STAT 481 -- Midterm I

Exam Time: 1:00 - 1:50 PM, February 18, 2015

Name: ______________________

UIN: ______________________

CRN Session: Graduate

Score Table

Problems	Score
1 – 28 pt	
2 – 22 pt	
3 – 18 pt	
4 – 32 pt	
Total	

1. [28pt]. An insurance company is reviewing its current policy rates. When originally setting the rates they believed that the average claim amount was $1,800. They are concerned that the true mean is actually higher than this, because they could potentially lose a lot of money. They randomly select 36 claims, and calculate a sample mean of $1,950. Assuming that the standard deviation of claims from a normal distribution is $500, and set significance level α=0.05, test to see if the insurance company should be concerned.

(a). [6pt] Identify the parameter of interest in this statistical study, and set up appropriate null and alternative hypotheses on the parameter accordingly.

Average claim μ

$H_0: \mu = 1800$

$H_1: \mu > 1800$ (right-sided test)

$\sigma = 500, \quad n = 36, \quad \alpha = 0.05, \quad \bar{x} = 1950$

(b). [8pt] Based on the sample statistic, determine the rejection region given level 0.05. What is your conclusion?

$\mathcal{C} = \{ Z > Z_\alpha \}$ sample mean $\bar{X}$

$= \left\{ \frac{\bar{X} - 1800}{500/\sqrt{36}} > 1.645 \right\}$, $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \overset{H_0}{\sim} N(0,1)$

$= \left\{ \bar{X} > 1800 + 1.645 \cdot \frac{500}{\sqrt{36}} \right\}$

$= \{ \bar{X} > 1937 \}$

Sample mean $\bar{x} = 1950$, i.e., $\bar{x} \in \mathcal{C}$

so we conclude that the average claim is higher than 1800.

or

$Z_0 = \frac{\bar{x} - 1800}{500/\sqrt{36}} = \frac{1950 - 1800}{500/\sqrt{36}} = 1.8 > Z_{0.05} = 1.645$

$\Rightarrow$ Reject H_0.

(c). [8pt] Calculate its p-value. Will you reach the same decision as in (b) given the same level 0.05?

$$\text{p-value} = P\{Z > z_0\} \qquad z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

$$= P\{Z > 1.8\} \qquad = \frac{1950 - 1800}{500/\sqrt{36}}$$

$$= 1 - 0.9641 \qquad = 1.8$$

$$= 0.0359$$

$$\text{p-value} < \alpha = 0.05$$

Same conclusion as in (b)

(d). [6pt] Compute the power at μ=1980.

$$P\{\bar{X} > 1937 \mid \mu = 1980\}$$

$$= P\left\{\frac{\bar{X} - 1980}{500/\sqrt{36}} > \frac{1937 - 1980}{500/\sqrt{36}}\right\}$$

$$= P\{Z > -0.516\}$$

$$= P\{Z < 0.516\} = 0.698$$

2. [22pt] Suppose a sample of $n=16$ students were given a diagnostic test before studying a particular module and then again after completing the module. We want to find out if our teaching leads to improvements in students' knowledge/skills (i.e. test scores). Assume that the diagnostic test score is following normal distribution. Let X be the test score before the module, Y be the test score after the module, and their differences be $D_i = Y_i - X_i, i = 1,\ldots,16$.

Sample means and standard deviations of the test scores and the differences are:
$\bar{x} = 18.4, s_x = 3.15, \bar{y} = 20.5, s_y = 4.06, \bar{D} = 2.05, s_D = 2.84$.

(a).[4 pt] State null and alternative hypotheses for this study.

$H_0: \mu_D = 0$ vs $H_1: \mu_D > 0$ paired data

or $\mu_{Y-X} = 0$ $\quad \mu_{Y-X} > 0$

(b). [12pt] What test will you use? Calculate the statistic and its p-value, then draw your conclusion.

paired t-test $\quad n = 16$

$$t_0 = \frac{\bar{D}}{S_D/\sqrt{n}} = \frac{2.05}{2.84/\sqrt{16}} = \frac{2.05}{0.71} = 2.89$$

$$\text{p value} = P\{t(n-1) > 2.89\} \qquad n-1 = 15$$

$0.005 < \text{p-value} < 0.01$, p-value $< \alpha = 0.05$

Reject H_0, i.e, the improvement is significant

(d). [6pt] Find its 95% confidence interval for the mean difference.

$$P\left\{-t_{\alpha/2}(n-1) < \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}} < t_{\alpha/2}(n-1)\right\} = 1 - \alpha$$

$$\therefore P\left\{\mu_D \in \bar{D} \pm t_{\alpha/2}(n-1) \cdot \frac{S_D}{\sqrt{n}}\right\} = 1 - \alpha$$

$$\bar{D} \pm t_{\alpha/2}(n-1) \cdot \frac{S_D}{\sqrt{n}}$$

$$= 2.05 \pm (2.13) \cdot \frac{2.84}{\sqrt{16}} = 2.05 \pm 1.51$$

3. [18 pt] There are three candidates are running for a political position. A pilot study was done by a survey company to investigate if they have the same voting rate. 120 people have been selected at random and asked for their preferences. The data is collected and listed in the table

Candidate	No. 1	No. 2	No. 3	Total
Count	44	36	40	120

(a).[4pt] Specify the parameter of interest, and state the null and alternative hypotheses.

Multinomial distribution $(n, P_{1,0}, P_{2,0}, P_{3,0})$

$P_{i,0}$ is the voting rate

$H_0: P_{1,0} = P_{2,0} = P_{3,0}$ vs $H_1: P_{i,0}$ not the same.

