Experiment Design for Computer Sciences (01CH740)

Topic 02 - Point and Interval Indicators

Claus Aranha
caranha@cs.tsukuba.ac.jp

University of Tsukuba, Department of Computer Sciences

April 13, 2021

Version 2021.2 (Updated April 13, 2021)

Part I - Introduction

Lecture Outline

In the last lecture, we talked about how experiments are important for the scientific endeavor. In this lecture, we will focus on basic concepts and techniques for representing experimental data.

- The importance of characterizing your data;
- Statistical Concepts: Population and Sample;
- Characterizing a Population: Point Estimators and Interval Estimators;
- Visualizing Estimators;

Part II - Indicators

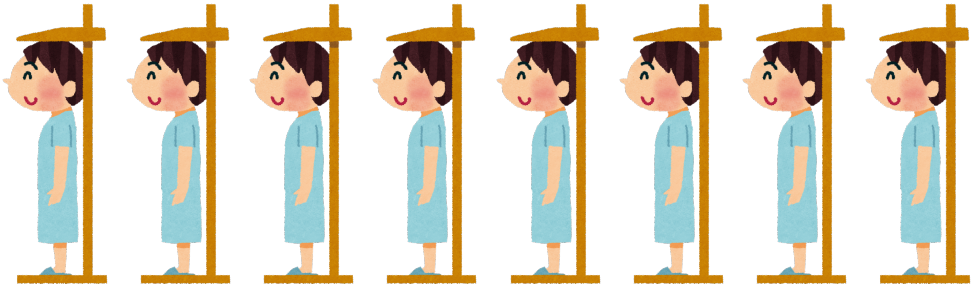
Using data to characterize a system

In the last lecture, we talked about using experiments to obtain data. How can we use the data to gain knowledge about a system?

For example, let's say that we measure the height of one student. Can I use this information to say something about the height of all students in the university?

Using data to characterize a system

From a single student's height, it seems hard to learn anything about the height of the students of the university **in general**. A better approach would be to take the height of several students.



If I calculate the average height of several students, what is the relationship between this value and height of students in general?

Population and Sample

This example introduces us to the concepts of statistical Population and Sample.

- **Population**: It is a large set of objects that are of interest as a whole. It can be a real set (all students in a university) or a theoretical set (all possible results of an experiment).
- **Observation**: It is one element from the population. One student from the university, or one execution of an experiment.
- **Sample**: It is a subset of the population. By examining the sample, we can make inferences about the population as a whole.

"Making inferences of the population from the sample". What does it mean?

Population and Sample

Making Inferences from the sample

Population

A pool has many colorful balls. If you knew exactly how many, you could calculate the probability of picking a ball of a certain color at random. **But you don't know!**

Sample

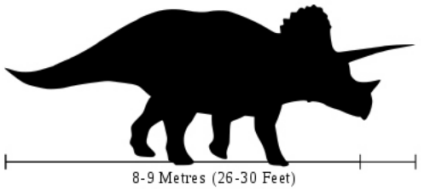
It is possible to **estimate** the proportion of balls in the pool. To do that, we pick up a number of balls, and examine the proportion of the sample.

Population and Parameters

Our usual goal when obtaining and analysing experimental data is to estimate values for parameters of the population.

A **parameter** θ of the population is an unknown value of interest that characterizes some important aspect for our research.

Because we cannot observe the population directly, we have to estimate the parameter's true value from information gathered from the sample.



Parmeter	Average Size	Maximal Size
Length	7.5–8.5 meters (25–28 ft)	9 meters (30 ft)
Height	2–2.9 meters (6.6–9.5 ft)	3 meters (9.8 ft)
Mass	6,000–8,500 kilograms (13,200–18,700 lb)	10–12 tonnes (22,000– 26,000 lb)
Skull Length	2–2.2 meters (6.6–7.2 ft)	2.5–2.8 meters (8.2– 9.2 ft)
Brow Horns Length (horn core)	70–100 centimetres (28–39 in)	115 centimetres (45 in)

Samples and Statistics

By observing data obtained from a sample, we can **characterize** (estimate) parameters from a population of interest. For example:

- We calculate the average of the running time of multiple executions of a program, and estimate the mean running time;
- We ask the age of several students in a school, and estimate the maximum and minimum age of the students;
- We estimate the efficacy of a remedy by counting what percentage of patients get better after drinking it;
- We determine which of two neural networks is more precise by subtracting the test error of the two networks from each other;

Statistic

A statistic is a **function** calculated from data obtained from a sample.

