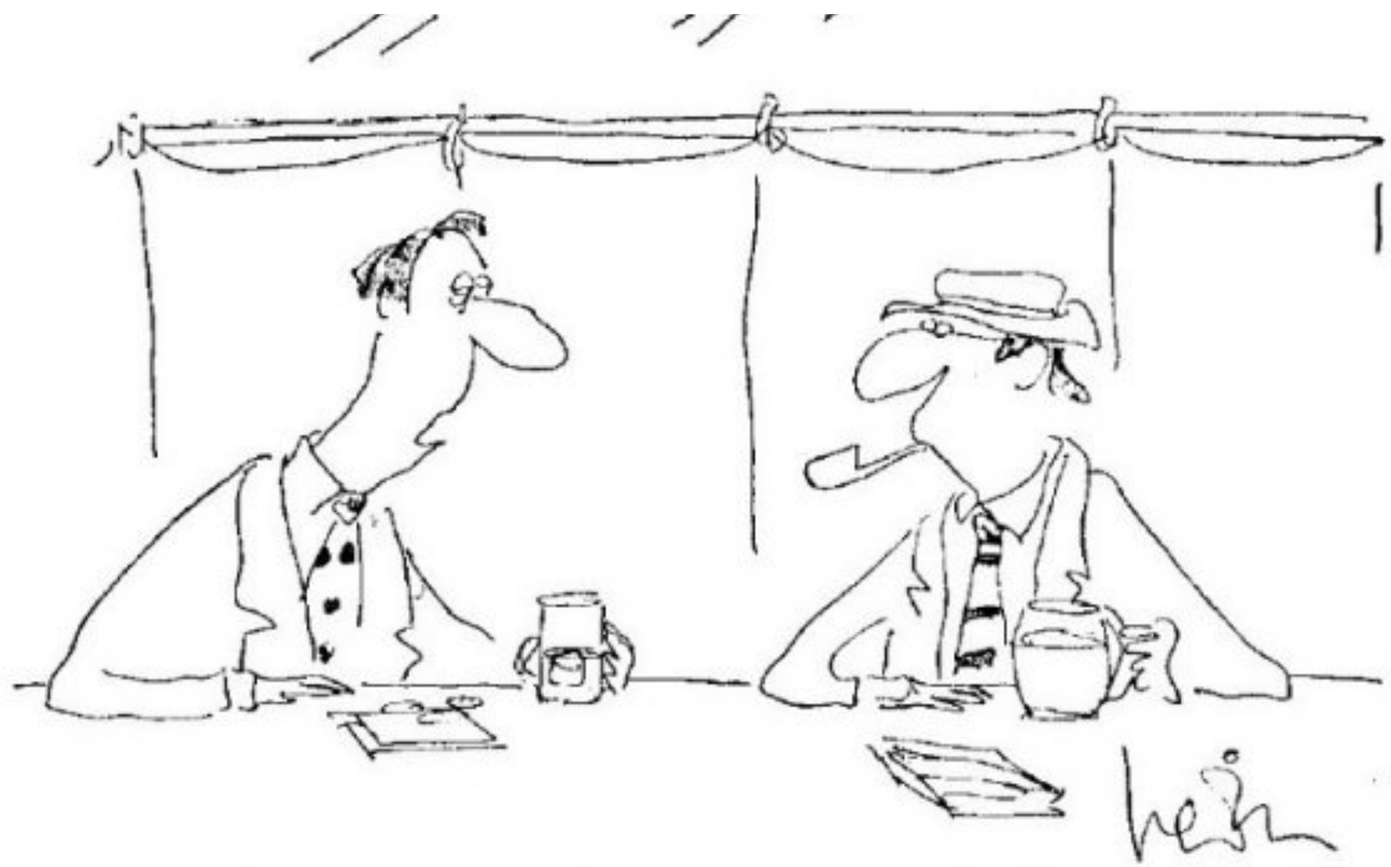


"Well, I'll be damned if I'll defend to the death your right to say something that's statistically incorrect."

Statistical Comparison of Algorithms — Part II

Leandro L. Minku
University of Birmingham, UK

Overview

- Recap of the general idea underlying statistical hypothesis tests.
- What to compare?
  - Two algorithms on a single problem instance.
  - Two algorithms on multiple problem instances.
  - Multiple algorithms on a single problem instance.
  - Multiple algorithms on multiple problem instances.
- How to design the comparisons?
  - Tests for 2 groups.
    - Observation corresponds to a single run.
    - Observation corresponds to the aggregation of multiple runs.
  - Tests for N groups.
    - Observation corresponds to a single run.
    - Observation corresponds to the aggregation of multiple runs.
- Commands to run the statistical tests.

Overview

- Recap of the general idea underlying statistical hypothesis tests.

- What to compare?
    - Two algorithms on a single problem instance.
    - Two algorithms on multiple problem instances.
    - Multiple algorithms on a single problem instance.
    - Multiple algorithms on multiple problem instances.

- How to design the comparisons?
    - Tests for 2 groups.
        - Observation corresponds to a single run.
        - Observation corresponds to the aggregation of multiple runs.
    - Tests for N groups.
        - Observation corresponds to a single run.
        - Observation corresponds to the aggregation of multiple runs.

- Commands to run the statistical tests.

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.

A.M. Mood, R.A. Graybill and D.C. Boes. Introduction to the Theory of Statistics. Third edition. Chapter 9 — Test of Hypotheses. McGraw-Hill, 1974.

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.

Runs

Performance for A1
0.6015110151
0.2947677998
0.9636589224
0.251976978
0.3701006544
0.9940754515
0.4283523627
0.1904817054
0.7377491128
0.5392380701
0.4230920852
0.7221442924
0.8882444038
0.3186565207
0.5532666035
0.8306283304
0.4488794934
0.6386464711
0.703989767
0.1133421799
0.9693252021
0.4042517894
0.6884307214
0.1627650897
0.5280297005
0.6990777731
0.020703112
0.580238106
0.5673830342
0.2294966863

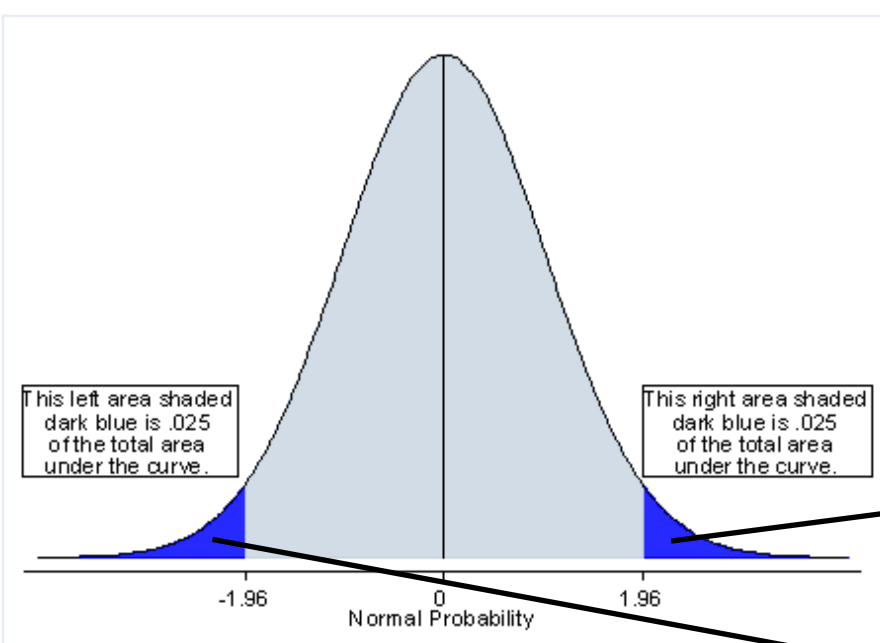
Performance for A2
0.0633347888
1.0930402922
0.1792341981
1.207096969
1.0606484322
0.6473818857
0.8043431063
0.658958582
1.0576089397
0.7364416374
0.1942901434
0.5849134532
0.4971571929
0.2973731101
0.9801976669
0.1366545414
0.258875354
1.3587444717
1.0901669778
0.5101653608
0.6768334243
1.3479477059
1.1339212937
1.154985441
1.0054153791
1.0128717172
0.5093192254
1.3938111293
0.790654944
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.

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

- Formulate Hypotheses:
  - H0: μ1 = μ2 —> μ1 - μ2 = 0
  - H1: μ1 ≠ μ2 —> μ1 - μ2 ≠ 0

- Level of significance α = 0.05 (probability of Type I error).
- Test statistic Z = $\frac{M1 - M2}{\sigma/\sqrt{N}}$
- Theoretical sampling distribution of the test statistic assuming H0 is true: normal distribution.



Probability of observing test statistic values ≤ -1.96 or ≥ 1.96 assuming that H0 is true is α = 0.05.

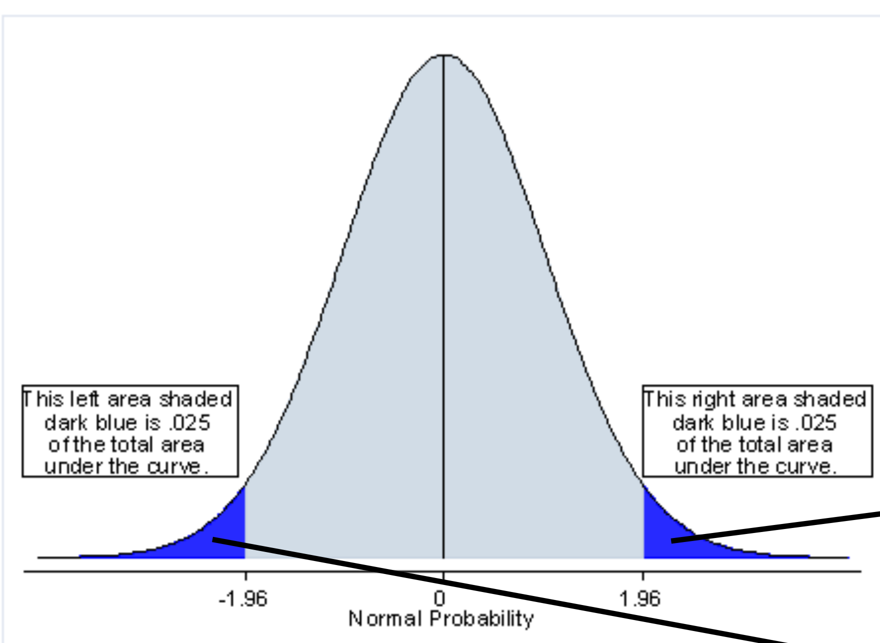
=1/2 level of significance of 0.05

=1/2 level of significance of 0.05

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

- Formulate Hypotheses:
  - H0: μ1 = μ2 —> μ1 - μ2 = 0
  - H1: μ1 ≠ μ2 —> μ1 - μ2 ≠ 0

- Level of significance α = 0.05 (probability of Type I error).
- Test statistic Z = $\frac{M1 - M2}{\sigma/\sqrt{N}}$
- Theoretical sampling distribution of the test statistic assuming H0 is true: normal distribution.



If the test statistic falls in this region, we will reject H0.

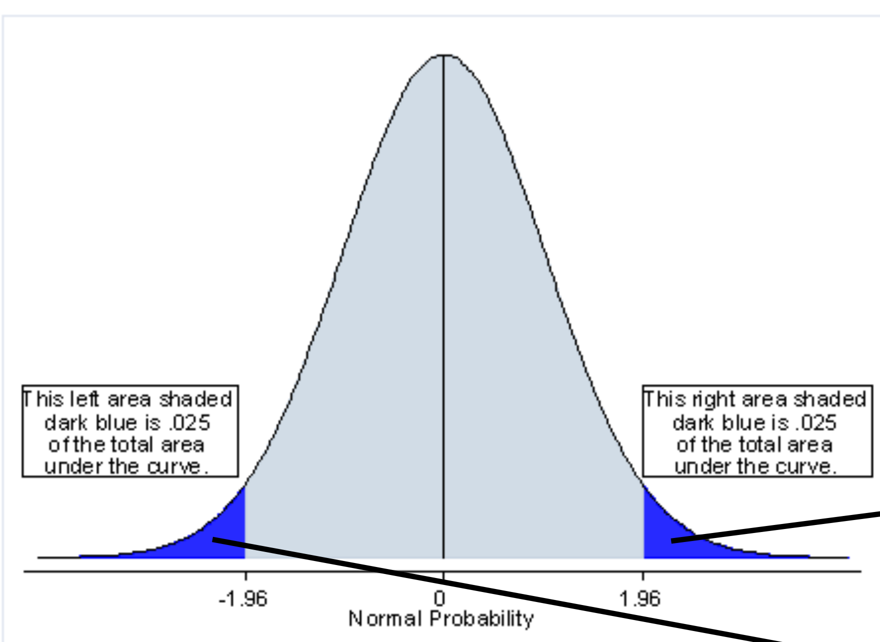
=1/2 level of significance of 0.05

=1/2 level of significance of 0.05

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

- Formulate Hypotheses:
  - H0: μ1 = μ2 —> μ1 - μ2 = 0
  - H1: μ1 ≠ μ2 —> μ1 - μ2 ≠ 0

- Level of significance α = 0.05 (probability of Type I error).
- Test statistic Z = $\frac{M1 - M2}{\sigma/\sqrt{N}}$
- Theoretical sampling distribution of the test statistic assuming H0 is true: normal distribution.



But there is still a small chance that H0 was true (Type I error).

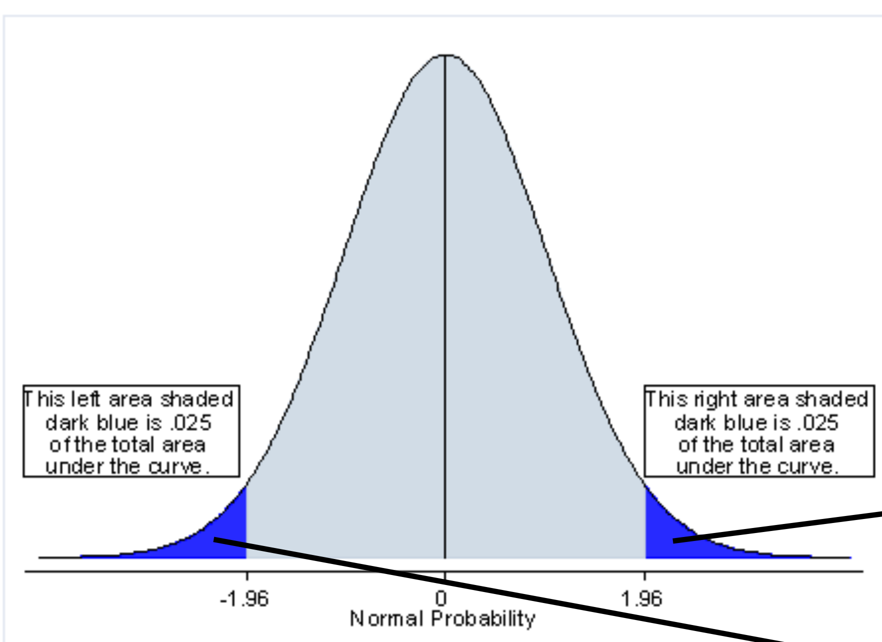
=1/2 level of significance of 0.05

=1/2 level of significance of 0.05

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

- Formulate Hypotheses:
  - H0: μ1 = μ2 —> μ1 - μ2 = 0
  - H1: μ1 ≠ μ2 —> μ1 - μ2 ≠ 0

- Level of significance α = 0.05 (probability of Type I error).
- Test statistic Z = $\frac{M1 - M2}{\sigma/\sqrt{N}}$
- Theoretical sampling distribution of the test statistic assuming H0 is true: normal distribution.



Critical region is the set of test statistic values that would lead to rejecting H0.

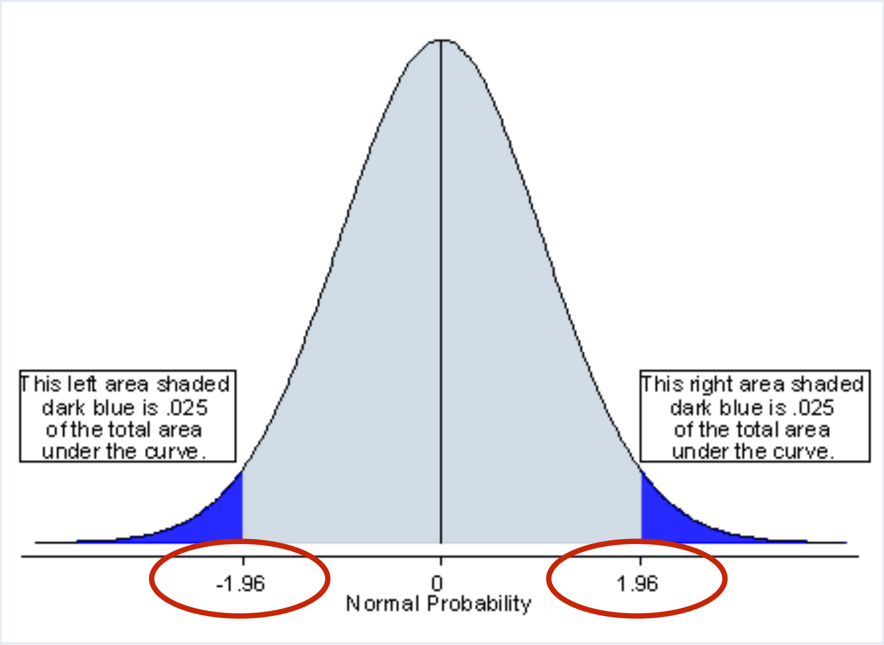
=1/2 level of significance of 0.05

=1/2 level of significance of 0.05

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

- Formulate Hypotheses:
  - H0: $\mu 1 = \mu 2 \longrightarrow \mu 1 - \mu 2 = 0$
  - H1: $\mu 1 \neq \mu 2 \longrightarrow \mu 1 - \mu 2 \neq 0$

- Level of significance $\alpha = 0.05$ (probability of Type I error).
- Test statistic Z = $\frac{M1 - M2}{\sigma/\sqrt{N}}$
- Theoretical sampling distribution of the test statistic assuming H0 is true: normal distribution.

