Example with CI: Effect of Treatment on Test Y

| Placebo | Treatment |
|---|---|
| 18 | 22 |
| 21 | 25 |
| 16 | 17 |
| 22 | 24 |
| 19 | 16 |
| 24 | 29 |
| 17 | 20 |
| 21 | 23 |
| 23 | 19 |
| 18 | 20 |
| 14 | 15 |
| 16 | 15 |
| 16 | 18 |
| 19 | 26 |
| 18 | 18 |
| 20 | 24 |
| 12 | 18 |
| 22 | 25 |
| 15 | 19 |
| 17 | 16 |

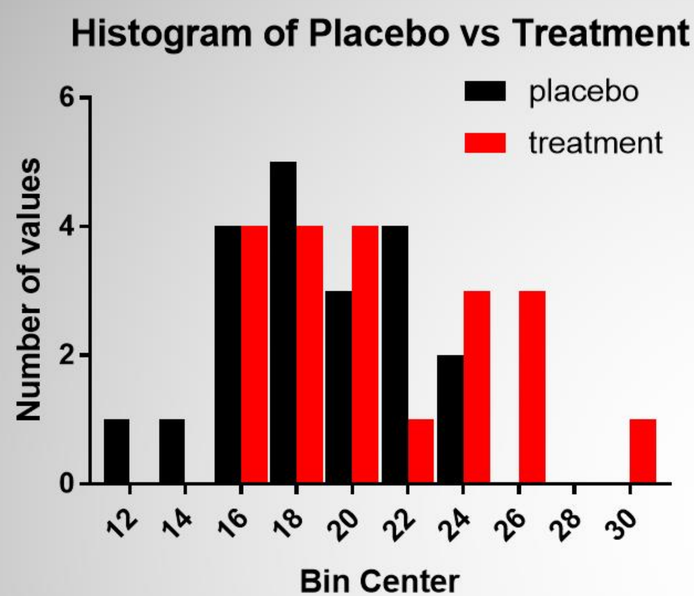
| | | |
|---|---|---|
| Mean | 18.40 | 20.45 |
| N | 20 | 20 |

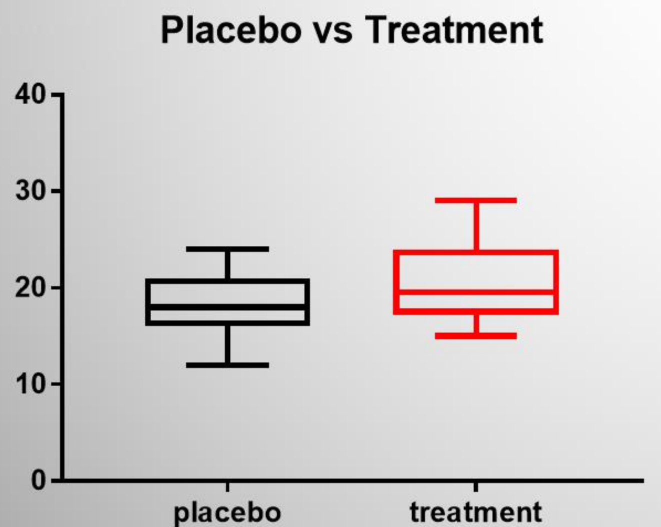
$H_0$: the means are equal
$H_A$: the means are not equal

Assumption of Independence

The observations are independent by treatment. The two groups are different people.

# Testing normality (symmetry)




| | | placebo | treatment |
|---|---|---|---|
| | | Y | Y |
| 1 | Number of values | 20 | 20 |
| 2 | | | |
| 3 | Minimum | 12 | 15 |
| 4 | 25% Percentile | 16 | 17.25 |
| 5 | Median | 18 | 19.5 |
| 6 | 75% Percentile | 21 | 24 |
| 7 | Maximum | 24 | 29 |
| 8 | | | |
| 9 | Mean | 18.4 | 20.45 |
| 10 | Std. Deviation | 3.152 | 4.058 |
| 11 | Std. Error of Mean | 0.7049 | 0.9075 |
| 12 | | | |
| 13 | Skewness | -0.05198 | 0.4398 |
| 14 | Kurtosis | -0.4966 | -0.7904 |
| 15 | | | |
| 16 | D'Agostino & Pearson normality test | | |
| 17 | K2 | 0.1646 | 1.545 |
| 18 | P value | 0.9210 | 0.4619 |
| 19 | Passed normality test (alpha=0.05)? | Yes | Yes |
| 20 | P value summary | ns | ns |
| 21 | | | |

| Col. stats | | A placebo | B treatment |
|---|---|---|---|
| 1 | Number of values | 20 | 20 |
| 2 | | | |
| 3 | Minimum | 12.00 | 15.00 |
| 4 | 25% Percentile | 16.00 | 17.25 |
| 5 | Median | 18.00 | 19.50 |
| 6 | 75% Percentile | 21.00 | 24.00 |
| 7 | Maximum | 24.00 | 29.00 |
| 8 | | | |
| 9 | Mean | 18.40 | 20.45 |
| 10 | Std. Deviation | 3.152 | 4.058 |
| 11 | Std. Error of Mean | 0.7049 | 0.9075 |
| 12 | | | |
| 13 | Skewness | -0.05198 | 0.4398 |
| 14 | Kurtosis | -0.4966 | -0.7904 |
| 15 | | | |
| 16 | **D'Agostino & Pearson normality test** | | |
| 17 | K2 | 0.1646 | 1.545 |
| 18 | P value | 0.9210 | 0.4619 |
| 19 | Passed normality test (alpha=0.05)? | Yes | Yes |
| 20 | P value summary | ns | ns |
| 21 | | | |
| 22 | **Shapiro-Wilk normality test** | | |
| 23 | W | 0.9820 | 0.9423 |
| 24 | P value | 0.9569 | 0.2654 |
| 25 | Passed normality test (alpha=0.05)? | Yes | Yes |
| 26 | P value summary | ns | ns |

# Testing Homoscedasticity

| | SD |
|---|---|
| Treatment SD | 4.058 |
| Placebo | 3.152 |
| *Ratio* | *1.28* |
| *Rule of thumb* | *<2* |

F test

| | | |
|---|---|---|
| 21 | F test to compare variances | |
| 22 | F,DFn, Dfd | 1.658, 19, 19 |
| 23 | P value | 0.2795 |
| 24 | P value summary | ns |
| 25 | Significantly different? (P < 0.05) | No |
| 26 | | |
| 27 | | |

F test p = 0.28, fail to reject the null of no differences in variances and conclude the variances are equal

| | |
|---|---|
| Normal/Symmetrical | Yes |
| Equal Variances | Yes |
| Independent | Yes |
| 2 groups of continuous data | Yes |

Independent t-test

**Unpaired t test**
Tabular results

| | | |
|---|---|---|
| 1 | Table Analyzed | Placebo vs Treatment |
| 2 | | |
| 3 | Column B | treatment |
| 4 | vs. | vs. |
| 5 | Column A | placebo |
| 6 | | |
| 7 | **Unpaired t test** | |
| 8 | P value | 0.0824 |
| 9 | P value summary | ns |
| 10 | Significantly different (P < 0.05)? | No |
| 11 | One- or two-tailed P value? | Two-tailed |
| 12 | t, df | t=1.784, df=38 |
| 13 | | |
| 14 | **How big is the difference?** | |
| 15 | Mean of column A | 18.40 |
| 16 | Mean of column B | 20.45 |
| 17 | Difference between means (B - A) ± SEM | 2.050 ± 1.149 |
| 18 | 95% confidence interval | -0.2762 to 4.376 |
| 19 | R squared (eta squared) | 0.07728 |
| 20 | | |
| 21 | **F test to compare variances** | |
| 22 | F, DFn, Dfd | 1.658, 19, 19 |
| 23 | P value | 0.2795 |
| 24 | P value summary | ns |
| 25 | Significantly different (P < 0.05)? | No |
| 26 | | |

## WRITE UP RESULTS

For this class:

The data for both groups passed the assumptions for an independent t-test (independence, normality/symmetry, homoscedasticity). With p=0.08 (two tailed independent t-test, t=1.78, df=38, $\alpha$=0.05), we fail to reject the null hypothesis of no difference in mean values of Test Y in placebo (18.4) and treatment groups (20.5). Further, the 95%CI for the difference in means contained 0 (-0.28 to 4.38). Therefore, we conclude treatment has no effect on the value of Test Y.

