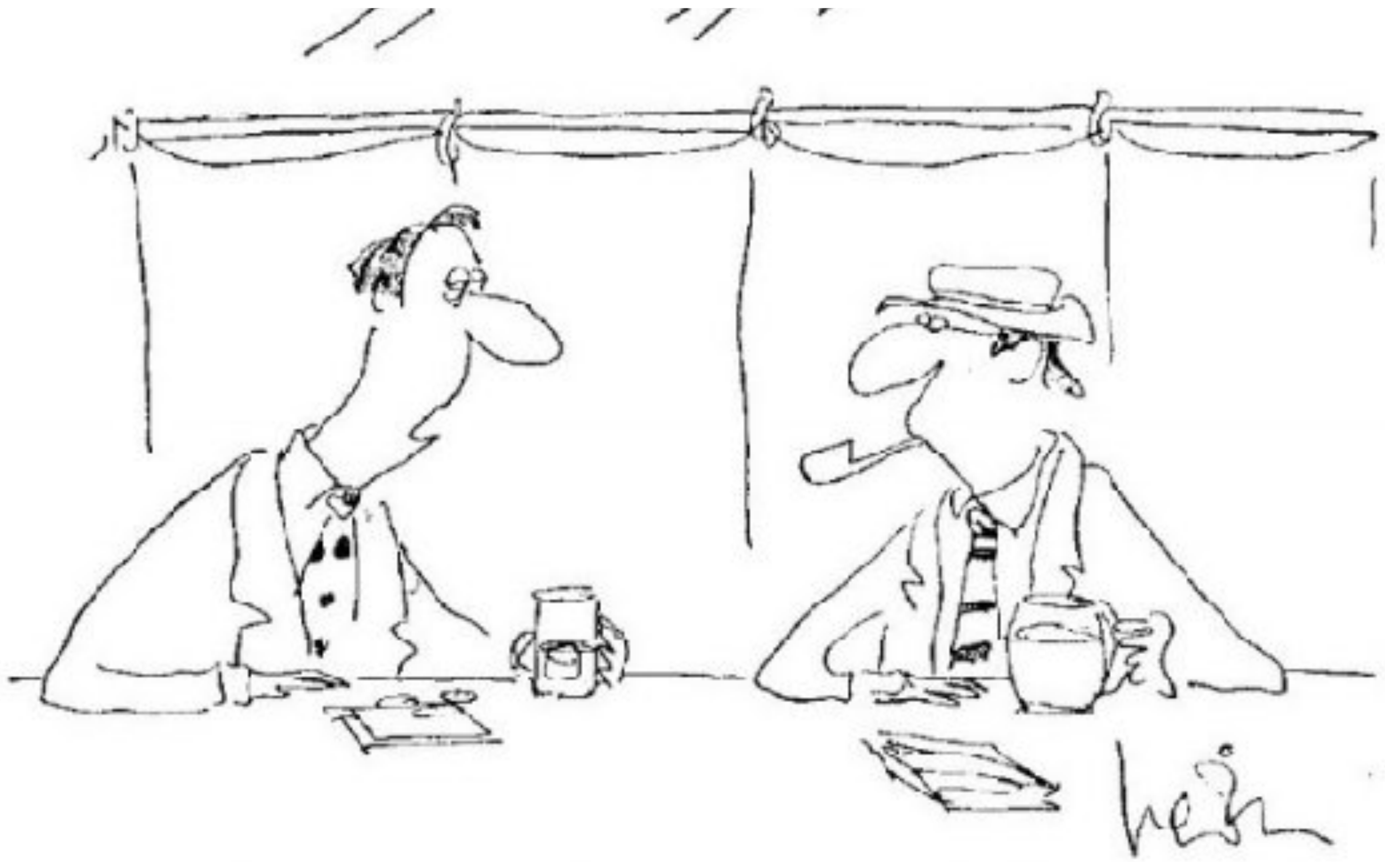
CO3091 - Computational Intelligence and Software Engineering

Lecture 08



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

Image from: http://online-behavior.com/sites/default/files/imagecache/Content/articles/statistical-truth.jpg

Evaluating and Comparing Algorithms - Part II

Leandro L. Minku

Overview

- Recap on the R commands for comparison of 2 groups.
- Comparisons of $N > 2$ groups.

