Experiment Design for Computer Sciences (0AL0400)

Review 02 - Course Review

Claus Aranha
caranha@cs.tsukuba.ac.jp

University of Tsukuba, Department of Computer Sciences

Version 2022.1 (Updated June 23, 2022)

Outline

This material summarizes the **key points** learned through the course.

The purpose of this material is to help you direct your review study. However, make sure to **review the original materials** and the **recommended reading** for detailed information on each topic.

Final examination rules and important points

The 2022 final examination will be held **online on manaba**, from 15:15 to 18:00 JST.

- During the exam, it is not necessary to log into the TEAMs channel;
- The teacher will not answer questions about the course subject during the exam;
- Reference materials during the exam;
  - You may prepare **One A4 page (front and back) of handwritten notes** and use it during the exam. **You must submit the notes at the end of the exam**
  - You may consult English dictionaries during the exam;
  - You are Not Allowed to consult any other materials during the exam, including the lecture notes;
- If you have difficulties accessing manaba during the exam, or any other problems, contact the teacher by e-mail or teams

Trial examination

Before the final examination, a trial examination will be published on manaba.
Please use the trial examination to test the manaba system.

- The trial examination will not be graded.
- The trial examination has a time limit of 1 hour.
  (the real exam has a time limit of 3 hours)
- It is not necessary to submit the trial examination.

Topic 01 – What is an experiment?

- The scientific method is a complex system. It includes not only experimentats and theories, but also community, communication, motivation and interaction with the society;
- Experiments are used to obtain data from the world in a methodical fashion;
- **Experiment Design** is the discipline of planning how to conduct an experiment to answer a scientific question;
- When designing an experiment, we must:
  - Choose which data is obtained from the experiment (return variable);
  - Choose the conditions to execute the experiment (controlled factors and noise factors);
  - Choose the objectives of the experimental analysis (experimental parameters)
  - Choose how to obtain the data of the experiment (number of observations, blocks)
  - Choose how to analyze the data of the experiment (statistical model, visualization)
- All these design choices must be defined before the experiment begin.

Topic 01 – Characteristics of a Good Experiment

A good experiment...

- ... examines a falsifiable hypothesis. (the hypothesis has clear criteria of what result would support or reject it)
- ... is reproducible. (all information necessary to repeat the experiment is available)
- ... controls the experimental environment to minimize the effects of factors unrelated to the scientific question (fairness of comparison; independence of conditions; etc)

Pre-registration of experiments can be used to reduce the effect of human bias: The experiment design fully is published before the experiment is executed.
Open Data and Reproducible Experiments are important to allow other scientists to double check and improve your work.

Topic 02 – Statistical Indicators

- A model is a mathematical description of something that we want to study;
- We use information from an experiment to specify aspects of a model;
  - The **population** of an experiment is the set of all possible results of that experiment;
  - An **observation** is a single data point from an experiment;
  - A **sample** is a set of observations;
- A **Statistical Indicator** is a function that uses data obtained from a model to calculate some of its parameters (characteristics).
  - For example: using data about the running time of a program, we use the mean as an estimate of the typical running time of that program.
- Point Estimators calculate specific values for a parameter, while Interval Estimators calculate a range of likely values.
- The value calculated by an Estimator may not be the real value of the parameter;
  - **Error** Difference between the value of the estimator and the value of the parameter;
  - **Bias** Systematic error caused by an Estimator;

Example of Statistical Indicators

- Mean
- Median
- Variance
- Correlation
- Confidence Interval
  - Confidence Interval is an Interval Estimator;
  - It calculates an interval that may contain the true value of the parameter with X probability;

Using Interval Estimators, such as the confidence interval, gives us more information about the model than just using point estimators such as the mean.

