Comparing Means of Two Groups

Elias Rizk MD

PennState

Comparing

- We often want to compare individuals (or other units) from two groups.

'Individuals'	Measurement	Groups	Question
Customers in a supermarket	Amount spent (dollars)	Male and female	Do male and female customers spend the same amounts?
Bank accounts	Number of transactions in month	Two types of account with different fee structures (one with lower per-transaction charge and the other with lower fixed charge)	Are there more transactions in accounts with lower per-transaction charges? By how much?
Milk containers filled in bottling factory	Volume of milk in container	Two different filling machines	Do both machines fill the containers with the same amount of milk on average?

Questions are often about underlying populations

- The questions in the above scenarios are not about the **specific** customers who entered the supermarket, the **specific** bank accounts that were sampled, etc.

- They ask about the differences between supermarket spending by males and females **in general**, the differences between the two types of bank account **in general**, etc.

Questions are often about underlying populations

- We are therefore usually interested in the characteristics of a population or process that we assume **underlies** the data that are collected.

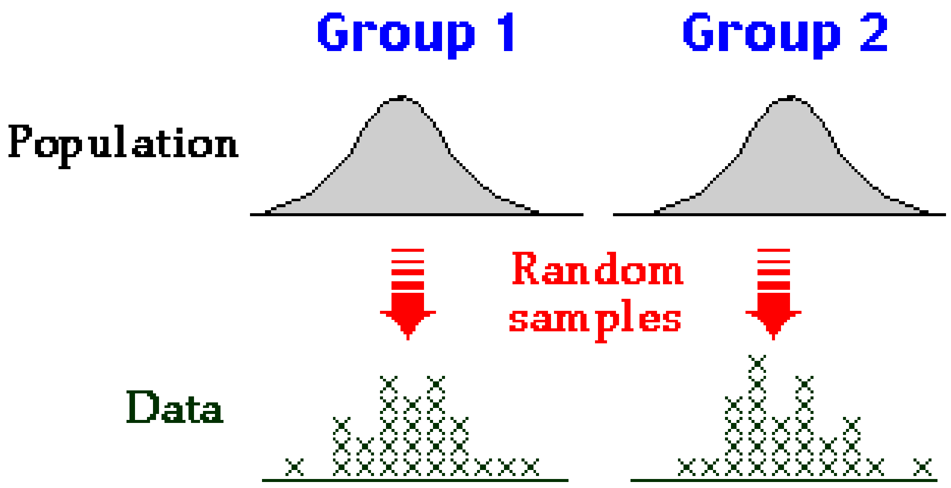
- The data provide information about the likely characteristics of the population.

Model for two groups

- A single batch of numerical values is usually modelled as a random sample from some population — often a normal distribution.

- In a similar way, data sets that consist of measurements from two groups are often modelled as two independent random samples from two underlying hypothetical infinite populations.
- Normal distributions are again commonly used as models.

Model for two groups



- The assumption of normality should be checked from graphical displays of the sample data. If the data are noticeably skewed, a transformation may provide values that can be adequately modelled by normal distributions

Region of Rejection and Retention

- Determining whether or not to reject the null depends on where the obtained t value falls within the t-distribution

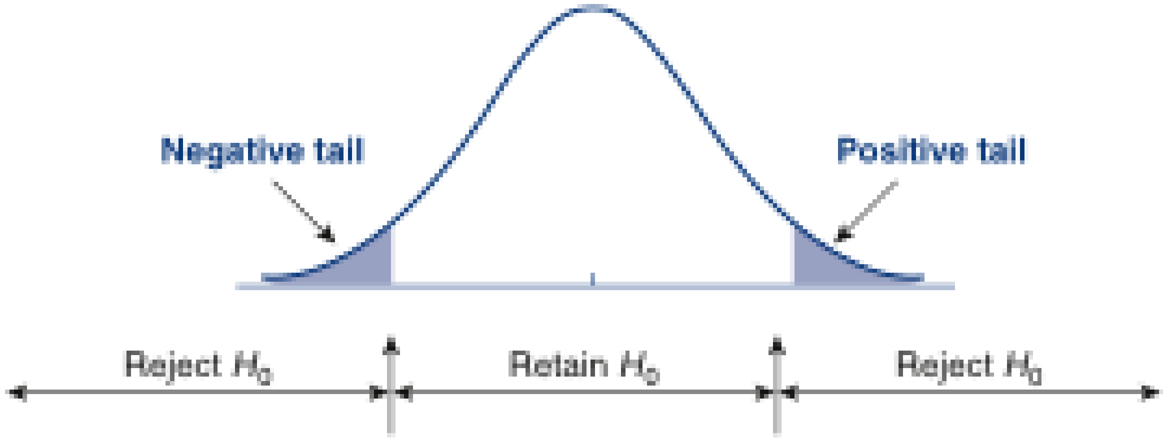
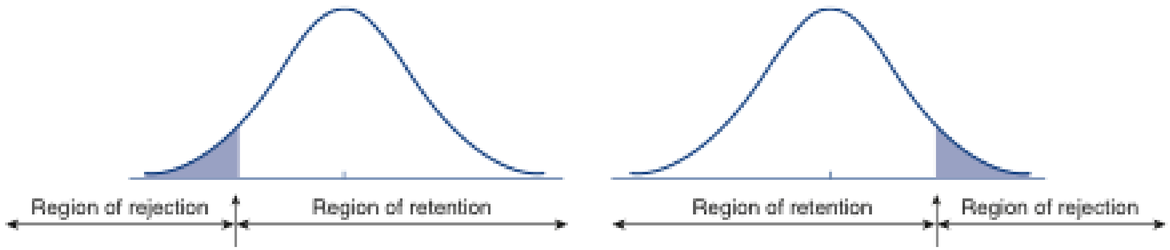


