MATH 4720 / MSSC 5720

Instructor: Mehdi Maadooliat

Review for
Final Exam

Department of Mathematical and Statistical Sciences

(1) A study of potential age discrimination considers promotions among middle managers in a large company. The data are as follows:

	Age			
	Under 30	30-39	40-49	50 and over
Promoted	9	29	32	10
Not promoted	41	41	48	40

Is there a statistically significant relationship between age and promotion at $\alpha = 0.05$?

(1) $\underline{\underline{H_0}}$: Age and Promotion are indep

$\underline{\underline{H_a}}$ " " " are NOT indep.

	Age				
	Under 30	30-39	40-49	50 and over	Total
Promoted	9	29	32	10	80
NOT "	41	41	48	40	170
Total	50	70	80	50	250

Observed (O_{ij}'s)

Assumption: All $E_{ij} > 5$ ✓

Expected Counts (E_{ij}'s)

	Under 30	30-39	40-49	50 and over	
Promoted	$\frac{50\times80}{250}=16$	$\frac{70\times80}{250}=22.4$	$\frac{80\times80}{250}=25.6$	$\frac{50\times80}{250}=16$	80
Not "	$\frac{50\times170}{250}=34$	$\frac{70\times170}{250}=47.6$	$\frac{80\times170}{250}=54.4$	$\frac{50\times170}{250}=34$	170
Total	50	70	80	50	250

T.S $\chi^2 = \sum_{i,j} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$

$$\chi^2 = \frac{(9-16)^2}{16} + \frac{(29-22.4)^2}{22.4} + \frac{(32-25.6)^2}{25.6} + \frac{(10-16)^2}{16}$$

$$+ \frac{(41-34)^2}{34} + \frac{(41-47.6)^2}{47.6} + \frac{(48-54.4)^2}{54.4} + \frac{(40-34)^2}{34}$$

$\Rightarrow \chi^2 = 13.025$

$\Rightarrow \chi^2 = 13.025$

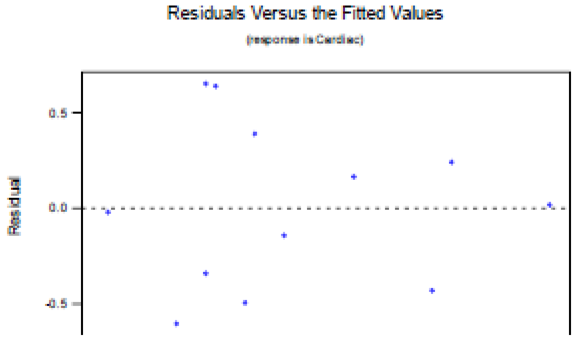
Rej H_0 if $\chi^2 > \underbrace{\chi^2_{0.05} \left(df = (4-1)(2-1)\right)}_{7.815}$

$\Rightarrow 13.025 > 7.815 \Rightarrow$ there is a significant relationship between age and promotion.

(2) A linear regression model is to be tested from a data on 12 male patients with congestive heart failures. The dependent variable is the cardiac index Y and the independent variable is the weight X. The Minitab output of the regression and the residual analysis is given below.

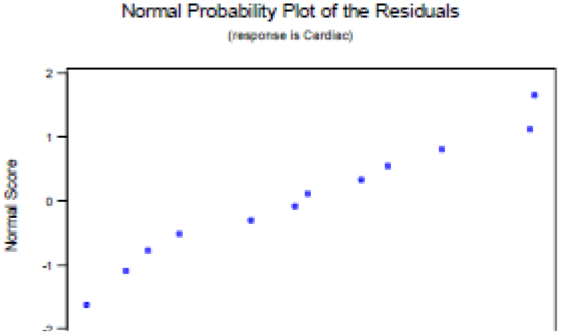
Regression Analysis

```
The regression equation is
Cardiac Index = 0.909 + 0.0126 Weight(kg)

Predictor        Coef       StDev          T        P
Constant       0.9090      0.7837       1.16    0.273
Weight(kg)    0.01256     0.01023       1.23    0.247

S = 0.4499      R-Sq = 13.1%     R-Sq(adj) = 4.4%
```



Residuals Versus the Fitted Values

(response is Cardiac)



Normal Probability Plot of the Residuals

(response is Cardiac)

(b) Intercept : the expected response (y) for $x=0$ $\Longrightarrow$ 0.909

A 95% CI of β_0 is $\hat{\beta}_0 \pm t_{\alpha/2} se(\hat{\beta}_0)$, i.e., $0.909 \pm 2.229 \times 0.7837$, i.e. (-.0838, 2.656).

slop : the amount of change in cardiac index for a unit change in weight.

