Data Science & ML Course

Lesson #14 Statistics Fundamentals I

Ivanovitch Silva

November, 2018

Agenda

- Sampling
  - Population and sampling
  - Sampling error
  - Simple random sampling (SRS)
  - Stratified sampling
  - Clustering sampling
- Variables in statistics
  - Quantitative and qualitative variables
  - Scale of measurements (nominal, ordinal, interval, ratio)

Update from repository

```
git clone https://github.com/ivanovitchm/datascience2machinelearning.git
```

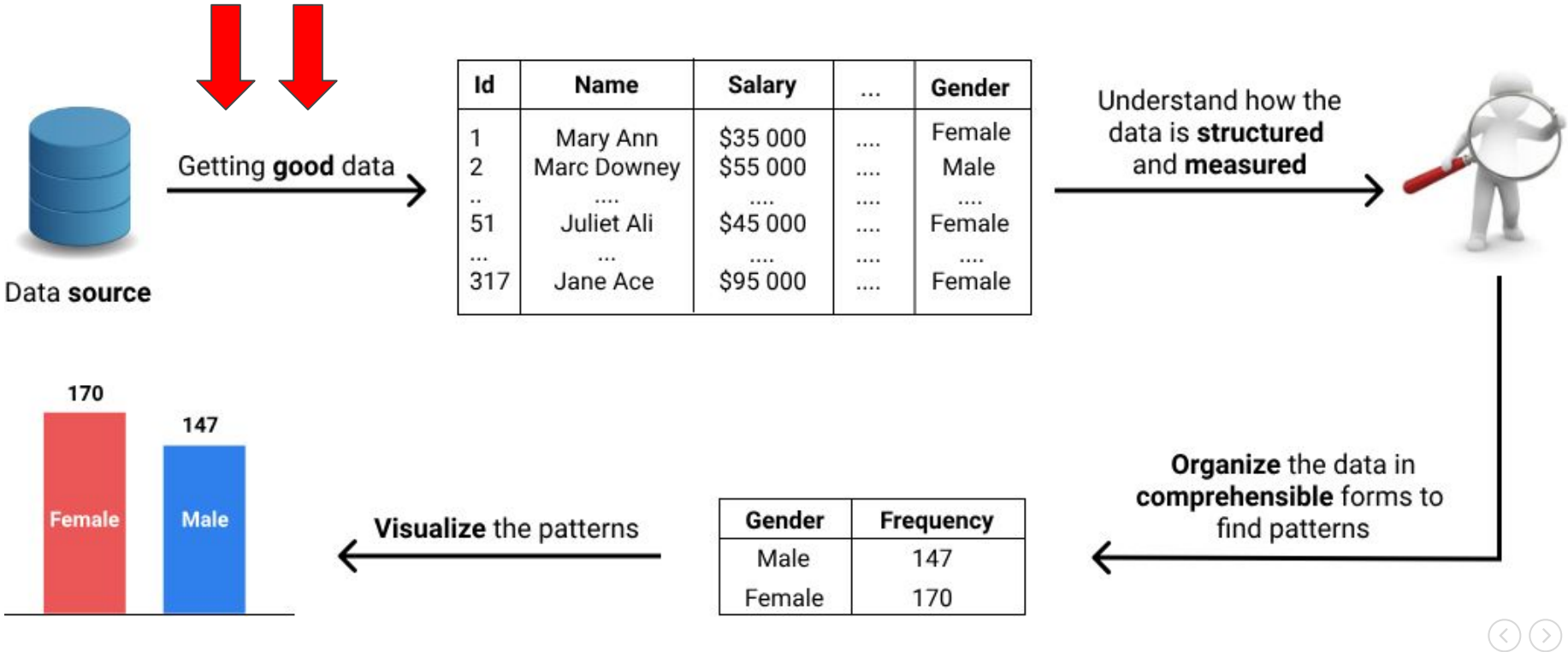
Or ....

```
git pull
```

PREVIOUSLY ON...

- Perform basic data analysis
- Data visualization
- Fundamental statistical metrics like the mean or the median, and we plotted histograms, bar graphs or line plots.

Go much deeper into the theory

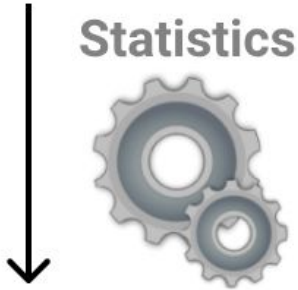


Solving problems with Statistics

Id	Name	Salary ($ / year)	Unexpected days off	Late at work	Extra hours worked
1	Macy Davidson	56000	6	Seldom	26
2	Jake Pugh	29000	0	Often	0
3	Draven Whitaker	43000			
4	Izabella Pratt	35000			
5	Gerardo Baker	25000			
6	Milo Norton	50000			
7	Fiona Benson	60000			

Id	Name	Department	Salary ($ / year)	....	Age
1	Alec Sullivan	IT	42000	....	26
2	Agustin Wang	PR	35000	....	31
3	Jocelyn Pruitt	Marketing	41000	....	32
....	....	....	....	....	....
72	Lainey Smith	IT	27000	....	21
....	....	....	....	....	....
231	Lexi Wilcox	IT	85000	....	57

Id	Name	Department	Salary ($ / year)	....	Age
1	Alec Sullivan	IT	42000	....	26
2	Agustin Wang	PR	35000	....	31
3	Jocelyn Pruitt	Marketing	41000	....	32
....	....	....	....	....	....
72	Lainey Smith	IT	27000	....	21
....	....	....	....	....	....
231	Lexi Wilcox	IT	85000	....	57



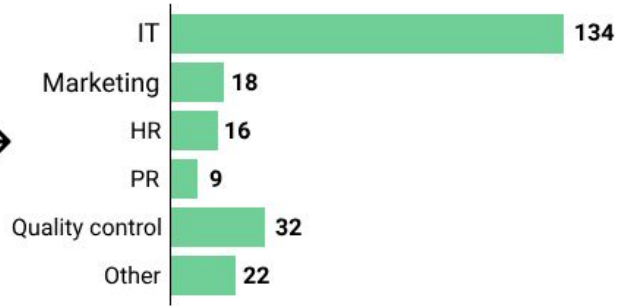
Summarize data

- **Most frequent value** in the *Department* column: IT

Organize data in a **comprehensible** form

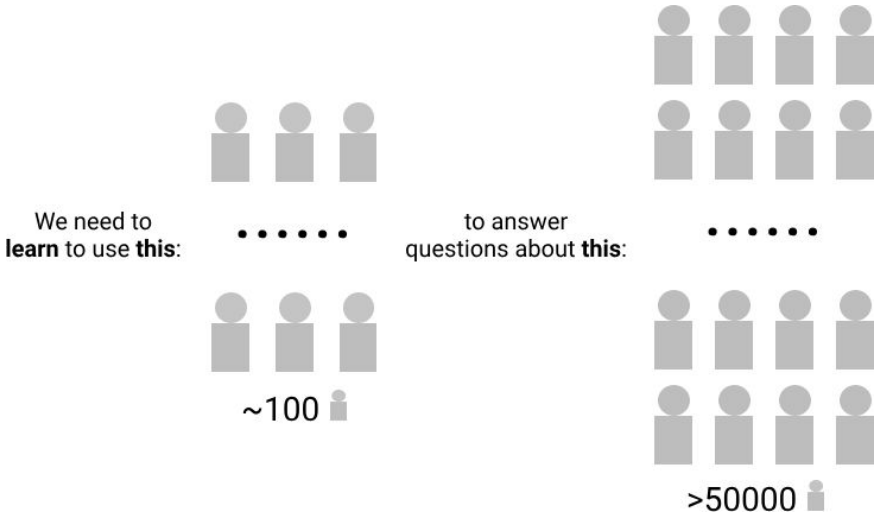
Department	Frequency
IT	134
Marketing	18
HR	16
PR	9
Quality control	32
Other	22

Visualize data

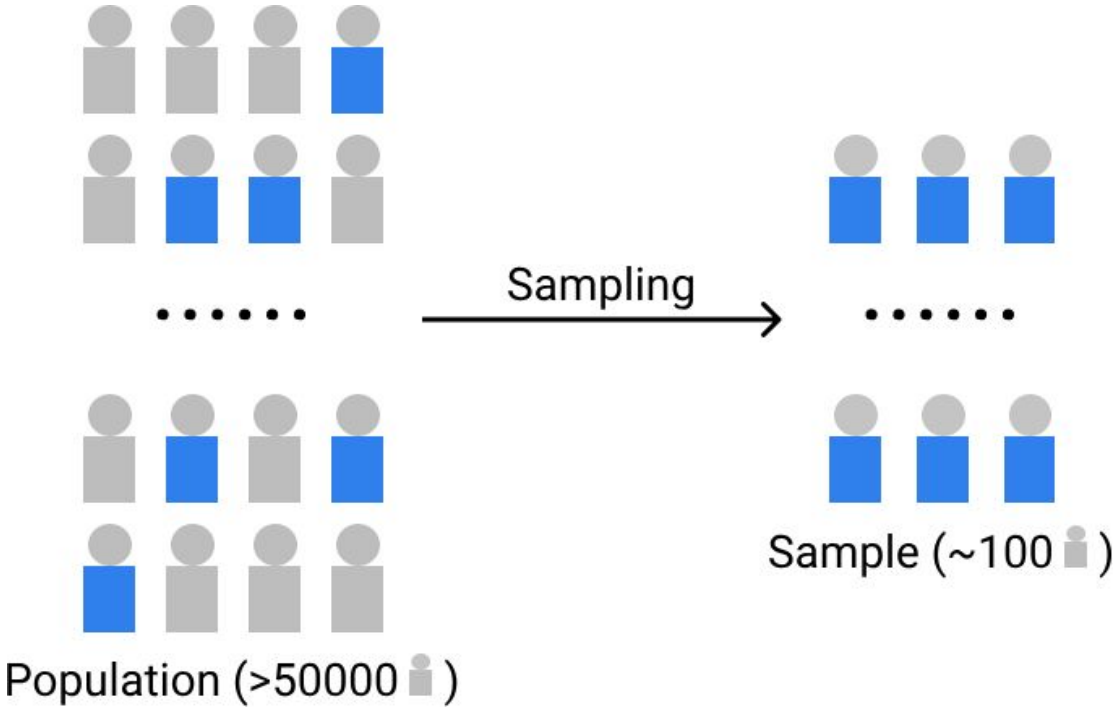


Solving problems with statistics

1. You run an international company with over **50000 employees.**
2. Now you want to determine whether the employees have been impacted negatively in any significant way.



