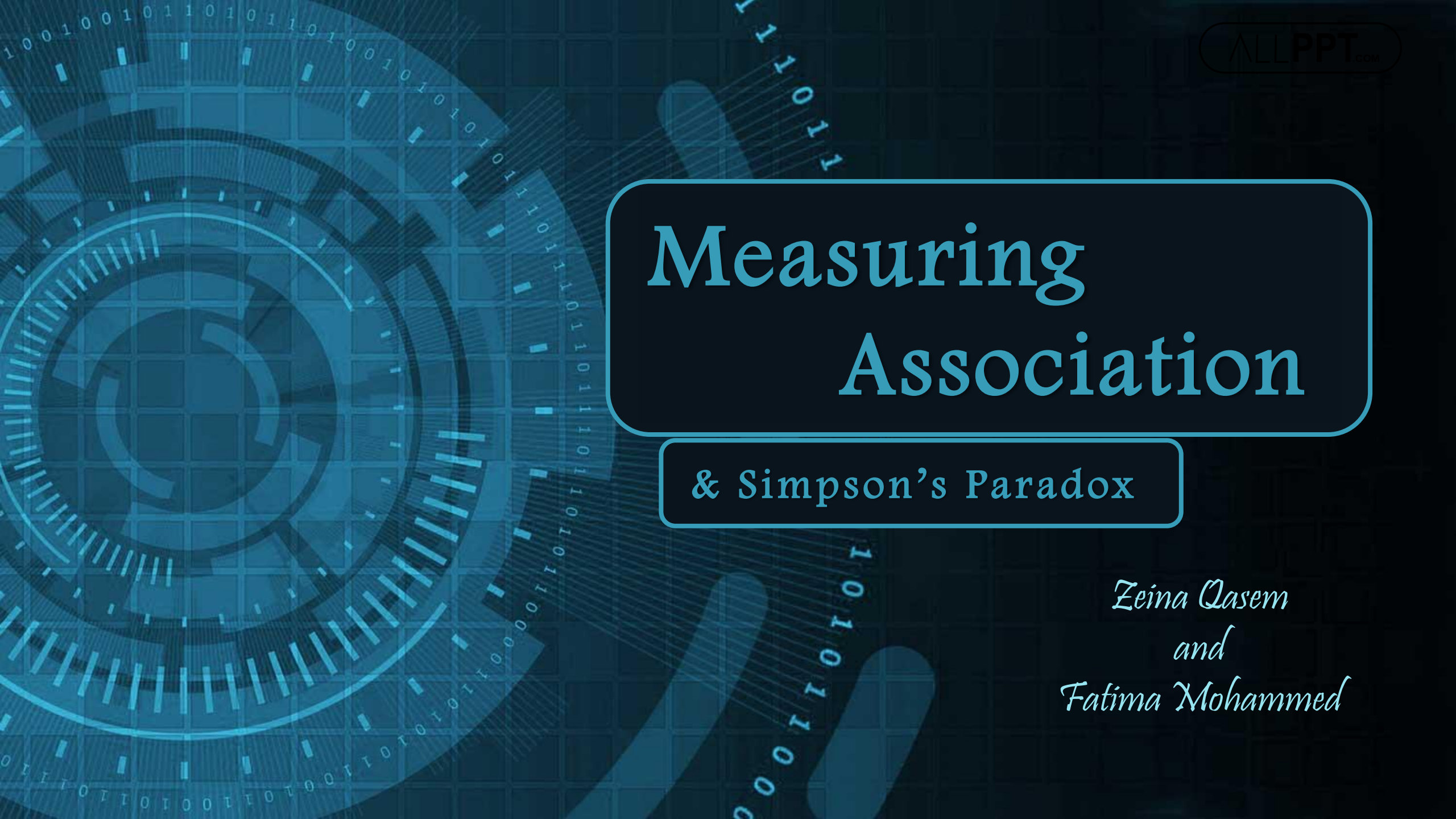


Measuring Association

& Simpson's Paradox

Zeina Qasem
and
Fatima Mohammed

Intro To Bivariate Data

The set of data where two variables are involved is referred to as bivariate data.
Analysis of bivariate data allows to study and understand the relation between the 2 variables (correlation).
The two variables could be in a linear, nonlinear relationship, monotonic or not related at all.

Nonlinear Correlation

Nonlinear correlation occurs when the two variables involved are in a relation that is not linear. This happens when a change in one variable does not lead to a proportional change in the other.

Linear Correlation

Linear correlation occurs when the two variables involved are in a relation that is linear, i.e. a change in one variable leads to a proportional change in the other.

Forms of Nonlinear Correlation

- Quadratic correlation
- Exponential correlation
- Logarithmic correlation
- Power-law correlation
- Sigmoid correlation

- **To study and better understand correlations:**
  - Spearman's rank correlation coefficient
  - Kendall's Tau
  - Goodman & Kruskal's Gamma
  - PP Score
  - Phi Correlation Coefficient
  - Karl Pearson's correlation coefficient
  - Pearson's Chi-Squared Test
  - Chatterjee Correlation Coefficient

Types of coefficients and their domain of function :

Pearson Correlation Coefficient :

- Linear data
- Parametric
- Continuous, categorical

Spearman correlation coefficient :

- linear (monotonic)
- Non – parametric measure
- Any data type, ordinal

Kendall Tau's Correlation Coefficient :

- Linear data (monotonic association)
- Non – parametric
- Ordinal

PP Score :

- Linear, non-linear
- Non-parametric
- Continuous, nominal, categorical

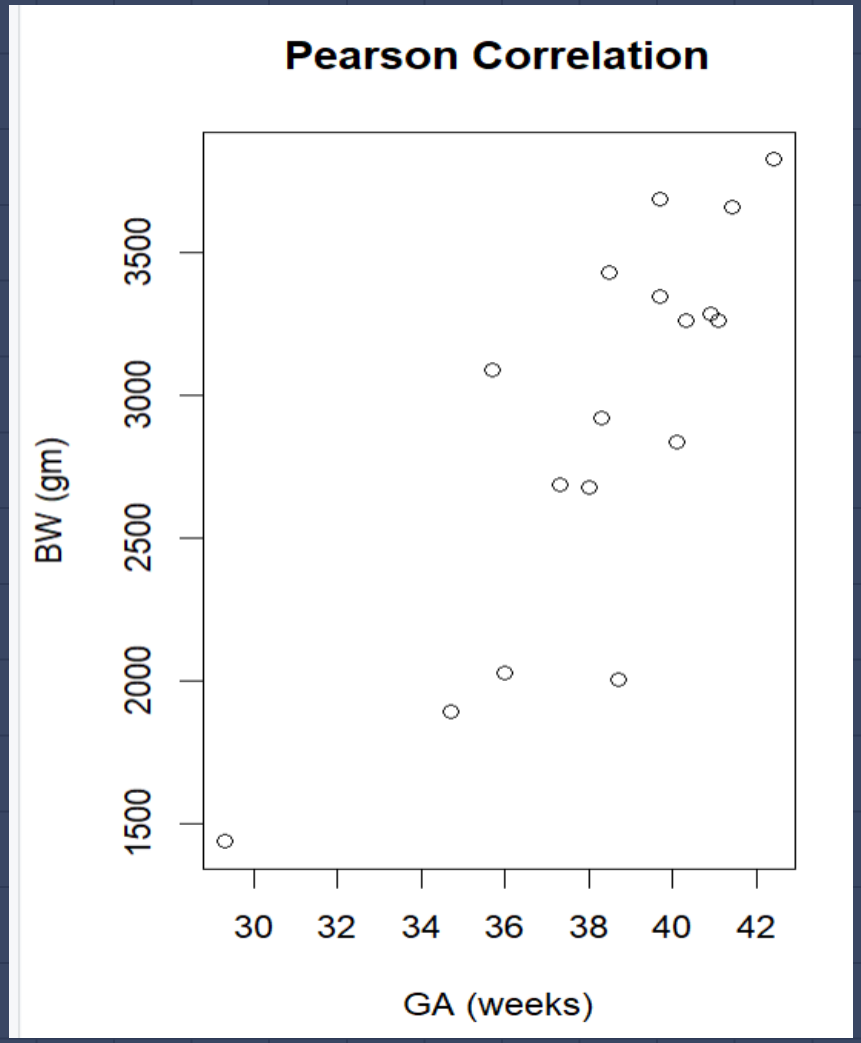
Chatterjee Correlation Coefficient :

