SESSIONS 4

Association

Data Science Program

Purwadhika
Startup and Coding School

Association

1) Association
2) Type of Association based on variable
3) How To Explore ?
4) Exploring Association
   - Smoking status and 20 year survival in women
   - Education and crime rate
   - Death penalty and race

Relationship Between Two Events

Two event often related to each other. For example:

- Air temperature and humidity
- Price and demand
- Fertilizer and plant height
- Weight and height
- Time and COVID-19 victim in daily

There are two types of relationship

- Association → correlation
- Causation → regression

Response Variable and Explanatory Variable

When analyzing relationship between two variable usually we must first distinguish between **response variable** (y) and **explanatory variable** (x).

Response Variable:

- Value in response variable depends on explanatory variable.

Explanatory Variable:

- Quantitative : how different value in explanatory relate to changes in response variable
- Qualitative : it is like grouping or aggregating. how is the comparison between group based on some aggregate function (mean, sum, count, percentage, etc)

Causation. If change in X cause change in Y, doesn’t imply that change in Y cause change in X.

Type of Cases Can happen in Association

Variable can be qualitative or quantitative. So, there are three possible cases:

1. **Qualitative Vs Qualitative:**

   ex. gender and education

2. **Quantitative Vs Qualitative:**

   ex. income and race, height and gender

3. **Quantitative Vs Quantitative:**

   ex. air temperature and humidity, weight and height

How to explore the relationship?

Qualitative vs Qualitative

- Graphical Summary: Barchart, Pie chart
- Numerical/Table Summary : Contingency table/cross tabulation, Odds ratio, Difference of proportion, Ratio of proportions, Chi-square Test.

Qualitative vs Quantitative

- Graphical Summary: Barplot
- Numerical/Table Summary: Aggregated table, Anova F-Test.

Quantitative vs Quantitative (we will focus on this)

- Graphical Summary: Scatterplot
- Numerical/Table Summary: Pearson Correlation or Spearman Correlation, Regression.

Purwadhika
Startup and Coding School

Correlation

- Correlation is about association and **association doesn’t imply causation.**
- Correlation **doesn’t differentiate response** (x) variable and **explanatory** variable (y).
- Correlation only **measure how strong relationship** and **the direction of relationship.**

- Correlation ranged by

  $-1 < r < 1$

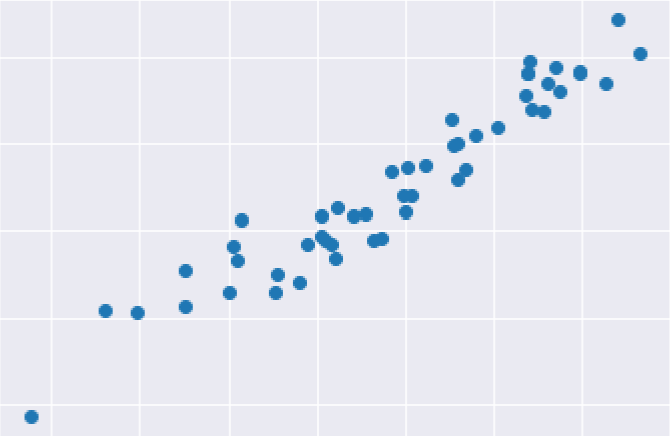
- Positive direction (+)
- Negative direction (-)
- The magnitude (absolute value)

