论文写作规范

七、研究中的实验及数据处理

刘帅

2022 年 10 月 25 日
软件学院 206

Outline

1. 14 Experimentation
   - Baselines
   - Persuasive Data
   - Interpretation
2. 15 Statistical Principles
   - Variables
   - Reporting variability
   - Statistical Tools
   - A "Statistical Principles" Checklist

Experiments are an essential part of sound science.
Tests should be **fair**.

Baselines

- Your results should be compared to **the best previous method**.
- Update the choice of baseline to the latest.
- What if you solve an existing problem in a novel way that is for some reason not comparable to previous work? Potential point of comparison: the first workable option that a reasonable person might suggest.

Persuasive Data

Put yourself in the position of the reader.
You want to be persuaded that an algorithm is the very best option available, for a certain class of problem.

- –[what data] their characteristics

Persuasive Data

Put yourself in the position of the reader.
You want to be persuaded that an algorithm is the very best option available, for a certain class of problem.

- –[what data] their characteristics
- –[mechanisms] need to be normalized or cleaned up

Persuasive Data

Put yourself in the position of the reader.
You want to be persuaded that an algorithm is the very best option available, for a certain class of problem.

- –[what data] their characteristics
- –[mechanisms] need to be normalized or cleaned up
- –[Sufficiency] larger data sets offer different statistical properties. data volumes need to be large enough to ensure that the experiment will be able to detect the effect that is being hypothesised.
  the number of data sets: A single data set may not be persuasive

Persuasive Data

- Domain knowledge: the problem has been abstracted or simplified in some way that makes it unrealistic

Persuasive Data

- Domain knowledge: the problem has been abstracted or simplified in some way that makes it unrealistic
  - such as in biological or medical areas.

Persuasive Data

- Domain knowledge: the problem has been abstracted or simplified in some way that makes it unrealistic
  - such as in biological or medical areas.
  - Unvalidated results

Persuasive Data

- Domain knowledge: the problem has been abstracted or simplified in some way that makes it unrealistic
  - such as in biological or medical areas.
  - Unvalidated results
- Limits: Knowledge in the data is needed to help assess the significance of the results, and to then, as far as possible, rectify 矫正 the data. This may involve, for example, careful manual data processing, following explicit guide-lines.

- If parameters have been derived by **tuning**, the only way to establish their validity is to see if they give good behaviour on other data. Choosing parameters to suit data, or choosing data to suit parameters, in all likelihood invalidates the research.

- If parameters have been derived by **tuning**, the only way to establish their validity is to see if they give good behaviour on other data. Choosing parameters to suit data, or choosing data to suit parameters, in all likelihood invalidates the research.
- **reference data sets**. Allow the direct comparison of work between institutions and between papers. It carries risks, that is, methods can become so specialized that they do not work on other data.

- If parameters have been derived by **tuning**, the only way to establish their validity is to see if they give good behaviour on other data. Choosing parameters to suit data, or choosing data to suit parameters, in all likelihood invalidates the research.
- **reference data sets**. Allow the direct comparison of work between institutions and between papers. It carries risks, that is, methods can become so specialized that they do not work on other data.
- Verify that what you are testing is what you intended to test, and an experiment should only succeed if the hypothesis is correct. Ask whether a single data set is sufficient.
  To what volumes of data should your claims apply?

Interpretation

- Consider whether there are other possible interpretations of the results; and, if so, design further tests to eliminate these possibilities.

Interpretation

- Consider whether there are other possible interpretations of the results; and, if so, design further tests to eliminate these possibilities.
- Conclusions should be sufficiently supported by the results. Success in a special case does not prove success in general, so be aware of factors in the test that may make it special.

Interpretation

- Consider whether there are other possible interpretations of the results; and, if so, design further tests to eliminate these possibilities.
- Conclusions should be sufficiently supported by the results. Success in a special case does not prove success in general, so be aware of factors in the test that may make it special.
- Don't draw undue 过度 conclusions or inferences 推论. If, say, one method is faster than another on a large data set, and they are of the same speed on a medium data set, that does not imply that the second is faster on a small data set; it only implies that different costs dominate at different scales.

- don’ t overstate your conclusions. For example, if a new algorithm is somewhat worse than an existing one, it is wrong to describe them as equivalent. A reader might infer that they are equivalent if the difference is small, but it is not honest for you to make that claim.

