Confidence Intervals

Parameter Vs Estimate

Aspect	Parameter	Estimate
Nature	A true, fixed value (unknown in practice).	A computed approximation based on sample data.
Source	Describes the entire population or model.	Derived from a sample, which is a subset of the population.
Examples	μ,σ2,β1 (true model slope)	$\bar{x}$, s^2, $\hat{\beta}_1$(sample mean, sample variance, estimated slope).
Variability	Constant (does not vary from sample to sample).	Varies depending on the sample drawn.
Usage in Inference	The target of estimation and hypothesis testing.	Used to make inferences or predictions about the parameter.
Measurement	Cannot be directly measured in most cases.	Computed using statistical methods from available data.

- Estimate is also known as **statistic.**

Inferential Statistics

- Inferential statistics is a branch of statistics that focuses on drawing conclusions about a population based on a sample of data.

Point Estimate

- **Single numerical value** that is calculated from sample data and is used as the best guess or approximation of an unknown population parameter.
- **"best guess"**

- Derived from Sample Data

Confidence Interval (CI)

- A confidence interval is a range of values, calculated from sample data, within which the true value of a population parameter is likely to fall.
- For example, a 95% confidence interval for a population mean suggests that **if we repeated our sampling many times, about 95% of those intervals would contain the true mean**.
    - A 95% confidence interval for the population mean might be [50,60], meaning we are 95% confident that the true mean lies between 50 and 60.

> 💡 ***A 95% CI does not mean there is a 95% probability that the true parameter lies within the interval. Instead, it means that 95% of such intervals would contain the true parameter if the sampling process were repeated.***

Margin of Error:

- The amount added and subtracted from the sample statistic to create the interval.
- Depends on the standard error and the critical value (e.g., z-score or t-score).

Formula for CI:

$$\text{Probability} = \frac{1 + \text{confidence\_level}}{2} = \frac{1 + 0.95}{2} = 0.975$$

Confidence Interval = Point Estimate ± Margin of Error

- Point estimate → Mean

1. **Confidence Interval for the Mean (Known Population Variance):**

$$\text{CI} = \bar{x} \pm z \cdot \frac{\sigma}{\sqrt{n}}$$

   - $\bar{x}$: Sample mean.
   - z: Critical value from the standard normal distribution (e.g., 1.96 for 95% confidence).
   - σ: Population standard deviation.
   - n: Sample size.

2. **Confidence Interval for the Mean (Unknown Population Variance):**

$$\text{CI} = \bar{x} \pm t \cdot \frac{s}{\sqrt{n}}$$

   - s: Sample standard deviation.
   - t: Critical value from the t-distribution (depends on degrees of freedom, $n - 1$).

z

- For a 95% confidence level, z∗ is approximately **1.96**.
- For a 90% confidence level, z∗ is approximately **1.645**.
- For a 99% confidence level, z∗ is approximately **2.576**.

$$\text{SE} = \frac{s}{\sqrt{n}}$$

Confidence level:

Usually expressed as a percentage like **95%**, indicates how sure we are that the true value lies within the interval.

- **Confidence Interval is created for Parameters and not statistics.**

Ways to calculate CI

1. z procedure → When you have SD of population
2. t procedure → when you don't have SD of population

Confidence Interval (Sigma Known)

- **When you draw a sample of n numbers 100 times, the mean of population will fall in this range 95 times.**

- Generally, you don't have population SD.

Assumptions:

- **Random** sampling
- Known **population standard deviation**
- **Normal distribution** or **large sample size (>30)**
  - If not normally distributed → Apply CLT

1. **Confidence Interval for the Mean (Known Population Variance):**

$$\text{CI} = \bar{x} \pm z \cdot \frac{\sigma}{\sqrt{n}}$$

   - $\bar{x}$: Sample mean.
   - z: Critical value from the standard normal distribution (e.g., 1.96 for 95% confidence).
   - σ: Population standard deviation.
   - n: Sample size.

