UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Design and Analysis of Experiments

05 - Statistical Inference

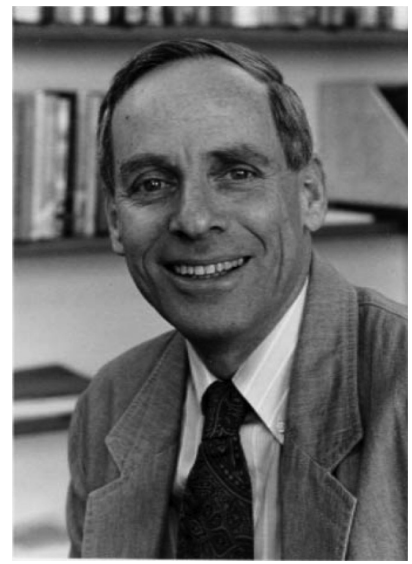
Version 2.11

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

Graduate Program in Electrical Engineering

Belo Horizonte
March 2015

"*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`

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

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;
- Point value for the parameter.

Example: $H_0 : \mu = 25$

Alternative hypothesis (H_1)

- Presence of some effect;
- Existence of something “new”.
- Interval value for the parameter.

Example: $H_1 : \mu \neq 25$

Statistical Hypotheses

Definitions

Determination of the reference value for the null hypothesis H_0:

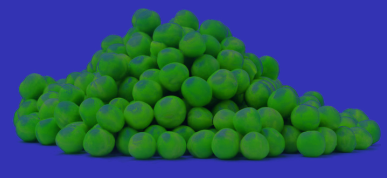
- Previous knowledge about the process (investigation of changes);
- Value obtained from theory or models (model validation);
- Project requirements (investigation of system compliance);

Hypothesis testing involves:

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

Statistical Hypotheses

Example



Suppose you 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* $500g$, 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;

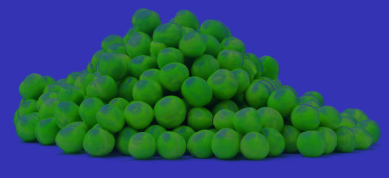
---

[a] We could use any other item on your usual grocery list, but why not pay a little tribute to Gregor Mendel?

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

Statistical Hypotheses

Example

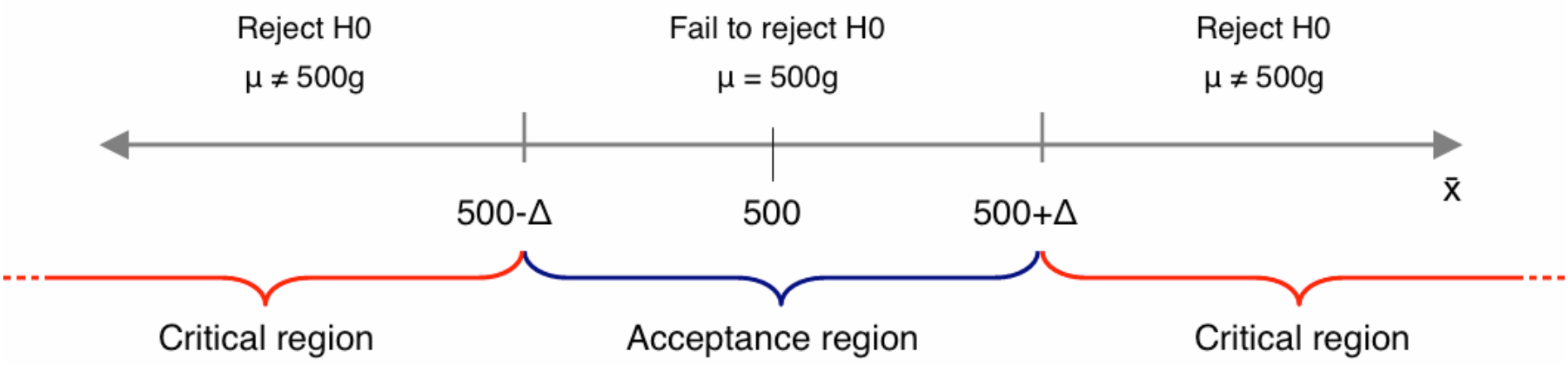


Since the sample mean $\bar{x}$ is a good estimator of the real mean μ, we can assume:

- If $\bar{x} \cong 500$g - corroboration of H_0;
- If $\bar{x} \ll 500$g or $\bar{x} \gg 500$g - 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 $(1 - \alpha)$.

