# Statistical Inference for two related populations

Session 22

07/05/2021

# Statistical Inference for two related populations

- In the previous section, hypotheses were tested about the difference in two population means when the samples are independent.
- In this section, a method is presented to analyze dependent samples or related samples.
- This test is also known as matched pairs, t test for related measures etc.

# Some types of situations which the two samples being studied are related or dependent?

- Before and After study :An experimental control mechanism, the same person or object is measured both before and after a treatment. Certainly the after measurement is NOT independent of the before measurement because the measurements are taken on the same person or object in both cases.
- Eg: Rating of a company before and after 1 week of viewing a 15-minute DVD of the company.
- Other examples of related measures samples include studies in which twins, siblings, or spouses which are placed in two different groups.

# Hypotheses Testing

- The matched pairs test for related samples requires that the two samples be the same size and that the individual related scores be matched.
- The following formula is used to test hypotheses about dependent populations.

$$t = \frac{\bar{d} - D}{\frac{s_d}{\sqrt{n}}}$$

$$df = n - 1$$

Where, n= number of pairs

d = sample difference in pairs

D = mean population difference

$S_d$= standard deviation of sample difference

$\bar{d}$ : mean sample difference

# Formulas for $\bar{d}$ and s$_d$

- One can use the following formulas for calculating mean sample difference and standard deviation of sample difference.

$$\bar{d} = \frac{\sum d}{n}$$

$$S_d = \sqrt{\frac{\sum(d-\bar{d})^2}{n-1}}$$

or

$$S_d = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}}$$

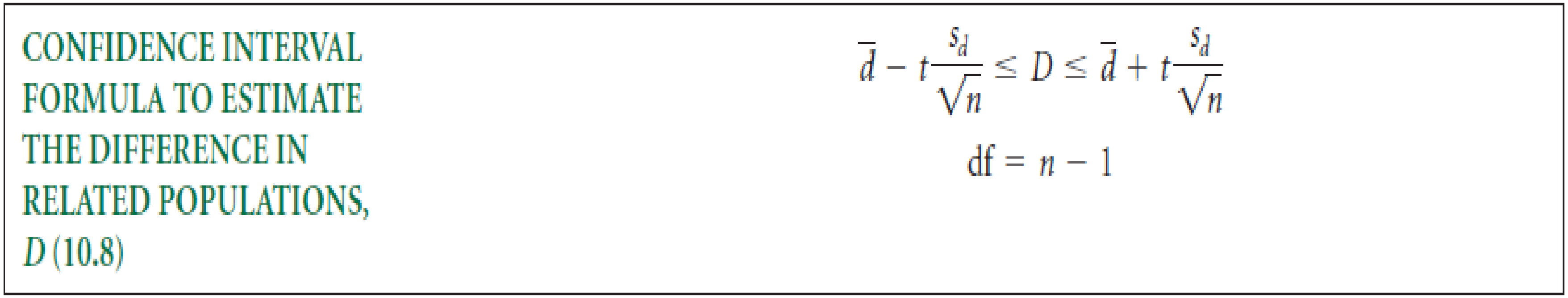
# Confidence Intervals

- Sometimes a researcher is interested in estimating the mean difference in two populations for related samples. A confidence interval for *D,* the mean population difference of two related samples, can be constructed by algebraically rearranging formula

$$t = \frac{\bar{d} - D}{\frac{s_d}{\sqrt{n}}}$$

which was used to test hypotheses about *D.* Again the assumption is that the differences are normally distributed in the population.