- don’ t overstate your conclusions. For example, if a new algorithm is somewhat worse than an existing one, it is wrong to describe them as equivalent. A reader might infer that they are equivalent if the difference is small, but it is not honest for you to make that claim.
- Numerical measures allow numerical manipulation, but such manipulation does not always make sense if applied to the qualitative goal we wish to achieve. cannot directly interpret the degree of difference between scores.

- don’ t overstate your conclusions. For example, if a new algorithm is somewhat worse than an existing one, it is wrong to describe them as equivalent. A reader might infer that they are equivalent if the difference is small, but it is not honest for you to make that claim.
- Numerical measures allow numerical manipulation, but such manipulation does not always make sense if applied to the qualitative goal we wish to achieve. cannot directly interpret the degree of difference between scores.
- Predictivity： the conclusions in our papers are usually about properties of systems, not their behaviour on the data we have already seen.

An "Experimentation" Checklist

design of the experiments

- Have appropriate baselines been identified? What makes them appropriate? Are they state-of-the-art?

An "Experimentation" Checklist

design of the experiments

- Have appropriate baselines been identified? What makes them appropriate? Are they state-of-the-art?
- What data has to be gathered, and where from?

An "Experimentation" Checklist

design of the experiments

- Have appropriate baselines been identified? What makes them appropriate? Are they state-of-the-art?
- What data has to be gathered, and where from?
- How will readers gather comparable data for themselves?

An "Experimentation" Checklist

design of the experiments

- Have appropriate baselines been identified? What makes them appropriate? Are they state-of-the-art?
- What data has to be gathered, and where from?
- How will readers gather comparable data for themselves?
- Is the data real? Is it sufficient in volume? What validation is required for artificial data?

An "Experimentation" Checklist

design of the experiments

- Have appropriate baselines been identified? What makes them appropriate? Are they state-of-the-art?
- What data has to be gathered, and where from?
- How will readers gather comparable data for themselves?
- Is the data real? Is it sufficient in volume? What validation is required for artificial data?
- Should the data be seeded with examples to test the validity of the outcomes?

- Is there reference data for the problem, and what are its limitations?

- Is there reference data for the problem, and what are its limitations?
- Will a domain expert be needed to interpret the results?

- Is there reference data for the problem, and what are its limitations?
- Will a domain expert be needed to interpret the results?
- What are the likely limitations on the results?

- Is there reference data for the problem, and what are its limitations?
- Will a domain expert be needed to interpret the results?
- What are the likely limitations on the results?
- Should the experimental results correspond to predictions made by a model?

- Is there reference data for the problem, and what are its limitations?
- Will a domain expert be needed to interpret the results?
- What are the likely limitations on the results?
- Should the experimental results correspond to predictions made by a model?
- Will the reported results be comprehensive or a selection? Will the selection be representative?

- What enduring properties might be observed by other people attempting to validate the work with different hardware, data, and implementation?

- What enduring properties might be observed by other people attempting to validate the work with different hardware, data, and implementation?
- Are the experiments feasible? Do you have the resources (time, machines, data, code, humans) required to undertake them to a reasonable standard?

An "Experimentation" Checklist

software to be developed

- Can baselines be obtained from elsewhere, or do they need to be implemented? Will they be of similar standard to the implementation of your system?

An "Experimentation" Checklist

software to be developed

- Can baselines be obtained from elsewhere, or do they need to be implemented? Will they be of similar standard to the implementation of your system?
- How much coding is required? What existing resources can be used? That is, is your coding effort being used effectively?

An "Experimentation" Checklist

software to be developed

- Can baselines be obtained from elsewhere, or do they need to be implemented? Will they be of similar standard to the implementation of your system?
- How much coding is required? What existing resources can be used? That is, is your coding effort being used effectively?
- Can the code (or the proposed contribution) be decomposed into components? How will the individual components be tested for correctness, and evaluated for significance?

An "Experimentation" Checklist

software to be developed

- Can baselines be obtained from elsewhere, or do they need to be implemented? Will they be of similar standard to the implementation of your system?
- How much coding is required? What existing resources can be used? That is, is your coding effort being used effectively?
- Can the code (or the proposed contribution) be decomposed into components? How will the individual components be tested for correctness, and evaluated for significance?
- How will you know that the code is correct?

