

*"Well, I'll be damned if I'll defend to the death your right to say something that's statistically incorrect."*

# Statistical Comparison of Algorithms — Part I

Leandro L. Minku  
University of Birmingham, UK

# Overview

- General idea underlying statistical hypothesis tests.
- Statistical hypotheses.
- Choosing a significance level.
- Choosing a statistical hypothesis test.
- Interpreting test results.
- Criticism over statistical hypothesis tests.
- Confidence intervals.
- Effect size.
- Box plots.

# Comparison of Experimental Results

- Computer science research studies may involve the **comparison** of different algorithms, configurations of an algorithm, methods or approaches.
- What exactly you wish to compare depends on your research questions, which **must be formulated** before deciding whether and what to compare.
- Comparisons are not always straightforward when our algorithms or their inputs present **stochastic** behaviour.
  - Evolutionary algorithms.
  - Machine learning.

If different runs may give different results, how to compare them?

# Groups of Observations

You can treat the performance of your algorithm as a random variable, and perform multiple runs to get an idea of its underlying distribution.

	Performance for A1	Performance for A2
Runs	0.6015110151	0.0633347888
	0.2947677998	1.0930402922
	0.9636589224	0.1792341981
	0.251976978	1.207096969
	0.3701006544	1.0606484322
	0.9940754515	0.6473818857
	0.4283523627	0.8043431063
	0.1904817054	0.658958582
	0.7377491128	1.0576089397
	0.5392380701	0.7364416374
	0.4230920852	0.1942901434
	0.7221442924	0.5849134532
	0.8882444038	0.4971571929
	0.3186565207	0.2973731101
	0.5532666035	0.9801976669
	0.8306283304	0.1366545414
	0.4488794934	0.258875354
	0.6386464711	1.3587444717
	0.703989767	1.0901669778
	0.1133421799	0.5101653608
0.9693252021	0.6768334243	
0.4042517894	1.3479477059	
0.6884307214	1.1339212937	
0.1627650897	1.154985441	
0.5280297005	1.0054153791	
0.6990777731	1.0128717172	
0.020703112	0.5093192254	
0.580238106	1.3938111293	
0.5673830342	0.790654944	
0.2294966863	1.3811101009	

In statistics, each of the cells is referred to as an **observation**, and each column is called a **group or sample**, the performance metric being monitored is the **response**, and the algorithms are **treatments**.

How to compare the random variables based on their corresponding groups of observations?

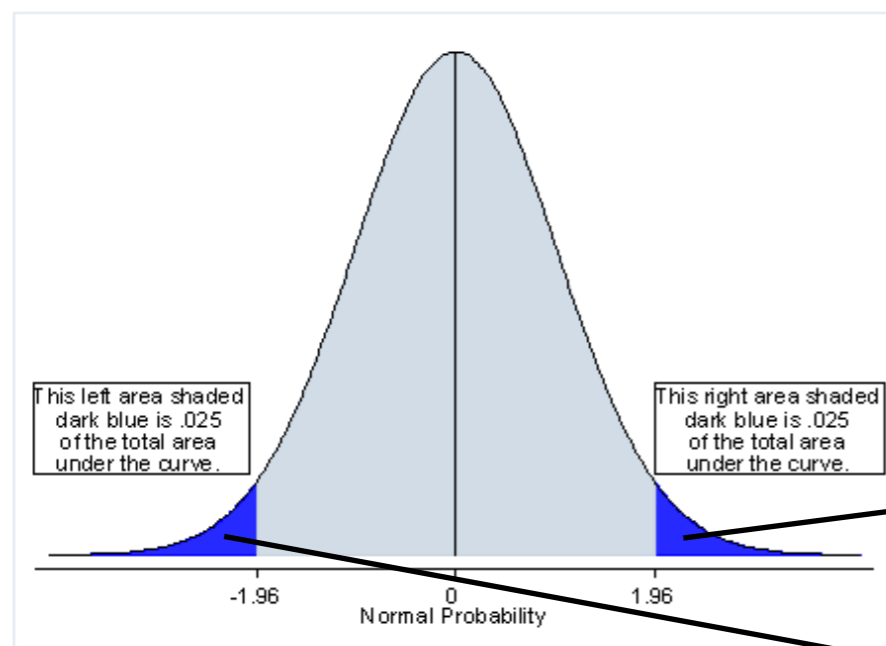
# Statistical Hypothesis Tests

**Statistical hypothesis:** assertion or conjecture about the distribution of one or more random variables.

**Statistical hypothesis test:** rule or procedure to decide whether to **reject** a hypothesis.

# General Idea — Z Test for Two Population Means, Variance Known

- Formulate Hypotheses:
  - $H_0: \mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0$
  - $H_1: \mu_1 \neq \mu_2 \rightarrow \mu_1 - \mu_2 \neq 0$
- Level of significance  $\alpha = 0.05$  (probability of Type I error).
- Test statistic  $Z = \frac{M_1 - M_2}{\sigma/\sqrt{N}}$
- Theoretical sampling distribution of the test statistic assuming  $H_0$  is true: normal distribution.



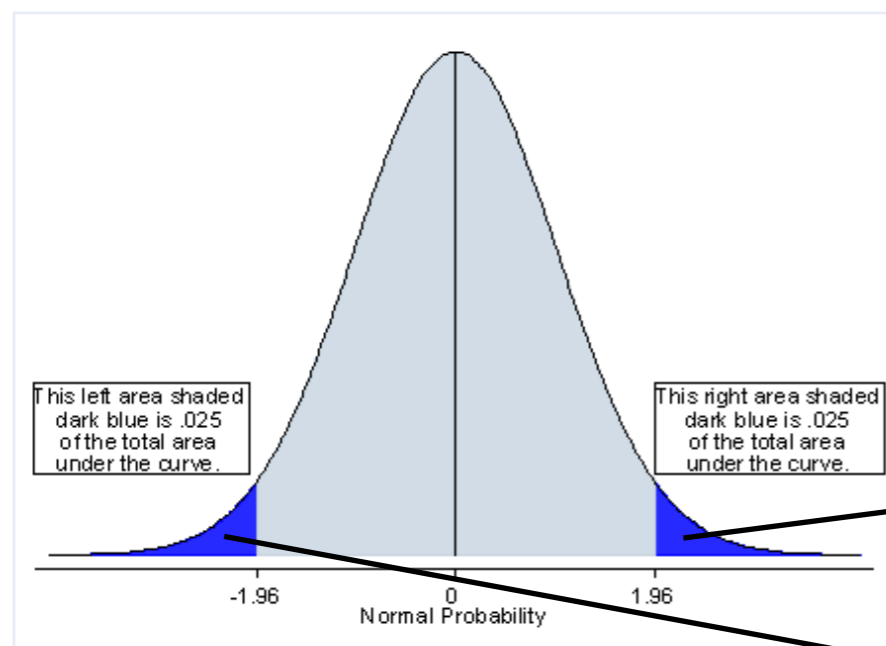
Probability of observing test statistic values  $\leq -1.96$  or  $\geq 1.96$  assuming that  $H_0$  is true is  $\alpha = 0.05$ .

= 1/2 level of significance of 0.05

= 1/2 level of significance of 0.05

# General Idea — Z Test for Two Population Means, Variance Known

- Formulate Hypotheses:
  - $H_0: \mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0$
  - $H_1: \mu_1 \neq \mu_2 \rightarrow \mu_1 - \mu_2 \neq 0$
- Level of significance  $\alpha = 0.05$  (probability of Type I error).
- Test statistic  $Z = \frac{M_1 - M_2}{\sigma/\sqrt{N}}$
- Theoretical sampling distribution of the test statistic assuming  $H_0$  is true: normal distribution.



If the test statistic falls in this region, we will reject  $H_0$ .

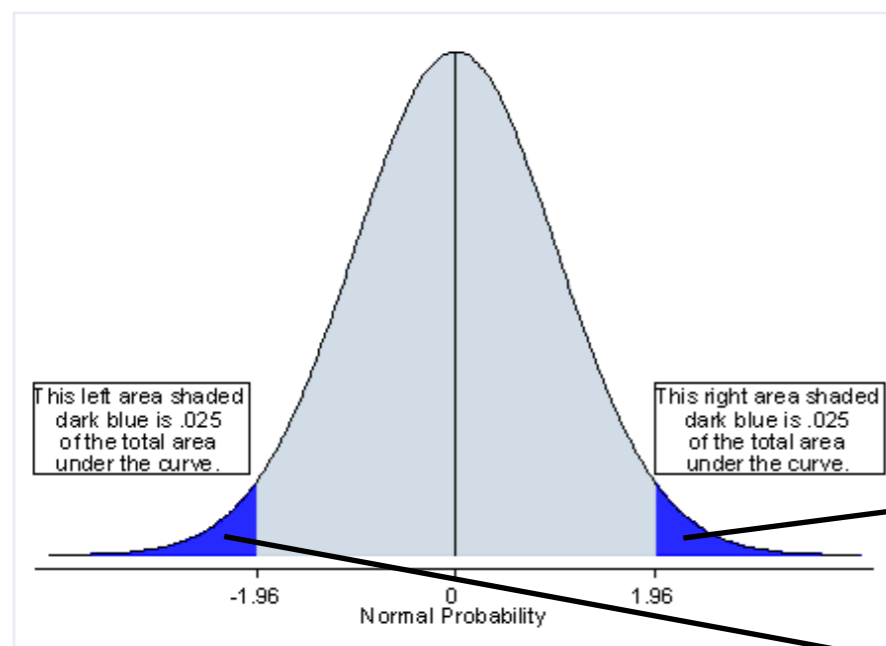
= 1/2 level of significance of 0.05

= 1/2 level of significance of 0.05



# General Idea — Z Test for Two Population Means, Variance Known

- Formulate Hypotheses:
  - $H_0: \mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0$
  - $H_1: \mu_1 \neq \mu_2 \rightarrow \mu_1 - \mu_2 \neq 0$
- Level of significance  $\alpha = 0.05$  (probability of Type I error).
- Test statistic  $Z = \frac{M_1 - M_2}{\sigma/\sqrt{N}}$
- Theoretical sampling distribution of the test statistic assuming  $H_0$  is true: normal distribution.