Figure 17.1 Regions of Rejection and Retention

Directional Tests

- One tailed tests place the entire region of rejection in a single tail



- Two tailed tests divide the region of rejection into portions for each tail

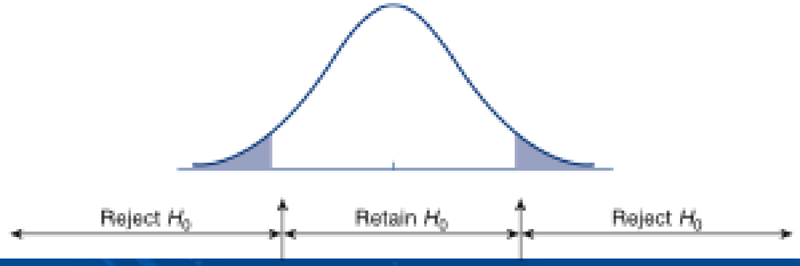


Figure 17.3 Regions of Rejection in a Two-Tailed Test

t-Distributions

- The shape of the t-distribution changes depending upon the size of your sample

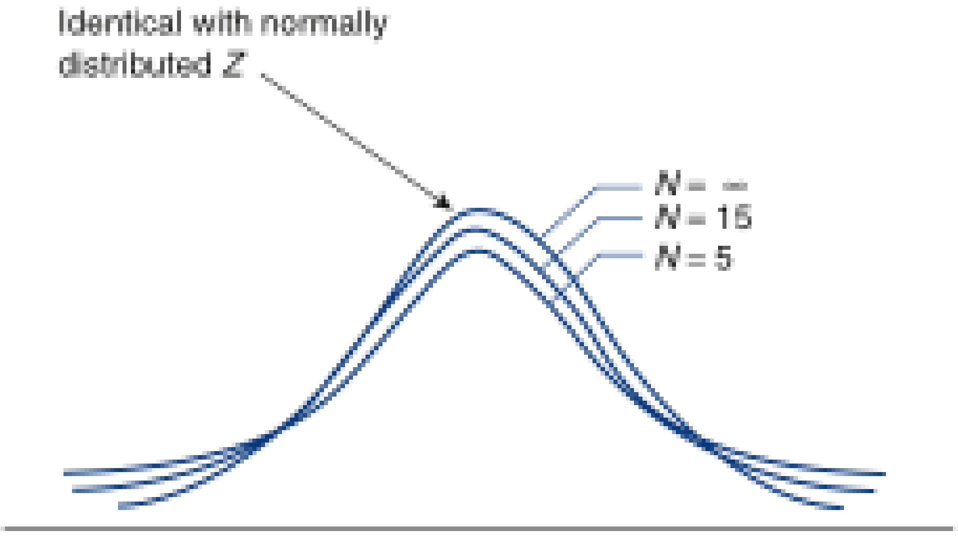
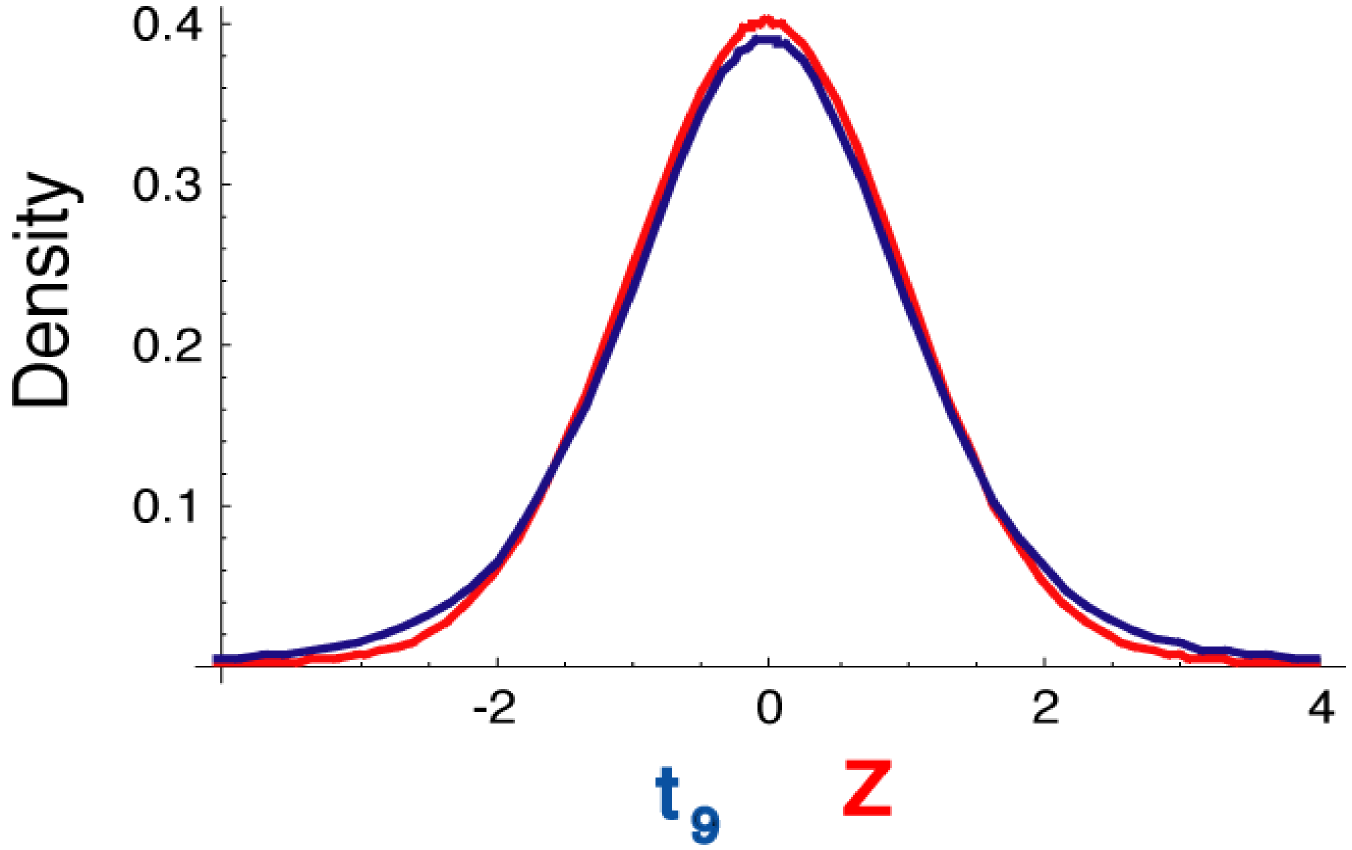
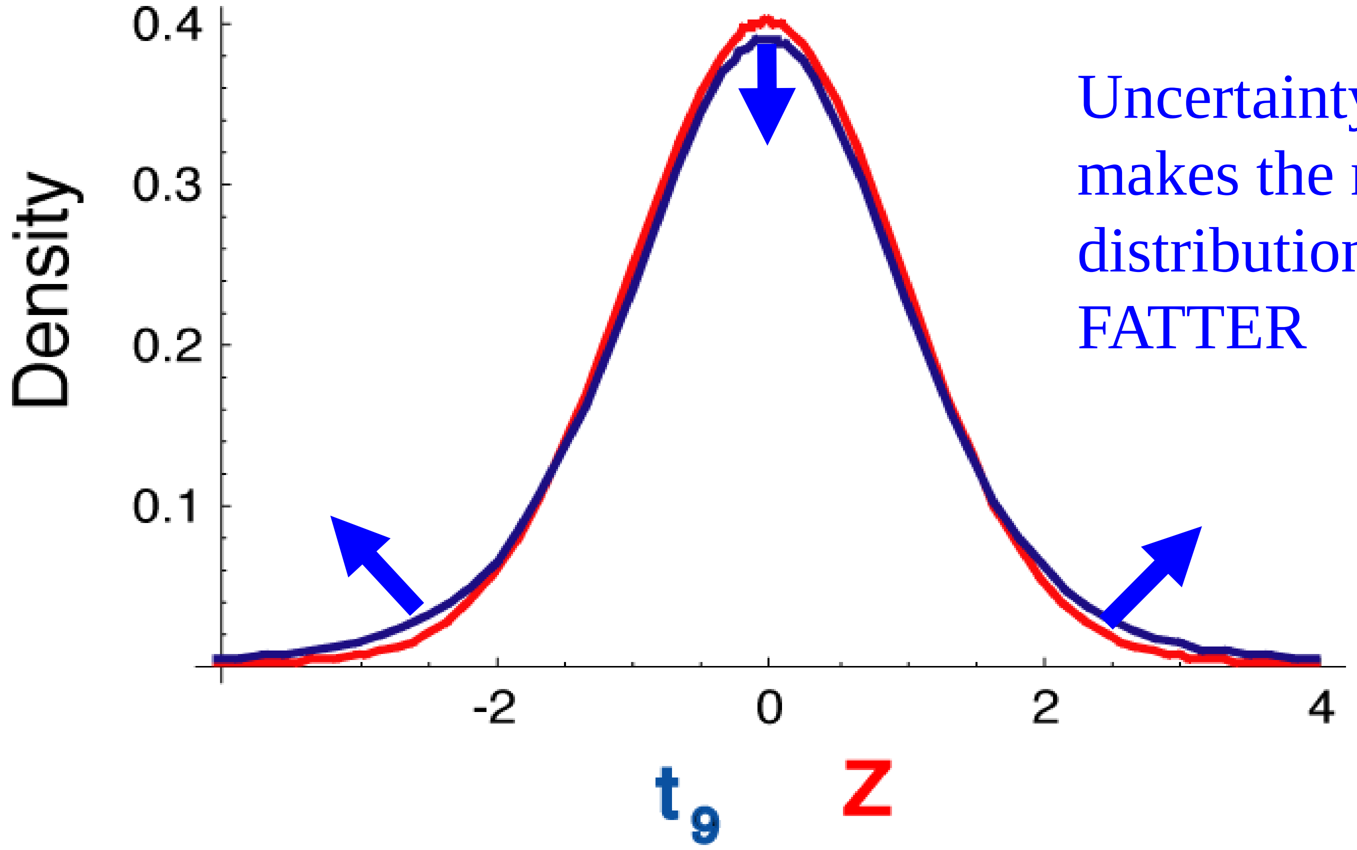


Figure 17.5 Change in t Distribution as a Function of Sample Size

t has a Student's *t* distribution*



t has a Student's *t* distribution*



Uncertainty makes the null distribution FATTER

* Under the null hypothesis

Difference between means

Comparing the populations

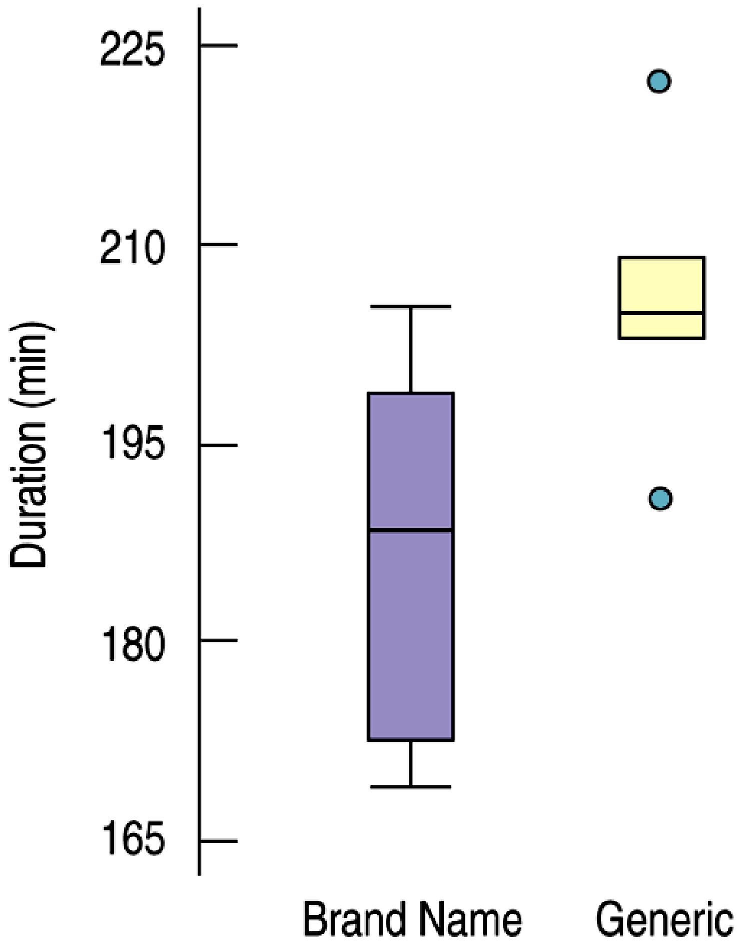
- For two-group data sets, we usually want to compare the underlying populations.