# Confidence Interval



CONFIDENCE INTERVAL FORMULA TO ESTIMATE THE DIFFERENCE IN RELATED POPULATIONS, *D* (10.8)

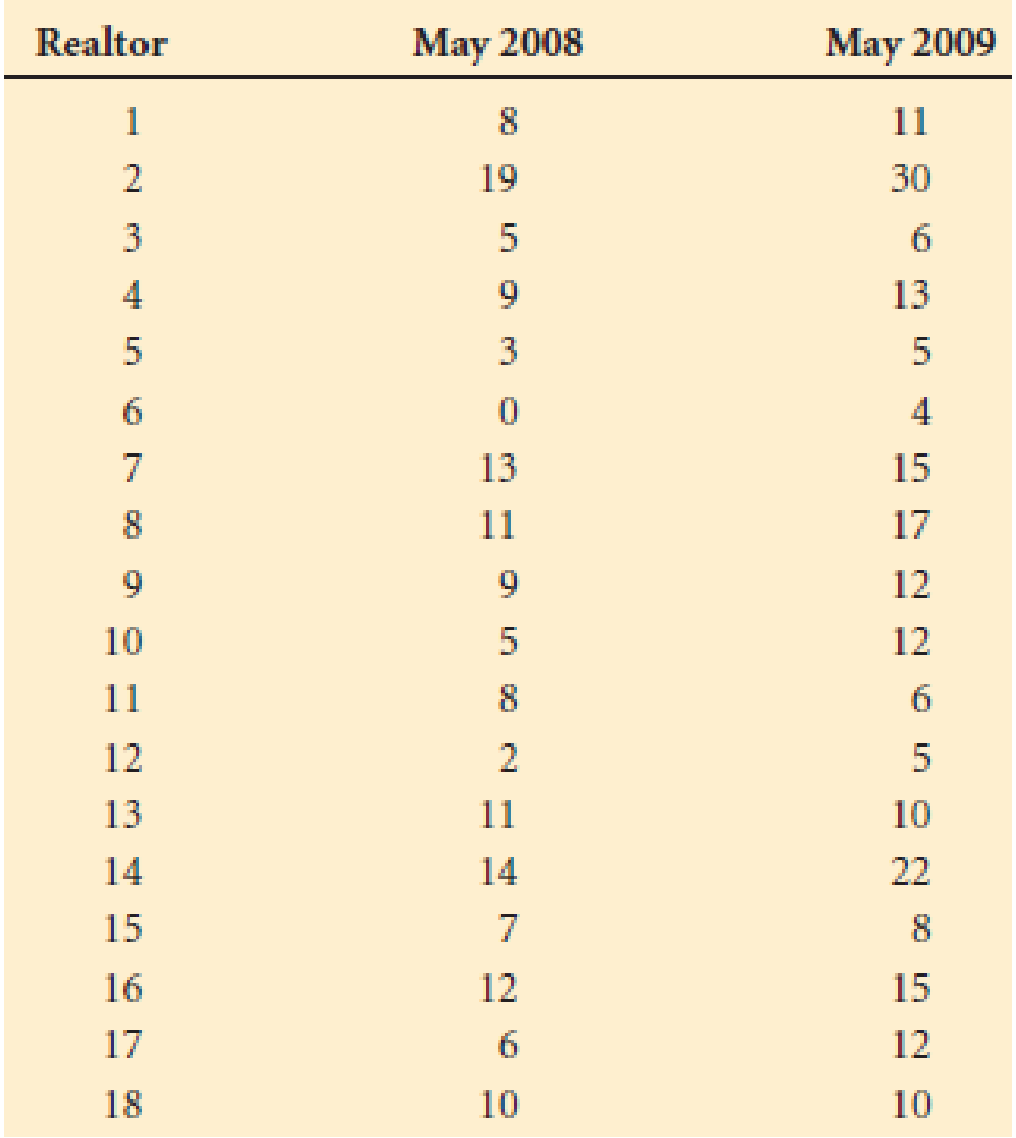
$$\bar{d} - t\frac{s_d}{\sqrt{n}} \le D \le \bar{d} + t\frac{s_d}{\sqrt{n}}$$