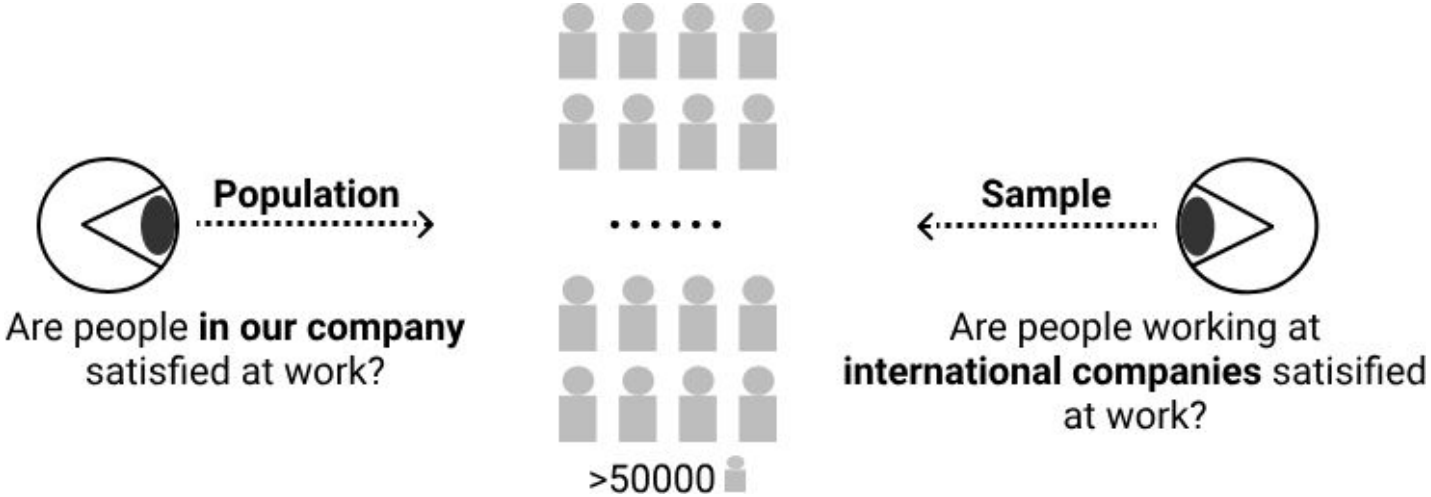
Population and Sampling



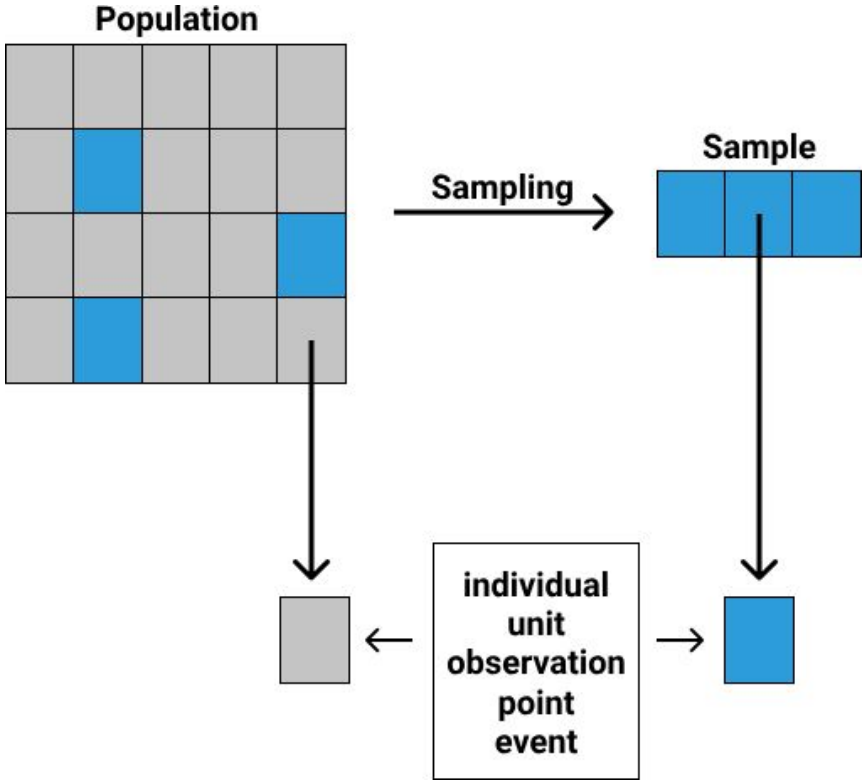
Whether a set of data is a sample or a population depends on the question we're trying to answer.

Population and Sampling

If we tried to find out whether people at international companies are satisfied at work



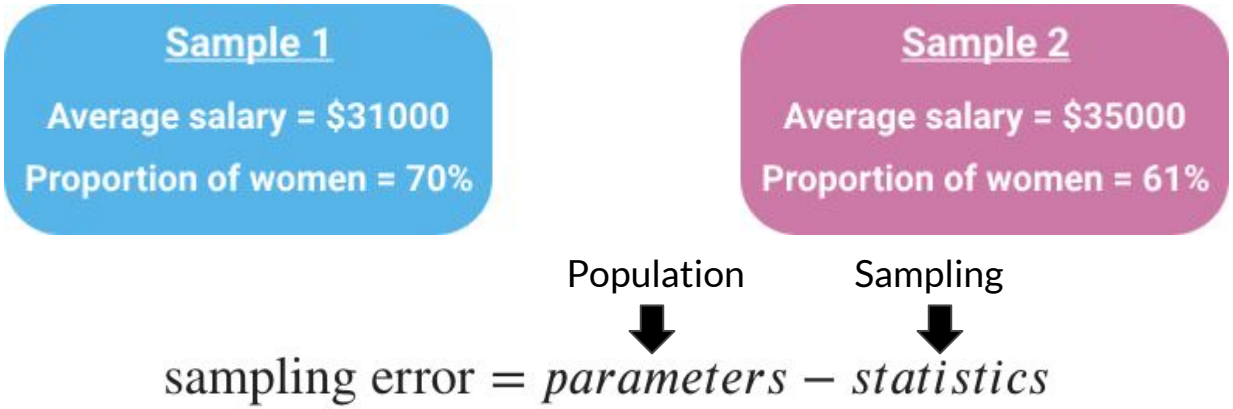
Population and Sampling



You'll often see this terminology used interchangeably: sample unit, sample point, sample individual, or sample observation.

Sampling error

For instance, let's say we know that the average salary in our company is **U$ 34500**, and the **proportion of women is 60%**.



Dataset

WNBA Player stats Season 2016-2017

Points, Assists, Height, Weight and other personal details and stats

Thomas De Jonghe • updated a year ago (Version 1)

7 voters

share

Data | Overview | Kernels | Discussion | Activity | Download (9 KB) | New Kernel

Data (9 KB) | API | `kaggle datasets download -d jinxbe/wnba-player-s…` | ? | Download All

Data Sources

WNBA Stats.csv 143 x 32

About this file

Personal Stats, bio from the 2016-2017 season WNBA.

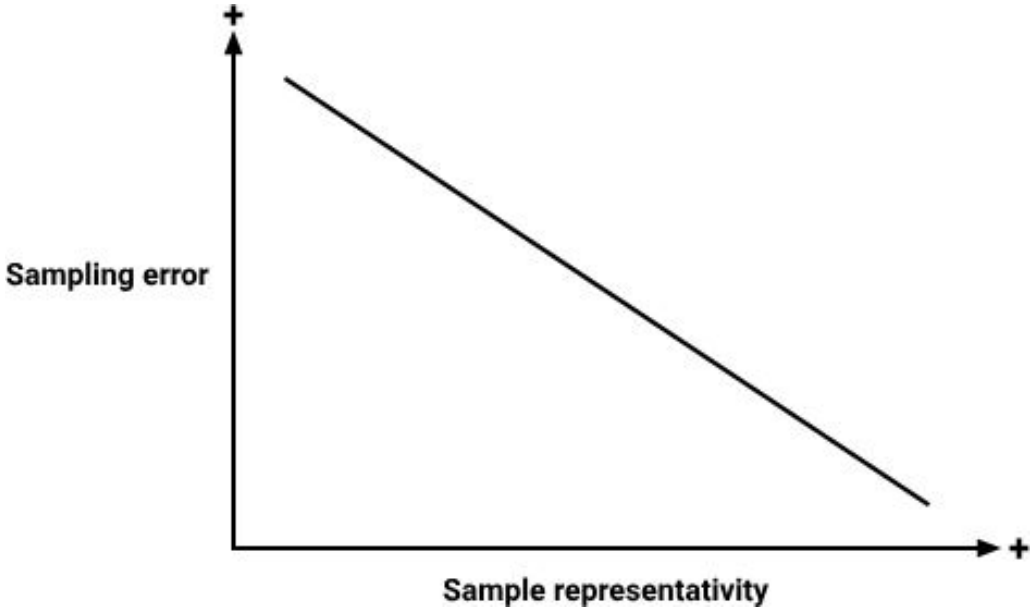
Columns

- A Name
- A Team
- A Pos
- # Height
- # Weight
- # BMI
- A Birth_Place
- Birthdate