- In particular, the main questions of interest are:
  - Are the two population distributions the same?
  - If the populations are different, how big is the difference?

Plot the Data

Comparing the populations

- The natural display for comparing two groups is boxplots of the data for the two groups, placed side-by-side. For example:



PennState

Comparing Two Means

- Once we have examined the side-by-side boxplots, we can turn to the comparison of two means.
- Comparing two means is not very different from comparing two proportions.
- This time the parameter of interest is the difference between the two means, $\mu_1 - \mu_2$.

Comparing Two Means

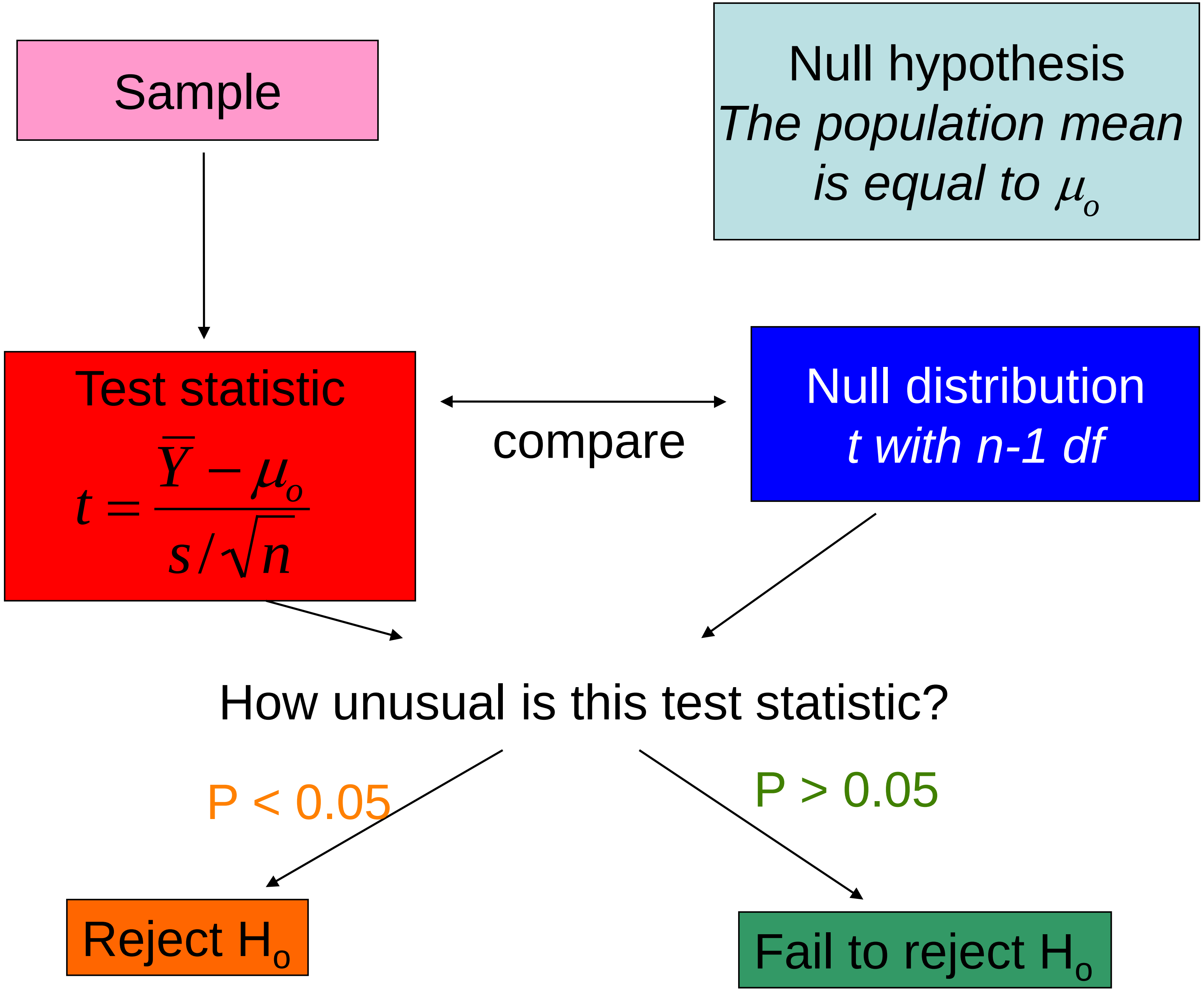
- A t-test may be used to evaluate whether a single group differs from a known value (a one-sample t-test)
- Whether there is a significant difference in paired measurements (a paired, or dependent samples t-test).
- Whether two groups differ from each other (an independent two-sample t-test)

One Sample T-tests

- One sample t-tests are used in the following two situations
  - The size of a sample is less than 25
  - The population standard deviation is unknown
  - The one-sample t-test is a statistical hypothesis test used to determine whether an unknown population mean is different from a specific value.

- The formula for a one sample test uses the estimated population standard deviation to calculate the standard error
  - $\sigma_M = \frac{\sigma_{est}}{\sqrt{n}}$
  - $t = \frac{M-\mu}{\sigma_M}$

