Inferential Statistics
Part-2

Sourav Karmakar

souravkarmakar29@gmail.com

What shall we learn?

- Estimator and estimates
- Two types of estimates
- Confidence Interval
- Confidence intervals with known population variance
- Statistical measurement: skewness and kurtosis
- Student’s T distribution
- Confidence intervals with unknown population variance
- Hypothesis testing
- P-value
- Z-test
- One Sample T-test
- Two sample T-test with dependent means (paired samples)
- Two sample T-test with independent means
- Chi-square test for categorical variable

Estimator and Estimates

Estimator and Estimate: Estimator provides an estimate of the certain population parameter solely on sample information. A specific value of an estimator is called and estimate.
For an estimator to be good it should have low bias and low variance.

For example: Sample Mean is an estimator of the population mean.

The formula for calculating sample mean:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

And, we can show that $\bar{x}$ is an un-biased estimator of population mean (μ). i.e.$\mathbb{E}[\bar{x}] = \mu$

The formula for calculating sample standard deviation:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

We can also show that sample variance is an unbiased estimator of population variance. i.e. $\mathbb{E}[s^2] = \sigma^2$

Two types of Estimates

There are two types of estimates:

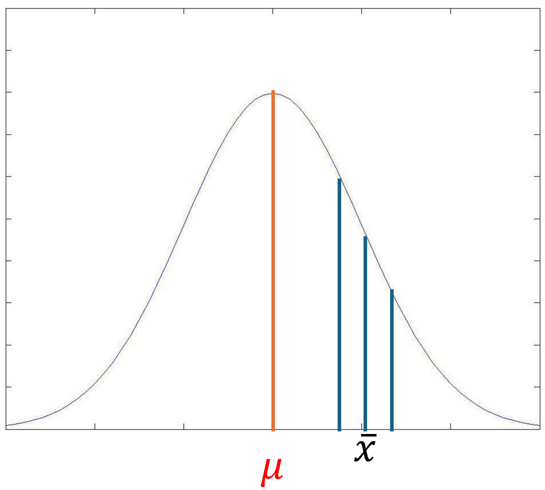
1. **Point estimate:** a single number computed from the sample data.
2. **Confidence Interval estimate:** an interval computed from the sample data within which the population parameter lies with certain confidence.

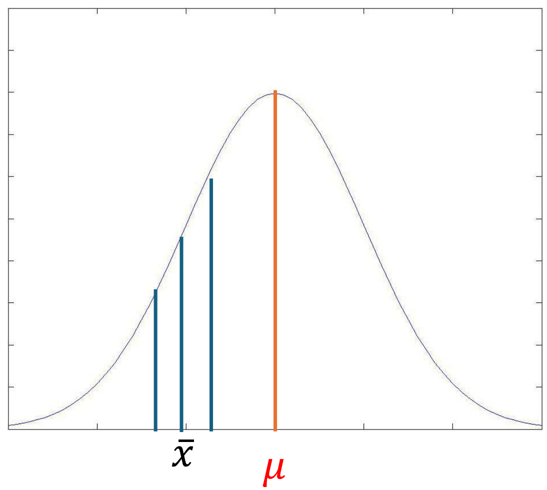
These two estimates are closely related. In fact,

$$confidence\ interval = point\ estimate \pm Margin\ of\ error.$$

Why point estimate is not enough?

Because of the inherent variability of the samples collected, the point estimates often varies widely from the true population parameter.





Confidence Interval

It is the range / interval estimated from the sample data. The population parameter lies within this interval with certain **confidence**.

Example: Say for a given sample the sample mean is 4 and the 95% confidence interval is [2.5, 5.5]. Here the confidence level is 95%. i.e. the probability that the true population mean lies within 2.5 to 5.5 is 95%.

Note:
1. You can never be 100% confident: because that means we know the entire population data (which is impossible / impractical in most cases)
2. When you say you are 95% confident, there is still a 5% chance that the population parameter is out of the interval.

Confidence Interval depends on: (1) Sample size , (2) How much confidence do we need (confidence level).

$$Confidence\ Level = 1 - \alpha$$

Where, α is known as the significance level.

In statistics, the significance level is decided even before the testing begins. Usually we fix the significance level based on the certainty we want in the output of the experiment and it depends on the use case.

Confidence Interval

Confidence Level (CL)	Significance Level (α)
90%	0.1
95%	0.05
99%	0.01

Confidence interval is calculated using the following formula

$$Confidence\ Interval = Point\ Estimate \pm Margin\ of\ error$$

The margin of error is calculated using the following formula:

$$Margin\ of\ error = Reliability\ factor \times standard\ error$$

The $reliability\ factor$ depends on the confidence level (or significance level). Higher the confidence level more will be the reliability factor for a given statistic.

The $standard\ error$ depends on the number of samples. Higher the sample size, less will be the standard error.

Confidence Interval with known population variance

Suppose, we want to estimate mean of a population (μ).

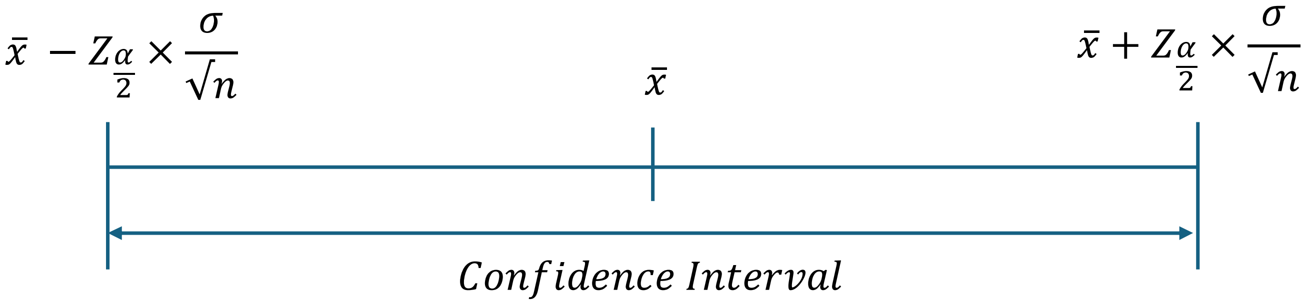
We collect a sample of sample size = n and calculate the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Now, we decide how much confidence we want. Say we want 95% confidence. So, $Confidence\ Level\ =\ 95\%$.

Then, we determine the significance level (α). Say, $\alpha = 0.05$

Now, lets assume the population variance (σ^2) is known. Then, the standard error , $SE = \frac{\sigma}{\sqrt{n}}$

The reliability factor is computed using the $z-$table. $Reliability\ factor = Z_{\alpha/2}$



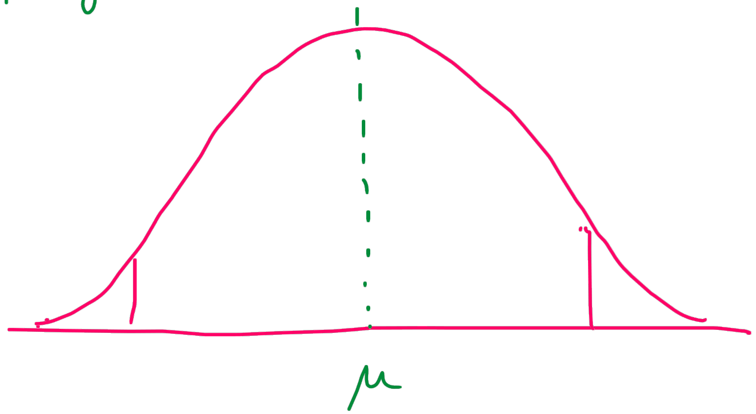
Why the reliability factor = $Z_{\alpha/2}$

$\mathbb{E}[\bar{x}] = \mu$

Central Limit Theorem

Sampling distribution

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$



$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad \therefore \quad z \sim N(0,1)$$

