Operations Research
and Complex Systems
Department of Electrical Engineering
Universidade Federal de Minas Gerais

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Design of Experiments

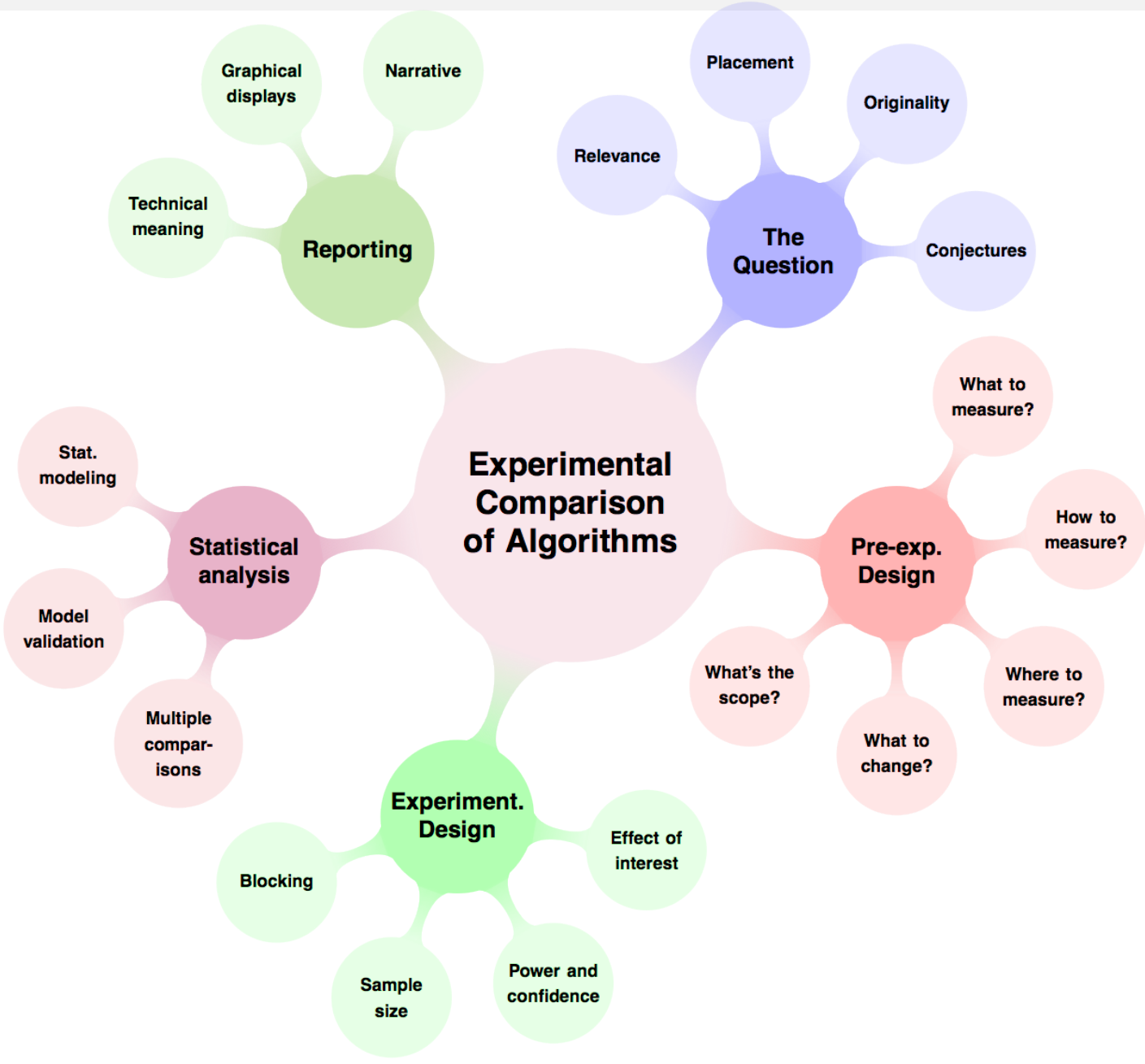
and Experimental Comparison of Evolutionary Algorithms

Prof. Felipe Campelo, Ph.D.

