

# Scientific Methodology in Computer Science

MO430A

Prof. Dr. Bruno B. P. Cafeo

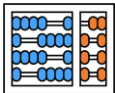
Institute of Computing  
University of Campinas



# Agenda

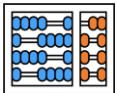
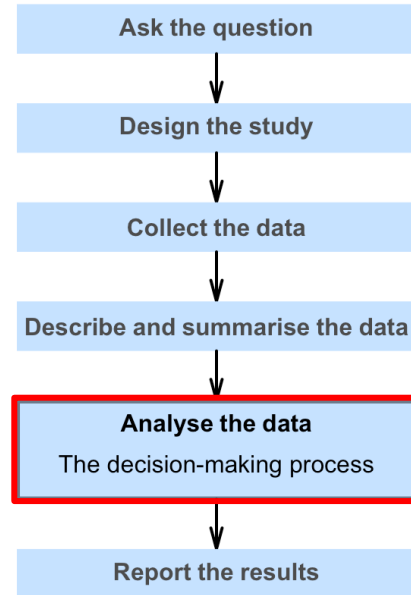
---

- Sampling variation
- The need for making decisions
- Assumption
- Expectation
- Observation
- Make a decision



# Where are we?

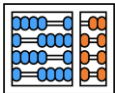
---



# Sampling

---

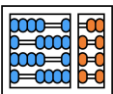
- Each sample is likely to be different.
- Our sample is just one of countless possible samples from the population.
- Each sample is likely to produce a different value for the sample statistic.
- Hence we only observe one of the many possible values for the sample statistic.
- Since many values for the sample statistic are possible, the possible values of the sample statistic vary (called sampling variation) and have a distribution (called a sampling distribution).



# Sampling Variation

---

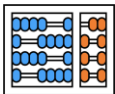
*Sampling variation* refers to how the sample estimates (statistics) vary from sample to sample, because each sample is different.



# Sampling Variation

---

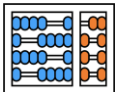
- Since we only have one value of the sample statistic, out of the many values of the sample statistic that are possible, what does this difference between the sample means imply about the difference between the population means?
- Two reasons could explain why the sample means are different:
  - The population means are the same; the difference is due to sampling variation. That is, we just happen to have, by chance, one of those samples where the difference between the means is quite noticeable. The sample means are different only because we have data from one of the many possible samples, and every sample is likely to be different.
  - Alternatively, the population means are different, and the sample means reflect this.
- How do we decide which of these explanations is supported by the data?



# Sampling Variation

---

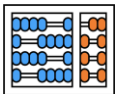
- The two possible explanations (*statistical hypotheses*) have special names:
  - There is no difference between the population parameters: the difference is simply due to sampling variation. This is the null hypothesis, or  $H_0$
  - There is a difference between the population parameters. This is the alternative hypothesis, or  $H_1$
- How do we decide which of these explanations is supported by the data? What is the decision-making process?



# Sampling Variation

---

- One approach to the decision-making process begins by assuming the null hypothesis is true.
- The data are examined to see if sufficient information exists to support the alternative hypothesis.
- However, conclusions drawn about the population from the sample can never be certain, since the sample studied is just one of many possible samples that could have been taken.





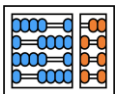
# The need for making decisions

---

- Suppose I draw a sample of 15 cards from the pack, and all are red cards. What should you conclude? How likely is it that this would happen simply by chance? Is this evidence that the pack of cards is somehow unfair, or rigged?



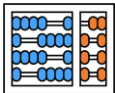
How likely is it that we get  
15 red cards in a row?



# The need for making decisions

---

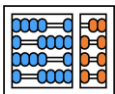
- Getting 15 reds cards out of 15 from a well-shuffled pack seems very unlikely, so you probably conclude that the pack is somehow unfair. But importantly, how did you reach that decision? Your unconscious decision-making process may have worked like this:
  - You assumed, quite reasonably, that I used a standard, well-shuffled pack of cards, where half the cards are red and half the cards are black. That is, you assumed the population proportion is  $p = 0.5$ .
  - Based on that assumption, you expected about half the cards in the sample of 15 to be red, and about half to be black. You wouldn't necessarily expect exactly half red and half black, but you'd probably expect something close to that. That is, you would expect that  $p$  would be close to 0.5.
  - But what you observed was nothing like that: All 15 cards were red. That is,  $p = 0$ .
  - You then made a decision: since what you observed (all red cards) was not like what you were expecting (about half red cards), the 15 red cards contradict what you were expecting, based on your assumption of a fair pack... so your assumption of a fair pack is probably wrong.