  0 - 0.3 : weak

  0.3 - 0.7 : medium

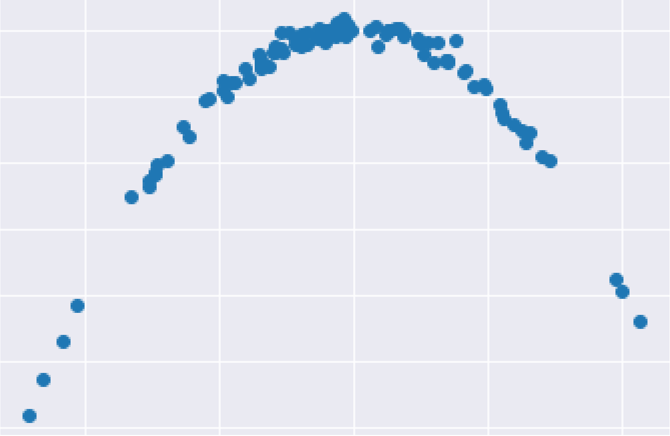
  0.7 - 1 : strong

Type of Numerical Relationship



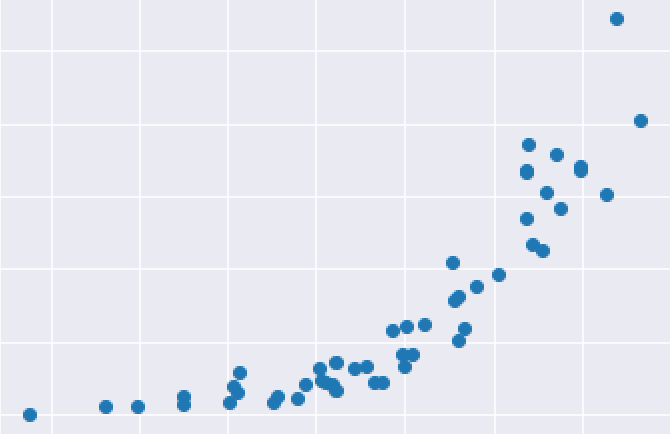
Linear:
- Use Pearson Correlation

Ex. height and weight



Non Linear and Non-monotonic : Strongly not recommended to measured by Pearson or Spearman.

Ex. fertilizer dose and plant height



Non-Linear or Monotonic:
- quadratic
- qube

Use Spearman Correlation

Ex. daily case of COVID-19

Purwadhika
Startup and Coding School

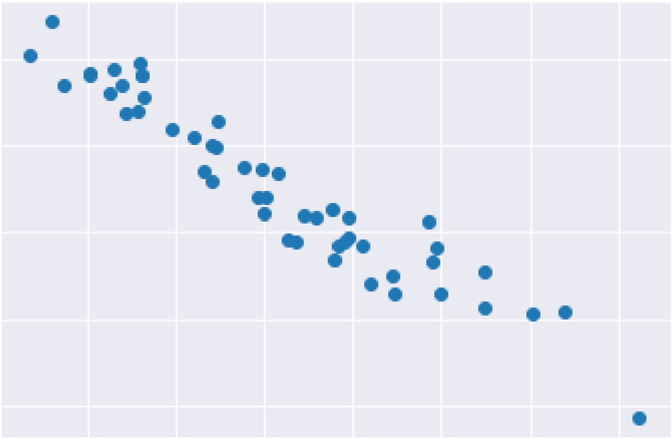
Pearson Correlation

1. Both of the variable should be quantitative
2. Relationship between two variable should be linear
3. Parametric method