- Point estimate → Mean ($\bar{x}$)

$$\text{Probability} = \frac{1 + \text{confidence\_level}}{2} = \frac{1 + 0.95}{2} = 0.975$$

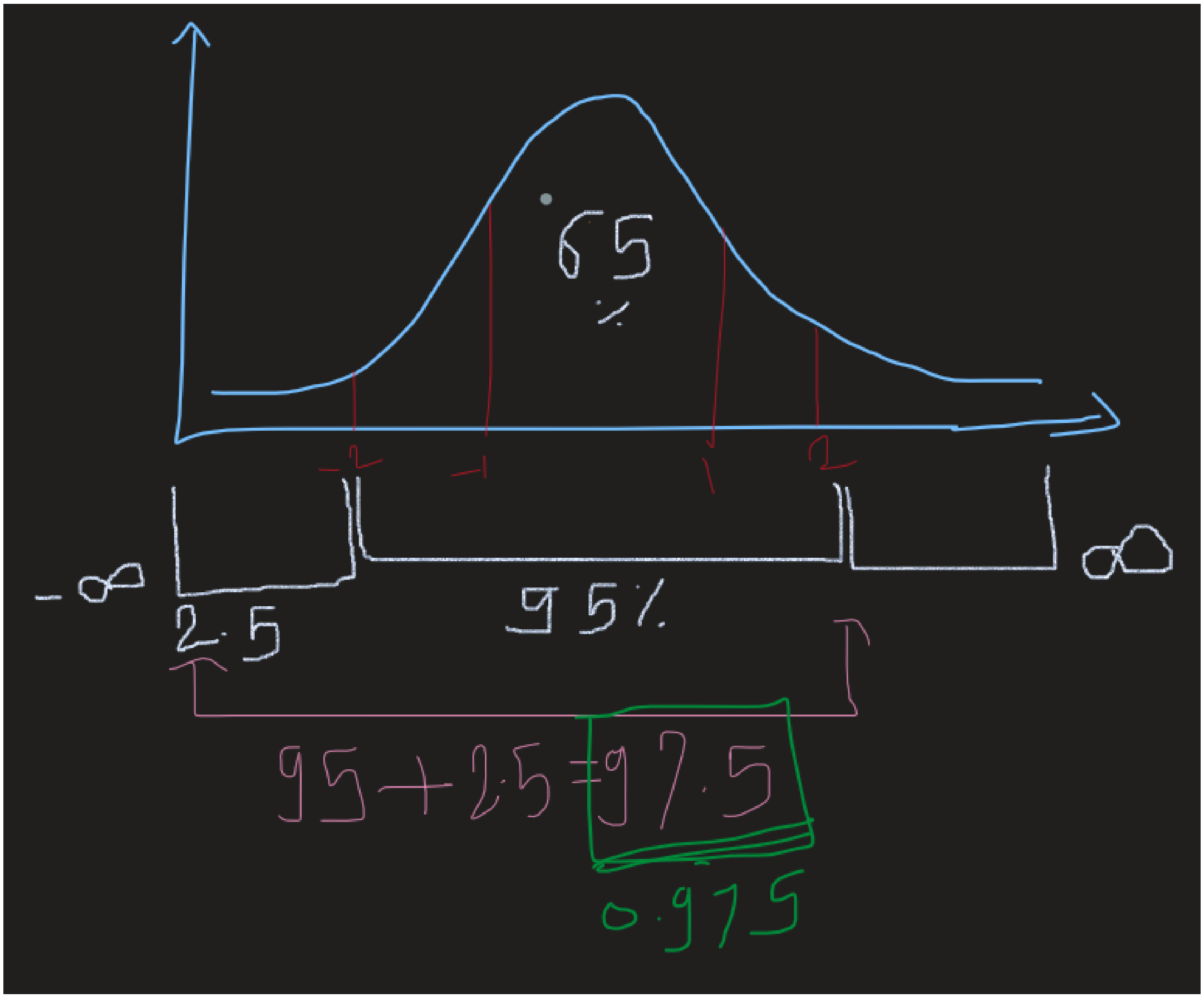
- We fan look for this probability in the Z-table to find out the z (critical value)

