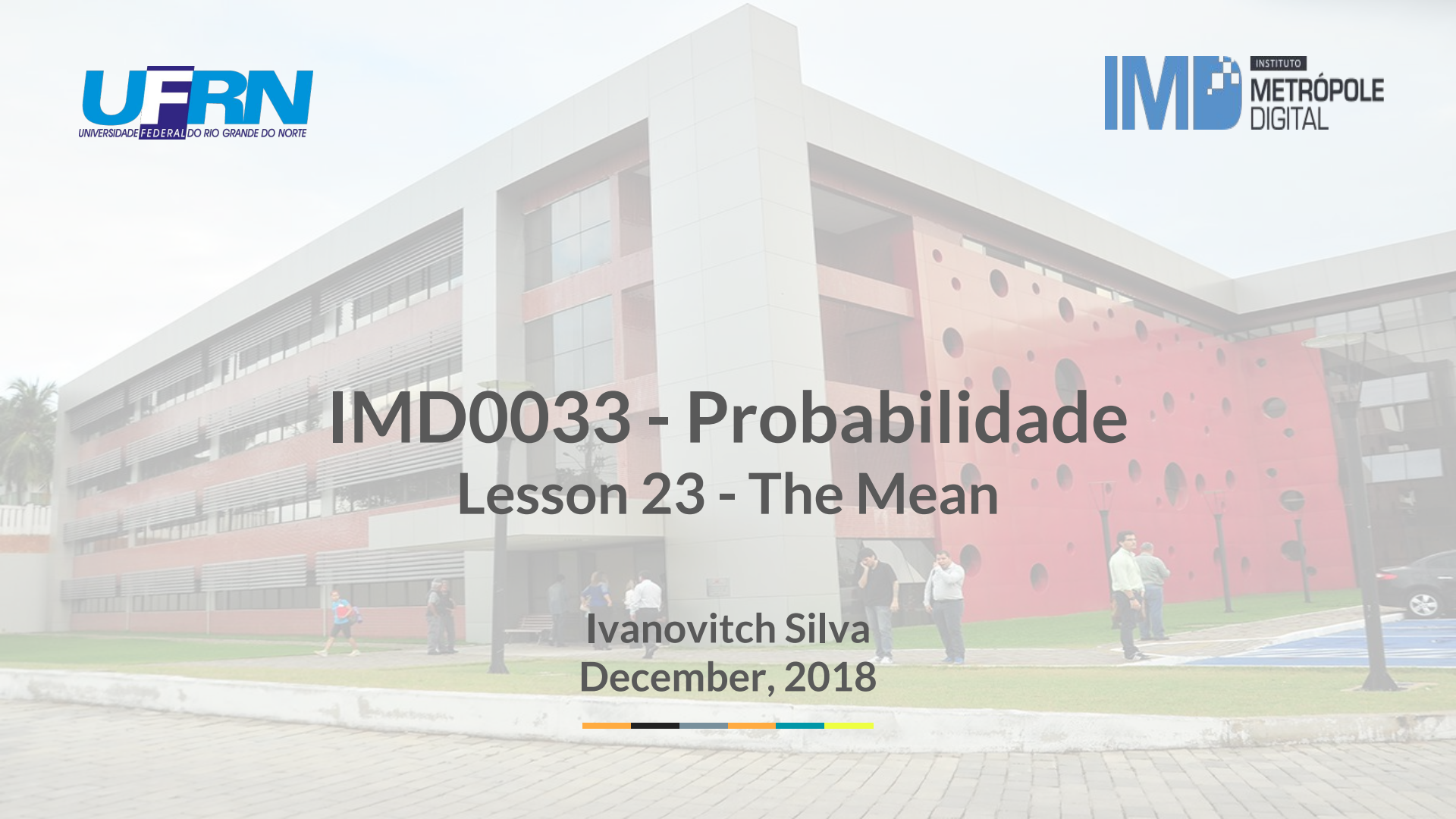


IMD0033 - Probabilidade

Lesson 23 - The Mean

Ivanovitch Silva

December, 2018

Agenda

- The mean
- The mean as a balance point
- Defining the mean algebraically
- Estimating the population mean
- Estimates from low-sized samples

Atualizar o repositório

```
git clone https://github.com/ivanovitchm/imd0033_2018_2.git
```

Ou ....

```
git pull
```

PREVIOUSLY ON...

