Lectures 2: Hypothesis Testing

Question

- Empirical Methods are about identifying empirically laws that govern the (mutual) behaviour of entities
- How can we perform such empirical evaluation?
- People typically say ... "Experiment!!"
- Indeed, experiments need to be properly organized in terms of GQM
- But, as a starting point, let's see the result of experimentations...

Galileo

What about this?

Source with modifications:
https://it.wikipedia.org/wiki/Aerostato#/media/File:2006_Ojiya_balloon_festival_011.jpg

Interesting experimentations (for us)...

- Are agile methods more effective than traditional methods?
- Are open source system more secure than closed source systems?
- Is Python easier to learn than C++?
- Are formal methods effective in reducing uncertainty?

Answers

We have no answer!

- We need to shape better the empirical question that we want to address:
  - With the GQM, to define with some degree of precision what we want to determine
  - Inside the GQM with Hypothesis Testing, to formulate properly the object of study, the focus of our experiment
  - With Experimental Design, to properly consider all factors that emerged initally (and partially) while defining the GQM
  - With Statistics, to compute the result of testing the hypotheses within the defined design of the experiment.
- We start with a sweet and soft introduction to statistics

Descriptive and Inferential Statistics (1/2)

- **Descriptive statistics** enables us to present the data in a more meaningful way, which allows simpler interpretation of the data.
- **Inferential statistics** are techniques that allow us to use samples to make generalizations about the populations from which the samples were drawn. It is, therefore, important that the sample accurately represents the population.

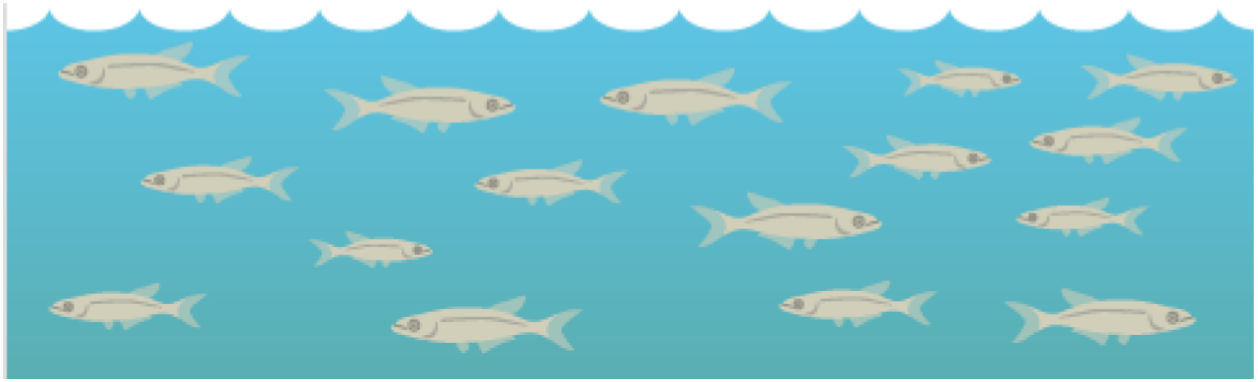
Descriptive and Inferential Statistics (2/2)

Examples of tools that we use in descriptive statistics are:

- Measures of central tendency:
  - the mean,
  - the median,
  - the mode,
  - ...
- Measures of spread:
  - the variance,
  - the range,
  - the 95-percentile,
  - ...