An "Experimentation" Checklist

software to be developed

- Can baselines be obtained from elsewhere, or do they need to be implemented? Will they be of similar standard to the implementation of your system?
- How much coding is required? What existing resources can be used? That is, is your coding effort being used effectively?
- Can the code (or the proposed contribution) be decomposed into components? How will the individual components be tested for correctness, and evaluated for significance?
- How will you know that the code is correct?
- Is the code going to be made publicly available? If not, why not?

An "Experimentation" Checklist

the use of human subjects

- In what ways might humans be needed? To annotate data? As test subjects? Can your question be meaningfully answered without human subjects?

An "Experimentation" Checklist

the use of human subjects

- In what ways might humans be needed? To annotate data? As test subjects? Can your question be meaningfully answered without human subjects?
- Who will the humans be and how will they be selected?

An "Experimentation" Checklist

the use of human subjects

- In what ways might humans be needed? To annotate data? As test subjects? Can your question be meaningfully answered without human subjects?
- Who will the humans be and how will they be selected?
- How many humans will be required, and how does that correspond to your budget? What compromises will be introduced if fewer humans are used, or for less time?

An "Experimentation" Checklist

the use of human subjects

- In what ways might humans be needed? To annotate data? As test subjects? Can your question be meaningfully answered without human subjects?
- Who will the humans be and how will they be selected?
- How many humans will be required, and how does that correspond to your budget? What compromises will be introduced if fewer humans are used, or for less time?
- How will objectivity and independence be maintained? What steps need to be taken to avoid introduction of bias?

An "Experimentation" Checklist

the use of human subjects

- In what ways might humans be needed? To annotate data? As test subjects? Can your question be meaningfully answered without human subjects?
- Who will the humans be and how will they be selected?
- How many humans will be required, and how does that correspond to your budget? What compromises will be introduced if fewer humans are used, or for less time?
- How will objectivity and independence be maintained? What steps need to be taken to avoid introduction of bias?
- Is ethics clearance required? 伦理审查：1. 参与者为自愿参与并且可以随时退出. 2. 已经给参与者提供了关于项目的相关信息. 3. 数据进行匿名及保密. 4. 对参与者权益的保护

- The ideal experiment examines the effect of one variable on the behaviour of an object being studied

- The ideal experiment examines the effect of one variable on the behaviour of an object being studied
- elimination of variables

- The ideal experiment examines the effect of one variable on the behaviour of an object being studied
- elimination of variables
- have a clear understanding of the relevant parameters

Samples and Populations.

- "algorithm new is typically faster than algorithm old ": to claim that new is faster on average.
- (Such a statement could as easily be made on the basis of a theoretical analysis as on the basis of experiments.)
- But an average of what? If the intended meaning is only that new is faster than old on average for **the runs undertaken in the experiments**, what is it about **these runs** that makes them representative or predictive?

- population: the set of all possible runs
- But in all likelihood the population is infinite, as it must contain all possible combinations of input data
- It is necessary to resort to taking a sample

The task may be qualitatively changed.

For instance, a search engine will achieve the same effectiveness on a given query and collection of Web pages each time it is run, but will achieve different scores for different queries or different collections—which may be due to the new data changing the nature of the task. In cases where the tasks change qualitatively, it may not be meaningful to talk about a population, sample, or average, so other strategies for interpretation of results may be appropriate.

Reporting variability

- Averaging provides valuable insight into typical behaviour.
- A descriptive statistic is **standard deviation**
- You can report quantiles, such as the 25th to 75th percentiles, combined with the median.
- An **odd number** of experimental runs is preferable to an even number, since the middle run will be the median

Statistical Tools

- Measures of correlation are used to determine whether two variables depend on each other.
- Regression is used to identify the relationship between two variables.
- hypothesis testing 统计假设检验 Hypothesis tests are used to investigate whether improvements are significant. It is often the case that, in a series of comparisons of two techniques for the same task, one is better than the other some but not all of the time.
- there are packages that do much of the hard work. Second, many statistical problems can be couched in terms of elementary probability and then resolved computationally.

