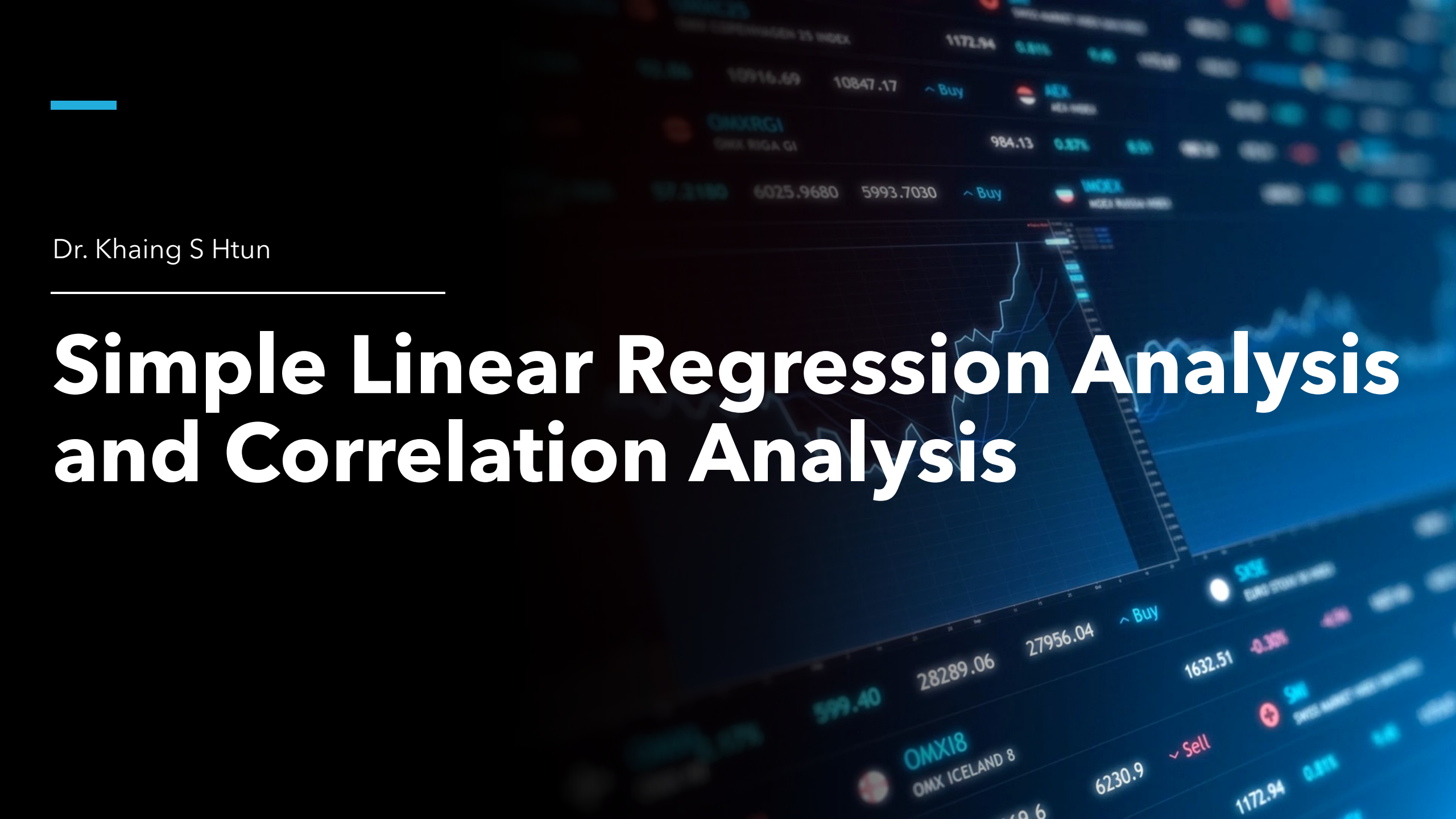