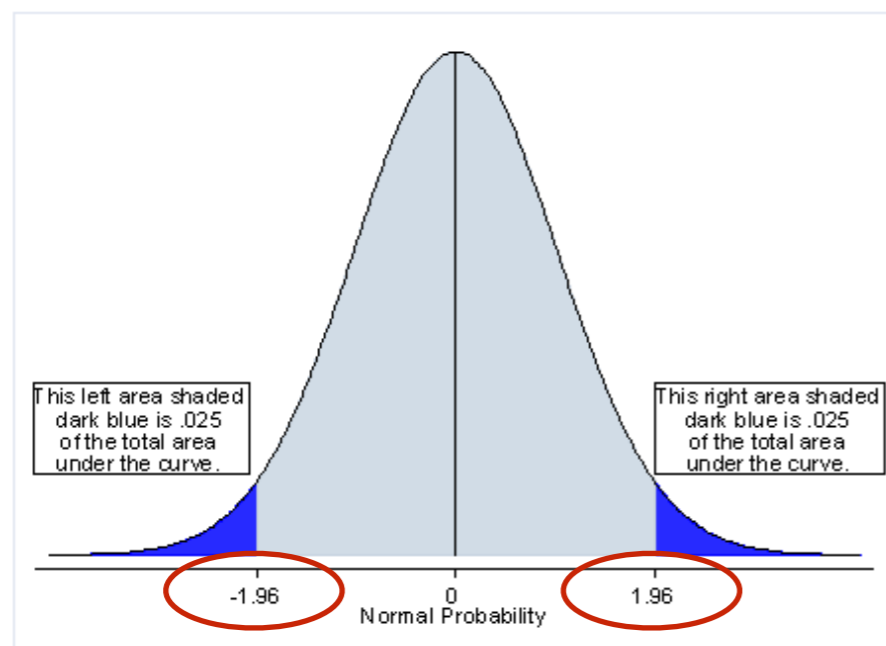
Critical region is the set of test statistic values that would lead to rejecting  $H_0$ .

= 1/2 level of significance of 0.05

= 1/2 level of significance of 0.05

# General Idea — Z Test for Two Population Means, Variance Known

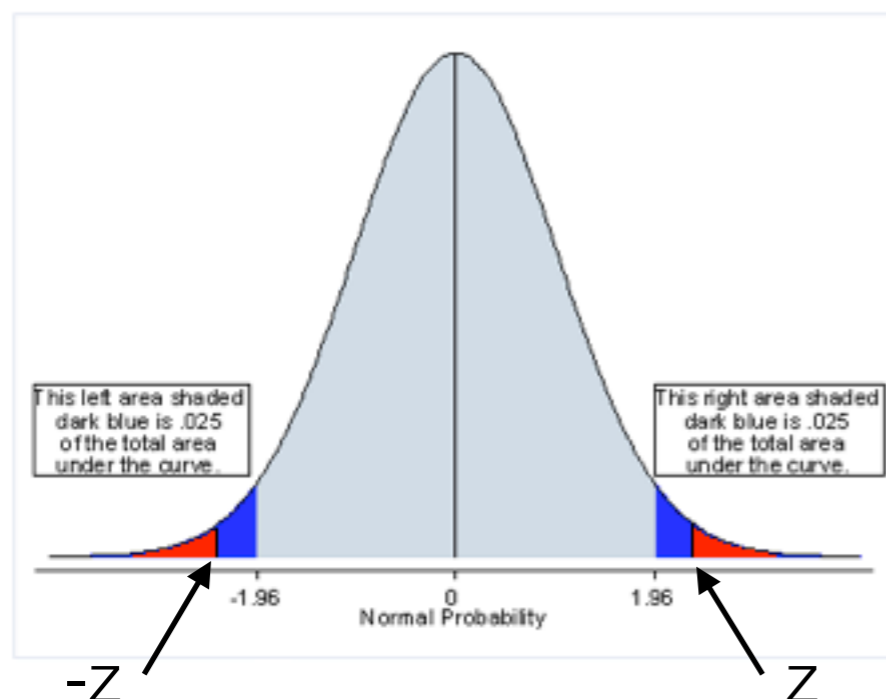
- Formulate Hypotheses:
  - $H_0: \mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0$
  - $H_1: \mu_1 \neq \mu_2 \rightarrow \mu_1 - \mu_2 \neq 0$
- Level of significance  $\alpha = 0.05$  (probability of Type I error).
- Test statistic  $Z = \frac{M_1 - M_2}{\sigma/\sqrt{N}}$
- Theoretical sampling distribution of the test statistic assuming  $H_0$  is true: normal distribution.



**Critical values** are the “boundary” values of the critical region.

# General Idea — Z Test for Two Population Means, Variance Known

- Formulate Hypotheses:
  - $H_0: \mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0$
  - $H_1: \mu_1 \neq \mu_2 \rightarrow \mu_1 - \mu_2 \neq 0$
- Level of significance  $\alpha = 0.05$  (probability of Type I error).
- Test statistic  $Z = \frac{M_1 - M_2}{\sigma/\sqrt{N}}$
- Theoretical sampling distribution of the test statistic assuming  $H_0$  is true: normal distribution.



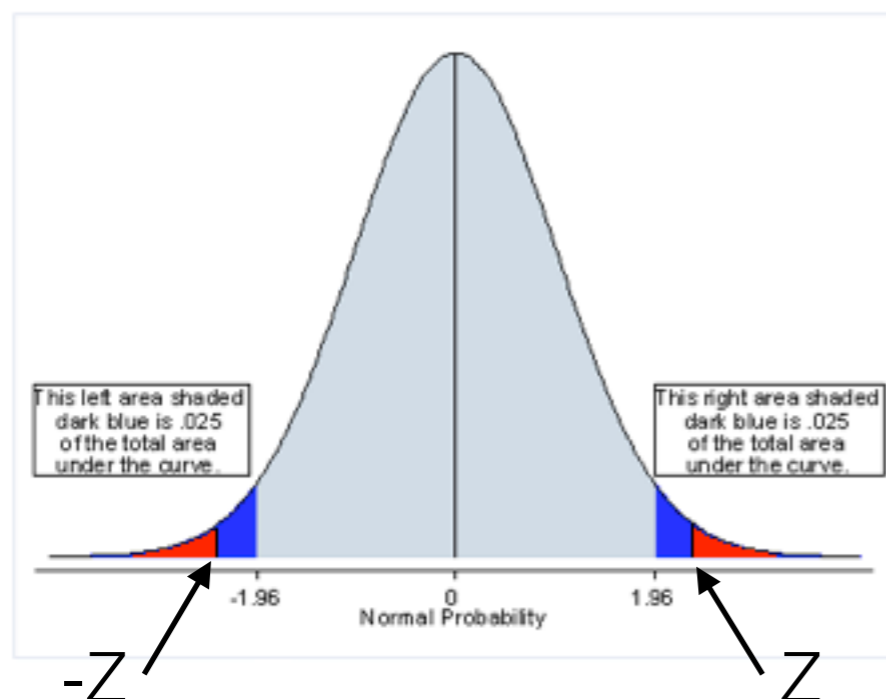
- **P-value:** probability of observing test statistic value at least as extreme as the value  $z$ , assuming  $H_0$ , is the AUC of the region starting at  $z$  and  $-z$ .
- If  $p\text{-value} \leq \alpha$ , reject  $H_0$ .
- Otherwise, do not reject  $H_0$

# General Idea — Z Test for Two Population Means, Variance Known

- **Formulate Hypotheses:**

- **$H_0: \mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0$**
- **$H_1: \mu_1 \neq \mu_2 \rightarrow \mu_1 - \mu_2 \neq 0$**

- Level of significance  $\alpha = 0.05$  (probability of Type I error).
- Test statistic  $Z = \frac{M_1 - M_2}{\sigma/\sqrt{N}}$
- Theoretical sampling distribution of the test statistic assuming  $H_0$  is true: normal distribution.



- **P-value:** probability of observing test statistic value at least as extreme as the value  $z$ , assuming  $H_0$ , is the AUC of the region starting at  $z$  and  $-z$ .
- If  $p\text{-value} \leq \alpha$ , reject  $H_0$ .
- Otherwise, do not reject  $H_0$

# Statistical Hypotheses for Two Groups — Two-Tailed Tests

- Two-tailed (two-sided) test, e.g.:

Null Hypothesis

$$H_0: \mu_1 = \mu_2 \longrightarrow \mu_1 - \mu_2 = 0$$

H<sub>0</sub> is the hypothesis being tested.

Alternative Hypothesis

$$H_1: \mu_1 \neq \mu_2 \longrightarrow \mu_1 - \mu_2 \neq 0$$

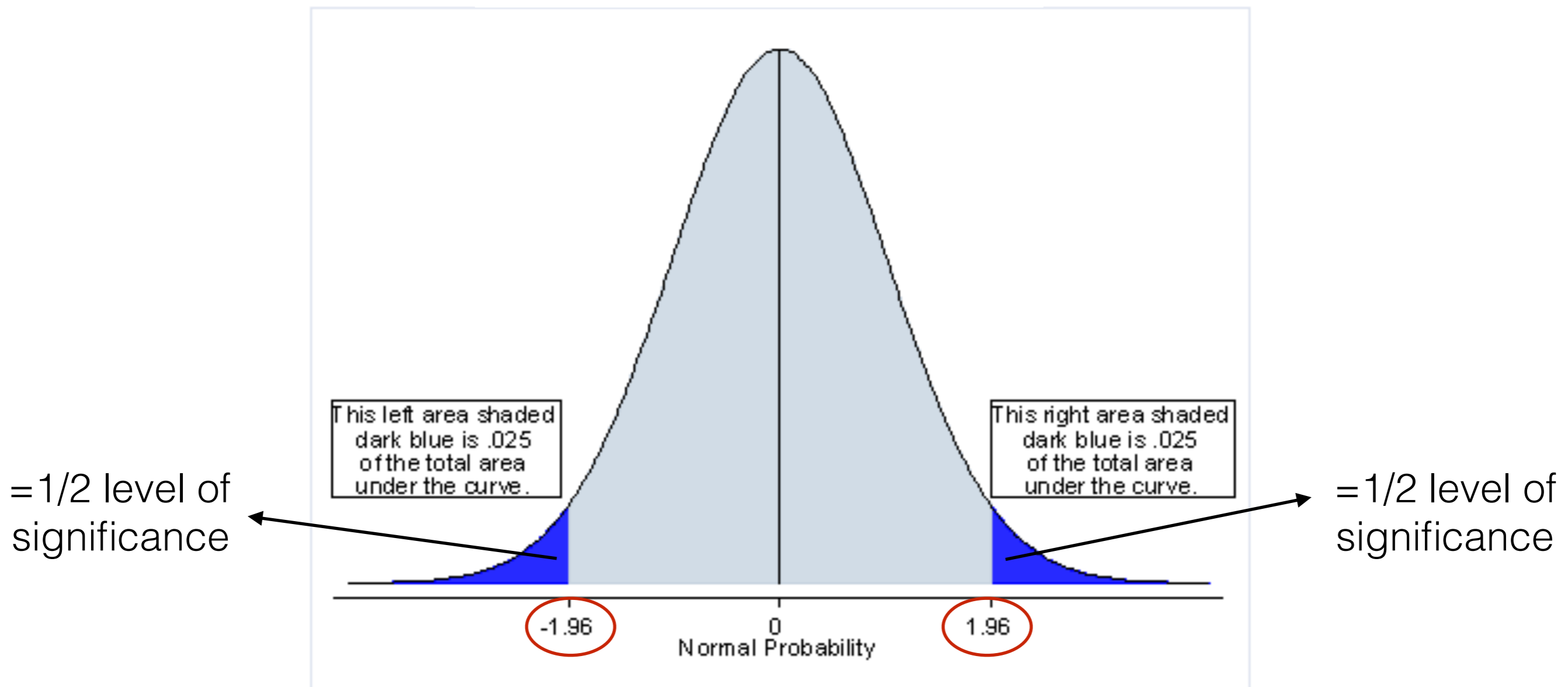
H<sub>1</sub> is usually the desirable outcome, that would lead to an action being taken.

