Experiment Design for Computer Sciences (0AL0400)

Topic 02 - Point and Interval Indicators

Claus Aranha
caranha@cs.tsukuba.ac.jp

University of Tsukuba, Department of Computer Sciences

April 21, 2022

Version 2022 (Updated April 21, 2022)

Lecture Outline

In the first lecture, we talked about what is science, and how experiments (carefully designed experiments) are important in science.

Starting from this lecture, we will talk about how we use statistics to understand the data that we gather from experiments, and how we can draw conclusions about them.

Topics for today:

- Experimental Data: Population, Observation and Sample;
- Point indicators: How we obtain information about the population from samples;
- Interval indicators: Indicators with quality info!

Lecture Outline

Data and "Experiment Data"

When we talk about "data" in Computer Science, the first thing that comes to mind is "information that we feed to a program". **For example**: images, network logs, user databases, etc.

In this course, we are talking about "Experiment Data", which should be understood as "Data about the result of an experiment". **For example**: How long did the experiment take? What is the success rate of my program?

In fact, we can use the tecniques of this lecture for the first kind of data too! But to make things simpler, let's concentrate on the second kind of data.

Example of Data Collection

Using Experiment Data to characterize a system

Tsukuba University is famous for many olympic athletes. Let's say we want to investigate if this is because some special characteristic of the students. After thinking a bit, I come up with the following question:

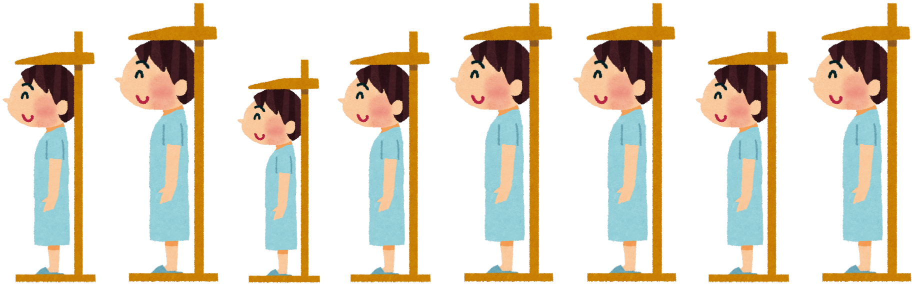
- Is the olympic performance related with the height of the students?
- Are students in Tsukuba taller than normal in Japan?

The 1st question seems hard to answer right now, but 2nd question seems easier.

Imagine I take ONE student from Tsukuba and measure their height. Can this information answer the second question? Why?

Using experiment data to characterize a system

From the height of only one student, I cannot really learn anything about the height of the students of the university **in general**. A better approach would be to measure the height of several students.



Now that I have the height of many students, what can I say about the height of the students in the university, **in general**?

Population and Sample

This example introduces us to some important concepts in statistics: Population, Observation, and Sample.

- **Population**: This is a set of objects that we want to learn more about, using experiments. It can be a real set (all students of the university), or a theoretical set (all possible results of running a program).
- **Observation**: This is one element from the population. One student from the university, or one execution of the program. One data point from an experiment.
- **Sample**: This is a set of observations. A subset of the population.

Our goal, when we analyse data from an experiment, is to **"learn something about the population, by examining the observations in the sample"**.

Population and Sample

Learn something about the population, by examining the observations in the sample

Population

You want to know the proportion of colorful balls in a pool (you like the red ones). Because we don't know exactly how many there are, we need to **make an estimation.**

Sample

To learn the proportion of red balls, we pick a number of balls, and **estimate that the proportion of red balls in the pool is equal to the proportion in my hand.**

Population, Model and Parameters

What is a model?

A model is a description of the population, focused on the scientific questions that we want to make.

- The balls are distributed evenly in the pool, so I can take my sample from any part.
- The height of the students in the university can be represented by a **normal curve**, with mean μ and standard deviation *sd*.
- The SIR infection model says that **susceptible** people become **infected**, and then **Recovering**, so if we can learn the number of people in each group and the transition probability, we can predict the progress of a disease.

