# Tutorial letter 203/1/2017

## Applied Statistics II

# STA2601

## Semester 1

## Department of Statistics

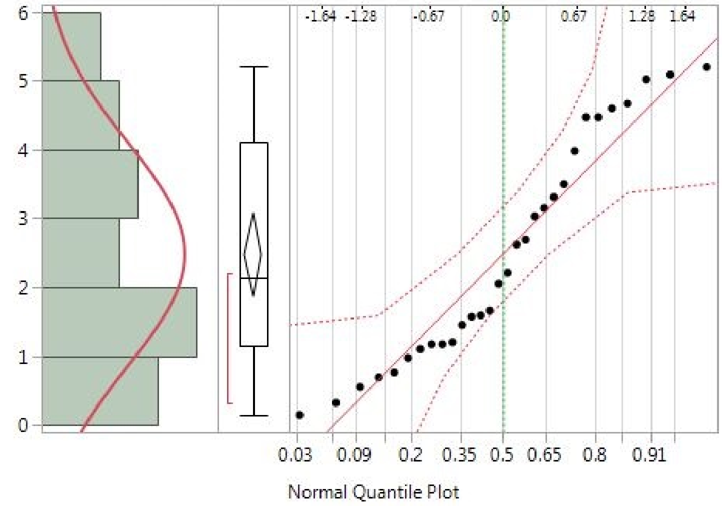
**Solutions to Assignment 03**

**QUESTION 1**

(a)

**Distributions**

**Percent of organic matter**




—— Normal(2.48,1.61418)

**Quantiles**

| | | |
|---|---|---|
| 100.0% | maximum | 5.2 |
| 99.5% | | 5.2 |
| 97.5% | | 5.2 |
| 90.0% | | 4.985 |
| 75.0% | quartile | 4.1025 |
| 50.0% | median | 2.13 |
| 25.0% | quartile | 1.1525 |
| 10.0% | | 0.564 |
| 2.5% | | 0.14 |
| 0.5% | | 0.14 |
| 0.0% | minimum | 0.14 |

**Summary Statistics**

| | |
|---|---|
| Mean | 2.48 |
| Std Dev | 1.614185 |
| Std Err Mean | 0.2947085 |
| Upper 95% Mean | 3.0827466 |
| Lower 95% Mean | 1.8772534 |
| N | 30 |

**Fitted Normal**

**Parameter Estimates**

| Type | Parameter | Estimate | Lower 95% | Upper 95% |
|---|---|---|---|---|
| Location | μ | 2.48 | 1.8772534 | 3.0827466 |
| Dispersion | σ | 1.614185 | 1.2855481 | 2.1699734 |

-2log(Likelihood) = 112.866121837

**Goodness-of-Fit Test**

Shapiro-Wilk W Test

| W | Prob<W |
|---|---|
| 0.920552 | 0.0277* |

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

(i) The assumptions are:

- observations are independent
- the data follows a normal distribution

Now based on the assumption of **independent observations** and the assumption that the percent of organic matter have a **normal distribution** (i.e., the sample comes from a normal population) we may assume that

$$T = \frac{\sqrt{n}\,(\bar{X} - \mu_0)}{S} \sim t_{n-1}.$$

**Are they met?** The soil specimens were drawn randomly, thus the assumption of **independent observations is met.**

The normality assumption is violated because from the JMP graphical output we see that the normal curve does not fit the histogram very well and there also seems to be a systematic deviation around the line in the Normal Quantile Plot. The histogram show that data is positively skewed and this is also supported by the boxplot with a longer tail to the right (Its subjective).

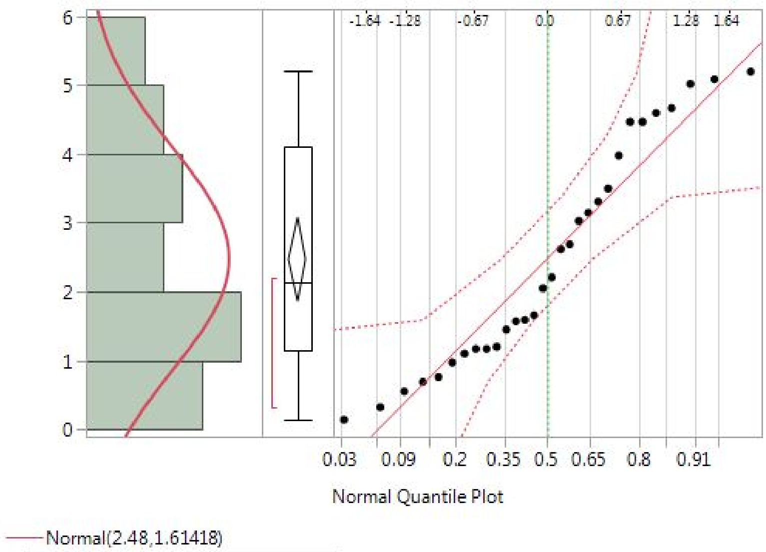
We need a proper test. The Shapiro-Wilk test for normality shows that the null hypothesis ($H_0$: Data comes from a normal distribution) would be rejected ($p$-value $= 0.0277$), indicating that we may assume the data does not come from a normal distribution. Luckily the test is not too sensitive and we may proceed.

(8)

(ii) We have to test $H_0 : \mu = 3$ against $H_1 : \mu \neq 3$.

**Distributions**

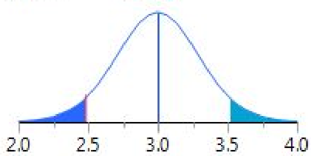
**Percent of organic matter**




**Test Mean**

| | |
|---|---|
| Hypothesized Value | 3 |
| Actual Estimate | 2.48 |
| DF | 29 |
| Std Dev | 1.61418 |

| | t Test |
|---|---|
| Test Statistic | -1.7645 |
| Prob > \|t\| | 0.0882 |
| Prob > t | 0.9559 |
| Prob < t | 0.0441* |



**Method 1: Using the critical value approach**

$$T = \frac{\sqrt{n}\left(\overline{X} - \mu_0\right)}{s} = \frac{\sqrt{30}\,(2.48 - 3)}{1.61418} \approx -1.7645$$

The critical value is $t_{\alpha/2;n-1} = t_{0.05;29} = 1.699$

We will reject $H_0$ if $T \geq 1.699$ or $T \leq -1.699$ or if $|T| \geq 1.699$.