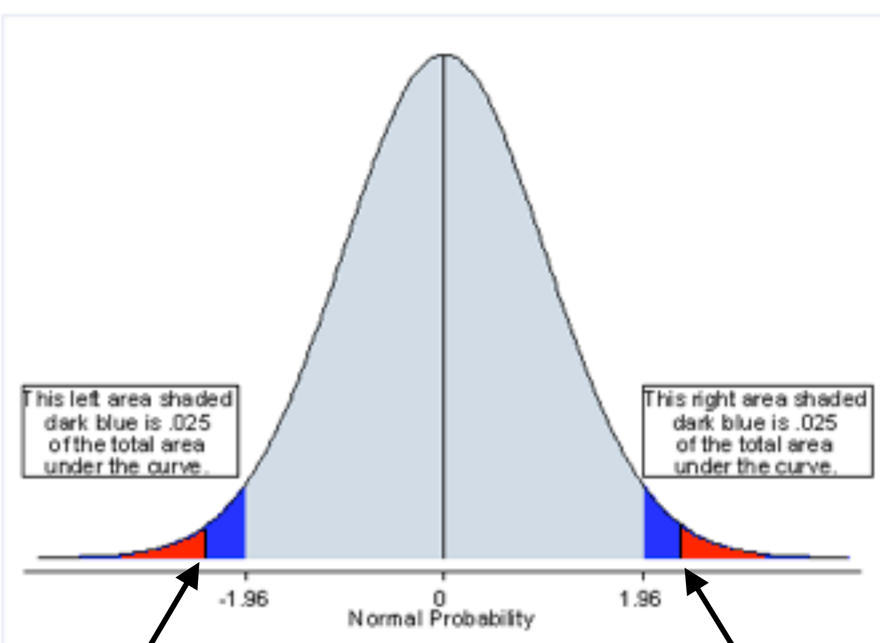


Critical values are the "boundary" values of the critical region.

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

- Formulate Hypotheses:
  - H0: $\mu1 = \mu2 \longrightarrow \mu1 - \mu2 = 0$
  - H1: $\mu1 \neq \mu2 \longrightarrow \mu1 - \mu2 \neq 0$

- Level of significance $\alpha = 0.05$ (probability of Type I error).
- Test statistic Z = $\frac{M1 - M2}{\sigma/\sqrt{N}}$
- Theoretical sampling distribution of the test statistic assuming H0 is true: normal distribution.



-Z

Z

- P-value: probability of observing test statistic value at least as extreme as the value z, assuming H0, is the AUC of the region starting at z and -z.
- If p-value ≤ α, reject H0.
- Otherwise, do not reject H0

Terminology

- For two tailed test (H0: $\mu 1 = \mu 2$, H1: $\mu 1 \neq \mu 2$):
    - Not rejecting H0: **no statistically significant difference** has been found between μ1 and μ2 at the level of significance of $\alpha = 0.05$ (p-value of …).
        - It doesn't mean that we accept H0, 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.

- Rejecting H0: **statistically significant difference** between μ1 and μ2 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.
        - μ1 is significantly larger than μ2.
        - μ1 is significantly smaller than μ2.

Choosing Statistical Tests

- Different statistical hypothesis tests use different test statistics, which make different assumptions about the population underlying the 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

Tests for comparing means of the underlying distributions.

Choosing Statistical Tests

- Different statistical hypothesis tests use different test statistics, which make different assumptions about the population underlying the 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

Tests for comparing medians of the underlying distributions.

Choosing Statistical Tests

- Different statistical hypothesis tests use different test statistics, which make different assumptions about the population underlying the 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	Tukey
	Paired (related)	Paired t-test	ANOVA	Tukey
Non-parametric (no normality)	**Unpaired (independent)**	Wilcoxon rank-sum test = Mann–Whitney U test	Kruskal-Wallis test	Dunn
	Paired (related)	Wilcoxon signed-rank test	Friedman test	Nemenyi

Overview

- Recap of the general idea underlying statistical hypothesis tests.
- What to compare?
  - Two algorithms on a single problem instance.
  - Two algorithms on multiple problem instances.
  - Multiple algorithms on a single problem instance.
  - Multiple algorithms on multiple problem instances.
- How to design the comparisons?
  - Tests for 2 groups.
    - Observation corresponds to a single run.
    - Observation corresponds to the aggregation of multiple runs.
  - Tests for N groups.
    - Observation corresponds to a single run.
    - Observation corresponds to the aggregation of multiple runs.
- Commands to run the statistical tests.

Overview

- Recap of the general idea underlying statistical hypothesis tests.
- What to compare?
    - Two algorithms on a single problem instance.
    - Two algorithms on multiple problem instances.
    - Multiple algorithms on a single problem instance.
    - Multiple algorithms on multiple problem instances.
- How to design the comparisons?
    - Tests for 2 groups.
        - Observation corresponds to a single run.
        - Observation corresponds to the aggregation of multiple runs.
    - Tests for N groups.
        - Observation corresponds to a single run.
        - Observation corresponds to the aggregation of multiple runs.
- Commands to run the statistical tests.

Runs for Comparing Two Algorithms on a Single Problem Instance

Runs

Performance for A1	Performance for A2
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

Overview

- Recap of the general idea underlying statistical hypothesis tests.
- What to compare?
  - Two algorithms on a single problem instance.
  - Two algorithms on multiple problem instances.
  - Multiple algorithms on a single problem instance.
  - Multiple algorithms on multiple problem instances.
- How to design the comparisons?
  - Tests for 2 groups.
    - Observation corresponds to a single run.
    - Observation corresponds to the aggregation of multiple runs.
  - Tests for N groups.
    - Observation corresponds to a single run.
    - Observation corresponds to the aggregation of multiple runs.
- Commands to run the statistical tests.

Comparing Two Algorithms on a Single Problem Instance Using a Test for 2 Groups

One Comparison

Runs

Performance for A1	Performance for A2
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

- An observation in a group may be, e.g.:
  - One run of the group’s EA with a given random seed.
  - One run of the group's ML algorithm with a given training / validation / testing partition.
  - One run of the group’s ML algorithm with a given random seed and training / validation / testing partition.

Which Statistical Test To Use?

Choose one of the statistical tests for two groups.

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

Overview

- Recap of the general idea underlying statistical hypothesis tests.
- What to compare?
  - Two algorithms on a single problem instance.
  - Two algorithms on multiple problem instances.
  - Multiple algorithms on a single problem instance.
  - Multiple algorithms on multiple problem instances.
- How to design the comparisons?
  - Tests for 2 groups.
    - Observation corresponds to a single run.
    - Observation corresponds to the aggregation of multiple runs.
  - Tests for N groups.
    - Observation corresponds to a single run.
    - Observation corresponds to the aggregation of multiple runs.
- Commands to run the statistical tests.

Runs for Comparing Two Algorithms on Multiple Problem Instances

Runs

Performance for A1 on Problem Instance 1	Performance for A2 on Problem Instance 1	Performance for A1 on Problem Instance 2	Performance for A2 on Problem Instance 2	Performance for A1 on Problem Instance 3	Performance for A2 on Problem Instance 3
0.6015110151	0.0633347888	0.760460255	0.6551929305	0.5476658046	0.9046872039
0.2947677998	1.0930402922	0.0572251119	0.3337481166	0.4137681613	0.9520324941
0.9636589224	0.1792341981	0.5574389137	0.0036406675	0.0806697314	0.7879171027
0.251976978	1.207096969	0.6322326728	0.178944475	0.9069706099	0.7637043188
0.3701006544	1.0606484322	0.3735014456	0.7309588448	0.1943163828	0.409963062
0.9940754515	0.6473818857	0.4563438955	0.9244792748	0.0127057396	0.8664534697
0.4283523627	0.8043431063	0.189285421	0.4301181359	0.6483924752	0.2972555845
0.1904817054	0.658958582	0.0110451456	0.2721486911	0.0711753396	0.3053791677
0.7377491128	1.0576089397	0.4170535561	0.7586322057	0.6792222569	0.2630606971
0.5392380701	0.7364416374	0.7564326315	0.0227292371	0.0306830725	0.9960538673
0.4230920852	0.1942901434	0.6220609574	0.4968550089	0.4738853995	0.2809200487
0.7221442924	0.5849134532	0.0501721525	0.5922216047	0.8292532503	0.5101169699
0.8882444038	0.4971571929	0.5578816063	0.9233305764	0.9567378471	0.3927596693
0.3186565207	0.2973731101	0.9426834162	0.6820758707	0.4673124996	0.0602585103
0.5532666035	0.9801976669	0.9013300173	0.0850999199	0.96967731	0.1907651876
0.8306283304	0.1366545414	0.6234262334	0.7930495869	0.1963517577	0.3978416505
0.4488794934	0.258875354	0.8931927863	0.8423898115	0.7760340429	0.8830631927
0.6386464711	1.3587444717	0.3288020403	0.6413379584	0.4379052422	0.9575326536
0.703989767	1.0901669778	0.6895393033	0.7447397911	0.1255642571	0.3187901091
0.1133421799	0.5101653608	0.7622498292	0.4499571978	0.6202795375	0.8254916123
0.9693252021	0.6768334243	0.0886043736	0.303599728	0.5320392225	0.8695490318
0.4042517894	1.3479477059	0.0628773789	0.1713403165	0.579999126	0.0869615532
0.6884307214	1.1339212937	0.024849294	0.2187812116	0.827169888	0.3043244402
0.1627650897	1.154985441	0.1848034125	0.3121568679	0.17672092	0.8562839972
0.5280297005	1.0054153791	0.5693529861	0.6661441082	0.8148790556	0.2333843976
0.6990777731	1.0128717172	0.6075816357	0.7424533118	0.0247170569	0.7947430999
0.020703112	0.5093192254	0.9308488478	0.8053636709	0.0813859012	0.5402830557
0.580238106	1.3938111293	0.0362369791	0.8241804624	0.9262922227	0.7284770885
0.5673830342	0.790654944	0.6035423176	0.3438211307	0.7991833945	0.2747318668
0.2294966863	1.3811101009	0.0712389681	0.5202705748	0.3406950799	0.8479146701

. . .

Overview

- Recap of the general idea underlying statistical hypothesis tests.
- What to compare?
  - Two algorithms on a single problem instance.
  - Two algorithms on multiple problem instances.
  - Multiple algorithms on a single problem instance.
  - Multiple algorithms on multiple problem instances.
- How to design the comparisons?
  - Tests for 2 groups.
    - Observation corresponds to a single run.
    - Observation corresponds to the aggregation of multiple runs.
  - Tests for N groups.
    - Observation corresponds to a single run.
    - Observation corresponds to the aggregation of multiple runs.
- Commands to run the statistical tests.

Comparing Two Algorithms on Multiple Problem Instances Using Multiple Tests for 2 Groups

First Comparison

Performance for A1 on Problem Instance 1	Performance for A2 on Problem Instance 1
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
[illegible]	[illegible]
0.2294966863	1.3811101009

Second Comparison

Performance for A1 on Problem Instance 2	Performance for A2 on Problem Instance 2
0.760460255	0.6551929305
0.0572251119	0.3337481166
0.5574389137	0.0036406675
0.6322326728	0.178944475
0.3735014456	0.7309588448
0.4563438955	0.9244792748
0.189285421	0.4301181359
0.0110451456	0.2721486911
0.4170535561	0.7586322057
[illegible]	[illegible]
0.0712389681	0.5202705748

Third Comparison

Performance for A1 on Problem Instance 3	Performance for A2 on Problem Instance 3
0.5476658046	0.9046872039
0.4137681613	0.9520324941
0.0806697314	0.7879171027
0.9069706099	0.7637043188
0.1943163828	0.409963062
0.0127057396	0.8664534697
0.6483924752	0.2972555845
0.0711753396	0.3053791677
0.6792222569	0.2630606971
[illegible]	[illegible]
0.3406950799	0.8479146701

…

- An observation in a group may be, e.g.:
  - One run of the group's EA on the group's problem instance with a given random seed.
  - One run of the group's ML algorithm on the group's dataset with a given training / validation / testing partition.
  - One run of the group's ML algorithm on the group’s dataset with a given random seed and training / validation / testing partition.

Which Statistical Test To Use?

You could potentially use one of the statistical tests for two groups and perform one test for each problem instance.

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

Comparing Two Algorithms on Multiple Problem Instances Using Multiple Tests for 2 Groups

- Advantage:
  - You know in which problem instances the algorithms performed differently and in which they didn't.
- Disadvantage:
  - Multiple comparisons lead to higher probability of at least one Type I error.
  - Requires p-values or level of significance to be corrected to avoid that (e.g., Holm-Bonferroni corrections).
    - Can in turn lead to weak tests (unlikely to detect differences).
    - Controversy in terms of how many comparisons to consider in the adjustment.

Overview

- Recap of the general idea underlying statistical hypothesis tests.
- What to compare?
  - Two algorithms on a single problem instance.
  - Two algorithms on multiple problem instances.
  - Multiple algorithms on a single problem instance.
  - Multiple algorithms on multiple problem instances.
- How to design the comparisons?
  - Tests for 2 groups.
    - Observation corresponds to a single run.
    - Observation corresponds to the aggregation of multiple runs.
  - Tests for N groups.
    - Observation corresponds to a single run.
    - Observation corresponds to the aggregation of multiple runs.
- Commands to run the statistical tests.

Comparing Two Algorithms on Multiple Problem Instance Using a Single Test for 2 Groups Consisting of Aggregated Runs

Runs

Performance for A1 on Problem Instance 1
0.6015110151
0.2947677998
0.9636589224
0.251976978
0.3701006544
0.9940754515
0.4283523627
0.1904817054
0.7377491128
0.5392380701
0.4230920852
0.7221442924
0.8882444038
0.3186565207
0.5532666035
0.8306283304
0.4488794934
0.6386464711
0.703989767
0.1133421799
0.9693252021
0.4042517894
0.6884307214
0.1627650897
0.5280297005
0.6990777731
0.020703112
0.580238106
0.5673830342
0.2294966863

Average Performance of A1 on Problem Instance 1

Problem Instance

Performance for A1	Performance for A2
0.5287501145	0.7941165821
0.4587431195	0.5156587096
0.4847217528	0.5633566591
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

J. Demsar. Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research 7 (2006) 1–30.

Comparing Two Algorithms on Multiple Problem Instance Using a Single Test for 2 Groups Consisting of Aggregated Runs

Runs

Performance for A1 on Problem Instance 2
0.6015110151
0.2947677998
0.9636589224
0.251976978
0.3701006544
0.9940754515
0.4283523627
0.1904817054
0.7377491128
0.5392380701
0.4230920852
0.7221442924
0.8882444038
0.3186565207
0.5532666035
0.8306283304
0.4488794934
0.6386464711
0.703989767
0.1133421799
0.9693252021
0.4042517894
0.6884307214
0.1627650897
0.5280297005
0.6990777731
0.020703112
0.580238106
0.5673830342
0.2294966863

Average Performance of A1 on Problem Instance 2

Problem Instance

Performance for A1	Performance for A2
0.5287501145	0.7941165821
0.4587431195	0.5156587096
0.4847217528	0.5633566591
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

Comparing Two Algorithms on Multiple Problem Instance Using a Single Test for 2 Groups Consisting of Aggregated Runs

Runs

Performance for A2 on Problem Instance 1
0.6015110151
0.2947677998
0.9636589224
0.251976978
0.3701006544
0.9940754515
0.4283523627
0.1904817054
0.7377491128
0.5392380701
0.4230920852
0.7221442924
0.8882444038
0.3186565207
0.5532666035
0.8306283304
0.4488794934
0.6386464711
0.703989767
0.1133421799
0.9693252021
0.4042517894
0.6884307214
0.1627650897
0.5280297005
0.6990777731
0.020703112
0.580238106
0.5673830342
0.2294966863

Average Performance of A2 on Problem Instance 1

Problem Instance

Performance for A1	Performance for A2
0.5287501145	0.7941165821
0.4587431195	0.5156587096
0.4847217528	0.5633566591
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

Comparing Two Algorithms on Multiple Problem Instance Using a Single Test for 2 Groups Consisting of Aggregated Runs

One Comparison

Problem Instance

Performance for A1	Performance for A2
0.5287501145	0.7941165821
0.4587431195	0.5156587096
0.4847217528	0.5633566591
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

- An observation in a group may be, e.g.:
  - The average of multiple runs of the group's EA on a given problem instance.
    - The multiple runs are performed by varying the EA's random seed.
  - The average of multiple runs of the group's ML algorithm on a given dataset.
    - The multiple runs are performed by varying the ML algorithm's random seed and/or training/validation/test sample.

Which Statistical Test To Use?

You could potentially use one of the statistical tests for two paired groups, most likely Wilcoxon Signed-Rank Test.

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

J. Demsar. Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research 7 (2006) 1–30.

Comparing Two Algorithms on Multiple Problem Instance Using a Single Test for 2 Groups Consisting of Aggregated Runs

- Advantages:
  - No issue with multiple comparisons.
- Disadvantages:
  - The test can still be weak if the number of problem instances (i.e., observations) is too small.
  - Ignores variability across runs — use only the combined (e.g., average) result for each set of runs.
  - When the two algorithms are not significantly different across problem instances, it does not mean that the two algorithms perform similarly on each individual problem instance.
    - It could be that one algorithm is better for some problem instances, and worse for others. So, overall, there is no winner **across** problem instances.

