



مجموعة جامعة
الجميع الذكية
للتعليم القابضة

MIDOCEAN
UNIVERSITY
BY EVERYONE'S SMART UNIVERSITY

Mathematical Foundations For Data Science and AI

Lecture 10

AI102

Statistics; Hypothesis Testing

By
Dr. Khaled ELMenshawy

www.midocean.ae

Lecture

10

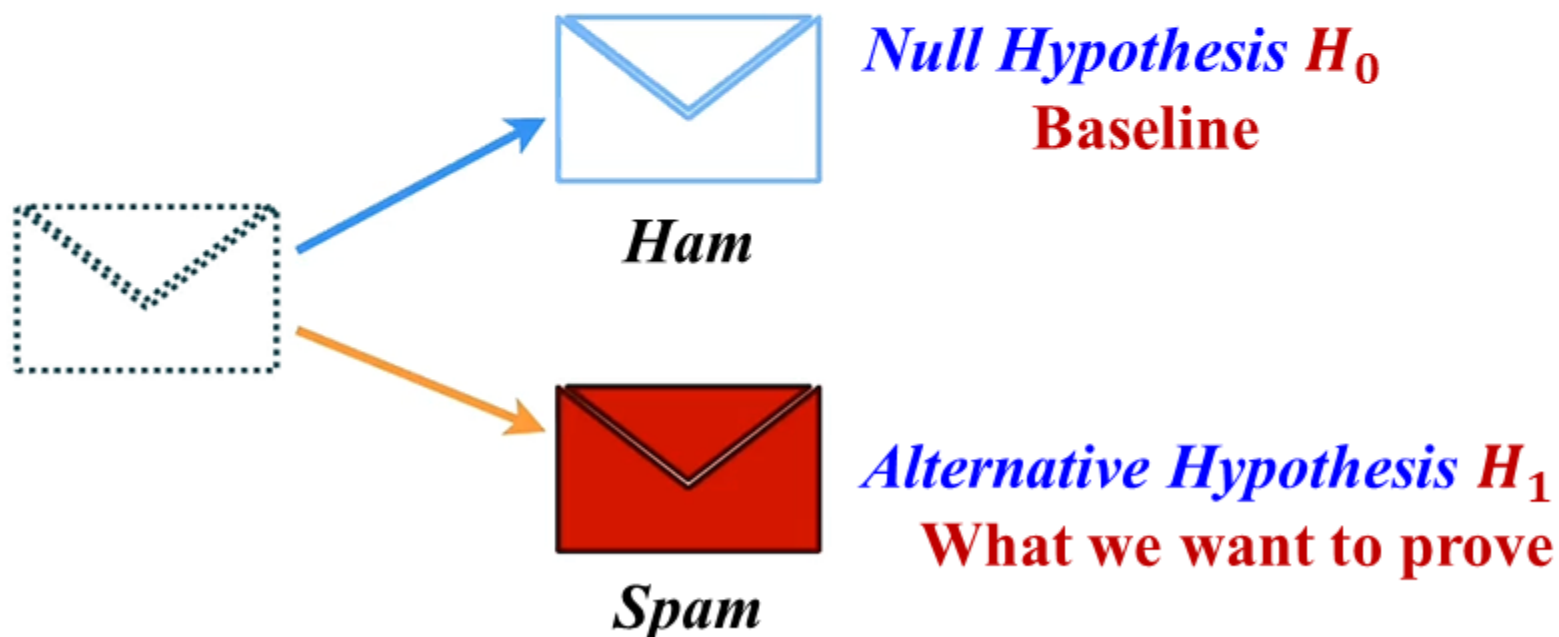
LEARNING OUTCOMES

- Understanding Hypothesis Test.
- Understanding Data Quality.
- Understanding Tailed Test.
- Understanding P-Value.
- Understanding t-Distribution.
- Understanding t-Test.
- Understanding Chi-Square Test

Hypothesis test is a procedure for deciding whether the results of a research study support a particular theory which applies to a population. **A population** is the collection of all items of interest to our study and is usually denoted with an uppercase **N**.



When rejecting that the email is not spam, accepting that the email is Spam. Otherwise, when accepting that the email is Spam, not accepting that the email is Ham, simply there are not enough evidence prove that this email is Spam.



Plenty of evidence against $H_0 \rightarrow$ Reject H_0 and accept H_1

Earn extra cash

Apply Now

Risk Free

Win Lottery

Very unlikely under $H_0 \rightarrow$ Reject H_0 and accept H_1

Ideally, we would like to always make the perfect decisions, Unfortunately, this is impossible to guarantee due to the randomness of the world and the fact that we only have partial information from the population.