- Linear, non-linear
- Non- parametric
- Categorical and continuous

1. Karl Pearson Correlation Coefficient:

Uses :

- Measures strength of linear relation between 2 variables
- Relation between 2 continuous variables
- Assumes data to be normally distributed for both variables
- Works best with continuous data
- Widely used in many fields..



Pearson coefficient = 0.82

Karl Pearson Correlation Coefficient

Advantages :

- Measuring linear relationships is very common in research
- Simple and interpretable
- Is scale – invariant

Disadvantages :

- Sensitive to outliers
- Inappropriate for non-normally distributed data
- Parametric
- Limited to bivariate analysis

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where:
n is the sample size
xi and yi represent sample points
$\bar{x}$ and $\bar{y}$ are mean values of the whole products

2. Spearman's Correlation Coefficient :

- Non – parametric
- Ranks the values of each variable, and calculates the difference between the ranks
- Ranges from -1 to 1
- Sensitive to outliers (more robust that Pearson)
- Large sample sizes

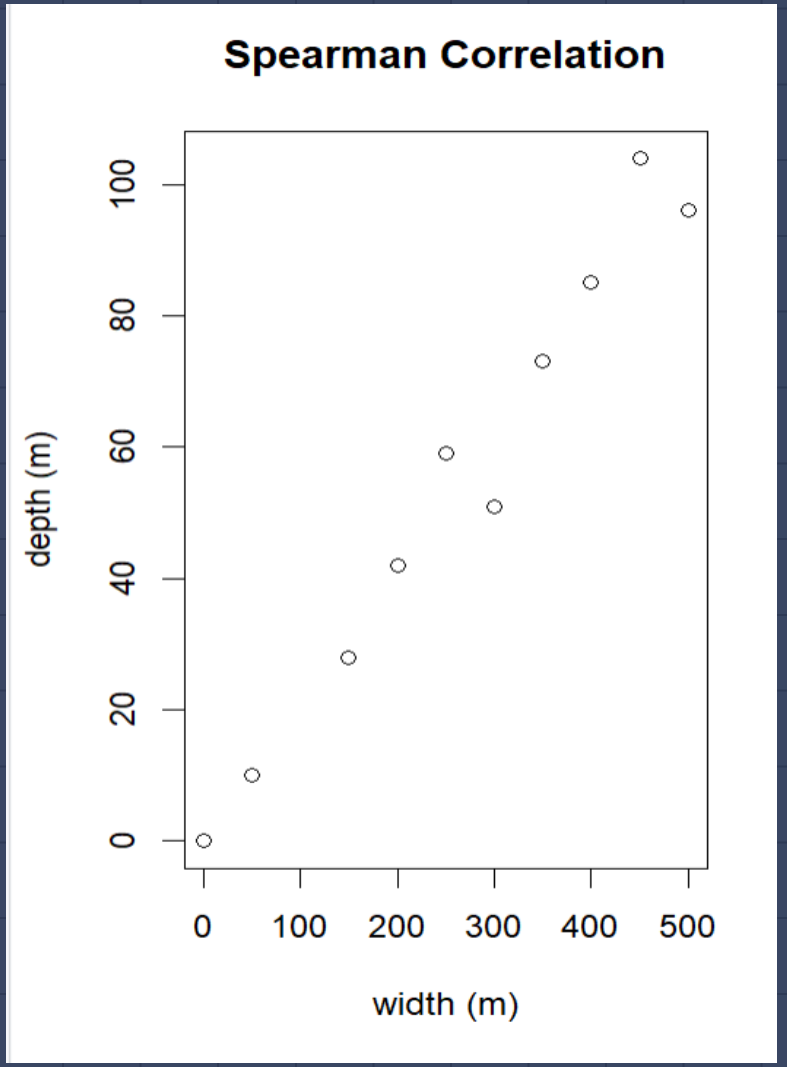
$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where:

ρ is Spearman's rank correlation coefficient

d_i is the difference between the two ranks of each observation

n is the number of observations



Spearman coefficient = 0.97598

Spearman's Correlation Coefficient

Uses:

- To determine degree of monotonic correlation between two variables in an ordinal dataset
- Often to prove or disprove a hypothesis
- Works with any type of data

Advantages:

- Works with continuous and ordinal datasets
- Does not require data to be normally distributed
- Works with ordinal/non-normally distributed data
- Useful in exploratory data analysis

Disadvantages:

- A monotonic relationship is assumed between the variables
- Assigning ranks to a large number of numerical values is a time-consuming and exhausting process
- Cannot detect nonlinear relations
- As it is based on rank, information can be lost (especially when with ties)
- Sensitive to (affected by) outliers
- Cannot establish causation

3. Kendall's Tau Correlation Coefficient:

- Based on number of concordant and discordant pairs of observations between the variables
- Ranges from -1 to 1
- Not as sensitive to outliers as Spearman's
- Small sample sizes

where:

n_c is the number of concordant pairs

n_d is the number of discordant pairs

$$n_0 = \frac{n(n-1)}{2}$$

m is the min (r, c)

$$\tau_A = \frac{n_c - n_d}{n_0}$$

Used with no ties

$$\tau_B = \frac{n_c - n_d}{\sqrt{(n_0 - n_c)(n_0 - n_d)}}$$

Used when some ties

$$\tau_C = \frac{(n_c - n_d)}{n^2 \frac{(m-1)}{m}}$$

Used when many ties

Kendall's Rank/Tau Correlation Coefficient

Uses:

- To determine the strength of correlation between two variables in an ordinal dataset
- To know the direction of the relationship between two variables
- The variables are continuous with outliers or ordinal
- Only when two variables are involved
- Works with tied datasets
- Small sample size
- Alternative to Spearman and Pearson CC's

Advantages:

- Non-parametric
- Continuous and ordinal datasets
- Detects non- linear relations

(as long as they're monotonic)

