COMP320 Research Practice

04 - Experimentation and Hypotheses

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.

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.

Experiments

Data gathering

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

Characteristics

- Use of historical data;
- Investigating correlations;

Problems

- Data representativeness;
- Availability of data;

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;

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

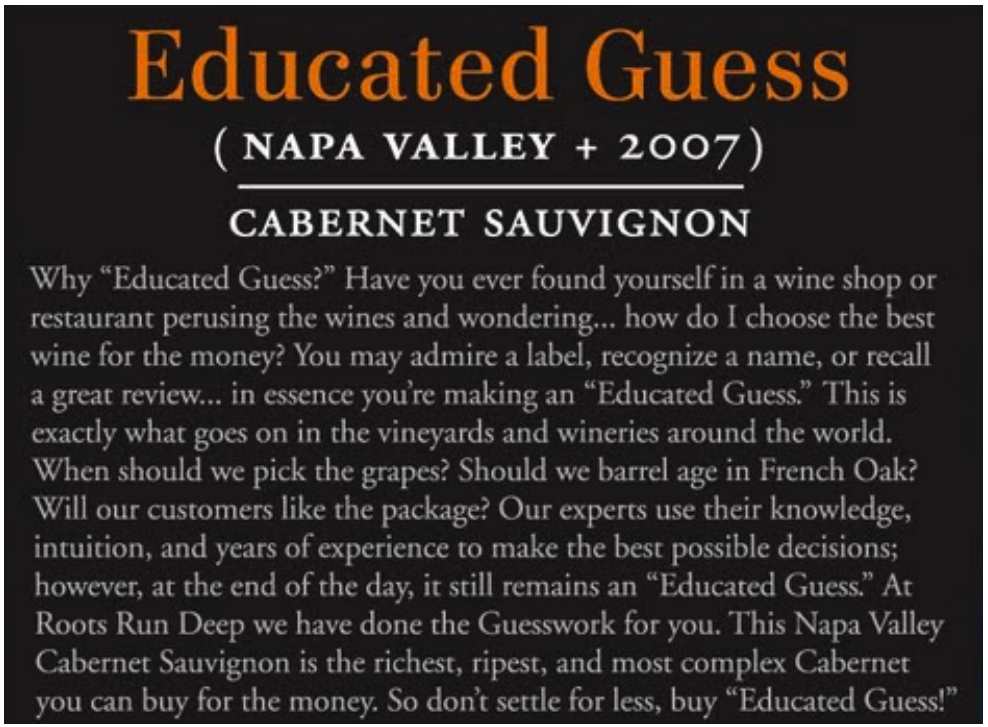
- Select arbitrary combination of levels for the factors;
- Test and observe behavior;
- Change one or two factors at a time, then re-test;

- 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;