Potential Solution to Mitigate Lack of Insights When The Algorithms Are Not Significantly Different Across Datasets: Effect Size

- Use measures of effect size for each problem instance separately.
- E.g.: non-parametric A12 effect size.
  - Represents the probability that running a given algorithm A1 yields better results than A2.
  - Big is |A12| >= 0.71
  - Medium is |A12| >=0.64
  - Small is |A12| >= 0.56
  - Insignificant is |A12| < 0.56

Problem Instance

Performance for A1	Performance for A2	Effect Size
0.5287501145	0.7941165821	0.3
0.4587431195	0.5156587096	-0.7
0.4847217528	0.5633566591	-0.4
0.251976978	1.207096969	0.8
0.3701006544	1.0606484322	0.25
0.9940754515	0.6473818857	-0.4
0.4283523627	0.8043431063	-0.9
0.1904817054	0.658958582	0.7
0.7377491128	1.0576089397	0.78
0.5392380701	0.7364416374	-0.3
0.4230920852	0.1942901434	-0.22
0.7221442924	0.5849134532	0.12
0.8882444038	0.4971571929	0.4

András Vargha and Harold D. Delaney. A Critique and Improvement of the "CL" Common Language Effect Size Statistics of McGraw and WongAuthor. Journal of Educational and Behavioral Statistics, Vol. 25, No. 2 (2000), pp.101-132

Effect Size

- Advantages:
  - Not affected by the number of runs.
  - Avoid multiple comparison issue of statistical tests.
  - Gives an idea of the size of the effect of the difference in performance.
- Disadvantages:
  - Completely ignores the number of runs.
    - Could have large effect sizes even if the experiment was based on very few runs.
    - So, it’s recommended to be used together with statistical tests, following a rejection of H0.

Overview

- Recap of the general idea underlying statistical hypothesis tests.
- What to compare?
  - Two algorithms on a single problem instance.
  - Two algorithms on multiple problem instances.
  - Multiple algorithms on a single problem instance.
  - Multiple algorithms on multiple problem instances.
- How to design the comparisons?
  - Tests for 2 groups.
    - Observation corresponds to a single run.
    - Observation corresponds to the aggregation of multiple runs.
  - Tests for N groups.
    - Observation corresponds to a single run.
    - Observation corresponds to the aggregation of multiple runs.
- Commands to run the statistical tests.

Runs for Comparing Multiple Algorithms On a Single Problem Instance

Runs

Performance for A1	Performance for A2	Performance for A3
0.6015110151	0.0633347888	0.7725776185
0.2947677998	1.0930402922	0.6037878711
0.9636589224	0.1792341981	0.2000145838
0.251976978	1.207096969	0.1124429684
0.3701006544	1.0606484322	0.0765464923
0.9940754515	0.6473818857	0.9356262246
0.4283523627	0.8043431063	0.893382197
0.1904817054	0.658958582	0.3686623329
0.7377491128	1.0576089397	0.0552056497
0.5392380701	0.7364416374	0.6485590856
0.4230920852	0.1942901434	0.686919529
0.7221442924	0.5849134532	0.956750494
0.8882444038	0.4971571929	0.8807609468
0.3186565207	0.2973731101	0.2476675087
0.5532666035	0.9801976669	0.3168956009
0.8306283304	0.1366545414	0.7664107613
0.4488794934	0.258875354	0.1607483861
0.6386464711	1.3587444717	0.1702079105
0.703989767	1.0901669778	0.1151715671
0.1133421799	0.5101653608	0.5060234619
0.9693252021	0.6768334243	0.6248869323
0.4042517894	1.3479477059	0.4384962961
0.6884307214	1.1339212937	0.8133689603
0.1627650897	1.154985441	0.0685902033
0.5280297005	1.0054153791	0.9532216617
0.6990777731	1.0128717172	0.7946400358
0.020703112	0.5093192254	0.1304510306
0.580238106	1.3938111293	0.3950510006
0.5673830342	0.790654944	0.6486004062
0.2294966863	1.3811101009	0.5494810601

. . .

Overview

- Recap of the general idea underlying statistical hypothesis tests.
- What to compare?
  - Two algorithms on a single problem instance.
  - Two algorithms on multiple problem instances.
  - Multiple algorithms on a single problem instance.
  - Multiple algorithms on multiple problem instances.
- How to design the comparisons?
  - Tests for 2 groups.
    - Observation corresponds to a single run.
    - Observation corresponds to the aggregation of multiple runs.
  - Tests for N groups.
    - Observation corresponds to a single run.
    - Observation corresponds to the aggregation of multiple runs.
- Commands to run the statistical tests.

Comparing Multiple Algorithms On a Single Problem Instance Using Multiple Tests for 2 Groups

Runs

First Comparison

Performance for A1	Performance for A2
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

Second Comparison

Performance for A1	Performance for A3
0.6015110151	0.7725776185
0.2947677998	0.6037878711
0.9636589224	0.2000145838
0.251976978	0.1124429684
0.3701006544	0.0765464923
0.9940754515	0.9356262246
0.4283523627	0.893382197
0.1904817054	0.3686623329
0.7377491128	0.0552056497
0.5392380701	0.6485590856
0.4230920852	0.686919529
0.7221442924	0.956750494
0.8882444038	0.8807609468
0.3186565207	0.2476675087
0.5532666035	0.3168956009
0.8306283304	0.7664107613

Third Comparison

Performance for A2	Performance for A3
0.0633347888	0.7725776185
1.0930402922	0.6037878711
0.1792341981	0.2000145838
1.207096969	0.1124429684
1.0606484322	0.0765464923
0.6473818857	0.9356262246
0.8043431063	0.893382197
0.658958582	0.3686623329
1.0576089397	0.0552056497
0.7364416374	0.6485590856
0.1942901434	0.686919529
0.5849134532	0.956750494
0.4971571929	0.8807609468
0.2973731101	0.2476675087
0.9801976669	0.3168956009
0.1366545414	0.7664107613

. . .

- An observation in a group may be, e.g.:
    - One run of the group's EA on the problem instance with a given random seed.
    - One run of the group's ML algorithm on the dataset with a given training / validation / testing partition.
    - One run of the group's ML algorithm on the dataset with a given random seed and training / validation / testing partition.

Which Statistical Test To Use?

You could potentially use one of the statistical tests for two groups and perform one test for each problem instance.

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

Comparing Multiple Algorithms On a Single Problem Instance Using Multiple Tests for 2 Groups

- Advantages and disadvantages
  - Similar to those of the pairwise comparisons of two algorithms on multiple problem instances.

Overview

- Recap of the general idea underlying statistical hypothesis tests.
- What to compare?
  - Two algorithms on a single problem instance.
  - Two algorithms on multiple problem instances.
  - Multiple algorithms on a single problem instance.
  - Multiple algorithms on multiple problem instances.
- How to design the comparisons?
  - Tests for 2 groups.
    - Observation corresponds to a single run.
    - Observation corresponds to the aggregation of multiple runs.
  - Tests for N groups.
    - Observation corresponds to a single run.
    - Observation corresponds to the aggregation of multiple runs.
- Commands to run the statistical tests.

Compare Multiple Algorithms On a Single Problem Instance Using a Test for N Groups

One Comparison

Runs

Performance for A1	Performance for A2	Performance for A3
0.6015110151	0.0633347888	0.7725776185
0.2947677998	1.0930402922	0.6037878711
0.9636589224	0.1792341981	0.2000145838
0.251976978	1.207096969	0.1124429684
0.3701006544	1.0606484322	0.0765464923
0.9940754515	0.6473818857	0.9356262246
0.4283523627	0.8043431063	0.893382197
0.1904817054	0.658958582	0.3686623329
0.7377491128	1.0576089397	0.0552056497
0.5392380701	0.7364416374	0.6485590856
0.4230920852	0.1942901434	0.686919529
0.7221442924	0.5849134532	0.956750494
0.8882444038	0.4971571929	0.8807609468
0.3186565207	0.2973731101	0.2476675087
0.5532666035	0.9801976669	0.3168956009
0.8306283304	0.1366545414	0.7664107613
0.4488794934	0.258875354	0.1607483861
0.6386464711	1.3587444717	0.1702079105
0.703989767	1.0901669778	0.1151715671
0.1133421799	0.5101653608	0.5060234619
0.9693252021	0.6768334243	0.6248869323
0.4042517894	1.3479477059	0.4384962961
0.6884307214	1.1339212937	0.8133689603
0.1627650897	1.154985441	0.0685902033
0.5280297005	1.0054153791	0.9532216617
0.6990777731	1.0128717172	0.7946400358
0.020703112	0.5093192254	0.1304510306
0.580238106	1.3938111293	0.3950510006
0.5673830342	0.790654944	0.6486004062
0.2294966863	1.3811101009	0.5494810601

- An observation in a group may be, e.g.:
  - One run of the group's EA on the problem instance with a given random seed.
  - One run of the group's ML algorithm on the dataset with a given training / validation / testing partition.
  - One run of the group's ML algorithm on the dataset with a given random seed and training / validation / testing partition.

Which Statistical Test To Use?

You could potentially use one of the statistical tests for N groups. Kruskal-Wallis and Friedman are non-parametric, but ANOVA enables comparison of multiple factors and their interactions.

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

Compare Multiple Algorithms On a Single Problem Instance Using a Test for N Groups

- Advantage:
  - More powerful.
- Disadvantages:
  - Doesn't tell which pair is different.
  - Relies on post-hoc tests for determining which pair is different.
    - Post-hoc tests are weaker.

ANOVA - Analysis of Variance

- Enables to analyse the impact of multiple factors and their interactions.

- Examples of factors:
  - Algorithms.
  - Each parameter of an algorithm.
  - Datasets given as inputs to algorithms.
  - Initial condition of an algorithm (when dealing with paired data).
  - …

- Each factor can have multiple levels.

- Each factor level and each combination of factors with their levels is a group.

Example of Factors and Corresponding Groups

- parameter β with levels β1, β2, β3
- parameter α with levels α1,α2.

Performance β1,α1	Performance β2,α1	Performance β3,α1
0.2427365435	0.8207683226	0.0068735215
0.2838782503	0.6193219672	0.7603308253
0.4728852466	0.718615256	0.991473224
0.1602043263	0.6568282119	0.9653211501
0.3113725667	0.003657249	0.9002240284
0.7092353466	0.9641411756	0.9044996039
0.1243187189	0.1916947681	0.5001887854
0.9923597255	0.4643217917	0.4644260767
0.1593878649	0.7075588114	0.6043046496
0.7137943972	0.7178264102	0.3684897267
0.4405825825	0.9738042639	0.6371247198
0.0546034079	0.1643357663	0.8491521557
0.130989165	0.8930972954	0.9200755227
0.4630962713	0.7359805298	0.7894468571
0.3653479179	0.0494488408	0.4480319903

Performance β1,α2	Performance β2,α2	Performance β3,α2
0.6513221103	0.7155298328	0.5250285096
0.4486536453	0.4544934118	0.2665807758
0.923068983	0.4432370842	0.0714614966
0.2180154489	0.8604004404	0.2213692251
0.871509453	0.4888057283	0.9734517445
0.5255328568	0.120754892	0.8236567329
0.7085732815	0.5772912123	0.7173770375
0.869020659	0.8938754508	0.6566561072
0.674964929	0.0196329623	0.5775361005
0.1924289421	0.5813982673	0.0571435841
0.3358277807	0.1917446121	0.0112761131
0.7760143983	0.2131797303	0.4513054562
0.5871792892	0.9556053877	0.1188456733
0.2420052565	0.50039103	0.7654434184
0.9896802846	0.1324466465	0.6181376898

Performance β1	Performance β2	Performance β3
0.2427365435	0.8207683226	0.0068735215
0.2838782503	0.6193219672	0.7603308253
0.4728852466	0.718615256	0.991473224
0.1602043263	0.6568282119	0.9653211501
0.3113725667	0.003657249	0.9002240284
0.7092353466	0.9641411756	0.9044996039
0.1243187189	0.1916947681	0.5001887854
0.9923597255	0.4643217917	0.4644260767
0.1593878649	0.7075588114	0.6043046496
0.7137943972	0.7178264102	0.3684897267
0.4405825825	0.9738042639	0.6371247198
0.0546034079	0.1643357663	0.8491521557
0.130989165	0.8930972954	0.9200755227
0.4630962713	0.7359805298	0.7894468571
0.3653479179	0.0494488408	0.4480319903
0.6513221103	0.7155298328	0.5250285096
0.4486536453	0.4544934118	0.2665807758
0.923068983	0.4432370842	0.0714614966
0.2180154489	0.8604004404	0.2213692251
0.871509453	0.4888057283	0.9734517445
0.5255328568	0.120754892	0.8236567329
0.7085732815	0.5772912123	0.7173770375
0.869020659	0.8938754508	0.6566561072
0.674964929	0.0196329623	0.5775361005
0.1924289421	0.5813982673	0.0571435841
0.3358277807	0.1917446121	0.0112761131
0.7760143983	0.2131797303	0.4513054562
0.5871792892	0.9556053877	0.1188456733
0.2420052565	0.50039103	0.7654434184
0.9896802846	0.1324466465	0.6181376898

Performance α1	Performance α2
0.2427365435	0.6513221103
0.2838782503	0.4486536453
0.4728852466	0.923068983
0.1602043263	0.2180154489
0.3113725667	0.871509453
0.7092353466	0.5255328568
0.1243187189	0.7085732815
0.9923597255	0.869020659
0.1593878649	0.674964929
0.7137943972	0.1924289421
0.4405825825	0.3358277807
0.0546034079	0.7760143983
0.130989165	0.5871792892
0.4630962713	0.2420052565
0.3653479179	0.9896802846
0.8207683226	0.7155298328
0.6193219672	0.4544934118
0.718615256	0.4432370842
0.6568282119	0.8604004404
0.003657249	0.4888057283
0.9641411756	0.120754892
0.1916947681	0.5772912123
0.4643217917	0.8938754508
0.7075588114	0.0196329623
0.7178264102	0.5813982673
0.9738042639	0.1917446121
0.1643357663	0.2131797303
0.8930972954	0.9556053877
0.7359805298	0.50039103
0.0494488408	0.1324466465
0.0068735215	0.5250285096
0.7603308253	0.2665807758
0.991473224	0.0714614966
0.9653211501	0.2213692251
0.9002240284	0.9734517445
0.9044996039	0.8236567329
0.5001887854	0.7173770375
0.4644260767	0.6566561072
0.6043046496	0.5775361005
0.3684897267	0.0571435841
0.6371247198	0.0112761131
0.8491521557	0.4513054562
0.9200755227	0.1188456733
0.7894468571	0.7654434184
0.4480319903	0.6181376898

ANOVA - Analysis of Variance

- Assumptions:
  - Normality*.
  - Equal variances (Levene test, F-test)*.
  - Independence of observations (in each group and between groups).
  - Possibly several others, depending on the type of ANOVA.

* violation to this may not be a big problem if equal no. observations are used for each group: http://vassarstats.net/textbook/ (chapter 14, part 1)

**Sensitivity to violations of sphericity: Gueorguieva; Krystal (2004). "Move Over ANOVA". Arch Gen Psychiatry 61: 310–317. doi:10.1001/archpsyc.61.3.310

ANOVA for Unpaired and Paired Comparisons

Source: www.design.kyushu-u.ac.jp/~takagi

unpaired

one-factor (one-way) ANOVA

column factor

group A	group B	group C
4.23	2.51	3.04
3.21	3.3	2.89
3.63	3.75	3.55
4.42	3.22	4.39
4.08	3.99	3.86
3.98	3.65	3.5
3.75	2.62	3.6
3.22	2.93	3.21

We cannot say that three groups are significantly different. (p=0.089)

paired

two-factor (two-way) ANOVA

column factor

sample factor

initial condition	group A	group B	group C
#1	4.23	2.51	3.04
#2	3.21	3.3	2.89
#3	3.63	3.75	3.55
#4	4.42	3.22	4.39
#5	4.08	3.99	3.86
#6	3.98	3.65	3.5
#7	3.75	2.62	3.6
#8	3.22	2.93	3.21

There are significant difference *somewhere* among three groups. (p<0.05)

Within vs Between Subject Factors

The type of ANOVA to be used will also depend on whether factors are within- or between-subject.

Between-subjects factor in medicine:

Consider a study of the treatment of a certain disease using drugs D1 and D2.

Factor: drug. Levels: D1, D2.

Contaminated persons (subjects) in group 1 were examined after being given drug D1, whereas other contaminated persons in group 2 were examined after being given drug D2.

We had to change subjects to vary the factor level.

Within vs Between Subject Factors

The type of ANOVA to be used will also depend on whether factors are within- or between-subject.

Within-subjects factor in medicine:

Consider a study of the treatment of a certain disease using different doses of a drug (dose D1 and D2).

Factor: drug dose. Levels: D1, D2.

Each contaminated person (subject) was examined twice, once after using dose D1 and once after using dose D2.

Different levels were investigated using the same subjects.

If different subjects were paired in some way, you may have to consider the factor as within-subject!

Within vs Between Subject Factors

In computational intelligence:

- If you are testing a neural network approach and you have to vary the dataset in order to vary the level of a factor, this factor is likely to be a between-subjects factor.
- Similar for an evolutionary algorithm and problem instances.
- Most other cases would be within-subject factors (?)