One-sample t-test

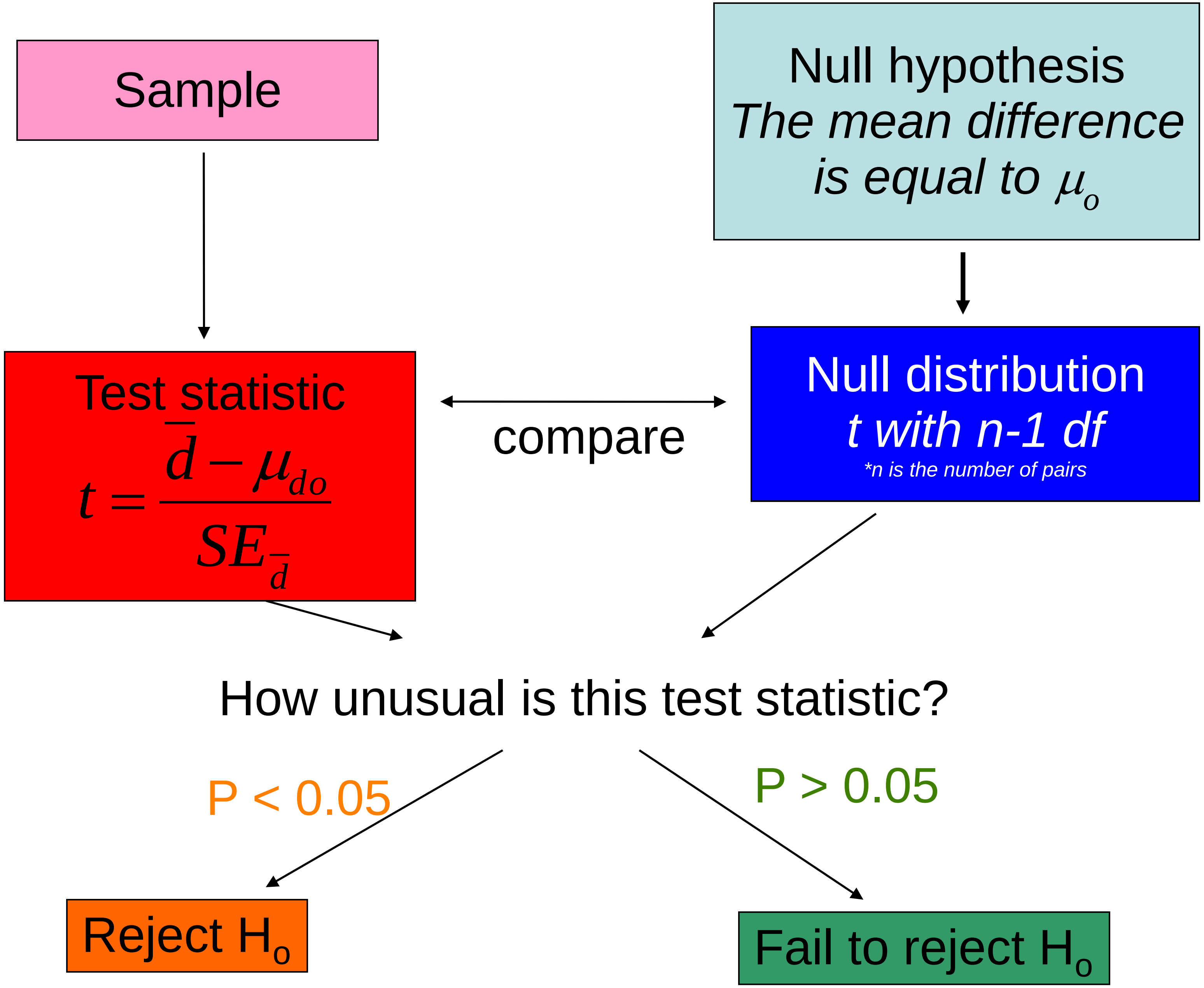


Quick reference summary: One-sample *t*-test

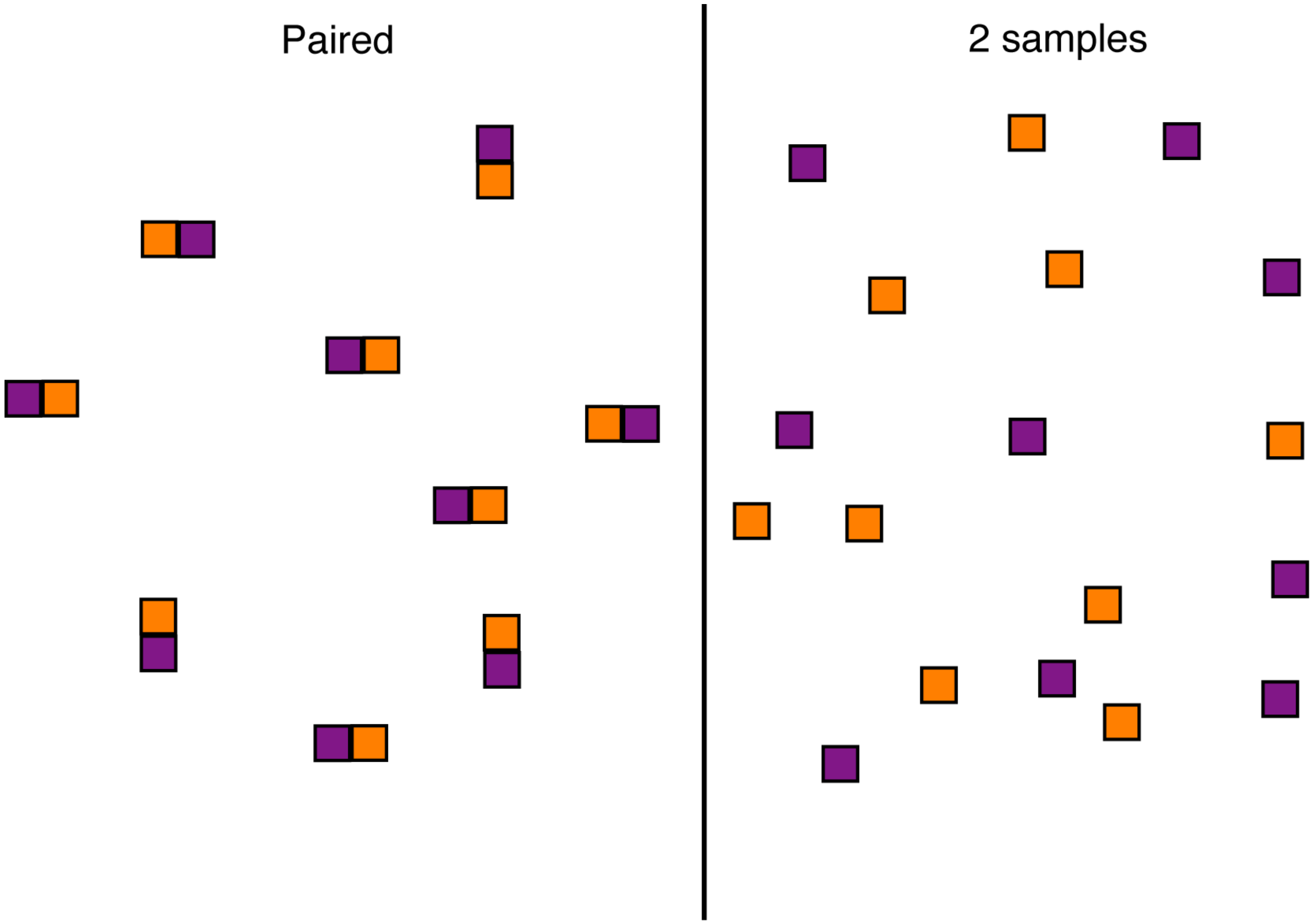
- What is it for? *Compares the mean of a numerical variable to a hypothesized value, μ_o*
- What does it assume? *Individuals are randomly sampled from a population that is normally distributed*
- Test statistic: *t*
- Distribution under H_o: *t-distribution with n-1 degrees of freedom*
- Formulae: $\bar{Y}$ = *sample mean, s = sample standard deviation*

$$t = \frac{\bar{Y} - \mu_o}{s/\sqrt{n}}$$

Paired t-test

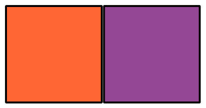


Paired vs. 2 sample comparisons



Paired designs

- Data from the two groups are paired
- There is a one-to-one correspondence between the individuals in the two groups



More on pairs

- Each member of the pair shares much in common with the other, *except* for the tested categorical variable
- Example: identical twins raised in different environments
- Can use the same individual at different points in time
- Example: before, after medical treatment

Paired design: Examples

- Same river, upstream and downstream of a power plant
- Tattoos on both arms: how to get them off? Compare lasers to dermabrasion

Paired comparisons

- To compare two groups, we use the mean of the *difference* between the two members of each pair

Example: National No Smoking Day

- Data compares injuries at work on National No Smoking Day (in Britain) to the same day the week before

- Each data point is a year

Data

Year	Injuries before No Smoking Day	Injuries on No Smoking Day
1987	516	540
1988	610	620
1989	581	599
1990	586	639
1991	554	607
1992	632	603
1993	479	519
1994	583	560
1995	445	515
1996	522	556

Calculate differences

Injuries before No Smoking Day	Injuries on No Smoking Day	Difference (d)
516	540	24
610	620	10
581	599	18
586	639	53
554	607	53
632	603	-29
479	519	40
583	560	-23
445	515	70
522	556	34

Paired *t* test

- Compares the mean of the differences to a value given in the null hypothesis

- For each pair, calculate the difference.

- The paired *t*-test is a one-sample *t*-test on the differences.

Hypotheses

Ho: Work related injuries do not change during
No Smoking Days ($\mu=0$)

Ha: Work related injuries change during
No Smoking Days ($\mu \neq 0$)

Calculate differences

Injuries before No Smoking Day	Injuries on No Smoking Day	Difference (d)
516	540	24
610	620	10
581	599	18
586	639	53
554	607	53
632	603	-29
479	519	40
583	560	-23
445	515	70
522	556	34

CAUTION!

- The number of data points in a paired *t* test is the number of *pairs*.   -- *Not* the number of individuals
- Degrees of freedom = Number of pairs - 1

Here, df = 10-1 = 9

Critical value of *t*

Test statistic: $t = 2.45$

So we can reject the null hypothesis: Stopping smoking increases job-related accidents in the short term.

Assumptions of paired *t* test

- Pairs are chosen at random
- The differences have a normal distribution

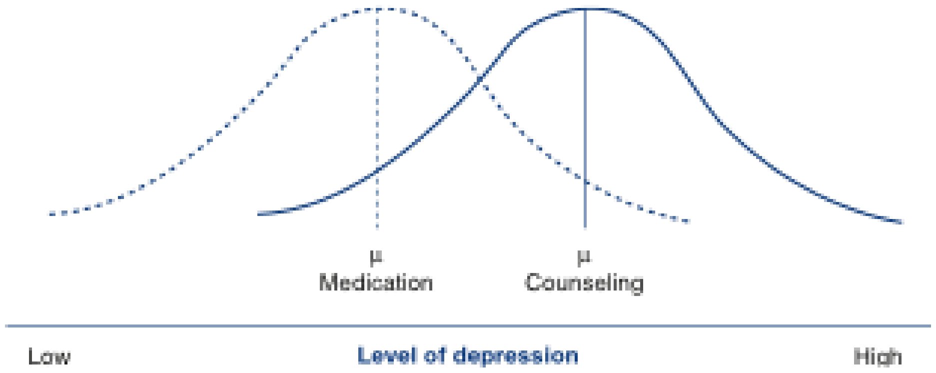
It does *not* assume that the individual values are normally distributed, only the differences.

Quick reference summary: Paired *t*-test

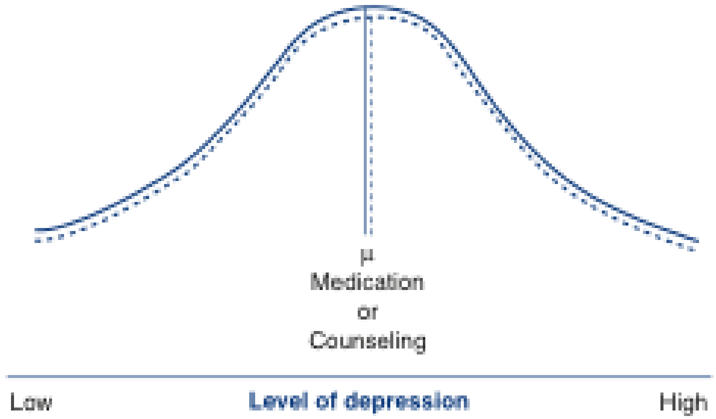
- What is it for? *To test whether the mean difference in a population equals a null hypothesized value,* μ_{do}
- What does it assume? *Pairs are randomly sampled from a population. The differences are normally distributed*
- Test statistic: *t*
- Distribution under H_o: *t-distribution with n-1 degrees of freedom, where n is the number of pairs*

Two Sample Studies

- Two samples can be compared when parameters for both populations are not available
- Research Hypothesis