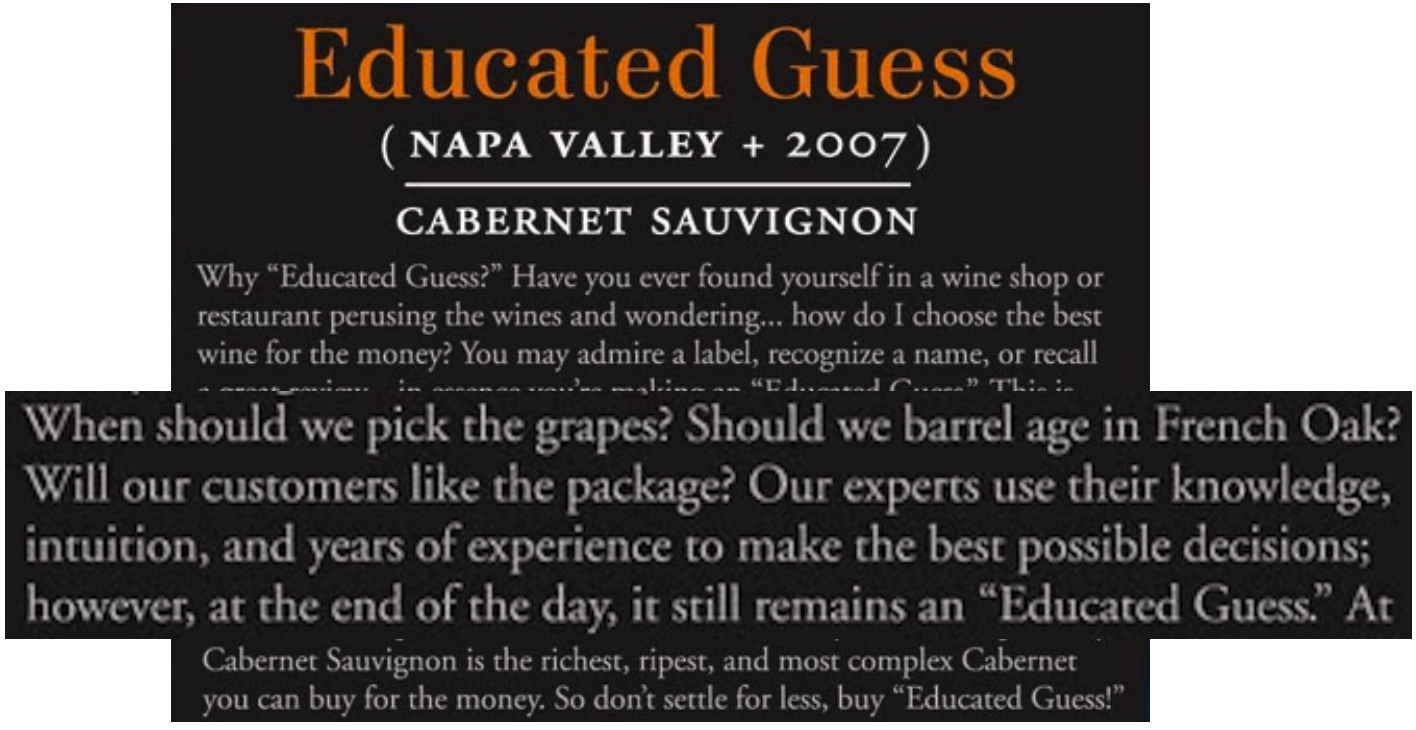


Images (c) Roots Run Deep Winery: `http://www.rootsrundeep.com/educated_guess.html`

Experimentation strategies

Educated guessing

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

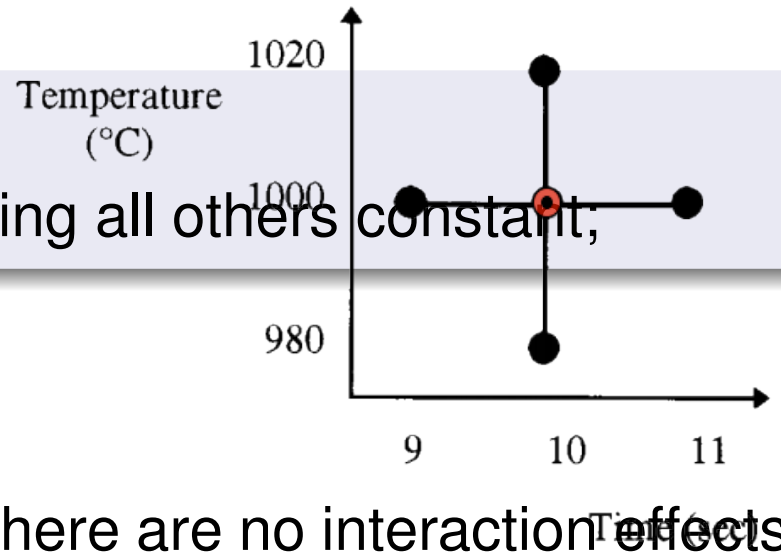


Images (c) Roots Run Deep Winery: `http://www.rootsrundeep.com/educated_guess.html`

Experimentation strategies

COST: Change One Separate factor at a Time

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



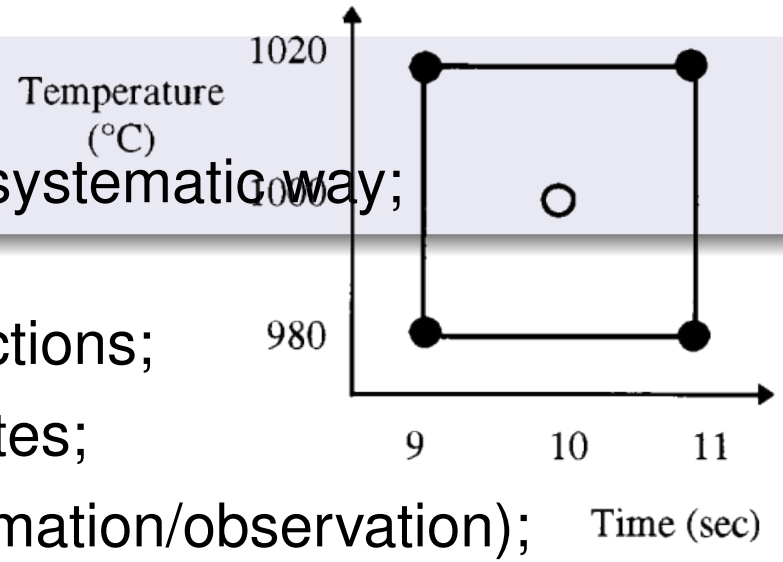
- Also widely used;
- Can achieve good results as long as there are no interaction effects;