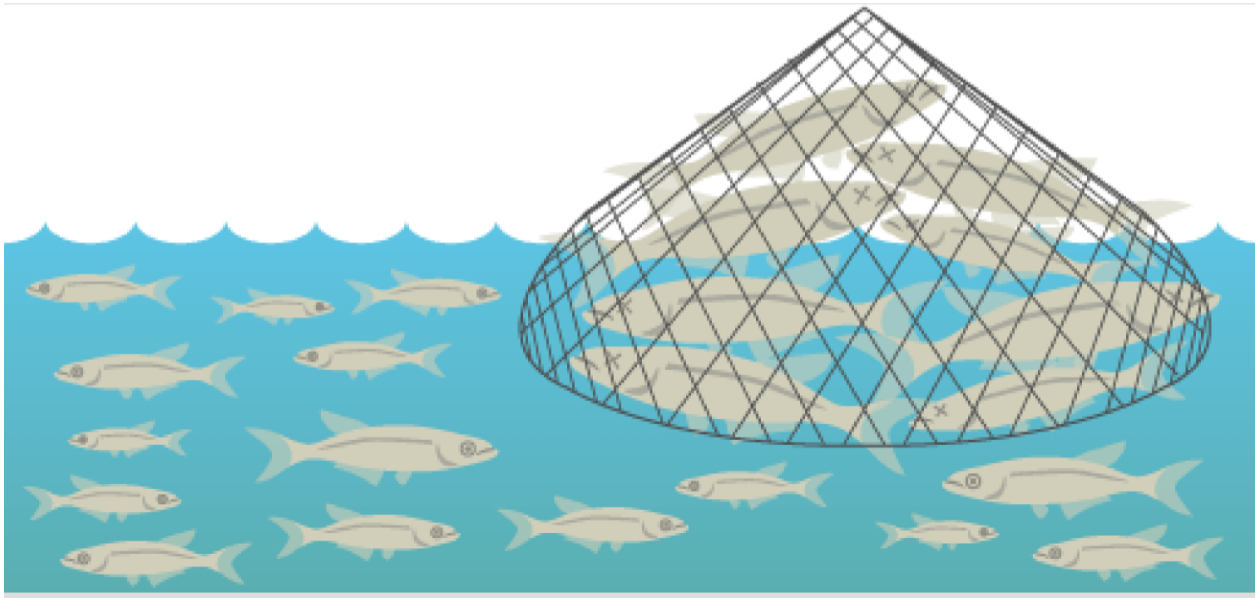
Statistical concepts – Population

Population – set of objects that are studied in a task.



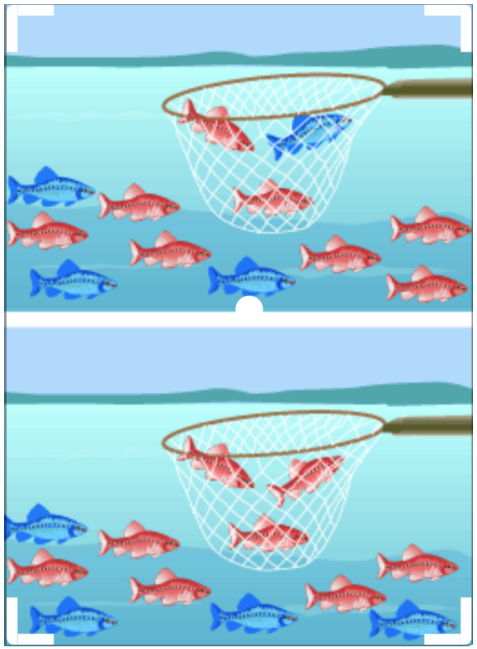
Statistical concepts – Sample

Sample – finite set of objects from the population.



Statistical concepts

Samples maybe different...