- Works very well for small sample size
- Less sensitive to outliers

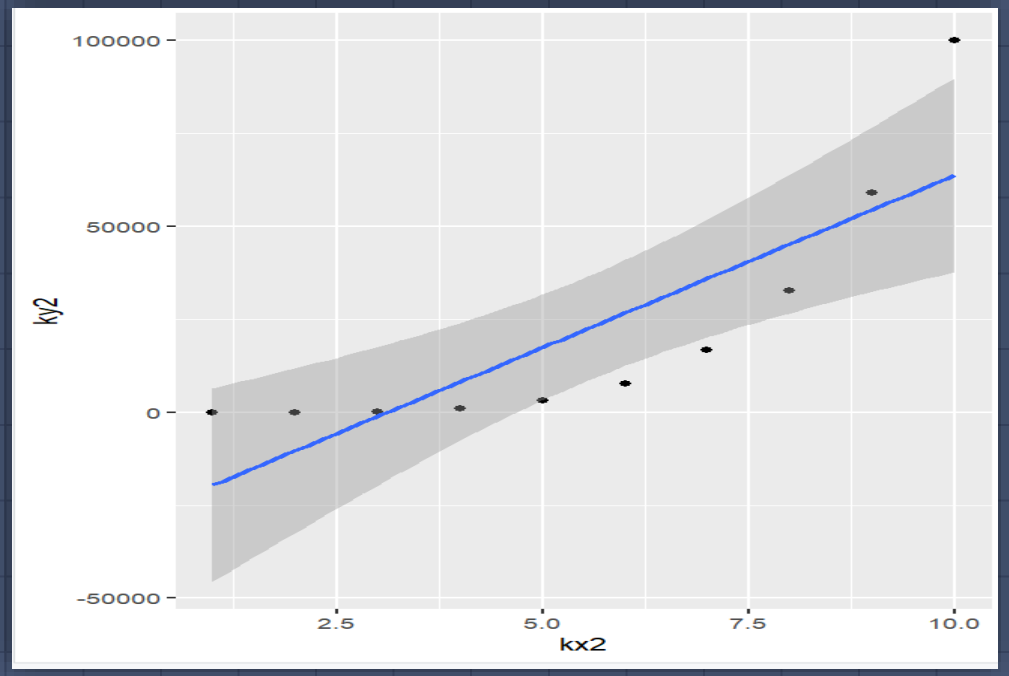
(compared to spearman)

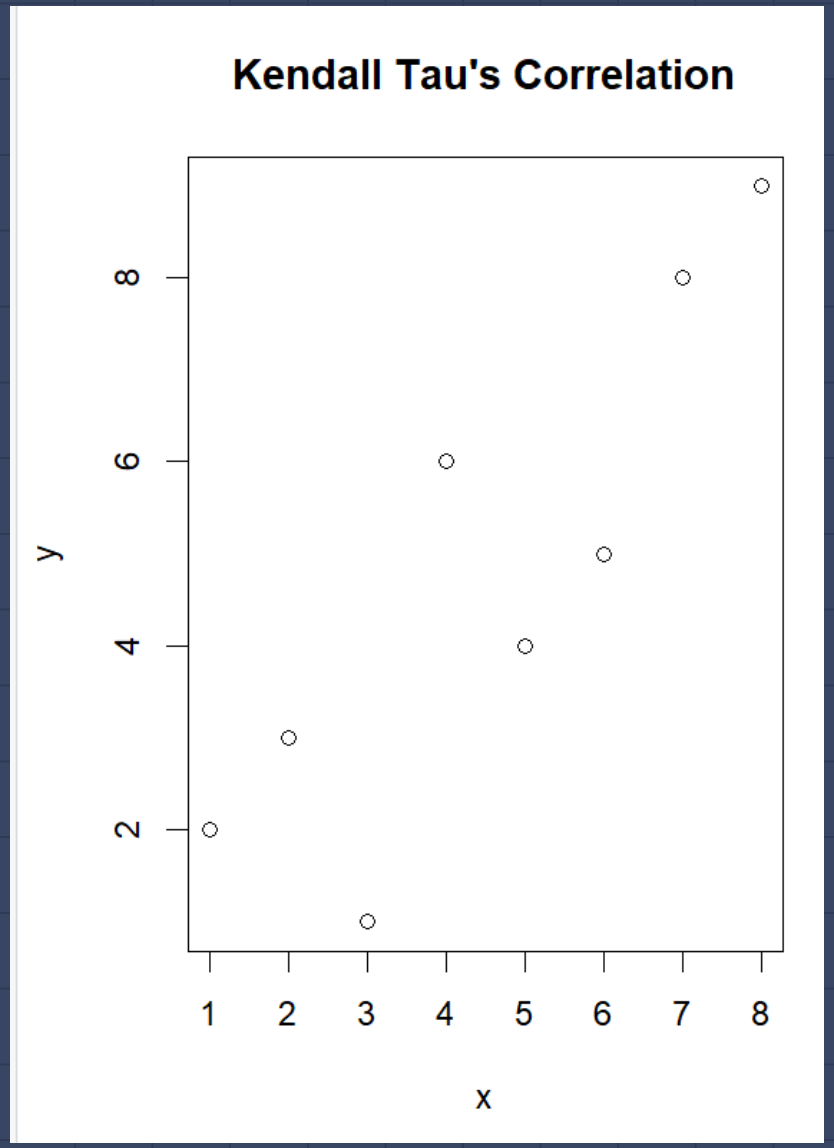
- Useful with datasets with ties
- Works with non-normally distributed datasets

Kendall's Rank/Tau Correlation Coefficient

Disadvantages:

- Unsuitable for highly skewed distributions
- Not used for grouped data
- Time-consuming
- Cannot establish causation





4. Chatterjee Correlation Coefficient (CCC):

- Measures strength of relations/degree of dependence between two variables (Xi and Yi)
- Rank-based correlation coefficient
- Can be used for both categorical and continuous variables
- Range of ξ_n is [-1,1]
- Used irrespective of the law of the variables
- Used in multivariate analysis, data visualization, machine learning, and more…

$$\xi_n(X,Y) = 1 - \frac{3\sum_{i=1}^{n-1}|r_{i+1} - r_i|}{n^2 - 1}$$

In the case of no ties

$$\xi_n(X,Y) = 1 - \frac{n\sum_{i=1}^{n-1}|r_{i+1} - r_i|}{2\sum_{i=1}^{n} l_i(n - l_i)}$$

In the case of ties

where:

r_i is the rank of Yi : the number of j such that Yj ≤ Yi

l_i is the number of j such that Yj ≥ Yi

Journal of the American Statistical Association ›
Volume 116, 2021 - Issue 536

Enter keywords, authors, DOI, ORCID etc

Submit an article | Journal homepage

5,893
Views
36
CrossRef citations to date
14
Altmetric

Theory and Methods

A New Coefficient of Correlation

Sourav Chatterjee

Pages 2009-2022 | Received 15 Oct 2019, Accepted 15 Mar 2020, Accepted author version posted online: 27 Apr 2020, Published online: 28 May 2020

Download citation | https://doi.org/10.1080/01621459.2020.1758115 | Check for updates

An article about the new coefficient of correlation, published by the 'Journal of the American Statistical Association'

Chatterjee Correlation Coefficient

- Consistently estimates a quantity that is 0 iff the variables are independent and 1 iff one is a measurable function of the other
  - Law of X and Y can be discrete or continuous

Disadvantages:

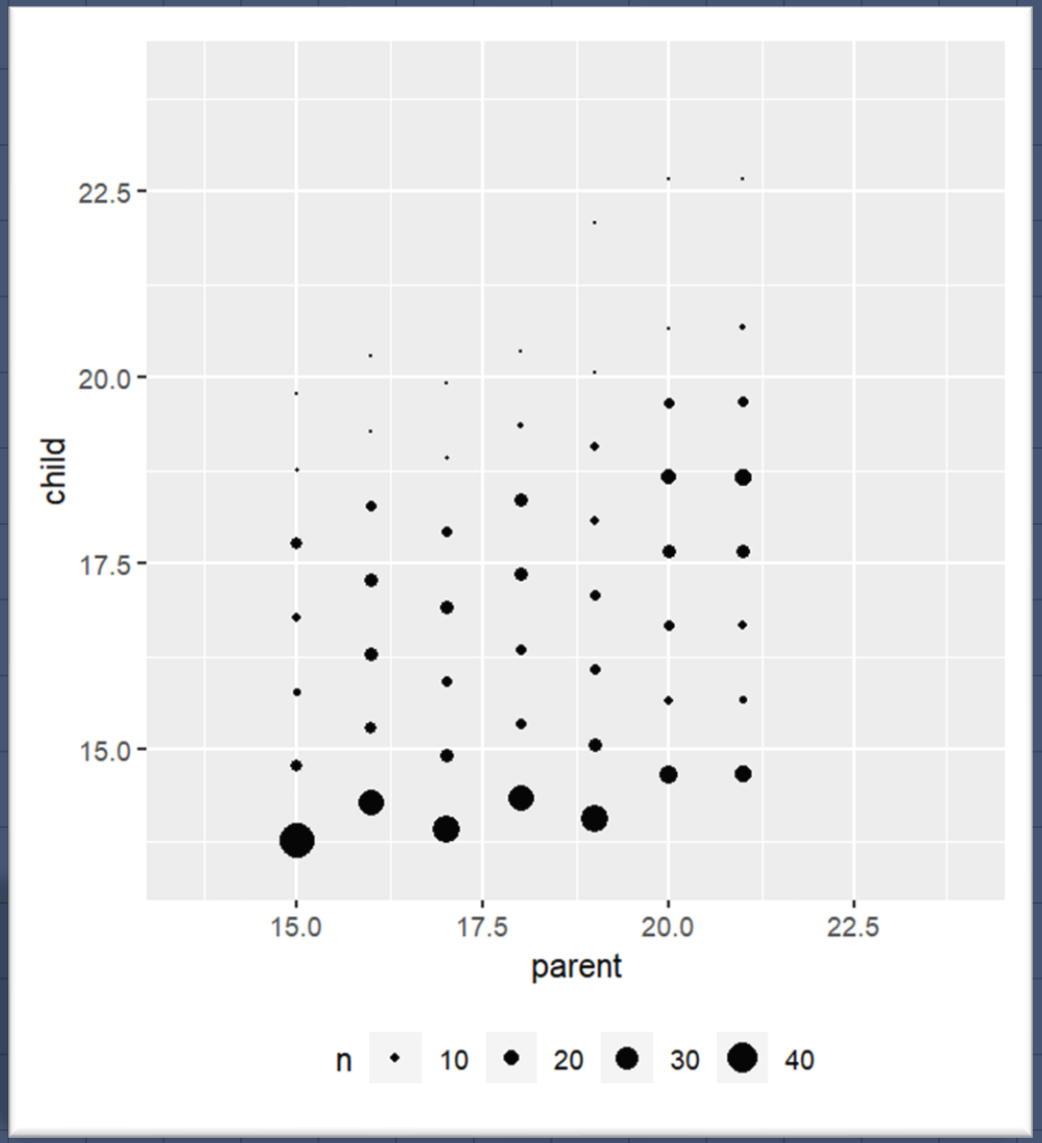
- Less widely known and used
- Requires paired data
- Not symmetric, $\xi_n(X,Y) \neq \xi_n(Y,X)$

Chatterjee Correlation Coefficient

Advantages:

- An easy to understand and simple to use formula
- No need for assumptions for the variables' distributions
- Robust to outliers
- Non - parametric
- Works well with non-linear relationships (as well as linear)
- Useful for small sample sizes (limited/rare data)
- No exceptions on the properties (easy to use)
- A robust alternative to other correlation coefficients

Chatterjee Correlation Coefficient



CCC : 0.9225

```
> xicor(peas$child,peas$parent)
[1] 0.9225
```

https://cran.r-project.org/web/packages/XICOR/XICOR.pdf

5. Predictive Power Score (PPS) :

Uses :

