00 - General Info

Monday, March 13, 2017 1:12 PM

- COURSERA: - R Programming
  - Reproducible Research

↳ possível fazer gratuitamente COM CERTIFICADO fazendo requisitos ("duvidoso")

- MATERIALS: github.com/fcampelo → Repo: design-and-analysis...

  ↳ Recommended readings after each class

  ↳ Book: - D.C. Montgomery, G.C. Runger — Applied Statistics and Probability for Engineers

  - M. J. Crawley — The R Book

  Rand Wilcox — Fundamentals of Modern Statistical Techniques

- EVALUATION

  - "Turning Student Groups into Effective Teams", Barbara Oakley

  - Case Studies (4) and Final Project IN GROUPS

  - Simple Test

  - Final Test AND/OR Seminars

- Mantra base: "Antes de tudo, é necessário entender muito bem da área na qual fará a análise estatística"

Course Plan

Tuesday, April 11, 2017 3:14 PM

Universidade Federal de Minas Gerais
Escola de Engenharia

EEE933 - TÓPICOS ESPECIAIS EM SISTEMAS DE COMPUTAÇÃO E TELECOMUNICAÇÕES
PLANEJAMENTO E ANÁLISE DE EXPERIMENTOS

---

1. Dados Gerais

Horário: Segunda-feira, 13:00 - 16:35
Carga Horária: 60 horas-aula (4 créditos)
Professor: Felipe Campelo, Departamento de Engenharia Elétrica, sala 2225
E-mail: fcampelo@ufmg.br

2. Ementa

Introdução ao método científico; o papel da experimentação na ciência; conceitos estatísticos; princípios de planejamento de experimentos; experimentos comparativos simples, inferência estatística e teste de hipóteses; experimentos de fator único, análise de variância, modelos fatoriais; cálculo de tamanho amostral; pseudoreplicação.

3. Proposta do Curso

Este curso tem como objetivo a capacitação dos alunos para o planejamento experimental, definição e teste de hipóteses, e análise estatística dos dados obtidos em suas respectivas áreas de atuação. Espera-se que os alunos tragam para a sala de aula problemas relacionados a suas áreas de atuação, os quais serão estudados ao longo do semestre a partir da elaboração e teste de hipóteses de trabalho. Ao final do curso, espera-se que o estudante tenha obtido conhecimento suficiente para a realização de experimentos planejados e análise estatística dos dados relativos à sua tese, dissertação ou trabalho final de curso.

4. Critérios de Avaliação

1) Estudos de Caso: 35 pontos
2) Prova: 25 pontos
3) Seminario: 15 pontos
4) Projeto Final: 25 pontos

5. Política de frequência

1. A aferição de frequência será realizada por meio de lista de presença, com eventual verificação por chamada oral. A lista será circulada em sala no máximo 20 minutos após o início da aula, e recolhida logo a seguir.
2. Alunos que não alcançarem 70 pontos terão a frequência computada, e caso seja atestada a ausência em mais de 25% do curso receberão reprovação por frequência.

É importante ressaltar que eu considero a presença nas aulas **muito** importante – muitos conceitos são apresentados, discutidos e esclarecidos em sala de aula, e a ausência pode prejudicar fortemente seu progresso e aprendizado. Contudo, é facultado ao aluno frequentar ou não as aulas, caso considere desnecessário (e assumindo que isso tenha sido previamente acertado com sua equipe).

Um último aviso (que espero ser desnecessário): tentativas de burlar o sistema de aferição de presença (por exemplo, pedindo a um colega para assinar em seu lugar) serão tratadas no maior rigor possível (reprovação por frequência e comunicação formal ao colegiado). A lista de presença é um documento público, e sua falsificação consiste de crime de falsidade ideológica (além de ser uma quebra da relação de confiança que deve nortear o relacionamento em sala de aula).

6. Bibliografia

Principais

[1] Felipe Campelo (2015), Lecture Notes on Design and Analysis of Experiments. Online: **http://git.io/v3Kh8** Version 2.11; Creative Commons BY-NC-SA 4.0.
[2] Michael J. Crawley, "The R Book", 1st ed., Wiley, 2007.
[3] Material dos cursos de *Data Science* da John Hopkins University - *https://github.com/rdpeng/courses*
[4] D.C. Montgomery, G.C. Runger, "Applied Statistics and Probability for Engineers", 4th ed., John Wiley & Sons Wiley, 2006.

Universidade Federal de Minas Gerais
Escola de Engenharia

EEE933 - TÓPICOS ESPECIAIS EM SISTEMAS DE COMPUTAÇÃO E TELECOMUNICAÇÕES
PLANEJAMENTO E ANÁLISE DE EXPERIMENTOS

Adicionais

[1] D.C. Montgomery, "Design and Analysis of Experiments", 6th ed., John Wiley & Sons, 2005.
[2] R.L. Mason, R.F. Gunst, J.L. Hess, "Statistical Design and Analysis of Experiments, With Applications to Engineering and Science", John Wiley & Sons, 2003.
[3] B.S. Everitt, T. Hothorn, "A Handbook of Statistical Analyses Using R", 1st ed., Chapman & Hall/CRC, 2006.
[4] R.E. Walpole, R.H. Myers, S.L. Myers, K. Ye, "Probabilidade e Estatística para Engenharia e Ciências", 8ª. Ed., Pearson, 2009.
[5] P. Murrell, "R Graphics", 1st. Ed., Chapman & Hall/CRC, 2006.
[6] D.J. Sheskin, "Handbook of Parametric and Nonparametric Statistical Procedures", 5th ed., Chapman & Hall/CRC, 2011.
[10] J.G.C. Da Silva, "Estatística Experimental: Planejamento de Experimentos" – **http://goo.gl/p8UJvZ**
[11] J.J. Faraway, "Practical Regression and Anova using R", 2002 – **http://goo.gl/ewMWL**
[12] W. Chang, "R Graphics Cookbook", O'Reilly 2013.

6. Programação Prevista (sujeita a alterações)

Dia	Assunto
13.3	Introdução ao curso / O método científico
20.3	O método científico
27.3	Revisão: conceitos estatísticos básicos
03.4	Inferência para uma amostra
10.4	Inferência para duas amostras
17.4	Inferência para duas amostras
24.4	Inferência para duas amostras
08.5	Inferência para múltiplas amostras
15.5	Inferência para múltiplas amostras
22.5	Esclarecimento de dúvidas / reservado para estudo.
29.5	Prova
05.6	Seminarios
12.6	Seminarios
19.6	Preparação dos trabalhos finais
26.6	Preparação dos trabalhos finais
03.7	Apresentações dos trabalhos finais

→ Dois horários: · 12:00 - 13:40 · 13:00 - 14:00 — chegar pontualmente!

· 4/5 questões
· Metade conceitual, metade interpretação
↳ $\alpha, \beta, \pi, \ldots$
↳ p-value...
↳ t-value plot...
↳ avaliação de casos
↳ planejamento de experimento

↳ TRABALHO FINAL
· Planejamento, execução e análise de experimentos

Team Policies

Tuesday, April 11, 2017 3:15 PM

ENG Departamento de Engenharia Elétrica UFMG

EEE933 - Planejamento e Análise de Experimentos
Prof. Felipe Campelo

Declaração de Políticas de Equipe[1]

Cada equipe será incumbida de determinadas responsabilidades na condução e realização de problemas e projetos. O presente documento apresenta as políticas de equipe que devem ser seguidas no âmbito da disciplina.

- Para cada trabalho, determinem um membro como *Coordenador*, um como *Relator*, e um como *Verificador* do trabalho. Em times compostos por quatro membros, determinem também um membro como *Monitor*. Estes papéis devem ser alternados a cada trabalho;
- Entrem em acordo sobre um horário comum para reuniões, e deixem claro as tarefas que cada um deve realizar até a reunião (leituras, trabalho preliminar em algum aspecto técnico, etc.). Os membros da equipe devem completar seus preparativos individuais antes de cada reunião;
- O *Coordenador* deve entrar em contato com os demais membros antes de cada reunião, se certificando de que todos estejam cientes do local e horário do encontro, bem como das tarefas alocadas a cada um;
- ***Reunião e trabalho***:
  - O papel do *Coordenador* é manter o grupo focado, e se certificar de que todos estão envolvidos no trabalho;
  - O *Relator* deve trabalhar no preparo da versão final do trabalho a ser entregue;
  - O *Verificador* realiza a verificação final do trabalho antes que o mesmo seja entregue;
  - O *Monitor* é responsável por se certificar que todos entendem tanto a solução encontrada quanto a estratégia utilizada para encontrá-la. Em equipes compostas por três membros os papéis de *Monitor* e *Verificador* devem ser desempenhados pelo mesmo membro;

  Ao final de cada encontro o grupo deve agendar a próxima reunião (data/local) e os papéis que cada um irá desempenhar no trabalho seguinte;
- O *Verificador* é responsável pela entrega da versão final do trabalho, com os nomes de todos os membros *que participaram ativamente no mesmo*. Se o *Verificador* tiver problemas de agenda e não puder comparecer à aula no dia e horário de entrega, esta tarefa pode ser delegada a outro membro da equipe.
- É importante revisar as atividades corrigidas, e se certificar de que cada um entenda por quê pontos foram perdidos, e como corrigir estes erros.

[1] Adaptado de B. Oakley *et al.*, "Turning Student Groups into Effective Teams", 2004.

- É importante revisar as atividades corrigidas, e se certificar de que cada um entenda por quê pontos foram perdidos, e como corrigir estes erros.

[1] Adaptado de B. Oakley *et al.*, "Turning Student Groups into Effective Teams", 2004.

- *Entrem em contato com o professor caso haja algum conflito que não possa ser resolvido pela própria equipe;*
- **Como lidar com membros não-cooperativos**:
  - Se algum membro da equipe se recusa a cooperar em suas tarefas (por qualquer que seja o motivo), seu nome não deve ser incluído no trabalho entregue;
  - Se o problema persistir, a equipe deve agendar um encontro com o professor de forma a tentar resolver o problema;
  - Caso não seja possível encontrar uma boa solução e o problema ainda persistir, o restante da equipe deve notificar o membro não-cooperativo (por escrito, com cópia para o professor) que ele está sob risco de ser excluído do grupo.
  - Caso as tentativas anteriores tenham sido em vão, o grupo deve então notificar o membro não-cooperativo (por escrito, com cópia para o professor) que o mesmo está excluído do grupo;
  - Alunos excluídos de alguma equipe possuem duas alternativas:
    - Encontrar um grupo que conte com somente três membros e que esteja disposto a incorporá-lo na equipe (neste caso os três membros da equipe devem se manifestar por escrito ao professor);
    - Realizar os trabalhos restantes de forma individual.
  - Da mesma forma, se algum dos membros da equipe julgar que está fazendo todo o trabalho do grupo, o mesmo deve comunicar por escrito ao grupo que pretende se desligar do grupo caso não haja uma maior cooperação, e um segundo comunicado por escrito declarando seu desligamento da equipe caso o primeiro não resulte em uma maior participação dos demais membros. Assim como no caso anterior, todos os comunicados devem ser feitos por escrito com cópia para o professor. As opções para o membro da equipe que se desligar de um grupo são as mesmas detalhadas anteriormente: encontrar um grupo que o receba ou realizar as tarefas individualmente.

dos os comunicados devem ser feitos por escrito com cópia para o professor. As opções para o membro da equipe que se desligar de um grupo são as mesmas detalhadas anteriormente: encontrar um grupo que o receba ou realizar as tarefas individualmente.

Trabalhar em equipes nem sempre é uma tarefa simples – membros da equipe por vezes tem outras responsabilidades e não conseguem preparar suas parcelas do trabalho ou participar das reuniões, e conflitos resultantes de níveis distintos de habilidade ou compreensão do conteúdo podem ocorrer, bem como de aspectos ligados à ética de trabalho. Entretanto, em equipes que trabalham e se comunicam bem os benefícios superam em muito as dificuldades.

Uma das formas de melhorar as chances de uma dada equipe funcionar bem consiste em determinar conjuntamente, e antes do início das atividades, quais são as expectativas coletivas dos membros da equipe em relação aos colegas. E este é o objetivo da primeira tarefa de cada equipe, a saber, a definição de um *Acordo de Expectativas da Equipe*.

01 - What Is Science

Monday, March 13, 2017 1:02 PM

· Thomas Khun - Estrutura das Revoluções Científicas , Popper, ??
↳ início séc XX

· ~~Prova~~ → Margem de Plausibilidade

* "Terra é uma esfera"

· PERIGO: Viés de Confirmação → "investigar se está correto" ✓
vs.
"provar que está correto" X

↳ Thinking Fast and Slow - Daniel Kahneman

↳ Metodologia de Planejamento de Experimentos:
salvaguarda contra viés

* CRISPR - CAS9
↳ tecnologia de edição genética

Slides

Monday, March 13, 2017 6:02 PM

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Design and Analysis of Experiments

01 - What is Science

Version 2.11

Felipe Campelo
http://www.cpdee.ufmg.br/~fcampelo

Graduate Program in Electrical Engineering

Belo Horizonte
March 2015

"*The most that can be expected from any model is that it can supply a useful approximation to reality. All models are wrong; some models are useful*"

George E.P. Box (1919 – 2013)
British statistician

Image: https://asq.org/about-asq/honorary-members/box

What is science?

Some common misconceptions

- Science is a collection of facts; ×
- Science is the creation of new gadgets; ×
- Scientific ideas are absolute and unchangeable; ×
- Scientific ideas are subject to change, therefore unreliable; ×
- Observations give answers directly to the scientists; ×
- Science **proves** stuff; ×
- Science can only **disprove** stuff; ×
- The scientist works to **show** that his/her theory is right; ×

Essential reading: Common Misconceptions About Science: `http://goo.gl/TN7k9B`
Image: `http://xkcd.com`

What is science?

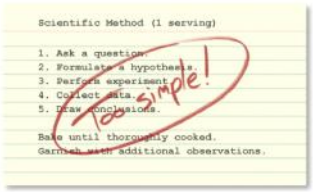
A good operational definition

> "*What do you think science is?*
> *There's nothing magical about science.*
> *It is simply a systematic way for carefully*
> *and thoroughly observing nature and*
> *using consistent logic to evaluate results.*"
> – Steven P. Novella

Image: `http://www.relativelyinteresting.com/definition-science-steven-novella/`

What is science?
The scientific process

- Normally shown as a flowchart or a sequence of steps;
- Oversimplification of a complex and iterative process;
- Suggests an “end” to the process.



- Actually includes:
  - Several activities, performed at different stages;
  - Interaction with the scientific community;
  - Creative, “outside the box” thinking;
  - Preliminary conclusions, subject to revision as new and better data become available;
  - Learning from failures as much as from successes.

Image: http://goo.gl/7cCGaz - (c) Understanding Science, 2015. Used with permission.

What is science?
The scientific process

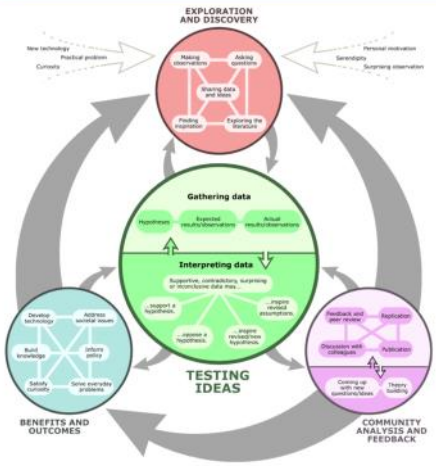
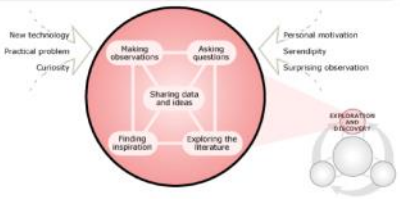


Image: http://goo.gl/VglXc5 - (c) Understanding Science, 2015. Used with permission.

What is science?
The scientific process

"*Dans les champs de l'observation le hasard ne favorise que les esprits préparés.*" – **Louis Pasteur** (Univ. Lille, France, 1854).

- Observations → **questions**;
- Exploratory experimentation;
- Preparation + serendipity.



Benzene (1865)

Kekule

Radioactivity (1896)

Becquerel

Penicillin (1928)

Fleming

Top image: http://goo.gl/fy8Glh - (c) Understanding Science, 2015. Used with permission.
Scientists: http://goo.gl/SG6sgp | http://goo.gl/rhLC9C | http://goo.gl/CFj8M1

What is science?
The scientific process

- Drawing and testing hypotheses;
- Comparing alternative explanations;
- Accepting / rejecting ideas based on **evidence**;
- **Predictions** *versus* **observation**: corroboration or refutation?

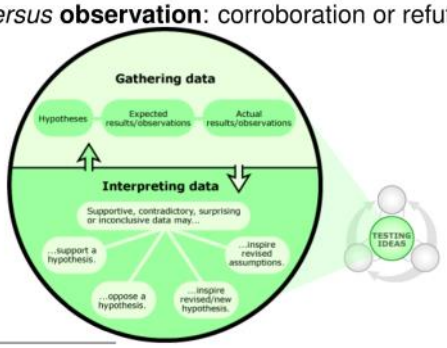


Image: http://goo.gl/aOgSqT - (c) Understanding Science, 2015. Used with permission.

What is science?
The scientific process

James Lind (1747):

- Observation: scurvy in sailors;
- Conjecture: Caused by the body rottenning;
- Idea: attempt to avoid/reverse effects with acidic substances;

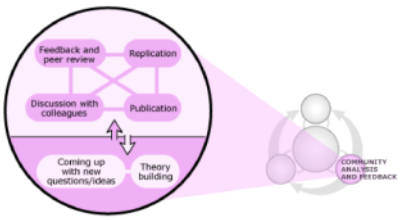
Separation of a group of 12 affected sailors in six groups with identical diets, except for the addition of a supplement:

Group 1 Cider.	Group 2 Vitriol.	Group 3 Vinegar.
Group 4 Sea water.	Group 5 Oranges and lemons.	Group 6 Tea.

Image: http://commons.wikimedia.org/wiki/File:James_Lind_by_Chalmers.jpg

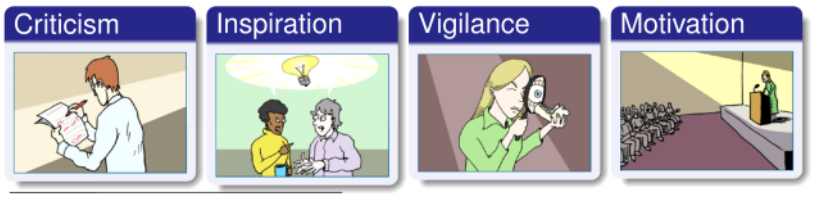
What is science?
The scientific process

Interaction with the scientific community is **fundamental**:



- Colleagues;
- Collaborators;
- Reviewers;
- Rivals;

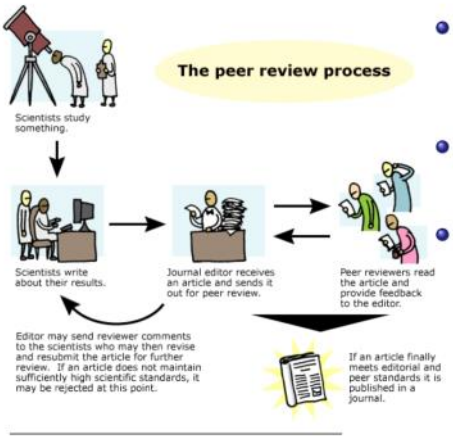
This interaction plays essential roles for the progress of research:



All images: http://goo.gl/9pSCTG - (c) Understanding Science, 2015. Used with permission.

What is science?
The scientific process

Publication and peer review.



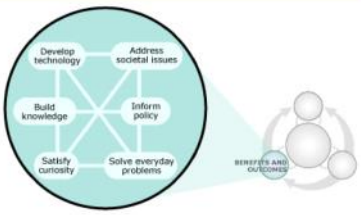
- Additionally, *post-publication review* by the wider scientific community;
- **Replication** and verification of results;
- **Reproducibility** is essential.

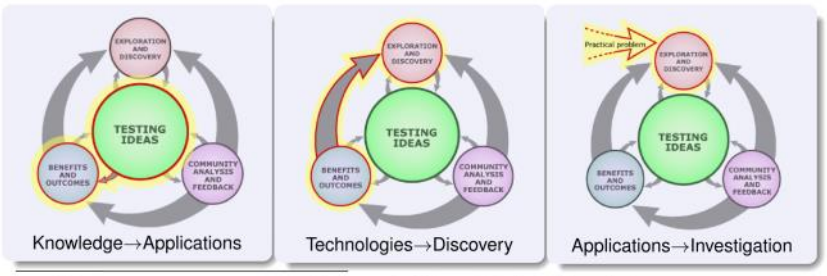
Image: http://goo.gl/VWCVkK - (c) Understanding Science, 2015. Used with permission.

What is science?
The scientific process

The scientific process is a way of building knowledge:

- Generate and test new ideas about how the world works;
- Iteratively increasing the reliability of the knowledge;





All images: http://goo.gl/IBRSoQ - (c) Understanding Science, 2015. Used with permission.

What is science?

To wrap it up

"*It is important to be literate in the scientific method, not only for the sake of your own research. We are also agents of change in the population and, as such, we need to be aware of good and bad science, and able to point the difference to the society.*"

– Claus C. Aranha

Image: http://lattes.cnpq.br/2897895256340893

Bibliography

Required reading

1. *Understanding Science*. 2014. University of California Museum of Paleontology. 3 January 2014. - http://www.understandingscience.org
2. F.L.H. Wolfs, *APPENDIX E: Introduction to the Scientific Method.* - http://goo.gl/osGpU

Recommended reading

1. Carl Sagan, *The demon-haunted world: science as a candle in the dark*, Random House, 1996.
2. The Skeptics Guide to the Universe. - http://www.theskepticsguide.org

http://undsci.berkeley.edu/

About this material

Conditions of use and referencing

Please reference this work as:

Felipe Campelo (2015), *Lecture Notes on Design and Analysis of Experiments*.
Online: https://github.com/fcampelo/Design-and-Analysis-of-Experiments
Version 2.11, Chapter 1; Creative Commons BY-NC-SA 4.0.

```
@Misc{Campelo2015-01,
  title={Lecture Notes on Design and Analysis of Experiments},
  author={Felipe Campelo},
  howPublished={\url{https://github.com/fcampelo/Design-and-Analysis-of-Experiments}},
  year={2015},
  note={Version 2.11, Chapter 1; Creative Commons BY-NC-SA 4.0.},
}
```

Introduction to Scientific Method

Monday, March 20, 2017 12:46 PM

APPENDIX E: Introduction to the Scientific Method

Introduction to the Scientific Method

The scientific method is the process by which scientists, collectively and over time, endeavor to construct an accurate (that is, reliable, consistent and non-arbitrary) representation of the world.

Recognizing that personal and cultural beliefs influence both our perceptions and our interpretations of natural phenomena, we aim through the use of standard procedures and criteria to minimize those influences when developing a theory. As a famous scientist once said, "Smart people (like smart lawyers) can come up with very good explanations for mistaken points of view." In summary, the scientific method attempts to minimize the influence of bias or prejudice in the experimenter when testing an hypothesis or a theory.

