SAS — Assignment 4

牛孟杰 11910901

Problem 1.

Note that $SSM = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\bar{y}_{ij}-\bar{y})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}[(\bar{y}_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y})+(\bar{y}_{i\cdot}-\bar{y})+(\bar{y}_{\cdot j}-\bar{y})]$

$$= SSAB + SSA + SSB + 2\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\bar{y}_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y})(\bar{y}_{i\cdot}-\bar{y})}_{1^\circ} + 2\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\bar{y}_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y})(\bar{y}_{\cdot j}-\bar{y})}_{2^\circ}$$

$$+2\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\bar{y}_{i\cdot}-\bar{y})(\bar{y}_{\cdot j}-\bar{y})}_{3^\circ}$$

let all of n_{ij} equal n_0

$$\Rightarrow 1^\circ = n_0\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y})(\bar{y}_{i\cdot}-\bar{y}) = n_0\sum_{i=1}^{a}\left[(\bar{y}_{i\cdot}-\bar{y})\sum_{j=1}^{b}(\bar{y}_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y})\right]$$

$$= n_0\sum_{i=1}^{a}(\bar{y}_{i\cdot}-\bar{y})(b\bar{y}_{i\cdot}-b\bar{y}_{i\cdot}-b\bar{y}+b\bar{y}) = 0$$

Similarly, $2^\circ = n_0\sum_{j=1}^{b}(\bar{y}_{\cdot j}-\bar{y})(a\bar{y}_{\cdot j}-a\bar{y}_{\cdot j}-a\bar{y}+a\bar{y}) = 0.$

$3^\circ = n_0\sum_{i=1}^{a}(\bar{y}_{i\cdot}-\bar{y})(b\bar{y}-b\bar{y}) = 0.$

from above, we have $SSM = SSA + SSB + SSAB$

Problem 2.

Problem 2(1)

From problem, we have $\bar{Y}_A = 55,\ S_A^2 = 300,\ \bar{Y}_B = 70,\ S_B^2 = 225,$

$\bar{Y}_C = 65,\ S_C^2 = 257.143,\ \bar{Y}_D = 80,\ S_D^2 = 175$

then, the overall mean $\bar{Y} = \dfrac{6\bar{Y}_A + 7\bar{Y}_B + 8\bar{Y}_C + 7\bar{Y}_D}{28} = 67.85$

$\Rightarrow SSW = \sum_{i=1}^{4}(n_i-1)S_i^2 = 5S_A^2+6S_B^2+7S_C^2+6S_D^2 = 1500+1350+1800+1050 = 5700$

$SSB = \sum_{i=1}^{4} n_i(\bar{Y}_i-\bar{Y})^2 = 6(55-67.85)^2+7(70-67.85)^2+8(65-67.85)^2+7(80-67.85)^2 = 2121.43$

$\Rightarrow$ ANOVA table:

Source	Df	Sum of Squares	Mean Squares	F value	P-value.
Between	3	2121.43	707.14	2.977	0.0516
Within	24	5700	237.5		
Total	27	7821.43			

Note that under H_0, $F = \dfrac{SSB/3}{SSW/24} \sim F(3,24)$

As p-value < 0.1, $H_0: \mu_A=\mu_B=\mu_C=\mu_D$ is rejected at signifiance level $\alpha=0.1$, $\Rightarrow$ which means the mean score of different classes are not all equal.

Problem 2(2)

First of all, we need to give the dispersion variable $z_{ij} = (y_{ij}-\bar{y}_i)^2$

$\rightarrow$

Factors	A	B	C	D
	625	100	225	100
z_{ij}	225	100	25	0
	25	100	625	25
	400	400	625	25
	0	25	25	100
	225	225	225	400
		400	25	400
			25	

$\Rightarrow$ then we need to perform the one-way ANOVA on z_{ij}:

We have $\bar{z}_A = 250$. $S^2_{z_A} = 55500$. $\bar{z}_B = 192.86$, $S^2_{z_B} = 23482.14$

$\bar{z}_C = 225$ $S^2_{z_C} = 68571.43$ $\bar{z}_D = 150$. $S^2_{z_D} = 30625$, $\bar{z} = 203.5714$

$$\Rightarrow SSW = \sum_{i=1}^{4} (n_i - 1) S^2_{z_i} = 5S^2_{z_A} + 6S^2_{z_B} + 7S^2_{z_C} + 6S^2_{z_D} = 1082143$$

$$SSB = \sum_{i=1}^{4} n_i (\bar{z}_i - \bar{z})^2 = 6(250-203.5714)^2 + 7(192.86-203.5714)^2 + 8(225-203.5714)^2$$
$$+ 7(150-203.5714)^2 = 3750$$

$\Rightarrow$ the ANOVA table based on z_{ij} is

Source	Df	Sum of Squares	Mean Squares	F value	P-value.
Between	3	37500	12500	0.277	0.841
Within	24	1082143	45089		
Total	27				

As p-value > 0.1, we can not reject $H_0: \sigma_A^2 = \sigma_B^2 = \sigma_C^2 = \sigma_D^2$ at the significance level $\alpha = 0.1$, which means that we cannot reject the equal-variance assumption.

