Story of Lady Tasting Tea

How statistical inference could help you

By Alamsyah Hanza

DATA SCIENCE INDONESIA

The Content

- Get to Know You
- The Lady who have different tea
- Statistics
- Daily Inference
- Statistical Inference
- Hypothesis Testing
- Can the Lady distinguish it?
- Conclusion

DATA SCIENCE INDONESIA

Hey I’m,

Alamsyah K Hanza

I am a,

- CURRENTLY | Head of Data Science Indonesia community and **Data Scientist at Gojek**
- 18 - 19 | **Business Intelligence at Gojek**
- 17 - 18 | **Data Science Engineer at PegiPegi**

RULES

DATA SCIENCE INDONESIA

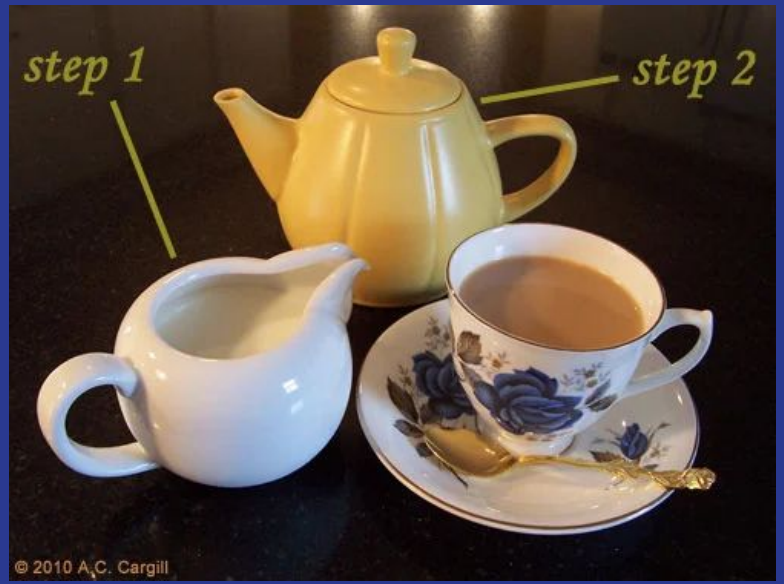
The Lady who have a Favorite Tea

DATA SCIENCE INDONESIA

In 1930 something at Rothamsted, **Ronald Fisher** offered **Mureal Bristol** a cup of hot tea that he had just drawn from an urn.

Bristol declined it, saying that she preferred the flavour when the **milk was poured into the cup before the tea**. Fisher scoffed that the order of pouring could not affect the flavour.

Bristol insisted that it did and that she could tell the difference. Overhearing this debate, William Roach (Mureal's Husband) said, "Let's test her."