*Sending the Ham email
to Spam folder*

Type I Error
False Positive



*Sending the Spam
email to Ham folder*

Type II Error
False Negative

Decision	Reality	
	H_0 True (Not Spam)	H_0 False (Spam)
Reject H_0 (Decide spam)	Type I Error <i>False Positive</i>	<i>Correct</i>
Do not Reject H_0 (Decide not spam)	<i>Correct</i>	Type II Error <i>False Negative</i>

Sending a ham email to the spam box is worse than sending a spam email to the inbox

*Type I Errors are **more severe** than Type II Errors*

How many times are you willing to incorrectly send a ham email to the spam box in order to have a good spam?

Type I Error
False Positive

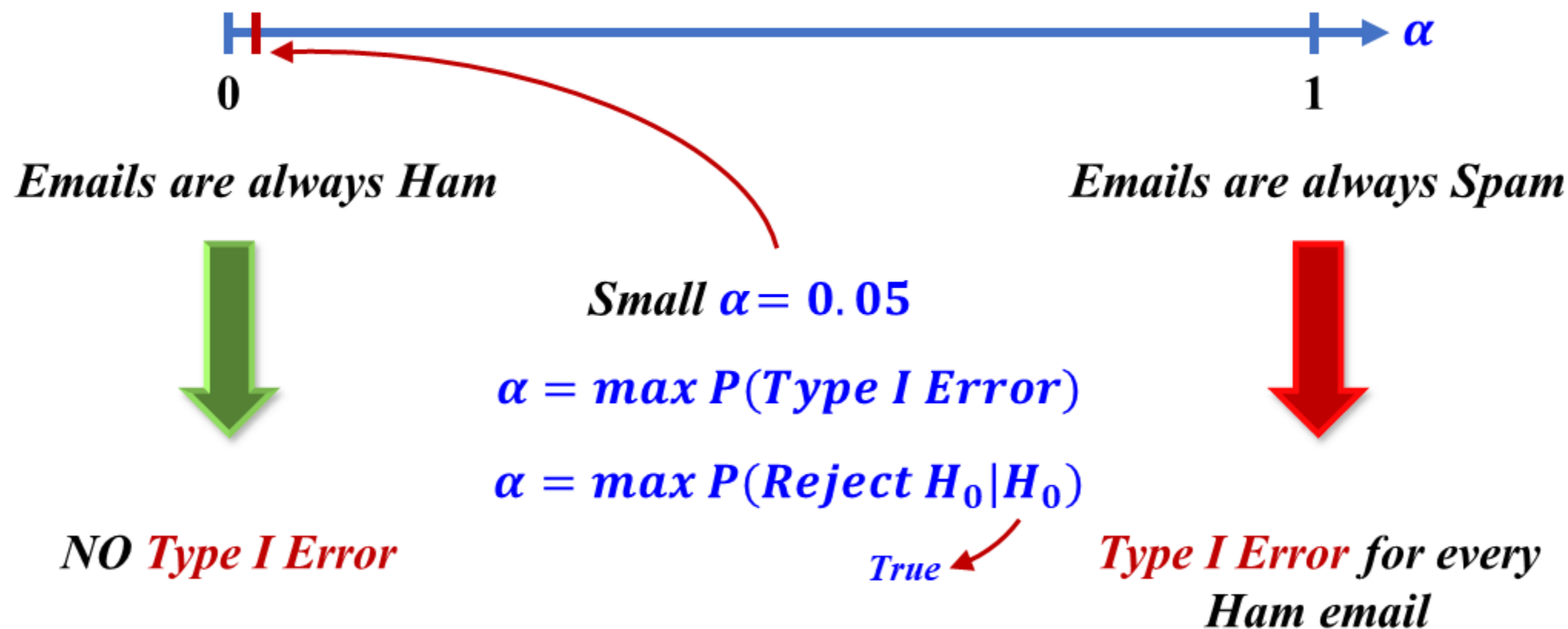
*The maximum probability of **Type I Error** → **significance level (α)***

***The significance level (α)** is the probability of rejecting the null hypothesis (H_0) when it the null hypothesis is true.*



*Ham email classified
to be Spam emails*





Reliable Data***Representative***

Represent the population of the experiment.