Since $-1.7645 < -1.699$, we reject $H_0$ at the $10\%$ level of significance and conclude that the average percentage of organic matter is something other than $3\%$.

**Method II: Using the p-value approach**

$p$-value $= 0.0882$. Since $0.0882 < 0.10$, we reject $H_0$ at the $10\%$ level of significance and conclude that the average percentage of organic matter is something other than $3\%$.

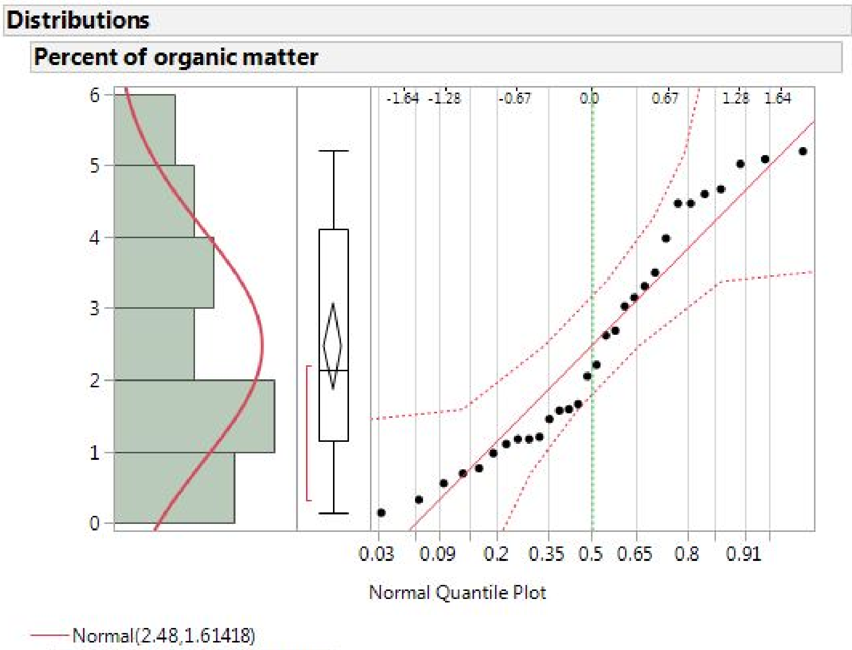
(10)

(iii) Yes. $p$-value $= 0.0882$. Since $0.0882 > 0.05$, we do not reject $H_0$ at the $5\%$ level of significance and conclude that the average percentage of organic matter is $3\%$.

(2)

(iv) We have to test $H_0 : \sigma = 1.5$

against $H_1 : \sigma \neq 1.5$




**Test Standard Deviation**

| | |
|---|---|
| Hypothesized Value | 1.5 |
| Actual Estimate | 1.61418 |
| DF | 29 |

| Test | ChiSquare |
|---|---|
| Test Statistic | 33.5832 |
| Min PValue | 0.5097 |
| Prob < ChiSq | 0.7452 |
| Prob > ChiSq | 0.2548 |

**Method 1: Using the critical value approach**

Assuming $\mu$ is unknown, i.e., $\widehat{\mu} = \overline{X}$, then the test statistic is

$$\begin{aligned} U &= \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{\sigma^2} \\ &= \frac{\left(\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}\right)}{1.5^2} \\ &= \frac{\left(260.0742 - \frac{(74.4)^2}{30}\right)}{2.25} \\ &= \frac{260.0742 - 184.512}{2.25} \\ &= \frac{75.5622}{2.25} \\ &= 33.5832 \end{aligned}$$

The critical values are

$$\begin{aligned} \chi^2_{1-\alpha/2;n-1} &= \chi^2_{0.95;29} & \chi^2_{\alpha/2;n-1} &= \chi^2_{0.05;29} \\ &= 17.7083 & &= 42.5569 \end{aligned}$$

Reject $H_0$ if $U < 17.7083$ or $U > 42.5569$

Since $17.7083 < 33.5832 < 42.5569$, we do not reject $H_0$ at the 10% level of significance and conclude that $\sigma = 1.5$.

**Method II: Using the p-value approach**

$p$-value $= 0.5097$. Since $0.5097 > 0.05$, we do not reject $H_0$ at the 10% level of significance and conclude that $\sigma = 1.5$.

(10)

(b) We have to test: $H_0 : \mu_X = \mu_Y$ against $H_1 : \mu_X > \mu_Y$

$n_X = 30 \quad \overline{X} = 2.48 \quad S_X^2 = 2.6056$

$n_Y = 20 \quad \overline{Y} = 2.43 \quad S_Y^2 = 2.4749$

The test statistic is

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$$

Now

$$\begin{aligned} S_p^2 &= \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2} \\ &= \frac{(30 - 1)2.6056 + (20 - 1)2.4749}{30 + 20 - 2} \\ &= \frac{75.5624 + 47.0231}{48} \\ &= \frac{122.5855}{48} \\ &= 2.553864583 \\ \Longrightarrow & \quad S_{pooled} = \sqrt{2.553864583} \approx 1.5981 \end{aligned}$$

Then

$$\begin{aligned} T &= \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \\ &= \frac{(2.48 - 2.43) - (0)}{1.5981\sqrt{\frac{1}{30} + \frac{1}{20}}} \\ &= \frac{0.05}{1.5981\sqrt{0.083333333}} \\ &= \frac{0.05}{0.461331732} \\ &\approx 0.1084 \end{aligned}$$

The critical value is

$$\begin{aligned} t_{\alpha;n_1-n_2-2} &= t_{0.05;48} \\ &= 1.684 + \frac{8}{20}(1.671 - 1.684) \\ &= 1.684 + \frac{2}{5}(-0.013) \\ &= 1.684 - 0.0052 \\ &\approx 1.6788 \end{aligned}$$

$\therefore$ Reject $H_0$ if $T \geq 1.6788$.

