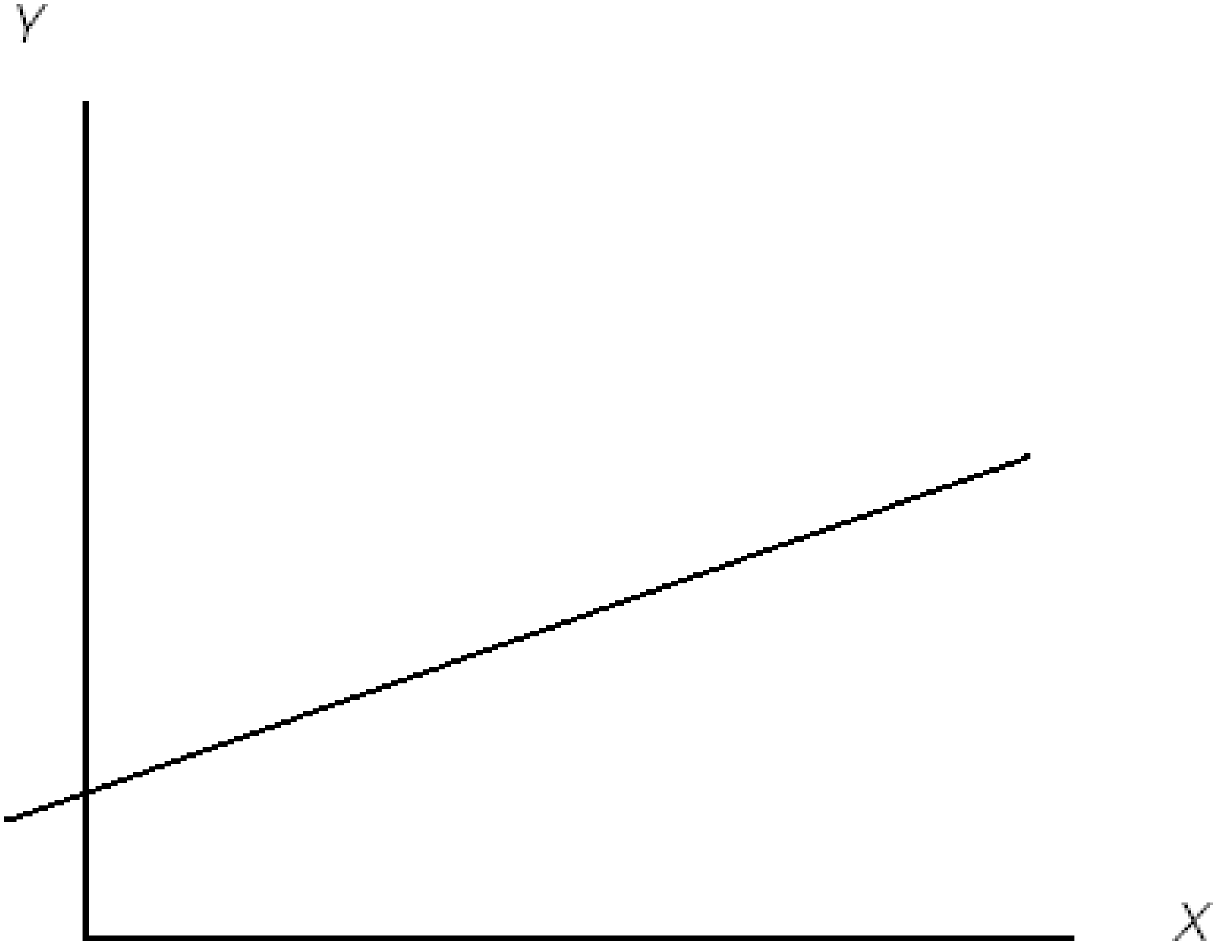
Warsaw Summer School 2023, OSU Study Abroad Program