: 0.01256

A 95% CI of β_1 is $\hat{\beta}_1 \pm t_{\alpha/2} se(\hat{\beta}_1)$, i.e., $0.01256 \pm 2.229 \times 0.01023$, i.e. (-0.01024, 0.03536).

(a) What is the least square line of best fit describing the relationship between Y and the X.?

(b) Interpret the intercept and the slope of the line, and find the 95% confidence intervals for the true intercept and the true slop parameters.

(c) Test the hypothesis that the cardiac index depends on the weight. Use $\alpha = 0.10$. You must also write the value of the test statistics and the p-value.

(d) What are the assumptions for the above regression analysis? From the Minitab output, explain if all these assumptions are satisfied. If not, then what would you suggest to modify the regression model.

(e) How much variability in the cardiac index is due the patients' weights?

(f) Based on the information presented here, predict the cardiac index for a patient of weight of 80 kg? Is this a reliable prediction? Explain.

(g) Assume that the 95% confidence interval for the mean cardiac index of patients of weight 80 kg is (-0.1, 3.8), and the prediction interval is (-1.1, 5.5). Interpret these intervals.

(c) $H_0: \beta_1 = 0$ TS. $t = \dfrac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \dfrac{0.01256}{0.01023} \simeq 1.23$

$H_a: \beta_1 \neq 0$

$p\text{-value} = 2P(t(df = n-2) > 1.23) = 2(0.1234)$

$= 0.247$ ⇒ Fail to rej H_0

Not enough evidence to support the liner relationship between cardiac index & weight

(d) the error (ϵ_i's) are iid normal (QQ-plot of residuals confirm that)

$Var(\epsilon_i) = \text{constant}$ (residual plot confirms that.)

(f) $y = 0.909 + 0.0126(80)$

it's not a reliable prediction since we fail to reject $H_0: \beta_1 = 0$

(g) - We are 95% confident that on average cardiac index of patients with 80 kg weight is $(-0.1, 3.8)$

- we predict with 95% chance that cardiac index a patient with 80 kg weight is $(-1.1, 5.5)$

(3) Researchers in the development of new treatments for cancer patients often evaluate the effectiveness of new treatments by reporting the proportion of patients who survive for a specific period of time after completion of the treatment. In a recent study on 600 patients for comparing an old treatment with a new treatment, the following data reports the number of patients surviving at least 5 years after the treatment.

	# of Patients	# survived
Old Treatment	450	200
New Treatment	150	74

(a) Find a 95% confidence interval of proportion of patients who will survive under the old treatment. Also find a 95% confidence interval of proportion of patients who will survive under the new treatment. You must check all the necessary assumptions.

(b) Is there a sufficient evidence that the new treatment is better than the old treatment at $\alpha = 0.05$? You must check all the necessary assumptions and state the null and the alternative hypotheses.

ⓐ 95% CI of survival under old treatment.

$$\hat{\pi}_o \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}_o(1-\hat{\pi}_o)}{n_o}} \quad \text{where} \quad \left\{ \begin{array}{l} \hat{\pi}_o = \frac{200}{450} = 0.4\overline{44} \\ n_o = 450 \\ Z_{\alpha/2} = 1.96 \end{array} \right.$$

$$\boxed{(0.399, 0.490)}$$

Assumption

$$\left. \begin{array}{l} n_o\hat{\pi}_o = 450(0.4\bar{4}) > 5 \\ n_o(1-\hat{\pi}_o) = 450(1-0.4\bar{4}) > 5 \end{array} \right\} \checkmark$$

$$0.4\bar{4} \pm 1.96\sqrt{\frac{0.4\bar{4}(1-0.4\bar{4})}{450}} = 0.44\bar{4} \pm 0.046$$