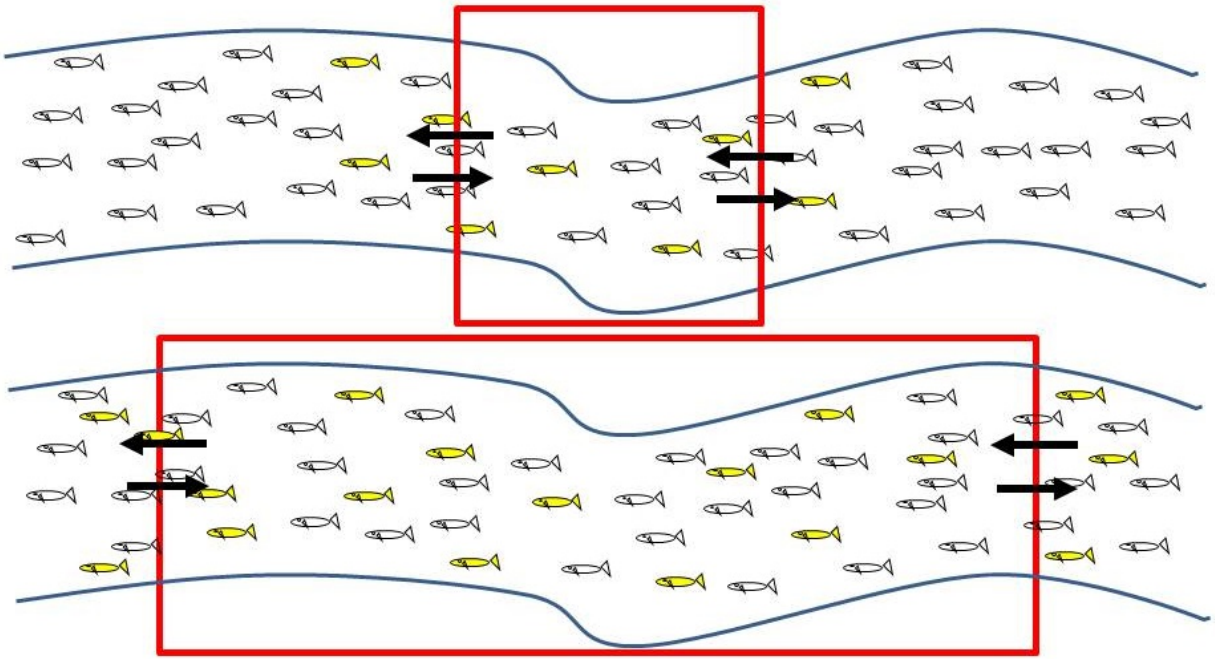


But sample should represent main properties of population (that are investigated in the research / analysis)
Reliable statistical analysis requires **representative samples**.

Statistical concepts – Sample Size

$$X^n = (X_1, \ldots, X_n).$$

n – sample size



Statistical concepts – Simple sample

X^n– **simple sample**, if $X_1, \ldots, X_n$ - independently identically distributed (i.i.d.) random variables. Each has the same density function $f(x)$.

Statistical concepts (1/4)

A statistic $T(X^n)$ – is a function of a sample.

For the central tendency,

- Let X be our sample:

$$X = \{X_1 \dots X_n\}$$

  - E.g., $X = \{1, 2, 4, 4, 4, 2, 0, 9, 0, 1\}$
- Notice that we write it as a set, but it is a bag and not a set, since there can be duplicates.

Statistical concepts (2/4)

- Sample mean:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

  - In our case it is 2.7

- Sample median:

$$median(X) = X_j : X_j \text{ is central}$$

  - In our case it is 2

Statistical concepts (3/4)

- Sample median more formally *(Is it useful? – we need to experiment it . . .)*:

$$median(X) = X_j : abs(card(\{X_k : X_k \in X, X_k \leq X_j\}) - card(\{X_h : X_h \in X, X_h \geq X_j\})) \leq 1$$

- Sample mode:

$$mode(X) = X_j : X_j \text{ is the most frequent value in } X$$

  - In our case it is 4

Statistical concepts (4/4)

- Sample mode more formally:

$$mode(X) = \arg\max_{x \in X}(occurrences(x, X))$$

where:

$$occurrences(a, S) = \text{ number of elements of a } \in S$$

Definition of statistic

A statistic $T(X^n)$ – is a function of a sample.

- We have seen the sample mean:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- We can have the sample variance:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X - \bar{X})^2$$

  - In our case it is 66.1

Back on kinds of statistics (1/2)

To remember #1: **Descriptive statistics computes** properties of the sample, the sample statistics.

To remember #2: **Inferential statistics estimates** the population parameters from the sample statistics.

To remember #3: Statistics are computed, parameters are estimated.

Back on kinds of statistics (2/2)

- Any time we move from sample to population we perform estimations,
  - i.e., we use statistics to estimate parameters of the population, and different statistics have different properties
- Estimations have properties and need to be as *precise* as possible.

Estimations

A good estimator must satisfy three conditions:

- Unbiased: The expected value of the estimator must be equal to the mean of the parameter.
- Consistent: The value of the estimator approaches the value of the parameter as the sample size increases.
- Relatively Efficient: The estimator has the smallest variance of all estimators which could be used.

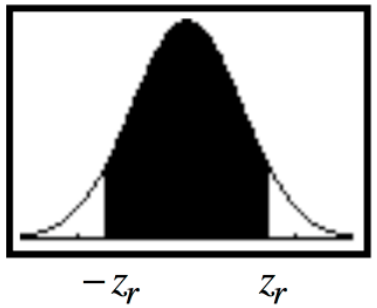
Estimates and point estimates

- We have two kinds of estimates:
  - Point estimates
  - Interval estimates
- A point estimate is the single best guess about the value of parameter.