Visualization of Results

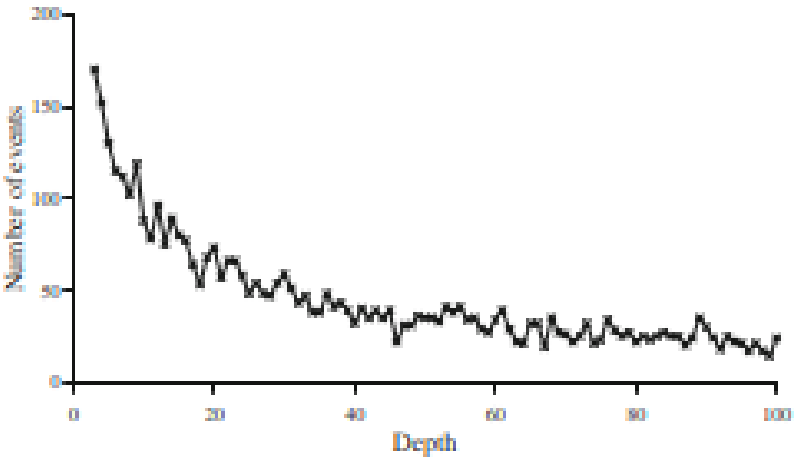
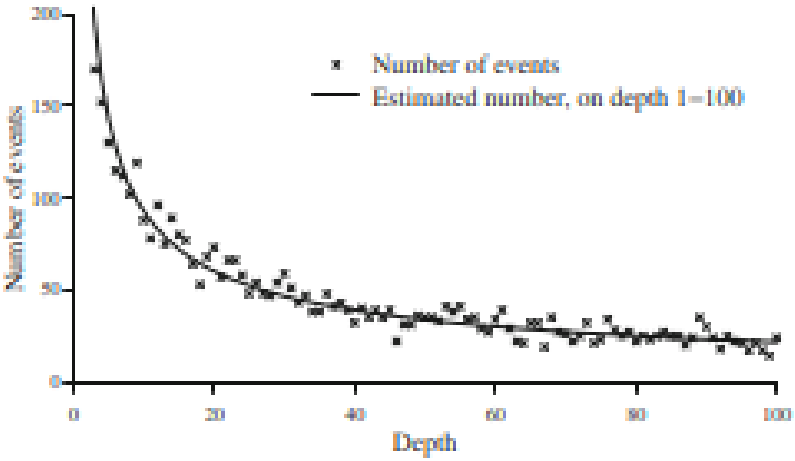


FIGURE 7. *The number of events observed at each depth; depths 1 and 2 have been omitted for reasons of scale.*



The solid line shows a best-fit to the points.

Visualization of Results

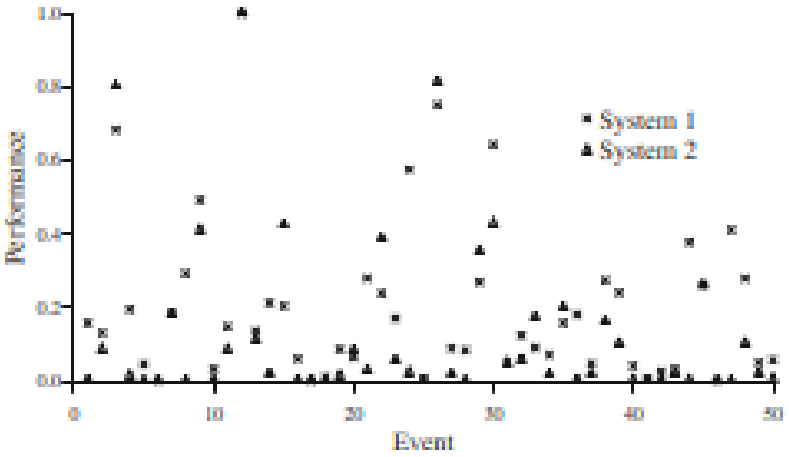
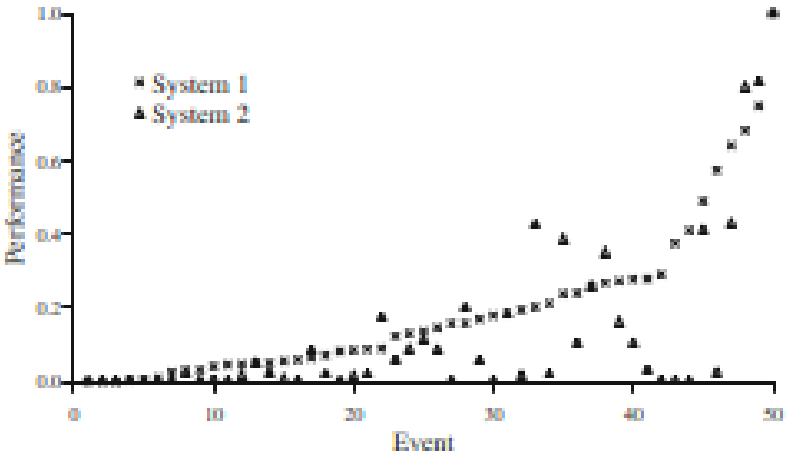