ANOVA

- One-way ANOVA:
  - 1-factor (1-way).
  - between-subjects.
- Repeated measures ANOVA:
  - 1-factor (1-way).
  - within-subjects.
  - Assumption of sphericity is important when factors have more than 2 levels*: variances of the differences between all possible pairs of groups are equal.
    (Check with Mauchly test, use Greenhouse-Geisser corrections if violated).
- **Factorial ANOVA:**
  - 2- or 3-factors (2- or 3- way) (more factors are allowed, but difficult to interpret).
  - allows to analyse interactions among factors.
  - between-subjects.
- **Multi-factor (multi-way) repeated measures ANOVA:**
  - Similar to repeated measures, but allow multiple factors.
  - If you choose GLM -> Repeated Measures in SPSS
- **Split-plot ANOVA:**
  - 2- or 3-factors (2- or 3- way) (more factors are allowed, but difficult to interpret).
  - allows to analyse interactions among factors.
  - both between and within-subjects are present.
  - Sphericity assumption*.
  - If you choose GLM -> Repeated Measures in SPSS, you can use a split-plot design.

*Sensitivity to violations of sphericity: Gueorguieva; Krystal (2004). "Move Over ANOVA". Arch Gen Psychiatry 61: 310–317. doi:10.1001/archpsyc.61.3.310

ANOVA

- Be careful with the possibility of people using different terminologies.
- Before using an ANOVA, double check what is said about its robustness to assumptions and possible corrections to violations.

Overview

- Recap of the general idea underlying statistical hypothesis tests.
- What to compare?
  - Two algorithms on a single problem instance.
  - Two algorithms on multiple problem instances.
  - Multiple algorithms on a single problem instance.
  - Multiple algorithms on multiple problem instances.
- How to design the comparisons?
  - Tests for 2 groups.
    - Observation corresponds to a single run.
    - Observation corresponds to the aggregation of multiple runs.
  - Tests for N groups.
    - Observation corresponds to a single run.
    - Observation corresponds to the aggregation of multiple runs.
- Commands to run the statistical tests.

Runs for Comparing Multiple Algorithms On Multiple Problem Instances

Performance A1,P1	Performance A2,P1	Performance A3,P1
0.2427365435	0.8207683226	0.0068735215
0.2838782503	0.6193219672	0.7603308253
0.4728852466	0.718615256	0.991473224
0.1602043263	0.6568282119	0.9653211501
0.3113725667	0.003657249	0.9002240284
0.7092353466	0.9641411756	0.9044996039
0.1243187189	0.1916947681	0.5001887854
0.9923597255	0.4643217917	0.4644260767
0.1593878649	0.7075588114	0.6043046496
0.7137943972	0.7178264102	0.3684897267
0.4405825825	0.9738042639	0.6371247198
0.0546034079	0.1643357663	0.8491521557
0.130989165	0.8930972954	0.9200755227
0.4630962713	0.7359805298	0.7894468571
0.3653479179	0.0494488408	0.4480319903

Performance A1,P2	Performance A2,P2	Performance A3,P2
0.6513221103	0.7155298328	0.5250285096
0.4486536453	0.4544934118	0.2665807758
0.923068983	0.4432370842	0.0714614966
0.2180154489	0.8604004404	0.2213692251
0.871509453	0.4888057283	0.9734517445
0.5255328568	0.120754892	0.8236567329
0.7085732815	0.5772912123	0.7173770375
0.869020659	0.8938754508	0.6566561072
0.674964929	0.0196329623	0.5775361005
0.1924289421	0.5813982673	0.0571435841
0.3358277807	0.1917446121	0.0112761131
0.7760143983	0.2131797303	0.4513054562
0.5871792892	0.9556053877	0.1188456733
0.2420052565	0.50039103	0.7654434184
0.9896802846	0.1324466465	0.6181376898

. . .

Performance A1,P3	Performance A2,P3	Performance A3,P3
0.3416006903	0.4970160131	0.4718756455
0.7381210078	0.3584098418	0.8155352564
0.1071763751	0.7864971575	0.7240501319
0.8742274034	0.1541535386	0.9032038082
0.7084579663	0.508243141	0.7062380635
0.1219630796	0.8280537131	0.9749030762
0.2974400269	0.3944554154	0.6101680766
0.729700828	0.8581229621	0.0641535632
0.7470682827	0.9125746179	0.0460176817
0.1673516291	0.0554353041	0.1263241582
0.3971516509	0.7514405253	0.4142972319
0.8030160547	0.0083224922	0.3836179054
0.6470250029	0.8022686257	0.8601624586
0.4209855006	0.442395957	0.5539153072
0.8114558498	0.1537486115	0.9410634711

. . .

Overview

- Recap of the general idea underlying statistical hypothesis tests.
- What to compare?
  - Two algorithms on a single problem instance.
  - Two algorithms on multiple problem instances.
  - Multiple algorithms on a single problem instance.
  - Multiple algorithms on multiple problem instances.
- How to design the comparisons?
  - Tests for 2 groups.
    - Observation corresponds to a single run.
    - Observation corresponds to the aggregation of multiple runs.
  - Tests for N groups.
    - Observation corresponds to a single run.
    - Observation corresponds to the aggregation of multiple runs.
- Commands to run the statistical tests.

Comparing Multiple Algorithms On Multiple Problem Instances Using Multiple Tests for 2 Groups

1st comparison

A1,P1	A2,P1
0.2427365	0.8207683
0.2838782	0.6193219
0.4728852	0.7186152
0.1602043	0.6568282
0.3113725	0.0036572
0.7092353	0.9641411
0.1243187	0.1916947
0.9923597	0.4643217
0.1593878	0.7075588
0.7137943	0.7178264
0.4405825	0.9738042
0.0546034	0.1643357
0.1309891	0.8930972
0.4630962	0.7359805
0.3653479	0.0494488

2nd comparison

A1,P1	A3,P1
0.2427365	0.0068735
0.2838782	0.7603308
0.4728852	0.9914732
0.1602043	0.9653211
0.3113725	0.9002240
0.7092353	0.9044996
0.1243187	0.5001887
0.9923597	0.4644260
0.1593878	0.6043046
0.7137943	0.3684897
0.4405825	0.6371247
0.0546034	0.8491521
0.1309891	0.9200755
0.4630962	0.7894468
0.3653479	0.4480319

3rd comparison

A2,P1	A3,P1
0.8207683	0.0068735
0.6193219	0.7603308
0.7186152	0.9914732
0.6568282	0.9653211
0.0036572	0.9002240
0.9641411	0.9044996
0.1916947	0.5001887
0.4643217	0.4644260
0.7075588	0.6043046
0.7178264	0.3684897
0.9738042	0.6371247
0.1643357	0.8491521
0.8930972	0.9200755
0.7359805	0.7894468
0.0494488	0.4480319

4th comparison

A1,P2	A2,P2
0.65132211	0.71552983
0.4486536	0.45449341
0.9230689	0.44323708
0.2180154	0.86040044
0.8715094	0.48880572
0.5255328	0.12075489
0.7085732	0.57729121
0.8690206	0.89387545
0.6749649	0.01963296
0.1924289	0.58139826
0.3358277	0.19174461
0.7760143	0.21317973
0.5871792	0.95560538
0.2420052	0.50039103
0.9896802	0.13244664

5th comparison

A1,P2	A3,P2
0.65132211	0.5250285
0.4486536	0.2665807
0.9230689	0.0714614
0.2180154	0.2213692
0.8715094	0.9734517
0.5255328	0.8236567
0.7085732	0.7173770
0.8690206	0.6566561
0.6749649	0.5775361
0.1924289	0.0571435
0.3358277	0.01127611
0.7760143	0.4513054
0.5871792	0.1188456
0.2420052	0.76544341
0.9896802	0.6181376

6th comparison

A2,P2	A3,P2
0.71552983	0.5250285
0.45449341	0.2665807
0.44323708	0.0714614
0.86040044	0.2213692
0.48880572	0.9734517
0.12075489	0.8236567
0.57729121	0.7173770
0.89387545	0.6566561
0.01963296	0.5775361
0.58139826	0.0571435
0.19174461	0.01127611
0.21317973	0.4513054
0.95560538	0.1188456
0.50039103	0.76544341
0.13244664	0.6181376

…

7th comparison

A1,P3	A2,P3
0.3416006	0.49701601
0.7381210	0.35840984
0.10717637	0.7864971
0.8742274	0.1541535
0.7084579	0.5082431
0.12196307	0.82805371
0.29744002	0.39445541
0.72970082	0.8581229
0.7470682	0.91257461
0.16735162	0.0554353
0.39715165	0.75144052
0.80301605	0.0083224
0.6470250	0.80226862
0.4209855	0.4423959
0.81145584	0.15374861

8th comparison

A1,P3	A3,P3
0.3416006	0.47187564
0.7381210	0.81553525
0.10717637	0.72405013
0.8742274	0.90320380
0.7084579	0.70623806
0.12196307	0.97490307
0.29744002	0.61016807
0.72970082	0.06415356
0.7470682	0.04601768
0.16735162	0.12632415
0.39715165	0.41429723
0.80301605	0.38361790
0.6470250	0.86016245
0.4209855	0.55391530
0.81145584	0.94106347

9th comparison

A2,P3	A3,P3
0.49701601	0.47187564
0.35840984	0.81553525
0.7864971	0.72405013
0.1541535	0.90320380
0.5082431	0.70623806
0.82805371	0.97490307
0.39445541	0.61016807
0.8581229	0.06415356
0.91257461	0.04601768
0.0554353	0.12632415
0.75144052	0.41429723
0.0083224	0.38361790
0.80226862	0.86016245
0.4423959	0.55391530
0.15374861	0.94106347

…

- An observation in a group may be, e.g.:
  - One run of the group's EA on the group's problem instance with a given random seed.
  - One run of the group's ML algorithm on the group's dataset with a given training / validation / testing partition.
  - One run of the group's ML algorithm on the group's dataset with a given random seed and training / validation / testing partition.

Which Statistical Test To Use?

You could potentially use one of the statistical tests for N groups. Kruskal-Wallis and Friedman are non-parametric, but ANOVA enables comparison of multiple factors and their interactions.

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

Comparing Multiple Algorithms On Multiple Problem Instances Using Multiple Tests for 2 Groups

- Advantages and disadvantages similar to:
  - comparison of two algorithms over multiple problem instances based on pairwise comparisons and
  - comparison of multiple algorithms over a single problem instance based on pairwise comparisons.

Overview

- Recap of the general idea underlying statistical hypothesis tests.
- What to compare?
  - Two algorithms on a single problem instance.
  - Two algorithms on multiple problem instances.
  - Multiple algorithms on a single problem instance.
  - Multiple algorithms on multiple problem instances.
- How to design the comparisons?
  - Tests for 2 groups.
    - Observation corresponds to a single run.
    - Observation corresponds to the aggregation of multiple runs.
  - Tests for N groups.
    - Observation corresponds to a single run.
    - Observation corresponds to the aggregation of multiple runs.
- Commands to run the statistical tests.

Example of Factors and Corresponding Groups

- parameter β with levels β1, β2, β3
- parameter P with levels P1,P2.

Performance β1,P1	Performance β2,P1	Performance β3,P1
0.2427365435	0.8207683226	0.0068735215
0.2838782503	0.6193219672	0.7603308253
0.4728852466	0.718615256	0.991473224
0.1602043263	0.6568282119	0.9653211501
0.3113725667	0.003657249	0.9002240284
0.7092353466	0.9641411756	0.9044996039
0.1243187189	0.1916947681	0.5001887854
0.9923597255	0.4643217917	0.4644260767
0.1593878649	0.7075588114	0.6043046496
0.7137943972	0.7178264102	0.3684897267
0.4405825825	0.9738042639	0.6371247198
0.0546034079	0.1643357663	0.8491521557
0.130989165	0.8930972954	0.9200755227
0.4630962713	0.7359805298	0.7894468571
0.3653479179	0.0494488408	0.4480319903

Performance β1,P2	Performance β2,P2	Performance β3,P2
0.6513221103	0.7155298328	0.5250285096
0.4486536453	0.4544934118	0.2665807758
0.923068983	0.4432370842	0.0714614966
0.2180154489	0.8604004404	0.2213692251
0.871509453	0.4888057283	0.9734517445
0.5255328568	0.120754892	0.8236567329
0.7085732815	0.5772912123	0.7173770375
0.869020659	0.8938754508	0.6566561072
0.674964929	0.0196329623	0.5775361005
0.1924289421	0.5813982673	0.0571435841
0.3358277807	0.1917446121	0.0112761131
0.7760143983	0.2131797303	0.4513054562
0.5871792892	0.9556053877	0.1188456733
0.2420052565	0.50039103	0.7654434184
0.9896802846	0.1324466465	0.6181376898

Performance β1	Performance β2	Performance β3
0.2427365435	0.8207683226	0.0068735215
0.2838782503	0.6193219672	0.7603308253
0.4728852466	0.718615256	0.991473224
0.1602043263	0.6568282119	0.9653211501
0.3113725667	0.003657249	0.9002240284
0.7092353466	0.9641411756	0.9044996039
0.1243187189	0.1916947681	0.5001887854
0.9923597255	0.4643217917	0.4644260767
0.1593878649	0.7075588114	0.6043046496
0.7137943972	0.7178264102	0.3684897267
0.4405825825	0.9738042639	0.6371247198
0.0546034079	0.1643357663	0.8491521557
0.130989165	0.8930972954	0.9200755227
0.4630962713	0.7359805298	0.7894468571
0.3653479179	0.0494488408	0.4480319903
0.6513221103	0.7155298328	0.5250285096
0.4486536453	0.4544934118	0.2665807758
0.923068983	0.4432370842	0.0714614966
0.2180154489	0.8604004404	0.2213692251
0.871509453	0.4888057283	0.9734517445
0.5255328568	0.120754892	0.8236567329
0.7085732815	0.5772912123	0.7173770375
0.869020659	0.8938754508	0.6566561072
0.674964929	0.0196329623	0.5775361005
0.1924289421	0.5813982673	0.0571435841
0.3358277807	0.1917446121	0.0112761131
0.7760143983	0.2131797303	0.4513054562
0.5871792892	0.9556053877	0.1188456733
0.2420052565	0.50039103	0.7654434184
0.9896802846	0.1324466465	0.6181376898

Performance P1	Performance P2
0.2427365435	0.6513221103
0.2838782503	0.4486536453
0.4728852466	0.923068983
0.1602043263	0.2180154489
0.3113725667	0.871509453
0.7092353466	0.5255328568
0.1243187189	0.7085732815
0.9923597255	0.869020659
0.1593878649	0.674964929
0.7137943972	0.1924289421
0.4405825825	0.3358277807
0.0546034079	0.7760143983
0.130989165	0.5871792892
0.4630962713	0.2420052565
0.3653479179	0.9896802846
0.8207683226	0.7155298328
0.6193219672	0.4544934118
0.718615256	0.4432370842
0.6568282119	0.8604004404
0.003657249	0.4888057283
0.9641411756	0.120754892
0.1916947681	0.5772912123
0.4643217917	0.8938754508
0.7075588114	0.0196329623
0.7178264102	0.5813982673
0.9738042639	0.1917446121
0.1643357663	0.2131797303
0.8930972954	0.9556053877
0.7359805298	0.50039103
0.0494488408	0.1324466465
0.0068735215	0.5250285096
0.7603308253	0.2665807758
0.991473224	0.0714614966
0.9653211501	0.2213692251
0.9002240284	0.9734517445
0.9044996039	0.8236567329
0.5001887854	0.7173770375
0.4644260767	0.6566561072
0.6043046496	0.5775361005
0.3684897267	0.0571435841
0.6371247198	0.0112761131
0.8491521557	0.4513054562
0.9200755227	0.1188456733
0.7894468571	0.7654434184
0.4480319903	0.6181376898

Which Statistical Test To Use?

You could potentially use one of the statistical tests for N groups. Kruskal-Wallis and Friedman are non-parametric, but ANOVA enables comparison of multiple factors and their interactions.

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

Remember that the problem instance can be a between-subjects factor in ANOVA.

Overview

- Recap of the general idea underlying statistical hypothesis tests.
- What to compare?
    - Two algorithms on a single problem instance.
    - Two algorithms on multiple problem instances.
    - Multiple algorithms on a single problem instance.
    - Multiple algorithms on multiple problem instances.
- How to design the comparisons?
    - Tests for 2 groups.
        - Observation corresponds to a single run.
        - Observation corresponds to the aggregation of multiple runs.
    - Tests for N groups.
        - Observation corresponds to a single run.
        - Observation corresponds to the aggregation of multiple runs.
- Commands to run the statistical tests.

Comparing Multiple Algorithms On Multiple Problem Instances Using a Test for N Groups

One Comparison

Problem Instance

Average Performance for A1	Average Performance for A2	Average Performance for A3	Average Performance for A4
0.6015110151	0.0633347888	0.0633347888	0.0633347888
0.2947677998	1.0930402922	1.0930402922	1.0930402922
0.9636589224	0.1792341981	0.1792341981	0.1792341981
0.251976978	1.207096969	1.207096969	1.207096969
0.3701006544	1.0606484322	1.0606484322	1.0606484322
0.9940754515	0.6473818857	0.6473818857	0.6473818857
0.4283523627	0.8043431063	0.8043431063	0.8043431063
0.1904817054	0.658958582	0.658958582	0.658958582
0.7377491128	1.0576089397	1.0576089397	1.0576089397
0.5392380701	0.7364416374	0.7364416374	0.7364416374
0.4230920852	0.1942901434	0.1942901434	0.1942901434
0.7221442924	0.5849134532	0.5849134532	0.5849134532
0.8882444038	0.4971571929	0.4971571929	0.4971571929

. . .

- An observation in a group may be, e.g.:
  - The average of multiple runs of the group's EA on a given problem instance.
    - The multiple runs are performed by varying the EA's random seed.
  - The average of multiple runs of the group's ML algorithm on a given dataset.
    - The multiple runs are performed by varying the ML algorithm's random seed and/or training/validation/test sample.

Comparing Multiple Algorithms On Multiple Problem Instances Using a Test for N Groups

- Similar to comparison of two algorithms over multiple problem instances, we can consider each observation to be the average results of a given algorithm on a given problem instance over multiple runs.