Image: (c) D.C. Montgomery

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);

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.

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;

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;

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;

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`

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;
- 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?
- Practical relevance?
  - Employ exploratory experiments;
- Placement within the literature;
  - Avoid repetition and irrelevance.

"*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

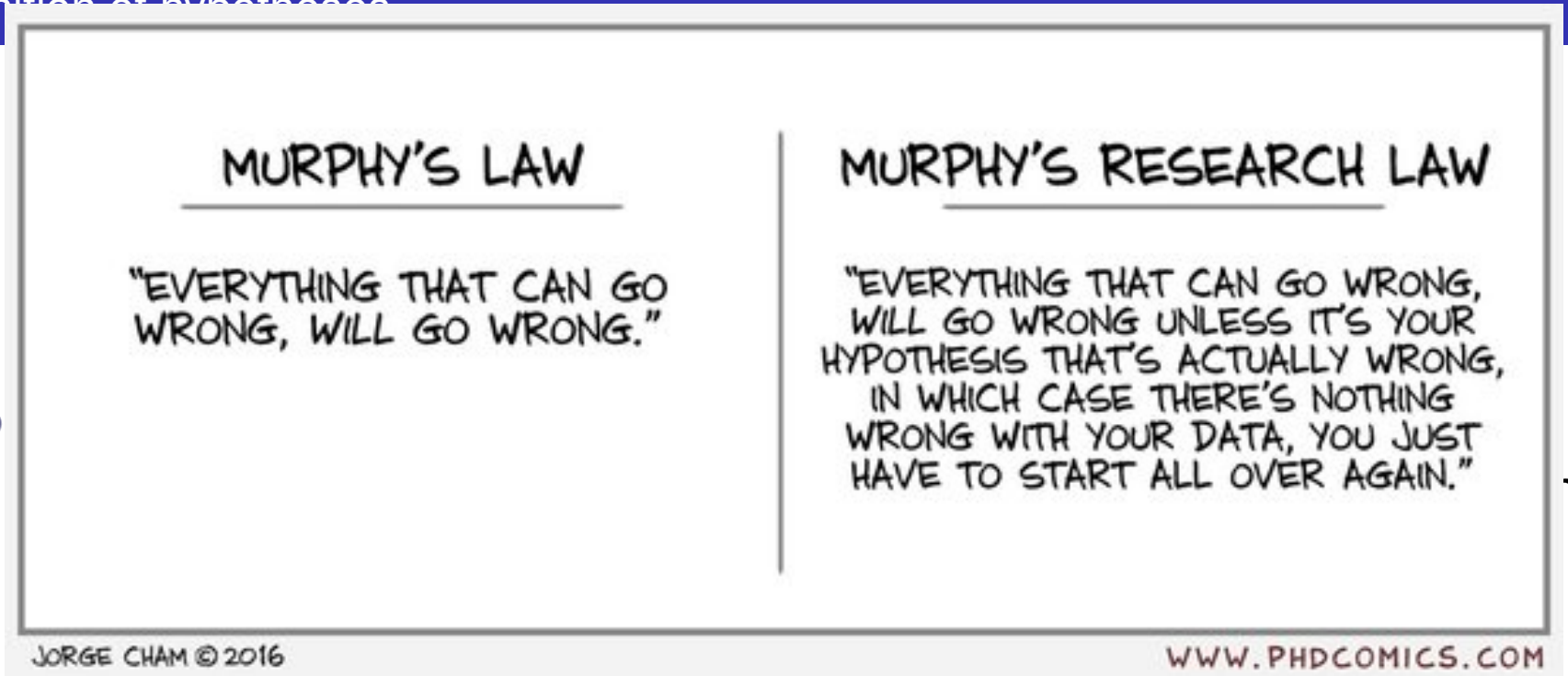


Image: `http://phdcomics.com/comics/archive.php?comicid=1867`

Choice of Experimental Design

Experimental design

- (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;

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.