I. The scientific method has four steps

1. Observation and description of a phenomenon or group of phenomena.

2. Formulation of an hypothesis to explain the phenomena. In physics, the hypothesis often takes the form of a causal mechanism or a mathematical relation.

3. Use of the hypothesis to predict the existence of other phenomena, or to predict quantitatively the results of new observations.

4. Performance of experimental tests of the predictions by several independent experimenters and properly performed experiments.

If the experiments bear out the hypothesis it may come to be regarded as a theory or law of nature (more on the concepts of hypothesis, model, theory and law below). If the experiments do not bear out the hypothesis, it must be rejected or modified. What is key in the description of the scientific method just given is the predictive power (the ability to get more out of the theory than you put in; see Barrow, 1991) of the hypothesis or theory, as tested by experiment. It is often said in science that theories can never be proved, only disproved. There is always the possibility that a new observation or a new experiment will conflict with a long-standing theory.

II. Testing hypotheses

As just stated, experimental tests may lead either to the confirmation of the hypothesis, or to the ruling out of the hypothesis. The scientific method requires that an hypothesis be ruled out or modified if its predictions are clearly and repeatedly incompatible with experimental tests. Further, no matter how elegant a theory is, its predictions must agree with experimental results if we are to believe that it is a valid description of nature. In physics, as in every experimental science, "experiment is supreme" and experimental verification of hypothetical predictions is absolutely necessary. Experiments may test the theory directly (for example, the observation of a new particle) or may test for consequences derived from the theory using mathematics and logic (the rate of a

radioactive decay process requiring the existence of the new particle). Note that the necessity of experiment also implies that a theory must be testable. Theories which cannot be tested, because, for instance, they have no observable ramifications (such as, a particle whose characteristics make it unobservable), do not qualify as scientific theories.

If the predictions of a long-standing theory are found to be in disagreement with new experimental results, the theory may be discarded as a description of reality, but it may continue to be applicable within a limited range of measurable parameters. For example, the laws of classical mechanics (Newton's Laws) are valid only when the velocities of interest are much smaller than the speed of light (that is, in algebraic form, when $v/c << 1$). Since this is the domain of a large portion of human experience, the laws of classical mechanics are widely, usefully and correctly applied in a large range of technological and scientific problems. Yet in nature we observe a domain in which v/c is not small. The motions of objects in this domain, as well as motion in the "classical" domain, are accurately described through the equations of Einstein's theory of relativity. We believe, due to experimental tests, that relativistic theory provides a more general, and therefore more accurate, description of the principles governing our universe, than the earlier "classical" theory. Further, we find that the relativistic equations reduce to the classical equations in the limit $v/c << 1$. Similarly, classical physics is valid only at distances much larger than atomic scales ($x >> 10^{-8}$ m). A description which is valid at all length scales is given by the equations of quantum mechanics.

We are all familiar with theories which had to be discarded in the face of experimental evidence. In the field of astronomy, the earth-centered description of the planetary orbits was overthrown by the Copernican system, in which the sun was placed at the center of a series of concentric, circular planetary orbits. Later, this theory was modified, as measurements of the planets motions were found to be compatible with elliptical, not circular, orbits, and still later planetary motion was found to be derivable from Newton's laws.

Error in experiments have several sources. First, there is error intrinsic to instruments of measurement. Because this type of error has equal probability of producing a measurement higher or lower numerically than the "true" value, it is called random error. Second, there is non-random or systematic error, due to factors which bias the result in one direction. No measurement, and therefore no experiment, can be perfectly precise. At the same time, in science we have standard ways of estimating and in some cases reducing errors. Thus it is important to determine the accuracy of a particular measurement and, when stating quantitative results, to quote the measurement error. A measurement without a quoted error is meaningless. The comparison between experiment and theory is made within the context of experimental errors. Scientists ask, how many standard deviations are the results from the theoretical prediction? Have all sources of systematic and random errors been properly estimated? This is discussed in more detail in the appendix on *Error Analysis* and in Statistics Lab 1.

III. Common Mistakes in Applying the Scientific Method

As stated earlier, the scientific method attempts to minimize the influence of the scientist's bias on the outcome of an experiment. That is, when testing an hypothesis or a theory, the scientist may have a preference for one outcome or another, and it is important that this preference not bias the results or their interpretation. The most fundamental error is to mistake the hypothesis for an explanation of a phenomenon, without performing experimental tests. Sometimes "common sense" and "logic" tempt us into believing that no test is needed. There are numerous examples of this, dating from the Greek philosophers to the present day.

Another common mistake is to ignore or rule out data which do not support the hypothesis. Ideally, the experimenter is open to the possibility that the hypothesis is correct or incorrect. Sometimes, however, a scientist may have a strong belief that the hypothesis is true (or false), or feels internal or external pressure to get a specific result. In that case, there may be a psychological tendency to find "something wrong", such as systematic effects, with data which do not support the scientist's expectations, while data which do agree with those expectations may not be checked as carefully. The lesson is that all data must be handled in the same way.

Another common mistake arises from the failure to estimate quantitatively systematic errors (and all errors). There are many examples of discoveries which were missed by experimenters whose data contained a new

phenomenon, but who explained it away as a systematic background. Conversely, there are many examples of alleged "new discoveries" which later proved to be due to systematic errors not accounted for by the "discoverers."

In a field where there is active experimentation and open communication among members of the scientific community, the biases of individuals or groups may cancel out, because experimental tests are repeated by different scientists who may have different biases. In addition, different types of experimental setups have different sources of systematic errors. Over a period spanning a variety of experimental tests (usually at least several years), a consensus develops in the community as to which experimental results have stood the test of time.

IV. Hypotheses, Models, Theories and Laws

In physics and other science disciplines, the words "hypothesis," "model," "theory" and "law" have different connotations in relation to the stage of acceptance or knowledge about a group of phenomena.

An hypothesis is a limited statement regarding cause and effect in specific situations; it also refers to our state of knowledge before experimental work has been performed and perhaps even before new phenomena have been predicted. To take an example from daily life, suppose you discover that your car will not start. You may say, "My car does not start because the battery is low." This is your first hypothesis. You may then check whether the lights were left on, or if the engine makes a particular sound when you turn the ignition key. You might actually check the voltage across the terminals of the battery. If you discover that the battery is not low, you might attempt another hypothesis ("The starter is broken"; "This is really not my car.")

The word model is reserved for situations when it is known that the hypothesis has at least limited validity. A often-cited example of this is the Bohr model of the atom, in which, in an analogy to the solar system, the electrons are described has moving in circular orbits around the nucleus. This is not an accurate depiction of what an atom "looks like," but the model succeeds in mathematically representing the energies (but not the correct angular momenta) of the quantum states of the electron in the simplest case, the hydrogen atom. Another example is Hook's Law (which should be called Hook's principle, or Hook's model), which states that the force exerted by a mass attached to a spring is proportional to the amount the spring is stretched. We know that this principle is only valid for small amounts of stretching. The "law" fails when the spring is stretched beyond its elastic limit (it can break). This principle, however, leads to the prediction of simple harmonic motion, and, as a model of the behavior of a spring, has been versatile in an extremely broad range of applications.

A scientific theory or law represents an hypothesis, or a group of related hypotheses, which has been confirmed through repeated experimental tests. Theories in physics are often formulated in terms of a few concepts and equations, which are identified with "laws of nature," suggesting their universal applicability. Accepted scientific theories and laws become part of our understanding of the universe and the basis for exploring less well-understood areas of knowledge. Theories are not easily discarded; new discoveries are first assumed to fit into the existing theoretical framework. It is only when, after repeated experimental tests, the new phenomenon cannot be accommodated that scientists seriously question the theory and attempt to modify it. The validity that we attach to scientific theories as representing realities of the physical world is to be contrasted with the facile invalidation implied by the expression, "It's only a theory." For example, it is unlikely that a person will step off a tall building on the assumption that they will not fall, because "Gravity is only a theory."

Changes in scientific thought and theories occur, of course, sometimes revolutionizing our view of the world (Kuhn, 1962). Again, the key force for change is the scientific method, and its emphasis on experiment.

V. Are there circumstances in which the Scientific Method is not applicable?

While the scientific method is necessary in developing scientific knowledge, it is also useful in everyday problem-solving. What do you do when your telephone doesn't work? Is the problem in the hand set, the cabling inside your house, the hookup outside, or in the workings of the phone company? The process you might go

through to solve this problem could involve scientific thinking, and the results might contradict your initial expectations.

Like any good scientist, you may question the range of situations (outside of science) in which the scientific method may be applied. From what has been stated above, we determine that the scientific method works best in situations where one can isolate the phenomenon of interest, by eliminating or accounting for extraneous factors, and where one can repeatedly test the system under study after making limited, controlled changes in it.

There are, of course, circumstances when one cannot isolate the phenomena or when one cannot repeat the measurement over and over again. In such cases the results may depend in part on the history of a situation. This often occurs in social interactions between people. For example, when a lawyer makes arguments in front of a jury in court, she or he cannot try other approaches by repeating the trial over and over again in front of the same jury. In a new trial, the jury composition will be different. Even the same jury hearing a new set of arguments cannot be expected to forget what they heard before.

VI. Conclusion

The scientific method is intricately associated with science, the process of human inquiry that pervades the modern era on many levels. While the method appears simple and logical in description, there is perhaps no more complex question than that of knowing how we come to know things. In this introduction, we have emphasized that the scientific method distinguishes science from other forms of explanation because of its requirement of systematic experimentation. We have also tried to point out some of the criteria and practices developed by scientists to reduce the influence of individual or social bias on scientific findings. Further investigations of the scientific method and other aspects of scientific practice may be found in the references listed below.

VII. References

1. Wilson, E. Bright. An Introduction to Scientific Research (McGraw-Hill, 1952).

2. Kuhn, Thomas. The Structure of Scientific Revolutions (Univ. of Chicago Press, 1962).

3. Barrow, John. Theories of Everything (Oxford Univ. Press, 1991).

Send comments, questions and/or suggestions via email to **wolfs@nsrl.rochester.edu**.

---

02 - The Role of Experimentation

Monday, March 20, 2017 12:42 PM

UFMG

Design and Analysis of Experiments
02 - The Role of Experimentation

Version 2.11

Felipe Campelo
http://www.cpdee.ufmg.br/~fcampelo

Graduate Program in Electrical Engineering

Belo Horizonte
March 2015

"O que é MELHOR, lasanha ou batata-frita?" → intrinsecamente subjetivo

"There may be some beliefs that cannot be decided by data, but such beliefs are dogmas that lie (double entendre intended) beyond the reach of evidence."

John K. Kruschke
American mathematician and cognitive psychologist

Image: https://www.amazon.com/John-K.-Kruschke/e/B00474QQRO

Experiments
Definition of experiment

An experiment can be characterized as a test (or a series of tests) wherein changes are introduced in the state of a system or process, enabling the observation and characterization of effects that can occur as a result of these changes.

↳ experimentos planejados

Usually performed with an objective in mind:

- Uncovering influential variables in a given system or process;
- Determining desired values for certain parameters
- Characterize behavior of the system or process under study.

→ Em algumas áreas, não é possível introduzir mudanças no sistema para verificar seu efeito, visto que o objetivo é conhecer o sistema *in natura* → i.e. Ecologia

Experiments
Data gathering

- Retrospective study; → muito bons para gerar questões p/ estudos posteriores
- Observational study;
- Designed experiment;

↳ mineração de dados

Characteristics

- Use of historical data;
- Investigating correlations;

Problems

- Data representativeness; → possivelmente não há registro de todos os dados necessários
- Availability of data;

⇒ ex: correlação simples entre nº de afogamentos e nº de grevistas
↳ explicado pelo tamanho da população (possivelmente não inclusa nos dados)

• Outliers e erros nos dados

Cuidado com falsas correlações!
↳ spuriouscorrelations.com

Experiments
Data gathering

- Retrospective study;
- Observational study;
- Designed experiment;

Characteristics

- Observation of the system with minimal disturbance; → ... medições em uma linha de produção
- Investigation of usual behaviors;

Problems

- Low representativeness of extreme cases;
- Low variability can affect observation of interesting effects; → sensibilidade afetada por ruídos

Experiments
Data gathering

- Retrospective study;
- Observational study;
- Designed experiment;

Characteristics

- Observation of the system with minimal disturbance;
- Investigation of usual behaviors;

Problems

- Low representativeness of extreme cases;
- Low variability can affect observation of interesting effects;

Experiments
Data gathering

- Retrospective study;
- Observational study;
- Designed experiment;

Foco da disciplina

Characteristics

- Introduction of deliberate changes in the system;
- Inference on the *causality* of the effects;

Problems

- Requires rigorous experimental design and data analysis;
- Usually more expensive.

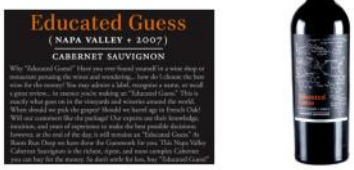
Experimentation strategies
Educated guessing

- Select arbitrary combination of levels for the factors;
- Test and observe behavior;
- Change one or two factors at a time, then re-test;

- Widely used in industry;
- Can achieve good results, but has a lot of limitations;

Experimentation strategies
Educated guessing

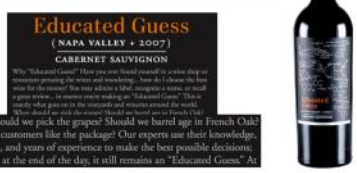
- Select arbitrary combination;
- Test an observe;
- Change and re-test;



Images (c) Roots Run Deep Winery: http://www.rootsrundeep.com/educated_guess.html

Experimentation strategies
Educated guessing

- Select arbitrary combination;
- Test an observe;
- Change and re-test;



When should we pick the grapes? Should we barrel age in French Oak? Will our customers like the package? Our experts use their knowledge, intuition, and years of experience to make the best possible decisions; however, at the end of the day, it still remains an "Educated Guess." At

Images (c) Roots Run Deep Winery: http://www.rootsrundeep.com/educated_guess.html

Experimentation strategies
COST: Change One Separate factor at a Time

- Select a reference point;
- Change each factor individually, keeping all others constant;

- Also widely used;
- Can achieve good results as long as there are no interaction effects;

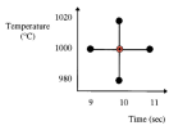


Image: (c) D.C. Montgomery

↳ nunca será alcançado

Experimentation strategies
Factorial designs

- Select **levels** for each factor;
- Vary the factors simultaneously, in a systematic way;

- Estimation of main effects and interactions;
- Greater precision in the effect estimates;
- More efficient use of resources (information/observation);

⊖ Requer muitas observações (maior custo) se há muitas variáveis

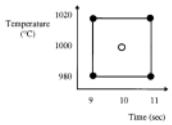


Image: (c) D.C. Montgomery

Fundamental principles
Design of experiments (DoE)

Process of designing data gathering protocols to enable accurate analyses by statistical tools, capable of supporting sound and objective conclusions.

- Applicable to systems and processes subject to noise, experimental errors, uncertainties, etc.
- Necessary for the conclusions to have a quantifiable meaning;
- Helpful in avoiding errors due to personal biases or other artifacts of experimentation and analysis.

→ Compromisso em descobrir a verdade, não provar um ponto

* Debi ① (autor em heurísticas multi objetivo)
↳ muita diferença entre artigo e código

Fundamental principles
Design of experiments (DoE)

Design of the experiment

- Scientific/technical question of interest;
- Selection of variables and values;
- Definition of the desired confidence level;
- Sample size calculations;
- Determination of protocols for data gathering;

→ S-TRIP
- tempo
- proximidade do Ótimo
- tempo real ?
- custo de obtenção dos dados | processamento

Statistical analyses of the data

- Calculation of a test statistic;
- Validation of the assumptions of the statistical model;
- Calculation of the magnitude of effects;
- Drawing of conclusions and recommendations;

Fundamental principles
Design of experiments (DoE)

- Repetition and replication;
- Randomization;
- Blocking.

- Repeated measurements - estimation of within-group variability;
- Replication - estimative of the experimental error;
- Greater precision in estimating the model parameters;

Fundamental principles
Design of experiments (DoE)

- Repetition and replication;
- Randomization;
- Blocking;

- Avoids contamination of the data by order-dependent effects such as:
  - Heating effects;
  - Wear and tear effects;
  - External interferences;

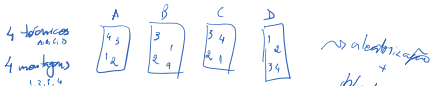
Use a medição não é pesforta (e.g. temp em 10 frascos)
⇒ medir em ORDEM aleatória

Fundamental principles
Design of experiments (DoE)

- Repetition and replication;
- Randomization;
- Blocking;

- Isolation of nuisance variables (those that influence the response, but are not interesting for the analyses) that can be controlled;
- Improvement in the estimation of effects for the factors of interest;
- Reduction or eliminations of inconvenient factor effects;

4 tratamentos A, B, C, D
4 manhãs 1, 2, 7, 4



↘ aleatorização + blocking

Fundamental principles
The role of experimental design

Experimental design is useful for avoiding the influence of spurious factors and personal biases on the results, by performing experiments in a impartial and objective way.

"Never have too much love for your hypotheses."

!

"The great tragedy of Science - the slaying of a beautiful hypothesis by an ugly fact."
– Thomas H. Huxley

Image: http://www.iep.utm.edu/huxley/

Discussion
Jacques Benveniste and the memory of water

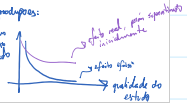
- Nature (1988);
- Investigation committee: Maddox, Stewart, Randi;
- Retracted by Nature due to evidence of misconduct.

Methodological problems

- Experimenter bias (absence of proper blinding);
- Cherrypicking (selective recording of results);
- Unaccounted sampling errors;
- Possible contamination;
- Complete lack of prior physical/ chemical plausibility;
- **Non-reproducibility**.

Image: http://www.jacques-benveniste.org/inmemoriam/photo1.html

· Reprodutores:

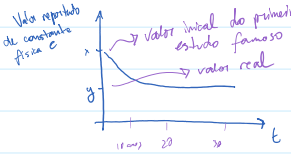


⇒ VANISHING EFFECT

artigo: "High Dilution Experiment: a Delusion"

↘ Sociologia da Ciência

↘ Ex: Efeito de Expectativa



→ bias dos pesquisadores e editores

→ mecanismo de autocorreção da ciência

Structure of Experimental Design
Main points

To enable the use of a scientific approach in the design of an experiment, it is important to have on a solid understanding of:

- The field where the experiment is to be conducted;
- The strategy for data collection;
- The way the data should be analyzed (at least qualitatively).

Structure of Experimental Design
Guidelines for a good design

- Pre-experimental design:
  - Identification and definition of the problem;
  - Selection of experimental and response variables of interest;
  - Choice of experimental protocols;
- Choice of the experimental design; → valorar das variáveis que devem contar nas observações
- Collection of the data;
- Statistical data analyses;
- Conclusions and recommendations; ★

Pre-experimental design
Before we start

- Is the investigation relevant?
- Would the results be interesting for the research community? por exemplo the general community (sociedade)
- Practical relevance?
  - Employ exploratory experiments;
- Placement within the literature;
  - Avoid repetition and irrelevance. ★

(RE-) READ: Some Modest Advice For Graduate Students

"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
Definition of hypotheses

- The translation *scientific question* → *test hypothesis* requires special attention, and a solid knowledge of the technical area in which the experiment is being performed;

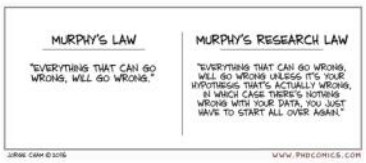


Image: http://phdcomics.com/comics/archive.php?comicid=1867

Choice of Experimental Design
Experimental design

→ Extremamente mais fácil com maior controlo sobre o experimento (e.g. experimentos computacionais)

- (Relatively) simple, as long as the pre-experimental part is well done;
- Dependent on what is being tested (statistical question);
- A sound design tends to determine the analyses technique to be used, at least qualitatively;
- Involves considerations about:
  - Sample size;
  - Ordering of observations;
  - Determination of restrictions to the randomization and the use of blocks, etc.
- Available in several statistical/mathematical packages; SAS SPSS MINITAB... (?)

Choice of Experimental Design
Problem-dependent

- Depending on the experimental question, different experimental designs are required
- A solid, statistically sound design tends to determine which statistical tests must be employed in the analysis step, at least qualitatively.
- Quantification of the proportion between intra-groups and inter-groups variability;

Actual Experiment
Data gathering

- Must be consistent with design, otherwise the validity of the results may be compromised - data collection must always follow the plan:
  - No premature stops;
  - *No-peeking rule*[a];
- Use of pilot experiments:
  - Gathering of preliminary information;
  - Practice with the experimental conditions;

[a] Except when planned, of course.

→ menor escala p/ teste

→ Tratamentos anteriores experimentais, por exemplo

Analysis of the experimental data
A consequence of design

- Analysis techniques are generally relatively simple, but *the devil is in the details*;
- Use of existing statistical tools and frameworks, such as

- Free, versatile, good graphical capabilities, relatively simple (but with one hell of a learning curve);

Analysis of the experimental data
Statistical modeling

- General procedure for testing the experimental hypotheses:
  - Definition of a *null-model* (absence of effects) and of a desired level of significance;
  - Determination of *P*(*data*|*null-model*);
  - Decision by rejection (or not) of the null hypothesis;
  - Validation of model assumptions;
  - Estimation of the *magnitude* of differences - **practical significance**;

*Statistical methods do not **prove** anything, but they allow an objective definition of margins of plausibility for certain statements.*

↳ "com muitos dados, qualquer coisa tem significância estatística"

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.

↳ "bíblia" da visualização de dados

Other great resources on graphical excellence:
Flowing Data (http://flowingdata.com/)
Information is Beautiful (http://www.informationisbeautiful.net/)

↳ experimento não foi planejado p/ verificar esse fato ⇒ requer investigações posteriores

↳ Grant Proposal Guide

Discussion

Some more relevant points

- Use of previous knowledge, theoretical or empirical;
- Iterative experimentation;
- Statistical × practical significance;
- Use of additional experiments to validate conclusions.

↳ buscar segunda linha de evidências

Bibliography

Required reading

1. Dept. Biochemistry & Cell Biology, Rice University , *Common Errors in Student Research Papers*.
2. T. Brady, *Reviewer's quick guide to common statistical errors in scientific papers.*
3. R.M. Szydlos et al., *Sign of the Zodiac as a predictor of survival for recipients of an allogeneic stem cell transplant for chronic myeloid leukaemia (CML): an artificial association*. Transplantation Proceedings 42 (8):3312-3315, 2010.
4. D.C. Montgomery, *Design and Analysis of Experiments*, Chapter 1. 5th ed., Wiley, 2005

Recommended reading

1. S.C. Stearns, *Some Modest Advice for Graduate Students*
2. A. Rendahl, *Experimental design and statistical analysis: questions to consider as you write your grant.*
3. J. Maddox et al., *"High-dilution" experiments a delusion*. Nature 334, 287-290, 1988
4. V. Czitron, *One-Factor-at-a-Time Versus Designed Experiments*. The American Statistician. 53(2) 126-131, 1999.
5. B. Dunning, *An Enthusiast's Primer on Study Types*. Skeptoid Podcast, Skeptoid Media, Inc., 2013.