What people may usually write:

We conclude treatment has no effect on the value of Test Y (p=0.08).

what is the type of the test and what is the result

## Box 2. Checklist for clear statistical reporting for t-tests and ANOVAs

Basic biomedical science studies often include several small experiments. Providing information about statistical tests in the legend of each table or figure makes it easy for readers to determine what test was performed for each set of data and confirm that the test is appropriate.

**t-tests**

# State whether the test was unpaired (for comparing independent groups) or paired (for non-independent data, including repeated measurements on the same individual or matched participants, specimens or samples).
# State whether the test assumed equal or unequal variance between groups.
# Report the t-statistic, degrees of freedom and exact p-value.
# To focus on the magnitude of the difference, it is strongly recommended to report effect sizes with confidence intervals.

**ANOVAs**

# Specify the number of factors included in the ANOVA (i.e., one- vs. two-way ANOVA).
# For each factor, specify the name and level of the factor and state whether the factor was entered as a within-subjects (i.e., independent) factor or as a between-subjects (i.e., non-independent) factor.
# If the ANOVA has two or more factors, specify whether the interaction term was included. If the ANOVA has three or more factors and includes interaction terms, specify which interaction terms were included.
# Report the F-statistic, degrees of freedom and exact p-value for each factor or interaction.
# Specify if a post-hoc test was performed. If post-hoc tests were performed, specify the type of post-hoc test and, if applicable, the test statistic and p-value.
# To focus on the magnitude of the difference, it is strongly recommended to report effect sizes with confidence intervals.

DOI: https://doi.org/10.7554/eLife.36163.005

Weissgerber et al. eLife 2018;7:e36163. DOI: https://doi.org/10.7554/eLife.36163

# How do you interpret the confidence intervals?

Confidence intervals for the independent t-test (difference in mean)

the confidence interval is a range of likely values for the difference in population means

the interval either contains the population value or it does not

| How big is the difference? | |
|---|---|
| Mean of column A | 18.40 |
| Mean of column B | 20.45 |
| Difference between means (B - A) ± SEM | 2.050 ± 1.149 |
| 95% confidence interval | -0.2762 to 4.376 |

Although our estimate of the difference in means is 2.05, because the 95%CI includes the null value (0; no difference), we conclude there is no difference in groups (based on our sample)

# What about the *non-parametric* tests?

Nonparametric tests make no assumptions about the distribution of the populations

Use **ranks** of the data values from low to high and analyzes the value of the ranks, ignoring the actual values of the data

They can be useful when:

Sample sizes are small and data distribution cannot be determined

Sample sizes are larger, but the data severely violate the normality assumption of standard statistical tests, *e.g.*, data are heavily skewed or have “outliers”

Nonparametric tests can still depend on other assumptions

e.g., Data are independent

# Properties of parametric and non-parametric datasets

| | **Parametric** | **Non-parametric** |
|---|---|---|
| Distribution | Assumed to be Gaussian (normal) | Any |
| Variance | Assumed to be homoscedastic | Any |
| Data type | Continuous | Any also for ordinal data |
| | | |
| Central measure | Mean | Distribution (or median in specific circumstances) |

# But, use parametric tests where possible!

Parametric tests usually have a higher statistical power than non-parametric tests and are thus likelier to avoid a false negative result (type 2 error)

Choices of non-parametric tests are limited: Compared to parametric methods, there are only a few non-parametric tests

The information in the data is not fully utilized in a non-parametric test

Confidence intervals from parametric tests are easier to interpret

# What Happens If the Wrong Type of Test Is Used?

When parametric test assumptions are met

parametric methods will have more power to detect differences than non-parametric methods

When parametric test assumptions are violated

parametric methods will lack power

analyzing data not normally distributed with parametric tests will increase Type II error (concluding no difference when there is one)

# What if I choose the 'wrong' test?
# Relationship to sample size

| Distribution | Test | Small Sample | Large Sample |
|---|---|---|---|
| Gaussian | Nonparametric | Nonparametric tests may have much lower power than parametric tests to detect differences with small sample sizes. P values will be too large. | Little problem. With large samples, nonparametric tests are nearly as powerful as parametric tests. P values will be very slightly larger. |
| Non-Gaussian | Parametric | Misleading. With small sample sizes, parametric tests are not very robust to violations of the normality assumption. P values may be inaccurate. | Little problem. With large sample sizes, parametric tests are robust to violations of the normality assumption. |

# Some notes…

Don’t do both parametric and non-parametric tests and pick up the result with the smaller p-value

Don’t refer to the data as “non-parametric”. Data are not “parametric” or “non-parametric”, the method or the test is

# Some commonly used parametric tests and their non-parametric equivalents

They do not compare the median, but the distribution

| | **Parametric test** | **Non-parametric test** |
|---|---|---|
| Correlation | Pearson | Spearman |
| One sample, Compare to another | One-sample t-test | Wilcoxon signed rank test |
| Two groups, independent data | Independent t –test | Mann-Whitney *U* test (the same as the Wilcoxon rank-sum test) |
| Paired data | Paired t-test | Sign test, Wilcoxon signed rank test |
| >two groups, independent data | One-way analysis of variance (ANOVA) | Kruskal-Wallis test |
| | | |

# Independent t-test non-parametric equivalent: Mann-Whitney U (Wilcoxon rank sum)

the Mann-Whitney test does not compare medians of the two groups but the distribution and location of ranks of the data in each group

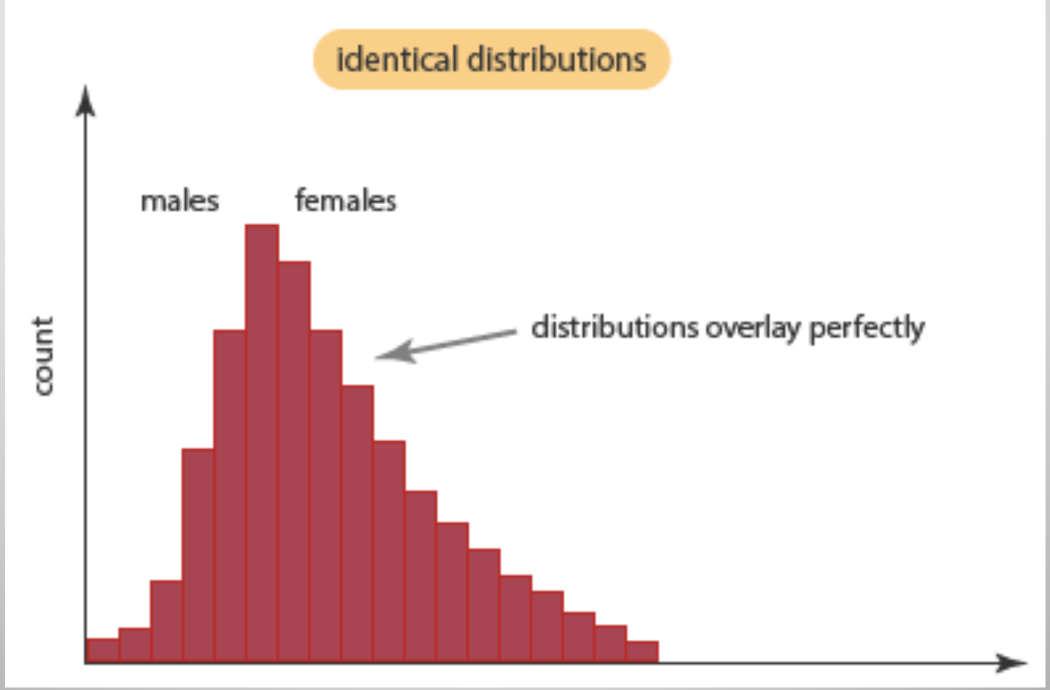
Data are grouped together and ranked from smallest to largest
Ranks for each group are summed and adjusted for sample size