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 $P(data|null\text{-}model)$;
  - Decision by rejection (or not) of the null hypothesis;
  - Validation of model assumptions;
  - Estimation of the *magnitude* of differences - **practical significance**;

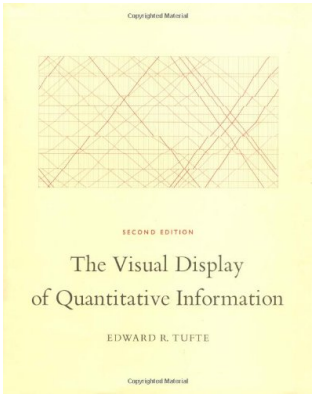
*Statistical methods do not **prove** anything, but they allow an objective definition of margins of plausibility for certain statements.*

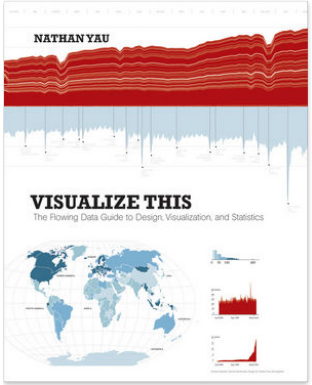
Reporting of results

Presentation

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

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





Other great resources on graphical excellence:

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

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

Conclusions

Drawing and reporting conclusions

- Conclusions should be based on solid evidence from the data;
- Be conservative - it is common to exaggerate the generality of the results;
- Report significance levels and the assumptions under which the results are valid;
- *Suggest explanations* to the observed results;
- Careful with *anomaly hunting*;

Always let the science drive the statistics. If you get a statistically significant result, go back and describe what it means in the scientific context.

– Aaron Rendahl

Discussion

Some more relevant points

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

Bibliography

Required reading

1. Dept. Biochemistry & Cell Biology, Rice University , *Common Errors in Student Research Papers.* - http://www.ruf.rice.edu/~bioslabs/tools/report/reporterror.html
2. T. Brady, *Reviewer's quick guide to common statistical errors in scientific papers.* https://goo.gl/IKvAYc
3. R.M. Szydloa 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* - http://goo.gl/jtT4hA
2. A. Rendahl, *Experimental design and statistical analysis: questions to consider as you write your grant.* - http://goo.gl/yT0TK7
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. - http://skeptoid.com/episodes/4381

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.

```
@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 2; Creative Commons BY-NC-SA 4.0.},
}
```

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:

- Falsifiability;
- Testability;

> "*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

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;

> "*Numquam ponenda est pluralitas sine necessitate.*"
>
> William of Ockham
>
> (1287-1347)
>
> English philosopher and theologian

Image: `http://www.philosophybasics.com/philosophers_ockham.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)

- Absence of effects;
- *Conservative* model.

Example: $H_0 : \mu = 25$

Alternative hypothesis (H_1)

- Presence of some effect;
- Existence of something “new”.

Example: $H_1 : \mu \neq 25$

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 own a company that sells green peas to large customers, and that you want to determine whether your 50kg sacks 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 sack is* $50kg$, and the alternative of interest could be expressed as the complementary inequality.

$$\begin{cases} H_0 : \mu = 50kg \\ H_1 : \mu \neq 50kg \end{cases}$$

Suppose still that $n = 10$ packs are randomly sampled, and their contents are weighted using a calibrated scale;