Point and Interval Indicators

The idea of estimating parameters of a population using information obtained from a sample is called **statistical inference**.

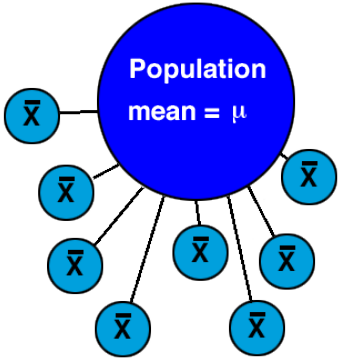
In this lecture, we will focus on two central concepts of statistical inference: **Point Estimators** and **Interval Estimators**.

- **Point Estimators**: are statistics that estimate the value of a population parameter from information in a sample;
- **Interval Estimators**: are statistics that estimate a range of values of a population parameter from information in a sample;

Statistics and Sampling Distributions

Suppose that you want to obtain a point estimate for an arbitrary parameter of the population (e.g. mean size);

Random samples of the population can be interpreted as a **random variable**, and any function of these samples (any statistic) will be a random variable as well.



As a random variable, any statistic also has its own probability distribution, called **sampling distribution**.

Most statistical tests use properties of the sampling distributions (which are not the same as the true distribution of the population). We will talk more about those later.

Definition of Point Indicator

A *point estimator* is a statistic which provides the value of maximum plausibility for an (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: $x = \{x_1, x_2, \ldots, x_n\}$;

A given function $\hat{\Theta} = h(x)$ 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.

Examples

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.

Errors and Biases

Note that we are being very careful to always use the word **estimate** when we talk about statistics. Why is that?

In all the examples that we mentioned, if we are unlucky[1], we could obtain an estimate that is very different from the true value of the population:

Bad statistics example

To estimate the height of the students of a school, we pick 10 students, and we measure the height of the youngest one.

- **Error**: The difference between an estimate and the true value of a population's parameter;
- **Bias**: The property of a statistic that systematically produces wrong estimates;

[1] or careless, or malicious

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.

Unbiased estimators

For example, the usual estimators for mean is an unbiased estimator;

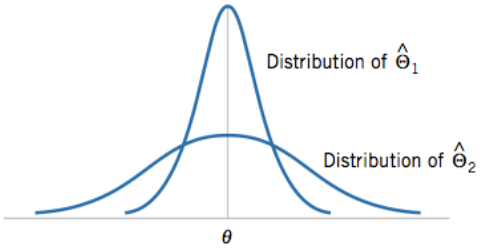
Let $x_1, \ldots, x_n$ be a random sample from a given population X, characterized by its mean μ and variance σ^2. In this situation, it is possible to show that:

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

(Remember that the expected value of one observation is the mean value of the population)

Unbiased estimators

Fora a population parameter θ, it is usually possible to define more than one unbiased estimator. The variances of these estimators may, however, be different



In these cases, we usually want 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.

Standard error of a point estimator

Remember that because a point estimator is a random variable, it has an associated distribution and error. For example, the standard error of an estimator $\hat{\Theta}$ is

$$\sigma_{\hat{\Theta}} = \sqrt{Var\left[\hat{\Theta}\right]}$$

However, we can't know this directly. We can **estimate** the standard error of the estimator from the data in the sample. In this case, 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).

Standard error of a point estimator

Examples

Assuming a random variable X from a gaussian distribution, and a sample error s, we can calculate the standard errors of several common point indicators[2]

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

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

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

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

Point Estimator Use Case

Consider an operation to produce coaxial cables[3]. The mean resistance of the production is 50Ω, with a standard deviation of 2Ω (Population mean, and population deviation).

Let's assume that the resistance value of the produced cables are distributed follow a normal distribution ($X \sim \mathcal{N}\left(\mu = 50, \sigma^2 = 4\right)$)

[3] Example inspired on `https://www.sas.com/resources/whitepaper/wp_4430.pdf`

Point estimator use case

Suppose that we take a random sample of 25 cables is taken from this production process (an experiment, to measure if the process is correct, for example).

The sample mean of the the observations is:

$$\bar{x} = \frac{1}{25} \sum_{i=1}^{25} x_i$$

The sample mean follows a normal distribution, with $E[\bar{x}] = \mu = 50\Omega$[4] and $\sigma_{\bar{x}} = \sqrt{\sigma^2/25} = 0.4\Omega$. The error depends on the sample size.

