Introduction to STATA
Course Code: STA L 4108

Prof. Dr. Md. Atiqul Islam
M.Sc. (SUST, BD), M.Sc. (UHasselt, BE), PhD (RuG, NL)

Professor
Department of Statistics
Jagannath University, Dhaka-1100
E-mail: atique@stat.jnu.ac.bd

Lecture: 2

- **Hypothesis Testing and Regression Analysis**

Hypothesis Test for a Population Mean (μ)

- Let $x_1, x_2, \dots, x_n$ be a random sample of size n from a random variable X with mean μ and variance σ^2. We would like to test the following hypotheses at $\alpha\%$ level of significance:

$$H_0: \mu = \mu_0 \ \ vs. \ \ H_1: \mu \neq \mu_0$$

$$or, \ \ H_0: \mu \leq \mu_0 \ \ vs. \ \ H_1: \mu > \mu_o$$

$$or, \ \ H_0: \mu \geq \mu_0 \ \ vs. \ \ H_1: \mu < \mu_o$$

- **Scenario-1:** If X has a normal distribution with ***unknown population variance*** (σ^2).

- **Scenario-2:** If X has a normal distribution with ***known population variance*** (σ^2).

- **Scenario-3:** If X has a general distribution but we have a large sample size $(n \geq 30)$.

Hypothesis Test for a Population Mean (μ)

▷ **Scenario-1:** If X has a normal distribution with ***unknown population variance*** (σ^2).

1) **Set up Hypothesis:** We would like to test the following hypotheses at α% level of significance:

$$H_0: \mu = \mu_0 \ \ vs. \ \ H_1: \mu \neq \mu_0$$

$$or, \ \ H_0: \mu \leq \mu_0 \ \ vs. \ \ H_1: \mu > \mu_o$$

$$or, \ \ H_0: \mu \geq \mu_0 \ \ vs. \ \ H_1: \mu < \mu_o$$

2) **Test Statistic:**

$$t_{cal} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

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

Hypothesis Test for a Population Mean (μ)

▷ **Scenario-1:** If X has a normal distribution with ***unknown population variance*** (σ^2).

3) Rejection Region: at α% level of significance

$\boldsymbol{When\ \ H_1: \mu \neq \mu_0\ (Two-sided\ Test)}$

$The\ rejection\ region\ is:$

$$(-\infty, -t_{\alpha/2,(n-1)}] \cup [t_{\alpha/2,(n-1)}, +\infty)$$

$$\therefore |t_{cal}| > t_{\alpha/2,(n-1)}$$

$\boldsymbol{When\ \ H_1: \mu > \mu_0\ (One-sided\ Test)}$

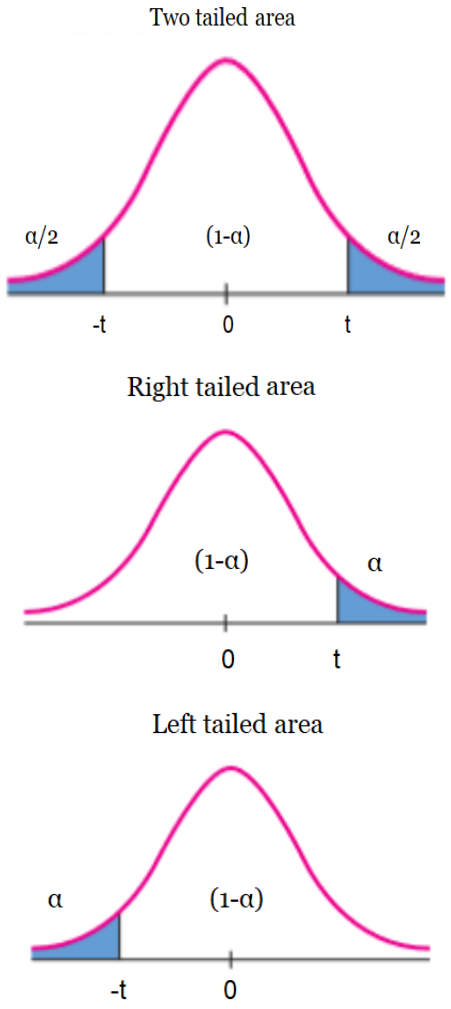
$The\ rejection\ region\ is: [t_{\alpha,(n-1)}, +\infty)$