- Null Hypothesis



PennState

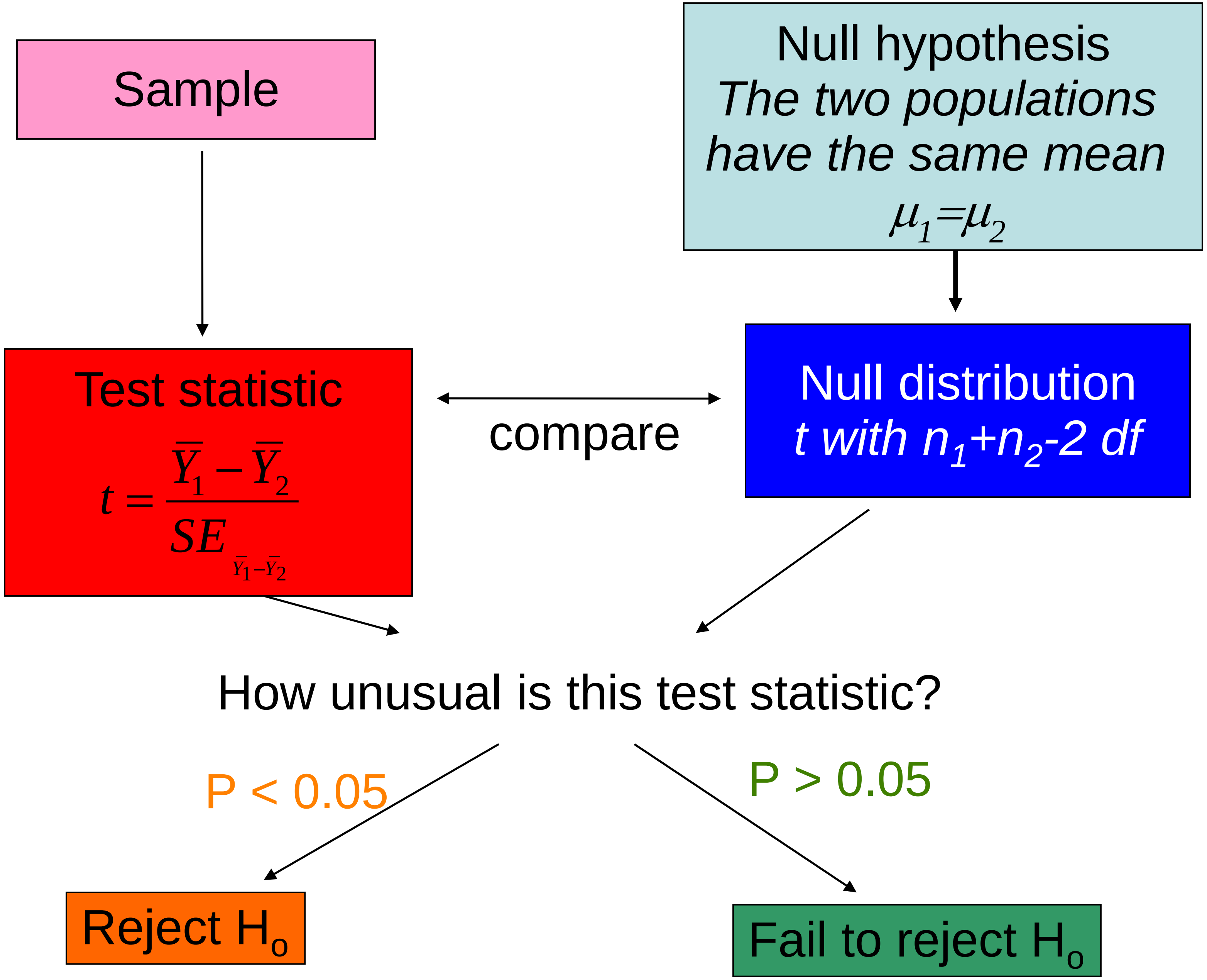
Assumptions and Conditions

- Independence Assumption (Each condition needs to be checked for both groups.):
  - Randomization Condition: Were the data collected with suitable randomization (representative random samples or a randomized experiment)?
  - 10% Condition: We don’t usually check this condition for differences of means. We will check it for means only if we have a very small population or an extremely large sample.
  - The variance of both populations is equal.

Assumptions and Conditions (cont.)

- Normal Population Assumption:
  - Nearly Normal Condition: This must be checked for *both* groups. A violation by either one violates the condition.
- Independent Groups Assumption: The two groups we are comparing must be independent of each other.

Two-sample t-test



Quick reference summary: Two-sample *t*-test

- What is it for? *Tests whether two groups have the same mean*
- What does it assume? *Both samples are random samples. The numerical variable is normally distributed within both populations. The variance of the distribution is the same in the two populations*
- Test statistic: *t*
- Distribution under H_o: *t-distribution with n_1+n_2-2 degrees of freedom.*
- Formulae:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{SE_{\bar{Y}_1 - \bar{Y}_2}}$$

$$SE_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$s_p^2 = \frac{df_1 s_1^2 + df_2 s_2^2}{df_1 + df_2}$$

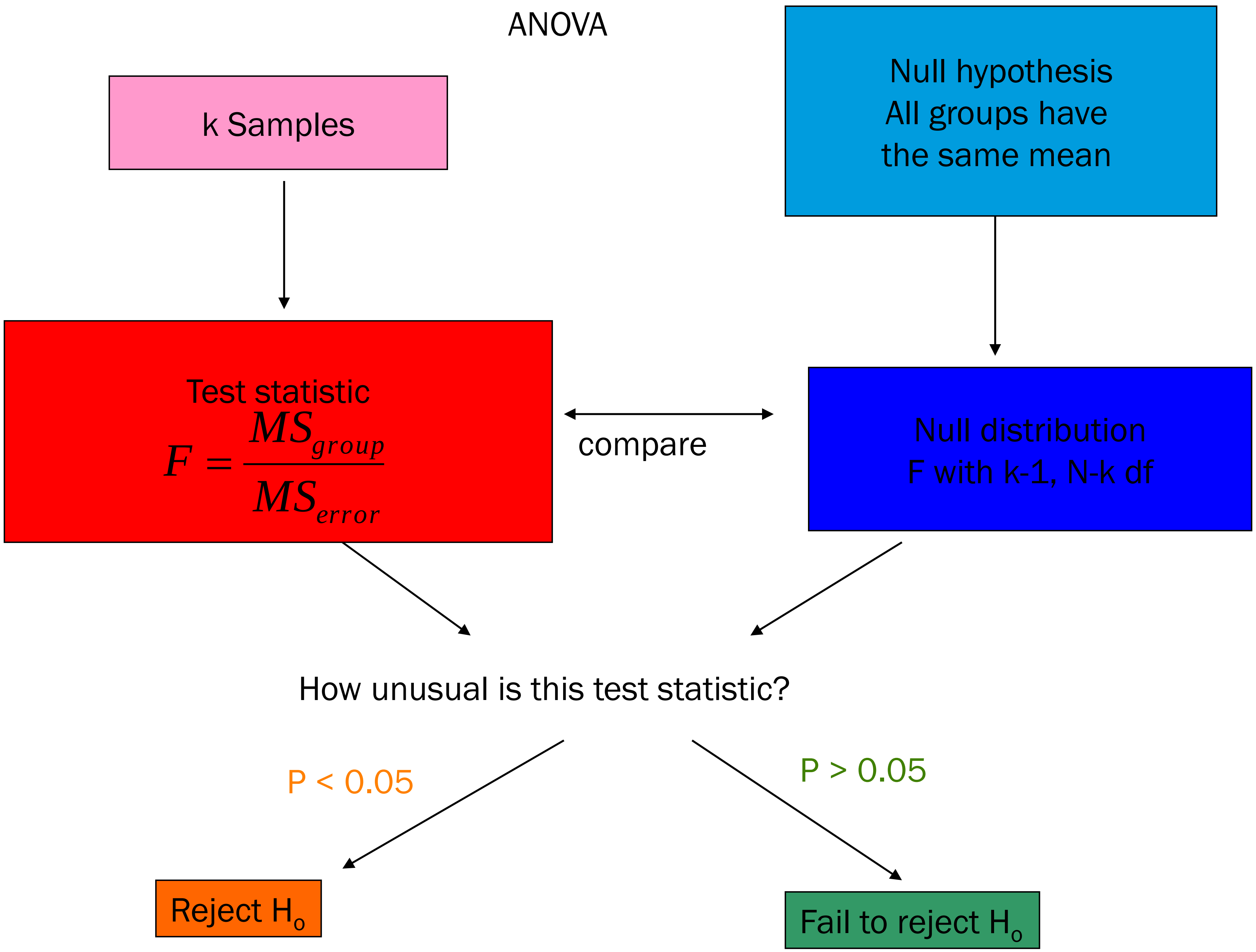
Comparing means when variances are not equal

Welch's *t* test

Welch's approximate t-test compares the means of two normally distributed populations that have unequal variances.