Correlation

Linear Relationship

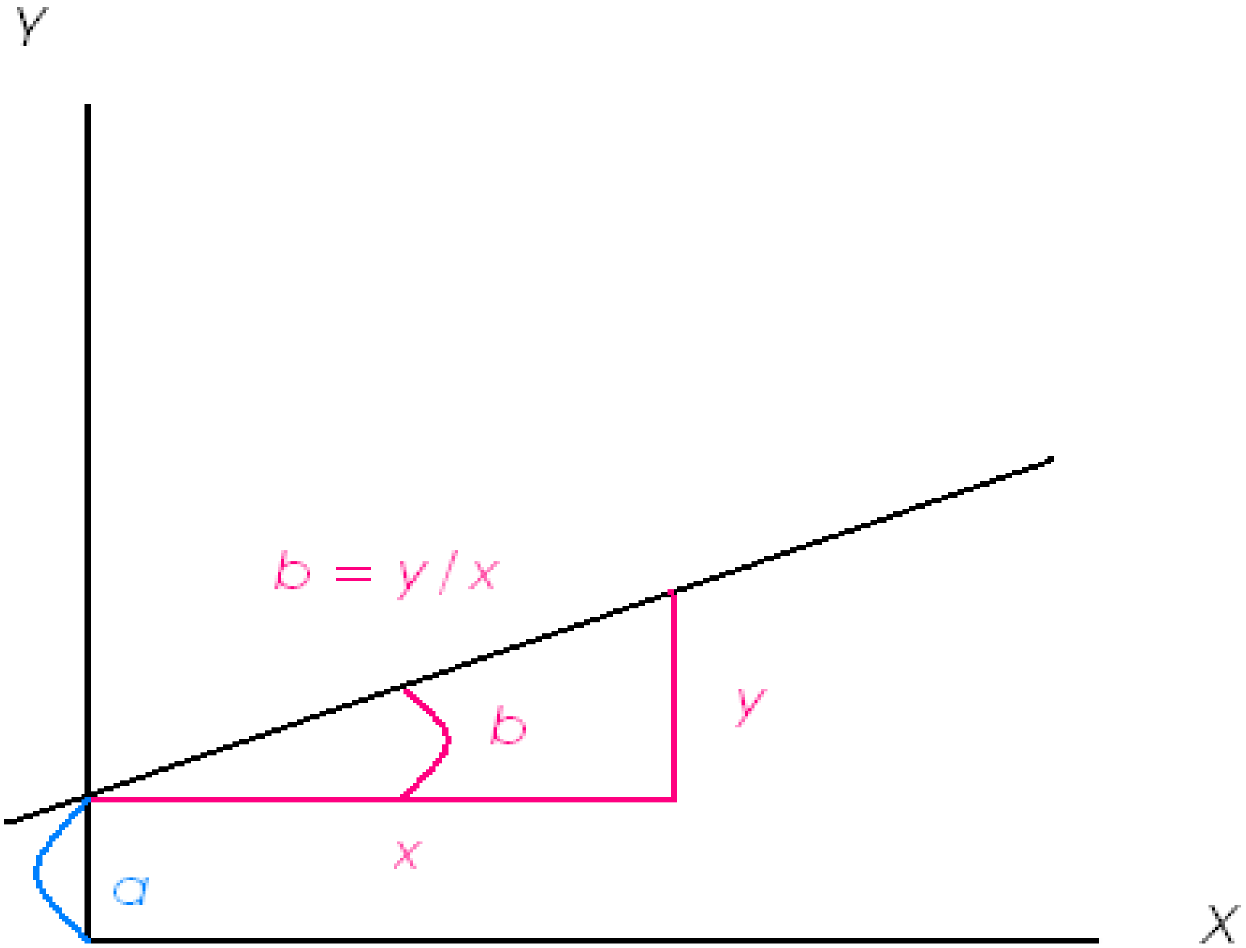


Linear Relationship

The line = a mathematical function that can be expressed through the formula $Y = a + bX$, where Y & X are our variables.

Y, the dependent variable, is expressed as a linear function of the independent (explanatory) variable X.

Linear Relationship



Linear Relationship

The constant a = value of Y at the point in which the line $Y = a + bX$ intersects the Y-axis (also called the intercept).

The slope b equals the change in Y for a one-unit increase in X (one-unit increase in X corresponds to a change of b units in Y). The slope describes the rate of change in Y-values, as X increases.

Verbal interpretation of the slope of the line:
"Rise over run": the rise divided by the run (the change in the vertical distance is divided by the change in the horizontal distance).