Image: `http://www.storko.eu/ed_files/image/green-peas.jpg`

Statistical Hypotheses

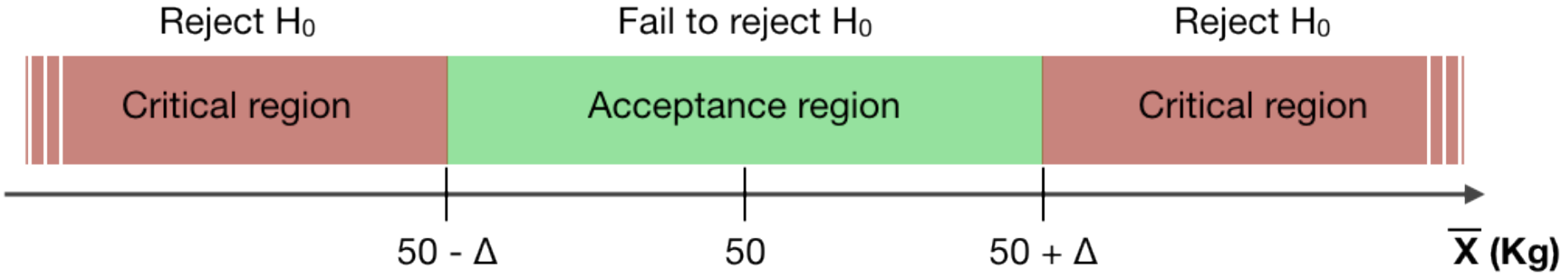
Example

Since the sample mean $\bar{x}$ is a good estimator of the real mean μ, common sense suggests that:

- If $\bar{x} \cong 50$kg - corroboration of H_0;
- If $\bar{x} \ll 50$kg or $\bar{x} \gg 50$kg - refutation of H_0;

That is, we can use $\bar{x}$ as the basis for a statistical test.

But how to define 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 $(1 - \alpha)$.

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 $\mu = 50$kg), the distribution of values of $\bar{x}$ is aproximately normal (assuming the CLT holds), with average 50kg and standard error $(\sigma/\sqrt{n})$ kg;

For a desired 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 under the null hypothesis is $1 - \alpha = 0.95$.

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:

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

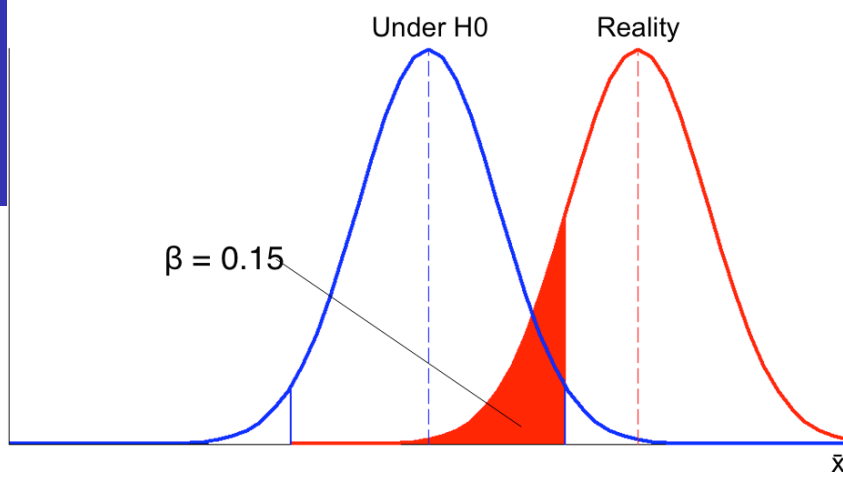
The quantity $(1 - \beta)$ is known as *power* of the test, and quantifies its sensitivity to effects that violate the null hypothesis.

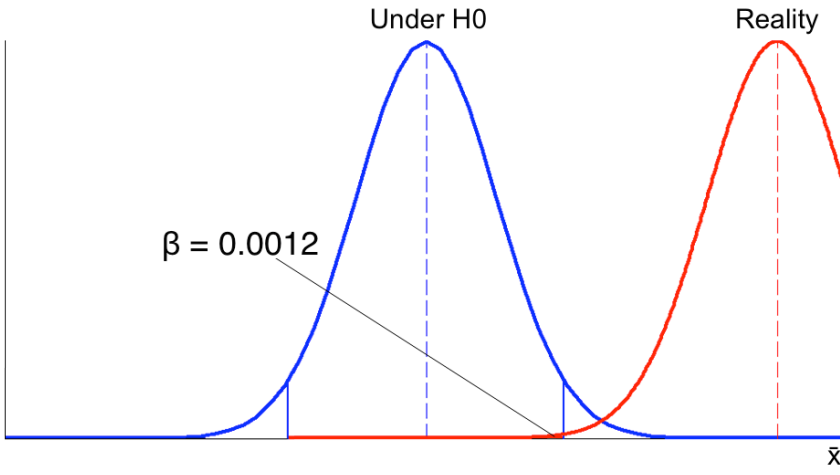
Inferential Errors

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.