ANOVA

- ANOVA is used when more than two groups are compared
- In order to conduct an ANOVA, several assumptions must be made
  - The population from which the samples are drawn are normally distributed
  - The populations from which the samples are drawn have equal variances
- Test statistic: F
- Distribution under H_o: F distribution with k-1 and N-k degrees of freedom


ANOVA
k Samples
Null hypothesis
All groups have
the same mean
Test statistic
$F = \frac{MS_{group}}{MS_{error}}$
compare
Null distribution
F with k-1, N-k df
How unusual is this test statistic?
P < 0.05
P > 0.05
Reject H_o
Fail to reject H_o

Partitioning Variances

- When conducting an ANOVA, the variances of the groups are partitioned into between-group variance and within-group variance

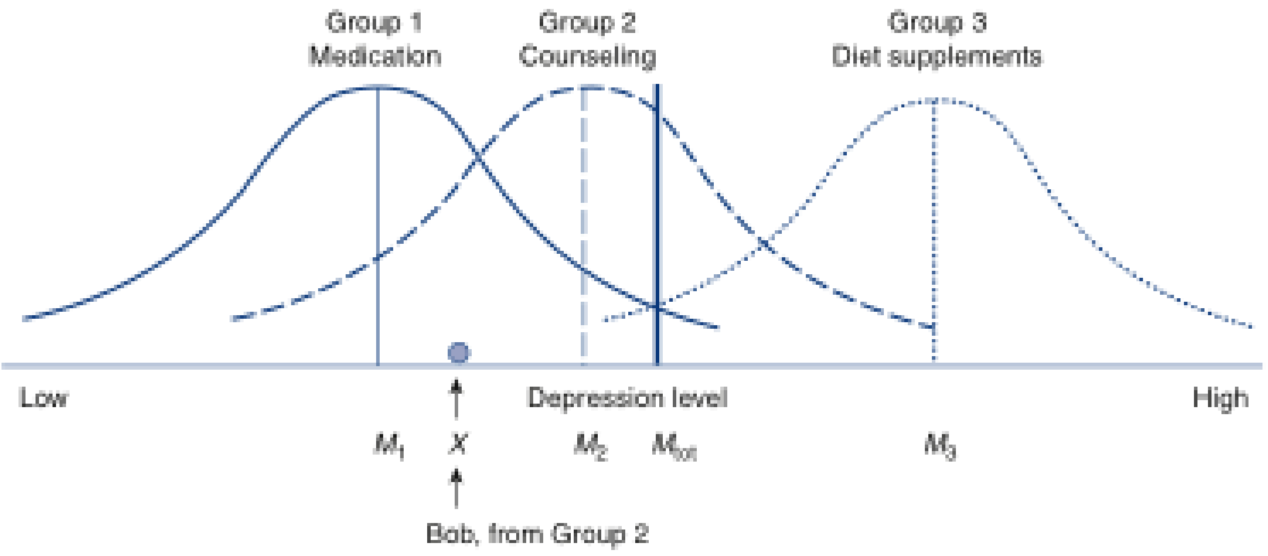


Figure 24.1 Three Different Treatment Populations

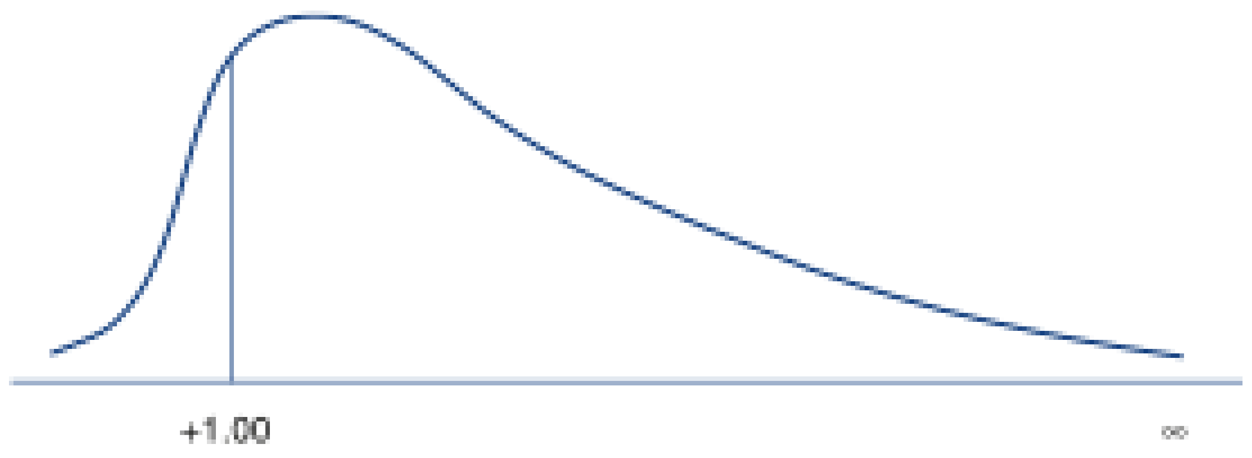
PennState

Mean Squares

- Mean Squares are variances
  - They are the average squared deviation score
- The test statistic for an ANOVA, F, is calculated by dividing two mean squares
- $F = \frac{MS_{between}}{MS_{within}}$

Decision Making for ANOVA

- The F-Distribution is used to obtain the critical value for an ANOVA



One Way ANOVA

- One Way ANOVA's involve a single independent variable

Depression Level After Treatment		
Medication	*Counseling*	*Diet Supplement*
23	38	40
16	32	28
15	29	33
32	42	42
26	25	35
18	17	35
22	37	41
14	26	30
14	19	34
22	22	39
202	287	357
$M_{med} = \frac{202}{10} = 20.20$	$M_{couns} = \frac{287}{10} = 28.70$	$M_{diet} = \frac{357}{10} = 35.70$

Logic for ANOVA

- The logic for an F-test (ANOVA) is the same as other hypotheses tests
- $F = \frac{MS_{between\ observed} - MS_{between\ expected}}{MS_{within}}$
- The $MS_{\text{Between expected}}$ is always believed to be zero

Calculating Sum of Squares

- These are the formulas for the variance partitions
- $SS_{bet} = \sum_1^k \frac{(\sum X_g)^2}{n_g} - \frac{(\sum X_{tot})^2}{N}$
- $SS_{with} = \sum_1^N X^2 - \sum_1^k \frac{(\sum X_g)^2}{n_g}$
- $SS_{tot} = \sum_1^N X^2 - \frac{(\sum X_{tot})^2}{N}$

Calculating df and MS

- df_{bet}= No. of Groups – 1 = k-1
- df_{with} = No. of Subjects – No. of groups = N-k
- df_{tot} = No. of Subjects – 1 = N-1

- $MS_{bet} = \frac{SS_{bet}}{df_{bet}}$
- $MS_{with} = \frac{SS_{with}}{df_{with}}$

ANOVA Tables

Source of variation	Sum of squares	df	Mean Squares	F ratio	P
Treatment					
Error					
Total					

ANOVA Tables

Source of variation	Sum of squares	df	Mean Squares	F ratio	P
Treatment	$SS_{group} = \sum n_i(\bar{Y}_i - \bar{Y})^2$				
Error	$SS_{error} = \sum s_i^2(n_i - 1)$				
Total	$SS_{group} + SS_{error}$				

ANOVA Tables

Source of variation	Sum of squares	df	Mean Squares	F ratio	P
Treatment	$SS_{group} = \sum n_i(\bar{Y}_i - \bar{Y})^2$	k-1			
Error	$SS_{error} = \sum s_i^2(n_i - 1)$	N-k			
Total	$SS_{group} + SS_{error}$	N-1			

ANOVA Tables

Source of variation	Sum of squares	df	Mean Squares	F ratio	P
Treatment	$SS_{group} = \sum n_i(\bar{Y}_i - \bar{Y})^2$	k-1	$MS_{group} = \frac{SS_{group}}{df_{group}}$		
Error	$SS_{error} = \sum s_i^2(n_i - 1)$	N-k	$MS_{error} = \frac{SS_{error}}{df_{error}}$		
Total	$SS_{group} + SS_{error}$	N-1			

PennState

ANOVA Tables

Source of variation	Sum of squares	df	Mean Squares	F ratio	P
Treatment	$SS_{group} = \sum n_i(\bar{Y}_i - \bar{Y})^2$	k-1	$MS_{group} = \frac{SS_{group}}{df_{group}}$	$F = \frac{MS_{group}}{MS_{error}}$	
Error	$SS_{error} = \sum s_i^2(n_i - 1)$	N-k	$MS_{error} = \frac{SS_{error}}{df_{error}}$		
Total	$SS_{group} + SS_{error}$	N-1			

ANOVA Tables

Source of variation	Sum of squares	df	Mean Squares	F ratio	P
Treatment	$SS_{group} = \sum n_i(\bar{Y}_i - \bar{Y})^2$	k-1	$MS_{group} = \frac{SS_{group}}{df_{group}}$	$F = \frac{MS_{group}}{MS_{error}}$	*
Error	$SS_{error} = \sum s_i^2(n_i - 1)$	N-k	$MS_{error} = \frac{SS_{error}}{df_{error}}$		
Total	$SS_{group} + SS_{error}$	N-1			