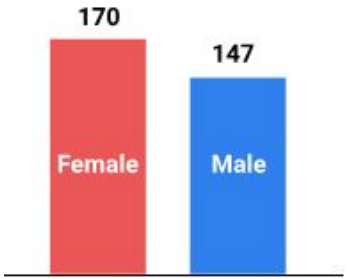
Data **source**

Getting **good** data

Id	Name	Salary	...	Gender
1	Mary Ann	$35 000	....	Female
2	Marc Downey	$55 000	....	Male
..	....	....	....	....
51	Juliet Ali	$45 000	....	Female
...	...	....	....	....
317	Jane Ace	$95 000	....	Female

Understand how the data is **structured** and **measured**

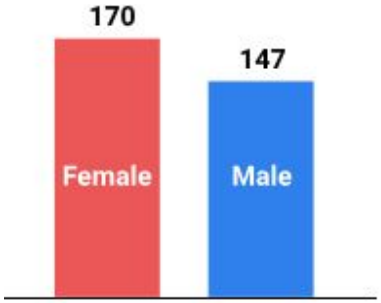
Frequency tables are not the only way of bringing data to a comprehensible form.



Visualize the patterns

Gender	Frequency
Male	147
Female	170

Organize the data in **comprehensible** forms to find patterns



Visualize the patterns

Gender	Frequency
Male	147
Female	170

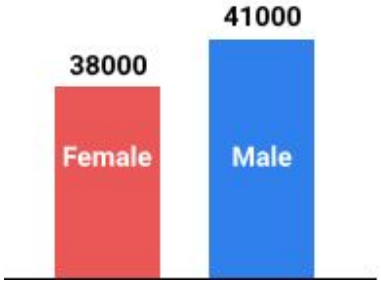
By measuring frequencies and build **frequency tables**

Organize the data in **comprehensible** forms

By summarizing it with **average values**

Gender	Average salary
Male	41000
Female	38000

Visualize the patterns



Depending on the particular characteristics of a distribution, we'll see that we can summarize it using the **mean**, the **weighted mean**, the **median**, or the **mode**.

