Aston University
Engineering & Applied Science

Design and Analysis of Experiments

05 - Statistical Inference

Version 2.2.2019

Felipe Campelo
f.campelo@aston.ac.uk

Computer Science Group

Birmingham
March 2019

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;
- Predictive power;
- Comparison between alternative hypotheses - principle of parsimony (Ockham's razor);

"*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

Comparing Competing Hypotheses

As multiple, competing hypotheses can be proposed to explain any given phenomenon, it is common to contrast these proposed explanations using a number of criteria.

Besides the required testability / falsifiability and the desired parsimony, common ones include:

- Scope: how applicable is the hypothesis to multiple instances of the phenomenon it addresses?
- External consistency: how well does the hypothesis fit with existing, well-accepted knowledge systems?

Statistical Hypotheses

Definitions

Statistical hypotheses are defined as objective statements about parameters of one or more populations;

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

The *null hypothesis significance testing* (NHST) approach 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 a sample;
- Calculating test statistics;
- Deciding based on the computed value;

Statistical Hypotheses

Example

Suppose that you have implemented an algorithm and you want to verify whether your implementation is correct. A theoretical model of the algorithm behaviour indicates that its convergence should happen on average at iteration 50, so a good initial test would be to run your algorithm a few times and check if it conforms with this prediction.

In this case the null hypothesis could be defined as: *the average iteration at which convergence is achieved is* 50, and the alternative of interest could be expressed as the complementary inequality.

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

Suppose still that $n = 10$ runs are performed and recorded.

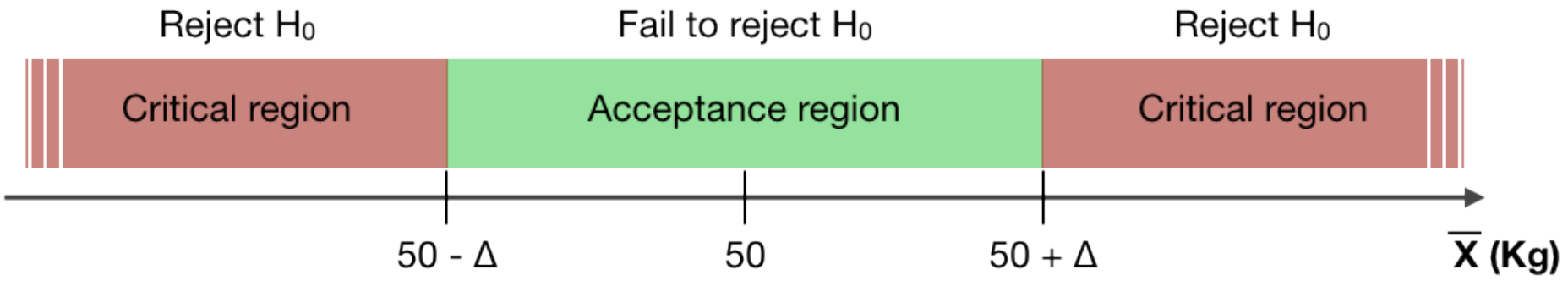
Statistical Hypotheses

Example

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

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

That is, we can use $\bar{x}$ as the basis for a 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 $(1 - \alpha)$ or, sometimes, as $100(1 - \alpha)\%$.

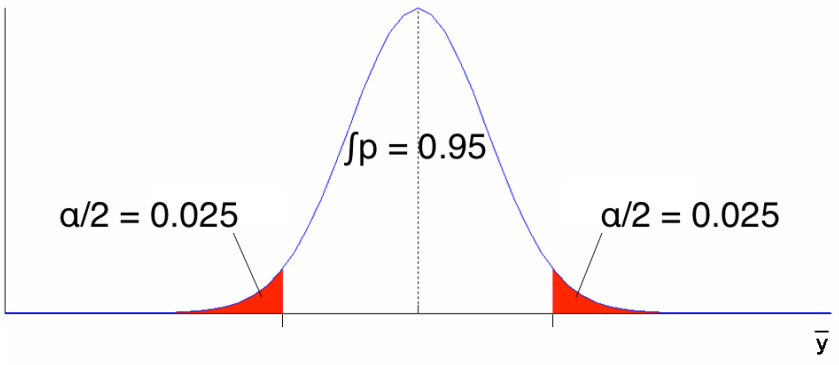
Inferential Errors

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 $\mu = 50$), the distribution of $\bar{X}$ is approximately Normal (assuming that the population is Normal, OR that the CLT conditions are met and N is suitably large), with average 50 and standard error $(\sigma/\sqrt{n})$;