The statistical hypothesis test will check if there is enough evidence to reject H<sub>0</sub> in favour of H<sub>1</sub>.

# Two-Tailed Tests

Level of significance  $\alpha = 0.05$

Distribution of the test statistic

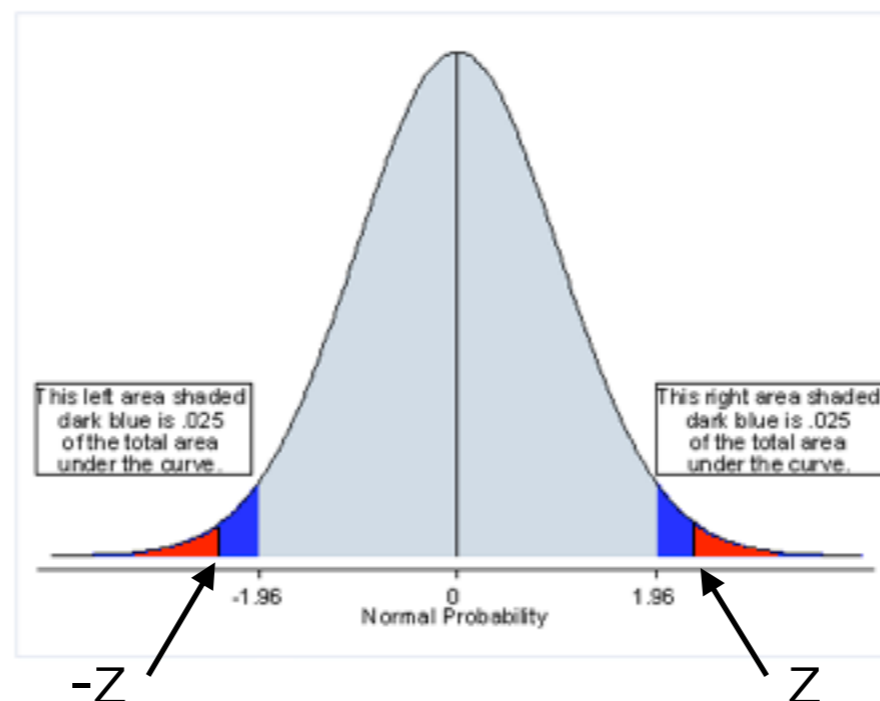


Extreme values of the test statistic may occur both in the left and in the right side/tail of the sampling distribution.

# Why H0 (and not H1) Contains Equality?

- In other words, why is the hypothesis being tested H0 and not H1?
  - H0:  $\mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0$
  - H1:  $\mu_1 \neq \mu_2 \rightarrow \mu_1 - \mu_2 \neq 0$

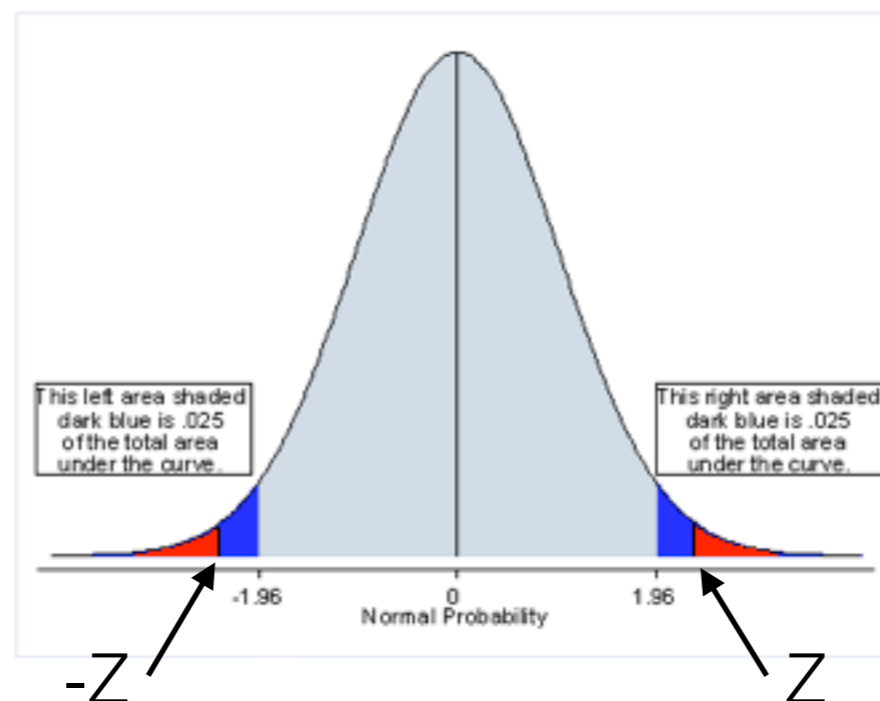
We will check the probability of observing a test statistic equal to or more extreme than  $z$ , assuming that H0 is true. For that, we can check how far  $z$  is from the mean of the sampling distribution for  $\mu_1 - \mu_2 = 0$ .



# Why H0 (and not H1) Contains Equality?

- In other words, why is the hypothesis being tested H0 and not H1?
  - H0:  $\mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0$
  - H1:  $\mu_1 \neq \mu_2 \rightarrow \mu_1 - \mu_2 \neq 0$

If we were testing H1 instead, we would need to check how far  $z$  is from the mean of the sampling distribution for  $\mu_1 - \mu_2 = X$ , where  $X$  is **unknown**!





# Statistical Hypotheses for Two Groups — One-Tailed Tests

- One-tailed (one-sided) test:

Null Hypothesis

$$H_0: \mu_1 \leq \mu_2 \longrightarrow \mu_1 - \mu_2 \leq 0$$

Alternative Hypothesis

$$H_1: \mu_1 > \mu_2 \longrightarrow \mu_1 - \mu_2 > 0$$

Null Hypothesis

$$H_0: \mu_1 \geq \mu_2 \longrightarrow \mu_1 - \mu_2 \geq 0$$

Alternative Hypothesis

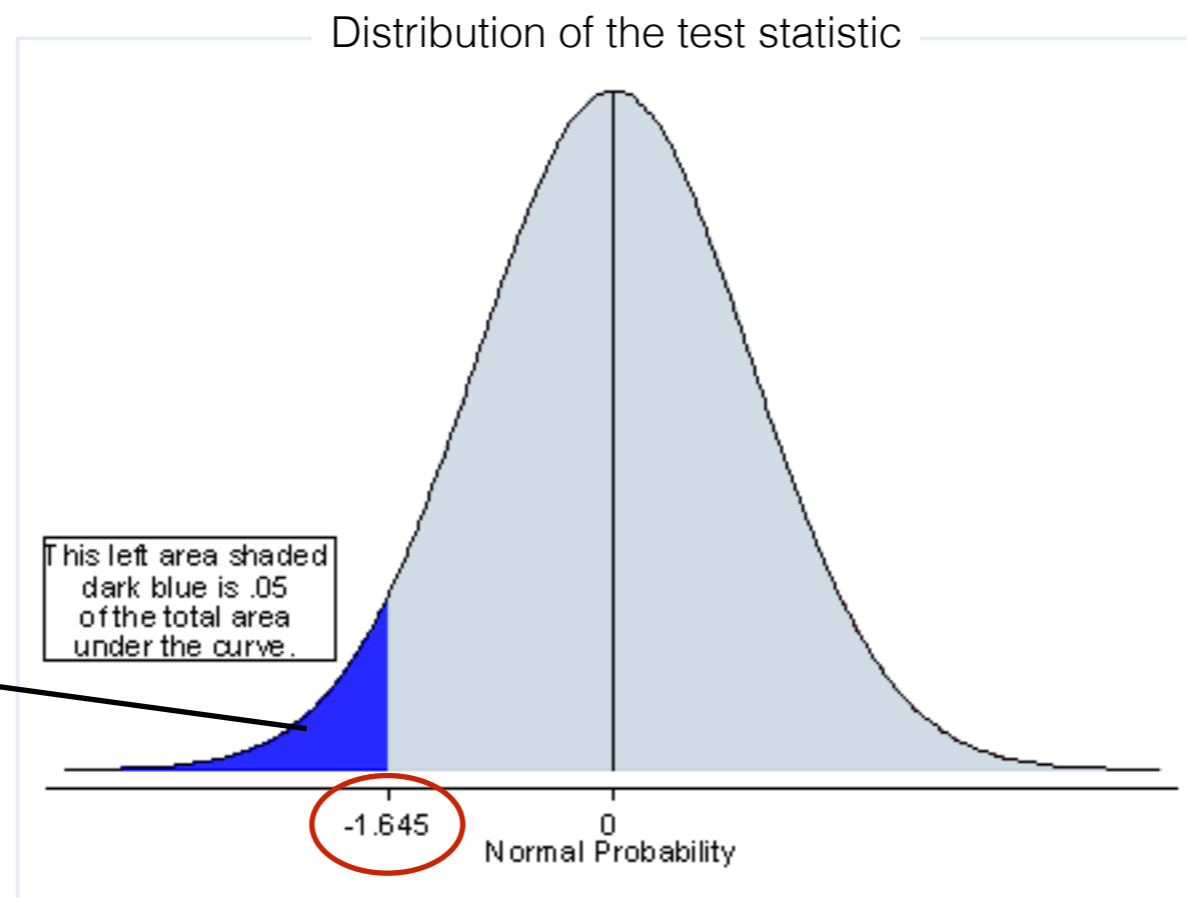
$$H_1: \mu_1 < \mu_2 \longrightarrow \mu_1 - \mu_2 < 0$$

# One-Tailed Tests

For one-tailed tests, deviations on only one side of a benchmark value are considered; Assumes that rare cases only occur to one side, **completely disregarding the possibility of deviations on the other side.**

$$H_0: \mu_1 \geq \mu_2 \longrightarrow H_0: \mu_1 - \mu_2 \geq 0$$

$$H_1: \mu_1 < \mu_2 \longrightarrow H_1: \mu_1 - \mu_2 < 0$$



= level of significance

Assumes that any occurrence of  $\mu_1 > \mu_2$  is purely by chance.