Recap on R Commands for Comparison of Two Groups

Reading Observations

- You can enter observations manually, or you can load observations from a .csv table. E.g.:
    - `observation = read.csv('/Users/llm11/Desktop/observations.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
```

Two-Tailed Wilcoxon Rank-Sum in R

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

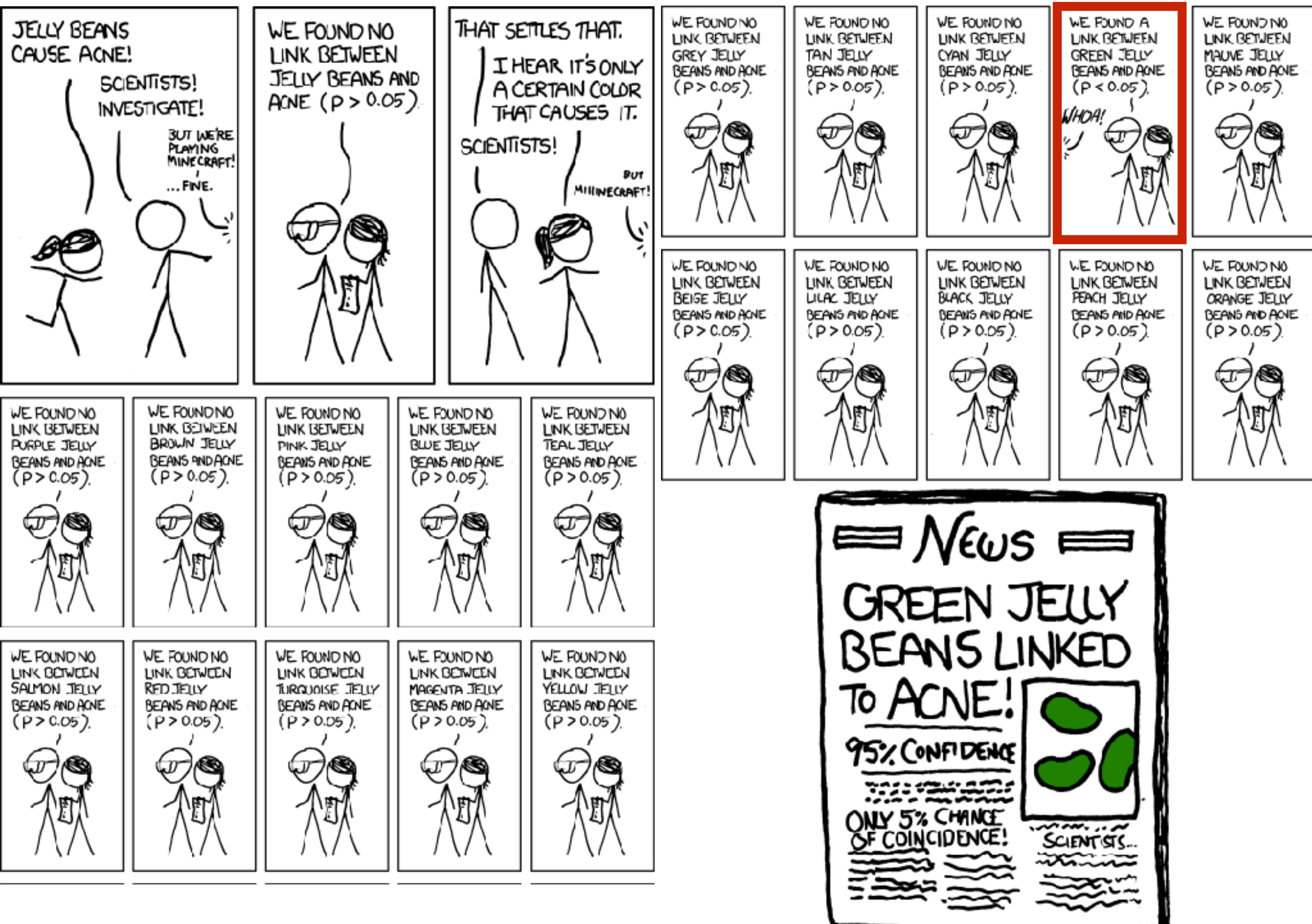
- Example:
  - H0: Group 1 = Group 2
  - H1: Group 1 ≠ Group 2
  - Level of significance = 0.05
  - `wilcox.test(observation[,1],observation[,2],alternative = "two.sided",paired=FALSE, conf.level = 0.95)`
  - p-value: 0.007647 ≤ 0.05 (reject H0)
    - **Groups 1 and 2 are statistically significantly different.**
    - **Median(group 1) = 0.3805, Median(group 2) = 0.7474**

Two-Tailed Wilcoxon Sign Rank in R

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

- Example:
  - H0: Group 1 = Group 2
    H1: Group 1 ≠ Group 2
    Level of significance = 0.05
  - `wilcox.test(observation[,1],observation[,2],alternative = "two.sided",paired=TRUE, conf.level = 0.95)`
  - p-value: 0.002766 ≤ 0.05 (reject H0)
    - **Groups 1 and 2 are statistically significantly different.**
    - **Median(group 1) = 0.3805, Median(group 2) = 0.7474**

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 MIIINECRAFT!
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 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).
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 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).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

The Problem of Multiple Comparisons

- Type 1 error: rejection of H0 when H0 is true, i.e., finding significant difference when there isn't.
- Statistical tests have some probability of presenting a type 1 error. Let's say this probability is 5%.

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-value ≤ adjusted level of significance, reject H0.

- 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-value ≤ adjusted level of significance, reject H0.

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

Dealing with Multiple Comparisons

- Holm-Bonferroni corrections:
  - Consider you have K comparisons.
  - Sort p-values

    (p-value-1 is the largest, p-value-K is the smallest).
  - For i=K to 1:
    - Adjusted level of significance = level of significance / i.
    - Compare p-value-i against adjusted level of significance.
    - If no significant difference is found, stop and make all comparisons from this onwards not significant.

Dealing with Multiple Comparisons

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

i	p-value-i	Adjusted Significance 0.05 / i	Reject H0?
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.

Comparison of N Groups

So far, we talked mainly about pairwise comparisons.

Example:
Compare EA1 vs EA2.

Best Fitness EA1
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

Best Fitness EA2
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