Topic 03 – Hypothesis Testing

- **Statistical Inference** procedures use data from an experiment to establish the probability of an statement being true.
- **Null Hypothesis Significance Testing** uses statistical inference to compare different hypothesis about the model under study.
  - Null Hypothesis: Standard assumptions about the model. No clear effect.
  - Alternate Hypothesis: Unexpected effects. Anomalies.
- The Hypothesis test procedure consists of collecting data through experiment, and then using that data to compare the probabilities of the null and alternate hypothesis.
- Possible outcomes:
  - Null Hypothesis cannot be rejected;
    - Data evidence towards the alternate hypothesis is not strong enough to reject NH;
    - Type II error: Null Hypothesis is actually false;
  - Reject the Null hypothesis;
    - Alternate hypothesis is more likely than NH by a large margin, given the data;
    - Type I error: Null Hypothesis is actually true;

Topic 03 – Hypothesis Parameters

- α: Desired probability of a type I error.
  - Test confidence level: 1-α
  - Used as probability threshold for rejecting the null hypothesis;
- β: Desired probability of a type II error
  - Test power: 1-β
  - Actual probability of a Type II error is controlled also by unknown factors;
- δ^*: Minimal Interesting effect size;
  - Minimal difference in the experiment result that has practical interest,
  - **regardless of the hypothesis test result**
- *n*: Sample size – Number of observations *for each experimental condition*

Topic 03 – Hypothesis Tests

- **Z test**: Compares the indicator against a fixed value. Calculates the probability of the sample when the Null hypothesis is true. Assumes the sample residuals come from a Normal distribution with known variance;
- **T test**: Compares the indicator against a fixed value. Calculates the probability of the sample when the Null hypothesis is true. Assumes the sample residuals come from a t distribution with $n - 1$ degrees of freedom. Estimates the variance from the sample error.
- **p-value**: Maximum value of α (lowerst significance level) that would reject the null hypothesis of a test.

Topic 04 – Two Sample Testing

- **t-test for two samples:** Perform Statistical Inference on the difference between the two samples;
- Assumptions:
  - Normality of residuals;
  - Equality of variances;
  - **Independence of Observations**
- To reduce the effect of different variance, sample sizes proportional to the variance can be used;
- The independence assumption must be guaranteed at the experiment design stage;

Topic 04 – Paired Testing

- The Assumption of Independence states that the experimental conditions of all observations are exactly the same;
- On the other hand, **Noise Factors** can influence the results of an experiment in a systematic manner;
- **Pairing** is a technique to reduce the effect of a noise source that affects **pairs of observations** in an experiment equally.
- The statistical model for paired test is very similar as the model for two sample testing, except on how the difference between samples is calculated;

Topic 05 – Non-normal data

The t-test assumes that the residuals follow a normal distribution.

- Large outliers;
- Extreme non-normal distributions;
- Multi-modal data;
- Non-continuous data;

When this assumption does not hold, some sort of treatment is necessary.

- Removing outliers;
- Log transformation, square root transformation;
- Bootstrap transformation;
- Non-parametric tests
  (Wilcoxon Rank Sum, Mann-Whitney U-test, Kruskal Wallis, etc)

Topic 05 – Multiple sample testing (ANOVA)

A comparison of multiple samples (for example, multiple algorithms) can be modeled as an experiment with one discrete control factor and multiple levels;

Testing is done in two stages:

- ANOVA – Tests all samples at the same time, indicates if at least one level has a significant effect;
- post-hoc testing – series of pairwise tests between the levels to identify which level has the significant effect;

The strategy for post-hoc testing (one-vs-all, all-vs-all) must be defined during the experiment design stage. The alpha value of the post-hoc tests must be adjusted to take multiple comparisons into account;

Topic 06 – Sample Size Calculation

The amount of data that we must collect from an experiment is an important decision for the experiment design stage:

- Sample size has a large influence on the power and confidence of an experiment;
- On the other hand, larger samples incur in large experimental costs;

Power calculations for statistical tests can be used to estimate the desired sample size of an experiment;

- Fix the sample size and estimate the power of an experiment;
- Fix the power and estimate the sample size;

Power calculations usually require an estimate for the variance of the model. This value can be obtained from domain knowledge or a preliminary experiment;

The power calculation for the ANOVA also depends on the strategy for ad-hoc testing;

Topic 07 – Factorial Analysis

Variables of an experiment:

- Controlled Factors; Factor Levels;
- Noise Factors / Nuisance Factors; (see Blocking and Pairing)
- Random noise / Background Noise / Residual Noise;
- Response Variable;

When an experiment has more than one controlled factor, we are interested in possible **interaction effects** between those factors.

If we don't expect strong interaction effects, we can use "One Variable At a Time" (OVAT) design. OVAT analyses each controlled factor in a different experiment.

Factorial Design treats every combination of factors as a different level of a "unified" factor. In this model ANOVA can be used to identify interaction effects.

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