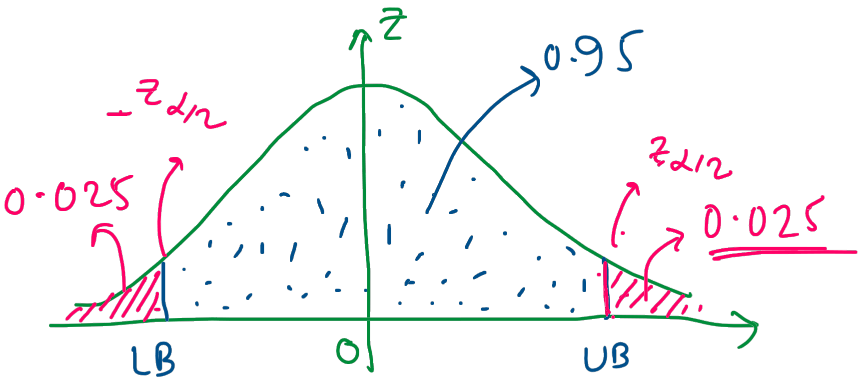
$\alpha = 0.05$

$\alpha/2 = 0.025$

95% confidence

$$LB = -z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} + \bar{x}$$

$$UB = z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} + \bar{x}$$



Interpreting Z - table

Z - Table

→ 2^{nd} decimal

first decimal ↓

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

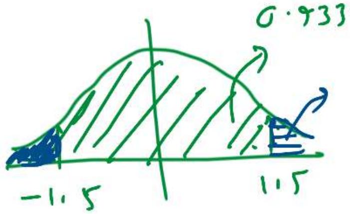
Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.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
→ 2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
3.0	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900
3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99996	.99997	.99997

standard normal distribution

$X \sim N(0,1)$

pdf : $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

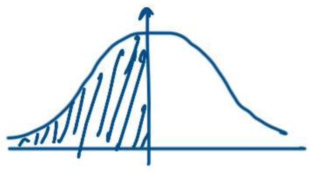
cdf : $P(x \leq k) = \int_{-\infty}^{k} f(x)\, dx$



$1 - 0\cdot9332$
$= \underline{0\cdot0668}$

$P(x \leq z) = \int_{-\infty}^{z} f(x)\, dx$: z table value

$z = 0\cdot00$ $\therefore P(x \leq 0\cdot00) = 0\cdot5$

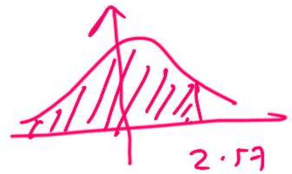


$z = 1\cdot64$

$P(x \leq 1\cdot64) = 0\cdot9495$

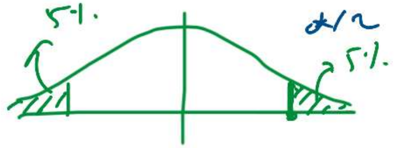
$z = 2\cdot5\underline{7}$

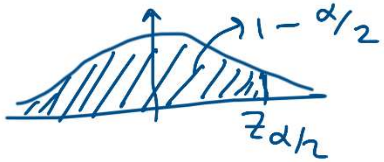
$\therefore P(x \leq 2\cdot57) = 0\cdot99492$



$\underline{P(x \leq -1\cdot5)} = 1 - P(x \leq 1\cdot5) = 0\cdot0668$

How to calculate $Z_{\alpha/2}$





$z(1.64) = 0.9495$
$z(1.65) = 0.9505$

Z - Table

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
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
3.0	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900
3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99996	.99997	.99997

Confidence Level (CL)	Significance Level (α)	$\frac{\alpha}{2}$	Z-table value $(1-\frac{\alpha}{2})$	$Z_{\frac{\alpha}{2}}$
90%	0.1	0.05	0.95	1.65
95%	0.05	0.025	0.975	1.96
99%	0.01	0.005	0.995	2.58

90% confidence interval: $\bar{x} \pm 1.65 \times \frac{\sigma}{\sqrt{n}}$

95% confidence interval: $\bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$

99% confidence interval: $\bar{x} \pm 2.58 \times \frac{\sigma}{\sqrt{n}}$

Calculate confidence interval with known population variance

You take a random sample of 100 light bulbs and find the average lifespan is 58 days, with a known population standard deviation of 8 days. Construct 90%, 95%, and 99% confidence interval.

$$CI = \bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$\bar{x} = 58$ days, $\sigma = 8$, $n = 100$

$$\text{Standard Error} = \frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{100}} = \frac{8}{10} = 0.8$$

CL	$Z_{\alpha/2}$	ME	LCB	UCB
90%	1.65	$1.65 \times 0.8 = 1.32$	56.68	59.32
95%	1.96	$1.96 \times 0.8 = 1.57$	56.43	59.57
99%	2.58	$2.58 \times 0.8 = 2.06$	55.94	60.06

Relationship between confidence interval and sample size

The confidence interval around a point estimate: $CI = \bar{x} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$, Margin of error = $Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

If we fix the confidence level (say 95%) , the reliability factor i.e. $Z_{\frac{\alpha}{2}}$ is fixed.

The standard error ($\frac{\sigma}{\sqrt{n}}$) depends on the sample size. So, the overall Margin of Error depends on sample size for fixed Confidence level.

$$margin\ of\ error \propto \frac{1}{\sqrt{n}}$$

So, increase in sample size decreases Margin of error and confidence interval for the same confidence level.

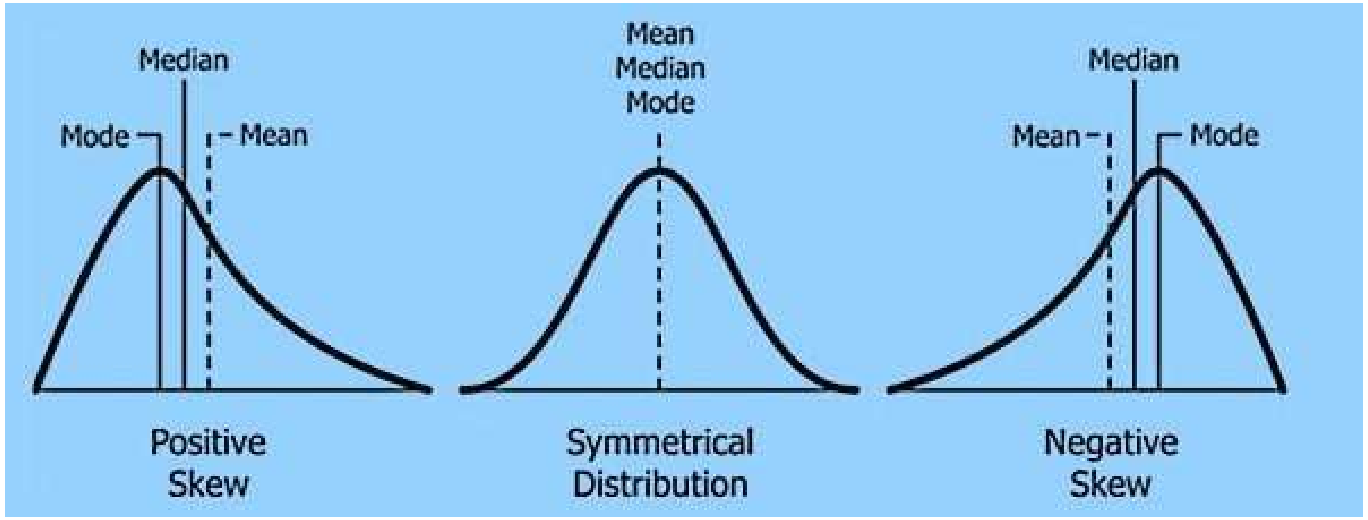
And decrease in sample size increases Margin of error and confidence interval for the same confidence level.

Skewness

Skewness is a statistical measure that describes the asymmetry or lack of symmetry in the distribution of data. It quantifies the degree and direction of deviation from a perfectly symmetric distribution, such as a normal distribution.

Positive vs. Negative Skew

1. **Positive Skew (Right Skew):** In a positively skewed distribution, the tail on the right-hand side (the larger values) is longer or fatter than the left-hand side (the smaller values). The mean is typically greater than the median, and the distribution has a longer right tail.
2. **Negative Skew (Left Skew):** In a negatively skewed distribution, the tail on the left-hand side (the smaller values) is longer or fatter than the right-hand side (the larger values). The mean is typically less than the median, and the distribution has a longer left tail.



Skewness

There are various ways to calculate skewness:

1. **Pearson's first coefficient of Skewness ($\boldsymbol{SK_1}$):** It uses mean, mode and standard deviation to calculate skewness

$$SK_1 = \frac{Mean\ - Mode}{Standard\ Deviation}$$

2. **Pearson's second coefficient of Skewness ($\boldsymbol{SK_2}$):** It uses mean, median and standard deviation

$$SK_2 = \frac{3(Mean\ - Median)}{Standard\ Deviation}$$

3. **Bowley's coefficient of Skewness ($\boldsymbol{SK_B}$):** It uses quartile values

$$SK_B = \frac{Q_1 + Q_3\ - 2Q_2}{Q_3 - Q_1}$$

If the skewness value computed above is between -0.5 to 0.5 then the distribution is approximately symmetric.
If it is between -1 to 0.5 or 0.5 to 1, then moderately skewed.
If it is less than -1 or more than 1, then highly skewed.