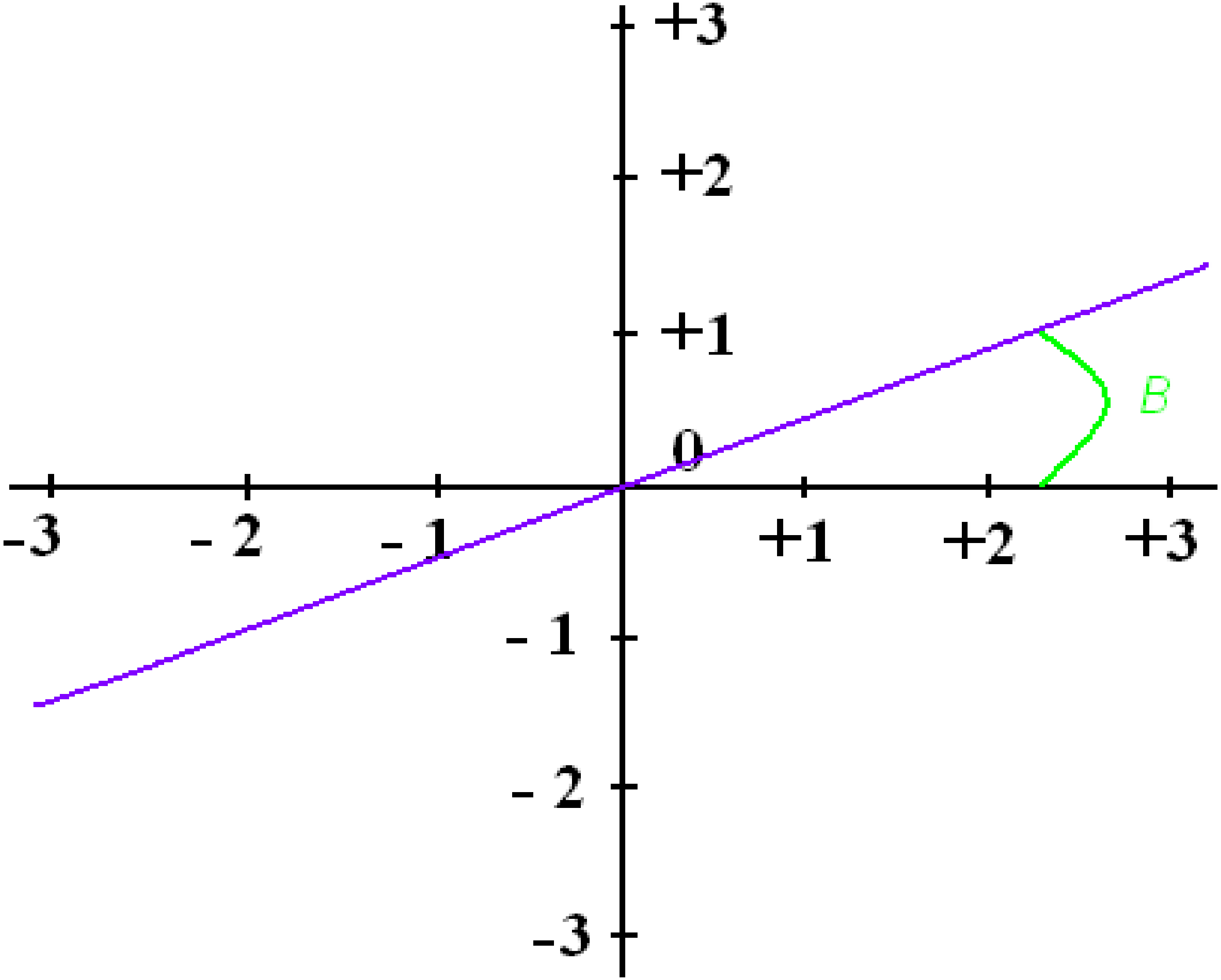
$$Y = a + bX$$

The formula $Y = a + bX$ maps out a strait-line graph with slope b and Y-intercept a.

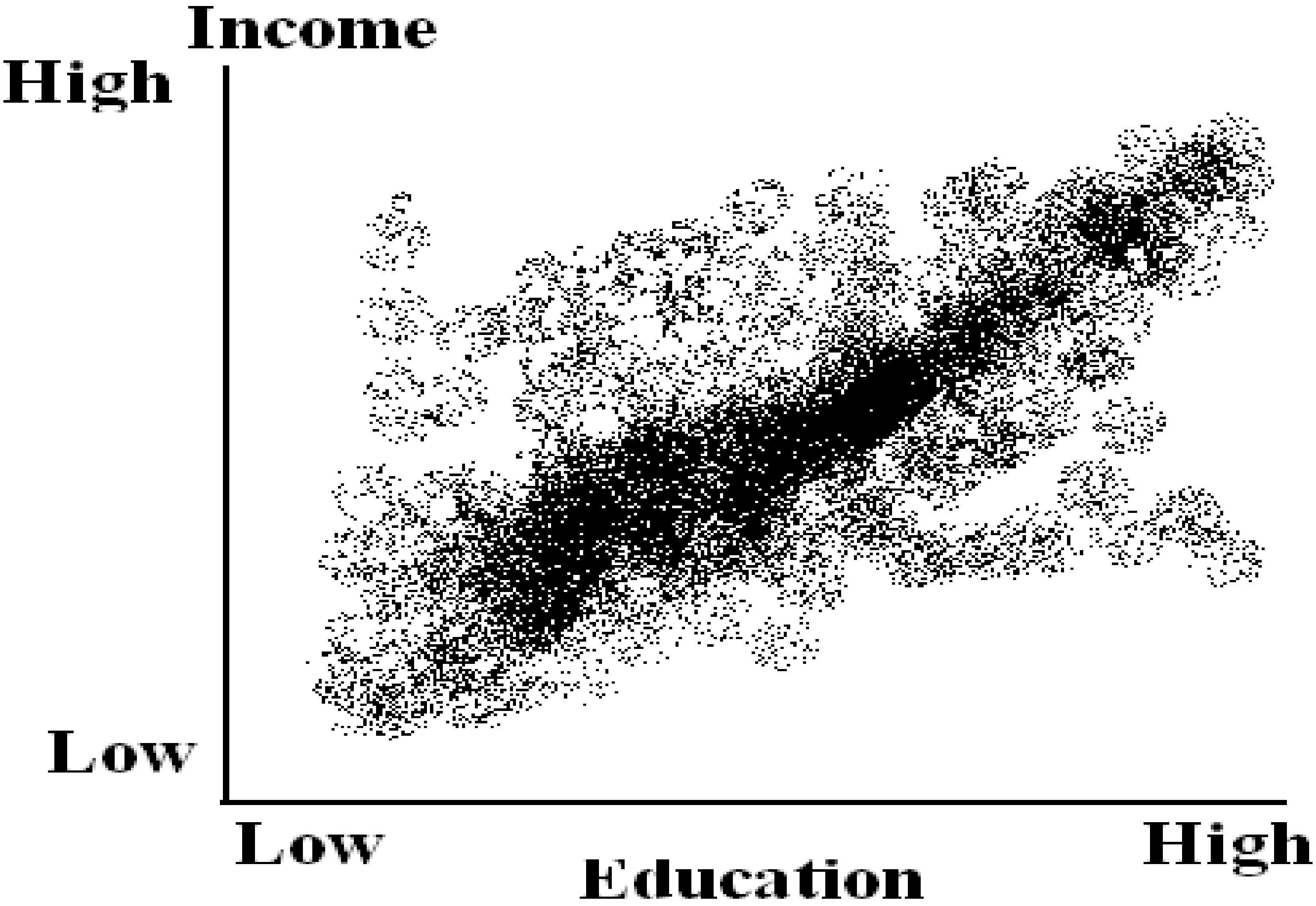
If Y and X are expressed in the standard scores (z-scores), then we have:

$$Z(y) = B*Z(x)$$

Z(y) = B * Z(x)



In reality data are scattered



r

Correlation is a statistical technique used to measure & describe a relationship between two variables (whether X and Y are related to each other).

How it works:

Correlation is represented by a line through the data points & by a number (the correlation coefficient) that indicates how close the data points are to the line.

For z-scores of Y & X the correlation equals B.

r

The stronger the relationship between X & Y, the closer the data points will be to the line. The weaker the relationship, the farther the data points will drift away from the line.

Correlation = a <u>unitless</u> statistic (it is standardized).

Thus, we can <u>directly compare</u> the strength of correlations for various pairs of variables.

Calculations and logic of Pearson's correlation

r = the sum of the products of the deviations from each mean, divided by the square root of the product of the sum of squares for each variable.

$$r = \frac{\Sigma\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{\sqrt{\Sigma\left(X - \overline{X}\right)^2 \Sigma\left(Y - \overline{Y}\right)^2}}$$

Scatterplot

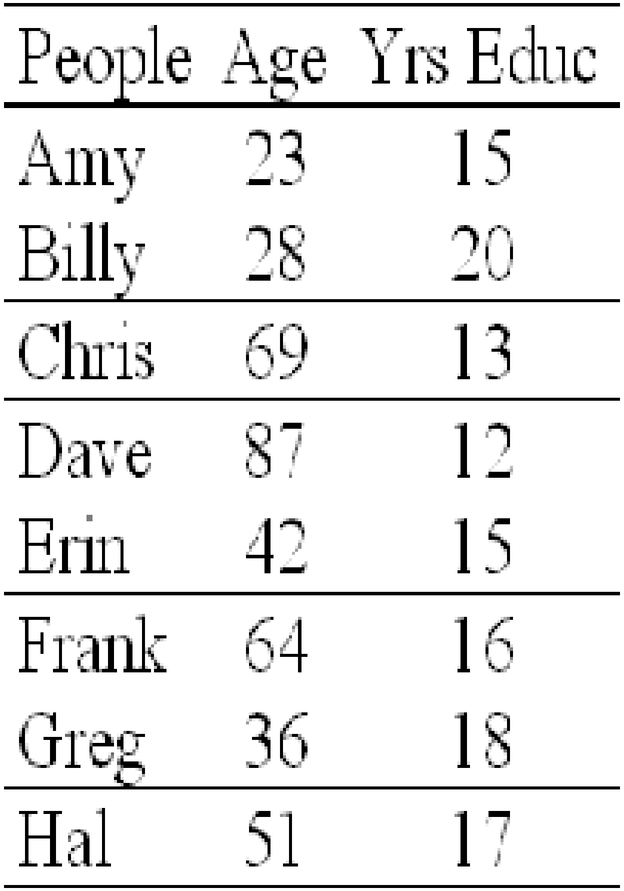
Scatterplot shows where each person falls in the distribution of X and the distribution of Y simultaneously.

Vertical Axis: the dependent variable (Y)

Horizontal Axis: the independent variable (X)

Each point is a person & their coordinates (their responses on the X variable and response on the Y variable) determine where the person goes in the graph.