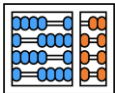
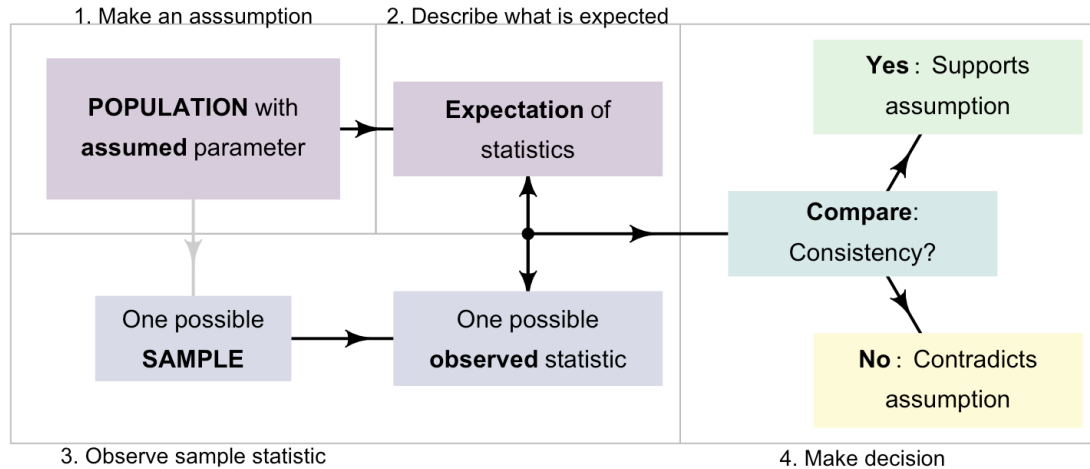
# How decisions are made

---

- **Assumption:** Make an assumption about the population parameter. Initially, assume that the sampling variation explains any discrepancy between the observed sample and assumed value of the population parameter. The initial assumption is that there has been 'no change, no difference, no relationship', depending on the context.
- **Expectation:** Based on the assumption about the parameter, describe what values of the sample statistic might reasonably be observed from all the possible samples that might be obtained (due to sampling variation).
- **Observation:** Observe the data from one of the many possible samples, and compute the observed sample statistic from this sample.
- **Decision:** If the observed sample statistic is:
  - unlikely to have happened by chance, it contradicts the assumption about the population parameter, and the assumption is probably wrong. The evidence suggests that the assumption is wrong (but it is not certainly wrong).
  - likely to have happened by chance, it is consistent with the assumption about the population parameter, and the assumption may be correct. No evidence exists to suggest the assumption is wrong (though it may be wrong).



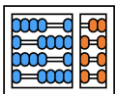
# How decisions are made



# How decisions are made

---

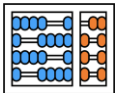
- **Assumption:** Make an assumption about the population parameter. Initially, assume that the sampling variation explains any discrepancy between the observed sample and assumed value of the population parameter. The initial assumption is that there has been 'no change, no difference, no relationship', depending on the context.
- **Expectation:** Based on the assumption about the parameter, describe what values of the sample statistic might reasonably be observed from all the possible samples that might be obtained (due to sampling variation).
- **Observation:** Observe the data from one of the many possible samples, and compute the observed sample statistic from this sample.
- **Decision:** If the observed sample statistic is:
  - unlikely to have happened by chance, it contradicts the assumption about the population parameter, and the assumption is probably wrong. The evidence suggests that the assumption is wrong (but it is not certainly wrong).
  - likely to have happened by chance, it is consistent with the assumption about the population parameter, and the assumption may be correct. No evidence exists to suggest the assumption is wrong (though it may be wrong).



# Making decisions in research

---

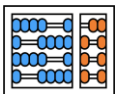
- The assumption about the parameter;
- The expectation of the statistic;
- The observations;
- Make a decision.



# Assumption about the population parameter

---

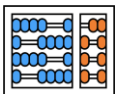
- The initial assumption is that there has been 'no change, no difference, no relationship', depending on the context. Using this idea, a reasonable assumption can be made about the population parameter:
  - We might assume that no difference exists between the parameter for two groups in the population, since we don't have any evidence yet to say there is a difference.
  - We might be interested in testing a claim, or evaluating a benchmark, about a population parameter, to determine if the evidence supports this claim or benchmark.
- These assumptions about the population parameter are called null hypotheses.



# Expectations of sample statistics

---

- Having assumed a value for the population parameter, the second step is to determine what values to expect from the sample statistic, based on this assumption.
- Since many samples are possible, and every sample is likely to be different (sampling variation), the value of the sample statistic depends on which one of the possible samples we obtain: the sample statistic is likely to be different for every sample.