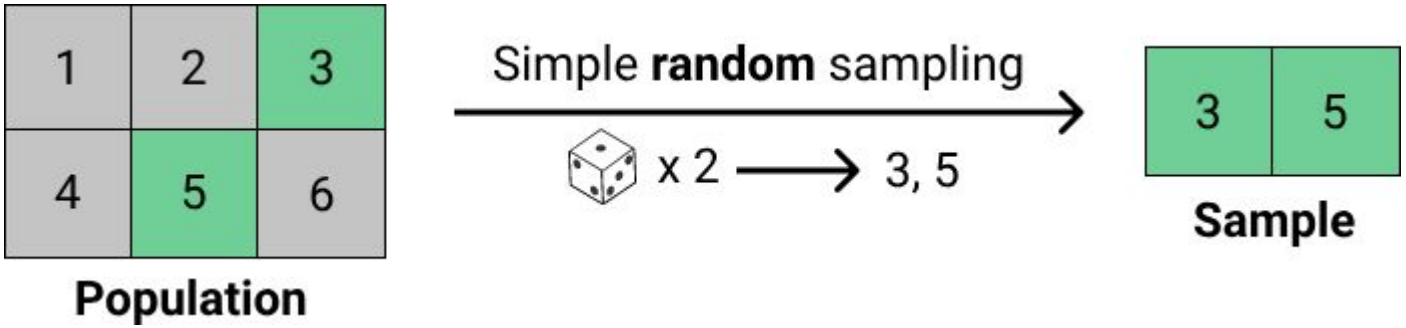
Dataset

	Name	Team	Pos	Height	Weight	BMI	Birth_Place	Birthdate	Age	College	Experience	Games Played	MIN	FGM	FGA
0	Aerial Powers	DAL	F	183	71.0	21.200991	US	January 17, 1994	23	Michigan State	2	8	173	30	85
1	Alana Beard	LA	G/F	185	73.0	21.329438	US	May 14, 1982	35	Duke	12	30	947	90	177
2	Alex Bentley	CON	G	170	69.0	23.875433	US	October 27, 1990	26	Penn State	4	26	617	82	218
3	Alex Montgomery	SAN	G/F	185	84.0	24.543462	US	December 11, 1988	28	Georgia Tech	6	31	721	75	195
4	Alexis Jones	MIN	G	175	78.0	25.469388	US	August 5, 1994	23	Baylor	R	24	137	16	50

Simple Random Sampling (SRS)

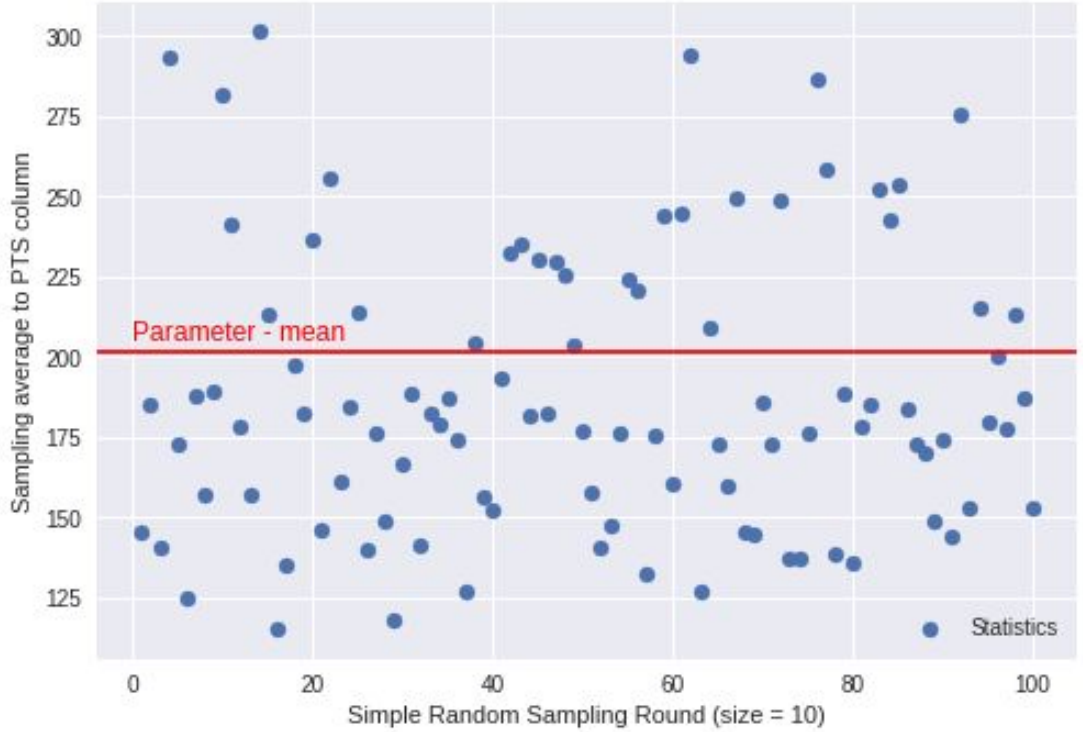


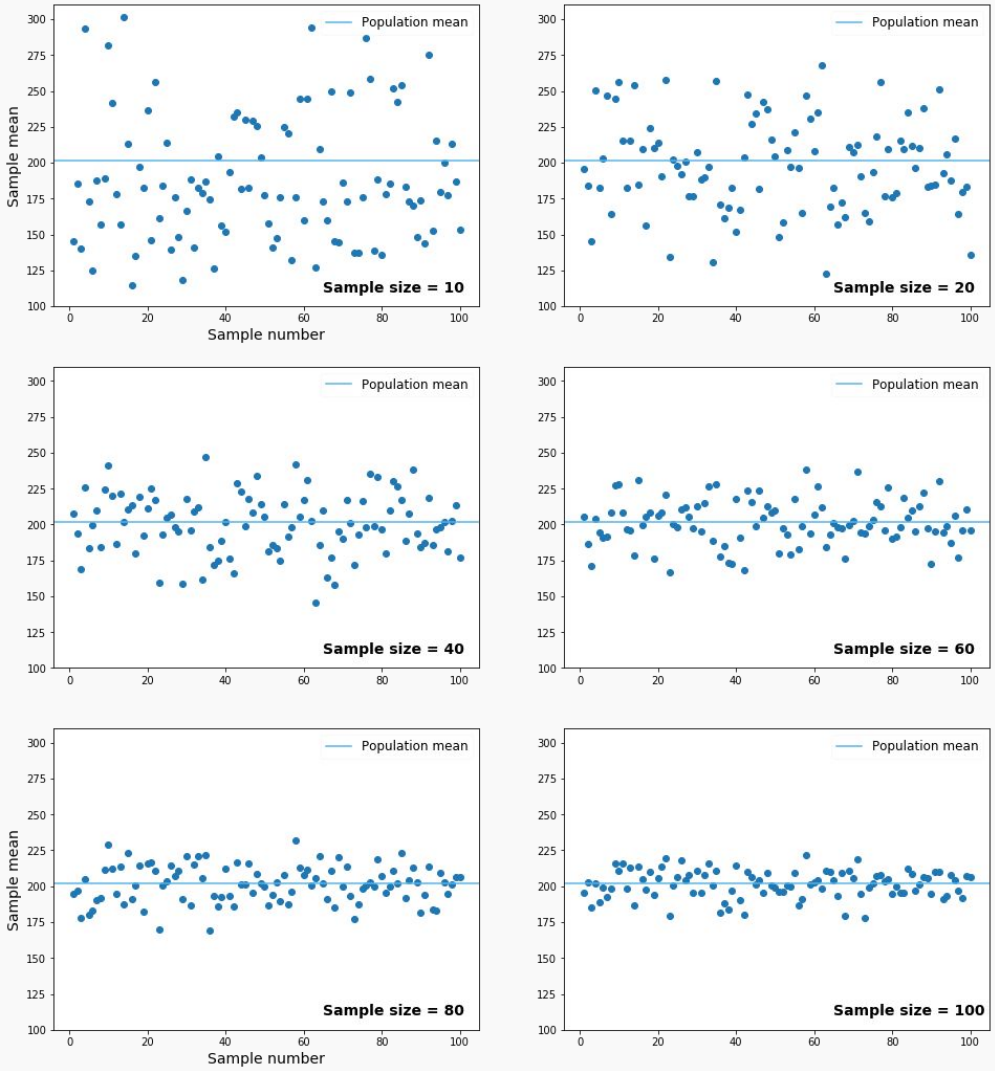
Simple Random Sampling (SRS)



Series.sample(2, random_state = 1)

Discrepancy between parameter and statistics

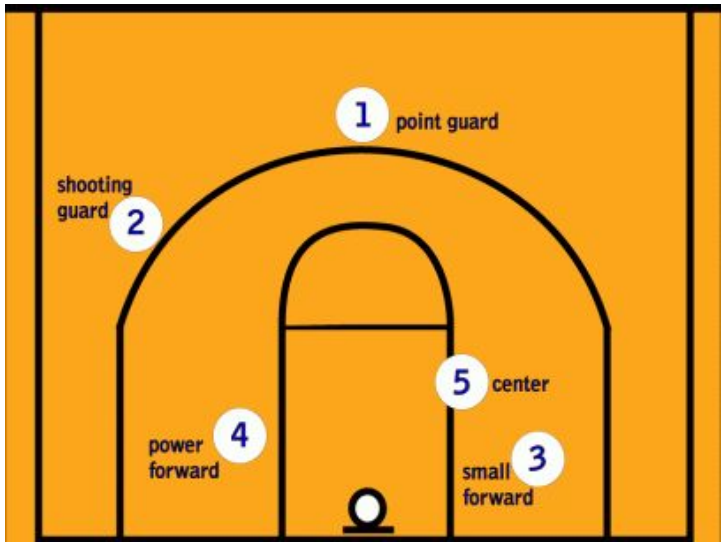




The Importance of Sample Size

Simple random sampling is not a reliable sampling method when the sample size is small.

