



# **Statistical tests for Quantitative variables (z-test, t-test & Correlation)**

---

**BY**

**Dr. Shaikh Shaffi Ahamed Ph.D.,**

**Professor**

**Dept. of Family & Community Medicine**

**College of Medicine,**


**King Saud University**



# Learning Objectives:

---

- (1) Able to understand the factors to apply for the choice of statistical tests in analyzing the data .
- (2) Able to apply appropriately Z-test, student's t-test & Correlation
- (3) Able to interpret the findings of the analysis using these three tests.



# Choosing the appropriate Statistical test

---

- Based on the three aspects of the data
  - Type of variables
  - Number of groups being compared &
  - Sample size

# Statistical Tests



---

## Z-test:

Study variable: Qualitative (Categorical)

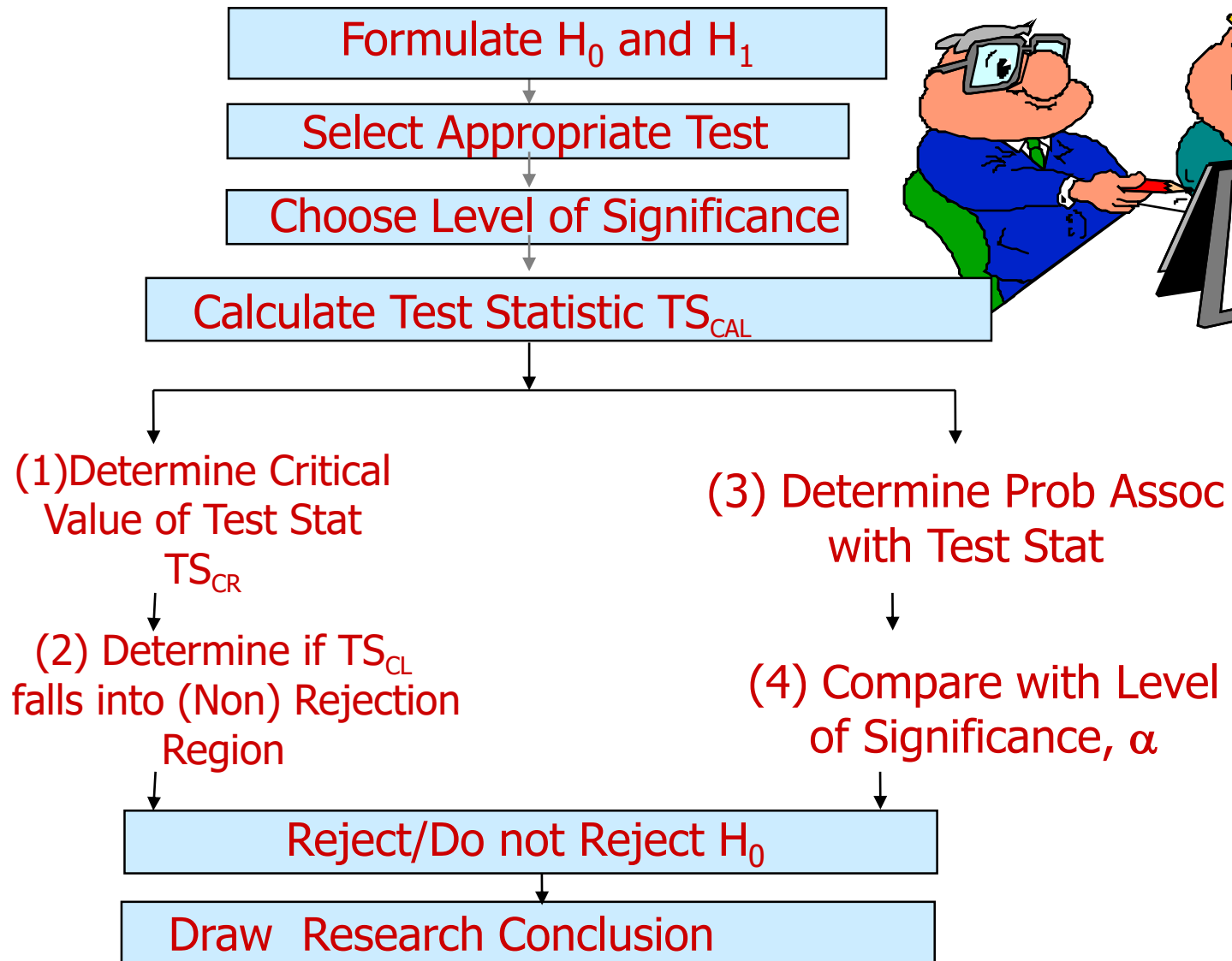
Outcome variable: Quantitative

Comparison:

- (i) sample mean with population mean
- (ii) two sample means

Sample size: larger in each group ( $>30$ ) & standard deviation is known

# Steps for Hypothesis Testing



# Example( Comparing Sample mean with Population mean):

- The education department at a university has been accused of “grade inflation” in medical students with higher GPAs than students in general.
- GPAs of all medical students should be compared with the GPAs of all other (non-medical) students.
  - There are 1000s of medical students, far too many to interview.
  - How can this be investigated without interviewing all medical students ?

# What we know:

- The average GPA for *all* students is 2.70. This value is a **parameter**.

$$\mu = 2.70$$

- To the right is the statistical information for a random sample of medical students:

$\bar{X} =$	3.00
$s =$	0.70
$n =$	117

# Questions to ask:

- Is there a difference between the parameter (2.70) and the statistic (3.00)?
- Could the observed difference have been caused by random chance?
- Is the difference real (significant)?



1. The sample mean (3.00) is the same as the pop. mean (2.70).
  - The difference is trivial and caused by random chance.
  
2. The difference is real (significant).
  - Medical students are different from all other students.

# Step 1: Make Assumptions and Meet Test Requirements

- Random sampling
  - Hypothesis testing assumes samples were selected using random sampling.
  - In this case, the sample of 117 cases was randomly selected from all medical students.
- Level of Measurement is Interval-Ratio
  - GPA is I-R so the mean is an appropriate statistic.
- Sampling Distribution is normal in shape
  - This is a “large” sample ( $n \geq 100$ ).

# Step 2 State the Null Hypothesis

- $H_0: \mu = 2.7$  (in other words,  $H_0: \bar{X} = \mu$ )
  - You can also state  $H_0$ : No difference between the sample mean and the population parameter
  - (In other words, the sample mean of 3.0 really the same as the population mean of 2.7 – the difference is not real but is due to chance.)
  - The sample of 117 comes from a population that has a GPA of 2.7.
  - The difference between 2.7 and 3.0 is trivial and caused by random chance.

# Step 2 (cont.) State the Alternat

## Hypothesis

- $H_1: \mu \neq 2.7$  (or,  $H_0: \bar{X} \neq \mu$ )
  - Or  $H_1$ : There is a difference between the sample mean and the population parameter
  - The sample of 117 comes a population that *does not* have a GPA of 2.7. In reality, it comes from a different population.
  - The difference between 2.7 and 3.0 reflects an actual difference between medical students and other students.
  - Note that we are testing whether the population the sample comes from is from a different population or is the same as the general student population.