We'll also learn to measure the **variability** in a distribution

$$A = [3, 3, 3, 3]$$

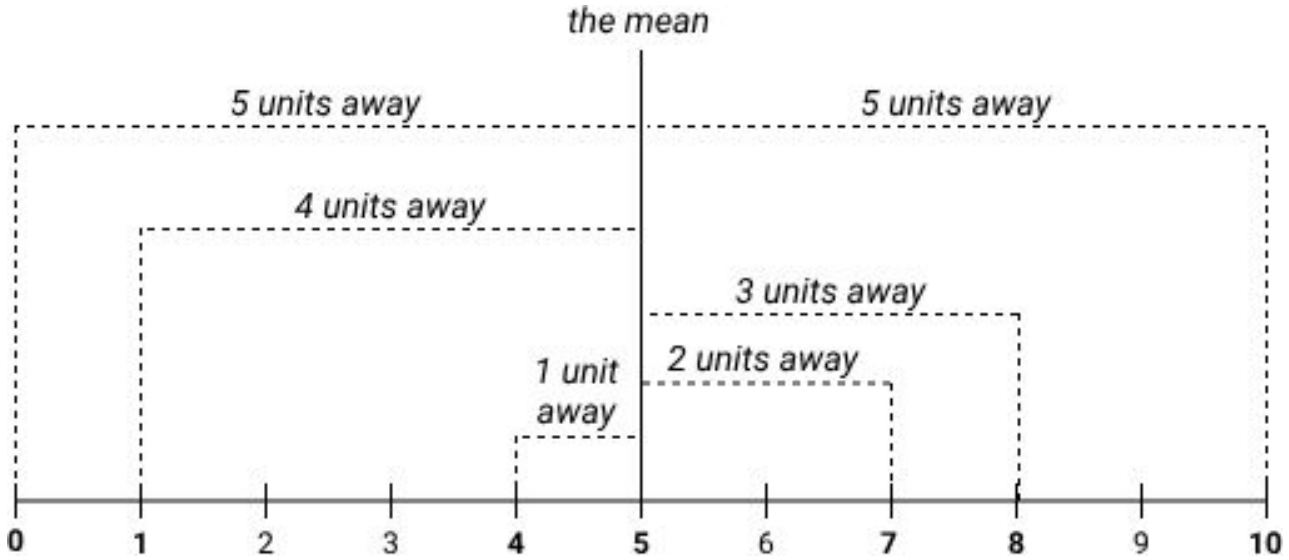
$$B = [30, 1, 15, 43]$$

We can clearly see that there's much more variability (diversity) in B. We'll learn to quantify variability using measures like **variance** and **standard deviation**.

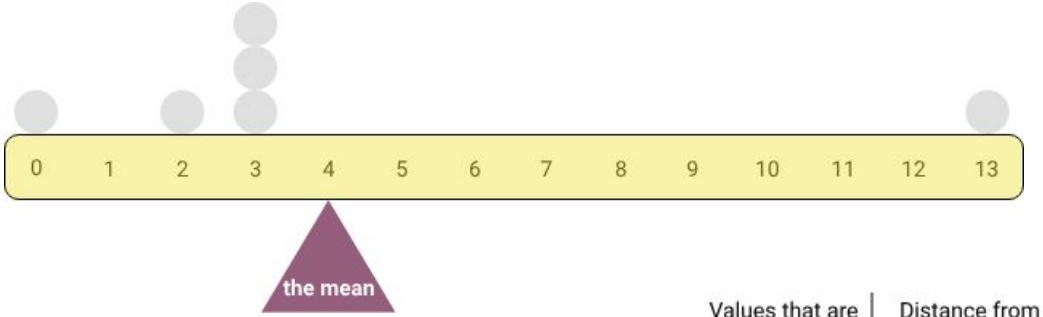
The Mean

$$[0, 1, 4, 7, 8, 10]$$

$$\frac{0+1+4+7+8+10}{6} = \frac{30}{6} = 5$$



The Mean as a Balance Point



[0,2,3,3,3,4,13]

Values that are **below** the mean	Distance from the mean
0	4 units
2	2 units
3	1 unit
3	1 unit
3	1 unit
	Total distance: ***9 units***

Values that are **above** the mean	Distance from the mean
13	9 units
	Total distance: ***9 units***

Defining the mean algebraically

$$\mu = \frac{x_1 + x_2 + \dots x_N}{N}$$

$N = 7$

[0 2 3 3 3 4 13]

$x_1\ x_2\ x_3\ x_4\ x_5\ x_6\ x_7$

$$\mu = \frac{0 + 2 + 3 + 3 + 3 + 4 + 13}{7} = \frac{28}{7} = 4$$

Population Mean

$$\bar{X} = \frac{x_1 + x_2 + \dots x_n}{n}$$

$n = 3$

[2 3 4]

$x_1\ x_2\ x_3$

$$\bar{X} = \frac{2 + 3 + 4}{3} = \frac{9}{3} = 3$$

Sample Mean

	Population	Sample
Mean	μ	$\bar{x}, \bar{x}_n, \bar{X}, M$
Number of values	N	n

Defining the mean algebraically

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{\sum X}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Population Mean

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum X}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Sample Mean

Journal of Statistics Education

Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project

Dean De Cock
Truman State University

Journal of Statistics Education Volume 19, Number 3(2011),
www.amstat.org/publications/jse/v19n3/decock.pdf

Key Words: Multiple Regression; Linear Models; Assessed Value; Group Project.