Assumptions: Observations are randomly sampled from a larger population and each value is obtained independently (iid)

# Independent t-test non-parametric equivalent: Mann-Whitney U (Wilcoxon rank sum)

Question: Are the rank scores of the two populations the same or different in terms of distribution and/or location (median)

$H_0$: the distribution of rank scores for the two groups are the same
$H_A$: the distribution of rank scores for the two groups are not the same




statistics.laerd.com

# Independent t-test non-parametric equivalent: Mann-Whitney U (Wilcoxon rank sum)

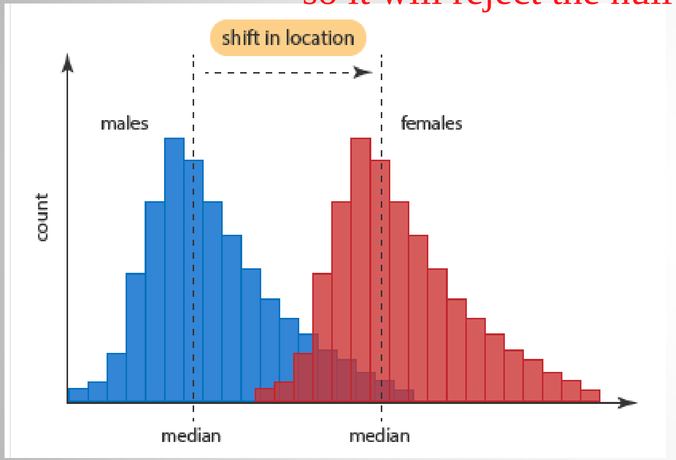
those two are the same

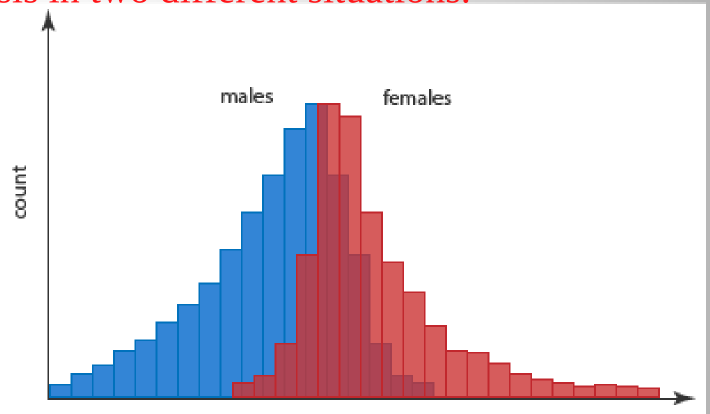
## Different distributions

Same shape, different median

Different shape, same median

so it will reject the null hypothesis in two different situations.




If p≤0.05 and the medians of the actual data (not ranks) are different, conclude that the medians are different.

If p≤0.05 and the medians of the actual data (not ranks) are the dsame, conclude that the distributions are different.

statistics.laerd.com

# Procedure for the Mann-Whitney U test

- Rank all the values in the dataset without paying attention to the group from which the value is drawn. If there are ties, assign the tied values the average of the ranks.
- Sum the ranks in the group 1 and name it $R_1$ and do the same for group 2 to get $R_2$.
- Calculate $U_1 = R_1 - \frac{n_1(n_1+1)}{2}$ and $U_2 = R_2 - \frac{n_2(n_2+1)}{2}$.
- The smaller of $U_1$ and $U_2$ is Mann-Whitney U statistic.
- Use Mann-Whitney test table to get the p-value or use Prism!

Old Rat Young Rat Data

# Compute Mann-Whitney statistic by hand

- Old: 20.8, 2.8, 50.0, 33.3, 29.4, 38.9, 29.4, 52.6,14.3.
- Young: 45.5, 55.0, 60.7, 61.5, 61.1, 65.5, 42.9, 37.5.
- Ranks:
  - Old: 3, 1, 11, 6, 4.5, 8, 4.5, 12, 2.
  - Young: 10, 13, 14, 16, 15, 17, 9, 7.
- $R_1 =$ 3+1+11+6+4.5+8+4.5+12+2=52
- $R_2 =$ 0+13+14+16+15+17+9+7=101

| Rank | Value |
|---|---|
| 1 | 2.8 |
| 2 | 14.3 |
| 3 | 20.8 |
| 4 | 29.4 |
| 5 | 29.4 |
| 6 | 33.3 |
| 7 | 37.5 |
| 8 | 38.9 |
| 9 | 42.9 |
| 10 | 45.5 |
| 11 | 50.0 |
| 12 | 52.6 |
| 13 | 55.0 |
| 14 | 60.7 |
| 15 | 61.1 |
| 16 | 61.5 |
| 17 | 65.5 |

Tie

# Compute Mann-Whitney statistic by hand

- $R_1 = 52, \quad R_2 = 101.$
- $n_1 = 9, n_2 = 8.$

- Therefore, $U_1 = 52 - \frac{9(9+1)}{2} = 7$ and $U_2 = 101 - \frac{8(8+1)}{2} = 65.$
- Mann-Whitney U-statistic is the smaller one, i.e. 7.

Table 3 Critical values of $U$ (5% significance).

| $n_1$ \ $n_2$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | |
| 2 | | | | | | | | 0 | 0 | 0 | 0 |
| 3 | | | | | 0 | 1 | 1 | 2 | 2 | 3 | 3 |
| 4 | | | | 0 | 1 | 2 | 3 | 4 | 4 | 5 | 6 |
| 5 | | | 0 | 1 | 2 | 3 | 5 | 6 | 7 | 8 | 9 |
| 6 | | | 1 | 2 | 3 | 5 | 6 | 8 | 10 | 11 | 13 |
| 7 | | | 1 | 3 | 5 | 6 | 8 | 10 | 12 | 14 | 16 |
| 8 | | 0 | 2 | 4 | 6 | 8 | 10 | 13 | 15 | 17 | 19 |
| 9 | | 0 | 2 | 4 | 7 | 10 | 12 | 15 | 17 | 20 | 23 |
| 10 | | 0 | 3 | 5 | 8 | 11 | 14 | 17 | 20 | 23 | 26 |
| 11 | | 0 | 3 | 6 | 9 | 13 | 16 | 19 | 23 | 26 | 30 |

Reject $H_0$ if $U_{smallest} \leq$ critical value

Reject $H_0$ if $U \leq 15$.

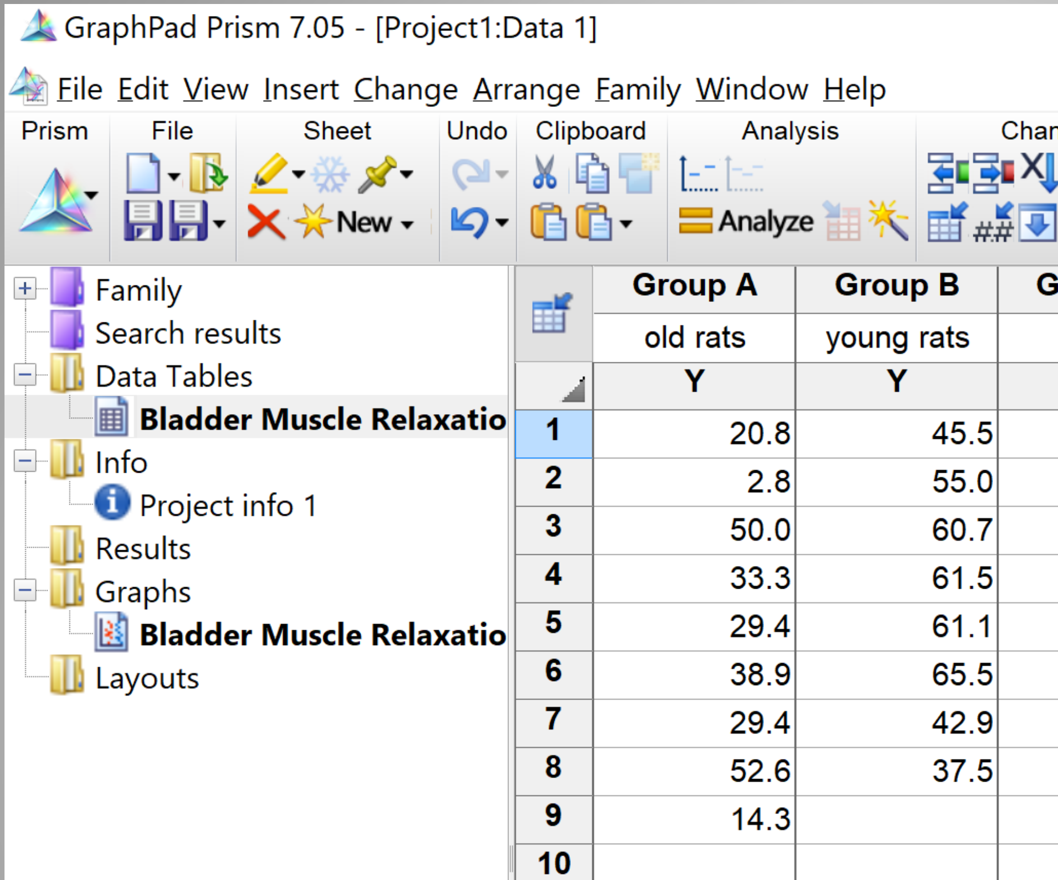
Fail to reject $H_0$ if $U > 15$.