$$\therefore t_{cal} > t_{\alpha,(n-1)}$$

$\boldsymbol{When\ \ H_1: \mu < \mu_0\ (One-sided\ Test)}$

$The\ rejection\ region\ is: (-\infty, -t_{\alpha,(n-1)}]$

$$\therefore t_{cal} < -t_{\alpha,(n-1)} \quad or \quad |t_{cal}| > |t_{\alpha,(n-1)}|$$



Hypothesis Test for a Population Mean (μ)

- **Scenario-1:** If X has a normal distribution with ***unknown population variance*** (σ^2).

4) **Decision:** If the calculated value is greater than the tabulated value, we may reject the null hypothesis (H_0), otherwise, we may fail to reject the (H_0), i.e.,

$$\boldsymbol{When\ \ H_1: \mu \neq \mu_0\ (Two-sided\ Test)}$$

$$\text{If } |t_{cal}| > t_{\alpha/2\,,(n-1)}; H_0 \text{ is rejected.}$$

$$\boldsymbol{When\ \ H_1: \mu > \mu_0\ (One-sided\ Test)}$$

$$\text{If } t_{cal} > t_{\alpha,(n-1)}; H_0 \text{ is rejected.}$$

$$\boldsymbol{When\ \ H_1: \mu < \mu_0\ (One-sided\ Test)}$$

$$\text{If } t_{cal} < -t_{\alpha,(n-1)} \quad or \quad |t_{cal}| > t_{\alpha,(n-1)}; H_0 \text{ is rejected.}$$

Tabulated value $t_{\alpha/2\,,(n-1)}$ can be calculated from $t-table$.

Exercise

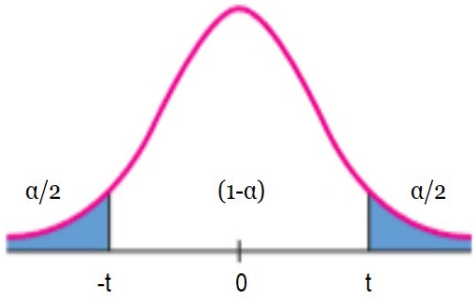
- **Example 1:** A machine is set to cut metal plates to a length of 44.350 mm. The lengths of a random sample of 24 metal plates have a sample mean of $\bar{x} = 44.364$ mm and a sample standard deviation of $s = 0.019$ mm. Is there any evidence that the machine is miscalibrated at 5% level of significance? What is the p-value of the test?
- **Solution:** Consider the following hypothesis:

$$H_0: \mu = 44.350 \ \ vs. \ \ H_1: \mu \neq 44.350$$

$Given\ that\ n = 24, \bar{x} = 44.364, s = 0.019$

The test statistic is

$$t_{cal} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{44.364 - 44.350}{\frac{0.019}{\sqrt{24}}} = 3.61$$



At 5% ($\alpha = 0.05$) level of significance, the rejection region is

$$\left(-\infty, -t_{0.025,\,23}\right] \cup \left[t_{0.025,\,23}, +\infty\right) = (-\infty, -2.069\,] \cup [2.069, +\infty)$$

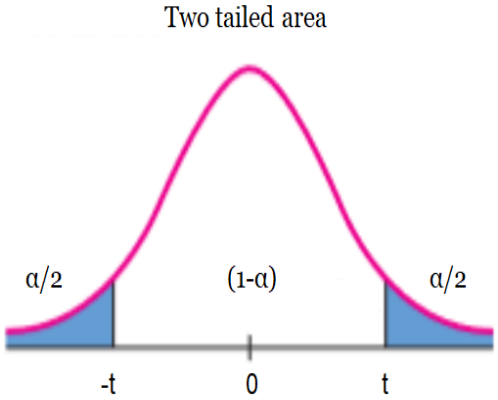
$$\therefore |t_{cal}| > 2.069$$

Comment: Since $|t_{cal}| > t_{tab}$, we may reject the H_0. Therefore, we may conclude that there is an evidence that the machine is miscalibrated at 5% level of significance.