Comparison of N Groups

We may wish to compare
EA1 vs EA2 vs EA3.

or
EA1 vs EA2 vs EA3 vs E4.

or
EA1 vs EA2 vs … vs EN.

Best Fitness EA1	Best Fitness EA2	Best Fitness EA3	Best Fitness EA4
0.8036808732	0.9442552933	0.9442552933	0.9442552933
0.1546026852	0.7277129425	0.7277129425	0.7277129425
0.1507085019	0.4319811615	0.4319811615	0.4319811615
0.9751186599	0.9379836847	0.9379836847	0.9379836847
0.4602321477	0.786503003	0.786503003	0.786503003
0.0132238786	0.8191139316	0.8191139316	0.8191139316
0.0175114877	0.9236880897	0.9236880897	0.9236880897
0.9041741739	0.8155635942	0.8155635942	0.8155635942
0.8697700955	0.7694358404	0.7694358404	0.7694358404
0.676352134	0.3217702059	0.3217702059	0.3217702059
0.5182328166	0.9849161406	0.9849161406	0.9849161406
0.0516411681	0.2586409871	0.2586409871	0.2586409871
0.5426649651	0.7945434749	0.7945434749	0.7945434749
0.4973629257	0.8179485709	0.8179485709	0.8179485709
0.4866079125	0.4132167082	0.4132167082	0.4132167082
0.2187455767	0.5915588229	0.5915588229	0.5915588229
0.8438274211	0.5936746635	0.5936746635	0.5936746635
0.2644009485	0.4386923753	0.4386923753	0.4386923753
0.256434446	0.743990941	0.743990941	0.743990941
0.0791214858	0.7951068189	0.7951068189	0.7951068189
0.2856093827	0.3314508633	0.3314508633	0.3314508633
0.3797759169	0.9218094004	0.9218094004	0.9218094004
0.5978962695	0.7508496968	0.7508496968	0.7508496968
0.0860532501	0.1372954398	0.1372954398	0.1372954398
0.2860286001	0.1251753599	0.1251753599	0.1251753599
0.2772790031	0.7858294814	0.7858294814	0.7858294814
0.7289846656	0.4592977329	0.4592977329	0.4592977329
0.3812438862	0.1583327209	0.1583327209	0.1583327209
0.114495351	0.4037452065	0.4037452065	0.4037452065
0.7128328204	0.8074019616	0.8074019616	0.8074019616

Pairwise Comparisons for N Groups

Potential way to compare EA1 vs EA2 vs EA3:

EA1 vs EA2
EA1 vs EA3
EA2 vs EA3

Best Fitness EA1	Best Fitness EA2	Best Fitness EA3	Best Fitness EA4
0.8036808732	0.9442552933	0.9442552933	0.9442552933
0.1546026852	0.7277129425	0.7277129425	0.7277129425
0.1507085019	0.4319811615	0.4319811615	0.4319811615
0.9751186599	0.9379836847	0.9379836847	0.9379836847
0.4602321477	0.786503003	0.786503003	0.786503003
0.0132238786	0.8191139316	0.8191139316	0.8191139316
0.0175114877	0.9236880897	0.9236880897	0.9236880897
0.9041741739	0.8155635942	0.8155635942	0.8155635942
0.8697700955	0.7694358404	0.7694358404	0.7694358404
0.676352134	0.3217702059	0.3217702059	0.3217702059
0.5182328166	0.9849161406	0.9849161406	0.9849161406
0.0516411681	0.2586409871	0.2586409871	0.2586409871
0.5426649651	0.7945434749	0.7945434749	0.7945434749
0.4973629257	0.8179485709	0.8179485709	0.8179485709
0.4866079125	0.4132167082	0.4132167082	0.4132167082
0.2187455767	0.5915588229	0.5915588229	0.5915588229
0.8438274211	0.5936746635	0.5936746635	0.5936746635
0.2644009485	0.4386923753	0.4386923753	0.4386923753
0.256434446	0.743990941	0.743990941	0.743990941
0.0791214858	0.7951068189	0.7951068189	0.7951068189
0.2856093827	0.3314508633	0.3314508633	0.3314508633