What is the value of the U statistic?

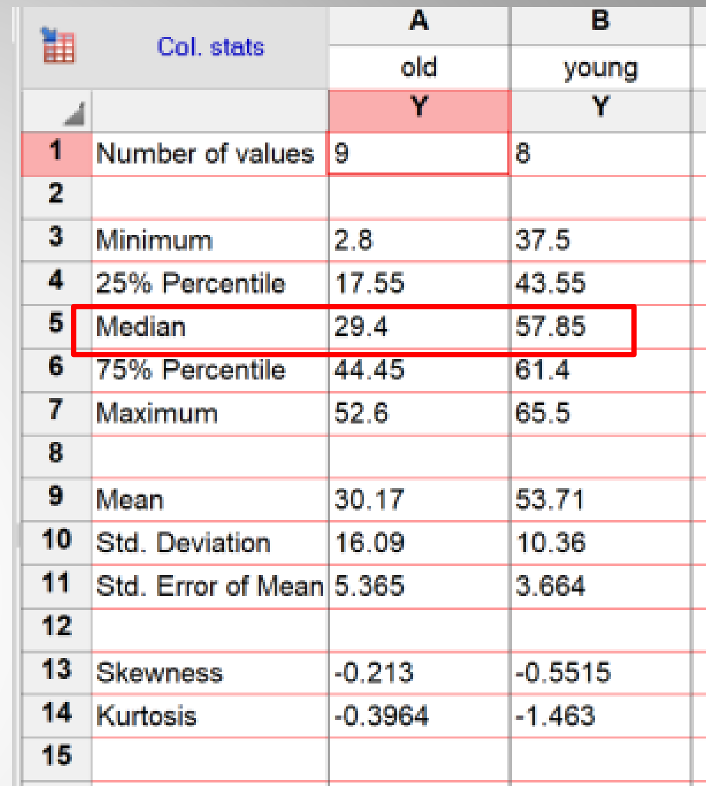
Do we reject or fail to reject the null hypothesis?

Can you conclude that the medians are different?

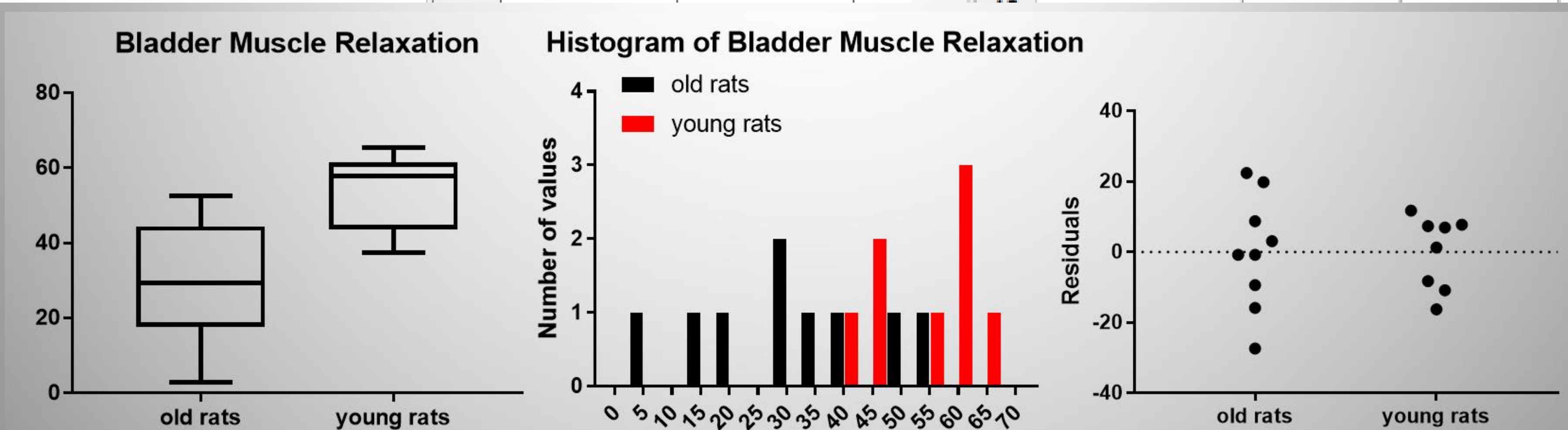
We saw this earlier…



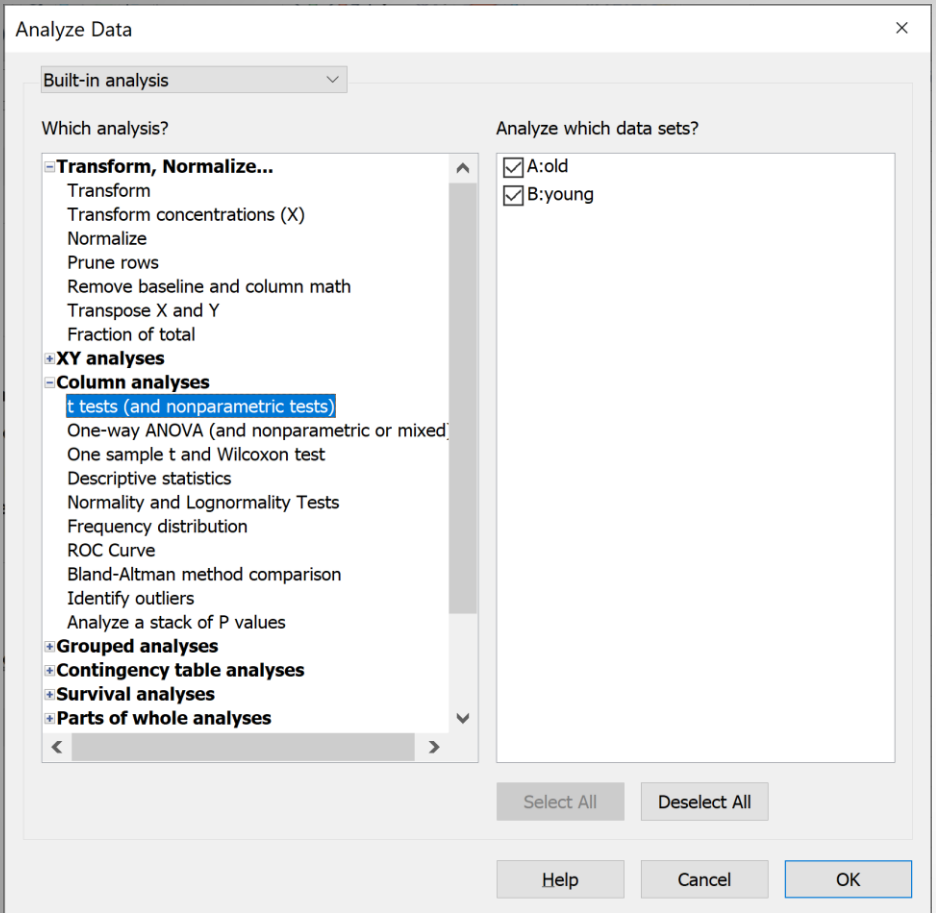

| | Group A | Group B |
|---|---|---|
| | old rats | young rats |
| | Y | Y |
| 1 | 20.8 | 45.5 |
| 2 | 2.8 | 55.0 |
| 3 | 50.0 | 60.7 |
| 4 | 33.3 | 61.5 |
| 5 | 29.4 | 61.1 |
| 6 | 38.9 | 65.5 |
| 7 | 29.4 | 42.9 |
| 8 | 52.6 | 37.5 |
| 9 | 14.3 | |
| 10 | | |