Since $0.1084 < 1.6788$, we do not reject $H_0$ at the $5\%$ level and conclude that the means are not significantly different from each other, i.e., $\mu_X = \mu_Y$. (10)

**[40]**

## QUESTION 2

| Group | Medication 1 | Medication 2 | Medication 3 |
|---|---|---|---|
| $n$ | 5 | 5 | 5 |
| $\sum X_{ij}$ | 35 | 18 | 34 |
| $\overline{X}_i$ | 7 | 3.6 | 6.8 |
| $\sum \left(X_{ij} - \overline{X}_i\right)^2$ | 16 | 23.2 | 2.8 |

(a)

$$\begin{aligned} S_1^2 &= \frac{1}{n_1 - 1} \sum \left(X_{1j} - \overline{X}_2\right)^2 \\ &= \frac{1}{5-1}(16) \\ &= \frac{1}{4}(16) \\ &= 4 \end{aligned} \qquad \begin{aligned} S_2^2 &= \frac{1}{n_2 - 1} \sum \left(X_{2j} - \overline{X}_2\right)^2 \\ &= \frac{1}{5-1}(23.2) \\ &= \frac{1}{4}(23.2) \\ &= 5.8 \end{aligned}$$

$$\begin{aligned} S_3^2 &= \frac{1}{n_3 - 1} \sum \left(X_{3j} - \overline{X}_3\right)^2 \\ &= \frac{1}{5-1}(2.8) \\ &= \frac{1}{4}(2.8) \\ &= 0.7 \end{aligned}$$

From the computations above it, follows that $S_1^2 = 4$; $S_2^2 = 5.8$ and $S_3^2 = 0.7$.

(9)

(b) (i) Ordinary average = $\dfrac{4 + 5.8 + 0.7}{3} = \dfrac{10.5}{3} = 3.5$ (2)

(ii) $MSE = \dfrac{SSE}{kn-k}$.

For this $ANOVA$ problem, we have $k = 3$ (there are four groups) and $n = 5$ (the number of observations in each sample).

$$\begin{aligned} SSE &= \sum_{i=1}^{k}\sum_{j=1}^{n}\left(X_{ij} - \overline{X}_i\right)^2 \\ &= 16 + 23.2 + 2.8 \\ &= 42 \end{aligned}$$

$$\begin{aligned} \therefore MSE &= \frac{42}{3\,(5) - 3} \\ &= \frac{42}{12} \\ &= 3.5. \qquad \text{The result in (i) = result in (ii).} \end{aligned}$$

This makes perfect sense! $MSE$ is like a pooled variance or an average variance, because the assumption of $ANOVA$ is that $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$ and if these variances are unknown, we estimate it by pooling.

(4)

(c) It is reasonable to assume that the three samples are **independent**. The people are different and were randomly selected and thus do not have influence on each other. (2)

(d) We have to test:

$H_0 : \mu_1 = \mu_2 = \mu_3$ against

$H_1 : \mu_p \neq \mu_q$ for at least one $p \neq q$.

The test statistic is $F = \dfrac{MST_r}{MSE} \sim F_{k-1;kn-k}$

$$MST_r = \frac{n\sum_{i=1}^{k}\left(\overline{X}_i - \overline{X}\right)^2}{k-1}$$

where $\overline{X} = \dfrac{\sum\sum X_{ij}}{N} = \dfrac{87}{15} = 5.8$ (overall mean);

$$\begin{aligned} \text{and } \sum \left(\overline{X}_i - \overline{X}\right)^2 &= (7-5.8)^2 + (3.6-5.8)^2 + (6.8-5.8)^2 \\ &= (1.2)^2 + (-2.2)^2 + (1)^2 \\ &= 1.44 + 4.84 + 1 \\ &= 7.28 \\ \therefore MST_r &= \frac{5\,(7.28)}{3-1} = \frac{36.4}{2} = 18.2 \end{aligned}$$

We already know that $MSE = 3.5$ (see question (b)(ii)).

$$\begin{aligned} \therefore F &= \frac{MST_r}{MSE} \\ &= \frac{18.2}{3.5} \\ &= 5.2. \end{aligned}$$

(Note that these computations are the same with the JMP output under the heading: "**Analysis of Variance**".)

The critical value is $F_{0,05;2;12} = 3.89$. Reject $H_0$ if $F > 3.89$.

Since $5.2 > 3.89$, we reject $H_0$ at the $5\%$ level of significance and conclude that the three medications produce different relief times that is, $\mu_p \neq \mu_q$ for at least one pair $p \neq q$.

(Note that we reach the same conclusion with the JMP output under the heading: "**Analysis of Variance**" if we consider "Prob $>$ F"$< 0.0236$)

(11)

(e) For each pair of means, we compute a test statistic

$$T_{pq} = \frac{\overline{X}_p - \overline{X}_q}{S_{pooled}\sqrt{1/n + 1/n}} = \frac{\sqrt{n}\left(\overline{X}_p - \overline{X}_q\right)}{\sqrt{2}S} = \frac{\sqrt{5}\left(\overline{X}_p - \overline{X}_q\right)}{\sqrt{2}\sqrt{MSE}}.$$

We reject $H_0(p;q)$ if

$$\left|T_{pq}\right| > \sqrt{(k-1)\,F_{\alpha;k-1;kn-k}} = \sqrt{2\,(3.89)} \approx 2.7893$$

This implies that we reject $H_0$ if

$$\frac{\sqrt{5}\left|\overline{X}_p - \overline{X}_q\right|}{\sqrt{2}\sqrt{3.5}} \geq 2.7893$$

$$\text{i.e. if } \left|\overline{X}_p - \overline{X}_q\right| \geq \frac{(2.7893)\sqrt{2}\sqrt{3.5}}{\sqrt{5}} = \frac{7.379794132}{2.236067977} = 3.3003$$

$$\left|\overline{X}_1 - \overline{X}_2\right| = |7.0 - 3.6| = 3.4 > 3.3003 \implies \mu_1 \neq \mu_2$$

$$\left|\overline{X}_1 - \overline{X}_3\right| = |7.0 - 6.8| \quad 0.2 < 3.3003 \implies \mu_1 = \mu_3$$

$$\left|\overline{X}_2 - \overline{X}_3\right| = |3.6 - 6.8| \quad 3.2 < 3.3003 \implies \mu_2 = \mu_3$$