[4] since the sample mean is an unbiased estimator

The Central Limit Theorem

In the previous example, the production operation followed a normal distribution. But even for a population with an arbitrary distribution, the sampling distribution of its mean tends to be approximately normal.

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

The Central Limit Theorem

This result is known as the *Central Limit Theorem*[5], and is one of the most useful properties for statistical inference. The CLT allows the use of techniques based on the Normal distribution, even when the population under study is not normal.

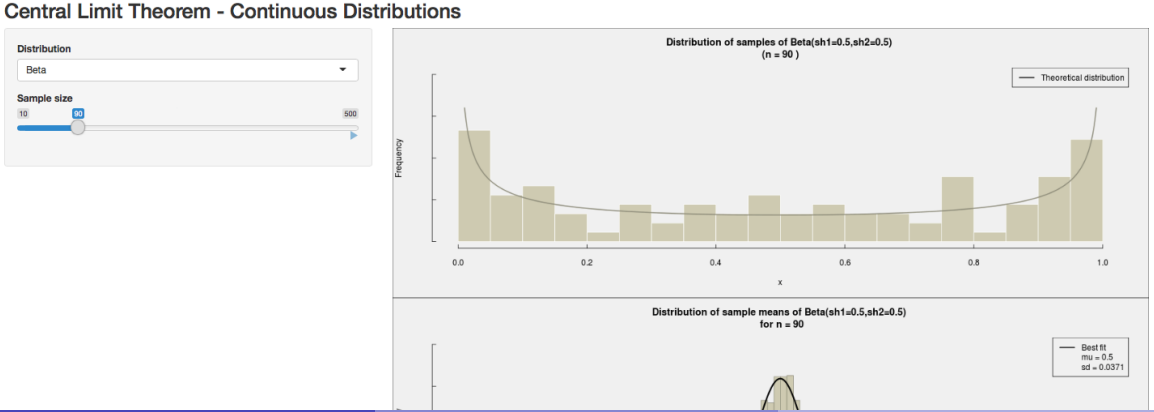
For "well-behaved" distributions (continuous, symmetrical, unimodal - the usual bell-shaped pdf we all know and love) even small sample sizes are commonly enough to justify invoking the CLT and using parametric techniques.

[5] For more details on the CLT, see
https://www.encyclopediaofmath.org/index.php/Central_limit_theorem

Sampling Distributions

The Central Limit Theorem

For an interactive demonstration of the CLT, download the files in `https://git.io/vnPj8` and run on RStudio.



Statistical Intervals

Statistical intervals are important in quantifying the uncertainty associated to a given estimate;

As an example, let's recap the coaxial cables example: *a coaxial cable manufacturing operation produces cables with a target resistance of* 50Ω *and a standard deviation of* 2Ω*. Assume that the resistance values can be well modeled by a normal distribution*.

Let us now suppose that a sample mean of $n = 25$ observations of resistance yields $\bar{x} = 48$. Given the sampling variability, it is very likely that this value is not exactly the true value of μ, but we are so far unable quantify how much uncertainty there is in this estimate.

Definition

Statistical intervals define regions that are likely to contain the true value of an estimated parameter.

More formally, it is generally possible to quantify the level of uncertainty associated with the estimation, thereby allowing the derivation of sound conclusions at predefined levels of certainty.

Three of the most common types of interval are:

- Confidence Intervals;
- Tolerance Intervals;
- Prediction Intervals;

How to Interpret a Confidence Interval

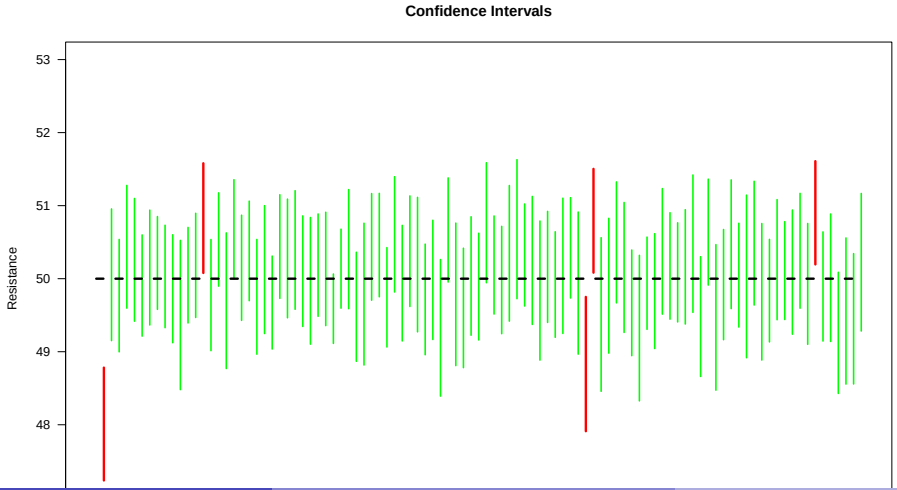
Confidence intervals quantify the degree of uncertainty associated with the estimation of population parameters such as the mean or the variance.