Skewness

There are other more robust method exist for measuring the skewness of a distribution. These are based on the standardized third moment about the mean.

For Population: $\gamma_1 = \frac{\mathbb{E}\left[(X-\mu)^3\right]}{\mathbb{E}\left[(X-\mu)^2\right]^{\frac{3}{2}}}$ $, where\ \mu\ is\ the\ population\ mean$

For sample: $g_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s}\right)^3$

Where,

- x_i is the i^{th} data point.
- $\bar{x}$ is the sample mean.
- s is the sample standard deviation.
- n is the sample size.

Kurtosis

Kurtosis is a statistical measure that quantifies the degree to which a probability distribution deviates from a normal distribution in terms of the shape and concentration of data in the tails. It provides insights into the presence of outliers and the **"tailed-ness"** of the distribution.

Kurtosis primarily measures the "tailed-ness" and not just the "peaked-ness" of a distribution. A distribution can have a high peak but also heavy tails, or a flat peak with light tails. The tails are the tapering ends on either side of the distribution, representing the extreme values.

Kurtosis is calculated using the formula: $\beta = \dfrac{\mathbb{E}\,[(X-\mu)^4]}{\mathbb{E}\,[(X-\mu)^2]^2}$ $, where\ \mu\ is\ the\ population\ mean$

The normal distribution has Kurtosis = 3 , when calculated using above formula. Hence statistician often used **excess kurtosis**: $\gamma_2 = \beta - 3$ as the measure.

Excess Kurtosis for Population: $\gamma_2 = \dfrac{\mathbb{E}\,[(X-\mu)^4]}{\mathbb{E}\,[(X-\mu)^2]^2} - 3$ $, where\ \mu\ is\ the\ population\ mean$

Excess Kurtosis for sample:

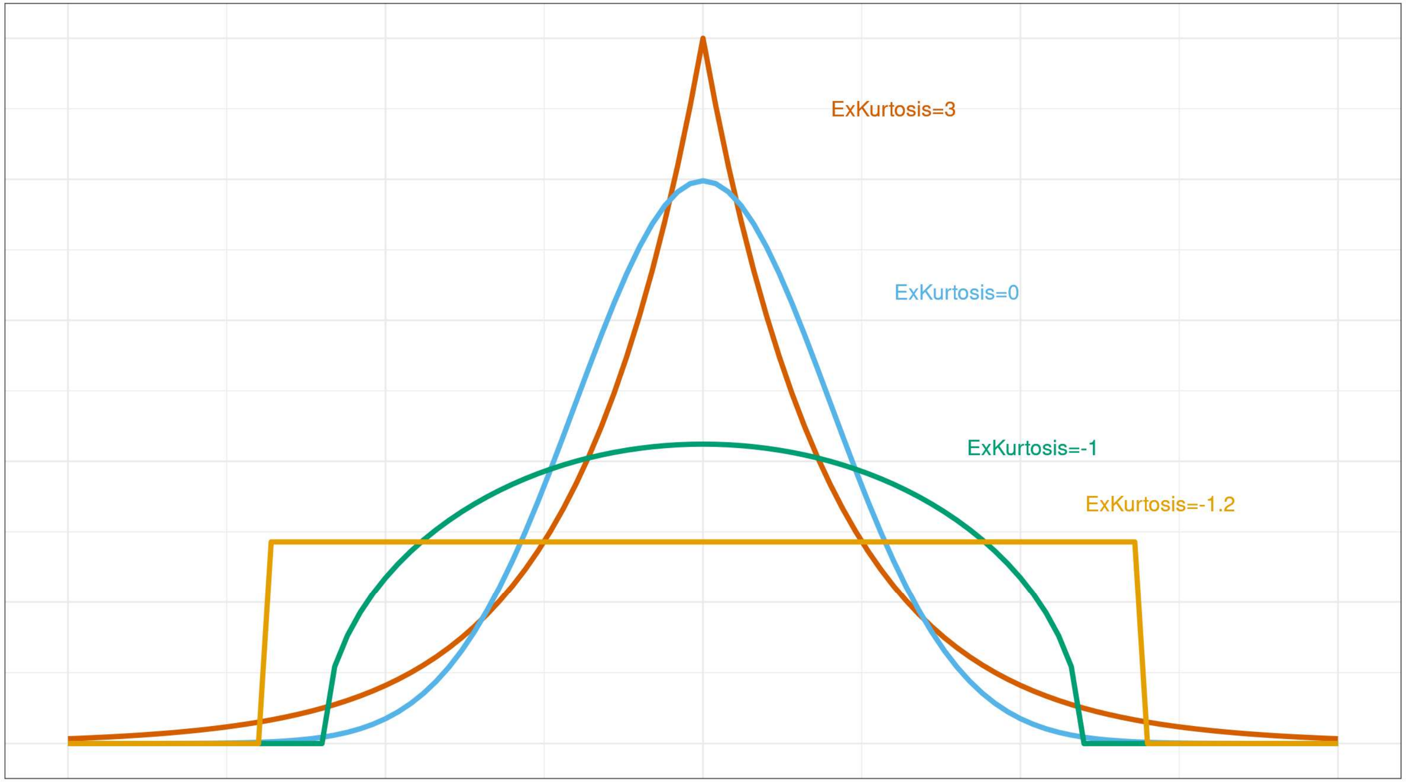
$$g_2 = \left[\frac{n\,(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s}\right)^4\right] - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Where,

- x_i is the i^{th} data point.
- $\bar{x}$ is the sample mean.
- s is the sample standard deviation.
- n is the sample size.

Kurtosis

1. **Mesokurtic Distribution** (excess Kurtosis ≈ 0): This distribution has a "medium" level of tailed- ness, similar to a normal (Gaussian) distribution.
2. **Leptokurtic Distribution** (excess Kurtosis > 0): This distribution has heavy tails. That indicates higher probability of extreme values (outliers)
3. **Platykurtic Distribution** (excess Kurtosis < 0): This distribution has light tails. It indicates a lower probability of extreme values.

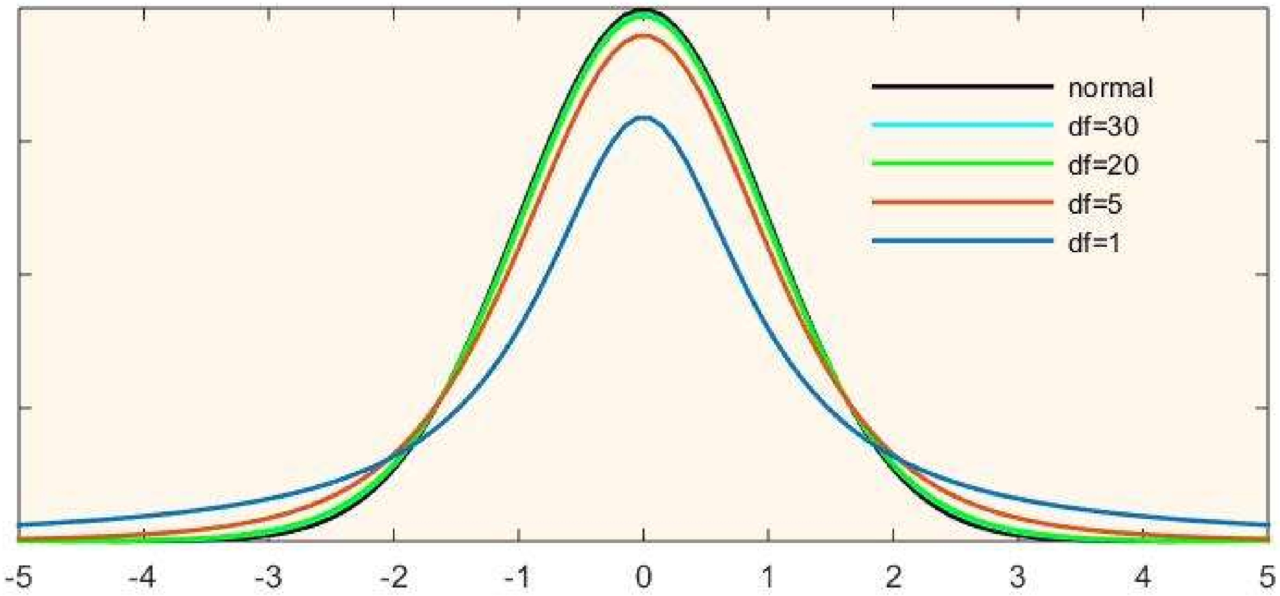


Student’s t-distribution and its application

A t-distribution, also known as Student's t-distribution, is a probability distribution that is used in statistics for making inferences about **population means when the sample size is small**, and the population's **standard deviation is unknown.**

[*Trivia: Student is the pen name of English statistician William Sealy Gosset*]

Visually the t-distribution looks much similar to normal distribution but has fatter tails (leptokurtic), allows for higher dispersion and variance associated with smaller sample size. For different degrees of freedom, the t-distribution looks following. Notice that with higher degrees of freedom t-distribution is almost similar to normal distribution.