# One-Tailed Tests

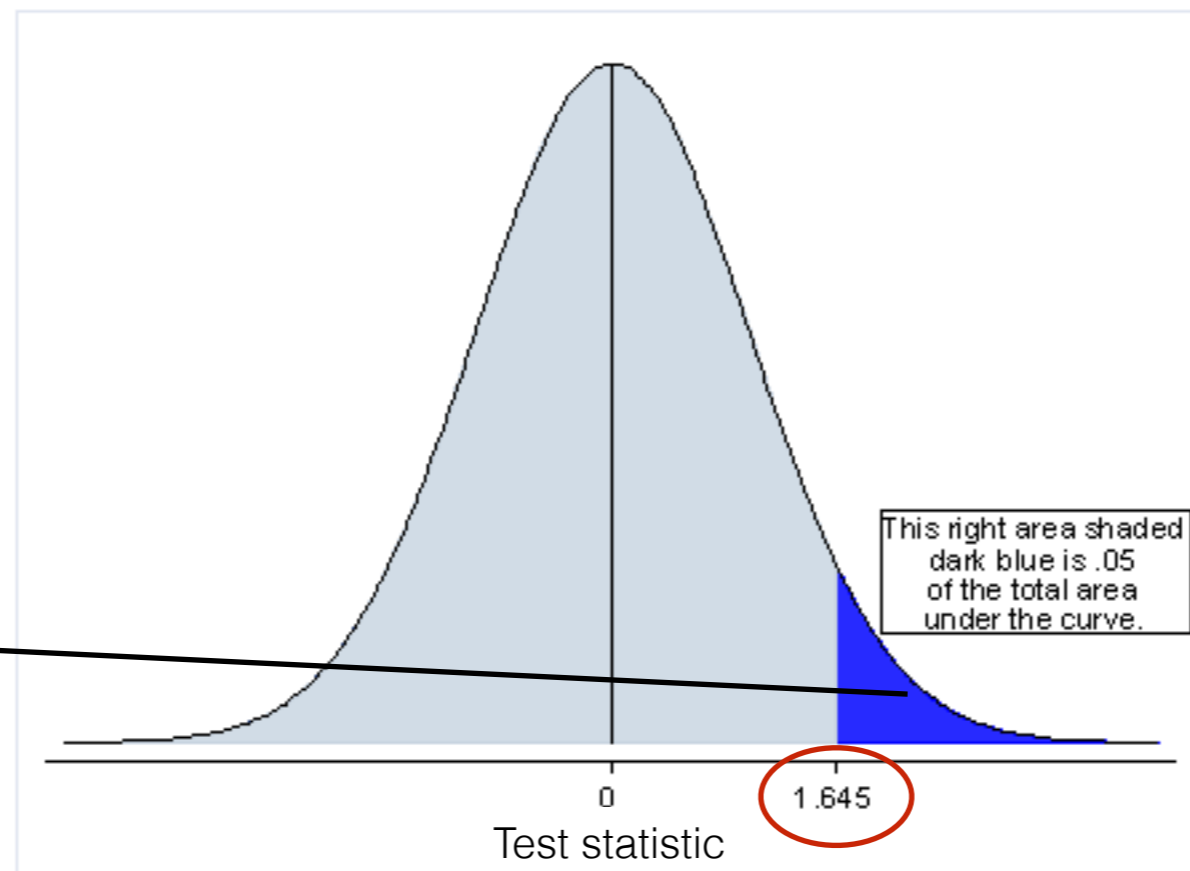
For one-tailed tests, deviations on only one side of a benchmark value are considered; Assumes that rare cases only occur to one side, **completely disregarding the possibility of deviations on the other side.**

$$H_0: \mu_1 \leq \mu_2 \longrightarrow H_0: \mu_1 - \mu_2 \leq 0$$

$$H_1: \mu_1 > \mu_2 \longrightarrow H_1: \mu_1 - \mu_2 > 0$$

Distribution of the test statistic

= level of significance



Assumes that any occurrence of  $\mu_1 < \mu_2$  is purely by chance.

**One-Tailed test is rarely used.**

# Concluding The Direction of the Difference When Using Two-Tailed Test

- It is still ok to reach a conclusion regarding the direction of the difference when using two-tailed test.
- This is because the two-tailed test can also be seen as two one-tailed tests, with level of significance of  $\alpha/2$  each.
- And it is impossible for both one-tailed tests to be significant simultaneously.
- So, you can use the position (left or right) of the test statistic to reach a conclusion on the direction of the difference.

# Statistical Hypotheses For N Groups

- Two-tailed (two-sided) test, e.g.:

Null  
Hypothesis

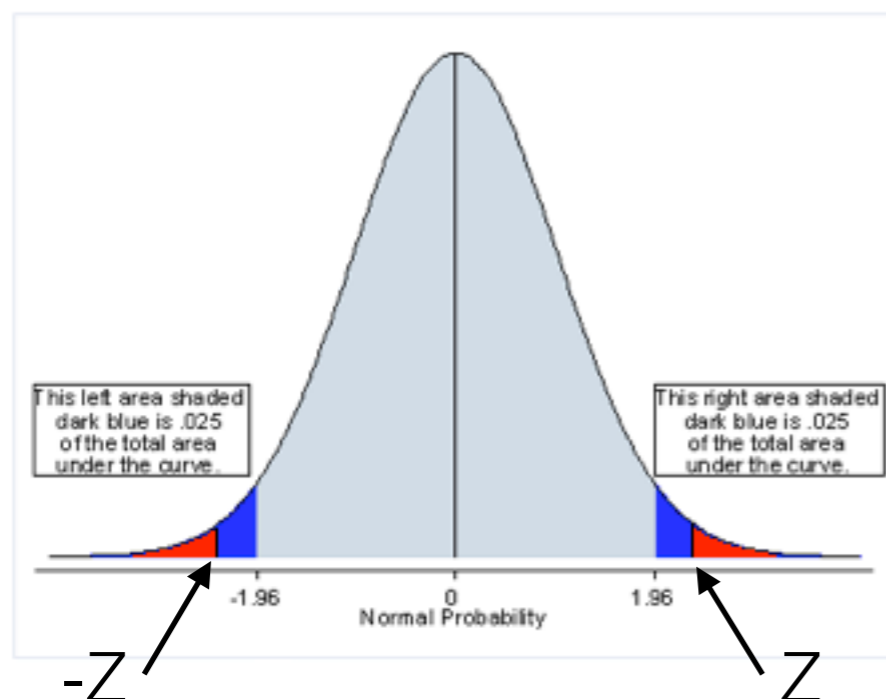
- $H_0: \mu_1 = \mu_2 = \dots = \mu_N$

Alternative  
Hypothesis

- $H_1: !(\mu_1 = \mu_2 = \dots = \mu_N)$

# General Idea — Z Test for Two Population Means, Variance Known

- Formulate Hypotheses:
  - $H_0: \mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0$
  - $H_1: \mu_1 \neq \mu_2 \rightarrow \mu_1 - \mu_2 \neq 0$
- **Level of significance  $\alpha = 0.05$  (probability of Type I error).**
- Test statistic  $Z = \frac{M_1 - M_2}{\sigma/\sqrt{N}}$
- Theoretical sampling distribution of the test statistic assuming  $H_0$  is true: normal distribution.



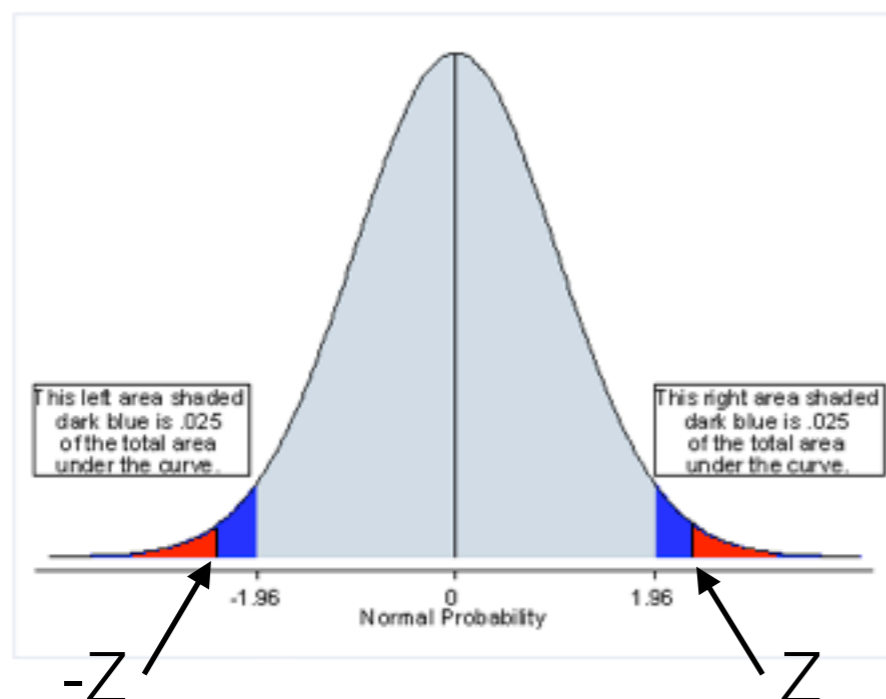
- **P-value:** probability of observing test statistic value at least as extreme as the value  $z$ , assuming  $H_0$ , is the AUC of the region starting at  $z$  and  $-z$ .
- If  $p\text{-value} \leq \alpha$ , reject  $H_0$ .
- Otherwise, do not reject  $H_0$

# Level of Significance ( $\alpha$ )

- $\alpha$  = probability of Type I error.
- Type I error: reject  $H_0$  when  $H_0$  is true.
- Usually,  $\alpha = 0.05$  is used.
- For more critical applications,  $\alpha = 0.01$  is typically used.
- Criticism tells us that the level of significance is chosen arbitrarily.
- Never choose a larger  $\alpha$  just to force the null hypothesis to be rejected.
- Ideally, choose  $\alpha$  before running the tests.