DATA SCIENCE INDONESIA

Ronald Fisher

Really? I’m not sure

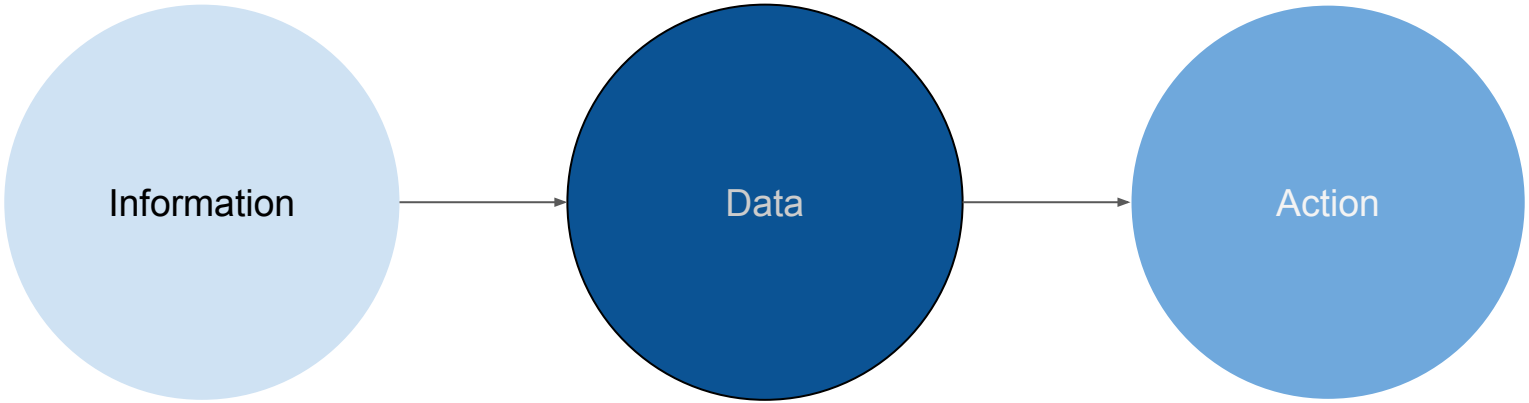
with british accent

DATA SCIENCE INDONESIA

Statistics

DATA SCIENCE INDONESIA

Based on the History ...



DATA SCIENCE INDONESIA

Unexpectedly, The Stats is in You

Statistics is not a new thing in your life. Its not the thing that you learn in school or collage only. Its already in you.

Remember? When you suddenly realize that he price of the spice spike up this day.. Even without checking the news..

DATA SCIENCE INDONESIA

Statistics Types

S. No	Descriptive Statistics	Inferential Statistics
1	Concerned with the describing the target population	Make inferences from the sample and generalize them to the population.
2	Organize, analyze and present the data in a meaningful manner	Compares, test and predicts future outcomes.
3	Final results are shown in form of charts, tables and Graphs	Final result is the probability scores.
4	Describes the data which is already known	Tries to make conclusions about the population that is beyond the data available.
5	Tools- Measures of central tendency (mean/median/ mode), Spread of data (range, standard deviation etc.)	Tools- hypothesis tests, Analysis of variance etc.

DATA SCIENCE INDONESIA

Daily Inference

DATA SCIENCE INDONESIA

How do you do the inference?

DATA SCIENCE INDONESIA

Discount Case

Let we have 1000 customers who used our discount.

The Demographic shows 750 of them comes from Kalimantan. We say that Kalimantan people love our discount.

Medicine Case

After the vaccine in a city, the transmission rate is gone down to 0.3 from 0.5 last month.

This significant difference shows how good the vaccine impacting the population.

Teacher Score

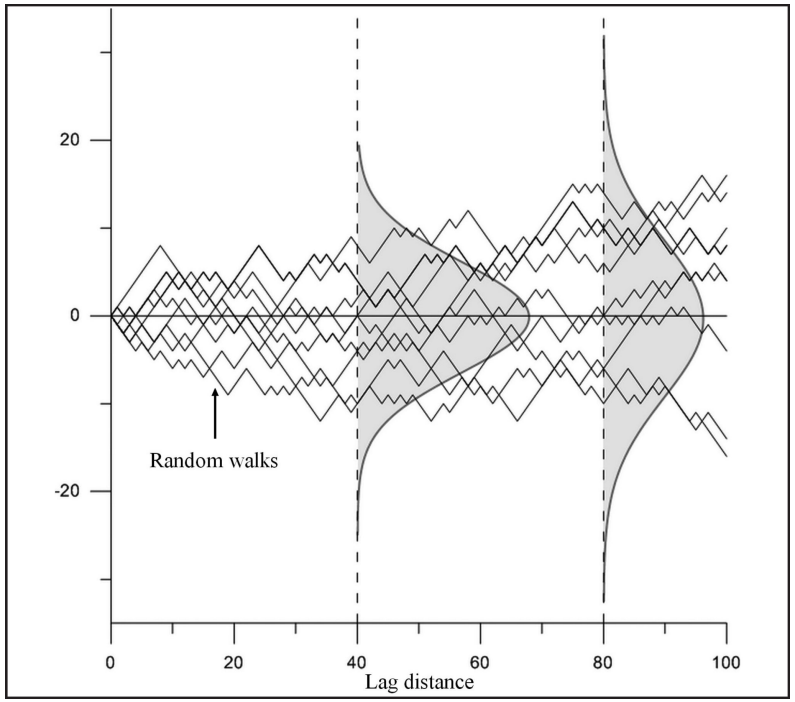
Every year a country evaluate their teacher based on their student score.