The goal of many experiments is to use data to **estimate the parameters of a model.**

Population, Model and Parameters

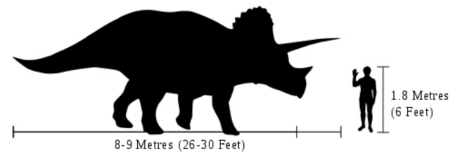
Example of model parameters

By analysing the sample data obtained from an experiment, we can estimate values for the parameters of the model.

The **true value** of the parameters in the model (what we don't know) is usually called θ. The values estimated from the experiment data is named $\hat{\theta}$. Note the difference!

The **model** is determined during the experiment design phase.

Every model is wrong, but some models are useful!



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)
Brow Horns Length	97–130 centimetres (38–51 in)	130–150 centimetres (51–59 in)

Sample Data, Statistics, and Parameters

Statistics are functions on the data

A statistic (singular) is a function, that takes experiment data as its input.

We estimate the value of the parameter of a model, by calculating a statistic, based on data we obtained from the sample. For example:

- How long does it take to run a program? Run the program many times (sample), and calculate the mean execution time (statistic).
- How effective is a drug? We give the drug to sick patients (sample), and count how many patients get better after two days (statistic);
- Is it better to add more perceptrons to a neural network? We run the network with different sizes (sample), and calculate the correlation between size and accuracy (statistic).

Point and Interval Indicators

In this lecture, we focus on two types of statistics: **Point Estimators** and **Interval Estimators**.

- **Point Estimators**: estimate one value for a parameter from a sample;
- **Interval Estimators**: estimate a range of values for a parameter;

Statistics are Random Variables

The estimated value depends on the population and the sample

- The value of an statistic is calculated from the sample, that we obtained from the experiment.
- The sample, in turn, is a subset of the population, which is what we want to study.

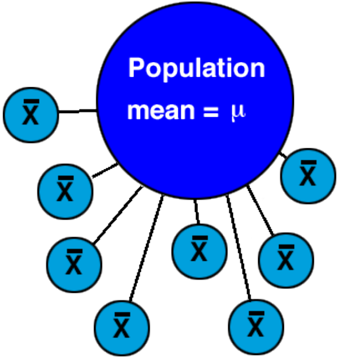
- The implication is that **every time we do a new experiment, the value calculated by an statistic will be a little bit different.**

Don't Worry!

These differences are expected and, if your experiment design is good, can be controlled. Just remember that a value you calculate from a sample is not necessarily **the truth**^{tm}

Statistics are Random Variables

What is measured by statistical tests?



A statistic can show different values, depending on the sample obtained. So it is useful to treat statistics as Random Variables.

As a Random Variable, a statistic has a **sampling distribution**. The sampling distribution is a model that describes what values we expect to obtain from a statistic, when we perform an experiment.

In general, statistical tests work by estimating the sampling distribution, and using the sampling distribution to understand the population under study. (We will study statistical tests next week).

Part II - Point Estimators

Definition of a Point Indicator

A more formal definition of Point Indicator is: *a statistic that provides the value of maximum plausibility for an (unknown) population parameter* θ.

In other words, a P.I. is a function that, given a sample obtained in an experiment, calculates the **more plausible** value for a parameter θ that describes the population of interest.

- Consider a random variable X, and a parameter θ that describes its distribution.
- Consider a sample $x = \{x_1, x_2, \ldots, x_n\}$ obtained from $X|\theta$
- The function $\hat{\Theta} = h(x)$ is a point estimator of θ, and the value $\hat{\theta}$ calulated from this function is a point estimate of θ.

Example of model parameters

We use point estimators to calculate several parameters for population models. Some common parameters that we are interested in estimating include:

- the population mean, μ;
- the population variance, σ^2;
- the population proportion, p;
- the difference in the means of two populations, $\mu_1 - \mu_2$;
- etc...