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361

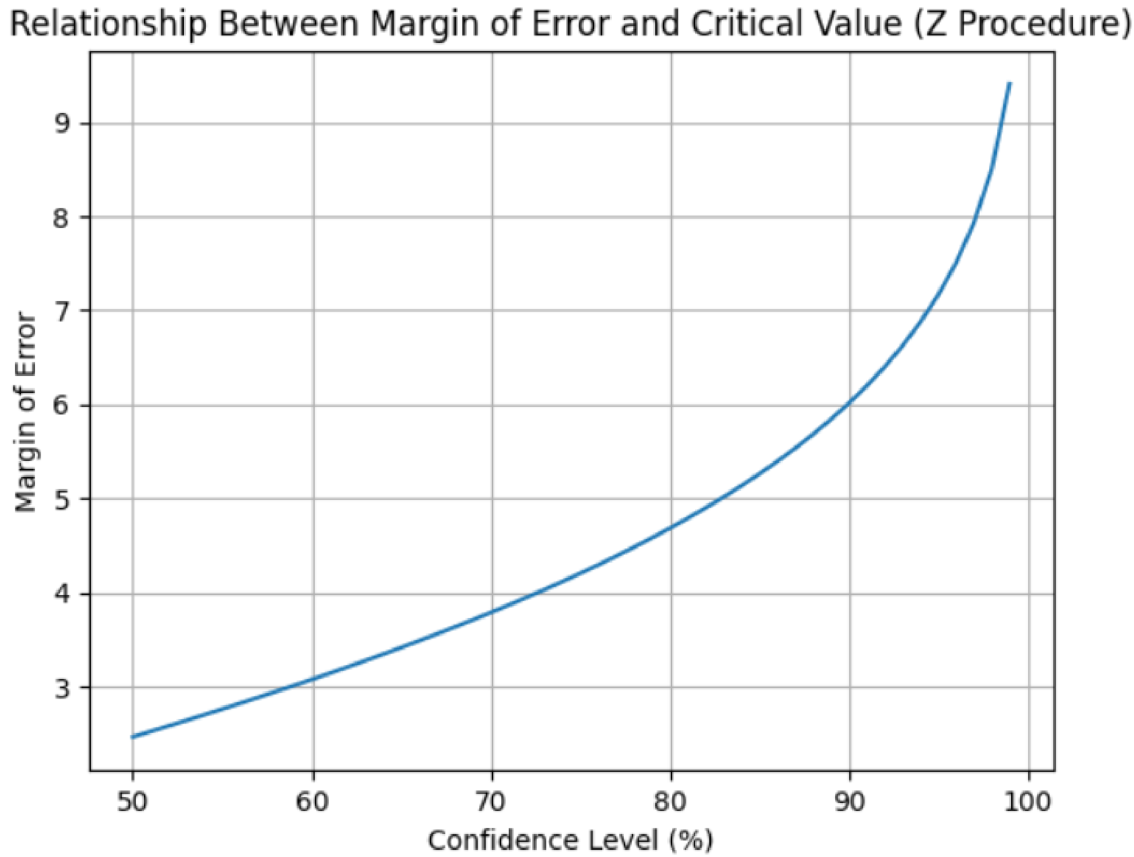