Dept. Electrical Engineering, Universidade Federal de Minas Gerais

Curitiba, Brazil - October 2015

Main Concepts



Motivation

Experimental analysis of algorithms

J.N. Hooker (1994): "*In other words, we should try to build an empirical science of algorithms*";

Theoretical analyses are elegant, but sometimes inadequate, impractical, or unrepresentative;

Experimental investigation can provide valuable information - **as long as it is done properly!**

Ideal situation: theory + experimentation

Motivation

Main question

How much of the observed difference in performance between algorithms is due to actual differences in behavior, and how much is random noise?

Motivation

Careless experimentation

> *"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of."*
>
> – Sir Ronald Fisher

Unfortunately still common in the EA literature, but top journals are increasingly requiring higher methodological standards.

Tends to generate strongly biased results in favor of the "proposed method" (unsurprisingly).

Image: `http://www.nndb.com/people/763/000196175/`

Motivation

Careless experimentation

Proposed method

- Careful implementation and debugging;
- Exhaustive tuning;
- Many experimental runs;

Competing algorithm

- Careless implementation;
- Use of “literature values” for parameters;
- Single experimental run;

Motivation

Careless experimentation

Proposed method

- Careful implementation and debugging;
- Exhaustive tuning;
- Many experimental runs;

Competing algorithm

- Careless implementation;
- Use of “literature values” for parameters;
- Single experimental run;

Description in the manuscript

“*The values presented here are the result of 30 independent runs of the algorithms*”.

Motivation

Some other common issues

Lack of a clear definition of the question one is trying to answer and of the hypotheses one is trying to test;

Inversion of the experimental rationale: desperate search to **demonstrate that my algorithm is the best**, instead of **investigating whether** it really is;

Lack of reproducibility.

Motivation

Other common issues

Lack of a clear definition of the question one is trying to answer and of the hypotheses one is trying to test;

Inversion of the experimental rationale: desperate search to **demonstrate that my algorithm is the best**, instead of **investigating whether** it really is;

Lack of reproducibility.

Image: http://vigilantmeadow.deviantart.com/art/DON-T-PANIC-165415311

Good practices

Important concepts in experimental algorithmics

- Relevant experiments;
- Placement within the literature;
- Adequate test instances;
- Good experimental desing;
- Efficient implementations;
- Reproducibility;
- Comparability;
- Telling the whole story;
- Support for the conclusions;
- Correct presentation of results;

Based on D.S. Johnson: *A Theoretician's Guide to the Experimental Analysis of Algorithms* - `http://plato.asu.edu/ftp/experguide.pdf`

Pre-experimental checklist

Before you even start...

Is the comparison relevant?

Will the (possible) results be of interest to anyone?

Does it have any practical implications?

How does it fit within the literature?

Pre-experimental checklist

Before you even start...

Is the comparison relevant?

Will the (possible) results be of interest to anyone?

Does it have any practical implications?

How does it fit within the literature?

However...

"*Sometimes one should do a completely wild experiment,*
like blowing the trumpet to the tulips every morning
for a month. Probably nothing would happen, but what if it did?"

– Sir George Howard Darwin

Girl playing: `http://www.film.queensu.ca/tulips/default.html`

Pre-experimental design

The quest for *The Question*

> *"- Would you tell me, please, which way I ought to go from here?*
> *- That depends a good deal on where you want to get to, said the Cat.*
> *- I don't much care where - said Alice.*
> *- Then it doesn't matter which way you go, the Cat replied*
>
> Lewis Carroll, **Alice in Wonderland**

The first (and possibly the most important) thing to determine is what exactly your experiment is intended to reveal / discover.

Image adapted from `https://www.pinterest.com/pin/381609768395834338/`

Pre-experimental design

The quest for *The Question*

What is the purpose of your experiment?

> "*I want to prove that my method is good!*"

Lets drop the exclamation mark and refine that a little, shall we?

Pre-experimental design

The quest for *The Question*

What is the purpose of your experiment?

> "*I want to prove that my method is good!*"

Lets drop the exclamation mark and refine that a little, shall we?

> "*I want to **discover if** my method is good.*"

That's much better.

Pre-experimental design

The quest for *The Question*

What is the purpose of your experiment?

> "*I want to prove that my method is good!*"

Lets drop the exclamation mark and refine that a little, shall we?

> "*I want to **discover if** my method is good.*"

That's much better. Now, "good" compared to what?

> "*I want to discover if my method is **better than algorithm A**.*"

We're making progress...