# General Idea — Z Test for Two Population Means, Variance Known

- Formulate Hypotheses:
  - $H_0: \mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0$
  - $H_1: \mu_1 \neq \mu_2 \rightarrow \mu_1 - \mu_2 \neq 0$
- Level of significance  $\alpha = 0.05$  (probability of Type I error).
- **Test statistic  $Z = \frac{M_1 - M_2}{\sigma/\sqrt{N}}$**
- **Theoretical sampling distribution of the test statistic assuming  $H_0$  is true: normal distribution.**



- **P-value:** probability of observing test statistic value at least as extreme as the value  $z$ , assuming  $H_0$ , is the AUC of the region starting at  $z$  and  $-z$ .
- If  $p\text{-value} \leq \alpha$ , reject  $H_0$ .
- Otherwise, do not reject  $H_0$



# Computing The Test Statistic and P-Value

- We don't normally need to compute them by ourselves.
- There are statistical tools that can be used to compute them for us.
  - R.
  - Matlab.
  - SPSS.
- Or, online resources can also compute them for you.
  - E.g.: <https://www.graphpad.com/quickcalcs/PValue1.cfm>

# Choosing Statistical Tests

- Different statistical hypothesis tests use different test statistics, which make different assumptions about the population underlying the collected observations (and consequently about the sampling distribution of the test statistic).

Data Distribution		2 groups	N groups (N>2)
Parametric (normality)	Unpaired (independent)	Unpaired t-test	ANOVA
	Paired (related)	Paired t-test	ANOVA
Non-parametric (no normality)	Unpaired (independent)	Wilcoxon rank-sum test = Mann–Whitney U test	Kruskal-Wallis test
	Paired (related)	Wilcoxon signed-rank test	Friedman test

# Which Test to Use?

Data Distribution		2 groups	N groups (N>2)
Parametric (normality)	Unpaired (independent)	Unpaired t-test	ANOVA
	Paired (related)	Paired t-test	ANOVA
Non-parametric (no normality)	Unpaired (independent)	Wilcoxon rank-sum test = Mann-Whitney U test	Kruskal-Wallis test
	Paired (related)	Wilcoxon signed-rank test	Friedman test

# Which Test to Use?

Data Distribution		2 groups	N groups (N>2)
Parametric (normality)	Unpaired (independent)	Unpaired t-test	ANOVA
	Paired (related)	Paired t-test	ANOVA
Non-parametric (no normality)	Unpaired (independent)	Wilcoxon rank-sum test = Mann-Whitney U test	Kruskal-Wallis test
	Paired (related)	Wilcoxon signed-rank test	Friedman test

# Comparison of N Groups

We may wish to compare  
A1 vs A2 vs A3.

or

A1 vs A2 vs A3 vs A4.

or

A1 vs A2 vs ... vs AN.

Result A1
0.8036808732
0.1546026852
0.1507085019
0.9751186599
0.4602321477
0.0132238786
0.0175114877
0.9041741739
0.8697700955
0.676352134
0.5182328166
0.0516411681
0.5426649651
0.4973629257
0.4866079125
0.2187455767
0.8438274211
0.2644009485
0.256434446
0.0791214858
0.2856093827
0.3797759169
0.5978962695
0.0860532501
0.2860286001
0.2772790031
0.7289846656
0.3812438862
0.114495351
0.7128328204

Result A2
0.9442552933
0.7277129425
0.4319811615
0.9379836847
0.786503003
0.8191139316
0.9236880897
0.8155635942
0.7694358404
0.3217702059
0.9849161406
0.2586409871
0.7945434749
0.8179485709
0.4132167082
0.5915588229
0.5936746635
0.4386923753
0.743990941
0.7951068189
0.3314508633
0.9218094004
0.7508496968
0.1372954398
0.1251753599
0.7858294814
0.4592977329
0.1583327209
0.4037452065
0.8074019616

Result A3
0.9442552933
0.7277129425
0.4319811615
0.9379836847
0.786503003
0.8191139316
0.9236880897
0.8155635942
0.7694358404
0.3217702059
0.9849161406
0.2586409871
0.7945434749
0.8179485709
0.4132167082
0.5915588229
0.5936746635
0.4386923753
0.743990941
0.7951068189
0.3314508633
0.9218094004
0.7508496968
0.1372954398
0.1251753599
0.7858294814
0.4592977329
0.1583327209
0.4037452065
0.8074019616

Result A4
0.9442552933
0.7277129425
0.4319811615
0.9379836847
0.786503003
0.8191139316
0.9236880897
0.8155635942
0.7694358404
0.3217702059
0.9849161406
0.2586409871
0.7945434749
0.8179485709
0.4132167082
0.5915588229
0.5936746635
0.4386923753
0.743990941
0.7951068189
0.3314508633
0.9218094004
0.7508496968
0.1372954398
0.1251753599
0.7858294814
0.4592977329
0.1583327209
0.4037452065
0.8074019616

# Pairwise Comparisons for N Groups

Potential way to compare  
A1 vs A2 vs A3:

A1 vs A2  
A1 vs A3  
A2 vs A3

Result A1
0.8036808732
0.1546026852
0.1507085019
0.9751186599
0.4602321477
0.0132238786
0.0175114877
0.9041741739
0.8697700955
0.676352134
0.5182328166
0.0516411681
0.5426649651
0.4973629257
0.4866079125
0.2187455767
0.8438274211
0.2644009485
0.256434446
0.0791214858
0.2856093827
0.3797759169
0.5978962695
0.0860532501
0.2860286001
0.2772790031
0.7289846656
0.3812438862
0.114495351
0.7128328204

Result A2
0.9442552933
0.7277129425
0.4319811615
0.9379836847
0.786503003
0.8191139316
0.9236880897
0.8155635942
0.7694358404
0.3217702059
0.9849161406
0.2586409871
0.7945434749
0.8179485709
0.4132167082
0.5915588229
0.5936746635
0.4386923753
0.743990941
0.7951068189
0.3314508633
0.9218094004
0.7508496968
0.1372954398
0.1251753599
0.7858294814
0.4592977329
0.1583327209
0.4037452065
0.8074019616

Result A3
0.9442552933
0.7277129425
0.4319811615
0.9379836847
0.786503003
0.8191139316
0.9236880897
0.8155635942
0.7694358404
0.3217702059
0.9849161406
0.2586409871
0.7945434749
0.8179485709
0.4132167082
0.5915588229
0.5936746635
0.4386923753
0.743990941
0.7951068189
0.3314508633
0.9218094004
0.7508496968
0.1372954398
0.1251753599
0.7858294814
0.4592977329
0.1583327209
0.4037452065
0.8074019616

# Pairwise Comparisons for N Groups

Potential way to compare  
A1 vs A2 vs A3 vs A4:

A1 vs A2

A1 vs A3

A1 vs A4

A2 vs A3

A2 vs A4

A3 vs A4

Result A1
0.8036808732
0.1546026852
0.1507085019
0.9751186599
0.4602321477
0.0132238786
0.0175114877
0.9041741739
0.8697700955
0.676352134
0.5182328166
0.0516411681
0.5426649651
0.4973629257
0.4866079125
0.2187455767
0.8438274211
0.2644009485
0.256434446
0.0791214858
0.2856093827
0.3797759169
0.5978962695
0.0860532501
0.2860286001
0.2772790031
0.7289846656
0.3812438862
0.114495351
0.7128328204

Result A2
0.9442552933
0.7277129425
0.4319811615
0.9379836847
0.786503003
0.8191139316
0.9236880897
0.8155635942
0.7694358404
0.3217702059
0.9849161406
0.2586409871
0.7945434749
0.8179485709
0.4132167082
0.5915588229
0.5936746635
0.4386923753
0.743990941
0.7951068189
0.3314508633
0.9218094004
0.7508496968
0.1372954398
0.1251753599
0.7858294814
0.4592977329
0.1583327209
0.4037452065
0.8074019616

Result A3
0.9442552933
0.7277129425
0.4319811615
0.9379836847
0.786503003
0.8191139316
0.9236880897
0.8155635942
0.7694358404
0.3217702059
0.9849161406
0.2586409871
0.7945434749
0.8179485709
0.4132167082
0.5915588229
0.5936746635
0.4386923753
0.743990941
0.7951068189
0.3314508633
0.9218094004
0.7508496968
0.1372954398
0.1251753599
0.7858294814
0.4592977329
0.1583327209
0.4037452065
0.8074019616

Result A4
0.9442552933
0.7277129425
0.4319811615
0.9379836847
0.786503003
0.8191139316
0.9236880897
0.8155635942
0.7694358404
0.3217702059
0.9849161406
0.2586409871
0.7945434749
0.8179485709
0.4132167082
0.5915588229
0.5936746635
0.4386923753
0.743990941
0.7951068189
0.3314508633
0.9218094004
0.7508496968
0.1372954398
0.1251753599
0.7858294814
0.4592977329
0.1583327209
0.4037452065
0.8074019616

**Problem:** multiple comparisons



JELLY BEANS CAUSE ACNE!

SCIENTISTS! INVESTIGATE!

BUT WE'RE PLAYING MINECRAFT! ... FINE.

WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE ( $P > 0.05$ ).

THAT SETTLES THAT.

I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.

SCIENTISTS!

BUT MINECRAFT!

WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE ( $P > 0.05$ ).

WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE ( $P > 0.05$ ).

WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE ( $P > 0.05$ ).

WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE ( $P < 0.05$ ).

WHOA!

WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE ( $P > 0.05$ ).

WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE ( $P > 0.05$ ).

WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE ( $P > 0.05$ ).

WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE ( $P > 0.05$ ).

WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE ( $P > 0.05$ ).

WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE ( $P > 0.05$ ).

WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE ( $P > 0.05$ ).

WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE ( $P > 0.05$ ).

WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE ( $P > 0.05$ ).

WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE ( $P > 0.05$ ).

WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE ( $P > 0.05$ ).

WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE ( $P > 0.05$ ).

WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE ( $P > 0.05$ ).

WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE ( $P > 0.05$ ).

WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE ( $P > 0.05$ ).

WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE ( $P > 0.05$ ).

NEWS

GREEN JELLY BEANS LINKED TO ACNE!

95% CONFIDENCE

ONLY 5% CHANCE OF COINCIDENCE!

SCIENTISTS...



# The Problem of Multiple Comparisons

- Statistical tests have some probability of presenting a Type 1 error. Let's say this probability is  $\alpha = 0.05$ .

Number of Tests	Probability of Getting At Least One Type 1 Error
1	0.05
2	$1 - (0.95^2) = 0.0975$
3	$1 - (0.95^3) = 0.1426$
...	...
100	$1 - (0.95^{100}) = 0.9941$

Probability of getting at least one type 1 error =  $1 -$  probability of getting no error

If we run multiple tests, we have increased chances of getting at least one Type 1 error.