z = 1.96

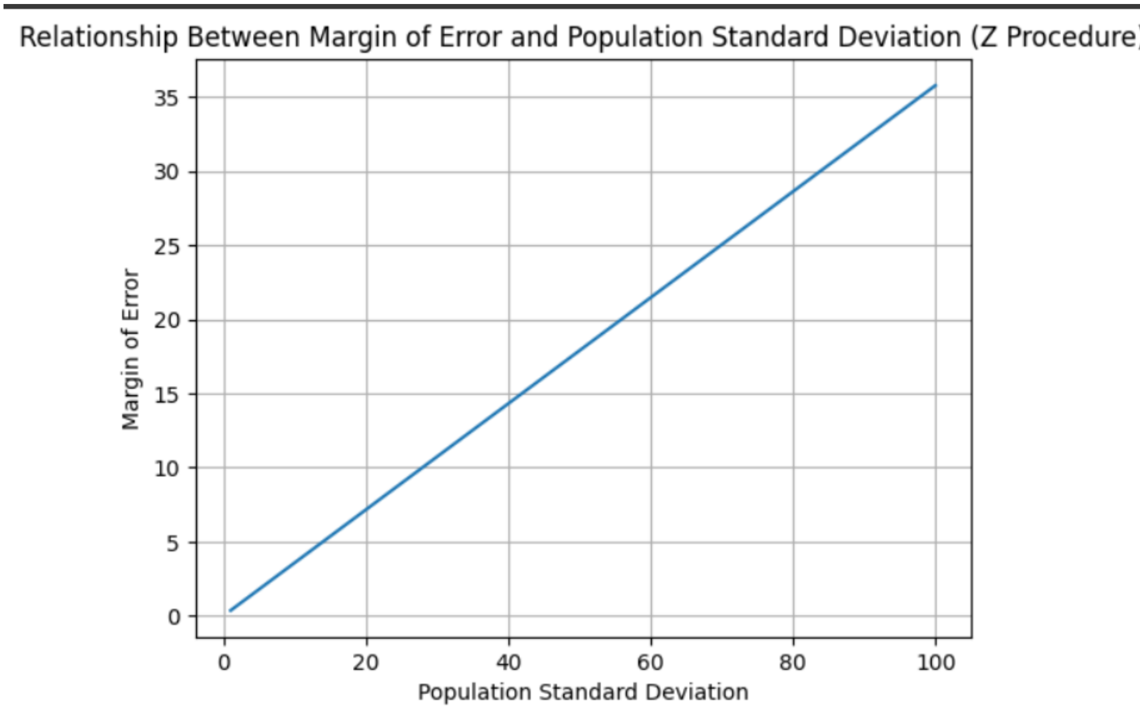
How did we get to 0.975?