↳ demonstração de relação absurda ("tortura de dados")

About this material

Conditions of use and referencing

Please reference this work as:

Felipe Campelo (2015), *Lecture Notes on Design and Analysis of Experiments*.
Online: https://github.com/fcampelo/Design-and-Analysis-of-Experiments
Version 2.11, Chapter 2; Creative Commons BY-NC-SA 4.0.

03 - Point Estimators

Monday, April 3, 2017 1:08 PM

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Design and Analysis of Experiments

03 - Point Estimators

Version 2.11

Felipe Campelo
http://www.cpdee.ufmg.br/~fcampelo

Graduate Program in Electrical Engineering

Belo Horizonte
March 2015

"A scientist must indeed be freely imaginative and yet skeptical, creative and yet a critic. There is a sense in which he must be free, but another in which his thought must be very precisely regimented; there is poetry in science, but also a lot of bookkeeping."

Sir Peter B. Medawar
1915-1987
British Immunologist

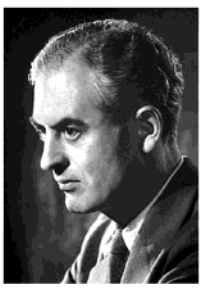


Image: http://commons.wikimedia.org/wiki/File:Peter_Brian_Medawar.jpgQ

Introduction
Probability vs. Statistics

Statistical inference: using *samples* to draw conclusion about *populations*;

Probability

Given the pool,what are the odds of drawing a certain combination of colors?

Statistics

Given the colors of a few balls drawn, what can I know about the pool?

Pool image: http://goo.gl/y8doaN

Population, Sample and Observation
Definitions

"A **population** is a large set of objects of a similar nature which is of interest as a whole"[1]. It can be an actual set (e.g., all balls in the pool) or a hypothetical one (e.g., all possible outcomes for an experiment).

A **sample** is a subset of a population. "A sample is chosen to make inferences about the population by examining or measuring the elements in the sample"[2].

An **observation** is a single element of a given sample, an individually collected data point. An observation can also be considered as a sample of size one.

Green ball: http://goo.gl/Fb8Z68

[1] Glossary of statistical terms: http://www.statistics.com/glossary&term_id=812

[2] Glossary of statistical terms: http://www.statistics.com/glossary&term_id=274

Point and Interval Estimates

Basic concepts

Two of the central concepts of statistical inference are *point estimators* and *statistical intervals*.

Both terms refer to using information obtained from a *sample* to infer probable values about *population* parameters;

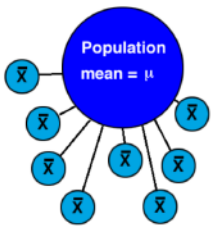
- **Point estimate**: estimated value for a given population parameter;
- **Statistical interval**: estimated interval of possible/probable values for a given population parameter;

Statistics and Sampling Distributions

Definition

Suppose one wants to obtain a point estimate for an arbitrary parameter, e.g. the mean of a given population;

Randomly sampling from a population results in a random variable, and any function of these observations - that is, any *statistic* - is consequently a random variable itself;



Being random variables means that statistics also have their own probability distributions, called *sampling distributions*[3]. Sampling distributions have specific characteristics that we'll explore later.

Image: http://www.philender.com/courses/intro/notes2/sample.html

[3] D.W. Stockburger: http://www.psychstat.missouristate.edu/introbook/sbk19.htm

Point Estimators
Definition

A *point estimator* is a statistic which provides the value of maximum plausibility for a given (unknown) population parameter θ.

Consider a random variable X distributed according to a given $f(X|\theta)$.

Consider also a random sample from this variable:
$\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$;

A given function $\hat{\Theta} = h(\mathbf{x})$ is called a *point estimator* of the parameter θ, and a value returned by this function for a given sample is referred to as a *point estimate* $\hat{\theta}$ of the parameter.

Point Estimators
Usual cases

Point estimation problems arise frequently in all areas of science and engineering, whenever there is a need for estimating, e.g.,:

- a population mean, μ;
- a population variance, σ^2;
- a population proportion, p;
- the difference in the means of two populations, $\mu_1 - \mu_2$;
- etc..

In each case there are multiple ways of performing the estimation task, and the decision about which estimators to use is based on the mathematical properties of each statistic.

Point Estimators
Unbiased estimators

A good estimator should consistently generate estimates that lie close to the real value of the parameter θ.

A given estimator $\hat{\Theta}$ is said to be *unbiased* for parameter θ if:

$$E\left[\hat{\Theta}\right] = \theta$$

or, equivalently:

$$E\left[\hat{\Theta}\right] - \theta = 0$$

The difference $E\left[\hat{\Theta}\right] - \theta$ is referred to as the *bias* of a given estimator.

Point Estimators
Unbiased estimators

The usual estimators for mean and variance are unbiased estimators;

Let $x_1, \ldots, x_N$ be a random sample from a given population X, characterized by its mean μ and variance σ^2. In this situation, it is possible to show that[4]:

$$E\left[\bar{x}\right] = E\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \mu$$

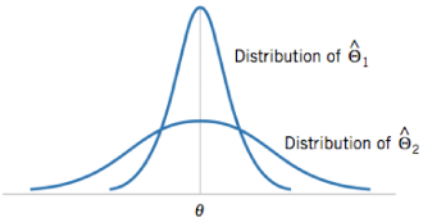
and:

$$E\left[s^2\right] = E\left[\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2\right] = \sigma^2$$

[4] For details see S.D. Anderson (1999), *Proof that Sample Variance is Unbiased*: http://git.io/vUn9N.

Point Estimators
Unbiased estimators

There usually exists more than one unbiased estimator for a given parameter θ. The variances of these estimators may, however, be different



A logical choice is to try to obtain the unbiased estimator of minimal variance. This is generally called the *minimal-variance unbiased estimator* (MVUE).

MVUE are generally chosen as estimators due to their ability of generating estimates $\hat{\theta}$ that are (relatively) close to the real value of θ.

Image: D.C.Montgomery,G.C. Runger, *Applied Statistics and Probability for Engineers*,Wiley 2003.

Point Estimators
Standard error

The *standard error* of an estimator $\hat{\Theta}$ corresponds to the standard deviation of that estimator,

$$\sigma_{\hat{\Theta}} = \sqrt{Var\left[\hat{\Theta}\right]}$$

When the standard error is estimated from a given sample we refer to it as the *estimated standard error*, $\hat{\sigma}_{\hat{\Theta}}$ (the notations $s_{\hat{\Theta}}$ and $se(\hat{\Theta})$ are also common).

Point Estimators
Standard error

For samples of Gaussian variables, it can be shown that:

$$\hat{\sigma}_{\hat{\Theta}} = \frac{S}{\sqrt{n}}$$

Since the distribution of many point estimators is approximately Normal for relatively large sample sizes (the usual rule of thumb is $N > 30$), this is a quite useful result, as it allows the use of the well-known properties of the Gaussian distribution as asymptotic results for the distributions of the estimators.

Sampling Distributions
Sampling distributions of means

Suppose a coaxial cable manufacturing operation that produces cables with a target resistance of 50Ω and a standard deviation of 2Ω[5], and assume that the resistance values can be well modeled by a normal distribution, i.e., $X \sim \mathcal{N}\left(\mu = 50, \sigma^2 = 4\right)$.

Also suppose a random sample of 25 cables is taken from this production process and their resistance is measured. The sample mean of the observations taken,

$$\bar{x} = \frac{1}{25}\sum_{i=1}^{25} x_i$$

is also normally distributed, with $E[\bar{x}] = \mu = 50\Omega$ (since the sample mean is an unbiased estimator) and $s_{\bar{x}} = \sqrt{\sigma^2/25} = 0.4$.

[5] Example inspired in https://www.sas.com/resources/whitepaper/wp_4430.pdf

Sampling Distributions
The Central Limit Theorem

Even for arbitrary population distributions the sampling distribution of means tends to be approximately normal (with $E[\bar{x}] = \mu$ and $s_{\bar{x}} = \sigma^2/N$).

More generally, let $x_1, \ldots, x_n$ be a sequence of *independent and identically distributed* (**iid**) random variables, with mean μ and finite variance σ^2. Then:

$$z_n = \frac{\sum_{i=1}^{n} (x_i) - n\mu}{\sqrt{n\sigma^2}}$$

is distributed approximately as a standard normal variable, that is, $z_n \sim \mathcal{N}(0, 1)$.

For more details on the CLT, see
http://www.encyclopediaofmath.org/index.php/Central_limit_theorem

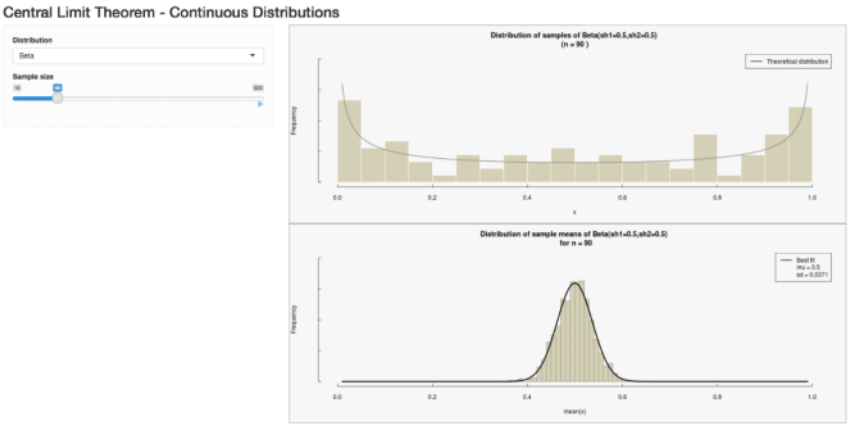
Sampling Distributions
The Central Limit Theorem

This result is known as the *Central Limit Theorem*, and is one of the most useful properties for statistical inference. The CLT allows the use of techniques based on the Gaussian distribution, even when the population under study is not normal.

For "well-behaved" distributions (continuous, symmetrical, unimodal - the usual bell-shaped pdf we all know and love) even small sample sizes are commonly enough to justify invoking the CLT and using parametric techniques.

Sampling Distributions
The Central Limit Theorem

For an interactive demonstration of the CLT, check
`http://drwho.cpdee.ufmg.br:3838/CLT/`



Bibliography

Required reading

1. D.C. Montgomery and G.C. Runger, *Applied Statistics and Probability for Engineers*, Chapter 7. 3rd Ed., Wiley 2005.
2. D.W. Stockburger, *Sampling Distributions*. In: *Introductory Statistics: Concepts, Models, and Applications* - `http://www.psychstat.missouristate.edu/introbook/sbk19.htm`

Recommended reading

1. R. Willett, *ECE 830 Estimation and Decision Theory, Spring 2014*, Chapters 13-15 - `http://willett.ece.wisc.edu/education.html`
2. S. Okasha, *Philosophy of Science - a very brief introduction*, Oxford Paperbacks, 2002.

About this material

Conditions of use and referencing

Please reference this work as:

Felipe Campelo (2015), *Lecture Notes on Design and Analysis of Experiments.*
Online: https://github.com/fcampelo/Design-and-Analysis-of-Experiments
Version 2.11, Chapter 3; Creative Commons BY-NC-SA 4.0.

```
@Misc{Campelo2015-01,
  title={Lecture Notes on Design and Analysis of Experiments},
  author={Felipe Campelo},
  howPublished={\url{https://github.com/fcampelo/Design-and-Analysis-of-Experiments}},
  year={2015},
  note={Version 2.11, Chapter 3; Creative Commons BY-NC-SA 4.0.},
}
```

04 - Statistical Intervals

Monday, April 3, 2017 1:13 PM

UFMG
UNIVERSIDADE FEDERAL DE MINAS GERAIS

Design and Analysis of Experiments
04 - Statistical Intervals

Version 2.11

Felipe Campelo
http://www.cpdee.ufmg.br/~fcampelo

Graduate Program in Electrical Engineering

Belo Horizonte
March 2015

"Science is an integral part of culture. It's not this foreign thing, done by an arcane priesthood. It's one of the glories of the human intellectual tradition."

Stephen Jay Gould
1941-2002
American paleontologist

Image: Harvard University / AP

Statistical Intervals
Introduction

Statistical intervals are important in quantifying the uncertainty associated to a given estimate;

As an example, let's recap the coaxial cables example: *a coaxial cable manufacturing operation produces cables with a target resistance of* 50Ω *and a standard deviation of* 2Ω. *Assume that the resistance values can be well modeled by a normal distribution.*

Let us now suppose that a sample mean of $N = 25$ observations of resistance yields $\bar{x} = 48$. Given the sampling variability, it is very likely that this value is not exactly the true value of μ, but we are so far unable quantify how much uncertainty there is in this estimate.

Statistical Intervals
Definition

Statistical intervals define regions that are likely to contain the true value of an estimated parameter.

More formally, it is generally possible to quantify the level of uncertainty associated with the estimation, thereby allowing the derivation of sound conclusions at predefined levels of certainty.

Three of the most common types of interval are:

- Confidence intervals;
- Tolerance intervals;
- Prediction intervals;

Confidence Intervals
Definition

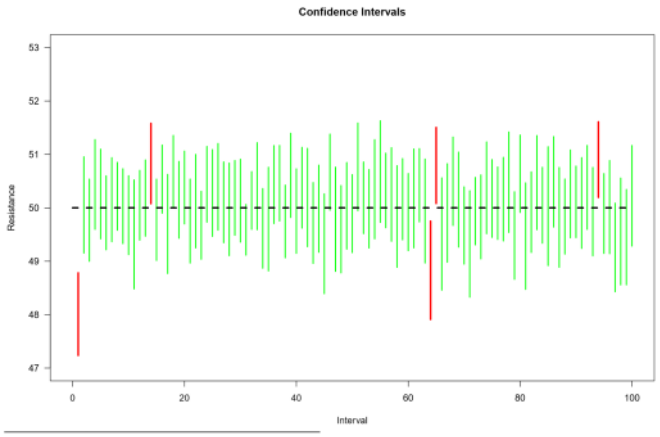
Confidence intervals quantify the degree of uncertainty associated with the estimation of population parameters such as the mean or the variance.

Can be defined as "*the interval that contains the true value of a given population parameter with a confidence level of* $100(1-\alpha)$";

- **Wrong**: "there is a 95% chance that the interval contains the true population mean."
- **right**: "The method used to derive the interval has a hit rate of 95%" - i.e., the interval generated has a 95% chance of "capturing" the true population parameter.

Easier to understand if you think about the confidence level as a confidence in the **method**, not in the interval.

Confidence Intervals
Example: 100 $CI_{.95}$ for a sample of 25 observations



For an interactive demonstration of the factors involved in the definition of a confidence interval, see
http://drwho.cpdee.ufmg.br:3838/CI/

Confidence Intervals

CI on the Mean of a Normal Variable

The two-sided $CI_{(1-\alpha)}$ for the mean of a normal population with known variance σ^2 is given by:

$$\bar{x} + \frac{\sigma}{\sqrt{N}} z_{(\alpha/2)} \leq \mu \leq \bar{x} + \frac{\sigma}{\sqrt{N}} z_{(1-\alpha/2)}$$

where $(1-\alpha)$ is the confidence level and $z_{(x)}$ is the x-quantile of the standard normal distribution.

For the more usual case with an unknown variance,

$$\bar{x} + \frac{s}{\sqrt{N}} t_{(\alpha/2;N-1)} \leq \mu \leq \bar{x} + \frac{s}{\sqrt{N}} t_{(1-\alpha/2;N-1)}$$

where $t_{(x;N-1)}$ is the x-quantile of the t distribution with $N-1$ degrees of freedom.

Confidence Intervals

CI on the Variance of a Normal Variable

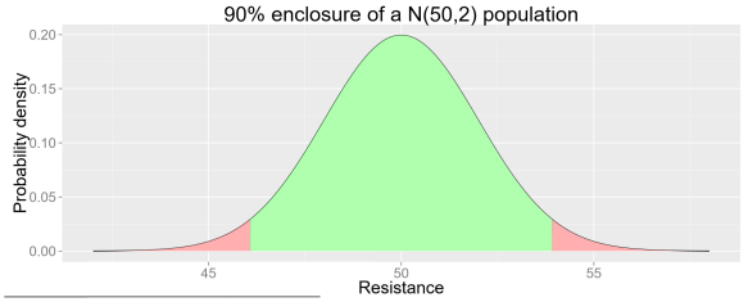
Similarly, a two-sided confidence interval on the variance of a normal variable can be easily calculated:

$$\frac{(N-1)s^2}{\chi^2_{\alpha/2;N-1}} \leq \sigma^2 \leq \frac{(N-1)s^2}{\chi^2_{1-\alpha/2;N-1}}$$

where $\chi^2_{\alpha/2;N-1}$ and $\chi^2_{1-\alpha/2;N-1}$ are the upper and lower $(\alpha/2)$-quantiles of the χ^2 distribution with $N-1$ degrees of freedom, respectively.

Tolerance Intervals
Definition

"*A tolerance interval is an* ***enclosure*** *interval for a specified proportion of the sampled population, not its mean or standard deviation. For a specified confidence level, you may want to determine lower and upper bounds such that a given percent of the population is contained within them.*"[1].



[1] J.G. Ramírez: http://goo.gl/NJz7ot

Tolerance Intervals
Definition

The common practice in engineering of defining specification limits by adding $\pm 3\sigma$ to a given estimate of the mean arises from this definition - for a normally-distributed population, aproximately 99.75% of the observations will fall within these limits.

However, as in most cases the true population variance is unknown, one has to use its estimate s^2 and compensate for the uncertainty in this estimation. The two-sided tolerance interval is then given as:

$$\bar{x} \pm \sqrt{\frac{(N-1)\left(N + z^2_{(\alpha/2)}\right)}{N\,\chi^2_{(\gamma;N-1)}}}$$

wherein γ is the proportion of the population to be enclosed, and $1-\alpha$ represents the desired confidence level for the interval.

Prediction Intervals

Definition

Prediction intervals quantify the uncertainty associated with forecasting the value of a future observation;

Essentially, one is interested in obtaining an interval within which he or she can declare that the next observation will fall with a given probability;

For a normal distribution, we have:

$$\bar{x} + t_{(\alpha/2;N-1)} s \sqrt{1 + \frac{1}{N}} \leq X_{N+1} \leq \bar{x} + t_{(1-\alpha/2;N-1)} s \sqrt{1 + \frac{1}{N}}$$

which is similar to the confidence interval for the mean, but adding 1 to the term within the square root to account for the prediction noise.

Statistical Intervals

Wrapping up

Statistical intervals quantify the uncertainty associated with different aspects of estimation;

Reporting intervals is always better that point estimates, as it provides to you (and your readers) the necessary information to quantify the location and spread of your estimated values;

The correct interpretation is a little tricky (although not that difficult)[2], but it is essential in order to derive the correct conclusions based on the statistical interval of interest.

[2] See the table at the end of: http://goo.gl/NJz7ot

Bibliography

Required reading

1. J.G. Ramírez, *Statistical Intervals: Confidence, Prediction, Enclosure*: `http://goo.gl/NJz7ot`
2. D.C. Montgomery and G.C. Runger, *Applied Statistics and Probability for Engineers*, Chapter 8. 3rd Ed., Wiley 2005.

Recommended reading

1. Simply Statistics (blog) - `http://simplystatistics.org`
2. R. Dawkins, *Climbing Mount Improbable*, W.W.Norton&Co.,1997.

About this material

Conditions of use and referencing

Please reference this work as:

Felipe Campelo (2015), *Lecture Notes on Design and Analysis of Experiments.* Online: `https://github.com/fcampelo/Design-and-Analysis-of-Experiments` Version 2.11, Chapter 4; Creative Commons BY-NC-SA 4.0.

```
@Misc{Campelo2015-01,
  title={Lecture Notes on Design and Analysis of Experiments},
  author={Felipe Campelo},
  howPublished={\url{https://github.com/fcampelo/Design-and-Analysis-of-Experiments}},
  year={2015},
  note={Version 2.11, Chapter 4; Creative Commons BY-NC-SA 4.0.},
}
```

05 - Statistical Inference

Monday, April 3, 2017 1:11 PM

UFMG

Design and Analysis of Experiments
05 - Statistical Inference

Version 2.11

Felipe Campelo
http://www.cpdee.ufmg.br/~fcampelo

Graduate Program in Electrical Engineering

Belo Horizonte
March 2015

"Nothing in life is to be feared,
it is only to be understood.
Now is the time to understand more,
so that we may fear less."

Marie Sklodowska Curie
1867–1934
Polish-French physicist and chemist.

Image: http://marie-curie.unblog.fr

Statistical Inference
Introduction

Definitions such as point estimators and statistical intervals belong to a branch of statistical theory known as *descriptive statistics*, that is, methods that are focused on accurately describing characteristics such as location or uncertainty about a given population parameter;

While these concepts are certainly important, in many cases description is not enough – one may need decision-making tools to deal with information from random samples, tools that allow a researcher to perform *inference* with a quantifiable degree of certainty.

Statistical hypotheses
Scientific Hypotheses

A *hypothesis* is a proposed explanation for an observable phenomenon.

Scientific hypotheses must satisfy (at least) two conditions:

- Testability;
- Falsifiability;

"The more we learn about the world, and the deeper our learning, the more conscious, specific, and articulate will be our knowledge of what we do not know, our knowledge of our ignorance."

Sir Karl R. Popper
(1902-1994)
Austro-British philosopher

Image: Copyright 2008 Ivan Jerónimo http://www.ivanjeronimo.com.br/images.php?imagem=33

Statistical Hypothesis
The hypothetico-deductive model

The *hypothetico-deductive model* of construction of scientific knowledge includes:

- Formulation of falsifiable hypotheses;
- Refutation or corroboration of the hypotheses by the data;
- Comparison between alternative hypotheses - principle of parsimony (Ockham's razor);
- Predictive power;

↳ Entre duas explicações igualmente boas para um fato, usualmente favorece-se aquela com menos premissas não verificadas

↳ Teste mais importante p/ hipótese

"Numquam ponenda est pluralitas sine necessitate."
"Never assume plurarity without meaning." William of Ockham
(1287-1347)
English philosopher and theologian

Image: http://www.philosophybasics.com/philosophers_ockham.html

---

Statistical Hypotheses

→ Ex: Porque caneta está na sala?

1) Fabricada na China, exportada p/ Brasil p/ distribuidora, licitada pela universidade, encaminhada p/ professor

2) Gnomos mágicos colocaram na gaveta

⇒ (1) aparentemente mais complexa, porém todas as premissas são sólidas e bem estabelecidas.

Image: https://www.philosophybasics.com/philosophers_locke.html

Statistical Hypotheses
Definitions

Statistical hypotheses are defined as objective statements about parameters of one ore more populations.

Attention: the statements in statistical hypotheses are about parameters of the *population or model*, **not the** *sample*.

On frequentist approaches, the formal test of hypotheses involves the contrast between *null* and *alternative* hypotheses.