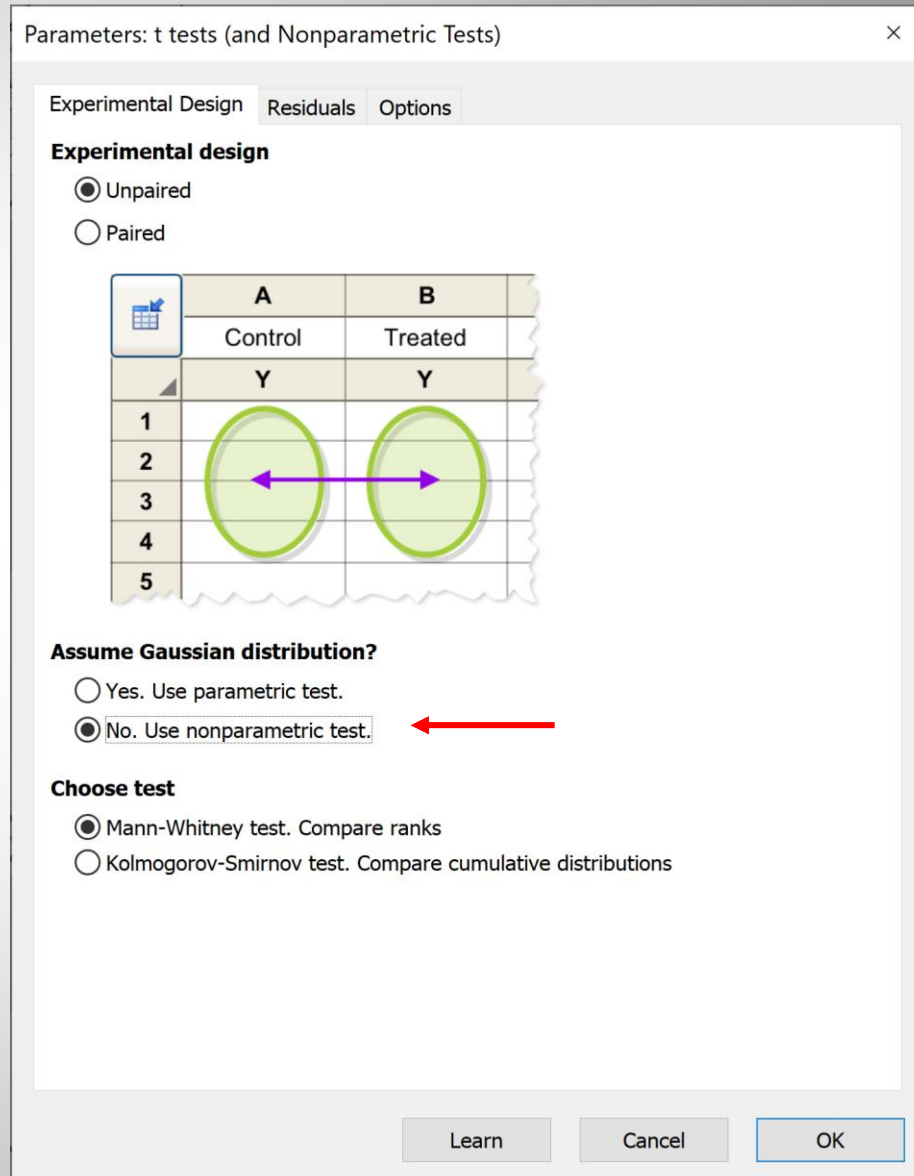


| | Col. stats | A | B |
|---|---|---|---|
| | | old | young |
| | | Y | Y |
| 1 | Number of values | 9 | 8 |
| 2 | | | |
| 3 | Minimum | 2.8 | 37.5 |
| 4 | 25% Percentile | 17.55 | 43.55 |
| 5 | Median | 29.4 | 57.85 |
| 6 | 75% Percentile | 44.45 | 61.4 |
| 7 | Maximum | 52.6 | 65.5 |
| 8 | | | |
| 9 | Mean | 30.17 | 53.71 |
| 10 | Std. Deviation | 16.09 | 10.36 |
| 11 | Std. Error of Mean | 5.365 | 3.664 |
| 12 | | | |
| 13 | Skewness | -0.213 | -0.5515 |
| 14 | Kurtosis | -0.3964 | -1.463 |
| 15 | | | |



To do the Mann-Whitney U test, start with the same Column table as for the t-test



Parameters: t tests (and Nonparametric Tests)

Experimental Design | Residuals | Options

**Calculations**

P value: ○ One-tailed ◉ Two-tailed (recommended)

Report differences as: young - old

Confidence level: 95%

Definition of statistical significance: P < 0.05

**Graphing options**

☐ Graph differences (paired)

☐ Graph ranks (nonparametric)

☐ Graph correlation (paired)

☑ Graph CI of difference between medians

**Additional results**

☐ Descriptive statistics for each data set

☐ t test: Also compare models using AICc

☑ Mann-Whitney: Also compute the CI of difference between medians

Assumes both distributions have the same shape.

☐ Wilcoxon: When both values on a row are identical, use method of Pratt

If this option is unchecked, those rows are ignored and the results will match prior version of Prism

**Output**

Show this many significant digits (for everything except P values): 4

P value style: GP: 0.1234 (ns), 0.0332 (*), 0.0021 (**), 0.0( N = 6

☐ Make options on this tab be the default for future tests.

Learn | Cancel | OK

Mann-Whitney test
Tabular results

less power for the test

| | | |
|---|---|---|
| 2 | | |
| 3 | Column B | young |
| 4 | vs. | vs. |
| 5 | Column A | old |
| 6 | | |
| 7 | **Mann Whitney test** | |
| 8 | P value | 0.0035 |
| 9 | Exact or approximate P value? | Exact |
| 10 | P value summary | ** |
| 11 | Significantly different (P < 0.05)? | Yes |
| 12 | One- or two-tailed P value? | Two-tailed |
| 13 | Sum of ranks in column A,B | 52 , 101 |
| 14 | Mann-Whitney U | 7 |
| 15 | | |
| 16 | **Difference between medians** | |
| 17 | Median of column A | 29.40, n=9 |
| 18 | Median of column B | 57.85, n=8 |
| 19 | Difference: Actual | 28.45 |
| 20 | Difference: Hodges-Lehmann | 23.95 |
| 21 | 95.36% CI of difference | 8.500 to 39.90 |
| 22 | Exact or approximate CI? | Exact |

Prism computed an exact P value (0.0035), which takes into account ties among values. Note that most other programs do not compute exact P values when there are tied values, but would instead report an approximate P value (0.0061).

Prism reports the **difference in medians** in two ways.

- Subtract the median of one group from the median of the other
- Compute the Hodges-Lehmann estimate.
    - Many think it is the best estimate for the difference between population medians

Since the nonparametric test works with ranks, it is usually not possible to get a confidence interval with exact 95% confidence.

Prism finds a close confidence level, and reports what it is. For example, you might get a 96.2% confidence interval when you asked for a 95% interval.

You can either report the precise confidence level ("96.2%") or just report the confidence level you requested ("95%"). The latter approach is used more commonly.