Interval estimates

- An interval estimate is an interval that contain the true value of the corresponding parameter with the specified probability.
- It is also called the confidence interval.
- *We will see later that if the distribution of a random variable Z is Normal (N(0,1)), then:*

$$\mathbb{P}(-z_r \leq Z \leq z_r) = r$$



Where:

- z_r - z-score
- r - level of confidence (usually, denoted as $1 - \alpha$)

Unbiased estimator

Any statistic whose mathematical expectation is equal to a parameter θ is called an **unbiased estimator** of the parameter θ. Otherwise, the statistic is said to be **biased**.

Bias of an estimator

The **bias** of an estimator $\hat{\theta_n}$ is

$$\mathbb{E}(\hat{\theta}) - \theta$$

Bias-variance decomposition

- The mean squared error MSE is:

$$MSE = \mathbb{E}(\hat{\theta} - \theta)^2$$

.

- The **bias-variance decomposition** for the MSE of an estimator $\hat{\theta_n}$

$$MSE = bias^2(\hat{\theta_n}) + \mathrm{Var}(\hat{\theta_n})$$

.

- We will prove this formula later in the course.

Hypothesis testing

- Now we introduce the following concepts:
  - (a) A statistical hypothesis.
  - (b) A test of a hypothesis against an alternative hypothesis and the associated concept of the critical region of the test.
  - (c) The power of a test.

Example

- Let X be an outcome of a random experiment, for example, some test score, $X \sim N(\theta, 100)$.
- Some past experience says that $\theta = 75$.
- After some changes we suspect that no longer $\theta = 75$, but $\theta > 75$
  - $\theta > 75$ is our **statistical hypothesis**
- However, our hypothesis might be false, hence $\theta \leq 75$ (this also may be considered as a hypothesis).
  - $\theta \leq 75$ may be considered as a hypothesis

Example

- Thus, we have two hypotheses:
  - An first hypothesis of no effect, which we represent as H_0, and which we call null hypothesis
  - Another hypothesis of a specific effect, which we represent as H_1, and which we call alternate hypothesis
- In our case we can say:
  - Null hypothesis:

$$H_0 : \theta \leq 75$$

  - Alternate hypothesis:

$$H_1 : \theta > 75$$

Example

- To check H_0 against H_1 we run the experiment
- We consider a random sample $X_1, \ldots, X_n$ from $N(\theta, 100)$
- We devise a test, i.e., a rule, that will tell us what decision to make once we get observations $x_1, \ldots, x_n$.
- For the sake of the example, we assume $n = 25$.

How do we construct the test – the rule?

- This will be the subject of the last lectures of the course
  - and the beginning of the twin course on advanced statistics
- Now we just provide an intuitive explanation

Procedure (1/2)

- We want to determine if the alternate hypothesis H_1 with a *negative* approach
- We check if we can reject the null hypothesis H_1 with a high level of probability, say, 95%
- We divide the sample in two parts, one of which can also be empty. Formally speaking we say that:
- We partition the sample space into a subset C and its complement C^*.
- Let us assume that the experiments lead to outcomes $X_1, \ldots, X_n$, which are reflected to the values $x_1, \ldots, x_n$

Procedure (2/2)

- If the point $(x_1, \ldots, x_n) \in C$, we reject the hypothesis H_0, otherwise we cannot reject H_0.
- If we reject the null hypothesis, we claim that the alternate hypothesis hold.
- If we cannot reject the null hypothesis, right now we cannot draw any conclusion.

C is the critical region of the test.

Example. Test 1 (1/2)

There are a lot of rules (tests). We start with **Test 1**

$$C = \{(x_1, \ldots, x_n); (x_1 + x_2 + \ldots + x_n > 25 * 75)\}$$

We shall reject H_0 if and only if $(x_1, \ldots, x_n) \in C$.

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

We accept $H_0 : \theta \leq 75$ if $\bar{x} \leq 75$.

Note: This will be all proved formally during the last lectures of of the course.

Example. Test 1 (1/2)

The rule: We shall **reject** the $H_0 : \theta \leq 75$ if the mean of the sample **exceeds the maximum value of the mean of the distribution** when the hypothesis H_0 is true.

$$\bar{X} \sim N(\theta, 100/n) = N(\theta, 100/25) = N(\theta, 4)$$

Note: This will be all proved formally during the last lectures of of the course.