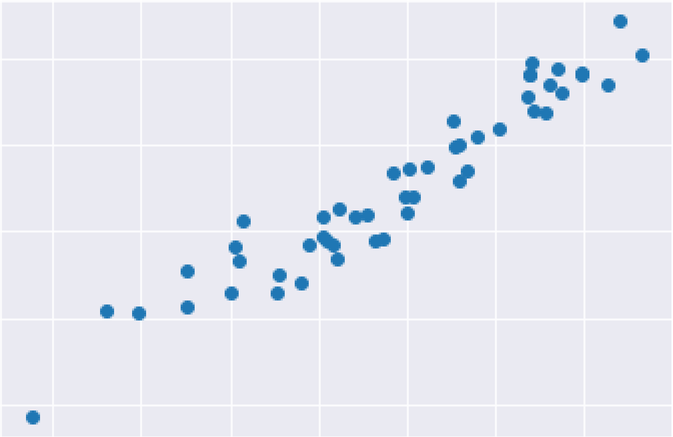
Formula

$$\rho_{X,Y} = \frac{E[(X-E[X])(Y-E[Y])]}{\sigma_X \sigma_Y}$$

Linear Negative



Linear Positive



Purwadhika
Startup and Coding School

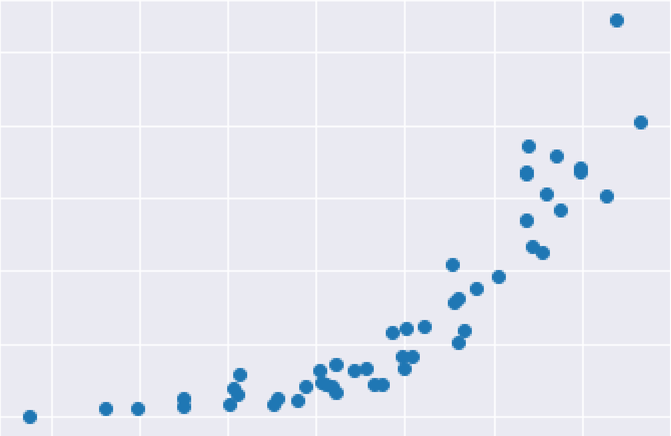
Spearman Correlation

1. Beside quantitative variable. it can be used to explore variable with ordinal scale.
2. Relationship between two variable should not be linear. It should be either positive monotonic or negative monotonic
3. Nonparametric version of Pearson

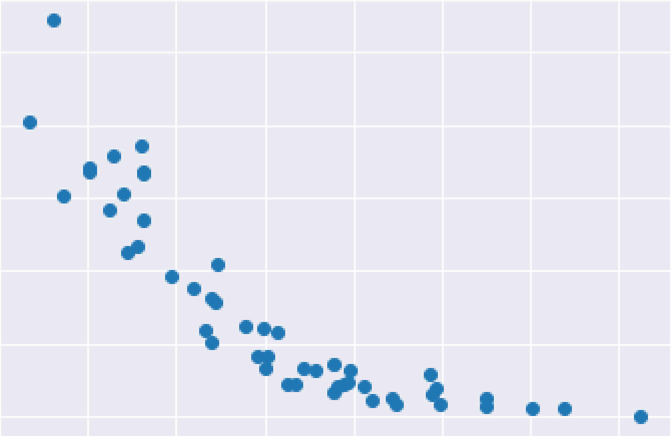
Formula

$$\rho_{rank_x}, \rho_{rank_y} = \frac{cov(rank_x, rank_y)}{\sigma_{rank_x}, \sigma_{rank_y}}$$