$$\text{df} = n - 1$$

# Example

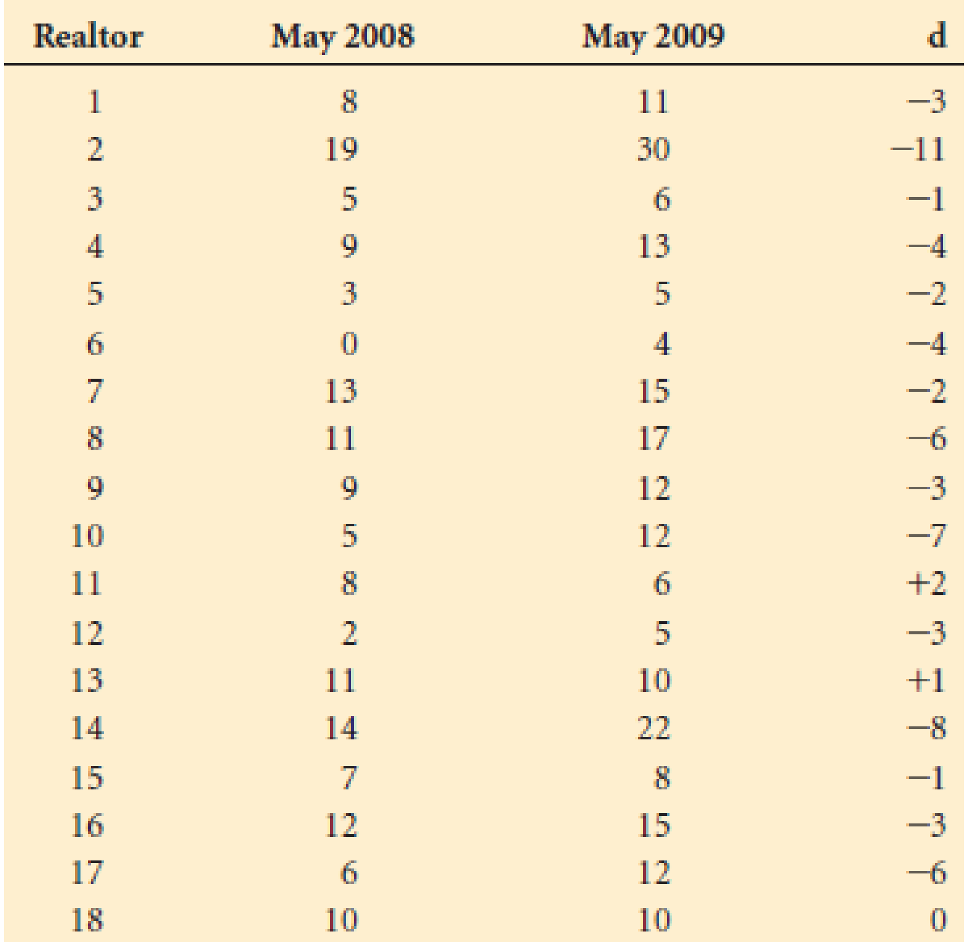
- The sale of new houses apparently fluctuates seasonally. Superimposed on the seasonality are economic and business cycles that also influence the sale of new houses. In certain parts of the country, new-house sales increase in the spring and early summer and drop off in the fall. Suppose a national real estate association wants to estimate the average difference in the number of new-house sales per company in a region between 2008 and 2009. To do so, the association randomly selects 18 real estate firms in the area and obtains their new-house sales figures for May 2008 and May 2009. The numbers of sales per company are shown. Using these data, the association's analyst estimates the average difference in the number of sales per real estate company in that region for May 2008 and May 2009. Construct a 99% confidence interval. The analyst assumes that differences in sales are normally distributed in the population.

# Data



| Realtor | May 2008 | May 2009 |
|---|---|---|
| 1 | 8 | 11 |
| 2 | 19 | 30 |
| 3 | 5 | 6 |
| 4 | 9 | 13 |
| 5 | 3 | 5 |
| 6 | 0 | 4 |
| 7 | 13 | 15 |
| 8 | 11 | 17 |
| 9 | 9 | 12 |
| 10 | 5 | 12 |
| 11 | 8 | 6 |
| 12 | 2 | 5 |
| 13 | 11 | 10 |
| 14 | 14 | 22 |
| 15 | 7 | 8 |
| 16 | 12 | 15 |
| 17 | 6 | 12 |
| 18 | 10 | 10 |