Hypothesis testing. Definitions

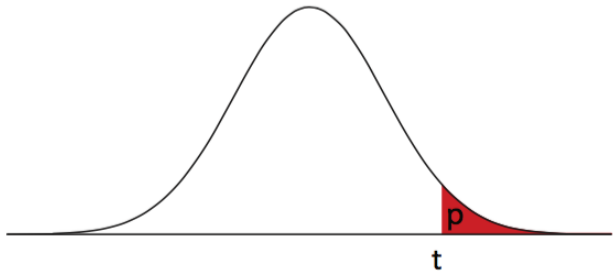
Definition 1. A **statistical hypothesis** is an assertion about the distribution of one or more random variables. If the statistical hypothesis completely specifies the distribution, it is called a simple statistical hypothesis; if it does not, it is called a composite statistical hypothesis.

Hypothesis testing. Definitions

Definition 2. A test of a statistical hypothesis is a rule which, when the experimental sample values have been obtained, leads to a decision to accept or to reject the hypothesis under consideration.

Hypothesis testing. Definitions

Definition 3. Let C be that subset of the sample space which, in accordance with a prescribed test, leads to the rejection of the hypothesis under consideration. Then C is called **the critical region** of the test.



Hypothesis testing. Definitions

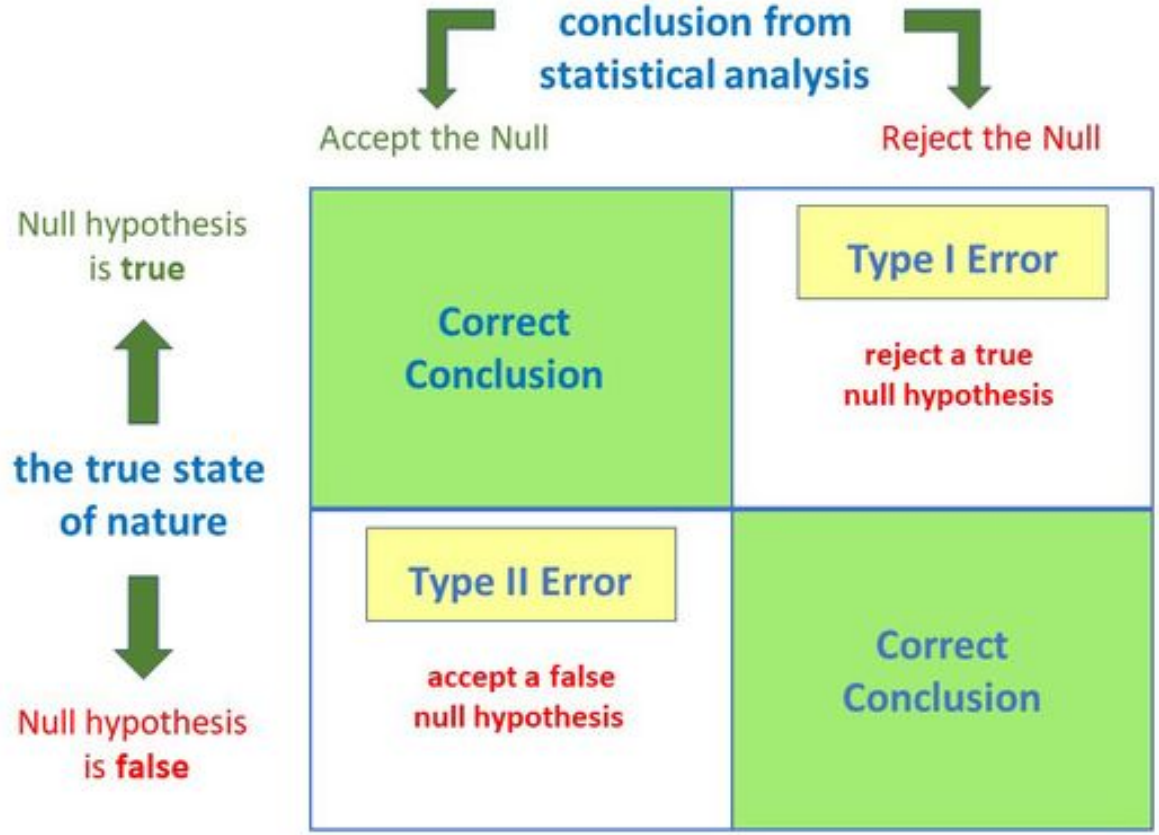
Definition 4. The p-value associated with a test (again) is the probability that we obtain a value of the test statistic that is at least as extreme as the observed value of our test statistic assuming the H_0 hypothesis is true.