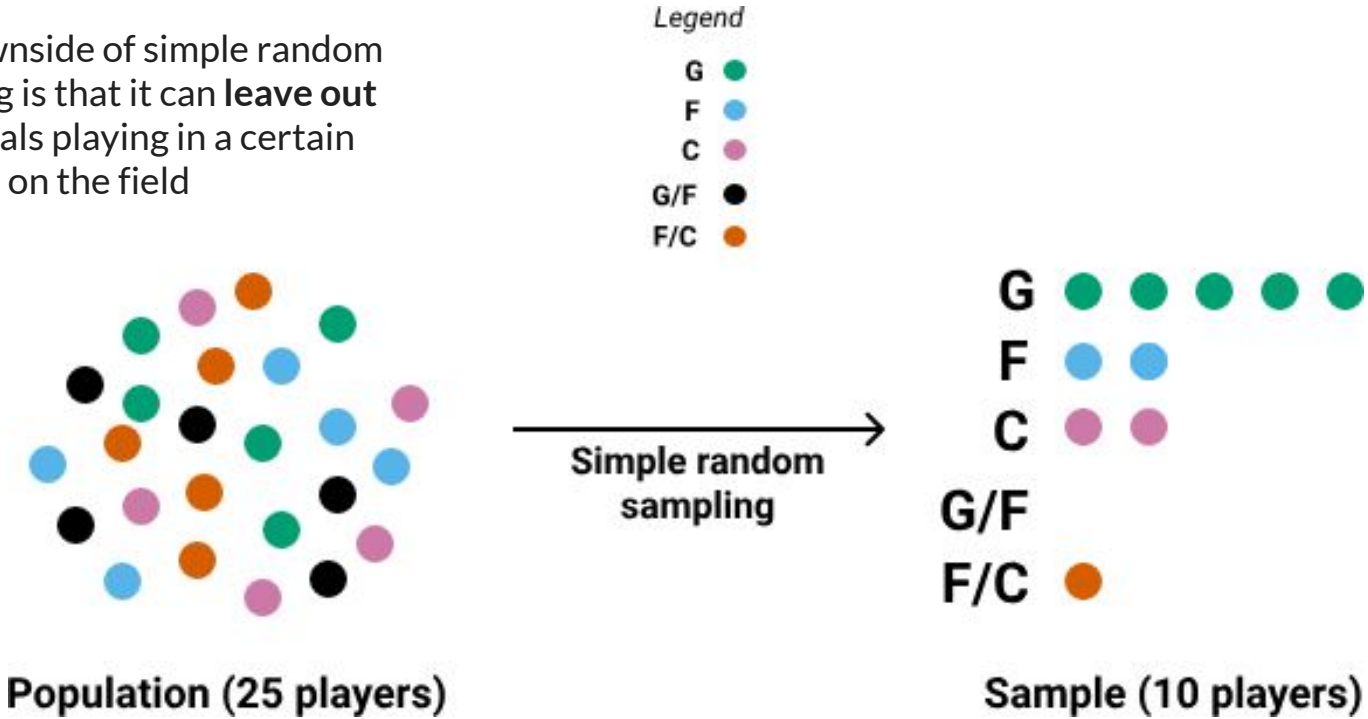
Stratified Sampling

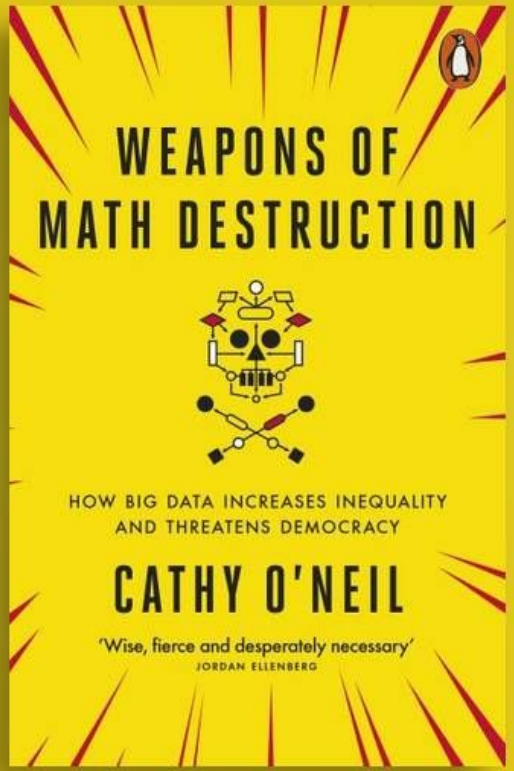


Abbreviation	Full name
F	Forward
G	Guard
C	Center
G/F	Guard/Forward
F/C	Forward/Center

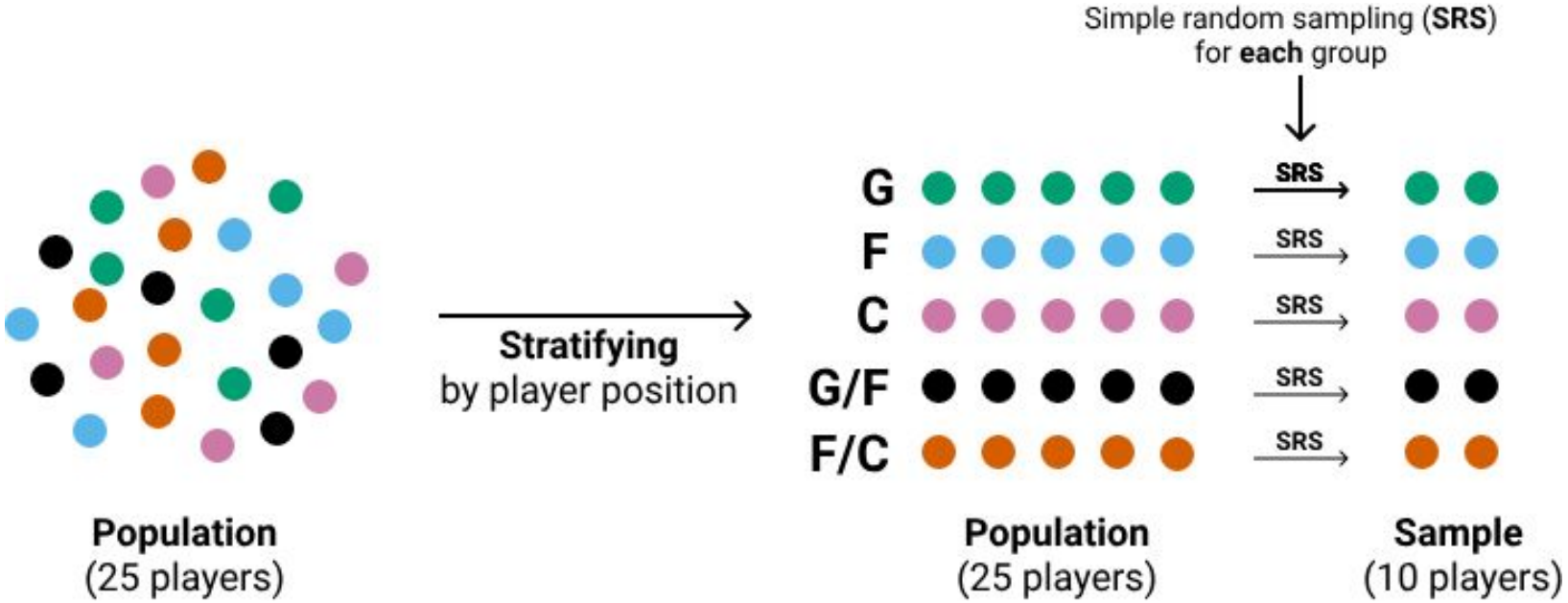
Stratified Sampling (SRS problem)

The downside of simple random sampling is that it can **leave out** individuals playing in a certain position on the field




WEAPONS OF
MATH DESTRUCTION
HOW BIG DATA INCREASES INEQUALITY
AND THREATENS DEMOCRACY
CATHY O'NEIL
'Wise, fierce and desperately necessary'
JORDAN ELLENBERG

Stratified Sampling (solution)



Proportional Stratified Sampling

```
1 wnba['Games Played'].value_counts()
```

```
29    30
30    25
28    10
27     8
25     7
23     7
22     6
26     6
31     5
24     5
14     4
20     4
18     3
4      3
7      3
17     2
16     2
15     2
10     2
21     2
32     1
5      1
19     1
8      1
9      1
12     1
2      1
```

```
1 wnba['Games Played'].value_counts(bins=3)
```