All pairs of means are not significantly different from each other except the pairs $\overline{X}_1$ and $\overline{X}_2$;, that is, $\mu_1 \neq \mu_2$. (7)

**[35]**

## QUESTION 3

(a)

| Twin set | First born | Second born | $Y_i$ =First - Second |
|---|---|---|---|
| 1 | 32 | 44 | 12 |
| 2 | 36 | 43 | 7 |
| 3 | 21 | 28 | 7 |
| 4 | 30 | 39 | 9 |
| 5 | 49 | 51 | 2 |
| 6 | 27 | 25 | $-2$ |
| 7 | 39 | 32 | $-7$ |
| 8 | 38 | 42 | 4 |
| 9 | 56 | 64 | 8 |
| 10 | 44 | 44 | 0 |

$n = 10 \quad \sum Y_i = 40 \quad \sum \left(Y_i - \overline{Y}\right)^2 = 300$

We have to test:

$H_0 : \mu_d = 0$ against

$H_1 : \mu_d \neq 0$

$$\begin{aligned}
\overline{Y} &= \frac{1}{n}\sum Y_i & S_y^2 &= \frac{1}{n-1}\sum\left(Y_i-\overline{Y}\right)^2 \\
&= \frac{1}{10}(40) & &= \frac{1}{9}(300) \\
&= 4 & &= 33.33333333 \\
& & \Longrightarrow S_y &= \sqrt{33.33333333} \\
& & &\approx 5.7735
\end{aligned}$$

The test statistic is

$$\begin{aligned}
T &= \frac{\sqrt{n}\left(\overline{Y}-\mu\right)}{S_y} \\
&= \frac{\sqrt{10}\,(4-0)}{5.7735} \\
&= \frac{12.64911064}{5.7735} \\
&\approx 2.1909
\end{aligned}$$

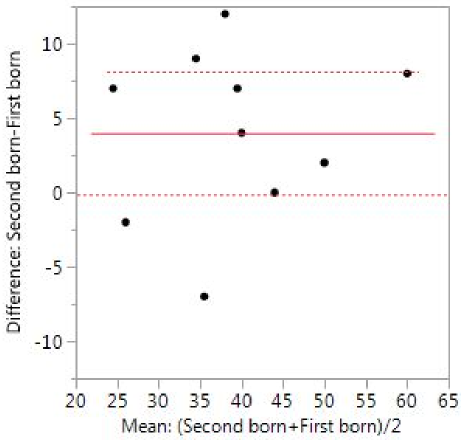
$t_{\alpha/2;(n-1)} = t_{0.025;9} = 2.262$. We will reject $H_0$ if $T \geq 2.262$ or $T \leq -2.262$ or if $|T| \geq 2.262$.

Since $-2.262 < 2.1909 < 2.262$, we do not reject $H_0$ at the $5\%$ level of significance and conclude that there is no difference in income between the twins. (13)

(b) The output is

**Matched Pairs**

**Difference: Second born-First born**




| | | | |
|---|---|---|---|
| Second born | 41.2 | t-Ratio | 2.19089 |
| First born | 37.2 | DF | 9 |
| Mean Difference | 4 | Prob > \|t\| | 0.0562 |
| Std Error | 1.82574 | Prob > t | 0.0281* |
| Upper 95% | 8.13012 | Prob < t | 0.9719 |
| Lower 95% | -0.1301 | | |
| N | 10 | | |
| Correlation | 0.86331 | | |

(5)

(c) Paired data since the pair of observations are twins. (2)

**[20]**

## QUESTION 4

(a) Start the *JMP* program

$>$ Enter *Company* in the first column and label it *Company*.

(make sure to change the scale to nominal)

$>$ Enter *Drying times* in the second column and label it *Drying times*.

This is a one-way ANOVA. To fit the model

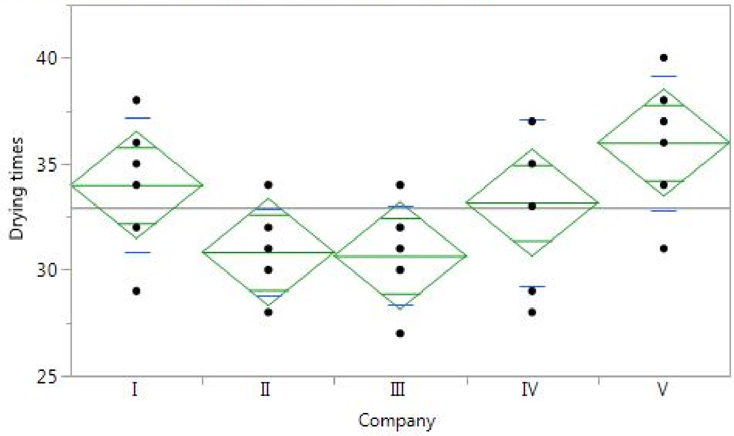
$>$ Choose Analyze$>$Fit $Y$ by $X$ with *Company* as $X$ factor and *Drying times* as $Y$ response.

$>$ Click Ok.

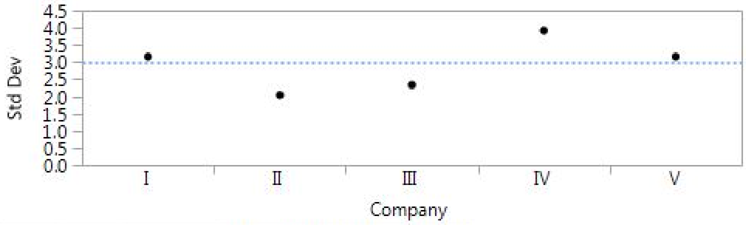
$\Longrightarrow$ Then on the Oneway Analysis of *Drying times* By *Company* click on the **Red** triangle

