8112

THE UNIVERSITY OF SYDNEY

FACULTIES OF ARTS AND SCIENCE

STAT 2012: Statistical Tests (Normal)

Semester 2, 2012 Time allowed: Two hours

Solution to examination

Part B: Extended answer questions Total 50 marks.

1. (a) Two independent samples:

(i) The 95% CI for $\mu_1 - \mu_2$ using two independent samples t-test is

$$\left(\bar{y}_1 - \bar{y}_2 - t_{\alpha/2,n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ \bar{y}_1 - \bar{y}_2 + t_{\alpha/2,n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\ \right)$$

$$=\left(12.333 - 7.8 - 2.262 \times 2.5849\sqrt{\frac{1}{6} + \frac{1}{5}},\ 12.333 - 7.8 + 2.262 \times 2.5849\sqrt{\frac{1}{6} + \frac{1}{5}}\right)$$

$$=(4.5333 - 2.262 \times 2.5849 \times 0.60553,\ 4.5333 + 2.262 \times 2.5849 \times 0.60553\ )$$

$$=(4.5333 - 2.262 \times 1.5652,\ 4.5333 + 2.262 \times 1.5652\ )$$

$$=(0.9925865,\ 8.0740802)$$

where

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{(6-1)7.467 + (5-1)5.7}{6+5-2}} = 2.5848562$$

Since the CI does not contain 0, the time until some reliefs were felt differs between the two formulas.

(ii) Wilcoxon rank sum test:

We have $n_x = 6$, $n_y = 5$ and $N = n_x + n_y = 11$. The ranks in the combined sample are

Sample 1: 11 (6.5) 8 (3.5) 12 (8) 16 (11) 13 (9) 14 (10)

Sample 2: 8 (3.5) 6 (2) 5 (1) 11 (6.5) 9 (5)

Since there are ties, normal approximation should be used and so there is NO need to just base on sample of lower size. Hence

$$
\begin{aligned}
W &= 6.5{+}3.5{+}8{+}11{+}9{+}10 = 48 \quad \text{(sample 1)} \qquad \text{or} \quad \tfrac{11(11+1)}{2} - 48 = 18 \quad \text{(sample 2)} \\
E(W) &= \frac{1}{2} n_x(n_x + n_y + 1) = \frac{1}{2} 6(11+1) = 36 \quad \text{(sample 1)} \qquad \text{or} \quad \frac{1}{2} 5(12) = 30 \quad \text{(sample 2)} \\
\sum_i r_i^2 &= 6.5^2 + 3.5^2 + 8^2 + 11^2 + 9^2 + 10^2 + 3.5^2 + 2^2 + 1^2 + 6.5^2 + 5^2 = 505 \\
Var(W) &= \frac{n_x \times n_y}{N(N-1)} \left(\sum_i r_i^2 - \frac{1}{4} N(N+1)^2 \right) = \frac{6(5)}{11(10)} \left(505 - \frac{1}{4} 11 \times 12^2 \right) = 29.72727 \\
\text{p-value} &= 2\Pr(W \geq 48) = 2\Pr\left(Z \geq \frac{48-36}{\sqrt{29.72727}} \right) = 2\Pr(Z \geq 2.20092) = 2(1 - 0.9861) = 0.0278 \\
&\overset{or}{=} 2\Pr(W \leq 18) = 2\Pr\left(Z \leq \frac{18-30}{\sqrt{29.72727}} \right) = 2\Pr(Z \leq -2.20092) = 0.0278 \quad \text{(sample 2)}
\end{aligned}
$$

(b) Two-way data:

(i) We have $N = 9$, $\bar{r} = \frac{1}{2}(1+3) = 2$. The ranks are

	Temperature		
Moisture	$15°C$	$25°C$	$35°C$
50%	63 (1)	74 (2)	81 (3)
70%	72 (1)	80 (2)	105 (3)
90%	57 (1)	65 (3)	61 (2)
$\bar{r}_{\cdot j}$	1	2.333	2.667

There are no ties. The Friedman test for the equality of mean germination rate across the three levels of temperature is

1. **Hypothesis:** $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ vs H_1: Not all β_j are zero.

2. **Test statistic:**

$$
\begin{aligned}
q_0 &= \frac{12r}{c(c+1)} \sum_{j=1}^{c} (\bar{r}_{\cdot j})^2 - 3r(c+1) \\
&= \frac{12(3)}{3(3+1)} (1^2 + 2.333^2 + 2.667^2) - 3(3)(3+1) = \frac{14}{3} = 4.667 \\
&\overset{or}{=} (rc - r) \frac{r \sum_j \bar{r}_j^2 - rc\bar{r}^2}{\sum_i \sum_j r_{ij}^2 - rc\bar{r}^2} = \frac{(3 \cdot 3 - 3)[\frac{1^2+7^2+8^2}{3} - 9(\frac{3}{2})^2]}{3(1^2 + 2^2 + 3^2) - 9(\frac{3}{2})^2} \\
&= \frac{6(4.667)}{6} = 4.667
\end{aligned}
$$

3. ***P*-value:** $\Pr(\chi_2^2 \geq \frac{14}{3}) \in (0.05, 0.1)$.