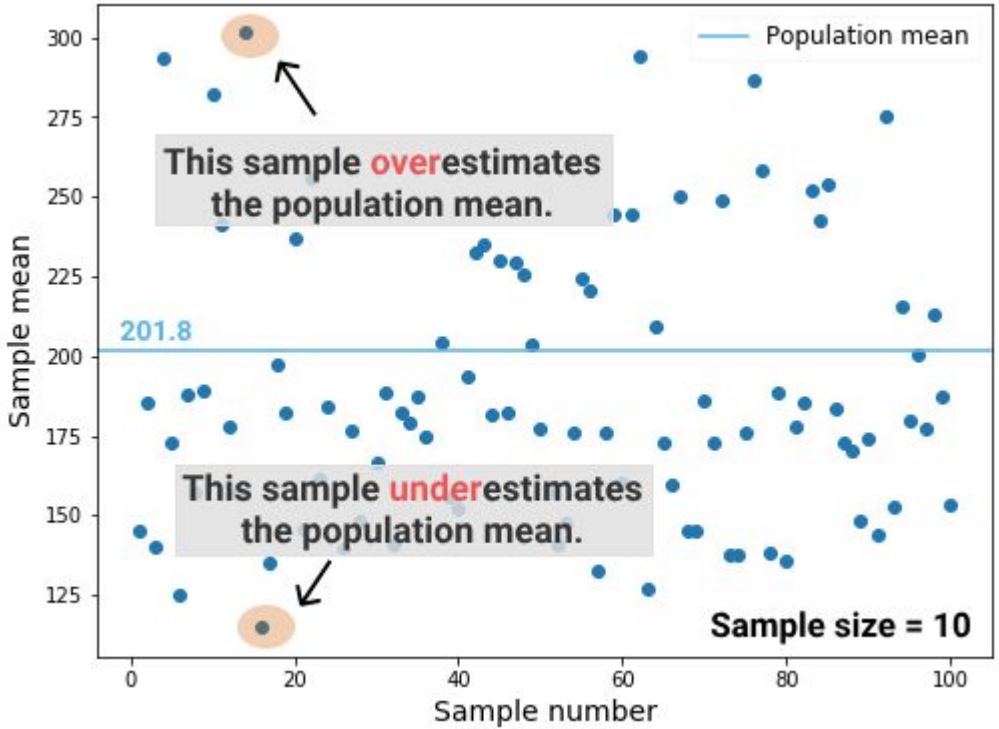
```
(22.0, 32.0]     104
(12.0, 22.0]      26
(1.969, 12.0]     13
```

```
1 wnba['Games Played'].value_counts(bins=3,
2                                   normalize=True)
```

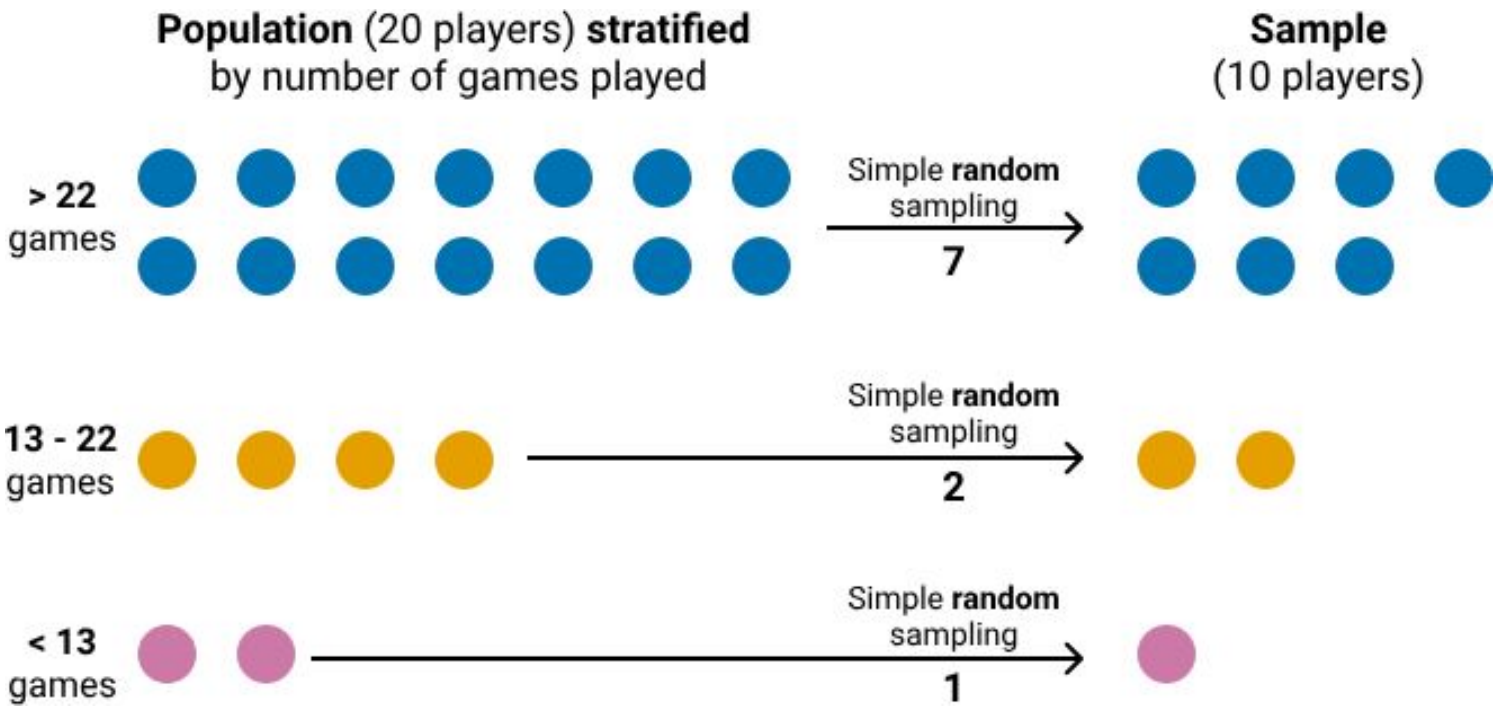
```
(22.0, 32.0]     0.727273
(12.0, 22.0]     0.181818
(1.969, 12.0]    0.090909
```

72.72% players who played more than 23 games

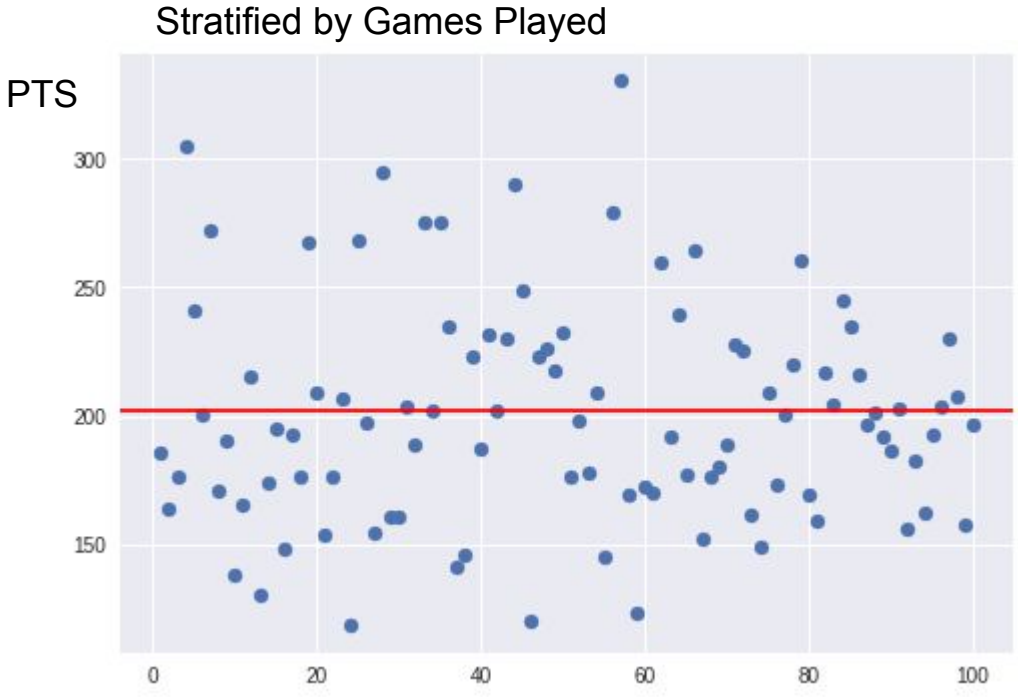
Proportional Stratified Sampling



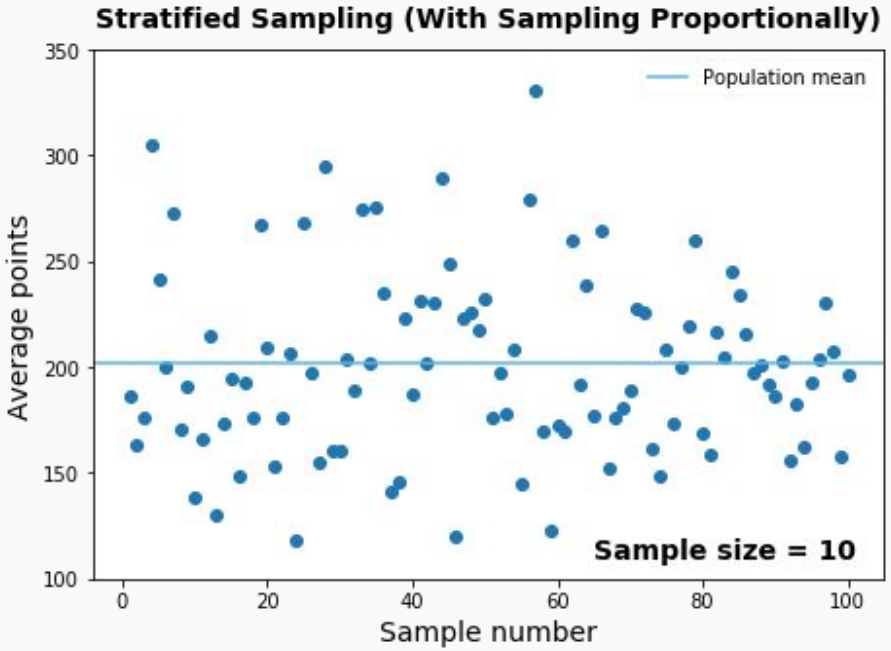
Proportional Stratified Sampling (solution)