Unpaired t test
Tabular results

| | | |
|---|---|---|
| 2 | | |
| 3 | Column B | young |
| 4 | vs. | vs. |
| 5 | Column A | old |
| 6 | | |
| 7 | **Unpaired t test** | |
| 8 | P value | 0.0030 |
| 9 | P value summary | ** |
| 10 | Significantly different (P < 0.05)? | Yes |
| 11 | One- or two-tailed P value? | Two-tailed |
| 12 | t, df | t=3.531, df=15 |
| 13 | | |
| 14 | **How big is the difference?** | |
| 15 | Mean of column A | 30.17 |
| 16 | Mean of column B | 53.71 |
| 17 | Difference between means (B - A) ± SEM | 23.55 ± 6.667 |
| 18 | 95% confidence interval | 9.335 to 37.76 |
| 19 | R squared (eta squared) | 0.4540 |
| 20 | | |
| 21 | **F test to compare variances** | |
| 22 | F, DFn, Dfd | 2.412, 8, 7 |
| 23 | P value | 0.2631 |
| 24 | P value summary | ns |
| 25 | Significantly different (P < 0.05)? | No |

| | | |
|---|---|---|
| 1 | Table Analyzed | Bladder relaxation |
| 2 | | |
| 3 | Column B | young |
| 4 | vs. | vs. |
| 5 | Column A | old |
| 6 | | |
| 7 | **Unpaired t test with Welch's correction** | |
| 8 | P value | 0.0028 |
| 9 | P value summary | ** |
| 10 | Significantly different (P < 0.05)? | Yes |
| 11 | One- or two-tailed P value? | Two-tailed |
| 12 | Welch-corrected t, df | t=3.624, df=13.78 |
| 13 | | |
| 14 | **How big is the difference?** | |
| 15 | Mean of column A | 30.17 |
| 16 | Mean of column B | 53.71 |
| 17 | Difference between means (B - A) ± SEM | 23.55 ± 6.497 |
| 18 | 95% confidence interval | 9.591 to 37.50 |
| 19 | R squared (eta squared) | 0.4881 |
| 20 | | |
| 21 | **F test to compare variances** | |
| 22 | F, DFn, Dfd | 2.412, 8, 7 |
| 23 | P value | 0.2631 |
| 24 | P value summary | ns |
| 25 | Significantly different (P < 0.05)? | No |

Mann-Whitney test
Tabular results

| | | |
|---|---|---|
| 2 | | |
| 3 | Column B | young |
| 4 | vs. | vs. |
| 5 | Column A | old |
| 6 | | |
| 7 | **Mann Whitney test** | |
| 8 | P value | 0.0035 |
| 9 | Exact or approximate P value? | Exact |
| 10 | P value summary | ** |
| 11 | Significantly different (P < 0.05)? | Yes |
| 12 | One- or two-tailed P value? | Two-tailed |
| 13 | Sum of ranks in column A,B | 52 , 101 |
| 14 | Mann-Whitney U | 7 |
| 15 | | |
| 16 | **Difference between medians** | |
| 17 | Median of column A | 29.40, n=9 |
| 18 | Median of column B | 57.85, n=8 |
| 19 | Difference: Actual | 28.45 |
| 20 | Difference: Hodges-Lehmann | 23.95 |
| 21 | 95.36% CI of difference | 8.500 to 39.90 |
| 22 | Exact or approximate CI? | Exact |

Prism computed an exact P value (0.0035), which takes into account ties among values. Note that most other programs do not compute exact P values when there are tied values, but would instead report an approximate P value (0.0061).

# Write up results for Mann-Whitney

For this class:
The data for both groups failed the assumptions for an independent t-test (independence, normality/symmetry, homoscedasticity) so we performed the non-parametric Mann Whitney U test. With $p=0.004$ (two-sided Mann-Whitney test, $\alpha=0.05$), we reject the null hypothesis of no difference in rank score distribution of muscle relaxation between old an young rats. We conclude that older rats have lower levels of muscle relaxation (median = 4.5) compared to younger rats (13.5).

What people may usually write:
We conclude that older rats have lower levels of muscle relaxation (median = 4.5) compared to younger rats (13.5; $p=0.004$).

# The One Sample t-Test

The one sample t test is used to compare a single sample to a population with a known mean.

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

$\overline{x}$ = sample means,
$\mu_0$ = population mean,
s = sample SD,
n = sample size

df = n-1

$H_0$: $\overline{x} = 0$
$H_A$: $\overline{x} \neq 0$

$H_0$: $\overline{x} = 100$
$H_A$: $\overline{x} \neq 100$

It doesn't have to be zero, it could be any number you want to compare to your data

- Assumptions -
The data should follow the normal distribution.
The sample is a random sample from its population (iid).

## Data in Excel

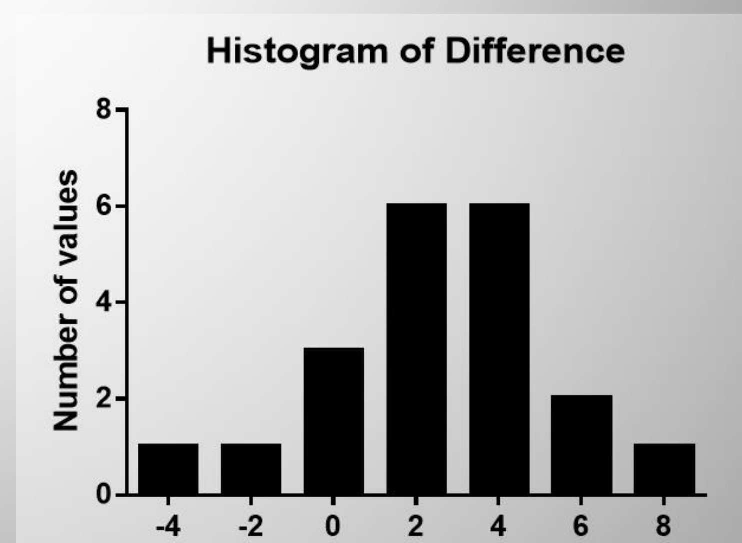
| | A | B | C | D |
|---|---|---|---|---|
| 1 | Patient | Before Treatment | After Treatment | Difference |
| 2 | 1 | 18 | 22 | 4 |
| 3 | 2 | 21 | 25 | 4 |
| 4 | 3 | 16 | 17 | 1 |
| 5 | 4 | 22 | 24 | 2 |
| 6 | 5 | 19 | 16 | -3 |
| 7 | 6 | 24 | 29 | 5 |
| 8 | 7 | 17 | 20 | 3 |
| 9 | 8 | 21 | 23 | 2 |
| 10 | 9 | 23 | 19 | -4 |
| 11 | 10 | 18 | 20 | 2 |
| 12 | 11 | 14 | 15 | 1 |
| 13 | 12 | 16 | 15 | -1 |
| 14 | 13 | 16 | 18 | 2 |
| 15 | 14 | 19 | 26 | 7 |
| 16 | 15 | 18 | 18 | 0 |