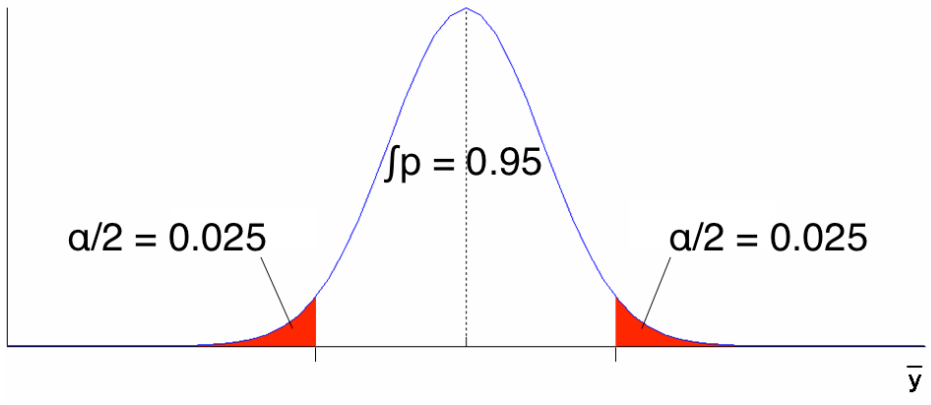
Inferential Errors

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 $1 - \alpha = 0.95$.



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:

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

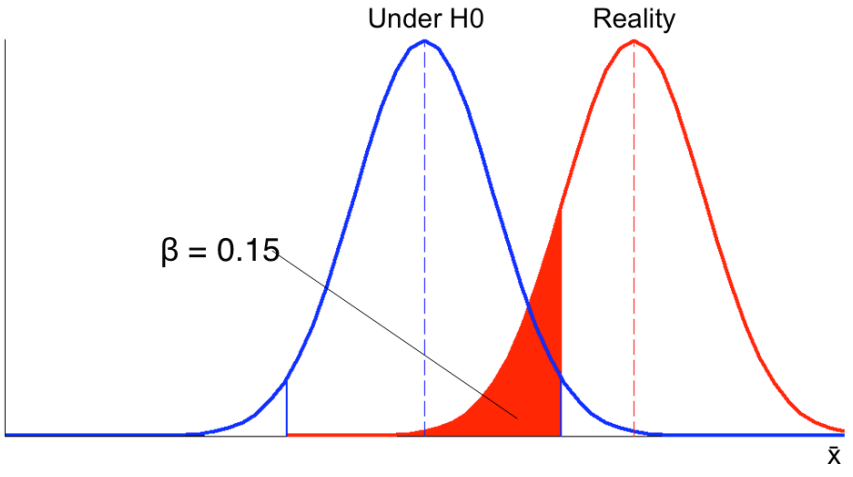
The quantity $(1 - \beta)$ is known as *power* of the test, and quantifies its sensitivity to effects that violate the null hypothesis.

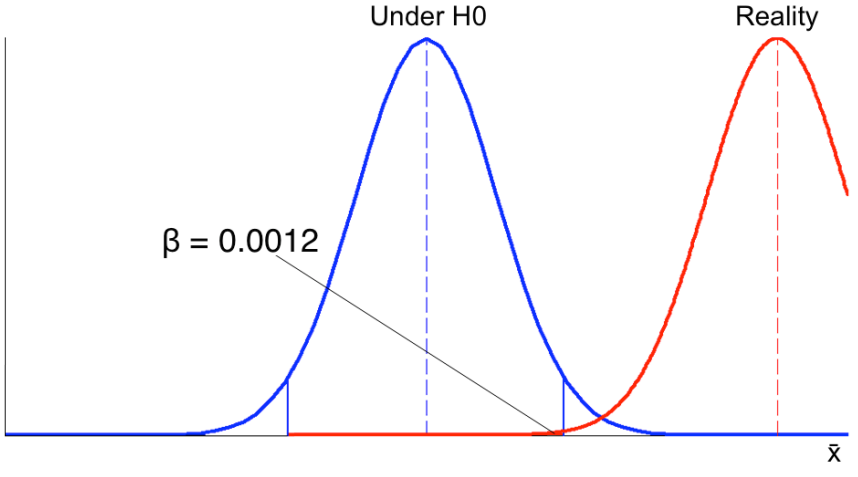
Inferential Errors

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;
- 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);

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;