Null hypothesis (H_0)	Alternative hypothesis (H_1)
• Absence of effects;	• Presence of some effect;
• *Conservative* model;	• Existence of something "new".
• Point value for the parameter.	• Interval value for the parameter.
Example: $H_0 : \mu = 25$	**Example:** $H_1 : \mu \neq 25$

necessário definir o grau de precisão é relevante

Statistical Hypotheses
Definitions

Determination of the reference value for the null hypothesis H_0:

- Previous knowledge about the process (investigation of changes);
- Value obtained from theory or models (model validation);
- Project requirements (investigation of system compliance);

Hypothesis testing involves:

- Obtaining the sample;
- Calculation of test statistics;
- Decision based on the computed value;

Statistical Hypotheses
Example

Suppose you are a large-scale customer of green peas[a], and that you want to determine if the 500g packages from a given food supplier really contain their nominal weight (at least on average).

In this case the null hypothesis could be defined as: *the average net weight of a package is* 500*g*, and the alternative of interest could be expressed as the complementary inequality.

$$\begin{cases} H_0 : \mu = 500g \\ H_1 : \mu \neq 500g \end{cases}$$

Suppose still that $n = 10$ randomly selected packs are obtained from this supplier, and their contents are weighted using a calibrated scale;

[a] We could use any other item on your usual grocery list, but why not pay a little tribute to Gregor Mendel?
Image: https://www.otsorko.eu/ed_files/image/green-peas.jpg

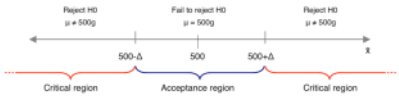
Statistical Hypotheses
Example

Since the sample mean $\bar{x}$ is a good estimator of the real mean μ, we can assume:

- If $\bar{x} \cong 500g$ - corroboration of H_0;
- If $\bar{x} \ll 500g$ or $\bar{x} \gg 500g$ - refutation of H_0;

Suggests the use of $\bar{x}$ as basis for a statistical test.

Definition of a *critical region* for the rejection of H_0:



Inferential Errors
Type I error

Type I error (false positive): rejecting the null hypothesis when it is true.

The probability of occurrence of a false positive in any hypothesis testing procedure is generally known as the *significance level* of the test, represented by Greek letter α:

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 | H_0 \text{ is true})$$

Another frequently used term is the *confidence level* of the test, given by $(1 - \alpha)$.

⇒ ① aparentemente mais complexa, porém todas as premissas são sólidas e bem estabelecidas.
② depende de mais premissas não verificadas.

POPULAÇÃO → parâmetros
AMOSTRA (Sample)
observação
ESTIMADORES AMOSTRAIS

→ Estimadores têm suas próprias características: forma da distribuição, média $E[\theta]$, variância $V[\theta]$

- Média μ: $\hat{\mu} = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
- Desvio Padrão σ^2: $\hat{\sigma}^2 = s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$

↝ Diferentes estimadores podem ser utilizados p/ calcular o mesmo parâmetro (e.g. média podada vs. média podada)

- Escolha depende de características do estimador: → viés → variância ⋮

→ média é o melhor estimador, mas pode ser mal indicador de caso típico em alguns problemas

* "Uma Senhora Bebe Chá" . livro sobre história da estatística

DECISÃO \ REALIDADE	H_0: V	H_0: F
H_0: V	✓	Tipo 2 (Falso negativo) (β)
H_0: F	Tipo 1 (Falso positivo) (α)	✓

↳ $(1-\beta)$: potência do teste

↳ nível de significância: usualmente $\alpha = 0.05$ (~2 desvios padrão) ou $\alpha = 0.01$ (~2.5 " ")

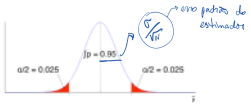
$(1-\alpha)$ → nível de significância

Inferential Errors
Type I error

For a given sample, the selected value of α defines the critical threshold for the rejection of H_0.

If H_0 is true (i.e., if $\mu = 500$g), the distribution of values of $\bar{x}$ is aproximately normal (remember the CLT), with average 500 and variance given by s^2/n;

For a Type-I error probability $\alpha = 0.05$, the critical values of the distribution of $\bar{x}$ are the ones for which the probability content within the acceptance region is $1 - \alpha = 0.95$.



- Decisão do nível de significância deve se basear no problema (custo de um falso positivo e de um falso negativo)

↳ NO ENTANTO, há resistências dos resistores para valores diferentes do padrão da área

↳ C.L.T. (Central Limit Theorem): a média de uma amostra de (n) elementos independentes tende a uma distribuição normal quando $(n \to \infty)$.

Inferential Errors
Type II error

Type II error (false negative): failure to reject the null hypothesis when it is false.

The probability of occurrence of a false negative in any hypothesis testing procedure is generally represented by the Greek letter β:

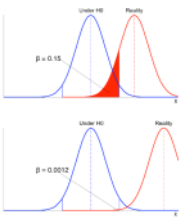
$$\beta = P(\text{type II error}) = P(\text{not reject } H_0 | H_0 \text{ is false})$$

The quantity $(1 - \beta)$ is known as *power* of the test, and quantifies its sensitivity to effects that violate the null hypothesis.

Inferential Errors
Type II error

Unlike the Type-I error, the definition of the Type-II error rate requires further specification of the value of the parameter being investigated under the alternative hypothesis;

The probability of failing to reject a false H_0 is strongly dependent on the magnitude of the difference between the value under H_0 and the real value of the parameter.



↳ Tamanho do efeito que tem importância prática

Inferential Errors
Type II error

The power of a test is governed by several factors:

- Controllable: significance level, sample size;
- Uncontrollable: real value of the parameter;

If H_0 is false, the smaller the magnitude of the difference between the real value of the parameter and the one under the null hypothesis, the greater the probability of a type II error - ***but the practical importance of the effect gets smaller***.

Inferential Errors
Considerations

Type I error (α) depends only on the distribution of the null hypothesis - easier to control;

Type II error (β) depends on the real value of the parameter - more difficult to specify and control;

These characteristics lead to the following classification of the conclusions obtained from the test of hypotheses:

- Rejection of H_0 - *strong* conclusion;
- Failure to reject H_0 - *weak* conclusion (but we can strengthen it);

→ Por isso é importante definir H_0 como a solução padrão e H_1 como a nova proposta

It is important to remember that failing to reject H_0 does not mean that there is evidence in favor of H_0 - it only suggests that it is a better model than the alternative.

Hypothesis Testing
General procedure

- Identify the parameter of interest;
- Define H_0 and H_1 (one- or two-sided);
- Determine desired α, β;
- Define minimally interesting effect δ^*;
- Calculate sample size;
- Determine the test statistic and critical region;
- Compute the statistic;
- Decide whether or not to reject H_0;

Image (c) Roots Run Deep Winery: https://www.rootsrundeep.com/hypothesis.html

- α: Erro tipo I
- β: Erro tipo II
- σ: Desvio padrão dos resíduos
- n: Tamanho amostral / D.O.F. dos resíduos
- δ: "Tamanho de efeito" / Magnitude da diferença

parâmetros do experimento

4 usados p/ calcular o S^*

desconhecido, mas δ^* pode ser determinado com conhecimento do problema

σ desconhecido: Como calcular (1) necessário?
- estimativa
- estudo piloto
- desvio padrão de interesse $d^* = \frac{\delta^*}{\sigma}$ → Cohen's d

controverso / não aplicável sempre

resultados não são exatos (100% seguros)

Hypothesis Testing
Mean of a normal distribution, variance known

Back to the green peas example, we want to determine if there is any significant deviation on the mean weight of the packages. Assume for now that the variance of the process is known. The test hypotheses are defined as:

$$\begin{cases} H_0 : \mu = 500g \\ H_1 : \mu \neq 500g \end{cases}$$

$n = 10$

Let the desired significance level be $\alpha = 0.05$;

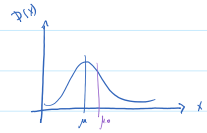
Given these characteristics, we expect that the sampling distribution of $\bar{X}$ is normal, with variance $Var(\bar{X}) = \sigma^2/n$ and, if H_0 is true – mean $\mu_{\bar{X}} = \mu_0 = 500$;

$$\bar{X} \sim \mathcal{N}\left(\mu, \sigma^2/n\right)$$

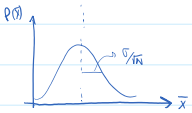
↳ TLC

$d < 0.3$ → "efeito pequeno"
$0.3 \leq d \leq 0.7$ → "efeito médio"
$d > 0.7$ → "efeito grande"

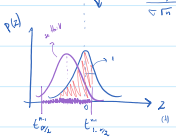
→ DEPENDE DA ÁREA / CONTEXTO !



↓ amostra (N)



$Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$



Hypothesis Testing
Mean of a normal distribution, variance known

Based on these characteristics, the variable

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

will be distributed according to the standard normal, $\mathcal{N}(0,1)$, but only **if H_0 is true**.

This result implies a probability of $(1-\alpha)$ that Z_0 will fall within the range $(\pm z_{\alpha/2})^b$ if H_0 is true, which provides a selection criterion between H_0 and H_1:

- If $|z_0| > z_{\alpha/2}$, we reject H_0 at the confidence level $1-\alpha$;
- Otherwise, there is not enough evidence to reject H_0;

[b] $z_{\alpha/2}$ is the superior $100(1-\alpha/2)\%$ percentile of the standard normal distribution;

Hypothesis Testing
Mean of a normal distribution, variance known

Assume that we got $\bar{x} = 496.48g$ from our $n = 10$ observations, and that $\sigma = 10g$. In this case,

$$z_0 = \frac{496.48 - 500}{10/\sqrt{10}} = -1.113$$

The critical values of the standard normal distribution are $\pm z_{\alpha/2} = \pm z_{0.025} = \pm 1.96$;

Since $|z_0| < z_{0.025}$, we can conclude that there is not enough evidence to reject H_0 at the 95% confidence level.

↳ Teste Z : rejeita ou não H_0
↳ pressupõe variância conhecida

Hypothesis Testing
Mean of a normal distribution, variance known

```
> if(!require(TeachingDemos)){
+   install.packages("TeachingDemos")
+   library(TeachingDemos)
+ }
>
> sample <- as.numeric(scan("../data files/greenpeas.txt"))
> z.test(sample,
+        mu=500,
+        stdev=10)

One Sample z-test
data:  sample
z = -1.1131, n = 10.000, Std. Dev. = 10.000,
Std. Dev. of the sample mean = 3.162,
p-value = 0.2657
alternative hypothesis: true mean is not equal to 500
95 percent confidence interval:
 490.282 502.678
sample estimates:
mean of as.numeric(sample)
                    496.48
```

Hypothesis Testing
Mean of a normal distribution, variance unknown

Suppose now a more realistic situation in which the real variance is unknown. Besides, assume that we are interested in detecting only negative deviations from the nominal contents of the package.

The test hypotheses can be defined as:

$$\begin{cases} H_0 : \mu = 500g \\ H_1 : \mu < 500g \end{cases}$$

→ melhor potência com mesmo poder amostral

In this second scenario we want to be more conservative, so we pick a significance level of $\alpha = 0.01$;

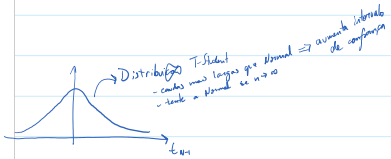
It can be shown that, **if H_0 is true**, then

$$T_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

única diferença em relação a Z_0

$$S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

estimador de variância



4/10/2017

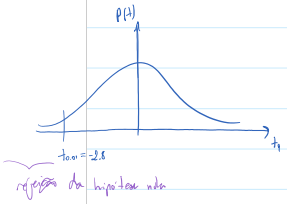
Hypothesis Testing
Mean of a normal distribution, variance unknown

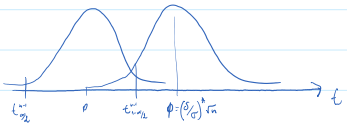
From the same data, $\bar{x} = 496.48g$, $n = 10$, $s = 6.97g$:

$$t_0 = \frac{496.48 - 500}{6.97/\sqrt{10}} = -1.5969$$

The critical value of this test statistic for the desired significance is $t_{\alpha,n-1} = t_{0.01,9} = -2.82$;

Given that $t_0 > -2.82$, we conclude that the evidence is insufficient to reject H_0 at the 99% confidence level;





Hypothesis Testing
Reporting results

Description of the results:

> *(In)Sufficient evidence for rejecting H_0 at the significance level α.*

Even thought it is correct, this description is relatively poor:

- It does not provide information on the intensity of the evidence for rejection/non-rejection;
- It imposes a predetermined significance level to the consumer of the information;
- Does not provide information the magnitude of the effect found or the sensitivity of the test.

Hypothesis Testing
The p-value

> **p-value**: *the lowest significance level that would lead to the rejection of H_0 for the available data.*

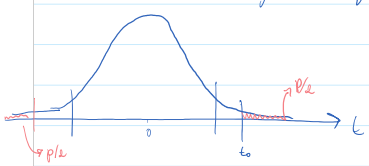
Can be interpreted as the probability under H_0 of the test statistic assuming a value at least as extreme as the one obtained;

For the previous example, the p-value could be calculated as:

$$p = P(t_0 \leq -1.597 | H_0 = \textbf{TRUE}) = \int_{-\infty}^{-1.597} (t_9)dt = 0.07237$$

> *A priori* definition of the significance level is still important!

→ p=0.003 ⇒ Qualquer intervalo de confiança ≤ 99.7% é suficiente p/ rejeita hipótese nula



→ probabilidade de H_0 = TRUE considerando os dados

Hypothesis Testing
p-values, significance and effect sizes

Statistical × practical significance: p-values can be made arbitrarily small, if n is big enough;

As an example, suppose a test of $H_0 : \mu = 500g$ against a two-sided alternative, with $n = 5000$, $\bar{x} = 499g$, $s = 5g$. In this case we would have:

- $t_0 = -14.142$;
- $p = 1.02 \times 10^{-23}$;

Is it really *that* significant?

→ p é muito pequeno porque o tamanho de efeito considerado é muito pequeno → possivelmente insignificante
→ chance de ser ruído, efeito da amostra, etc.

Hypothesis Testing
p-values, significance and effect sizes

To "tell the whole story" of the experiment, it is necessary to use effect size estimators alongside the tests of statistical significance;

While there are whole books on the subject [c], the main idea is quite simple - to quantify the magnitude of the observed deviation from the null hypothesis.

Examples of effect size estimators include the simple point estimator for the difference $\bar{x} - \mu_0$, or the dimensionless d estimator:

$$d = \frac{\bar{x} - \mu_0}{s}$$

which quantifies the difference in terms of sample standard deviations.

[c] See, for instance, Paul D. Ellis' *The Essential Guide to Effect Sizes*, Cambridge University Press, 2010.

→ Barnett, around

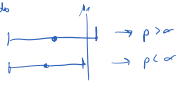
Hypothesis Testing
p-values, effects sizes and confidence intervals

Point estimators + confidence intervals quantify the magnitude and accuracy of effects, and must be reported alongside the results of significance testing whenever possible.

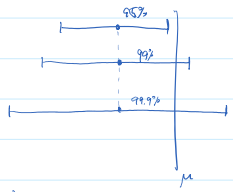
Suppose we are testing $H_0 : \mu = 500$ against the two-sided alternative hypothesis, with $n = 10$ and $\alpha = 0.01$. Assume that the population is known to be normal, with unknown variance. We'll use the same data as before:

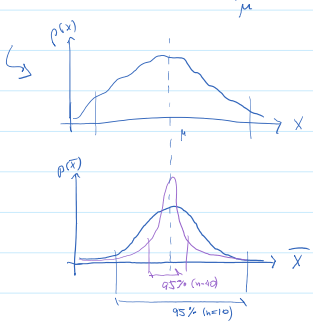
```
> t.test(sample, mu = 500, conf.level = 0.99)
        (...)
t = -1.5969, df = 9, p-value = 0.1447
alternative hypothesis: true mean is not equal to 500
99 percent confidence interval:
 489.3166 503.6434
sample estimates:
mean of x
   496.48
```

→ graus de liberdade
→ α desejado: 0.01
→ menor α que rejeita H_0: 0.14
⇒ H_0 não rejeitada ($p > \alpha$)
⇒ H_0 não rejeitado



→ Intervalo de confiança: $\bar{x} + t^{n-1}_{\alpha/2}\, s/\sqrt{n} < \mu < \bar{x} + t^{n-1}_{1-\alpha/2}\, s/\sqrt{n}$





⇒ Intervalo de confiança se refere à média, não a distribuição da população !

Sample size and Type-II error
Some considerations

The probability of Type-II error can be easily evaluated *a posteriori*, but its definition *a priori* requires some care;

The power of a test is essentially a function of 4 elements:

- Actual size of the difference;
- Variability of the observations;
- Significance level;
- Sample size.

The experimenter generally have little control over the first two.

Sample size and Type-II error
Some considerations

A strategy for estimating an effective lower bound for the power of a test includes a definition of an *minimally interesting effect* δ^*.

This value must be derived from technical and scientific knowledge about the phenomenon or system under experimentation.

It is essential to have a good understanding of the field in which the experiment will be conducted.

Once δ^* is defined, the experimenter can obtain an estimate of the variability of observations (e.g., by a pilot study), which can then be used to obtain an approximate power value for the experiment;

$d^* = \left(\frac{\delta^*}{\sigma}\right)$

Sample size and Type-II error
Some considerations

Having obtained this estimation of the Type-II error probability, one can run his/her experiment with a better understanding of its ability to detect effects of interest.

- The test will have lower power for differences smaller than δ^*, but these differences are below the minimally interesting effect; any effect greater than δ^* will result in a higher power for the test;
- This technique can also be used as a way to compute the maximum necessary sample size for the experiment.

Sample size and Type-II error
Example

Suppose that on the green peas example one is really interested in detecting deviations from the nominal value greater than 1%, i.e., $\delta^* = 0.01 * 500 = 5g$. The researcher defines that, for this minimally interesting effect, a test power of 0.85 is desired. The test will again be performed with $\alpha = 0.01$.

The same sample of $n = 10$ packs is used. The estimated standard deviation for this sample is $s = 6.97g$. From this data, we can compute the power of this test as:

```
> s<-sd(sample)                        One-sample t test power calculation
> power.t.test(n=10,                   n = 10
+              delta=5,                delta = 5
+              sd=s,                   sd = 6.970382
+              sig.level=0.01,         sig.level = 0.01
+              type = "one.sample",    power = 0.3474724 <-- Power
+              alternative = "one.sided") alternative = one.sided
```

$\delta^* = 5g$
$\alpha = 0.01$
$\beta = 0.15$ (desejado)
$n = 10$
$\sigma^2 = 10$
↳ estimado historicamente

Sample size and Type-II error
Example

What is the smallest sample size needed to obtain the desired power of 0.85?

```
> power.t.test(power=0.85, delta=5, sd=s, sig.level=0.01,
               type = "one.sample", alternative = "one.sided")

One-sample t test power calculation
n = 24.76091 <-- (round this value up)
delta = 5
sd = 6.970382
sig.level = 0.01
power = 0.85
alternative = one.sided
```

We need at least 25 observations to detect a $-5g$ (1%) or larger deviation on the mean weight of the green peas packages with a power level of 0.85.

Model validation
The normality assumption

The assumption of normality, required for the **z** and **t** tests, needs to be validated.

> "*The Assumption of Normality (note the upper case) that underlies parametric stats does not assert that the observations within a given sample are normally distributed, nor does it assert that the values within the population (from which the sample was taken) are normal. This core element of the Assumption of Normality asserts that* ***the distribution of sample means (across independent samples) is normal.***"
>
> – J. Toby Mordkoff, 2011.[a]

↳ Dados não precisam ser normais, apenas a distribuição das médias.

No entanto, dados normais induzem médias normais.

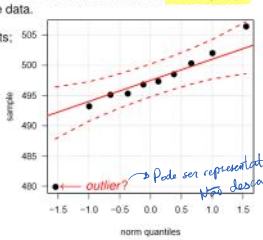
[a] Check J.T. Mordkoff's "the assumption(s) of normality for a nice discussion on this topic: (http://goo.gl/2x4ku)

Model validation
The normality assumption

If the conditions for the CLT cannot be assumed, then normality tests can be performed on the data.

Graphical/qualitative tests;

```
> library(car)
> qqPlot(sample,
+        pch=16,
+        cex=1.5,
+        las=1)
```



↳ Pode ser representativo. Não descartar p/ dados normais

Model validation
The normality assumption

Analytical tests of normality (choose **one**):

- Shapiro-Wilk;
- Anderson-Darling;
- Lilliefors / Kolmogorov-Smirnov;

Mais utilizado, porém todos são bons

↳ não é necessário fazer mais

These procedures use different aspects of the sample distribution to test the following hypotheses:

$$\begin{cases} H_0 : \text{population is normal} \\ H_1 : \text{population is not normal} \end{cases}$$

In this case, rejection of the null hypothesis suggests evidence that the sample came from a non-normal population. Generally we use a strict threshold (e.g., $\alpha = 0.01$ or 0.001) for these tests, and consider their results together with a graphical analysis.

$\begin{cases} p < \alpha : \text{pop not normal } (H_0 \text{ rejected with } \alpha \text{ conf.}) \\ p > \alpha : \text{pop normal } (H_0 \text{ not rejected}) \end{cases}$

Model validation
The normality assumption

Even thought the Lilliefors / Kolmogorov-Smirnov test is possibly the most widely used for normality testing, the Shapiro-Wilk test is recommended as a better alternative in Michael Crawley's *The R Book*, and will be used throughout this course.

```
> shapiro.test(sample)

Shapiro-Wilk normality test
data:  sample
W = 0.8809, p-value = 0.1335
```

Model validation
The independence assumption

Possibly the strongest assumption of the statistical model used for the t-test is that of independence, that is, of the absence of unmodeled biases contaminating the data.

While I know of no procedure to test independence in the general case, serial autocorrelations in the residual data (which can emerge, for instance, as a consequence of heating effects or equipment degradation) can be tested by a procedure known as the Durbin-Watson test:

```
> library(lmtest)
> dwtest(sample~1)

Durbin-Watson test
data:  sample ~ 1
DW = 2.1117, p-value = 0.573
alternative hypothesis: true autocorrelation is greater than 0
```

Model validation
The independence assumption

The Durbin-Watson test depends on the ordering of the data, so observations should be ordered (either in the data file or by manipulating the data vector) according to covariates that are suspected to introduce dependencies, e.g., order of collection, placement criteria, etc.

Violations of the independence assumption tend to be the hardest to weed out, so extra care is recommended in the design of the experiment in order to prevent, control, or at least document all variables that could introduce dependencies in the data.

The green peas experiment
Going over the process

After examining the green peas example, it is interesting to go back and follow the recommended sequence for this kind of experiment:

- Formulate question of interest;
- Define minimally interesting effect;
- Define desired confidence and power;
- Calculate required sample size;
- Collect data;
- Perform statistical analysis;
- Draw conclusions and recommendations.

Bibliography

Required reading

- D.C. Montgomery, G.C. Runger, *Applied Statistics and Probability for Engineers*, Ch. 9. 5th ed., Wiley, 2010.
- W. Thalheimer and S. Cook, *How to calculate effect sizes from published research articles: A simplified methodology* - http://goo.gl/c0g1oK
- J.T. Mordkoff (2011), *The assumption(s) of normality* - http://goo.gl/Z3w8ku

Recommended reading

- A. Reinhart, *Statistics Done Wrong: the woefully complete guide* - http://www.statisticsdonewrong.com
- E. Ernst and S. Singh, *Trick or Treatment*, Norton & Company, 2009.