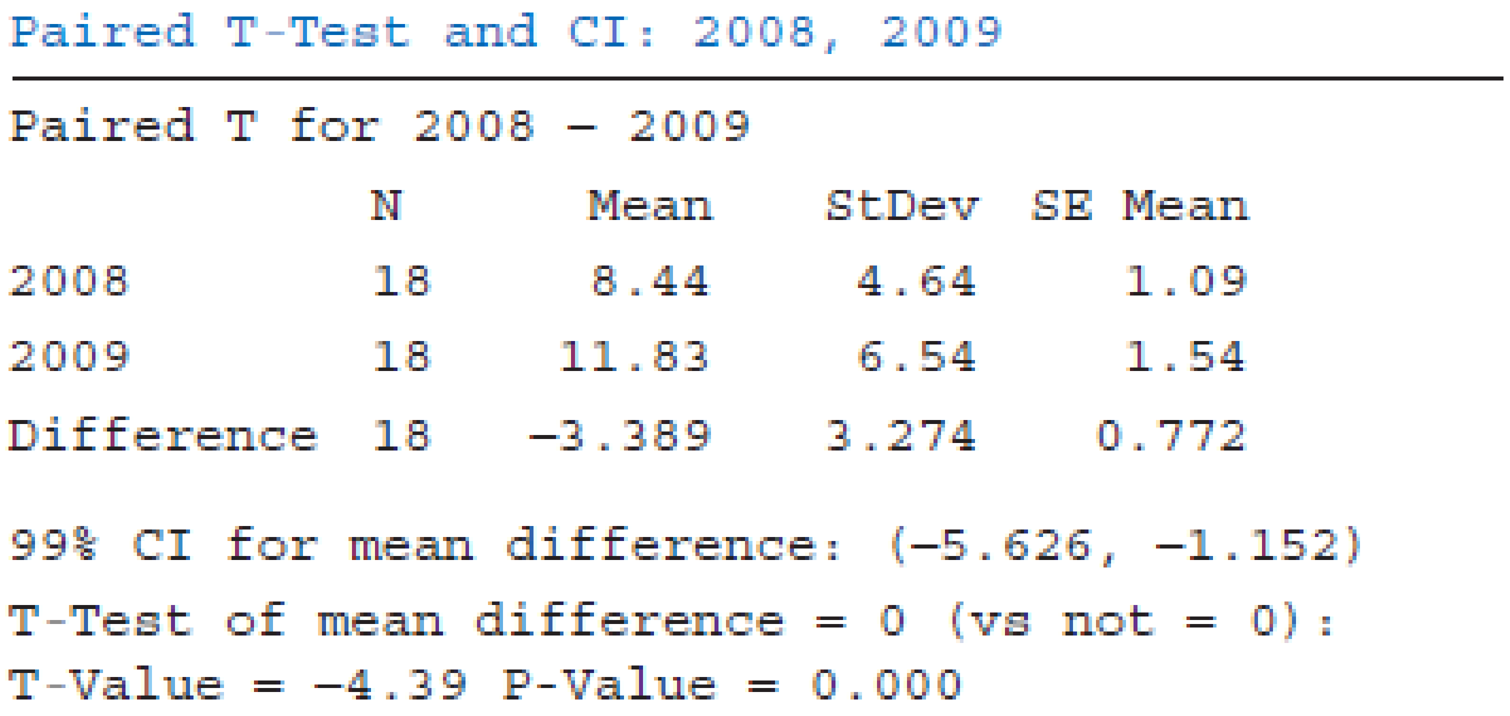
# Differences in Number of New House Sales, 2008–2009



| Realtor | May 2008 | May 2009 | d |
|---|---|---|---|
| 1 | 8 | 11 | −3 |
| 2 | 19 | 30 | −11 |
| 3 | 5 | 6 | −1 |
| 4 | 9 | 13 | −4 |
| 5 | 3 | 5 | −2 |
| 6 | 0 | 4 | −4 |
| 7 | 13 | 15 | −2 |
| 8 | 11 | 17 | −6 |
| 9 | 9 | 12 | −3 |
| 10 | 5 | 12 | −7 |
| 11 | 8 | 6 | +2 |
| 12 | 2 | 5 | −3 |
| 13 | 11 | 10 | +1 |
| 14 | 14 | 22 | −8 |
| 15 | 7 | 8 | −1 |
| 16 | 12 | 15 | −3 |
| 17 | 6 | 12 | −6 |
| 18 | 10 | 10 | 0 |