Monotonic Positive

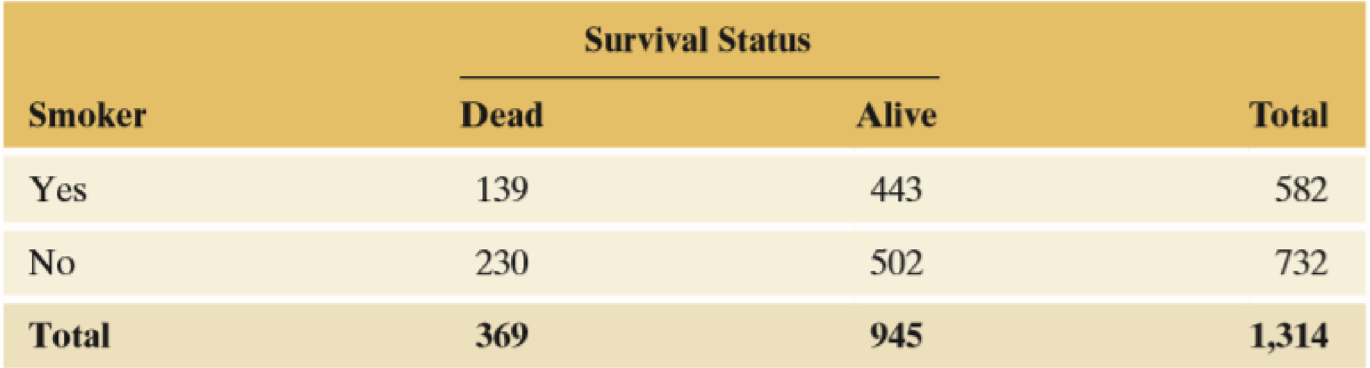


Monotonic Negative



Smoking Status and 20-year survival in Women

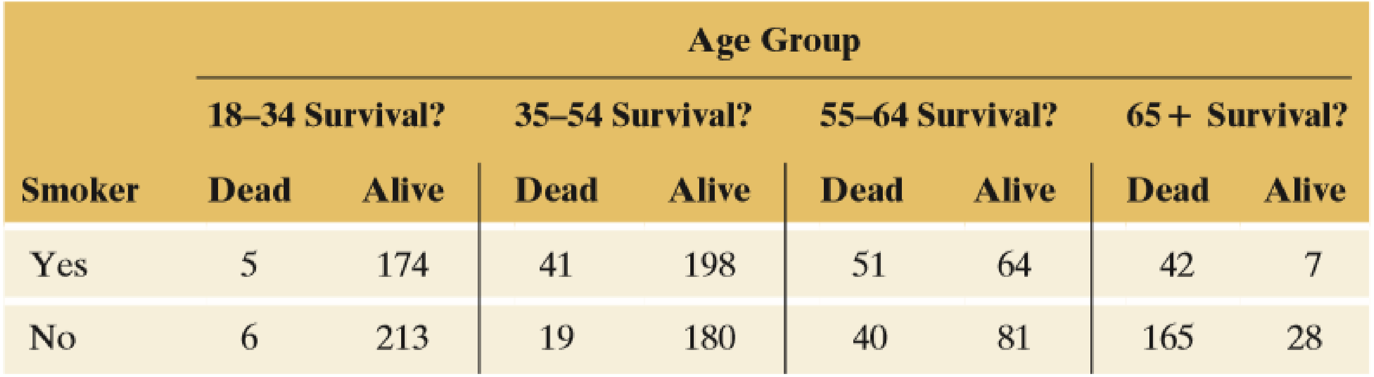
A survey of 1,314 women in the United Kingdom that asked each woman whether she was a smoker. Twenty years later, a follow-up survey observed whether each woman was dead or still alive



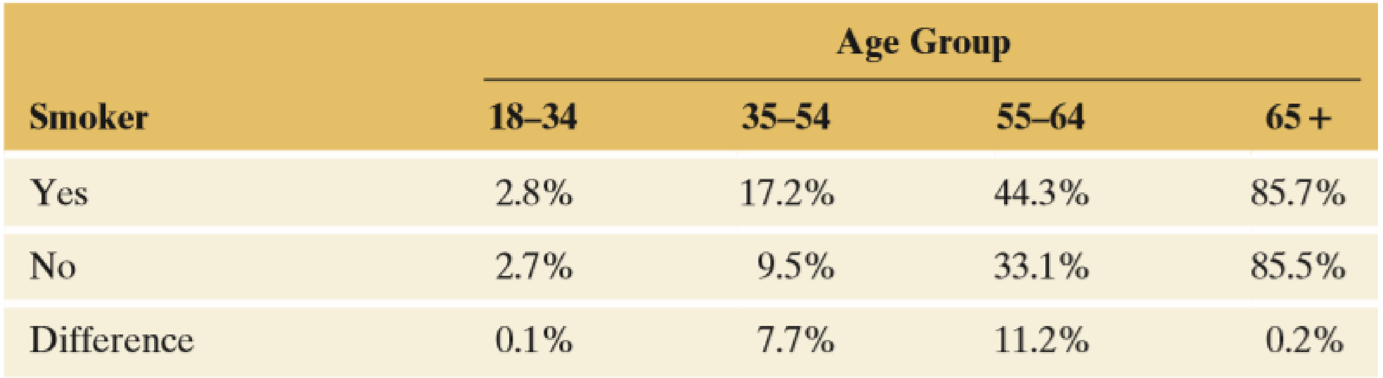
	Survival Status		
Smoker	**Dead**	**Alive**	**Total**
Yes	139	443	582
No	230	502	732
Total	**369**	**945**	**1,314**