60

About this material

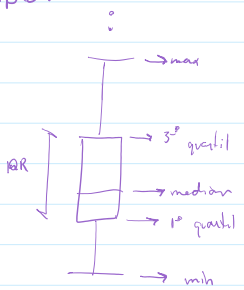
Conditions of use and referencing

Please reference this work as:

Felipe Campelo (2015), *Lecture Notes on Design and Analysis of Experiments*.
Online: https://github.com/fcampelo/Design-and-Analysis-of-Experiments
Version 2.11, Chapter 5; Creative Commons BY-NC-SA 4.0.

```
@Misc{Campelo2015-DA,
  title={Lecture Notes on Design and Analysis of Experiments},
  author={Felipe Campelo},
  howpublished={\url{https://github.com/fcampelo/Design-and-Analysis-of-Experiments}},
  year={2015},
  note={Version 2.11, Chapter 5; Creative Commons BY-NC-SA 4.0.},
}
```


* Boxplot:



→ min e max a uma distância máxima de 1.5 IQR da dos quartis

Case Study 01

Monday, April 17, 2017 2:16 PM

- Exploratory Data Analisys
  - plot em ordem com linha de regressão ⇒ análise de efeito de ordem
  - dot plot → observar outliers
  - Estimated PDF
  - QQ plot → padrão em "U"/"∩" tende a normalizar com transf em log

- Intervalos de confiança
  - Paramétrico : assume dist. conhecida (t.test)
  - Não-paramétrico : Bootstraping (library boot), pegar média de todas ⇒ médias normais, premissa "validada"
    ↳ intervalo de conf. das médias

bootstrap pode ser usado p/ construir hipótese nula → sob Ho (bootstrap) → real

- Teste de autocorrelação
  - Durbin Watson
    - lmtest : teste unilateral → multiplicar por 2 se necessário bilateral
    - car (dwt/durbinWatsonTest) → bilateral, usa bootstraping
  - Regressão linear e verificar valor (P) do slope

- VALORES REAIS: 80 moedas
  R$ 6.40

- Potência: considerar pior caso de (S) com 95/99%

de confiances

The Assumption(s) of Normality

Monday, April 24, 2017 1:09 PM

The Assumption(s) of Normality

This is very complicated, so I'll provide two versions. At a minimum, you should know the short one. It would be great if you knew them both.

Short version: in order to do something as magical as provide a specific probability for observing a particular mean or a particular difference between two means, our statistical procedures must make some assumptions. One of these assumptions is that the sampling distribution of the mean is normal. That is, if you took a sample, calculated its mean, and wrote this down; then took another (independent) sample and got its mean and wrote it down; and did this an infinite number of times; then the distribution of the means you'll write down will always be a perfect bell curve. While maybe surprising, this assumption turns out to be relatively uncontroversial, at least when large samples are used, such as $N \geq 30$. But in order to use the same statistical procedures for all sample sizes and in order for the underlying procedures to be as straightforward as they are, we must expand this assumption to saying that all populations from which we take samples are normal. In other words, we have to assume that the data inside each sample are normal, not just that the means across samples are normal. This is a very strong assumption and it probably isn't always true, but we have to assume this to use our procedures. Luckily, there are simple ways to protect ourselves from the problems that would arise if these assumptions are not true.

Now, the long version….

Nearly all of the inferential statistics that psychologists use (e.g., *t*-tests, ANOVA, simple regression, and MRC) rely upon something that is called the "Assumption of Normality." In other words, these statistical procedures are based on the assumption that the value of interest (which is calculated from the sample) will exhibit a bell-curve distribution function if oodles of random samples are taken and the distribution of the calculated value (across samples) is plotted. This is why these statistical procedures are called *parametric*. By definition, parametric stats are those that make assumptions about the shape of the sampling distribution of the value of interest (i.e., they make assumptions about the skew and kurtosis parameters, among other things; hence the name). The shape that is assumed by all of the parametric stats that we will discuss is *normal* (i.e., skew and kurtosis are both zero). The only statistic of interest that we will discuss here is the mean.

What is assumed to be normal?

When you take the parametric approach to inferential statistics, the values that are assumed to be normally distributed are the means across samples. To be clear: the Assumption of Normality (note the upper case) that underlies parametric stats does not assert that the observations within a given sample are normally distributed, nor does it assert that the values within the population (from which the sample was taken) are normal. (At least, not yet.) This core element of the Assumption of Normality asserts that the distribution of sample means (across independent samples) is normal. In technical terms, the Assumption of Normality claims that *the sampling distribution of the mean is normal* or that *the distribution of means across samples is normal.*

Example: Imagine (again) that you are interested in the average level of anxiety suffered by graduate students. Therefore, you take a group of grads (i.e., a random sample) and measure their levels of anxiety. Then you calculate the mean level of anxiety across all of the subjects. This final value is the sample mean. The Assumption of Normality says that if you repeat the above sequence many many many times and plot the sample means, the distribution would be normal. Note that I never said anything about the distribution of anxiety levels within given samples, nor did I say anything about the distribution of anxiety levels in the population that was sampled. I only said that the distribution of sample means would be normal. And again, there are two ways to express this: "the distribution of sample means is normal" and/or "the sampling distribution of the mean is normal." Both are correct as they imply the same thing.

Why do we make this assumption?

As mentioned in the previous chapter, in order to know how wrong a best guess might be and/or to set up a confidence interval for some target value, we must estimate the sampling distribution of the characteristic of interest. In the analyses that we perform, the characteristic of interest is almost always the mean. Therefore, we must estimate the sampling distribution of the mean.

The sample, itself, does not provide enough information for us to do this. It gives us a start, but we still have to fill in certain blanks in order to derive the center, spread, and shape of the sampling distribution of the mean. In parametric statistics, we fill in the blanks concerning shape by assuming that the sampling distribution of the mean is normal.

Why do we assume that the sampling distribution of the mean is normal, as opposed to some other shape?

The short and flippant answer to this question is that we had to assume something, and normality seemed as good as any other. This works in undergrad courses; it won't work here.

The long and formal answer to this question relies on **Central Limit Theorem** which says that: *given random and independent samples of N observations each, the distribution of sample means approaches normality as the size of N increases, regardless of the shape of the population distribution*. Note that the last part of this statement removes any conditions on the shape of population distribution from which the samples are taken. No matter what distribution you start with (i.e., no matter what the shape of the population), the distribution of sample means becomes normal as the size of the samples increases. (I've also seen this called "the Normal Law.")

✂ The long-winded, technical version of Central Limit Theorem is this: if a population has finite variance σ^2 and a finite mean μ, then the distribution of sample means (from an infinite set of independent samples of N independent observations each) approaches a normal distribution (with variance σ^2/N and mean μ) as the sample size increases, regardless of the shape of population distribution.

In other words, as long as each sample contains a very large number of observations, the sampling distribution of the mean ***must*** be normal. So if we're going to assume one thing for all situations, it has to be a normal, because the normal is always correct for large samples.

The one issue left unresolved is this: how big does *N* have to be in order for the sampling distribution of the mean to always be normal? The answer to this question depends on the shape of the population from which the samples are being taken. To understand why, we must say a few more things about the normal distribution. As a preview: if the population is normal to start with, than any size sample will work, but if the population is outrageously non-normal, you'll need a decent-sized sample.

The **First Known Property** of the Normal Distribution says that: *given random and independent samples of N observations each (taken from a normal distribution), the distribution of sample means is normal and unbiased (i.e., centered on the mean of the population), regardless of the size of N.*

✂ The long-winded, technical version of this property is: if a population has finite variance σ^2 and a finite mean μ and is normally distributed, then the distribution of sample means (from an infinite set of independent samples of *N* independent observations each) must be normally distributed (with variance σ^2/N and mean μ), regardless of the size of *N*.

Therefore, if the population distribution is normal, then even an N of 1 will produce a sampling distribution of the mean that is normal (by the First Known Property). As the population is made less and less normal (e.g., by adding in a lot of skew and/or messing with the kurtosis), a larger and larger *N* will be required. In general, it is said that Central Limit Theorem "kicks in" at an N of about 30. In other words, as long as the sample is based on 30 or more observations, the sampling distribution of the mean can be safely assumed to be normal.

✂ If you're wondering where the number 30 comes from (and whether it needs to be wiped off and/or disinfected before being used), the answer is this: Take the worst-case scenario (i.e., a population distribution that is the farthest from normal); this is the exponential. Now ask: how big does *N* have to be in order for the sampling distribution of the mean to be close enough to normal for practical purposes (when the population is exponential)? Answer: around 30. Given that empirical distributions are rarely as non-normal as the exponential, the value of 30 is a conservative criterion. (Note: this is a case where extensive computer simulation has proved to be quite useful. No-one ever "proved" that 30 is sufficient; this rule-of-thumb was developed by having a computer do what are called "Monte Carlo simulations" for a month or two.) (Note, also: observed data in psychology and neuroscience are never as "bad" as a true exponential and, so, Ns of 10 or more are almost always enough to correct any problems, but we still talk about 30 to cover every possibility.)

At this point let's stop for a moment and review. 1. Parametric statistics work by making an assumption about the shape of the sampling distribution of the characteristic of interest; the particular assumption that all of our parametric stats make is that the sampling distribution of the mean is normal. (To be clear: we assume that if we took a whole bunch of samples, calculated the mean for each, and then made a plot of these values, the distribution of these means would be normal.) 2. As long as the sample size, *N*, is at least 30 and we're making an inference about the mean, then this assumption must be true (by Central Limit Theory plus some simulations), so all's well if you always use large samples to make inferences about the mean.

The remaining problem is this: we want to make the same assumption(s) for all of our inferential procedures and we sometimes use samples that are smaller than 30. Therefore, as of now, we are not guaranteed to be safe. Without doing more or assuming some more, our procedures might not be warranted when samples are small.

This is where the second version of the Assumption of Normality (caps again) comes in. By the First Known Property of the Normal, if the population is normal to start with, then the means from samples of any size will be normally distributed. In fact, when the population is normal, then even an N of 1 will produce a normal distribution (since you're just reproducing the original distribution). So, if we assume that our populations are normal, then we're always safe when making the parametric assumptions about the sampling distribution, regardless of sample size.

To prevent us from having to use one set of statistical procedures for large (30+) samples and another set of procedures for smaller samples, the above is exactly what we do: we assume that the population is normal. (This removes any reliance on the Monte Carlo simulations.) The one thing about this that (rightfully) bothers some people is that we know -- from experience -- that many characteristics of interest to psychologists are not normal. This leaves us with three options: 1. Carry on regardless, banking on the idea that minor violations of the Assumption of Normality (at the sample-means level) will not cause too much grief -- the fancy way of saying this is "we capitalize of the robustness of the underlying statistical model," but it really boils down to looking away and whistling. 2. Remember that we only need a sample size as big as 30 to guarantee normality if we started with the worst-case population distribution -- viz., an exponential -- and psychological variables are rare this bad, so a sample size of only 10 or so will probably be enough to "fix" the non-normalness of any psych data; in other words, with a little background knowledge concerning the shape of your raw data, you can make a good guess as to how big your samples need to be to be safe (and it never seems to be bigger than 10 and is usually as small as 2, 3, or 4, so we're probably always safe since nobody I know collects samples this small). 3. Always test to see if you are notably violating the Assumption of Normality (at the level of raw data) and do something to make the data normal (if they aren't) before running any inferential stats. The third approach is the one that I'll show you.

Another Reason to Assume that the Population is Normal

Although this issue is seldom mentioned, there is another reason to expand the Assumption of Normality such that it applies down at the level of the individual values in the population (as opposed to only up at the level of the sample means). As hinted at in the previous chapter, the mean and the standard deviation of the sample are used in very different ways. In point estimation, the sample mean is used as a "best guess" for the population mean, while the sample standard deviation (together with a few other things) is used to estimate how wrong you might be. Only in the final step (when one calculates a confidence interval or a probability value), do these two things come back into contact. Until this last step, the two are kept apart.

In order to see why this gives us another reason to assume that populations are normal, note the following two points. First, it is assumed that any error in estimating the population mean is independent of any error in estimating how wrong we might be. (If this assumption is not made, then the math becomes a nightmare ... or so I've been told.) Second, the **Second Known**

Property of the Normal Distribution says that: *given random and independent observations (from a normal distribution), the sample mean and sample variance are independent*. In other words, when you take a sample and use it to estimate both the mean and the variance of the population, the amount by which you might be wrong about the mean is a completely separate (statistically independent) issue from how wrong you might be about the variance. As it turns out, the normal distribution is the only distribution for which this is true. In every other case, the two errors are in some way related, such as over-estimates of the mean go hand-in-hand with either over- or under-estimates of the variance.

Therefore, if we are going to assume that our estimates of the population mean and variance are independent (in order to simplify the mathematics involved, as we do), and we are going to use the sample mean and the sample variance to make these estimates (as we do), then we need the sample mean and sample variance to be independent. The only distribution for which this is true is the normal. Therefore, we assume that populations are normal.

Testing the Assumption of Normality

If you take the idea of "assuming" seriously, then you don't have to test the shape of your data. But if you happen to know that your assumptions are sometimes violated -- which, starting now, you do, because I'm telling you that sometimes our data aren't normal -- then you should probably do something before carrying on.

There are at least two approaches to this. The more formal approach is to conduct a statistical test of the Assumption of Normality (as it applies to the shape of the sample). This is most-often done using either the Kolmogorov-Smirnov or the Shapiro-Wilk Test, which are both non-parametric tests that allow you to check the shape of a sample against a variety of known, popular shapes, including the normal. If the resulting *p*-value is under .05, then we have significant evidence that the sample is not normal, so you're "hoping" for a *p*-value of .05 or above.

✂ Some careful folks say that you should reject the Assumption of Normality if the *p*-value is anything under .10, instead of under .05, because they know that the K-S and S-W tests are not very good at detecting deviations from the target shape (i.e., these tests are not very powerful). I, personally, use the .10 rule, but you're not obligated to join me. Just testing for normality at all puts you in the 99th percentile of all behavioral researchers.

So which test should you use … K-S or S-W? This is a place where different sub-fields of psychology and neuroscience have different preferences and I'll discuss this in class. For now, I'll explain how you can get both using SPSS.

The easiest way to conduct tests of normality (and a good time to do this) is at the same time that you get the descriptive statistics. Assuming that you use **Analyze… Descriptive Statistics… Explore…** to do this, all you have to do is go into the Plots sub-menu and (by clicking Plots on the upper right side of the Explore window) and then put a check-mark next to Normality plots with tests. Now the output will include a section labeled **Tests of Normality**, with both the K-S and S-W findings.

If you would like to try the K-S test now, please use the data in *Demo11A.sav* from the first practicum. Don't bother splitting up the data by Experience; for now, just rerun Explore with Normality plots with tests turned on. The *p*-values for mACC_DS1 are .125 for K-S and .151 for S-W. The *p*-values for mACC_DS5 are .200 for K-S and .444 for S-W. All of this implies that these data are normal (enough) for our standard procedures, no matter which test or criterion you use.

As to what you're supposed to do when your data aren't normal, that's next….

How to calculate effect sizes from published research: A simplified methodology

Monday, April 24, 2017 1:14 PM

How to calculate effect sizes from published research: A simplified methodology

Will Thalheimer
Samantha Cook

A Work-Learning Research Publication

Published August 2002

How to calculate effect sizes from published research articles: A simplified methodology

Will Thalheimer
Work-Learning Research

Samantha Cook
Harvard University

Overview

This article provides a simplified methodology for calculating Cohen's ***d*** effect sizes from published experiments that use *t*-tests and F-tests. Accompanying this article is a Microsoft Excel Spreadsheet to speed your calculations. Both the spreadsheet and this article are available as free downloads at www.work-learning.com/effect_sizes.htm.

Why we use effect sizes

Whereas statistical tests of significance tell us the likelihood that experimental results differ from chance expectations, effect-size measurements tell us the relative magnitude of the experimental treatment. They tell us the ***size*** of the experimental ***effect***. Effect sizes are especially important because they allow us to compare the magnitude of experimental treatments from one experiment to another. Although percent improvements can be used to compare experimental treatments to control treatments, such calculations are often difficult to interpret and are almost always impossible to use in fair comparisons across experimental paradigms.

A simple methodology

Although extensive articles have been written detailing methods for calculating effect sizes from published research articles (e.g., Rosnow & Rosenthal, 1996; Rosnow, Rosenthal, & Rubin, 2000), at least some of us—the first author included—require a simpler approach. This article provides a method to calculate Cohen's ***d*** from both t-tests and some F-tests of significance. Accompanying this article is a Microsoft Excel Spreadsheet that can be used to compute Cohen's ***d*** from published data.

Cohen's ***d*** has two advantages over other effect-size measurements. First, its burgeoning popularity is making it the standard. Thus, its calculation enables immediate comparison

to increasingly larger numbers of published studies. Second, Cohen's (1992) suggestion that effect sizes of .20 are small, .50 are medium, and .80 are large enables us to compare an experiment's effect-size results to known benchmarks. The simple methodology offered below is not new but is drawn from previously published articles, most notably Rosnow and Rosenthal (1996) and Rosnow, Rosenthal, and Rubin (2000). We have simplified the methodology not by changing the formulas and calculations but by discarding as much as possible the jargon and computational rationales typically included in articles written for research audiences. This article is an attempt to provide a practical methodology to enable the calculation of effect sizes.

What is an effect size?

In essence, an effect size is the difference between two means (e.g., treatment minus control) divided by the standard deviation of the two conditions. It is the division by the standard deviation that enables us to compare effect sizes across experiments. Because t-tests and F-tests utilize different measures of standard deviation, two separate calculations are required. You will find it useful to keep this distinction in mind as you read this document and utilize the accompanying spreadsheet.

Table of Contents

Calculating Cohen's *d* from t-tests

(1) $$d = \frac{\bar{x}_t - \bar{x}_c}{s_{pooled}}$$

Key to symbols:

d = Cohen's ***d*** effect size
$\bar{x}$ = mean (average of treatment or comparison conditions)
s = standard deviation
Subscripts: ***t*** refers to the treatment condition and ***c*** refers to the comparison condition (or control condition).

How to calculate:

The article should list the means ($\bar{x}$) of the treatment condition and the comparison condition. Use those numbers in the formula and calculate the pooled standard deviation by using Formula 1a below. After you use Formula 1a, simply finish calculating Formula 1 to get Cohen's ***d***.

(1a) $$s_{pooled} = \sqrt{\frac{(n_t - 1)s_t^{\,2} + (n_c - 1)s_c^{\,2}}{n_t + n_c}}$$

Key to symbols:

s = standard deviation
n = number of subjects
Subscripts: ***t*** refers to the treatment condition and ***c*** refers to the comparison condition (or control condition).

How to calculate:

The article should list the number of subjects (***n***) and the standard deviations (***s***) of the treatment condition and the comparison condition. Use those numbers to make your calculations. If the article does not list the standard deviations, use either Formula 2 or Formula 3 below if possible.

Calculating Cohen's d from t-tests:

When you don't have standard deviations or standard errors.

When an experiment that uses a t-test does not list standard deviations, you can calculate Cohen's ***d*** as follows using the t statistic:

(2) $$d = t\sqrt{\left(\frac{n_t + n_c}{n_t\, n_c}\right)\left(\frac{n_t + n_c}{n_t + n_c - 2}\right)}$$

Key to symbols:

d = Cohen's ***d*** effect size
t = ***t*** statistic
n = number of subjects
Subscripts: ***t*** refers to the treatment condition and ***c*** refers to the comparison condition (or control condition).

How to calculate:

The article should list the ***t*** statistic, which it will usually do, for example, with the following notation: ***t*** (29) = 3.12, where 29 is the degrees of freedom and 3.12 is the ***t*** statistic. The article should also list the number of subjects (***n***) within each condition. Use those numbers to make your calculations. If the article does not list the number of subjects in each condition but does list the total number of subjects—and if you can assume that both conditions have roughly equal numbers of subjects—you can estimate Cohen's ***d*** by using Formula 2a below.

Warning: Some studies using repeated-measure designs (where each subject is measured several times within the same condition) incorrectly use experimental trials, instead of subjects, as the units of analysis. The formulas on this page cannot be used for these studies because the t-statistic is not relevant to the number of subjects (***n***) in the study. These studies are often easy to spot because they have outrageously high degrees of freedom.

(2a) $$d \approx \frac{2t}{\sqrt{n-2}}$$

Calculating Cohen's *d* from t-tests:

When you have standard errors instead of standard deviations.

When an experiment that uses a t-test does not list standard deviations but does list standard errors (SE), you can calculate the standard deviations as follows and then use the resulting numbers in Formula 1a:

(3) $$s = SE\sqrt{n}$$

Key to symbols:

s = standard deviation
SE = standard error
n = number of subjects

How to calculate:

This formula assumes that the article lists the standard error (***SE***) and number of subjects (***n***) within each condition. Use those numbers to make your calculations.

Calculating Cohen's *d* from F-tests

(4) $$d = \frac{\bar{x}_t - \bar{x}_c}{\sqrt{MSE\left(\frac{n_t + n_c - 2}{n_t + n_c}\right)}}$$

Key to symbols:

d = Cohen's ***d*** effect size
$\bar{x}$ = mean (average of treatment or comparison condition)
n = number of subjects
MSE = mean squared error
Subscripts: ***t*** refers to the treatment condition and ***c*** refers to the comparison condition (or control condition).

How to calculate:

If standard deviations are available, use Formulas 1 and 1a above because ***MSE***'s will not produce a precise Cohen's ***d*** when the F-test is a comparison among more than two conditions. Otherwise, continue.

The article should list the means ($\bar{x}$) of the treatment condition and the comparison condition, and the mean squared error (***MSE***). Use those numbers in the formula to get Cohen's ***d***. Be careful to select the correct ***MSE*** if many are listed. Note that only when the F-test numerator degrees of freedom are equal to 1—when the F-test compares one condition to one other condition—will the ***MSE*** produce an exact Cohen's ***d*** effect size. In this case, the F-test is equivalent to a t-test. Selecting other ***MSE***'s may not produce valid results.

Calculating Cohen's *d* from F-tests:

When you don't have MSE's.

When an experiment that uses an F-test does not list the MSE, you can calculate Cohen's ***d*** as follows using the F statistic. This calculation should only be used when the F-test compares one condition to one other condition.

(5) $$d = \sqrt{F\left(\frac{n_t + n_c}{n_t\, n_c}\right)\left(\frac{n_t + n_c}{n_t + n_c - 2}\right)}$$

Key to symbols:

d = Cohen's ***d*** effect size
F = F statistic
n = number of subjects
Subscripts: ***t*** refers to the treatment condition and ***c*** refers to the comparison condition (or control condition).

How to calculate:

This formula can ONLY be used when the F-test compares two conditions (when the first degrees of freedom is equal to one). The article should list the ***F*** statistic, which it will usually do, for example, with the following notation: ***F*** (1,39) = 3.12, where 1 is the degrees of freedom based on the number of conditions, and 39 is the degrees of freedom based on the number of subjects. The article should also list the number of subjects ***(n)*** within each condition.