Pre-experimental design

The quest for *The Question*

Lets proceed with our quest for *The Question*: you want to compare your method to another, but on what exactly?

Pre-experimental design

The quest for *The Question*

Lets proceed with our quest for *The Question*: you want to compare your method to another, but on what exactly?

> "*I want to discover if my method is better than algorithm A* ***in terms of solution quality***."

Pre-experimental design

The quest for *The Question*

Lets proceed with our quest for *The Question*: you want to compare your method to another, but on what exactly?

> "*I want to discover if my method is better than algorithm A* ***in terms of solution quality***."

And how are you going to measure solution quality?

Pre-experimental design

The quest for *The Question*

Lets proceed with our quest for *The Question*: you want to compare your method to another, but on what exactly?

> "*I want to discover if my method is better than algorithm A* ***in terms of solution quality***."

And how are you going to measure solution quality?

> "*I want to discover if my method is better than algorithm A in terms of solution quality,* ***measured using indicator*** $\mathcal{F}$."

Pre-experimental design

The quest for *The Question*

Lets keep it up just a little more, we're almost there. When you say "better", what exactly do you mean? Typical, best, worst case?

Pre-experimental design

The quest for *The Question*

Lets keep it up just a little more, we're almost there. When you say "better", what exactly do you mean? Typical, best, worst case?

> "*I want to discover if my method is better than algorithm A in terms of* ***average*** *solution quality, measured using* ***the mean of*** *indicator* $\mathcal{F}$."

Pre-experimental design

The quest for *The Question*

Lets keep it up just a little more, we're almost there. When you say "better", what exactly do you mean? Typical, best, worst case?

> "*I want to discover if my method is better than algorithm A in terms of* ***average*** *solution quality, measured using* ***the mean of*** *indicator* $\mathcal{F}$."

And you want to investigate if your method is good for what?

Pre-experimental design

The quest for *The Question*

Lets keep it up just a little more, we're almost there. When you say "better", what exactly do you mean? Typical, best, worst case?

> "*I want to discover if my method is better than algorithm A in terms of* ***average*** *solution quality, measured using* ***the mean of*** *indicator* $\mathcal{F}$."

And you want to investigate if your method is good for what?

> "*I want to discover if my method is better than algorithm A, in terms of average solution quality (measured using the mean of indicator* $\mathcal{F}$*),* ***for the solution of a class of problems*** $\mathcal{Q}$."

Pre-experimental design

The quest for *The Question*

The process of determining *The Question* is an important one, which is often overlooked in experimental algorithmics.

Besides escaping the mockery of the Cheshire Cat, there are good reasons not to ignore this step:

- *HARKing* tends to greatly increase the rate of false positives in favor of the “proposed approach”;
- Thinking about *The Question* forces the experimenter to consider important aspects of his or her research, such as scope and performance measurement;

Harking = *Hypothesizing After the Results are Known*

Pre-experimental design

Selection of test instances

The name of the game is *Representativeness*!

Benchmark sets can result in *overfitting* of the algorithms to specific instances;

Randomly generated instances may not be representative of real problems;

Arbitrary problem selection can introduce experimenter biases;

Some ideas

- Random sampling of representative instances (*random factor* approach)
- Training and validation (*machine learning* approach)

Pre-experimental design

Non-experimental parameters

How to deal with non-experimental parameters?

- Fixed values? (generalization problem)
- Randomized values? (representativeness problem)
- Literature values? (adequacy problem)

Tuning approach

- Use a portion of the computational budget of the experiment to tune these parameters;
- Balanced effort for *all* algorithms.

Design of Experiments

What is it?

Definition of data collection protocols that enable a correct analysis by means of statistical tools capable of supporting sound and objective conclusions.

Necessary for conclusions to have some quantifiable *meaning*;

Useful for preventing mistakes due to personal biases or other experimental artifacts.

Design of Experiments

Sample size, replication and pseudoreplication - a short detour

Suppose that we want to investigate the question: "*Is the average hair length different between students and professors in the Computational Intelligence field?*"

Lets assume that the audience of this tutorial is a representative sample of our population of interest;

If we take 5 professors and 5 students from the audience and measure 1 hair from each head, what is our sample size?

Example adapted from B. Shipley, *Cause and Correlation in Biology*. Cambridge University Press, 2000.

