— Solution —

STAT 481 -- Midterm I

Time: 1:00 - 1:50 PM, 02/19, 2016 **CRN:** 19396 (Graduate)

Name: _____ **UIN:** _____

1. [30pt] The diameters of steel shafts produced by a certain manufacturing process should have a mean diameter of 25.5 units (1 unit = 0.01 inch). The diameter is known to follow normal distribution with ~~standard deviation~~ variance 0.01 unit. A random sample of 100 shafts is chosen to test the claim above, and it has an average diameter of 25.45 units.

$n=10$, $\sigma^2=0.01$, $\bar{x}=25.45$

(a). [4pt] What is the parameter of interest in this study? Set up appropriate hypotheses for the parameter.

Parameter: mean diameter μ (of the steel shafts)

$H_0: \mu = 25.5$ vs $H_1: \mu \neq 25.5$

(b). [4pt] What sample statistic will you use to estimate the parameter in (a)? What is the sampling distribution of the sample statistic under null hypothesis?

Sample mean $\bar{x}$ can be used to estimate μ, population mean of diameter of the steel shafts

σ^2 is know, Sampling distribution of $\bar{x}$: $\dfrac{\bar{x}-\mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$,

X: diameter of steel shaft

$X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, $\sigma^2=0.01$.

(b). [4pt] Based on the data information given above, calculate the test statistic. Determine the rejection region and draw conclusion accordingly. Significance level α = 0.05.

$$Z_0 = \frac{\bar{x}-\mu_0}{\sqrt{\sigma^2/n}} = \frac{25.45-25.50}{\sqrt{0.01^2/10}} = \frac{0.05}{0.1/\sqrt{100}} = \frac{\sqrt{100}}{2} = 5$$

Two-sided test:

$$C = P\{|Z_0| > Z_{\alpha/2}\} = P\{|Z_0| > 1.96\} = \left\{\frac{\bar{x}-25.5}{0.01} > 1.96\right\}$$

$$= \{\bar{x} > 25.52 \text{ or } \bar{x} < 25.48\}$$

$\because Z_0 \in C$ reject H_0

(c). [4pt] Find the p-value for this test.

$$\text{P-value} = 2 * P\{Z > 5\}$$
$$= 2 * 0$$
$$= 0$$

(d). [4pt] Construct a 95% confidence interval on the mean shaft diameter.

95% C.I. for μ: $\bar{x} \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma^2}{n}}$

$$25.45 \pm 1.96 \cdot \sqrt{\frac{0.01}{100}}$$

$$(25.43, 25.47)$$

(e). [5pt] Compute the power at ~~$\mu = 0.255$~~. $\mu = 25.55$

$$\text{power} = P\left(\left| \frac{\bar{x} - 25.5}{0.01} \right| > 1.96 \;\middle|\; \mu = 25.55 \right)$$

$$= P(\bar{x} > 25.52 \text{ or } \bar{x} < 25.48 \mid \mu = 25.55)$$

$$= P\left(Z > \frac{25.52 - 25.55}{0.01}\right) + P\left(Z < \frac{25.48 - 25.55}{0.01}\right)$$

$$= P(Z > -3) + P(Z < -7) \cong 1$$

i.e. $\text{power}(\mu = 25.55) = 1$

2. [20pt] Trace metals in drinking water affect the flavor and an unusually high concentration can pose a health hazard. Ten pairs of data were taken measuring zinc concentration in bottom water and surface water. Does the data suggest that the true average concentration in the bottom water exceeds that of surface water? (Assume normal distribution for the concentration)

Bottom(X)	0.43	0.266	0.567	0.531	0.707	0.716	0.651	0.589	0.469	0.723
Surface(Y)	0.415	0.238	0.39	0.41	0.605	0.609	0.632	0.523	0.411	0.612

Summary Statistics (sample mean and sample standard deviation):
$\overline{X} = 0.565, S_X = 0.147; \overline{Y} = 0.485, S_Y = 0.131; \{D_i = X_i - Y_i, i = 1,...,10\}, S_D = 0.052$

(a).[4pt] State your hypotheses for the parameter of interest.

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

$\mu_D = \mu_{X-Y} = \mu_X - \mu_Y$ is the mean of the difference of the average concentration between the bottom and the surface of the water.

(b). [6pt] Determine the appropriate test statistic and its sampling distribution.

paired t-test

Under $H_0: \mu_D = 0$, $T = \dfrac{\overline{D}}{\sqrt{S_D^2/n}} \sim t(n-1)$, $n = 10$

(b). [5pt] Calculate the observed statistic value.

$$t_0 = \frac{(0.565 - 0.485)}{0.052/\sqrt{10}} = \frac{0.08}{\frac{0.052}{\sqrt{10}}} = 4.865$$

(c). [5pt] Draw your conclusion based on the p-value.

$t_{0.05}(9) = 1.833$ $t_{0.005}(9) = 4.781$

$\therefore t_0 > t_{0.05}(9)$ right-sided test

$\therefore$ p-value < 0.05, Actually p-value < 0.005

$\therefore$ Reject H_0

3. [20 pt] An institute asked 100 people which of four candidates they will vote for.