95% CI for survival under new treatment

$$\hat{\pi}_n \pm Z_{\alpha/2}\sqrt{\frac{\hat{\pi}_n(1-\hat{\pi}_n)}{n_n}}$$ wehere $$\begin{cases} \hat{\pi}_n = \frac{74}{150} = 0.49\bar{3} \\ n_n = 150 \\ Z_{\alpha/2} = 1.96 \end{cases}$$

Assumption:

$$\left.\begin{aligned} n\hat{\pi}_n &= 150(0.49\bar{3}) > 5 \\ n(1-\hat{\pi}_n) &= 150(1-0.49\bar{3}) > 5 \end{aligned}\right\} \checkmark$$

95% CI for π_n is $0.49\bar{3} \pm \overset{1.96}{}\sqrt{\frac{0.49\bar{3}(1-0.49\bar{3})}{150}} = 0.49\bar{3} \pm 0.080$

$$\boxed{(0.413,\ 0.573)}$$

(b) $\underline{H_o}: \pi_o = \pi_n$

$\underline{H_a}: \pi_o < \pi_n$

$\underline{\text{Assumptions:}}$ $$\left.\begin{array}{l} n_o\hat{\pi}_o > 5, \quad n_o(1-\hat{\pi}_o) > 5 \\ n_n\hat{\pi}_n > 5, \quad n_n(1-\hat{\pi}_n) > 5 \end{array}\right\} \checkmark$$

$\underline{\text{T.S}}$ $$Z = \frac{\hat{\pi}_o - \hat{\pi}_n}{\sqrt{\frac{\hat{\pi}_o(1-\hat{\pi}_o)}{n_o} + \frac{\hat{\pi}_n(1-\hat{\pi}_n)}{n_n}}} = \frac{0.44\overline{4} - 0.49\overline{3}}{\sqrt{\frac{0.44\overline{4}(1-0.44\overline{4})}{450} + \frac{0.49\overline{3}(1-0.49\overline{3})}{150}}} = -1.039$$

Reject H_o if $\underbrace{Z}_{-1.039} < \underbrace{-Z_\alpha}_{-1.645}$ $\Longrightarrow$ Fail to reject $(H_o: \pi_o = \pi_n)$

NOT suff evidence that the new treatment is better

(4) A local doctor suspects that there is a seasonal trend in the occurrence of the common cold. The doctor collected the following information from a sample of 800 cases of patients with the common cold over the past few years. The data is given below

Season	Frequency
Winter	258
Spring	215
Summer	170
Fall	157

(a) State the null hypothesis to test the seasonal trend on the occurrence of the common cold.

(b) Is there sufficient evidence that there is a seasonal trend in the occurrence of the common cold at $\alpha = 0.01$?

④ ⓐ $H_0: \pi_1 = \pi_2 = \pi_3 = \pi_4 = \frac{1}{4}$

$H_a: \pi_i \neq \frac{1}{4}$ for some $i = 1, 2, 3, 4$

$$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$$

ⓑ $N = 258 + 215 + 170 + 157 = 800$

$\hat{\pi}_1 = \frac{258}{800} = 0.323$ $\quad \hat{\pi}_2 = \frac{215}{800} = 0.269$ $\quad \hat{\pi}_3 = \frac{170}{800} = 0.212$ $\quad \hat{\pi}_4 = \frac{157}{800} = 0.196$

$$\Rightarrow \chi^2 = \frac{(258-200)^2}{200} + \frac{(215-200)^2}{200} + \frac{(170-200)^2}{200} + \frac{(157-200)^2}{200}$$

$= 31.69$

Reject H_0 if $\underbrace{\chi^2}_{31.69} > \underbrace{\chi^2_{0.01}\,(df = 4-1)}_{11.34}$

Suff Evid that, there is a seasonal trend in the occurrence of the common cold.