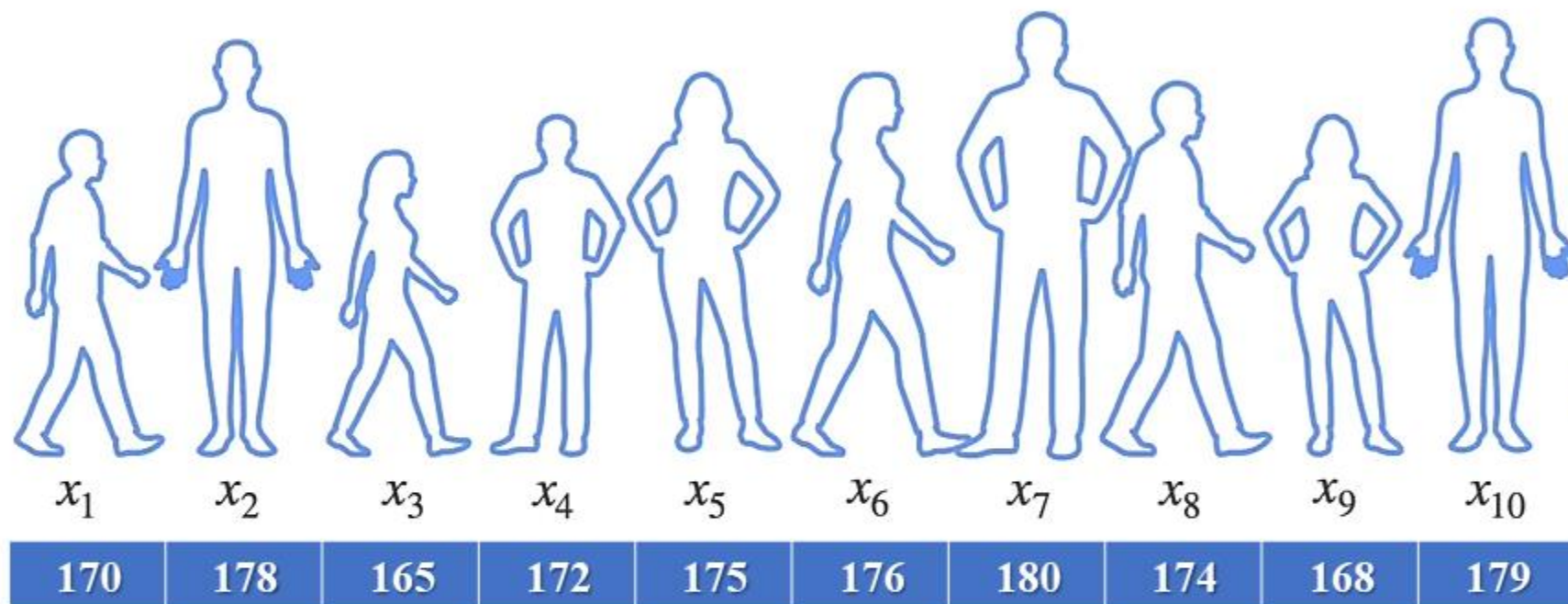
Randomized

Choose in random order from the population of the experiment.

Sample Size

Sample size is enough to obtain an accurate results.

Tailed Test

*observed statistics*

$$\bar{X} = 173.7$$

The mean heights of the 18-year-old of the Egyptian in 90's was 170.5 cm

$$H_0: \mu = 170.5 \text{ vs } H_1: \mu > 170.5$$

$$\bar{X} = \frac{1}{10} \sum_{i=1}^{10} X_i$$

Test statistics

The mean heights of the 18-year-old of the Egyptian in 90's was 170.5 cm

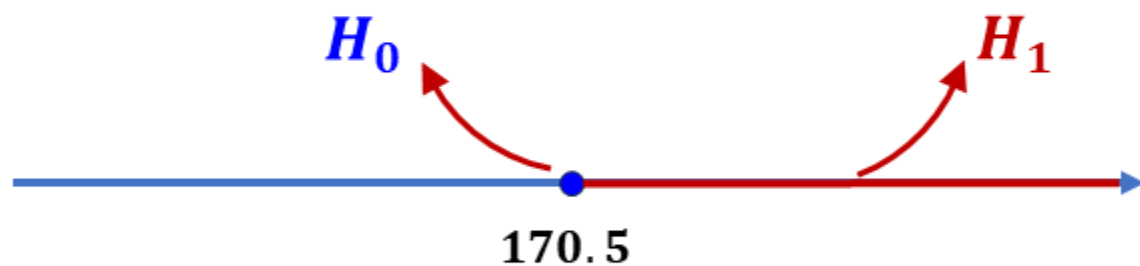
Right-Tailed Test $\longrightarrow H_0: \mu = 170.5$ vs $H_1: \mu > 170.5$