It shows that the top 100 teachers comes from urban cities. We say that the urban city teacher is better than suburban.

Consider This

DATA SCIENCE INDONESIA

Consider This



DATA SCIENCE INDONESIA

How do you do the inference Now?

DATA SCIENCE INDONESIA

Discount Case

Let we have 1000 customers who used our discount.

The Demographic shows 750 of them comes from Jabodetabek. What is the conclusion?

Medicine Case

After the vaccine in a city, the transmission rate is gone down to 0.3 from 0.5 last month.

What does this significant differences show to us?

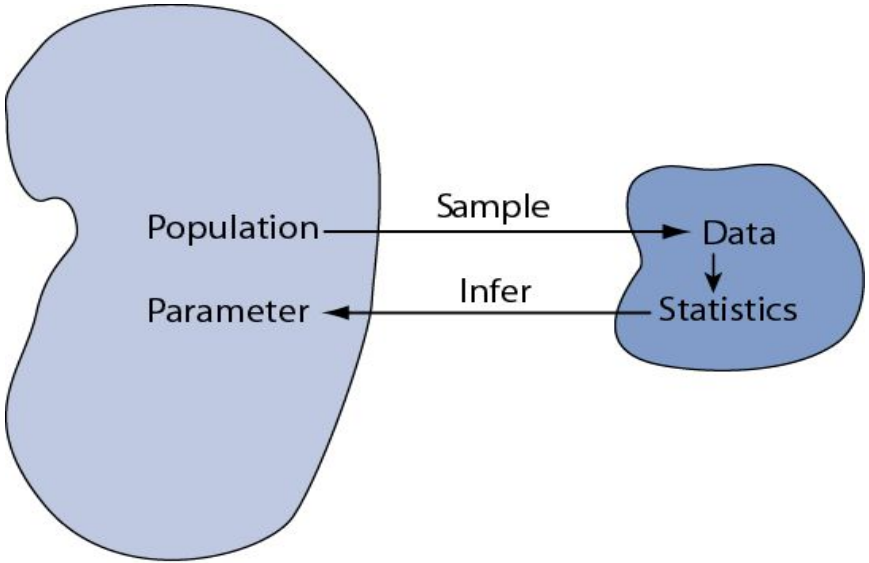
Teacher Score

Every year a country evaluate their teacher based on their student score.

It shows that the top 100 teachers comes from urban cities. What does it mean?

Statistical Inference

DATA SCIENCE INDONESIA



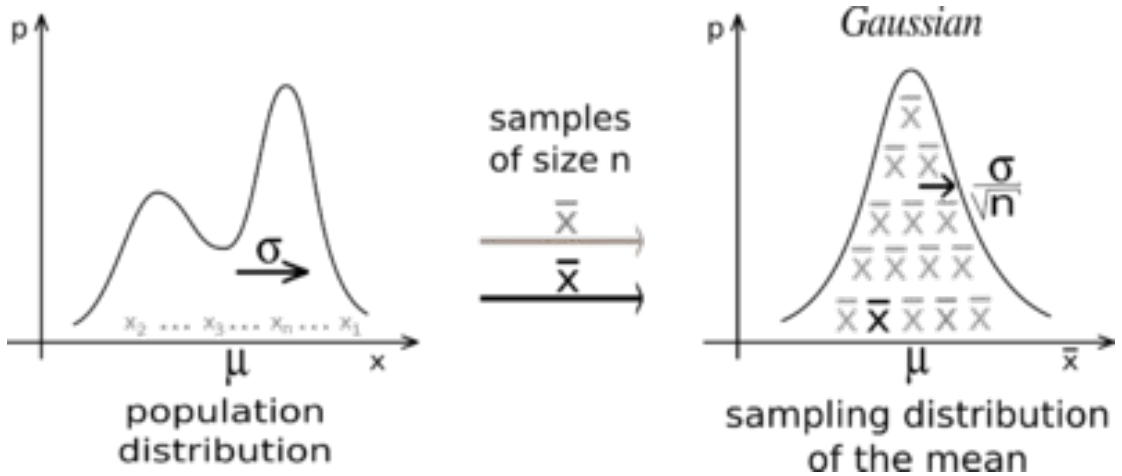
Inferencing

Statistical inference is the act of generalizing from a sample to a population with calculated degree of certainty. In other words, Use a random sample to learn something about a larger population.