# Dealing with Multiple Comparisons

- We can correct the level of significance (or p-value).
  - If  $p\text{-value} \leq \text{adjusted level of significance}$ , reject  $H_0$ .
- Bonferroni corrections:
  - Divide level of significance by number of comparisons.  
Example:
    - Level of significance = 0.05.
    - Number of comparisons = 10.
    - Adjusted level of significance = 0.005.
  - If  $p\text{-value} \leq \text{adjusted level of significance}$ , reject  $H_0$ .

**Problem:** very weak, i.e., likely to miss significant differences.

# Dealing with Multiple Comparisons

- Holm-Bonferroni corrections:
  - Example: level of significance = 0.05, number of comparisons = 4.

<b>i</b>	<b>p-value<sub>i</sub></b>	<b>Adjusted Significance 0.05 / i</b>	<b>Reject H0?</b>
4	0.0010	0.0125	Yes
3	0.0020	0.0167	Yes
2	0.0400	0.0250	No
1	0.0410	0.0500	No

Holm-Bonferroni corrections can still be weak, even though not so weak as Bonferroni.

# Statistical Tests For N Groups

Tests for N groups are **stronger** than pairwise comparisons with corrections, i.e., more likely to detect significant differences when they exist.

Data Distribution		2 groups	N groups (N>2)
Parametric (normality)	Unpaired (independent)	Unpaired t-test	ANOVA
	Paired (related)	Paired t-test	ANOVA
Non-parametric (no normality)	Unpaired (independent)	Wilcoxon rank-sum test = Mann-Whitney U test	Kruskal-Wallis test
	Paired (related)	Wilcoxon signed-rank test	Friedman test

# Statistical Tests For N Groups

- $H_0$ : all  $\mu$  are equal
- $H_1$ : at least one pair of  $\mu$  is different
- **Problem:**
  - Require post-hoc tests to find which pair of random variables is different.
  - Post-hoc tests are typically weaker than the tests for N groups.

# Scott-Knott



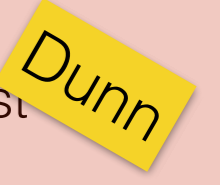

- Scott-Knott with non-parametric Bootstrap Sampling test can be used to cluster groups and then rank these clusters.
- May be stronger, as it avoids unnecessary comparisons.
- Not widely used in evolutionary computation or machine learning literature yet.
- Recently highly recommended by the software analytics literature.

N. Mittas and L. Angelis, Ranking and clustering software cost estimation models through a multiple comparisons algorithm, IEEE TSE, 39(4):537–551, 2013.

T. Menzies, Y. Yang, G. Mathew, B. Boehm, and J. Hihn, Negative results for software effort estimation, EMSE, 22(5):2658–2683, 2017.

Code at: <https://github.com/txt/ase16/blob/master/doc/stats.md>

# Post-hoc Tests

Data Distribution		2 groups	N groups (N>2)
Parametric (normality)	Unpaired (independent)	Unpaired t-test	ANOVA 
	Paired (related)	Paired t-test	ANOVA 
Non-parametric (no normality)	Unpaired (independent)	Wilcoxon rank-sum test = Mann-Whitney U test	Kruskal-Wallis test 
	Paired (related)	Wilcoxon signed-rank test	Friedman test 

# Which Test to Use?

Parametric tests make strong assumptions about the population being sampled, e.g., normality.

Data Distribution		2 groups	N groups (N>2)
Parametric (normality)	Unpaired (independent)	Unpaired t-test	ANOVA
	Paired (related)	Paired t-test	ANOVA
Non-parametric (no normality)	Unpaired (independent)	Wilcoxon rank-sum test = Mann-Whitney U test	Kruskal-Wallis test
	Paired (related)	Wilcoxon signed-rank test	Friedman test

The parametric tests above compare means of the underlying distributions.



# Which Test to Use?

Parametric tests are more powerful (better at detecting differences), than non-parametric ones, but can be highly affected by violations of assumptions.

Data Distribution		2 groups	N groups (N>2)
Parametric (normality)	Unpaired (independent)	Unpaired t-test	ANOVA
	Paired (related)	Paired t-test	ANOVA
Non-parametric (no normality)	Unpaired (independent)	Wilcoxon rank-sum test = Mann-Whitney U test	Kruskal-Wallis test
	Paired (related)	Wilcoxon signed-rank test	Friedman test

# Which Test to Use?

Factorial or split-plot ANOVA can be used to analyse the interaction between different factors used to create the groups, e.g., interaction between different parameters of an algorithm. No non-parametric tests are available for that.

Data Distribution		2 groups	N groups (N>2)
Parametric (normality)	Unpaired (independent)	Unpaired t-test	ANOVA
	Paired (related)	Paired t-test	ANOVA
Non-parametric (no normality)	Unpaired (independent)	Wilcoxon rank-sum test = Mann-Whitney U test	Kruskal-Wallis test
	Paired (related)	Wilcoxon signed-rank test	Friedman test

# Which Test to Use?

Non-parametric tests do not make strong assumptions about the population distribution, but are weaker than parametric tests when the assumptions are met.

Data Distribution		2 groups	N groups (N>2)
Parametric (normality)	Unpaired (independent)	Unpaired t-test	ANOVA
	Paired (related)	Paired t-test	ANOVA
Non-parametric (no normality)	Unpaired (independent)	Wilcoxon rank-sum test = Mann-Whitney U test	Kruskal-Wallis test
	Paired (related)	Wilcoxon signed-rank test	Friedman test

The non-parametric tests above compare medians of the underlying distributions.

# Normality Tests

E.g., Anderson-Darling test, D'Agostino-Pearson test, Kolmogorov-Smirnov test, Shapiro-Wilk test, Jarque–Bera test, etc.

Should we use normality tests to decide whether to use parametric tests?

- Controversy.
- Small number of observations, too weak: **more likely not to reject normality hypothesis.**
  - You may use parametric tests when you shouldn't.
  - Parametric tests are usually more sensitive to violations when the number of observations is small!
- Large number of observations: **very likely to reject normality hypothesis.**
  - You may opt for non-parametric tests exactly in the cases where the parametric tests are more robust to violations, i.e., when the number of observations is large.

# Should We Use Normality Tests to Decide Whether to Use Parametric Tests?

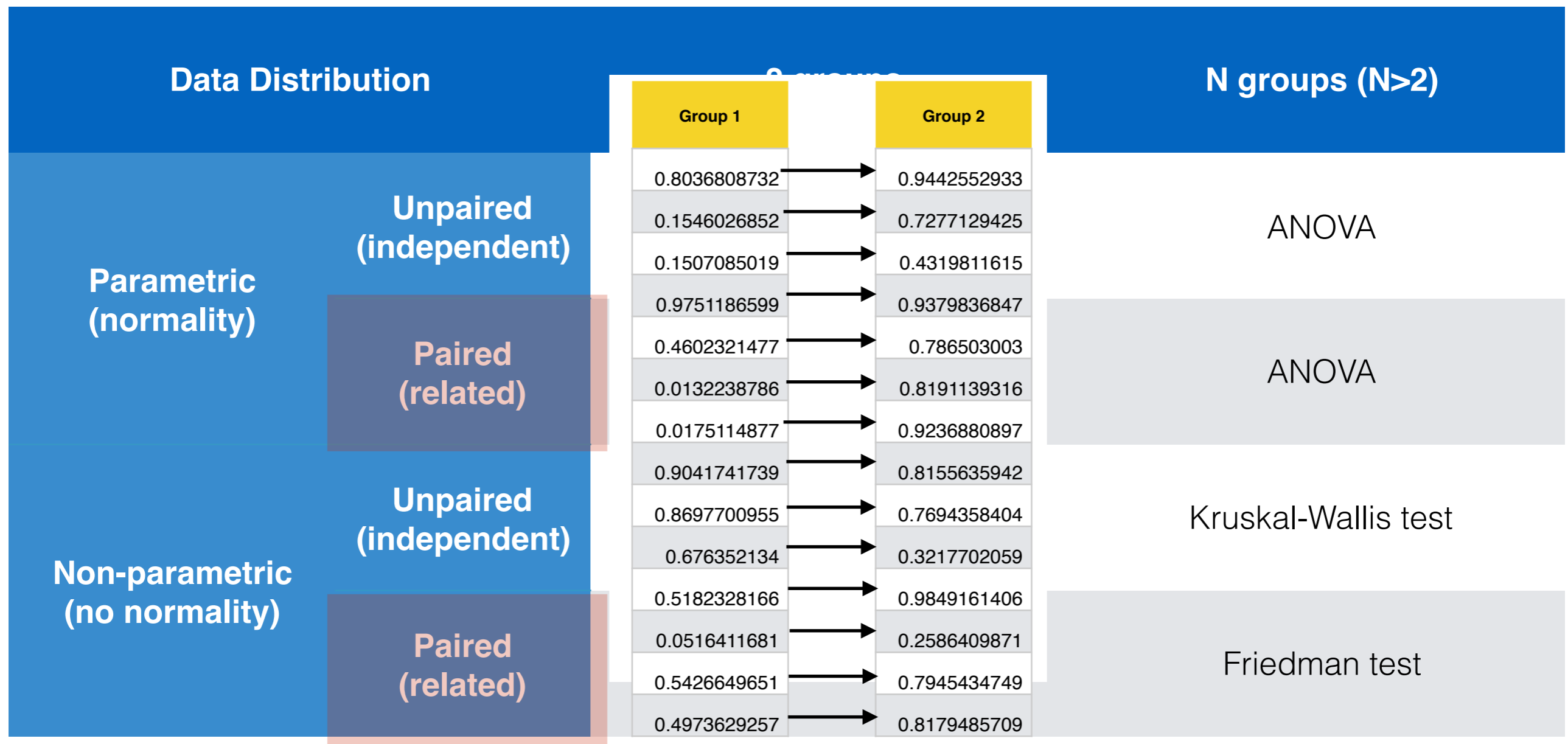
- Looking at histograms of the groups may be more informative than performing normality tests.
  - **Problem:** may not sound very "scientific".
- One may opt for using parametric tests when sample size is large.
  - **Problem:** some may argue that the test will still be affected.
- Many researchers opt for directly using non-parametric tests.
  - **Problem:** you may lose a bit of power.
- If you need to analyse the interactions between factors using factorial or split-plot ANOVA, p-value corrections can be applied to deal with certain violations of key assumptions.