Student t – *distribution* Table

Degrees of Freedom	Area in Upper Tail .20	.10	.05	.025	.01	.005
1	1.376	3.078	6.314	12.706	31.821	63.656
2	1.061	1.886	2.920	4.303	6.965	9.925
3	.978	1.638	2.353	3.182	4.541	5.841
4	.941	1.533	2.132	2.776	3.747	4.604
5	.920	1.476	2.015	2.571	3.365	4.032
6	.906	1.440	1.943	2.447	3.143	3.707
7	.896	1.415	1.895	2.365	2.998	3.499
8	.889	1.397	1.860	2.306	2.896	3.355
9	.883	1.383	1.833	2.262	2.821	3.250
10	.879	1.372	1.812	2.228	2.764	3.169
11	.876	1.363	1.796	2.201	2.718	3.106
12	.873	1.356	1.782	2.179	2.681	3.055
13	.870	1.350	1.771	2.160	2.650	3.012
14	.868	1.345	1.761	2.145	2.624	2.977
15	.866	1.341	1.753	2.131	2.602	2.947
16	.865	1.337	1.746	2.120	2.583	2.921
17	.863	1.333	1.740	2.110	2.567	2.898
18	.862	1.330	1.734	2.101	2.552	2.878
19	.861	1.328	1.729	2.093	2.539	2.861
20	.860	1.325	1.725	2.086	2.528	2.845
21	.859	1.323	1.721	2.080	2.518	2.831
22	.858	1.321	1.717	2.074	2.508	2.819
23	.858	1.319	1.714	2.069	2.500	2.807
24	.857	1.318	1.711	2.064	2.492	2.797
25	.856	1.316	1.708	2.060	2.485	2.787
26	.856	1.315	1.706	2.056	2.479	2.779
27	.855	1.314	1.703	2.052	2.473	2.771

$\frac{\alpha}{2}$ or α



Entries in the table give t values for an area or probability in the upper tail of the t distribution. For example:

$t_{\alpha/2,(n-1)} = t_{0.025,\,23} = 2.069$

Student $t-distribution$ Table

TABLE D *t* Distribution Critical Values

	Upper-tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

p-value = 0.001

- **Finding the p-value**
- At 23 d.f. with the test statistic value (3.61), the approximate p-value is

  $P(t_{cal} > 3.61) = 0.001$
- Since $p < \alpha$, H_0 is rejected.

Solution in STATA

- **Solution:** Consider the following hypothesis:

$$H_0: \mu = 44.350 \ \ vs. \ \ H_1: \mu \neq 44.350$$

$$Given\ that\ n = 24, \bar{x} = 44.364, s = 0.019, \alpha = 0.05$$

STATA Command: `ttesti` #obs #mean #sd #val [, level(#)]

```
. ttesti 24 44.364 0.019 44.350, level (95)

One-sample t test
```

	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
x	24	44.364	.0038784	.019	44.35598	44.37202

```
    mean = mean(x)                                t =   3.6098
Ho: mean = 44.350                 degrees of freedom =       23

  Ha: mean < 44.350               Ha: mean != 44.350               Ha: mean > 44.350
 Pr(T < t) = 0.9993         Pr(|T| > |t|) = 0.0015          Pr(T > t) = 0.0007
```

p-value = 0.0015

Comment: Since $p < \alpha$, we may reject the H_0. Therefore, we may conclude that there is an evidence that the machine is miscalibrated at 5% level of significance.