## Step 3 Select Sampling Distribution and Establish the Critical Region

- Sampling Distribution= Z
  - Alpha ( $\alpha$ ) = .05
  - $\alpha$  is the indicator of “rare” events.
  - Any difference with a probability less than  $\alpha$  is rare and will cause us to reject the  $H_0$ .

# Step 3 (cont.) Select Sampling Distribution and Establish the Critical Region

- Critical Region begins at  $Z = \pm 1.96$ 
  - This is the critical Z score associated with  $\alpha = .05$ , two-tailed test.
  - If the obtained Z score falls in the Critical Region, or “the region of rejection,” then we would reject the  $H_0$ .

**When the Population  $\sigma$  is not known,  
use the following formula:**

$$Z = \frac{\bar{X} - \mu}{s / \sqrt{n - 1}}$$

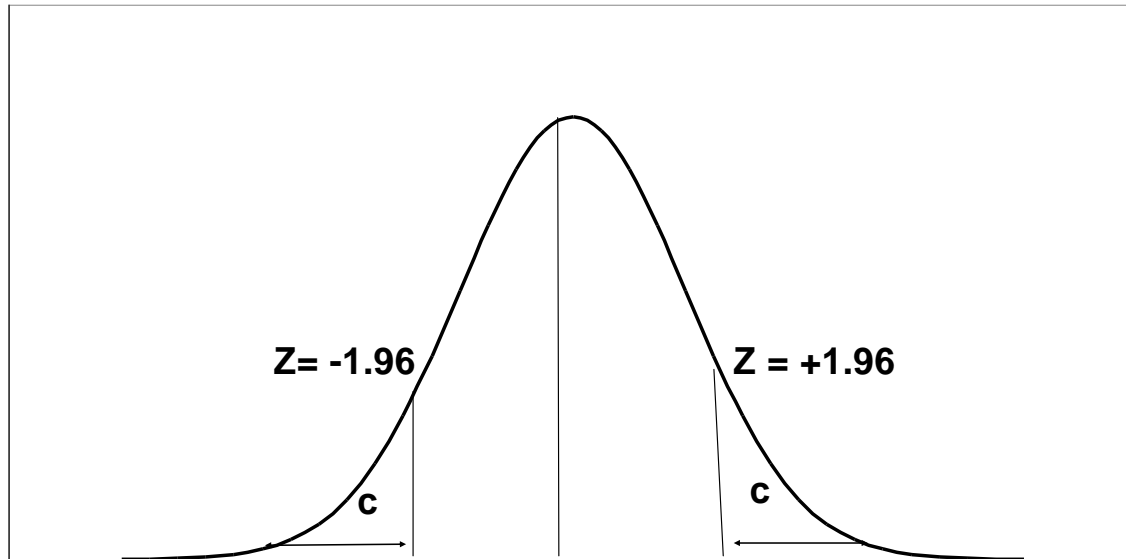
# Test the Hypotheses

$$Z = \frac{3.0 - 2.7}{\frac{.7}{\sqrt{117-1}}} = 4.62$$

- Substituting the values into the formula, we calculate a Z score of 4.62.



# Two-tailed Hypothesis Test



When  $\alpha = .05$ , then .025 of the area is distributed on either side of the curve in area **(C)**

The .95 in the **middle section** represents **no significant difference** between the population and the sample mean.

The cut-off between the middle section and +/- .025 is represented by a **Z-value of +/- 1.96**.

# Step 5 Make a Decision and Interpret Results

- The obtained Z score fell in the Critical Region, so we *reject* the  $H_0$ .
  - If the  $H_0$  were true, a sample outcome of 3.00 would be unlikely.
  - Therefore, the  $H_0$  is false and must be rejected.
- Medical students have a GPA that is significantly different from the non-medical students ( $Z = 4.62$ ,  $p < 0.05$ ).

# Summary:

- The GPA of medical students is *significantly* different from the GPA of non-medical students.
- In hypothesis testing, we try to identify statistically significant differences that did not occur by random chance.
- In this example, the difference between the parameter 2.70 and the statistic 3.00 was large and unlikely ( $p < .05$ ) to have occurred by random chance.

# Comparison of two sample means

## Example : Weight Loss for Diet vs Exercise

*Did dieters lose more fat than the exercisers?*

### Diet Only:

sample mean = 5.9 kg

sample standard deviation = 4.1 kg

sample size =  $n = 42$

standard error =  $SEM_1 = 4.1 / \sqrt{42} = 0.633$

### Exercise Only:

sample mean = 4.1 kg

sample standard deviation = 3.7 kg

sample size =  $n = 47$

standard error =  $SEM_2 = 3.7 / \sqrt{47} = 0.540$

measure of variability =  $\sqrt{[(0.633)^2 + (0.540)^2]} = 0.83$

## Example : Weight Loss for Diet vs Exercise

**Step 1. Determine the null and alternative hypotheses.**

**Null hypothesis:** No difference in average fat lost in population for two methods. Population mean difference is *zero*.

**Alternative hypothesis:** There is a difference in average fat lost in population for two methods. Population mean difference is not *zero*.

**Step 2. Sampling distribution: Normal distribution (z-test)**

**Step 3. Assumptions of test statistic ( sample size > 30 in each group)**

**Step 4. Collect and summarize data into a test statistic.**

The sample mean difference =  $5.9 - 4.1 = 1.8$  kg  
and the standard error of the difference is 0.83.

So the *test statistic*: 
$$z = \frac{1.8 - 0}{0.83} = 2.17$$

## Example : Weight Loss for Diet vs Exercise

### Step 5. Determine the $p$ -value.

*Recall the alternative hypothesis was two-sided.*

$p$ -value =  $2 \times$  [proportion of bell-shaped curve above 2.17]

Z-test table  $\Rightarrow$  proportion is about  $2 \times 0.015 = 0.03$ .

### Step 6. Make a decision.

The  $p$ -value of **0.03** is less than or equal to **0.05**, so ...

- If really no difference between dieting and exercise as fat loss methods, would see such an extreme result only 3% of the time, or 3 times out of 100.
- Prefer to believe truth does not lie with null hypothesis.  
We conclude that there is a *statistically significant difference between average fat loss for the two methods.*

# Student's t-test:

Study variable: Qualitative ( Categorical)

Outcome variable: Quantitative

Comparison:

- (i) sample mean with population mean
- (ii) two means (independent samples)
- (iii) paired samples

Sample size: each group  $< 30$  ( can be used even for large sample size)

# Student's t-test

## 1. Test for single mean

Whether the sample mean is equal to the predefined population mean ?

---

## 2. Test for difference in means

Whether the CD4 level of patients taking treatment A is equal to CD4 level of patients taking treatment B ?

## 3. Test for paired observation

Whether the treatment conferred any significant benefit ?