- $\bar{d} = -3.389 \; S_d = 3.274$

# Comments from the output



```
Paired T-Test and CI: 2008, 2009

Paired T for 2008 - 2009

             N      Mean    StDev  SE Mean
2008        18      8.44     4.64     1.09
2009        18     11.83     6.54     1.54
Difference  18    -3.389    3.274    0.772

99% CI for mean difference: (-5.626, -1.152)
T-Test of mean difference = 0 (vs not = 0):
T-Value = -4.39 P-Value = 0.000
```

# Test-7:Statistical inferences for related populations (Differences in the values are normally distributed, dependent samples )

| | Left tail Test | Right tail Test | Two tail Test |
|---|---|---|---|
| Hypotheses | $H_0$: D = 0<br>$H_1$: D < 0 | $H_0$: D = 0<br>$H_1$: D > 0 | $H_0$: D = 0<br>$H_1$: D ‡ 0 |
| Test Statistic | $t = \frac{\bar{d} - D}{\frac{s_d}{\sqrt{n}}}$ | $t = \frac{\bar{d} - D}{\frac{s_d}{\sqrt{n}}}$ | $t = \frac{\bar{d} - D}{\frac{s_d}{\sqrt{n}}}$ |
| Rejection Rule | Reject $H_0$ if<br>$t \leq -t_{\alpha,n-1}$ | Reject $H_0$ if<br>$t \geq t_{\alpha,n-1}$ | Reject $H_0$ if<br>$\lvert t \rvert \geq t_{\frac{\alpha}{2},n-1}$ |

- Example : Suppose a stock market investor is interested in determining whether there is a significant difference in the P/E (price to earnings) ratio for companies from one year to the next. In an effort to study this question, the investor randomly samples nine companies from the *Handbook of Common Stocks* and records the P/E ratios for each of these companies at the end of year 1 and at the end of year 2. The data are shown in Table. Assume that differences in P/E ratios are normally distributed in the population.

| Company | P/E ratio year 1 | P/E ratio year 2 |
|---|---|---|
| 1 | 8.9 | 12.7 |
| 2 | 38.1 | 45.4 |
| 3 | 43 | 10 |
| 4 | 34 | 27.2 |
| 5 | 34.5 | 22.8 |
| 6 | 15.2 | 24.1 |
| 7 | 20.3 | 32.3 |
| 8 | 19.9 | 40.1 |
| 9 | 61.9 | 106.5 |