Exercise

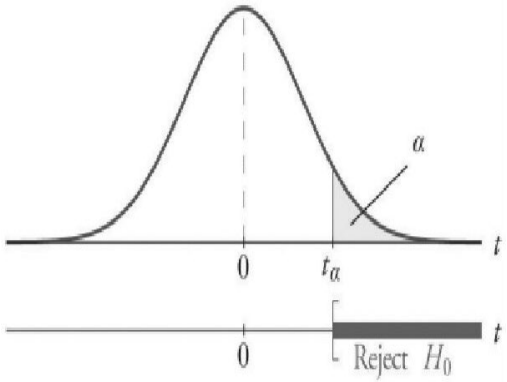
- **Example 2:** A chemical plant is required to maintain ambient sulfur levels in the working environment atmosphere at an average level of no more than 12.50. The results of 15 randomly timed measurements of the sulfur level produced a sample mean of $\bar{x} = 14.82$ and a sample standard deviation of $s = 2.91$. What is the evidence that the chemical plant is in violation of the working code at 5% level of significance? What is the p-value of the test?
- **Solution:** Consider the following hypothesis:

$$H_0: \mu \leq 12.50 \;\; vs. \;\; H_1: \mu > 12.50$$

$Given\ that\ n = 15, \bar{x} = 14.82, s = 2.91$

The test statistic is

$$t_{cal} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{14.82 - 12.50}{\frac{2.91}{\sqrt{15}}} = 3.09$$



At 5% ($\alpha = 0.05$) level of significance, the rejection region is

$$\left[t_{0.05,\,14}, +\infty\right) = [1.761, +\infty)$$

$$\therefore t_{cal} > 1.761$$

Comment: Since $t_{cal} > t_{tab}$, we may reject the H_0. Therefore, we may conclude that there is an evidence that the chemical plant is in violation of the working code at 5% level of significance.

Student $t-distribution$ Table

TABLE D *t* Distribution Critical Values

	Upper-tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

p-value = 0.005

- **Finding the p-value**
- At 14 d.f. with the test statistic value (3.09), the approximate p-value is

  $P(t_{cal} > 3.09) = 0.005$
- Since $p < \alpha$, H_0 is rejected.

Solution in STATA

▷ **Solution 2:** Consider the following hypothesis:

$$H_0: \mu \leq 12.50 \ \ vs. \ \ H_1: \mu > 12.50$$

$$Given\ that\ n = 15, \bar{x} = 14.82, s = 2.91, \alpha = 0.05$$

STATA Command: `ttesti` #obs #mean #sd #val [, level(#)]

```
. ttesti 15 14.82 2.91 12.50, level(95)

One-sample t test
```

	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
x	15	14.82	.7513588	2.91	13.2085	16.4315

```
    mean = mean(x)                                t =   3.0877
Ho: mean = 12.50                     degrees of freedom =     14

  Ha: mean < 12.50               Ha: mean != 12.50                Ha: mean > 12.50
 Pr(T < t) = 0.9960         Pr(|T| > |t|) = 0.0080          Pr(T > t) = 0.0040
```

p-value = 0.004

Comment: Since $p < \alpha$, we may reject the H_0. Therefore, we may conclude that there is an evidence that the chemical plant is in violation of the working code at 5% level of significance.

Student $t-$ *distribution* Table

TABLE D *t* Distribution Critical Values

	Upper-tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

p-value = 0.02

- **Finding the p-value**
- At 15 d.f. with the test statistic value (-2.282), the approximate p-value is

  $P(t \leq |-2.282|) = 0.02$
- Since $p < \alpha$, H_0 is rejected.

C.W: *t*-test: One sample

- Example: An outbreak of Salmonella-related illness was attributed to ice cream produced at a certain factory. Scientists measured the level of Salmonella in 9 randomly sampled batches of ice cream. The levels (in MPN/g) were:

  0.593, 0.142, 0.329, 0.691, 0.231, 0.793, 0.519, 0.392, 0.418

  Is there evidence that the mean level of Salmonella in the ice cream is greater than 0.3 MPN/g?

