

Scientific Methodology in Computer Science

MO430A

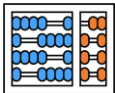
Prof. Dr. Bruno B. P. Cafeo

Institute of Computing
University of Campinas

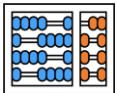
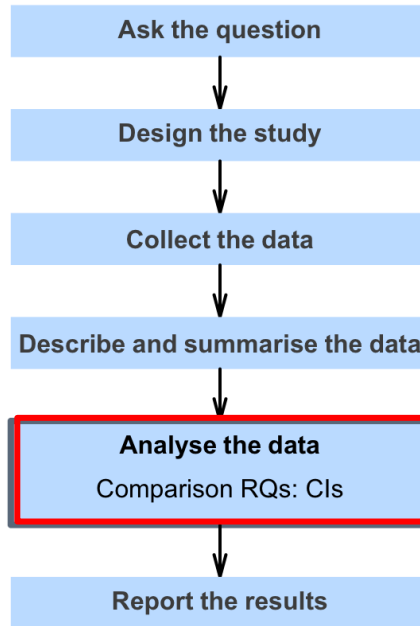


Agenda

- Sample distribution
- Sampling means and standard errors
- Confidence Intervals
 - Known proportions
 - 68-95-99.7
 - Unknown proportions
- CI for one mean
- CIs for mean differences (paired data)
- CIs for two independent means

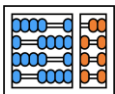


Where are we?



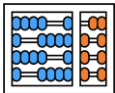
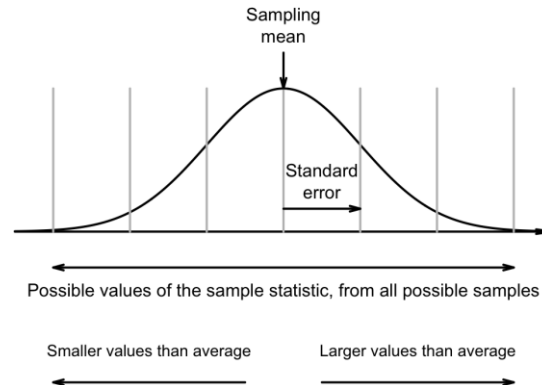
Sample proportions have a distribution

A **sampling distribution** is the distribution of some sample statistic, showing how its value varies from sample to sample.



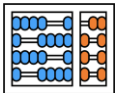
Sampling Means and Standard Errors

- **Sampling mean:** The sampling mean is the mean of the sampling distribution of a statistic.
- **Standard error:** A standard error is the standard deviation of the sampling distribution of a statistic.



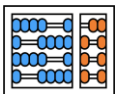
Sampling Variation

- A **sampling distribution** describes how all possible values of a sample statistic is likely to vary from sample to sample.
- Under certain circumstances, the sampling distribution often can be described by a **normal distribution**.
- The standard deviation of this normal distribution is called a **standard error**.
- The standard error is the name specifically given to the standard deviation that describes the variation in the sample statistic *across all possible samples*.



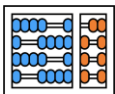
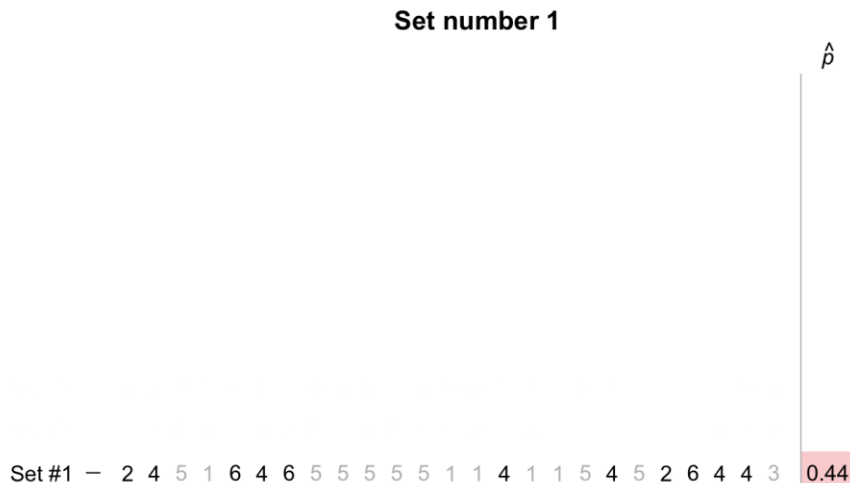
Introducing confidence intervals