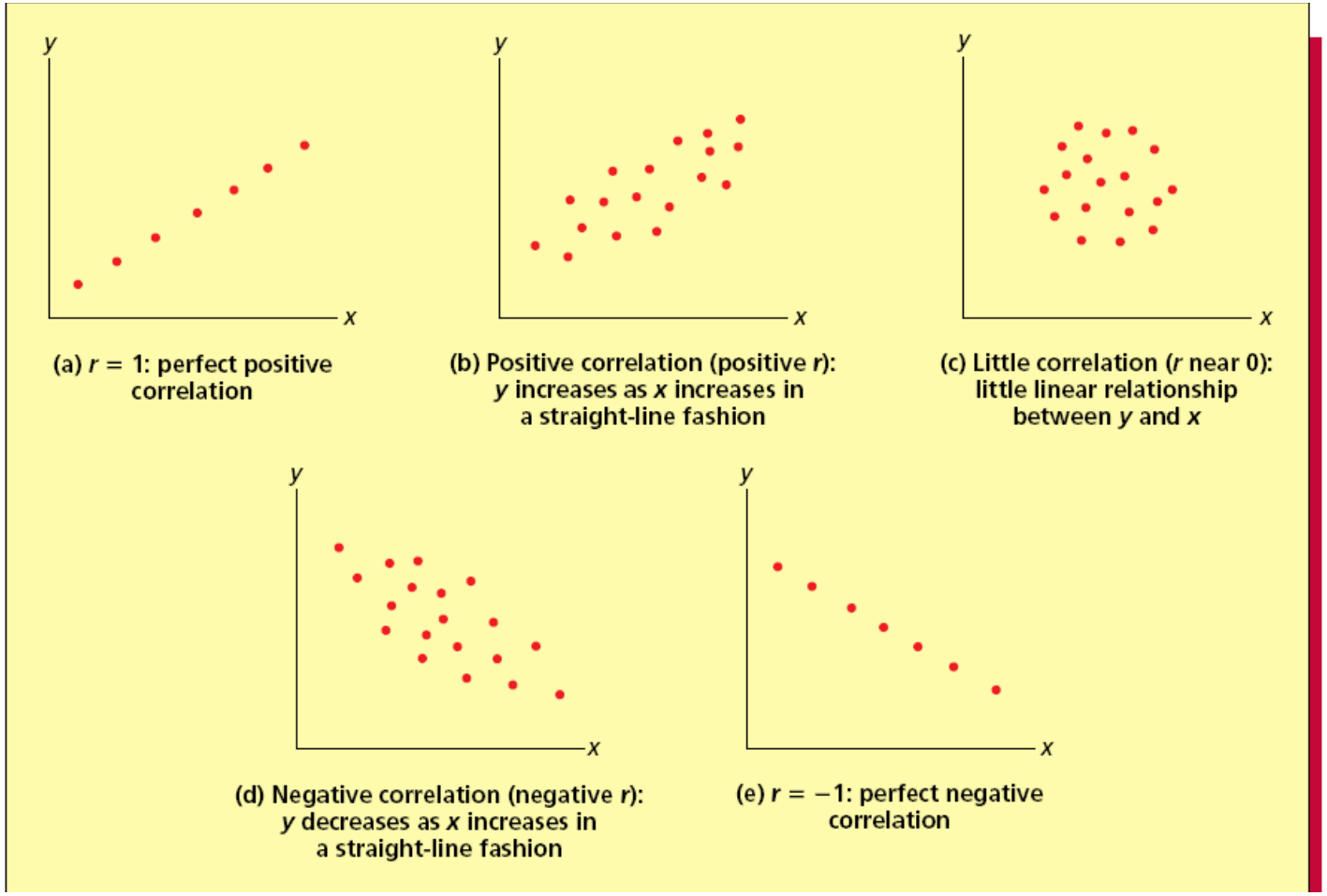