We want to learn about population parameters, but we can only calculate sample statistics.

DATA SCIENCE INDONESIA

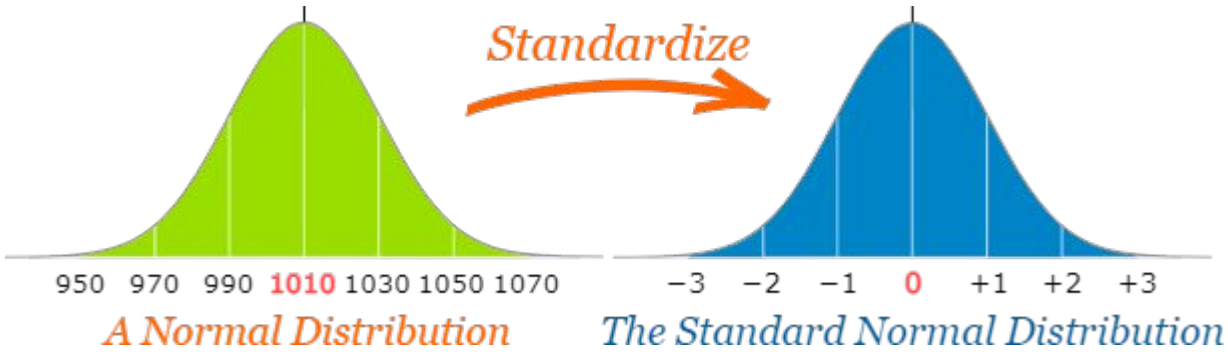
Central Limit Theorem



The Central Limit Theorem states that the sampling **distribution of the sample means approaches a normal distribution** as the sample size gets larger — no matter what the shape of the population distribution. This fact holds especially true for sample sizes over 30.