- detec
  relati
- Rang
- Calcu
  over
- Num
  varia
- Used

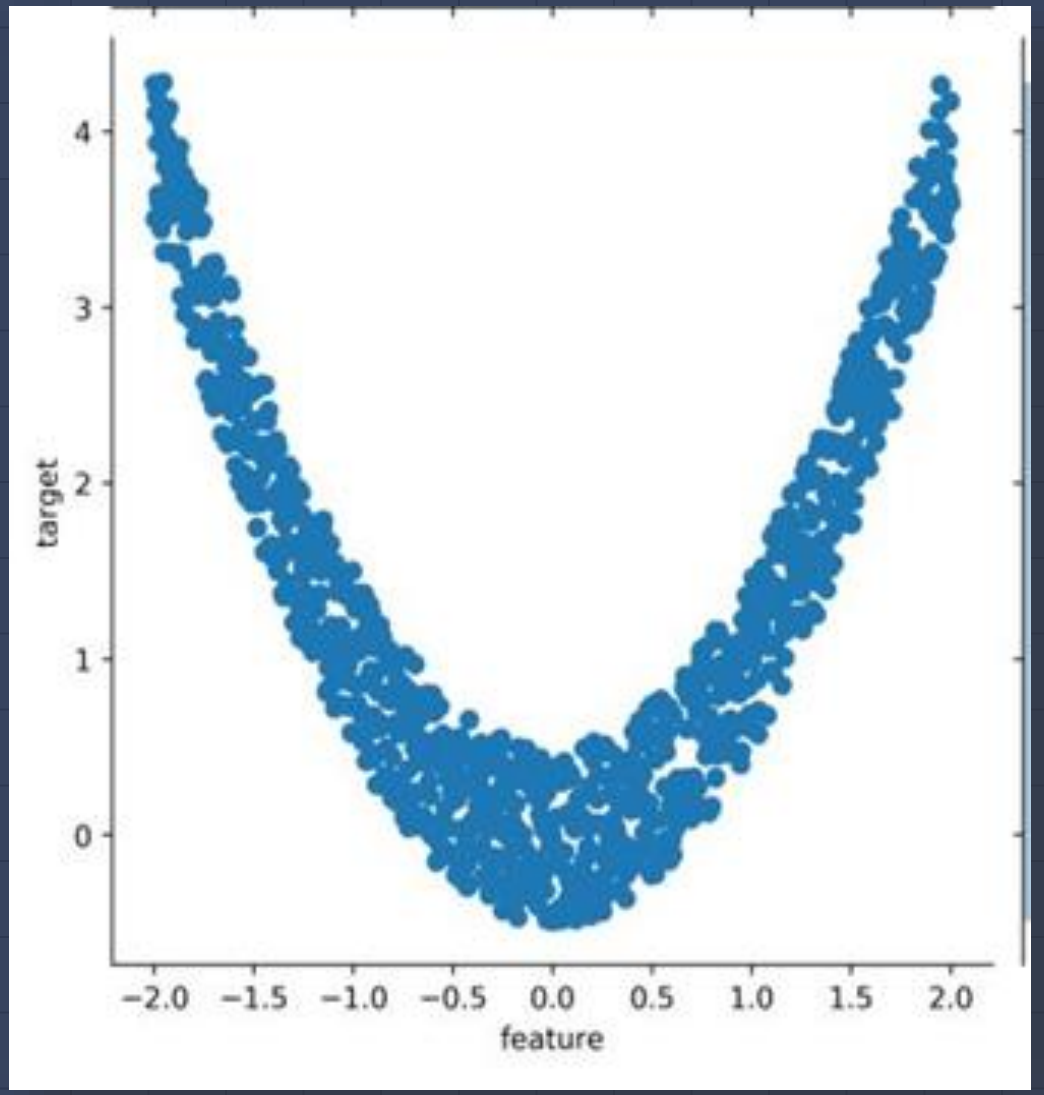
bles

l

$$y(x) = x^2 + \varepsilon,$$
$$x \sim \mathcal{U}(-2, 2),$$
$$\varepsilon \sim \mathcal{U}(-0.5, 0.5)$$

Correlation: 0

PPS x to y: 0.67
PPS y to x: 0



5. Predictive Power Score (PPS) :

Advantages :

- Non – parametric
- Flexible and works with wide range of data and data types
- Simple, easy to interpret

PPS = (F1_model – F1_naive) / (1 – F1_naive)

where:
F1_model is the classification score for the regression model

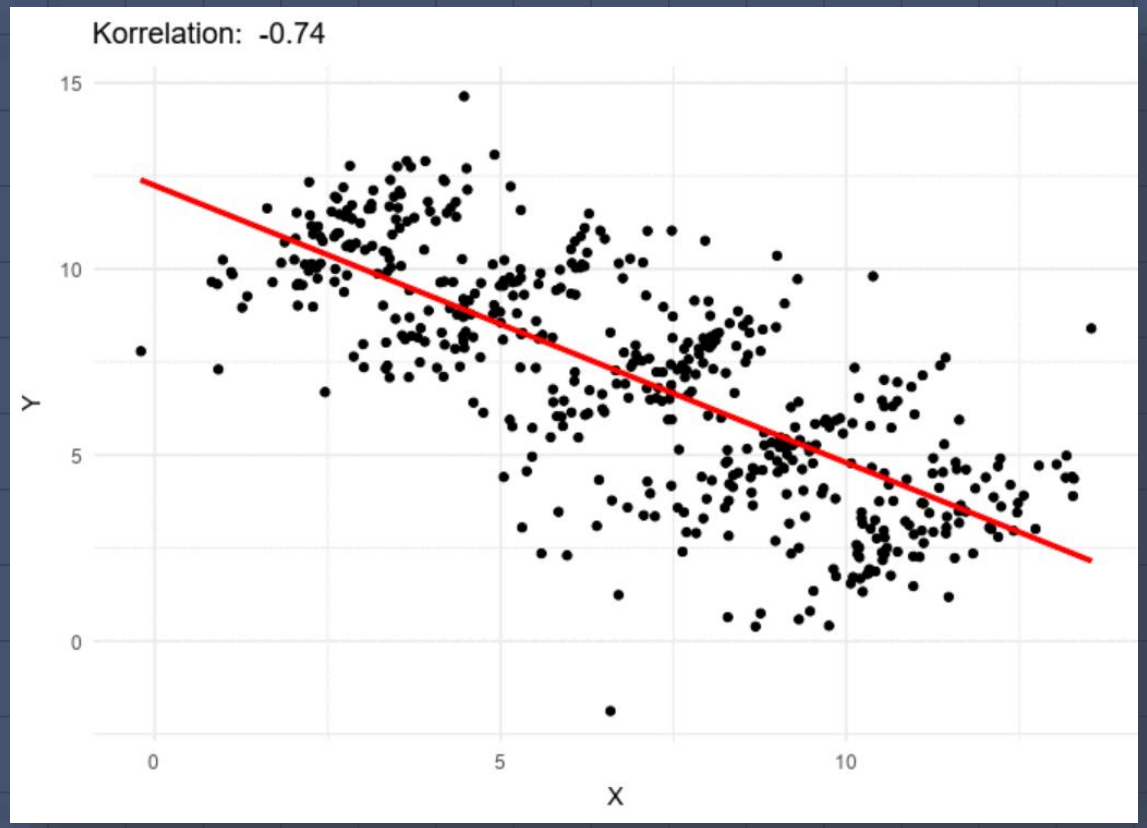
Disadvantages:

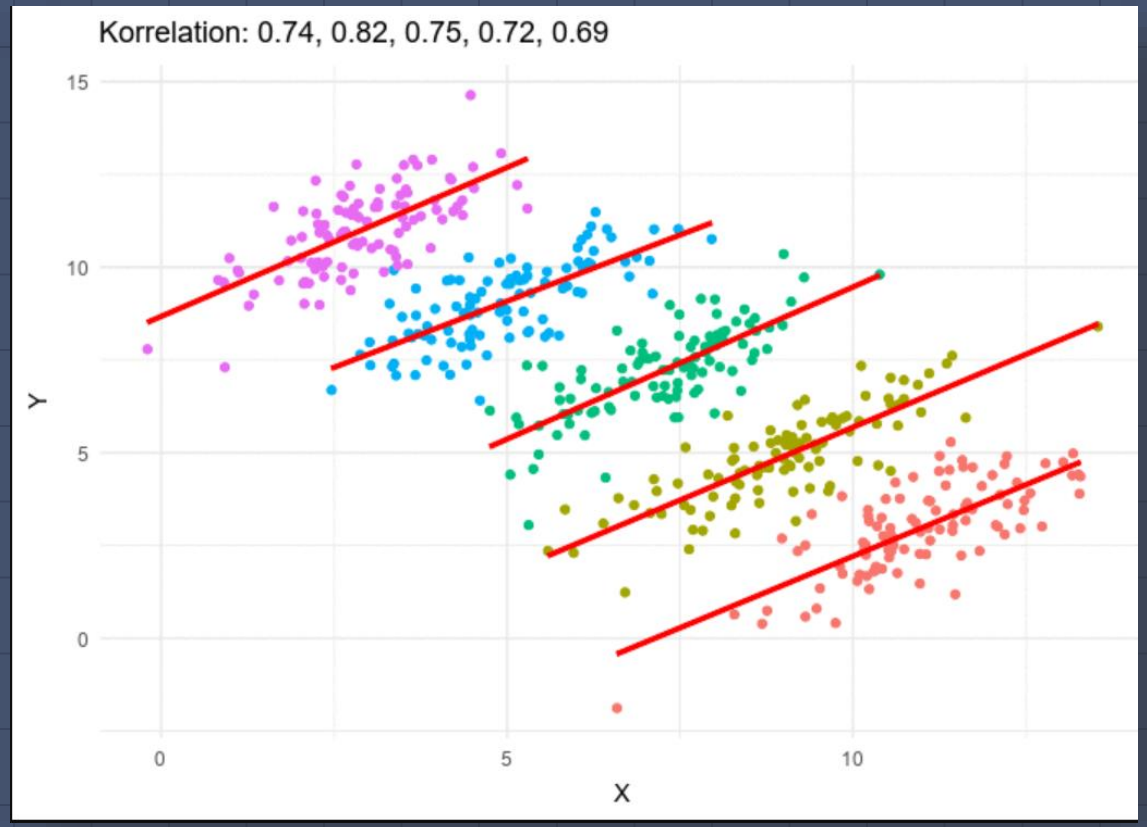
• Predictive Power Score can have different algorithms or methods for producing the output, making the results depend on several factors ..