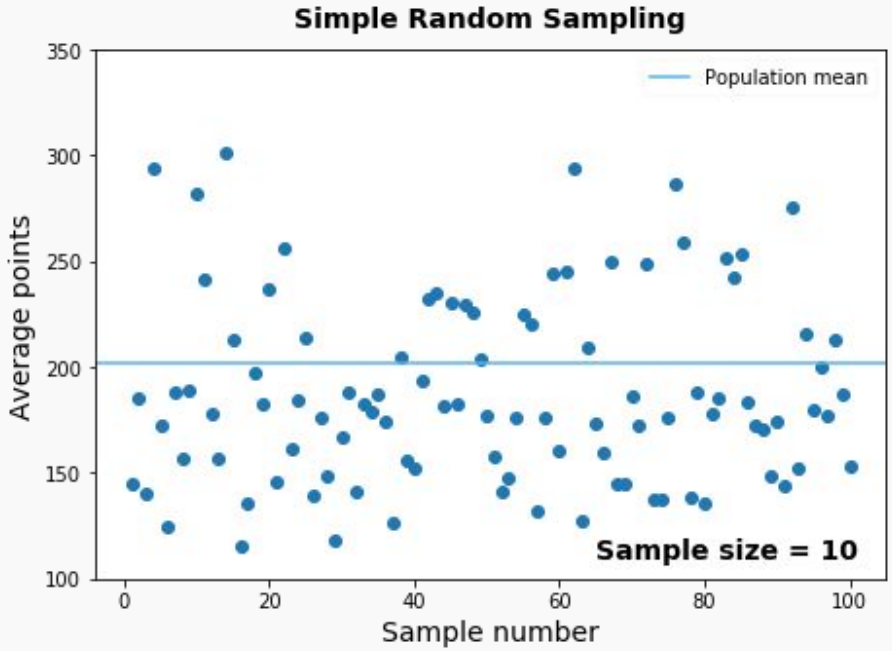


Proportional Stratified Sampling (solution)

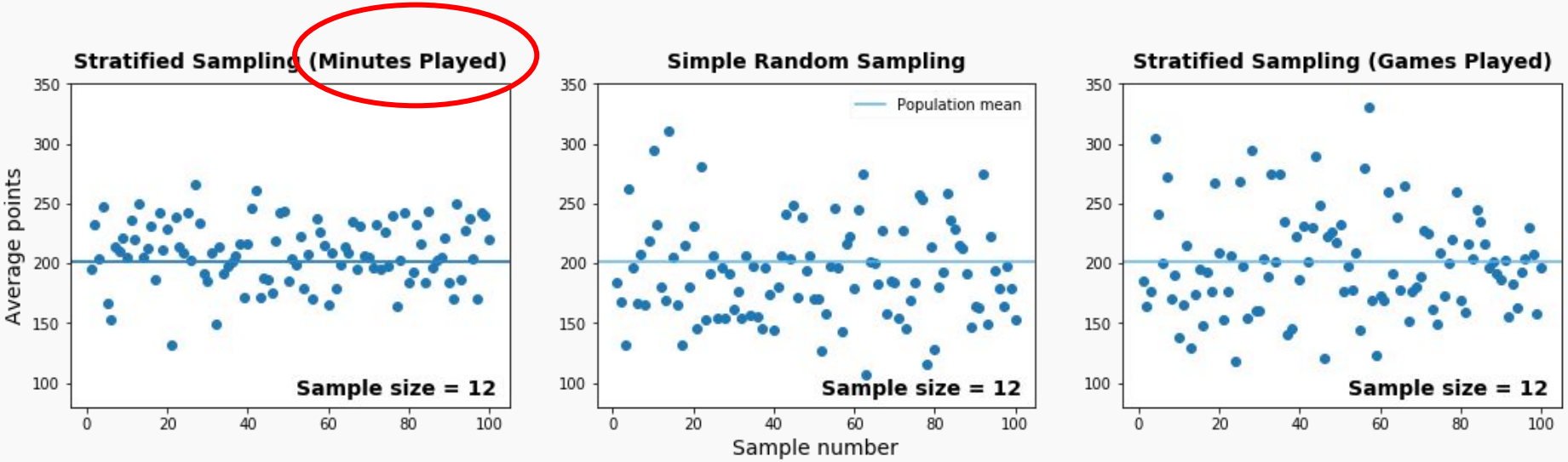


Choosing right strata



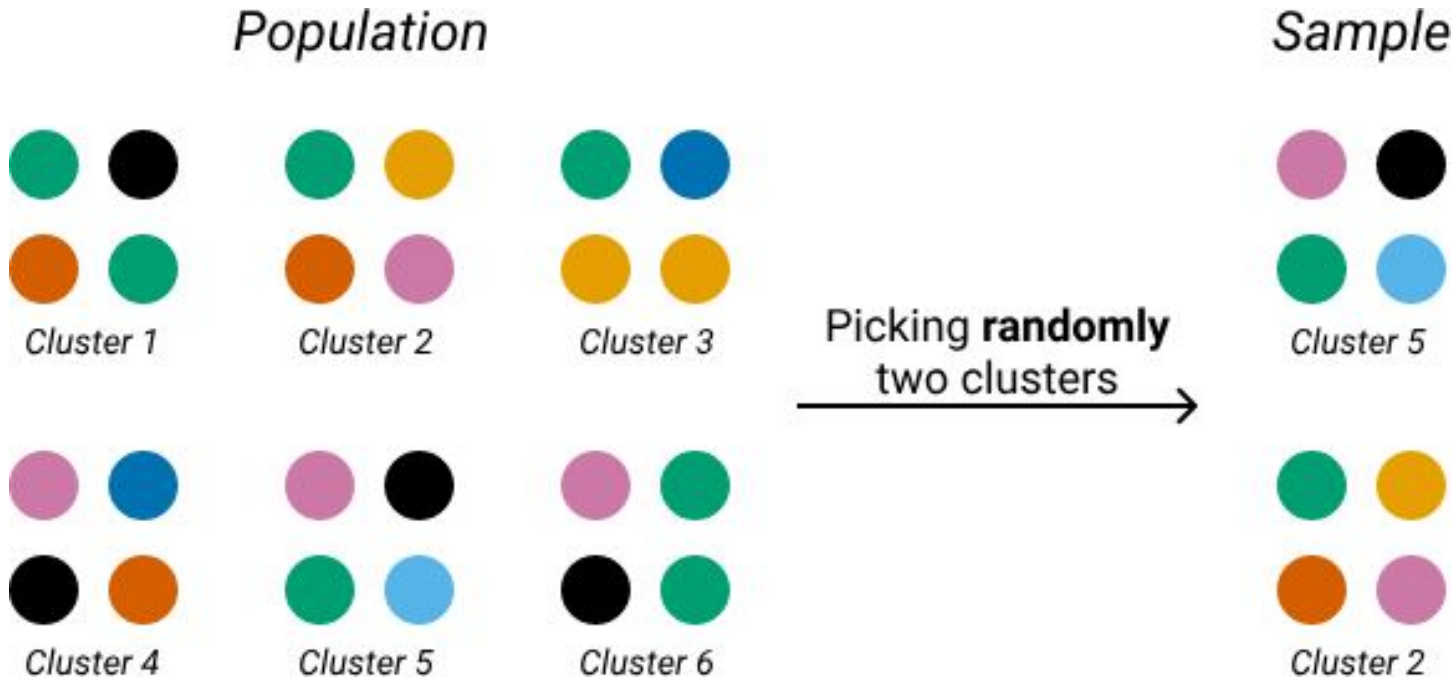


Choosing right strata



Minimize the variability within each stratum.
Maximize the variability between strata.
The stratification criterion should be strongly correlated with the property you're trying to measure.

Cluster Sampling



Sampling in Data Science

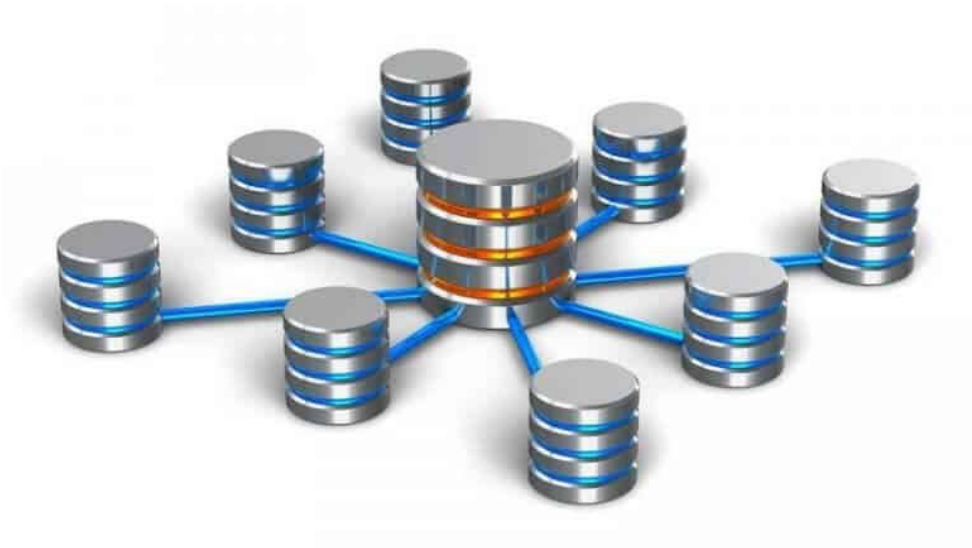
e-commerce

1. Company that has a table in a database with more than 10 million rows of online transactions.
2. The marketing team asks you to analyze the data and find categories of customers with a low buying rate, so that they can target their marketing campaigns at the right people
3. Instead of working with more than 10 million rows at each step of your analysis, you can save a lot of code running time by sampling several hundred rows