$>$ Choose Unequal Variances

**Oneway Analysis of Drying times By Company**




**Tests that the Variances are Equal**




| Level | Count | Std Dev | MeanAbsDif to Mean | MeanAbsDif to Median |
|---|---|---|---|---|
| I | 6 | 3.162278 | 2.333333 | 2.333333 |
| II | 6 | 2.041241 | 1.500000 | 1.500000 |
| III | 6 | 2.338090 | 1.666667 | 1.666667 |
| IV | 6 | 3.920034 | 3.166667 | 3.166667 |
| V | 6 | 3.162278 | 2.333333 | 2.333333 |

| Test | F Ratio | DFNum | DFDen | Prob > F |
|---|---|---|---|---|
| O'Brien[.5] | 0.9233 | 4 | 25 | 0.4661 |
| Brown-Forsythe | 0.8459 | 4 | 25 | 0.5095 |
| Levene | 0.9435 | 4 | 25 | 0.4553 |
| Bartlett | 0.6053 | 4 | . | 0.6588 |

**Welch's Test**

Welch Anova testing Means Equal, allowing Std Devs Not Equal

| F Ratio | DFNum | DFDen | Prob > F |
|---|---|---|---|
| 3.4937 | 4 | 12.332 | 0.0400* |

**For your own information**:

The standard deviation column shows the estimates you are testing. The $p$-values are listed under the column called $Prob > F$ and are testing the assumption that the variances are equal. Small $p$-values suggest that the variance are not equal.

**Interpretation**:

We have to test:

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$, against $H_1 : \sigma_p^2 \neq \sigma_q^2$ for at least one $p \neq q$

Using the Levene's test, $p$-value $= 0.4553$. Since $0.4553 > 0.05 \Longrightarrow$ we can not reject $H_0$ at the $5\%$ level of significance. The assumption of equal variances is not violated.

(10)

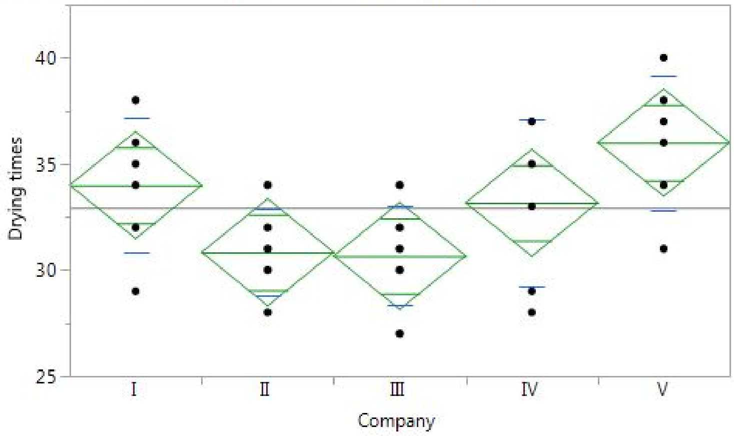
(b) $\Longrightarrow$ Click on the triangle "Tests that the variances are equal" to hide the output.

$\Longrightarrow$ Then click on the **Red** triangle on Oneway Analysis of *Drying times* by *Company*.

> Choose <u>Means/ANOVA</u>

$\Longrightarrow$ Click again on the **Red** triangle and choose <u>Means and Std dev</u>.

**Oneway Analysis of Drying times By Company**




**Oneway Anova**

**Summary of Fit**

| | |
|---|---|
| Rsquare | 0.34946 |
| Adj Rsquare | 0.245374 |
| Root Mean Square Error | 3 |
| Mean of Response | 32.93333 |
| Observations (or Sum Wgts) | 30 |

**Analysis of Variance**

| Source | DF | Sum of Squares | Mean Square | F Ratio | Prob > F |
|---|---|---|---|---|---|
| Company | 4 | 120.86667 | 30.2167 | 3.3574 | 0.0249* |
| Error | 25 | 225.00000 | 9.0000 | | |
| C. Total | 29 | 345.86667 | | | |

**Means for Oneway Anova**

| Level | Number | Mean | Std Error | Lower 95% | Upper 95% |
|---|---|---|---|---|---|
| I | 6 | 34.0000 | 1.2247 | 31.478 | 36.522 |
| II | 6 | 30.8333 | 1.2247 | 28.311 | 33.356 |
| III | 6 | 30.6667 | 1.2247 | 28.144 | 33.189 |
| IV | 6 | 33.1667 | 1.2247 | 30.644 | 35.689 |
| V | 6 | 36.0000 | 1.2247 | 33.478 | 38.522 |

Std Error uses a pooled estimate of error variance

**For your information:**

On the plot, the dots shows the response for each *Company*. The line across the middle is the grand mean. The diamonds give a $95\%$ confidence interval for each *Company* with the middle line of each diamond showing the group mean. If the groups are significantly different, then the diamonds do not overlap.

**Interpretation**:

(i) $H_0 : \mu_1 = \mu_2 = \mu_3$ against
$H_1 : \mu_p \neq \mu_q$ for at least one $p \neq q$.

(ii) The test statistic is $F = \frac{MSTr}{MSE} \sim F_{k-1;n-k}$

(iii) From the output: Computations for ANOVA we see that $F = 3.3574$ which is significant with a $p$–value of $0.0249$. Since $0.0249 < 0.05$ we reject $H_0$ in favour of $H_1$ at the $5\%$ level of significance and conclude that $u_p \neq \mu_q$ for at least one pair $p \neq q$, that is, the mean drying times of the companies are not the same.

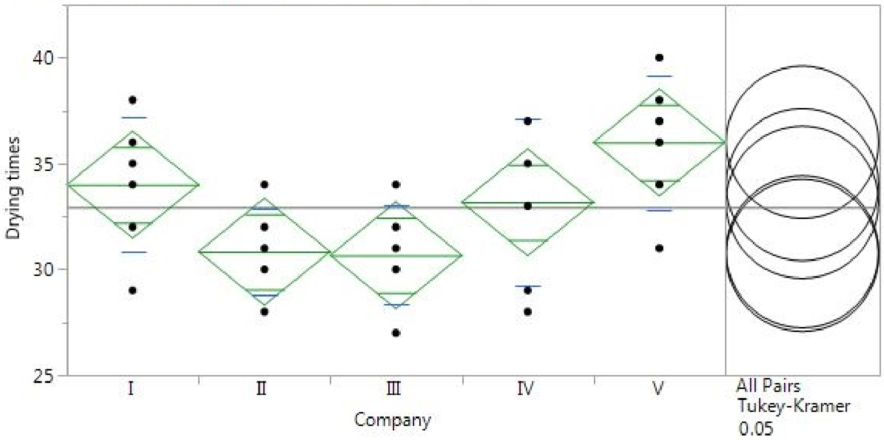
(10)

(c) $\Longrightarrow$ Hide the output "Oneway ANOVA" and "Means and Std deviations" by clicking the triangles.

