

Scientific Methodology in Computer Science

MO430A

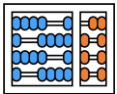
Prof. Dr. Bruno B. P. Cafeo

Institute of Computing
University of Campinas

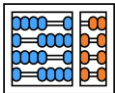
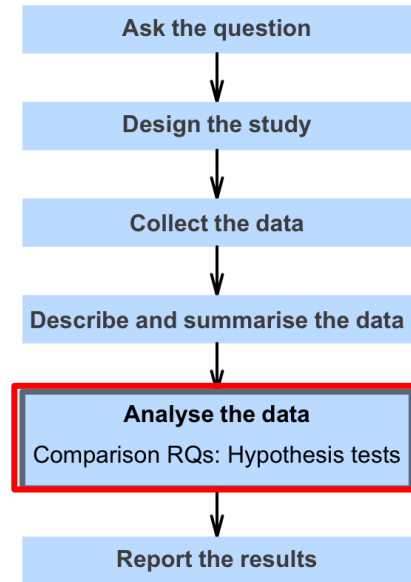


Agenda

- Hypothesis
- P-value
- Test for one proportion
- Test for one mean
- Test for mean difference (paired data)
- Test for means of two independent groups



Where are we?



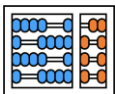
RQ - Overview

- RQs can also be written with one of two purposes in mind:
 - **Estimation:** These RQs ask how precisely a value in the population is estimated by using the sample, and are answered using confidence intervals.

Among {the population}, what is {the outcome}?

- **Making decisions:** These RQs are concerned with making a decision about a population, and are answered using hypothesis testing.

Among {the population}, is {the outcome} equal to {some value}?



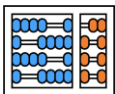
RQ - Overview

- RQs can also be written with one of two purposes in mind:
 - **Estimation:** These RQs ask how precisely a value in the population is estimated by using the sample and are answered using confidence intervals.

Among {the population}, what is {the outcome}?

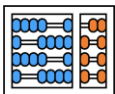
- **Making decisions:** These RQs are concerned with making a decision about a population, and are answered using hypothesis testing.

Among {the population}, is {the outcome} equal to {some value}?



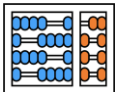
How decisions are made... again

- **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).



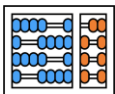
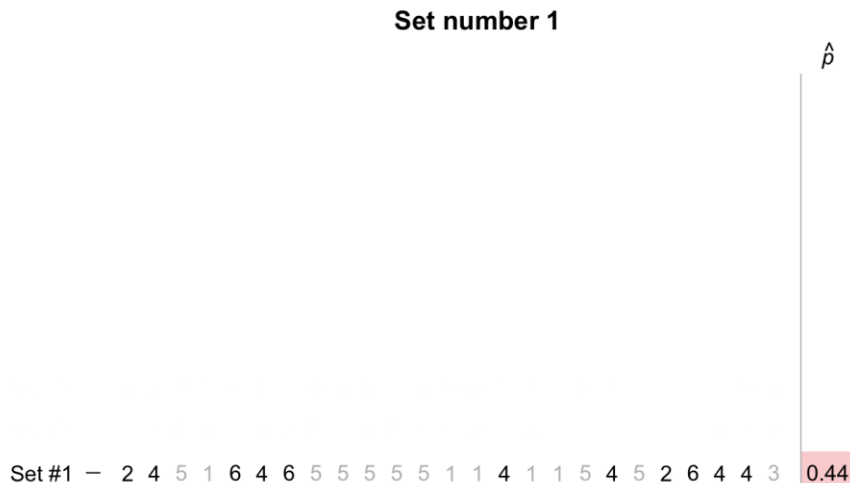
Hypothesis

- The word hypothesis means “a possible explanation”
- **Scientific hypotheses** refer to potential scientific explanations that can be tested by collecting data. For example, an engineer may hypothesise that replacing sand with glass in the manufacture of concrete will produce desirable characteristics (Devaraj et al. 2021).
- **Statistical hypotheses** refer to statistical explanations that are required to determine whether the evidence (i.e., data) supports the scientific hypotheses. The statistical hypotheses are the foundation of the logic of hypothesis testing.



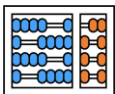
Dice problem... again

- Suppose a fair, six-sided dice is rolled 25. What proportion of the rolls will produce an even number? That is, what will be the sample proportion of even numbers?