Abstract

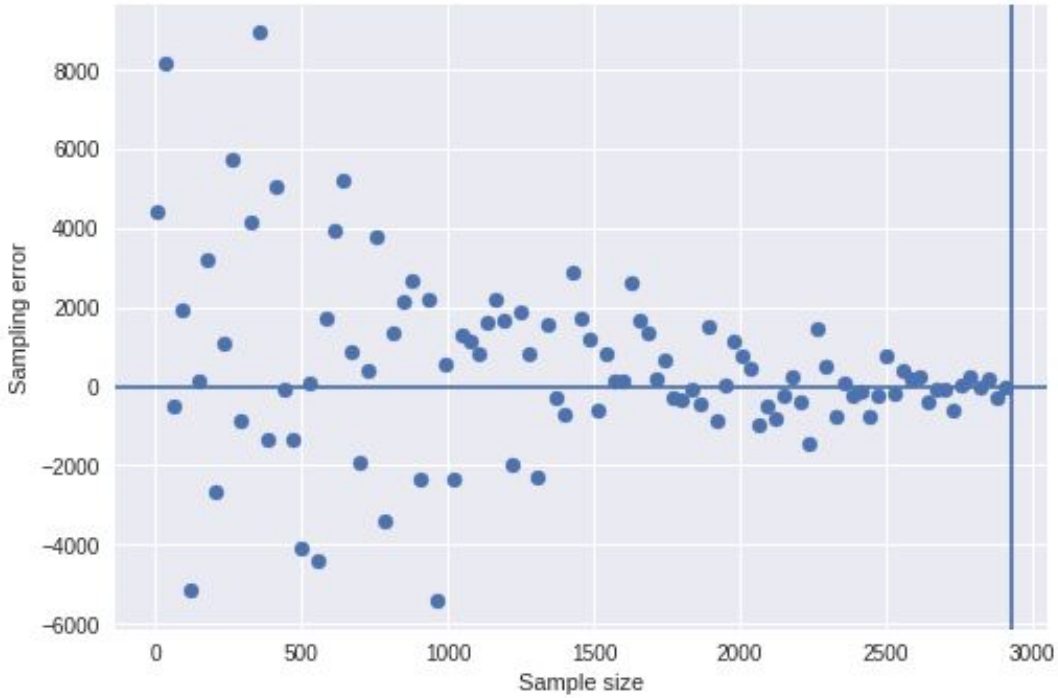
This paper presents a data set describing the sale of individual residential property in Ames, Iowa from 2006 to 2010. The data set contains 2930 observations and a large number of explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) involved in assessing home values. I will discuss my previous use of the Boston Housing Data Set and I will suggest methods for incorporating this new data set as a final project in an undergraduate regression course.

```
import pandas as pd
pd.set_option('display.max_columns', 500)

houses = pd.read_csv("AmesHousing_1.txt",sep='\t')
houses.shape
```