- Confidence intervals help answer estimation-type RQ, where the precision of a statistic is of interest.
 - Estimation: These RQs ask how precisely a value in the population is estimated by using the sample, and are answered using confidence intervals.
 - Making decisions: These RQs are concerned with making a decision about a population, and are answered using hypothesis testing.
- Descriptive RQs:
 - One proportion where the response variable is qualitative.
 - One mean where the response variable is quantitative.
 - Mean difference for paired quantitative data.
- Relational or interventional RQs with a comparison:
 - Comparing means in two groups
 - Comparing odds in two groups



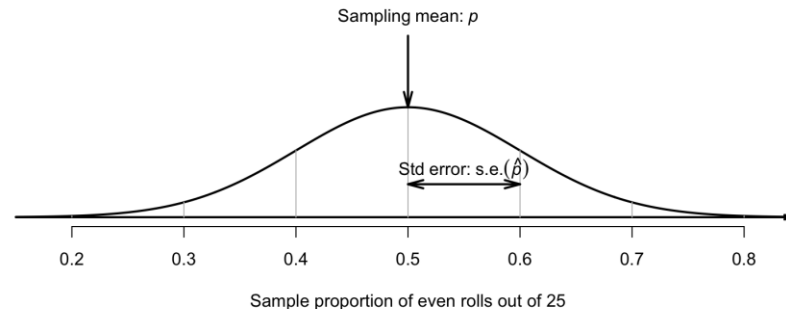
Sampling distribution: known proportion

- 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?

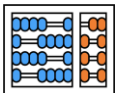


Sampling distribution: known proportion

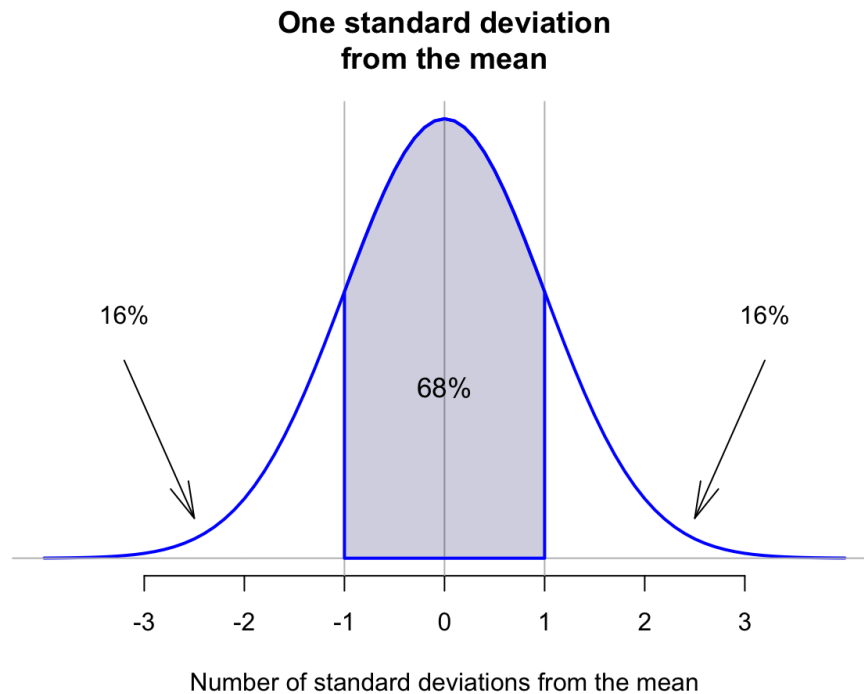
- The proportion of even rolls varies from set to set. For these 10 sets of $n=25$ rolls, the percentage of even rolls ranged from $\hat{p}=0.32$ even rolls to $\hat{p}=0.60$ even rolls.
- The mean of this distribution is the sampling mean, and its value is p .
- The standard deviation for this distribution is the standard error denoted $s.e.(\hat{p})$.