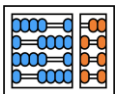
Dice problem... again

- If the dice was fair, I would expect about one-sixth of rolls to produce *1*, but not necessarily exactly one-sixth of the rolls, due to sampling variation.
- However, by initially assuming the population proportion of ones would be $1/6$, the possible values of the sample proportion from all possible rolls of the fair dice could be determined.
- This is the beginning of the decision-making process



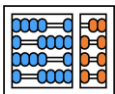
Dice problem... again

- More formally, the initial assumption about the population is that the dice is fair (I have no evidence against this), and hence that the population proportion of rolling a is $p=1/6$, or approximately $p=0.16667$.
- Then, the values of the sample proportion that are reasonable to expect from all possible sample is described, and compared to the observed value of \hat{p} from just one of those possible samples



Test for one proportion

- If the sample proportion of rolls that show a certain outcome is not exactly $1/6$, two possibilities exist:
 - The population proportion is $1/6$, and the sample proportion is not exactly $p = 1/6$ due to sampling variation.
 - The population proportion is **not** $1/6$; in other words, the sample proportion is not exactly $p = 1/6$ because the dice is not fair.
- These two potential explanations are referred to as statistical hypotheses. Formally, the two statistical hypotheses above are written as:
 - H_0 : $p = 1/6$, the *null hypothesis*; and
 - H_1 : $p \neq 1/6$, the *alternative hypothesis*.

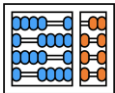


Sampling Distribution

- When the proportion of rolls that show a really is $p=1/6 = 0.1666$, what values of the sample proportion are reasonable to expect from all possible samples, given sampling variation?

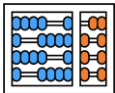
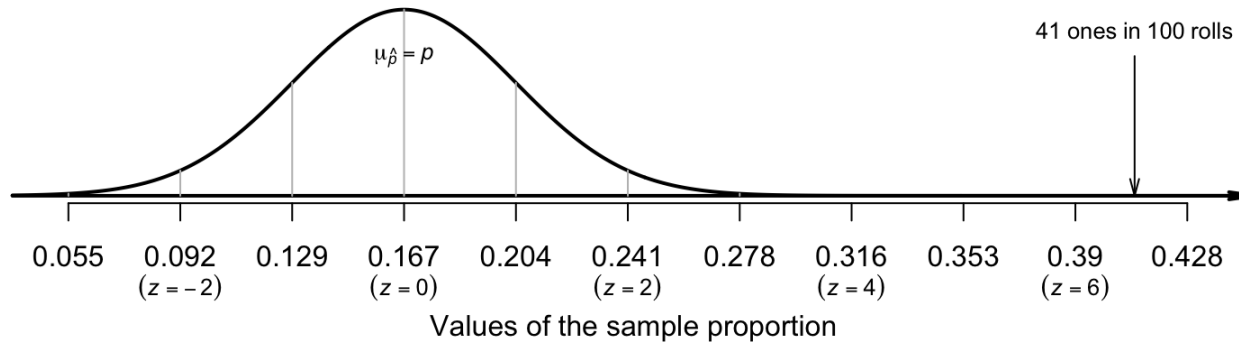
$$\text{s.e.}(\hat{p}) = \sqrt{\frac{p \times (1 - p)}{n}}$$

- Considering 100 rolls
- An approximate normal distribution
- With mean $\mu_p=1/6$
- $\text{s.e.}(\hat{p})=0.037267$



Sampling Distribution

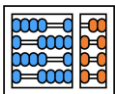
**Sampling distribution of the sample proportion of ones
in 100 rolls**



Computing the test statistic and z-scores

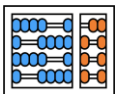
- One way to measure how far the sample proportion $\hat{p}=0.41$ is from the population proportion $p=1/6$ in 100 rolls is to use a z-score

$$\begin{aligned} z &= \frac{\text{sample statistic} - \text{mean of the distribution}}{\text{standard deviation of the distribution}} \\ &= \frac{\hat{p} - p}{\text{s.e.}(\hat{p})} \\ &= \frac{0.41 - 0.1666...}{0.037267} = 6.53. \end{aligned}$$