- 31 % non-smoker died and 24% smoker died
- Smoker has lower death rate

Smoking Status and 20-year survival in Women



	Age Group							
	18–34 Survival?		35–54 Survival?		55–64 Survival?		65 + Survival?	
Smoker	Dead	Alive	Dead	Alive	Dead	Alive	Dead	Alive
Yes	5	174	41	198	51	64	42	7
No	6	213	19	180	40	81	165	28



	Age Group			
Smoker	18–34	35–54	55–64	65 +
Yes	2.8%	17.2%	44.3%	85.7%
No	2.7%	9.5%	33.1%	85.5%
Difference	0.1%	7.7%	11.2%	0.2%

- Percentage of survival rate is vary for each age group
- Non-smoker always has lower death rate when age group taken into account
- The association very different than before

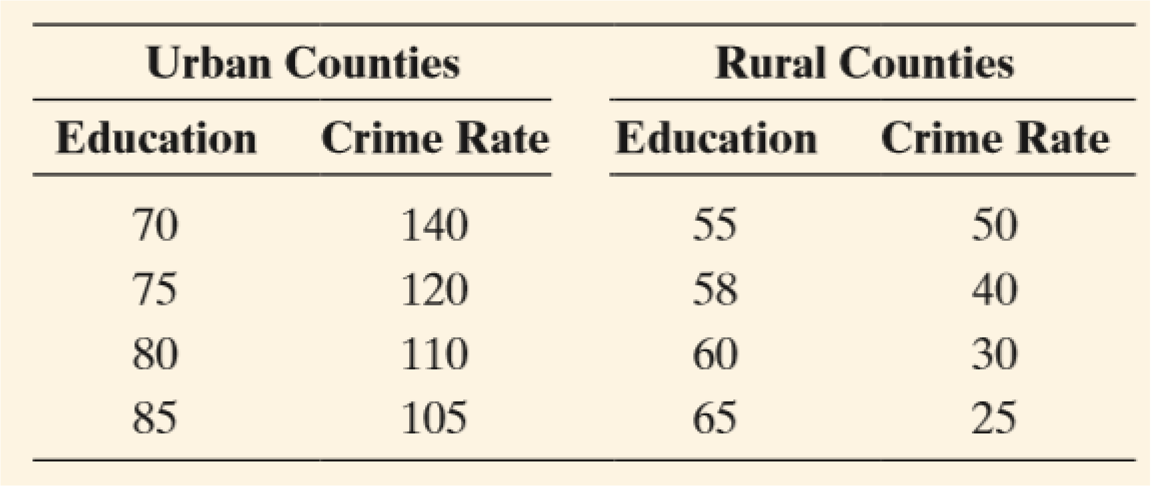
For instance, for smokers of age 18-34, from the first table the proportion who died was 5/(5 + 1742) = 0.028, or 2.8%

Purwadhika
Startup and Coding School

Simpson's Paradox

Beware of the **Simpson's Paradox** when analyzing relationship : **Education and Crime Rate**