Dr. Khaing S Htun

Simple Linear Regression Analysis and Correlation Analysis

The Simple Correlation Analysis



- A measure of the linear relationship between variables and its strength

$$y = f(x)$$

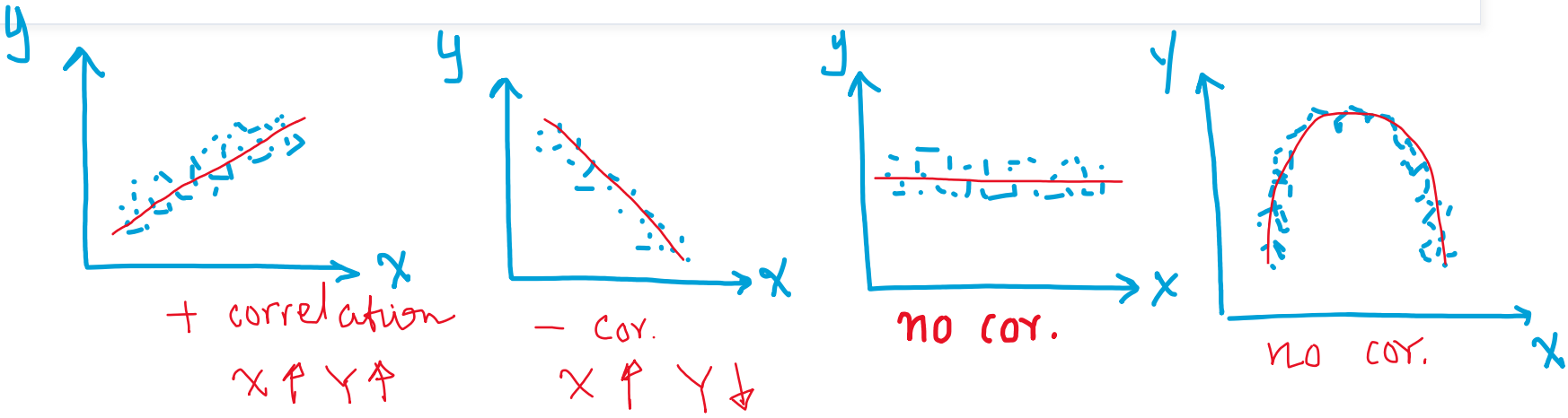
- Scatter diagram
  - Observed values (x_i, y_i),

$$i = 1, 2, 3, 4, \ldots, n$$

The Simple Correlation Analysis

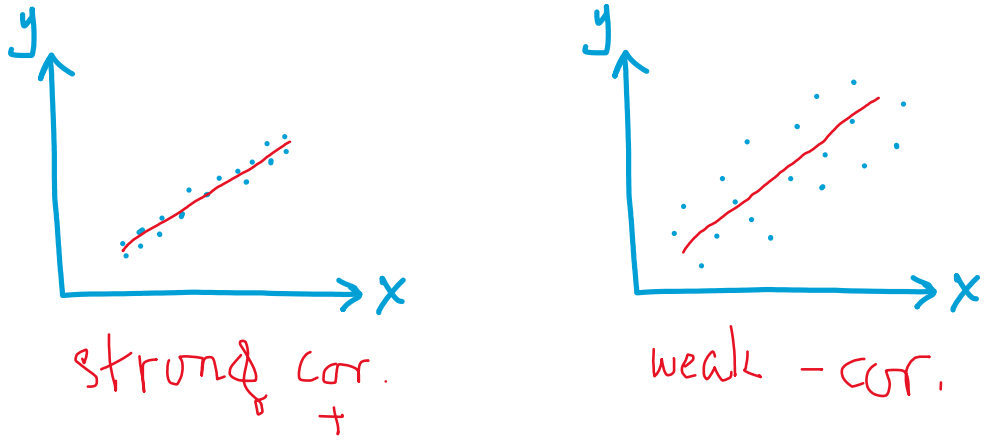
Direction

- Positive
- Negative
- No correlation



Strength

- Strong
- Medium
- Weak



The Simple Correlation Coefficient, r

- Karl Pearson about 1900

- **Correlation Coefficient, r**
  - -1 < **r** < 1
    - +1 = perfect positive linear relationship
    - 0 = no linear relationship
    - -1 = perfect negative linear relationship

Strength

- r= ±0.8 or higher → strong correlation
- r= ±0.5 – 0.8 → medium correlation
- r= ±0.4 or lower → weak correlation