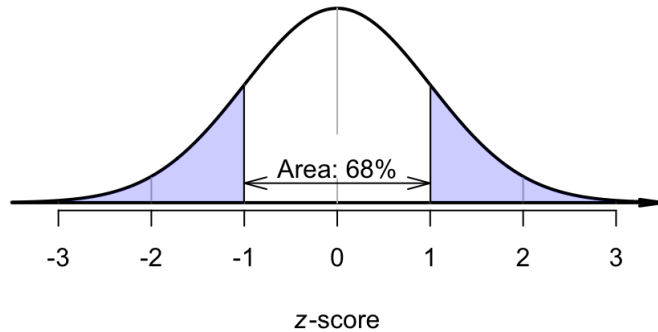
Determining P-values

- The value of the z-score shows that the value of \hat{p} is highly very unusual... but how unusual?
- Quantifying how unusual is assessed more precisely using a P-value, which is used widely in scientific research.
- **The P-value is a way of measuring how unusual an observation is, when H_0 is assumed to be true**

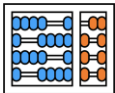
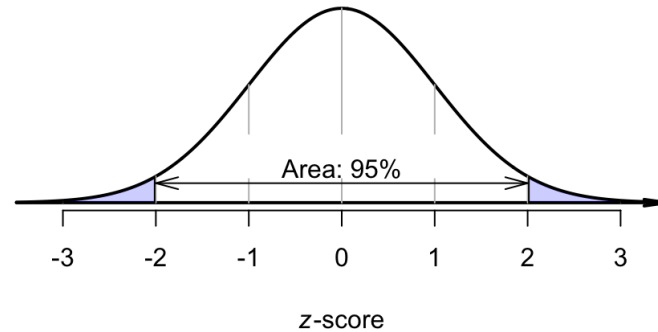


Determining P-values

The P -value if $z = 1$

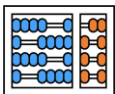


The P -value if $z = 2$



Making decisions with P-values

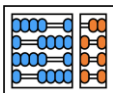
- P-values tells us the probability of observing the sample statistic (or something even more extreme), assuming the null hypothesis is true.
- In this context, the P-value tells us the probability of observing the value of \hat{p} (or something more extreme), just through sampling variation (chance) if $p=0.1666$
- In this die-rolling example, where the P-value is very small, the data contradict the null hypothesis (that $p=1/6$), suggesting that the dice may not be fair



Summary

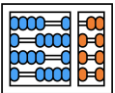
1. **Assumption:** Write the *null hypothesis* and *alternative hypothesis* about the *parameter* (based on the RQ):
 - $H_0: p = 0.1666\dots$, and
 - $H_1: p \neq 0.1666\dots$ (this is a two-tailed alternative hypothesis).
2. **Expectation:** The sampling distribution describes what values to expect reasonably expect from the sample statistic across all possible samples, *if* the null hypothesis is true. Under certain circumstances, the sample proportions will vary with an approximate normal distribution around a mean of $p = 0.1666\dots$ with a standard deviation of $\text{s.e.}(\hat{p}) = 0.0372678$.
3. **Observation:** Compute the z -score: $z = 6.53$ to measure the distance between the assumed population value, and the observed sample value.
4. **Consistency?:** Determine if the data are consistent with the assumption, by computing the P -value. Here, the P -value is (much) less than 0.001 . The P -value can be computed by software, or approximated using the 68--95--99.7 rule.

The **conclusion** is that very strong evidence exists that p is *not* 0.16667 , based on the evidence.



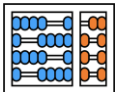
Test for one mean

- The average internal body temperature is commonly believed to be 37.0 °C, based on data over 150 years old (Wunderlich 1868).
- More recently, researchers wanted to re-examine this claim (Mackowiak, Wasserman, and Levine 1992) to see if this benchmark is still appropriate.
- The null hypothesis (H_0): ???, and
- The alternative hypothesis (H_1): ???.



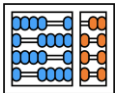
Test for one mean

- The average internal body temperature is commonly believed to be 37.0 °C, based on data over 150 years old (Wunderlich 1868).
- More recently, researchers wanted to re-examine this claim (Mackowiak, Wasserman, and Levine 1992) to see if this benchmark is still appropriate.
- The null hypothesis (H_0): $\mu = 37$, and
- The alternative hypothesis (H_1): $\mu \neq 37$.



