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

CC BY NC SA

"*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

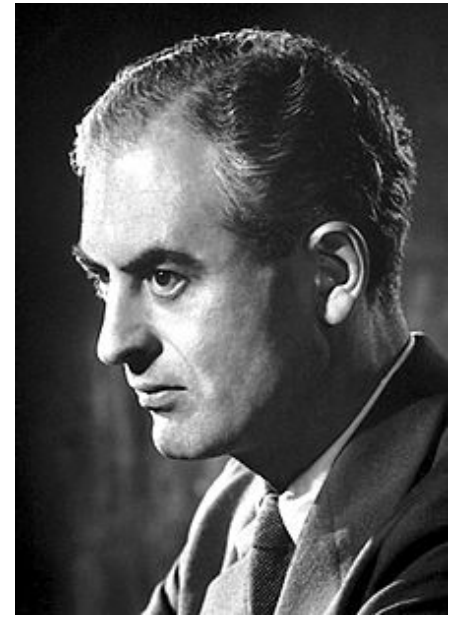


Image: http://commons.wikimedia.org/wiki/File:Peter_Brian_Medawar.jpgQ

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;

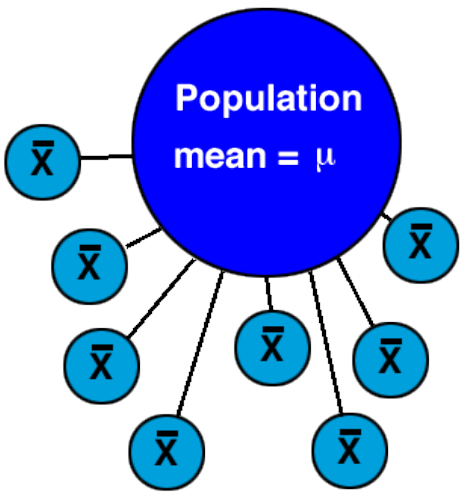
- **Point estimate**: estimated value for a given population parameter;
- **Statistical interval**: estimated interval of possible/probable values for a given population parameter;

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]:

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

and:

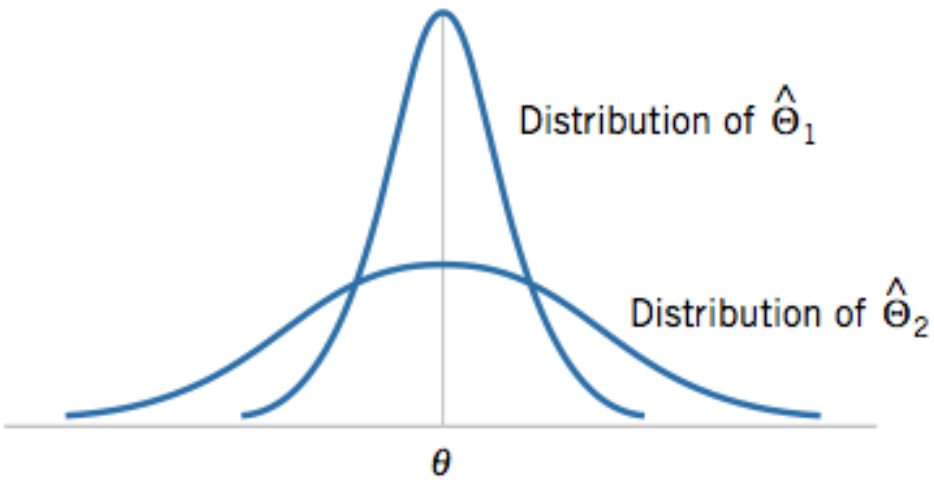
$$E\left[s^2\right] = E\left[\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2\right] = \sigma^2$$

[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 the most usual point estimates used with Gaussian variables, we have:[5]

$$\hat{\sigma}_{\bar{X}} = \frac{s}{\sqrt{n}}$$

$$\hat{\sigma}_{s^2} = s^2 \sqrt{\frac{2}{n-1}}$$

$$\hat{\sigma}_s = \frac{s}{\sqrt{2(n-1)}} + O\left(\frac{1}{n\sqrt{n}}\right) \approx \frac{s}{\sqrt{2(n-1)}}$$

More general estimates of the standard error for different distributions (or different statistics) can usually be obtained using resampling strategies or asymptotic results.

---

[5] See Ahn and Fessler (2003), *Standard Errors of Mean, Variance, and Standard Deviation Estimators*: https://git.io/v5Z5v

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:

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

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.