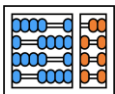
$$s.e.(\hat{p}) = \sqrt{\frac{p \times (1 - p)}{n}}$$



68 – 95 – 99.7



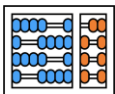
$$p \pm (\text{multiplier} \times \text{s.e.}(\hat{p}))$$



Sampling distribution: unknown proportion

- In the dice example (Sect. 20.1), the sampling distribution for the sample proportion was given, including an equation for computing the standard error for the sample proportion for samples of size n , when the value of p was known.
- When p is unknown, the best available estimate can be used, which is \hat{p} .

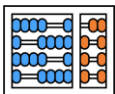
$$\text{s.e.}(\hat{p}) = \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}.$$



Sampling distribution: unknown proportion

- Let's pretend for the moment that the proportion of even rolls of a fair die is unknown (to demonstrate ideas).
- In this case, an estimate of the proportion of even rolls can be found by rolling a dice $n = 25$ times and computing \hat{p} .
- Suppose 11 of the $n=25$ rolls produced an even number, so that $\hat{p}=11/25=0.44$

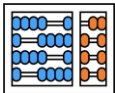
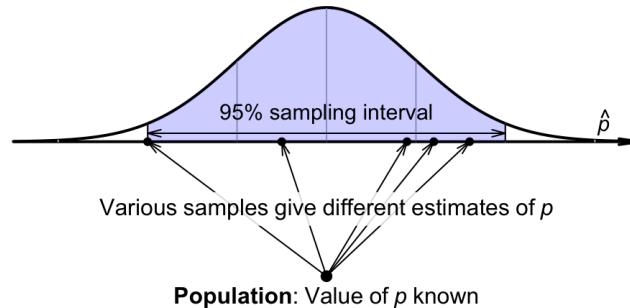
$$\text{s.e.}(\hat{p}) = \sqrt{\frac{0.44 \times (1 - 0.44)}{25}} = 0.099277.$$



68 – 95 – 99.7

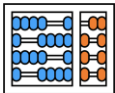
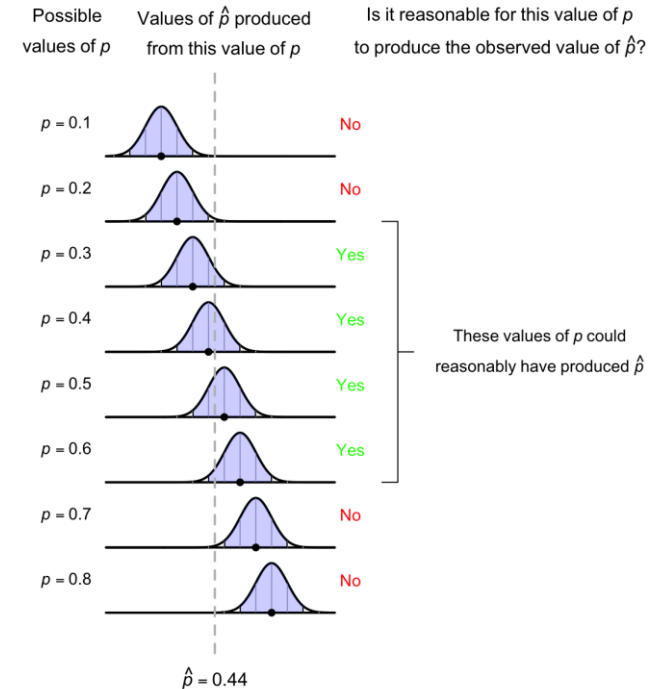
- Using the 68 - 95 - 99.7 rule again: about 95% of the values of \hat{p} are expected to be between $p - 0.199$ and $p + 0.199$.