References

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Rosnow, R. L., & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods, 1*, 331-340.

Rosnow, R. L., Rosenthal, R., & Rubin, D. B. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science, 11*, 446-453.

How to cite this article

Thalheimer, W., & Cook, S. (2002, August). *How to calculate effect sizes from published research articles: A simplified methodology.* Retrieved November 31, 2002 from http://work-learning.com/effect_sizes.htm.

(NOTE: You should replace the fictional November 31 date with the date on which the article was downloaded.)

Acknowledgements

We would like to thank Allison Stieber for copyediting this document and Don Rubin for supporting the second author's involvement in this effort.

06 - Simple Comparisons

Monday, April 24, 2017 1:04 PM

UFMG

UNIVERSIDADE FEDERAL DE MINAS GERAIS

Design and Analysis of Experiments

06 - Simple Comparisons

Version 2.11

Felipe Campelo
http://www.cpdee.ufmg.br/~fcampelo

Graduate Program in Electrical Engineering

Belo Horizonte
March 2015

"Science is simply common sense at its best, that is, rigidly accurate in observation, and merciless to fallacy in logic."

Thomas H. Huxley
1825-1895
English biologist

Image: http://www.iep.utm.edu/huxley/

Simple Comparative Experiments

Statistical inference for two samples

The concepts of comparison between two populations based on information obtained from their samples follow the same principles used for testing hypotheses about a single population;

Inferences for two samples frequently arise when comparing the effect of a technique (treatment) against a *control group*: placebo, classical technique, random search, etc;

Usual questions involve:

- Comparison of means;
- Comparison of variances;
- Comparison of proportions;
- etc.

Comparison of two means

Example: Length of steel rods

One of the critical aspects of manufacturing steel rods is cutting the bars with a precise length, which is expected by the customers.

This process is prone to errors, which result in additional costs for standardizing and reprocessing the rods.

An engineer is interested in comparing the current controller of the cutting scissors with a new method that could potentially improve the performance of the process.

Adapted from D.F. Carvalho Jr.'s course project for the Design and Analysis of Experiments Course, PPGEE-UFMG, June 2012. The data used in this example is not necessarily the original one.

Image: http://www.shutterstock.com/pic-73207399/

Comparison of two means
Example: Length of steel rods

A possible statistical model for this kind of data would be:

$$y_{ij} = \mu_i + \epsilon_{ij} \begin{cases} i = 1, 2 \\ j = 1, \ldots, n_i \end{cases}$$

Lets initially assume that the residuals ϵ_{ij} are iid $\mathcal{N}(0, \sigma_i^2)$, which implies:

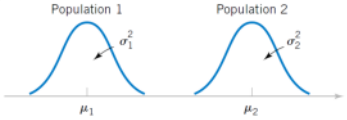


Image: D.C.Montgomery,G.C. Runger, *Applied Statistics and Probability for Engineers*,Wiley 2003.

Comparison of two means
Definitions

What we wish is to perform an inference about the difference in the mean values of constructive deviations for the two controllers. In this case, a reasonable response variable would be the *absolute error*, from which we could test our hypotheses.

The statistical hypotheses can be stated as:

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases} \quad \textbf{or, equivalently,} \quad \begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

Suppose a desired significance level $\alpha = 0.05$, and that the engineer is interested in detecting any difference larger than $15mm$ in the mean absolute error with a power $(1 - \beta) = 0.8$.

Also, lets assume that the variance of the process is unknown but similar for both controllers.

Seria mais adequado testar se o novo método é melhor que o atual:

$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 > \mu_2 \end{cases}$ ⇒ Teste unilateral dá maior sensibilidade p/ o teste

(erro absoluto atual)

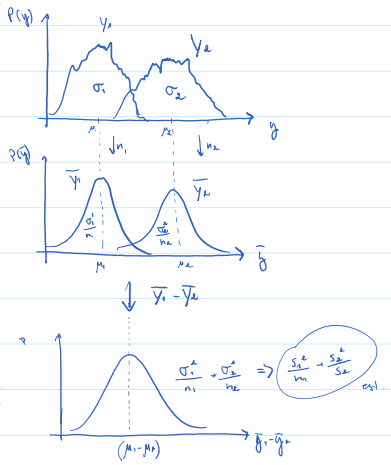
Comparison of two means
Definitions

Since the variance is unknown, it will have to be estimated from the data. As we are assuming $\sigma_1^2 \approx \sigma_2^2$, we can use the pooled variance estimator:

$$S_p^2 = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2} = wS_1^2 + (1 - w) S_2^2$$

Based on this estimator and the stated assumptions, we have that:

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1 + n_2 - 2)}$$



$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \Rightarrow \left(\frac{S_1^2}{n_1} + \frac{S_2^2}{S_2}\right)$ est.

Comparison of two means
Rejection threshold

If we recall our working hypotheses:

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases}$$

we have that, under H_0:

$$t_0 = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)^{*0}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(\bar{y}_1 - \bar{y}_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1 + n_2 - 2)}$$

We'll reject H_0 at the $(1 - \alpha)$ confidence level if $|t_0| \geq t_{\alpha/2, (n_1 + n_2 - 2)}$

$1 - \frac{\alpha}{2}$

$\frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t^{dif.}$

$t^{min\,n_1+n_2-2}_{\alpha/2}$ $\quad$ $t^{min\,n_1+n_2-2}_{1-\alpha/2}$

→ teste bilateral

→ teste unilateral

↓ se $\sigma_1^2 \approx \sigma_2^2$, $S_1^2 \approx S_2^2 \approx S_p^2$

$T = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $\quad \overset{H_0: T}{\Rightarrow} \quad \frac{(\bar{y}_1 - \bar{y}_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1 + n_2 - 2)}$

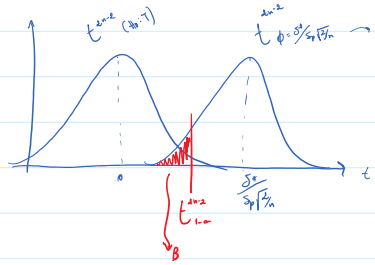
Comparison of two means
Sample sizes

Now recall that the process engineer was interested in some very specific characteristics for his test:

- Significance: $\alpha = 0.05$;
- Power: $(1 - \beta) = 0.8$;
- Minimally interesting effect: $\delta^* = 15mm$

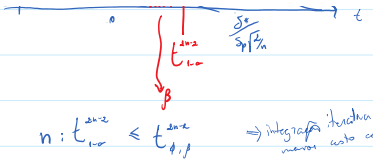
From these specifications, we can obtain the required sample sizes. The derivation of the sample size formulas is not particularly difficult, but we'll concentrate only on the results. More details can be easily found in the literature[a].

→ Estimativa de (n) com dist. balanceada $(n_1 = n_2 = n)$



The derivation of the sample size formulas is not particularly difficult, but we'll concentrate only on the results. More details can be easily found in the literature[a].

[a] Check, for instance, Paul Mathews' *Sample Size Calculations*, MMB, 2010.



$n : t^{2n-2}_{1-\alpha} \leqslant t^{2n-2}_{\phi,\beta}$ ⇒ integração iterativa, menor custo computacional

Comparison of two means
Sample sizes

For the general case of unequal sample sizes, we have:

$$n_1 = \left(1 + \frac{n_1}{n_2}\right)\left(\frac{s_p}{\delta^*}\right)^2 (t_{\alpha/2} + t_\beta)^2$$

where s_p is the estimated common standard deviation, and $t_{\alpha/2}$ and t_β are the $\alpha/2$ and β quantiles of the $t_{(n1+n2-2)}$ distribution. The sample size n_2 can be calculated by simply substituting (n_1/n_2) by (n_2/n_1).

aproximadas

For equal sample sizes ($n_1 = n_2 = n$) the expression is simplified to:

$$n = 2\left(\frac{s_p}{\delta^*}\right)^2 (t_{\alpha/2} + t_\beta)^2$$

* Facas Gavinsky e meias vivalina

Comparison of two means
Sample sizes

These formulas are very convenient, but leave us with a riddle: we need variance estimate in order to calculate the sample size, but we need observations to be able to estimate the variance.

There are a few ways to proceed in this case. The most practical are:

- Use process knowledge or historical data to obtain an (initial) estimate of the variance;
- Perform a pilot study and collect samples to estimate the variance.

The first method is almost always preferable since it does not imply additional costs for the experiment.

- Experimentação iterativa: partindo de um valor inicial, realizar experimentos até atingir os valores desejados
- Expressar $d = \delta^*/\sigma \Rightarrow n \geq 2\left(\frac{1}{d}\right)^2 \left(t^{2n-2}_{\alpha} + t^{2n-2}_{\beta}\right)^2$ → Pode prejudicar replicabilidade → evitado quando é muito caro ou demorado em cond. experimentais definidas

Comparison of two means
Sample sizes

If no information is available to estimate the variance, a pilot study must be performed to obtain this value. The sample size required for this pilot study is given by:

$$n_{pilot} \approx 2\left(\frac{z_{\alpha_n/2}}{e_n}\right)^2$$

$\alpha_n = 0.05$, $e_n = 0.1$, $z_{\alpha/2} \approx 2$ ⇒ $2\left(\frac{2}{0.1}\right)^2 = 800$ → muito maior que o esperado para o experimento em si

where $(1 - \alpha_n)$ is the desired confidence level for the sample size estimate of the main study, and e_n is the maximum relative error allowed for the sample size.

This calculation can yield some scarily large sample sizes for a pilot study (much larger than would be actually required for the main study itself), so use this with caution.

Comparison of two means
Sample sizes

For the steel rods experiment, suppose that the engineer uses data available from the controller manuals, as well as historical measurements, to estimate the common standard deviation for the cutting process as $\hat{\sigma} \approx 15mm$.

Assuming that equal sample sizes are desired, we can simply use the formula:

$$n = 2\left(\frac{\hat{\sigma}}{\delta^*}\right)^2 (t_{\alpha/2} + t_\beta)^2$$

Easy, right?

2, se teste pareado

α, se teste unilateral

Comparison of two means
Sample sizes

For the steel rods experiment, suppose that the engineer uses data available from the controller manuals, as well as historical measurements, to estimate the common standard deviation for the cutting process as $\hat{\sigma} \approx 15mm$.

Assuming that equal sample sizes are desired, we can simply use the formula:

$$n = 2\left(\frac{\hat{\sigma}}{\delta^*}\right)^2 (t_{\alpha/2} + t_\beta)^2$$

Easy, right?

Comparison of two means
Sample sizes

The last problem we have to solve is that the values of $t_{\alpha/2}$ and t_β are also dependent of n, which makes the equation

$$n = 2\left(\frac{\hat{\sigma}}{\delta^*}\right)^2 (t_{\alpha/2} + t_\beta)^2$$

transcendental in n. We'll have to iterate until we find the smallest n that satisfies:

$$n \geq 2\left(\frac{\hat{\sigma}}{\delta^*}\right)^2 (t_{\alpha/2} + t_\beta)^2$$

Usually $t_{\alpha/2} \approx z_{\alpha/2}$ is used for the first iteration. Easy, right?

Comparison of two means
Sample sizes

The last problem we have to solve is that the values of $t_{\alpha/2}$ and t_β are also dependent of n, which makes the equation

$$n = 2\left(\frac{\hat{\sigma}}{\delta^*}\right)^2 (t_{\alpha/2} + t_\beta)^2$$

transcendental in n. We'll have to iterate until we find the smallest n that satisfies:

$$n \geq 2\left(\frac{\hat{\sigma}}{\delta^*}\right)^2 (t_{\alpha/2} + t_\beta)^2$$

Usually $t_{\alpha/2} \approx z_{\alpha/2}$ is used for the first iteration. Easy, right?

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

Comparison of two means
Example: Length of steel rods

Required sample size:

```
> ss.calc<-power.t.test(delta=15,
                        sd=15,
                        sig.level=0.05,
                        power=0.8,
                        type="two.sample",
                        alternative="two.sided")

Two-sample t test power calculation
n = 16.71477
delta = 15
sd = 15
sig.level = 0.05
power = 0.8
alternative = two.sided

NOTE: n is number in *each* group
```

Comparison of two means
Example: Length of steel rods

Computationally, we can perform the t-test for comparing the means of two independent populations by:

```
> y<-read.table("../data files/steelrods.txt",
+               header=T)

> with(y,
+      t.test(Length.error~Process,
+             alternative = "two.sided",
+             mu = 0,
+             var.equal = TRUE,
+             conf.level = 0.95))

Two Sample t-test
data:  Length.error by Process
t = -14.312, df = 32, p-value = 1.849e-15
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.09272982 -0.06962312
sample estimates:
mean in group new mean in group old
       0.07782353        0.15900000
```

notação de fórmula: (y~x) variável de resposta ~ variável de medida

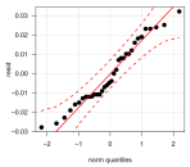
confere: df = 2n-2

Comparison of two means
Example: Length of steel rods

The assumptions of the test must be verified. In this particular case:

- Normality of the residuals; → Normalidade dos MEDIAS dos resíduos
- Equality of variance of the residuals;
- Independence of the residuals.

```
> resid<-y$Length.error - rep(means[2:1],
+                             each=n)

> shapiro.test(resid)
Shapiro-Wilk normality test
data:  resid
W = 0.9552, p-value = 0.176

> library(car)
> qqPlot(resid,
+        pch=16,
+        cex=1.5,
+        las=1)
```



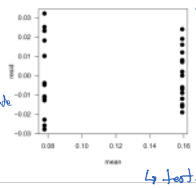
Comparison of two means
Example: Length of steel rods

The assumptions of the test must be verified. In this particular case:

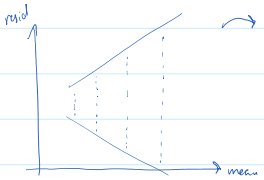
- Normality of the residuals;
- Equality of variance of the residuals;
- Independence of the residuals.

```
> with(y,
+      fligner.test(Length.error~Process))
Fligner-Killeen test of homogeneity of variances
data:  Length.error by Process
Fligner-Killeen:med chi-squared = 1.6837,
df = 1, p-value = 0.1944

> plot(x=rep(means[2:1],
+            each=n),
+      y=resid,
+      xlab="mean",
+      pch=16,
+      cex=1.5,
+      las=1)
```

não há indícios de desigualdade das variâncias



↳ espalhamento com para diferenças



↳ indicador de heterocedastidade (dependência da variância da média)

↓ requer/sugere transf. nos dados (ex: log)

↳ test-t é relativamente robusto a pequenas variações

Comparison of two means
Example: Length of steel rods

The assumptions of the test must be verified. In this particular case:

- Normality of the residuals;
- Equality of variance of the residuals;
- Independence of the residuals.

As mentioned in an earlier lecture, there is no general test for the independence assumption, and it has to be guaranteed in the design phase.

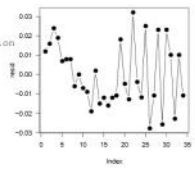
One can at most test for serial autocorrelation in the residuals using Durbin-Watson's test, but this test is absolutely dependent on the ordering of the observations - very useful to detect ordering-related trends in the residuals, but not much more than that.

Comparison of two means
Example: Length of steel rods

The assumptions of the test must be verified. In this particular case:

- Normality of the residuals;
- Equality of variance of the residuals;
- Independence of the residuals.

```
> library(lmtest)
> with(y,
+      dwtest(Length.error~Process))
Durbin-Watson test
data:  Length.error ~ Process
DW = 2.2215, p-value = 0.6838
alternative hypothesis: true autocorrelation
is greater than 0

> plot(resid,
+      pch=16,
+      cex=1.5,
+      type="b",
+      las=1)
```



Comparison of two means
Unequal variances

Suppose now a more general case, in which the variances of the two populations are unknown and cannot be assumed equal.

For this cases, a modification on the t-test called *Welch's t test* is usually employed. The Welch statistic can be calculated as:

$$t_0^* = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Under the null hypothesis t_0^* is distributed approximately as a t_ν distribution, with:

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

graus de liberdade

Premissas: 1) Independência → design e coleta
2) Norm. da dist. das médias (não dos dados!) → investigada com bootstrap

↓
fundamenta vários testes futuros

solução aproximada, ligeiramente conservador

menor df ⇒ caudas mais largas na dist. de Student ⇒ mais conservador

→ Premissas abrandadas ⇒ teste mais geral com menor sensibilidade

Bibliography

Required reading

1. D.C. Montgomery, G.C. Runger, *Applied Statistics and Probability for Engineers*, Ch. 10. 5th ed., Wiley, 2010.; **OR**
2. D.C. Montgomery, *Design and Analysis of Experiments*, Ch. 2. 5th ed., Wiley, 2005;
3. R. Nuzzo, *Scientific method: Statistical errors*, Nature 506(7487) - http://goo.gl/Ebq6Rc

Recommended reading

1. P. Mathews, *Sample Size Calculations: Practical Methods for Engineers and Scientists*, Ch. 1-2. 1st ed.. MMB, 2010.
2. Radiolab (podcast): http://zaais.sb.org

difícil de encontrar

↳ ciência, arte, sociedade ...

↳ TED Radio Hour ⇒ podcast da TED

About this material

Conditions of use and referencing

Please reference this work as:

Felipe Campelo (2015), *Lecture Notes on Design and Analysis of Experiments*.
Online: https://github.com/fcampelo/Design-and-Analysis-of-Experiments
Version 2.11, Chapter 6; Creative Commons BY-NC-SA 4.0.

```
@M{oc:Campelo2015-DE,
  title={Lecture Notes on Design and Analysis of Experiments},
  author={Felipe Campelo},
  howPublished={\url{https://github.com/fcampelo/Design-and-Analysis-of-Experiments}},
  year={2015},
  note={Version 2.11, Chapter 6; Creative Commons BY-NC-SA 4.0.},
}
```

07 - Paired Design

Monday, April 24, 2017 3:49 PM

UFMG
UNIVERSIDADE FEDERAL DE MINAS GERAIS

Design and Analysis of Experiments
07 - Paired Design

Version 2.11

Felipe Campelo
http://www.cpdee.ufmg.br/~fcampelo

Graduate Program in Electrical Engineering

Belo Horizonte
March 2015

"I am driven by two main philosophies: know more today about the world than I knew yesterday and lessen the suffering of others. You'd be surprised how far that gets you."

Neil deGrasse Tyson
1958 -
American astrophysicist and author.

Image: http://goo.gl/hzsFUJ

Comparison of two means
Dependent populations

Suppose the following situation: a young researcher develops an optimization algorithm (A) for a given family of problems, and wants to compare its convergence speed against a method that represents the state-of-the-art (B).

The researcher implements both methods and wants to determine whether the proposed one has a better average performance for problems of that particular family, represented by a given benchmark set.

The measurements are made under homogeneous conditions (same computer, same operational conditions, etc.) and the time is measured in a way that is not sensitive to other processes running in the system.

Image: http://goo.gl/xwqiq9

Comparison of two means
Dependent populations

This problem has some important questions worth considering:

- What is the actual question of interest? → Método **A** melhor que **B** ?
- What is the *population* for which that question is relevant? → Classe/família de problemas (não apenas as instâncias testadas)
- What are the independent observations for that population? → Obs. p/ cada instância de prob. na família
  ↳ Mesmas execuções p/ mesma instância não são independentes
- What is the relevant sample size for the experiment?

Inst.	A	B
1	x	x
2	x	x
⋮	x	x
n	x	x

$y_{ij} = \mu_i + \tau_j + \varepsilon_{ij}$ (média do algoritmo i; efeito da instância j; erro)

→ $y_{Aj} = \mu_A + \tau_j + \varepsilon_{Aj}$, $y_{Bj} = \mu_B + \tau_j + \varepsilon_{Bj}$

Se τ_j não é modelado, ele é incluído nos resíduos → não há independência

diferença entre instâncias não é considerada

Solução:

$$D_j = (y_{Aj} - y_{Bj}) = (\mu_A - \mu_B) + \underbrace{(\varepsilon_{Aj} + \varepsilon_{Bj})}_{\varepsilon_j}$$

Pareamento

Comparison of two means
Paired design

The variability due to the different test problems is a strong source of spurious variation that can and must be controlled;

An elegant solution to eliminate the influence of this nuisance parameter is the *pairing* of the measurements by problem:

- Observations are considered in pairs (A, B) for each problem;
- Hypothesis testing is done on the sample of *differences*;

$$\begin{cases} H_0 : \mu_D = 0 \\ H_1 : \mu_D < 0 \end{cases}$$

↳ nº de obs. ind. = nº inst.

Comparison of two means
Paired design

Let y_{Aj} and y_{Bj} denote paired observations of average time for methods A and B, for each problem instance j. The *paired differences* of the observations are simply $d_j = y_{Aj} - y_{Bj}$.

If we model our observations as an additive process:

$$y_{ij} = \underbrace{\mu + \tau_i}_{\mu_i} + \beta_j + \varepsilon_{ij}$$

where μ is the grand mean, τ_i is the effect of the i-th algorithm on the mean, β_j is the effect of the j-th problem, and ε_{ij} is the model residual, then:

$$\begin{aligned} d_j &= (\mu + \beta_j - \mu - \beta_j) + \tau_A - \tau_B + \varepsilon_{Aj} - \varepsilon_{Bj} \\ &= \mu_D + \varepsilon_j \end{aligned}$$

($\mu + \beta_j - \mu - \beta_j \to 0$)

Comparison of two means
Paired design

The hypotheses of interest can now be defined in terms of μ_D, e.g.:

$$\begin{cases} H_0 : \mu_D = 0 \\ H_1 : \mu_D \neq 0 \end{cases}$$

which can now be treated as a test of hypotheses for a single sample: the population of interest is the differences in average times until convergence for the problems under investigation. The test statistic is given by:

$$T_0 = \frac{\bar{D}}{S_D/\sqrt{N}}$$

which is distributed under the null hypothesis as a Student-t variable with $N - 1$ degrees of freedom (where N is the number of test problem instances in the experiment);

Comparison of two means
Paired design

Some other important questions worth considering:

- In this example the minimally interesting effect size δ^* must be expressed in terms of *average time gains across problems* (not within individual instances);
- The most important sample size to consider in this situation refers to the *number of problem instances*, and not necessarily to the number of within-problems repeated measures;
- The number of repetitions within each problem will have an impact on the uncertainty associated to each observation (that is, to each value of mean time to convergence for each algorithm on each problem), and should be selected with some care[a].

[a] Alternatively, we can set it as arbitrarily large, particularly in cases where the cost of repetitions is small. As much as I hate to admit it, the lazy heuristic of setting it as ≥ 30 should be enough in most algorithmic studies. A more methodologically sound approach to setting this is under development, and will be included in future versions of these lecture notes.

Fernanda Takahashi

1) Ajuda interpretar efeitos em cada instâncias útil p/ estudo de parâmetros
2) Pode ser útil para convencer revisores
↳ Maximizar nº de instâncias é mais útil!

Comparison of two means
Paired design

Some other important questions worth considering:

- Pairing removes the effects of controllable nuisance factors from the analysis.
- Strongly indicated in cases with strong correlations between samples (e.g., heterogeneous experimental conditions).

