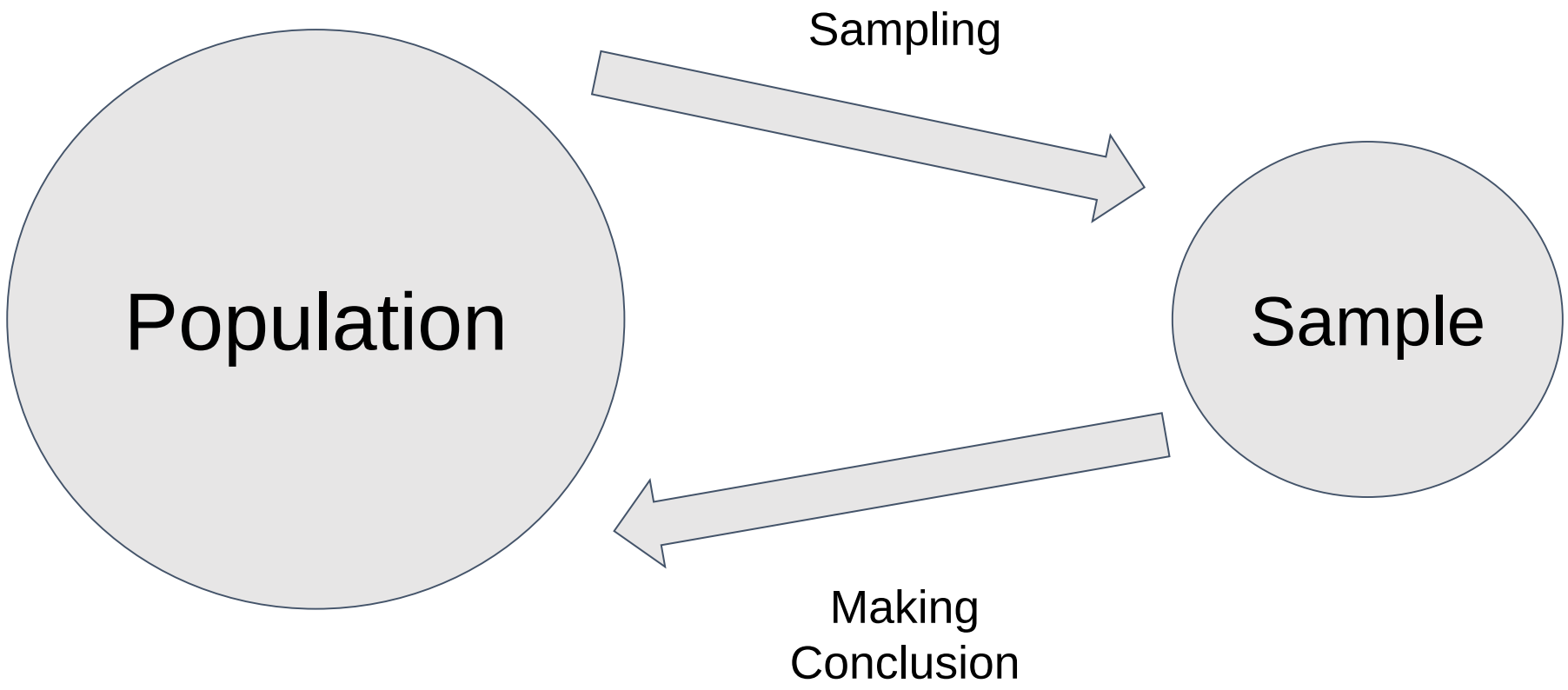
SESSIONS 4

Inferential Statistics: Confidence Interval

Data Science Program

Purwadhika
Startup and Coding School

Inferential Statistics



Purwadhika
Startup and Coding School

Inferential Statistic

- Inferential statistics uses **samples** to **generalizations** any phenomenon in the **population**.
- Inferential statistics uses **probability** to measure the **reliability of conclusion**.
- Method usually used to do inference are:
    - Point estimation
    - Interval estimation
    - Hypothesis testing