Confidence Interval with unknown population variance

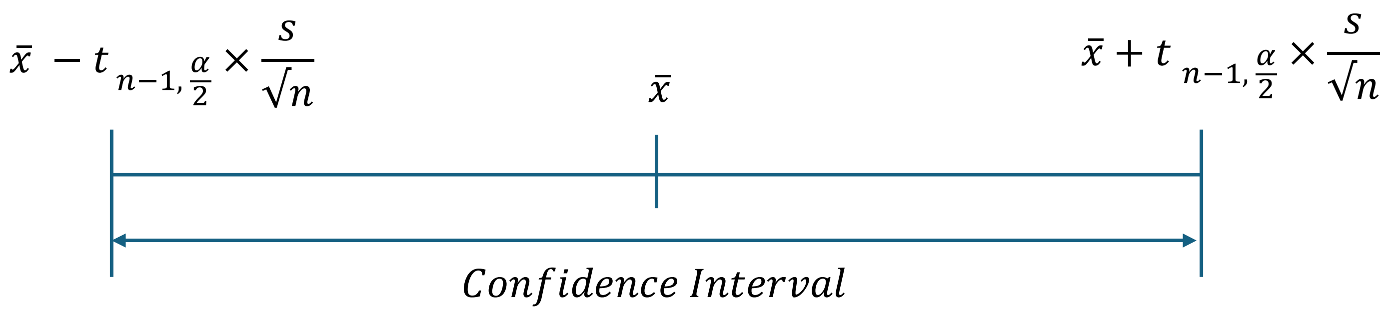
We collect a sample of sample size = n and calculate the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Then we calculate sample standard deviation using the formula: $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x - \bar{x})^2}$

Now, we decide how much confidence we want. Say we want 95% confidence. So, $Confidence\ Level = 95\%$.

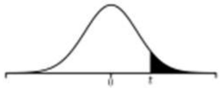
Then, we determine the significance level (α). Say, $\alpha = 0.05$

Then we find out the value of $t_{n-1,\frac{\alpha}{2}}$ from the t-table and calculate the confidence interval as follows.



How to calculate $t_{n-1\,,\,\alpha/2}$

t-table



Critical Values for Student's t-Distribution.

	Upper Tail Probability: $\Pr(T > t)$									
df	0.2	0.1	0.05	0.04	0.03	0.025	0.02	0.01	0.005	0.0005
1	1.376	3.078	6.314	7.916	10.579	12.706	15.895	31.821	63.657	636.619
2	1.061	1.886	2.920	3.320	3.896	4.303	4.849	6.965	9.925	31.599
3	0.978	1.638	2.353	2.605	2.951	3.182	3.482	4.541	5.841	12.924
4	0.941	1.533	2.132	2.333	2.601	2.776	2.999	3.747	4.604	8.610
5	0.920	1.476	2.015	2.191	2.422	2.571	2.757	3.365	4.032	6.869
6	0.906	1.440	1.943	2.104	2.313	2.447	2.612	3.143	3.707	5.959
7	0.896	1.415	1.895	2.046	2.241	2.365	2.517	2.998	3.499	5.408
8	0.889	1.397	1.860	2.004	2.189	2.306	2.449	2.896	3.355	5.041
9	0.883	1.383	1.833	1.973	2.150	2.262	2.398	2.821	3.250	4.781
10	0.879	1.372	1.812	1.948	2.120	2.228	2.359	2.764	3.169	4.587
11	0.876	1.363	1.796	1.928	2.096	2.201	2.328	2.718	3.106	4.437
12	0.873	1.356	1.782	1.912	2.076	2.179	2.303	2.681	3.055	4.318
13	0.870	1.350	1.771	1.899	2.060	2.160	2.282	2.650	3.012	4.221
14	0.868	1.345	1.761	1.887	2.046	2.145	2.264	2.624	2.977	4.140
15	0.866	1.341	1.753	1.878	2.034	2.131	2.249	2.602	2.947	4.073
16	0.865	1.337	1.746	1.869	2.024	2.120	2.235	2.583	2.921	4.015
17	0.863	1.333	1.740	1.862	2.015	2.110	2.224	2.567	2.898	3.965
18	0.862	1.330	1.734	1.855	2.007	2.101	2.214	2.552	2.878	3.922
19	0.861	1.328	1.729	1.850	2.000	2.093	2.205	2.539	2.861	3.883
20	0.860	1.325	1.725	1.844	1.994	2.086	2.197	2.528	2.845	3.850
21	0.859	1.323	1.721	1.840	1.988	2.080	2.189	2.518	2.831	3.819
22	0.858	1.321	1.717	1.835	1.983	2.074	2.183	2.508	2.819	3.792
23	0.858	1.319	1.714	1.832	1.978	2.069	2.177	2.500	2.807	3.768
24	0.857	1.318	1.711	1.828	1.974	2.064	2.172	2.492	2.797	3.745
25	0.856	1.316	1.708	1.825	1.970	2.060	2.167	2.485	2.787	3.725
26	0.856	1.315	1.706	1.822	1.967	2.056	2.162	2.479	2.779	3.707
27	0.855	1.314	1.703	1.819	1.963	2.052	2.158	2.473	2.771	3.690
28	0.855	1.313	1.701	1.817	1.960	2.048	2.154	2.467	2.763	3.674
29	0.854	1.311	1.699	1.814	1.957	2.045	2.150	2.462	2.756	3.659
30	0.854	1.310	1.697	1.812	1.955	2.042	2.147	2.457	2.750	3.646
31	0.853	1.309	1.696	1.810	1.952	2.040	2.144	2.453	2.744	3.633
32	0.853	1.309	1.694	1.808	1.950	2.037	2.141	2.449	2.738	3.622
33	0.853	1.308	1.692	1.806	1.948	2.035	2.138	2.445	2.733	3.611
34	0.852	1.307	1.691	1.805	1.946	2.032	2.136	2.441	2.728	3.601
35	0.852	1.306	1.690	1.803	1.944	2.030	2.133	2.438	2.724	3.591
36	0.852	1.306	1.688	1.802	1.942	2.028	2.131	2.434	2.719	3.582
37	0.851	1.305	1.687	1.800	1.940	2.026	2.129	2.431	2.715	3.574
38	0.851	1.304	1.686	1.799	1.939	2.024	2.127	2.429	2.712	3.566
39	0.851	1.304	1.685	1.798	1.937	2.023	2.125	2.426	2.708	3.558
40	0.851	1.303	1.684	1.796	1.936	2.021	2.123	2.423	2.704	3.551

CL	α	$\alpha/2$	(n−1) DOF	$t_{n-1,\alpha/2}$
90%	0.1	0.05	10	1.812
90%	0.1	0.05	20	1.725
95%	0.05	0.025	10	2.228
95%	0.05	0.025	20	2.086
99%	0.01	0.005	10	3.169
"	"	"	20	2.845

Calculate confidence interval with unknown population variance

Suppose we want to know what is the average starting salary of a data scientist in India. We collect the salary of 9 early-stage data scientist from Glassdoor.
The salary data looks following (all values are in Lakhs Per Annum, where, lakhs = 100,000)
8.5, 8.0, 7.2, 5.3, 6.7, 9.8, 5.7, 9.3, 6.2.

(1) Calculate the 90%, 95%, and 99% confidence interval around the estimated mean.
(2) If you know that the population standard deviation in this case is 1.6, how the confidence interval looks like?