↳ e.g. instâncias diferentes

Comparison of two means
Paired design

Going back to our example, assume the following facts about the desired comparison:

- The benchmark set is composed of seven problems ($N = 7$);
- The researcher is interested in finding differences in mean time to convergence greater than ten seconds ($\delta^* = 10$) with a power of at least $(1 - \beta) = 0.8$, using a significance level $\alpha = 0.05$;
- The researcher performs $n = 30$ repeated runs[b] of each algorithm in each problem, from random initial conditions.

[b] Not that I necessarily recommend this number, but it is generally an easy alternative if you don't want to keep justifying your choices to less statistically-savvy reviewers.

Comparison of two means
Paired design

Step 1: load and precondition the data.

```
> # Read data
> data<-read.table("../data files/soltimes.csv",
+                  header=T)

# "Problem" is a categorical variable, not a continuous one
> data$Problem<-as.factor(data$Problem)

# Summarize within-problem observations by mean
> aggdata<-aggregate(Time~Problem:Algorithm,
+                    data=data,
+                    FUN=mean)
> summary(aggdata)
 Problem Algorithm      Time
 1:2     A:7       Min.   : 37.63
 2:2     B:7       1st Qu.:109.45
 3:2               Median :178.73
 4:2               Mean   :175.48
 5:2               3rd Qu.:245.25
 6:2               Max.   :296.79
 7:2
```

Comparison of two means
Paired design

Step 2: analysis

```
> # Perform paired t-test
> t.test(Time~Algorithm,
+        paired=T,
+        data=aggdata)

Paired t-test
data:  Time by Algorithm
t = -9.1585, df = 6, p-value = 9.54e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -21.85862 -12.64118
sample estimates:
mean of the differences
                -17.2499
```

→ tempo modelado por algoritmo

Comparison of two means
Paired design

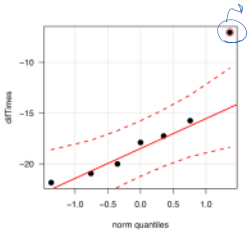
Alternatively, we could have done:

```
> difTimes<-with(aggdata,
+              Time[1:7]-Time[8:14])
> t.test(difTimes)


One Sample t-test
data:  difTimes
t = -9.1585, df = 6, p-value = 9.54e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -21.85862 -12.64118
sample estimates:
mean of x
 -17.2499
```

Comparison of two means
Paired design

Verify assumptions:

```
> shapiro.test(difTimes)

Shapiro-Wilk normality test
data:  difTimes
W = 0.8387, p-value = 0.09655

# Redo test without outlier
> indx<-which(difTimes==max(difTimes))
> t.test(difTimes[-indx])$p.value
[1] 6.179743e-06
> t.test(difTimes[-indx])$conf.int
[1] -21.41856 -16.48037
```



→ pode ser removido temporariamente p/ verificar se há diferença

Comparison of two means
Paired design

What happens if we fail to consider the problem effects?

```
> t.test(Time~Algorithm,data=aggdata)

Welch Two Sample t-test
data:  Time by Algorithm
t = -0.3609, df = 11.993, p-value = 0.7245
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -121.40320   86.90341
sample estimates:
mean in group A mean in group B
       166.8527        184.1026
```

diminui

variância entre instâncias foi p/ resíduo

. Independência violada ⇒ taxas de erro aumentam (no ex., erro II)

Comparison of two means
Paired design

Paired designs can require smaller sample sizes for equivalent power in cases where the between-units (in our example, the between-problems) variation is relatively high;

More specifically, if the within-level variation is given by σ_ϵ and the between-units variation is σ_u, we have that, for large enough N (e.g., $N \geq 10$),

$$\frac{N_{\text{unpaired}}}{N_{\text{paired}}} \approx \sqrt{2\left[\left(\frac{\sigma_u}{\sigma_\epsilon}\right)^2 + 1\right]}$$

Failure to consider inter-unit variability can result in the masking of relevant effects by the nuisance factor.

Similarly, failure in recognizing the dependence structure of within-unit measurements yields tests with artificially inflated degrees of freedom, which results in the inflation of the effective value of α.

Bibliography

Required reading

1. D.C. Montgomery, G.C. Runger, *Applied Statistics and Probability for Engineers*, Ch. 10. 5th ed., Wiley, 2010.
2. M.J. Crawley, *The R Book*, Ch. 8. 1st ed., Wiley, 2007;

Recommended reading

1. L. Lehe and V. Powell, *Simpson's Paradox* - http://vudlab.com/simpsons/
2. J.P. Simmons, L.D. Nelson, and U. Simonsohn, *False-Positive Psychology : Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant*, Psychological Science 22(11):1359-1366, 2011 - http://goo.gl/9e0cdw

About this material
Conditions of use and referencing

Please reference this work as:

Felipe Campelo (2015), *Lecture Notes on Design and Analysis of Experiments.* Online: https://github.com/fcampelo/Design-and-Analysis-of-Experiments Version 2.11, Chapter 7; Creative Commons BY-NC-SA 4.0.

```
@Misc{Campelo2015-01,
  title={Lecture Notes on Design and Analysis of Experiments},
  author={Felipe Campelo},
  howPublished={\url{https://github.com/fcampelo/Design-and-Analysis-of-Experiments}},
  year={2015},
  note={Version 2.11, Chapter 7; Creative Commons BY-NC-SA 4.0.},
}
```

08 - Testing Equivalence and Non-Inferiority

Monday, May 8, 2017 12:56 PM

UFMG
UNIVERSIDADE FEDERAL DE MINAS GERAIS

Design and Analysis of Experiments
08 - Testing Equivalence and Non-Inferiority

Version 2.11

Felipe Campelo
http://www.cpdee.ufmg.br/~fcampelo

Graduate Program in Electrical Engineering

Belo Horizonte
April 2016

" Science makes people reach selflessly for truth and objectivity; it teaches people to accept reality, with wonder and admiration, not to mention the deep awe and joy that the natural order of things brings to the true scientist."

Lise Meitner
1878 - 1968
Austrian Physicist

Image: http://www.alltomvetenskap.se/nyheter/vem-var-86

Testing equivalence
Introduction

The tests introduced in the preceding chapters deal with situations in which one is interested in detecting *differences* between a population parameter θ – e.g., a population mean μ or a difference between population means $(\mu_1 - \mu_2)$ – and its nominal value θ_0 under a null hypothesis;

Another useful class of experiments in engineering and science is one in which the experimenter is interesting in investigating *equivalence* (within a given margin of error), for instance:

- Conformity/compliance testing (industrial certification);
- Equivalence of effects (pharmaceutical industry);

Image adapted from: http://goo.gl/iXeCiY

Testing equivalence
Introduction

In principle, one could express this as a shift in focus from trying to establish whether a population parameter is different from a given reference to trying to determine whether it is equal to that reference.

In usual (two-sided) comparative studies, the alternative hypothesis (i.e., the one that presents novelty in relation to the current state of knowledge) is the one of difference between the parameters of interest - that is, unless there is strong evidence of differences, one cannot rule out the null hypothesis of equality;

Testing equivalence
Introduction

In equivalence testing, the situation is reversed: the (approximate) equality of two parameters is the novelty one hopes to establish. Consequently, the burden of proof shifts to providing evidence that there is no difference.

The term *equivalent* is not used strictly, but to mean the absence of practical differences - that is, any differences that might exist fall within an *equivalence margin* or *limit of practical significance* δ^*.

Using this approach, the equivalence of two parameters can be established if a sample provides enough evidence that the true difference is smaller than δ^* units.

Testing Non-inferiority
Definition

A similar concept to equivalence testing is the definition of non-inferiority of a given treatment/ process/ method in relation to another (e.g., a standard solution).

In non-inferiority tests, one can declare that a given process is not worse than a standard one only if enough evidence is provided to conclude that the performance of the proposed process is no more than δ^* units worse than that of the standard.

In the case of non-inferiority tests, one can in principle use a regular test of differences with a one-sided alternative (which would be equivalent to setting $\delta^* = 0$), or define the null hypothesis in a way that includes δ^* in its formulation.

Comparison of studies

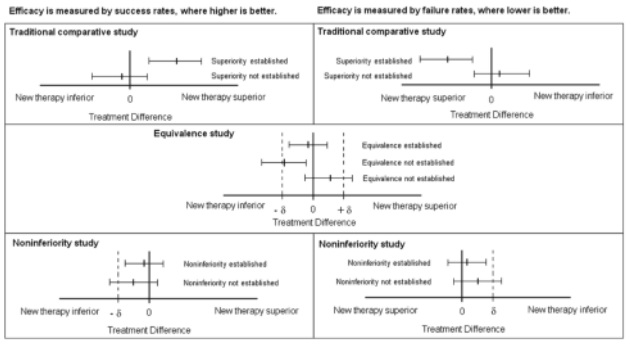


Image: Walker and Nowacki (2011), J. General Internal Medicine 26(2):192-196.

Testing Equivalence

Quick-and-dirty approach

A simple way of thinking about testing equivalence of two means is to observe confidence intervals instead of p-values:

> "*Equivalence can be established at the α significance level if a $(1-2\alpha)$-confidence interval for the difference between the two means is contained within a interval $\pm\delta^*$.*"

The difference between testing for differences and for equivalence can be easily illustrated using this approach:

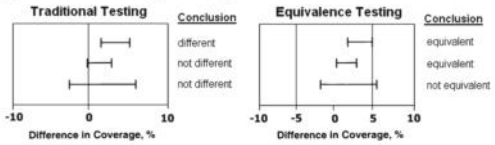


Image: Walker and Nowacki (2011), J. General Internal Medicine 26(2):192-196.

Equivalence test for a single mean

Hypotheses

An equivalence test for a single population mean can be expressed by the hypotheses:

$$\begin{cases} H_0 : |\mu - \mu_0| = & \Delta\mu \geq \delta^* \\ H_1 : & \Delta\mu < \delta^* \end{cases}$$

The most usual way of testing these hypotheses is the TOST (*two one-sided tests*) method. As the name suggests, two one-sided significance tests are constructed so that the desired statistical properties can be achieved. Using our standard notation:

$$\begin{cases} H_0^1 : & \Delta\mu = -\delta^* \\ H_1^1 : & \Delta\mu > -\delta^* \end{cases} \qquad \begin{cases} H_0^2 : & \Delta\mu = \delta^* \\ H_1^2 : & \Delta\mu < \delta^* \end{cases}$$

If both tests reject their respective H_0, then equivalence (within the equivalence margin δ^*) can be declared with significance level α.

Equivalence test for a single mean
Graphical interpretation

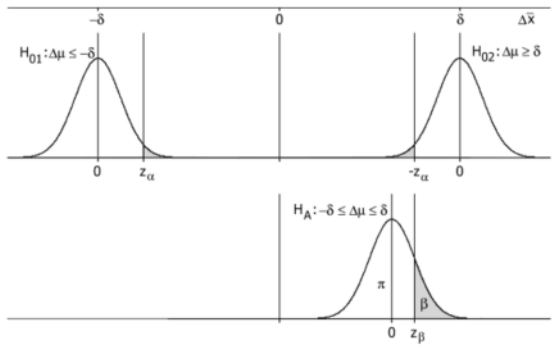


Image: Matthews (2010), Sample Size Calculations, MMB. pg. 46

Equivalence of a single mean
Sample size

Sample sizes for testing equivalence of a single mean can be derived using essentially the same considerations used for the usual tests. In the case of a single sample:

$$n \geq \left(\frac{(t_\alpha + t_\beta)\hat{\sigma}}{\delta^* - \Delta\mu} \right)^2$$

As in the previous cases, iteration is needed to solve for n (since the quantiles of the t distribution depend on n). Use $t_x = z_x$ for the first iteration.

Equivalence of two means
Hypotheses

Analogously to the single sample test of equivalence, the hypotheses for testing the equivalence of two population means can be described as:

$$\begin{cases} H_0: & \mu_1 - \mu_2 \geq \delta^* \\ H_1: & \mu_1 - \mu_2 < \delta^* \end{cases}$$

$$\begin{cases} H_0^1: & \mu_1 - \mu_2 = -\delta^* \\ H_1^1: & \mu_1 - \mu_2 > -\delta^* \end{cases} \qquad \begin{cases} H_0^2: & \mu_1 - \mu_2 = \delta^* \\ H_1^2: & \mu_1 - \mu_2 < \delta^* \end{cases}$$

Just as in the previous case, both hypotheses are tested at the desired α value, and the rejection of both H_0 indicates evidence of equivalence.

Equivalence of two means
Sample size

Sample size for the $n_1 = n_2 = n$ case can be approximated based on the Zhang formula[a]:

$$n \geq (t_{\alpha;\nu} + t_{(1-c)\beta;\nu})^2 \left(\frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{\delta^* - \Delta\mu^*} \right)^2$$

with $\Delta\mu^* < \delta^*$ as the maximum real difference between the two means for which a power of $(1 - \beta)$ is desired, and:

$$c = \frac{1}{2} \exp\left(-7.06 \frac{\Delta\mu^*}{\delta^*} \right)$$

The degrees of freedom ν of the t-quantiles are given by the Welch t-test formula (see Chapter 6).

[a] Zhang (2003), J. Biopharm. Stat. 13(3):529-538.

Example
Laboratory certification

A ballistics laboratory is in the process of being certified for the evaluation of shielding technology, and needs to provide evidence of equivalence of a given callibration procedure with the reference equipment;

The certification authority demands that the mean hole area generated by this procedure in the lab be the same as the one from the reference equipment, and tolerates deviations no greater than $4mm^2$;

From previous measurements, the standard deviations can be roughly estimated as $\hat{\sigma}_{Lab} = 5mm^2$ and $\hat{\sigma}_{ref} = 10mm^2$.

The desired error levels for the comparison are $\alpha = 0.01$ and $\beta = 0.1$.

Image: http://www.everydaynodaysoff.com/2013/08/05/ballistic-shield-for-operators-only/

Example
Laboratory certification

To calculate the required sample size, assume that $\Delta\mu^* = 0.5$. Then:

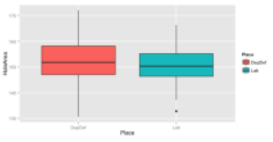
```
> # load functions to calculate sample size for TOST
> source("calcN_tost.R")
>
> # Calculate sample size
> calcN_tost2(alpha = 0.01,
+             beta = 0.1,
+             diff_mu = 0.5,
+             tolmargin = 4,
+             s1 = 5,
+             s2 = 10)
[1] 144.1999
```

We'll need 145 observations from each group to test for equivalence with the desired experimental properties.

Example
Laboratory certification

After collecting the observations, we proceed to the analysis:

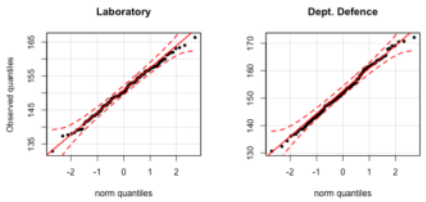
```
> data<-read.table("../data files/labdata-example.csv",
+                  header = T, sep = ",")

> # Two one-sided t-tests
> t.test(HoleArea~Place,  data = data,  alternative = "less", mu = 4,
+        conf.level = 0.99)$p.value
[1] 0.00304124
> t.test(HoleArea~Place, data = data, alternative = "greater", mu = -4,
+        conf.level = 0.99)$p.value
[1] 6.586193e-10

> # Get (1-2*alpha) CI
> t.test(HoleArea~Place, data = data,  conf.level = 0.98)$conf.int
[1] -0.5117627  3.6244386
```



Example
Laboratory certification

Verification of test assumptions:

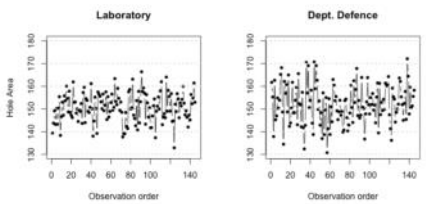
```
> par(mfrow=c(1,2))
> qqPlot(subset(data, Place=="Lab")[,2],
+        pch=20,
+        main = "Laboratory",
+        ylab = "Observed quantiles")
> qqPlot(subset(data, Place=="DepDef")[,2],
+        pch=20,
+        main = "Dept. Defence",
+        ylab = " ")
```



Example
Laboratory certification

Verification of test assumptions:

```
> dwtest(HoleArea~Place, data=data)
DW = 1.8116, p-value = 0.04757

> par(mfrow=c(1,2))
> plot(seq_along(subset(data, Place=="Lab")[,2]),
+      subset(data, Place=="Lab")[,2], ...)
> plot(seq_along(subset(data, Place=="DepDef")[,2]),
+      subset(data, Place=="DepDef")[,2], ...)
```



Bibliography

Required reading

1. E. Walker, A.S. Nowacki, *Understanding Equivalence and Noninferiority Testing*, Journal of General Internal Medicine 26(2):192-196, 2011.

Recommended reading

1. P. Mathews, *Sample Size Calculations: Practical Methods for Engineers and Scientists*, Ch. 2.4, 1st ed., MMB, 2010.
2. P. Zhang, *A Simple Formula for Sample Size Calculation in Equivalence Studies*, Journal of Biopharmaceutical Statistics 13(3):529-538, 2003.

About this material

Conditions of use and referencing

Please reference this work as:

Felipe Campelo (2015), *Lecture Notes on Design and Analysis of Experiments.* Online: https://github.com/fcampelo/Design-and-Analysis-of-Experiments Version 2.11, Chapter 8; Creative Commons BY-NC-SA 4.0.

```
@Misc{Campelo2015-01,
  title={Lecture Notes on Design and Analysis of Experiments},
  author={Felipe Campelo},
  howPublished={\url{https://github.com/fcampelo/Design-and-Analysis-of-Experiments}},
  year={2015},
  note={Version 2.11, Chapter 8; Creative Commons BY-NC-SA 4.0.},
}
```

Case Study 02

Monday, May 15, 2017 1:22 PM

- Desbalanceamento de amostras não impede a inferência

↓

diferente número de repetições por amostra/algoritmo

* Repetições em uma mesma instância devem ser sumarizadas em um único valor ⇝ Não necessariamente da mesma forma que a análise final! (e.g. pode ser interesse comparar a média do pior caso)

- Teste pareado:

$-\ \mu_D = \mu_A - \mu_B$ (padrão)

teste $\begin{cases} \cdot\ H_0 : \mu_D^t = 0 \\ \cdot\ H_1 : \mu_D^t > 0 \end{cases}$ ⇝ teste unilateral consegue maior potência p/ mesmo (n)

acc $\begin{cases} \cdot\ H_0 : \mu_D^a = \delta_a^* \\ \cdot\ H_1 : \mu_D^a < \delta_a^* \end{cases}$

- Número de amostras:

$$\cdot\ N_t \geq \left(t^{n_1}_{1-\alpha} + t^{n-1}_{1-\beta}\right)^2 \left(\frac{\sigma}{\delta^*}\right)^2$$

$$= \frac{\left(t^{n_1}_{1-\alpha} + t^{n-1}_{1-\beta}\right)^2}{d^{*2}} = (8)$$

→ · Usar p/ piloto
· Obter σ_A, σ_B
· Recalcular (n)

} necessário corrigir (α) devido à aplicação de hipóteses sucessivas

$\cdot\ d = \frac{\delta}{\sigma}$

↳ σ_{w1} (across instances)
↳ σ_{w1} (within instances)

$\sigma^2 = \sigma_{A1}^2 + \sigma_{w1}^2$

↳ $\sigma_{w1} = \sqrt{\frac{\sigma_{Aj}^2}{n} + \frac{\sigma_{Bj}^2}{n}} = \sqrt{\frac{\sigma_{Aj}^2 + \sigma_{Bj}^2}{n}}$

⇓

O número de repetições por instância (n) ajuda diminuir a variância total

- Mauro Birattari (2004-05)
  How many instances and how many runs ?

⇒ Para orçamento de esforço fixo, melhor configuração é $n=1$ e max instâncias

10 - Analysis of Variance

Monday, May 15, 2017 3:38 PM

UFMG
UNIVERSIDADE FEDERAL DE MINAS GERAIS

Design and Analysis of Experiments
10 - Analysis of Variance

Version 2.11

Felipe Campelo
http://www.cpdee.ufmg.br/~fcampelo

Graduate Program in Electrical Engineering

Belo Horizonte
April 2015

"Chance is commonly viewed as a self-correcting process in which a deviation in one direction induces a deviation in the opposite direction to restore the equilibrium. In fact, deviations are not "corrected" as a chance process unfolds, they are merely diluted."

Amos Nathan Tversky
1937-1996
Israeli cognitive and mathematical psychologist

Image: http://upload.wikimedia.org/wikipedia/he/2/2b/Amos_Tversky.jpg

---

Comparison of multiple means
Introduction

In the previous sections, we have (hopefully) developed a solid understanding of the main concepts associated with comparing the means of two groups;

There are many cases, however, in which one may want to perform inferences about differences of the means of multiple populations;

We will develop the main concepts and ideas related with this kind of test by examining a simple example, related to a paper manufacturing operation.

Image: http://goo.gl/xYWGN

→ Poderia comparar todos contra todos
- requer correção de α,
  o que incorre em perda de potência
- mais trabalhoso

Example: paper manufacturing
Problem definition

Tensile strength (TS) is an important characteristic for certain types of paper for industrial use;

A reasonable conjecture is that this characteristic is influenced by the kind of hardwood used in the manufacturing process.

The process engineer wants to investigate whether four different hardwoods result in papers with relevant differences of TS, using a pilot plant as his experimental unit.

Example inspired by Montgomery & Runger (2010), Ch. 13.

Example: paper manufacturing
Problem definition

The total budget allocated for the experiment allows the execution of six production runs for each kind of hardwood.

Under these specifications, the experiment has a single experimental *factor* (*Hardwood type*) with $a = 4$ *levels* (hardwood types *A*, *B*, *C* and *D*) and $n = 6$ *replicates* at each level.

The response variable will be the tensile strength of paper (measured in kPa). The engineering team is interested in finding out whether any hardwood leads to an increase in the mean TS value of the paper.

The minimum difference of practical meaning is defined as $5kPa$, and a reasonable upper estimate for the standard deviation of this process is $\hat{\sigma} = 6kPa$. Desired error levels are defined as $\alpha = 0.1$ and $\beta = 0.2$.

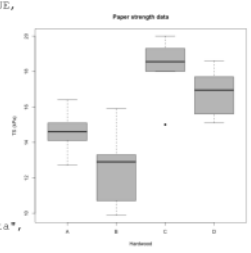
→ 1 factor ~ independentes

4 niveis (tratamentos)

- A
- B
- C
- D

Example: paper manufacturing
Exploratory data analysis

```
> paper   <- read.table(file = "../data files/paper_strength.csv",
+                       header = TRUE,
+                       sep = ",")
>
> summary(paper)
 Hardwood      TS_kPa
 A:6      Min.   : 9.88
 B:6      1st Qu.:13.90
 C:6      Median :15.35
 D:6      Mean   :15.56
          3rd Qu.:17.77
          Max.   :20.00
>
> boxplot(TS_kPa~Hardwood,
+         data = paper,
+         xlab = "Hardwood",
+         ylab = "TS (kPa)",
+         main = "Paper strength data",
+         pch  = 16,
+         col  = "gray")
```

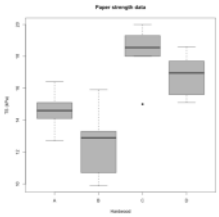


Example: paper manufacturing
Exploratory data analysis

The boxplot suggests the existence of differences among the factor levels;

Besides, we can also observe a small variability in the spread of different levels; some suggestion of asymmetry in level *B*; and a possible outlier in level *C*.