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:

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

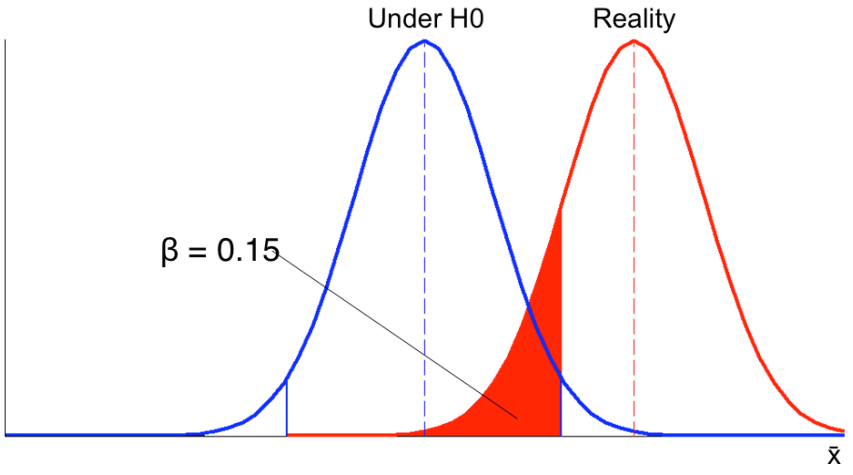
The quantity $(1 - \beta)$ is known as *power* of the test, and quantifies its sensitivity to effects that violate the null hypothesis.

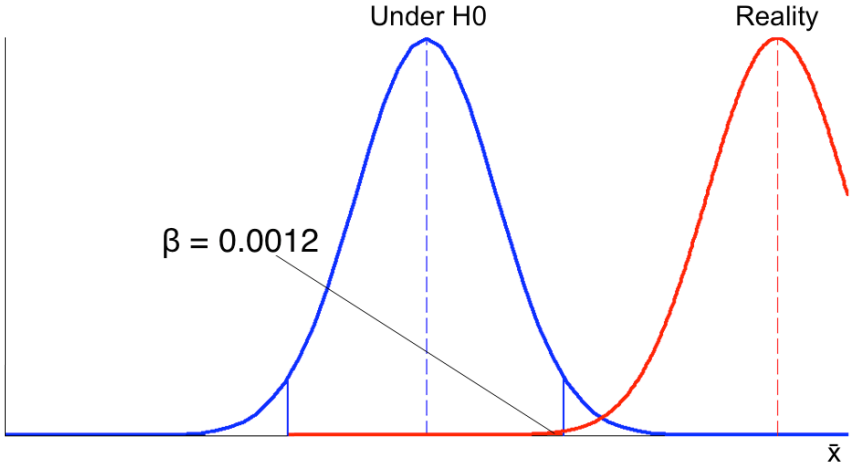
Inferential Errors

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 strengthen 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^*;
- Determine the test statistic and critical region;
- Calculate sample size;
- Compute the statistic;
- Decide whether or not to reject H_0;

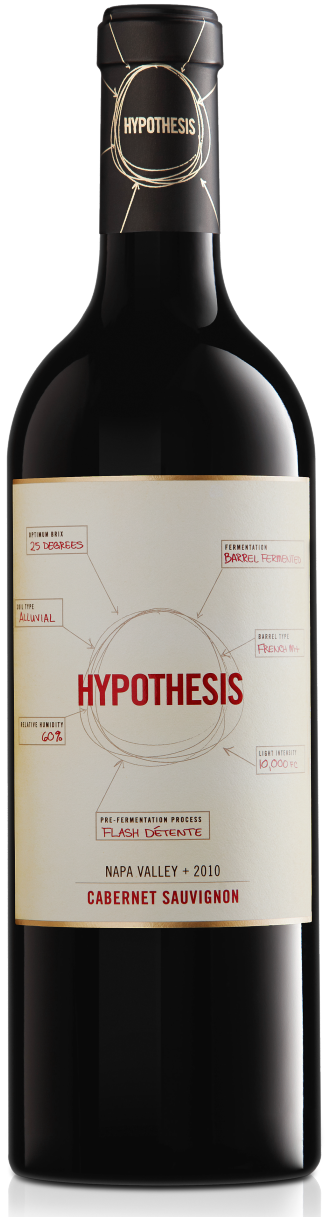


Image (c) Roots Run Deep Winery: `http://www.rootsrundeep.com/hypothesis.html`

Hypothesis Testing

Mean of a Normal distribution, variance known

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

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

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, **under the null hypothesis**, according to the standard Normal, $\mathcal{N}(0, 1)$.