• May reflect different types of relationships on a single score, which may lead to misleading and quite complex patterns

Simpson's Paradox

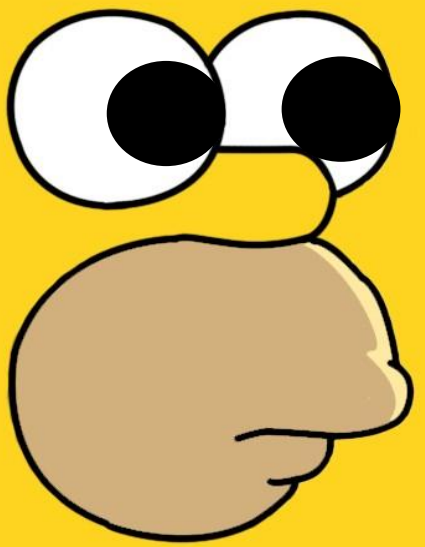
- Statistical phenomenon where a trend appears in different groups of data but disappears or reverses when the groups are combined
- When an omitted variable is not taken into account in the analysis, leading to a contradictory interpretation of the results
- Simpson's Paradox often arises due to the presence of lurking variables, also known as confounding variables



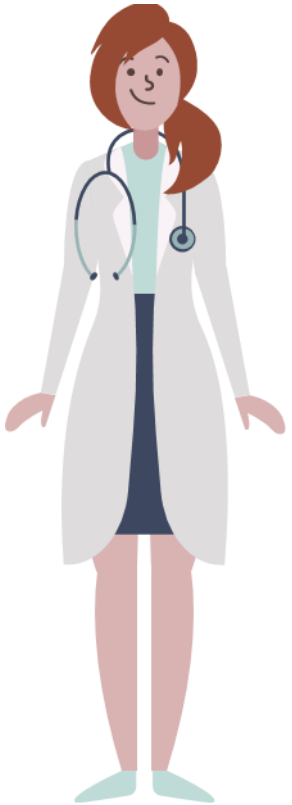


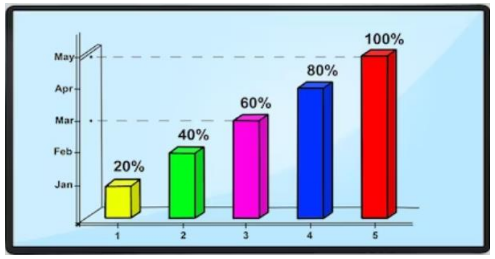
SIMPSON'S EFFECT IN THE " PALMER PENGUIN " DATASET