# Which Test to Use?

Data Distribution		2 groups	N groups (N>2)
Parametric (normality)	Unpaired (independent)	Unpaired t-test	ANOVA
	Paired (related)	Paired t-test	ANOVA
Non-parametric (no normality)	Unpaired (independent)	Wilcoxon rank-sum test = Mann-Whitney U test	Kruskal-Wallis test
	Paired (related)	Wilcoxon signed-rank test	Friedman test

# Which Test to Use?

Example: using the same initial conditions to both groups (e.g., same initial weights or initial population).



# Which Test to Use?

Paired tests use more information, being more powerful (i.e., better at detecting significant differences if such differences exist).

Data Distribution		2 groups	N groups (N>2)
Parametric (normality)	Unpaired (independent)	Unpaired t-test	ANOVA
	Paired (related)	Paired t-test	ANOVA
Non-parametric (no normality)	Unpaired (independent)	Wilcoxon rank-sum test	Kruskal-Wallis test
	Paired (related)	Wilcoxon signed-rank test	Friedman test



# Which Test to Use?

Data Distribution		2 groups	N groups (N>2)
Parametric (normality)	Unpaired (independent)	Unpaired t-test	ANOVA
	Paired (related)	Paired t-test	ANOVA
Non-parametric (no normality)	Unpaired (independent)	Wilcoxon rank-sum test	Kruskal-Wallis test
	Paired (related)	Wilcoxon signed-rank test	Friedman test

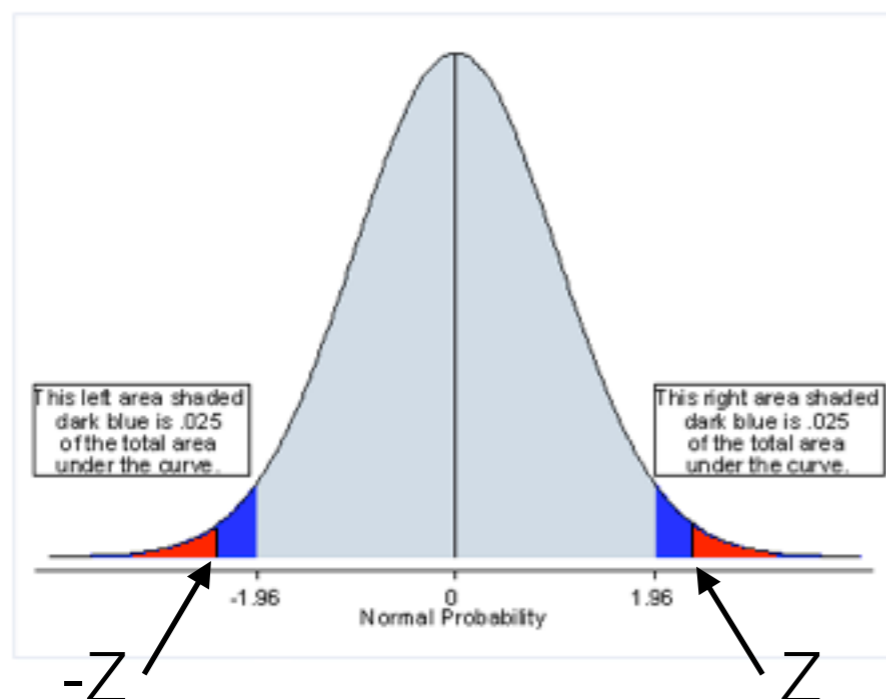
# Which Test to Use?

Example: using different random seeds or number of runs.

Data Distribution		2 groups		N groups (N>2)
		Group 1	Group 2	
Parametric (normality)	Unpaired (independent)	0.8036808732	0.9442552933	ANOVA
		0.1546026852	0.7277129425	
		0.1507085019	0.4319811615	
		0.9751186599	0.9379836847	
Non-parametric (no normality)	Paired (related)	0.4602321477	0.786503003	ANOVA
		0.0132238786	0.8191139316	
		0.0175114877	0.9236880897	Kruskal-Wallis test
	Unpaired (independent)	0.9041741739	0.8155635942	
		0.8697700955	0.7694358404	
		0.676352134	0.3217702059	
		0.5182328166		
	Paired (related)	0.0516411681		Friedman test
0.5426649651				
0.4973629257				

# General Idea — Z Test for Two Population Means, Variance Known

- Formulate Hypotheses:
  - $H_0: \mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0$
  - $H_1: \mu_1 \neq \mu_2 \rightarrow \mu_1 - \mu_2 \neq 0$
- Level of significance  $\alpha = 0.05$  (probability of Type I error).
- Test statistic  $Z = \frac{M_1 - M_2}{\sigma/\sqrt{N}}$
- Theoretical sampling distribution of the test statistic assuming  $H_0$  is true: normal distribution.



- **P-value:** probability of observing test statistic value at least as extreme as the value  $z$ , assuming  $H_0$ , is the AUC of the region starting at  $z$  and  $-z$ .
- **If p-value  $\leq \alpha$ , reject  $H_0$ .**
- **Otherwise, do not reject  $H_0$ .**

# Test Output

- Test statistic.
  - Used to reach the conclusion of whether or not to reject  $H_0$ .
  - Conclusion is reached by comparing it to the critical values, which depend on:
    - the theoretical sampling distribution,
    - the number of tails (1 or 2), and
    - the desired level of significance.
- P-Value.
  - Used to reach the conclusion of whether or not to reject  $H_0$ .
  - Conclusion is reached by comparing it to the level of significance.
- Whether or not  $H_0$  is rejected, given a level of significance.

# Terminology

- For two tailed test ( $H_0: \mu_1 = \mu_2$ ,  $H_1: \mu_1 \neq \mu_2$ ):
  - Not rejecting  $H_0$ : **no statistically significant difference** has been found between  $\mu_1$  and  $\mu_2$  at the level of significance of  $\alpha = 0.05$  (p-value of ...).
    - It doesn't mean that we accept  $H_0$ , it just means that we have not found enough evidence to reject it.
- G.K. Kanji. 100 Statistical Tests.  
Chapter "Introduction to Statistical Testing". SAGE Publications, 1993.
- Note that if the data are consistent with a given null hypothesis, they are also consistent with other similar hypotheses! So, we can't accept this specific hypothesis.
- Rejecting  $H_0$ : **statistically significant difference** between  $\mu_1$  and  $\mu_2$  has been found at the level of significance of  $\alpha = 0.05$  (p-value of ...).
  - Once we know they are significantly different, we can look at the **direction** of the differences to gain an insight into which of the algorithms is better.
    - $\mu_1$  is significantly larger than  $\mu_2$ .
    - $\mu_1$  is significantly smaller than  $\mu_2$ .

# Is The Magic Number 30 Good Enough?

- The idea that we should create groups of size 30 is argued to be inspired by the [Central Limit Theorem](#).
- The Central Limit Theorem states that the distribution of sample means approaches a normal distribution as the sample size approaches infinity.
- The idea of 30 comes from computer simulation experiments presented in introductory textbooks taking idealised computer samples from a normal, sometimes a skewed, distribution.
  - It showed that 30 was enough for the sampling distribution to be an approximation of the normal distribution.
- In reality, less normal samples may need to be larger. Samples that were not drawn perfectly randomly (i.e., biased samples) will also need to be larger. The sample size will also depend on the level of significance, power of the study, effect size, variability and desired precision.

Chuck Chakrapani. Statistical Reasoning vs. Magical Thinking. Vue 2011.

Sitanshu Sekhar Kar, Archana Ramalingam. Is 30 the Magic Number? Issues in Sample Size Estimation. National Journal of Community Medicine 4(1):175-179, 2013.

# Criticism Over Statistical Hypothesis Tests

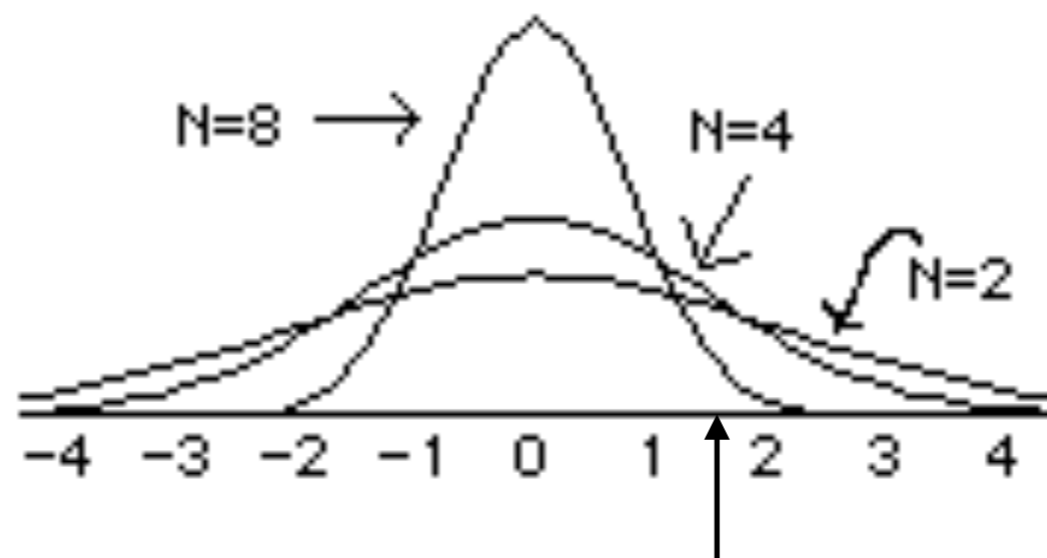
- $H_0$  means that there is no difference. However, it is virtually impossible for two populations in the real world to be identical.

S.L. Chow. Statistical Significance: Rationale, Validity and Utility.  
Chapter 1 (A Litany of Criticisms of NHSTP). SAGE Publications, 1996.

- In fact, the more data we collect the more likely it is for us to find a difference.
- People could use this to manipulate conclusions.

# Example — Sampling Distribution of the Mean

- The spread of the sampling distribution of the mean decreases as the sample size increases.
- Standard error of the mean reduces — we are estimating the true mean more confidently.
- Many more possible values of the test statistic would fall within the critical region.
- Therefore, it becomes easier to detect a significant difference with larger sample sizes.



A test statistic here may reject  $H_0$  when  $N=8$ , and not when  $N=4$  or 2.



# Effect Size

- Measures of effect size can be used to check how large the effect of the differences in performance are, independent of the sample size.
  - E.g., based on difference between means: Glass effect size (parametric)  $(\text{mean1} - \text{mean2}) / \text{std1}$ .
    - Problem: assumption of normal distribution.
  - E.g.: A12 (non-parametric).
    - Problem: effect size of the difference in ranks. Large differences in ranks don't mean that they are practically large.
- Do not use the p-value as an indication of effect size!