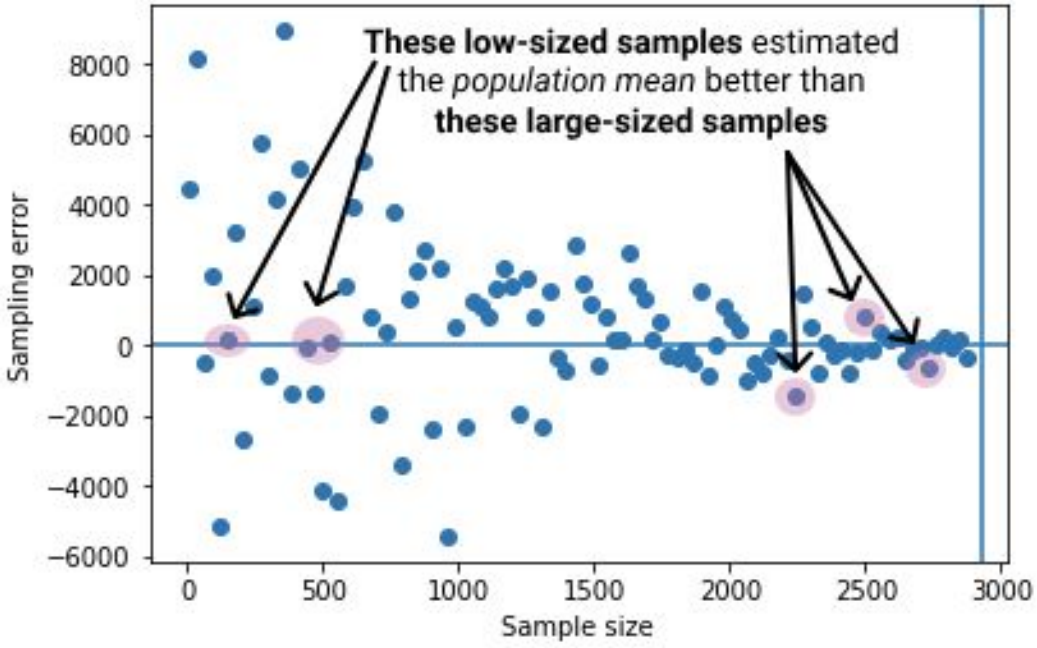
__	Order	PID	MS SubClass	MS Zoning	Lot Frontage	Lot Area	Street	Alley	Mo Sold	Yr Sold	Sale Type	Sale Condition	SalePrice
0	1	526301100	20	RL	141.0	131770	Pave	0	5	2010	WD	Normal	215000
1	2	526350040	20	RH	80.0	11622	Pave	0	6	2010	WD	Normal	105000
2	3	526351010	20	RL	81.0	14267	Pave	12500	6	2010	WD	Normal	172000
3	4	526353030	20	RL	93.0	11160	Pave	0	4	2010	WD	Normal	244000
4	5	527105010	60	RL	74.0	13830	Pave	0	3	2010	WD	Normal	189900