ANOVA Table: Example

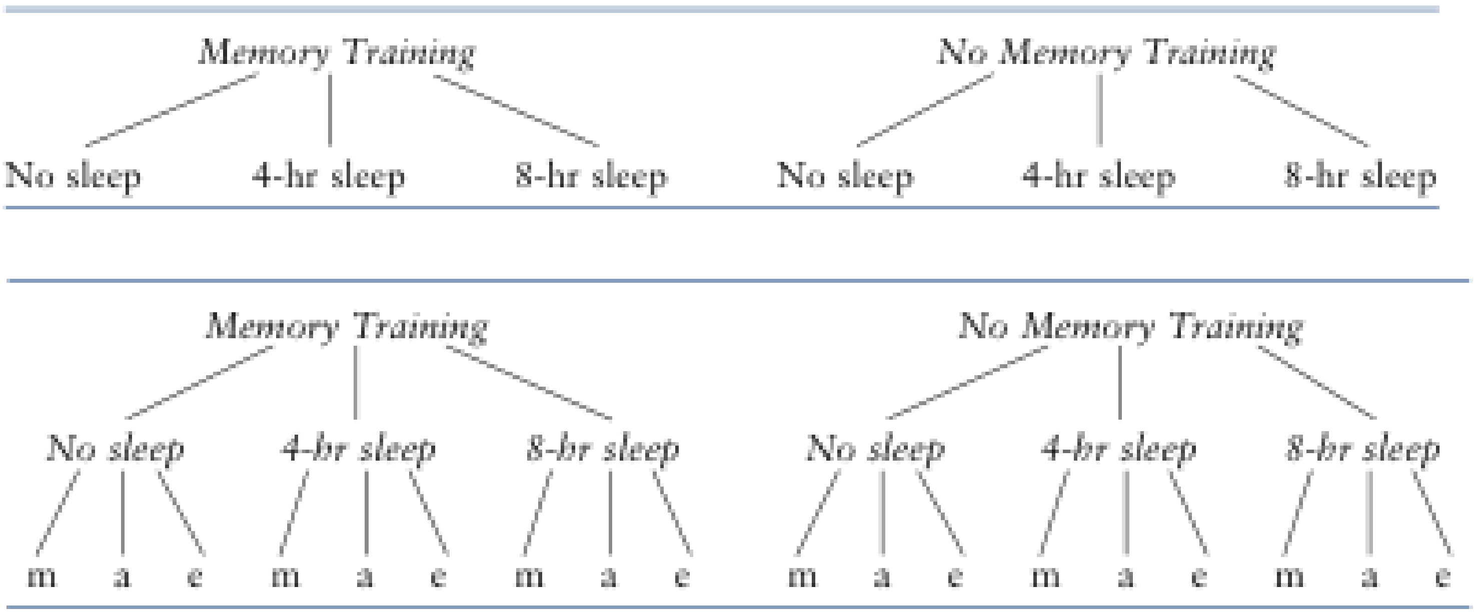
Source of variation	Sum of squares	df	Mean Squares	F ratio	P
Treatment	7.22	2			
Error	9.41	19			
Total					

ANOVA Table: Example

Source of variation	Sum of squares	df	Mean Squares	F ratio	P
Treatment	7.22	2	3.61	7.29	0.004
Error	9.42	19	0.50		
Total	16.64	21			

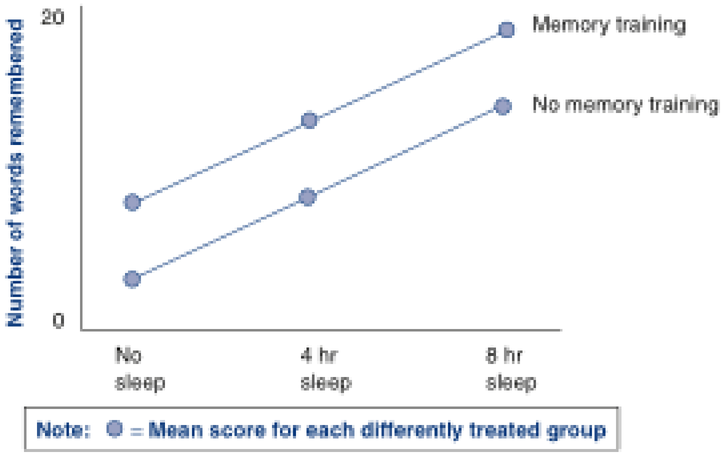
Factorial ANOVA

- Factorial ANOVAs contain multiple independent variables



Main Effects

- Main effects indicate significant differences within a single independent variable



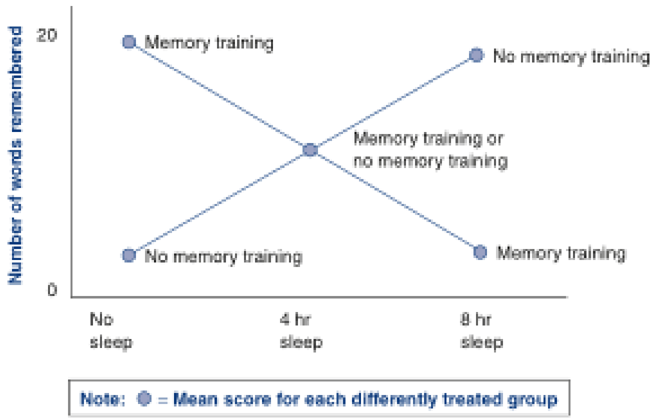
Connected Group Means for a Two-Way ANOVA With Two Main Effects and No Interaction Effect

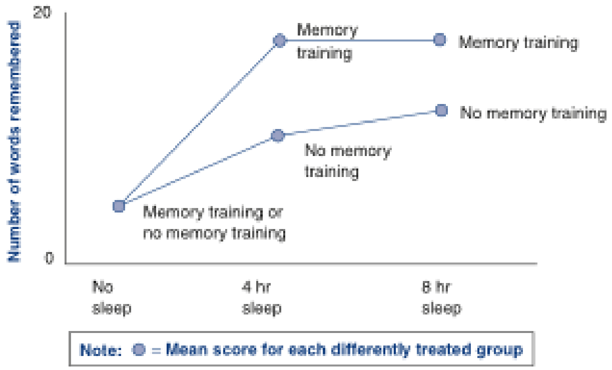
Group Means for a Two-Way ANOVA With Two Main Effects and No Interaction Effect

	No Sleep	*4-hr Sleep*	*8-hr Sleep*	*Memory Condition Row Means*
Memory training	10	15	20	15
No memory training	5	10	15	10
Sleep Condition Column Means	7.5	12.5	17.5	

Interaction Effects

- Interaction effects indicate significant differences across two independent variables





PennState

Two-factor ANOVA Table

Source of variation	Sum of Squares	df	Mean Square	F ratio	P
Treatment 1	SS_1	$k_1 - 1$	$\frac{SS_1}{k_1 - 1}$	$\frac{MS_1}{MSE}$	
Treatment 2	SS_2	$k_2 - 1$	$\frac{SS_2}{k_2 - 1}$	$\frac{MS_2}{MSE}$	
Treatment 1 * Treatment 2	SS_{1*2}	$(k_1 - 1)*(k_2 - 1)$	$\frac{SS_{1*2}}{(k_1 - 1)*(k_2 - 1)}$	$\frac{MS_{1*2}}{MSE}$	
Error	SS_{error}	XXX	$\frac{SS_{error}}{XXX}$		
Total	SS_{total}	N-1			