Cousin Itt: `http://kawiku.deviantart.com/art/Cousin-Itt-334962933`

Design of Experiments

Sample size, replication and pseudoreplication - a short detour

Suppose that we want to investigate the question: "*Is the average hair length different between students and professors in the Computational Intelligence field?*"

Lets assume that the audience of this tutorial is a representative sample of our population of interest;

If we take 5 professors and 5 students from the audience and measure 1 hair from each head, what is our sample size?

What if we take 30 hairs from each head?

Example adapted from B. Shipley, *Cause and Correlation in Biology*. Cambridge University Press, 2000.

Cousin Itt: `http://kawiku.deviantart.com/art/Cousin-Itt-334962933`

Design of Experiments

Sample size, replication and pseudoreplication - a short detour

Sampling more hairs from a given head improves the precision of our estimate for that particular head, but ***it does not increase our effective sample size!***

In this example, performing statistical tests considering the 300 individual measurements (10 heads, 30 hairs per head) as independent values would falsely inflate our degrees of freedom.

This common mistake is called *pseudoreplication*, and results in a much higher rate of false positives in statistical tests.

A simple solution is to use the average hair length per head as the individual data points.

Cousin Itt: `http://kawiku.deviantart.com/art/Cousin-Itt-334962933`

Design of Experiments

Sample size, replication and pseudoreplication - a short detour

The analogy between the hair example and the comparison of EAs is quite straightforward:

	Hair example	**Algorithm comparisons**
Population of interest	People in CI	Problem class $\mathcal{Q}$
Comparison	Student $\times$ Prof.	Alg. A $\times$ Alg. B
Observation unit	Head	Problem instance
Within-unit replicate	Individual hairs	Individual runs

More runs will buy you *some* power by reducing the uncertainties associated with each instance.

For greater power, *more instances* is the way to go.

Check M. Birattari's *On the estimation of the expected performance of a metaheuristic on a class of instances. How many instances, how many runs?* (`http://goo.gl/TVo6YI`) for a deeper discussion.

Design of Experiments

Minimally relevant difference

More instances (and, to a certain extent, more runs per instance) will provide increased power in statistical comparisons.

Just don't overdo it!

To avoid falling victim to *p-hacking*, it is important to define (prior to running the experiment) what is the smallest difference that would have any practical implication.

When analysing and describing the experiment, this practical threshold can provide a much needed reality check for the eager researcher.

---

p-hacking: tweaking your experiment (or torturing your data) until it gives you statistical significance.

Design of Experiments

Statistical power - a(nother) short detour

The earlier discussion gave us some insight on *how to think* about comparative experiments in evolutionary computation;

The power (i.e., sensitivity) of a given experiment to detect a certain difference in performance between two algorithms is a function of some factors:

- *Sample size* ✓

Design of Experiments

Statistical power - a(nother) short detour

The earlier discussion gave us some insight on *how to think* about comparative experiments in evolutionary computation;

The power (i.e., sensitivity) of a given experiment to detect a certain difference in performance between two algorithms is a function of some factors:

- *Sample size* ✓
- Significance level ✓

Design of Experiments

Statistical power - a(nother) short detour

The earlier discussion gave us some insight on *how to think* about comparative experiments in evolutionary computation;

The power (i.e., sensitivity) of a given experiment to detect a certain difference in performance between two algorithms is a function of some factors:

- *Sample size* ✓
- Significance level ✓
- Effect size ✓

Design of Experiments

Statistical power - a(nother) short detour

The earlier discussion gave us some insight on *how to think* about comparative experiments in evolutionary computation;

The power (i.e., sensitivity) of a given experiment to detect a certain difference in performance between two algorithms is a function of some factors:

- *Sample size* ✓
- Significance level ✓
- Effect size ✓
- Residual variance (i.e., unnacounted variability)

Design of Experiments

Statistical power - a(nother) short detour

Suppose that you want to investigate the mean efficiency of different fuel mixtures (in terms of km/$) for a fleet of vehicles.

Between-vehicle variation is a potentially large source of variability (possibly larger than the differences due to different fuels);

This variability can be modeled by considering each vehicle as an *experimental unit*, and isolating its effects in the statistical model.

For known and controllable sources of spurious variation (such as the vehicles in this example) the technique used to model this variation out of our inference is called *blocking*.