Note that for each of these parameters, there are many different statistics that we could use to calculate an estimate. Which statistic / which point estimator should be used, depends on the mathematical properties of the statistic AND the model under study.

Statistical Errors and Biases

We said before that the value calculated by a point estimator (the estimate, $\hat{\theta}$) depends on the sample. So if we are unlucky with our sample, the estimated value can be very different from the true parameter value, θ.

Talking about this difference, we want to consider **Statistical Error** and **Statistical Bias**:

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

It is important to note that besides Statistical Bias, we also have Structural Bias, which depends a lot on the design of the experiment. Today, we will focus on the first one.

Statistical Errors and Biases

Example of a biased statistic

Imagine that we want to know the typical height of the students of the CS Program.

Our experiment design is to get a random sample of 10 students, measure their height, and use a statistic to calculate the point estimate of the height, from the sample data. Let's consider two statistics:

- The point estimate of the height is **the minimum height** from the sample.
- The p. e. of the height is **the height of a random observation** from the sample;

The first statistic will give you a result that is **usually smaller** than what we could consider a representative value for the students' height.

To think about: What could we say (mathematically) about the second statistic?

Unbiased estimators

- A good estimator should consistently generate estimates that are close to the real value of the parameter θ.
- An unbiased estimator will generate estimates where the errors, when they happen, are equally distributed above and below the real value of θ.

Mathematically, we say that an estimator $\hat{\Theta}$ is said to be *unbiased* for parameter θ if its expected value is θ ($E\left[\hat{\Theta}\right] = \theta$). 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

Example: sample's average

Let's show briefly that the average of a sample of values is an unbiased estimator for the mean of the population:

Let $x_1, \ldots, x_n$ be a random sample from a given population X, and $\bar{x}$ be the average of the sample. The **value of interest** in this population is modeled by a distribution with mean μ.

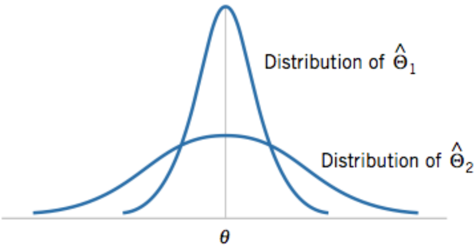
Noting that the expected value of one observation x_i is the mean μ of the population, the expected value of the average of the sample will be:

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

Unbiased estimators

If statistical bias is the same, what else is important for a statistic?

It is usually possible to define multiple unbiased estimators for a parameter θ. However, the variance of these estimators may be different.



For example, the average value of a sample and the value of a single observation are both unbiased estimators of the mean. The first one has smaller variance.

We usually want to obtain the *minimal-variance unbiased estimator* (MVUE).

This is because they will usually generate estimates $\hat{\theta}$ that are closer to the real value of θ.

Standard error of a point estimator

In the last slide we showed that a point estimator $\hat{\Theta}$ will have an associated distribution, with its mean ($\mu_{\hat{\Theta}}$) and variance ($Var\left[\hat{\Theta}\right]$).

So we can talk about the standard error of an estimator $\hat{\Theta}$ as:

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

Because the values of the estimator depend on the parameters of the population, we cannot know precisely the value of $\sigma_{\hat{\Theta}}$ or $\mu_{\hat{\Theta}}$. But we can **estimate the error of the estimator**, using data from 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 under a gaussian distribution, and a sample error s, we can calculate the standard errors of several common point indicators[1]

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

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

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

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

Point Estimator Use Case

Consider a factory that produces coaxial cables.

Let's assume that the resistance values of the cables produced by this factory are distributed following a normal distribution, with mean 50Ω and variance 4Ω.

$$X \sim \mathcal{N}\left(\mu = 50, \sigma^2 = 4\right)$$

Point Estimator Use Case

Let's imagine an experiment, where we take a random sample of 25 cables from the production process (to see if the resistance of the cables is as expected).