0.3797759169	0.9218094004	0.9218094004	0.9218094004
0.5978962695	0.7508496968	0.7508496968	0.7508496968
0.0860532501	0.1372954398	0.1372954398	0.1372954398
0.2860286001	0.1251753599	0.1251753599	0.1251753599
0.2772790031	0.7858294814	0.7858294814	0.7858294814
0.7289846656	0.4592977329	0.4592977329	0.4592977329
0.3812438862	0.1583327209	0.1583327209	0.1583327209
0.114495351	0.4037452065	0.4037452065	0.4037452065
0.7128328204	0.8074019616	0.8074019616	0.8074019616

Pairwise Comparisons for N Groups

Potential way to compare EA1 vs EA2 vs EA3 vs EA4:

EA1 vs EA2
EA1 vs EA3
EA1 vs EA4
EA2 vs EA3
EA2 vs EA4
EA3 vs EA4

Problem: have to apply corrections for multiple comparisons, which can result in weak tests.

Best Fitness EA1	Best Fitness EA2	Best Fitness EA3	Best Fitness EA4
0.8036808732	0.9442552933	0.9442552933	0.9442552933
0.1546026852	0.7277129425	0.7277129425	0.7277129425
0.1507085019	0.4319811615	0.4319811615	0.4319811615
0.9751186599	0.9379836847	0.9379836847	0.9379836847
0.4602321477	0.786503003	0.786503003	0.786503003
0.0132238786	0.8191139316	0.8191139316	0.8191139316
0.0175114877	0.9236880897	0.9236880897	0.9236880897
0.9041741739	0.8155635942	0.8155635942	0.8155635942
0.8697700955	0.7694358404	0.7694358404	0.7694358404
0.676352134	0.3217702059	0.3217702059	0.3217702059
0.5182328166	0.9849161406	0.9849161406	0.9849161406
0.0516411681	0.2586409871	0.2586409871	0.2586409871
0.5426649651	0.7945434749	0.7945434749	0.7945434749
0.4973629257	0.8179485709	0.8179485709	0.8179485709
0.4866079125	0.4132167082	0.4132167082	0.4132167082
0.2187455767	0.5915588229	0.5915588229	0.5915588229
0.8438274211	0.5936746635	0.5936746635	0.5936746635
0.2644009485	0.4386923753	0.4386923753	0.4386923753
0.256434446	0.743990941	0.743990941	0.743990941
0.0791214858	0.7951068189	0.7951068189	0.7951068189
0.2856093827	0.3314508633	0.3314508633	0.3314508633
0.3797759169	0.9218094004	0.9218094004	0.9218094004
0.5978962695	0.7508496968	0.7508496968	0.7508496968
0.0860532501	0.1372954398	0.1372954398	0.1372954398
0.2860286001	0.1251753599	0.1251753599	0.1251753599
0.2772790031	0.7858294814	0.7858294814	0.7858294814
0.7289846656	0.4592977329	0.4592977329	0.4592977329
0.3812438862	0.1583327209	0.1583327209	0.1583327209
0.114495351	0.4037452065	0.4037452065	0.4037452065
0.7128328204	0.8074019616	0.8074019616	0.8074019616

Statistical Tests For N Groups

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-Wallist test
	Paired (related)	Wilcoxon signed-rank test	Friedman test

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

Statistical Tests For N Groups

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-Wallist test
	Paired (related)	Wilcoxon signed-rank test	Friedman test

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

Statistical Hypotheses

Null Hypothesis

H0: all groups are equal

Alternative Hypothesis

H1: at least one pair of groups is different

Example:

- H0: Fitness(EA1) = Fitness (EA2) = Fitness(EA3)
  H1: !(Fitness(EA1) = Fitness (EA2) = Fitness(EA3))