Left -Tailed Test $\longrightarrow H_0: \mu = 170.5$ vs $H_1: \mu < 170.5$

Two -Tailed Test $\longrightarrow H_0: \mu = 170.5$ vs $H_1: \mu \neq 170.5$

The mean heights of the 18-year-old of the Egyptian in 90's was 170.5 cm

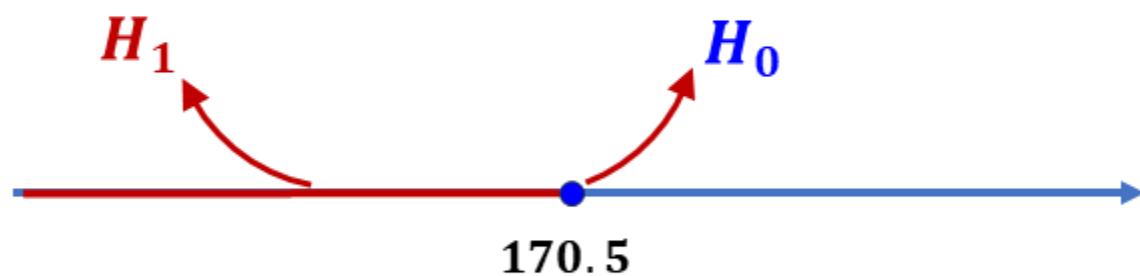
Right-Tailed Test $\longrightarrow H_0: \mu = 170.5$ vs $H_1: \mu > 170.5$



$$\bar{X} \gg 170.5 \xRightarrow{\text{reject}} H_0$$

The mean heights of the 18-year-old of the Egyptian in 90's was 170.5 cm

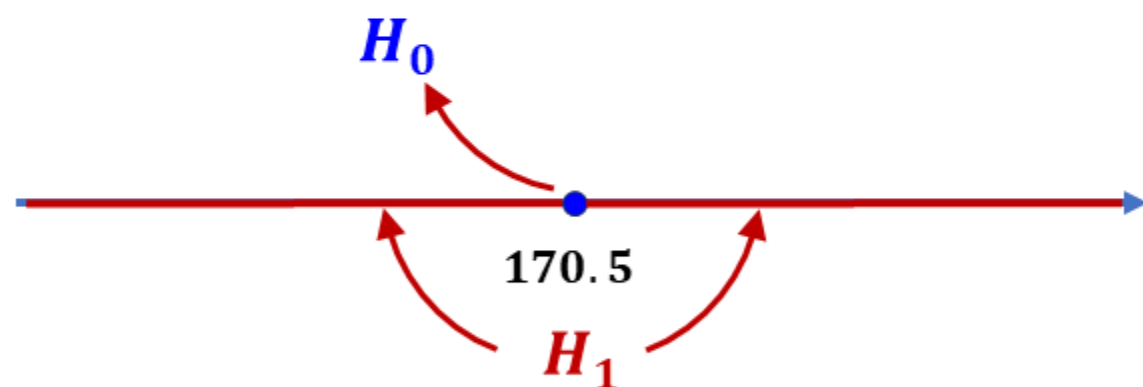
Left-Tailed Test $\longrightarrow H_0: \mu = 170.5$ vs $H_1: \mu < 170.5$



$$\bar{X} \ll 170.5 \xRightarrow{\text{reject}} H_0$$

The mean heights of the 18-year-old of the Egyptian in 90's was 170.5 cm

Two-Tailed Test $\longrightarrow H_0: \mu = 170.5$ vs $H_1: \mu \neq 170.5$



$$\bar{X} \ll 170.5 \text{ or } \bar{X} \gg 170.5 \xRightarrow{\text{reject}} H_0$$

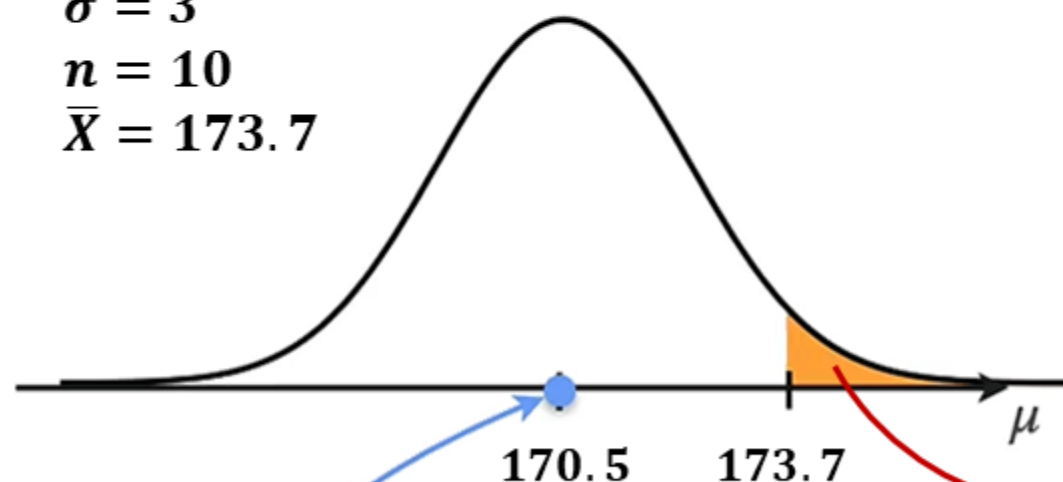
P-Value is the probability under the assumption of no effect or no difference (null hypothesis), of obtaining a result equal to or more extreme than what was actually observed.