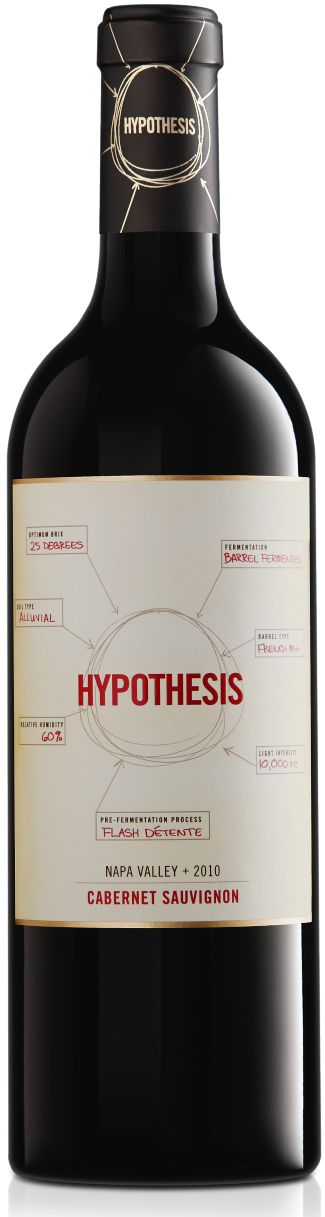
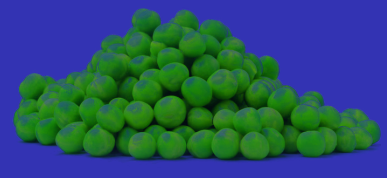


Image (c) Roots Run Deep Winery: `http://www.rootsrundeep.com/hypothesis.html`

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

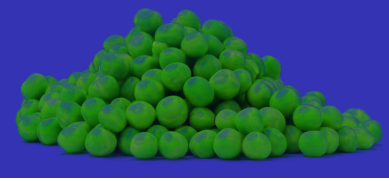
$$\begin{cases} H_0 : \mu = 500g \\ H_1 : \mu \neq 500g \end{cases}$$

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 $\mu_{\bar{X}} = \mu_0 = 500$;

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**.

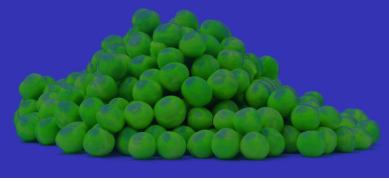
This result implies a probability of $(1 - \alpha)$ that Z_0 will fall within the range $(\pm z_{\alpha/2})$[b] if H_0 is true, which provides a selection criterion between H_0 and H_1:

- 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,

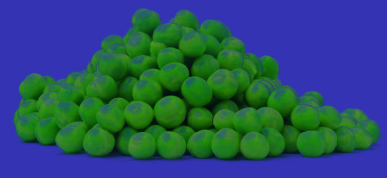
$$z_0 = \frac{496.48 - 500}{10/\sqrt{10}} = -1.113$$

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.

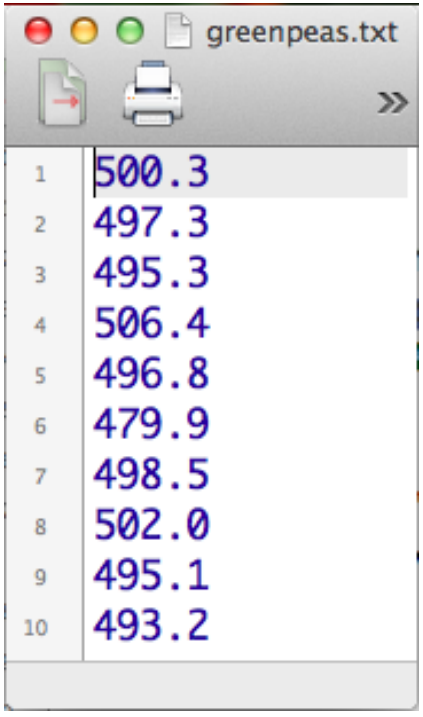
Hypothesis Testing

Mean of a normal distribution, variance known



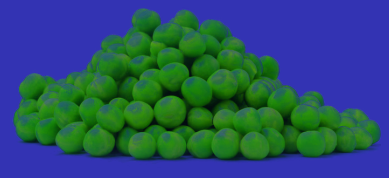
```
> if(!require(TeachingDemos)){
+  install.packages("TeachingDemos")
+  library(TeachingDemos)
+ }
>
> sample<-scan("../data files/greenpeas.txt")
> z.test(as.numeric(sample),
+        mu=500,
+        stdev=10)

One Sample z-test
data:  as.numeric(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:

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

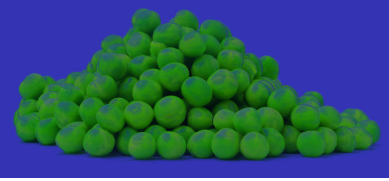
In this second scenario we want to be more conservative, so we pick a significance level of $\alpha = 0.01$;

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

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

Hypothesis Testing

Mean of a normal distribution, variance unknown



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

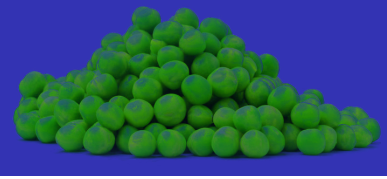
$$t_0 = \frac{496.48 - 500}{6.97/\sqrt{10}} = -1.5969$$

The critical value of this test statistic for the desired significance is $-t_{\alpha,n-1} = -t_{0.01,9} = -2.54$;

Given that $t_0 > -2.54$, 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