Can be defined as "*the interval that contains the true value of a given population parameter with a confidence level of* $100(1-\alpha)$";

Another useful definition is to think about confidence intervals in terms of confidence *in the method*: "The method used to derive the interval has a hit rate of 95%" - i.e., the interval generated has a 95% chance of 'capturing' the true population parameter."

Confidence Intervals

Example: 100 $CI_{.95}$ for a sample of 25 observations



CI on the Mean of a Normal Variable

The two-sided $CI_{(1-\alpha)}$ for the mean of a normal population with known variance σ^2 is given by:

$$\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where $(1-\alpha)$ is the confidence level and z_x is the x-quantile of the standard normal distribution.

For the more usual case with an unknown variance,

$$\bar{x} + t^{(n-1)}_{\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t^{(n-1)}_{1-\alpha/2}\frac{s}{\sqrt{n}}$$

where $t^{(n-1)}_x$ is the x-quantile of the t distribution with $n-1$ degrees of freedom.

CI on the Variance and Standard Deviation of a Normal Variable

A two-sided confidence interval on the variance of a normal variable can be easily calculated:

$$\frac{(n-1)s^2}{\chi^{2(n-1)}_{1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^{2(n-1)}_{\alpha/2}}$$

where $\chi^{2(n-1)}_{x}$ represents the x-quantile of the χ^2 distribution with $n-1$ degrees of freedom. For the standard deviation one simply needs to take the squared root of the confidence limits.

Wrapping up

Statistical intervals quantify the uncertainty associated with different aspects of estimation;

Reporting intervals is always better than point estimates, as it provides the necessary information to quantify the location and uncertainty of your estimated values;

The correct interpretation is a little tricky (although not very difficult)[7], but it is essential in order to derive the correct conclusions based on the statistical interval of interest.

[7] See the table at the end of `https://git.io/v5ZFh`

Part III - Outro

Summary

Descriptive Statistics

Experiment Data can be used to estimate facts about the world:

- Point estimators: Sample Means, Variance, Correlation, etc.
  - Give us specific information about the model we want to study
  - "What is the average height of a student?"
  - **An estimator is not the real value!**
- Interval estimators: Confidence Interval, IQR, etc.
  - Give us more information than point estimators.
  - "How certain should I be about this point estimator".
  - Size of interval estimator depends on the number of samples.

Recommended Reading

- *D.W. Stockburger*, The Sampling Distribution. In: Introductory Statistics: Concepts, Models, and Applications - `http://psychstat3.missouristate.edu/Documents/IntroBook3/sbk17.htm`
- *J.G. Ramírez*, Statistical Intervals: Confidence, Prediction, Enclosure: `https://git.io/v5ZFh`
- Crash Course Statistics Playlist, in particular videos #3 to #7: `https://www.youtube.com/playlist?list=PL8dPuuaLjXtNM_Y-bUAhblSAdWRnmBUcr`

Programming in R

The material for this week includes some coding examples. These examples are written in the **R** language.

Although we will have an R tutorial in the future, you can read the following material to get yourself acquainted with R:

- R for beginners: `https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf`
- Rstudio: `https://rstudio.com`

About these Slides

These slides were made by Claus Aranha, 2021. You are welcome to copy, re-use and modify this material.

These slides are a modification of "Design and Analysis of Experiments (2018)" by Felipe Campelo, used with permission.

Individual images in some slides might have been made by other authors. Please see the following references for those cases.

Image Credits I

[Page 5] Height image from `https://www.irasutoya.com`
[Page 9] Triceratops information table CC by Zachi Evenor and MathKnight
[Page 18] Image: D.C.Montgomery,G.C. Runger, *Applied Statistics and Probability for Engineers*,Wiley 2003.
[Page 21] Coaxial cable image from `https://pixabay.com`