1) Sample mean, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{9} \times 66.7 = 7.4$

$ME = t_{n-1, \alpha/2} \times \frac{s}{\sqrt{n}}$

2) Sample Std. deviation $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - 7.4)^2} = 1.59$

$\therefore SE = \frac{1.59}{3} = 0.53$

3) DoF = 9 − 1 = 8,

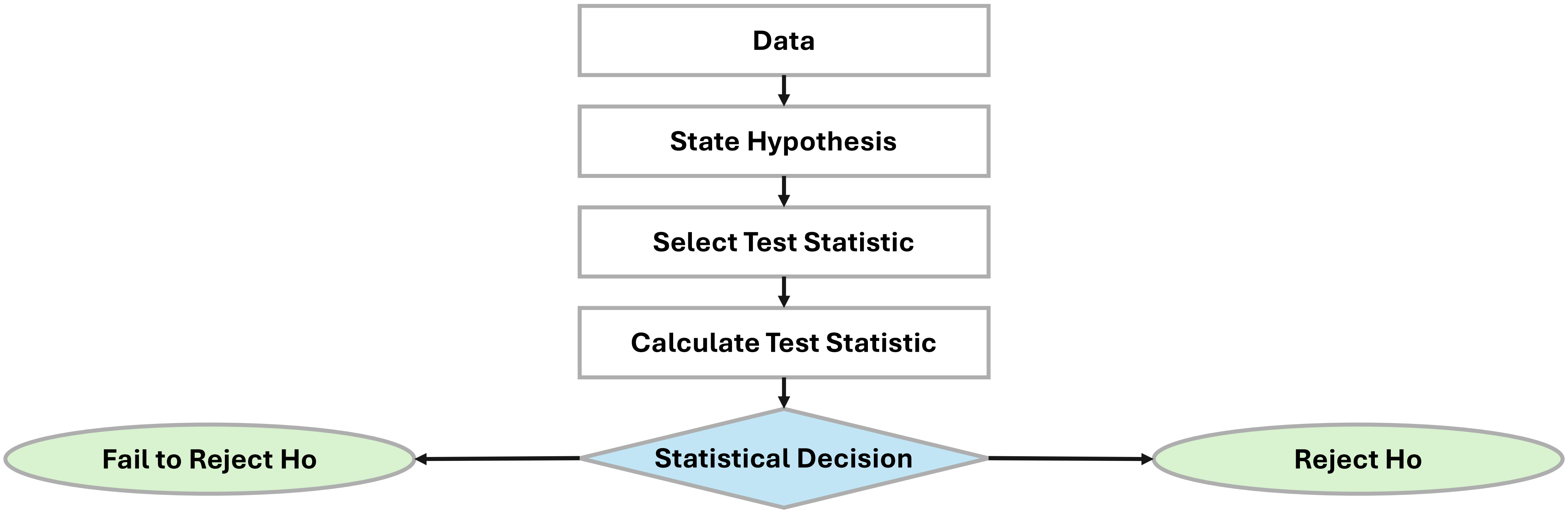
CL	α	$\alpha/2$	$t_{8,\alpha/2}$	ME	LCB	UCB
90%	0.1	0.05	1.86	0.986	6.41	8.39
95%	0.05	0.025	2.306	1.22		
99%	0.01	0.005	3.355	1.77		

Hypothesis testing

Hypothesis testing is a fundamental statistical method used to assess claims or hypotheses about populations based on sample data. It involves two key hypotheses:

1. Null Hypothesis (H_0): This is the default assumption or status quo. It suggests that there is no significant effect, difference, or relationship in the population.

2. Alternative Hypothesis (H_a): This contradicts the null hypothesis and suggests that there is a significant effect, difference, or relationship in the population.



Type-I and Type-II Error

Type-I Error (False Positive): A Type-I error occurs when we reject the null hypothesis (H0) when it is actually true.

Example: In a medical test, a Type I error would be incorrectly diagnosing a healthy person as having a disease.

Probability: The probability of making a Type I error is denoted by alpha (α). The maximum acceptable probability of making a type-I error is called the significance level of the test. A common significance level is 0.05, meaning there's a 5% chance of making a Type I error.

Type-II Error (False Negative): A Type-II error occurs when you fail to reject the null hypothesis when it is actually false.

Example: In a medical test, a Type II error would be failing to diagnose a person who actually has the disease.

Probability: The probability of making a Type-II error is denoted by beta (β). **Power of a test** is (1 - β), which represents the probability of correctly rejecting a false null hypothesis.

	Null Hypothesis is TRUE	Null Hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct Outcome! (True positive)
Fail to reject null hypothesis	Correct Outcome! (True negative)	Type II Error (False negative)

Examples of Null and Alternate Hypothesis

Scenario-1: A company claims that its light bulbs have an average lifespan of 10,000 hours. A consumer group tests a sample of bulbs to see if the claim is accurate.
$\boldsymbol{H_0}$: The average lifespan of the light bulbs is 10,000 hours. (μ=10,000)
$\boldsymbol{H_a}$: The average lifespan of the light bulbs is not 10,000 hours. (μ ≠10,000)

Scenario-2: A political candidate claims that at least 50% of voters support them. A survey is conducted to verify this claim.
$\boldsymbol{H_0}$: The proportion of voters who support the candidate is 50% or more. (p≥0.50)
$\boldsymbol{H_a}$: The proportion of voters who support the candidate is less than 50%. (p<0.50)

Scenario-3: A new teaching method is introduced, and researchers want to see if it improves student test scores compared to the old method.
$\boldsymbol{H_0}$: There is no significant difference in the average test scores between students taught with the new method and students taught with the old method. ($\mu_{new} = \mu_{old}$)
$\boldsymbol{H_a}$: Students taught with the new method have significantly higher average test scores than students taught with the old method. ($\mu_{new} > \mu_{old}$)

One-sided and two-sided hypothesis test

One-Sided Hypothesis Test (One-Tailed Test)
A one-sided test, also known as a one-tailed test, is used when the alternative hypothesis specifies a **direction** for the difference or relationship. This means you are interested in whether a parameter is either *greater than* a certain value or *less than* a certain value, but not both. The rejection region for the null hypothesis is entirely in one tail of the sampling distribution (either the left or the right tail).

Example:

1. A drug is newly developed to treat a certain disease and claims that it does so in less than 15 days. In this case the null hypothesis $H_0 : \mu \geq 15$ and the alternate hypothesis is $H_a : \mu < 15$. This is a left-sided hypothesis test as the rejection region of null hypothesis is entirely on the left tail of the sampling distribution.

2. A fertilizer is developed to increase the yield of a crop. A group of farmers uses that fertilizer. In this case the null hypothesis $H_0 : \mu_{new} = \mu_{old}$ i.e. there is no significant difference in using the fertilizer. The alternate hypothesis is $H_a : \mu_{new} > \mu_{old}$. In this case we are only interested to know whether the yield has increased or not. This is a right-sided hypothesis test as the rejection region of null hypothesis is entirely on the right tail of the sampling distribution.