Hypothesis testing. Summary

The following 5 steps are followed when testing hypotheses:

- Specify H_0 and H_1 – the null and alternative hypotheses
- Determine the appropriate test statistic
- Determine the critical region
- Compute the value of the test statistic
- Make decision

Errors of a test (1/2)



Taken with modifications from: `https://www.simplypsychology.org/type_I_and_type_II_errors.html`

Errors of a test (2/2)

There could be errors of two kind:

- A type 1 error is when we reject the null hypothesis when the null hypothesis is true, that is we think that something is going on, but nothing is really there.
  - The probability of committing a type 1 error is typically referred to as α.
- A type 2 error is when we fail to reject the null hypothesis when actually we should reject it, that is, we fail to perceive a phenomena.
  - The probability of committing a type 2 error is typically referred to as β.

The power function of a test informally is the probability of not committing a type 2 error, that is, $(1 - \beta)$

Back to the example (1/9)

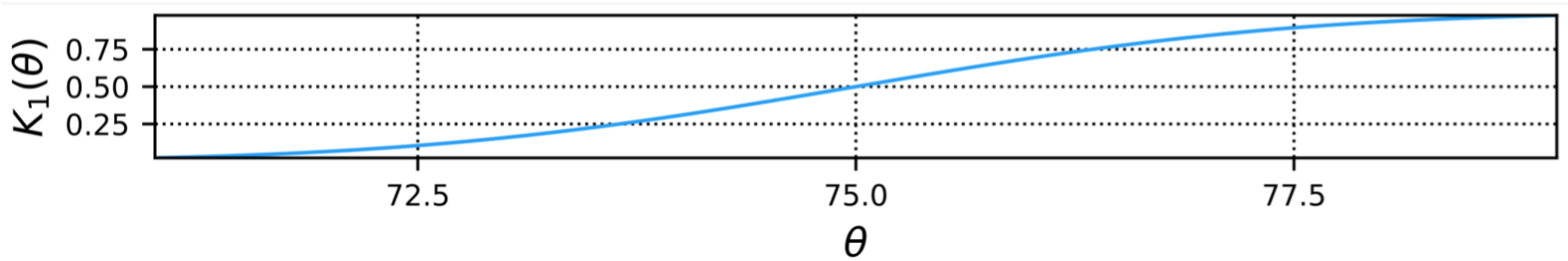
Is the Test 1 good?

We are now going to see only a trailer.

- The full movie will be played throughout the course
- But perceive the big picture now before progressing...

Back to the example (2/9)

Let us construct **the power function** of the Test 1:
$K_1(\theta) = \mathbb{P}(\bar{X} > 75)$; $K_1(75) = 0.5$. See the standard normal distribution table.



Is it good to reject H_0, when H_0 is true (in 50% of cases)?

Remember, this is just a trailer, you are not required to fully learn it right now ...

Back to the example (3/9)

Let us try another one, Test 2:

$$C = \{(x_1, \ldots, x_n) : (x_1 + x_2 + \ldots + x_n > 25 * 78)\}$$

We shall reject H_0 if and only if $(x_1, \ldots, x_n) \in C$.

Remember, this is just a trailer, you are not required to fully learn it right now ...

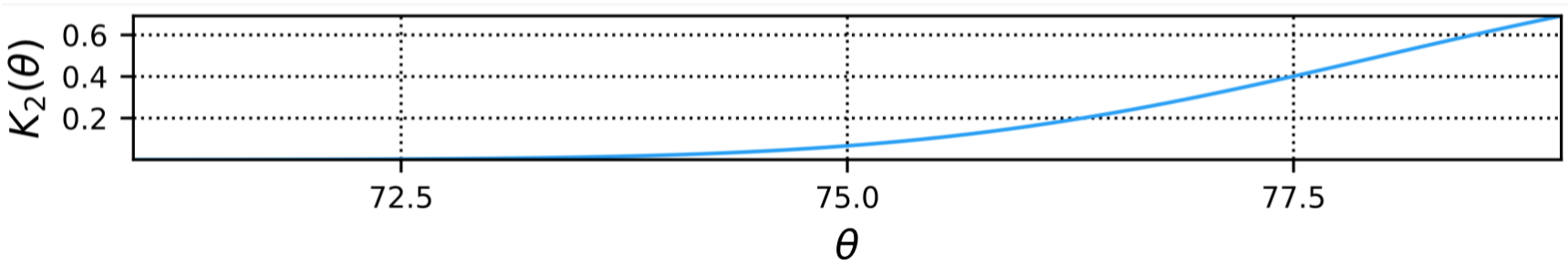
Back to the example (4/9)