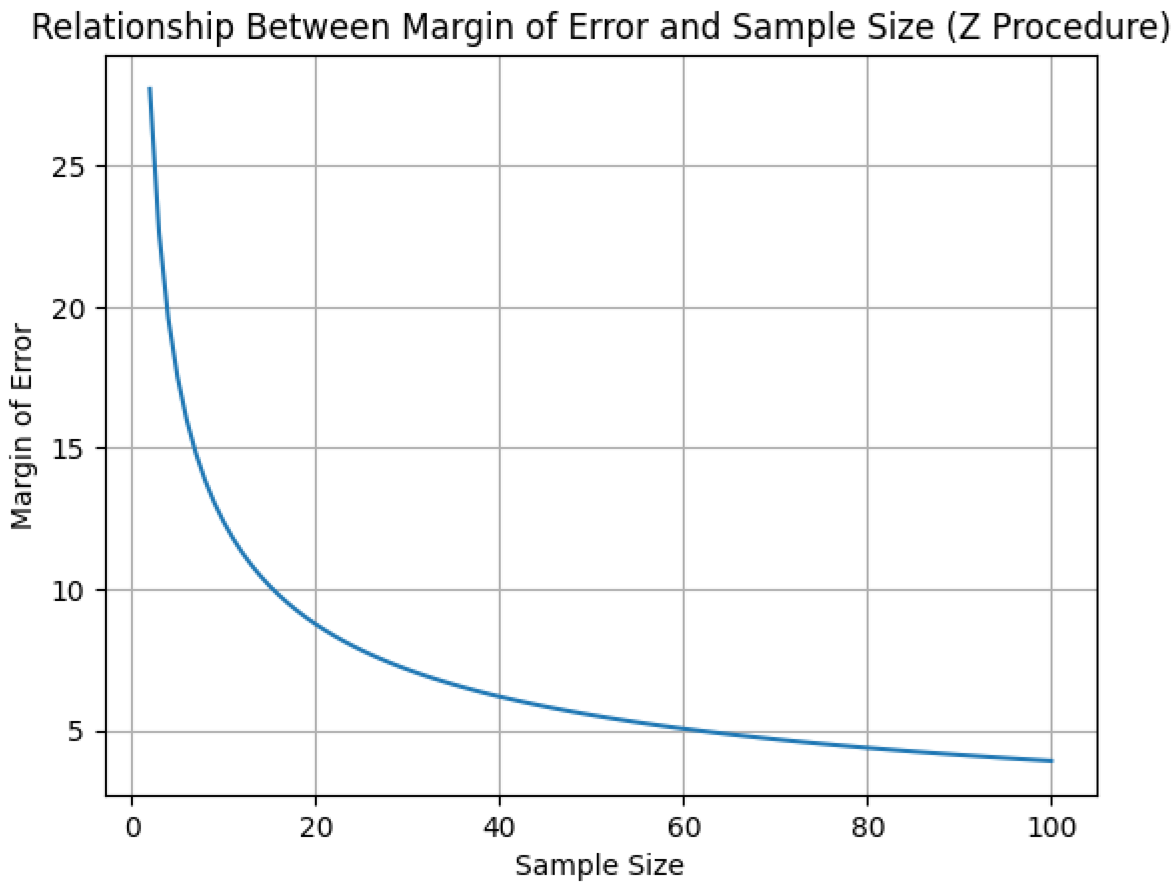


Factors Affecting Margin of Error

- Confidence Level (1-alpha)
- Sample Size
- Population Standard Deviation


Relationship Between Margin of Error and Critical Value (Z Procedure)
Margin of Error
Confidence Level (%)
3
4
5
6
7
8
9
50
60
70
80
90
100


Relationship Between Margin of Error and Population Standard Deviation (Z Procedure)
Margin of Error
Population Standard Deviation
0
5
10
15
20
25
30
35
0
20
40
60
80
100



Confidence Interval (Sigma not known) (t-Procedure)

> 💡 The **t-distribution** is similar to the **normal distribution** (bell-shaped and symmetric), but it has **heavier tails**

- The t-distribution accounts for this added uncertainty by making the tails "wider," allowing for larger margin of errors and more conservative estimates of the confidence interval.

Assumptions:

1. Random sampling
2. Sample standard deviation
3. Approximately normal distribution
4. Independent observations
5. Could be useful in sample size less than 30

Here, we don't have a population mean → So, take sample mean

Formula:

$$\mathrm{CI} = \bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

- $\bar{x}$ = sample mean
- $t_{\alpha/2}$ = critical value from the **t-distribution** for a given confidence level (based on the degrees of freedom)
- s = sample standard deviation
- n = sample size
- α = significance level, where $1 - \alpha$ is the confidence level (e.g., for a 95% confidence interval, $\alpha = 0.05$)

- When we use s, the distribution is called → **Student's t-distribution**

- It is similar to normal distribution, but not a normal distribution

Steps to Calculate a Confidence Interval when the Population Mean is Unknown:

1. Calculate the sample mean $\bar{x}$ and sample standard deviation s.
2. Find the t-critical value $t_{\alpha/_2}$ from the t-distribution. This depends on:
    - The confidence level (e.g., for 95%, α=0.05, and you'll use $t_{0.025}$.
    - The **degrees of freedom**, which is df=n−1, where n is the sample size.
    - you would look up the value in a t-distribution table
        - If n=30, look for→ df=29
        - The **t-critical value** is approximately 2.045.
3. Compute the margin of error