(5) Four methods of training individuals to use a special piece of equipment have been developed. To compare the methods, 24 employees with similar experience were selected and six assigned to each training method. After the training, the employees were asked to perform a number of tasks and a score of efficiency assigned to each employee. The ANOVA and Means for the results are shown below.

Source	df	MS
Method	3	142.041
Error	20	24.615

Method	A	B	C	D
Mean	30.333	24.167	33.667	23.667
St.Dev.	5.0	6.1	4.0	4.5

(a) Is there sufficient evidence that there is a difference in the four training methods at $\alpha = 0.05$.? State also the null and the alternative hypotheses.

(b) Using a multiple comparisons procedures (LSD or Tukey's), determine the pairs of methods that are significantly different at $\alpha = 0.05$. **Show your work.**

(c) What assumptions did you make for the above analyses?

(a) $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

$H_a: \mu_i \neq \mu_j$ for some (i,j)

Reject H_0 $\Rightarrow$ suff. evid. that there is a diff in the four training methods.

T.S

$$F = \frac{MS_{Trt}}{MS_{Err}} = \frac{SS_{Trt}/df_{Trt}}{SS_{Err}/df_{Err}} = \frac{142.041}{24.615} = 5.77$$

Rej H_0 if

$$\underbrace{F}_{5.77} > \underbrace{F_{0.05}(df_1 = 4-1;\ df_2 = 24-4)}_{F_{0.05}(3, 20) = 3.09}$$

Method	A	B	C	D
Mean	30.333	24.167	33.667	23.667
St.Dev.	5.0	6.1	4.0	4.5

(b) Fisher's LSD : $|\bar{y}_i - \bar{y}_j| > t_{\alpha/2}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$

$$\left.\begin{array}{l} t_{\alpha/2}(df=20) = 2.086 \\ MSE = 24.615 \\ \frac{1}{n_i}+\frac{1}{n_j} = \frac{1}{6}+\frac{1}{6} = \frac{1}{3} \end{array}\right\} \quad t_{\alpha/2}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} = 2.086\sqrt{24.615\left(\frac{1}{3}\right)} = \boxed{5.975}$$

Amethod A vs B : $|30.333 - 24.167| = 6.166 > 5.975$

A vs C : $|30.333 - 33.667| = 3.334 < 5.975$

A vs D : $|30.333 - 23.667| = 6.666 > 5.975$

} Methods B and D are different from method A

B vs C : $|24.167 - 33.667| = 9.5 > 5.975$

B vs D : $|24.167 - 23.667| = 0.5 < 5.975$

} B and C are different

C vs D : $|33.667 - 23.667| = 10 > 5.975$ ⇒ C and D are different

ⓒ - equallity of the variances, (Homogeneity of variances):

$$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$$

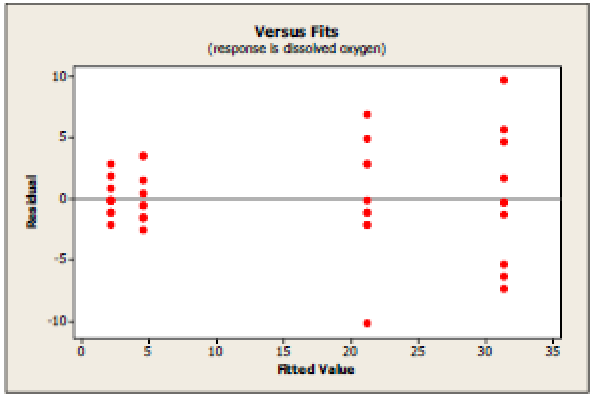
- Data for each method is following a normal distribution.

(6) An ANOVA is to be performed to compare four groups. The means and the standard deviations are as follows:

	Mean	St. Dev
Group 1	2.2	1.476
Group 2	4.6	2.119
Group 3	21.2	4.733
Group 4	31.4	5.522

1. What assumption of ANOVA is not satisfied?

2. What you need to do so that a valid ANOVA can be performed to compare the four groups (You may refer to the graph on the right)?



ⓐ Homogeneity of variances

ⓑ transform the data.