- **Solution:** The null and alternative hypotheses are

$$H_0: \mu \leq 0.3$$
$$H_1: \mu > 0.3$$

n=9
xbar= 0.456444
s=0.21284
aplha=0.05
pval= 0.0293
Reject null hypothe

Two Independent Samples $t-test$

- Let $x_{11}, x_{12}, \dots, x_{1n_1}$ be a random sample of size n_1 from a random variable X_1 with mean μ_1 and variance σ_1^2, and $x_{21}, x_{22}, \dots, x_{2n_2}$ be a random sample of size n_2 from a random variable X_2 with mean μ_2 and variance σ_2^2 . We would like to test the following hypotheses at $\alpha\%$ level of significance:

$$\boldsymbol{H_0: \mu_1 = \mu_2 \ \ vs. \ \ H_1: \mu_1 \neq \mu_2 \quad [Two-sided\ test\}}$$

$$\boldsymbol{or, \ \ H_0: \mu_1 \leq \mu_2 \ \ vs. \ \ H_1: \mu_1 > \mu_2 \quad [One-sided\ test]}$$

$$\boldsymbol{or, \ \ H_0: \mu_1 \geq \mu_2 \ \ vs. \ \ H_1: \mu_1 < \mu_2 \quad [One-sided\ test]}$$

- **Assumptions:**

1) Samples are come from normal population.
2) They are independent.
3) Population variances are **unknown** and assumed that they are equal, i.e., $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Two Independent Samples $t-test$

1) **Set up Hypothesis:** We would like to test the following hypotheses at α% level of significance:

$$H_0: \mu_1 = \mu_2 \ \ vs. \ \ H_1: \mu_1 \neq \mu_2 \quad [Two-sided\ test]$$

$$or, \ \ H_0: \mu_1 = \mu_2 \ \ vs. \ \ H_1: \mu_1 > \mu_2 \quad [One-sided\ test]$$

$$or, \ \ H_0: \mu_1 = \mu_2 \ \ vs. \ \ H_1: \mu_1 < \mu_2 \quad [One-sided\ test]$$

2) **Test Statistic:**

$$t_{cal} = \frac{\bar{x}_1 - \bar{x}_2}{s_p \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

$$where, \ \ s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}, Pooled\ variance$$

$$s_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2 \ \text{ and } \ s_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(x_{2i}-\bar{x}_2)^2$$

Two Independent Samples $t-test$

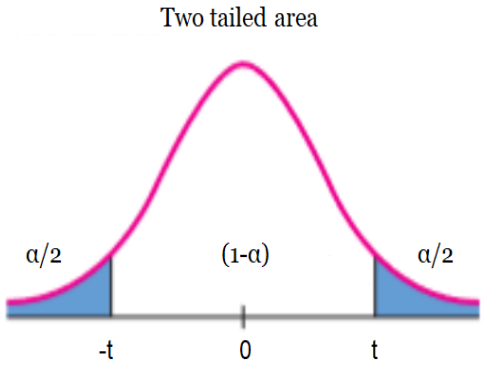
3) Rejection Region: at $\alpha\%$ level of significance

$\boldsymbol{When\ \ H_1: \mu_1 \neq \mu_2\ (Two-sided\ Test)}$

The rejection region is:

$$\left(-\infty,\ -t_{\frac{\alpha}{2},(n_1+n_2-2)}\right] \cup \left[t_{\frac{\alpha}{2},(n_1+n_2-2)},\ +\infty\right)$$

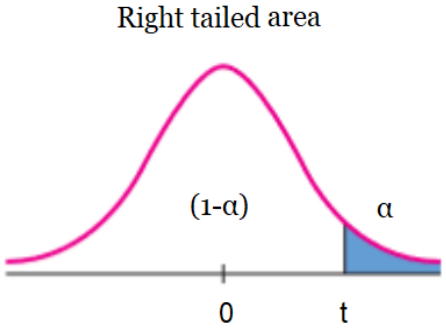
$$\therefore |t_{cal}| > t_{\alpha/2,(n_1+n_2-2)}$$