DATA SCIENCE INDONESIA

Standardization

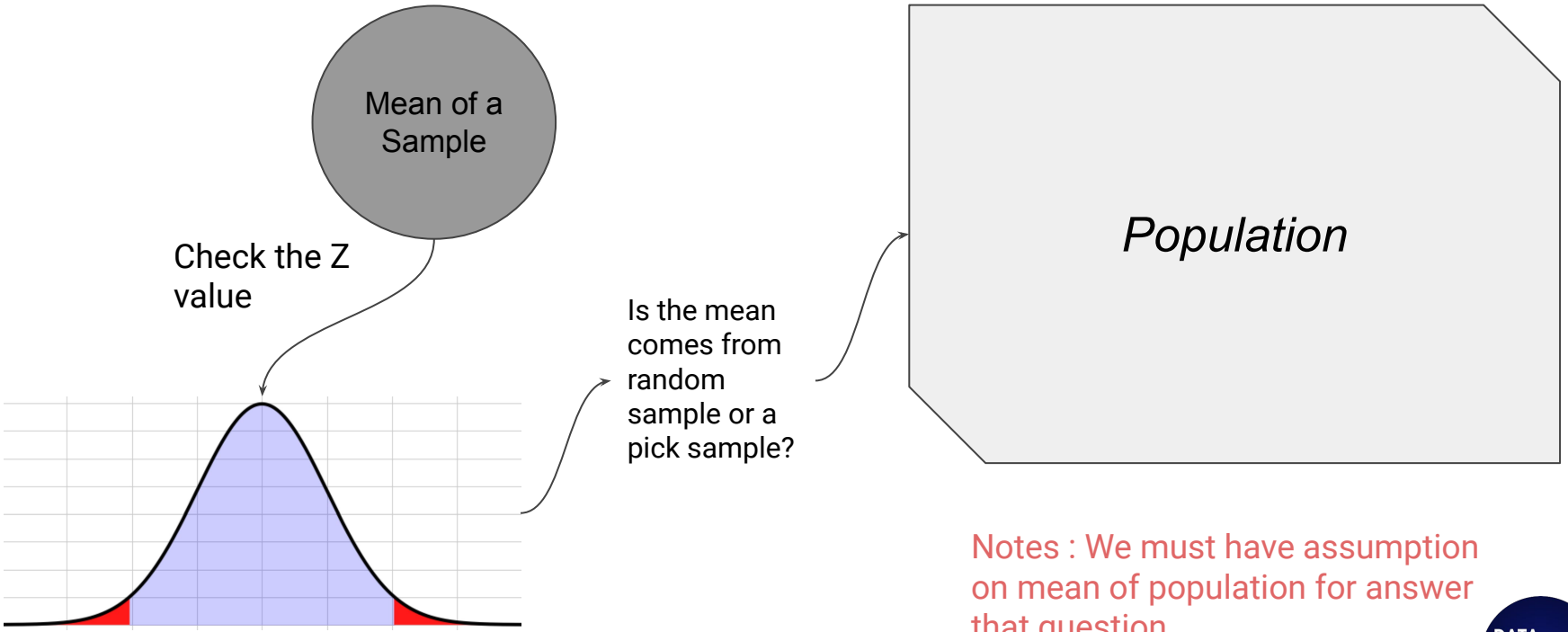


Any point (x) from a normal distribution can be converted to the standard normal distribution (z) with the following formula :