4. **Decision:** Since the p-value > 0.05, the data are consistent with H_0 that the germination rate is the across the 3 temperature levels.

(ii) Since there is an outlier, the Friedman test is preferred as the test uses ranks which are less affected by outliers.

2. Two-way ANOVA with replicates:

The number of data $n = 36$, the number of blocks $r = 4$, the number of treatments $c = 3$ and the number of replicates $m = 3$. Summary of data is

$$\begin{aligned}
\bar{y}_{i.} &= 48.667, \quad 56.0, \quad 24.667, \quad 41.0 \\
\bar{y}_{.j} &= 42.25, \quad 47.75, \quad 37.75 \\
\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{m} y_{ijk}^2 &= 47^2 + 50^2 + 53 + \ldots + 30^2 = 71751 \\
\bar{y}_{...} &= 42.583 \\
\sum_{i=1}^{r}\sum_{j=1}^{c} s_{ij}^2 &= 177
\end{aligned}$$

(a) We also have the following sums:

$$\begin{aligned}
\sum_{i=1}^{r} \bar{y}_{i..}^2 &= 48.667^2 + 56.0^2 + 24.667^2 + 41.0^2 = 7793.889 \\
\sum_{j=1}^{c} \bar{y}_{.j.}^2 &= 42.25^2 + 47.75^2 + 37.75^2 = 5490.188
\end{aligned}$$

$$\begin{aligned}
CM &= n\bar{y}^2 = 36(42.583^2) = 65280.25 \\
SST_o &= \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{m} y_{ijk}^2 - n\bar{y}^2 = 71751 - 65280.25 = 6470.75 \\
SST &= rm\sum_{j=1}^{c} \bar{y}_{.j.}^2 - n\bar{y}^2 = 4(3)(5490.188) - 65280.25 = 602 \\
SSB &= cm\sum_{i=1}^{r} \bar{y}_{i..}^2 - n\bar{y}^2 = 3(3)(7793.889) - 65280.25 = 4864.75 \\
SSR &= (m-1)\sum_{i=1}^{r}\sum_{j=1}^{c} s_{ij}^2 = 2(177) = 354 \\
SSI &= SST_o - SST - SSB - SSR = 6470.75 - 602 - 4864.75 - 354 = 650
\end{aligned}$$

The ANOVA table for two-way data with replicate is

ANOVA table

Source	df	SS	MS	F
Treatments (Method)	2	602	$\frac{602}{2} = 301$	$\frac{301}{14.75} = 20.40678$
Blocks (Food type)	3	4864.75	$\frac{4864.75}{3} = 1621.5833$	$\frac{1621.5833}{14.75} = 109.9379$
Interaction	6	650	$\frac{650}{6} = 108.3333$	$\frac{108.3333}{14.75} = 7.344633$
Residuals	24	354	$\frac{354}{24} = 14.75$	
Total	35	6470.75		

(b) The two-way ANOVA tests for day effects are

1. **Hypothesis:** $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ vs $H_1:$ Not all β_j are the same
2. **Test statistic:** $f_{t0} = \dfrac{SST/(c-1)}{SSR/(r-1)(c-1)} = \dfrac{602/2}{354/24} = 20.40678$
3. **Assumption:** $Y_{ijk} \sim \mathcal{N}(\mu + \alpha_i + \beta_j + \delta_{ij}, \sigma^2)$ and Y_{ijk} are independent.
4. *P*-**value:** $\Pr(F_{2,24} \geq 20.40678) < 0.001 \quad (F_{2,24,0.999} = 9.34)$.
5. **Decision:** Since p-value < 0.05, there is strong evidence in the data against H_0. The three methods of irradiation to reduce bacteria for food preservation are not all the same.

(c) The revised test statistic is

$$f'_{t0} = \frac{SST/(c-1)}{SSR'/(rcm - r - c + 1)} = \frac{602/2}{(354+650)/(6+24)} = \frac{301}{} = 8.994024$$

3. Regression analysis:

(a) Given

$$\sum_{i=1}^{10} x_i = 43.3, \quad \sum_{i=1}^{10} y_i = 401, \quad n = 10,$$
$$\sum_{i=1}^{10} x_i^2 = 238.77, \quad \sum_{i=1}^{10} y_i^2 = 19633, \quad \sum_{i=1}^{10} x_i y_i = 2161.2,$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)^2 = 238.77 - \frac{1}{10} 43.3^2 = 51.281,$$
$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)\Big(\sum_{i=1}^{n} y_i\Big) = 2161.2 - \frac{1}{10}(43.3)(401) = 424.870,$$

$$\begin{aligned}\hat{\beta} &= \frac{S_{xy}}{S_{xx}} = \frac{424.87}{51.281} = 8.2851348\\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} = \frac{401}{10} - 8.2851348\,\frac{43.3}{10} = 4.2253661.\end{aligned}$$