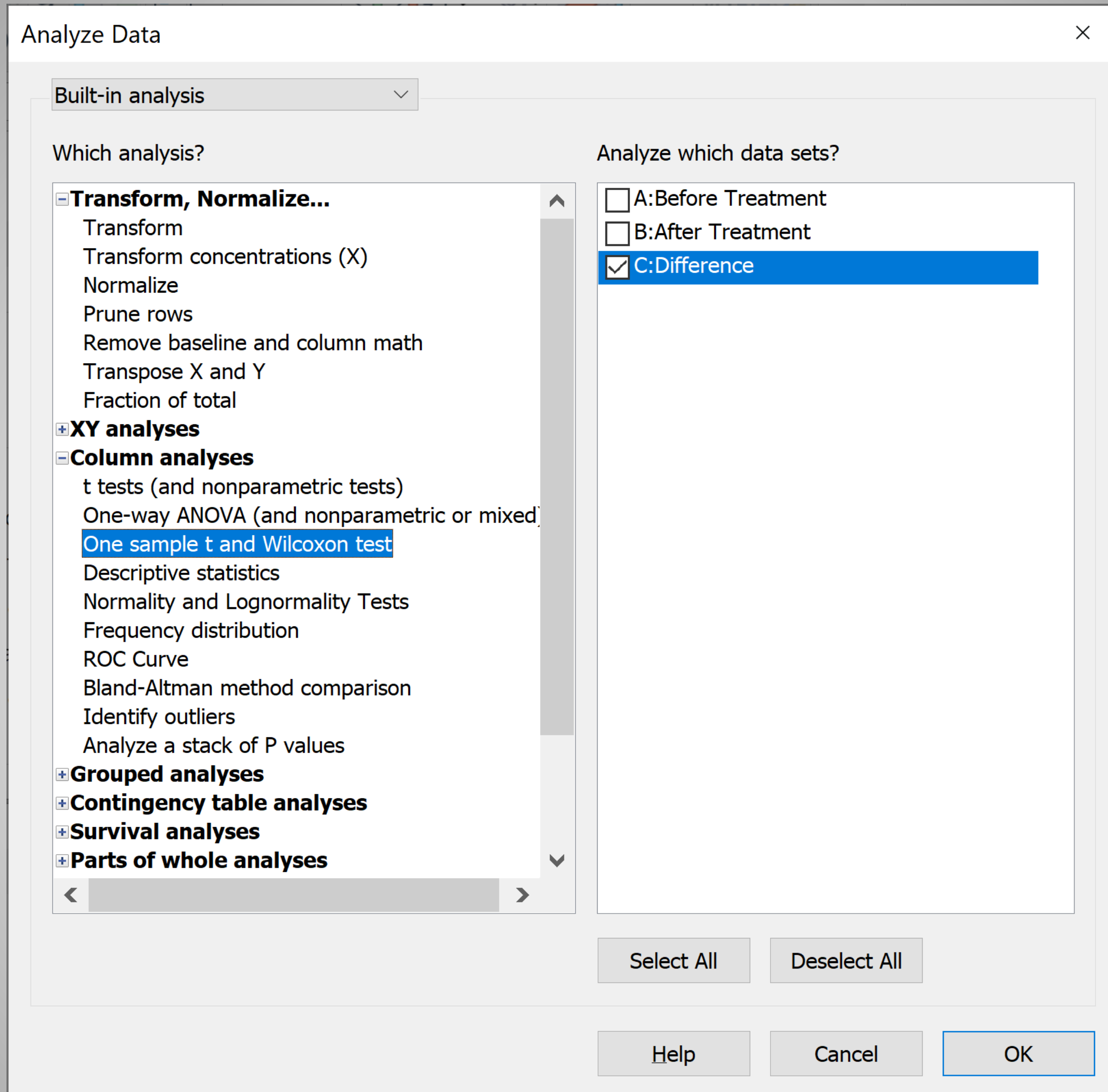
## Data in Prism

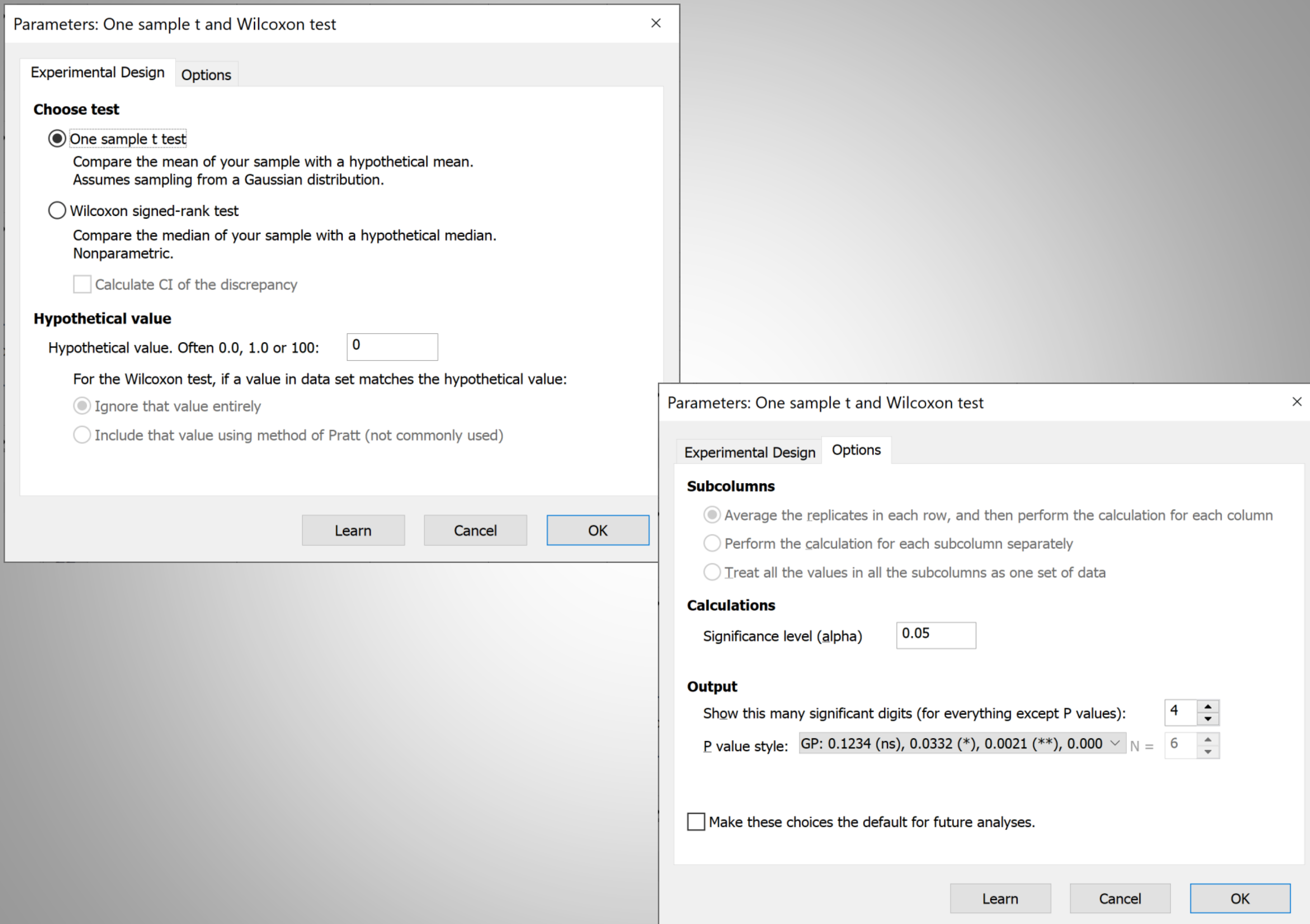
| Group A | Group B |
|---|---|
| Before Treatment | After Treatment |
| 18 | 22 |
| 21 | 25 |
| 16 | 17 |
| 22 | 24 |
| 19 | 16 |
| 24 | 29 |
| 17 | 20 |
| 21 | 23 |
| 23 | 19 |
| 18 | 20 |
| 14 | 15 |
| 16 | 15 |
| 16 | 18 |
| 19 | 26 |
| 18 | 18 |
| 20 | 24 |
| 12 | 18 |
| 22 | 25 |
| 15 | 19 |
| 17 | 16 |

| | Group A | G |
|---|---|---|
| | Difference | |
| | Y | |
| 1 | 4 | |
| 2 | 4 | |
| 3 | 1 | |
| 4 | 2 | |
| 5 | -3 | |
| 6 | 5 | |
| 7 | 3 | |
| 8 | 2 | |
| 9 | -4 | |
| 10 | 2 | |
| 11 | 1 | |
| 12 | -1 | |

$$\boxed{\begin{array}{l} H_0: \overline{x} = 0 \\ H_A: \overline{x} \neq 0 \end{array}}$$