Inferential Errors

Type II error

The power of a test is governed by several factors:

- Controllable: significance level, sample size, directionality of H_1;
- Uncontrollable: real value of the parameter, variance;

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 fortify it);

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;

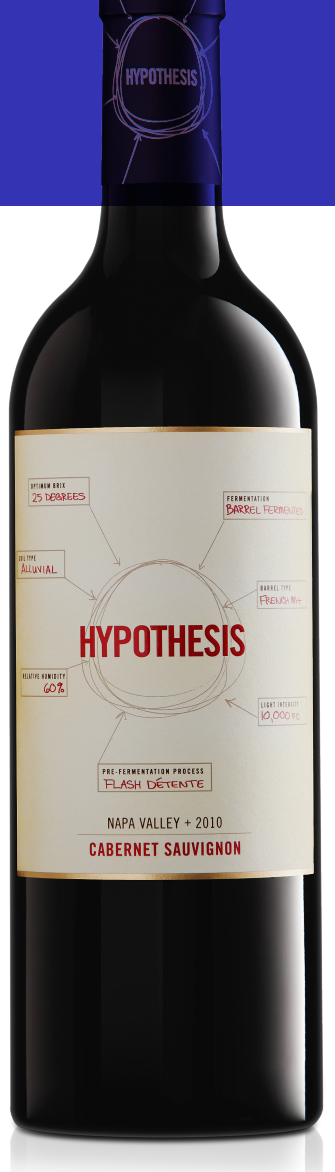


Image (c) Roots Run Deep Winery: `http://www.rootsrundeep.com/hypothesis.html`

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 sacks. Assume (for now) that the variance of the process is known. The test hypotheses are defined as:

$$\begin{cases} H_0 : \mu = 50kg \\ H_1 : \mu \neq 50kg \end{cases}$$

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

Given these characteristics, we expect that the sampling distribution of $\bar{X}$ is normal, with $Var(\bar{X}) = \sigma^2/n$ and – **if H_0 is true** – a mean of $\mu_{\bar{X}} = \mu_0 = 50$;

Hypothesis Testing

Mean of a normal distribution, variance known

Based on these characteristics, the standardized variable

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

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

This result implies that:

$$P\left(z_{\alpha/2} \leq Z_0 \leq z_{1-\alpha/2} \mid H_0 \text{ is true}\right) = 1 - \alpha$$

which provides a selection criterion between H_0 and H_1:

- If $z_{\alpha/2} > Z_0$ or $z_{1-\alpha/2} < Z_0$, reject H_0 with confidence $(1 - \alpha)$;
- Otherwise, there is not enough evidence to reject H_0 at this confidence level;

Hypothesis Testing

Mean of a normal distribution, variance known

Assume that we got $\bar{x} = 49.65$kg from our $n = 10$ observations, and that we know that $\sigma = 1$kg. In this case,

$$z_0 = \frac{49.65 - 50}{1/\sqrt{10}} = -1.113$$

The critical values of the standard normal distribution at the significance level $\alpha = 0.05$ are $[z_{0.025}, z_{0.975}] = [-1.96, 1.96]$;

Since $z_0 \in [-1.96, 1.96]$, we can conclude that there is not enough evidence to reject H_0 at the 95% confidence level.

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:

$$\begin{cases} H_0 : \mu \geq 50kg \\ H_1 : \mu < 50kg \end{cases}$$

Assume also that we want to be more conservative, so we pick a significance level $\alpha = 0.01$;

In this case, **if H_0 is true**, we have that

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

Hypothesis Testing

Mean of a normal distribution, variance unknown

From the same data used earlier, $\bar{x} = 49.65\text{kg}$, $n = 10$, $s = 0.697\text{kg}$:

$$t_0 = \frac{49.65 - 50}{0.697/\sqrt{10}} = -1.597$$

The critical value of this test statistic for the desired significance is $t_{\alpha,n-1} = t_{0.01,9} = -2.82$, which means that under H_0 there is a 99% change that the test statistic will yield a value greater than this number.