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 $\bar{x} = 49.65$ from our $n = 10$ observations, and that we know that $\sigma = 1$. 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 positive deviations from the theoretical predictions.

The test hypotheses can be defined as:

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

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 $t_\alpha^{(n-1)} = t_{0.99}^{(9)} = 2.82$, which means that under H_0 there is a 99% change that the test statistic will yield a value smaller than this threshold.

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

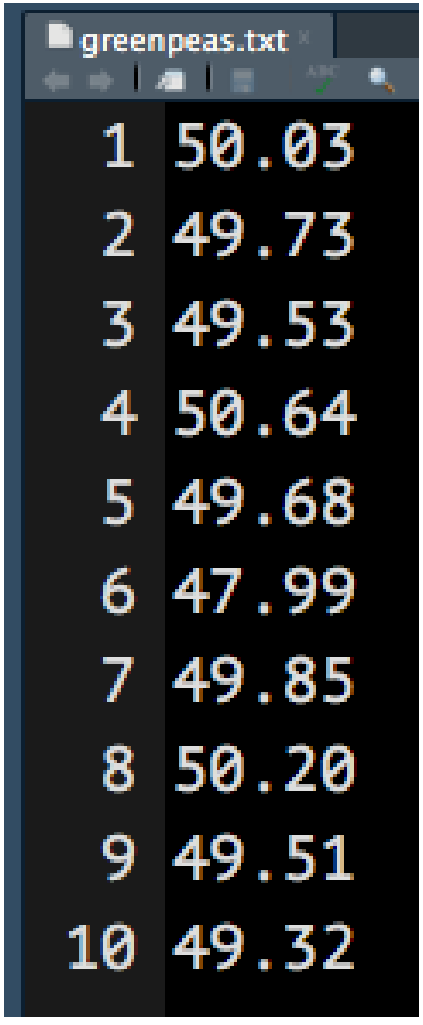
Hypothesis Testing

Mean of a Normal distribution, variance unknown

```
> my.sample <- read.table("../data files/algorithm.txt")

> t.test(my.sample,
+        alternative = "greater",
+        mu = 50,
+        conf.level = 0.99)

One Sample t-test
data:  my.sample
t = -1.5969, df = 9, p-value = 1
alternative hypothesis: true mean is less than 50
99 percent confidence interval:
-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:

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

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(my.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 has 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 is most useful to compute the required sample size for the experiment.

Sample size and Type-II error

Example

Suppose that on the algorithm validation example we are really interested in detecting positive deviations from the nominal value greater than 1%, i.e., $\delta^* = 0.01 \times 50 = 0.5$. 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 $n = 10$ runs is used. Assume that a reasonable upper bound for the standard deviation can be estimated as $\sigma = 1$. From this data, we can compute the power of this test as:

```
> 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")