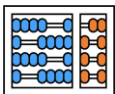




# Expectations of sample statistics

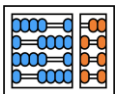
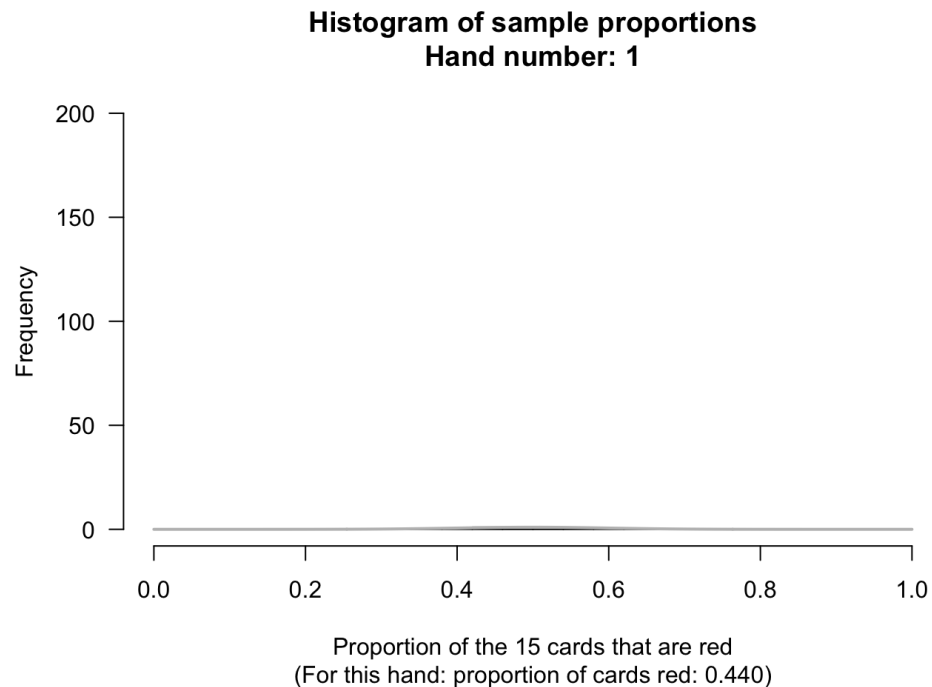
---

- Let`s consider the cards previously mentioned. Assuming a fair pack, then half the cards are red in the population (the pack of cards), so the population proportion is assumed to be  $p=0.5$ . In a sample of 15 cards, what values could be reasonably expected for the sample proportion  $p$  of red cards (the statistic)?
- How would  $p$  vary from sample to sample? Perhaps 15 red cards out of 15 cards happens reasonably frequently... or perhaps it doesn't. How could we find out? We could:
  - use mathematical theory.
  - shuffle a pack of cards and deal 15 cards many hundreds of times, then count how often we see 15 red cards out of 15 cards.
  - simulate (using a computer) dealing 15 cards many hundreds of times, and count how often we get 15 red cards out of 15 cards.



# Expectations of sample statistics

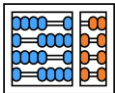
---



# Observations about our sample

---

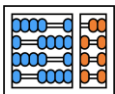
- We then take a sample (one of the many samples that are possible), and observe the sample statistic. In this situation, we observe 15 red cards out of 15 cards.



# Making a decision

---

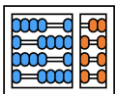
- Using the sample data, we make a decision.
- We note that observing 15 red cards out of 15 cards is quite rare: it never happened once in the 1000 simulations. So based on simulating one thousand hands, we could conclude that we would almost never find 15 red cards in 15 cards... if the assumption of a fair pack was true.
- But we did find 15 red cards in 15 cards... so the assumption (a fair pack) is probably wrong.



# Making a decision

---

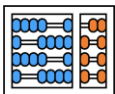
- What if we had observed 4 red cards in a hand of 15 cards (a sample proportion of  $p=4/15=0.267$ ), rather than 15 red cards out of 15?
- The conclusion is not quite so obvious then: these values of  $p$  are uncommon, but they certainly do happen when  $p=0.5$ .
- In these situations, a more sophisticated approach for making a decision is needed.
- Special tools are needed to describe what to expect from the sample statistic after making assumptions about the population parameter.

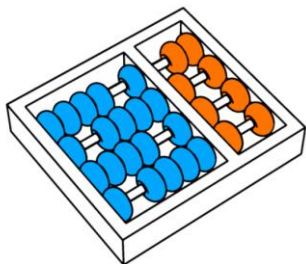


# The “tools”

---

- Tools to describe the random nature of what happens with sample statistics, and so determine if the sample statistic is consistent with the assumption: **Probability**.
- Tools to describe the distribution of the population and the sample: **Distributions and models**.
- Tools to describe how sample statistics vary from sample to sample (sampling variation), and hence what to expect from the sample statistic: **Sampling variation**.





**INSTITUTO DE  
COMPUTAÇÃO**



**Prof. Dr. Bruno B. P. Cafeo**

Sala 04  
Instituto de Computação - Unicamp  
Av. Albert Einstein, 1251  
Cidade Universitária  
Campinas – SP  
13083-852

<https://ic.unicamp.br/~cafeo/>  
[cafeo@ic.unicamp.br](mailto:cafeo@ic.unicamp.br)