# Steps for test for single mean

## 1. Questioned to be answered

Is the Mean weight of the sample of 20 rats is 24 mg?

$N=20$ ,  $\bar{x}=21.0$  mg,  $sd=5.91$  ,  $\mu=24.0$  mg

## 2. Null Hypothesis

The mean weight of rats is 24 mg. That is, The sample mean is equal to population mean.

## 3. Test statistics

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \quad \text{--- } t_{(n-1)} \text{ df}$$

## 4. Comparison with theoretical value

if tab  $t_{(n-1)} < \text{cal } t_{(n-1)}$  reject  $H_0$ ,

if tab  $t_{(n-1)} > \text{cal } t_{(n-1)}$  accept  $H_0$ ,

## 5. Inference

# t -test for single mean

- Test statistics

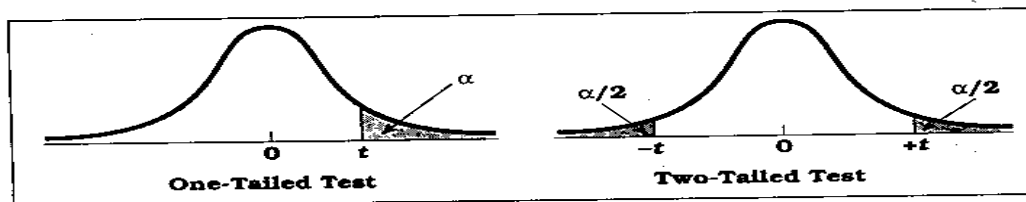
$n=20$ ,  $\bar{x}=21.0$  mg,  $sd=5.91$  ,  
 $\mu=24.0$  mg

$$t = \frac{|21.0 - 24|}{5.91/\sqrt{20}} = 2.30$$

$t_{\alpha} = t_{.05, 19} = 2.093$  Accept  $H_0$  if  $t < 2.093$

Reject  $H_0$  if  $t \geq 2.093$

**Inference : We reject  $H_0$ , and conclude that the data is not providing enough evidence, that the sample is taken from the population with mean weight of 24 gm**



**Table D.6 Percentage Points of the *t* Distribution** (Source: The entries in this table were computed by the author.)

df	Level of Significance for One-Tailed Test								
	.25	.20	.15	.10	.05	.025	.01	.005	.0005
	Level of Significance for Two-Tailed Test								
	.50	.40	.30	.20	.10	.05	.02	.01	.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	63.662
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	3.496
100	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.390
$\infty$	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

## t-test for difference in means

Given below are the 24 hrs total energy expenditure (MJ/day) in groups of lean and obese women. Examine whether the obese women's mean energy expenditure is significantly higher ?.

Lean			Obese		
6.1	7.0	7.5	8.8	9.2	9.2
7.5	5.5	7.6	9.7	9.7	10.0
7.9	8.1	8.1	11.5	11.8	12.8
8.1	8.4	10.2			
10.9					

## Null Hypothesis

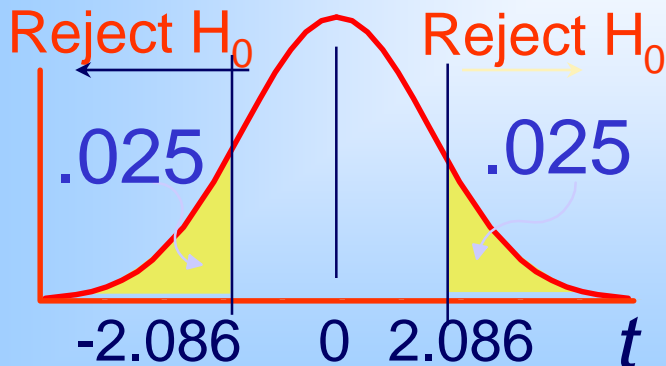
Obese women's mean energy expenditure is equal to the lean women's energy expenditure.

### Data Summary

	<b>lean</b>	<b>Obese</b>
<b>N</b>	<b>13</b>	<b>9</b>
	<b>8.10</b>	<b>10.30</b>
<b>S</b>	<b>1.38</b>	<b>1.25</b>

# Solution

- $H_0: \mu_1 - \mu_2 = 0$  ( $\mu_1 = \mu_2$ )
- $H_1: \mu_1 - \mu_2 \neq 0$  ( $\mu_1 \neq \mu_2$ )
- $\alpha = 0.05$
- $df = 13 + 9 - 2 = 20$
- **Critical Value(s):**



# Calculating the Test Statistic:

## •Compute the Test Statistic:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

**Hypothesized  
Difference**  
(usually zero  
when testing for  
equal means)

$$df = n_1 + n_2 - 2$$

$$S_p^2 = \frac{(n_1 - 1) \times S_1^2 + (n_2 - 1) \times S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

# Developing the Pooled-Variance t Test

- Calculate the Pooled Sample Variances as an Estimate of the Common Populations Variance:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

$S_p^2$  = Pooled-Variance

$n_1$  = Size of Sample 1

$S_1^2$  = Variance of Sample 1

$n_2$  = Size of Sample 2

$S_2^2$  = Variance of sample 2



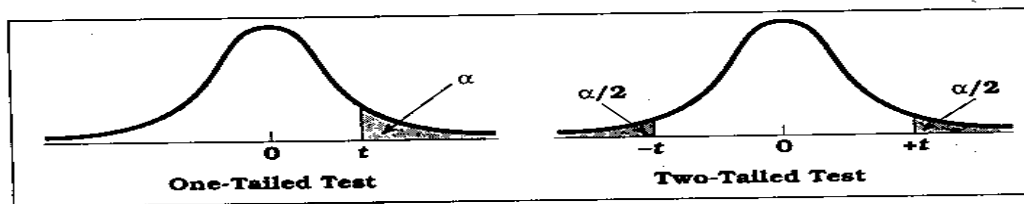
First, estimate the common variance as a weighted average of the two sample variances using the degrees of freedom as weights

$$S_P^2 = \frac{(n_1 - 1) \times S_1^2 + (n_2 - 1) \times S_2^2}{(n_1 - 1) + (n_2 - 1)}$$
$$= \frac{(13 - 1) \times 1.38^2 + (9 - 1) \times 1.25^2}{(13 - 1) + (9 - 1)} = 1.765$$

# Calculating the Test Statistic:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_P^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{|8.1 - 10.3| - 0}{1.76 \sqrt{\frac{1}{13} + \frac{1}{9}}} = 3.82$$

tab t  $9+13-2=20$  dff = t  $0.05,20=2.086$



**Table D.6 Percentage Points of the *t* Distribution** (Source: The entries in this table were computed by the author.)

df	Level of Significance for One-Tailed Test								
	.25	.20	.15	.10	.05	.025	.01	.005	.0005
	Level of Significance for Two-Tailed Test								
	.50	.40	.30	.20	.10	.05	.02	.01	.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	63.662
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	3.496
100	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.390
$\infty$	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