$$Z = \frac{(X - \mu)}{\sigma}$$

DATA SCIENCE INDONESIA

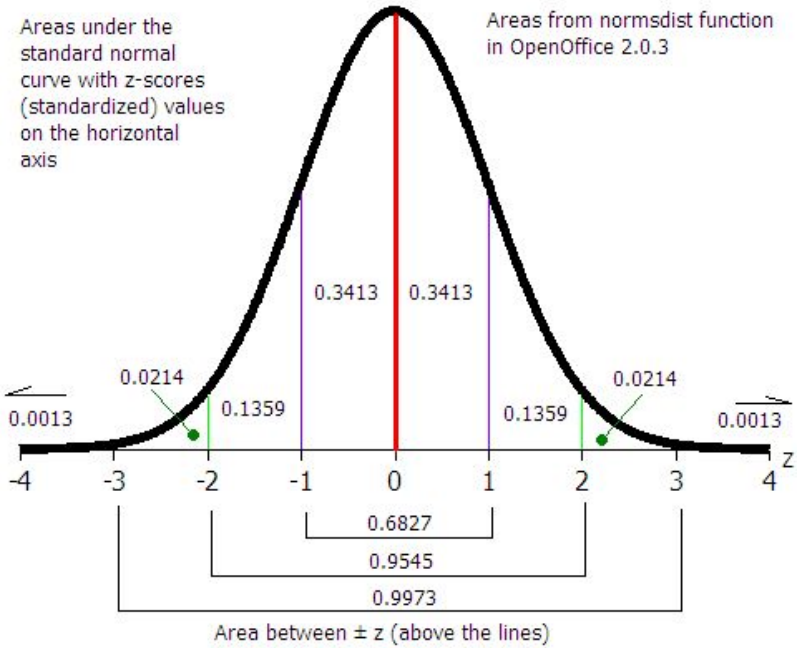
Reverse CLT for our usage

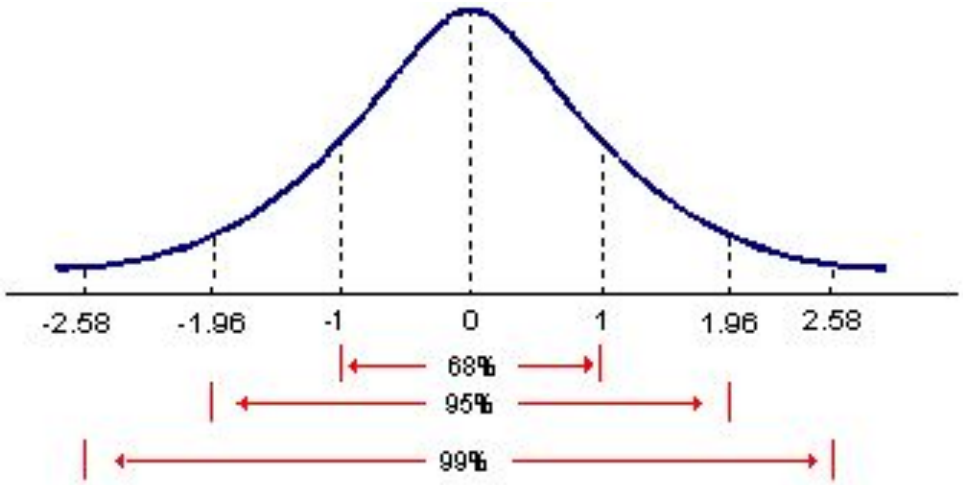


Notes : We must have assumption on mean of population for answer that question

DATA SCIENCE INDONESIA

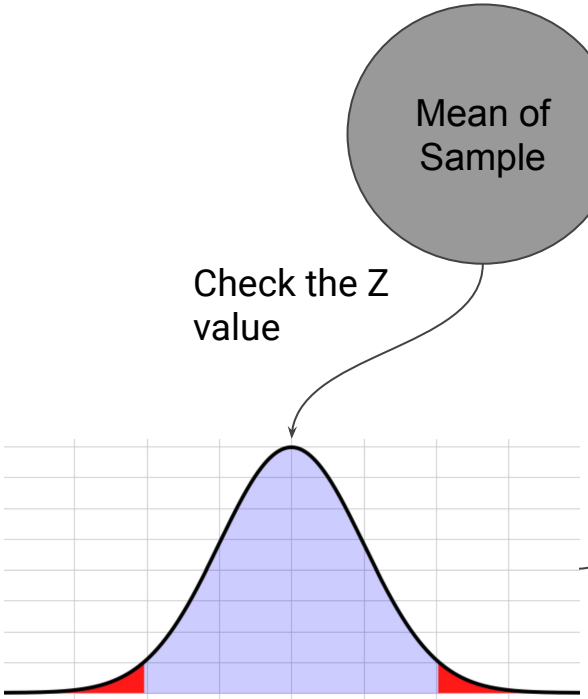
Remember





DATA SCIENCE INDONESIA

Reverse CLT for our usage



$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

DATA SCIENCE INDONESIA

Hypothesis Testing

DATA SCIENCE INDONESIA

Definition

A statistical method that uses sample data to evaluate a hypothesis about a population parameter. It is intended to help researchers differentiate between real and random patterns in the data. Remember The goal of hypothesis testing is to see if **there is enough evidence against the null hypothesis**. General steps :

1. State the research question.
2. State the statistical hypothesis.
3. Calculate the test statistic.
4. Decide if result is significant and compare.
5. Interpret result as it relates to your research question.

DATA SCIENCE INDONESIA

Sample Case (Mean)

Problem

Our cases

Mean of a value for customer who have used our promotion code = 175.000 (denotes as Mu).

Let say we know the mean value from population = 150.000 (denotes as M). Question: is the customer who have used promotion significantly different with others?

Test Statistic

Transform to Standard

H0 : Mu = M

H1 : Mu != M

Use the Z formula :

$$Z = \frac{(X - \mu)}{\sigma}$$

Conclusion

Check our Ruler

If the Z less than -2 or more than 2, then the Mu happen is not coincidence.

So, Mu is different with M (enough evidence to reject H0).

Remember,

Pr(Z<-2) = 2.14%

It means the Mu is not coincidence since it must be happen only 2.14% of the time.

DATA SCIENCE INDONESIA

Quick Big Question

Which one is the *p-value*? What is it exactly?

DATA SCIENCE INDONESIA

Can the Lady distinguish it?

DATA SCIENCE INDONESIA

The Set Up

A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. We will consider the problem of designing an experiment by means of which this assertion can be tested. […] [It] consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgment in a random order. The subject has been told in advance of that the test will consist, namely, that she will be asked to taste eight cups, that these shall be four of each kind […]. — Fisher, 1935.