These characteristics will need to be taken into account during the analysis.



Example: paper manufacturing
Statistical model

This data can be described by a linear statistical model of the form:

$$y_{ij} = \underbrace{\mu_i + \epsilon_{ij}}_{\text{Means model}} = \underbrace{\mu + \tau_i + \epsilon_{ij}}_{\text{Effects model}} \begin{cases} i = 1, \ldots, a \\ j = 1, \ldots, n \end{cases}$$

where μ is the overall mean, τ_i represents the effect of the i-th level, and ϵ_{ij} is the residual (random error, or unmodeled variability);

In the derivation of the statistical test for the existence of differences in the group means, we will employ the effects model, and initially consider a few assumptions about the residuals:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \begin{cases} i = 1, \ldots, a \\ j = 1, \ldots, n \end{cases} \quad , \quad \text{with } \epsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$$

- $\sum_i \tau_i = 0$
- Única forma de termos todos os fatores iguais é se $\tau_i = 0 \; \forall i$

Cada observação já é um somatório de várias operações

	Nível			
	A	B	C	D
①	x	x		
	x	x		
	x	x		
	x	x		
	x	x		
	x	x		

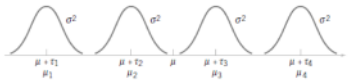
Replicações

$y_{ij} = \mu_i + \varepsilon_{ij}$

$= \mu + \tau_i + \varepsilon_{ij}$

Example: tensile strength
Statistical model

If these assumptions are correct, the populations are expected to be distributed as:



Since we are interested in testing our data for differences in the mean values of each population, the test hypotheses can be described as:

$$\begin{cases} H_0 : \tau_i = 0, \;\; \forall i \in \{1, 2, \ldots, a\} \\ H_1 : \exists\, \tau_i \neq 0 \end{cases} \quad == \quad \begin{cases} \cdot H_0 : \mu_1 = \mu_2 = \ldots = \mu_a = \mu \\ \cdot H_a : \exists \mu_i \neq \mu \end{cases}$$

If data collection is performed in random order, under constant experimental conditions, we have a *completely randomized design*.

Image: Montgomery&Runger (2010), Ch. 13

→ evitar efeito de desgastes, perda de calibração, etc.

→ poderia ser feita aleatorizações estratificada (permite análise de efeito de ordem)

$\varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$

- normalidade → Homocedasticidade
- independência
- igualdade das variâncias nos resíduos

The Fixed Effects Model
Definition

This approach to modeling the mean effects of specific factor levels is known as the *fixed effects model*; → mais comum p/ efeitos categóricos

This approach is appropriate to testing hypotheses in situations when factor levels are arbitrarily defined by the experimenter;

For these cases, the inference is made over the mean values for each level, and cannot be extended to similar levels that were not tested (e.g., other types of hardwood);

Other situations may require different kinds of modeling, such as *random* or *mixed effects models*, but these will not be explored here.

The Fixed Effects Model
Development

As mentioned earlier, we will use the *effects model* for describing the development of the statistical test:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \begin{cases} i = 1, \ldots, a \\ j = 1, \ldots, n \end{cases}$$

where treatment effects are seen as deviations from the grand mean μ.

By construction, we have that:

$$\sum_{i=1}^{a} \tau_i = 0;$$

The Fixed Effects Model
Development

The total variability of the data can be expressed by the *total sum of squares*, which represents the sum of the squared deviations between each observation and the overall sample mean:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\bullet\bullet})^2$$

↳ estimador da média global $\hat{\mu}$

With some relatively simple algebra, the SS_T can be divided into two terms, representing the within-group and the between-group variability:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\bullet\bullet})^2 = \underbrace{n\sum_{i=1}^{a}(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2}_{SS_{Levels}} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\bullet})^2}_{SS_E}$$

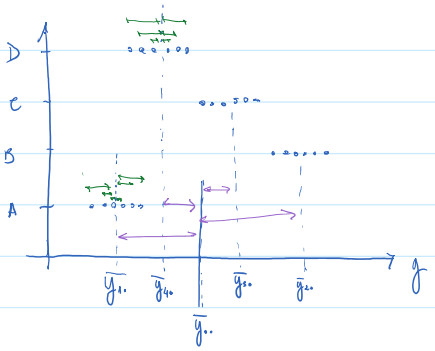
where $\bullet$ indicates the summation over an index, and $\bar{}$ indicates an averaging operation.

$\bar{y}_{\bullet\bullet} = \frac{1}{an}\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij} = \hat{\mu}$

$\bar{y}_{i\bullet} = \frac{1}{n}\sum_{j=1}^{n} y_{ij} = \hat{\mu}_i = (\hat{\mu + \tau_i})$

→ soma das diferenças entre níveis e média global ⟺

→ soma das diferenças intra-níveis ⟺



The Fixed Effects Model
Development

Dividing the sums of squares by their respective number of degrees of freedom yields quantities known as *mean squares*.

The relevant means squares for our test will be the *levels mean square* and the *residual mean square*:

$$MS_E = \frac{SS_E}{a(n-1)} \qquad MS_{Levels} = \frac{SS_{Levels}}{a-1}$$

The expected values of these quantities are:

$$E[MS_E] = \sigma^2 \qquad E[MS_{Levels}] = \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1}$$

estimador não enviesado de variância (excluindo efeito de τ_i)

depende de τ_i

The Fixed Effects Model
Development

$$E[MS_E] = \sigma^2 \qquad E[MS_{Levels}] = \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1}$$

Notice that MS_E is an unbiased estimator for the common variance of the residuals, while MS_{Levels} is biased by a term that is proportional to the squared values of the τ_i coefficients.

However, under H_0 we have that $\tau_i = 0$ for all i, that is, $E[MS_{Levels}] = E[MS_E] = \sigma^2$. But only if the null hypothesis is true.

The Fixed Effects Model
Development

It can be shown that, if H_0 is true, the statistic

$$F_0 = \frac{MS_{Levels}}{MS_E}$$

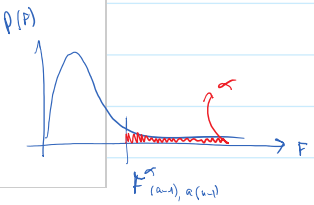
→ teste F

is distributed according to an F distribution with $a-1$ degrees of freedom for the numerator and $a(n-1)$ for the denominator. The usual notation is $F_{(a-1),a(n-1)}$

↳ sob H_0

If H_0 is false, the expected value of MS_{Levels} is larger than that of MS_E, which results in larger values of F_0 and defines the critical region for our test:

Reject H_0 at the α significance level if $f_0 > F_{\alpha;(a-1),a(n-1)}$



Só é necessário teste unilateral à direita

Example: paper manufacturing
Computational analysis

Example: paper manufacturing
Computational analysis

```
> model <- aov(TS_kPa~Hardwood, data = paper)
> summary.aov(model)
            Df Sum Sq Mean Sq F value   Pr(>F)
Hardwood     3 110.77   36.92   13.62 4.56e-05 ***
Residuals   20  54.24    2.71
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The *ANOVA table* provides information on the sources of variation, together with their corresponding degrees of freedom, sums of squares and mean square values. The table also informs the values of the test statistic and the corresponding p-value of the test ($Pr(> F)$).

In this case, the low p-value ($p = 4.56 \times 10^{-5}$) suggests the rejection of the null hypothesis in favor of the alternative. But what does that mean?

Example: paper manufacturing
Computational analysis

Recall the null and alternative hypotheses for the ANOVA:

$$\begin{cases} H_0 : \tau_i = 0, \ \ \forall i \\ H_1 : \exists\, \tau_i \neq 0 \end{cases}$$

The rejection of the null hypothesis leads to the conclusion that *there is at least one level with an effect significantly different from zero.* But which one?

For this analysis to be complete, we still need to answer two questions:

- Can we verify the assumptions of the test?
- Which means are different from which, and by how much?

Assumptions
Model validation

As mentioned earlier, the ANOVA model is based on three assumptions on the behavior of the residuals:

- *Independence*;
- *Homoscedasticity*, i.e., equality of variances across groups;
- *Normality*;

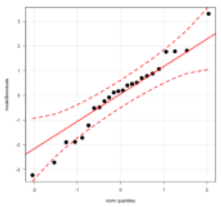
The residuals of the model can be easily obtained as:

$$e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - (\hat{\mu} + \hat{\tau}_i) = y_{ij} - \bar{y}_{i\bullet}$$

Assumptions
Model validation

The normality assumption can be tested using the Shapiro-Wilk test coupled with a normal QQ plot of the residuals:

```
> shapiro.test(model$residuals)
Shapiro-Wilk normality test
data:  model$residuals
W = 0.9722, p-value = 0.7225

> library(car)
> qqPlot(model$residuals,
pch = 16,
lwd = 3,
cex = 2,
las = 1)
```

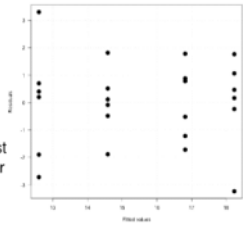


The ANOVA is relatively robust to moderate violations of normality, as long as the other assumptions are verified or the sample size is large enough.

Assumptions
Model validation

The homoscedasticity assumption can be verified by the Fligner-Killeen test, together with plots of residuals by fitted values:

```
> fligner.test(TS_kPa~Hardwood, data = paper)
Fligner-Killeen test of homogeneity of
variances
data:  TS_kPa by Hardwood
Fligner-Killeen:
med chi-squared = 1.0622, df = 3,
p-value = 0.7862
>
> plot(x = model$fitted.values,
+ y = model$residuals,
+ ...)
```



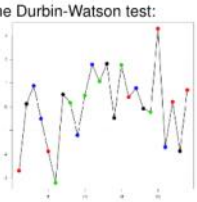
ANOVA is relatively robust to modest violations of homoscedasticity, as far as the sample is *balanced*.

Assumptions
Model validation

As usual, the independence assumption should be guaranteed (to the best of the experimenter's knowledge) on the design phase;

To test for serial correlations, we can use the Durbin-Watson test:

```
> durbinWatsonTest(model)
 lag Autocorrelation D-W Statistic p-value
   1     -0.008996302      1.872801   0.868
 Alternative hypothesis: rho != 0

> plot(x    = seq_along(model$residuals),
+      y    = model$residuals,
+      type = "l", ...)
> points(x    = seq_along(model$residuals),
+        y    = model$residuals,
+        type = "p",
+        col  = as.numeric(paper[,1]), ...)
```



The ANOVA is sensitive to violations of independence. Randomization and attention to possibly influential factors can help avoiding violations of this assumption.

Multiple comparisons
The need for multiple comparisons

If the ANOVA assumptions are verified (i.e., if we have solid grounds for trusting the result of the test), we usually need to determine *which* levels of the factor are significantly different[a];

Whenever possible, the planning of which comparisons will be after an analysis of variance procedure should be defined *a priori*. Post-hoc definition of hypotheses (a.k.a. HARKing[b]) are a common entry point for researcher biases into the analysis, and should be dealt with very carefully.

[a] Of course this is only necessary if we rejected H_0 in the original test. For more on how to proceed with nonsignificant results, see Ellis(2010).

[b] **H**ypothesizing **A**fter the **R**esults are **K**nown. See Kerr(1998).

Multiple comparisons
Types of comparisons

The planning of multiple comparisons must be guided by the technical question underlying the experiment.

Whenever possible, the researcher should opt to perform the smallest number of comparisons needed to adequately answer his or her question. This will require the smallest sample size, or result in the largest power for a given experimental setup.

Usual questions involve (but are not limited to):

- *How does one level compare to the others?*
- *How does each level compare to the grand mean?*
- *How do the levels compare to each other (all vs. all)?*

Multiple comparisons
MHT considerations

The multiple comparisons performed after an ANOVA are essentially composed of a series of t-tests for the difference between two population means, with some slight modifications;

If the assumptions of the ANOVA are verified, we already have some information about the data: we know, for instance, that the groups are homoscedastic, and that their common variance is estimated by MS_E (with $a(n-1)$ degrees of freedom);

We also know that, if we are going to perform multiple tests on the same data set, that the probability of a type-I error on each test is α. If we want to to keep our overall error rate controlled at a given level, we will need to correct the α value used for each test.

Multiple comparisons
MHT corrections

There are a number of ways of adjusting the α value of the pairwise comparisons in order to maintain the *familywise error rate* (FWER) at a controlled level[c].

Two of the most common (and most conservative) are the Bonferroni and the Šidák corrections.

Assuming K planned comparisons, the Bonferroni method tests each individual hypothesis with:

$$\alpha_{adj} = \frac{\alpha_{family}}{K}$$

while the Šidák correction uses:

$$\alpha_{adj} = 1 - \left(1 - \alpha_{family}\right)^{1/K}$$

[c] The methods presented here work well for a relatively small number of comparison. For more on MHT, see Schaffer(1995)'s discussion on controlling the False Discovery Rate.

Multiple comparisons
All vs. all

Pairwise comparisons of the *all vs. all* type appear whenever we are simply interested in detecting which levels are significantly different from which, without any prior information or special interest in one specific level or ordering.

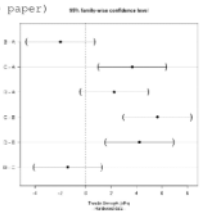
In these cases, the number of comparisons is $K = a(a-1)/2$, where a is the number of levels.

The sample size calculations for this case follow the same equations used for the t test for two independent samples, but with the α value corrected for multiple hypotheses and the number of degrees of freedom of the reference distribution equal to those of the residual term in the ANOVA, i.e., $a(n-1)$.

Multiple comparisons
All vs. all

Suppose that was the case for our hardwood example:

```
> library(multcomp)
> paper_tukey <- glht(model, linfct = mcp(Hardwood = "Tukey"))
> paper_tukey_CI <- confint(paper_tukey, level = 0.95)

Simultaneous Confidence Intervals
Multiple Comparisons of Means: Tukey Contrasts
Fit: aov(formula = TS_kPa ~ Hardwood, data = paper)
Quantile = 2.7989
95% family-wise confidence level
Linear Hypotheses:
           Estimate lwr     upr
B - A == 0 -1.9867  -4.6478  0.6745
C - A == 0  3.6500   0.9889  6.3111
D - A == 0  2.2333  -0.4278  4.8945
C - B == 0  5.6367   2.9755  8.2978
D - B == 0  4.2200   1.5589  6.8811
D - C == 0 -1.4167  -4.0778  1.2445

> plot(paper_tukey_CI, ...)
```



Multiple comparisons
All vs. one

Pairwise comparisons of the *all vs. one* type usually emerge in the context of comparing levels against a control:

- Comparison of a proposed approach vs. existing ones;
- Comparison of different approaches vs. a standard one (or a placebo-like group);

In these cases, the number of comparisons is $K = a - 1$, where a is the number of levels. Each test can again be performed using the t_0 test statistic:

$$t_0^i = \frac{\bar{y}_i - \bar{y}_0}{\hat{\sigma}\sqrt{\left(\frac{1}{n_i} + \frac{1}{n_0}\right)}}$$

Multiple comparisons
All vs. one

There are two main approaches to calculating sample size for *all vs. one* comparisons:

- Balanced design;
- Optimal allocation of units.

With a balanced design (that is, all levels have the same number of observations), the calculation of n follows the same approach as the *all vs. all* comparisons, but correcting α for only $a - 1$ comparisons.

For the optimal allocation of units, an unbalanced design is used.

Multiple comparisons
All vs. one - optimal allocation

As several levels will be compared against the single control group, the relative importance of the latter is greater and therefore it should have a larger sample size.

To maximize the power of this multiple comparisons procedure, the sample size of the control group should be:

$$n_0 = n_i\sqrt{K}$$

where n_i is the common sample size for the non-control levels:

$$n_i = \left(1 + \frac{1}{\sqrt{K}}\right)\left(\frac{(t_{(\alpha_{adj}/2)} + t_{\beta})\hat{\sigma}}{\delta^*}\right)^2$$

A good free software for doing sample size calculations and power analysis in nontrivial contexts such as this one is G*Power 3, http://www.gpower.hhu.de/. It is also not difficult to implement these calculations in R.

Multiple comparisons
All vs. one - **Dunnett's test**

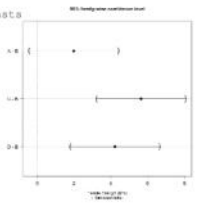
As in the case of *all vs. all* comparisons, there is a test that is usually employed for its superior sensitivity: Dunnett's test.

The control group sample size n_0 calculated assuming that Bonferroni-corrected t-tests will be used is slightly overestimated in relation to the required n_0 for Dunnett's test, but in practice the differences are small enough not to matter;

Multiple comparisons
All vs. one

Assuming that in our example the *B* level is the standard one, against which the other ones are to be compared:

```
> paper$Hardwood   <- relevel(paper$Hardwood, ref = "B")
> model2           <- aov(TS_kPa~Hardwood, data = paper)
> paper_dunnett    <- glht(model2, linfct = mcp(Hardwood = "Dunnett"))
> paper_dunnett_CI <- confint(paper_dunnett, level = 0.95)

Simultaneous Confidence Intervals
Multiple Comparisons of Means: Dunnett Contrasts
Fit: aov(formula = TS_kPa ~ Hardwood,
         data = paper)
Quantile = 2.5394
95% family-wise confidence level
Linear Hypotheses:
           Estimate lwr     upr
A - B == 0  1.9867  -0.4277  4.4011
C - B == 0  5.6367   3.2223  8.0511
D - B == 0  4.2200   1.8056  6.6344

> plot(paper_dunnett_CI, ...)
```



Multiple comparisons
Some final considerations

The kind of comparisons that are to be performed after an ANOVA should be planned in advance, as it influences your data collection and sample size calculations. There are of course sample size formulas for the pure ANOVA, but these are usually of limited use since researchers frequently want to know where the detected differences lie.

There are a myriad of approaches for post-ANOVA multiple comparisons[e], but in general the formulas for sample size calculation will follow he ideas outlined above: correct the α value to account for type-I error inflation and calculate *n* based on formulas for two-sample *t* tests.

[e] Check Hothorn *et al.* (2008) for an idea on how varied this can get.

More on sample sizes
Sample size formulas for ANOVA

If one is interested in calculating the required sample size for the ANOVA procedure (without worrying about the eventual post-hoc testing), the formulas are almost as simple as those used for the t tests.

Essentially, the power/sample size calculations for the ANOVA boil down to the equality:

$$F_{(1-\alpha)} = F_{\beta;\phi}$$

with both *F* distributions having $(a-1)$ degrees of freedom in the numerator and $a(n-1)$ in the denominator. The noncentrality parameter ϕ is given by:

$$\phi = \frac{n\sum_{i=1}^{a}\tau_i^2}{\hat{\sigma}^2}$$

More on sample sizes
Sample size formulas for ANOVA

To ilustrate the sample size calculation procedure, imagine an experimental design with $a = 4$, $\alpha = 0.05$, $\hat{\sigma} = 7$, and suppose that the researcher wants to be able to detect whether any two means present differences of magnitude $\delta^* = 12$ with power $(1 - \beta) = 0.8$.

Under these conditions, two scenarios tend to be of interest: the first is if we have two levels biased symmetrically about the grand mean, and all the others equal to zero:

$$\tau = \left\{-\frac{\delta^*}{2}, \frac{\delta^*}{2}, 0, 0\right\}$$

and the second is if we have one level biased in relation to all others:

$$\tau = \left\{-\frac{(a-1)\delta^*}{a}, \frac{\delta^*}{a}, \frac{\delta^*}{a}, \frac{\delta^*}{a}\right\}$$

More on sample sizes
Sample size formulas for ANOVA

For the first case we have a noncentrality parameter of:

$$\phi = \frac{4\left(6^2 + 6^2 + 0 + 0\right)}{7^2} = 5.88$$

Which allows us to calculate the required sample size by iterating on n until:

$$F_{(1-\alpha)} \leq F_{\beta;\phi}$$

More on sample sizes
Sample size formulas for ANOVA

Doing it manually:

```
> a       <- 4
> alpha   <- 0.05
> sigma   <- 7
> delta   <- 12
> beta    <- 0.2
>
> tau <- c(-delta/2, delta/2, rep(0, a - 2)) # define tau vector
> n   <- 2
> while (qf(1 - alpha, a - 1, a*(n - 1)) >
+        qf(beta, a - 1, a*(n - 1), n*sum(tau^2)/sigma^2)) n <- n + 1
> print(n)
[1] 9
```

Using `power.anova.test()`:

```
> vartau <- var(tau)
> power.anova.test(groups = 4, between.var = vartau,
+                  within.var = sigma^2, sig.level = alpha,
+                  power = 1 - beta)$n
[1] 8.463358
```

More on sample sizes
Sample size formulas for ANOVA

The second case (one level biased in relation to all others) is also quite easy to calculate manually, but lets keep it simple:

```
> tau <- c(-delta*(a - 1)/a, rep(delta/a, a - 1))
> vartau <- var(tau)
> power.anova.test(groups = 4, between.var = vartau,
+                  within.var = sigma^2, sig.level = alpha,
+                  power = 1 - beta)$n
[1] 6.018937
```

It is important to remember that these are the sample sizes required for the ANOVA only - any multiple comparisons procedure executed afterwards to pinpoint the significant differences will have smaller power for same-sized effects (unless more observations are added). This is one reason why it is common to design experiments calculating the sample sizes based on the multiple comparisons procedure, instead of using the ANOVA formulas.

Bibliography

Required reading

1. D.C. Montgomery, G.C. Runger, *Applied Statistics and Probability for Engineers*, Ch. 13. 5th ed., Wiley, 2010.
2. N.L. Kerr, *HARKing: Hypothesizing After the Results are Known*, Personality and Social Psychology Reviews 2(3): 196–217, 1998.

Recommended reading

1. P. D. Ellis, *The Essential Guide to Effect Sizes*. 1st ed., Cambridge, 2010.
2. J.P. Schaffer, *Multiple Hypothesis Testing*, Annual Reviews on Psychology 46, 561–584, 1995.
3. P. Mathews, *Sample Size Calculations: Practical Methods for Engineers and Scientists*. Ch. 8, 1st ed., MMB, 2010.
4. T. Hothorn, F. Bretz, P. Westfall, *Simultaneous Inference in General Parametric Models*. Biometrical Journal 50(3), 346–363, 2008.

About this material

Conditions of use and referencing

Please reference this work as:

Felipe Campelo (2015), *Lecture Notes on Design and Analysis of Experiments*. Online: https://github.com/fcampelo/Design-and-Analysis-of-Experiments Version 2.11, Chapter 10; Creative Commons BY-NC-SA 4.0.

```
@Misc{Campelo2015-01,
  title={Lecture Notes on Design and Analysis of Experiments},
  author={Felipe Campelo},
  howPublished={\url{https://github.com/fcampelo/Design-and-Analysis-of-Experiments}},
  year={2015},
  note={Version 2.11, Chapter 10; Creative Commons BY-NC-SA 4.0.},
}
```