Sampling in Data Science

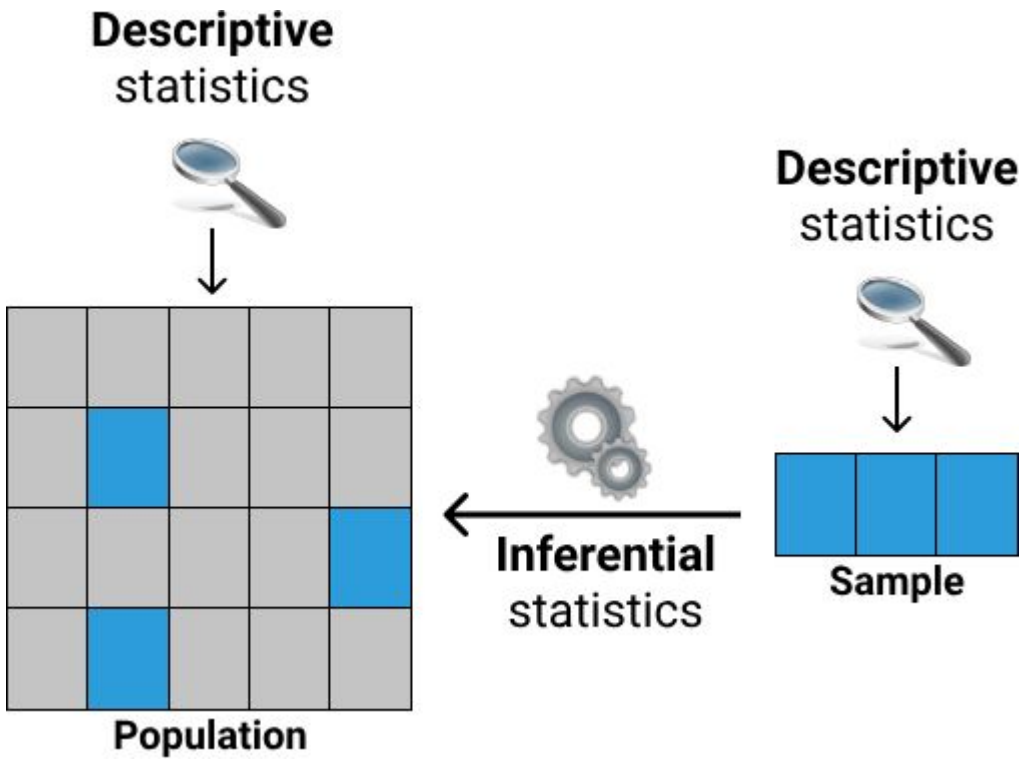
1. It could be that you need to collect data from an API that either has usage limit, or is not free.
2. In this case, you are more or less forced to sample. Knowing how and what to sample can be of great use.

Sampling in Data Science



1. Another common use case of sampling is when the data is scattered across different locations (different websites, different databases, different companies, etc.).
2. As we've discussed in the previous screen, cluster sampling would be a great choice in such a scenario.

Descriptive & Inferential Statistics



Next Steps

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

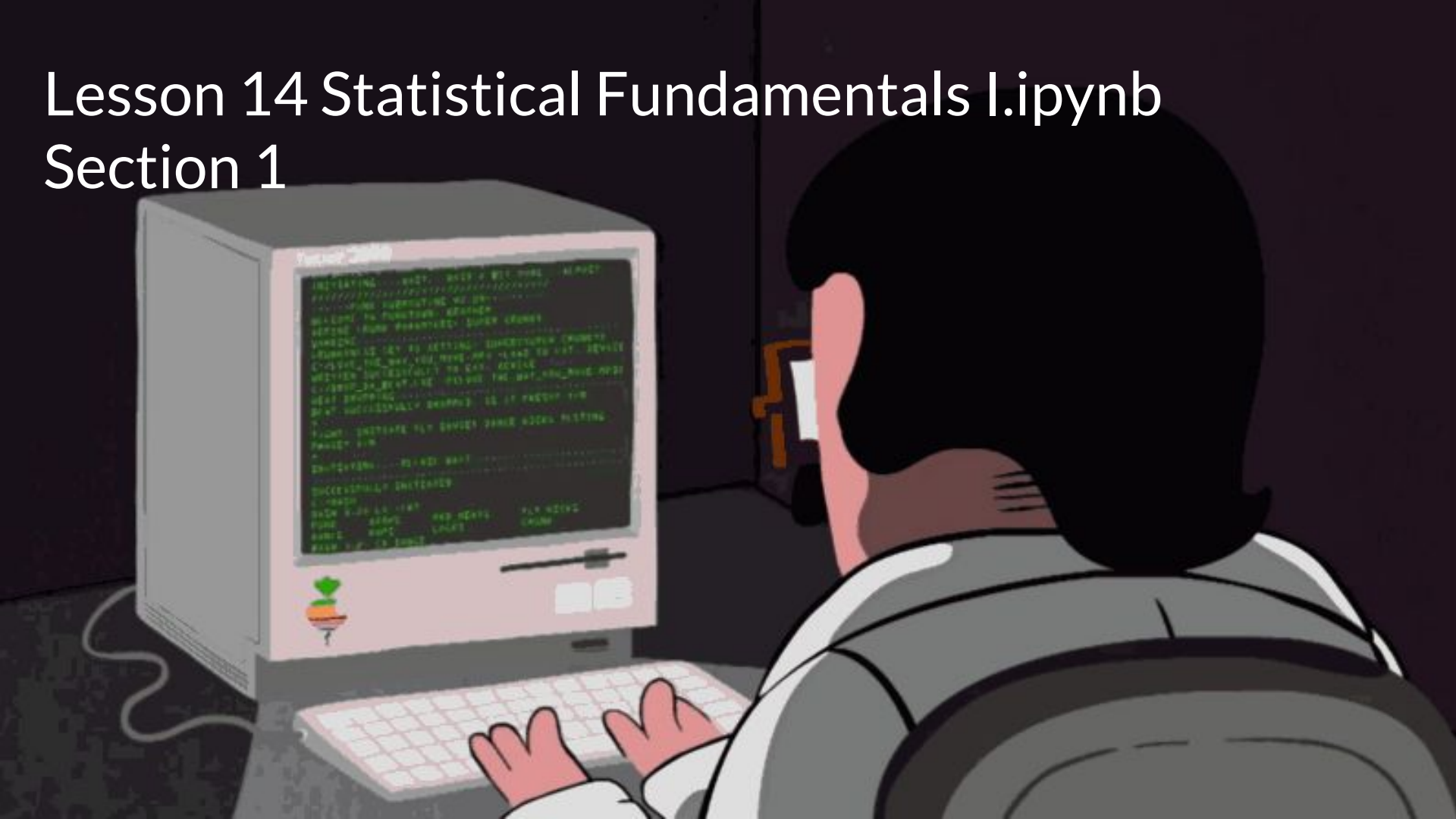
Understand how the data is **structured** and **measured**

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

Gender	Frequency
Male	147
Female	170

Visualize the patterns

Lesson 14 Statistical Fundamentals I.ipynb Section 1



Variables in Statistics

	Name	Team	Pos	Height	Weight	BMI	Birth_Place	Birthdate
39	Crystal Langhorne	SEA	F/C	188	84.0	23.766410	US	October 27, 1986
52	Érika de Souza	SAN	C	196	86.0	22.386506	BR	September 3, 1982
102	Nia Coffey	SAN	F	185	77.0	22.498174	US	May 21, 1995

The properties with varying values we call **variables**

Quantitative and Qualitative Variables

	Quantitative variables	**Qualitative** variables
Describe **quantities**	YES	NO
Describe **qualities**	NO	YES
Use **numbers**	YES	YES
The numbers are **actual quantities**	YES	NO
Use **words**	YES	YES
The words express a **quantity**	YES	NO

Height	Name
BMI	Team
Age	Pos
Birth_Data	Birth_Place
Weight	College

Scale of Measurements

The system of rules that define how each variable is measured is called **scale of measurement**

	Team	Height
We can tell whether two individuals are **different**	YES	YES
We can tell **the size** of the difference	NO	YES
We can tell the **direction** of the difference	NO	YES

- Nominal
- Ordinal
- Interval
- Ratio

Nominal Scale

	Nominal
We can tell whether two individuals are **different**	YES
We can tell the **direction** of the difference	NO
We can tell **the size** of the difference	NO
We can measure **quantitative** variables	NO
We can measure **qualitative** variables	YES

	Name	Team	Pos	Birth_Place	College
0	Aerial Powers	DAL	F	US	Michigan State
1	Alana Beard	LA	G/F	US	Duke
2	Alex Bentley	CON	G	US	Penn State
3	Alex Montgomery	SAN	G/F	US	Georgia Tech
4	Alexis Jones	MIN	G	US	Baylor

Ordinal Scale (ranking)

	Nominal	Ordinal
We can tell whether two individuals are **different**	YES	YES
We can tell the **direction** of the difference	NO	YES
We can tell **the size** of the difference	NO	NO
We can measure **quantitative** variables	NO	YES
We can measure **qualitative** variables	YES	NO

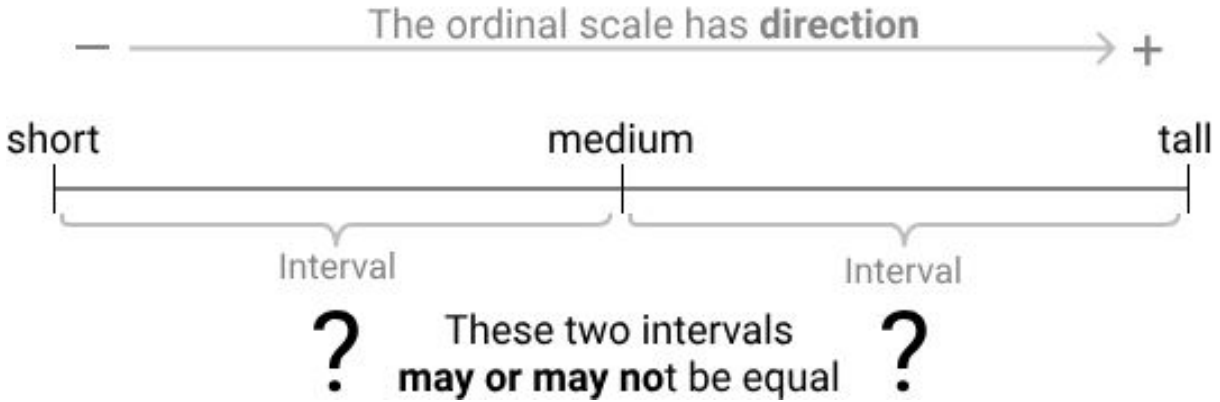
	Height	Height_labels
0	183	tall
1	185	tall
2	170	short
3	185	tall
4	175	medium

Nominal or Ordinal??