```
> t.test(sample,
+        alternative = "less",
+        mu = 500,
+        conf.level = 0.99)

One Sample t-test
data:  sample
t = -1.5969, df = 9, p-value =0.07237
alternative hypothesis: true mean is less than 500
99 percent confidence interval:
     -Inf 502.6991
sample estimates:
mean of x
   496.48
```

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 $H_0 : \mu = 500g$ 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?

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:

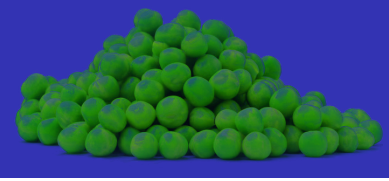
$$d = \frac{\bar{x} - \mu_0}{s}$$

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 $H_0 : \mu = 500$ 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
```

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;

Sample size and Type-II error

Some considerations

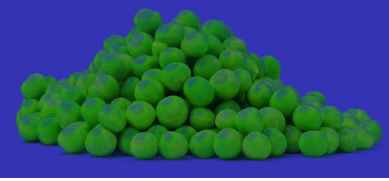
Having obtained this estimation of the Type-II error probability, one can run his/her experiment with a better understanding of its ability to detect effects of interest.

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:

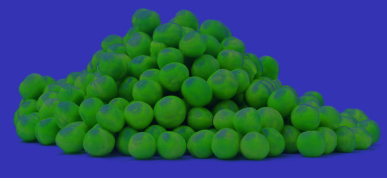
```
> s<-sqrt(var(sample))
> power.t.test(n=10,
+         delta=5,
+         sd=s,
+         sig.level=0.01,
+         type = "one.sample",
+         alternative = "one.sided")
```

```
One-sample t test power calculation
n = 10
delta = 5
sd = 6.970382
sig.level = 0.01
power = 0.3474724
alternative = one.sided
```

⟵ **Power**

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

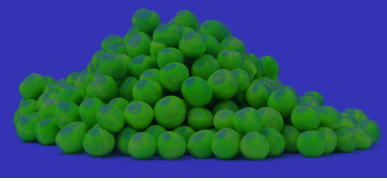
The assumption of normality, required for the **z** and **t** tests, needs to be validated.

- Graphical/qualitative tests;
- Analytical/quantitative tests;

It is interesting to use both whenever possible.