Image: `https://financialsensei.wordpress.com/tag/save-money-on-gas/`

Design of Experiments

Blocking

Blocking is the principle behind well-known techniques for comparing algorithms on multiple instances, such as *Friedman's test* and *Wilcoxon-Mann-Whitney test*;

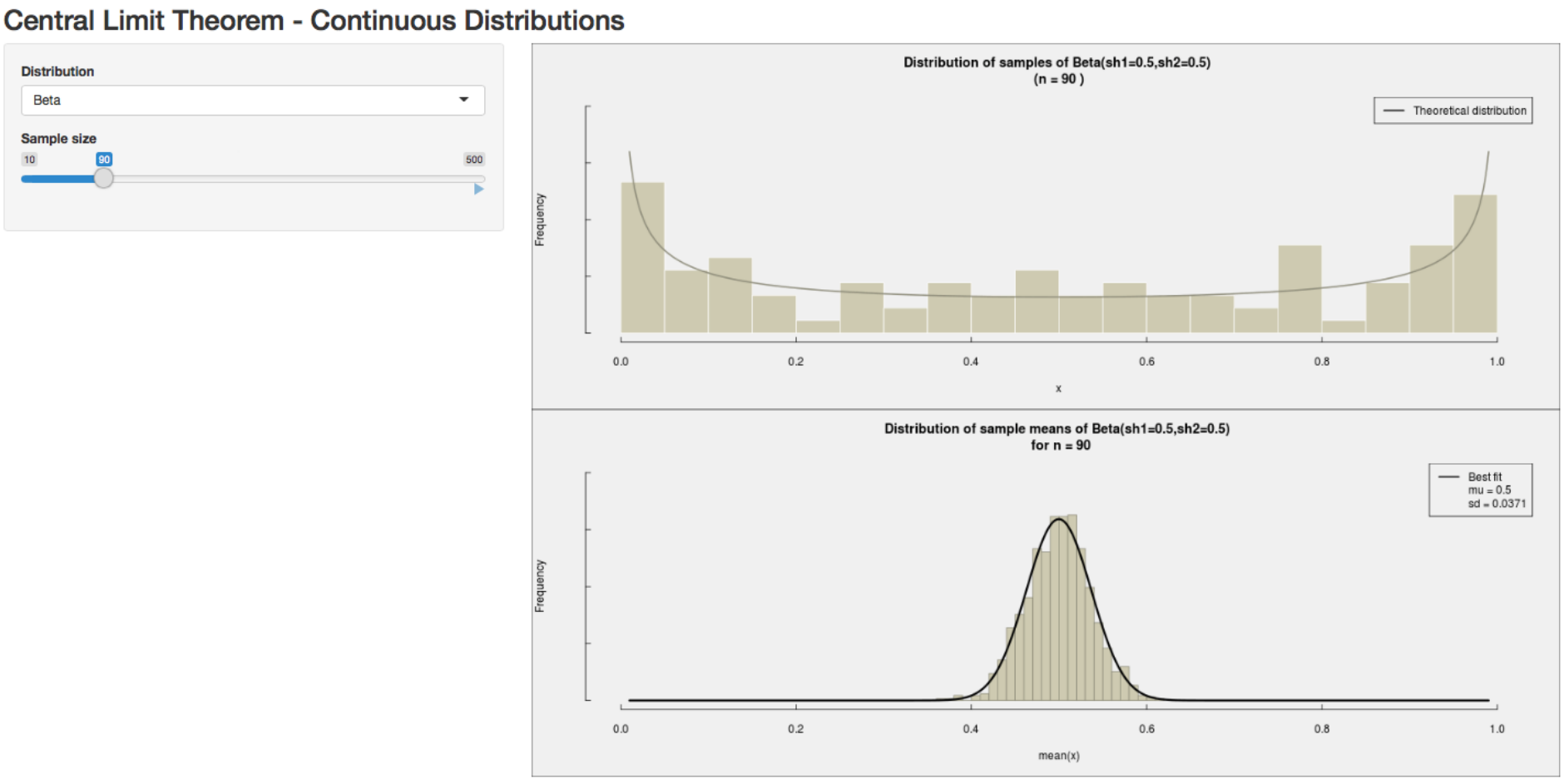
However, the usually shunned parametric counterparts (*blocked ANOVA* and *paired t-tests*) also deserve some attention;

The assumption of normal *sampling distribution of the means* is generally well covered by the Central Limit Theorem even for modest sample sizes (important exceptions: ***truncated observations***, ***extreme skewness***, and ***outliers***).