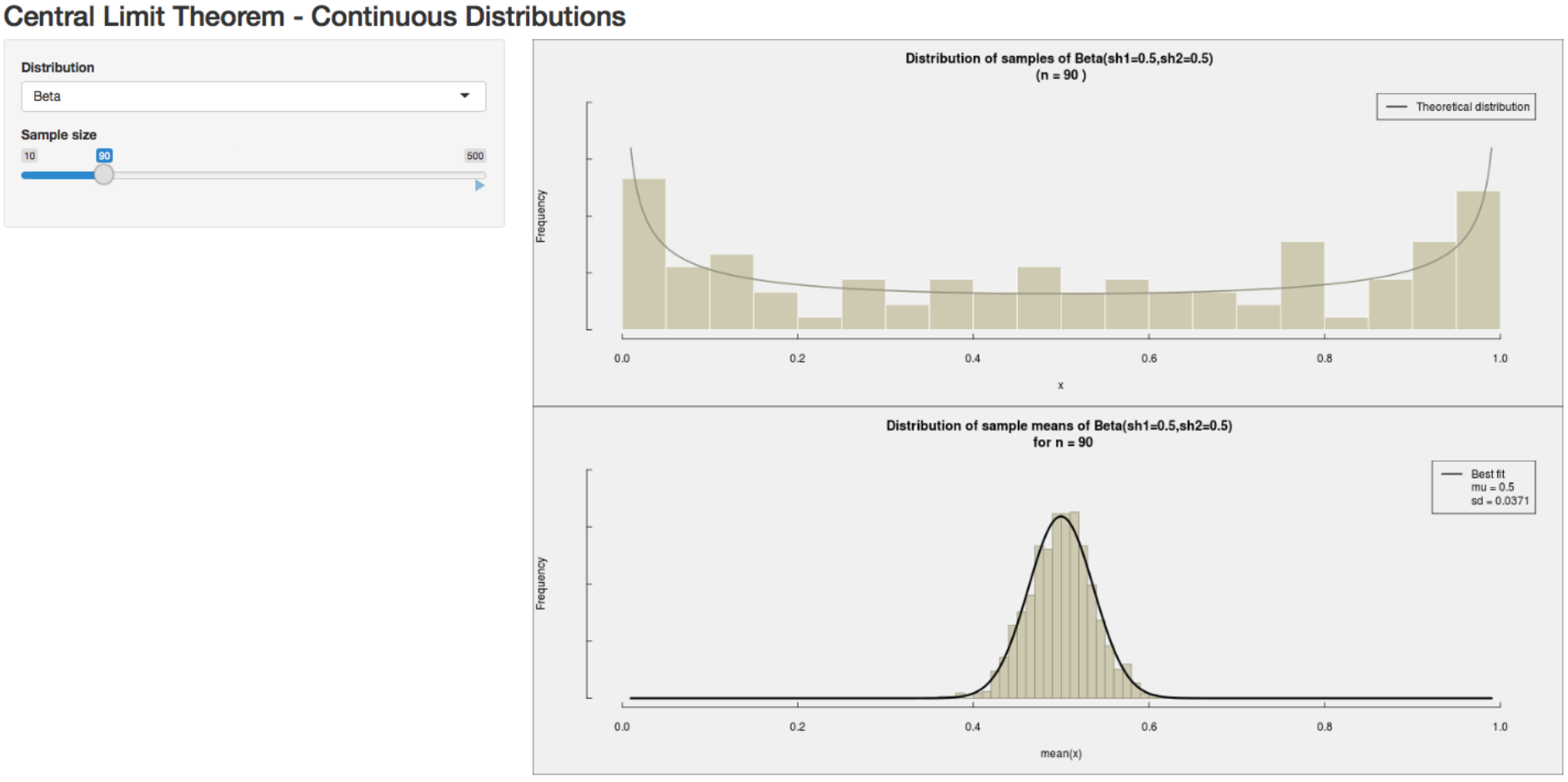
For “well-behaved” distributions (continuous, symmetrical, unimodal - the usual bell-shaped pdf we all know and love) even small sample sizes are commonly enough to justify invoking the CLT and using parametric techniques.

Sampling Distributions

The Central Limit Theorem

For an interactive demonstration of the CLT, check

`http://orcslab.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

This work is licensed under the Creative Commons CC BY-NC-SA 4.0 license (Attribution Non-Commercial Share Alike International License version 4.0).

http://creativecommons.org/licenses/by-nc-sa/4.0/

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.},
}
```