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 $E[\bar{x}] = \mu = 50\Omega$ and $\sigma_{\bar{x}} = \sqrt{\sigma^2/25} = 0.4\Omega$.

Note that the error of the estimator depends on the sample size. So if we collect a bigger or smaller sample, the error will decrease or increase, respectively.

The Central Limit Theorem

We love Normal distributions.

In the previous example, the resistance of cables in our imaginary factory follows a normal distribution.

So the sample average statistic for this model will also follow a normal distribution.

Normal distributions have some very nice properties (that we will explore in the next class), so it is nice when our statistics follow them.

Does this only happen when the population ("real") distribution is real?

The Central Limit Theorem

Sample-based statistics follow a normal distribution under certain conditions

The **Central Limit Theorem** states that, under certain conditions, the sampling distribution of the mean will tend to be approximately normal, even for populations with arbitrary distributions.

More formally, 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. We define the statistic z_n as:

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

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

The Central Limit Theorem

The **Central Limit Theorem** 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 techniques that assume a model with a normal distribution.

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

Part III - Interval Estimators

Statistical Intervals

How sure are you?

As we saw, point indicators are statistics (functions) that estimate values of a population parameter, from data in the sample.

Statistical intervals are functions that quantify **the uncertainty** associated to an estimate;

Let's remember the coaxial cable factory example:

- A cable factory produces cable, and we are interested in their resistance.
- The cable producion can be described as a normal distribution, with $\mu = 50, \sigma = 2$.

Let's now suppose that we obtained a sample with $n = 25$ observations, and calculated a sample mean of $\bar{x} = 48$. Because of the sample variability, it is likely that this estimate is not the true value of μ, but how much error can we assume exist in this statistic?

Statistical Intervals

Definition

A **statistical interval** defines a region that is likely to contain the true value of an estimated parameter.

The statistic interval allows us to quantify the **level of uncertainty** associated with the estimation. This information helps us arrive at sound conclusions about the parameter, at **predefined** levels of certainty.

Three of the most common types of interval are:

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

What is 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."

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.

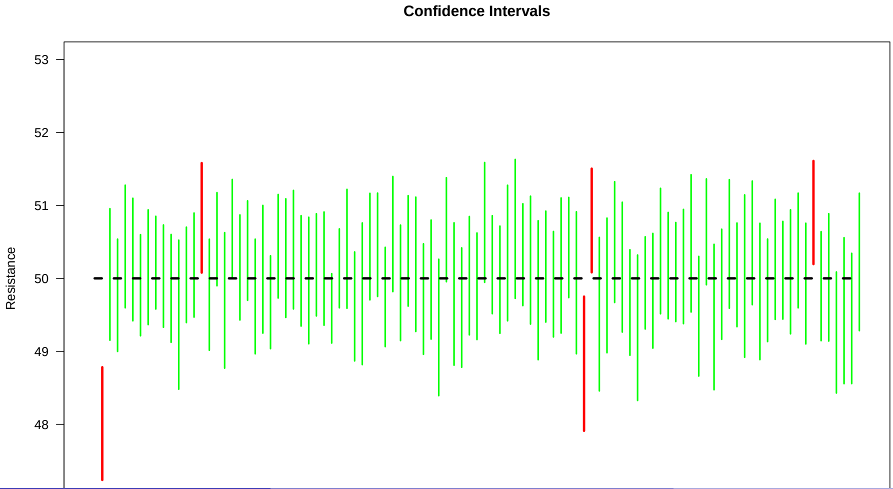
For the more usual case with an unknown variance,

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

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

Confidence Intervals

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



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.

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, 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, 2022. 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 4] Height image from `https://www.irasutoya.com`
[Page 9] Triceratops information table CC by Zachi Evenor and MathKnight
[Page 21] Image: D.C.Montgomery,G.C. Runger, *Applied Statistics and Probability for Engineers*,Wiley 2003.
[Page 24] Coaxial cable image from `https://pixabay.com`