# T-test for difference in means

**Inference : The cal t (3.82) is higher than tab t at 0.05, 20. ie 2.086 . This implies that there is a evidence that the mean energy expenditure in obese group is significantly ( $p < 0.05$ ) higher than that of lean group**

# Example

**Suppose we want to test the effectiveness of a program designed to increase scores on the quantitative section of the Graduate Record Exam (GRE). We test the program on a group of 8 students. Prior to entering the program, each student takes a practice quantitative GRE; after completing the program, each student takes another practice exam. Based on their performance, was the program effective?**

- **Each subject contributes 2 scores: repeated measures design**

<b>Student</b>	<b>Before Program</b>	<b>After Program</b>
<b>1</b>	<b>520</b>	<b>555</b>
<b>2</b>	<b>490</b>	<b>510</b>
<b>3</b>	<b>600</b>	<b>585</b>
<b>4</b>	<b>620</b>	<b>645</b>
<b>5</b>	<b>580</b>	<b>630</b>
<b>6</b>	<b>560</b>	<b>550</b>
<b>7</b>	<b>610</b>	<b>645</b>
<b>8</b>	<b>480</b>	<b>520</b>

- **Can represent each student with a single score: the difference (D) between the scores**

<b>Student</b>	<b>Before Program</b>	<b>After Program</b>	<b>D</b>
<b>1</b>	<b>520</b>	<b>555</b>	<b>35</b>
<b>2</b>	<b>490</b>	<b>510</b>	<b>20</b>
<b>3</b>	<b>600</b>	<b>585</b>	<b>-15</b>
<b>4</b>	<b>620</b>	<b>645</b>	<b>25</b>
<b>5</b>	<b>580</b>	<b>630</b>	<b>50</b>
<b>6</b>	<b>560</b>	<b>550</b>	<b>-10</b>
<b>7</b>	<b>610</b>	<b>645</b>	<b>35</b>
<b>8</b>	<b>480</b>	<b>520</b>	<b>40</b>

- **Approach: test the effectiveness of program by testing significance of D**
- **Null hypothesis: There is no difference in the scores of before and after program**
- **Alternative hypothesis: program is effective → scores after program will be higher than scores before program → average D will be greater than zero**

$$\mathbf{H_0: \mu_D = 0}$$

$$\mathbf{H_1: \mu_D > 0}$$



**So, need to know  $\sum D$  and  $\sum D^2$ :**

<b>Student</b>	<b>Before Program</b>	<b>After Program</b>	<b>D</b>	<b>D<sup>2</sup></b>
<b>1</b>	<b>520</b>	<b>555</b>	<b>35</b>	<b>1225</b>
<b>2</b>	<b>490</b>	<b>510</b>	<b>20</b>	<b>400</b>
<b>3</b>	<b>600</b>	<b>585</b>	<b>-15</b>	<b>225</b>
<b>4</b>	<b>620</b>	<b>645</b>	<b>25</b>	<b>625</b>
<b>5</b>	<b>580</b>	<b>630</b>	<b>50</b>	<b>2500</b>
<b>6</b>	<b>560</b>	<b>550</b>	<b>-10</b>	<b>100</b>
<b>7</b>	<b>610</b>	<b>645</b>	<b>35</b>	<b>1225</b>
<b>8</b>	<b>480</b>	<b>520</b>	<b>40</b>	<b>1600</b>
			<b><math>\sum D = 180</math></b>	<b><math>\sum D^2 = 7900</math></b>

**Recall that for single samples:**

$$t_{obt} = \frac{\bar{X} - \mu}{s_{\bar{X}}} = \frac{\text{score} - \text{mean}}{\text{standard error}}$$

**For related samples:**

$$t_{obt} = \frac{\bar{D} - \mu_D}{s_{\bar{D}}}$$

where:

$$s_{\bar{D}} = \frac{s_D}{\sqrt{N}} \quad \text{and} \quad s_D = \sqrt{\frac{\sum D^2 - \frac{(\sum D)^2}{N}}{N - 1}}$$

## Mean of D:

$$\bar{D} = \frac{\sum D}{N} = \frac{180}{8} = 22.5$$

## Standard deviation of D:

$$s_D = \sqrt{\frac{\sum D^2 - \frac{(\sum D)^2}{N}}{N - 1}} = \sqrt{\frac{7900 - \frac{(180)^2}{8}}{8 - 1}} = 23.45$$

## Standard error:

$$s_{\bar{D}} = \frac{s_D}{\sqrt{N}} = \frac{23.45}{\sqrt{8}} = 8.2908$$

$$t_{obt} = \frac{\bar{D} - \mu_D}{s_{\bar{D}}}$$

Under  $H_0$ ,  $\mu_D = 0$ , so:

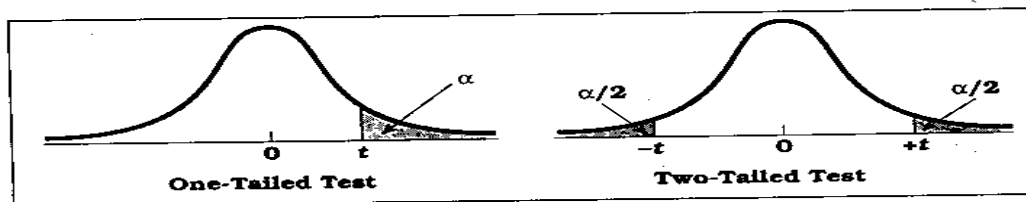
$$t_{obt} = \frac{\bar{D}}{s_{\bar{D}}} = \frac{22.5}{8.2908} = 2.714$$

From Table B.2: for  $\alpha = 0.05$ , one-tailed, with  $df = 7$ ,

$$t_{critical} = 1.895$$

$$2.714 > 1.895 \rightarrow \text{reject } H_0$$

The program is effective.



**Table D.6 Percentage Points of the *t* Distribution** (Source: The entries in this table were computed by the author.)

df	Level of Significance for One-Tailed Test								
	.25	.20	.15	.10	.05	.025	.01	.005	.0005
	Level of Significance for Two-Tailed Test								
	.50	.40	.30	.20	.10	.05	.02	.01	.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	63.662
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	3.496
100	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.390
$\infty$	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

# Z- value & t-Value

“Z and t” are the measures of:

How difficult is it to believe the null hypothesis?

**High z & t values**

Difficult to believe the null hypothesis -  
accept that there is a real difference.

**Low z & t values**

Easy to believe the null hypothesis -  
have not proved any difference.

# Karl Pearson Correlation Coefficient

# Working with two variables (parameter)

As Age	↑	BP	↑
As Height	↑	Weight	↑
As Age	↑	Cholesterol	↑
As duration of HIV	↑	CD4 CD8	↓



A number called the **correlation** measures both the direction and strength of the linear relationship between two related sets of quantitative variables.