DATA SCIENCE INDONESIA

The Ruler/Measurement

Success count	Combinations of selection	Number of Combinations
0	oooo	1 × 1 = 1
1	ooox, ooxo, oxoo, xooo	4 × 4 = 16
2	ooxx, oxox, oxxo, xoxo, xxoo, xoox	6 × 6 = 36
3	oxxx, xoxx, xxox, xxxo	4 × 4 = 16
4	xxxx	1 × 1 = 1
	Total	70

DATA SCIENCE INDONESIA

The Results

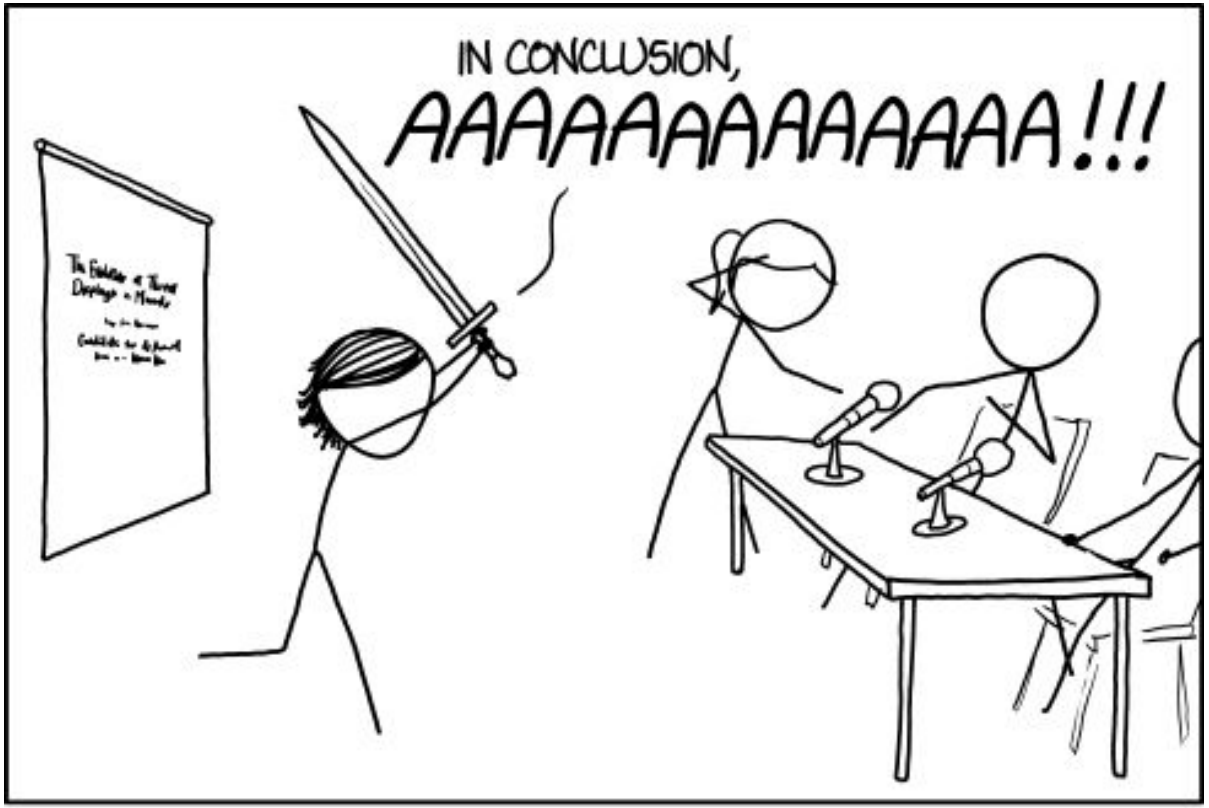
David Salsburg reports that a colleague of Fisher, H. Fairfield Smith, revealed that in the actual experiment the lady succeeded in identifying all eight cups correctly.

The chance of someone who just guesses of getting all correct, assuming **she guesses that any four had the tea put in first and the other four the milk**, would be only 1 in 70 (the combinations of 8 taken 4 at a time).

DATA SCIENCE INDONESIA

Conclusion

DATA SCIENCE INDONESIA



THE BEST THESIS DEFENSE IS A GOOD THESIS OFFENSE.

DATA SCIENCE INDONESIA

Thank you