Hence the fitted least squares line is

$$[1]\ \hat{y} = \hat{\alpha} + \hat{\beta}\,x = 4.2253661 + 8.2851348\,x.$$

(b) The test for the regression model in (a) is

1. **Hypotheses:** H_0: $\beta = 0$ vs $H_1: \ \beta \neq 0$.
2. **Test statistic:** $t_0 = \dfrac{\hat{\beta}}{\sqrt{\dfrac{s^2}{S_{xx}}}} = \dfrac{8.2851348}{\sqrt{\dfrac{4.099345}{51.281}}} = 29.30357$, where

$$\begin{aligned}S_{yy} &= \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} = 19633 - \frac{401^2}{10} = 3552.9\\ SSR &= S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 3552.9 - \frac{424.87^2}{51.281} = 32.79476\\ s^2 &= \frac{SSR}{n-2} = \frac{32.79476}{8} = 4.099345\end{aligned}$$

3. **Assumption:** $Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$. Y_i are independent.
4. **P-value:** [0.5] p-value $= 2\Pr(t_8 > 29.30357) < 0.002$ ($t_{8,0.001} = 4.501$)
5. **Decision:** [0.5] Since p-value < 0.05, there is strong evidence in the data that a linear relationship exists between Y, gross revenue, and X, the payment to two highest paid actors/actresses.

(c) The predicted gross revenue when the paid to the two highest-paid actors/actresses in the movie:

$$\begin{aligned}\widehat{y}|\, x_0 = 8 &= \hat{\alpha} + \hat{\beta}x_0 = 4.2253661 + 8.2851348\,(8) = 70.50644.\\ \text{s.e.}(\widehat{y}|\, x_0 = 8) &= \sqrt{s^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}\\ &= \sqrt{4.099345\left(1 + \frac{1}{10} + \frac{(8 - 4.33)^2}{51.281}\right)} = 2.363465.\end{aligned}$$

The 95% Prediction Interval for the gross revenue when the is $x_0 = 12$ lb:

$$\left[(\hat{\alpha} + \hat{\beta}x_0) - t_{\alpha/2,n-2}\ s_e\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},\ (\hat{\alpha} + \hat{\beta}x_0) + t_{\alpha/2,n-2}\ s_e\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\right]$$

$$= \quad (70.50644 \ - \ 2.306 \times 2.363465, \ 70.50644 \ + \ 2.306 \times 2.363465)$$
$$= \quad (65.05628, 75.95661).$$

(d)

$$\sum_{i=1}^{n} \hat{e}_i = \sum_{i=1}^{n}[y_i-(\hat{\alpha}+\hat{\beta}x_i)] = \sum_{i=1}^{n} y_i - n\hat{\alpha} - \hat{\beta}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i - n(\bar{y}-\hat{\beta}\bar{x}) - \hat{\beta}\sum_{i=1}^{n} x_i = 0$$

4. Chi-square GOF test for the binomial distribution:

(a) The sample estimate of π is

$$\hat{\pi} = \frac{1}{400}\sum_{i=1}^{4} i \times O_i = \frac{1}{400}[0(22) + 1(32) + 2(20) + 3(16) + 4(10)] = 0.4.$$

(b) (5 marks) The expected frequencies:

i	O_i	$c_i = \sum_{j=0}^{i} p_i$	p_i	$E_i = 100p_i$
0	22	0.1296	0.1296	12.96
1	32	0.4752	0.3456	34.56
2	20	0.8208	0.3456	34.56
3	16	0.9744	0.1536	15.36
4	10	1.0000	0.0256	2.56
Sum	100		1.0000	100

i	O_i	E_i	$\frac{(O_i-E_i)^2}{E_i}$
0	22	12.96	$\frac{(22-12.96)^2}{12.96} = 6.306$
1	32	34.56	$\frac{(32-34.56)^2}{34.56} = 0.190$
2	20	34.56	$\frac{(20-34.56)^2}{34.56} = 6.134$
≥ 3	26	17.92	$\frac{(26-17.92)^2}{17.92} = 3.642$
Sum	100	100	16.273

(c) The Chi-square test is

1. **Hypothesis:** H_0: $i \sim$ Bin(4,π) vs H_1: i do not follow Bin(4,π)
2. **Test statistic:** $\chi_0^2 = \sum_{i=0}^{4} \frac{(y_i - 100p_i)^2}{100p_i} = 16.273$
3. **P-value:** $\Pr(\chi_2^2 > 16.273) < 0.01 \quad (\chi^2_{2,0.01} = 9.210)$
4. **Conclusion:** The data are against H_0. The data of the number of members i with blood type A in a family of 4 do not follow the Bin(4,π).