Problem 2(3)

We can rank this observations at first, the following is the rank table.

y_{ij}	35	40	40	50	50	50	55	55	60	60	60	60	70	70	70	70	75	75
group	A	A	C	A	B	C	A	B	B	C	C	D	A	C	C	D	B	D
r_{ij}	1	2.5	2.5	5	5	5	7.5	7.5	10.5	10.5	10.5	10.5	14.5	14.5	14.5	14.5	17.5	17.5

(continue)

y_{ij}	80	80	80	80	80	85	90	90	90	100
group	A	B	B	C	D	D	B	C	D	D
r_{ij}	21	21	21	21	21	24	26	26	26	28

$\bar{r}_A = \frac{1+2.5+5+7.5+14.5+21}{6} = \frac{103}{12}$ $\bar{r}_B = \frac{5+7.5+10.5+17.5+21+21+26}{7} = 15.5$

$\bar{r}_C = \dfrac{2.5+5+10.5+10.5+14.5+14.5+21+26}{8} = 13.0625$ $\bar{r}_D = \dfrac{10.5+14.5+17.5+21+24+26+28}{7} = \dfrac{283}{14}$

$\Rightarrow$ test statistic of Kruskal-Wallis test is $W = \dfrac{(28-1)\sum n_i(\bar{r}_i - \bar{r})^2}{\sum\sum(r_{ij}-\bar{r})^2} = 6.9264$

Note that $KW \sim \chi^2_3$ approximately under H_0. $\Rightarrow$ P-value $= 0.0743 > 0.05$. We can not reject that group medians of these groups are equal at the significance level $\alpha = 0.05$.

Problem 3

Problem 3(1)

The ANOVA table is as follows

源	自由度	平方和	均方	F 值	Pr > F
模型	3	137.7023590	45.9007863	23.22	<.0001
误差	35	69.1920000	1.9769143		
校正合计	38	206.8943590			

The p-value<0.0001, which indicates the data provide sufficient evidence that worker productivity under the 4 levels of component arrival rate are different. (at siginificant level $\alpha = 0.05$)

Problem 3(2)

The ANOVA table is as follows

源	自由度	平方和	均方	F 值	Pr > F
模型	3	62.4494701	20.8164900	5.04	0.0052
误差	35	144.4448889	4.1269968		
校正合计	38	206.8943590			

The p-value=0.0052<0.05, which indicates the data provide sufficient evidence that worker productivity under the 4 levels of room temperature are different. (at siginificant level $\alpha = 0.05$)

Problem 3(3)

The simultaneous confidence intervals using the Tukey-Kramer method is given below:

Comparisons significant at the 0.05 level are indicated by ***.				
X2 比较	均值间差值	Simultaneous 95% 置信限		
70 - 75	1.8200	-0.6302	4.2702	
70 - 65	2.7000	0.2498	5.1502	***
70 - 80	3.3889	0.8716	5.9062	***
75 - 70	-1.8200	-4.2702	0.6302	
75 - 65	0.8800	-1.5702	3.3302	
75 - 80	1.5689	-0.9484	4.0862	
65 - 70	-2.7000	-5.1502	-0.2498	***
65 - 75	-0.8800	-3.3302	1.5702	
65 - 80	0.6889	-1.8284	3.2062	
80 - 70	-3.3889	-5.9062	-0.8716	***
80 - 75	-1.5689	-4.0862	0.9484	
80 - 65	-0.6889	-3.2062	1.8284	

From the above figure, we have

- The Ci of $\mu_A - \mu_B$ is $[-5.1502, -0.2498]$, which a indicates there exists significant difference between 65°F and 70°F at significant level $\alpha = 0.05$.
- The Ci of $\mu_A - \mu_C$ is $[-3.3302, 1.5702]$, which indicates no significant difference between 65°F and

75°F at significant level $\alpha = 0.05$.

- The Ci of $\mu_A - \mu_D$ is $[-1.8284, 3.2062]$, which indicates no significant difference between 65°F and 80°F at significant level $\alpha = 0.05$.
- The Ci of $\mu_B - \mu_C$ is $[-0.6302, 4.2702]$, which indicates no significant difference between 70°F and 75°F at significant level $\alpha = 0.05$.
- The Ci of $\mu_B - \mu_D$ is $[0.8716, 5.9062]$, which indicates there exists significant difference between 70°F and 80°F at significant level $\alpha = 0.05$.
- The Ci of $\mu_C - \mu_D$ is $[-0.9484, 4.0862]$, which indicates no significant difference between 75°F and 80°F at significant level $\alpha = 0.05$.