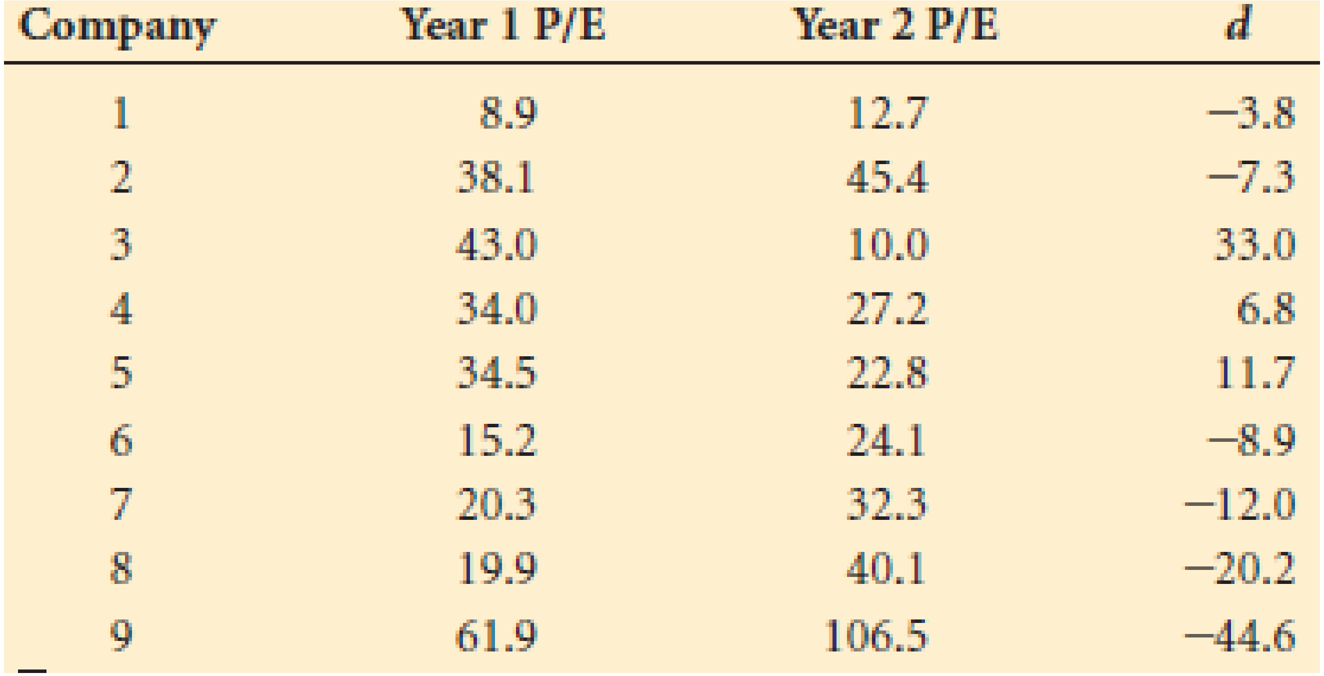
# Solution

- The null and alternative hypothesis are
- $H_0$: D=0 against $H_1$: D≠0, where D : difference in the P/E (price to earnings) ratio for companies from one year to the next in the population
- Assumptions :differences in P/E ratios are normally distributed in the population and the samples are dependent.
- Let α=0.01
- The test statistic is

$$t = \frac{\bar{d} - D}{\frac{s_d}{\sqrt{n}}}$$

with (n-1) d.f.