(b).[4pt] What test statistic will you use? Check if required assumptions are met.

Use χ^2-test statistic, assumption

① count follows multinomial distribution ✓

② count number in each cell ≥ 5 ✓

$nP_i = 120 \times \frac{1}{3} = 40$, $i = 1, 2, 3$

(c).[10pt] Compute the observed statistic based on the data, then draw conclusion based on significance level 0.05. $k=3$

$$\chi^2 = \sum_{i=1}^{k} \frac{(y_i - n \cdot P_{i,0})^2}{nP_{i,0}}$$

$P_{1,0} = P_{2,0} = P_{3,0} = \frac{1}{3}$

$nP_{i,0} = 120 \cdot \frac{1}{3} = 40$, $i = 1, 2, 3$

$$= \frac{(44-40)^2}{40} + \frac{(36-40)^2}{40} + \frac{(40-40)^2}{40}$$

$$= \frac{4^2}{40} + \frac{4^2}{40} = 0.8$$

$df = k-1 = 3-1 = 2$ $\qquad \chi^2_{0.05}(2) = 5.99$

$\therefore \chi^2 < \chi^2_{0.05}(2)$ or pvalue > 0.05

$\therefore$ Fail to reject H_0.

4. [32pt]. A study is run to investigate the relationship between the tail lengths (in inches) and the weights (in pounds) of wolves. The idea is predict weight from tail lengths. Here are data information of n=10 wolves .

Tail Length (x)	Weight (y)
10	79
13	72
19	100
19	116
20	85
20	88
23	100
24	80
25	160
27	120

Some useful summaries:

$$\sum_{i=1}^{10} x_i = 200, \sum_{i=1}^{10} y_i = 1000, \sum_{i=1}^{10} x_i^2 = 4250, \sum_{i=1}^{10} y_i^2 = 106{,}250$$

$$S_{xx} = \sum_{i=1}^{10}(x_i - \bar{x})^2 = 250, S_{xy} = \sum_{i=1}^{10}(x_i - \bar{x})(y_i - \bar{y}) = 750, S_{yy} = \sum_{i=1}^{10}(y_i - \bar{y})^2 = 6250$$

(1). [6pt] Simple linear regression method will be used to evaluate the relationship, please write down the regression model and necessary model assumptions.

$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, $i = 1, \dots, n$

x_i: Tail Length $\quad Y_i$: Weight

(2). [6pt] What is the least square criterion for the linear regression model in (1)? What steps need to get the least square estimators? You don't need to solve the equation.

Least Square Criterion

$$\min\{Q(\beta_0, \beta_1)\} = \min\left\{\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2\right\}$$

$$\begin{cases} \partial Q / \partial \beta_0 = 0 \\ \partial Q / \partial \beta_1 = 0 \end{cases} \Rightarrow \begin{cases} 2\sum_i (y_i - \beta_0 - \beta_1 x_i) = 0 \\ 2\sum_{i=1}^{n} x_i \cdot (y_i - \beta_0 - \beta_1 x_i) = 0 \end{cases}$$

(3). [6pt] Calculate their least squares estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$, given data in the table.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{750}{250} = 3$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \left(\frac{1000}{10}\right) - 3 \cdot \left(\frac{200}{10}\right)$$
$$= 100 - 60 = 40$$

(4). [6pt] Complete the ANOVA table based on the summary information.

Source	DF	SS	MS	F
Regression	1	2250	2250	F=4.5
Error	8	4000	500	
Total	9	6250		

(5). [4pt] Check if the data provides sufficient evidence to support a linear relationship given significance level 0.05. State hypotheses first. Do you have another way to evaluate the linear relationship?
[Quantile values: $F_{0.05}(1, 8) = 5.32$, $F_{0.05}(1, 9) = 5.12$, $F_{0.05}(1, 10) = 4.96$.]

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

① $F = 4.5 < F_{0.05}(1, 8) = 5.32$

Fail to reject H_0

② $R^2 = \frac{SSR}{SSTO} = \frac{2250}{6250} = 0.36$, not strong linear relationship

③ $se(\hat{\beta}_1) = \sqrt{\frac{MSE}{S_{xx}}} = \sqrt{\frac{500}{250}} = \sqrt{2} = 1.414$

$\hat{\beta}_1 \pm t_{\alpha/2}(n-2) \cdot se(\hat{\beta}_1) = 3 \pm (2.306) \cdot (1.414) = (-0.26, 6.26)$

$0 \in (-0.26, 6.26) \Rightarrow$ Fail to reject $H_0: \beta_1 = 0$

(6). [4pt]. Show that $E(SSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$, given that $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$.

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad \begin{cases} \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \\ \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \end{cases} \quad \hat{y}_i - \bar{y} = \hat{\beta}_1 (x_i - \bar{x})$$
$$= \sum_{i=1}^{n} \hat{\beta}_1^2 \cdot \left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)$$

$$E\hat{\beta}_1^2 = Var(\hat{\beta}_1) + (E\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} + \beta_1^2$$

$$\therefore E(SSR) = (E\hat{\beta}_1^2) \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$$
$$= \left[\frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} + \beta_1^2\right] \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$$
$$= \sigma^2 + \beta_1^2 \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Brief Formula Sheet

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$\sum_{i=1}^{k} \frac{(y_i - np_{i,0})^2}{np_{i,0}} \qquad \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{(y_{ij} - n\hat{p}_{i.}\hat{p}_{.j})^2}{n\hat{p}_{i.}\hat{p}_{.j}}$$

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \qquad \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$