UF _m_ G

UNIVERSIDADE FEDERAL
DE MINAS GERAIS

# Design and Analysis of Experiments
## 03 - Point Estimators

Version 2.11

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

Graduate Program in Electrical Engineering

Belo Horizonte
March 2015

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

Sir Peter B. Medawar
1915-1987
British Immunologist
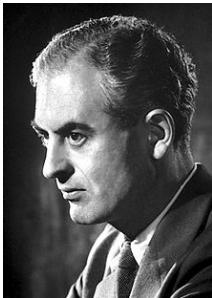
**Statistical inference**: using *samples* to draw conclusion about *populations*;

## 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

"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

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;

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;
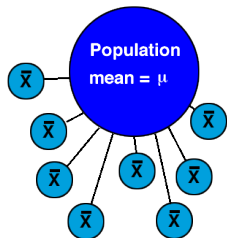


Being random variables means that statistics also have their own probability distributions, called *sampling distributions*[3]. Sampling distributions have specific characteristics that we'll explore later.

---

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

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

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

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

Considera also a random sample from this variable:
$\mathbf{y} = \{y_1, y_2, \ldots, y_N\}$;

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

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

- a population mean, $\mu$;
- a population variance, $\sigma^2$;
- 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.

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

A given estimator $\hat{\Theta}$ is said to be *unbiased* for parameter $\theta$ 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 $y_1, \ldots, y_N$ be a random sample from a given population $Y$, characterized by its mean $\mu$ and variance $\sigma^2$. In this situation, it is possible to show that[4]:

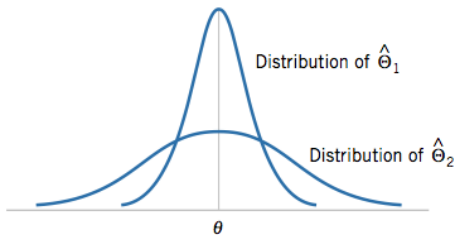$$E[\bar{y}] = E\left[\frac{1}{N}\sum_{i=1}^{N} y_i\right] = \mu$$

and:

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

---

[4] See http://goo.gl/i7bg3V for details.

# Point Estimators
Unbiased estimators

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



Distribution of $\hat{\Theta}_1$

Distribution of $\hat{\Theta}_2$

$\theta$

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 $\theta$.

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).

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

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

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

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

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,

$$\bar{y} = \frac{1}{25} \sum_{i=1}^{25} y_i$$

is also normally distributed, with $E[\bar{y}] = \mu = 50\Omega$ (since the sample mean is an unbiased estimator) and $s_{\bar{y}} = \sigma^2/25 = 0.16$.

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

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

More generally, let $y_1, \ldots, y_n$ be a sequence of *independent and identically distributed* (**iid**) random variables, with mean $\mu$ and finite variance $\sigma^2$, and let $x = \sum y_i$. Then:

$$z_n = \frac{x - n\mu}{\sqrt{n\sigma^2}}$$

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

---

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

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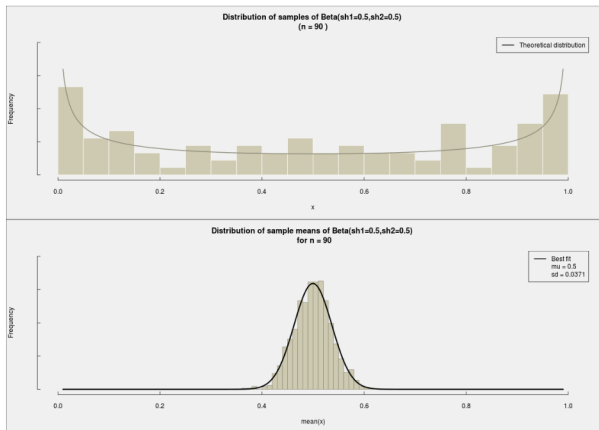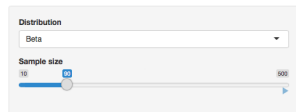
For an interactive demonstration of the CLT, check
`http://drwho.cpdee.ufmg.br:3838/CLT/`

# Bibliography

**Required reading**

1. D.C. Montgomery and G.C. Runger, *Applied Statistics and Probability for Engineers*, Chapter 7. 3rd Ed., Wiley 2005.

2. D.W. Stockburger, *Sampling Distributions*. In: *Introductory Statistics: Concepts, Models, and Applications* - http://www.psychstat.missouristate.edu/introbook/sbk19.htm

**Recommended reading**

1. R. Willett, *ECE 830 Estimation and Decision Theory, Spring 2014*, Chapters 13-15 - http://willett.ece.wisc.edu/education.html

2. S. Okasha, *Philosophy of Science - a very brief introduction*, Oxford Paperbacks, 2002.

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.},
}
```