$$H_0: \mu = 170.5 \text{ vs } H_1: \mu > 170.5$$

$$\sigma = 3$$

$$n = 10$$

$$\bar{X} = 173.7$$



Null Hypothesis

GOAL: Type I Error probability $< \alpha = 0.05$

Type I Error determine $\mu > 170.5$, when population mean did not change.

$$P(\bar{X} > 173.7 \mid \mu = 170.5)$$

$$= 0.00038 < \alpha$$

Conclusion: Reject H_0

$$Z = 3.37$$

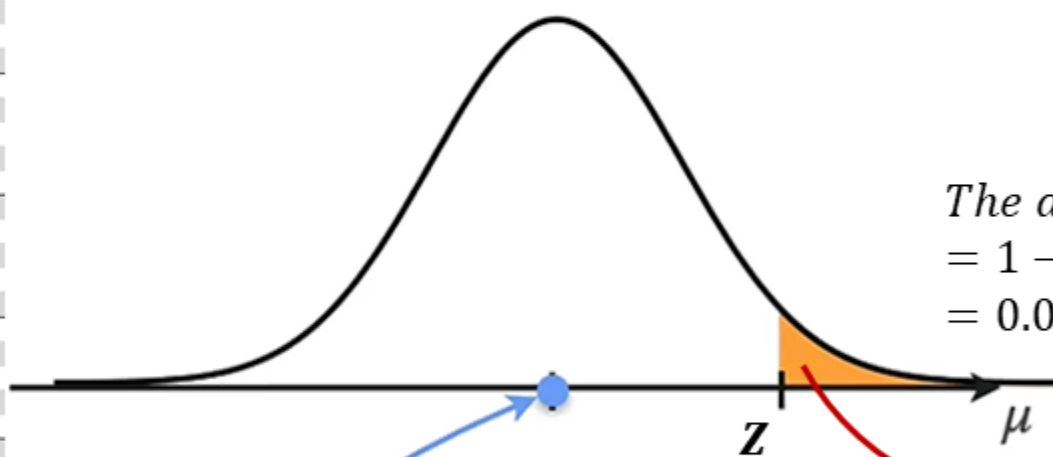
$$= 1 - 0.99962 = 0.00038$$

Z statistics

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99987	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.99990	0.99990	0.99991	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997	0.99997

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} = \frac{173.7 - 170.5}{3 / \sqrt{10}} = 3.37$$