- But also similar to comparison of multiple algorithms over a single problem instance, instead of using a statistical test for 2 groups, we use for N groups.

- Advantages and disadvantages can be derived as before.

Examples of Statistical Hypothesis Tests

You could potentially use one of the statistical tests for paired N groups, most likely Friedman.

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

Overview

- Recap of the general idea underlying statistical hypothesis tests.
- What to compare?
    - Two algorithms on a single problem instance.
    - Two algorithms on multiple problem instances.
    - Multiple algorithms on a single problem instance.
    - Multiple algorithms on multiple problem instances.
- How to design the comparisons?
    - Tests for 2 groups.
        - Observation corresponds to a single run.
        - Observation corresponds to the aggregation of multiple runs.
    - Tests for N groups.
        - Observation corresponds to a single run.
        - Observation corresponds to the aggregation of multiple runs.
- Commands to run the statistical tests.

Software or Programming Languages With Statistical Support

- Many available:
    - R, Matlab, SPSS, etc.
- R:
    - Programming language for statistical computing.
    - Can be used to run statistical tests.

Reading Observations

- You can enter observations manually, or you can load observations from a .csv table. E.g.:
  - `observations2 = read.csv('/Users/minkull/Desktop/observations-two-groups.csv', header = TRUE, sep = ",")`
- For help with a command:
  - `help(command)`

```
Group 1,Group 2
0.803680873,0.944255293
0.154602685,0.727712943
0.150708502,0.431981162
0.97511866,0.937983685
0.460232148,0.786503003
0.013223879,0.819113932
0.017511488,0.92368809
0.904174174,0.815563594
0.869770096,0.76943584
0.676352134,0.321770206
0.518232817,0.984916141
0.051641168,0.258640987
0.542664965,0.794543475
0.497362926,0.817948571
0.486607913,0.413216708
0.218745577,0.591558823
0.843827421,0.593674664
0.264400949,0.438692375
0.256434446,0.743990941
0.079121486,0.795106819
0.285609383,0.331450863
0.379775917,0.9218094
0.59789627,0.750849697
0.08605325,0.13729544
0.2860286,0.12517536
0.277279003,0.785829481
0.728984666,0.459297733
0.381243886,0.158332721
0.114495351,0.403745207
0.71283282,0.807401962
```

Accessing Observations

- observations2[1,2]
- observations2[,2]
- observations2[1,]

- You can type observations2[1,2], observations2[,2] and observations2[1,] in R to see their content.

Group 1	Group 2
0.803680873	0.944255293
0.154602685	0.727712943
0.150708502	0.431981162
0.97511866	0.937983685
0.460232148	0.786503003
0.013223879	0.819113932
0.017511488	0.92368809
0.904174174	0.815563594
0.869770096	0.76943584
0.676352134	0.321770206
0.518232817	0.984916141
0.051641168	0.258640987
0.542664965	0.794543475
0.497362926	0.817948571
0.486607913	0.413216708
0.218745577	0.591558823
0.843827421	0.593674664
0.264400949	0.438692375
0.256434446	0.743990941
0.079121486	0.795106819
0.285609383	0.331450863
0.379775917	0.9218094
0.59789627	0.750849697
0.08605325	0.13729544
0.2860286	0.12517536
0.277279003	0.785829481
0.728984666	0.459297733
0.381243886	0.158332721
0.114495351	0.403745207
0.71283282	0.807401962

Accessing Observations

- observations2[1,2] —> take the observation from the first row and second column

Group 1	Group 2
0.803680873	0.944255293
0.154602685	0.727712943
0.150708502	0.431981162
0.97511866	0.937983685
0.460232148	0.786503003
0.013223879	0.819113932
0.017511488	0.92368809
0.904174174	0.815563594
0.869770096	0.76943584
0.676352134	0.321770206
0.518232817	0.984916141
0.051641168	0.258640987
0.542664965	0.794543475
0.497362926	0.817948571
0.486607913	0.413216708
0.218745577	0.591558823
0.843827421	0.593674664
0.264400949	0.438692375
0.256434446	0.743990941
0.079121486	0.795106819
0.285609383	0.331450863
0.379775917	0.9218094
0.59789627	0.750849697
0.08605325	0.13729544
0.2860286	0.12517536
0.277279003	0.785829481
0.728984666	0.459297733
0.381243886	0.158332721
0.114495351	0.403745207
0.71283282	0.807401962

Accessing Observations

- observations2[1,2] —> take the observation from the first row and second column
- observations2[,2] —>

Group 1	Group 2
0.803680873	0.944255293
0.154602685	0.727712943
0.150708502	0.431981162
0.97511866	0.937983685
0.460232148	0.786503003
0.013223879	0.819113932
0.017511488	0.92368809
0.904174174	0.815563594
0.869770096	0.76943584
0.676352134	0.321770206
0.518232817	0.984916141
0.051641168	0.258640987
0.542664965	0.794543475
0.497362926	0.817948571
0.486607913	0.413216708
0.218745577	0.591558823
0.843827421	0.593674664
0.264400949	0.438692375
0.256434446	0.743990941
0.079121486	0.795106819
0.285609383	0.331450863
0.379775917	0.9218094
0.59789627	0.750849697
0.08605325	0.13729544
0.2860286	0.12517536
0.277279003	0.785829481
0.728984666	0.459297733
0.381243886	0.158332721
0.114495351	0.403745207
0.71283282	0.807401962

Accessing Observations

- observations2[1,2] —> take the observation from the first row and second column
- observations2[,2] —> take all the observations from the second column

Group 1	Group 2
0.803680873	0.944255293
0.154602685	0.727712943
0.150708502	0.431981162
0.97511866	0.937983685
0.460232148	0.786503003
0.013223879	0.819113932
0.017511488	0.92368809
0.904174174	0.815563594
0.869770096	0.76943584
0.676352134	0.321770206
0.518232817	0.984916141
0.051641168	0.258640987
0.542664965	0.794543475
0.497362926	0.817948571
0.486607913	0.413216708
0.218745577	0.591558823
0.843827421	0.593674664
0.264400949	0.438692375
0.256434446	0.743990941
0.079121486	0.795106819
0.285609383	0.331450863
0.379775917	0.9218094
0.59789627	0.750849697
0.08605325	0.13729544
0.2860286	0.12517536
0.277279003	0.785829481
0.728984666	0.459297733
0.381243886	0.158332721
0.114495351	0.403745207
0.71283282	0.807401962

Accessing Observations

- observations2[1,2] —> take the observation from the first row and second column
- observations2[,2] —> take all the observations from the second column
- observations2[1,] —> ?

Group 1	Group 2
0.803680873	0.944255293
0.154602685	0.727712943
0.150708502	0.431981162
0.97511866	0.937983685
0.460232148	0.786503003
0.013223879	0.819113932
0.017511488	0.92368809
0.904174174	0.815563594
0.869770096	0.76943584
0.676352134	0.321770206
0.518232817	0.984916141
0.051641168	0.258640987
0.542664965	0.794543475
0.497362926	0.817948571
0.486607913	0.413216708
0.218745577	0.591558823
0.843827421	0.593674664
0.264400949	0.438692375
0.256434446	0.743990941
0.079121486	0.795106819
0.285609383	0.331450863
0.379775917	0.9218094
0.59789627	0.750849697
0.08605325	0.13729544
0.2860286	0.12517536
0.277279003	0.785829481
0.728984666	0.459297733
0.381243886	0.158332721
0.114495351	0.403745207
0.71283282	0.807401962

Accessing Observations

- observations2[1,2] —> take the observation from the first row and second column
- observations2[,2] —> take all the observations from the second column
- observations2[1,] —> take all the observations from the first row

Group 1	Group 2
0.803680873	0.944255293
0.154602685	0.727712943
0.150708502	0.431981162
0.97511866	0.937983685
0.460232148	0.786503003
0.013223879	0.819113932
0.017511488	0.92368809
0.904174174	0.815563594
0.869770096	0.76943584
0.676352134	0.321770206
0.518232817	0.984916141
0.051641168	0.258640987
0.542664965	0.794543475
0.497362926	0.817948571
0.486607913	0.413216708
0.218745577	0.591558823
0.843827421	0.593674664
0.264400949	0.438692375
0.256434446	0.743990941
0.079121486	0.795106819
0.285609383	0.331450863
0.379775917	0.9218094
0.59789627	0.750849697
0.08605325	0.13729544
0.2860286	0.12517536
0.277279003	0.785829481
0.728984666	0.459297733
0.381243886	0.158332721
0.114495351	0.403745207
0.71283282	0.807401962

Statistical Hypothesis 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

Two-Tailed Wilcoxon Sign Rank in R

```
wilcox.test(x, y, alternative = "two.sided", paired = TRUE, conf.level = 0.95)
```

- Example:
  - H0: μ1 = μ2
  - H1: μ1 ≠ μ2
  - Level of significance = 0.05
  - `result = wilcox.test(observations2[,1], observations2[,2],alternative = "two.sided",paired=TRUE, conf.level = 0.95)`
  - p-value: 0.002766 ≤ 0.05
    - Reject H0.
    - Statistically significantly difference between μ1 and μ2 has been found at the level of significance of 0.05 (p-value = 0.002766).
    - median(observations2[,1]) = 0.3805, median(observations2[,2]) = 0.7474
    - μ1 is significantly smaller than μ2

Completely Equal Pairs of Observations

- `observationnull = read.csv('/Users/minkull/Desktop/observations_null.csv', header = TRUE, sep = ",")`
- `wilcox.test(observationnull[,1],observationnull[,2],alternative = "two.sided",paired=TRUE, conf.level = 0.95)`
  - `p-value = NA`

```
Group 1,Group 2
1,1
2,2
3,3
4,4
5,5
6,6
7,7
8,8
9,9
10,10
11,11
12,12
13,13
14,14
15,15
16,16
17,17
18,18
19,19
20,20
21,21
22,22
23,23
24,24
25,25
26,26
27,27
28,28
29,29
30,30
```

Statistical Hypothesis 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

Two-Tailed Wilcoxon Rank Sum in R

```
wilcox.test(x, y, alternative = "two.sided", paired = FALSE, conf.level = 0.95)
```

- Example:
  - H0: μ1 = μ2
  - H1: μ1 ≠ μ2
  - Level of significance = 0.05
  - `result = wilcox.test(observations2[,1],observations2[,2],alternative = "two.sided",paired=FALSE, conf.level = 0.95)`
  - p-value: 0.007647 ≤ 0.05
    - Reject H0.
    - Statistically significantly difference between μ1 and μ2 has been found at the level of significance of 0.05 (p-value = 0.007647).
    - median(observations2[,1]) = 0.3805, median(observations2[,2]) = 0.7474
    - μ1 is significantly smaller than μ2

Statistical Hypothesis 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

Unpaired (Welch) T-Test in R

```
t.test(x, y, alternative = "two.sided", paired = FALSE)
```

- Example:
  - H0: μ1 = μ2
  - H1: μ1 ≠ μ2
  - Level of significance = 0.05
  - `result = t.test(observations2[,1],observations2[,2],alternative = "two.sided",paired=FALSE)`
  - p-value: 0.006003 ≤ 0.05
    - Reject H0.
    - Statistically significantly difference between μ1 and μ2 has been found at the level of significance of 0.05 (p-value = 0.006003).
    - mean(observations2[,1]) = 0.4211538, mean(observations2[,2]) = 0.6263828
    - μ1 is significantly smaller than μ2

Statistical Hypothesis 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

Paired T-Test in R

```
t.test(x, y, alternative = "two.sided", paired = TRUE)
```

- Example:
  - H0: $\mu1 = \mu2$
  - H1: $\mu1 \neq \mu2$
  - Level of significance = 0.05
  - `result = t.test(observations2[,1],observations2[,2],alternative = "two.sided",paired=TRUE)`
  - p-value: $0.00185 \leq 0.05$
    - Reject H0.
    - Statistically significantly difference between μ1 and μ2 has been found at the level of significance of 0.05 (p-value = 0.00185).
    - mean(observations2[,1]) = 0.4211538, mean(observations2[,2]) = 0.6263828
    - μ1 is significantly smaller than μ2

Statistical Hypothesis 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

Friedman Test for Paired Comparisons in R

- R command:
  - `result = friedman.test(matrix_observationsn)`

    matrix_observationsn contains a matrix of groups to be compared.

  - When reading from a .csv file, read.csv reads data into an observations “frame”. E.g.:

    ```
    observationsn <- read.csv('/Users/minkull/
    Desktop/observations-n-groups.csv')
    ```

  - To convert from a frame to a matrix, you can use the list command. E.g.:

    ```
    matrix_observationsn =
    data.matrix(observationsn)
    ```

Friedman Test for Paired Comparisons

- Example:
  - H0: all groups are equal
  - H1: at least one pair of groups is different

  - p-value = 8.935e-09 < 0.05 (Reject H0)

Post-Hoc Tests in R

- You need to install the following package: PMCMRPlus
  - install.packages(“PMCMRplus”)

- Once installed, load package:
  - library(PMCMRplus)

PMCMR Package's Nemenyi Post-Hoc Test for All Pairs

- R command:
  - `result = frdAllPairsNemenyiTest(observationsn)`
- This test already accounts for multiple comparisons. So, no further corrections are needed.
- Example:

	Group 1	**Group 2**
Group 2	0.16711	—
Group 3	8.6E-09	0.00011

PMCMR Package's Nemenyi Post-Hoc Test Against Control Group

- R command:
  - `result = frdManyOneNemenyiTest(observationsn)`
- This test already accounts for multiple comparisons. So, no further corrections are needed.
- Example:

	Group 1
Group 2	0.13
Group 3	5.7E-09

Tsutils Package's Nemenyi with Plot Options in R

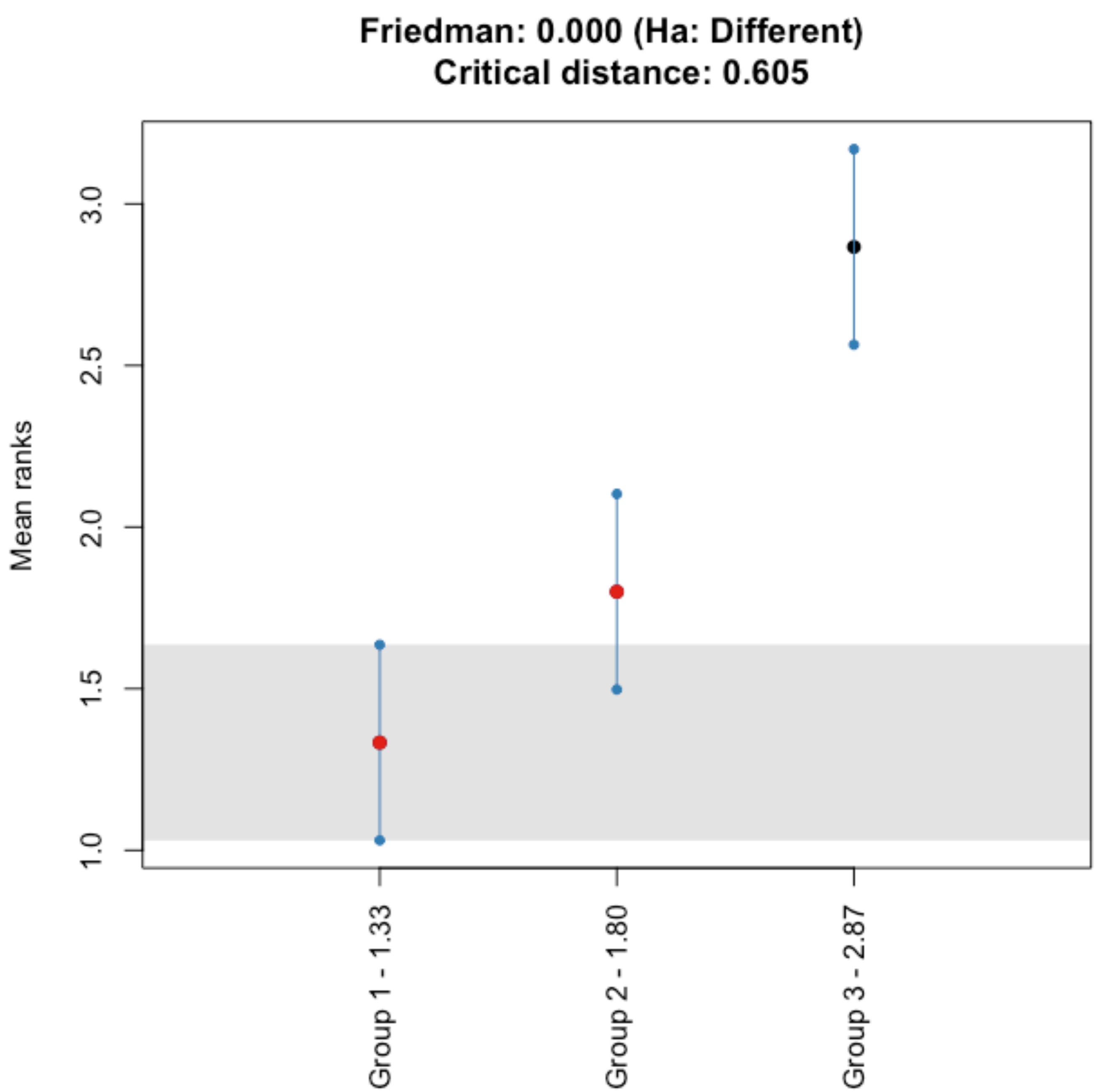
- install.packages("tsutils")
- library(tsutils)