Tom	Bill	Mary	John	Total
23	20	27	30	100

The claim to test is that the candidates get same votes from the group?

(a) [6pt] State both hypotheses to test the claim. p_i: voting rate for candidate

H_0: $p_i = \frac{1}{4}$, $i = 1, 2, 3, 4$ vs H_1: at least one $p_i \neq \frac{1}{4}$.

[H_1: the voting rates are different among the candidates]

(b) [10pt] Which test statistic will you use for this test? What is your conclusion based on the data given level 0.05? $n = 100$ $np_i = 25$

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(y_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{(y_i - 25)^2}{25}$$

$$= \frac{1}{25}\left[2^2 + 5^2 + 2^2 + 5^2\right] = \frac{58}{25} = 2.32$$

Under H_0: $\chi^2 \sim \chi^2(k-1) = \chi^2(3)$

$\chi^2_{0.05}(3) = 7.81$, $\chi_0^2 < \chi^2_{0.05}(3)$ $\Rightarrow$ Fail to reject H_0

(c) [4pt] What assumptions do you need for the distribution of the data in order to follow its sampling distribution on which you reply to draw your conclusion in (b)?

(1) Data follow multinomial $Y \sim (n, p_1, p_2, p_3, p_4)$

(2) Each cell / category has expected count ≥ 5

4. [30pt]. A shipping company offers customers the opportunity to purchase damage insurance for shipped packages. The shipping company is interested in whether or not a simple linear predictive model for package value can be constructed based on the package weight. Eight packages have been randomly sampled, and their weight (in pound) and value (in dollar) recorded in the following table:

Weight (X)	Value (Y)
51.67	34.22
106.08	78.48
86.28	45.86
93.74	51.03
94.46	46.06
98.96	60.34
74.76	41.06
81.38	47.44

$\bar{x} = 85.92, \bar{y} = 50.56$

$\sum_{i=1}^{8}(x_i - \bar{x})^2 = 2{,}028.90$

$\sum_{i=1}^{8}(x_i - \bar{x})(y_i - \bar{y}) = 1333.78$

$\sum_{i=1}^{8}(y_i - \bar{y})^2 = 1{,}284.72$

(1) [6pt] If we try to fit the data with a linear regression model, what model assumption are needed?

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

Y_i: value, response

x_i: weight, covariate

(2) [8pt] Derive the normal equation for the least square estimates for the intercept and the slope in the linear model in (1). Specify the objective function first. [You don't need to solve the equation.]

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

$$\begin{cases} \frac{\partial Q}{\partial \beta_0} = 0 \\ \frac{\partial Q}{\partial \beta_1} = 0 \end{cases} \Rightarrow \begin{cases} 2\sum_i (y_i - \beta_0 - \beta_1 x_i) = 0 \\ 2\sum_i (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \end{cases}$$

or

$$\begin{cases} n\beta_0 + \sum_i x_i \cdot \beta_1 = \sum_i y_i \\ \beta_0 \cdot \sum_i x_i + \beta_1 \cdot \sum_i x_i^2 = \sum_i x_i y_i \end{cases}$$

Solution of the normal equation

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

(3) [8pt] Calculate the least squares estimates for the intercept and the slope in the regression line based on the data given in the table.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sum_i (x_i-\bar{x})^2} = \frac{1333.78}{2028.9} = 0.6574$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 50.56 - (0.6574)\cdot 85.92 = -5.92$$

(4). [8pt] Show that SSTO = SSR + SSE, i.e. $\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$.
What are the degrees of freedom for these Sum of Squares respectively?

a) $\sum_i (y_i-\bar{y})^2 = \sum_i (y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2$

$= \sum_i (y_i-\hat{y}_i)^2 + \sum_i (\hat{y}_i-\bar{y})^2 + 2\sum_i (y_i-\hat{y}_i)(\hat{y}_i-\bar{y})$

Need to show that $\sum_i (y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) = 0$

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$\therefore \hat{y}_i = (\bar{y} - \hat{\beta}_1\bar{x}) + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1(x_i-\bar{x}), \quad \bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}, \quad \hat{y}_i - \bar{y} = \hat{\beta}_1(x_i-\bar{x})$

$\sum_i (y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) = \sum_i (y_i - \bar{y} - \hat{\beta}_1(x_i-\bar{x}))\cdot \hat{\beta}_1(x_i-\bar{x})$

$= \left[\sum_i (y_i-\bar{y})(x_i-\bar{x}) - \sum_i (x_i-\bar{x})^2 \cdot \hat{\beta}_1\right]\cdot \hat{\beta}_1$

$= [S_{xy} - S_{xx}\cdot \hat{\beta}_1]\cdot \hat{\beta}_1$

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \therefore S_{xy} = S_{xx}\cdot \hat{\beta}_1 \quad \Rightarrow \sum_i (y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) = 0$

b) SSTO = SSR + SSE

DF: $(n-1) = 1 + (n-2)$

$7 = 1 + 6$

$SSTO = S_{yy} = 1284.72$

DF(SSTO) = 7