Estimate

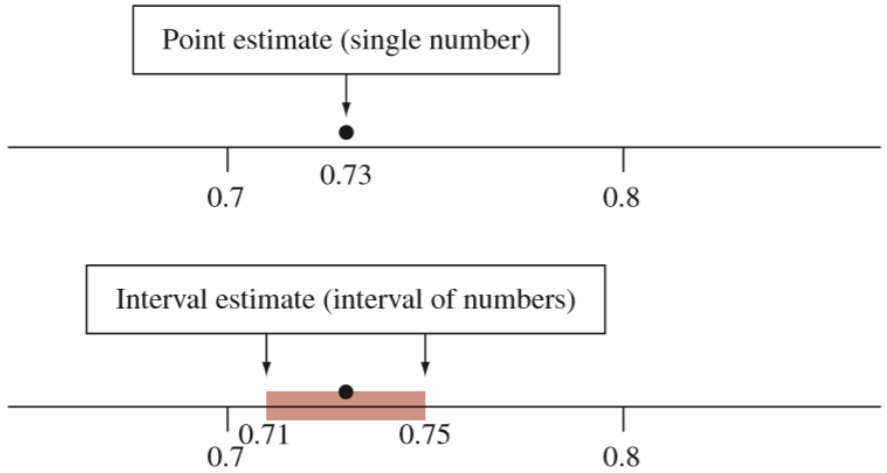
- Estimation type
- Point estimate
- Interval estimate
- Different kind of interval estimate
  - Mean
  - Proportion

Estimation

We already know that we could use the sample data to estimate the unknown population parameter. Statistical inference use sample data to forms two types of estimators of parameters:

- **Point Estimate** (Single Number)
- **Confidence Interval** (Interval)

A study about people in Bogor who agree about Bogor Station Renovation.
People who agree with the renovation estimated 73%.



Purwadhika
Startup and Coding School

Point Estimate

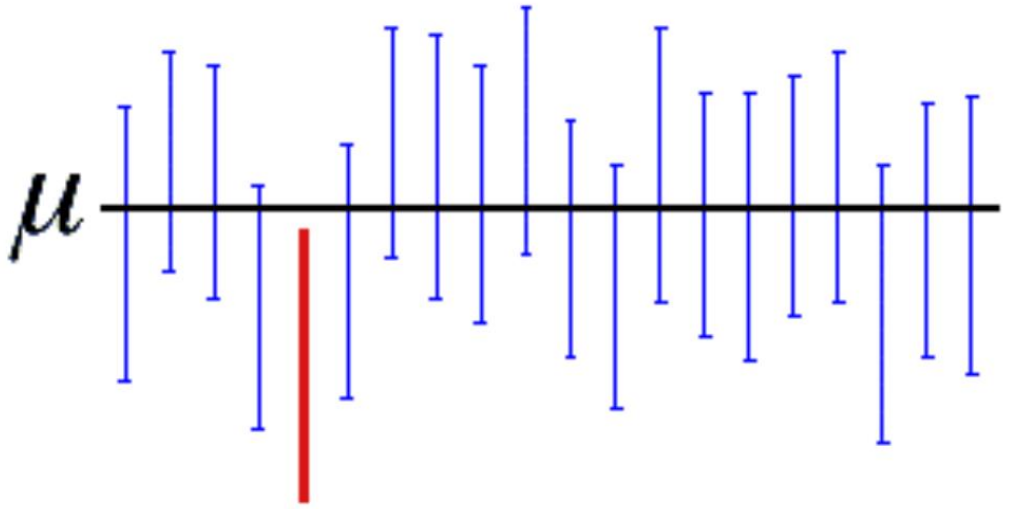
- Point estimate consist of a **single number**, calculated from the **sample**, that is the **best guess** for the unknown parameter.
- The **sample** (X1,X2,….Xn of size n) is assumed taken **randomly** from the **population**.
- For example:
  - to estimate the **μ** (Population Mean) we could use estimator **$\bar{x}$**.
  - to estimate the ***P*** (Population Proportion) we could use estimator **p**.
- A good point estimator of a parameter is one with sampling distribution that is centered around parameter, and smallest Standard Error.

Interval Estimate

We construct an **interval** that we hope **contains** the **parameter** with certain **level** of **confidence**.

The **confidence level** is a number from 0 to 1, which represent the **probability**. (Often it is 0.90 (90%), 0.95(95%), or 0.99(99%))

95 % Confidence level means that if we take sample size n and we repeat for 20 times. We believed that the parameter covered in confidence interval 19 times.