Kruskal-Wallist Test for Unpaired Comparisons

- R command:
  - `kruskal.test(list_observations)`

    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/llm11/Desktop/observations2.csv')
    ```

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

    ```
    list_observations = list(observations[,1],observations[,2],observations[,3])
    ```

Mathematical Notation

- What exactly a number like 4.802e-11 means?
  - Here, e does not mean the Napier's or Euler's constant 2.71828182845904523536028747135266249775724709…
  - e-11 means 10^{-11}
  - So, 4.802e-11 means $4.802 * 10^{-11}$
  - Sometimes, in computers, 4.802 E-11 is also used.

Friedman Test for Paired Comparisons

- R command:
  - `friedman.test(matrix_observations)`

    matrix_observations contains a matrix 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/llm11/
    Desktop/observations2.csv')
    ```

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

    ```
    matrix_observations =
    data.matrix(observations)
    ```

Post-Hoc Tests

- Problem of tests for N groups:
  - You don't know which of the pairs is different!
- To decide which pair is different, we need to run post-hoc tests.
- Kruskal Wallist's post hoc test is usually Dunn.
- Friedman's post hoc test is usually Nemenyi.

Post-Hoc Tests in R

- You need to install the following package: PMCMR
  - Menu Packages -> setCRAN mirror -> UK (London 2)
  - Menu Packages -> install package -> PMCMR

or

  - install.packages(“PMCMR”)
  - choose UK (London 2) mirror when prompted

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

Dunn Post-Hoc Test

- R command:
  - `posthoc.kruskal.dunn.test(list_observations,p.adjust.method="holm")`
- This test requires corrections to account for multiple comparisons (e.g., holm-bonferroni).

Nemenyi Post-Hoc Test

- R command:
  - `posthoc.friedman.nemenyi.test(matrix_observations)`
- This test already accounts for multiple comparisons. So, no further corrections are needed.

Post-hoc tests are weaker than the N-group tests — it could happen that you run the N-group test and find a significant difference, but find no significant differences in the post-hoc tests.

Summary

- Comparison of 2 groups.
- Performing multiple comparisons.
    - Issues presented by multiple comparisons.
    - Corrections that can be used to account for multiple comparisons.
    - Problem of applying corrections.
- Comparison of *N* groups.
    - Tests for *N* groups are usually stronger than performing pairwise comparisons.
    - Do not tell us which pairs are different.
    - Post-hoc tests can be use for telling which pairs are different, but they are weaker.

Further Reading

- Check the following R help pages:
  - https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kruskal.test.html
  - https://stat.ethz.ch/R-manual/R-devel/library/stats/html/friedman.test.html

- `posthoc.kruskal.dunn.test and posthoc.friedman.nemenyi.test in the following:`
  - https://cran.r-project.org/web/packages/PMCMR/PMCMR.pdf