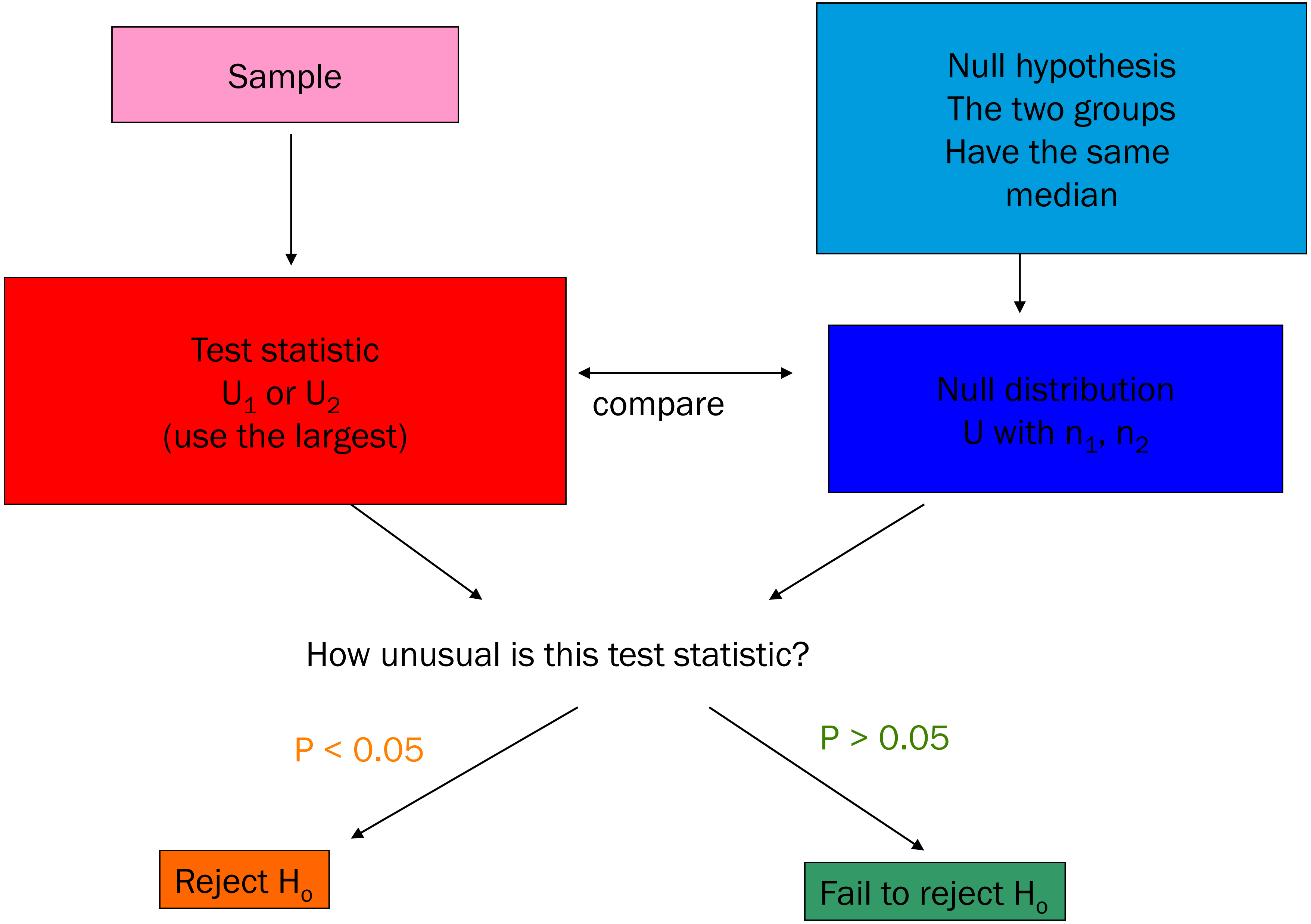
Nonparametric Tests

Nonparametric Tests

- Nonparametric tests are those that do not rely on probability distributions for population parameters
- They are used when
  - Data are badly skewed
  - Sample sizes are small
  - Data are not on an interval or ratio scale

Parametric	Nonparametric
One-sample and Paired t-test	Sign test
Two-sample t-test	Mann-Whitney U-test

Mann-Whitney U test



Mann-Whitney U test

- Large-sample approximation:

$$Z = \frac{2U - n_1 n_2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/3}}$$

Use this when n_1 & n_2 are both > 10
Compare to the standard normal distribution

Mann-Whitney U Test

- If you have ties:
  - Rank them anyway, pretending they were slightly different
  - Find the average of the ranks for the identical values, and give them all that rank
  - Carry on as if all the whole-number ranks have been used up

Example

Data

14
2
5
4
2
14
18
14

Example

Data	Sorted Data
14	2
2	2
5	4
4	5
2	14
14	14
18	14
14	18

Example

Data	Sorted Data
14	2
2	2
5	4
4	5
2	14
14	14
18	14
14	18

TIES (arrows point to the two 2s and the three 14s in Sorted Data)

Example

Data	Sorted Data
14	2
2	2
5	4
4	5
2	14
14	14
18	14
14	18

TIES

Rank them anyway, pretending they were slightly different

Example

Data	Sorted Data	Rank A
14	2	1
2	2	2
5	4	3
4	5	4
2	14	5
14	14	6
18	14	7
14	18	8

Example

Data	Sorted Data	Rank A
14	2	1
2	2	2
5	4	3
4	5	4
2	14	5
14	14	6
18	14	7
14	18	8

Find the average of the ranks for the identical values, and give them all that rank

Example

Data	Sorted Data	Rank A	
14	2	1	Average = 1.5
2	2	2	
5	4	3	
4	5	4	
2	14	5	
14	14	6	
18	14	7	Average = 6
14	18	8	

Example

Data	Sorted Data	Rank A	Rank
14	2	1	1.5
2	2	2	1.5
5	4	3	3
4	5	4	4
2	14	5	6
14	14	6	6
18	14	7	6
14	18	8	8

Example

Data	Sorted Data	Rank A	Rank
14	2	1	1.5
2	2	2	1.5
5	4	3	3
4	5	4	4
2	14	5	6
14	14	6	6
18	14	7	6
14	18	8	8

These can now be used for the Mann-Whitney U test

Benefits and Costs of Nonparametric Tests

- Main benefit:
  - Make fewer assumptions about your data
  - E.g. only assume random sample
- Main cost:
  - Reduce statistical power
  - Increased chance of Type II error

When Should I Use Nonparametric Tests?

- When you have reason to suspect the assumptions of your test are violated
  - Non-normal distribution
  - No transformation makes the distribution normal
  - Different variances for two groups

Quick Reference Summary: Sign Test

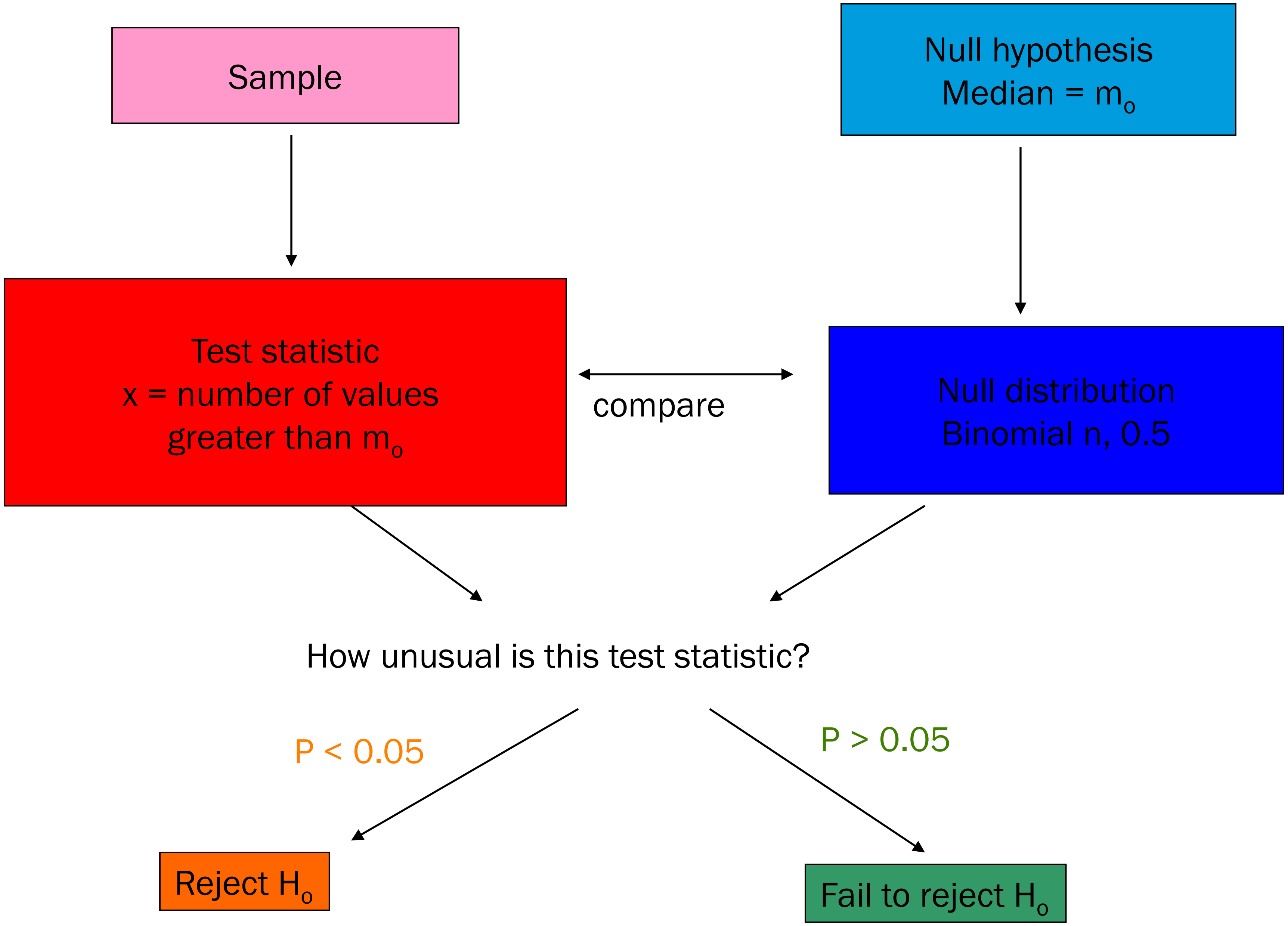
- What is it for? A non-parametric test to compare the medians of a group to some constant
- What does it assume? Random samples
- Formula: Identical to a binomial test with $p_o = 0.5$. Uses the number of subjects with values greater than and less than a hypothesized median as the test statistic.

P(x) = probability of a total of x successes
p = probability of success in each trial
n = total number of trials

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

P = 2 * Pr[x≥X]

Sign test


Sample
Null hypothesis
Median = m_o
Test statistic
x = number of values greater than m_o
compare
Null distribution
Binomial n, 0.5
How unusual is this test statistic?
$P < 0.05$
$P > 0.05$
Reject H_o
Fail to reject H_o

Quick Reference Summary: Mann-Whitney U Test

- What is it for? A non-parametric test to compare the central tendencies of two groups
- What does it assume? Random samples
- Test statistic: U
- Distribution under H_o: U distribution, with sample sizes n_1 and n_2
- Formulae:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

n_1= sample size of group 1
n_2= sample size of group 2
R_1= sum of ranks of group 1

$$U_2 = n_1 n_2 - U_1$$

Use the larger of U1 or U2 for a two-tailed test