Tutorial Letter 203/1/2016

Applied Statistics II

STA2601

Semester 1

Department of Statistics

Solutions to Assignment 3

BAR CODE

QUESTION 1

(a) Based on the assumption of **independent observations** and the assumption that the weights have a **normal distribution** (i.e. that the sample comes from a normal population) we may assume that

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

Are they met? If we assume that there are no twins involved and that the babies were born to sixteen different mothers then the assumption of **independent observations are OK.**

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 negatively skewed and this is also supported by the boxplot with a longer tail to the left. Luckily the test is not too sensitive and we may proceed.

(4)

(b) We have to test $H_0 : \mu = 3.6$ against $H_0 : \mu > 3.6$.

$$n = \qquad \sum X_i = 56 \qquad \sum \left(X_i - \overline{X}\right)^2 = 3.9708$$

$$\begin{aligned}
\bar{X} &= \frac{\sum X_i}{n} = \frac{56}{16} = 3.5. \\
S_X^2 &= \frac{\sum \left(X_i - \overline{X}\right)^2}{n-1} = \frac{3.9708}{15} = 0.26472 \\
\therefore S_X &= 0.5145. \\
\therefore T &= \frac{\sqrt{16}\,(3.5 - 3.6)}{0.5145} \\
&= \frac{-0.4}{0.5145} \\
&\approx -0.7775
\end{aligned}$$

We will reject H_0 if $T \geq t_{0.05;\,16-1} == T \geq t_{0.05;\,15} = 1.753$ (Stoker, Table III).

Since $-0.7775 < 1.753 \Longrightarrow$ we do not reject H_0 at the 5% level of significance. The mean birth weight of a new born baby is 3.6 kg.

(8)

(c) If we know that $\sigma = 0.50$ we will use the test statistic

$$Z = \frac{\sqrt{n}\left(\bar{X} - \mu_0\right)}{\sigma} \sim n\,(0;\,1)\,.$$

For this specific sample, it becomes

$$\begin{aligned} Z &= \frac{\sqrt{16}\,(3.5 - 3.6)}{0.50} \\ &= -0.8. \end{aligned}$$

We will reject H_0 if $Z \geq z_{0.05} = 1.645$.

Since $-0.8 > 1.645 \Longrightarrow$ we do not reject H_0 at the 5% level of significance. The mean birth weight of a new born baby is 3.6 kg.

(6)

(d) A 90% two-sided confidence interval is computed as $\overline{X} - \left(\frac{S}{\sqrt{n}}\right)(t_{0.05;19}) < \mu < \overline{X} + \left(\frac{S}{\sqrt{n}}\right)(t_{0.05;19})$

where $3.5 \pm \left(\frac{0.5145}{\sqrt{16}}\right)(1.753) = 3.5 \pm (0.128625)\,(1.753) = 3.5 \pm 0.2255 = (3.2745;\ 3.7255)$
$\Longrightarrow 3.2745 \leq \mu \leq 3.7255.$

Yes. This means we do not reject $H_0 : \mu = 3.6$ since 3.6 is contained in the interval which confirms our conclusion in part (b).

(5)

(e) We have to test:

$H_0 : \quad \mu_d = 0$ against

$H_1 : \quad \mu_d \neq 0$

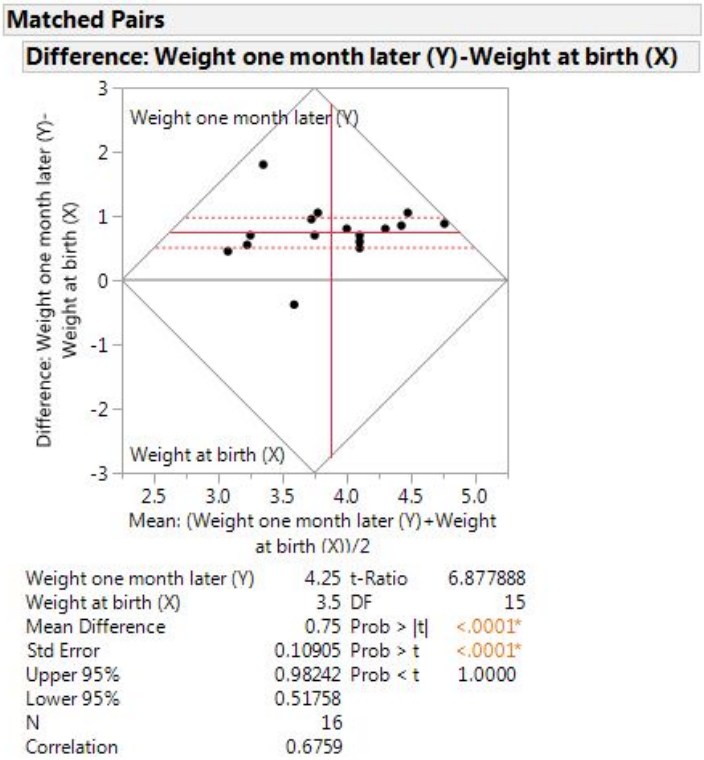


Figure 1: Paired T-Tests

Method 1: Using the critical value approach

From the output the test statistic is

$$t = \frac{(\overline{x} - \mu_d)}{\frac{s_d}{\sqrt{n}}} = \frac{(0.75 - 0)}{0.10905} \approx 6.8776$$

The critical value is $t_{\alpha/2;n-1} = t_{0.025;15} = 2.131$. Reject H_0 if $T \leq -2.131$ or $T \geq 2.131$.

Since $6.8776 > 2.131$, we reject H_0 in favour of H_1 at the 5% level of significance and conclude that $\mu \neq 0$. The baby has gained weight.

Method II: Using the p-value approach

p-value < 0.0001. Since $0.0001 << 0.05$, we can reject H_0 in favour of H_1 at the 5% level of significance and conclude that hat $\mu_d \neq 0$. The baby has gained weight. (7)

[30]

QUESTION 2

(a) JMP worked with the difference **male - female** which means the testing of $H_0 : \mu_{\text{Male}} = \mu_{\text{Female}}$ against $H_1 : \mu_{\text{Male}} > \mu_{\text{Female}}$.

Method 1: Using the critical value approach

From the output the test statistics is

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{3.56929}{0.96120} \approx 3.713348$$

The critical value is

$$\begin{aligned} t_{\alpha;n_1+n_2-2} &= t_{0.01;31} \\ &= 2.457 + \frac{1}{5}(2.438 - 2.457) \\ &= 2.457 + 0.2(-0.019) \\ &= 2.457 - 0.0038 \\ &\approx 2.4532 \end{aligned}$$

Reject H_0 if $T \geq 2.4532$.

Since $3.713348 > 2.4532$, we reject H_0 at the 1% level of significance and conclude that females have a lower average score.

Method II: Using the p-value approach

The t-test statistic is $t = 3.713348$ which is highly significant with a one-sided p-value $= 0.0004$). Since $0.0004 < 0.01$ we reject H_0 at the 1% level of significance and conclude that females have a lower average score.

(4)

(b) The t-test statistic is based on the assumptions that we have two independent groups from normal populations and that the variances are equal, that is,

1) Independent sample
2) Equal population variances, i.e. $\sigma_1^2 = \sigma_2^2$

3) Samples are from normal populations.

Assumption 1:

We cannot deduce this from the output and simply have to assume that the two groups are not connected.

Assumption 2:

We have to test:

$H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 \neq \sigma_2^2$

The output shows 5 different tests that the variances are equal and none of them are significant. The $F-$test has a p-value = 0.4489 which means we cannot reject $H_0 : \sigma_1^2 = \sigma_2^2$. Using the Levene's test, p-value $= 0.1448$. Since $0.1448 > 0.05 \Longrightarrow$ we can not reject H_0 at the 5% level of significance. The assumption of equal variances is not violated.

Assumption 3:

The output does not show any tests for normality and again we simply have to assume this.

(5)

(c) If we assume that $(\mu_1 - \mu_2) = (\mu_{\text{Female}} - \mu_{\text{Male}})$ a 95% two-sided confidence interval for $(\mu_1 - \mu_2)$ is $(-5.52968;\ -1.60890)$.

However, a 95% two-sided confidence interval for $(\mu_{\text{Male}} - \mu_{\text{Female}}) = (1.60890;\ 5.52968)$.

(3)

[12]

QUESTION 3

We are testing $H_0 : \mu = 20$ against $H_1 : \mu \neq 20$.

The power of the test is a function of Φ which is defined as $\Phi = \dfrac{\delta}{\sqrt{2}}$

$$\begin{aligned} \delta &= \frac{\sqrt{n}(\mu - \mu_0)}{\sigma} \\ &= \frac{\sqrt{n}(20 \pm \sqrt{2}\,\sigma - 20)}{\sigma} \\ &= \pm\sqrt{n}\sqrt{2} \\ \Longrightarrow \Phi = \frac{\delta}{\sqrt{2}} &= \pm\sqrt{n} \end{aligned}$$

From table F:

For $n = 7.$ $\upsilon = 6.$ $\Phi = \sqrt{7} = 2.645$ with power 0.86

For $n = 8.$ $\upsilon = 7.$ $\Phi = \sqrt{8} = 2.828$ with power 0.92

Thus, the **smallest sample size** to ensure that the power of the test, at the 5% level of significance, will be at least 0.90 is $n = 8.$

[5]

QUESTION 4

We have to use the t-statistic for independent samples with *unequal variances* if we want to test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2.$

$$\begin{aligned} \therefore T &= \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})}} \\ &= \frac{(40 - 42) - (0)}{\sqrt{\left(\frac{30}{15} + \frac{56}{18}\right)}} \\ &= -0.88465 \end{aligned}$$

The formula for the degrees of freedom is horrific to compute!

$$\begin{aligned} \upsilon &= \frac{\left[\frac{30}{15} + \frac{56}{18}\right]^2}{\frac{30^2}{15^2(15-1)} + \frac{56^2}{18^2(18-1)}} \\ &= \frac{(5.111111111)^2}{\frac{900}{3\,150} + \frac{3136}{5\,508}} \\ &= \frac{26.12345679}{0.855067953} \\ &\approx 30.5513 \end{aligned}$$

The critical region to test H_0 against H_1 is where $|T| \geq t_{0.025;\,30.55}$

Now if we interpolate, $t_{0.025;\,30.55} = 2.042 + \frac{0.55}{5}(2.030 - 2.042) = 2.0407.$

Since $-0.88465 < 2.0407$ we cannot reject H_0 at the 5% level of significance and conclude that $\mu_1 = \mu_2.$

[11]

QUESTION 5

(a) Yes, it is reasonable to assume that the five groups may be considered as *independent groups* because cars in one group cannot influence cars in the other groups. (2)

(b) No formal tests for normality are included in the output and the graphical output shows only the "Means Diamonds" which is not a graphical test for normality. To perform the ANOVA we simply have to assume that the five groups may be considered as *coming from normal populations.* (2)

(c) We have to test $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2$ against $H_1 : \sigma_p^2 \neq \sigma_q^2$ for at least one $p \neq q$.

From **Figure 8** we conclude that all the tests for the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2$ are significant at the 5% level of significance **but not at the 1% level of significance**. (Only one p-value is < 0.01, which is Bartlett's test with a **p-value = 0.0074**). It looks as if the assumption of *equal population variances* could be violated. (3)

(d) ANOVA Test:

(i) We have to test $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ against $H_1 : \mu_p \neq \mu_q$ for at least one pair $p \neq q$. (2)

(ii) We compute F manually by computing $MST_r = \dfrac{n\sum_{i=1}^{k}\left(\overline{X}_i - \overline{X}\right)^2}{k-1}$ which is defined for **sub-samples of equal sizes**. Our formula assumes that $n_1 = n_2 = ... = n_5$ In this case the groups are not equal. (2)

(iii) The **mean weights** for the different types of cars differ significantly (at any level of significance) because $F = 73.1036$ with a **p-value** $= \mathbf{0.0001}$ **which is highly significant.** This implies that $\mu_p \neq \mu_q$ for at least one pair $p \neq q$. (2)

(e) All pairs of means differ significantly except for the two groups "*Compact cars*" and "*Sporty cars*". This is graphically confirmed by the "Means Diamonds" where we can see that "*Compact cars*" and "*Sporty cars*" have almost identical pictures. On the "All Pairs Tukey-Kramer" display their two circles overlap completely.

From the output of the formal statistical test we see that the confidence interval for the difference of the mean weight (*Compact* - *Sporty*) $= (-162.59 : 168.751)$. This is the only interval **which includes zero** and means we cannot reject $\mu_{Compact} = \mu_{Sporty}$. Confirming this is the p-value for the t-statistic $= 0.9707 >> \alpha = 0.05$.

All the other intervals for the difference of the means are (positive value; positive value) **which excludes zero** and means **we reject** $\mu_p = \mu_q \Longrightarrow \mu_p \neq \mu_q$.

(4)

[17]

QUESTION 6

(a) For the manual solution we need the following computations:

	America(X)	Worldwide(Y)	(X-mean)**2	(Y-mean)**2	(X-mean)(Y-mean)
	100	215	16129	63001	31877
	102	152	15625	98596	39250
	170	340	3249	15876	7182
	90	122	18769	118336	47128
	110	236	13689	52900	26910
	250	539	529	5329	1679
	254	575	729	11881	2943
	280	574	2809	11664	5724
	359	796	17424	108900	43560
	410	857	33489	152881	71553
	320	697	8649	53361	21483
	105	193	14884	74529	33306
	190	356	1369	12100	4070
	300	600	5329	17956	9782
	280	564	2809	9604	5194
	180	420	2209	2116	2162
	370	752	20449	81796	40898
	210	396	289	4900	1190
	190	380	1369	7396	3182
	270	556	1849	8100	3870
Total	4540	9320	181646	911222	402943

$$\bar{x} = \frac{\sum x_i}{20} = \frac{4\,540}{20} = 227 \quad \text{and}$$

$$\bar{y} = \frac{\sum y_i}{20} = \frac{9\,320}{20} = 466.$$

We also compute

$$\sum (x_i - \bar{x})^2 = 181\,646 = d^2;$$

$$\sum (y_i - \bar{y})^2 = 911\,222 \quad \text{and}$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 402\,943.$$

$$\text{Hence, } \hat{\beta}_1 = \frac{\sum y_i (x_i - \bar{x})}{d^2}$$

$$= \frac{402\,943}{181\,646}$$

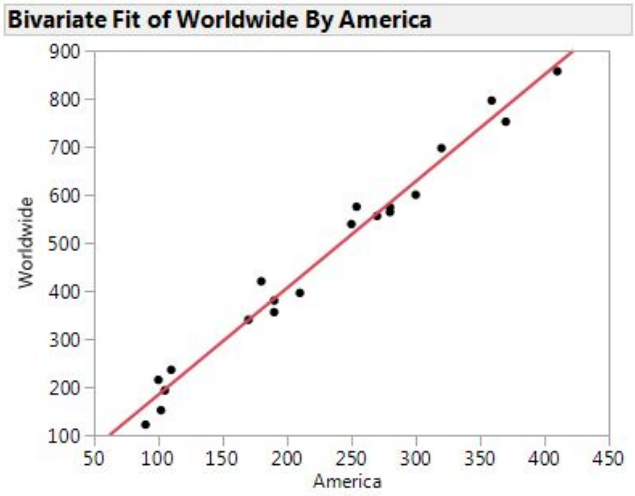
$$= 2.2183.$$

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ &= 466 - 2.2183(227) \\ &= -37.55.\end{aligned}$$

The equation of the regression line of y on x is: $y = -37.55 + 2.2183x$.

OR

Using SAS JMP, the output is



Linear Fit

Worldwide = -37.5512 + 2.2182872*America

Summary of Fit

RSquare	0.980928
RSquare Adj	0.979869
Root Mean Square Error	31.07222
Mean of Response	466
Observations (or Sum Wgts)	20

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	893843.31	893843	925.7991
Error	18	17378.69	965	Prob > F
C. Total	19	911222.00		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-37.5512	17.94883	-2.09	0.0509
America	2.2182872	0.072905	30.43	<.0001*

Figure 2: Simple Linear Regression Model

The equation of the regression line of y on x is: $y = -37.55 + 2.2183x$.

(9)

(b) If $x = 150 \implies y = -37.55 + 2.2183(150)$

$= 295.20$ billion dollars

(2)

(c) We need S^2 in order to compute the error of the estimate.

$$S^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} \quad \text{or} \quad (n-2)S^2 = \sum_{i=1}^{n}[Y_i - \overline{Y}]^2 - \frac{\left[\sum_{i=1}^{n}[Y_i - \overline{Y}][X_i - \overline{X}]\right]^2}{\sum_{i=1}^{n}[X_i - \overline{X}]^2}$$

$$= 911\,222 - \frac{(402\,943)^2}{181\,646} = 17\,378.69352$$

$$\text{Hence} \quad S^2 = \frac{17\,378.69352}{18} = 965.4830$$

Now if $X = 150$ is considered a "**future observation**" then $\hat{Y}_0(X) = \hat{\beta}_0 + \hat{\beta}_1 X$

and the error of the estimate is (see section 8.5 of the study guide) $S\sqrt{1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{d^2}}$.

$$S\sqrt{1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{d^2}} = 31.0722\sqrt{1 + \frac{1}{20} + \frac{(150 - 227)^2}{181\,646}}$$

$$= 31.0722\sqrt{1.08264041}$$

$$= 32.3306$$

This means that the predicted Box-office amount Worldwide is 295.20 ± 32.33 billion dollars for a Box-office amount of 150 billion dollars in America. (7)

(d) We have to test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$.

$$T = \frac{\hat{\beta}_1 - \beta_1}{s/d}$$

$$= \frac{2.2183 - 0}{31.0722/\sqrt{181646}}$$

$$\approx 30.4271.$$

The critical value is $t_{0.05/2;n-2} = t_{0.025;18} = 2.101$.

We will reject H_0 if $T \geq 2.101$ or if $T \leq -2.101$.

Since $30.427 > 2.101$ we reject H_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.)

(7)

[25]

[**100]**