The Simple Correlation Coefficient, r

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

n = the number of paired observations

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

where

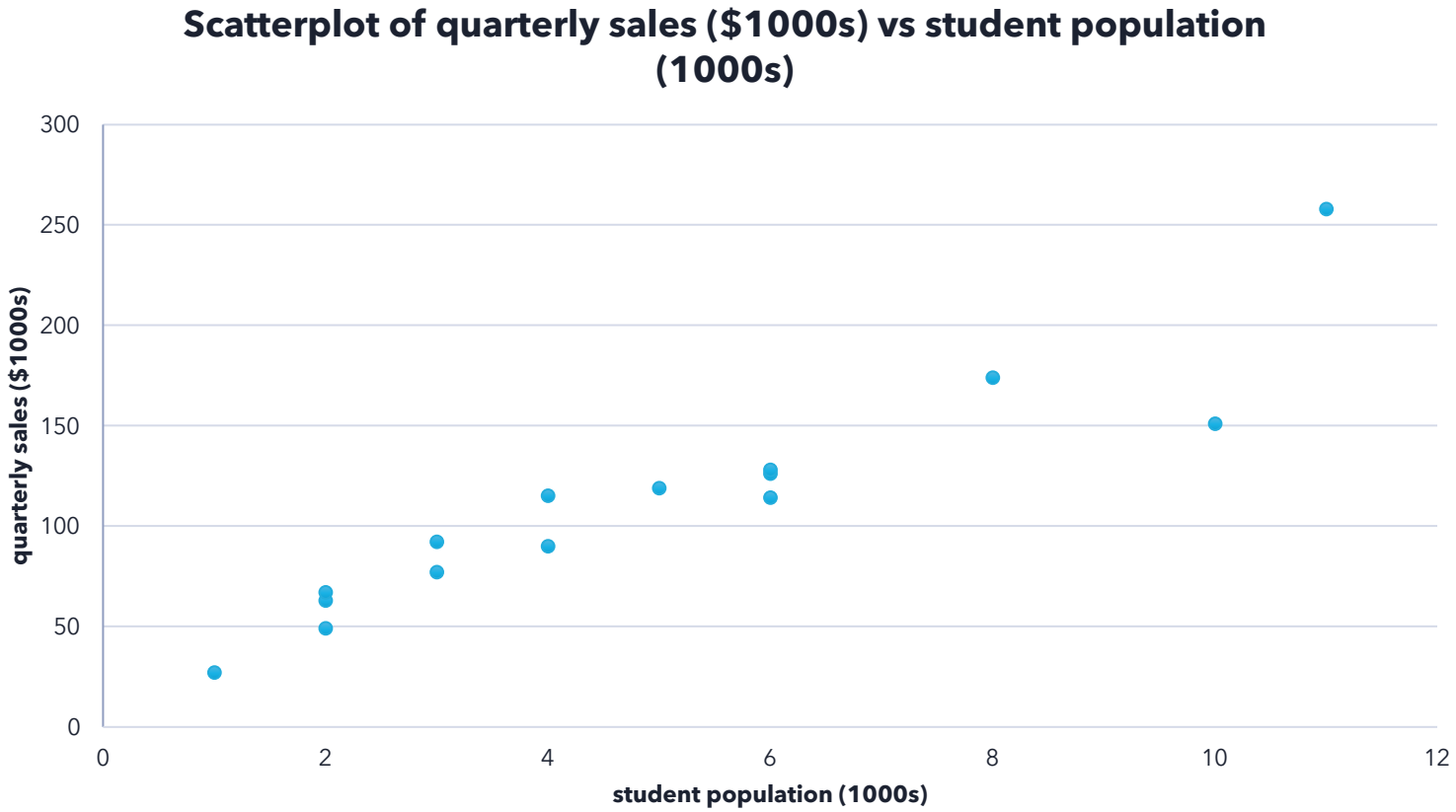
$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$SS_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

The Simple Correlation Coefficient, r
(example)

Restaurant	student population (1000s), x	quarterly sales ($1000s), y
1	3	92
2	2	63
3	6	126
4	8	174
5	2	49
6	4	90
7	5	119
8	6	114
9	2	67
10	4	115
11	6	128
12	11	258
13	3	77
14	10	151
15	1	27



The Simple Correlation Coefficient, r
(example)

Restaurant	student population (1000s), x	quarterly sales ($1000s), y	x^2	y^2	xy
1	3	92	9	8464	276
2	2	63	4	3969	126
3	6	126	36	15876	756
4	8	174	64	30276	1392
5	2	49	4	2401	98
6	4	90	16	8100	360
7	5	119	25	14161	595
8	6	114	36	12996	684
9	2	67	4	4489	134
10	4	115	16	13225	460
11	6	128	36	16384	768
12	11	258	121	66564	2838
13	3	77	9	5929	231
14	10	151	100	22801	1510
15	1	27	1	729	27
Sum, Σ	**73**	**1,650**	**481**	**226,364**	**10,255**