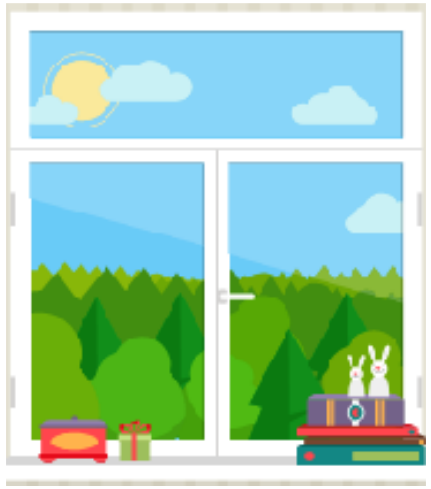
SIMPSON'S PARADOX

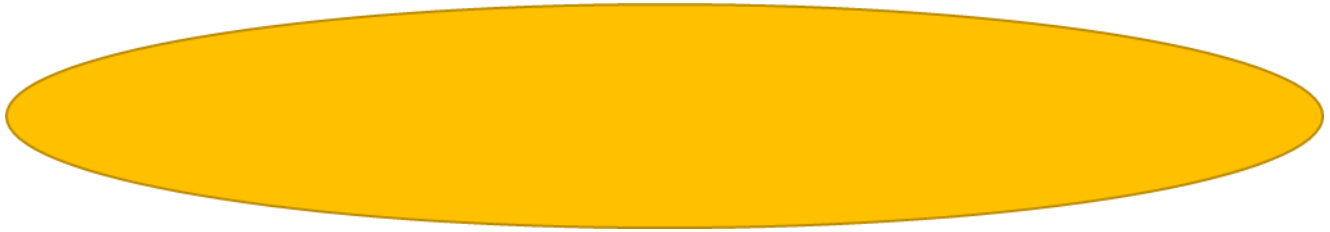


MEET
Dr. SARA
And
ANNA



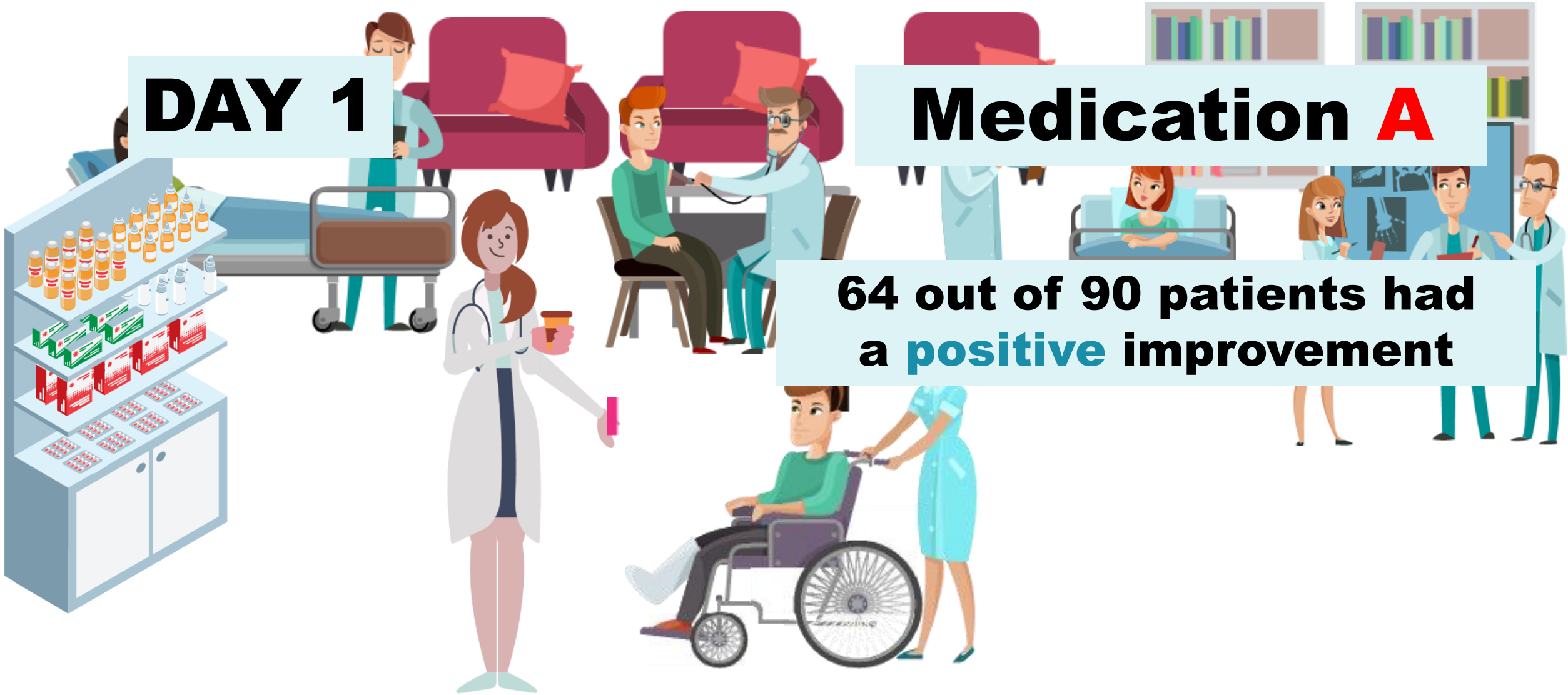

May
Apr
Mar
Feb
Jan
20%
40%
60%
80%
100%
1
2
3
4
5



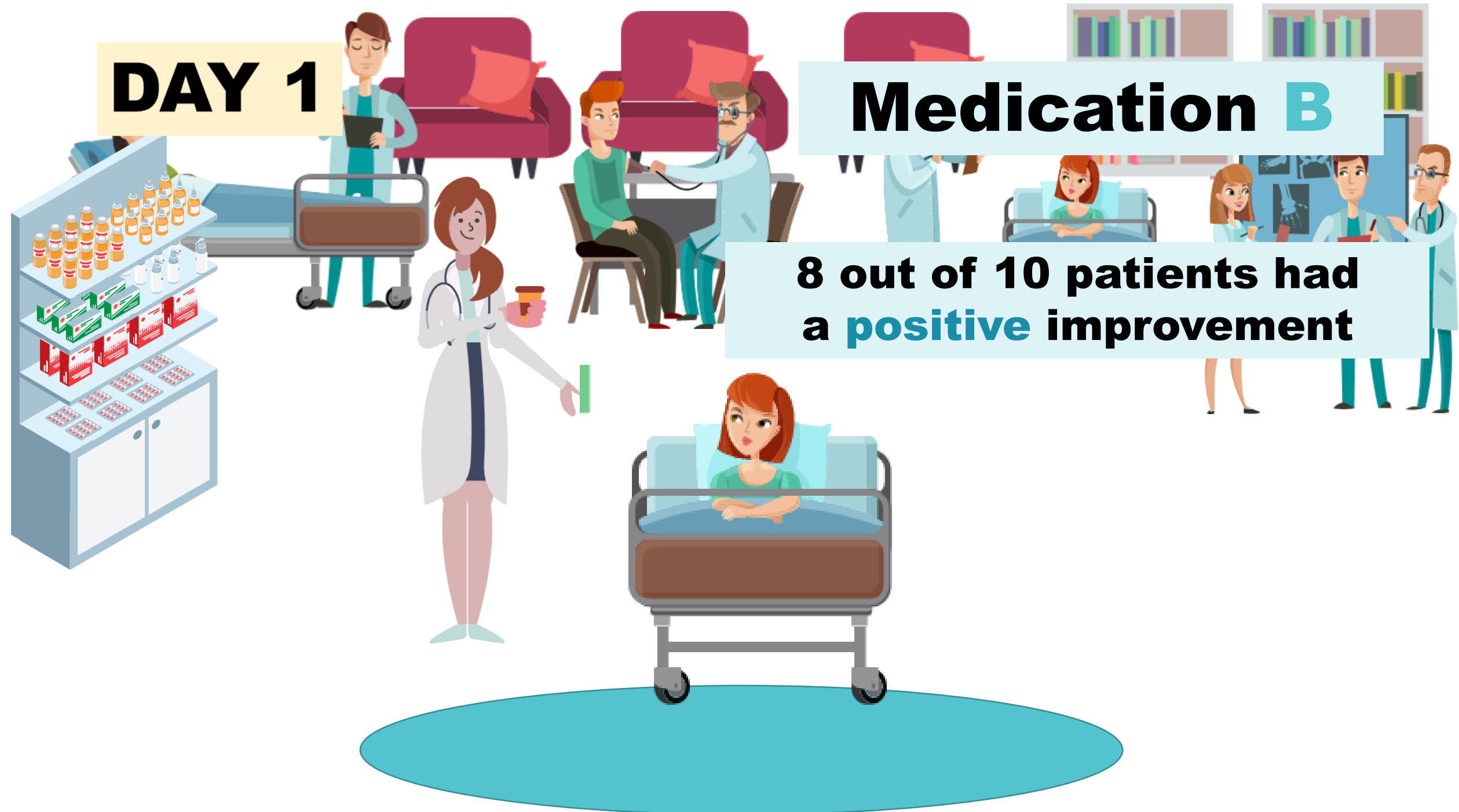


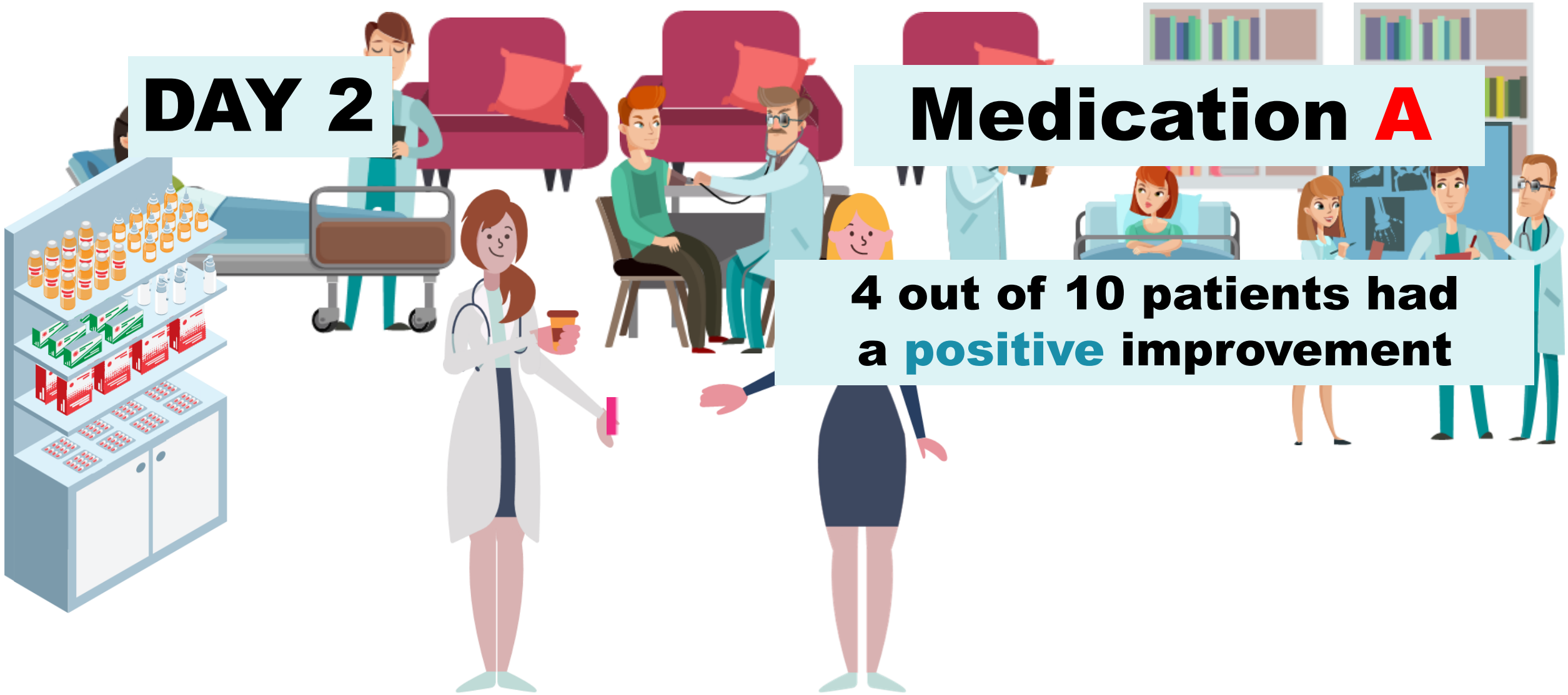
DAY 1

Medication A



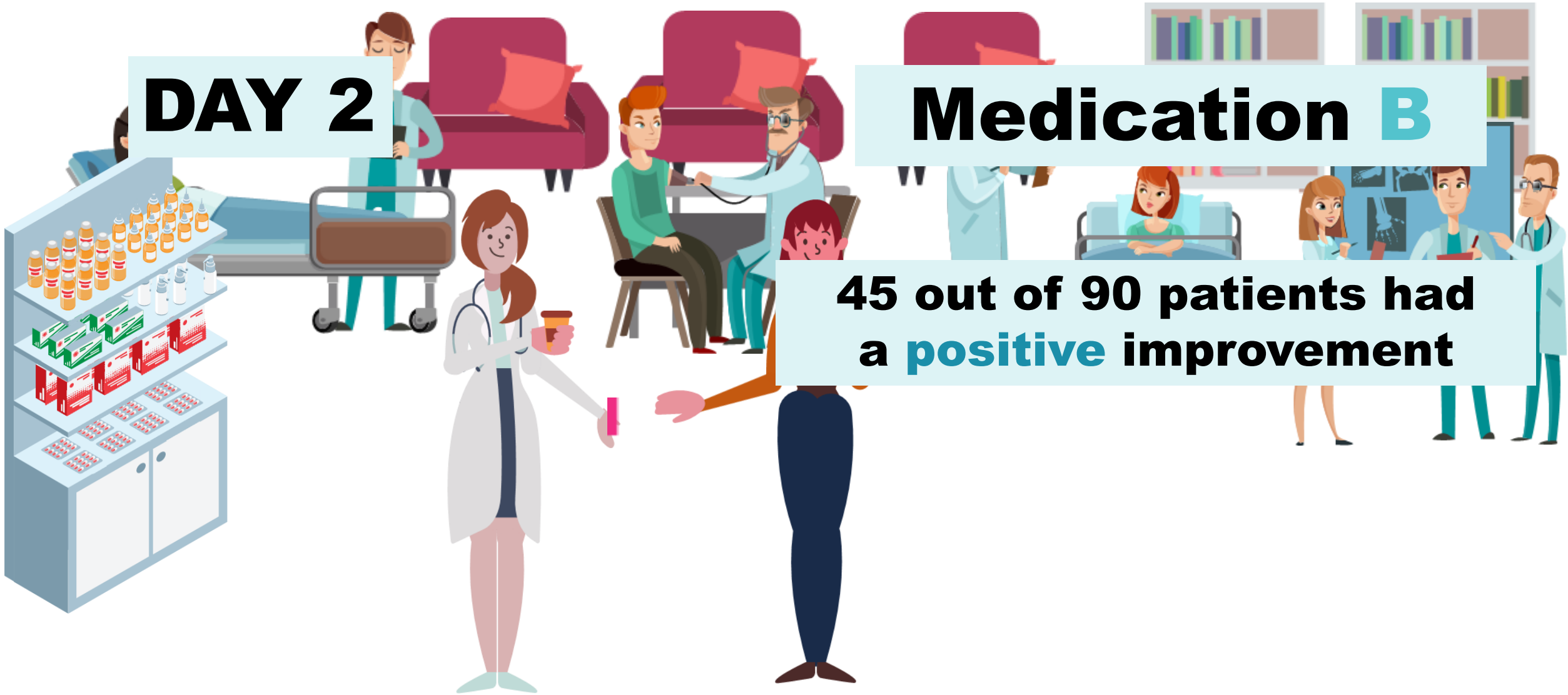
64 out of 90 patients had a positive improvement


DAY 1
Medication B
8 out of 10 patients had
a positive improvement


DAY 2
Medication A
4 out of 10 patients had
a positive improvement

DAY 2

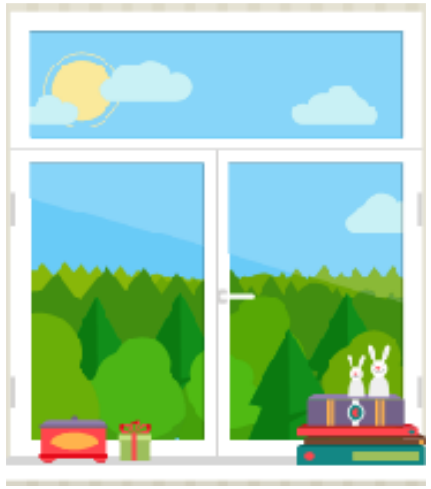
Medication B

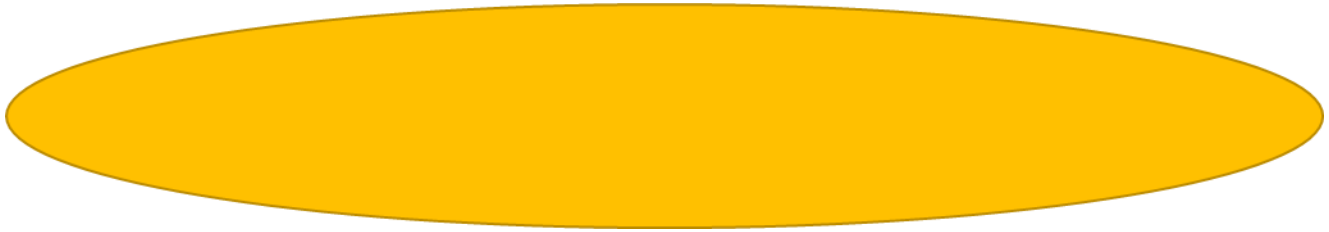


45 out of 90 patients had a positive improvement


May
Apr
Mar
Feb
Jan
20%
40%
60%
80%
100%
1
2
3
4
5

medication type	A	B
Day 1	64/90 = 70%	8/10 = 80%
Day 2	4/10 = 40%	45/90 = 50%
over all	67/100	53/100





In conclusion,

The overall cure rate for medication A is 0.67, while the overall cure rate for medication B is 0.38.
However, when we look at the cure rates by day, we see that medication A has a higher cure rate than medication B on each day individually.

This is an example of Simpson's paradox, where the relationship between medication and cure rate changes as we take each day into account.

Therefore, we cannot conclude that medication B is better than medication A.


THANK YOU
Zeina Qasem & Fatima Mohammed