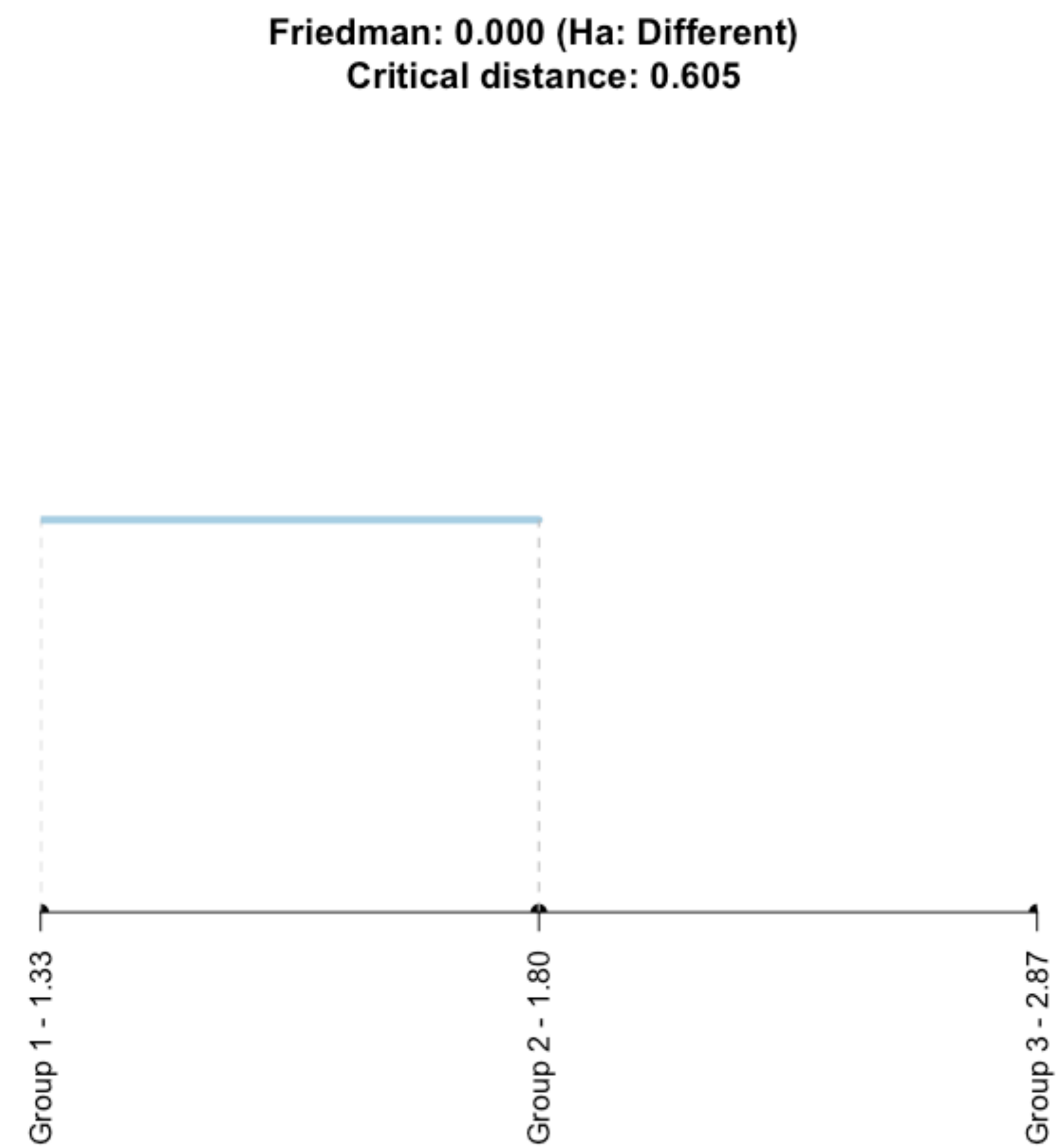
- result = nemenyi(matrix_observationsn,conf.level=0.95,plottype='mcb',labels=c('Group 1','Group 2','Group 3'))

- result = nemenyi(matrix_observationsn,conf.level=0.95,plottype='line',labels=c('Group 1','Group 2','Group 3'))

- Rankings assume that smaller values have smaller ranks.

Tsutils Package's Nemenyi with Plot Options in R

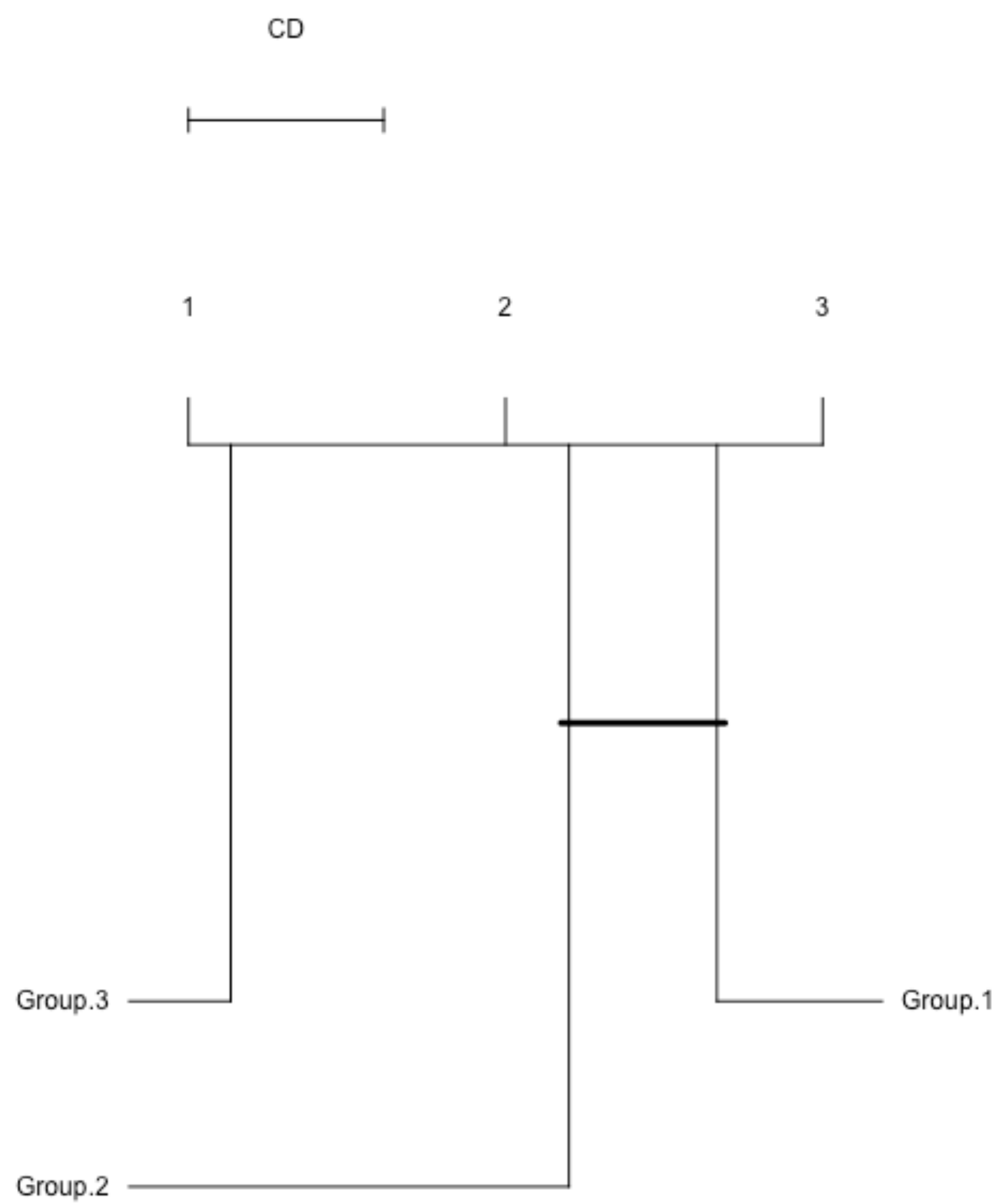




Critical Distance Plot from Package scmamp in R

- How to install latest version: https://rdrr.io/cran/scmamp/f/README.md
- if (!require("devtools")) {
-     install.packages("devtools")
- }
- devtools::install_github("b0rxa/scmamp")
- library("scmamp")
- result = plotCD(matrix_observationsn,alpha=0.05)
- Rankings assume that larger values have smaller ranks.

Critical Distance Plot from Package scmamp in R

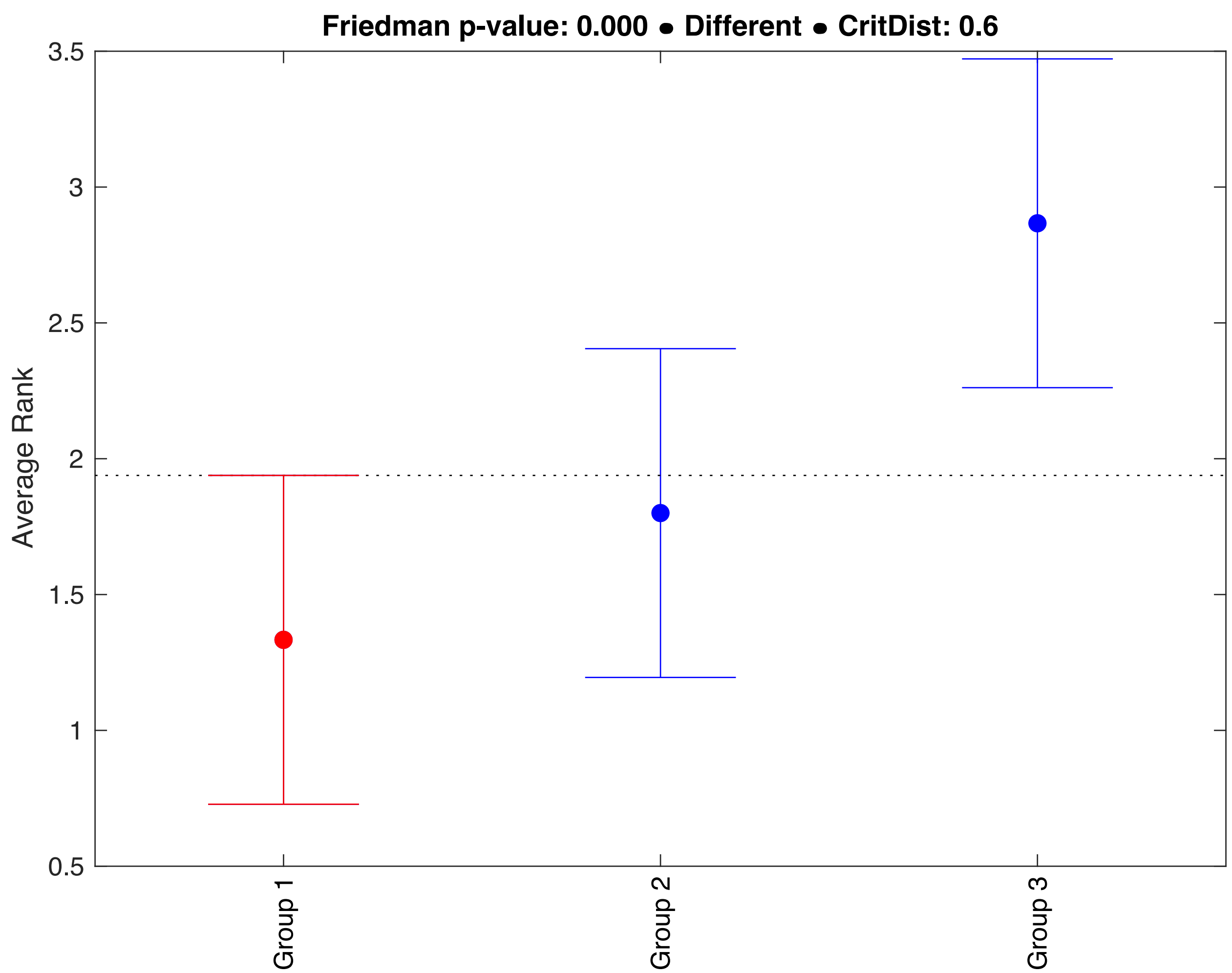


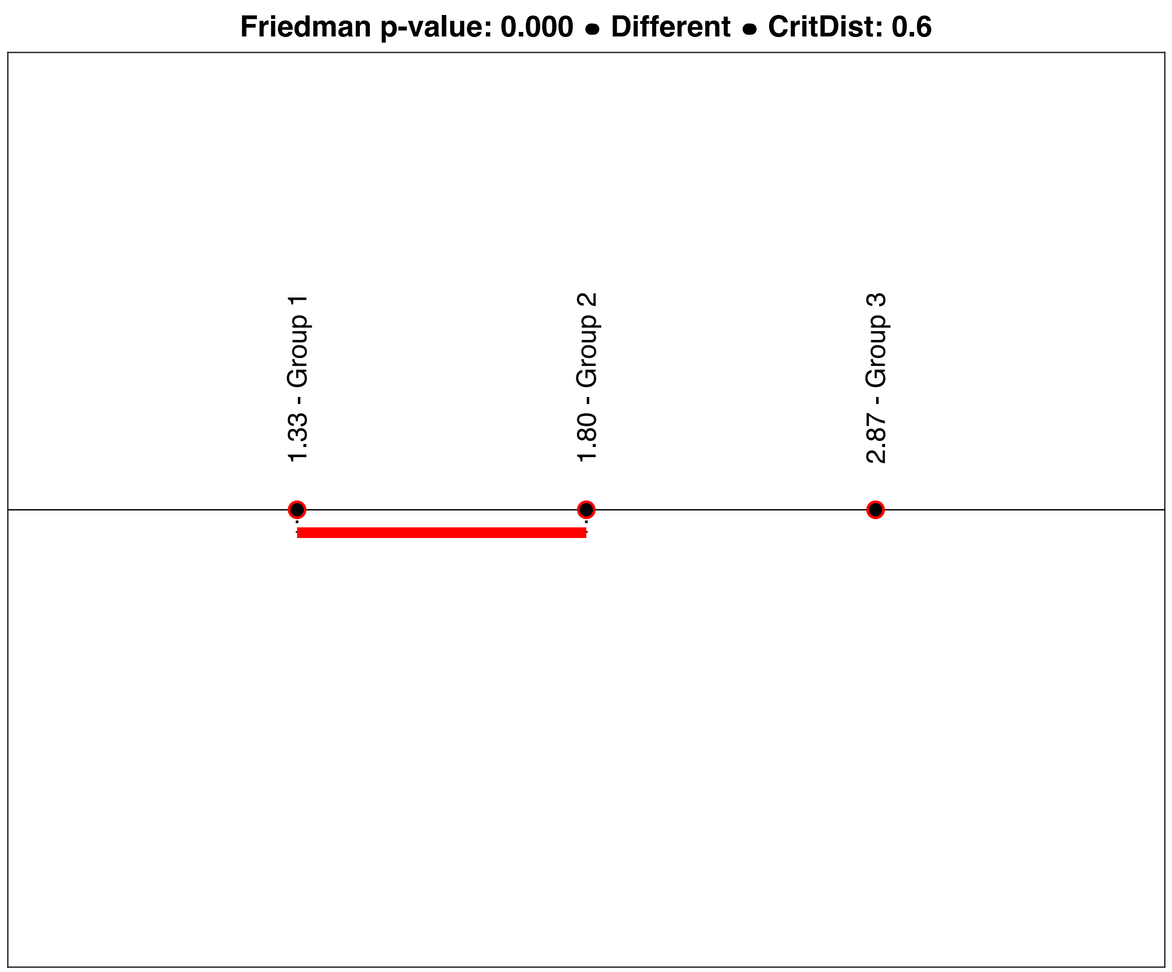
Nikolaos Kourentzes' Nemenyi Code for Matlab

- Download Nikolaos Kourentzes code at: http://kourentzes.com/forecasting/wp-content/uploads/2016/08/anom_nem_tests_matlab.zip

- Example:
  - observationsn = readtable('observations-n-groups.csv','HeaderLines', 1)
  - obsn = table2array(observationsn)
  - labels=["Group 1","Group 2","Group 3"]
  - [p, testresult, meanrank, CDa, rankmean] = nemenyi(obsn, 1,'alpha',0.05,'labels',labels,'ploton','mcb');
  - [p, testresult, meanrank, CDa, rankmean] = nemenyi(obsn, 1,'alpha',0.05,'labels',labels,'ploton','line');

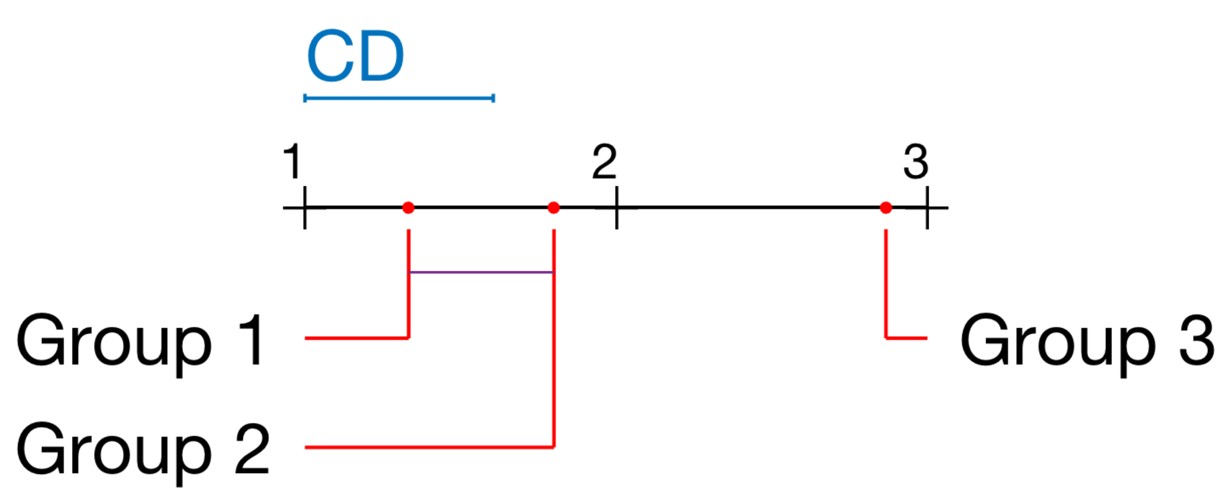
Nikolaos Kourentzes' Nemenyi Code for Matlab





Farshid Sepehrband’s Matlab Nemenyi Code

- Download the following code for Nemenyi and useful plot style:
  - https://zenodo.org/badge/latestdoi/45722511
  - Example:
  - drawNemenyi(obsn,labels,’~/Desktop','tmp-plot')



Statistical Hypothesis 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

Kruskal-Wallis Test for Unpaired Comparisons

- R command:
  - `result = kruskal.test(list_observationsn)`

    list_observations contains a list of groups to be compared.