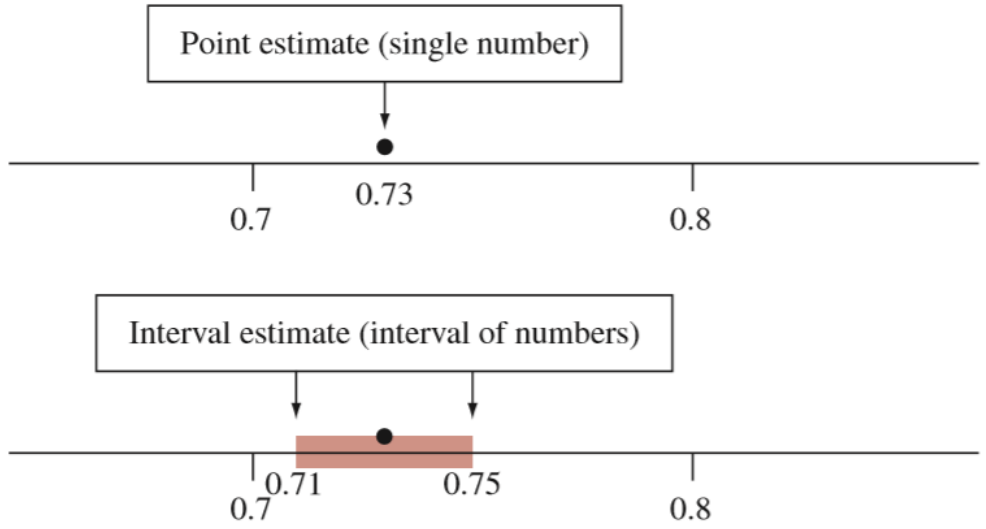
A 95% confidence interval indicates that 19 out of 20 samples (95%) from the same population will produce confidence intervals that contain the population parameter.

Purwadhika
Startup and Coding School

Why Is Point Estimate Not Sufficient?

- A point estimate is only an estimate, it could be wrong or far from the actual value.
- It also doesn't tell us how close the estimate likely to the parameter
- We need to quantify how close our estimation is and how far we will miss.

We 95 % believed that the true proportion of people in Bogor who agree with the new renovation of Bogor Station fall between 0.71 and 0.75.



Purwadhika
Startup and Coding School

Confidence Interval for Proportion

A confidence interval for a population proportion p, using the sample proportion $\hat{p}$ and the standard error $se = \sqrt{\hat{p}(1-\hat{p})/n}$ for sample size n, is

$$\hat{p} \pm z(se), \text{ which is } \hat{p} \pm z\sqrt{\hat{p}(1-\hat{p})/n}.$$

For 90%, 95%, and 99% confidence intervals, z equals 1.645, 1.96, and 2.58. This method assumes

- Data obtained by randomization (such as a random sample or a randomized experiment).
- A large enough sample size n so that the number of successes and the number of failures, that is, $n\hat{p}$ and $n(1-\hat{p})$, are both at least 15.

Confidence Interval for Mean

A 95% confidence interval for the population mean μ is

$$\bar{x} \pm t_{.025}(se), \text{ where } se = s/\sqrt{n}.$$

Here, $df = n - 1$ for the t-score $t_{.025}$ that has right-tail probability 0.025 (total probability 0.05 in the two tails and 0.95 between $-t_{.025}$ and $t_{.025}$). To use this method, you need

- Data obtained by randomization (such as a random sample or a randomized experiment).
- An approximately normal population distribution.

Confidence Interval

- Let’s take a further example from our previous Purwadhika student age. We already have the sample mean 25, the number of sample is 10, and if we calculate the Standard Error it would be 1.16.
- Often, we take the 95% confidence level, it means the **α** is 0.05 and we divide it by 2 and which would result 0.025 (this is for one-tail test). As the sample size is 10, we would have df equal to 9. Now, lets calculate the T-score from the table.
- If we put everything in the Confidence Interval of the mean, it would be 25 ± 2.2621*1.16. This give our confidence interval of Purwadhika Student age between 22.38 and 27.62, with 95% confidence and 5% chance the population parameter is outside of the range.
- This interval would be more accurate with more data