$\Longrightarrow$ Click on the **Red** triangle on Oneway Analysis of *Drying times* by *Company*.

$\Longrightarrow$ Choose Compare Means $>$ All Pairs, Tukey HSD.

**Oneway Analysis of Drying times By Company**




**Means Comparisons**

**Comparisons for all pairs using Tukey-Kramer HSD**

**Confidence Quantile**

| q* | Alpha |
|---|---|
| 2.93687 | 0.05 |

**HSD Threshold Matrix**

Abs(Dif)-HSD

| | V | I | IV | II | III |
|---|---|---|---|---|---|
| V | -5.0868 | -3.0868 | -2.2535 | 0.0799 | 0.2465 |
| I | -3.0868 | -5.0868 | -4.2535 | -1.9201 | -1.7535 |
| IV | -2.2535 | -4.2535 | -5.0868 | -2.7535 | -2.5868 |
| II | 0.0799 | -1.9201 | -2.7535 | -5.0868 | -4.9201 |
| III | 0.2465 | -1.7535 | -2.5868 | -4.9201 | -5.0868 |

Positive values show pairs of means that are significantly different.

**Connecting Letters Report**

| Level | | | Mean |
|---|---|---|---|
| V | A | | 36.000000 |
| I | A | B | 34.000000 |
| IV | A | B | 33.166667 |
| II | | B | 30.833333 |
| III | | B | 30.666667 |

Levels not connected by same letter are significantly different.

**Ordered Differences Report**

| Level | - Level | Difference | Std Err Dif | Lower CL | Upper CL | p-Value |
|---|---|---|---|---|---|---|
| V | III | 5.333333 | 1.732051 | 0.24652 | 10.42015 | 0.0366* |
| V | II | 5.166667 | 1.732051 | 0.07985 | 10.25348 | 0.0452* |
| I | III | 3.333333 | 1.732051 | -1.75348 | 8.42015 | 0.3310 |
| I | II | 3.166667 | 1.732051 | -1.92015 | 8.25348 | 0.3805 |
| V | IV | 2.833333 | 1.732051 | -2.25348 | 7.92015 | 0.4896 |
| IV | III | 2.500000 | 1.732051 | -2.58681 | 7.58681 | 0.6067 |
| IV | II | 2.333333 | 1.732051 | -2.75348 | 7.42015 | 0.6654 |
| V | I | 2.000000 | 1.732051 | -3.08681 | 7.08681 | 0.7762 |
| I | IV | 0.833333 | 1.732051 | -4.25348 | 5.92015 | 0.9884 |
| II | III | 0.166667 | 1.732051 | -4.92015 | 5.25348 | 1.0000 |

Manually, we should have computed for each pair of means, a test statistic

$$T_{pq} = \frac{\overline{X}_p - \overline{X}_q}{S_{\text{pooled}}\sqrt{\frac{1}{n} + \frac{1}{n}}}$$

where we have samples of equal sizes if we want to incorporate the principle of the Bonferroni equality.

The Turkey–Kramer HSD that are shown in the JMP out perform individual comparisons that make adjustments for multiple test.

Confirming this is the **Abs(Dif)-LSDs** which are $0.0799$ and $0.2465$ respectively. Since they are positive, the means are significantly different. (Recall a negative value of **Abs(Dif)-LSD** means the groups are not significantly different from each other.)

Companies that share the same letter are not significantly different from each other. Companies I, IV and V share the same letter A whilst companies I, II, III and IV share the same letter B.

Confidence intervals that do not include zero imply that the pairs of means differ significantly. All pairs include zero except the pair $III-V$ and $II-V$. The confidence interval for the pairs are $(0.2465 : 10.4202)$ and $(0.0799 : 10.2535)$. These are the only intervals that do not include zero and it means we reject the null hypothesis of equal means and conclude that $\mu_3 \neq \mu_5$ and $\mu_2 \neq \mu_5$. The $p-$values are $0.0366$ and $0.0452$ respectively which are less than $0.05$ and thus leading to the rejection of the null hypothesis of equal means.

(18)

(d) No. There are two paints from company II and III with almost equal times. We need further tests to test paints from company II and III to determine which has the shortest drying time.

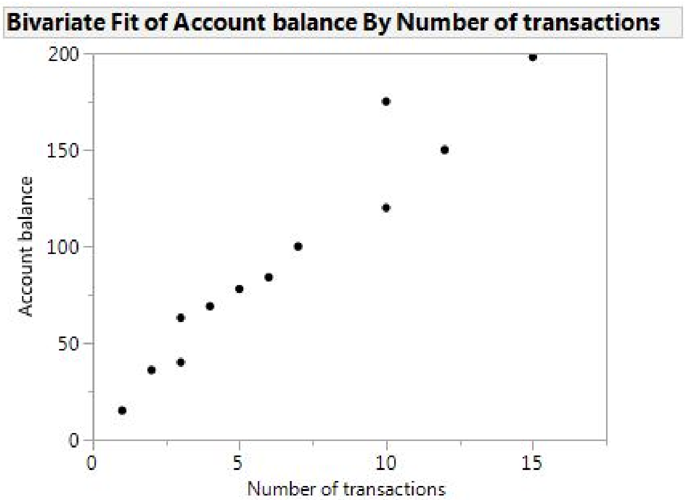
(2)

**[40]**

**QUESTION 5**

(a) The independent variable is the *number of transactions* and the dependent variable is *account balance*. (2)

(b) The scatter diagram is:



There is a strong positive relationship between *account balance* and *number of transactions*. (6)

(c) $n = 12$ $\qquad \Sigma X_i = 78$ $\qquad \Sigma X_i^2 = 718$

$\Sigma X_i Y_i = 9\,986$ $\qquad \Sigma Y_i = 1\,128$ $\qquad \Sigma Y_i^2 = 141\,720$

$$\begin{aligned} b &= \frac{n\Sigma X_i Y_i - (\Sigma X_i)(\Sigma Y_i)}{n\Sigma X_i^2 - (\Sigma X_i)^2} \\ &= \frac{12\,(9\,986) - (78)\,(1\,128)}{12\,(718) - (78)^2} \\ &= \frac{119\,832 - 87\,984}{8\,616 - 6\,084} \\ &= \frac{31\,848}{2\,532} \\ &\approx 12.5782 \end{aligned}$$

$$\begin{aligned} a &= \frac{\Sigma Y_i - b\,(\Sigma X_i)}{n} \\ &= \frac{1\,128 - 12.5782\,(78)}{12} \\ &= \frac{1\,128 - 981.0996}{12} \\ &= \frac{146.9004}{12} \\ &= 12.2417 \end{aligned}$$

The estimated regression equation is $\widehat{Account\ balance} = 12.2417 + 12.5782 No.\ of\ transactions$.