$$\begin{aligned}\sigma &= 3 \\ n &= 10 \\ \bar{X} &= 173.7\end{aligned}$$



The area under the curve
 $= 1 - 0.99962$
 $= 0.00038 = P - \text{Value}$

Null Hypothesis

P-Value < α reject H_0 P-Value > α accept H_0

Steps of Performing Hypothesis Test

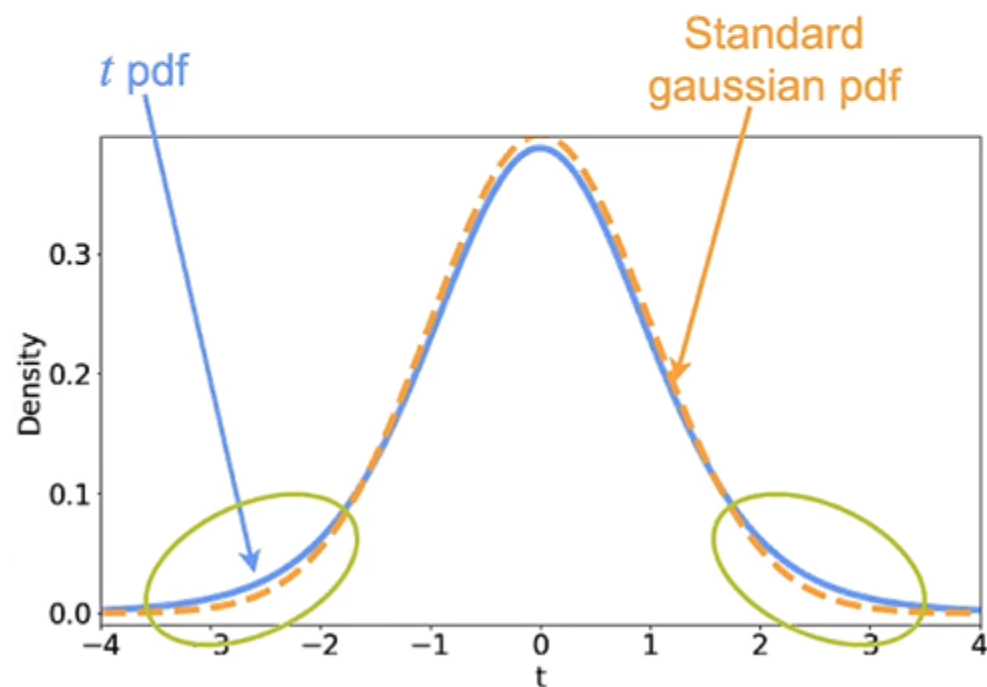
1. *State the hypothesis:*
 - *Null Hypothesis: the baseline $\rightarrow H_0$*
 - *Alternative Hypothesis: the statement we want to prove $\rightarrow H_1$*
2. *Design the Test:*
 - *Decide the test statistic to work with $\rightarrow \bar{X}$*
 - *Set significance level usually $\rightarrow \alpha = 0.05$*
3. *Compute the observed statistic based on samples $\rightarrow \bar{x}$*
4. *Reach a conclusion:*
 - *$P\text{-Value} < \alpha$ reject H_0*
 - *$P\text{-Value} > \alpha$ accept H_0*

***The T-distribution**, also known as the Student's T-distribution, is a type of probability distribution that is similar to the normal distribution with its bell shape but has heavier tails.*

Used when σ is Unknown $T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$ replace σ with its estimate S

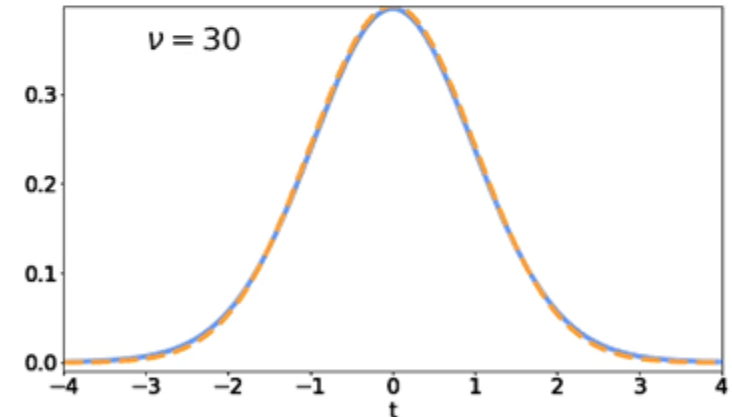
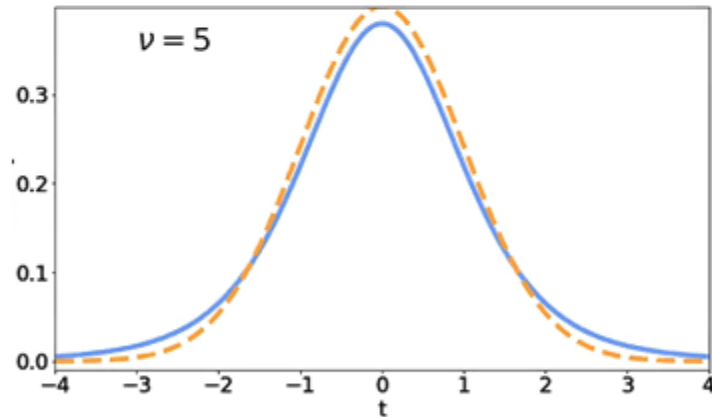
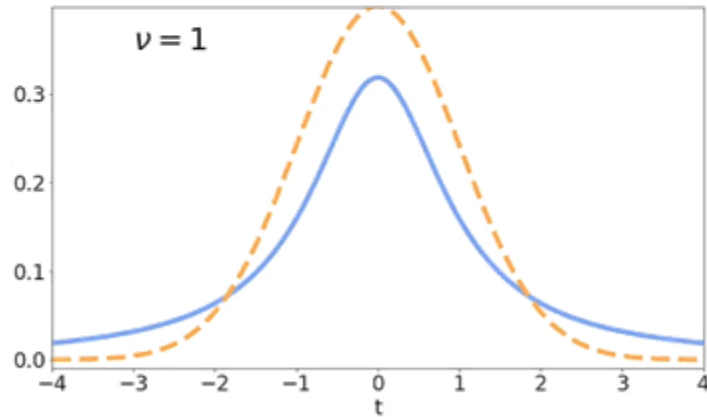
$\swarrow$
T statistics