$\boldsymbol{When\ \ H_1: \mu_1 > \mu_2\ (One-sided\ Test)}$

The rejection region is: $\left[t_{\alpha,(n_1+n_2-2)},\ +\infty\right)$

$$\therefore t_{cal} > t_{\alpha,(n_1+n_2-2)}$$

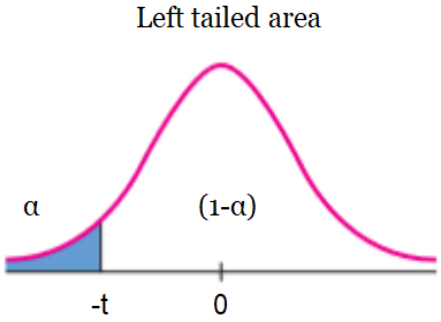


$\boldsymbol{When\ \ H_1: \mu_1 < \mu_2\ (One-sided\ Test)}$

The rejection region is: $\left(-\infty,\ -t_{\alpha,(n_1+n_2-2)}\right]$

$$\therefore t_{cal} < -t_{\alpha,(n_1+n_2-2)}$$

$$\Rightarrow |t_{cal}| > |t_{\alpha,(n_1+n_2-2)}|$$



Two Independent Samples $t-test$

4) **Decision:** If the calculated value falls in the rejection region, we may reject the null hypothesis (H_0), otherwise, we may fail to reject the (H_0), i.e.,

$$\boldsymbol{When\ \ H_1: \mu_1 \neq \mu_2\ (Two-sided\ Test)}$$

$$\text{If } |t_{cal}| > t_{\alpha/2,(n_1+n_2-2)};\ H_0\ is\ rejected.$$

$$\boldsymbol{When\ \ H_1: \mu_1 > \mu_2\ (One-sided\ Test)}$$

$$\text{If } t_{cal} > t_{\alpha,(n_1+n_2-2)};\ H_0\ is\ rejected.$$

$$\boldsymbol{When\ \ H_1: \mu_1 < \mu_2\ (One-sided\ Test)}$$

$$\text{If } t_{cal} < -t_{\alpha,(n_1+n_2-2)};\ H_0\ is\ rejected.$$

Tabulated value$-t_{\alpha/2,(n_1+n_2-2)}$can be calculated from $t-table$.

Exercise

- **Example:** The purpose of a study by Tam et al. (A-6) was to investigate wheelchair maneuvering in individuals with lower-level spinal cord injury (SCI) and healthy controls (C). Subjects used a modified wheelchair to incorporate a rigid seat surface to facilitate the specified experimental measurements. Interface pressure measurement was recorded by using a high-resolution pressure-sensitive mat with a spatial resolution of four sensors per square centimeter taped on the rigid seat support. During static sitting conditions, average pressures were recorded under the ischial tuberosities (the bottom part of the pelvic bones). The data for measurements of the left ischial tuberosity (in mm Hg) for the SCI and control groups are shown in Table 1.We wish to know if we may conclude, on the basis of these data, that, in general, healthy subjects exhibit lower pressure than SCI subjects at 5% level of significance.

Table 1: Pressures (mm Hg) Under the Pelvis during Static Conditions

Control	131	115	124	131	122	117	88	114	150	169
SCI	60	150	130	180	163	130	121	119	130	148

Exercise

- **Solution:** Consider the following hypothesis:

$$H_0: \mu_C \geq \mu_{SCI} \quad vs. \quad H_1: \mu_C < \mu_{SCI}$$

$$Given\ that\ n_C = 10, \bar{x}_C = 126.1, s_C^2 = 475.24,$$

$$n_{SCI} = 10, \bar{x}_{SCI} = 133.1, s_{SCI}^2 = 1036.84, \alpha = 0.05$$

$$Here,\ \ s_p^2 = \frac{(n_C - 1)s_C^2 + (n_{SCI} - 1)s_{SCI}^2}{n_C + n_{SCI} - 2}$$

$$= \frac{(9 \times 475.24) + (9 \times 1036.84)}{18} = 756.04$$

The test statistic is

$$t_{cal} = \frac{\bar{x}_C - \bar{x}_{SCI}}{s_p \times \sqrt{\frac{1}{n_C} + \frac{1}{n_{SCI}}}} = \frac{126.1 - 133.1}{27.4962 \times \sqrt{\frac{1}{10} + \frac{1}{10}}} = -0.5693$$

At 5% ($\alpha = 0.05$) level of significance, the rejection region is

$$t_{tab} = \left(-\infty, -t_{\alpha,\, n_c + n_{SCI} - 2}\right] = \left(-\infty, -t_{0.05,\, 18}\right] = (-\infty, -1.734]$$