Table



People	Age	Yrs Educ
Amy	23	15
Billy	28	20
Chris	69	13
Dave	87	12
Erin	42	15
Frank	64	16
Greg	36	18
Hal	51	17

Pearson's Correlation coefficient

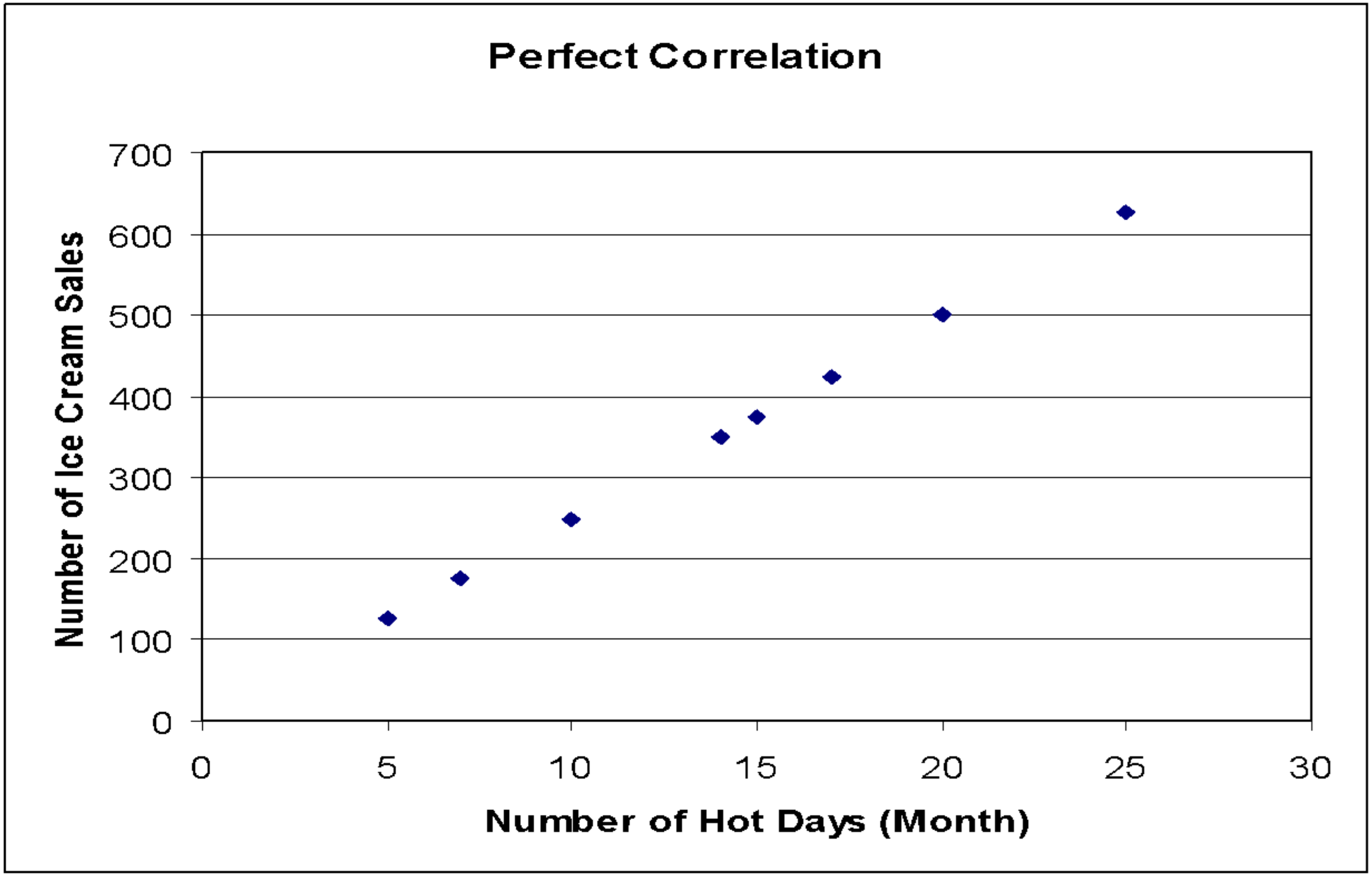
Pearson's Correlation coefficient

- **Denoted with "r" or "ρ" (Greek letter "rho").**