$$t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}.$$

4. **Construct the confidence interval**: $\bar{x} \pm$ margin of error.

```
import numpy as np
import scipy.stats as stats

# Sample data
data = np.array([10, 12, 14, 15, 16, 18, 20, 22, 24, 25])  # Replace with your data
confidence_level = 0.95  # 95% confidence level

# Calculate sample statistics
```

```
n = len(data)  # Sample size
sample_mean = np.mean(data)  # Sample mean
sample_std = np.std(data, ddof=1)  # Sample standard deviation (ddof=1 for n-1)

# Find the critical value (t-score)
degrees_of_freedom = n - 1
critical_value = stats.t.ppf((1 + confidence_level) / 2, df=degrees_of_freedom)

# Calculate the margin of error
margin_of_error = critical_value * (sample_std / np.sqrt(n))

# Construct the confidence interval
ci_lower = sample_mean - margin_of_error
ci_upper = sample_mean + margin_of_error

print(f"Sample Mean: {sample_mean:.2f}")
print(f"Sample Standard Deviation: {sample_std:.2f}")
print(f"Critical Value (t): {critical_value:.2f}")
print(f"Margin of Error: {margin_of_error:.2f}")
print(f"Confidence Interval: [{ci_lower:.2f}, {ci_upper:.2f}]")
```

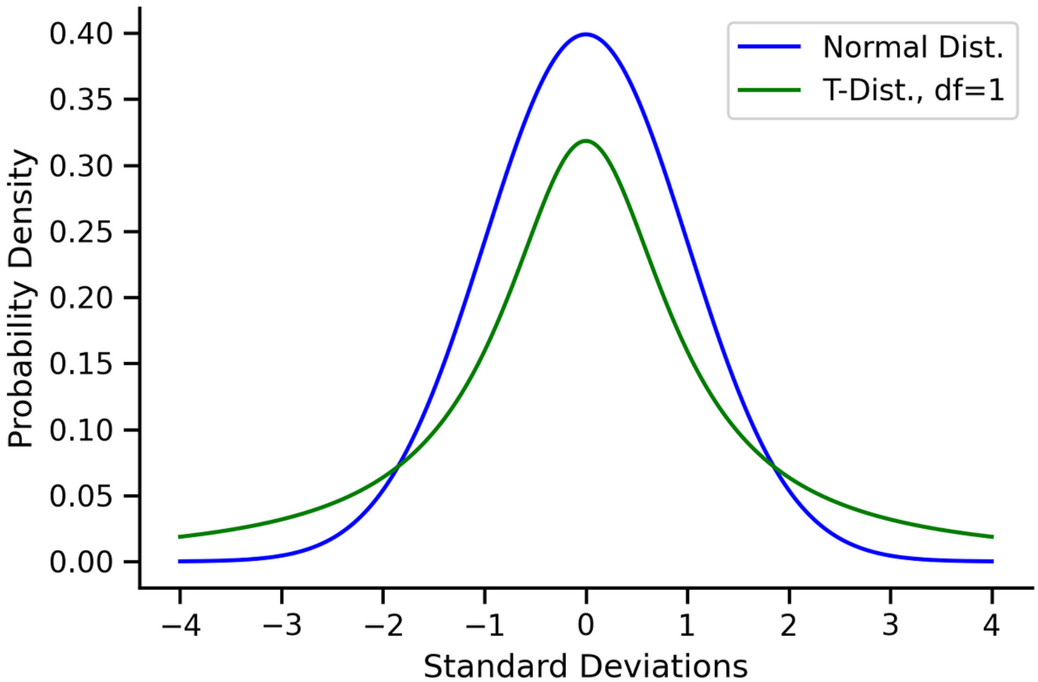
`ddof=1` → Degree of freedom =1.

- This is used because we divide by $n-1$ instead of n in case of sample SD

`stats.t.ppf(probability, df)` → Calculates the t-critical value

$$\text{Probability} = \frac{1 + \text{confidence\_level}}{2} = \frac{1 + 0.95}{2} = 0.975$$

Normal vs Student’s t-distribution:



- In Student’s t-distribution, **tails are fatter.**