FIGURE 3.3. *The ability of each system to respond to an event, for each of 50 events.*



In the lower graph in the figure, the events have been sorted by the performance on System 1.

A “Statistical Principles” Checklist

- What variables might influence your results? Will analysis of these variables mean that you need to make use of statistics?

A "Statistical Principles" Checklist

- What variables might influence your results? Will analysis of these variables mean that you need to make use of statistics?
- Can you predict the effect of altering each variable? How do they interact? Are they independent?

A "Statistical Principles" Checklist

- What variables might influence your results? Will analysis of these variables mean that you need to make use of statistics?
- Can you predict the effect of altering each variable? How do they interact? Are they independent?
- How do the experiments distinguish between the effects of the variables?

A "Statistical Principles" Checklist

- What variables might influence your results? Will analysis of these variables mean that you need to make use of statistics?
- Can you predict the effect of altering each variable? How do they interact? Are they independent?
- How do the experiments distinguish between the effects of the variables?
- Are effects random or systematic? How are they to be controlled?

A "Statistical Principles" Checklist

- What variables might influence your results? Will analysis of these variables mean that you need to make use of statistics?
- Can you predict the effect of altering each variable? How do they interact? Are they independent?
- How do the experiments distinguish between the effects of the variables?
- Are effects random or systematic? How are they to be controlled?
- What method will be used to investigate outliers?

A "Statistical Principles" Checklist

- What variables might influence your results? Will analysis of these variables mean that you need to make use of statistics?
- Can you predict the effect of altering each variable? How do they interact? Are they independent?
- How do the experiments distinguish between the effects of the variables?
- Are effects random or systematic? How are they to be controlled?
- What method will be used to investigate outliers?
- What is the population? How is a sample to be taken? What is the argument that demonstrates that a sample will be representative?

A "Statistical Principles" Checklist

- What variables might influence your results? Will analysis of these variables mean that you need to make use of statistics?
- Can you predict the effect of altering each variable? How do they interact? Are they independent?
- How do the experiments distinguish between the effects of the variables?
- Are effects random or systematic? How are they to be controlled?
- What method will be used to investigate outliers?
- What is the population? How is a sample to be taken? What is the argument that demonstrates that a sample will be representative?
- How precise will the individual measurements be? How important is it to achieve a particular level of precision?

- What is the right way to summarize your results—an average? A median? A minimum?

- What is the right way to summarize your results—an average? A median? A minimum?
- What form of variance will be present, and how will it be captured in your results?

- What is the right way to summarize your results—an average? A median? A minimum?
- What form of variance will be present, and how will it be captured in your results?
- How large is the effect you are hoping to observe, and how many measurements will be required in order to reliably observe it?

- What is the right way to summarize your results—an average? A median? A minimum?
- What form of variance will be present, and how will it be captured in your results?
- How large is the effect you are hoping to observe, and how many measurements will be required in order to reliably observe it?
- Is a hypothesis test appropriate, and if so, which one?

- What is the right way to summarize your results—an average? A median? A minimum?
- What form of variance will be present, and how will it be captured in your results?
- How large is the effect you are hoping to observe, and how many measurements will be required in order to reliably observe it?
- Is a hypothesis test appropriate, and if so, which one?
- Do the results make sense? Are they consistent with any obvious points of comparison?

- What is the right way to summarize your results—an average? A median? A minimum?
- What form of variance will be present, and how will it be captured in your results?
- How large is the effect you are hoping to observe, and how many measurements will be required in order to reliably observe it?
- Is a hypothesis test appropriate, and if so, which one?
- Do the results make sense? Are they consistent with any obvious points of comparison?
- What visualizations might help provide insight into the pattern or behaviour of the results?