- **Ranges from –1 to +1 (cannot be outside this range)**

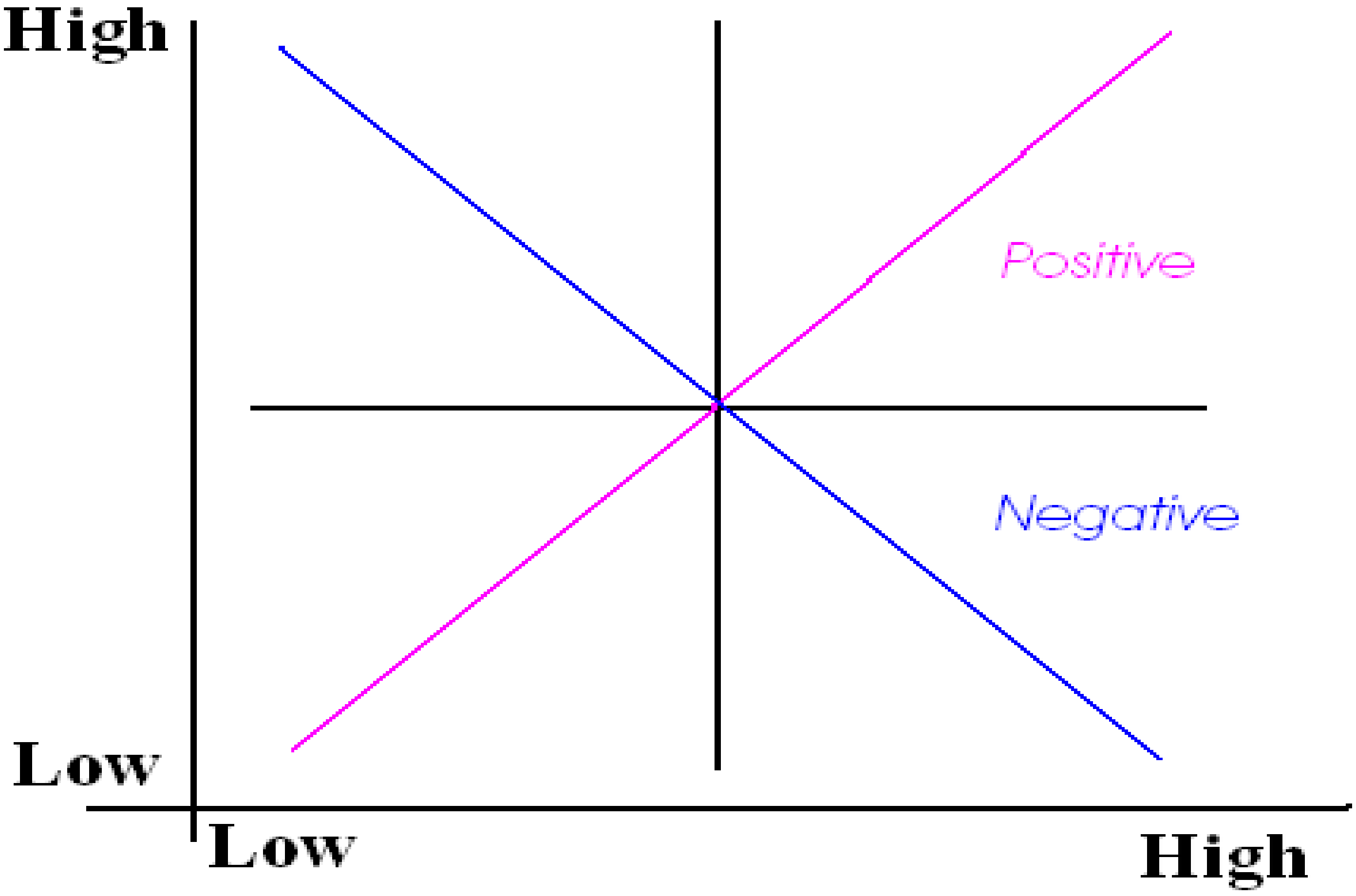
  r = +1 denotes perfect positive correlation between X &Y (all points on the line)

r = +1


Perfect Correlation
Number of Ice Cream Sales
700
600
500
400
300
200
100
0
0
5
10
15
20
25
30
Number of Hot Days (Month)