# Correlation Contd....

- Types of correlation –
- Positive – Variables move in the same direction
  
- Examples:
- Height and Weight
- Age and BP

# Correlation contd...

- Negative Correlation
- Variables move in opposite direction
  
- Examples:
  - Duration of HIV/AIDS and CD4 CD8
  - Price and Demand
  - Sales and advertisement expenditure

# Correlation contd.....

- Measurement of correlation
  1. Scatter Diagram
  2. Karl Pearson's coefficient of Correlation

# Graphical Display of Relationship

- Scatter diagram
- Using the axes
  - X-axis horizontally
  - Y-axis vertically
  - Both axes meet: origin of graph: 0/0
  - Both axes can have different units of measurement
  - Numbers on graph are (x,y)

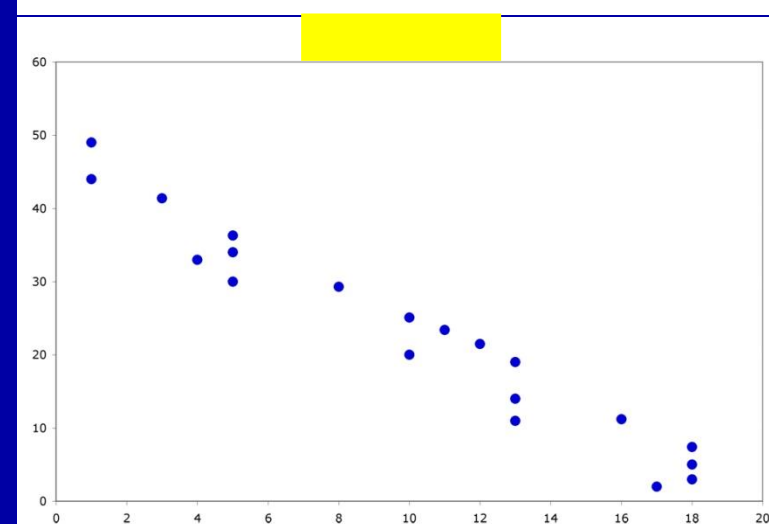
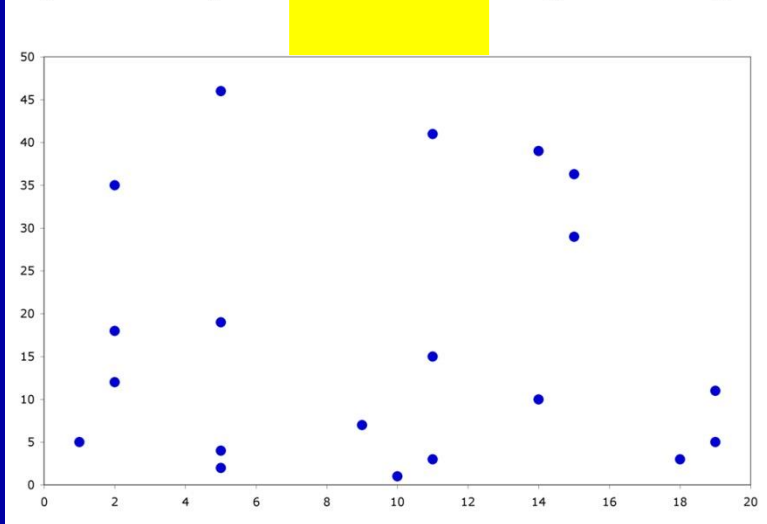
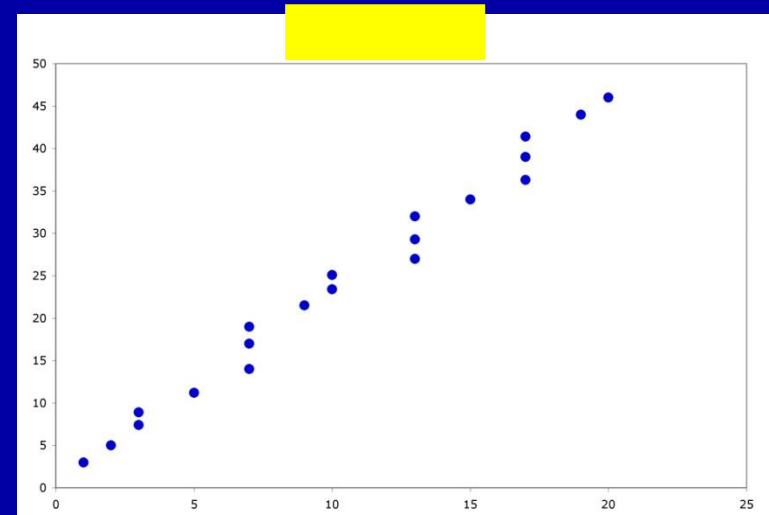
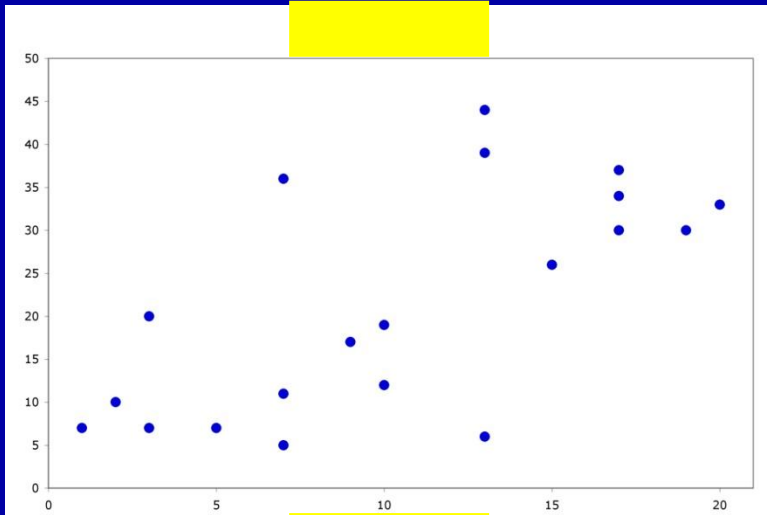
# Guess the Correlations:

.67

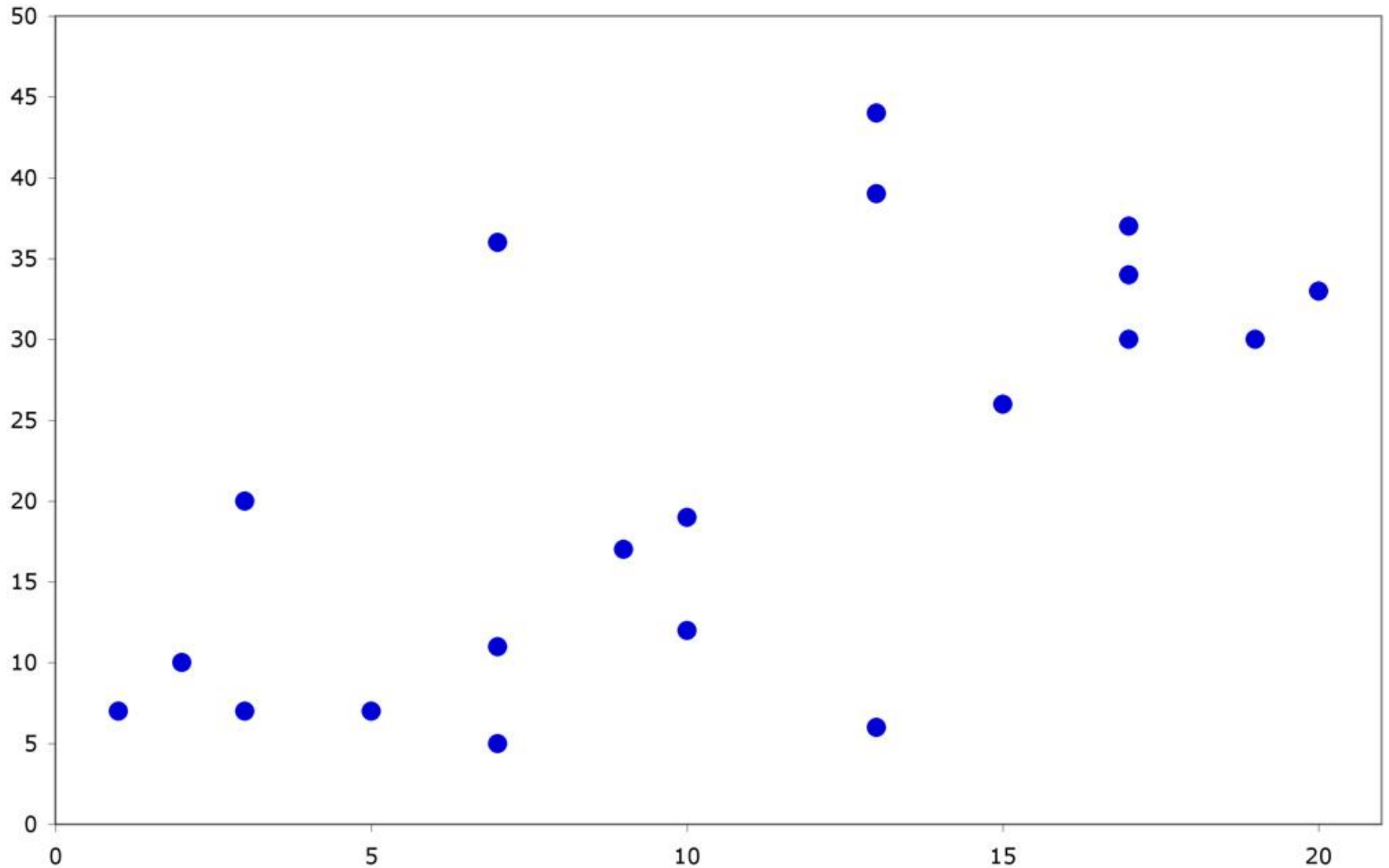
.993

.003

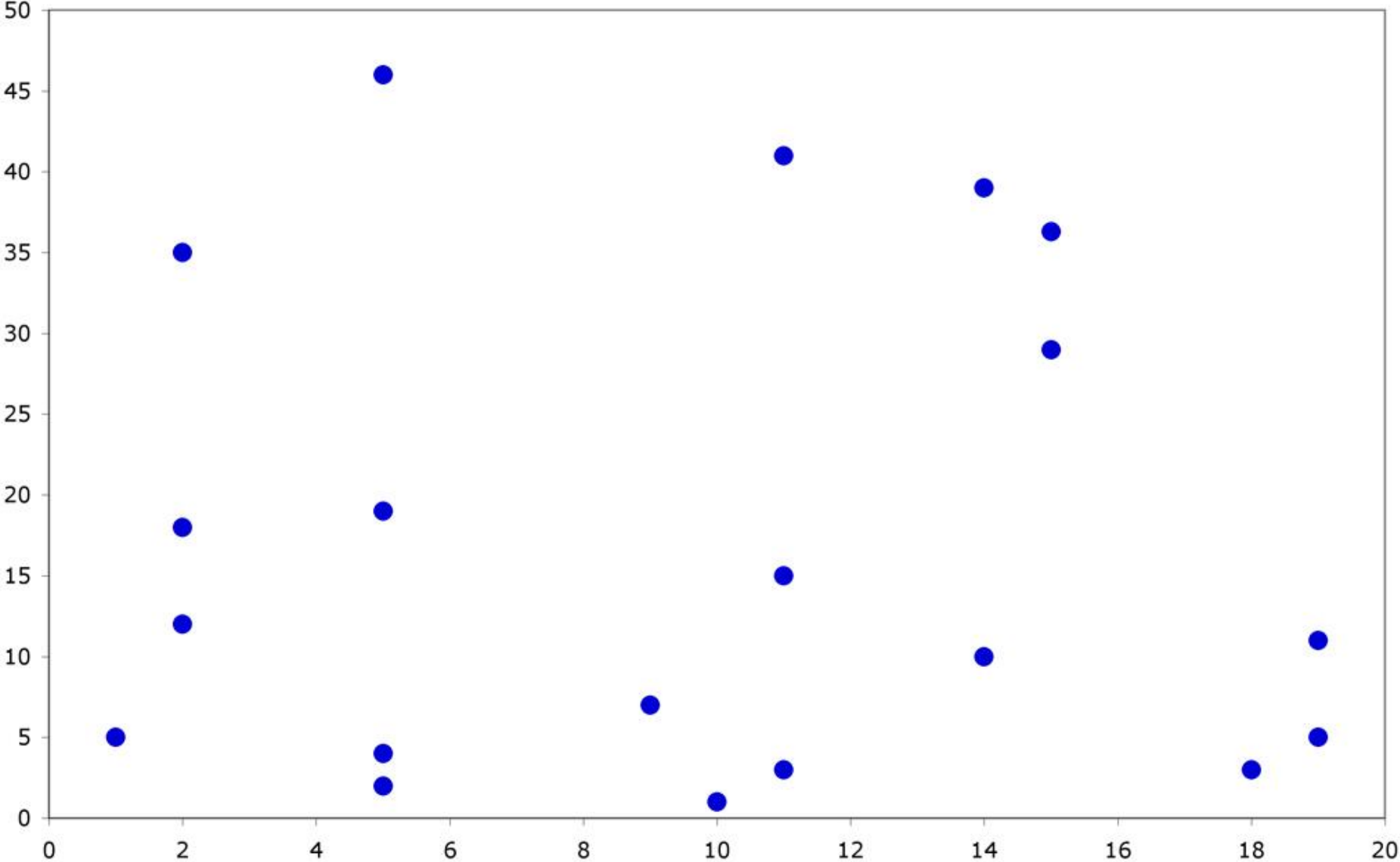
-.975



**Correlation = .67**

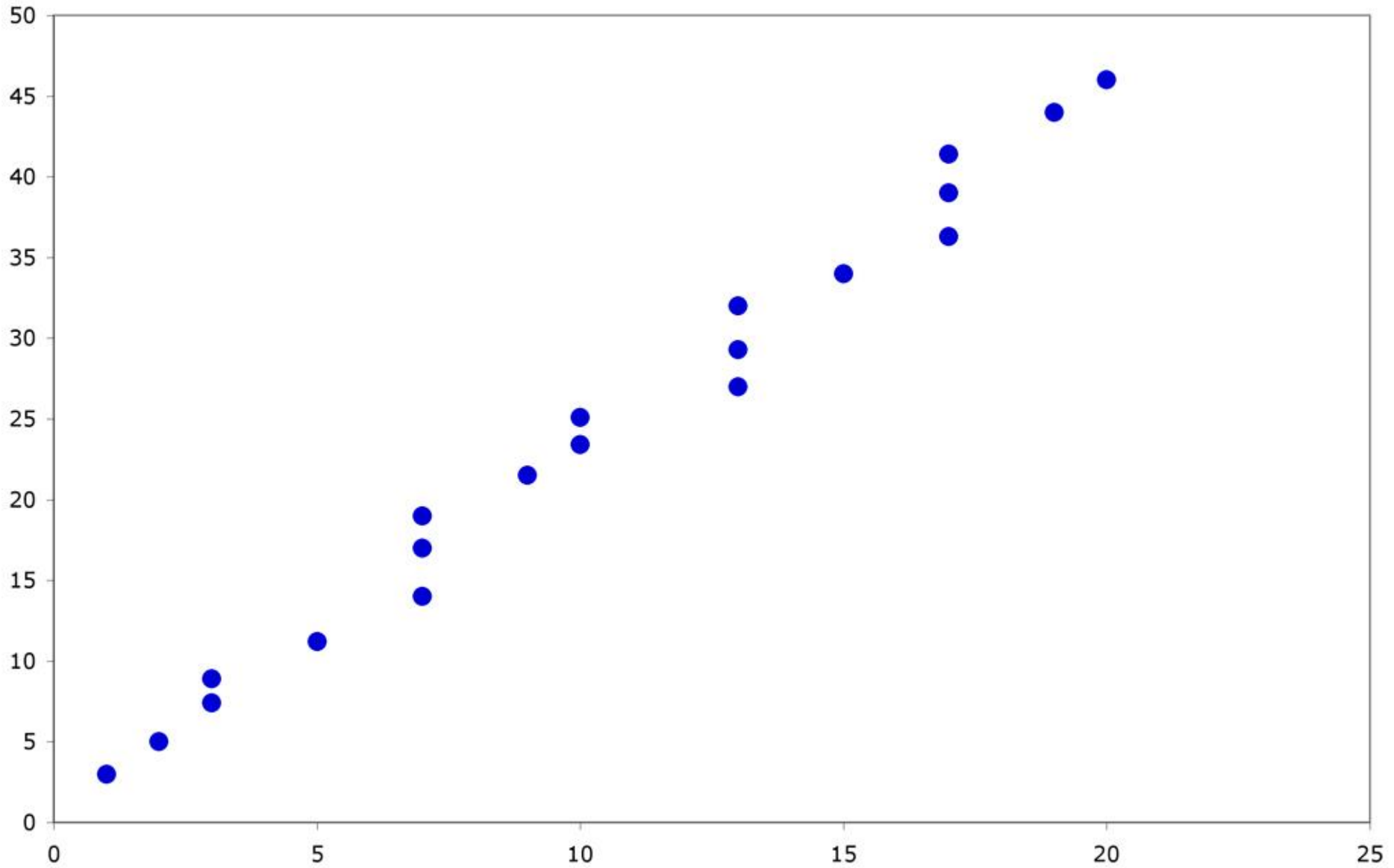


**Correlation = .003**

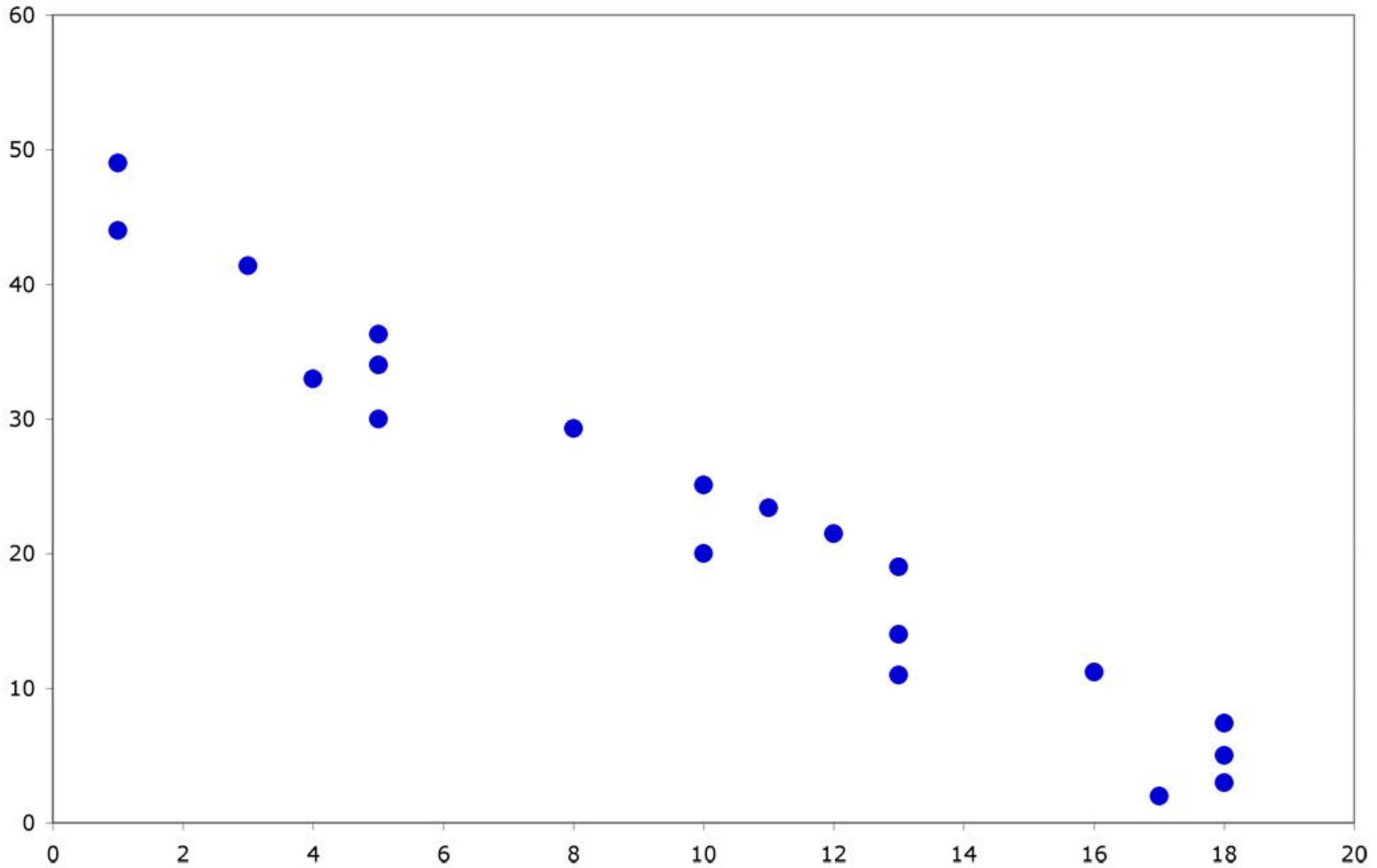




**Correlation = .993**



**Correlation = - .975**



# *The Pearson r*

$$r = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{\sqrt{\left[ \Sigma X^2 - \frac{(\Sigma X)^2}{N} \right] \left[ \Sigma Y^2 - \frac{(\Sigma Y)^2}{N} \right]}}$$