	Height_labels	College	Games Played	Experience
0	tall	Michigan State	8	2
1	tall	Duke	30	12
2	short	Penn State	26	4
3	tall	Georgia Tech	31	6
4	medium	Baylor	24	R

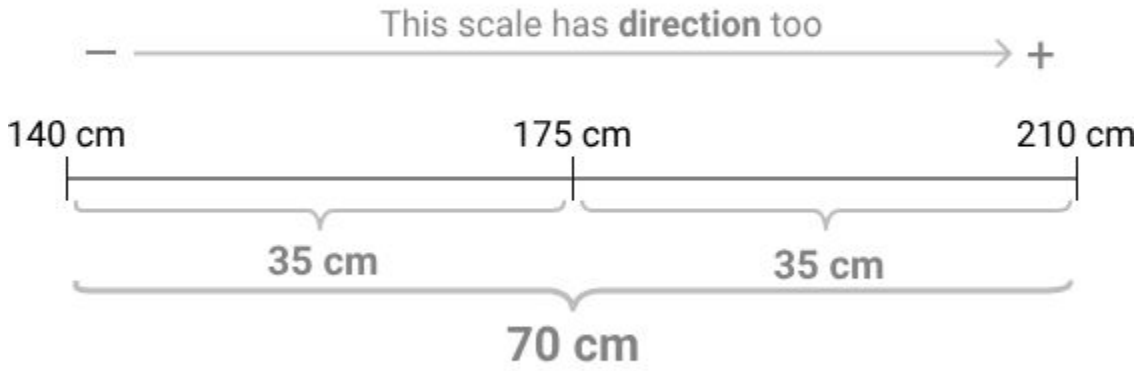
The interval and ratio scales

The height variable measured
on an **ordinal scale**



The interval and ratio scales

The height variable measured
on a scale that uses **real numbers**



We know the value of each interval, which means **we can compute the size of the difference** between any two points.

	Nominal	Ordinal	Interval	Ratio
We can tell whether two individuals are **different**	YES	YES	YES	YES
We can tell the **direction** of the difference	NO	YES	YES	YES
We can tell **the size** of the difference	NO	NO	YES	YES
We can measure **quantitative** variables	NO	YES	YES	YES
We can measure **qualitative** variables	YES	NO	NO	NO

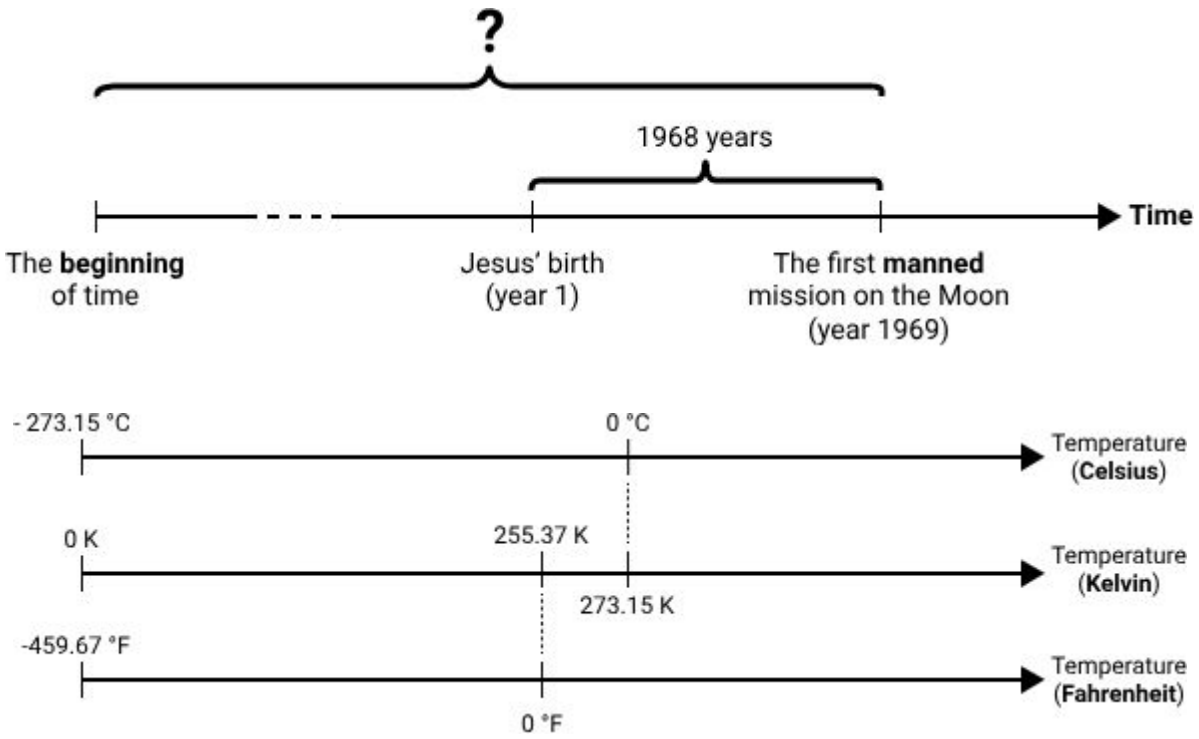
The difference between ratio and interval scales

What sets apart ratio scales from interval scales is the nature of the zero point.

_	Name	Weight	Weight_deviation
35	Clarissa dos Santos	89.0	10.021127
3	Alex Montgomery	84.0	5.021127
111	Renee Montgomery	63.0	-15.978873
85	Layshia Clarendon	64.0	-14.978873
128	Sugar Rodgers	75.0	-3.978873

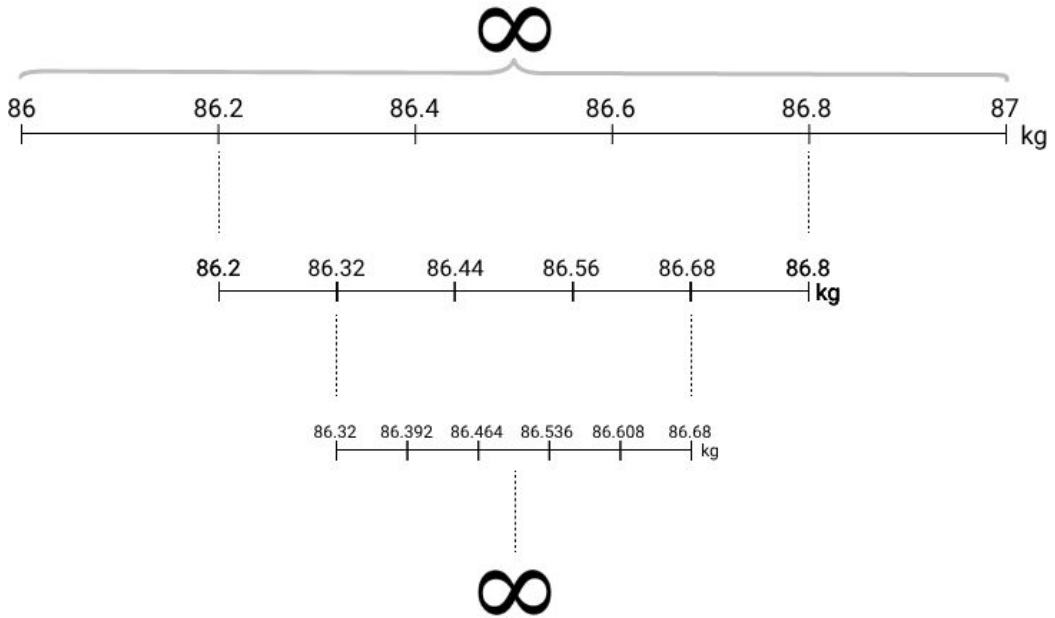
	Interval	Ratio
Well-defined **intervals**	YES	YES
The **zero point** indicates the **absence of a quantity**	NO	YES
Difference measured in terms of **distance**	YES	YES
Difference measured in terms of **ratios**	NO	YES

Common Examples of Interval Scales



Discrete and Continuous Variable

_	Name	Weight	PTS
77	Kayla Thornton	86.0	32
16	Asia Taylor	76.0	31
80	Kia Vaughn	90.0	134
137	Tierra Ruffin-Pratt	83.0	225
12	Amanda Zahui B.	113.0	51



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**

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

Gender	Frequency
Male	147
Female	170

Visualize the patterns

Lesson 14 Statistical Fundamentals I.ipynb Section 2