Describing the sampling distribution

- The sample mean is $\bar{x} = 36.8051$ °C
- The sample standard deviation is $s = 0.40732$ °C
- The sample size is $n = 130$

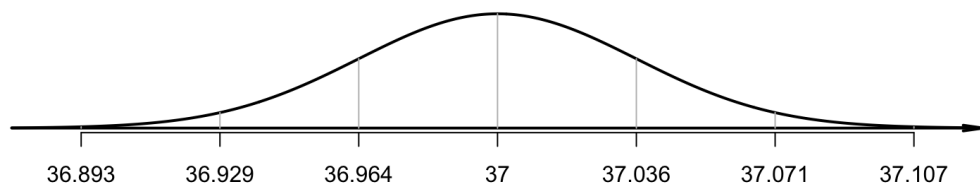


Expected sample means

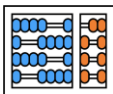
An approximate normal distribution;

With a sampling mean whose value is $\mu = 37.0^\circ\text{C}$ (from H_0);

With standard deviation of $\text{s.e.}(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{0.40732}{\sqrt{130}} = 0.035724$. This is the *standard error* of the sample means.

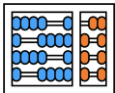
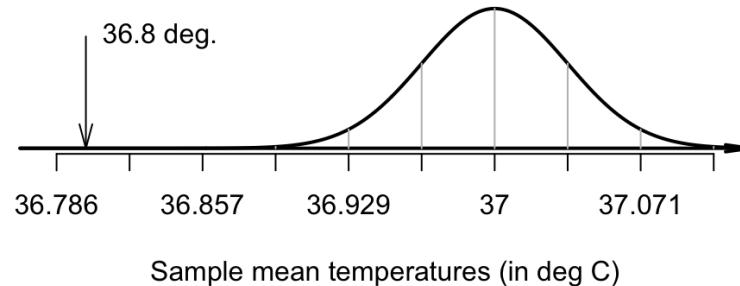


Sample means from sample of size 130 (deg C)



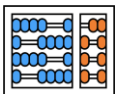
Computing the test statistic

- The sampling distribution describes how the sample means varies; that is, what to expect from the sample means, after assuming $\mu=37.0$ °C.
- How likely is it that such a value could occur in our sample by chance (by sampling variation)?



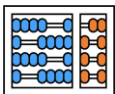
Making decisions with P-values

- P-values measure the likelihood of observing the sample statistic (or something more extreme), based on the assumption about the population parameter being true.
- The P-value tells us the likelihood of observing the value of \bar{x} (or something more extreme), just through sampling variation if $\mu=37$.
- There is strong evidence that population mean body temperature is not 37 °C



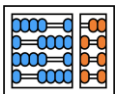
Test for the mean difference (paired data)

- μ_d : The mean *difference* in the *population*.
 - \bar{d} : The mean *difference* in the *sample*.
 - s_d : The *sample* standard deviation of the *differences*.
 - n : The number of *differences*.
- $H_0: \mu_d = 0$.



Test for the mean difference (paired data)

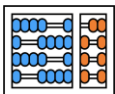
- μ_d : The mean *difference* in the *population*.
 - \bar{d} : The mean *difference* in the *sample*.
 - s_d : The *sample* standard deviation of the *differences*.
 - n : The number of *differences*.
- $H_0: \mu_d = 0$.



Test for means of two independent groups

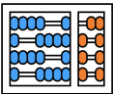
	Group A	Group B
Population means:	μ_A	μ_B
Sample means:	\bar{x}_A	\bar{x}_B
Standard deviations:	s_A	s_B
Standard errors:	$\text{s.e.}(\bar{x}_A) = \frac{s_A}{\sqrt{n_A}}$	$\text{s.e.}(\bar{x}_B) = \frac{s_B}{\sqrt{n_B}}$
Sample sizes:	n_A	n_B

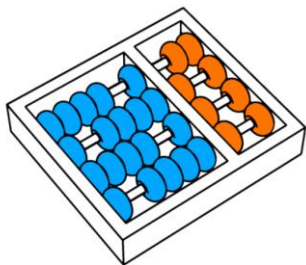
$$H_0: \mu_P - \mu_C = 0 \text{ (or } \mu_P = \mu_C \text{)}.$$



Selecting a test

- <https://www.intro2r.info/unit3/which-test.html>
- <https://www.scribbr.com/statistics/statistical-tests/>





**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