## *We Need:*

- *Sum of the Xs*      $\Sigma X$
- *Sum of the Ys*      $\Sigma Y$
- *Sum of the Xs squared*      $(\Sigma X)^2$
- *Sum of the Ys squared*      $(\Sigma Y)^2$
- *Sum of the squared Xs*      $\Sigma X^2$
- *Sum of the squared Ys*      $\Sigma Y^2$
- *Sum of Xs times the Ys*      $\Sigma XY$
- *Number of Subjects*      $(N)$

## Example:

A sample of 6 children was selected, data about their age in years and weight in kilograms was recorded as shown in the following table . Find the correlation between age and weight.

serial No	Age (years)	Weight (Kg)
1	7	12
2	6	8
3	8	12
4	5	10
5	6	11
6	9	13

<b>Serial n.</b>	<b>Age (years) (x)</b>	<b>Weight (Kg) (y)</b>	<b>xy</b>	<b>X<sup>2</sup></b>	<b>Y<sup>2</sup></b>
<b>1</b>	<b>7</b>	<b>12</b>	<b>84</b>	<b>49</b>	<b>144</b>
<b>2</b>	<b>6</b>	<b>8</b>	<b>48</b>	<b>36</b>	<b>64</b>
<b>3</b>	<b>8</b>	<b>12</b>	<b>96</b>	<b>64</b>	<b>144</b>
<b>4</b>	<b>5</b>	<b>10</b>	<b>50</b>	<b>25</b>	<b>100</b>
<b>5</b>	<b>6</b>	<b>11</b>	<b>66</b>	<b>36</b>	<b>121</b>
<b>6</b>	<b>9</b>	<b>13</b>	<b>117</b>	<b>81</b>	<b>169</b>
<b>Total</b>	<b><math>\sum x =</math> 41</b>	<b><math>\sum y =</math> 66</b>	<b><math>\sum xy =</math> 461</b>	<b><math>\sum x^2 =</math> 291</b>	<b><math>\sum y^2 =</math> 742</b>

$$r = \frac{461 - \frac{41 \times 66}{6}}{\sqrt{\left[291 - \frac{(41)^2}{6}\right] \cdot \left[742 - \frac{(66)^2}{6}\right]}}$$

$$r = 0.759$$

strong direct correlation

# EXAMPLE: Relationship between Anxiety and Test Scores

Anxiety (X)	Test score (Y)	X <sup>2</sup>	Y <sup>2</sup>	XY
10	2	100	4	20
8	3	64	9	24
2	9	4	81	18
1	7	1	49	7
5	6	25	36	30
6	5	36	25	30
$\sum X = 32$	$\sum Y = 32$	$\sum X^2 = 230$	$\sum Y^2 = 204$	$\sum XY = 129$



## Calculating Correlation Coefficient

$$r = \frac{(6)(129) - (32)(32)}{\sqrt{(6(230) - 32^2)(6(204) - 32^2)}} = \frac{774 - 1024}{\sqrt{(356)(200)}} = -.94$$

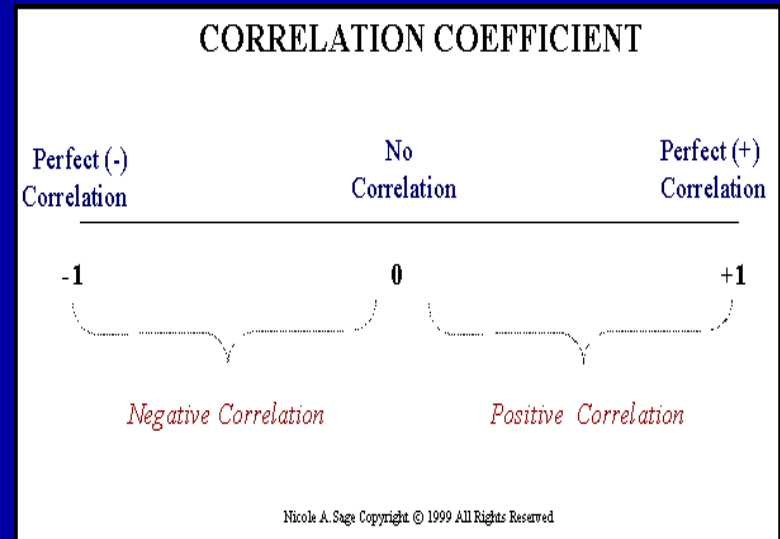
$$r = -0.94$$

**Indirect strong correlation**

# Correlation Coefficient

a correlation coefficient ( $r$ ) provides a quantitative way to express the degree of linear relationship between two variables.

- Range:  $r$  is always between -1 and 1
- Sign of correlation indicates direction:
  - high with high and low with low  $\rightarrow$  positive
  - high with low and low with high  $\rightarrow$  negative
  - no consistent pattern  $\rightarrow$  near zero
- Magnitude (absolute value) indicates strength (-.9 is just as strong as .9)
  - .10 to .40 weak
  - .40 to .80 moderate
  - .80 to .99 high
  - 1.00 perfect



# About “r”

- r is not dependent on the units in the problem
- r ignores the distinction between explanatory and response variables
- r is not designed to measure the strength of relationships that are not approximately straight line
- r can be strongly influenced by outliers

# Correlation Coefficient: Limitations

1. Correlation coefficient is appropriate measure of relation only when relationship is linear
2. Correlation coefficient is appropriate measure of relation when equal ranges of scores in the sample and in the population.
3. Correlation doesn't imply causality
  - Using U.S. cities as cases, there is a strong positive correlation between the number of churches and the incidence of violent crime
  - Does this mean churches cause violent crime, or violent crime causes more churches to be built?
  - More likely, both related to population of city (3d variable -- lurking or confounding variable)

***Ice-cream sales are strongly  
correlated with crime rates.***

***Therefore, ice-cream causes  
crime.***

Without proper interpretation,  
causation **should not** be  
assumed, or even implied.

# **In conclusion !**

Z-test will be used for both categorical(qualitative) and quantitative outcome variables.

Student's t-test will be used for only quantitative outcome variables.

Correlation will be used to quantify the linear relationship between two quantitative variables