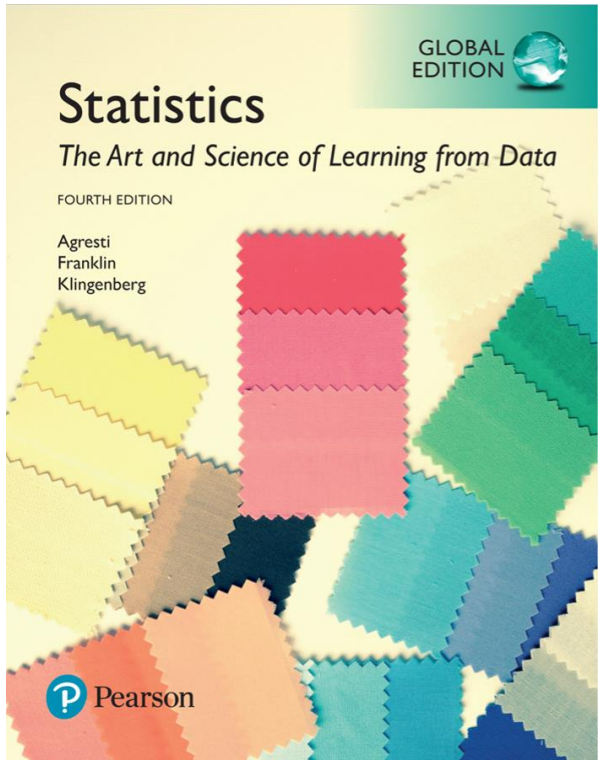
T Distribution Table

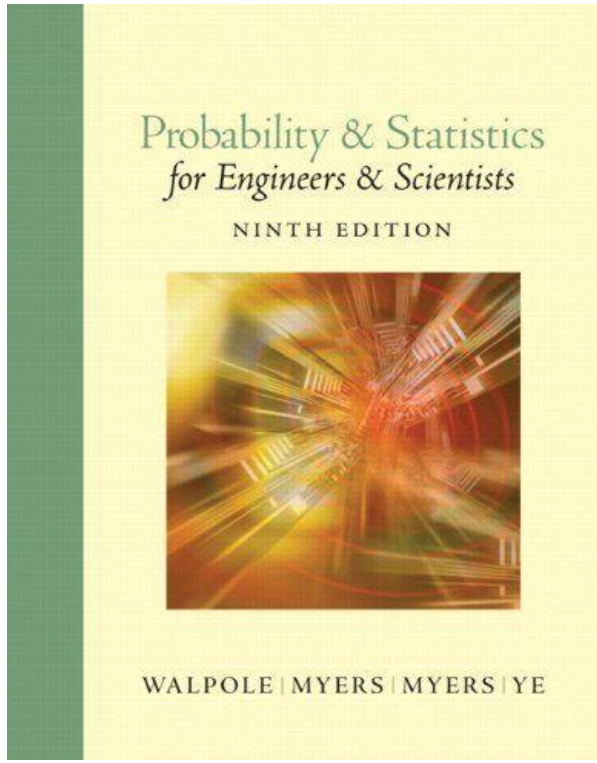
α (1 tail)	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
α (2 tail)	0.1	0.05	0.02	0.01	0.005	0.002	0.001
df							
1	6.3138	12.7065	31.8193	63.6551	127.3447	318.4930	636.0450
2	2.9200	4.3026	6.9646	9.9247	14.0887	22.3276	31.5989
3	2.3534	3.1824	4.5407	5.8408	7.4534	10.2145	12.9242
4	2.1319	2.7764	3.7470	4.6041	5.5976	7.1732	8.6103
5	2.0150	2.5706	3.3650	4.0322	4.7734	5.8934	6.8688
6	1.9432	2.4469	3.1426	3.7074	4.3168	5.2076	5.9589
7	1.8946	2.3646	2.9980	3.4995	4.0294	4.7852	5.4079
8	1.8595	2.3060	2.8965	3.3554	3.8325	4.5008	5.0414
9	1.8331	2.2621	2.8214	3.2498	3.6896	4.2969	4.7809
10	1.8124	2.2282	2.7638	3.1693	3.5814	4.1437	4.5869
11	1.7959	2.2010	2.7181	3.1058	3.4966	4.0247	4.4369

Our T-score

Purwadhika
Startup and Coding School

Reference





Purwadhika
Startup and Coding School

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/

Purwadhika
Startup and Coding School