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$



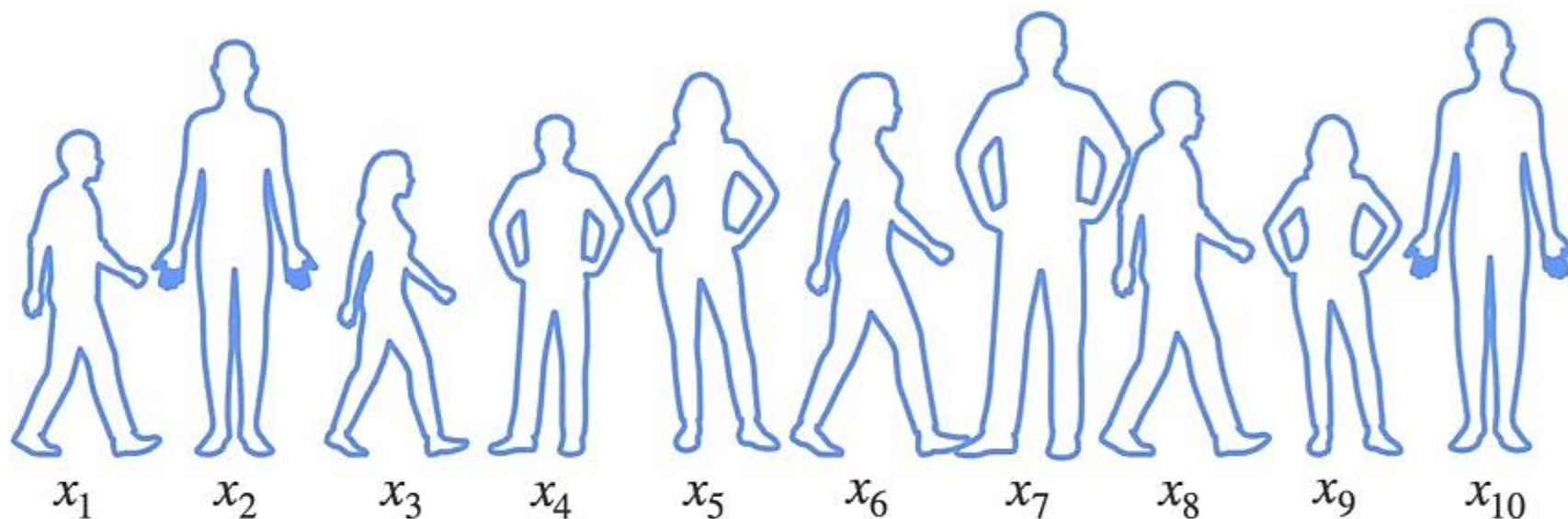
The degree of freedom (ν) illustrate the number of values involved in a calculation that has the freedom to vary.

(ν) is a parameter to control how heavy the tails are.



Increasing (ν)



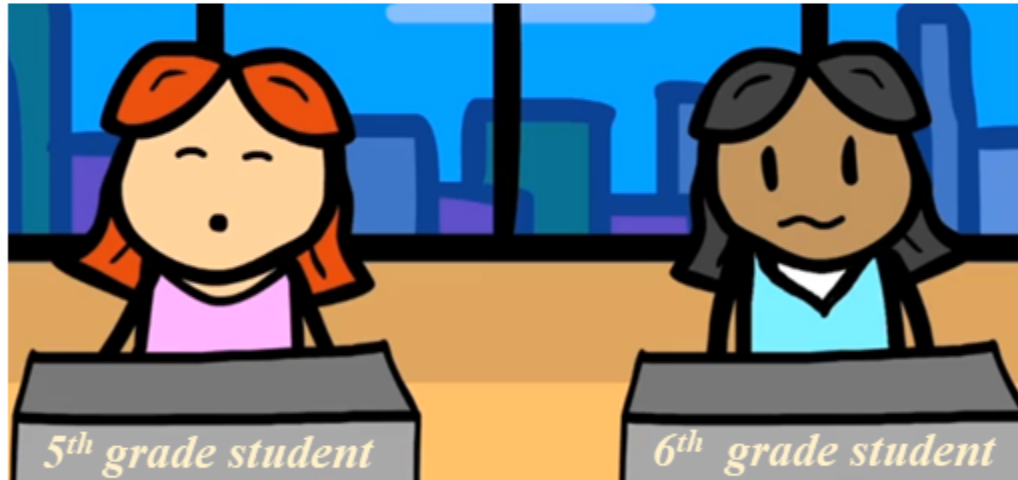


$$n = 10 \rightarrow v = 10 - 1$$

The degree of freedom (v) = sample size - 1 = $n - 1$

*T-Statistics is used when the population has a normal distribution with **unknown** (σ)*