  - When reading from a .csv file, read.csv reads data into an observations “frame”. E.g.:

    ```
    observations <- read.csv('/Users/minkull/
    Desktop/observations-n-groups.csv')
    ```

  - To convert from a frame to a list, you can use the list command. E.g.:

    ```
    list_observationsn = list(observationsn[,
    1],observationsn[,2],observationsn[,3])
    ```

Kruskal-Wallis Test for Unpaired Comparisons

- Example:
  - H0: all groups are equal
  - H1: at least one pair of groups is different

  - p-value = 1.338e-11 < 0.05 (Reject H0)

Dunn Post-Hoc Test

- R command:
    - `library("PMCMRplus")`
    - `result = kwAllPairsDunnTest(observationsn, p.adjust.method = "holm")`
- This test requires corrections to account for multiple comparisons (e.g., holm-bonferroni).
- Example:

	Group 1	Group 2
Group 2	0.052	—
Group 3	2.0E-11	1.7E-06

Dunn Post-Hoc Test

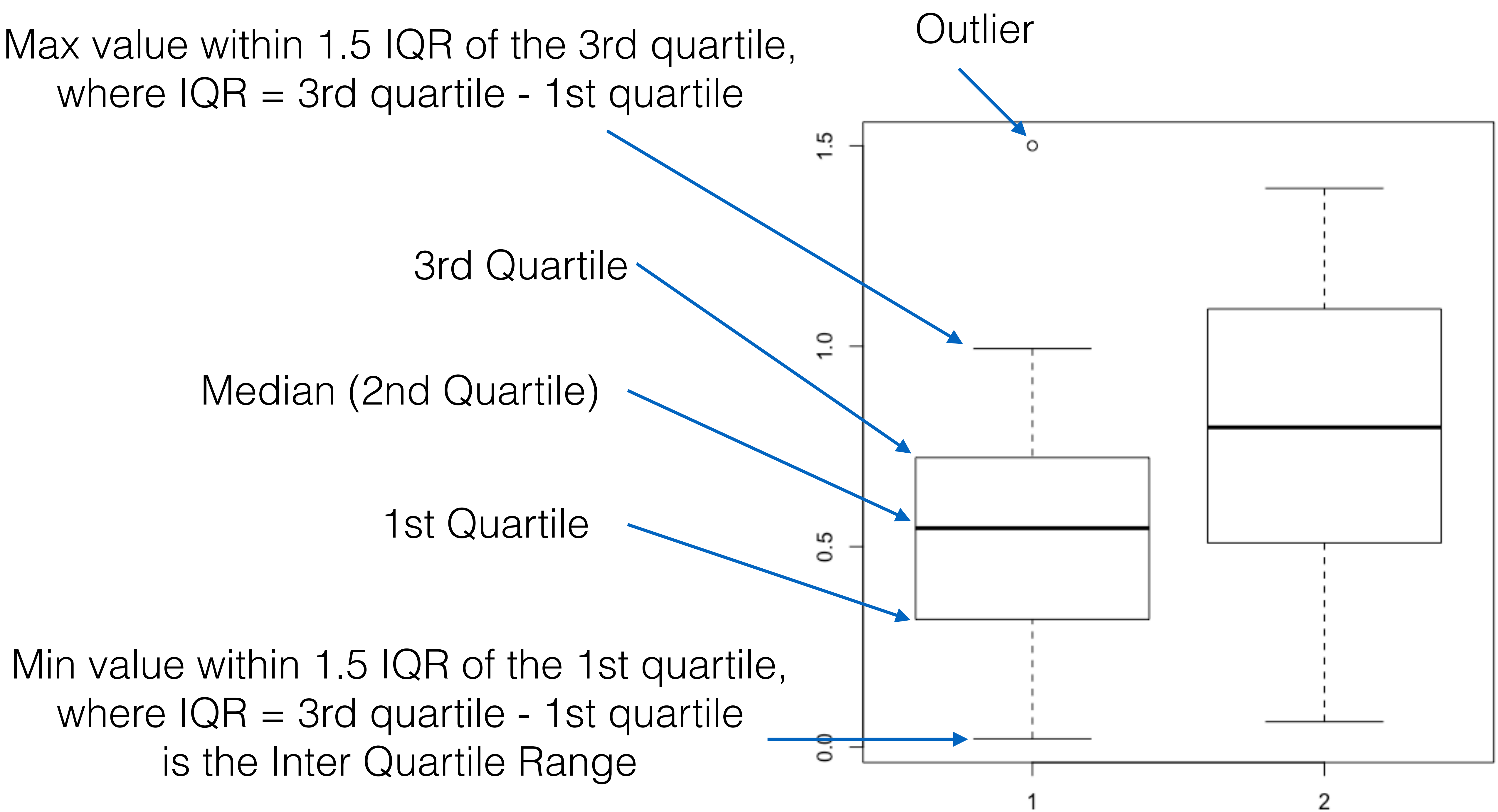
- R command:
  - `library("PMCMRplus")`
  - `result = kwManyOneDunnTest(observationsn, p.adjust.method = "holm")`
- This test requires corrections to account for multiple comparisons (e.g., holm-bonferroni).
- Example:

	Group 1
Group 2	0.094
Group 3	1.3E-11

A12 Effect Size in Matlab

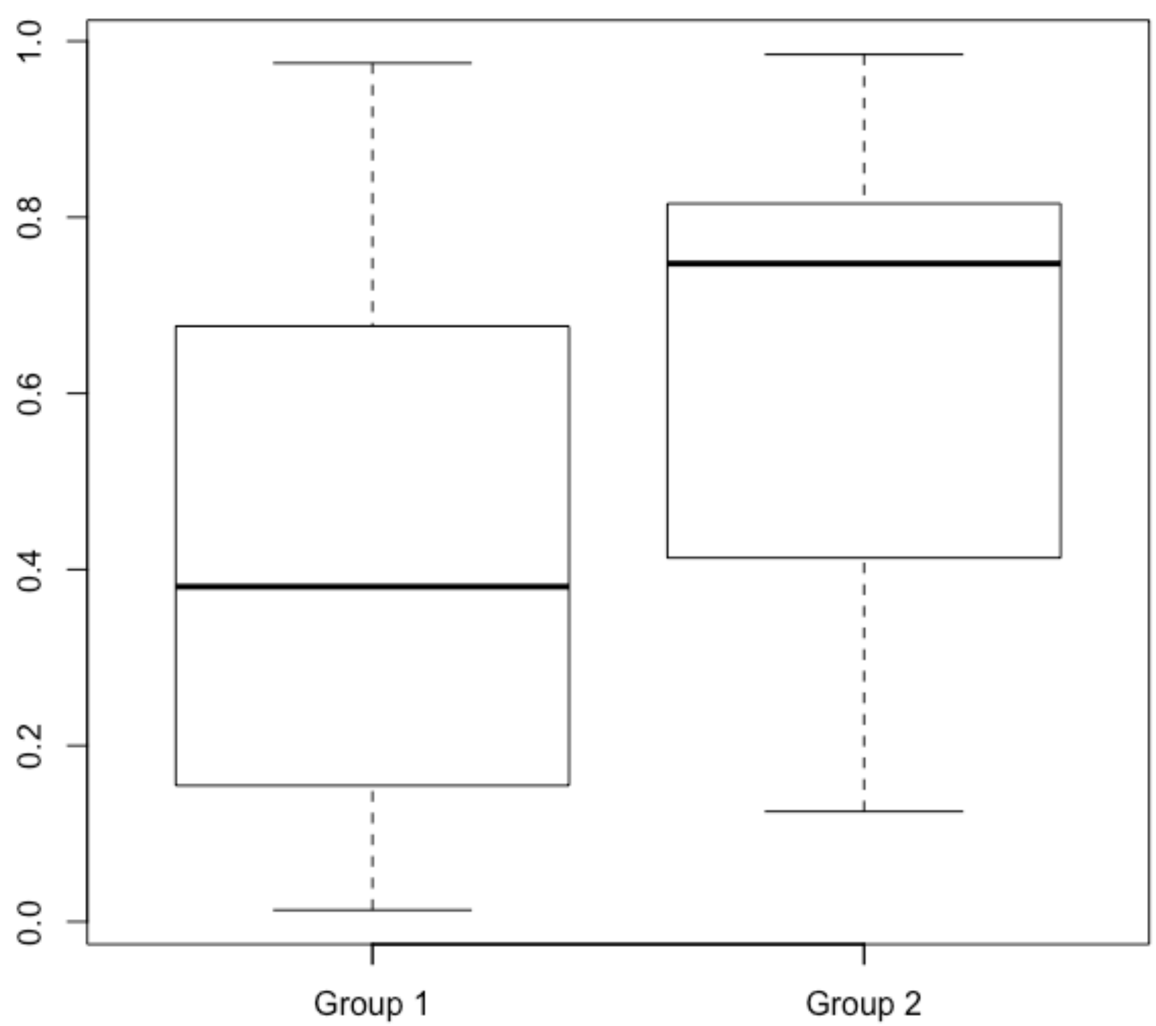
- Matlab implementation of A12 available at: https://github.com/minkull/A12-Effect-Size
- Example:
  - observations = readtable('observations-two-groups.csv','HeaderLines', 1)
  - obs = table2array(observations)
  - a12(obs(:,1),obs(:,2))
    - -0.6989
  - The "-" sign is here to indicate that obs(:,1) are smaller than obs(:,2).

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
1.5
1.0
0.5
0.0
1
2

Creating Boxplots in R

- boxplot(observations2[,1], observations2[,2],labels="")
- axis(1, at=c(1,2), labels=c("Group 1","Group 2"))



Statistical Hypothesis 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

Multi-Factor Repeated Measures ANOVA

- Open SPSS
- Load observations-anova-within-subject.sav
- Analyse -> General Linear Model -> Repeated Measures
- Create within-subject factors
  - B with 3 levels
  - D with 2 levels
- Select the columns corresponding to the observations of each factor.
- Click on Plot to decide which plots to create.
  - It's easier to decide which plots to create after running the test.
  - Normally, plots for significant factors and interactions are created.
- Click on options to select to print descriptive statistics and effect size.

Mauchly's Test of Sphericity[a]

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	p-value Sig.	Epsilon[b] Greenhouse-Geisser	 Huynh-Feldt	 Lower-bound
B	.848	2.147	2	.342	.868	.980	.500
D	1.000	.000	0	.	1.000	1.000	1.000
B * D	.965	.460	2	.794	.966	1.000	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept
Within Subjects Design: B + D + B * D

b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

Sphericity assumption is satisfied, as p-value > 0.05.

Tests of Within-Subjects Effects

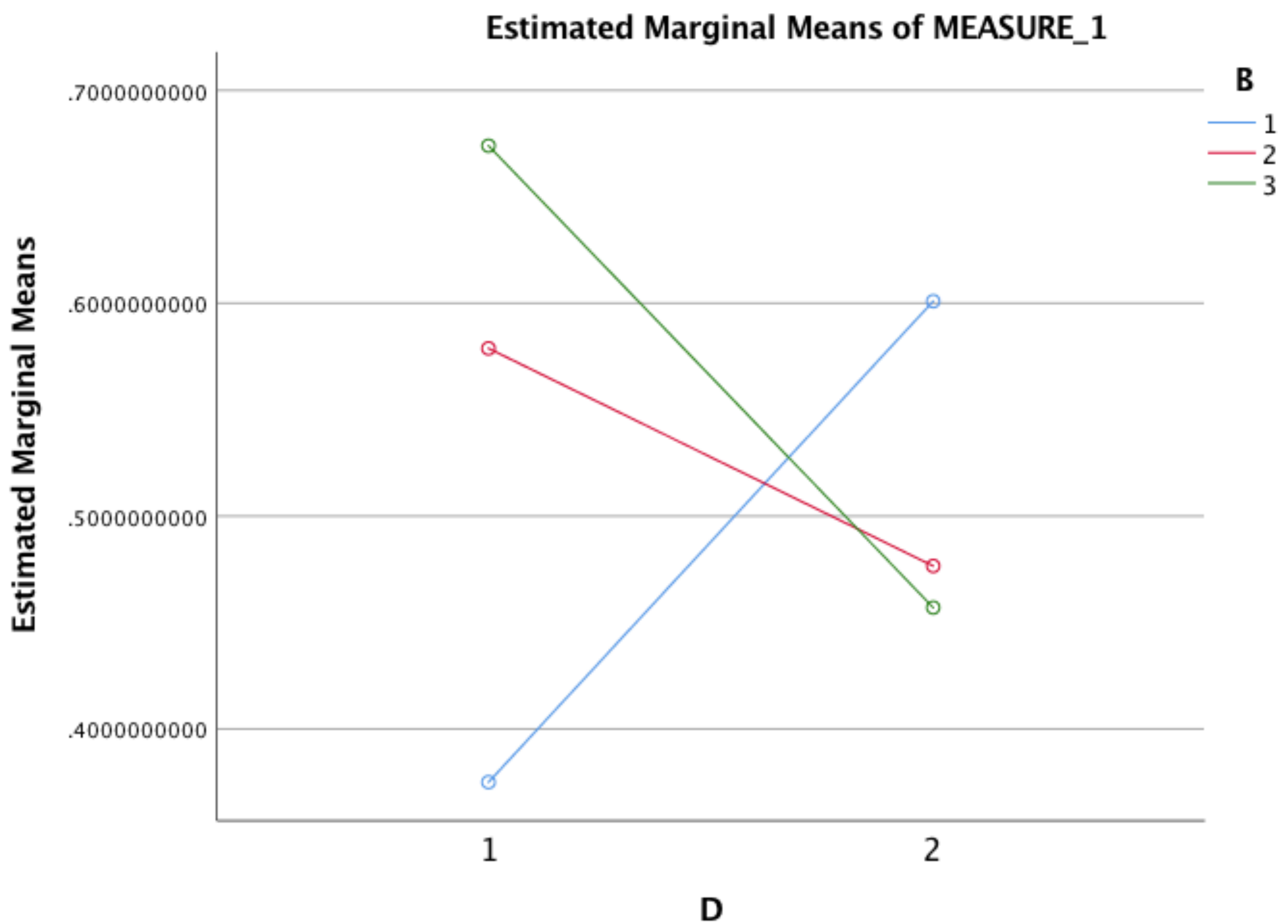
Measure: MEASURE_1

p-value

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
B	Sphericity Assumed	.090	2	.045	.451	.641	.031
	Greenhouse-Geisser	.090	1.736	.052	.451	.615	.031
	Huynh-Feldt	.090	1.960	.046	.451	.637	.031
	Lower-bound	.090	1.000	.090	.451	.513	.031
Error(B)	Sphericity Assumed	2.798	28	.100			
	Greenhouse-Geisser	2.798	24.301	.115			
	Huynh-Feldt	2.798	27.438	.102			
	Lower-bound	2.798	14.000	.200			
D	Sphericity Assumed	.022	1	.022	.179	.679	.013
	Greenhouse-Geisser	.022	1.000	.022	.179	.679	.013
	Huynh-Feldt	.022	1.000	.022	.179	.679	.013
	Lower-bound	.022	1.000	.022	.179	.679	.013
Error(D)	Sphericity Assumed	1.700	14	.121			
	Greenhouse-Geisser	1.700	14.000	.121			
	Huynh-Feldt	1.700	14.000	.121			
	Lower-bound	1.700	14.000	.121			
B * D	Sphericity Assumed	.793	2	.396	5.955	.007	.298
	Greenhouse-Geisser	.793	1.933	.410	5.955	.008	.298
	Huynh-Feldt	.793	2.000	.396	5.955	.007	.298
	Lower-bound	.793	1.000	.793	5.955	.029	.298
Error(B*D)	Sphe						
	Gree						
	Huyı						
	Lowe						

If sphericity was violated, we would use the p-value with Greenhouse-Geisser corrections.

Interaction Between B and D Is Significant



Effect Size Eta Squared

- Percentage of the variance accounted for a factor or interaction.

- Calculated as follows:
  - Total = Sum the Type III Sum of Squares for all factors, interactions and errors.
  - Divided the Type III Sum of Squares of a given factor or interaction by Total.

- Rule of thumb:
  - Small: 0.01
  - Medium: 0.06
  - Large: 0.14

Miles J and Shevlin M (2001) Applying Regression and Correlation: A Guide for Students and Researchers. Sage:London.