Education and Crime Rate



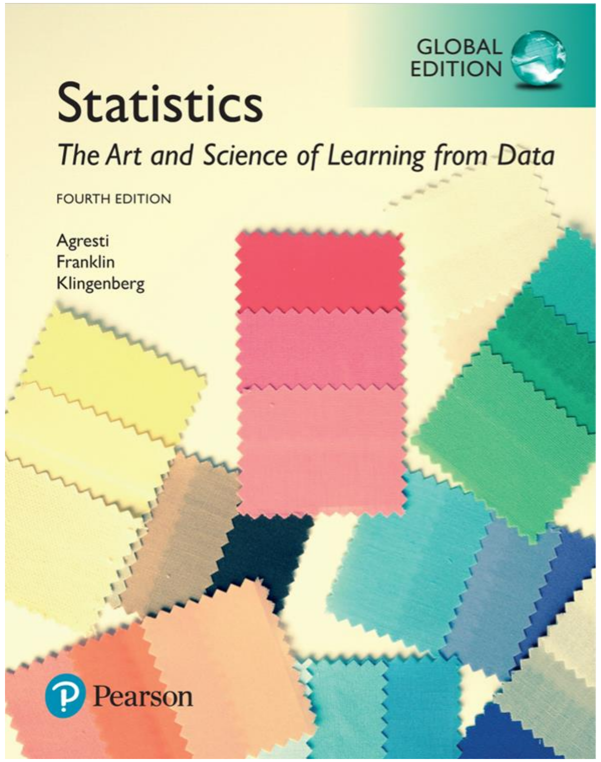
Urban Counties		Rural Counties	
Education	**Crime Rate**	**Education**	**Crime Rate**
70	140	55	50
75	120	58	40
80	110	60	30
85	105	65	25

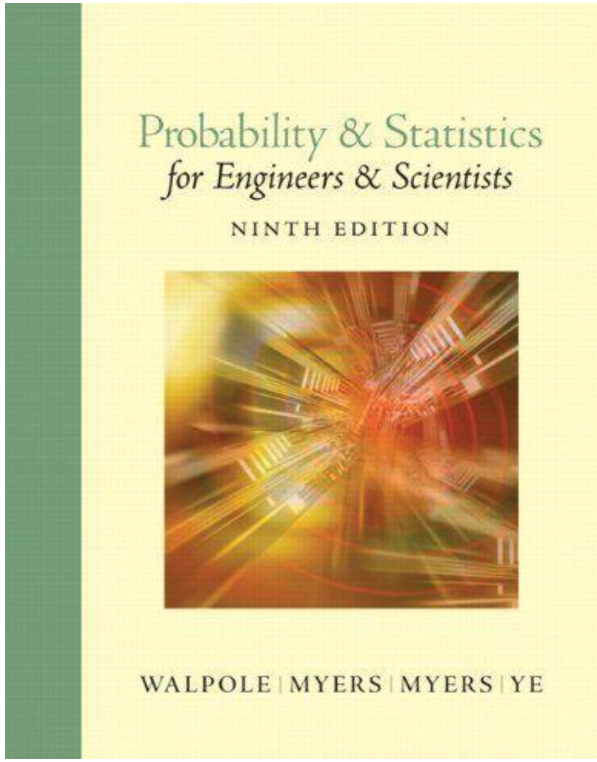
Let's Analyze this data:

1. Make The Dataframe In Python For Whole Dataset
2. Analyze Marginally
3. Analyze Partially

Purwadhika
Startup and Coding School

Reference


GLOBAL EDITION
Statistics
The Art and Science of Learning from Data
FOURTH EDITION
Agresti
Franklin
Klingenberg
Pearson


Probability & Statistics
for Engineers & Scientists
NINTH EDITION
WALPOLE | MYERS | MYERS | YE

Reference

https://towardsdatascience.com/data-science-you-need-to-know-a-b-testing-f2f12aff619a

https://towardsdatascience.com/data-science-fundamentals-a-b-testing-cb371ceecc27

https://www.niagahoster.co.id/blog/ab-testing-adalah/

https://vwo.com/blog/ab-testing-examples/

https://www.scribbr.com/methodology/sampling-methods/