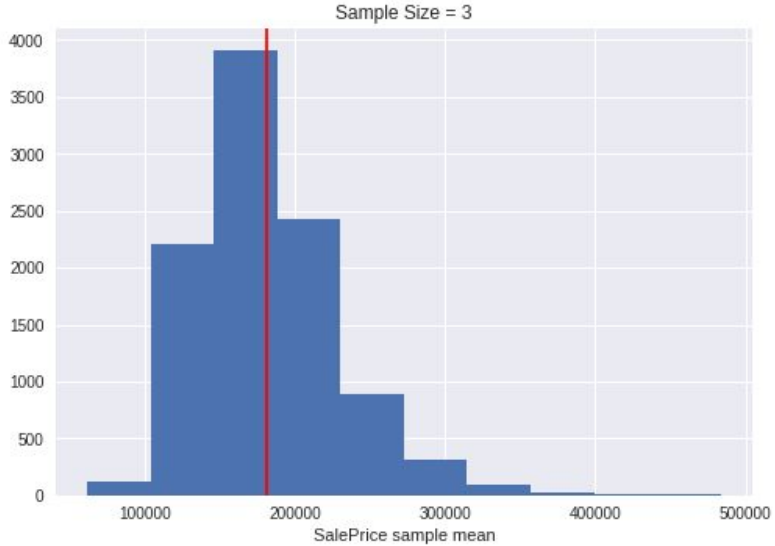
Estimating the population mean

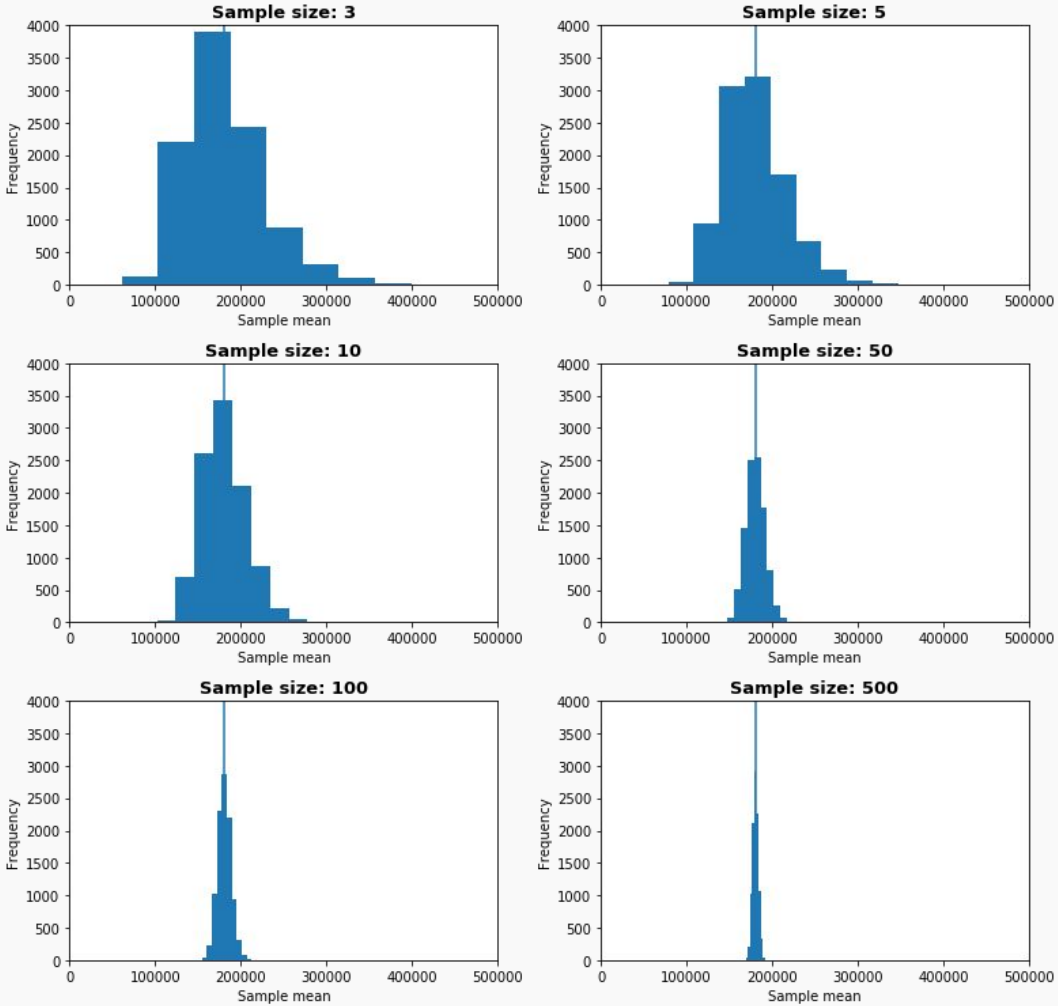


The general tendency for the sampling error is to decrease as the sample size increases.

Estimates from low-sized samples






Sample size: 3
Sample size: 5
Sample size: 10
Sample size: 50
Sample size: 100
Sample size: 500
Frequency
Sample mean
0
500
1000
1500
2000
2500
3000
3500
4000
0
100000
200000
300000
400000
500000

Next Steps

In the next mission, we'll explore a few edge cases where it's either impossible to compute the mean, or it's possible but not theoretically sound.


MacBook Pro