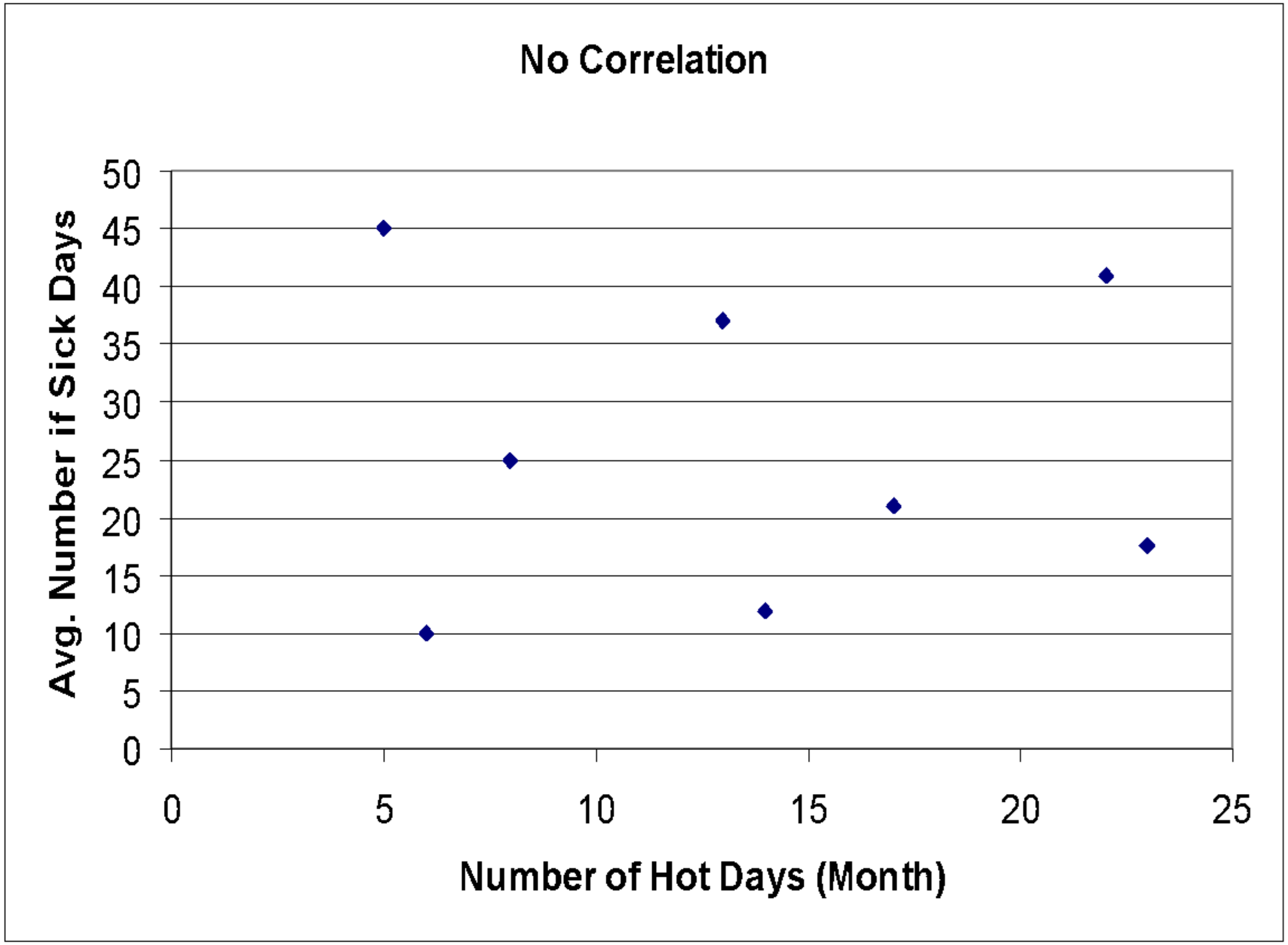
r = -1

r = -1 denotes perfect negative correlation between X & Y (all points on the line)



r = 0

r = 0 denotes no relationship between X & Y (points all over the place, no line is decipherable)



Interpretation

Information the Pearson's r gives us:

- **1. Strength of relationship**
- **2. Direction of relationship**

Interpretation

2. <u>Direction</u> of relationship:

a) Positive correlation: whenever value of r = positive.

<u>Interpretation</u>: the two variables move in the same direction.

- **- as X variable goes up, Y increases as well;**
- **- as X variable decreases, Y decreases as well.**

b) Negative correlation: whenever value of r = negative.

<u>Interpretation</u>: the two variables move in opposite direction:

- **- as X variable goes up, Y decreases;**
- **- as X variable decreases, Y increases.**