# Criticism Over Statistical Hypothesis Tests

- We can't conclude that we accept  $H_0$ , and many people don't realise that.

G.K. Kanji. 100 Statistical Tests.  
Chapter "Introduction to Statistical Testing". SAGE Publications, 1993.

- In fact, many books use the term "accept  $H_0$ " to say that we do not have enough evidence to reject it. This can be misleading.
- And what to do if we actually want to reach a conclusion about two algorithms performing similarly, i.e., accepting  $H_0$ ?

# Confidence Intervals

- **Range of values** computed in such a way that it contains the estimated parameter of the population with high probability.
- E.g., a 95% confidence interval could be  $0.1 \leq \mu_1 - \mu_2 \leq 0.3$ .
  - This means that there is a 95% chance that the interval contains the true value of  $\mu_1 - \mu_2$ .
- **Relationship with statistical tests:**
  - $\alpha = 1 - \text{confidence level}$ .
  - The confidence interval should contain zero if there is no evidence to reject  $H_0: \mu_1 - \mu_2 = 0$ .
  - But the confidence interval will also contain several other possible values.
- So, confidence intervals may be **more informative** when we wish to analyse a specific scenario.
- Statistical hypothesis tests could still be more helpful when there is a need for **summarising results** across several scenarios.

# Criticism Over Statistical Hypothesis Tests

- When we fail to reject  $H_0$ , there is still some chance that  $H_1$  was actually true, and people often ignore that.
  - This can lead to rejecting new algorithms that may in fact be quite useful.

# Boxplot

Max value within 1.5 IQR of the 3rd quartile,  
where  $IQR = 3rd\ quartile - 1st\ quartile$

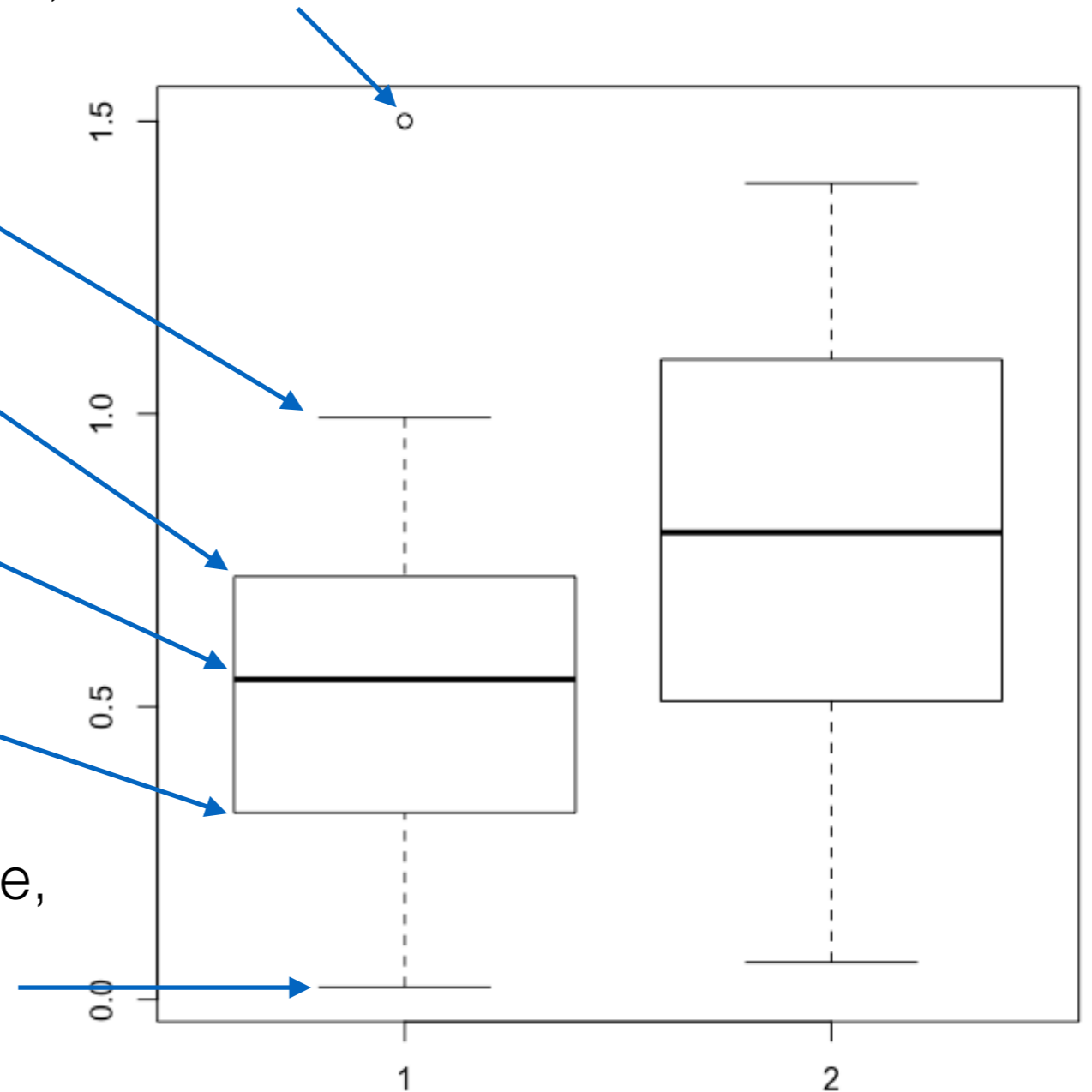
Outlier

3rd Quartile

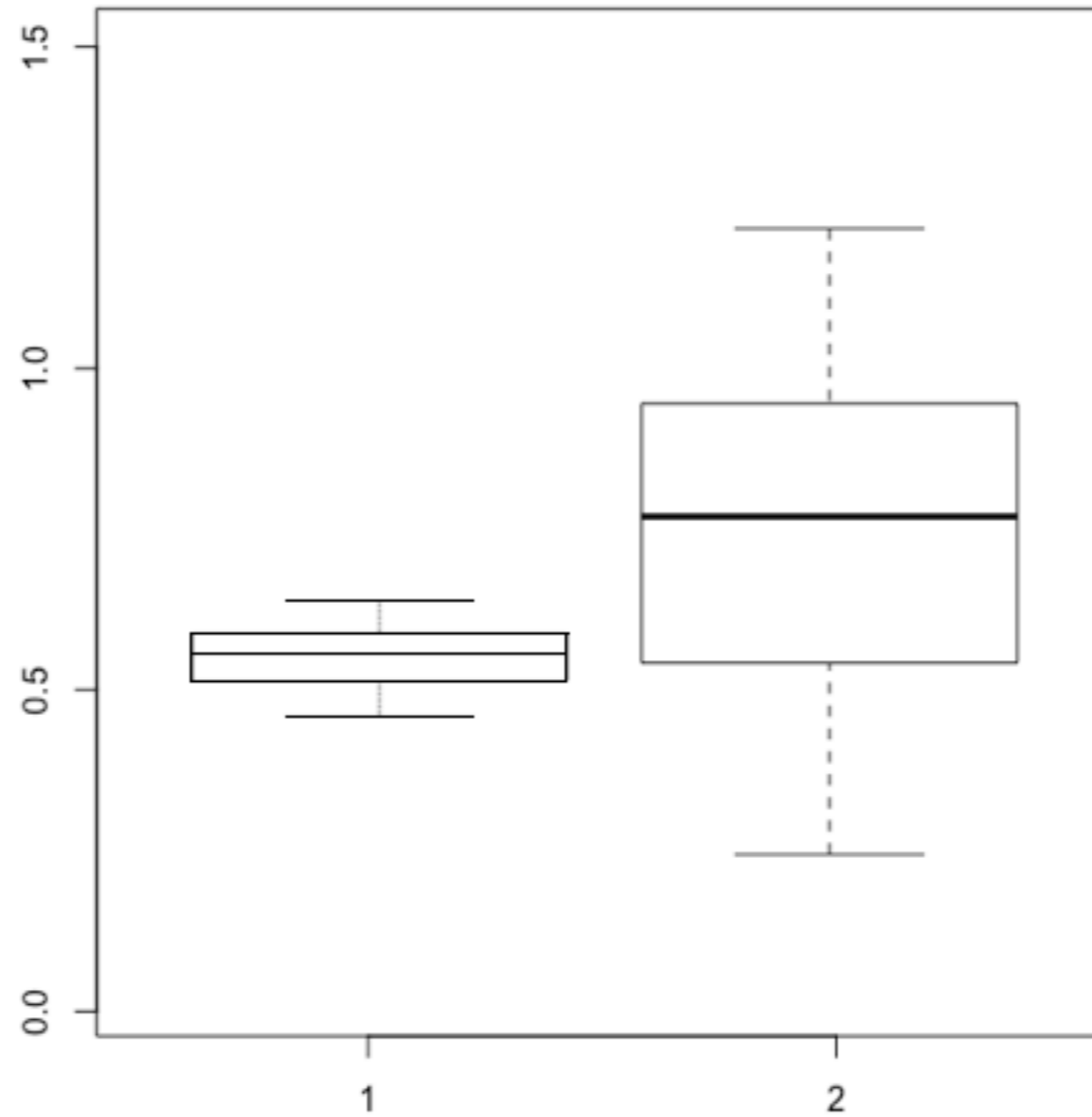
Median (2nd Quartile)

1st Quartile

Min value within 1.5 IQR of the 1st quartile,  
where  $IQR = 3rd\ quartile - 1st\ quartile$   
is the Inter Quartile Range



# Boxplot



# Summary

- Statistical hypothesis tests can be used to check if there is enough evidence to reject  $H_0$  in favour of  $H_1$ .
- Several statistical tests are available.
  - Each test has different strengths and limitations.
  - You need to decide which of these is more appropriate for your analysis.
  - Possibly, make the decision of test before designing your experiments, as the decision may influence the experimental design (e.g., paired samples)
- Given a choice of statistical test:
  - Decide hypotheses to be tested.
  - Decide level of significance  $\alpha$ .
  - Compute test statistic (and p-value).
  - Decide whether to reject  $H_0$ .

# Summary

- Criticism over statistical tests argues that, among others:
  - Statistical tests cannot accept  $H_0$ .
  - $H_0$  is very unlikely to be true in the real world.
  - Results are influenced by sample size.
  - $H_1$  may still be true, even if  $H_0$  is not rejected.
- Confidence intervals can be used if  $H_0$  is the hypothesis of interest.
- Effect size can be used to check the size of the effect of the differences.
- Box plots can be very useful tools to compare random variables.