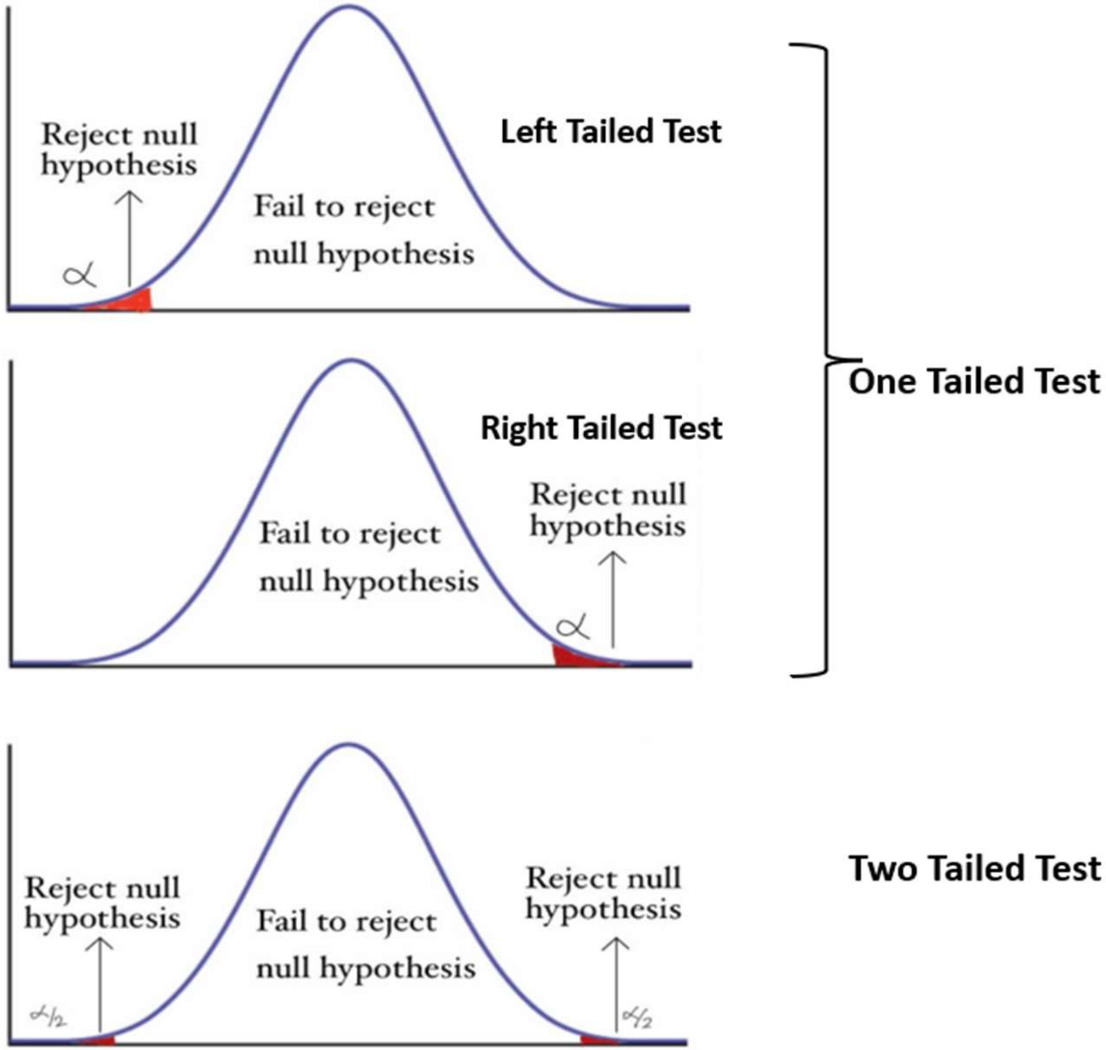
3. An instrument manufacturer claims the average lifespan of his instrument is 1000 hours. In this case the null hypothesis is $H_o : \mu = 1000$ and the alternate hypothesis is $H_a : \mu \neq 1000$. This is an example of two-sided hypothesis testing, where the actual value can be in either direction of claimed value and the rejection region of null hypothesis falls in both the tails of sampling distribution.

One-sided and two-sided hypothesis test

Null Hypothesis(H0) : $\mu \geq value$
Alternative Hypothesis(H_A) : $\mu < value$

Null Hypothesis(H0) : $\mu \leq value$
Alternative Hypothesis(H_A) : $\mu > value$

Null Hypothesis(H0) : $\mu = value$
Alternative Hypothesis(H_A) : $\mu \neq value$



P-value

The p-value is a statistical measure that quantifies the strength of evidence against the null hypothesis.

It represents the probability of observing sample data as extreme as, or more extreme than, the data observed, assuming the null hypothesis is true. A smaller p-value indicates stronger evidence against the null hypothesis.

Definition: A p-value measures the strength of evidence against the null hypothesis in hypothesis testing.

Interpretation: Smaller p-values indicate stronger evidence against the null hypothesis.

Decision: If $p \leq \alpha$ (chosen significance level, e.g., 0.05), you reject the null hypothesis; if $p > \alpha$, you fail to reject it.

Example: In order to test whether a coin is biased or not, I tossed the coin 5 times and obtained "head" all the time. Given significance level of 0.05, can we conclude that the coin is biased?

H_0 : The coin is unbiased $\quad p_H = p_T = 0.5$

H_a : The coin is biased

$$P(H, H, H, H, H) = \frac{1}{2^5} = \frac{1}{32} \quad ; \quad P(T, T, T, T, T) = \frac{1}{32}$$

$$\therefore \text{p-value} = P(H \times 5) + P(T \times 5) = \frac{1}{32} + \frac{1}{32} = \frac{1}{16} = 0.0625 > 0.05$$

We don't have enough statistical evidence to say the coin is unbiased or not given the significance level 0.05.

Ex: 2 :

H_0 : coin is unbiased

H_a : coin is biased

6 coin toss and observed T T H T T T

significance level is 0.05

$P(TTHTTT) = P(5T, 1H) = \frac{1}{64}$

$P(5H, 1T) = \frac{1}{64}$; $P(6T) = \frac{1}{64}$, $P(6H) = \frac{1}{64}$

$p\text{-value} = \frac{4}{64} = \frac{1}{16} = 0.0625 > 0.05$

Hypothesis testing with Z-test

Z-Test: The Z-test is a statistical hypothesis test used to assess whether a sample mean is significantly different from a known population mean when the sample size is sufficiently large ($n \geq 30$). It relies on the standard normal distribution and is suitable when **the population standard deviation is known**.

Steps in z-test:

Step 1: Formulate Hypotheses, i.e. state the null and alternate hypothesis

Step 2: Set Significance Level, usually significance level is 0.05 unless otherwise specified

Step 3: Calculate the Sample Mean

Step 4: Set Up the Z-Test Statistic using the formula: $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

Step 5: Find Critical Value or P-Value

Step 6: Make a Decision

Example-1: Testing the effect of new fertilizer on Crop Yield.

A farmer traditionally gets an average yield of **500 kg per acre** for a certain crop, with a known population standard deviation of **50 kg**. He tries a new fertilizer on a sample of **40 acres**. The average yield from these 40 acres is **520 kg.** The farmer wants to know if the new fertilizer has significantly increased the crop yield at a **5% level of significance**.

Hypothesis testing with Z-test

Assumptions for Z-test:

- The sample is randomly selected.
- The population standard deviation (σ) is known (50 kg).
- The sample size (n=40) is large enough for the Central Limit Theorem to apply, making the sampling distribution of the mean approximately normal.

Step-1: State the null and alternate hypothesis

H_0 : The new fertilizer has no significant effect on the crop yield i.e. $\mu = 500\ kg$
H_a : The new fertilizer has significantly *increased* the crop yield. i.e. $\mu > 500\ kg$

Clearly it is a one-tailed test, because the rejection region falls on the right tail.

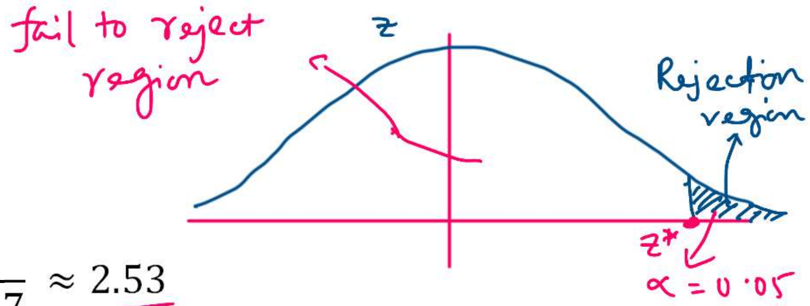
Step-2: State the significance level

The significance level is 5% as stated in the problem. This means we are willing to accept a 5% chance of incorrectly rejecting the null hypothesis (Type I error) if it is actually true.

Step-3: Calculate the sample mean

In this case it is given to be 520 kg. i.e. $\bar{x} = 520\ kg$

Hypothesis testing with Z-test



Step-4: Calculate the test statistic (z – statistic)

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{520 - 500}{\frac{50}{\sqrt{40}}} = \frac{20}{50/6.3245} = \frac{20}{7.9057} \approx 2.53$$

$(1-\alpha) = 0.95$

$z^* = 1.645$ $\quad z > z^*$

Step-5: Calculate the critical value or p-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
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861

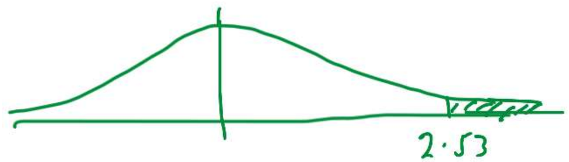
For right sided test. The critical value corresponding to significance level = 5% is 1.645.

Alternatively, the p-value for $z = 2.53$ is obtained as:

$$p = P(Z > 2.53) = 1 - \underbrace{P(Z <= 2.53)}_{z\text{-table}} = 1 - 0.9943 = 0.0057$$

Step-6: make a decision.

As p-value < significance level (0.05), or z-statistic > critical value, we can reject the null hypothesis. i.e. the farmer can be reasonably confident that the new fertilizer is effective.



Hypothesis testing with Z-test

Example-2: Testing the lifespan of a shoe.

A shoe manufacturer claims that the average lifespan of their new line of running shoes is 500 miles. A consumer protection agency wants to test this claim. They randomly select 100 pairs of these running shoes and find that the average lifespan for this sample is 485 miles. The manufacturer states that the population standard deviation for the lifespan of these shoes is 60 miles. The agency wants to conduct the test at a 5% level of significance.