# Differences in in the P/E (price to earnings) ratio for companies in the sample

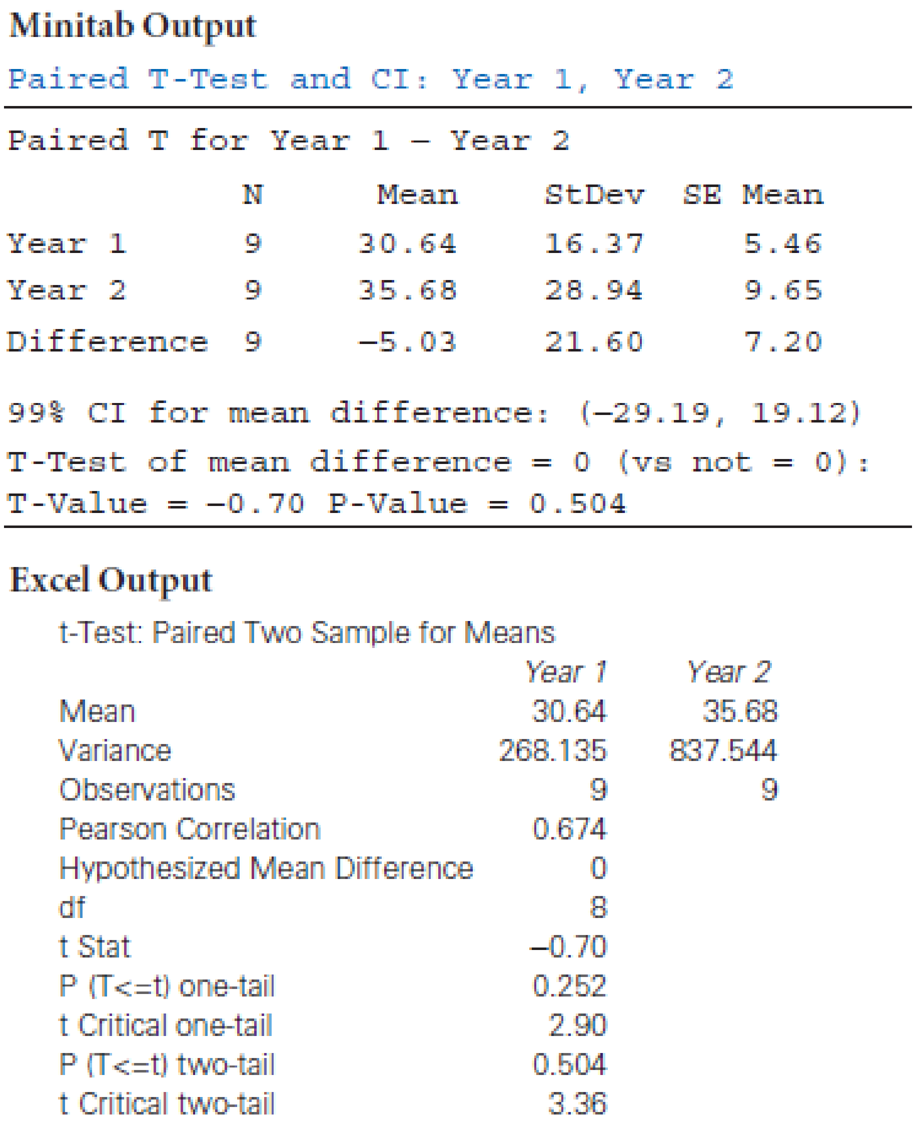


| Company | Year 1 P/E | Year 2 P/E | *d* |
|---|---|---|---|
| 1 | 8.9 | 12.7 | −3.8 |
| 2 | 38.1 | 45.4 | −7.3 |
| 3 | 43.0 | 10.0 | 33.0 |
| 4 | 34.0 | 27.2 | 6.8 |
| 5 | 34.5 | 22.8 | 11.7 |
| 6 | 15.2 | 24.1 | −8.9 |
| 7 | 20.3 | 32.3 | −12.0 |
| 8 | 19.9 | 40.1 | −20.2 |
| 9 | 61.9 | 106.5 | −44.6 |

- $\bar{d} = -5.033 \; S_d = 21.599$

- Calculated t = $-0.70$
- For α=0.01, $\frac{\alpha}{2} = 0.005$, $t_{0.005,8}$=3.355
- Rejection Rule is Reject $H_0$ if
- $|-0.70|$ < 3.355, we do not reject $H_0$. $|t| \geq t_{\frac{\alpha}{2},n-1}$
- There is not enough evidence from the data to declare a significant difference in the average P/E ratio between year 1 and year2.

# Output interpretation



**Minitab Output**

```
Paired T-Test and CI: Year 1, Year 2

Paired T for Year 1 – Year 2
             N      Mean     StDev  SE Mean
Year 1       9     30.64     16.37     5.46
Year 2       9     35.68     28.94     9.65
Difference   9     –5.03     21.60     7.20

99% CI for mean difference: (–29.19, 19.12)
T-Test of mean difference = 0 (vs not = 0):
T-Value = –0.70 P-Value = 0.504
```

**Excel Output**

```
t-Test: Paired Two Sample for Means
                                  Year 1    Year 2
Mean                               30.64     35.68
Variance                         268.135   837.544
Observations                           9         9
Pearson Correlation                0.674
Hypothesized Mean Difference           0
df                                     8
t Stat                             –0.70
P (T<=t) one-tail                  0.252
t Critical one-tail                 2.90
P (T<=t) two-tail                  0.504
t Critical two-tail                 3.36
```