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

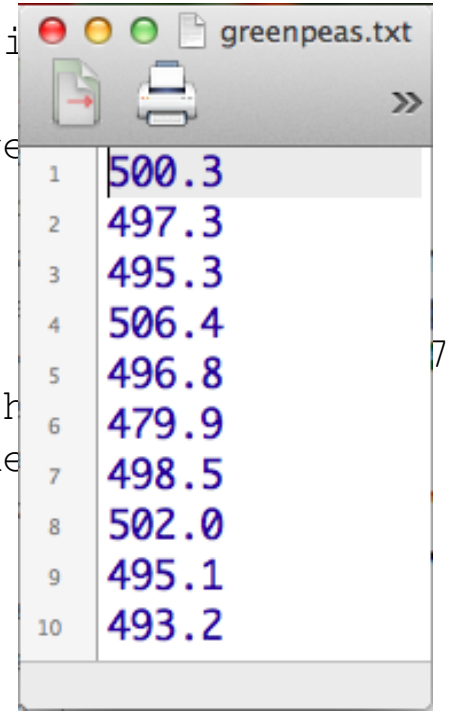
Hypothesis Testing

Mean of a normal distribution, variance unknown

```
> sample <- as.numeric(scan("../data files/greenpeas.txt"))

> t.test(sample,
+        alternati
+        mu = 50,
+        conf.leve

One Sample t-test
data:  sample
t = -1.5969, df =                    7237
alternative hypoth                   is less than 50
99 percent confide
-Inf 50.2699
sample estimates:
mean of x
49.648
```



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:

$$p = P(t_0 \leq -1.597 | H_0 = \textbf{TRUE}) = \int_{-\infty}^{-1.597} t_{(df=9)} dt = 0.07237$$

A priori definition of the significance level is still important!

Hypothesis Testing

p-values, significance and effect sizes

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

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

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

Is it really *that* significant?

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:

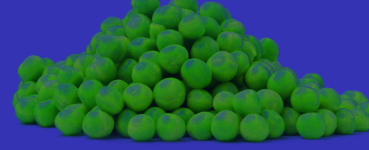
$$d = \frac{\bar{x} - \mu_0}{s}$$

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.

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 $H_0 : \mu = 50$ 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 = 50, conf.level = 0.99)
(...)
t = -1.5969, df = 9, p-value = 0.1447
alternative hypothesis: true mean is not equal to 50
99 percent confidence interval:
 48.93166 50.36434
sample estimates:
mean of x
   49.648
```

Sample size and Type-II error

Some considerations

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

Given a desired test, its power is essentially a function of 4 elements:

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

The experimenter generally have very 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;

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 we are really interested in detecting negative deviations from the nominal value greater than 1%, i.e., $\delta^* = 0.01 \times 50 = 0.5$kg. The researcher defines that, for this minimally interesting effect, a test power of 0.85 is desired. The desired significance is $\alpha = 0.01$.

The same sample of $n = 10$ sacks is used. Assume that you estimated a reasonable standard deviation for this sample as $s = 1$kg. From this data, we can compute the power of this test as:

```
> s <- sd(sample)
> power.t.test(n = 10,
+         delta = 0.5,
+         sd = 1,
+         sig.level = 0.01,
+         type = "one.sample",
+         alternative = "one.sided")
```

```
One-sample t test power calculation
n = 10
delta = 0.5
sd = 1
sig.level = 0.01
power = 0.1654013
alternative = one.sided
```

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 = 0.5, sd = 1, sig.level = 0.01,
               type = "one.sample", alternative = "one.sided")
```

← **(round this value up)**

```
One-sample t test power calculation
n = 47.98044
delta = 0.5
sd = 1
sig.level = 0.01
power = 0.85
alternative = one.sided
```

We need at least 48 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]

[a] Check J.T. Mordkoff's *The assumption(s) of normality* for a nice discussion on this topic: `http://goo.gl/Z3w8ku`