Problem 3(4)

The consequence two-way ANOVA is given below:

源	自由度	III 型 SS	均方	F 值	Pr > F
X1	3	116.6247752	38.8749251	155.59	<.0001
X2	3	46.1385116	15.3795039	61.55	<.0001
X1*X2	9	17.0887224	1.8987469	7.60	<.0001

Note that the P-value of the interaction effect between component arrival rate and room temperature is 0.0001 < 0.05, so that the data provide sufficient evidence to indicate an interaction effect between between component arrival rate X_1 and room temperature X_2 on worker productivity at significant level $\alpha = 0.05$.

Problem 3(5)

The consequence of test normality of the residuals are given below:

正态性检验				
检验	统计量		p 值	
Shapiro-Wilk	W	0.990597	Pr < W	0.9830
Kolmogorov-Smirnov	D	0.055452	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.01374	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.1015	Pr > A-Sq	>0.2500

Since the sample size is small here, we choose to look at the Shapiro-Wilk test, where the corresponding p-value is 0.9830>0.05, which indicates that the data does provide enough evidence to show the normal distribution of residuals at significant level $\alpha = 0.05$.

Problem 3(6)

Treating X_1 and X_2 as continuous variables, perform a regression analysis. The consequence is given below:

源	自由度	III 型 SS	均方	F 值	Pr > F
X1	1	6.26193717	6.26193717	8.78	0.0057
X1*X1*X1	1	9.91584598	9.91584598	13.90	0.0007
X2	1	10.79998894	10.79998894	15.13	0.0005
X1*X1	1	8.18512907	8.18512907	11.47	0.0019
X2*X2	1	10.22778710	10.22778710	14.33	0.0006
X2*X2*X2	1	9.65762864	9.65762864	13.53	0.0009

Note that

- The p-value of quadratic form of X_1 is 0.0019<0.05, which indicates that we need to add quadratic term of X_1 into the model. (at significant level $\alpha = 0.05$)
- The p-value of cubic form of X_1 is 0.0007<0.05, which indicates that we need to add cubic term of X_1 into the model. (at significant level $\alpha = 0.05$.)
- The p-value of quadratic form of X_2 is 0.0006<0.05, which indicates that we need to add quadratic term of X_2 into the model. (at significant level $\alpha = 0.05$.)
- The p-value of cubic form of X_2 is 0.0009<0.05, which indicates that we need to add cubic term of X_2 into the model. (at significant level $\alpha = 0.05$.)

Problem 3(7)

Two-way ANOVA is better.

- Even if quadratic and cubic terms of X_1, X_2 are added to the model in the problem 3(6) to improve the fitness, this model still does not consider the interaction terms of X_1 and X_2, and the effect of interaction terms can be proven to be significant in the problem 3(4).
- The R^2 of regression model is 0.88963 even after adding quadratic and cubic term of both X_1 and X_2 into the model, which is less than the R^2 of the two-way ANOVA model: 0.97222.

Problem 4

Problem 4(1)

The distribution and some statistics of 1000 p-values are given below:

矩			
数目	1000	权重总和	1000
均值	0.49865877	观测总和	498.658767
标准差	0.28896303	方差	0.08349963
偏度	-0.0001669	峰度	-1.1838695
未校平方和	332.0767	校正平方和	83.4161342
变异系数	57.9480501	标准误差均值	0.00913781

We filter out those with p-value less than 0.05, which is the probability of making type I error. We found a total of 49 elements.

总行数: 49 总列数: 1	行 1-49
	PROB
1	0.0124746849
2	0.0405512619
3	0.0149561583
4	0.0171768839
5	0.0086566925

Result: The Type I error rate based on the 1000 p-values ($\alpha = 0.05$) is 0.049.

Problem 4(2)

The distribution and some statistics of 1000 p-values are given below:

矩			
数目	1000	权重总和	1000
均值	0.50617449	观测总和	506.174488
标准差	0.30061263	方差	0.09036795
偏度	-0.0902094	峰度	-1.2519511
未校平方和	346.490198	校正平方和	90.2775858
变异系数	59.3891313	标准误差均值	0.00950621

We filter out those with p-value less than 0.05, which is the probability of making type I error. We found a total of 71 elements

总行数: 71 总列数: 1	行 1-71
	PROB
1	0.000947404[illegible]
2	0.008876120[illegible]
3	0.0351019971
4	0.008928820[illegible]
5	0.034281516[illegible]

Result: The Type I error rate based on the 1000 p-values ($\alpha = 0.05$) is 0.071.

Conclusion: When different groups have unequal variances (heteroscedasticity), using the traditional F-test in one-way ANOVA to compare group means would result in “Type I error increase” at a given significance level α, i,e,. at given significance level α, the rate of making type I errors will be significantly greater than α.