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

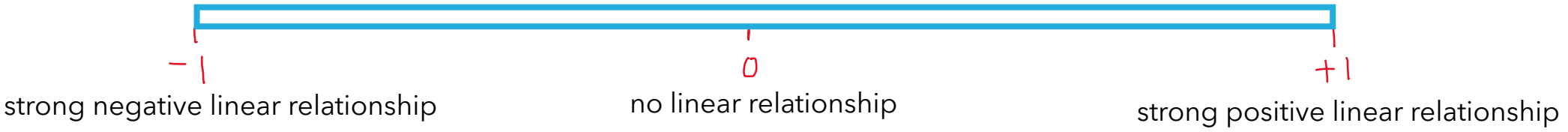
$$= \frac{15(10{,}255) - (73)(1{,}650)}{\sqrt{[15(481) - (73)^2][15(226{,}364) - (1{,}650)^2]}}$$

$$= \frac{153{,}825 - 120{,}450}{\sqrt{[1{,}886]\ [672{,}960]}} = \frac{33{,}375}{35{,}625.87} = +0.9368$$

relationship is **strong**

Testing the Significance of the Population Correlation Coefficient, ρ

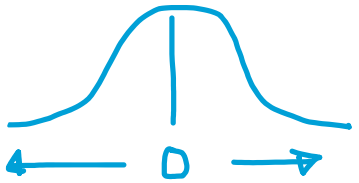
- r – **sample** correlation coefficient *(NOT population)*



$$the\ value\ of\ r\ is\ reliable\ or\ not?$$

Testing the Significance of the Population Correlation Coefficient, ρ

- ρ - **population** correlation coefficient *(NOT sample)*



Hypothesis testing:

1. State the **hypothesis** $H_0: \rho = 0$ $H_1: \rho \neq 0$
2. Find the **critical values** ⟹ from the table (α and d.f.)
3. Compute the **test value** (**t** test) formula
4. Make the **decision** with $\pm$ value
5. Summarize the result

Testing the Significance of the Population Correlation Coefficient, ρ

Hypotheses are:

Null hypothesis - H_0: $\rho = 0$

(There is no linear relationship between *y* and *x*, **or** the population correlation coefficient is zero.)

Alternative hypothesis - H_1: $\rho \neq 0$

(There is a linear relationship between *y* and *x*, **or** the population correlation coefficient is different from zero.)

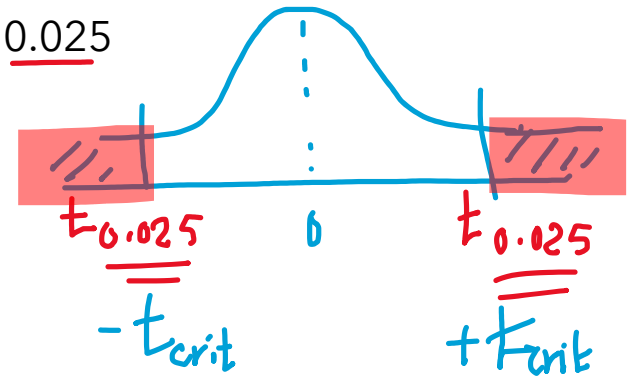
- *Reject* null hypothesis → significant difference between the value of r and 0
- *Fail to reject* → the value of r is **NOT** significantly different from 0

Testing the Significance of the Population Correlation Coefficient, ρ

Critical value

from the table

- degree of freedom, $d.f. = n - 2$
- α is the significant level
  - α = 1 – confidence level *(95%)*
  - α is 0.05 or 5%
    - 0.025 and 0.025



Degrees of Freedom	level of significant, α					
	.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