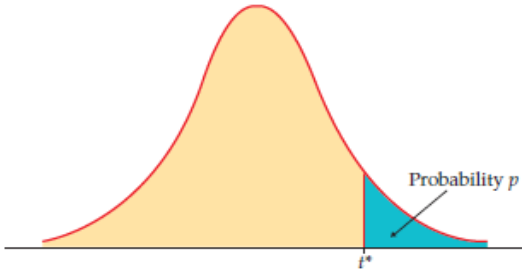
$$\therefore |t_{cal}| < |t_{tab}|$$

Comment: Since the calculate value do not fall in the rejection region, so we may not reject the H_0. Therefore, we may conclude that healthy subjects do not exhibit lower pressure than SCI subjects.

Student t – *distribution* Table

TABLE D *t* Distribution Critical Values

	Upper-tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											



Solution in STATA

- **Solution:** Consider the following hypothesis:

$$H_0: \mu_C \geq \mu_{SCI} \quad vs. \quad H_1: \mu_C < \mu_{SCI}$$

Using Calculator, we have $n_C = 10, \bar{x}_C = 126.1, s_C^2 = 475.24,$

$$n_{SCI} = 10, \bar{x}_{SCI} = 133.1, s_{SCI}^2 = 1036.84$$

Command: ***ttesti #obs1 #mean1 #sd1 #obs2 #mean2 #sd2***

```
. ttesti 10 126.1 21.8 10 133.1 32.2

Two-sample t test with equal variances
```

	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
x	10	126.1	6.893765	21.8	110.5052	141.6948
y	10	133.1	10.18253	32.2	110.0655	156.1345
Combined	20	129.6	6.037976	27.00265	116.9624	142.2376
diff		-7	12.29667		-32.83434	18.83434

```
    diff = mean(x) - mean(y)                                  t =  -0.5693
H0: diff = 0                     Degrees of freedom =       18

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.2881         Pr(|T| > |t|) = 0.5762          Pr(T > t) = 0.7119
```

Comment: Since $p > \alpha$, we may not reject the H_0. Therefore, we may conclude that healthy subjects do not exhibit lower pressure than SCI subjects.

Exercise

- **Example:** In an experiment, we compare the result of treatment A and treatment B by seeing the survival time of mouses.

  Treatment A: 17, 19, 15, 18, 21, 18

  Treatment B: 18, 15, 13, 16, 13

  Investigate whether treatment A and treatment B gives different survival time at 5% level of significance.

- **Solution:** Consider the following hypothesis:

$$H_0: \mu_1 = \mu_2 \ \ vs. \ \ H_1: \mu_1 \neq \mu_2$$

Using the calculator, we have, $n_1 = 6, \bar{x}_1 = 18, s_1^2 = 4,$

$$n_2 = 5, \bar{x}_2 = 15, s_2^2 = 4.5, \alpha = 0.05$$

Solution in STATA

- **Solution:** Consider the following hypothesis:

$$H_0: \mu_1 = \mu_2 \ \ vs. \ \ H_1: \mu_1 \neq \mu_2$$

$$Using\ calculator, \quad n_1 = 6, \bar{x}_1 = 18, s_1^2 = 4,$$

$$n_2 = 5, \bar{x}_2 = 15, s_2^2 = 4.5, \alpha = 0.05$$

```
. ttesti 6 18 2 5 15 2.1213

Two-sample t test with equal variances
```

	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
x	6	18	.8164966	2	15.90113	20.09887
y	5	15	.9486742	2.1213	12.36606	17.63394
Combined	11	16.63636	.7540503	2.500902	14.95623	18.31649
diff		3	1.24424		.1853327	5.814667

```
    diff = mean(x) - mean(y)                                      t =   2.4111
H0: diff = 0                                     Degrees of freedom =        9

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9804         Pr(|T| > |t|) = 0.0392          Pr(T > t) = 0.0196
```

Comment: Since the $p < \alpha$, so we may reject the H_0. Therefore, we may conclude that there is an evidence that treatment A and treatment B gives different survival time.

t-test: Two independent samples

- **Body Temperature data description (.dta)**: The data represent the 100 people with heart rate and body temperature.