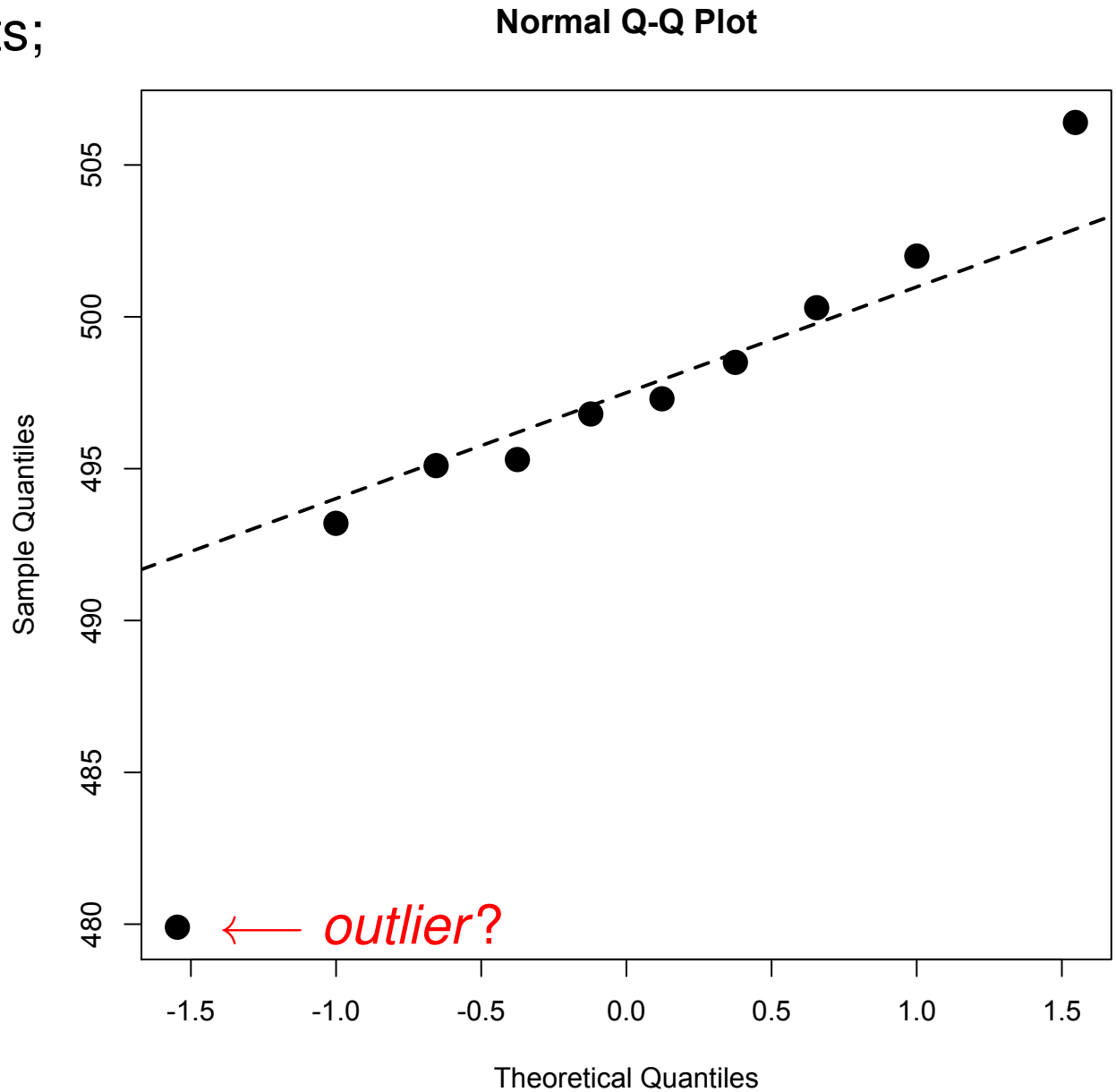
Model validation

The normality assumption



Graphical/qualitative tests;

```
> qqnorm(sample,
+        pch=16, cex=2)
> qqline(sample,
+        lty=2, lwd=2)
```



Model validation

The normality assumption

There is a great variety of analytical tests of normality;

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

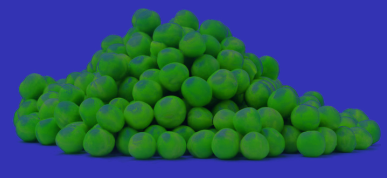
These procedures use different aspects of the sample distribution to test the following hypotheses:

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

In this case, rejection of the null hypothesis suggests evidence that the sample came from a non-normal population. Generally we will use a low α (e.g., 0.01) for these tests, and will consider their results together with a graphical analysis.

Model validation

The normality assumption



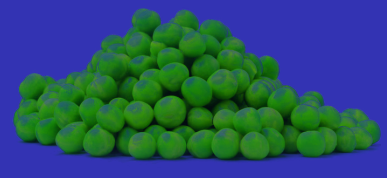
Even thought the Lilliefors test is possibly the most widely used for normality testing, the Shapiro-Wilk test tends to be more sensitive and has some other interesting properties, so we will be using it throughout the course.

```
> shapiro.test(sample)

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

Model validation

The independence assumption



Another assumption of the statistical model used for the t-test is the independence of the residuals, that is, the absence of external (unmodeled) biases contaminating the data.

While there is no procedure to test independence in the general case, the Durbin-Watson test can be used to evaluate serial autocorrelations in the residual data, which can be useful to investigate contamination by order-dependent effects.

```
> 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

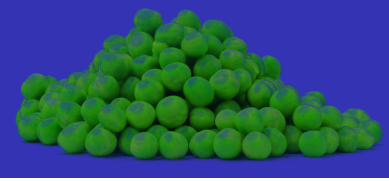
The independence assumption

As a test of serial autocorrelation, 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 deal with during the analysis, 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. W. Thalheimer and S. Cook, *How to calculate effect sizes from published research articles: A simplified methodology* - `http://goo.gl/c0g1oK`

Recommended reading

1. A. Reinhart, *Statistics Done Wrong: the woefully complete guide* - `http://www.statisticsdonewrong.com`
2. S. Okasha, *Philosophy of Science: A Very Short Introduction*, Oxford University Press, 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 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.},
}
```