Step-1: State the null and alternate hypothesis

H_0 : The average lifespan of the running shoes is 500 miles i.e. $\mu = 500\ miles$
H_a : The average lifespan of the running shoes is *not* 500 miles. i.e. $\mu \neq 500\ miles$

Clearly it is a two-tailed /two-sided test, because the average lifespan could be greater or less than 500 miles.

Step-2: State the significance level

The significance level is 5% as stated in the problem.

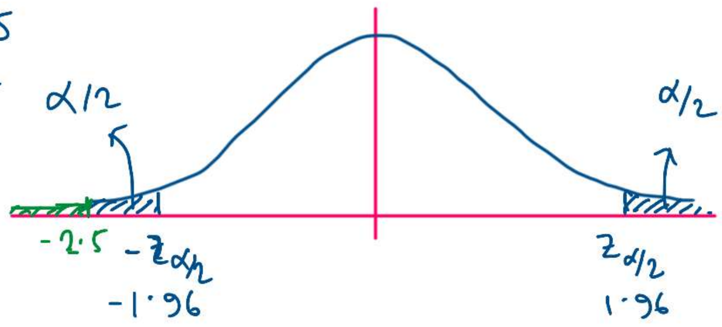
Step-3: Calculate the sample mean

In this case it is given to be 485 miles. i.e. $\bar{x} = 485\ miles$

Hypothesis testing with Z-test

$\alpha/2 = 0.025$

$1 - \alpha/2 = 0.975$



$z < -2.5$ $\quad$ $z > 2.5$

Step-4: Calculate the test statistic (z – statistic)

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{485 - 500}{\frac{60}{\sqrt{100}}} = \frac{-15}{60/10} = \frac{-15}{6} = -2.5$$

Step-5: Calculate the critical value or p-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
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861

For two-sided test, the critical value corresponding to significance level = 5% is ± 1.96.
Since, -2.5 < -1.96 our calculated z-statistic falls within the rejection region.

Alternatively, the p-value for $z = 2.5$ for two-sided test is obtained as:

$$\begin{aligned} p &= P(Z < -2.5) + P(Z > 2.5) \;\; [as\ two\ sided] \\ &= 2 \times P(Z > 2.5) \;\; [because\ of\ symmetry\ of\ distribution] \\ &= 2 \times \big(1 - P(Z \le 2.5)\big) \; [to\ find\ the\ value\ from\ the\ table] \\ &= 2 \times (1 - 0.9938) = 2 \times 0.0062 = 0.0124 \end{aligned}$$

Step-6: make a decision.

As p-value < 0.05, we can reject the null hypothesis. i.e. the average lifespan of shoe is not 500 miles.

T-test

Aspect	T-Test	Z-Test
Purpose	Compare means of two groups (small samples) or Compare a single sample mean with known / hypothesized population mean.	Compare a sample mean to a known/ hypothesized population mean (large samples)
Sample Size	Small sample sizes (typically < 30)	Large sample sizes (typically ≥ 30)
Population SD	Unknown and estimated from the sample	Known or estimated from a large sample
Distribution	Follows the t-distribution	Follows the standard normal (z) distribution
Formula	$t = (\bar{x} - \mu) / (s/\sqrt{n})$	$z = (\bar{x} - \mu) / (\sigma/\sqrt{n})$
Critical Values	Use t-distribution tables	Use standard normal (z) distribution tables
Common Use Case	Comparing means when SD is unknown	Comparing a sample mean to a known population mean
Example	Compare test scores of two groups	Compare the average height of a sample to a known population mean

Hypothesis testing with t-test

Example-1:
A coffee shop owner claims that their new automatic coffee machine dispenses, on average, 250 ml of coffee per cup. A customer suspects that the machine is not dispensing exactly 250 ml. To investigate, the customer randomly collects 20 cups of coffee dispensed by the machine and measures the volume in each cup.
The sample data (volumes in ml) is as follows: 248, 252, 249, 251, 247, 253, 250, 249, 252, 246, 254, 250, 248, 251, 249, 253, 250, 247, 252, 251. The customer wants to test this at a 5% level of significance.

Step-1: State the null and alternate hypothesis

H_0 : The average volume of coffee dispensed by the machine is 250 ml i.e. $\mu = 250\ ml$
H_a : The average volume of coffee dispensed by the machine is *not* 250 ml. i.e. $\mu \neq 250\ miles$

Clearly it is a two-tailed /two-sided test.

Step-2: State the significance level

The significance level is 5% as stated in the problem.

Step-3: Calculate the sample mean and the standard deviation

In this case the calculated sample mean. i.e. $\bar{x} = 250.1\ ml$.
The calculated sample standard deviation. i.e. $s = 2.458\ ml$.

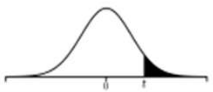
Hypothesis testing with t-test

Step-4: Calculate the test statistic (t – statistic)

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{250.1 - 250}{\frac{2.458}{\sqrt{20}}} = \frac{0.1}{2.458/4.472} = \frac{0.1}{0.5496} = 0.182$$

Step-5: Calculate the critical value or p-value. The degrees of freedom = 20 – 1 = 19

Critical Values for Student's *t*-Distribution.



df	0.2	0.1	0.05	0.04	0.03	0.025	0.02	0.01	0.005	0.0005
	Upper Tail Probability: $\Pr(T > t)$									
1	1.376	3.078	6.314	7.916	10.579	12.706	15.895	31.821	63.657	636.619
2	1.061	1.886	2.920	3.320	3.896	4.303	4.849	6.965	9.925	31.599
3	0.978	1.638	2.353	2.605	2.951	3.182	3.482	4.541	5.841	12.924
4	0.941	1.533	2.132	2.333	2.601	2.776	2.999	3.747	4.604	8.610
5	0.920	1.476	2.015	2.191	2.422	2.571	2.757	3.365	4.032	6.869
6	0.906	1.440	1.943	2.104	2.313	2.447	2.612	3.143	3.707	5.959
7	0.896	1.415	1.895	2.046	2.241	2.365	2.517	2.998	3.499	5.408
8	0.889	1.397	1.860	2.004	2.189	2.306	2.449	2.896	3.355	5.041
9	0.883	1.383	1.833	1.973	2.150	2.262	2.398	2.821	3.250	4.781
10	0.879	1.372	1.812	1.948	2.120	2.228	2.359	2.764	3.169	4.587
11	0.876	1.363	1.796	1.928	2.096	2.201	2.328	2.718	3.106	4.437
12	0.873	1.356	1.782	1.912	2.076	2.179	2.303	2.681	3.055	4.318
13	0.870	1.350	1.771	1.899	2.060	2.160	2.282	2.650	3.012	4.221
14	0.868	1.345	1.761	1.887	2.046	2.145	2.264	2.624	2.977	4.140
15	0.866	1.341	1.753	1.878	2.034	2.131	2.249	2.602	2.947	4.073
16	0.865	1.337	1.746	1.869	2.024	2.120	2.235	2.583	2.921	4.015
17	0.863	1.333	1.740	1.862	2.015	2.110	2.224	2.567	2.898	3.965
18	0.862	1.330	1.734	1.855	2.007	2.101	2.214	2.552	2.878	3.922
19	0.861	1.328	1.729	1.850	2.000	2.093	2.205	2.539	2.861	3.883
20	0.860	1.325	1.725	1.844	1.994	2.086	2.197	2.528	2.845	3.850

For two-sided test, the critical value corresponding to significance level = 5% and dof = 19 is ± 2.093.
Since, -2.093 < 0.182 < 2.093 our calculated z-statistic falls within the fail to reject region.

Alternatively, the p-value calculated (using the software) = 0.8575

Step-6: make a decision.

As p-value > 0.05, we **fail to reject** the null hypothesis. At the 5% level of significance, there is not sufficient evidence to conclude that the average volume of coffee dispensed by the machine is significantly different from 250 ml. The observed difference of 0.1 ml is likely due to random sampling variability. The customer should conclude that the machine *is* dispensing, on average, 250 ml of coffee per cup.

T-test for two dependent means (paired samples)

This test is used when you have two sets of observations from the same subjects or matched pairs. This often occurs in "before and after" studies, or when comparing two different treatments applied to the same individuals. The key idea is that the observations in one group are directly related to the observations in the other group.

Example: Evaluating the effectiveness of a new meditation program on stress levels.
A researcher wants to know if a 4-week meditation program significantly reduces stress levels in participants. They measure the stress levels of 10 individuals before starting the program and again after completing the program.