Let's import the data first!

- Question 1: Is there any significant changes in the heart rate of male and female?
- Recode the Gender variable into numerical coding (0=Male, 1=Female)

encode Gender, gen(tempvar)
recode tempvar (1=0 "Male") (2=1 "Female"), gen(Sex) label(Gender)
drop tempvar

- Check the distribution of heart rate variable!
- **Solution:** The null and alternative hypotheses are

$H_0: \mu_1 = \mu_2$ $(mean\ HR\ of\ male\ and\ female\ are\ same)$

$H_1: \mu_1 \neq \mu_2$ $(mean\ HR\ of\ male\ and\ female\ are\ different)$

Paired t-test

- Weight of the mice before treatment

  200.1, 190.9, 192.7, 213, 241.4, 196.9, 172.2, 185.5, 205.2, 193.7

- Weight of the mice after treatment

  392.9, 393.2, 345.1, 393, 434, 427.9, 422, 383.9, 392.3, 352.2

- **Question:** Is there any significant changes in the weights of mice before and after treatment?

Let's Import the data in STATA

Paired *t*-test

- **Solution:** Consider the following hypothesis:

The null hypothesis (two-sided) is:

$$H_0: \mu_d = 0$$

$(The\ population\ average\ weight\ loss\ is\ zero)\ \ vs.$

$$H_1: \mu_d \neq 0$$

$(The\ population\ average\ weight\ loss\ is\ not\ zero)$

$Given\ that\ n = 10, d_i = before_i - after_i\ \ (i = 1,2, \dots, 10)$

The test statistic is

$$t_{cal} = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} \sim t_{\frac{\alpha}{2},(n-1)}$$

- **Comment:** If $t_{cal} > t_{tab}\ (or,\ p < \alpha)$, we may reject the H_0.

Multiple Linear Regression Analysis

- The multiple linear regression model is

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} \ldots + \beta_p X_{ip} + \varepsilon_i}$$

Where,

- Y_i is the outcome for i
- β_0 is the intercept
- $\beta_1, \beta_2 \ldots, \beta_p$ are the slopes/regression coefficients
- $X_{i1}, X_{i2,\ldots,} X_{ip}$ are the predictors for i
- ε_i is the residual variation for i

$$\varepsilon_i \sim N(0, \sigma^2)$$

Example: Regression Analysis

- First import Heart Rate data (Data_HR.csv) data in STATA.
- The data represent 106 individuals with recorded heart rate and its associated factors. The information in the dataset is as follows:
  - Age: Years
  - Weight_kg: Body weight in kilograms
  - Height_cm: Height in centimeters
  - Fitness_Score: 0–100 scale (higher = more fit)
  - Stress_Level: 1–10 scale (higher = more stress)
  - Caffeine_mg: Caffeine consumed in mg (past 24 hours)
  - Medication: 1 = taking beta-blockers/heart meds, 0 = none
  - Resting_HR: Target variable - Resting heart rate (beats per minute).
  - sex: Gender (1=Male, 2=Female)
  - activity_level: Activity Level (1=Sedentary, 2=Moderate, 3= Active)
  - smoker: Skmoking Status (0=No, 1=Yes)

Example: Regression Analysis

- First import Heart Rate data (Data_HR.csv) data in STATA.
- The data represent 106 individuals with recorded heart rate and its associated factors.
  - Q1. Calculate BMI using the height and weight of the individuals.
  - Q2. Which factors most strongly correlate with resting heart rate? Comments on your results. (Use only continuous variables.)
  - Q3. Identify the risk factors associated with resting heart rate (in beats per minute). Comments on your results.

BMI Calculation

- Q1:

- In **STATA**, the new variable can be generated by

  gen BMI = weight*703/height^2

  generate BMI=weight*703/height^2

- If the units are in lbs and inches, the US customary system (USC) is used.

- In **STATA**, the new variable can be generated by

  gen BMI = weight/height^2

  generate BMI=weight/height^2

- If the units are in kg and meters, the International System of Units (SI) is used.

Thank You!