Tests of Within-Subjects Effects

Measure: MEASURE_1

p-value

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
B	Sphericity Assumed	.090	2	.045	.451	.641	.031
	Greenhouse-Geisser	.090	1.736				
	Huynh-Feldt	.090	1.960				
	Lower-bound	.090	1.000				
Error(B)	Sphericity Assumed	2.798	28				
	Greenhouse-Geisser	2.798	24.301				
	Huynh-Feldt	2.798	27.438				
	Lower-bound	2.798	14.000				
D	Sphericity Assumed	.022	1				
	Greenhouse-Geisser	.022	1.000				
	Huynh-Feldt	.022	1.000				
	Lower-bound	.022	1.000				
Error(D)	Sphericity Assumed	1.700	14				
	Greenhouse-Geisser	1.700	14.000				
	Huynh-Feldt	1.700	14.000				
	Lower-bound	1.700	14.000				
B * D	Sphericity Assumed	.793	2				
	Greenhouse-Geisser	.793	1.933				
	Huynh-Feldt	.793	2.000				
	Lower-bound	.793	1.000				
Error(B*D)	Sphericity Assumed	1.863	28				
	Greenhouse-Geisser	1.863	27.059				
	Huynh-Feldt	1.863	28.000				
	Lower-bound	1.863	14.000	.133			

Example: eta squared for factor B

Total = .090 + 2.7898 + .022 + 1.700 + .793 + 1.863 = 7.2578

Eta squared = .090 / 7.2578 = 0.01240

Example: eta squared for interaction B*D

Total = .090 + 2.7898 + .022 + 1.700 + .793 + 1.863 = 7.2578

Eta squared = .793 / 7.2578 = 0.1093

Split Plot ANOVA

- Open SPSS
- Load observations-anova-split-plot.sav
  - Here, the problem instance is considered to be a between-subjects factor.
- Analyse -> General Linear Model -> Repeated Measures
- Create within-subject factors
  - B with 3 levels
  - D with 2 levels
- Select the columns corresponding to the observations of each within-subject factor.
- Select the column corresponding to the levels of the between-subjects factor.
- Click on Plot to decide which plots to create.
  - It’s easier to decide which plots to create after running the test.
  - Normally, plots for significant factors and interactions are created.
- Click on options to select to print descriptive statistics and effect size.

Mauchly's Test of Sphericity[a]

Measure: MEASURE_1

				p-value	Epsilon[b]		
Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Greenhouse-Geisser	Huynh-Feldt	Lower-bound
B	.974	.689	2	.708	.975	1.000	.500
D	1.000	.000	0	.	1.000	1.000	1.000
B * D	.924	2.054	2	.358	.929	1.000	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept + Problem
Within Subjects Design: B + D + B * D

b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

Sphericity assumption is satisfied, as p-value > 0.05.

Tests of Within-Subjects Effects

Measure: MEASURE_1

p-value

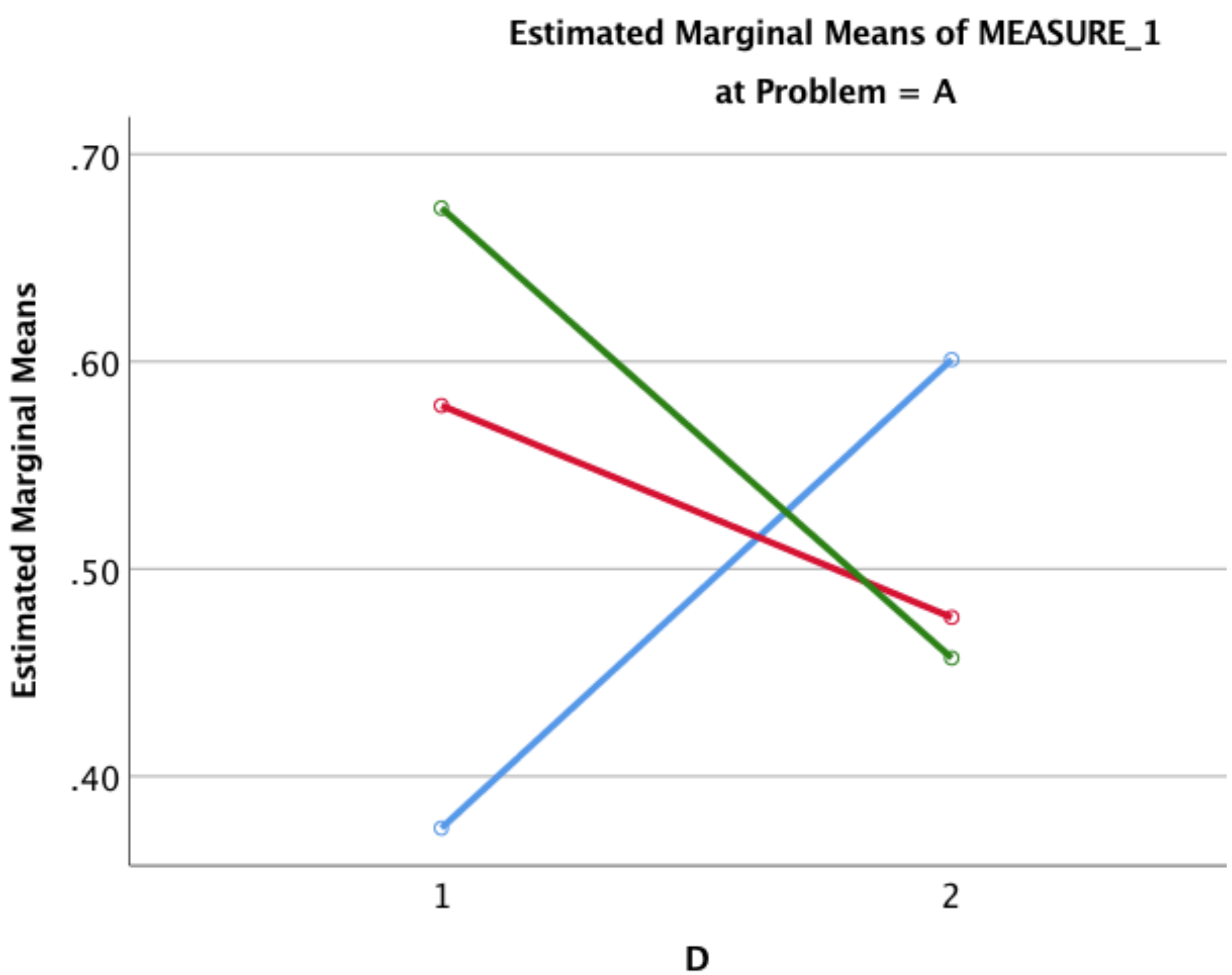
Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
B	Sphericity Assumed	.009	2	.004	.051	.951	.002
	Greenhouse-Geisser	.009	1.949	.004	.051	.948	.002
	Huynh-Feldt	.009	2.000	.004	.051	.951	.002
	Lower-bound	.009	1.000	.009	.051	.824	.002
B * Problem	Sphericity Assumed	.150	2	.075	.865	.427	.031
	Greenhouse-Geisser	.150	1.949	.077	.865	.424	.031
	Huynh-Feldt	.150	2.000	.075	.865	.427	.031
	Lower-bound	.150	1.000	.150	.865	.361	.031
Error(B)	Sphericity Assumed	4.673	54	.087			
	Greenhouse-Geisser	4.673	52.623	.089			
	Huynh-Feldt	4.673	54.000	.087			
	Lower-bound	4.673	27.000	.173			
D	Sphericity Assumed	.018	1	.018	.185	.670	.007
	Greenhouse-Geisser	.018	1.000	.018	.185	.670	.007
	Huynh-Feldt	.018	1.000	.018	.185	.670	.007
	Lower-bound	.018	1.000	.018	.185	.670	.007
D * Problem	Sphericity Assumed	.005	1	.005	.055	.817	.002
	Greenhouse-Geisser	.005	1.000	.005	.055	.817	.002
	Huynh-Feldt	.005	1.000	.005	.055	.817	.002
	Lower-bound	.005	1.000	.005	.055	.817	.002
Error(D)	Sphericity Assumed	2.562	27	.095			
	Greenhouse-Geisser	2.562	27.000	.095			
	Huynh-Feldt	2.562	27.000	.095			
	Lower-bound	2.562	27.000	.095			
B * D	Sphericity Assumed	1.008	2	.504	8.817	.000	.246
	Greenhouse-Geisser	1.008	1.859	.542	8.817	.001	.246
	Huynh-Feldt	1.008	2.000	.504	8.817	.000	.246
	Lower-bound	1.008	1.000	1.008	8.817	.006	.246
B * D * Problem	Sphericity Assumed	.247	2	.124	2.161	.125	.074
	Greenhouse-Geisser	.247	1.859	.133	2.161	.129	.074
	Huynh-Feldt	.247	2.000	.124	2.161	.125	.074
	Lower-bound	.247	1.000	.247	2.161	.153	.074
Error(B*D)	Sphericity Assumed	3.088	54	.057			
	Greenhouse-Geisser	3.088	50.188	.062			
	Huynh-Feldt	3.088	54.000	.057			
	Lower-bound	3.088	27.000	.114			

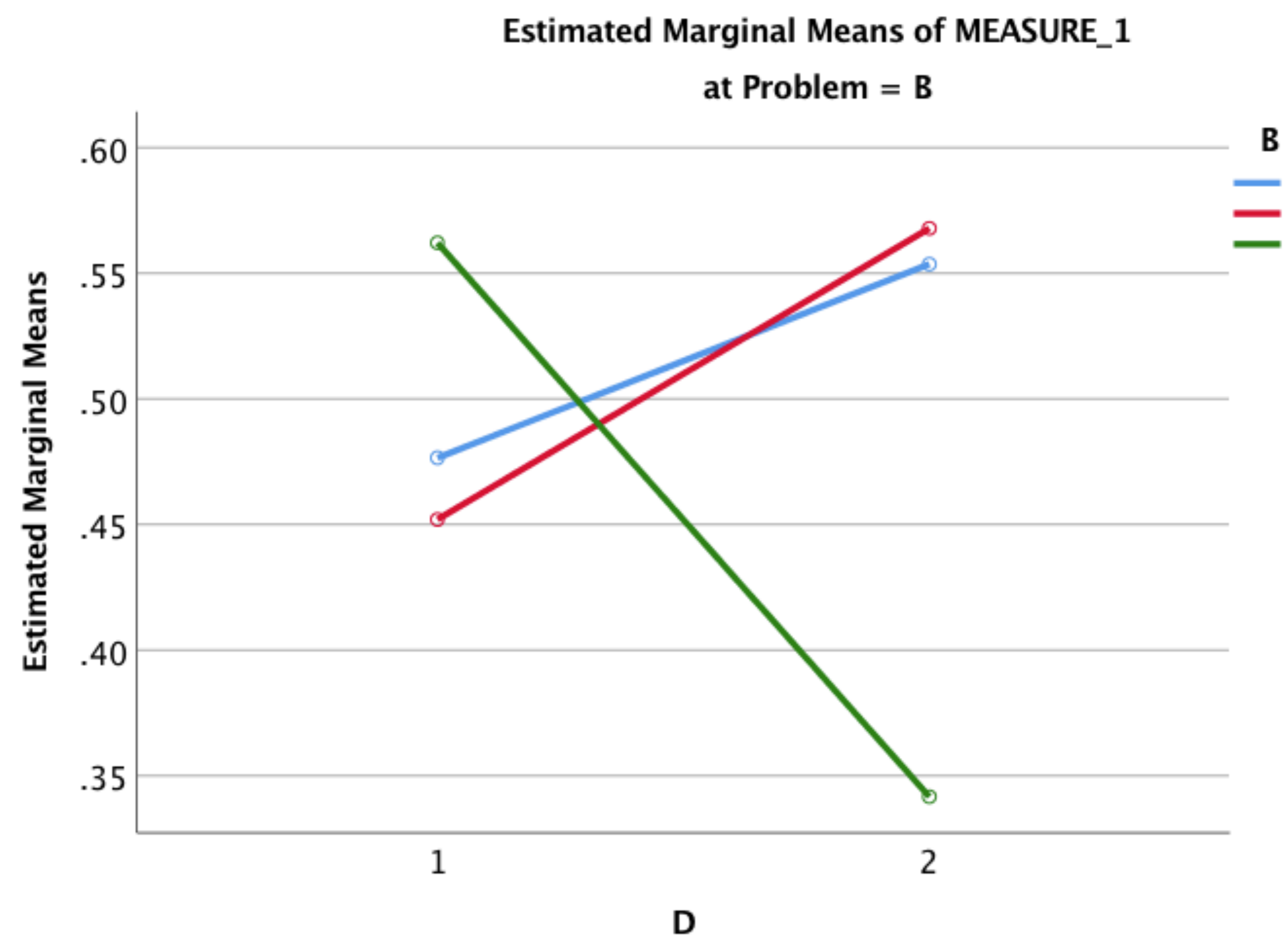
Tests of Between-Subjects Effects

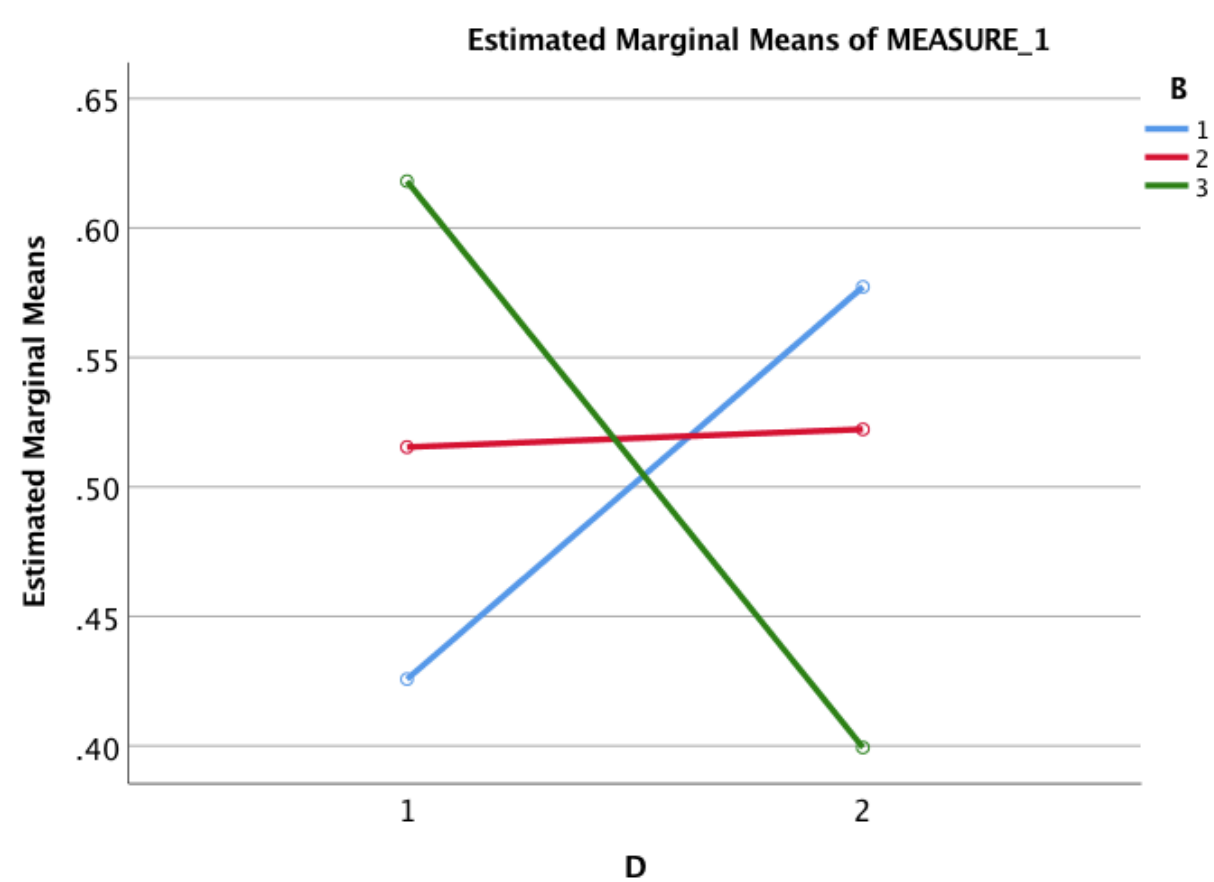
Measure: MEASURE_1

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	p-value Sig.	Partial Eta Squared
Intercept	45.145	1	45.145	541.447	.000	.953
Problem	.052	1	.052	.629	$\boxed{.435}$	.023
Error	2.251	27	.083			


Estimated Marginal Means of MEASURE_1
at Problem = A
Estimated Marginal Means
.70
.60
.50
.40
1
2
D


Estimated Marginal Means of MEASURE_1
at Problem = B
B
1
2
3
Estimated Marginal Means
.60
.55
.50
.45
.40
.35
1
2
D


Estimated Marginal Means of MEASURE_1
B
1
2
3
Estimated Marginal Means
.65
.60
.55
.50
.45
.40
1
2
D

Summary

- Recap of the general idea underlying statistical hypothesis tests.
- What to compare?
  - Two algorithms on a single problem instance.
  - Multiple algorithms on a single problem instance.
  - Two algorithms on multiple problem instances.
  - Multiple algorithms on multiple problem instances.
- How to design the comparisons?
  - Tests for 2 groups.
  - Test for N groups.
  - Groups are the algorithms.
  - Each observation can be an individual run on a given problem instance.
  - Each observations can be an aggregation of multiple runs on a given problem instance.
  - To avoid problems with test assumptions, we can use non-parametric tests.
  - But if we are interested in the interactions among multiple factors, ANOVA can be very useful.
- Commands to run the statistical tests.

Exercise 1

- Download the observations used in this presentation from:
  - www.cs.bham.ac.uk/~minkull/opensource/observations.csv
- Download this presentation from: www.cs.bham.ac.uk/~minkull/publications/presentation-statistical-tests-2.pdf
- Try out all the R commands from the presentation.

Exercise 2

- Pair up with your colleagues and discuss:
  - Research questions that you are currently investigating or about to investigate.
  - Whether you need to use statistical tests to answer these questions.
  - What statistical tests you would use.
- We will wrap up with a general discussion about these.
- Download previous presentation from: www.cs.bham.ac.uk/~minkull/publications/presentation-statistical-tests-1.pdf