Participants	Stress Level (before)	Stress Level (after)
1	75	68
2	82	75
3	60	58
4	90	80
5	70	72
6	65	60
7	88	85
8	78	70
9	72	65
10	85	78

T-test for two dependent means (paired samples)

Step-1: State the null and alternate hypothesis

H_0 : There is no significant difference in stress levels before and after the meditation program. In other words, the mean difference (μ_D) between the paired observations is zero. $\mu_D = 0$

H_a : There is a significant difference in stress levels before and after the meditation program. Specifically, we hypothesize that the program reduces stress, so the mean difference is greater than zero (Before - After > 0). This is a one-tailed test. $\mu_D > 0$

Participants	Stress Level (before)	Stress Level (after)	Difference (D)
1	75	68	7
2	82	75	7
3	60	58	2
4	90	80	10
5	70	72	-2
6	65	60	5
7	88	85	3
8	78	70	8
9	72	65	7
10	85	78	7

T-test for two dependent means (paired samples)

Step-2: State the significance level The significance level is 5%.

Step-3: Calculate the mean of the observed differences: $\overline{D} = 5.4$

Calculate the standard deviation of the observed differences: $S_D = 3.502$

Step-4: Calculate the test statistic (t – statistic)

$$t = \frac{\overline{D} - \mu_D}{\frac{S_D}{\sqrt{n}}} = \frac{5.4 - 0}{\frac{3.502}{\sqrt{10}}} = \frac{5.4}{3.502/3.162} = \frac{5.4}{1.107} = 4.878$$

Step-5: Calculate the critical value. The degrees of freedom = 10 – 1 = 9

For one-sided test, the critical value corresponding to significance level = 95% and dof = 9 is 1.833

Step-6: make a decision.

As the t-statistic > 1.833 (critical value), hence we reject the null hypothesis. There is statistically significant evidence at the 0.05 level to conclude that the meditation program significantly reduces stress levels.

T-test for two independent means

This test is used when you have two separate, unrelated groups of subjects, and you want to compare their means on a particular variable. This is widely used in A/B testing with small samples.

Example: Comparing the average exam scores of two different teaching methods.
A teacher wants to compare the effectiveness of two different teaching methods (Method A vs. Method B) on students' exam scores. They randomly assign 20 students to Method A and 22 students to Method B. After the course, all students take the same exam.

Data:

Method – A:

$$Sample\ Size\ (n_1) = 20$$
$$Mean\ Score\ (\overline{\boldsymbol{x_1}}) = 78$$
$$Standard\ Deviation\ (s_1) = 7$$

Method – B:

$$Sample\ Size\ (n_2) = 22$$
$$Mean\ Score\ (\overline{\boldsymbol{x_2}}) = 72$$
$$Standard\ Deviation\ (s_2) = 8$$

Step-1: State the null and alternate hypothesis

H_0 : There is no significant difference in the average exam scores between Method A and Method B. i.e. $\mu_1 = \mu_2$
H_a : There is a significant difference in the average exam scores between Method A and Method B.. i.e. $\mu_1 \neq \mu_2$

Clearly it is a two-tailed /two-sided test.

T-test for two independent means

Step-2: State the significance level   The significance level is 5%.

Step-3: Calculate the sample means and sample standard deviations. In this case both are given.

Step-4: Calculate the pooled standard deviation using the following formula:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(20-1)7^2 + (22-1)8^2}{(20+22-2)}} = 7.541$$

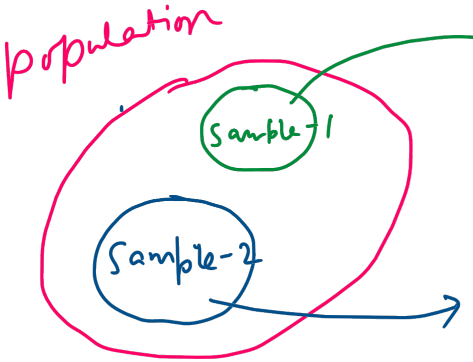
Pooled sample variance is a **weighted average of the individual sample variances** used in an independent t-test when we assume the two populations have the **same underlying variance**. It provides a more precise estimate of this shared population variance by combining data from both groups.

Step-5: Calculate the test statistic (t – statistic) using the following formula:

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(\overline{x_1} - \overline{x_2})}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

(as under the null hypothesis $\mu_1 = \mu_2$)

$$t = \frac{(78 - 72)}{7.541 \times \sqrt{\frac{1}{20} + \frac{1}{22}}} = \frac{6}{2.329} = 2.676$$



sample size: n_1 — DOF : $n_1 - 1$
sample std. deviation : s_1 — s_1^2 — $\frac{s}{\sqrt{n}}$

sample size : n_2 — DOF: $n_2 - 1$
sample std. deviation : s_2 — s_2^2

weights here are DOF

weighted sum of the variances : $(n_1 - 1) \cdot s_1^2 + (n_2 - 1) s_2^2$

weighted average of the " : $\frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{(n_1 - 1) + (n_2 - 1)}$

$$= \frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2}$$

pooled sample std. deviation $= \sqrt{\frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2}}$ $\qquad s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

T-test for two independent means

Step-6: Calculate the critical value. The degrees of freedom = $n_1 + n_2 - 2 = 20 + 22 - 2 = 40$

For two-sided test, The critical value corresponding to significance level = 5% and dof = 40 is ± 2.021

Step-7: make a decision.

As the t-statistic > 2.021 (critical value), hence we reject the null hypothesis. There is statistically significant evidence at the 0.05 significance level to conclude that there is a difference in average exam scores between students taught by Method A and Method B. Method A appears to lead to higher scores.

Chi-square test for categorical data

The Chi-Square Test of Independence is used to determine if there is a statistically significant association between two categorical variables. In simpler terms, it helps us figure out if two characteristics (like gender and political preference, or smoking status and lung disease) are related or independent of each other.

Example: Is there a relationship between gender and preferred type of social media?

A social media company wants to know if there's a relationship between a person's gender and their preferred social media platform (e.g., Facebook, Instagram, TikTok). They survey a random sample of 200 people and record their gender and their most frequently used social media platform.

Observed Data (Contingency Table):

Preferred Platform	Male	Female	Total (Row)
Facebook	45	55	100
Instagram	30	20	50
TikTok	25	25	50
Total (Column)	100	100	**200**

Chi-square test for categorical data

Step-1: State the null and alternate hypothesis

Null Hypothesis (H_0): Gender and preferred social media platform are independent. (There is no relationship between them.)
Alternative Hypothesis (H_1): Gender and preferred social media platform are not independent. (There is a relationship between them.)

Step-2: State the significance level Let, the significance level is 5%.

Step-3: Calculate the expected frequencies (E).

Under the assumption of independence (the null hypothesis), we need to calculate the expected frequency for each cell in the table. The formula for the expected frequency of a cell is:

$$E_{row,col} = \frac{Row\ total \times Column\ Total}{Grand\ Total}$$

Expected frequency table:

Preferred Platform	Male	Female	Total (Row)
Facebook	50	50	100
Instagram	25	25	50
TikTok	25	25	50
Total (Column)	100	100	**200**

Chi-square test for categorical data

Step-4: Calculate the Chi-square test statistic (χ^2) using the following formula:

$$\chi^2 = \sum \frac{(O\ - E)^2}{E}$$

Where O means observed frequency,
E means expected frequency
$\sum$ means sum across all the cells.

In our example: Following table shows the computed values of $\frac{(O - E)^2}{E}$ for all the cells.

Preferred Platform	Male	Female
Facebook	0.5	0.5
Instagram	1.0	1.0
TikTok	0.0	0.0

Hence, $\chi^2 = 0.5 + 0.5 + 1.0 + 1.0 + 0.0 + 0.0 = 3.0$

Chi-square test for categorical data

Step-5: Determine the degrees of freedom

The degrees of freedom for a chi-square test of independence are calculated as:

$$dof = (Number\ of\ Rows - 1) \times (Number\ of\ Columns - 1)$$

In our example: $dof = (3-1) \times (2-1) = 2 \times 1 = 2$

Step-6: Calculate the critical Chi-square value

Using χ^2 distribution table with significance level = 0.05 and dof = 2, the critical value = 5.991

Step-7: make a decision.

Since the calculated χ2 (3.0) is **less than** the critical χ2 (5.991), we **fail to reject the null hypothesis**.

there is **no statistically significant evidence** to suggest a relationship between gender and preferred social media platform. We conclude that gender and preferred social media platform are independent.

Thank You