Parametric methods usually present larger power (i.e., greater sensitivity) and are simpler to understand.

Design of Experiments

Central Limit Theorem



Interactive demo: `http://drwho.cpdee.ufmg.br:3838/CLT/`

Source code: `http://git.io/vnPj8`

Statistical modeling and inference

A consequence of design

If the experiment is well designed, its planning essentially determines the statistical model to be used (at least qualitatively);

The analysis techniques are usually simple (but the devil is in the details);

Use of existing tools and techniques;

Inference on the *statistical significance* of the results;

Statistical modeling and inference

A consequence of design

If the experiment is well designed, its planning essentially determines the statistical model to be used (at least qualitatively);

The analysis techniques are usually simple (but the devil is in the details);

Use of existing tools and techniques;

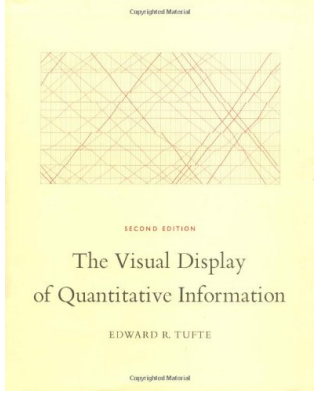
Inference on the *statistical significance* of the results;

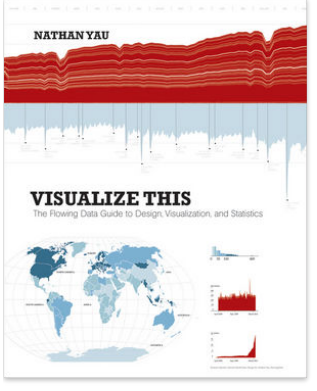
Reporting of results

Presentation

Combine textual, numeric and graphical elements to tell a story with your data. It simplifies the understanding and analysis of the results.

- Strive to achieve graphical excellence;
- Coherence of notation - special attention to figures and tables;
- Display simultaneous confidence intervals and other graphical indicators of effect size.





Other great resources on graphical excellence:

Flowing Data (`http://flowingdata.com/`)

Information is Beautiful (`http://www.informationisbeautiful.net`)

Reporting of results

Tell the whole story

Avoid *cherrypicking* your results;

Report and describe anomalous results and outliers (even if they are discarded in the modeling phase);

Exercise *extreme* caution when discarding outliers!

Detail stop criteria, computational cost, and any other relevant information for the understanding and reproducibility of your results.

Whenever possible, share your code! It's good for the field as a whole (for reproducibility), and it is good for your paper (for citations)!

R.D. Peng, *Reproducible Research in Computational Science*, Science 334(6060):1226-1227, 2011

P. Vandewalle, *Code Sharing Is Associated with Research Impact in Image Processing*, Computer Science and Engineering 14(4):42-47, 2012

Conclusions

Drawing and reporting conclusions

Conclusions should be based on solid evidence from the data;

Be conservative - don't exaggerate the generality of the results;

Report significance levels, effect sizes, and the assumptions under which the results are valid;

Suggest explanations to the observed results;

Be careful with *anomaly hunting*;

> "*Always let the science drive the statistics. If you get a statistically significant result, go back and describe what it means in the scientific context.*'"
>
> – Aaron Rendahl

Want more?

Some shameless self-promotion

If you liked this short tutorial, check the online materials available at:

`https://github.com/fcampelo/Design-and-Analysis-of-Experiments`

It's free, and I think you'll like it too!

Questions?

Images: http://kawiku.deviantart.com/art/Cousin-Itt-334962933

https://commons.wikimedia.org/wiki/File:Question-mark-blackandwhite.png

About this material

Conditions of use and referencing

Please reference this work as:

Felipe Campelo, *Design of Experiments and Statistical Comparison of Evolutionary Agorithms*. Online: http://git.io/vZph7, Latin American School on Computational Intelligence, October 13, 2015, Curitiba, Brazil; Creative Commons BY-NC-SA 4.0.

```
@Misc{Campelo2015-LASCI,
  title={Design of Experiments and Statistical Comparison of Evolutionary Agorithms},
  author={Felipe Campelo},
  howPublished={http://git.io/vZph7},
  year={2015},
  month={October 13},
  note={Latin American School on Computational Intelligence,
        Curitiba, Brazil; Creative Commons BY-NC-SA 4.0.}}
```