Sample: Values of \hat{p} likely to be produced with the given value of p



68 – 95 – 99.7

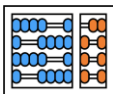
- Using the 68 - 95 - 99.7 rule again: about 95% of the values of \hat{p} are expected to be between $p - 0.199$ and $p + 0.199$.
- This is equivalent to saying that we are reasonably sure that a population with a value of 0.24 and 0.64 could reasonably have produced the observed value of $\hat{p}=0.44$.
- This interval is called a **confidence interval**



CI for one mean and for mean differences (paired data)

- Data are paired when two observations about the same variable are recorded for each unit of analysis. Paired data come from within individual comparisons.

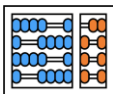
	One sample mean	Mean of paired data
The observations:	Values: x	Differences: d
Sample mean:	\bar{x}	\bar{d}
Standard deviation:	s	s_d
Standard error of sample mean:	$\text{s.e.}(\bar{x}) = \frac{s}{\sqrt{n}}$	$\text{s.e.}(\bar{d}) = \frac{s_d}{\sqrt{n}}$
Sample size:	Number of <i>observations</i> : n Number of <i>differences</i> : n	



CI for one mean and for mean differences (paired data)

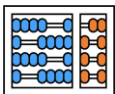
- Data are paired when two observations about the same variable are recorded for each unit of analysis. Paired data come from within individual comparisons.

	One sample mean	Mean of paired data
The observations:	Values: x	Differences: d
Sample mean:	\bar{x}	\bar{d}
Standard deviation:	s	s_d
Standard error of sample mean:	$\text{s.e.}(\bar{x}) = \frac{s}{\sqrt{n}}$	$\text{s.e.}(\bar{d}) = \frac{s_d}{\sqrt{n}}$
Sample size:	Number of <i>observations</i> : n Number of <i>differences</i> : n	



CI for two independent means

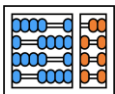
- A study examined the reaction times of students while driving (Strayer and Johnston 2001; Agresti and Franklin 2007).
- In one study, students were randomly allocated to one of two groups: one to use a mobile phone while driving, and one to not use a mobile phone while driving.
- This is a between-individuals comparison, since different students are in each group. The reaction time for each student was measured in a driving simulator.
- What are P, I, C, O in in this study?



CI for two independent means

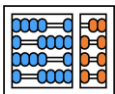
- For the reaction-time data, we use the subscript **P** for the phone-users group, and **C** for the control (non-phone users) group.
- Using this notation, the difference between population means (the parameter) is $\mu_P - \mu_C$. Since the population values are unknown, this parameter is estimated using the statistic $\bar{x}_P - \bar{x}_C$.

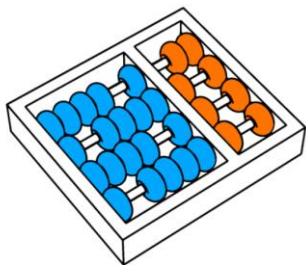
	Phone users: Group P	Non-phone users: Group C	Difference ($P - C$)
Population means:	μ_P	μ_C	$\mu_P - \mu_C$
Sample means:	\bar{x}_P	\bar{x}_C	$\bar{x}_P - \bar{x}_C$
Standard deviations:	s_P	s_C	
Sample sizes:	n_P	n_C	
Standard errors:	$\text{s.e.}(\bar{x}_P) = \frac{s_P}{\sqrt{n_P}}$	$\text{s.e.}(\bar{x}_C) = \frac{s_C}{\sqrt{n_C}}$	$\text{s.e.}(\bar{x}_P - \bar{x}_C)$



CI for two independent means

- A study examined the reaction times of students while driving (Strayer and Johnston 2001; Agresti and Franklin 2007).
- In one study, students were randomly allocated to one of two groups: one to use a mobile phone while driving, and one to not use a mobile phone while driving.
- This is a between-individuals comparison, since different students are in each group. The reaction time for each student was measured in a driving simulator.
- What are P, I, C, O in in this study?





**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