The power function of the Test 2:

$$K_2(\theta) = \mathbb{P}(\bar{X} > 78) = 1 - N\left(\frac{78-\theta}{2}\right)$$

Remember, this is just a trailer, you are not required to fully learn it right now ...

Back to the example (5/9)



$$K_2(75) = 0.067; but K_2(77) = 0.309$$

If true $\theta = 77$ we accept $H_1 : \theta = 75$ only with probability 0.309.

Remember, this is just a trailer, you are not required to fully learn it right now . . .

Back to the example (6/9)

Finally, we try to generalize with Test 3:

$$C = \{(x_1, \ldots, x_n) : (x_1 + x_2 + \ldots + x_n > n * c)\}$$

where c is some constant, n sample size, again $\bar{X} \sim N(\theta, 100/n)$.

Remember, this is just a trailer, you are not required to fully learn it right now ...

Back to the example (7/9)

But now we want the power function of the Test 3 to meet conditions:

$$K_3(\theta) = \mathbb{P}(\bar{X} > c) = 1 - N\left(\frac{c-\theta}{10/\sqrt{n}}\right)$$

$$K_3(75) = 0.159; K_3(77) = 0.841;$$

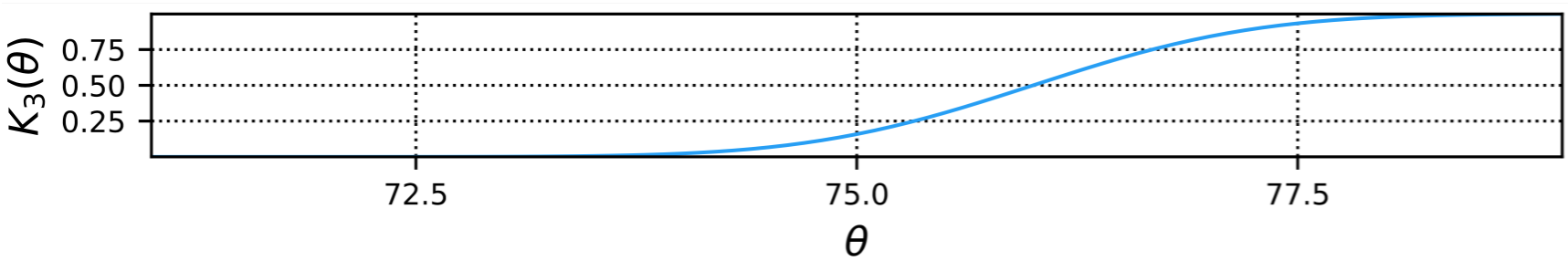
Remember, this is just a trailer, you are not required to fully learn it right now ...

Back to the example (8/9)

$$N\left(\frac{c-75}{10/\sqrt{n}}\right) = 0.159; N\left(\frac{c-77}{10/\sqrt{n}}\right) = 0.841;$$