(6)

(d) For each additional transaction, the account balance increase on the average by $12.5782$ million, that is, $R12\,578\,200$. (2)

(e)

| $X_i$ | $Y_i$ | $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X$ | $e_i = Y_i - \widehat{Y}_i$ | $e_i^2 = \left(Y_i - \widehat{Y}_i\right)^2$ |
|---|---|---|---|---|
| | 36 | 37.3981 | −1.3981 | 1.954684 |
| | 63 | 49.9763 | 13.0237 | 169.616762 |
| | 175 | 138.0237 | 36.9763 | 1 367.246762 |
| | 69 | 62.5545 | 6.4455 | 41.544470 |
| | 15 | 24.8199 | −9.8199 | 96.430436 |
| | 198 | 200.9147 | −2.9147 | 8.495476 |
| | 40 | 49.9763 | −9.9763 | 99.526562 |
| | 120 | 138.0237 | −18.0237 | 324.853762 |
| | 84 | 87.7109 | −3.7109 | 13.770779 |
| | 150 | 163.1801 | −13.1801 | 173.715036 |
| | 78 | 75.1327 | 2.8673 | 8.221409 |
| | 100 | 100.2891 | −0.2891 | 0.083579 |
| | | | | 2 305.459716 |

Thus $\sum_{i=1}^{12} \left(Y_i - \widehat{Y}_i\right)^2 = 2\,305.459716$

Now

$$\begin{aligned} MSE &= s^2 \\ &= \frac{\sum (y_i - \widehat{y}_i)^2}{n-2} \\ &= \frac{1}{n-2} \sum \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X\right)^2 \\ &= \frac{2\,305.459716}{10} \\ &\approx 230.546 \\ \Longrightarrow\ & s = \sqrt{230.546} \approx 15.1837 \end{aligned}$$

$$\begin{aligned} d^2 &= \sum \left(X - \overline{X}\right)^2 \\ &= \sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n} \\ &= 718 - \frac{(78)^2}{12} \\ &= 718 - 507 \\ &= 211 \end{aligned}$$

$H_0 : \beta_1 = 0 \qquad\qquad H_1 : \beta_1 \neq 0$

$\alpha = 0.05 \qquad \alpha/2 = 0.025 \qquad t_{\alpha/2;n-2} = t_{0.025;10} = 2.228$. We will reject $H_0$ if $T \geq 2.228$ or $T \leq -2.228$ or if $|T| \geq 2.228$.

Now

$$\begin{aligned} T &= \frac{\hat{\beta}_1 - B_1}{s/d} \\ &= \frac{12.5782 - 0}{15.1837/\sqrt{211}} \\ &= \frac{12.5782}{1.045289016} \\ &\approx 12.0332 \end{aligned}$$

Since $12.0332 > 2.228$ we reject $H_0$ at the 5% level significance and conclude that $\beta_1 \neq 0$. This means that the regression line is significant to explain the variability in $y$. (Only when $\beta_1 = 0$, does it imply that regression is meaningless.)

(10)

(f) $X_i = 5$

Then

$$\begin{aligned} \widehat{Account\ balance} &= 12.2417 + 12.5782 No.of transactions \\ &= 12.2417 + 12.5782(5) \\ &= 12.2417 + 62.891 \\ &= 75.1327 \end{aligned}$$

$\therefore$ The predicted account balance is $R75\,132\,700$.

(2)

(g) The standard error of the estimate is given by $SE = S\sqrt{\dfrac{1}{n} + \dfrac{\left(X_i - \overline{X}\right)^2}{d^2}}$.

Then

$$\begin{aligned}
SE &= S\sqrt{\frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{d^2}} \\
&= 15.1837\sqrt{\frac{1}{12} + \frac{(5-6.5)^2}{211}} \\
&= 15.1837\sqrt{0.083333333 + \frac{2.25}{211}} \\
&= 15.1837\sqrt{0.083333333 + 0.010663507} \\
&= 15.1837\sqrt{0.09399684} \\
&\approx 4.6552
\end{aligned}$$

(5)

(h) The $95\%$ confidence interval for the slope $\beta_1$ is

$$\widehat{\beta}_1 \pm t_{\alpha/2;n-2} \times \frac{s}{d}$$

$\widehat{\beta}_1 = 12.5782$ $\qquad$ $t_{\alpha/2;n-2} = t_{0.025;10} = 2.228$

$d = \sqrt{211} \approx 14.5258$ $\qquad$ $s = 15.1837$

Thus, the 95% confidence interval for the slope $\widehat{\beta}_1$ is $0.0261$

$$\begin{array}{lcl}
\widehat{\beta}_1 & \pm & t_{\alpha/2;n-2} \times \dfrac{s}{d} \\
12.5782 & \pm & 2.228 \times \dfrac{15.1837}{14.5258} \\
12.5782 & \pm & 2.228 \times 1.045291826 \\
12.5782 & \pm & 2.3289 \\
(12.5782 - 2.3289 & ; & 12.5782 + 2.3289) \\
(10.2493 & ; & 14.9071)
\end{array}$$

(5)