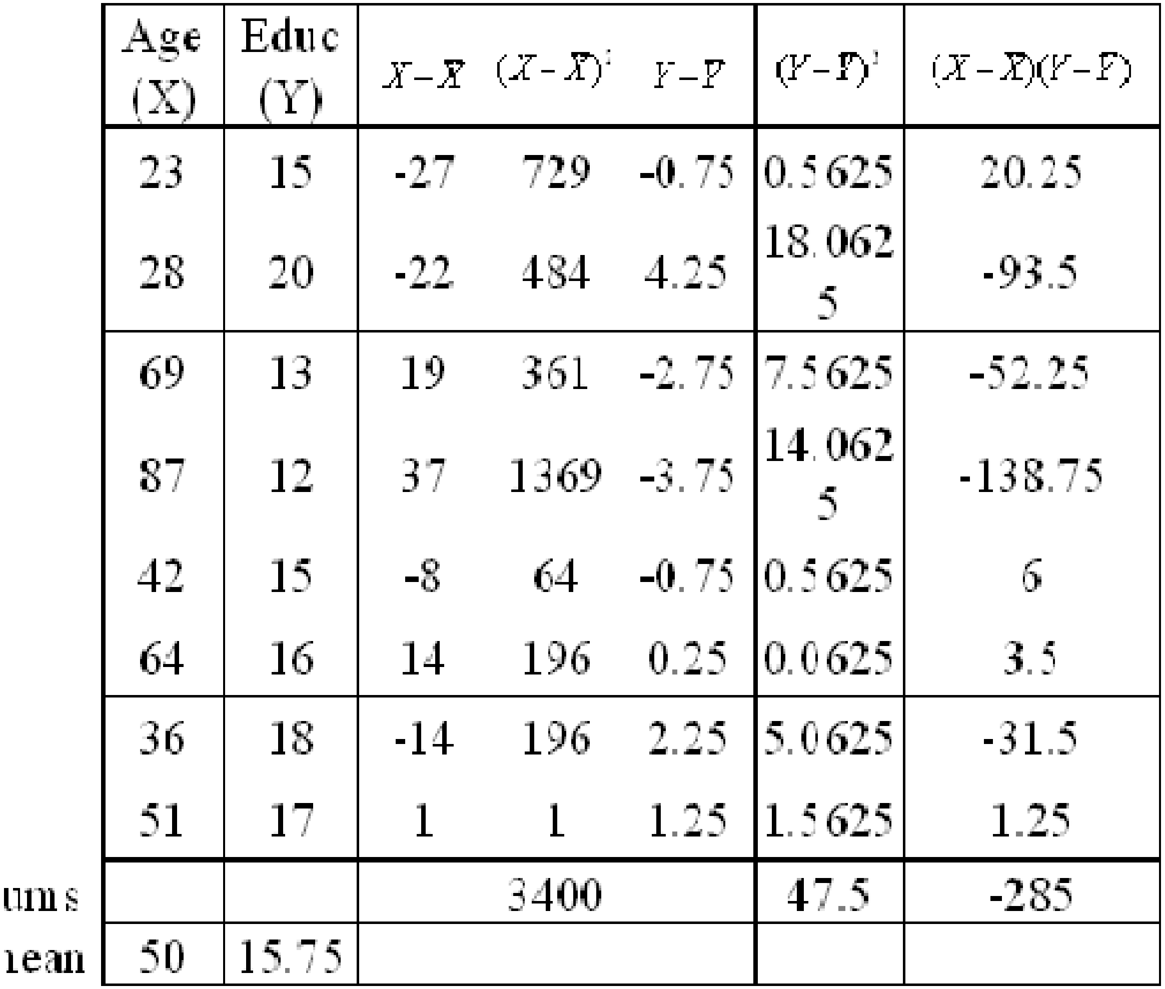
Interpretation

1. <u>Strength</u> of relationship

Correlations are on a <u>continuum</u>, thus, any value of r between –1 & 1 we need to describe:

- **weak □ 0.10**
- **moderate □ 0.30**
- **strong □ 0.60**

Computational table



	Age (X)	Educ (Y)	$X-\bar{X}$	$(X-\bar{X})^2$	$Y-\bar{Y}$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
	23	15	-27	729	-0.75	0.5625	20.25
	28	20	-22	484	4.25	18.0625	-93.5
	69	13	19	361	-2.75	7.5625	-52.25
	87	12	37	1369	-3.75	14.0625	-138.75
	42	15	-8	64	-0.75	0.5625	6
	64	16	14	196	0.25	0.0625	3.5
	36	18	-14	196	2.25	5.0625	-31.5
	51	17	1	1	1.25	1.5625	1.25
ums				3400		47.5	-285
ean	50	15.75					

Calculations and logic of Pearson’s correlation

r = the sum of the products of the deviations from each mean, <u>divided</u> by the square root of the product of the sum of squares for each variable.

$$r = \frac{\Sigma\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{\sqrt{\Sigma\left(X - \overline{X}\right)^2 \Sigma\left(Y - \overline{Y}\right)^2}}$$

Calculation

- r = – 0.709

$$r = \frac{\Sigma\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{\sqrt{\Sigma\left(X - \overline{X}\right)^2 \Sigma\left(Y - \overline{Y}\right)^2}}$$

$$r = \frac{-285}{\sqrt{161500}}$$

Test

Interpretation:

- **Strength: Strong correlation.**
- **Direction: negative: younger people have relatively higher education**

Is this correlation statistically significant?

- **We need Hypothesis Test:**
- **Test the Null Hypothesis that r = 0 (that there is no relationship/correlation between X and Y), using a two-tailed t-test.**

Testing

Step 1: Assumptions: Interval data, random samples, normal population distributions, linear relationship (& not many outliers).

Thus: We have to look at the scatterplot before going on with the test

What to search for in the scatter plot:

- **Is the relation linear (do points follow a straight line, or not)?**
- **Are there outliers?**

Test

Step 2: Hypotheses:

- **H0: $r = 0$ There is No relationship btw. age (X) & education (Y)**
- **H1: $r \neq 0$ There is a relationship btw. X & Y**

Test

Step 3: **two-tailed test t-test, but it’s set up a bit differently from the other t-tests**

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

$$df = N - 2$$

Test

Step 4: alpha = .05; get t critical (use the same table as before)

- **Critical values? (+/- 2.447)**

Test

<u>Step 5</u>: Get t calculated (we already have r = -.709)

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} = \frac{-.709\sqrt{8-2}}{\sqrt{1-(-.709)^2}} = \frac{-1.7367}{0.7052} = -2.463$$

Test

Step 6: Decision & Interpretation:

- **Reject the null hypothesis.**

- **Interpretation: we are 95% confident that the correlation between age & education exists in the population**