Analyze Data
Built-in analysis
Which analysis?
Transform, Normalize...
Transform
Transform concentrations (X)
Normalize
Prune rows
Remove baseline and column math
Transpose X and Y
Fraction of total
XY analyses
Column analyses
t tests (and nonparametric tests)
One-way ANOVA (and nonparametric or mixed)
One sample t and Wilcoxon test
Descriptive statistics
Normality and Lognormality Tests
Frequency distribution
ROC Curve
Bland-Altman method comparison
Identify outliers
Analyze a stack of P values
Grouped analyses
Contingency table analyses
Survival analyses
Parts of whole analyses
Analyze which data sets?
A:Before Treatment
B:After Treatment
C:Difference
Select All
Deselect All
Help
Cancel
OK


Parameters: One sample t and Wilcoxon test
Experimental Design
Options
Choose test
One sample t test
Compare the mean of your sample with a hypothetical mean.
Assumes sampling from a Gaussian distribution.
Wilcoxon signed-rank test
Compare the median of your sample with a hypothetical median.
Nonparametric.
Calculate CI of the discrepancy
Hypothetical value
Hypothetical value. Often 0.0, 1.0 or 100:
0
For the Wilcoxon test, if a value in data set matches the hypothetical value:
Ignore that value entirely
Include that value using method of Pratt (not commonly used)
Learn
Cancel
OK
Parameters: One sample t and Wilcoxon test
Experimental Design
Options
Subcolumns
Average the replicates in each row, and then perform the calculation for each column
Perform the calculation for each subcolumn separately
Treat all the values in all the subcolumns as one set of data
Calculations
Significance level (alpha)
0.05
Output
Show this many significant digits (for everything except P values):
4
P value style:
GP: 0.1234 (ns), 0.0332 (*), 0.0021 (**), 0.000
N =
6
Make these choices the default for future analyses.
Learn
Cancel
OK

| | One sample t test | A |
|---|---|---|
| | | Difference |
| 1 | Theoretical mean | 0.000 |
| 2 | Actual mean | 2.050 |
| 3 | Number of values | 20 |
| 4 | | |
| 5 | **One sample t test** | |
| 6 | t, df | t=3.231, df=19 |
| 7 | P value (two tailed) | 0.0044 |
| 8 | P value summary | ** |
| 9 | Significant (alpha=0.05)? | Yes |
| 10 | | |
| 11 | **How big is the discrepancy?** | |
| 12 | Discrepancy | 2.050 |
| 13 | SD of discrepancy | 2.837 |
| 14 | SEM of discrepancy | 0.6344 |
| 15 | 95% confidence interval | 0.7221 to 3.378 |
| 16 | R squared (partial eta squared) | 0.3546 |

## WRITE UP RESULTS

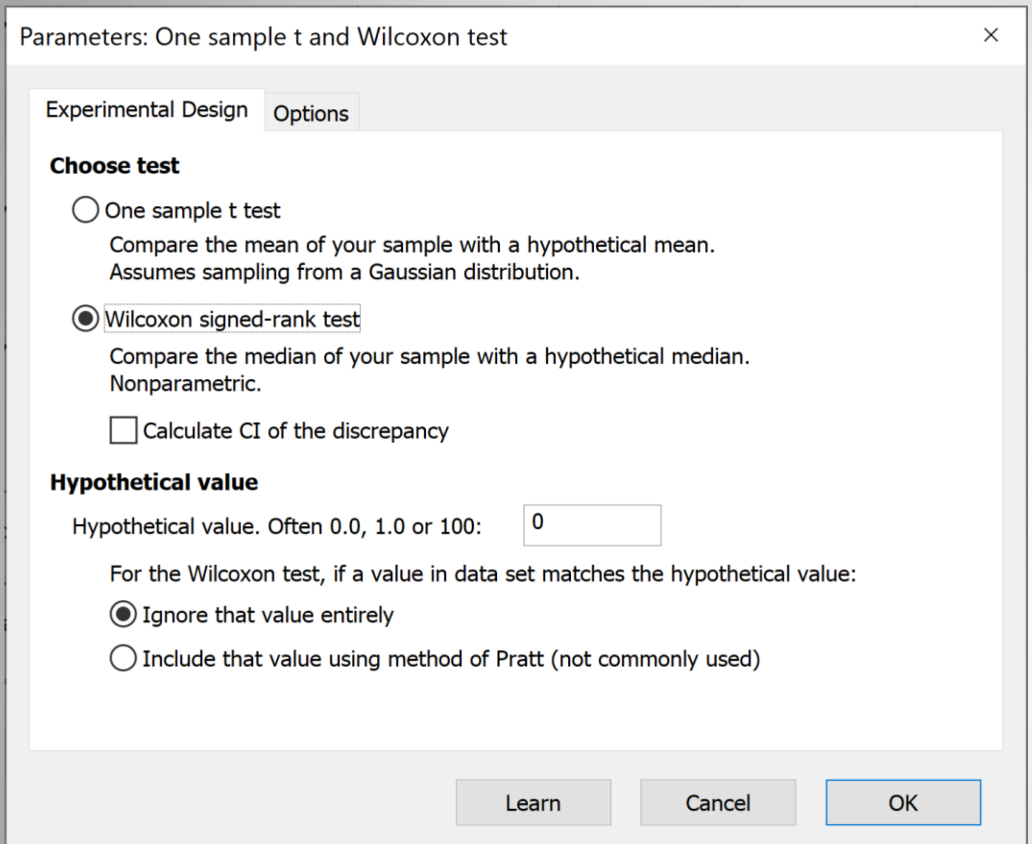
For this class:

The data passed the assumptions for a one sample t-test (data are normal/symmetrical). With p=0.004 (two-tailed one sample t-test, t=3.231, df=19, $\alpha$=0.05), we reject the null hypothesis that the mean of our data is zero. Further, the 95%CI do not include 0 (0.7 to 3.4). We conclude that with a mean increase of 2.05 over placebo, treatment increases the value of Test Y.

What people may usually write:

We found that treatment increases the value of Test Y by 2.05 (p=0.004).

# The non-parametric alternative to the one-sample t-test: the Wilcoxon signed rank test




$H_0$: the median = 0

$H_A$: the median ≠ 0

It doesn't have to be zero, it could be any number you want to compare to your data

| One sample Wilcoxon test | | A<br>Difference |
|---|---|---|
| 1 | Theoretical median | 0.000 |
| 2 | Actual median | 2.000 |
| 3 | Number of values | 20 |
| 4 | | |
| 5 | **Wilcoxon Signed Rank Test** | |
| 6 | Sum of signed ranks (W) | 132.0 |
| 7 | Sum of positive ranks | 161.0 |
| 8 | Sum of negative ranks | -29.00 |
| 9 | P value (two tailed) | 0.0058 |
| 10 | Exact or estimate? | Exact |
| 11 | P value summary | ** |
| 12 | Significant (alpha=0.05)? | Yes |
| 13 | | |
| 14 | **How big is the discrepancy?** | |
| 15 | Discrepancy | 2.000 |

| One sample t test | | A<br>Difference |
|---|---|---|
| 1 | Theoretical mean | 0.000 |
| 2 | Actual mean | 2.050 |
| 3 | Number of values | 20 |
| 4 | | |
| 5 | **One sample t test** | |
| 6 | t, df | t=3.231, df=19 |
| 7 | P value (two tailed) | 0.0044 |
| 8 | P value summary | ** |
| 9 | Significant (alpha=0.05)? | Yes |
| 10 | | |
| 11 | **How big is the discrepancy?** | |
| 12 | Discrepancy | 2.050 |
| 13 | SD of discrepancy | 2.837 |
| 14 | SEM of discrepancy | 0.6344 |
| 15 | 95% confidence interval | 0.7221 to 3.378 |
| 16 | R squared (partial eta squared) | 0.3546 |

# Write up results

For this class:
Because the data did not pass the assumptions for a one-sample t-test (data are normal/symmetrical), a Wilcoxon signed rank test was used. With $p=0.006$ (two-tailed Wilcoxon signed rank test, $\alpha=0.05$), we reject the null hypothesis that the median of our data is zero. We conclude that with a median increase of 2 points over placebo, treatment increases the value of Test Y.

What people may usually write:
We found that treatment increases the value of Test Y by 2 points ($p=0.006$)