(i) The corrrelation coefficient $r$ is

$$
\begin{aligned}
r &= \frac{\Sigma X_i Y_i - \dfrac{(\Sigma X_i)\,(\Sigma Y_i)}{n}}{\sqrt{\left(\Sigma X_i^2 - \dfrac{(\Sigma X_i)^2}{n}\right)\left(\Sigma Y_i^2 - \dfrac{(\Sigma Y_i)^2}{n}\right)}} \\
&= \frac{9\,986 - \dfrac{(78)\,(1\,128)}{12}}{\sqrt{\left(718 - \dfrac{(78)^2}{12}\right)\left(141\,720 - \dfrac{(1\,128)^2}{12}\right)}} \\
&= \frac{9\,986 - 7\,332}{\sqrt{(718 - 507)\,(141\,720 - 106\,032)}} \\
&= \frac{2\,654}{\sqrt{(211)\,(35\,688)}} \\
&= \frac{2\,654}{\sqrt{7\,530\,168}} \\
&= \frac{2\,654}{2\,744.115158} \\
&\approx 0.9672
\end{aligned}
$$

(5)

(j) $H_0 : \rho = 0.8$ against $H_1 : \rho \neq 0.8$

$n = 12$ $r = 0.97$

$$
\begin{aligned}
U &= \frac{1}{2}\log_e \frac{1+r}{1-r} & \eta &= \frac{1}{2}\log_e \frac{1+\rho}{1-\rho} \\
&= \frac{1}{2}\log_e \frac{1+0.97}{1-0.97} & &= \frac{1}{2}\log_e \frac{1+0.8}{1-0.8} \\
&= \frac{1}{2}\log_e \frac{1.97}{0.03} & &= \frac{1}{2}\log_e \frac{1.8}{0.2} \\
&= \frac{1}{2}\log_e 65.66666667 & &= \frac{1}{2}\log_e 9 \\
&\approx 2.0923 & &\approx 1.0986
\end{aligned}
$$

**Note: You can read the values from Table X Stoker**.

The test statistic is

$$\begin{aligned} z &= \sqrt{n-3}(U-\eta) \\ &= \sqrt{12-3}(2.0923-1.0986) \\ &= \sqrt{9}\times 0.9937 \\ &\approx 2.9811 \end{aligned}$$

$\alpha = 0.05$, $\alpha/2 = 0.025$ and $Z_{0.025} = 1.96$. Reject $H_0$ if $Z > 1.96$ or $Z < -1.96$ or $|Z| > 1.96$

Since $2.9811 > 1.96$, we reject $H_0$ at the $5\%$ level of significance and conclude that $\rho \neq 0.8$, that is, the correlation is significantly different from $\rho = 0.8$.

(10)

(k) $R^2 = (0.9672)^2 = 0.9354 \Longrightarrow 93.54\%$ of the variability in *account balance* is being explained / or accounted for by the least squares line.

(2)

(l) Model fitted is $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$

Commands for the Output:

*Start the JMP program*

$>$ *Enter* number of transactions *in the first column and label it* $\underline{Number\ of\ transactions\ (x)}$.

$>$ *Enter* account balance *in the second column and label it* $\underline{account\ balance\ (y)}$

*To plot:*

$>$ *Choose* Analyze$>$Fit $Y$ by $X$ *with* $\underline{Number\ of\ transactions\ (x)}$ *as* $X$ *factor and* $\underline{account\ balance\ (y)}$ *as* $Y$ *response.*

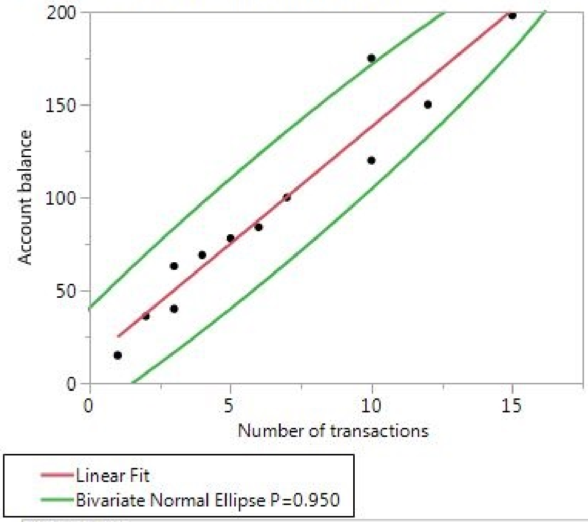
$>$ Click Ok.

*Click on the **Red** triangle on Bivariate Fit of* $account\ balance\ (y)$ *by* $Number\ of\ transactions\ (x)$.

$>$ *Choose* Fit Line

The JMP output is

### Bivariate Fit of Account balance By Number of transactions




### Linear Fit

Account balance = 12.241706 + 12.578199*Number of transactions

#### Summary of Fit

| | |
|---|---|
| RSquare | 0.9354 |
| RSquare Adj | 0.92894 |
| Root Mean Square Error | 15.18374 |
| Mean of Response | 94 |
| Observations (or Sum Wgts) | 12 |

#### Lack Of Fit

| Source | DF | Sum of Squares | Mean Square | F Ratio |
|---|---|---|---|---|
| Lack Of Fit | 8 | 528.4597 | 66.057 | 0.0743 |
| Pure Error | 2 | 1777.0000 | 888.500 | Prob > F |
| Total Error | 10 | 2305.4597 | | 0.9972 |
| | | | | Max RSq |
| | | | | 0.9502 |

#### Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Ratio |
|---|---|---|---|---|
| Model | 1 | 33382.540 | 33382.5 | 144.7978 |
| Error | 10 | 2305.460 | 230.5 | Prob > F |
| C. Total | 11 | 35688.000 | | <.0001* |

#### Parameter Estimates

| Term | Estimate | Std Error | t Ratio | Prob>\|t\| |
|---|---|---|---|---|
| Intercept | 12.241706 | 8.085542 | 1.51 | 0.1610 |
| Number of transactions | 12.578199 | 1.045292 | 12.03 | <.0001* |

#### Bivariate Normal Ellipse P=0.950

| Variable | Mean | Std Dev | Correlation | Signif. Prob | Number |
|---|---|---|---|---|---|
| Number of transactions | 6.5 | 4.379705 | 0.967161 | <.0001* | 12 |
| Account balance | 94 | 56.95932 | | | |

(10)

**[65]**

**[200]**