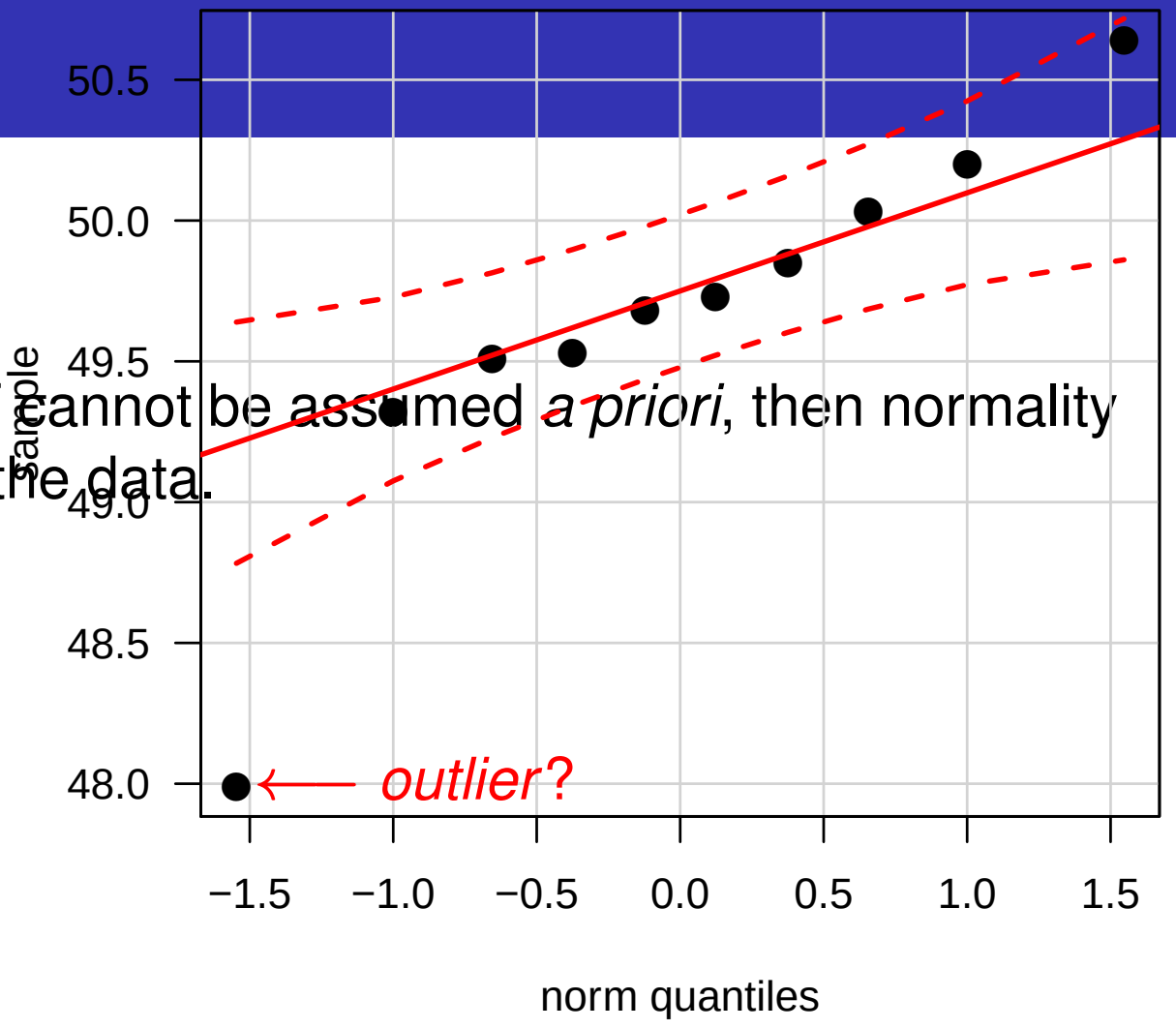
Model validation

The normality assumption

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

Graphical/qualitative tests;

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



Model validation

The normality assumption

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

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

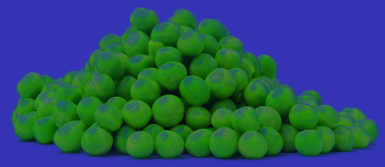
These procedures use different aspects of the sample distribution to test the following hypotheses:

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

In this case, rejection of the null hypothesis suggests evidence that the **sample** came from a non-normal population. Generally we use a strict α threshold for these tests, and consider their results together with a graphical analysis.

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 in 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, the special case of serial autocorrelations in the 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(car)
> durbinWatsonTest(lm(sample ~ 1))

lag Autocorrelation D-W Statistic p-value
1      -0.0848535      2.111733   0.898
Alternative hypothesis: rho != 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

1. D.C. Montgomery, G.C. Runger, *Applied Statistics and Probability for Engineers*, Ch. 9. 5th ed., Wiley, 2010.
2. J.T. Mordkoff (2011), *The assumption(s) of normality* - `http://goo.gl/Z3w8ku`

Recommended reading

1. A. Reinhart, *Statistics Done Wrong: the woefully complete guide* - `http://www.statisticsdonewrong.com`
2. W. Thalheimer and S. Cook, *How to calculate effect sizes from published research articles: A simplified methodology* - `http://goo.gl/c0g1oK`
3. E. Ernst and S. Singh, *Trick or Treatment*, Norton & Company, 2009.

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{Campelo2015-01,
  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 5; Creative Commons BY-NC-SA 4.0.},
}
```