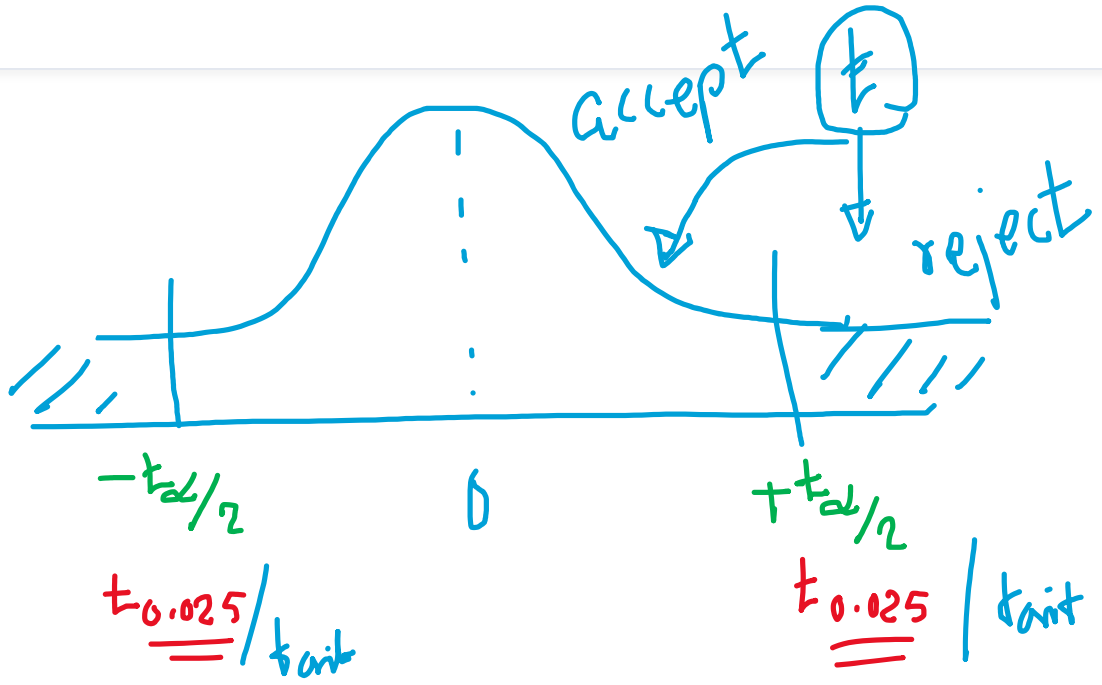
t_{crit} value

Testing the Significance of the Population Correlation Coefficient, ρ

from the formula

Test statistics,

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$



Decision rule: We will **reject** H_0 at α, if the value of $\mathbf{t \leq -t_{\alpha/2}}$ or $\mathbf{t \geq +t_{\alpha/2}}$

Testing the Significance of the Population Correlation Coefficient, ρ *(example)*

Restaurant	student population (1000s), x	quarterly sales ($1000s), y	x^2	y^2	xy
1	3	92	9	8464	276
2	2	63	4	3969	126
3	6	126	36	15876	756
4	8	174	64	30276	1392
5	2	49	4	2401	98
6	4	90	16	8100	360
7	5	119	25	14161	595
8	6	114	36	12996	684
9	2	67	4	4489	134
10	4	115	16	13225	460
11	6	128	36	16384	768
12	11	258	121	66564	2838
13	3	77	9	5929	231
14	10	151	100	22801	1510
15	1	27	1	729	27
Sum, Σ	**73**	**1,650**	**481**	**226,364**	**10,255**

n

r = +0.9368 – strong correlation

with $\alpha = 0.05$ and $d.f. = n - 2 = 15 - 2 = 13$

Test:

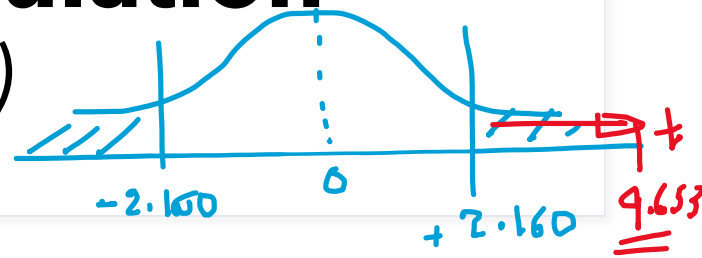
H_0: There is no a linear relationship between the quarterly sales(y) and the student population (x)

H_1: There is a linear relationship between the quarterly sales(y) and the student population(x)

test statistics,

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.9368\sqrt{15-2}}{\sqrt{1-(0.9368)^2}} = \frac{3.3777}{0.3499} = 9.653$$

Testing the Significance of the Population Correlation Coefficient, ρ *(example)*



Critical value

$\alpha = 0.05$ and $d.f. = n - 2 = 15 - 2 = 13$

- the critical value is $\pm t_{0.025} = \pm 2.160$
- computed t value, $t = 9.653$

Reject H_0 when $t \leq -t_{\alpha/2}$ or $t \geq +t_{\alpha/2}$

Therefore, H_0 is rejected at $\alpha = 0.05$.

Degrees of Freedom	level of significant, α					
	.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

we can conclude that there is a linear relationship between the student population(x) and the quarterly sales (y).

In-class Assignment

In-class Assignment: Correlation Coefficient

Find the value of the correlation coefficient, **r**, from the following table and test the significant of ρ

SUBJECT	AGE X	GLUCOSE LEVEL Y
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81