One-sample t test power calculation
n = 47.98044 ⟵ (round this value up)
delta = 0.5
sd = 1
sig.level = 0.01
power = 0.85
alternative = one.sided
```

We need at least 48 observations to detect a 0.5 (1%) or greater deviation from the theoretical mean value 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`

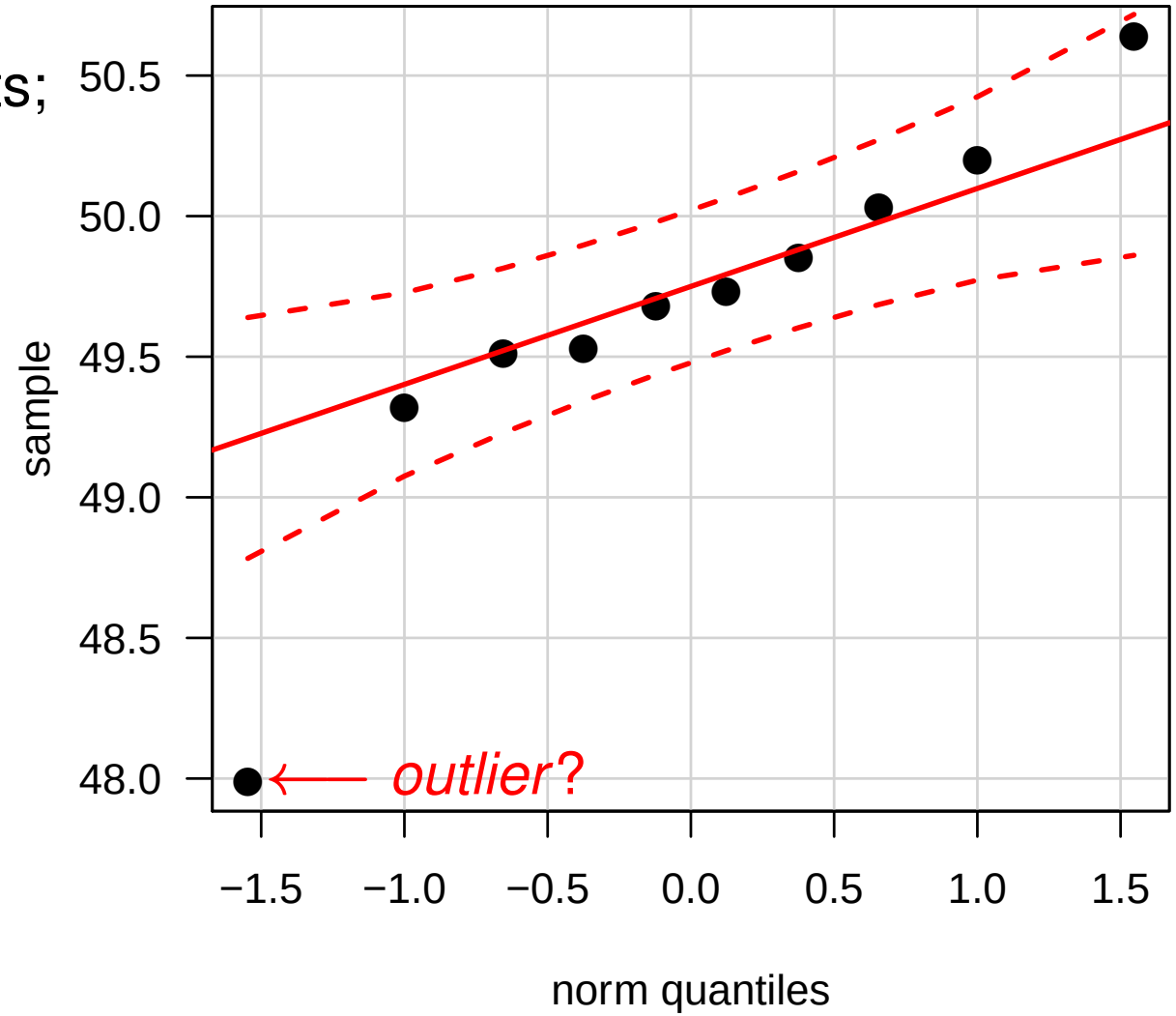
Model validation

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(my.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(my.sample)

Shapiro-Wilk normality test
data:  my.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 unmodelled 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(my.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 algorithm validation experiment

Going over the process

After examining the algorithm validation 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

Recommended 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`
3. A. Reinhart, *Statistics Done Wrong: the woefully complete guide*. No Starch Press, 2015 (`http://www.statisticsdonewrong.com`)

About this material

Conditions of use and referencing

Please reference this work as:

Felipe Campelo (2019), *Lecture Notes on Design and Analysis of Experiments*.
Online: https://github.com/fcampelo/Design-and-Analysis-of-Experiments
Version 2.2. Creative Commons BY-NC-SA 4.0.

```
@Misc{Campelo2019,
  title={Lecture Notes on Design and Analysis of Experiments},
  author={Felipe Campelo},
  howPublished={\url{https://github.com/fcampelo/Design-and-Analysis-of-Experiments}},
  year={2019},
  note={Version 2.2. Creative Commons BY-NC-SA 4.0.},
}
```