*The **t-Test** is a statistical test that is used to compare the means of two groups.*



H_0 : There is no difference between two groups

H_1 : There is a difference between two groups

Group 1	Group 2
5	2
6	3
4	4
5	3
2	7
$\mu_1 = 4.4$	$\mu_2 = 3.8$
$\sigma = 1.52$	$\sigma = 1.92$
$n_1 = 5$	$n_2 = 5$

How much variation in the sample (σ^2):

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$\frac{1.52^2}{5} + \frac{1.92^2}{5} = 0.462 + 0.737 = 1.199$$

Standard deviation (σ) = $\sqrt{\sigma^2} = \sqrt{1.199} = 1.095$

$$t - \text{Value} = \frac{\mu_1 - \mu_2}{\sigma^2} = \frac{4.4 - 3.8}{1.199} = 0.5$$

Degrees of freedom (<i>df</i>)	0.2	0.15	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	4.165	6.314	12.706	25.452	63.657	127.321	636.619	1273.239
2	1.886	2.282	2.920	4.303	6.205	9.925	14.089	31.599	44.705
3	1.638	1.924	2.353	3.182	4.177	5.841	7.453	12.924	16.326
4	1.533	1.778	2.132	2.776	3.495	4.604	5.598	8.610	10.306
5	1.476	1.699	2.015	2.571	3.163	4.032	4.773	6.869	7.976
6	1.440	1.650	1.943	2.447	2.969	3.707	4.317	5.959	6.788
7	1.415	1.617	1.895	2.365	2.841	3.499	4.029	5.408	6.082
8	1.397	1.592	1.860	2.306	2.752	3.355	3.833	5.041	5.617
9	1.383	1.574	1.833	2.262	2.685	3.250	3.690	4.781	5.291
10	1.372	1.559	1.812	2.228	2.634	3.169	3.581	4.587	5.049

Degree of freedom = $10 - 2 = 8$

From the t-table, t-value must be over 2.306

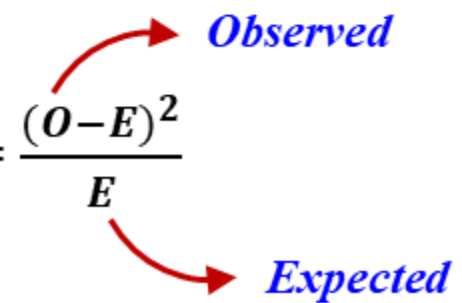
t-value = $0.5 < 2.306$

Then we can not reject null hypothesis

Chi-Square Test is a statistical method of calculating whether variations in data are due to one of the tested variables, or due to chance.

The purpose of chi-square testing is to compare categorical data and determine if the observed experimental variables are independent of each other by calculating their statistical significance

$$\text{Chi-Square} = \frac{(O-E)^2}{E}$$



df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Subject	Description
<i>The complement rule</i>	$P(A') = 1 - P(A)$
<i>The Sum of probability (Disjoint-Events)</i>	<i>The sum of all probabilities</i>
<i>Sum of Probabilities (Joint-Events)</i>	<i>The sum of probabilities subtracted from Joint events</i>
<i>Conditional Probability</i>	$P(A \cap B) = P(A) \times P(B A)$
<i>Bayes Theorem</i>	$P(A B) = \frac{P(A) \times P(B A)}{P(A) \times P(B A) + P(A') \times P(B A')}$
<i>Bernoulli Distribution</i>	$X \sim \text{Bernoulli "Bern"} (p)$
<i>Binomial Distribution</i>	$X \sim \text{Binomial "B"} (n, p)$
<i>Poisson Distribution</i>	$P(y) = \frac{\lambda^y e^{-\lambda}}{y!}$

Subject	Description
<i>The Mean</i>	$\frac{1}{n} \sum_{i=1}^n X_i$
<i>Variance σ^2</i>	<i>The average of the squared differences from the Mean</i>
<i>The Standard Deviation σ</i>	$\sqrt{\sigma^2}$
<i>Normal distribution</i>	$X \sim \text{Normal "N"} (\mu, \sigma^2)$
<i>Standardization</i>	$\frac{X - \mu}{\sigma}$
<i>Z-Score</i>	$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$

Subject	Description
<i>T-distribution</i>	$T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$
<i>Estimate of SD</i>	$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$