Solution: $n = 100; c = 76.$

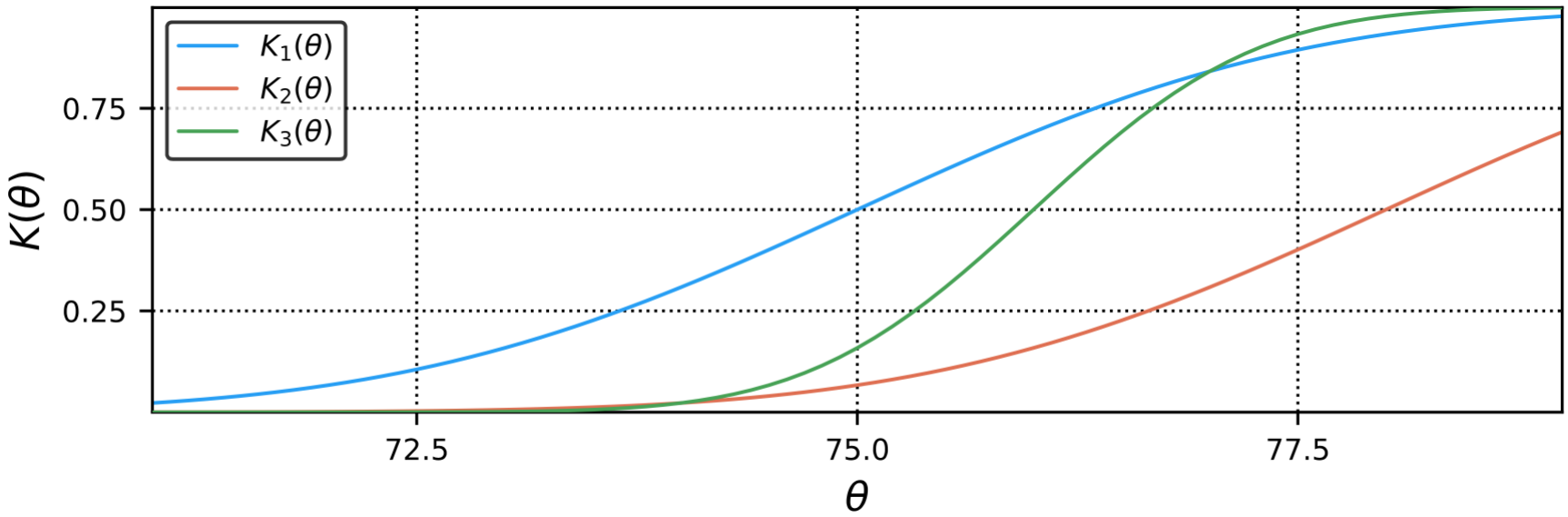
Note, $K_3(73) = 0.001; and K_3(79) = 0.999.$



Remember, this is just a trailer, you are not required to fully learn it right now ...

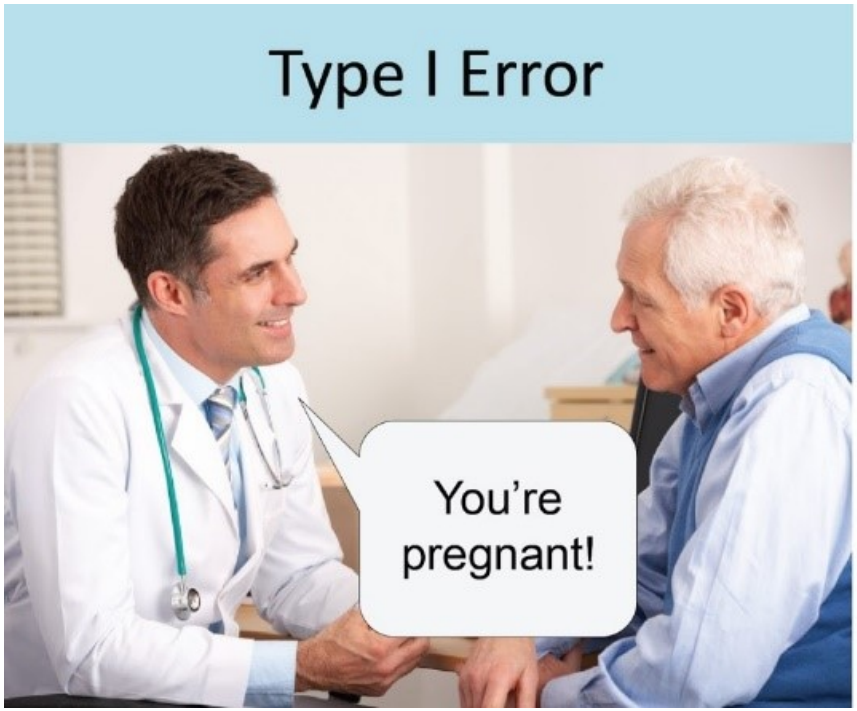
Back to the example (9/9)

Comparison of tests (K_1, K_2, and K_3):



Remember, this is just a trailer, you are not required to fully learn it right now ...

Hypothesis testing. Type I and II errors



Taken with modifications from:
https://www.statisticssolutions.com/wp-content/uploads/2017/12/rachnovblog.jpg

The following concepts are all equivalent:

- "significance level"
- "size of the critical region"
- "power of the test when H_0 is true,"
- "the probability of committing an error of type I"